| le savoir vivant |

# Handbook of Theoretical and Quantitative Geography

UNIL • FGSE Workshop series N° 2
Editors: F. Bavaud & C. Mager

Antunes • Badariotti • Banos • Caglioni • Charlton • Clarke • Dujardin • Foley • Foresti • Fotheringham • Grasland • Hynes • Kaiser • Kanevski • Mélançon • Moreno • Morrissey • Nijkamp • O'Donoghue • Pagliara • Peeters • Pinto • Pozdnoukhov • Rabino • Roca • Rozenblat • Thomas • Timmermans • Timonin • Tuia • Van Leeuwen

Unil
UNIL | Université de Lausanne
Faculté des géosciences
et de l'envir onnement

| le savoir vivant |

# Handbook of Theoretical and Quantitative Geography

UNIL • FGSE Workshop series N° 2
Editors: F. Bavaud & C. Mager

Antunes • Badariotti • Banos • Caglioni • Charlton • Clarke • Dujardin • Foley • Foresti • Fotheringham • Grasland • Hynes • Kaiser • Kanevski • Mélançon • Moreno • Morrissey • Nijkamp • O'Donoghue • Pagliara • Peeters • Pinto • Pozdnoukhov • Rabino • Roca • Rozenblat • Thomas • Timmermans • Timonin • Tuia • Van Leeuwen

UNIL | Université de Lausanne
Faculté des géosciences
et de l'envir onnement

**FGSE – Workshop series N° 2**
Faculty of geosciences and environment
University of Lausanne
Quartier Sorge – Amphipôle
CH- 1015 Lausanne
Switzerland

Handbook of Theoretical and Quantitative Geography
Editors: François Bavaud & Christophe Mager
Publisher: FGSE, University of Lausanne, Lausanne, Switzerland
2009, 457pp
ISBN-13: 978-2-9403-6808-2

**To order:**

University of Lausanne
Institute of Geography
Quartier Dorigny – Anthropole
CH-1015 Lausanne
Switzerland
http://www.unil.ch/igul

## FOREWORD

Modern geography is alive, robust and flourishing. This is what I conclude from the recently published 'Handbook of Theoretical and Quantitative Geography', edited by François Bavaud and Christophe Mager of the University of Lausanne. The book contains twelve substantive papers in 457 pages. I say substantive because they average 38 pages each, with appropriate empirical data, and treat subjects both old and new. The old are topics of long standing interest; the new are à la mode, so to speak. All are based on a conference of the same name held in 2007 in Montreux, Switzerland, on the lovely north shore of the Lake of Geneva. The European Colloquia for Theoretical and Quantitative Geography have now been held for many years, but, in my opinion, are not sufficiently well known outside of Europe. Unexpectedly there is one presentation that cogently argues the merits of qualitative approaches to information.

Papers from scientific conferences are most often a heterogeneous lot and often of widely variable quality. That is not the case here. The materials were obviously carefully selected and clearly warrant the 'Handbook' title. The authors came from Belgium, France, Great Britain, Italy, Ireland, The Netherlands, Portugal, Spain, and Switzerland. One third of the papers actually came from authors residing in two different countries, which speaks well for inter-country collaboration, an objective of the European Union. But who is missing? Well I missed the conference and know that the informal exchanges during free time are an exciting part of such events. Adding to this regret is that it has been over sixty years since I traveled by bicycle through the south facing vineyards of the area with a school friend from Chexbres, with appropriate tastings of course.

Waldo Tobler
Santa Barbara
November 2009

## ACKNOWLEDGMENTS

In 2007, the Institute of Geography of the University of Lausanne (IGUL) organized the Fifteenth European Colloquium for Quantitative and Theoretical Geography (ECTQG07), held for the first time in Switzerland (Montreux). The Colloquium also constituted the annual Workshop of the Faculty of Geosciences and Environment (FGSE), which is situated at the cross-section between natural and social sciences, and focuses on planetary changes of natural or anthropogenic origin, particularly in alpine and metropolitan contexts. The high-quality and variety of the contributions at the Colloquium reflected the broad range of issues and methods investigated in Europe, as well as the vigor and scope of contemporary research in quantitative and theoretical geography. To deepen the questions raised, several European research teams were asked to write articles presenting and evaluating, for various geographical fields, the progress achieved, the methods used, and the scientific prospects. This Handbook reflects their long-term efforts. Six themes were selected, by alphabetical order:

- Cellular automata, multi-agent systems and cooperative phenomena (Banos; Moreno et al.; Pinto et al.)
- Epistemological issues in geography (Grasland; Rabino & Caglioni)
- Geographical flows and networks (Banos; Pagliara & Timmermans; Rozenblat & Mélançon)
- Health geography and epidemiology (Foley et al.; Morrissey et al.)
- Resource management and risk analysis (Kanevski et al.)
- Urban economic geography and spatial economic analysis (Dujardin et al.; van Leeuwen & Nijkamp; Pagliara & Timmermans)

This Handbook shows that the thematic borders within which we wanted to restrict the authors did not systematically make – scientifically – sense, in particular because, by nature, geography, as a scientific object, and the geographers, as researchers, are wanderers. This is why we recommend traveling through this book as on a country lane, being attentive to unexpected approaches, new topics, innovative methods and cross-section fertility. We thank Prof. Waldo Tobler – to whom we owe so much – for his interest and also express our gratitude to the authors for the quality of their contributions. We hope that they will give birth to new vocations.

François Bavaud & Christophe Mager

# SIMULATING PEDESTRIAN BEHAVIOR IN COMPLEX AND DYNAMIC ENVIRONMENTS: AN AGENT BASED PERSPECTIVE

**Arnaud BANOS**

Géographie-Cités, UMR 8504 CNRS/Paris 1, France

## ABSTRACT

Agent-based simulation offers multiple advantages when dealing with complex phenomena like pedestrian movement, characterized by a possibly large number of locally interacting entities. The main goal of this article is to illustrate this key point, through the simulation of pedestrian movements in dynamic environments. The key point we will try to defend here, is that pedestrian movement should not only be considered as a specific phenomenon, but should also be included in a much more global and complex perspective, the urban system as a whole. Pedestrian motion indeed occurs in an ever-changing environment, defined by constraints and opportunities, but also nuisances and dangers.

## KEYWORDS

Agent based simulation, Complex systems, Pedestrian movement, Urban mobility

## INTRODUCTION

Despite the central and fundamental role pedestrian walking plays within the urban transport system, it still remains a poorly known transportation mode. Generally speaking, while most of the developed countries have been developing, for the last 40 years, a wide variety of sophisticated methods and tools aimed at studying urban mobility, only a few of them were really designed to deal with pedestrian movement, especially in interaction with the other transportation modes. Noticing the longstanding tremendous asymmetry in the scientific literature between traffic flow and pedestrian movement, Weifeng and his colleagues (Weifeng, 2003: 634) also regret that "till now, most pedestrian movement models are established based on the rules used for traffic flow and consider little of the special characteristics of pedestrian movement itself".

It is a paradox that walking, the very essence of all movement, is the one mode of transportation about which we know the least. Given its very nature - diffuse, almost isotropic, often short-lived and largely stochastic - pedestrian movement is a difficult field to both observe and model. In recent years, however, it has been receiving greater interest thanks to the emergence of two contingent phenomena. The first is the steady growth in pro-environmental politics, which by bringing the merits of soft modes of transport to the fore has highlighted the lack of knowledge to hand. Elsewhere, research into the phenomenon of self-organization has turned the notion of pedestrian movement into a rich field of research.

Exploring the behavior of pedestrians in interaction with their environment, in virtual cities where most phenomena can be mastered and studied, has thus become a main concern. Generally speaking, this idea of designing “virtual laboratories” (Batty & Torrens, 2001), within which “artificial societies” can be grown (for example Epstein & Axtell, 1995; Batty 2005) has become very popular in recent years and is largely related to two other fields, complex systems and agent based modeling. Moreover, it is firmly embedded in a microscopic approach of urban mobility, where the world is represented as closely as possible in a one-to-one way, which means that “people should be represented as people, cars should be represented as cars and traffic lights should be represented as traffic lights and not as, say, departure rates, traffic streams and capacities respectively” (Nagel et al., 2000: 152). Reaching such a modeling level, without being flooded with microscopic details, thus requires an ad-hoc procedure for designing agent-based computer simulations. The two crucial principles of reductionism and parsimony (Batty, 2005) may therefore constitute main guidelines, in our quest for the identification of the micro-specifications sufficient to generate macrostructures of interest (Epstein, 1999). We will illustrate this point with two different models we contributed to develop in the recent years.

## PEDESTRIAN MOVEMENT MODELLING

Since the first quantitative approaches to pedestrian movement in the 1950s (Mayne, 1954; Schweitzer, 2003), a considerable amount of work has been done in the field, from queuing models (Lovas, 1994) to traffic-like models (Test, 1976; Timms, 1992; Hine 1995; Timms & Cavalho, 1991), with some incursions towards more behavioral ones (Goldsmith, 1977; Griffiths et al., 1984; Hunt & Williams, 1982; Hunt & Griffiths, 1992) and more recently large investments in micro-simulation approaches (Blue & Adler, 1998; Helbing & Monnar, 1997; Helbing et al., 2001; Batty, 2003; Haklay et al., 2001; Kerridge et al., 2001).

The last family is of special interest for our purpose, as it specifically deals with interactions at the pedestrian level, a key point in our project. For example, focusing on the emergence of collective pedestrian behavior from very simple specifications made at the level of individuals, Blue & Adler (1998) managed to reproduce, with a cellular automata, collective characteristics such as formation of pedestrian lanes or speed-flow diagrams, while the social force model proposed by Helbing (Helbing & Monnar, 1997; Helbing et al., 2001) and extended recently by Pelechano et al. (2007) focuses more directly on the influence of the environment and other pedestrians on individual behavior, the whole being formalized as a fluid-dynamic pedestrian model.

While being fundamental steps, these works have nonetheless the drawback of relying on an excessive simplification of the urban environment (a corridor, a place or a room). Motivations and goals of pedestrians are also particularly simplified, reduced to the couple destination to reach / obstacles to avoid. Theses limitations encouraged other researchers to explore agent-based models of pedestrian movement (Batty, 2005; Haklay et al., 2001; Kerridge et al., 2001; Paris, 2007; Paris et al. 2006; Paris et al., 2007; Zachariadis, 2005). In this last family, each pedestrian/agent is defined by a set of capacities and tries to achieve a set of goals, interacting locally with its environment and with other agents.

The diversity and richness of individual capacities mobilized during pedestrian movement have for long limited any attempts at formalization and generalization. Let’s think for example that, nowadays, it is largely accepted that at least four cognitive capacities are required for a single road crossing. First, the capacity to detect traffic, which involves the activation of a strategy of visual search, able to identify and discriminate elements of the environment that are relevant for the current task. Second, the capacity to visually define appropriate timings, that is to compare the current speed of a given vehicle with the expected time needed for crossing. And finally, each pedestrian has to organize, coordinate and update in real-time flows of information coming from different sources, such as traffic on different lanes, traffic signals, other pedestrians…which involves two complementary capacities, short-term memory and the capacity to divide attention in order to handle information coming from multiple sources.

If we recognize the unavoidable heterogeneity of the distribution of these capacities amongst pedestrians, and replace them in their local urban environment, then we have a good idea of the kind of complexity we are dealing with. And it becomes quite evident, in such a perspective, that decomposing this complexity into more simple components, which could in turn be decomposed into smaller and smaller parts, is a dead-end issue. As Gilbert & Troitzsch (1999) recall, even if we were able to know exactly each of the factors implied during individual actions, we still wouldn't be able to predict the resulting behavior of a group composed by the individuals at study, given the tremendous amount and diversity of interactions at work. This strong limitation of the analytical perspective encouraged us to move towards a more appealing paradigm, broadly embedded in the science of complex systems, and methodologically anchored in geosimulation (Benenson & Torrens, 2002).

## AGENT-BASED MODELLING OF COMPLEX SYSTEMS

Defining cities as complex adaptive systems is not a new issue (Krugman, 1996; Portugali, 2000; Pumain et al., 1989). However, the rather recent growth in interest for the science of complex systems now regroups a huge variety of works being most of the time transversal to several disciplines and relying, despite their diversity, on a few basic principles. Generally speaking, in its broader acceptation, a complex system consists of a large number of localized interacting entities, operating within an environment. These entities, being human or not, act and are influenced by the local environment they are situated in. While the behavior of these entities may be inspired, guided or limited by various global trends (think for example of the Highway code), it is usually admitted that they are not directly controlled by upper-level instances but operate on their own, having some "self-control" over their actions and internal states. From that perspective, the study of complex systems requires the development of new scientific tools, nonlinear models, out-of equilibrium descriptions and computer simulation, agent-based modeling being one of these tools. As Doyne Farmer recalls in his foreword of Franck Schweitzer 's book on brownian agents, "the basic idea is to encapsulate the behavior of the interacting units of a complex system in simple programs that constitute self-contained interacting modules" (Schweitzer, 2003: VI). The features' list of agents Schweitzer provides in his book is so complete that we will rapidly recall its main components. Agents can indeed be depicted by their internal structure and their capacities of action. Two main components define the internal agent structure: a) the internal degrees of freedom, defined by the various variables defining the characteristics of each agent; b) autonomy, as agents act in accordance with their internal structure and the environmental conditions they are facing, without being controlled by upper-level instances.

Agent activities can then be described by four features of specific interest for our purpose: a) spatial mobility, as agents are situated in space and time and are able to move, a crucial capacity in everyday urban life; b) reactivity, as agents can perceive their environment and other agents using various kinds of sensors, and are able to respond to it, in accordance with the specification of their internal structure; c) proactivity, as agents can alter, modify their environment or other agents; and d) locality as agents act locally, according to the environment and the other agents they are directly interacting with.

A multi-agent system (MAS) may then consist in a large number of agents, being different or not from the perspective of their internal structure and their capacities of action. Designing and implementing such a system implies that some specific features be respected. First, modularity is fundamental for flexibility: MAS are usually composed of interacting modules, represented as agents themselves. Second, redundancy is a key feature: MAS are indeed composed of a large number of agents, eventually similar in function and design. The main idea is that such systems exploit this redundancy as a factor of robustness: the failure of any given agent can hardly lead to a system breakdown, balanced as it is by the large number of agents performing correctly. The third feature is called decentralization, as there is no (or little) global control in MAS, capacities being distributed across agents. This bottom-up organization of capacities leads to a kind of self-organized control, resulting or emerging from the interactions occurring between agents. Emergent behavior is also an important feature: as agents interact locally, following rules defined according to their internal structure and capacities. Anyway, being observed from an upper-level, groups of agents may exhibit collective behaviors that are not specified at the level of the agent, nor at the level of the group, but that emerge spontaneously. This new quality, therefore, is not deducible from the agents' characteristics, and due to non-linear effects is hardly predictable from the constitutive parts of the MAS. The fifth feature, the functionality of the system, for example in problem solving, is not explicitly defined but rather emerges from the interactions between agents. Adaptation can be seen as the next key characteristic of MAS: modularity, decentralization and emergent functionality are key features which, being combined one with each other, may let MAS adapt to changing situations. And finally, a last feature may be added to Schweitzer's list: Parallelism as, in order to allow agents to act in an autonomous way, a specific parallel architecture is needed, within a single computer and/or between connected computers (grid computing).

From this point, the key idea is to design a MAS allowing to run simulations and to compare results obtained with various sets of parameters, in order to test “what-if” scenarios. “Unfortunately, however, agent-based modelers often lack self-restraint and create agents that are excessively complicated. This results in models whose behavior can be seen as difficult to understand as the systems they are intended to study. One ends up not knowing what properties are generic and which properties are unwanted side-effects” (Farmer in Schweitzer, 2003: VII). It is fairly obvious that models do not have to be as complex as the reality they are designed to depict. But this general statement is of little value if it does not come with a more precise procedure, allowing the modeler to really define what a simple enough model is. Unfortunately, such a procedure does not exist, and should be replaced instead by a couple of principles serving as general guidelines. The first principle is reductionism, as defined in a very pragmatic way by Franck Schweitzer (2003: 5): “instead of incorporating as much detail as possible, we want to consider only as much detail as is necessary to produce a certain emergent behavior”. The second principle, called parsimony – or “Occam's razor” after the mediaeval philosopher William of Occam – suggests us to choose between several possible explanations of a phenomenon the simplest one, the one that requires the fewest leaps of Logic. This very general principle is essential for model building because of what is known as the under-determination of theories by data: “for a given set of observation or data, there is always an infinite number of possible models explaining those same data. This is because a model normally represents an infinite number of possible cases, of which the observed cases are only a finite subset” (Heylighen, 1997). This principle of parsimony is not as trivial as it seems to be. Indeed, it is truly acceptable that any model will always contain more assumptions about the reality than are testable (Batty & Torrens, 2001).

Combined together, reductionism and parsimony lead to a strategy that has been called “generative”, and which seeks to identify the set of rules that are sufficient to generate the structures of interest, that is the minimum set of micro-specifications needed to generate a given macrostructure (Epstein, 1999). Two recent examples developed in collaboration will allow us to illustrate this critical point. The first one focuses on pedestrian movement in a subway station (Banos & Charpentier, 2007), while the second one associates pedestrians and car drivers in a virtual city (Banos et al., 2005, Godara et al., 2007).

## SIMULATING PEDESTRIAN MOVEMENTS IN A SUBWAY STATION

Every subway station may be seen as an open complex system, part of a wider system dedicated to mobility. Simulating pedestrian's behavior in such a peculiar system faces different challenges, related to equipments and location of facilities (ticket dispensers, shops, services), traffic management (speed/density/flow of users) or Emergency procedures (evacuation).

### Design of the model

A map of the station was rasterized and divided into a large number (nearly 7000) of small cells (0.4 m$^2$). Then, a distinction was made between what we called "physical space" and "direction graph". The physical space is defined as the set of cells $E_p = \{C_1, C_2, \ldots C_i, C_n\}$, each cell $C_{i,[1,n]}$ being defined by a set of attributes $V = \{V_1, V_2, \ldots V_i, V_n\}$. Secondly, we sought to enrich the space, by distinguishing specific landmarks and directions. To this end, we took inspiration from classic research demonstrating the important role played by certain nodes and landmarks on individuals' mental representations of their environments (Lynch, 1960). We drew a direction graph $G\theta=(S, \theta)$, with S being the nodes, defined as localised decision points (Timmermans & Djikstra, 1999), and $\theta$ the edges defined as headings between adjacent nodes. A decision point can be seen as a pivot, a location where a choice of heading must be made. This point may correspond, in a given physical space $E_p$, to a landmark (signalling), a particular service (store, ticket office or machine, etc.), or simply to a bifurcation.

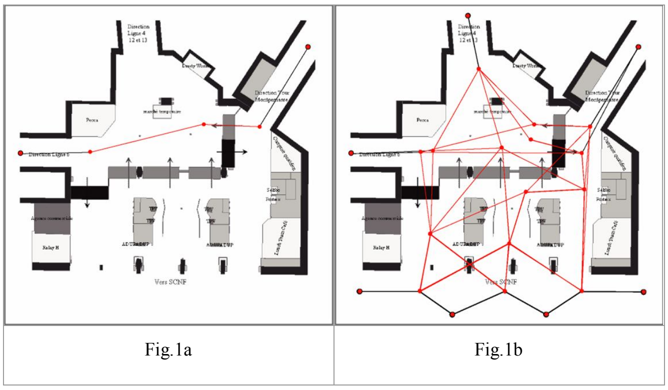


Fig.1a

Fig.1b

Figure 1: Montparnasse station and the direction graph built for a given O/D (Figure 1a) and for every O/D (Figure 1b)

Therefore, the direction graph may be seen as an underlying invisible skeleton, approximating individuals long distance vision field. Figure 1 shows the kind of graph drawn for Montparnasse station in Paris, France. Note that in this case, for simplicity, the nodes of these graphs are solely restricted to bifurcation points. In other words, we assume that agents are unfamiliar with the availability and location of services (ticket offices, ticket machines, shops), they shall identify locally when passing nearby through a vision sensor.

The complexity of this graph depends largely on the assumptions we make on pedestrians. If we accept the idea of an average pedestrian with complete information and boundless rationality (Homo Oeconomicus), then $G\theta^{i}_{(OD)} = G\theta_{(OD)}$ with $G\theta_{(OD)}$ the minimum energy path between a given origin and destination. Each pedestrian may then have the same optimized orientation guide for a given O/D. On the contrary, if we move towards more heterogeneous agents, with limited information and bounded rationality, then $G\theta^{i}_{(OD)} = (S',\theta') \ni S' \subset E_p$ and $\theta' \subset E_p$ . Each agent may then possess his own graph for a given O/D.

**Designing pedestrians**

In MAGE (Modélisation Agent à Grande Echelle / Modeling with Agents at Fine Scales), each pedestrian is represented by an agent $A_i$ , described by a vector of attributes $X(A_i) = \{G\theta^{i}_{(OD)};\delta_i;\phi_i;l^t_i;\theta^t_i;v^t_i\}$ with: $G\theta^{i}_{(OD)}$ his orientation graph for a given origin/destination; $\delta_i$ his vision field; $\phi_i$ his action field; $l^t_i$ his location *(x, y)* at time *t*; $\theta^t_i$ his orientation at time *t*, $0° \leq \theta^t_i \leq 360°$ and $v^t_i$ his speed at time t (Figure 2).

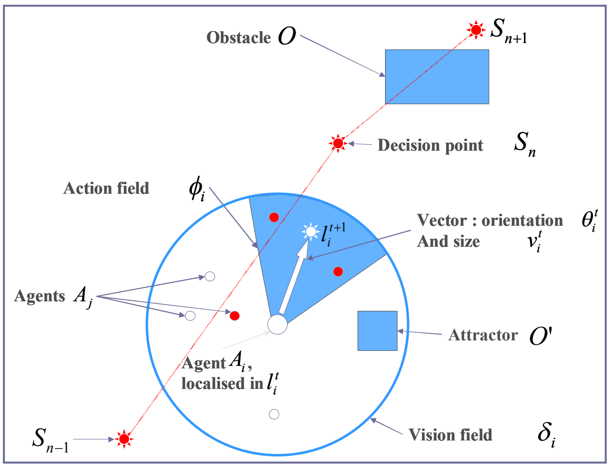


Figure 2: Pedestrian as a brownian agent

Each agent then updates at each time *t* his location $l_i^t \to l_i^{t+1}$, by modifying his orientation and/or his speed, depending on the various constraints and opportunities encountered locally :

$$\theta_i^t = f\left(S_n, O \subset \delta_i, O' \subset \delta_i, A_{i,j\neq i} \subset \delta_i, \varepsilon\right)$$

and

$$v_i^t = f\left(\# A_{i,j\neq i} \subset \phi_i\right)$$

In the absence of interaction with environment and other agents, each agent computes at each time *t* his next location $l_i^{t+1}$ :

$$x_i^{t+1} = x_i^t + \Delta x, \Delta x = v_i^t \times \sin\left(\theta_i^t + \varepsilon\right)$$
$$y_i^{t+1} = y_i^t + \Delta y, \Delta y = v_i^t \times \cos\left(\theta_i^t + \varepsilon\right)$$

with: $\theta_i^t$ the angle location $l_i^t$ of agent $A_i$ and his next decision point $S_n$; $v_i^t$ speed of $A_i$ (m/s); ε an angular value to be chosen by user in the interval [0, 360].

Then, following the set of decision points $S_n \to S_{n+1}$ constituting his orientation graph $G\theta^i_{(OD)}$, each agent may face obstacles and then try to bypass them:

if $l^{t+1}_{i,\theta^t_i,v^t_i} \neq O$

then $l^t_i \rightarrow l^{t+1}_{i,\theta^t_i,v^t_i}$ ,

else $\theta^t_i = \min\left|(\theta^t_i \pm \alpha) - \theta^t_i\right| \ni l^{t+1}_{i,\theta^t_i,v^t_i} \neq O,\ 0 \leq \alpha \leq 180°$

On the contrary, various equipments (ticket dispensers, shops…) may attract an agent with a given probability:

if $O' \subset \delta_i$

and if $\beta_i \leq \mathrm{B}$ , with $\mathrm{B}$ a constant belonging to interval [0,1] and $\beta_i$ a random (uniform) value chosen in the same interval [0,1],

then $O' \rightarrow S_n, S_n \rightarrow S_{n+1}$ and more generally $n \rightarrow n+1$.

Then, the presence of agents in the vicinity of agent *A<sub>i</sub>* may cause interferences. Indeed, if every agent is supposed to avoid any agent on his way, he may change his behavior when the level of pedestrian density increases. In this case, we implemented specific crowd behaviors, inspired from Boids as defined by Reynolds (1987) and extended by Hartman et al., (2006), Shao et al. (2005) and Thalmann et al. (1999). Each agent then adopts a crowd behavior when the density of agents in his neighborhood reaches a given value and when some of these neighbors share the same decision point than agent $A_i$:

$$\left(\# A_{j,j\neq i} \subset \delta_i\right) \geq \lambda \text{ and } \left(\# A_{j,j\neq i} \subset \delta_i \ni S_{nj} = S_{ni}\right) \neq 0$$

In this situation, agent $A_i$ tries to align himself to neighboring agents sharing the same decision point:

$$\theta^t_i = \frac{1}{\# A_{j,S_{nj}=S_{ni}}} \sum_{j=1}^{n} \theta^t_j$$

At the same time, he also tries to move closer to the barycenter $\overline{A}_{j,S_{nj}=S_{ni}}$ of this group:

$$x = \frac{1}{\# A_{j,S_{nj}=S_{ni}}} \sum_{j=1}^{n} x_j \text{ and } y = \frac{1}{\# A_{j,S_{nj}=S_{ni}}} \sum_{j=1}^{n} y_j$$

Finally, each agent modifies his speed according to the number of agents in his action field:

$$v_i^t = f\left(\# A_{i,j\neq i} \subset \phi_i\right)$$

A non-linear model fitted on data collected by Fruin (Fruin, 1987) was then used:

$$v_i^t = \begin{cases} v_{\max} \text{ if } \# A_{i,j\neq i} \subset \phi_i \ = \ 0 \\ \text{-}0{,}376 * S^{-1{,}5} + 1{,}372 \text{ if } \mathrm{S} > 0{,}4\mathrm{m}^2 \\ 0 \text{ if } \mathrm{S} \leq 0{,}4\mathrm{m}^2 \end{cases}$$

with $v_i^t$ the walking speed (m/s) of agent $A_i$ at time *t* and *S* the free surface (m$^2$/pedestrian) available, i.e.

$$\frac{S(\phi_i)}{\# A_{i,j\neq i} \subset \phi_i}$$

**Implementation and first results**

A first version of MAGE was implemented within the simulation platform Netlogo (http://ccl.northwestern.edu/netlogo/). Figures 3a and 3b show snapshots of the kind of environment created.

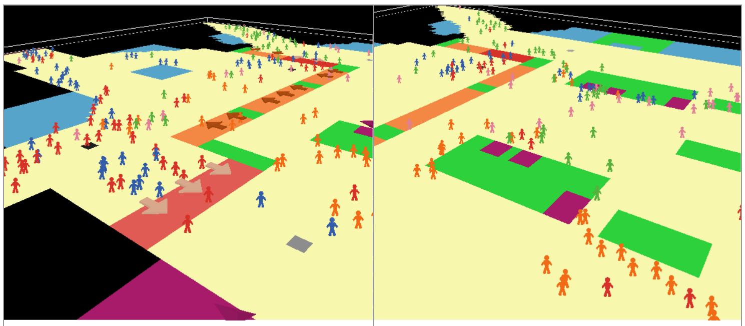


Fig.3a Virtual Montparnasse station

Fig.3b Agents queuing at turnstiles, exit doors and tickets dispensers

Figure 3: An overview of MAGE

The focus may also be put on global structures, through real time display of density maps (Figure 4). It may be noticed that Map 4a (no interactions with local environment and little noise) is very close to the orientation graph in Figure 1b.

The more interactions and noise we add, the more deviations we obtain from this initial skeleton (Map 4c).

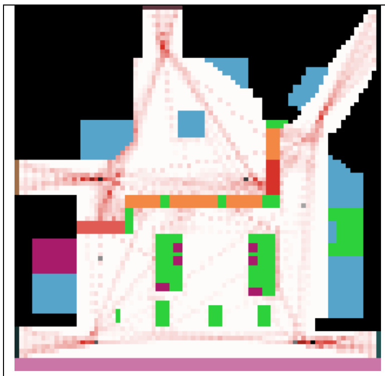
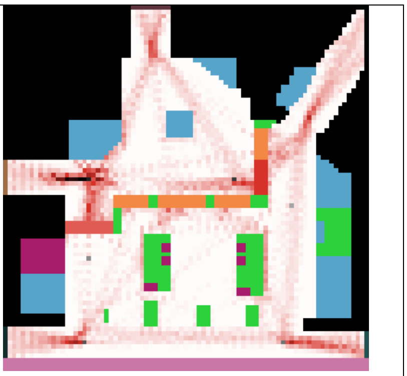
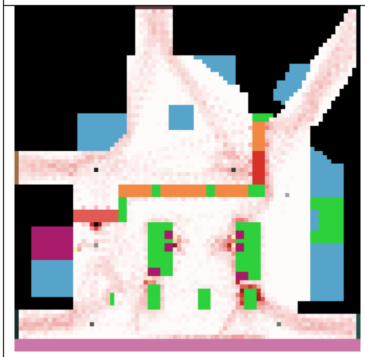
| | |
|---|---|
|  |  |
| Fig. 4a $\varepsilon = 0$, B = 0 | Fig. 4b $\varepsilon = 30$, B = 0 |
|  | Density maps obtained for different values of parameters e (random deviation from straight line) and B (probability of using a given equipment).<br>Each cell $C_{i,[1,n]}$ records at each time t the number of pedestrians passing. This cumulated number is then displayed using a gradation of reds (the darker the higher). |
| Fig. 4c $\varepsilon = 30$, B = 0.5 | |

Figure 4: Revealing macrostructures: density of agents

### Exploring the role of interaction through the fundamental diagram

One of the objectives of the MAGE model is to allow, further ahead, the introduction of an experimental approach to pedestrian behavior in the confines of a subway station. Using individual behavior, the aim is to achieve a clearer understanding of the emergence of observable collective behavior, which is fundamental to the design of public transport areas.

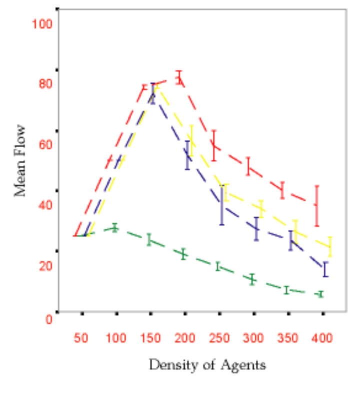


Fig. 5a Flow-Density diagrams in the absence of constraints (no ticket gates / no exit doors)

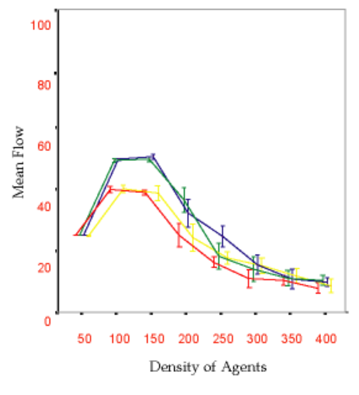


Fig. 5b Flow-Density diagrams in the presence of constraints (ticket gates / exit doors)

| Scenario | Curve color | Avoid other agents | Crowd behaviour |
|---|---|---|---|
| 1 | 🟦 | No | No |
| 2 | 🟩 | Yes | No |
| 3 | 🟨 | No | Yes |
| 4 | 🟥 | Yes | Yes |

Fig. 5c Scenarios tested

Figure 5: Flow-Density diagram showing the impact of local interactions on bi-directional flows ("Density" is the number of agents created, half at each entrance of the path highlighted Figure 1a; "Mean Flow" is the number of agents reaching the opposite exit during a one minute period, averaged over 10 simulations; vertical bars show the variability of this last output variable over the 10 simulations)

We defined an experimental protocol in order to illustrate the relevance of an individual-centered approach based on computer simulation. Within Montparnasse station, a space characterized by a multitude of flows leading to every direction, a benchmark path was selected (see Figure 1.a), to explore the influence of individual interactions on the flow-density diagram in the case of bi-directional flows. The diagram was constructed by way of an exit variable (flow) where values change depending on the values of the entry variable 'density'. This last variable is defined by the number of pedestrians following the itinerary, while flow corresponds to the number of passengers reaching their destination during one minute (Figures 5). These two variables are defined (density) and measured (flow) in a systematic way, for several combinations of behavior and spatial configurations. Agents' ability to avoid collision and adopt crowd behavior was favored. Note also the presence or absence of facilities constricting pedestrian movement (ticket gates and exit doors). Figure 5 highlights the variation of bi-directional flow relative to the density parameter, based on different scenarios.

In both diagrams, curves are characterized by two distinct variation ranges, on both sides by a so-called critical density value, after which pedestrian flow decreases as the number of agents increases. When there are no doors or gates (Figure 5a), the capacity of agents to avoid collision and adopt crowd behavior after certain density values increases pedestrian flow significantly (no overlapping of confidence intervals). Under this configuration, avoidance without crowd behavior has a penalizing effect, as each agent wastes time avoiding all other agents he comes across. Adding doors and gates alters the hierarchy of the lines in a remarkable fashion (Figure 5b).

Under this configuration, the critical density value is systematically replaced by an interval (between 100 and 150 agents), while the maximum flow values obtained are relatively lower than in a situation where there are no facilities to constrain agents. The change in hierarchy of the lines highlights the regulating role that these facilities play in one direction, which automatically encourages a certain level of flow separation.

This first example illustrates how complex environments can be formalized and introduced into an agent based model. This environment, being composed of both static and dynamic components, imposes constraints and provides opportunities to pedestrians, which have to be taken into account. However, what may happen if we want to deal with more active and reactive environments, defined not only by the built infrastructure but also by other categories of agents, having a strong influence on pedestrian behaviors? We will address that specific issue in the next example, focusing on the interactions between cars and pedestrians in a virtual city.

## SIMULATION OF CARS/PEDESTRIANS INTERACTIONS IN A VIRTUAL CITY

The SAMU prototype (Simulation Agents et Modélisation Urbaine) has been designed to explore the behavior of pedestrians in interaction with the motorized traffic (Banos et al., 2005; Godara et al., 2007). It is a hybrid model, combining characteristics of both cellular automata and agent-based models.

### The Basic Principle

Cars and pedestrians are defined as agents, situated on an active grid with which they interact. In order to do so, they perceive their local environment through a sensor (vision), which is of limited range (Figure 6).

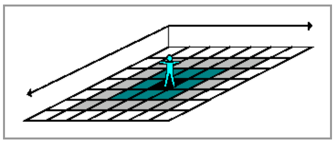
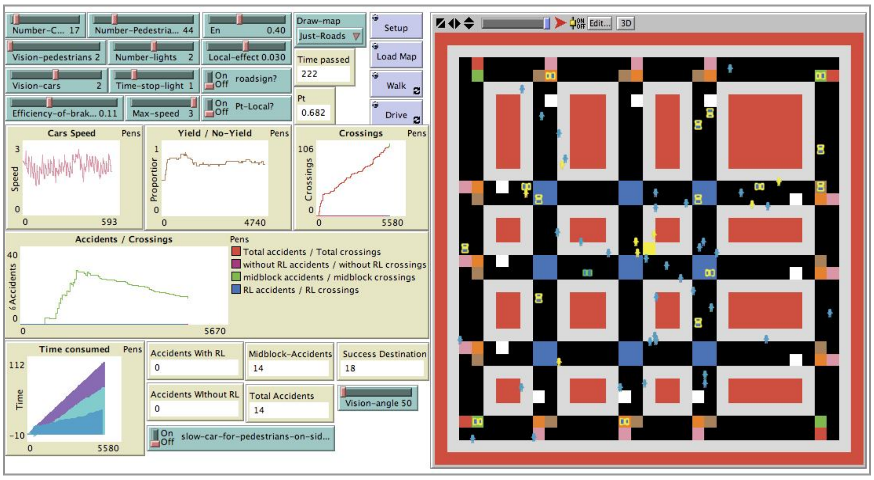


Figure 6: A situated agent, perceiving his local environment (source: internet)

Then, agents have to perform specific tasks, interacting locally with other agents and with their environment. Figure 7 shows the prototype developed in order to observe and test these interactions, as well as emerging parameters, such as speed of cars or proportion of cars/pedestrians collisions.



Number-C... 17
Number-Pedestria... 44
En 0.40
Draw-map: Just-Roads
Vision-pedestrians 2
Number-lights 2
Local-effect 0.030
Time passed: 222
Vision-cars 2
Time-stop-light 1
roadsign?: Off
Pt: 0.682
Efficiency-of-brak... 0.11
Max-speed 3
Pt-Local?: Off
Accidents With RL: 0
Midblock-Accidents: 14
Success Destination: 18
Accidents Without RL: 0
Total Accidents: 14
Vision-angle 50
slow-car-for-pedestrians-on-sid...: Off

Figure 7: The SAMU prototype
(http://arnaudbanos.perso.neuf.fr/geosimul/samu/samu_english.html)

### Model of cars behaviors

In SAMU, we developed a new model of car's behavior by appropriately modifying earlier NaSch/ChSch rules (Nagel & Schreckenberg, 1992) to take into account pedestrian and also turning movements. We also considered two-way traffic with turning movements to bring the model closer to the real world.

Following the prescription of the NaSch model, we allow the speed *V* of each vehicle to take one of the *Vmax + 1* integer values *V = 0, 1, 2….Vmax.* For urban systems we do not want to have *Vmax* more than 70 km/h. So we are taking maximum speed as 3 (22.5 m/s as each cell is 7.5 m in length as in NaSch model). Suppose *Vn* is the speed of the *nth* vehicle at time *t* while moving in any direction (different from NaSch/ChSch/BML model where vehicles move either towards east or towards north and number of cars is fixed on a given road). To emphasize the effect of turning movements and pedestrians we are considering only not signalled intersections in our model and also we want each car to stop at the intersection and decide regarding the turning movements to get homogeneity. The above assumption is true considering the fact that drivers become more cautious and reduce their speed at intersections to avoid any kind of collisions with other vehicles.

At each discrete time step *t* → *t*+1, the arrangement of N vehicles is updated in parallel according to the following *driving rules*:

*Step 1: Acceleration.*

If $Vn < Vmax$, the speed of the nth vehicle is increased by one, i.e., $Vn \rightarrow Vn+1$.

*Step 2: Deceleration (due to other vehicles/pedestrians).*

Suppose *Dn* is the gap in between the nth vehicle and the vehicle in front of it, and *Dpn* is the minimum gap between the car under consideration and the pedestrian in front of it on the road (if any).

```
if min($Dn,Dpn$) <= $Vn$

and

if $Dn < Dpn$,

  then $Vn \rightarrow (Dn - 1)$

  else $Vn \rightarrow [\ \sqrt{}(Vn^2 - (2*g*Dpn*e))\ ]$
```

using $Vf^2=Vo^2+2ad$, with Vf = Final velocity; Vo= Initial velocity; a = Acceleration rate; d = Distance traversed during acceleration. Furthermore, [ ] denotes the integer part and √ denotes the square root, *g* is the gravitational acceleration conventional with standard value of exactly 9.80665 m/s² and *e* is a parameter which includes driver's reaction time and braking efficiency of the car.

The motivation for this choice comes from the fact that we want to have the "physical braking phenomenon" in our model for accidents, i.e. if the gap *Dpn* and braking efficiency is not sufficient for the car to come to a complete stop before hitting the pedestrian then this will result in an accident (which actually happens in real world scenario).

*Step 3: Randomization.*

If $Vn > 0$, the speed of the car under consideration is decreased randomly by unity (i.e., $Vn \rightarrow Vn - 1$) with probability $p$ ($0 \leq p \leq 1$); p the random deceleration probability is identical for all the vehicles, and does not change during updating. Three different behavioral patterns are then embedded in that single computational rule: fluctuations at maximum speed, retarded acceleration and over-reaction at braking

*Step 4: Movement.*

Each vehicle moves forward with the given speed i.e. $Xn \rightarrow Xn + Vn$. Where *Xn* denotes the position of the nth vehicle at any time *t*.

While being very simple, this model allows reproducing some of the basic traffic dynamics. For example, Figure 8 shows the emergence of localised traffic jams when the density of vehicles increases.

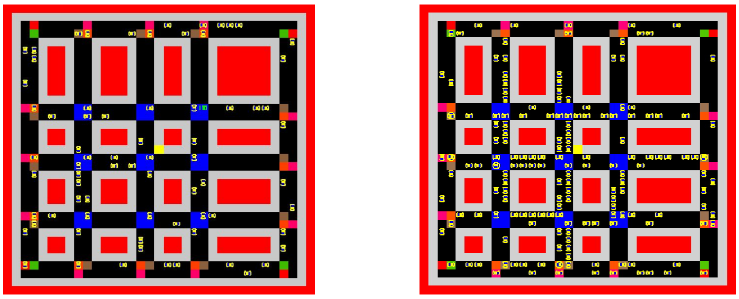


Figure 8: Emergence of localised traffic jams for a given density of vehicles

By simulation, we can explore this phenomenon in an even more detailed way and identify the so-called critical density point. In such a perspective, a closed system was defined (Figure 9) and several simulations were run, in order to explore the variations in traffic flow for different densities of vehicles.

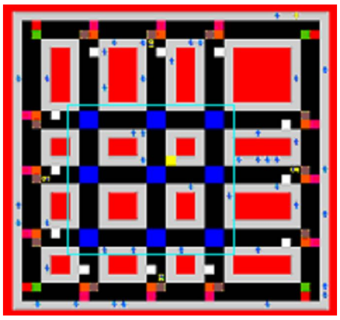
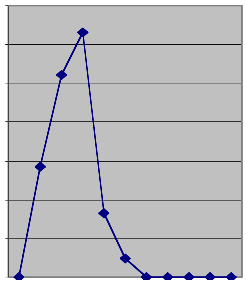
| | |
|---|---|
|  |  |
| *Definition of a closed sub-system* | *Y axis : rate of change of flux (in cars/sec)*<br>*X-axis: vehicular density* |

Figure 9: Exploring by simulation the flow-density relation

Here, the flow is approached by the rate of change of flux (cars/sec.), which is simply the difference of rate of incoming flux (cars entering in system per unit time) and rate of outgoing flux (cars leaving the system per unit time). The simulations for each vehicular density (c = .1, .2,.....1 ) are run for 1800 sec. and finally this rate of change of flux is plotted against varying vehicular density to obtain the fundamental diagram. At sufficiently high densities (c > .5), the dynamical phase of “free-flowing” traffic becomes unstable against the spontaneous formation of jams and the entire traffic system self-organizes so as to reach a completely jammed state.

### 5.3 Model of pedestrian’s behavior

Pedestrians are first located at random then have to reach their destination. A path-finding algorithm combining global and local attraction fields is used. The first one refers to a ground field based on distance to destination (gradient descending), while localized fields are emitted by marked-crosswalks.

On this basis, Godara et al. (2007) introduced modeling crossing behavior as a coordination game, which assumes that the probability of acting in a given manner increases as a function of the proportion of people in proximity to a given agent and acting in that manner (Taillard, 2006). In such a perspective, two possibilities of crossing are defined:

- Yield: before crossing, a pedestrian will stop and check for any moving cars on the street in his/her radius of vision and if the pedestrian detects any moving cars, he will stop and allow the cars to pass;
- No-Yield: the pedestrian will just cross without taking care of cars, i.e. blindly.

Defining the behavior of each pedestrian, for each crossing, is then addressed in a dynamic way.

*Step 1: Decision-making*

In this step a pedestrian will decide to Yield or not (No-Yield). This will depend on the value of a random Bernoulli’s variable (< 1). If the value of this variable is less than *En (Noise in the system)* then $Y = 0$, else $Y = 1$.

If $Y = 1$

Then

If $P_{tl} > .33$, the pedestrian will Yield

If $P_{tl} = .33$, the pedestrian will continue to do whatever he was doing earlier i.e. Yield or No-Yield

If $P_{tl}$ < .33, the pedestrian will not Yield (No-Yield)

Else the pedestrian will choose randomly from the option Yield/No-Yield.

*Step 2: Movement*

The pedestrian will move forward with his walking speed.

This very simple “metamimetic” process (Chavalarias, 2006) allows the proportion of yield/no-yield crossings to evolve in a punctuated equilibrium way (Figure 10).

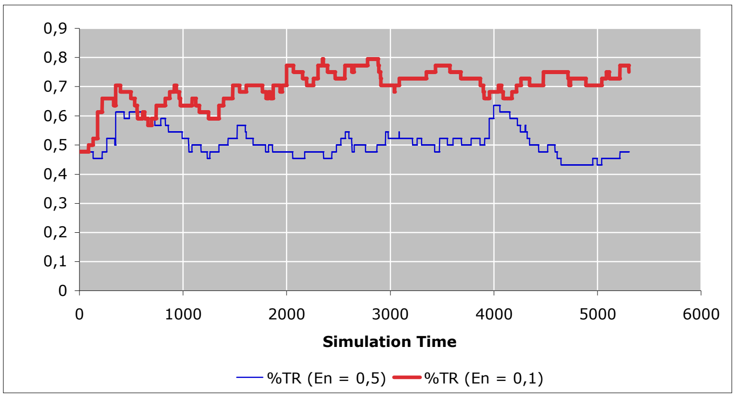


Figure 10: Proportion of yielding pedestrians during a given simulation, for varying Noise levels (En) in the system

Instead of a slow, continuous movement, fluctuations will occur, characterized by long periods of virtual standstill ("equilibrium"), "punctuated" by episodes of very fast development of new forms. Increasing the level of noise in the system reduces its stability (as more randomization takes place) and the proportion of carefully crossing (yielding) pedestrians decreases accordingly, which gives rise to higher number of accidents (Figure 11).

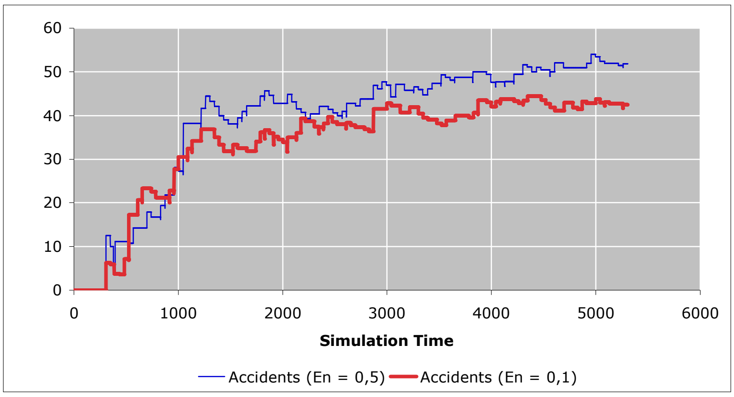


Figure 11: Number of accidents at midblock for varying noise in the system (En)

### Risk modelling

Despite its simplicity, SAMU provides a simulation platform useful to explore interactions between the two types of agents (pedestrians and cars) and possible accidents occurring from such localized interactions. Indeed, agents' movements can be recorded in a very simple way: at each time t, each cell increments its stock values based on the presence of a given entity: car, pedestrian or accident. Figure 12 shows the kind of maps obtained, which are updated dynamically during the simulation.

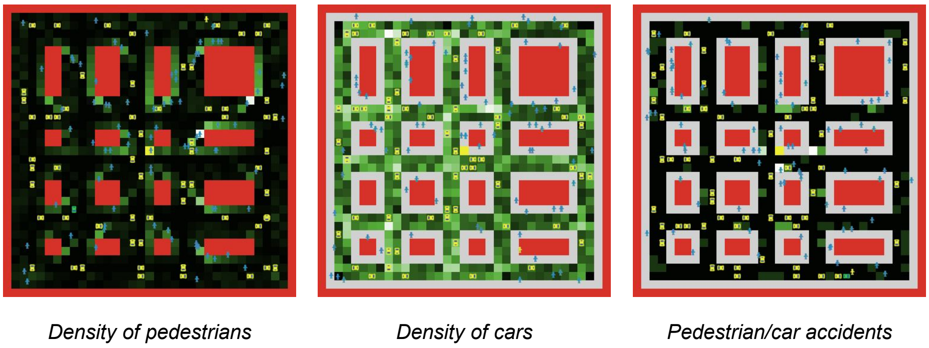

*Density of pedestrians* *Density of cars* *Pedestrian/car accidents*

Figure 12: Dynamic mapping of simulation outputs (pedestrians, cars and accidents densities)

From that point, the idea is to link these three variables, in order to estimate the risk of accident for a given set of parameters. As expected, the risk may be seen as a function of a combination of hazards (cars) and people exposed (pedestrian). Statistical models may then be defined and estimated, as Lassarre & Godara (2007) showed. As an example, Figure 13 shows, for a given simulation and for each cell, the relation between the number of pedestrians recorded and the total number of accidents having occurred.

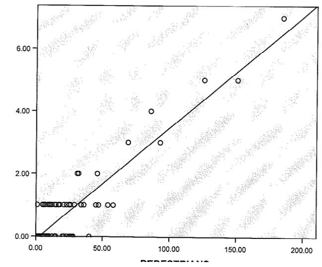


Figure 13: Accident risk modelling (Y-axis: number of accidents per cell / X-axis: number of pedestrians per cell)

On that basis, a Poisson model of the form

$$N = e^{\alpha \ln(P) + \beta \ln(F)}$$

could be defined, with:

- N the number of accidents having occurred on each cell
- P the number of pedestrians having passed on each cell
- F the number of cars which passed on each cell

Fitting by WLS method gave the following values for the two parameters ($\alpha$ = 1.154; $\beta$ = 1.130), for a surprisingly good quality of adjustment ($R^2$ = 0.814) given the simplicity of the model. However, being able to reproduce or at least approach an observed reality may not guarantee that the model grasps some parts of the essence of that reality. Evaluating and validating agent-based models is indeed much wider and much more complex than this apparently simple and evident task (Phan & Amblard, 2007). We believe a more reflexive approach is required, taking inspiration from the work of Ginot (2002), Kleijnen (1987, 2000) or Laperrière (2009) for example and going beyond traditional objectives of finding recognizable or expected structures.

This implies that we have to tackle the issues of uncertainty and surprise when simulating complex systems (McDaniel, 2005), these being prerequisites for the distribution and practical application of these models.

## CONCLUSION

These two examples, aimed at modeling and simulating pedestrian behavior in complex and dynamic environments, allow to test “what if scenarios” and to explore some fundamental aspects of complex spatial systems. However, reaching such a modeling level - even still simplified - without being flooded with microscopic details, is not that easy. As Leombruni & Richiardi (2005) see it, the complexity of agent-based modeling is also its strength, implying that we should accept this state of affairs and work towards setting up innovative protocols for simulation and validation. The principle of parsimony, suggests that we should constantly question the relationship between complexity and efficiency (Batty & Torrens, 2005). Would simpler models with no or fewer interactions enable similar results to be produced? Our initial investigations demonstrate the major role played by individual interactions and their influence on the overall behavior of the system. However, a certain amount of fine-tuning is needed before this type of application can really be generalized, in order to include for example: a wider variety of individual behaviors (heterogeneous agents), more complex spaces and finally, larger populations in the simulations. Determining values for these parameters and their combinations poses a formidable problem, which is aggravated by their already high number at this stage of the model’s development.

## REFERENCES

**Anderson, P.**, 1972, More is different. In: Science, 177: 393-396

**Banos, A.; Charpentier, A.**, 2007, Simulating pedestrian behavior in subway stations with agents. In: Proceedings of the 4th European Social Simulation Association. Toulouse, France, 10-14 September: 611-621

**Banos, A.; Godara, A.; Lassarre, S.**, 2005, SAMU: contribution to the study or urban "anthill". In: Proceedings of the Second Indian International Conference on Artificial Intelligence. Pune, India, 20-22 December: 2876-2888

**Banos, A.; Godara, A.; Lassarre, S.**, 2005, Simulating pedestrians and cars behaviors in a virtual city: an agent-based approach. In: Proceedings of the European Conference on Complex Systems. Paris, 14-18 November. http://arnaudbanos.perso.neuf.fr/papers/eccso5.pdf

**Batty, M.**, 2005, Cities and complexity. MIT Press: Cambridge

**Batty, M.; Torrens, P.**, 2005, Modelling and prediction in a complex world. In: Futures, 37: 745-766

**Benenson, I.; Torrens, P.**, 2002, Geosimulation: Automata-based modelling of urban phenomena. Wiley: Chichester

**Berrou, J.L.; Beecham, J.; Quaglia, P.; Kagarlis, M.A.; Gerodimos, A.**, 2007, Calibration and validation of the Legion simulation model using empirical data. In: Waldau, N.; Gattermann, P.; Knoflacher, H.; Schreckenberg, M. (Eds), Pedestrian and Evacuation Dynamics. New York: Springer Verlag: 155-166

**Blue, V.; Adler, J.**, 1998, Emergent fundamental pedestrian flows from cellular automata microsimulation. In: Transportation Research Record, 1644: 29-36

**Chavalarias, D.**, 2006, Metamimetic Games: Modeling Metadynamics in Social Cognition. In: Journal of Artificial Societies and Social Simulation 9, 2, 5. http://jasss.soc.surrey.ac.uk/9/2/5.html

**Cunningham, P.; Cullen, D.**, 1993, Pedestrian flow data collection and analysis. In: Proceedings of the Institution of Civil Engineers, Transport, 100:59-69

**Daamen, W.; Hoogendoorn, S.; Bovy, P.**, 2005, First order pedestrian traffic flow theory. In: Transportation research board annual meeting 2005, Washington: 1-14

**Daamen, W.; Hoogendoorn, S.**, 2004, Pedestrian traffic flow operations on a platform: observations and comparison with tool SimPed, Congress Proceedings of CompRail. Dresden, Germany, May: 125-134

**Daamen, W.**, 2002, SimPed: A pedestrian simulation tool for Large Pedestrian Area, Conference proceedings EuroSIW, 24-26 june

**Daly, P-N.; McGrath, F.; Annesley, T-J.**, 1991, Pedestrian speed/flow relationships for underground stations. In: Traffic engineering and control: 75-77

**Djikstra, J.; Timmermans, H.** 1999, Towards a multi-agent model for visualizing simulated user behavior to support the assessment of design performance. In: Ataman, O.; Bermúdez, J. (Eds), ACADIA99 – Media and design process. Acadia: 226-237

**Epstein, J.**, 1999, Agent-based computational models and generative social science. In: Complexity, 4, 5: 41-60

**Epstein, J.; Axtell, R.**, 1996, Growing artificial societies: social science from the bottom up. Brookings Institution Press, MIT Press: Washington DC

**Feurtey, F.**, 2000, Simulating the collision avoidance behavior of pedestrians, Master Thesis. University of Tokyo, School of Engineering. Directed by Takashi Chikayama. http://www.logos.t.u-tokyo.ac.jp/~franck/files/thesis.pdf

**Fruin, J.**, 1987, Pedestrian planning and design. Elevator world's publication : New York

**Gimblett, R.**, 2002, Integrating GIS and agent-based modeling techniques for simulating social and ecological processes. Oxford University Press: London

**Ginot, V.; Le Page, Ch.; Souissi, S.**, 2002, A multi-agents architecture to enhance end-user individual based modelling. In: Ecological Modelling, 157: 23-41

**Godara, A.; Lassarre, S.; Banos, A.**, 2007, Simulating Pedestrian-Vehicle Interaction in an Urban Network Using Cellular Automata and Multi-Agent Models. In: Schadschneider, A. et al. (Ed.), Traffic and granular flow'05. Berlin: Springer: 411-418

**Hacklay, M.; Thurstain-Goodwin, M.; O'Sullivan, D.; Schelhorn, T.**, 2001, So go downtown: simulating pedestrian movement in town centers. In: Environment and Planning B, 28: 343-359

**Hankin, B-D.; Wright, R-A.**, 1958, Passenger flow in subways. In: Operational Research Quaterly, 9, 2, 1958: 81-88

**Hartman, C.; Benes, B.**, 2006, Autonomous boids. In: Computer Animation and Virtual Worlds, 17: 199-206

**Helbing, D.; Johansson, A.**, 2007, Quantitative agent-based modelling of human interactions in space and time. In: Proceedings of the 4th European Social Simulation Association, 10-14 September. Toulouse, France: 623-637

**Helbing, D.; Molnar, P.; Farkas, I.; Bolay, K.**, 2001, Self-organizing pedestrian movement. In: Environment and Planning B, 28: 361-383

**Helbing, D.; Buzna, L.; Werner, T.**, 2003, Self-organised pedestrian crowd dynamics and design solutions. In: Transportation Science: 1-24

**Heylighen, F.**, 1997, Occam's Razor. In: Principia Cybernetica Web. http://pespmc1.vub.ac.be/occamraz.html

**Jiang, B.**, 1999, SimPed: Simulating pedestrian flows in a virtual urban environment. In: Journal of Geographic Information and Decision Analysis, 3, 1: 21-30

**Kerridge, J.; Hine, J.; Wigan, M.**, 2001, Agent-based modelling of pedestrian movements: the questions that need to be asked and answered. In: Environment and Planning B, 28: 327-341

**Kleijnen, J.P.C.**, 1987, Statistical Tools for Simulation Practitioners. Marcel Dekker, Inc.: New York

**Krugman, P.**, 1996, The self-organizing economy. Blackwell Publishers: New York

**Laperrière, V.; Badariotti, D.; Banos, A.; Muller, J.P.**, 2009, Structural validation of an individual-based model for plague epidemics simulation. In: Ecological Complexity, 6, 2: 102-112

**Lassarre, S.; Godara, A.**, 2007, Pedestrian Accident Risk Assessment in Urban Systems Using Virtual Laboratory SAMU. In: 11th World Conference on Transport Research in Berkeley, 23-28

**Leombruni, R.; Richiardi, M.**, 2005, Why are economist sceptical about agent-based simulation ? In: Physica A, 355: 103-109

**Lovas, G.**, 19954, Modeling and simulation of pedestrian traffic flow. In: Transportation Research B, 28, 6: 429-443

**Lynch, K.**, 1960, The image of the city. MIT Press: Boston

**Manheim, M.**, 1979, Fundamentals of transportation systems analysis. Volume 1: basic concepts. The MIT Press: Cambridge

**McDaniel, R.; Driebe, J.**, 2005, Uncertainty and surprise in complex systems: questions on working with the unexpected. Springer Verlag: New York

**Nagel, K.; Esser, J.; Rickert, M.**, 2000, Large-scale traffic simulations for transportation planning. In: Annual Reviews of Computational Physics, VII: 151-202

**Nagel, K.; Schreckenberg, M.**, 1992, Cellular automaton models for freeway traffic, Physics. I-2: 2221-2229

**Paris, S.**, 2007, Caractérisation des niveaux de services et modélisation des circulations de personnes dans les lieux d'échanges, PhD Thesis, Université de Rennes 1, France

**Paris, S.; Donikian, S.; Bonvalet, N.**, 2006, Environmental abstraction and path planning techniques for realistic crowd simulation. In: Computer Animation and Virtual Worlds, 17: 325-335

**Paris, S.; Pettré, J. S.; Donikian, S.**, 2007, Pedestrian reactive navigation for crowd simulation: a predictive approach. In: Eurographics 2007, 26, 3: 665-674

**Pelechano, N.; Allbeck, J.M.; Badler, N.I.**, 2007, Controlling individual agents in high-density crowd simulation, Eurographics 2007, ACM SIGGRAPH Symposium on Computer Animation

**Pettré, J.; Heras Ciechomski, P.; Maïm, J.; Yersin, B.; Laumond, J.P.; Thalmann, D.**, 2006, Real-time navigating crowds: scalable simulation and rendering. In: Computer Animation and Virtual Worlds, 17: 445-455

**Phan, D.; Amblard, F.**, 2007, Agent-based modelling and simulation in the social and human sciences. The Bardwell Press: Oxford

**Portugali, J.**, 2000, Self-organization and the city. Springer-Verlag: New York

**Pumain, D.; Sanders, L.; Saint-Julien, T.**, 1989, Villes et auto-organisation. Economica: Paris

**Reynolds, C.**, 1987, Flocks, Herds, and Schools: A Distributed Behavioral Model. In: Computer Graphics, 21, 4: 25-34

**Schweitzer, F.**, 2003, Brownian Agents and active particles. Springer-Verlag: New York

**Shao, W.; Terzopoulos, D.**, 2005, Autonomous pedestrians. In: Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation: 19-28

**T.R.B.,** 2000, Highway capacity manual, USA, Transportation Research Board

**Taillard, P.,** 2006, Quel degré de rationalité favorise la coordination des agents ? In: Revue d'Economie Politique, 116, 1: 23-42

**Thalmann, D.; Musse, S.R.; Kallmann, M.**, 1999, Virtual Humans' behavior: individuals, groups and crowds. In: Proceedings of Digital Media Futures: 13-15

**Weifeng, F.; Lizhong, Y.; Weicheng, F.**, 2003, Simulation of bi-direction pedestrian movement using a cellular automata model. In: Physica A, 321: 633-640

**Young, H.P.**, 1996, The Economics of Conventions. In: Journal of Economic Perspectives, 10, 2: 105-122

**Zachariadis, V.**, 2005, An agent-based approach to the simulation of pedestrian movement and factors that control it. In: CUPUM'05: Computers in Urban Planning and Urban Management, University College of London. 29 June-1 July. http://128.40.111.250/cupum/searchpapers/index.asp

## AUTHOR INFORMATION

**Arnaud BANOS**
arnaud.banos@parisgeo.cnrs.fr
Géographie-Cités
UMR 8504 CNRS/Paris 1, France

# NEIGHBOURHOOD EFFECTS AND ENDOGENEITY ISSUES

**Claire DUJARDIN, Dominique PEETERS and Isabelle THOMAS**

Université catholique de Louvain, CORE and Department of Geography, Belgium

## ABSTRACT

A recent body of research in the fields of geography, economics and sociology suggests that the spatial structure of cities might influence the socioeconomic characteristics and outcomes of their residents. In particular, the literature on neighbourhood effects emphasises the potential influence of the socioeconomic composition of neighbourhoods in shaping individual's behaviours and outcomes, through social networks, peer influences or socialisation effects. However, empirical work has not reached yet a consensus regarding the existence and magnitude of such effects. This is mainly because the study of neighbourhood effects raises important methodological concerns that have not often been taken into account. Notably, as individuals with similar socio-economic characteristics tend to sort themselves into certain parts of the city, the estimation of neighbourhood effects raises the issue of location-choice endogeneity. Indeed, it is difficult to distinguish between neighbourhood effects and correlated effects, i.e. similarities in behaviours and outcomes arising from individuals having similar characteristics. This problem, if not adequately corrected for, may yield biased results.

In the first part of this chapter, neighbourhood effects are defined and some methodological problems involved in measuring such effects identified. Particular attention is paid to the endogeneity issue, giving a formal definition of the problem and reviewing the main methods that have been used in the literature to try to solve it. The second part is devoted to an empirical illustration of the study of neighbourhood effects, in the case of labour-market outcomes of young adults in Brussels. To this end, the effect of living in a deprived neighbourhood on the unemployment probability of young adults residing in Brussels is estimated using logistic regressions. The endogeneity of neighbourhood is addressed by restricting the sample to young adults residing with their parents. However, this method is an imperfect solution. Therefore, a sensitivity analysis is used to assess the robustness of the results to the presence of both observed and unobserved parental covariates. Results show that living in a deprived neighbourhood significantly increases the unemployment probability of young adults. This result is not sensitive to the presence of observed and unobserved parental characteristics.

## KEYWORDS

Neighbourhood effects, Endogeneity, Self-selection, Sensitivity analysis, Brussels

## INTRODUCTION

The last twenty-five years have seen a rising interest among economists, social scientists and geographers in the study of the way in which neighbourhood context may affect individual behaviours and outcomes. The work of sociologist W.J. Wilson has been particularly influential in suggesting that social influences in the neighbourhood, i.e. neighbourhood effects, could be part of the explanation for the social problems experienced by poor inner-city residents in American metropolitan areas.

> *...in a neighbourhood with a paucity of regularly employed families and with the overwhelming majority of families having spells of long term joblessness, people experience a social isolation that excludes them from the job network system that permeates other neighbourhoods and that is so important in learning about or being recommended for jobs... In such neighbourhoods, the chances are overwhelming that children will seldom interact on a sustained basis with people who are employed or with families that have a steady breadwinner. The net effect is that joblessness, as a way of live, takes on a different meaning...*
>
> Wilson, *The Truly Disadvantaged* (1987: 57)

Studies aiming at estimating neighbourhood effects are numerous and have been the focus of several extensive surveys (see, among others, Jencks & Mayer, 1990; Ellen & Turner, 1997; Sampson et al., 2002; Dietz, 2002; Durlauf, 2004). These studies focus on a wide variety of outcomes such as teenagers' educational attainment and school attendance, delinquency, drug consumption, out-of-wedlock pregnancy, as well as labour-market outcomes (unemployment and welfare participation) and health status (for example, well-being and mental health).

Despite the bulk of empirical studies, there is still considerable debate over the existence and magnitude of neighbourhood effects. Some authors consider that neighbourhoods play a significant role in explaining individual outcomes while others are convinced that their role – if any – is of marginal importance compared to that of personal characteristics and backgrounds. For example, Jencks & Mayer (1990: 176) conclude their survey on the role of neighbourhood effects in shaping children and adolescents' behaviours by saying that "the literature we reviewed does not […] warrant any strong generalizations about neighbourhood effects".

According to several authors (Ginther et al., 2000; Dietz, 2002), the primary reason for this lack of consensus is the great diversity regarding the methods, data and variables used to test for the existence of neighbourhood effects. Moreover, the estimation of neighbourhood effects is a difficult task. Recent studies have highlighted the existence of important methodological problems inherent to the estimation of such effects, which have rarely been accounted for in most earlier empirical works (Manski, 1993; Durlauf, 2004; Blume & Durlauf, 2006). Of paramount importance is the endogeneity bias that results from the self-selection of individuals into neighbourhoods (Blume & Durlauf, 2006). Indeed, trying to explain individual outcomes by neighbourhood characteristics in a simple regression analysis does not lead to concluding results as residential locations are not exogenously determined. On the contrary, individuals having similar characteristics tend to sort themselves in some parts of the urban space. There is thus a two-way causality: on the one hand, residential location influences individual socioeconomic outcomes, and on the other hand, individual outcomes influence the choice of a residential location. Standard econometric methods are unable to distinguish between two-way causality and may consequently yield biased results.

This chapter seeks to contribute to the field by defining neighbourhood effects and giving an overview of existing solutions to overcome endogeneity problems. The first section gives a conceptual definition of what lies behind the generic term of “neighbourhood effects” in order to precisely define what this chapter is concerned with. The second section briefly lists some of the more recurrent methodological problems that has encountered most empirical work on neighbourhood effects. The third section then focuses on the most prominent problem in this field: the endogeneity associated with residential location. It provides a precise formulation of the endogeneity issue and critically reviews the various methods that have been proposed to solve it. The rest of the chapter is devoted to an empirical application of the study of neighbourhood effects. Instead of trying to solve the endogeneity issue, the importance of endogeneity biases is assessed through a sensitivity analysis, in order to evaluate the robustness of the results. This method is illustrated with estimates of neighbourhood effects on labour-market outcomes of young adults residing in Brussels.

## NEIGHBOURHOOD EFFECTS: DEFINITION AND IDENTIFICATION ISSUES

There are numerous reasons why residential location should matter in explaining individual behaviours and outcomes (see Ellen & Turner, 1997, Jencks & Mayer, 1990 for extensive surveys). For example, *peer influences* refer to a contagion effect in which the propensity to adopt a socially-deviant behaviour (like dropping out of school, consuming drugs or being unemployed) depends critically on the proportion of peers exhibiting the same behaviour in a community (this is known as the “epidemic theory of ghettos” developed by Crane, 1991). *Socialisation* or *role model effects* refer to the influence on the behaviour of teenagers of the socioeconomic success of adults in their neighbourhood of residence, those serving as models to which young people can identify and for what they may aspire to become (Wilson, 1987). The density and composition of *social networks* inside the residential neighbourhood may also influence individual outcomes by conditioning the quantity and quality of information available to individuals regarding access to social services and economic opportunities (Reingold, 1999). Finally, the *physical disconnection* and *isolation* may influence accessibility to services or economic opportunities (see for example the spatial mismatch hypothesis in which distance to job locations explains the high unemployment rates of American inner-city black residents; Ihlanfeldt & Sjoquist, 1998). These few mechanisms all point to a more general concept, which has generically been labelled “neighbourhood effects” or “contextual effects”. However, they cover very different types of residential location influences. It is thus helpful to give some conceptual definitions and to identify what kind of effects will be studied in this chapter.

### Spatial versus neighbourhood effects

First, the effects of residential location can be decomposed into pure spatial effects and neighbourhood effects. *Spatial effects* are pure locational effects and refer to the influence on individual outcomes of residing in a particular location in the city, which may give some advantages and disadvantages in terms of accessibility to economic and social opportunities located in the metropolitan area. Clearly, the spatial mismatch hypothesis which postulates that the spatial disconnection between residence and job locations might negatively affect individual labour-market outcomes is an example of such a spatial effect (Ihlanfeldt & Sjoquist, 1998). On the contrary, *neighbourhood effects* refer to the effects of belonging to a group on the behaviours and socioeconomic outcomes of individuals (Dietz, 2002), independently of the geographical location of the group within the city. Neighbourhood effects refer in general to the effects of belonging to a group defined on a geographical basis. In this case, neighbourhood effects concern influences on an individual's behaviour and outcomes due to the characteristics and behaviours of his/her neighbours.

Although spatial effects and neighbourhood effects are not mutually exclusive, they have often been considered separately. Therefore and for the sake of simplicity, pure spatial effects will be left aside from now and the remainder of this chapter will consider only neighbourhood effects. However, most of the methodological problems that will be highlighted in the next sections also apply to pure spatial effects.

### Endogenous versus contextual effects

In a widely cited article, Manski (1993) identifies two types of neighbourhood effects (or social interactions, following its own terminology): *endogenous effects*, whereby the propensity of an individual to behave in some way varies with the average behaviour of the group, and *contextual effects* whereby the propensity of an individual to behave in some way varies with the average background characteristics of the group. The distinction between endogenous and contextual effects is subtle as their definitions differ only by one word: a person's behaviour (or choice) is influenced either by the *behaviour* (endogenous effect) or by the *background characteristics* (contextual effect) of those in his/her group. To clarify, consider the example of school achievement. There is an endogenous effect if a teenager's achievement is influenced by the average achievement of his/her peers (i.e. other students in his/her school or neighbourhood). There is a contextual effect if his achievement is influenced by the socioeconomic composition of his neighbourhood, for example by the employment status observed among adults in the neighbourhood, which may provide role models and affect teenagers' aspirations. The relevance of this distinction lies in the fact that endogenous effects generate a social multiplier: acting on an individual's behaviour does not only influence this sole individual, but also exerts an influence on the aggregate behaviour of the group. For example, if a teenager's school achievement increases with the average achievement of the students in the school, then tutoring some students in the school does not only help the tutored students but indirectly helps all students in the school. Contextual effects do not generate such social multipliers (Manski, 1993).

Endogenous and contextual effects may be formalised using the following equation (adapted from Manski, 1993):

$$y_i = \alpha + \beta' X_i + \theta y^e{}_{n(i)} + \eta' X^e{}_{n(i)} + \varepsilon_i \qquad (1)$$

where $y_i$ is the outcome of interest of individual *i* (for example, school achievement), $X_i$ is a vector of individual characteristics (gender, age, family characteristics), $y^e{}_{n(i)}$ is the mean outcome of those individuals in the neighbourhood *n* in which individual *i* resides (mean educational attainment of peers), $X^e{}_{n(i)}$ is a vector of variables describing neighbourhood composition (including mean values of individual characteristics $X_i$ such as mean parental outcome), and $\varepsilon_i$ is a classical error term. The parameter $\theta$, when significantly different from zero, measures an endogenous effect, while $\eta$ measures a contextual effect.

### Neighbourhood versus correlated effects

Besides endogenous and contextual effects (which taken together will be labelled neighbourhood effects), Manski (1993) identifies a third explanation for the fact that individuals belonging to the same group have similar outcomes: *correlated effects*. In correlated effects, individuals belonging to the same group (in our case to the same neighbourhood) exhibit a similar behaviour simply because they have similar unobserved individual characteristics, and not because of the influence of the group's behaviour and composition. Taking the example of labour-market outcomes, individuals residing in the same neighbourhood may have high unemployment probabilities because living in a neighbourhood with a high proportion of unemployed or low-educated workers generates pervasive effects such as peer effects or poor social networks (a neighbourhood effect). On the contrary, this might simply reflect the fact that individuals living in the same neighbourhood share common unobserved factors that are detrimental in finding a job, such as low ability (a correlated effect).

The following equation amends equation 1 in order to reflect correlated effects:

$$y_i = \alpha + \beta' X_i + \theta y^e{}_{n(i)} + \eta' X^e{}_{n(i)} + \nu_{n(i)} + \varepsilon_i \qquad (2)$$

where the error term is decomposed in two parts, one is the classical error term $\varepsilon_i$ reflecting unobserved characteristics which are peculiar to individual *i*, and the second reflects unobserved characteristics that individual *i* shares in common with other individuals in his neighbourhood $v_{n(i)}$.

Endogenous and contextual effects both express an influence of the social environment on the behaviours of individuals (i.e. group/neighbourhood matters), whereas correlated effects express a non-social phenomenon (Manski, 1995). Therefore, distinguishing between neighbourhood effects (i.e. endogenous and/or contextual effects) on the one hand and correlated effects on the other hand is important for the design of social policies. Indeed, if neighbourhood effects exist, policies which aim to achieve a more even distribution of individuals across neighbourhoods (for example, by relocating some categories of residents in more socio-economically diverse neighbourhoods) may have an impact on individual outcomes.

### Identification issues

Distinguishing between endogenous, contextual and correlated effects using standard observational data is not straightforward and two identification issues have been highlighted in the theoretical literature on social interaction and neighbourhood effects (Blume & Durlauf, 2006).

#### *The reflection problem*

The first issue pertains to the difficulty in distinguishing between endogenous and contextual effects and is named the reflection problem (Manski, 1993). Indeed, when researchers observe a correlation between an individual outcome $y_i$ and the average outcome in a neighbourhood $y^e{}_{n(i)}$, they cannot determine whether this correlation is due to the causal influence of aggregate outcomes or to the fact that aggregate outcomes simply reflect the role of average background characteristics in influencing individuals $X^e{}_{n(i)}$. In very simple terms, the average background characteristics influence average outcomes, which in turn affect individual outcome.

In this framework, identification amounts to distinguishing between the direct effect of average background characteristics on individual outcome (the contextual effect) and the indirect effect of background characteristics, as reflected through the endogenous effect generated by the average outcome of the group (Durlauf, 2004).

Identification of endogenous and contextual effects relies on strong assumptions regarding how the characteristics of the individuals vary with the characteristics of the group, and these assumptions may not be credible according to the nature of the data (see Blume & Durlauf, 2006, for a useful synthesis of formal conditions under which statistical identification is possible in the case of basic linear models as well as binary choice models). Very few studies try to properly disentangle endogenous and contextual effects (Ioannides & Zabel, 2008 is a recent example). Most empirical studies either estimate one of the two effects, positing the assumption that the other is absent, or instead estimate an aggregate of both effects. The same caveat applies to this study, its goal being to prove the existence of neighbourhood effects, whether they arise from endogenous or contextual effects.

*The self-selection or endogeneity issue*

The second problem pertains to identifying neighbourhood effects (i.e. contextual and endogenous effects taken together) and disentangling these from correlated effects (i.e. similarities in behaviours and outcomes arising because of unobserved characteristics shared by individuals in the neighbourhood). Correlated effects arise because individuals are not randomly distributed across the urban space. On the contrary, individuals sort themselves into neighbourhoods on the basis of their personal and family background characteristics (for example, income) and some of these characteristics also influence the outcome of interest. These background characteristics are either observed by the researcher, and might be controlled for (i.e. included in $X_i$), or unobserved. Because of these *unobserved* characteristics, distinguishing neighbourhood effect estimates from any correlated effects is difficult. As will become clear later in this chapter, if correlated effects are ignored, estimated neighbourhood effects will be biased (Dietz, 2002; Durlauf, 2004). This problem is referred to as the endogeneity issue or the self-selection issue in the literature, as it arises from the fact that individuals choose their neighbourhood of residence (i.e. "self-select" into neighbourhoods). It is important to note that "endogeneity" is used here in its econometric sense, without any relation to Manski's endogenous effect.

The first set of studies on neighbourhood effects completely ignored self-selection issues (see for example Datcher, 1982). However, more recent studies have paid particular attention to developing strategies for coping with this problem and disentangling neighbourhood effects from correlated effects (see for example Dietz, 2002, and Durlauf, 2004, for useful reviews). The purpose of this chapter is to review these strategies and to highlight their respective advantages and shortcomings. Doing this, we deliberately ignore questions relating to the distinction of underlying mechanisms through which neighbourhood effects operate (for example the distinction between endogenous and contextual effects). While distinguishing Manski's endogenous and contextual effects would clearly be of interest, we consider solving the self-selection issue as a precondition, and leave the identification of particular mechanisms for future research.

## SOME METHODOLOGICAL ISSUES

Despite an increasing interest, there is still no consensus regarding the magnitude and even the existence of neighbourhood effects in previous empirical work (Jencks & Mayer, 1990; Dietz, 2002). This may be due to the great diversity regarding the data, methods and variables used to test for the existence of neighbourhood effects, in particular regarding the way researchers correct or not for the endogeneity of residential locations (Dietz, 2002). Table 1 illustrates this lack of consensus by comparing results obtained by ten selected papers focusing on the potential impact of neighbourhood effects on labour-market outcomes. These papers have been chosen to reflect pioneering works that do not control for endogeneity (the first four studies) as well as more recent studies proposing various ways to deal with this problem. Among earlier works, Datcher (1982) and Case & Katz (1991) both find evidence of neighbourhood effects on the labour-force participation of young men. Osterman (1991) also finds convincing support for the existence of neighbourhood effects on the welfare participation of single mothers. However, Corcoran et al. (1992), extending Datcher's pioneering study, find no evidence of neighbourhood effects and attribute this contradictory result to the wider range of individual controls used. Among studies trying to deal with the endogeneity issue, three find evidence of neighbourhood effects (O'Regan & Quigley, 1996; Cutler & Glaeser, 1997; Fieldhouse & Gould, 1998). However, Katz et al. (2001), using quasi-experimental data from a government housing relocation program, find no impact of residential changes on adults' labour-market outcomes.

Finally, Plotnick & Hoffman (1999) and Weinberg et al. (2004) both find evidence of neighbourhood effects when endogeneity is not accounted for, but much smaller or non-existent effects when endogeneity is accounted for.

Before moving to the formal exposition of the endogeneity issue and to the enumeration of solutions that have been proposed, it is useful to briefly review two of the other methodological concerns raised when one intends to initiate an empirical evaluation of neighbourhood effects: the choice of spatial scale and the choice of measures to characterise neighbourhoods (see Dujardin, 2006, for more details on these methodological issues).

| Study | Outcome [1] and population | Context [2] | Neighbourhood: definition and characteristics | Method | Main findings |
|---|---|---|---|---|---|
| Datcher (1982) | Earnings (education); young adult men | SMSAs | Zip-code areas; average income, racial composition | OLS regressions | Neighbourhood variables are significant |
| Case & Katz (1991) | Labour-market activity (several other outcomes); young adult men | Boston | Street blocks; average outcome measures | Probit models | Strong evidence of peer effects |
| Osterman (1991) | Welfare participation; single mothers | Boston | Zip-code areas; zip-code dummies | Logit models | Neighbourhood effects are powerful; 5 neighbourhoods exert a significant effect |
| Corcoran et al. (1992) | Earnings, hours of work; young male household heads | SMSAs | Zip-code areas; median income, unemployment rate, single mothers, public assistance | GLS regressions; wide range of individual controls | No evidence of neighbourhood effects |
| O'Regan & Quigley (1996) | Employment idleness; youth (16-19) at home | 4 New Jersey cities | Census tracts; adult employment, racial composition, public assistance, job access | Logistic regression; sample restricted to youth living with parents | Neighbourhood variables have a clear effect |
| Cutler & Glaeser (1997) | Idleness, earnings (education); young people (20-30) | MSAs | City-level segregation indices | OLS and instrumental variables estimates; inter-city scale | Blacks are significantly worse-off in more segregated cities |
| Fieldhouse & Gould (1998) | Unemployment | Great Britain | Travel-to-Work areas; unemployment rate, social and racial composition | Multilevel models | Geographical variables are important |
| Plotnick & Hoffman (1999) | Income-needs ratio (pregnancy, schooling); sisters aged 17-24 | SMSAs | Census tracts; income, single mothers, public assistance | Siblings data and family fixed-effects models | Evidence of neighbourhood effects in OLS but not in fixed-effect models |
| Katz et al. (2001) | Economic self-sufficiency (child well-being); MTO participants | Boston | Census tracts; poverty rate, welfare receipt, single mothers | Comparison of outcomes between Experimental, Section 8 and Control groups | No impact of vouchers on employment, earnings or welfare participation |
| Weinberg et al. (2004) | Annual hours worked; young adult men | MSAs | Census tracts; adult employment rate, job access | Panel regressions; individual fixed-effect models | Neighbourhood effects are present but naïve estimates are overestimated by a factor 2 to 5 |

[1] This table is limited to neighbourhood effects on labour-market outcomes. If the cited paper studies in addition other types of outcomes, these are indicated in brackets but findings are only reported for labour-market outcomes. [2] (S)MSAs: (Standard) Metropolitan Statistical Areas.

Table 1: Comparison of 10 selected studies on neighbourhood effects on labour-market outcomes

### Some scale concerns

The first question one has to answer when initiating a research project on neighbourhood effects is the choice of a database on which to conduct analyses. In this respect, two important concerns are the level of data aggregation and the definition of neighbourhood size.

Addressing the issue of data aggregation first, empirical studies can be divided into two groups according to whether they use individual-level data or spatially-aggregated data (Sampson et al., 2002). In the first group, studies generally regress an individual outcome measure (for example an individual's employment status) on individual and family characteristics (age, gender, level of education, etc) and on a set of variables describing the social composition of the neighbourhood (for example, mean educational level). In the second group, researchers focus on data at a spatially-aggregated level (for example, census tracts) and try to explain spatial variations of the outcome of interest (mean unemployment rate) using indicators of the local social composition. The main problem with such an approach is the risk of ecological fallacy, an interpretation error arising when one tries to infer individual-level relationships from results obtained at an areal level of aggregation (Wrigley, 1995). Robinson (1950) was the first to provide empirical evidence for this potential problem. He showed that literacy and foreign birth in the US were positively associated at the State level (suggesting that foreigners were more likely to be literate than the native-born). In contrast, at the individual-level the correlation was negative. Moreover, in the context of neighbourhood effects, studies that use spatially-aggregated data are unable to distinguish neighbourhood effects from simple compositional effects (Oakes, 2004). Indeed, a correlation between an area's unemployment rate and its mean level of education can either be due to the fact that the unskilled are more often unemployed or the fact that the spatial concentration of unskilled generates poor social networks. Dujardin et al. (2004) provide evidence of the superiority of individual level data in neighbourhood effect studies.

A second and related scale concern is the choice of appropriate neighbourhood delineations. Due to data availability, virtually all studies rely on geographic boundaries defined administratively (such as census tracts or zip-code areas) and assume that two individuals living in the same administrative unit have more contacts than individuals living in different units. While it is unclear whether these administrative boundaries accurately represent the neighbourhood conditions that really make a difference in people's lives (Ellen & Turner, 1997), due to confidentially restrictions, researchers often have no other choice but to use such artificial neighbourhood definitions.

### How to characterise neighbourhoods

The second issue that arises is the choice of a set of indicators which may be used to measure and characterise the social composition of neighbourhoods. In this respect, several tendencies are observed. First, some authors do not characterise the social composition of neighbourhoods *per se* but use several dummy variables to reflect the individual's location in a particular area/neighbourhood (see for example, Osterman, 1991 which uses 16 dummies to reflect different areas in Boston). This method seems quite simple but becomes relatively untreatable as soon as the number of areas increases. More often, researchers use one or several quantitative measures of the social composition of neighbourhoods (for example, average family income, unemployment rate, racial composition, percentage of single-mother households or poverty rates; see column 4 in Table 1). However, while it is likely that individual outcomes are influenced by a wide variety of neighbourhood characteristics, introducing all of them into a regression analysis may cause collinearity problems as many indicators of neighbourhoods' social composition are highly correlated (Johnston et al., 2004). Usually the symptoms of such collinearity problems are instability in parameter values and significance levels (O'Regan & Quigley, 1996, present an illustration of this parameter instability). For this reason, some authors prefer using a composite measure of the social composition of neighbourhoods. For example, Buck (2001) characterises a neighbourhood's level of social exclusion by using a composite indicator of housing tenure, housing density, unemployment and car ownership. Similarly, Duncan & Aber (1997) use a factorial analysis to summarise thirty-four socio-demographic variables into a small number of composite factors. Dujardin et al. (2008) also use factorial analyses combined with clustering techniques to define five types of neighbourhoods and use these categories as indicators of different socioeconomic environments.

## THE ENDOGENEITY OF RESIDENTIAL LOCATIONS

The most difficult methodological problem in the study of neighbourhood effects is probably the endogeneity of residential choices which may lead the researcher to confound neighbourhood effects with simple correlated effects. The objective of the following subsections is first to provide a precise formulation of the problem and then to review the various solutions that have been proposed – sometimes wrongly – in order to try to solve it.

**Definition**

Endogeneity arises because residential locations are not exogenously determined (Dietz, 2002). On the contrary, individuals/households have some degree of choice regarding the neighbourhoods in which they live. Therefore, individuals with similar socioeconomic characteristics, notably similar labour-market outcomes, tend to sort themselves in certain areas across urban space. For example, individuals with well-paid jobs will choose to reside in better-off neighbourhoods in order to benefit from a social environment of better quality. There is thus a two-way causality: on the one hand, individual outcomes influence the choice of a residential location, while, on the other hand, residential location influences in turn individual socioeconomic outcomes. Of course, standard approaches to the estimation of neighbourhood effects allow one to control for some individual and household characteristics that might influence both neighbourhood choices and individual outcomes, as illustrated by the $X_i$ in the following equation (the same notation as in equation 1 is used but, for the sake of simplicity, Manski's endogenous and contextual effects are not distinguished anymore; $N_{n(i)}$ is used instead for indicating neighbourhood characteristics more generally):

$$y_i = \alpha + \beta' X_i + \gamma' N_{n(i)} + \varepsilon_i \tag{3}$$

However, it is likely that some individual and household characteristics are in fact unobserved (and therefore not included in $X_i$) and influence both the outcome of interest $y_i$ (for example labour-market outcome) and neighbourhood choice $N_{n(i)}$. For example, individuals with a low labour-market attachment or with low abilities (which decreases their labour-market performance) may choose to reside in deprived neighbourhoods for economic or social reasons. As a consequence, what the researcher perceives as a neighbourhood effect through the estimated $\gamma$ parameter may simply be a spurious correlation reflecting common residential choice (Weinberg et al., 2004). Such unobserved individual/household characteristics are in fact incorporated in the error term $\varepsilon_i$, thus generating a correlation between $\varepsilon_i$ and the observed regressor $N_{n(i)}$ (recall that in equation 2, this error term was decomposed into an individual specific error term $\varepsilon_i$ and a group specific error term $v_{n(i)}$; equation 3 now assumes that $v_{n(i)}$ being unobserved, it cannot be distinguished from $\varepsilon_i$). Therefore, the consistency assumption in OLS estimation methods stating that regressors must be uncorrelated with the error term is not valid as

$$E\left(\varepsilon_i \middle| X_i, N_{n(i)}\right) \neq 0 \tag{4}$$

and results of studies based on standard methods that do not control for these unobserved characteristics will be biased (a formulation of the bias can be found in Greene, 2008). As it arises from the presence of unobserved individual and household characteristics, the endogeneity bias might also be understood as an omitted variable bias and a consistent estimation of equation 3 would require constructing a consistent estimate of $E(\varepsilon_i|X_i,N_{n(i)})$ and using it as an additional regressor (Durlauf, 2004).

The direction of the bias is difficult to predict as it depends on the relationship between the unobserved factors that determine neighbourhood choices and the unobserved factors that determine the outcome of interest (Evans et al., 1992). However, researchers generally assume that neighbourhood effects are *overestimated*. For example, in the context of neighbourhood effects on children outcomes, parents that lack the financial resources to move to better neighbourhoods often lack the qualities to help their children to perform well in school; then the true $\gamma$ parameter will be overestimated leading to an *upward* bias. However, a *downward bias* can also arise if parents choose a single-earner strategy and earn less, therefore being forced to live in poor neighbourhoods, but allowing the stay-at-home parent to spend more time with their children (Duncan et al., 1997). The following subsections review the methods that have been proposed in order to cope with this issue. The first method is based on experimental data, where exogeneity is reached through random assignment of individuals in different neighbourhoods, whereas the other methods use standard observational data collected through conventional surveys (such as census data).

### Quasi-experiments

Perhaps the best way to correctly identify the causal effects of neighbourhoods would be to conduct a kind of controlled experiment in which individuals would randomly be assigned to neighbourhoods (Durlauf, 2004). One could then compare the socioeconomic outcomes of similar individuals with respect to the characteristics of the neighbourhood in which they are assigned. While it is difficult to translate such controlled experiments to the study of neighbourhood effects, data provided by government-subsidised relocation programs in the United States can be considered as *quasi-experiments* (see Oreopoulos, 2003, for a review of such programs). Indeed, government interventions into the residential choices of households can be used to assess neighbourhood effects, as households that would normally belong to one neighbourhood are moved to another through an exogenous intervention. The best-known example is the *Moving To Opportunity* Program conducted by the Department of

Housing and Urban Development in five American cities (Baltimore, Boston, Chicago, Los Angeles and New York). In this program, eligible families (public housing families with children residing in census tracts with at least 40% of poverty) who volunteered for the program were randomly drawn from the waiting list and randomly assigned into three groups: the Experimental group, in which families received housing subsidies and search assistance to move to private-market housings in low-poverty census tracts, the Comparison group, in which families received geographically unrestricted housing subsidies but no search assistance, and the Control group, which received no special assistance. Households were periodically surveyed in years following relocation, thus allowing researchers to analyse the resulting evolution of their socioeconomic characteristics (see for example Katz et al., 2001, and Kling et al., 2007).

While it is clear that such data are well-suited for removing the endogeneity bias in neighbourhood effects studies, they do present some disadvantages. The first issue is undoubtedly the cost and difficulty of implementation as well as the ethical concerns surrounding such experiments (Harding, 2003). Moreover, eligibility conditions, screening of families and participation willingness generate sample selection problems (Dietz, 2002; Durlauf, 2004). Indeed, the estimates give the effects of the relocation program on the types of people who were authorised and chose to participate, and may thus be different from the effects obtained from relocating a randomly selected group of families from poor areas. Brock & Durlauf (2001) also suggest that the implementation of such programs to a wider scale would result in moves of large numbers of households, thus modifying the social composition of neighbourhoods. Therefore, estimation of neighbourhood effects without taking into account induced changes in the neighbourhood composition would be misleading and would not allow the generalisation of results to wide-scale programs.

#### Sample restriction

The second method is perhaps the easiest to implement and the most frequently used. It consists in restricting the studied sample to a group of individuals for whom residential choices are limited and might thus be considered as fairly exogenous. For example, in their study of neighbourhood effects on the employment outcomes of young adults, O'Regan & Quigley (1996) limit themselves to a sample of youngsters still living with at least one parent.

They justify their approach by the fact that location choice has been made previously by the parents and can thus be thought of as exogenous to the employment status of their children.

However, the sample restriction method presents some disadvantages (Glaeser, 1996; Ihlanfeldt & Sjoquist, 1998). First, it is only applicable to very limited sub-populations (generally people under 25) and one cannot use it when the focus is on elderly adults, nor can results be generalised to young adults that have moved out of parental home. Moreover, it can create a sample selection bias. Indeed, in the case of neighbourhood effects on labour-market outcomes, young adults obtaining a job are more likely to leave their parents' home than still-unemployed young adults, and this leaving rate might differ according to the perceived characteristics of neighbourhoods. Finally, it does not completely eliminate endogeneity bias. Indeed, the household's residential choice depends on observed as well as unobserved parental characteristics and some of these unobserved parental characteristics might also influence children employment outcomes. For example, lack of commitment to work or social norms may induce parents to locate in high poverty neighbourhoods and also probably influence youngster's motivation and intensity of job search (Glaeser, 1996).

**Siblings data and family fixed effects**

Some studies resort to siblings data to solve endogeneity, i.e. data on individual children from the same family (Aaronson, 1998; Plotnick & Hoffman, 1999). It consists in finding siblings pairs in which one sibling was exposed to a particular neighbourhood and the other one to another neighbourhood (because of family residential moves). For each sibling *i*=1, 2 belonging to family *f*, the outcome of interest ($y_{if}$) is then expressed as:

$$y_{if} = \alpha' X_f + \beta' X_{if} + \gamma' N_{n(i)} + \varepsilon_f + \varepsilon_{if} \tag{5}$$

where $X_f$ is a vector of family-specific variables, $X_{if}$ is a vector of individual characteristics for sibling *i* belonging to family *f*, $N_{n(i)}$ is the standard vector of neighbourhood characteristics, $\varepsilon_{if}$ is an error term associated with sibling *i* (individual unobserved characteristics) and $\varepsilon_f$ is an error term associated with family *f* (family unobserved characteristics). Then, by assuming that observed and unobserved family characteristics are constant across time (and thus for the two siblings), differentiating outcomes between siblings eliminates family effects and $\gamma$ is interpreted as an unbiased estimate of neighbourhood effects:

$$y_{1f} - y_{2f} = \beta'(X_{1f} - X_{2f}) + \gamma'(N_{n(1)} - N_{n(2)}) + (\varepsilon_{1f} - \varepsilon_{2f}) \quad (6)$$

The application of this method is quite limited due to data availability as it requires data on families that have raised two or more children in different neighbourhood conditions. Moreover, the assumption on which it rests is quite strong as it requires that parents do not change their behaviour when moving (Durlauf, 2004). However, residential moves might reflect changes in parents' unobserved characteristics that may also influence children outcomes (change in income, divorce). Therefore, such estimates of neighbourhood effects might in fact reflect changes in family unobservables.

#### Longitudinal data and individual fixed effects

Recent studies have used longitudinal data that allow researchers to track individuals over time and follow their successive residential moves (see for example Weinberg et al., 2004). By comparing individual outcomes before (*t*=1) and after the move (*t*=2), one can assess neighbourhood effects using the equations below:

$$y_{it} = \alpha + \beta' X_{it} + \gamma' N_{n(it)} + \varepsilon_{it} \quad (7)$$

$$y_{i2} - y_{i1} = \beta'(X_{i2} - X_{i1}) + \gamma'(N_{n(i2)} - N_{n(i1)}) + (\varepsilon_{i2} - \varepsilon_{i1}) \quad (8)$$

$X_{it}$ is a vector of individual characteristics for individual *i* at time *t*, some of these being constant across time (such as gender) while others may vary across time (such as level of education or marital status). $N_{n(it)}$ is a vector of characteristics of the neighbourhood in which the individual *i* resides at time *t*. However, this method suffers from the same problems as siblings studies. Data availability concerns limit its usefulness and it relies on the assumption that unobserved individual characteristics ($\varepsilon_{it}$) that influence socioeconomic outcomes are constant across time. Yet, it is likely that residential moves are the consequence of previous changes in individual characteristics, including observables such as the family composition as well as unobservable determinants.

#### Multilevel models

Multilevel models (also known as hierarchical, mixed or random-coefficient models) have also been used to study the influence of neighbourhoods on individual outcomes, particularly in the context of health studies (Blakely and Subramanian, 2006). Without going into details, the principle of multilevel modelling is that the parameters of the model are not constant but are allowed to vary across neighbourhoods:

$$y_i = \alpha_{n(i)} + \beta_{n(i)}{}' X_i + \varepsilon_i \text{ with } \alpha_{n(i)} = \alpha + \upsilon_{n(i)} \text{ and } \beta_{n(i)} = \beta + \mu_{n(i)} \qquad (9)$$

In this formulation, $\alpha_{n(i)}$ and $\beta_{n(i)}$ are random variables, with means $\alpha$ and $\beta$. $\upsilon_{n(i)}$ and $\mu_{n(i)}$ are neighbourhood's deviations from mean values and reflect the fact that the relationship between the outcome and the individual-level variables varies across neighbourhoods. Note that these random coefficients can further be explained by neighbourhood-level explanatory variables (see Goldstein, 2003, for more detailed formulations).

In other contexts, multilevel models present numerous advantages. Notably, they allow one to take into account the fact that, because of the grouping of individuals within neighbourhoods, individuals from the same neighbourhood are generally more similar than individuals drawn randomly from the population, which violates the assumption of standard OLS regressions that individual observations must be independent (Goldstein, 2003). However, in the context of epidemiologic studies, it has been argued that, by explicitly considering the grouping of individuals into neighbourhoods, multilevel models allow one to obtain unbiased estimates of neighbourhood effects and to perfectly distinguish neighbourhood effects from the effects of individual characteristics (which are often termed "compositional" effects in this context). As is argued by Oakes (2004), this assertion is a little bit presumptuous. Indeed, as they do not take into account the reciprocal nature of the relationship between neighbourhood and outcome, multilevel models are unable to specify whether the estimated spatial variations of the relationship between $y_i$ and $X_i$ are the result of neighbourhood effects or the result of non-random neighbourhood selection (i.e. correlated effects). Indeed, spatial variations in $\alpha_{n(i)}$ and $\beta_{n(i)}$ may simply indicate that some unobserved individual characteristics have an impact on the outcome and simultaneously determine neighbourhood choice.

### Inter-city studies

A sixth group of studies have used inter-city data in order to evaluate the effects of residential segregation on individual outcomes, and to achieve exogeneity. The best known example of such a study is Cutler & Glaeser (1997). In their study of the effect of racial and income segregation on individual outcomes, these authors argue that the selection bias arising from the fact that more successful African-Americans will choose to live in richer and whiter neighbourhoods is difficult to solve using an intra-urban approach. Instead, by focusing on inter-city data, it is possible to evaluate if African-Americans in more segregated cities on average have worse or better outcomes than African-Americans in less segregated cities. By doing this, one avoids the within-city sorting of individuals along abilities and other unobserved characteristics.

However such studies do present some shortcomings. First, one can question the comparability of segregation measures across different metropolitan areas. Indeed, the effect of size of spatial units on the value taken by such indicators is well documented (see for example Wong, 2004, for a discussion of the Modifiable Areal Unit Problem in the context of segregation indicators). By comparing segregation indicators across metropolitan areas, one implicitly assumes that spatial units in different cities are of comparable size and shape. This seems a rather strong assumption. Moreover, the constructed segregation index applies to all people residing in the same metropolitan area, regardless of their specific location within the city. Yet, there can be substantive variations in social composition within the city itself (Ihlanfeldt & Sjoquist, 1998). In the extreme case where every metropolitan area has the same degree of residential segregation but where important disparities exist between neighbourhoods in the same city, such an analysis would identify no segregation effects at all. A huge cost in terms of loss of information is thus attached to such strategies (Glaeser, 1996).

#### Instrumental variables

More generally, in econometrics, endogeneity is dealt with by using instrumental variable techniques. Broadly speaking, this method consists in replacing the endogenous regressor by an instrument, i.e. a variable which is highly correlated with the endogenous regressor but not with the unobserved determinants of the outcome. This therefore eliminates the correlation between the regressor and the error term (see Greene, 2008 for an introduction to instrumental variables). The most common form of Instrumental Variable Estimation (IVE) is two-stage least squares (2SLS). In the first stage, the endogenous regressor (in the case of neighbourhood effect studies, this would be the neighbourhood characteristic $N_{n(i)}$) is regressed on a set of chosen instruments. The resulting coefficient estimates are used to generate a predicted value for the neighbourhood characteristic. In the second stage, this predicted value is used in the outcome equation, in place of the actual neighbourhood characteristic. If the instruments are correctly specified (i.e. they are highly correlated with the neighbourhood characteristics but are not direct determinants of the outcome), then the predicted value from the first-stage equation is uncorrelated with the error term of the outcome equation.

Instrumental variables are rather common in econometric studies. However, they have rarely been applied to the study of neighbourhood effects. This is mainly because it is hard to find a good instrument in this particular context (Dietz, 2002). Indeed, one has to find a variable that is a true determinant of neighbourhood choice but that is not correlated with the outcome measure. One can easily figure out how tricky it is to find such an instrument, and the few examples of instrumental variable estimates of neighbourhood effects have received some criticism (Durlauf, 2004). For example, in their study of neighbourhood influences on educational outcomes, Duncan et al. (1997) use the characteristics of the mother's neighbourhood of residence after the child has left the parental home. This assumption relies on the fact that as long as the child is at home, the neighbourhood of residence reflects both the mother's preferences as well as her (unobserved) concerns about the effects of neighbourhood on the child's development. After the child has left home, the mother's neighbourhood choice no longer reflects these concerns but only her own preferences. However, the authors themselves criticize their instrument choice on the basis that inertia may cause the mother to stay in the same neighbourhood even after the children have left home. Evans et al. (1992) also use IVs to study the effect of peer groups on teenage pregnancy and school dropout rates. They assess the percentage of disadvantaged pupils in the school by measures of unemployment and poverty rates at the metropolitan level, based on the assumption that adolescents living in a metropolitan area with a high poverty rate are more likely to attend a school with a higher percentage of disadvantaged students. This choice of instrument is debatable as it implies exogeneity in parent's choice of a metropolitan area and it is not clear how such instruments can account for neighbourhood effects within cities (Durlauf, 2004).

While the validity of the instruments used by Evans et al. (1992) may be questioned, their results are quite interesting as these authors provide peer group effects estimates obtained with and without treatment for endogeneity. First, they regress the propensity for a teenage girl to become pregnant on a set of family characteristics as well as on the log of the percentage of disadvantaged students in her school. They find a positive and significant coefficient for this last variable, thus indicating the presence of peer group effects. Then, using an instrumental variable specification, they find no evidence at all for the existence of peer group effects, concluding that the significant effect in the first specification could in fact be attributed to unobserved family determinants affecting the choice of a school for their girl as well as the girl's propensity to become pregnant.

Their findings thus strongly advocate in favour of an explicit treatment of endogeneity in all studies of peer group and neighbourhood influences on individual outcomes.

### Conclusion

While initial studies of neighbourhood effects rarely mentioned the problems associated with endogeneity, nearly all recent empirical studies are aware of its existence and attempts to solve it are numerous. Results of comparative studies suggest that the biases arising from not taking the endogeneity of residential location choices into account are important (Evans et al., 1992; Plotnick & Hoffman, 1999; Weinberg et al., 2004). Unfortunately, the perfect solution does not exist and the various methods reviewed in this chapter either rely on very particular datasets that are rarely available (for example experimental data or siblings data) or present several important shortcomings. Durlauf (2004) suggests that future empirical works should attempt to simultaneously model neighbourhood effects and neighbourhood configurations via structural models: "Structural models will allow for a full exploration of self-selection in neighbourhoods models" (Durlauf, 2004: 2232). To our knowledge, there have been only few attempts at such structural modelling in the context of neighbourhood effects and this will have to be on the research agenda for the next years.

## EMPIRICAL APPLICATION: LABOUR-MARKET OUTCOMES OF YOUNG ADULTS IN BRUSSELS

The remainder of this chapter is devoted to an empirical illustration of the study of neighbourhood effects and how to cope with endogeneity issues. It presents results of an empirical exercise aiming at estimating neighbourhood effects on labour-market outcomes of young adults residing in the Brussels urban area. Testing all the methods presented in the previous section to cope with endogeneity would be an impossible task as most of these methods rely on particular datasets that are not available for Brussels. Instead, the endogeneity of residential choices will be treated with the sample restriction method, i.e. restricting the sample to young adults residing with their parents. In addition, a step-by-step model-building strategy will be used in order to provide an assessment of the sensitivity of the results in the presence of observed and unobserved parental covariates. This allows evaluating the robustness of the results. This method had already been applied by the authors on 1991 census data in a previously published paper (Dujardin et al., 2008). The analyses presented here provide an update of the results of this previous paper with 2001 census data. The following subsections proceed by briefly describing our research question (i.e. why residential location should influence labour-market outcomes), the study area and the data used. Then, the methods and results will be presented in detail.

### Research question

Estimates of neighborhood effects on individual outcomes are illustrated in the particular context of labour-market outcomes. There are numerous reasons why residential locations might influence labour-market outcomes. First, the socioeconomic composition of the neighbourhood can influence *human capital acquisition*, especially for young adults, which may in turn deteriorate their employability in later years. Indeed, as the success of a given student depends on the results of other students in his class, in neighbourhoods with a high concentration of low-ability students, peer effects can deteriorate school achievements and employability (Benabou, 1993). Social problems which deteriorate employability (like dropping out of school, consuming drugs or having illicit activities) also spread through social interactions within the neighbourhood (Crane, 1991). This contagion is all the more prevalent as adults are themselves unemployed and do not provide a role model to which youngsters could identify (Wilson, 1987). Moreover, the socioeconomic composition of the neighbourhood influences the quality of *social networks*. This is a crucial point since a significant portion of jobs are usually found through personal contacts and since low-skilled workers, young adults, and ethnic minorities often resort to such informal search methods (Holzer, 1988). Therefore, in neighbourhoods where local unemployment rates are higher than average, local residents know fewer employed workers that could refer them to their own employer or provide them with professional contacts. Finally, employers may be reluctant to hire workers residing in disadvantaged neighbourhoods (a practice known as *territorial discrimination*; Zenou & Boccard, 2000). Note that the spatial mismatch hypothesis, which emphasises the role of physical disconnection between job opportunities and residential locations, has also been put forward to explain the poor labour-market outcomes in the context of black ghettos in American cities (see Gobillon et al., 2007 for a survey). Previous work has suggested this hypothesis has no support in Brussels; we therefore will not elaborate further on this issue (Dujardin et al., 2008).

## Data and studied area

### *Studied area and neighbourhood size*

In institutional terms, Brussels comprises 19 municipalities and hosts around 1 million inhabitants on a 163 km² area. However, as in most cities, Brussels' functional metropolitan area extends far beyond its institutional limits. Therefore, the so-called Extended Urban Area is used to reflect its functional metropolitan area, which comprises 41 municipalities that together host 1.4 million inhabitants and cover a 723 km² area.

The smallest spatial unit for which census data are officially available is the statistical ward (in French, "secteur statistique"), a subdivision of the municipality defined according to social, economic and architectural similarities (Brulard & Van der Haegen, 1972). Statistical wards that present a common functional or structural nature (for example a common attraction pole like a school or a church) can further be grouped into larger entities, which constitute an intermediate level between the statistical ward and the municipality. The limits of these units were generally defined by geographical obstacles, like important roads, railways or waterways. Because of this delineation criterion and their functional definition, these units can be considered as appropriate to reflect social influences and are used here to define the local environment that may potentially matter for individuals. There are 328 such neighbourhoods in the Extended Urban Area, grouping on average 4,250 inhabitants each. For statistical reasons, neighbourhoods with fewer than 200 inhabitants are not considered in this analysis (i.e. 19 neighbourhoods were left aside).

### *Data and neighbourhood characteristics*

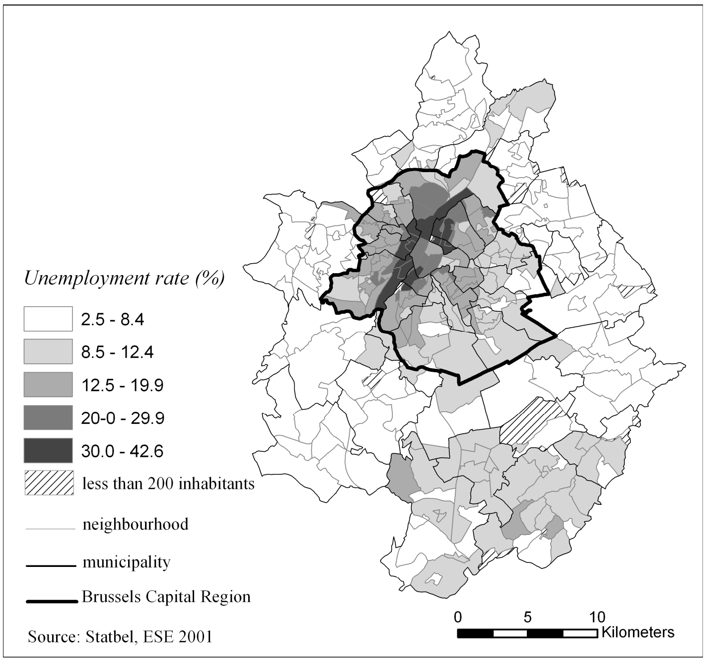
The statistical analyses are based on data extracted from the 2001 Socioeconomic Survey carried out by the Belgian National Institute of Statistics, which provides data for all individuals residing in Belgium (i.e. this is a 100% sample). Analyses were restricted to members of private households. For each individual, detailed information on personal characteristics (including age, gender, education, citizenship, employment status, kinship with household's head) are provided, along with family and housing characteristics (for instance, type of family or car ownership) as well as statistical ward (and neighbourhood) of residence.

These statistical analyses aim at explaining an individual's unemployment probability by taking into account personal and household characteristics as well as the possible role played by the neighbourhood of residence. It is therefore necessary to choose one or several measures of the socioeconomic composition of neighbourhoods. As previously mentioned, even though it is likely that individual outcomes are determined by a wide variety of neighbourhood characteristics (such as neighbourhood household mean income, unemployment rate or racial composition), considering all these together into a single regression may cause colinearity problems (Johnston et al., 2004). Therefore, a typology of neighbourhoods is used, which is intended to reflect different types of social environments within Brussels. This typology is built on a set of eleven neighbourhood characteristics, chosen in order to reflect various aspects of the social composition of neighbourhoods likely to affect labour-market outcomes. These variables concern educational levels, professional statuses, unemployment rates, percentages of foreigners, of single-mother households as well as average household income. First, a Principal Component Analysis is run in order to define a limited number of non-correlated factors summarizing the information carried by this set of neighbourhood variables. Then, neighbourhoods are grouped according to their coordinates on the factorial axes, using a hierarchical ascending classification (with the Ward method which minimizes intra-group variance), in order to define deprived neighbourhoods versus non-deprived neighbourhoods (in a previous paper, a 5-class typology was used; the 2-class typology is used here for the sake of simplicity; see Dujardin et al., 2008 for more details on this classification).[1]

*Spatial structure of Brussels*

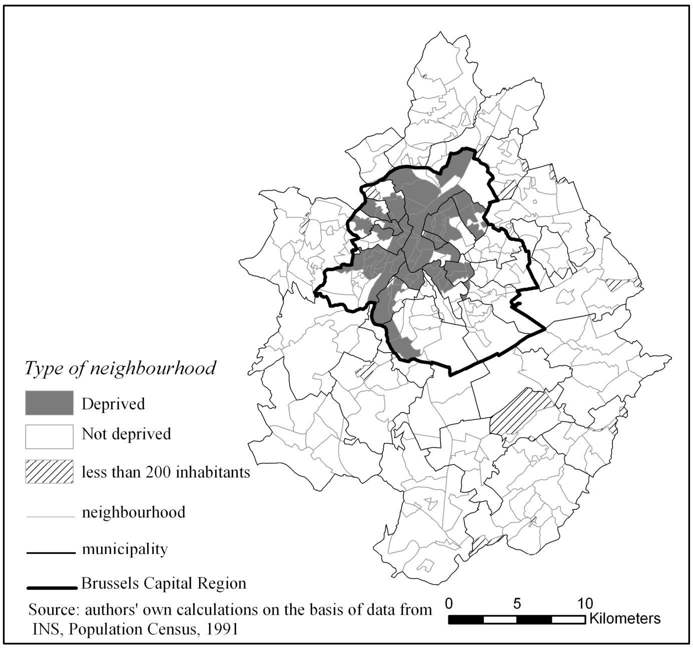
The Brussels Extended Urban Area presents a well-marked spatial structure characterised by important disparities opposing its city centre to the periphery. Figure 1 maps the percentage of unemployed workers among labour-force participants aged 19 to 64 in 2001, highlighting a zone of very high unemployment rates (above 20%, and even above 30% for some neighbourhoods) in the central part of the urban area, along the former industrial corridor. On the contrary, unemployment is much lower (below 12.5% or even 8.5%) in the suburbs. Figure 2 maps deprived neighbourhoods and table 2 summarises their characteristics. Deprived neighbourhoods are located in the centre of Brussels. They are characterised by high unemployment rates (2.5 times as high as the average unemployment rate in non-deprived neighbourhoods), high proportions of North-Africans and Turks and single-mother households, low educational levels, high percentage of blue-collars, and low income levels.

[1] The neighbourhood typology used here was defined by Dujardin et al. (2008) on the basis of 1991 census data. This 1991 typology is used here to explain individual unemployment propensities in 2001. Because of inertia in housing prices and residential choices, it is likely that similar results would have been obtained for a typology of neighbourhoods on the basis of 2001 data.



Source: Statbel, ESE 2001

Figure 1: Percentage of unemployed workers among labour-force participants in the Brussels E.U.A. in 2001



Source: authors' own calculations on the basis of data from INS, Population Census, 1991

Figure 2: Location of deprived neighbourhoods in the Brussels E.U.A.

| | Deprived | Not deprived | *Total* |
|---|---|---|---|
| Demography | | | |
| % North-Africans and Turks | 16.1 | 0.7 | *4.3* |
| % single-mother households | 23.5 | 15.8 | *17.6* |
| Average household income | 772 | 1,113 | *1,033* |
| Education | | | |
| % of students in technical classes | 39.7 | 25.5 | *28.8* |
| % with lower education | 53.9 | 43.2 | *45.7* |
| % with at least intermediate education | 28.6 | 46.2 | *42.1* |
| % with higher educational levels | 13.0 | 23.8 | *21.3* |
| Professional status | | | |
| % blue-collars | 31.7 | 16.2 | *19.8* |
| % executives | 7.4 | 14.0 | *12.4* |
| Unemployment | | | |
| Unemployment rate (19-64) | 18.8 | 7.4 | *10.0* |
| Youth unemployment rate (19-25) | 26.5 | 15.0 | *17.7* |
| Total population | 563,704 | 829,106 | *1,392,810* |
| Number of neighbourhoods | 72 | 237 | *309* |

*Source: calculations based on data from the 1991 Population Census (INS)*

Table 2: Mean characteristics of neighbourhood types

### A logistic model of unemployment probability with sample restriction

Individual-level data are used to estimate a logistic model of unemployment probability while taking into account both personal and household characteristics as well as the neighbourhood of residence, as illustrated in the following equation:

$$Log\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_1' I_i + \beta_2' H_i + \gamma DN_i \tag{10}$$

where $P_\text{i}$ is the unemployment probability of individual *i*, $I_i$ is a vector of personal characteristics, $H_i$ is a vector of household characteristics and $DN_i$ is a dummy variable indicating whether the neighbourhood in which individual *i* resides is deprived ($DN_i$=1) or not ($DN_i$=0). $\alpha$, $\beta_1$, $\beta_2$ and $\gamma$ are vectors of parameters that will be estimated using Maximum Likelihood Estimation (MLE). In particular, $\gamma$, when significantly different from zero identifies an impact of neighbourhood deprivation on unemployment, i.e. *neighbourhood effects*. Using (10), the individual probability of unemployment $P_i$ is given by:

$$P_i = \frac{e^{(\alpha + \beta_1' I_i + \beta_2' H_i + \gamma DN_i)}}{1 + e^{(\alpha + \beta_1' I_i + \beta_2' H_i + \gamma DN_i)}} \tag{11}$$

In this equation, the parameter $\gamma$ is potentially subject to an endogeneity bias. As already explained in the first part of this chapter, residential locations are partly determined by labour-market outcomes as individuals choose where they live and sort themselves in the urban space along their socioeconomic characteristics. In other words, the right-hand side variable $DN_i$ in equation 11 is an endogenous regressor, which is determined jointly with the left-hand side variable $P_i$, and $\gamma$ cannot be estimated without any bias using standard methods. The first part of this chapter discussed various strategies used in the literature to correct for the endogeneity of neighbourhood choice. Although it is an imperfect solution (Ihlanfeldt and Sjoquist, 1998), the sample restriction method is used here in this purpose and all statistical analyses are restricted to young labour-force participants (aged 19 to 25) residing with at least one parent (as in O'Regan & Quigley, 1996). This rests on the assumption that the choice of a residential location has been made previously by the parents and is thus fairly exogenous to the employment status of their children. In addition, focusing on at-home young adults will enable the use of parental explanatory variables as the individual census database allows identifying members of the same household. Parental characteristics are indeed important determinants of children outcomes and it is important to take these into account, when possible.

Setting aside individuals for whom important personal and family characteristics are missing as well as individuals living in neighbourhoods of fewer than 200 inhabitants, the studied sample consists of 27,044 individuals.

In a first step, equation 11 is estimated using only individual-level explanatory variables as well as neighbourhood type. Parental characteristics will be introduced in a second step, in the sensitivity analysis. The set of individual explanatory variables $I_i$ includes gender, age, level of education and citizenship. Three levels of education are distinguished: lower for individuals with at most a diploma of junior secondary education (normally corresponding to an age of 15); intermediate for those with a diploma of senior secondary education (normally aged 18); and higher for those with a higher diploma. Concerning citizenship, four main groups are defined: Belgians, foreigners from the European Union, North-African and Turkish foreigners, and other foreigners. Furthermore, the nationality of the household head is used to approximate the concept of ethnicity, by distinguishing Belgians with Belgian parents, Belgians with EU parents, Belgians with North-African or Turkish parents, and Belgians with parents of other nationality.

Results from this model are presented in Model I of Table 3, while Model II adds a dummy variable for living in a deprived neighbourhood or not. Note that all results are presented in terms of odds ratios. This means that the reference value is one (which indicates no effect) and that a value above one indicates that the corresponding variable increases the unemployment probability, while an odds ratio between zero and one indicates that the corresponding variable decreases the unemployment probability. Results show that men or educated workers are less likely to be unemployed than women or workers with a lower education. The probability of unemployment also decreases with the age of the individual. Moreover, citizenship plays a key role: North-Africans and Turks and other foreigners are more disadvantaged than UE citizens and Belgians. This is consistent with discrimination on the labour market, but may simply reflect differences in competence and qualification that are not taken into account by the educational level and which may differ between ethnic groups (such as experience or language fluency). Interestingly, young Belgian adults born to foreign parents are more likely to be unemployed than young Belgian adults of Belgian parents, suggesting that besides citizenship, the name or visible characteristics associated with foreign origin are a handicap on the labour market. This is consistent with both labour-market discrimination as well as social networks of lower quality for individuals with foreign parents.

Introducing neighbourhood deprivation in the regression (Model II) significantly increases the fit of the model (see the likelihood ratio and the Akaike Information Criterion). The odds ratio of the neighbourhood deprivation variable is 1.491 and is significant at a 1% level, indicating that all else being equal, residing in a deprived neighbourhood significantly increases the odds of being unemployed by nearly 1.5. This indicates that, all else being equal, young adults living in neighbourhoods characterised by the worst combination of social characteristics are more likely to be unemployed, thus confirming the importance of the social environment on labour-market outcomes (through mechanisms such as peer effects, role models or poor social networks).

As was already mentioned, the sample restriction used here to solve endogeneity presents some shortcomings, the most important being that it does not completely eliminate endogeneity. Indeed, the assumption on which it rests (that residential choice is made previously by the parents and is thus exogenous to the employment status of their young adult children) is questionable. Indeed, it is likely that parental characteristics determining residential choices also influence children's future employment outcomes (Glaeser, 1996). In this context, Model II does not allow distinguishing neighbourhood effects from the effect of parental characteristics on the unemployment probability of young adults living with their parents.

Instead of attempting to completely remove endogeneity, it may be useful to evaluate potential remaining bias by conducting a sensitivity analysis, in order to assess the robustness of estimated neighbourhood effects (as suggested by Glaeser, 1996: 62). A two-step strategy is therefore used. Firstly, several models of unemployment probability are estimated, which incorporate various sets of parental characteristics. The estimated neighbourhood effects from these models are compared in order to test the robustness of the results in the presence of *observed* parental covariates. The second step of the sensitivity analysis consists in generating random variables that are correlated to a certain degree with both the unemployment probability of young adults and with parental residential choice. Introducing these random variables in the unemployment probability model tests the sensitivity of the results to *unobserved* parental covariates.

### Sensitivity to observed parental covariates

In their study of neighbourhood effects on children outcomes, Ginther et al. (2000) argue that the disparity in past estimates of neighbourhood effects is mainly the result of differences in the set of included family characteristics. They specify a number of models that vary to the extent to which family characteristics are introduced as statistical controls and show that estimated neighbourhood effects tend to fall in value (and sometimes become insignificant) as the set of family controls becomes more complete. Building on their framework, several models of youth unemployment probability are estimated, which incorporate various sets of household characteristics, moving away from a model with no household variable towards a model including an extensive set of parental and household controls. These parental characteristics were built by assigning to each young adult the characteristics of the household head or that of the household head's spouse (when data was missing for the household head). For each young adult, the parental employment status and educational level were computed, as well as the household car ownership. Single-mother households were also identified, as these households are more frequently prone to social problems detrimental to finding a job. Table 3 includes these parental characteristics step by step in a model including only youth personal characteristics and the characteristics of their neighbourhood of residence (Models III to VI). Comparing estimated neighbourhood effects in these different models allows one to test the robustness of the results to the omission of observed parental covariates.

| Model | I | II | III | IV | V | VI |
|---|---|---|---|---|---|---|
| *Likelihood ratio* | 690.91 | 842.48 | 964.55 | 1040.05 | 1109.54 | 1130.33 |
| *Akaike Information Criterion* | 28,791 | 28,641 | 28,523 | 28,452 | 28,384 | 28,365 |
| Neighbourhood Type | | | | | | |
| *Deprived* | | 1.491$^{***}$ | 1.415$^{***}$ | 1.470$^{***}$ | 1.414$^{***}$ | 1.418$^{***}$ |
| Individual characteristics | | | | | | |
| *Male* | 0.937$^{**}$ | 0.946$^{*}$ | 0.951$^{NS}$ | 0.946$^{*}$ | 0.949$^{*}$ | 0.950$^{*}$ |
| *Age* | 0.914$^{***}$ | 0.915$^{***}$ | 0.908$^{***}$ | 0.905$^{***}$ | 0.906$^{***}$ | 0.906$^{***}$ |
| *Education* | | | | | | |
| *Lower* | Ref. | Ref. | Ref. | Ref. | Ref. | Ref. |
| *Intermediate* | 0.641$^{***}$ | 0.661$^{***}$ | 0.674$^{***}$ | 0.656$^{***}$ | 0.667$^{***}$ | 0.669$^{***}$ |
| *Higher* | 0.695$^{***}$ | 0.747$^{***}$ | 0.780$^{***}$ | 0.705$^{***}$ | 0.724$^{***}$ | 0.729$^{***}$ |
| *Citizenship* | | | | | | |
| *Belgian (of Belgian parents)* | Ref. | Ref. | Ref. | Ref. | Ref. | Ref. |
| *Belgian (of EU parents)* | 1.111$^{NS}$ | 1.069$^{NS}$ | 1.067$^{NS}$ | 1.110$^{NS}$ | 1.120$^{NS}$ | 1.136$^{*}$ |
| *Belgian (of North-African or Turkish par.)* | 1.916$^{***}$ | 1.517$^{***}$ | 1.351$^{***}$ | 1.424$^{***}$ | 1.397$^{***}$ | 1.443$^{***}$ |
| *Belgian (of parents of other citizenship)* | 1.893$^{***}$ | 1.689$^{***}$ | 1.574$^{***}$ | 1.501$^{**}$ | 1.469$^{***}$ | 1.491$^{***}$ |
| *EU* | 1.337$^{***}$ | 1.220$^{***}$ | 1.198$^{***}$ | 1.246$^{***}$ | 1.256$^{***}$ | 1.268$^{***}$ |
| *North African and Turkish* | 3.203$^{***}$ | 2.543$^{***}$ | 2.267$^{***}$ | 2.363$^{***}$ | 2.297$^{***}$ | 2.374$^{***}$ |
| *Other* | 2.815$^{***}$ | 2.548$^{***}$ | 2.435$^{***}$ | 2.342$^{***}$ | 2.159$^{***}$ | 2.175$^{***}$ |
| Parental and household characteristics | | | | | | |
| *Employment status and professional status* | | | | | | |
| *Employed* | | | Ref. | Ref. | Ref. | Ref. |
| *Not participating to labour force* | | | 1.393$^{***}$ | 1.443$^{***}$ | 1.393$^{***}$ | 1.381$^{***}$ |
| *Unemployed* | | | 1.571$^{***}$ | 1.614$^{***}$ | 1.544$^{***}$ | 1.511$^{***}$ |
| *Education* | | | | | | |
| *Lower* | | | | Ref. | Ref. | Ref. |
| *Intermediate* | | | | 1.049$^{NS}$ | 1.063$^{NS}$ | 1.053$^{NS}$ |
| *Higher* | | | | 1.371$^{***}$ | 1.398$^{***}$ | 1.386$^{***}$ |
| *Possession of an automobile* | | | | | 0.701$^{***}$ | 0.736$^{***}$ |
| *Single-mother household* | | | | | | 1.185$^{***}$ |

$^{***}$ significant at a 1% level; ** significant at a 5% level; * significant at a 10% level; NS not significant at a 10% level. Number of observations: 27,044.

Table 3: Logistic regression of unemployment probability (odds ratios)

Regarding parental characteristics, models III to VI show that the unemployment probability of a young adult is higher when the household head (or spouse) is not participating in the labour-force or is unemployed than when he/she is employed. This effect is highly significant and is consistent with social network theories (at the household level, unemployed parents being little able to help their job-seeking children) and socialisation considerations (unemployed parents failing to provide their children with an image of social success to whom they could identify; Wilson, 1987). Living in a single-mother household (Model VI) also significantly increases the likelihood that a young adult be unemployed, suggesting that these households are more frequently prone to social problems detrimental to finding a job. Living in a household which does not own a car (an indirect measure of lack of financial resources) also significantly increases the unemployment probability, suggesting it is more difficult for individuals from poorer households to find a job. Surprisingly, the effect of parental educational level is not always significant, and when significant, seems counter-intuitive: all other things being equal, having a parent with a higher level of education increases the unemployment propensity of young adults. This counter-intuitive effect illustrates one shortcoming of the sample restriction method used to solve endogeneity. As already mentioned, restricting the sample to young adults living with parents may create a sample selection bias in the sense that young adults obtaining a job are more likely to leave parental home than young adults still unemployed. Indeed, if having parents with a high socioeconomic status (as reflected by parental educational level) in fact increases the chances to find a well-paid job, it is likely that young adults originating from these families will move out of their parents' dwelling more rapidly. This would leave an over-representation of unemployed young adults among families with higher parental socioeconomic status. Another possible explanation is that children from more educated (and richer) families do not feel pressured to intensively search for a job in order to move out of unemployment if they get financial support from their parents. In the absence of longitudinal data in which young adults are followed after they move out of parental home, it is not possible to distinguish between these two potential explanations.

By comparing the parameters and significance levels of the neighbourhood types across the different models, one can assess the sensitivity of the estimated neighbourhood effects to the inclusion of a more comprehensive set of parental controls. Table 3 shows that although the inclusion of parental and household characteristics significantly increases the fit of the model (see the likelihood ratios), the estimated neighbourhood effects change little (ranging from 1.491 in Model II to 1.418 in Model VI) and all parameters remain significant at a 1% level.

To provide a more intuitive interpretation, marginal effects of neighbourhood type were computed for Models II, III and VI. These give the change in predicted probability of unemployment associated with a change of neighbourhood type on the average individual of our sample. Results indicate that living in a deprived neighbourhood in comparison with non-deprived neighbourhoods, increases the unemployment probability of the average young adult by 7.3 percentage points in Model II, 6.3 points in Model III and 6.3 points in Model VI. Thus, adding an extensive set of parental characteristics to a model including only individual characteristics makes the estimated neighbourhood effect fall by 1 point. Moreover, estimating neighbourhood effects including only parental employment status (Model III) gives the same result as in the more comprehensive specification (Model VI). By way of comparison, the observed unemployment rate for young adults living with parents is 31% in deprived neighbourhoods and 19% in the rest of the agglomeration. The estimated marginal effect for Model VI indicates that 6.3 points of this gap (i.e. approximately 50%) are due to neighbourhood effects. The remaining would be due to the sorting of individuals with similar personal and parental characteristics into neighbourhoods.

**Sensitivity to unobserved covariates**

After having evaluated the influence of observed parental characteristics on estimated neighbourhood effects, it may be useful to test the sensitivity of the results to the endogeneity bias which results from the omission of an unobserved parental covariate which is correlated to both the probability of unemployment among young adults and parental residential choice. The approach used here is based on the method developed by Rosenbaum & Rubin (1983) and recently applied by Harding (2003) in the context of neighbourhood effects. The goal of this analysis is to assess how an unobserved binary covariate which affects both the probability of unemployment among young adults and the choice of parents to reside in a deprived or a non-deprived neighbourhood would alter the conclusions of this research about the magnitude and significance of neighbourhood effects. This is done by generating a series of unobserved binary variables *U* that vary according to their degree of association with the neighbourhood dummy variable *DN* and with the binary outcome measure *Y*. The degrees of association between *U* and *Y* and between *U* and *DN* are measured by parameters *k* and *l*, both expressed in terms of odds ratios.

In practice, the method implemented consists in generating a binary variable *U* sampled according to the following logistic model:

$$\mathrm{Log}\left(\frac{\mathrm{Prob}(U_i=1)}{1-\mathrm{Prob}(U_i=1)}\right)=\alpha+\kappa y_i+\lambda DN_i \qquad (12)$$

where Prob($U_i$=1) is the probability that the unobserved variable *U* takes a value 1 for individual *i*, $y_i$ is a binary variable indicating whether *i* is unemployed or not and $DN_i$ is a binary variable indicating whether *i* resides in a deprived neighbourhood or not. In this formulation, $\kappa$=log(*k*) and $\lambda$=log(*l*), with *k* and *l* being *chosen* odds ratios that measure the strength of the association that is *imposed* by the researcher between *U* and *Y* and *U* and *DN* respectively. $\alpha$ is determined so that the overall prevalence of $U_i$=1 is 0.5. More precisely, the previous equation and the three imposed constraints (on the two sensitivity parameters *k* and *l* and the overall prevalence of $U_i$=1) are used to determine the proportion of $U_i$=1 in each one of the four subgroups defined by *DN* and *Y*, i.e. $p_{11}$=P($U_i$=1|$DN_i$=1,$y_i$=1), $p_{10}$=P($U_i$=1|$DN_i$=1,$y_i$=0), $p_{01}$=P($U_i$=1|$DN_i$=0,$y_i$=1) and $p_{00}$=P($U_i$=1|$DN_i$=0,$y_i$=0). A random variable *U* is then generated in each one of these subgroups, following a Bernoulli distribution of parameter $p_{11}$ for subgroup ($ND_i$=1,$y_i$=1), $p_{10}$ for subgroup ($ND_i$=1,$y_i$=0), etc. Once each individual has received a value for *U*, new estimates of neighbourhood effects are obtained by including *U* in the unemployment probability model given by (11). This is repeated for increasing values of the sensitivity parameters *k* and *l* in order to investigate what level of endogeneity bias (i.e. the strength of the association between *U* and *Y* and between *U* and *DN*) would be needed to invalidate the results and render the estimated neighbourhood effects not significant.

Table 4 presents the estimated odds ratios associated with living in a deprived neighbourhood on unemployment probability, for two different models: (*i*) the model including only individual characteristics (and neighbourhood type) and no parental characteristics (as in Table 3's Model II) and (*ii*) the model including the full set of parental controls (as in Table 3's Model VI). In each panel of Table 4, the odds ratio in the extreme top-left cell (corresponding to *k* and *l* equals to one) gives the baseline odds ratio associated with living in a deprived neighbourhood, without introducing any amount of selection on unobservables. This odds ratio is 1.491 in Model II and 1.418 in Model VI. For each one of these models, an artificially created binary variable *U* is included and a sensitivity matrix is obtained by varying the sensitivity parameters *k* and *l*, which measure the associations of the unobserved parental characteristic *U* with the neighbourhood type and the employment status.

As expected, Table 4 shows that in both specifications, the estimated neighbourhood effect decreases as the values of *k* and *l* increase. This means that accounting for a previously omitted variable correlated both with the neighbourhood type and with the employment status does indeed reduce the intensity of neighbourhood effects estimates. However, the effect seems fairly robust since the values of *k* and *l* would have to be very high to make the neighbourhood effect not significant. Indeed, in the model including no parental or household characteristic (Model II), an unobserved covariate which multiplies the odds of living in a deprived neighbourhood by 3.5 and the odds of being unemployed by 4.0 would be required to totally erase the neighbourhood effect. This is also true when the full set of observed parental characteristics is included in the analysis (Model VI).

In this case, the neighbourhood effect becomes insignificant when at least one of the two odds ratios reaches 3.5 and the other one equals 3.0.

| ***Model II*** | Sensitivity parameter *k* | | | | | | |
|---|---|---|---|---|---|---|---|
| Sensitivity parameter *l* | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |
| 1.0 | 1.491$^{***}$ | 1.485$^{***}$ | 1.488$^{***}$ | 1.482$^{***}$ | 1.494$^{***}$ | 1.503$^{***}$ | 1.490$^{***}$ |
| 1.5 | 1.496$^{***}$ | 1.437$^{***}$ | 1.391$^{***}$ | 1.366$^{***}$ | 1.352$^{***}$ | 1.308$^{***}$ | 1.319$^{***}$ |
| 2.0 | 1.491$^{***}$ | 1.397$^{***}$ | 1.327$^{***}$ | 1.301$^{***}$ | 1.250$^{***}$ | 1.217$^{***}$ | 1.187$^{***}$ |
| 2.5 | 1.494$^{***}$ | 1.380$^{***}$ | 1.274$^{***}$ | 1.211$^{***}$ | 1.200$^{***}$ | 1.164$^{***}$ | 1.123$^{***}$ |
| 3.0 | 1.501$^{***}$ | 1.365$^{***}$ | 1.263$^{***}$ | 1.184$^{***}$ | 1.137$^{***}$ | 1.112$^{***}$ | 1.070$^{**}$ |
| 3.5 | 1.506$^{***}$ | 1.333$^{***}$ | 1.229$^{***}$ | 1.135$^{***}$ | 1.093$^{***}$ | 1.062$^{*}$ | 1.032$^{NS}$ |
| 4.0 | 1.514$^{***}$ | 1.320$^{***}$ | 1.205$^{***}$ | 1.150$^{***}$ | 1.073$^{**}$ | 1.025$^{NS}$ | 0.992$^{NS}$ |

| ***Model VI*** | Sensitivity parameter *k* | | | | | | |
|---|---|---|---|---|---|---|---|
| Sensitivity parameter *l* | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |
| 1.0 | 1.418$^{***}$ | 1.411$^{***}$ | 1.418$^{***}$ | 1.412$^{***}$ | 1.420$^{***}$ | 1.433$^{***}$ | 1.417$^{***}$ |
| 1.5 | 1.423$^{***}$ | 1.367$^{***}$ | 1.325$^{***}$ | 1.297$^{***}$ | 1.287$^{***}$ | 1.246$^{***}$ | 1.256$^{***}$ |
| 2.0 | 1.417$^{***}$ | 1.328$^{***}$ | 1.265$^{***}$ | 1.237$^{***}$ | 1.187$^{***}$ | 1.163$^{***}$ | 1.132$^{***}$ |
| 2.5 | 1.420$^{***}$ | 1.309$^{***}$ | 1.212$^{***}$ | 1.155$^{***}$ | 1.143$^{***}$ | 1.106$^{***}$ | 1.069$^{*}$ |
| 3.0 | 1.431$^{***}$ | 1.300$^{***}$ | 1.198$^{***}$ | 1.126$^{***}$ | 1.119$^{***}$ | 1.085$^{**}$ | 1.058$^{NS}$ |
| 3.5 | 1.431$^{***}$ | 1.265$^{***}$ | 1.170$^{***}$ | 1.077$^{**}$ | 1.040$^{NS}$ | 1.011$^{NS}$ | 0.982$^{NS}$ |
| 4.0 | 1.438$^{***}$ | 1.256$^{***}$ | 1.147$^{***}$ | 1.093$^{**}$ | 1.018$^{NS}$ | 0.973$^{NS}$ | 0.946$^{NS}$ |

$^{***}$ significant at a 1% level; $^{**}$ significant at a 5% level; $^{*}$ significant at a 10% level; $^{NS}$ not significant at a 10% level. Number of observations: 27,044. Sensitivity parameters *k* and *l* are expressed in terms of odds ratios and measure the effect of an artificially created binary variable *U* on the probability of unemployment and the probability of living in a deprived neighbourhood respectively.

Table 4: Sensitivity of neighbourhood effect estimates to the presence of an unobserved covariate (odds ratios)

In order to provide a more substantive interpretation, one may compare values for *k* and *l* (i.e. the amount of selection on unobservables that would be needed to make neighbourhood effects disappear) to the odds ratios estimated on observed covariates. Let's assume for example that the unobserved variable *U* is parental involvement with their children (this is indeed the most-often mentioned source of bias in neighbourhood effects studies; see O'Regan & Quigley, 1996; Harding, 2003). Though parental involvement is not measurable using standard census surveys, it is likely to affect both their children labour-market outcomes and their residential location. Indeed, whatever their income, social status or educational level, more involved parents will try to help their children as much as they can, by offering them support and advices in their job search effort or by monitoring their peer relations. At the same time, these involved parents may anticipate the detrimental effects of the social environment on their children and make some financial sacrifices to afford housing outside distressed areas. The question asked by this sensitivity analysis is then: “How big would the effect of parental involvement need to be to completely wipe out the effect of neighbourhood?”

The sensitivity analysis shows that having involved parents would need to increase the odds of being unemployed by 4.0 in order to totally erase the neighbourhood effect in Model II (or 3.0 in Model VI). As can be seen from Table 3, none of the observed parental and household characteristics already included in the model results in odds ratios as high as 4.0 or even 3.0, the highest odds ratio for a parental characteristics being 1.571 in Model VI (for having an unemployed parent). Among individual characteristics, the highest odds ratio is 2.374 for North-African or Turkish citizenship, and it is still below the 3.0 limit obtained by the sensitivity analysis. In other words, for the neighbourhood effect to become insignificant when considering parental involvement, the effect of this unobserved characteristic on unemployment would have to be stronger than that of parental employment status or that of citizenship, which seems not realistic. This provides some relatively strong evidence on the robustness of estimated neighbourhood effects.

The statistical analyses conducted here thus confirm that residential location does influence the labour-market outcomes of young adults residing with their parents in the Brussels agglomeration. Living in one of the deprived neighbourhoods of Brussels, i.e. characterised by the worst combination of social characteristics, significantly increases the unemployment probability of young adults. Although it may be feared that focusing on young adults residing with their parents creates some sample selection problems, the sensitivity analysis developed here suggests that estimated neighbourhood effects are robust in the presence of both observed and unobserved parental covariates. Indeed, the amount of selection on unobservables would have to be unreasonably high to make the estimated neighbourhood effect not significant.

## CONCLUSION

The objective of this chapter was to investigate the endogeneity issue, which is one of the prominent problems encountered in neighbourhood effects studies. Indeed, despite the huge amount of empirical studies, there is still considerable debate about the existence and magnitude of neighbourhood effects (Ellen & Turner, 1997). This is mainly because empirical studies are subject to several methodological problems (Dietz, 2002, Durlauf, 2004), in particular to an endogeneity bias arising from the fact that individuals are not randomly distributed into the urban area but instead “self-select” into neighbourhoods on the basis of their personal characteristics. In other words, some individual characteristics that influence individual outcomes (for example their employment status) also influence their residential choice. This means that what the researcher perceives as an effect of neighbourhood on individual outcomes may simply stem from a correlated effect reflecting a common residential choice.

This chapter contributes to the literature on neighbourhood effects and endogeneity issues by firstly reviewing the methods that have been proposed in the literature to try to solve the endogeneity issue. Secondly this chapter shows (on the basis of one empirical example) how to cope with endogeneity and how to evaluate endogeneity biases in neighbourhood effects estimates. To this end, the effect of living in a deprived neighbourhood on the unemployment probability of young adults residing in Brussels is estimated by means of logistic regressions. The endogeneity of residential choices is addressed by means of the sample restriction method, which consists in restricting the sample to young adults residing with their parents. This is the simplest and most often used method in the literature (e.g. O’Regan & Quigley, 1996). It is based on the argument that residential choices have been made previously by the parents and can thus be considered as fairly exogenous to the employment status of their children. However, it is an imperfect solution as unobserved parental characteristics may still influence both the residential choice of parents and the employment status of their children (Ihlanfeldt & Sjoquist, 1998). In this context, the methodological originality of this chapter is to evaluate the potential remaining endogeneity biases by conducting a sensitivity analysis to the presence of both observed and unobserved parental characteristics.

The results of this empirical application mainly showed that living in a deprived neighbourhood significantly increases the unemployment probability of young adults in Brussels, which confirm previous findings on 1991 data (Dujardin et al., 2008). This result is robust in the presence of both observed and unobserved parental covariates. Indeed, neighbourhood effects remain statistically significant when an extensive set of parental controls is introduced in the regression. Moreover, the sensitivity analysis based on artificially created unobserved covariates developed for this analysis shows that the amount of selection on unobservables would have to be unreasonably high to render neighbourhood effects estimates not significant.

However, results also suggest that, while it seems that it is enough to remove much endogeneity biases, the sample restriction method presents other shortcomings. Indeed, besides the fact that it does not allow generalising the results to other age groups, restricting the sample to young adults residing with parents is likely to generate sample selection problems (for example, if the leaving rate of young adults once they have found a job differs according to the characteristics of parents and/or according to the type of neighbourhood they live in). Some counter-intuitive findings of the statistical analyses conducted in this chapter tend to indicate that this is probably the case in Brussels. In this context, it seems important that future research on neighbourhood effects focus on exploring other ways to solve the endogeneity issues. As suggested by Durlauf (2004), the solution to endogeneity will probably have to be searched in the simultaneous modelling of neighbourhood effects and neighbourhood choices, through structural models. This will require a better understanding of residential choices as well as feedback mechanisms between individual outcomes and the characteristics of the social environment.

## ACKNOWLEDGMENTS

Claire Dujardin ackowledges the F.R.S.-F.N.R.S. (Fonds National de la Recherche Scientifique) for having financed her doctoral and postdoctoral research. This work is also associated with the ATLAS (Housing) project financed by the Belgian Federal Science Policy.

## INFORMATIONS

The sensitivity analysis to unobserved covariates developed in this chapter makes use of the *Sensuc* function, which is part of the Design S library written by F. Harrell. Documentation and programs can be found at the following web links:

http://lib.stat.cmu.edu/S/Harrell/Design.html (general introduction to the Design library)
http://lib.stat.cmu.edu/S/Harrell/help/Design/html/sensuc.html (specific documentation on the Sensuc function)

## REFERENCES

**Aaronson, D.,** 1998, Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. In: Journal of Human Resources, 33, 4: 915-946

**Benabou, R.,** 1993, Workings of a city: location, education and production. In: Quarterly Journal of Economics, 108, 3: 619-652

**Blakely, T.; Subramanian, S.V.,** 2006, Multilevel studies. In: Oakes, M.; Kaufman, J. (Eds), Methods in Social Epidemiology. San Francisco: Jossey Bass: 316-340

**Blume, L.E.; Durlauf, S.N.,** 2006, Identifying social interactions: a review. In: Oakes, M.; Kaufman, J. (Eds), Methods in Social Epidemiology. San Francisco: Jossey Bass: 287-315

**Brock, W.A.; Durlauf, S.N.,** 2001, Interaction-based models. In: Heckman, J.J.; Leamer, E.E. (Eds), Handbook of Econometrics, 5. Amsterdam: North-Holland: 3297-3380

**Brulard, T.; Van der Haegen, H.,** 1972, La division des communes belges en secteurs statistiques. Le point de vue des géographes. In: Acta Geographica Lovaniencia, 10: 21-36

**Buck, N.,** 2001, Identifying neighborhood effects on social exclusion. In: Urban Studies, 38, 12: 2251-2275

**Case, A.C.; Katz, L.,** 1991, The company you keep: the effects of family and neighborhood on disadvantaged youths. In: NBER Working Paper N°3705

**Corcoran, M.; Gordon, R.; Laren, D.; Solon, G.,** 1992, The association between men's economic status and their family and community of origins. In: The Journal of Human Resources, 26, 4: 575-601

**Crane, J.,** 1991, The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. In: American Journal of Sociology, 96, 5: 1226-1259

**Cutler, D.M.; Glaeser, E.L.,** 1997, Are ghettos good or bad? In: Quarterly Journal of Economics, 112: 827-872

**Datcher, L.,** 1982, Effects of community and family background on achievement. In: The review of Economic and Statistics, 64, 1: 32-41

**Dietz, R.D.,** 2002, The estimation of neighborhood effects in the social sciences: an interdisciplinary approach. In: Social Science Research, 31, 4: 539-575

**Dujardin, C.,** 2006, The role of residential location on labour-market outcomes. The urban agglomerations of Lyon and Brussels, PhD Thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium

**Dujardin, C.; Selod, H.; Thomas, I.,** 2004, Le chômage dans l'agglomération bruxelloise: une explication par la structure urbaine. In: Revue d'Economie Régionale Urbaine, 1: 3-28

**Dujardin, C.; Selod, H.; Thomas, I.,** 2008, Residential segregation and unemployment: the case of Brussels. In: Urban Studies, 45, 1: 89-113

**Duncan, G.J.; Aber, J.L.,** 1997, Neighborhood models and measures. In: Brooks-Gunn, J. et al (Eds), Neighborhood Poverty. Volume 1: Context and Consequences for Children. Russel Sage Foundation: New York: 62-78

**Duncan, G.J.; Connell, J.P.; Klebanov, P.K.,** 1997, Conceptual and methodological issues in estimating causal effects of neighborhoods and family conditions on individual development. In: Brooks-Gunn, J. et al. (Eds), Neighborhood Poverty. Volume 1: Context and Consequences for Children. Russel Sage Foundation: New York: 219-249

**Durlauf, S.N.,** 2004, Neighborhood effects. In: Henderson, J.V.; Thisse, J.-F. (Eds), Handbook of Regional and Urban Economics, Volume 4. Amsterdam: North-Holland: 2173-2242

**Ellen, I.G.; Turner, M.A.,** 1997, Does neighborhood matter? Assessing recent evidence. Housing Policy Debate, 8, 4: 833-866

**Evans, W.N.; Oates, W.E.; Schwab, R.M.,** 1992, Measuring peer group effects: a study of teenage behavior. In: Journal of Public Economics, 100, 5: 966-991

**Fieldhouse, E.A.; Gould, M.I.,** 1998, Ethnic minority unemployment and local labour market conditions in Great Britain. In: Environment and Planning A, 30, 5: 833-853

**Ginther, D.; Haveman, R.; Wolfe, B.,** 2000, Neighborhood attributes as determinants of children’s outcomes. In: The Journal of Human Resources, 35, 4: 603-642

**Glaeser, E.L.,** 1996, Spatial effects upon unemployment outcomes: the case of New Jersey teenagers. Discussion. In: New England Economic Review, May/June: 58-64

**Gobillon, L.; Selod, H.; Zenou, Y.,** 2007, The mechanisms of spatial mismatch. In: Urban Studies, 44, 12: 2401-2427

**Goldstein, H.,** 2003, Multilevel Statistical Models. Edward Arnold: London

**Greene, W.H.,** 2008, Econometric Analysis. Sixth Edition. Prentice Hall: London

**Harding, D.J.,** 2003, Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. In: American Journal of Sociology, 109, 3: 676-719

**Holzer, H.,** 1988, Search method used by unemployed youth. In: Journal of Labor Economics, 6, 1: 1-20.

**Ihlanfeldt, K.R.; Sjoquist, D.L.,** 1998, The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. In: Housing Policy Debate, 9, 4: 849-892

**Ioannides, Y.M.; Zabel, J.E.,** 2008, Interactions, neighborhood selection and housing demand. In: Journal of Urban Economics, 63; 229-252

**Jencks, C.; Mayer, S.E.**, 1990, The social consequences of growing up in a poor neighbourhood. In: Lynn, L.E.; McGeary, M.G.H. (Eds), Inner-city Poverty in the United States. Washington D.C.: National Academy Press: 111-186

**Johnston, R.; Jones, K.; Burgess, S.; Propper, R.; Sarker, R.; Bolster, A.**, 2004, Scale, factor analyses, and neighborhood effects. In: Geographical Analysis, 36, 4: 350-368

**Katz, L.F.; Kling, J.R.; Liebman, J.B.,** 2001, Moving to Opportunity in Boston: early results of a randomized mobility experiment. In: Quarterly Journal of Economics, 116, 2: 607-654

**Kling, J.R.; Liebman, J.B.; Katz, L.F.,** 2007, Experimental analysis of neighborhood effets. In: Econometrica, 75, 1, 83-119

**Manski, C.F.,** 1993, Identification of endogenous social effects: the reflection problem. In: Review of Economic Studies, 60: 531-543

**Manski, C.F.,** 1995, Identification Problems in the Social Sciences. Harvard University Press: Chicago

**Oakes, J.M.,** 2004, The (mis)estimation of neighbourhood effects: causal inference for a practical social epidemiology. In: Social Science and Medicine, 58, 10: 1929-1952

**O'Regan, K.M.; Quigley, J.M.,** 1996, effects upon unemployment outcomes: the case of New Jersey teenagers. In: New England Economic Review, May/June: 41-58

**Oreopoulos, P.,** 2003, The long-run consequences of living in a poor neighborhood. In: Quarterly Journal of Economics, 118, 4: 1533-1575

**Osterman, P.,** 1991, Welfare participation in a full employment economy: the impact of neighborhood. In: Social Problems, 32, 4: 475-491

**Plotnick, R.D.; Hoffman, S.D.,** 1999, The effect of neighborhood characteristics on young adult outcomes: alternative estimates. In: Social Science Quarterly, 80, 1: 1-18

**Reingold, D.,** 1999, Social networks and the employment problem of the urban poor. In: Urban Studies, 35, 11: 1907-1932

**Robinson, W.S.,** 1950, Ecological correlations and the behavior of individuals. In: American Sociological Review, 15: 351-357

**Rosenbaum, P.R.; Rubin, D.B.,** 1983, Assessing the sensitivity to an unobserved binary covariate in an observational study with binary outcome. In: Journal of the Royal Statistical Society B, 45, 2: 212-218

**Sampson, R.J.; Morenoff, J.D.; Gannon-Rowley, T.,** 2002, Assessing neighborhood effects: social processes and new directions in research. In: Annual Review of Sociology, 28: 443-478

**Weinberg, B.A.; Reagan, P.B.; Yankow, J.J.,** 2004, Do neighborhoods affect work behaviour? Evidence from the NLSY79. In: Journal of Labor Economics, 22, 4: 891-924

**Wilson, W.J.,** 1987, The Truly Disadvantaged: the Inner City, the Underclass, and Public Policy. The University of Chicago Press: Chicago

**Wong, D.W.S.,** 2004, Comparing traditional and spatial segregation measures: a spatial scale perspective. In: Urban Geography, 25, 1: 66-82

**Wrigley, N.,** 1995, Revisiting the MAUP and the ecological fallacy. In: Cliff, A.D. et al. (Eds), Diffusing  Geography Essays for Peter Haggett. The Institute of British Geographers, Special Publications Series

**Zenou, Y.; Boccard, N.,** 2000, Labor discrimination and redlining in cities. In: Journal of Urban Economics, 48, 2: 260-285

## AUTHORS INFORMATION

**Claire DUJARDIN**
claire.dujardin@uclouvain.be
F.R.S.-F.N.R.S.
Center for Operations Research and Econometrics (CORE) and Department of Geography, Université catholique de Louvain (UCL), B-1348 Louvain-la-Neuve, Belgium

**Dominique PEETERS**
dominique.peeters@uclouvain.be
Center for Operations Research and Econometrics (CORE) and Department of Geography, Université catholique de Louvain (UCL), B-1348 Louvain-la-Neuve, Belgium

**Isabelle THOMAS**
isabelle.thomas@uclouvain.be
F.R.S.-F.N.R.S.
Center for Operations Research and Econometrics (CORE) and Department of Geography, Université catholique de Louvain (UCL), B-1348 Louvain-la-Neuve, Belgium

# GIS IN HEALTH AND SOCIAL CARE PLANNING

**Ronan FOLEY [*,**], Martin C. CHARLTON [*] and A. Stewart FOTHERINGHAM [*]**

[*] National Centre for GeoComputation, NUI Maynooth, Ireland, [**]Department of Geography, NUI Maynooth, Ireland

## ABSTRACT

The application of GIS and GI science approaches within health and social care has been an established area for research for over twenty years. The chapter identifies core theoretical concerns with topics such as: supply, demand, need and choice in health care, and examines some of the ways in which these notions have been modelled quantitatively in applied settings. In addition, summaries are provided of previous research in the sub-themes of accessibility & utilisation, health inequalities, location-allocation modelling, epidemiology, service planning and health informatics. The second part of the chapter examines three research case studies carried out by the authors in the UK and Ireland around the areas of: a) cross-border hospital accessibility, b) geographically-weighted regression modelling of illness data and, c) the planning of social care services. The final section identifies some future directions for work under the wider headings of spatial data, analysis and visualisation.

## KEYWORDS

Health care, Social care, Modelling, Spatial data, Visualisation

## INTRODUCTION

Given the inherent spatiality of health and the organisation of structures within society to manage health, it is no surprise that geographers have developed an interest in the subject in a variety of ways. Driven primarily by relationships between health, place and space, geographers and other subject specialists from computing, medicine, mathematics and statistics have researched spatial modelling aspects of the management of health care (Gatrell & Löytönen, 1998; McClafferty, 2003). Prior to the 1980s much of this exploration was carried out using quantitative models but with improvements in computing power, new forms of geocomputational modelling and analysis have been developed (Joseph & Phillips, 1984; Gatrell & Senior, 1999). As an analytical tool, GIS have been at the core of those explorations, while the increasingly common term, GI Science also frames that analysis as a process and a paradigm (Wilson & Fotheringham, 2008). One focus of this chapter will be on exploring the quantitative and theoretical aspects of how health care is modelled and theorised in spatial analytical terms. This is also linked to an improved reputation within health for the potential of spatial approaches as GIS becomes increasingly embedded in health care systems (Croner & Sperling, 1996; Gatrell & Senior, 1999; Melnick, 2002; McLafferty, 2003; Smith, Higgs & Gould, 2005). At the same time social scientists have also become interested in other aspects of the organisation of health in society and the more qualitative elements that shape that structure. These more theoretical and applied aspects have in part led analysts to consider some of the more formative aspects of health care, including social care. They also suggest the need to link people more fully into the systems to provide more effective modelling (Couclelis, 2003; Elwood, 2006).

The core concerns for these particular branches of theoretical and quantitative geographies have strong foundations within economic, biomedical and to a lesser extent, social geographies. As state systems in all these areas developed after the First World War and in to the 1960s, the quantitative revolution in the subject became a strong driving force for early research on these topics (Hubbard et al., 2002). Well-known examples of the kinds of quantitative models developed would include Christaller’s Central Place Theory and Hägerstrand’s early models of diffusion (Dicken & Lloyd, 1990). The former theoretically modelled the hierarchical nature of urban networks and their spatial organisation while the latter was interested in how spatial phenomena varied and developed over time and across space. Both were applied to an extent within medical geography and indeed acted as markers for the main initial divisions of interest in the subject, namely epidemiology and health care planning (Joseph & Phillips, 1984; Cliff & Haggett, 1988). Much of the early modelling, while theoretically very sound, was also partially hamstrung by two core barriers: access to detailed spatial data and the lack of processing power. Despite this, there was some effective cartographic research in the area of mapping disease, using examples such as measles, influenza and later, HIV/AIDS (Cliff, Haggett & Ord, 1986; Cliff & Haggett, 1988). With the advent of desktop computing power and improved access to electronic data records, these problems were partially tackled in the 1970s and early 1980s. A third issue related to how these data could be effectively visualised so as to bring out the inherently spatial aspects of the modelling. Traditional cartographic approaches were applied in the medical/health arena and as a result there was a reasonably strong tradition of medical mapping including some good historical, if contested examples from the work of John Snow and William Booth around cholera and poverty and Mourant on blood group distributions (Mourant, 1954; Cliff & Haggett, 1988; Koch & Dennike, 2006). How these three main elements, spatial data, spatial analysis and spatial visualisation, remain at the heart of all subsequent work in this field will be explored throughout the chapter and will form its primary aim (Fotheringham, Brunsdon & Charlton, 2000).

As a subject, medical geography emerged primarily from Germany (as *GeoMedizin*) and Russia in the 1940s and 1950s. In turn taken up in the English-speaking world, especially in the US, UK and Canada, the subject developed along two broad lines from the 1970s onwards. Focused firmly on health care planning, a concern with modelling effective spatial structures for service planning was to lead into much of the early quantitative work around modelling optimal locations of health services as well as looking into the equitable distribution of those services (Jones & Moon, 1987). The other main strand within medical geography related to epidemiology, the distribution and management of disease, driven by ecological theory and using diffusion modelling as one effective way of mapping and predicting the spread of disease (Gatrell, 2002). While this latter strand will be discussed briefly in this chapter, the focus is very much on the planning and service aspects of the subject. The two are clearly linked in that services are put in place to combat disease and the location and supply of services are in some ways determined by the demand created by the location and spread of disease. In addition, social care, in the form of support services for the poor, the elderly and people with disabilities, also has a strong service dimension which requires planning for both the delivery of those services and their integration with health care. Nonetheless, current understandings of health and social care planning spread well beyond the curative aspect of the subject. Indeed it could be argued that a strong focus on the planning aspects, in terms of improvements in access, social supports, preventative health and health promotion initiatives, are key steps in reducing disease in the first place and require more attention (Curtis, 2004).

The rest of this chapter will consist of two main sections. The first one will provide a critical literature review which looks at how theoretical and quantitative geography perspectives have been developed and used within health and social care planning. It will look at the range of theories, problems, methodologies and models developed around the broad themes of data, analysis and visualisation and seek to critically examine where the subject situates itself in late 2008. The second half of the chapter will summarise some examples of work that the authors have been engaged in for the past decade and will be drawn from the UK and Ireland to demonstrate how those themes have been addressed. A final section will look at where the future may lie and suggest some ways in which the attention of quantitative and theoretical geographers and computer scientists can be effectively directed into the future of health and social care planning. At all times, the work will have a very strong emphasis on services and their associated planning.

## RECENT ADVANCES IN HEALTH AND SOCIAL CARE PLANNING

### Introduction

Before embarking on a survey of work in the areas of health and social care planning, it is helpful to briefly list some of the core theoretical issues for researchers working in this area and identify the kinds of problems researchers have been grappling with over the past three decades. This is helpful in framing how subsequent quantitative approaches have sought to tackle these problems. The discussion will focus on empirical issues and not on the underlying political and ontological ideas which shape health care systems in the first place, though any meaningful study of how health and social care planning functions is shaped by how those services are organised within any single country. Nevertheless the same core issues commonly emerge no matter what type of health care system operates. In a summary document on the then current themes in the US, McLafferty (2003) identified four headings around which GIS and Health Care more generally might be studied, namely: need, access, utilisation and the evaluation and planning of health services. This provides a useful starting point for this work as well.

At the core of any health and social care planning discussions are the simple theoretical constructs of supply and demand. In a perfect system, supply would match demand. For geographers, this would be represented spatially by the correct placing of services in urban and rural settings to exactly match demand in those areas. It is probably fair to say that such a system does not exist anywhere in the world. Indeed in many countries, especially in the global South, systems barely exist and huge demands swamp limited supplies (Meade & Earickson, 2000; Curtis, 2004). Even in the more developed global North, similar inequalities exist, albeit on nothing like the same scale. Where those mismatches occur and at what scale remains a primary interest for geographers and identifies a role for spatial approaches in health and social care planning. In addition, the ways in which supply and demand operate in the real world are multiple and complex. One might consider a number of additional factors which add to the theoretical difficulties spatial planners face, namely equity or effectiveness, need and time (Rushton, 1998; Goddard & Smith, 2001; Guagliardo, 2004; Yuan, 2008). Equity within health care systems relates to notions of fairness and equal supply to meet equal demand. But this is skewed so that supply is often structured unfairly whereby unequal supplies operate in some areas when measured against need. Some examples of this might be the over-supply of general practitioners in urban areas as opposed to rural areas, a phenomenon observed in the US, China, France and the UK (Vigneron, 1997; Gatrell, 2002). Another example which is particularly relevant to geography is the notion of the Inverse Care Law, where proximity to a supply seems to be a more marked explanatory factor than need-based demand (Curtis, 2004). A further complicating factor is the linked concept of effectiveness (often mentioned in conjunction with efficiency), whereby health and social care planners need to make some judgements on how best to provide supply to meet what they perceive as realistic demand while at the same time spending limited funding in as efficient a way as possible (Oliver & Mossialos, 2004). Clearly there are conflicts between planning and the operation of equity or effectiveness based systems. Many of these are outside the control of theoretical and quantitative modelling but it does suggest that such models should be flexible and complex enough to take these structural elements into account.

Linked very strongly to the term demand is the notion of need. Demand is an all-encompassing term so that within any planning system, provision must be made for different forms of need. In most societies, the supply of services may have to be weighted in favour of those groups within society, namely, the very young, the elderly, the poor or the excluded, who are most likely to be in need. These considerations add considerably to the simple supply versus demand model but must be considered in any theoretical and quantitative model of service planning (Weiner & Harris, 2008). A final complicating factor is time. Clearly it is one thing matching supply to demand in a fixed time period but planning is very strongly associated with predictive capabilities. Any meaningful planning system should have the capacity and indeed aim, to try and model future supply and demand as well. This is often difficult and subject to many additional factors which are again, often outside the control of the planners or modellers (Curtis, 2004; Charlton, 2008). All three of these areas provide challenges, but also opportunities for health and social care planners and some of the ways in which these challenges have been met will be outlined in the rest of the literature review below.

Clearly for theoretical and quantitative geographers, the kinds of problems listed above provide a rich framework for their fertile imaginations. When need, demand, supply and additional elements such as choice and volume need to be expressed spatially, the use of geo-computational approaches seems simultaneously ideal and under-exploited. Plotting the relationships between these different concepts has been led by a number of statistical approaches which in turn have helped develop the medium of spatial statistics (Fotheringham, Brunsdon & Charlton, 2000). Regression and correlation have been used to look at and understand underlying relationships between expected and observed health patterns and derived some useful explanations from these. Some of the more planning-specific aspects are also driven by spatial theory especially around identifying optimal locations for services when weighted by existing patterns of supply and demand. In addition, the increasing technical power of computing has led to much more rapid and sophisticated modelling of different scenarios so that multiple choices can be modelled and best results identified. Within the geo-computational models, much of the research has been around improving the quality and precision of how that modelling is carried out. For example the development of improved estimations of kernels and associated aspects of spatial statistics has continued to drive work in this area (Fotheringham, Charlton & Brunsdon, 2000; Jacquez, 2008; Wilson & Fotheringham, 2008).

Mindful of the ontological and epistemological backgrounds and the technical concerns discussed above, the remainder of this section will explore recent work in a number of sub-themes of health and social care planning. Within each sub-theme there will be an implicit link to the core themes of data, analysis and visualisation. Accessibility and utilisation has been a common theme in health and social care planning which in turn has links with ways of studying and measuring health inequalities. Theoretical work around the areas of location-allocation and optimal location modelling is then examined followed by a more epidemiologically focused section. The following section discusses a summative theme, namely that of service planning and how policy is theorised, modelled and achieved. A final section focuses on what is an increasingly public and open interface between health geographers, medics, statisticians and the general public. With the increasing embeddedness of GIS in public life and the role of the internet in broadening both knowledge and expectations around health information, the potential for GI Science to take its potential into the public arena is considerable but must be approached with caution.

### Accessibility and Utilisation

While the subject of modelling accessibility to health care is more comprehensively and fully explored in a separate chapter in this book (Morrissey et al., 2009), it still has important links to other aspects of health and social care planning and will be briefly discussed here. One of the original core texts in the subject by Joseph & Phillips (1984), explicitly studied the twin terms accessibility and utilisation. In addition, any study of access to and utilisation of health care needs to be aware also of the core concepts referred to previously such as need, equity, supply and demand. A large number of authors have incorporated some of these elements into their spatial data requirements which in many ways are fundamental to the modelling of accessibility and utilisation. For example, geographies of supply are expressed through counts of the number of hospitals, their physical location and configuration and the relative size of those hospitals and numbers of services provided (Luo & Wang, 2003). Demand is often measured through utilisation but there are issues here in terms of how fully demand identifies need in a setting where waiting lists and staffing shortages are not unknown and where the structure of the system itself shapes utilisation rates. Need is also a complex term with a number of different definitions relating to expressed need in the form of presenting patients and unexpressed need within the wider population (Joseph & Phillips, 1984). Finally equity can be expressed in a number of ways, depending on whether one uses a vertical or horizontal definition (Goddard & Smith, 2001) or even whether one takes a measure based on a population or catchment based approach (Christie & Fone, 2003). The work of Khan & Bhardwaj (1994) is particularly useful in developing a fuller understanding of what they refer to as spatial and aspatial aspects of accessibility. The aspatial aspects include a wide and complex set of variables including income, education, social class, insurance and other social and economic factors, which affect how people access and utilise health care. They identify these as being separate but linked elements to the more purely spatial aspects of location, distance, time and supply which provide the other part of the equation. Together these provide a fuller integrated model. In addition, utilisation can often act as a confounding factor when combined with accessibility measures in that it sometimes merges measures of ‘supply’ with proxy measures of ‘demand’, especially for use in a health policy setting (Oliver & Mossialos, 2004).

The core traditional approach used by medical/health geographers has been to focus on a number of core spatial datasets and use these in the modelling of accessibility. Some of these approaches have been used before the widespread use of GIS and digital spatial data (Horner &

Taylor, 1979; Joseph & Phillips, 1984). The arrival of the latter has, however, allowed for more efficient and effective modelling using a number of new spatial analytical techniques. In general terms, modelling spatial accessibility has involved the following key spatial data. Firstly, the location and distribution of health care facilities often forms a first layer of information. While much of the research has focused on secondary and tertiary care, other services associated with primary care, community care and even voluntary services have also been modelled in this way (Bullen & Moon, 1994; Foley, 2002). A second core element is a layer that incorporates demographic data and the distributions of different populations. These function as proxies for demand/need and can be broken down into sub-populations depending on the services being modelled (Love & Lindqvist, 1995; Teljeur, Barry & Kelly, 2006). A final layer of information needed is information on the transportation network used to model the spatial linkages between patients or potential patients and services. This was traditionally modelled as Euclidian or straight-line distance which often enabled planners to quickly see buffers or catchments zones around hospitals and to visualise quickly those areas or groups who fell outside those areas (Kumar, 2004). Other visualisation approaches which have been enhanced by GIS include spider-graph approaches showing linkages between patient and service as a series of straight lines (Bullen & Moon, 1994; Campbell & Parker, 1998).

With the advent of GIS the ability to overlay and merge these three different layers within a single automated information system has been recognised as providing an important new evidence base for health care planning (Gatrell & Löytönen, 1998). Brabyn & Skelly (2002) took these core elements and combined them in a vector (linear) GIS to model access to public hospitals by travel time across New Zealand. They identified an effective accessibility score by area, weighted by population which also incorporated a locally relevant remoteness factor. Such a concern for local topography was also reflected by Lin et al. (2002) who studied access to hospitals in British Columbia in an area also characterised by mountains and a limited road network. They produced a tabular output by district which also incorporated existing utilisation data to weight the model.

A final example of this straightforward spatial modelling was provided by Christie & Fone (2003) who used the same methodologies to look at access to tertiary hospital services in Wales. In their example, they used GIS to help tabulate travel time by social groups (weighted by deprivation score) to look at equity of access to very specialist services, which unsurprisingly was biased towards the more affluent urban dweller, a finding consistent with much accessibility modelling work.

Some of the studies above used GIS to produce more robust forms of spatial modelling, by moving beyond Euclidian distance to include consideration of travel distance along road networks and even more usefully, travel time. Additional sophistication was provided by studies which looked at incorporating public as well as private transport into the models (Martin et al., 2002; Lovett et al., 2002; Jordan et al., 2004). These studies looked respectively at Cornwall and Norfolk and added public bus and train networks to the previously modelled road network to try and get a fuller picture of levels of access. Martin et al. (2002) also developed their model as a raster (surface) approach which is the other method feasible within a GIS and this was replicated by Luo & Wang (2003) who developed a model of access to primary care in the Chicago area based on a raster travel friction approach. They also used the power of the GIS to develop a gravity model approach which incorporated floating catchments weighted by location and population. This approach was also developed by Guagliardo (2003) in a similar study of access to primary care in Washington DC with the additional spatial modelling element of kernel density which effectively modelled the effect of service clustering and the subsequent impact on access. Other GIS approaches have focused on specific services within hospital provisions with interests in maternity and other specialist services (Beere & Brabyn, 2006; Schuurman et al., 2006).

Another valuable example is the work of Damiani et al. (2005) who used a surface model approach to measure access to acute hospitals which in turn was weighted by a measure of choice, as expressed through levels of available beds. In this way, a normal mapped result, which identified rural and remote areas as poorly served, was considerably modified when the impact of waiting times and available beds was fed in. In the modified version, areas around the South East and South West of England showed up as under-served, an interesting example of how demand and supply affect the original models. In all cases, the ability of GIS to effectively visualise accessibility in the form of vector time maps and raster accessibility surfaces have allowed for ready analysis and interpretation as both a pointer to and a time-saving short-cut for subsequent explanatory analysis. Finally the work of Robitaille et al. (2008) points to ways in which accessibility can be measured not just to services, but rather to its opposite, sites of health risk. In this case, the location of gambling machines in Montréal is used to model risk and vulnerability to health-endangering behaviours, within a GIS. Each small (dissemination) area was mapped with a gambling vulnerability score which in turn was spatially weighted using gravity model and LISA spatial autocorrelation methodologies, to identify locations for targeted health promotion interventions (Anselin, 1995).

**Equity and Health inequalities**

Apart from the more service-specific areas, much of the work of medical geographers and cartographers has also focused on the related issue of health inequalities. Some of the early cartographic work in the UK by the likes of Howe (1972) focused on mapping spatial variations in morbidity and mortality to draw attention to inequalities across regions and even within cities. This work was taken up by a number of early applied users in the late 1980s and early 1990s though it took till the late 1990s for it to become mainstream in the National Health Service (NHS) in the UK (Gordon & Womersley, 1997). In the US the use of GIS in visualising and analysing health inequalities was developed at an earlier stage principally through the work of Gerard Rushton, Chuck Croner and Robert Earickson (Gatrell & Senior, 1999; Rushton, 2000; McLafferty, 2003). In the same period European research, often with an epidemiological flavour was also developing strongly in areas such as water quality (Kistemann, Dangendorf & Exner 2001; Crabbe, Hamilton & Machin, 2004), and motor neurone disease (Sabel, 2000).

Quantitative measurements of health inequalities have taken a number of forms. There is a clear relationship between wealth and poverty and the different measures of health inequalities. Put simply, the rich traditionally have good indicators for health while the poor bear the burden of ill-health. A simple starting point for many geographers is to look at how these relationships are expressed spatially by mapping income against health indicators (Wilkinson et al., 1998; Driedger et al., 2007). For many countries, spatial data on income is hard to access. This is less of a problem in some Western European and Scandinavian countries but elsewhere the mapping of poverty and deprivation is more problematic. To get around the absence of direct measures of poverty and wealth, the development and creation of mathematically derived deprivation indices has been a common response. This was initiated in the UK with an explicit health focus to measure the extra income needed by general practitioners

(GPs) working in deprived areas but has since developed a much wider suite of applications in both health and social care planning (Jarman, 1984; Noble et al., 2004). Typically a series of relevant census indicators are chosen, standardised, weighted and then combined, often based on simple statistics but sometimes using principal components analysis to come up with a final score (Townsend & Davidson, 1982; Haase & Pratchske, 2005). Other innovative approaches include the use of associated multi-criteria analysis to develop the creation of deprivation scores (Bell, Schuurman & Hayes, 2007). These scores can then be applied as either an absolute or relative measure, with the latter being used more widely in policy interventions (Kelly & Teljeur, 2004; Philibert et al., 2007). There are a number of technical issues with the ways in which deprivation indices are created, especially as new versions have begun to combine global scores with domain specific subsets such as employment, access and health (Noble et al., 2004). Principal amongst those are the choice, type and relative weighting of indicators, technical issues with standardisation and the shrinkages applied in creating final scores and the geographic scale of their application within administrative units (Pringle et al., 2000). All of these are and have been contested, but despite the fact that there are ongoing issues with their construction, deprivation indices remain popular proxies of poverty. In turn they also provide a valuable denominator for the spatial modelling of health inequalities and geographical areas with high relative and absolute deprivation scores have been shown to have a persistent and strong association with indicators of poor health (Curtis, 2004). This is a common finding whether it is a study of the suburbs of Glasgow, the *banlieue* of Lyon or downtown Montreal. Other ways in which the associations between place and inequalities in health outcomes have been measured include the afore-mentioned standardised approaches, in this case through the production of single indicator measures. The relationships between ‘poor places’ and poor health is a strong one and GI Science provides an ability to measure morbidity and mortality in more sophisticated ways. Clearly to enable effective scientific comparison between places, there is a need to provide age and gender standardisation with the raw morbidity or mortality data (Bailey & Gatrell, 1995; Longley et al., 2005).

From an exploratory spatial data analysis perspective, other theoretical problems relate to the scale and size of the units for which comparisons are made. Detailed spatial units, whether wards in the UK, *Départements* in France, *Gemeinde* in Germany or mesh blocks in New Zealand, are rarely homogenous in either size or population. In addition rural units are typically larger, but with smaller populations, than their urban equivalents.

A number of statistical techniques have been applied to try and account for this spatial heterogeneity. Bayesian shrinkage approaches have been used to try and bring extreme values and outliers closer to the mean values (Langford, 1994; Johnson, 2004). A persistent issue in mapping health data to areal units is the modifiable areal unit problem (MAUP) whereby the distribution of observations may be aggregated to different areal units and thereby exhibit completely different patterns of density. A number of researchers have studied, and continue to study, this problem. One approach is to increase the levels of homogeneity in the basic building blocks and work in this area is being carried out by the National Centre for GeoComputation (NCG) in Ireland whereby a new geography of 'atomic small areas' has been designed to improve the homogeneity and level of detail of census units for the forthcoming 2011 Census. Other approaches return to the raw data and use a number of interpolation-based surface models to essentially bypass the MAUP problem through the creation of health and population surfaces (Openshaw & Taylor, 1981; Marceau, 1999; Shuttleworth & Lloyd, 2005). A final issue relates to comparative work between jurisdictions which may use different types of spatial units. In these cases a variety of approaches to unit modification in forms such as weighted aggregation and splitting are necessary to enable meaningful comparisons to be made (Gleeson et al., 2008).

The use of multi-level modelling has also developed apace in the last decade or so with an increased interest in the spatial variation of health-place relationships when examined at a number of different scales. This is not unlike concerns about using global relationships to model at a local level expressed in other statistical settings, but essentially the use of multi-level modelling allows analysts to look at regional or local variation against a national standard. One example was the work carried out in the north of the UK by Jones et al. (1997) which explored national patterns in the area of dental health and tried to see how these varied by local authority area. A further twist was to pick three areas where different models of flouridisation (none, artificial and natural) were practiced to see which provided the best outcomes. In the end, the best outcomes were achieved in the region where natural fluoridation was practiced. However in all cases, a strong positive linear relationship between levels of deprivation and poor dental health were observed. A final linked area, and one with a long tradition within medical & health geographies is the use of location quotients, whereby each local/regional measure is weighted by the national average to identify its relative performance against that global measure (Joseph & Phillips, 1984).

### Location-allocation and Optimal location modelliing

One theoretical problem which was particularly difficult to tackle in the pre-digital period, was that of location-allocation modelling and the associated area of optimal location modelling. Most health and social care planning systems need to plan the location of services in as efficient a way as possible. There is a spatial dimension to this problem in that decisions need to be made as to the locations of services. This decision is usually based on one of a number of different scenarios which can include: the location of a brand new service, the addition of a new service to existing services, the removal of a service or the multiple alteration of existing services. While it was possible to manually examine this problem, the constant drawing of different scenarios coupled with the mathematical calculations of summed distances and catchments meant it was slow work which often limited the number of different scenarios that could be modelled. With the arrival of GIS and geo-computational methods working in tandem with improved spatial data, the calculation of multiple scenarios became much quicker and the arrival at optimal locations for services based on available data was easier to achieve.

In addition some of the complexity of real-world choices could be built in to the system. Typically location-allocation models can factor in different location criteria, aimed at minimizing the aggregated distance travelled from home to health centre. The relative sizes of centres are important as well, as they are taken, in a very broad sense, as measures of attractiveness and the catchments and drawing power of larger centres can be factored in to the models as well (Fotheringham & Rogerson, 1994). Some examples of the kinds of work done is this area include the work of Hodgson in Goa, whereby the existing locations of clinics were remodelled and overlaid with the optimal locations given the distribution of the demand population (Bailey & Gatrell, 1995). What emerged was clear evidence of a need to shift the existing services from the over-supplied south to the under-supplied north. Oppong & Hodgson (1994) also carried out some local level planning in Ghana to observe the differing distributions between an existing and a modelled solution for rural health clinics in Northern Ghana. Indeed the application of location-allocation modelling may be particularly effective in mobile settings in the global South, as opposed to the more fixed health facility locations in the global North.

A third area where the modelling of optimal locations has been demonstrated to have power is in the area of system re-organisation, a common issue in most jurisdictions dependent on often fluid patterns of demand and supply. Clarke's work on the options and possible outcomes of the closure of a cardio-thoracic service in one hospital in Yorkshire and its spatial implications is a good example of such work (Martin, 1996). In this example, the knock-on effect of the re-distribution of service supply was modelled to examine where revised catchments would emerge and effectively modelled the new geographies of demand against three different hospitals, thereby enabling the health authorities in their choice of where to relocate services. Another example was the work of Charlton, Fotheringham & Brunsdon (2001) who looked at the spatial implications of hospital service closure in the UK Midlands. More recently this type of work has been extended specifically into the planning of community and health promotion areas through the work of Tomintz, Clarke & Rigby (2008) around smoking cessation services in Leeds, England. In the paper, the technique of micro-simulation is used to take existing survey and area data on 'at-risk' groups and use this to provide an intelligent estimate of smoking rates within the smallest available data units (output areas). This new knowledge is then used to look at the location of existing smoking-cessation services to see if they are effectively located given the modelled demand. In turn, location-allocation approaches were used to model potentially more effective locations for those services.

### Epidemiological planning

Much of the early work of quantitative and theoretical geographers on health and social care planning focused on the modelling of disease (Meade & Earickson, 2000). One of the earliest examples of disease modelling from a spatial dimension was Cliff, Haggett and Ord's work on the spread of influenza in Iceland (Cliff, Haggett & Ord, 1986). Prior to 1900 Iceland had no history of the disease but its outbreak and spread in three separate decades provided a very useful test-bed for the mapping of its arrival and diffusion. It was found in the last epidemic in 1957-8, that the infection arrived by air and was then spread, not in a gradual surface spread form but rather in a hierarchical way, moving from the capital Rekyavik to the smaller provincial towns and then out to smaller towns and then the countryside. Once visualised in cartographic form, the hierarchical and spatial aspects of these flows were clearer to planners and as a result more effective disease prevention measures were put into place.

This hierarchical process was also observed by Gould & Wallace (1994) with the spread of HIV/Aids in the US in the early 1980s. At a key time for the management of the spread of the virus, the advantage of being able to map existing and model potential future outbreaks of the disease was both providential for the containment and understanding of the disease and as a by-product, established the potential of GI science for epidemiological purposes. The Centre for Disease Control (CDC) in the US, continues to use GIS in disease mapping. Clearly for epidemiologists, the specific statistical and spatial statistical techniques used for mapping and monitoring disease are important. As there is no single definitive or perfect statistical approach, this has led to a variety of quantitative approaches across a range of health settings. A useful first collection on GIS and Health, Gatrell & Löytönen (1999) focused primarily on the epidemiological strand. Within that text a number of important issues were identified in terms of the effect of using different methodologies to map the same datasets. Braga et. al's work in the mapping of lung cancer clusters in the vicinity of the Italian cities of Lucca and Viterbo showed the benefits of using kernel and Bayesian methods for classifying rates of stomach cancer over more traditional standardised methods (Braga et. al, 1998). However while the results were more effective in visualising clusters they continue to raise some issues around whether the method drove the patterns rather than vice versa. In addition, critical concerns with the nature and meaning of disease clusters also inform ongoing epidemiological work in this area (Jacquez, 2008).

Again the place of statistical modelling of spatial data is central to studies like that of Johnson (2004) on the modelling of prostate cancer rates in New York State. In this work, standardised incidence rates (SIR) for prostate cancer by zip code were subjected to full Bayesian hierarchical modelling to identify the relative performance of spatially smoothed data. The result showed that the spatial scan statistic was effective in mirroring observed SIRs and that relative density of population was behaving consistently within the model with smaller rural zip codes being associated with greater levels of uncertainty. Innovative data approaches used to try and link personal spatial data with measured air quality data was the subject of work by Crabbe, Hamilton & Machin (2000) whereby the health status of individuals and their daily time-space patterns of exposure was compared with observed air quality data to see if relationships could be established. It was also a comparative piece of work between the cities of London and Barcelona.

### Policy and service planning

The location and distribution of health services, both primary and secondary, has long been a core concern of medical and health geographers. In particular the location of services in relation to the distribution of populations has focused attention on issues of equity, location quotients and notions of inverse care (Joseph & Phillips, 1984; Boyle et al. 2002). Additionally the relocation, reallocation and in some cases, closure of services has also been a highly contested area of health care planning debate. While such debates are often related to political and cultural factors, they are also debates which are regularly couched in spatial terms (Meade & Earickson, 2000; Jordan et. al, 2006). Similarly the geography of service administration and the tensions between administrative health region boundaries and the more natural and even non-spatial aspects of patient choice and mobility remain significant issues (Curtis, 2004; Gatrell & Rigby, 2004). All of these debates have a spatial dimension and clearly implicit in all, the notion of access to services is paramount. Indeed accessibility is often used as a proxy measure of service equity and it is with this in mind that the ways in which geographers have measured and modelled accessibility are explored below.

An area of study which in a sense links accessibility and utilisation with the location of services in a very specific spatial sense is the planning of emergency services. Such services include a range of ambulance services as well as fire, air-sea rescue and other paramedic services. A key aspect of such services is their mobility and spatial spread. Unlike most user-service interactions, which involve the former travelling to the latter, the reverse is often the case in emergency situations, though the patient is ultimately brought back to the service. Time and space modelling within GIS had been carried out in a number of emergency scenarios (Jones & Bentham, 1995). Examples include the work of Jones & Jørgenson (2003) who used multilevel modelling to explore risk factors associated with road traffic accidents in Norway and Zerger & Smith (2003) who looked at how GIS could effectively support disaster-scenario planning in the event of cyclones in Queensland, Australia.

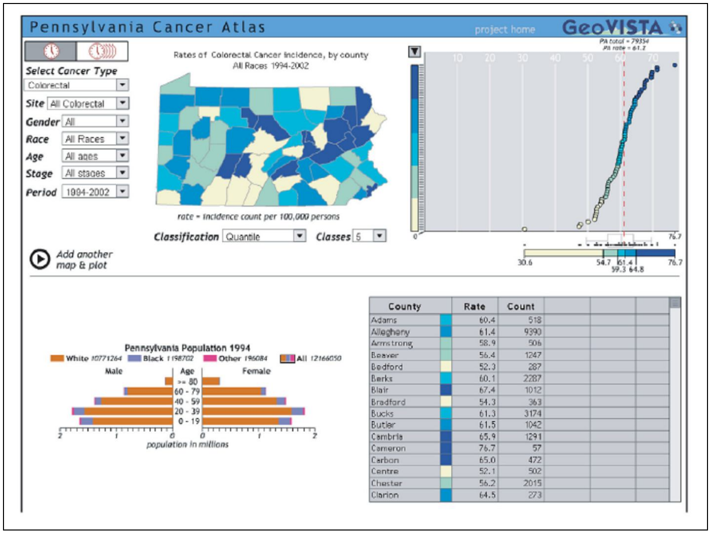
GIS has also been applied to the financial modelling aspects of health and social care planning. The funding and resourcing of health care systems is a complex mix of econometrics, demography, case-mix data and forecasting. In the 1980s in the UK, there was a realisation that health care resourcing was likely to vary over space and a number of approaches, broadly named, resource allocation models, were applied. These methods set in place the process whereby funding was moved from over-resourced areas to under-resourced areas and similar patterns were observed under the period of communist rule in the USSR as more equitable resourcing in the different republics led to an evening out (and improvement) in a range of health indicators.

More recently, the spatial dimension of resource allocation modelling has been more fully incorporated in countries such as Wales and New Zealand (Gordon et al., 2001; Senior & Rigby, 2001; Rigby et al., 2005).

The use of GIS in the planning of social care services, is, as stated previously, much less developed than health care. Depending on the jurisdiction, social care was traditionally managed separately from health care (as in the UK and Scandinavia) or embedded within health care structures (as in Ireland). With a growing realisation that a more integrated approach to social and health care planning may be more effective, some of the GI Science approaches used in the health sector are being slowly moved across to the social services. Many of these are still quite close to health services such as in the areas of community health, exemplified by the work of Driedger et al. (2007) whereby GIS-based spatial decision support systems were introduced into a public/community health arena related to early years services. The work aimed at seeing how easy or difficult it was to engage non-experts using web-based spatial tools and a number of difficulties arose, specifically around spatial literacy and the lack of spatial skills amongst key staff. Other work has been linked again into more specifically social settings such as the work of Zhang et al. (2006) on monitoring obesity and physical activity levels in public schools in Chicago. Here, the data was used to link spatial data across two different sorts of boundaries, namely census tracts and school catchments to better target educational initiatives to improve the future health of the children. Again some of these studies, while having a social/community dimension, are still within the health arena. More social-specific studies include the work of Foley (2002) on using GIS to plan services for informal carers (described in more detail below) and Milligan & Fyfe (2005) on planning for volunteering services as well as developing work in the fields of disability (Moss, Shell & Goins, 2006).

**Public Health and information management**

As noted previously, the gradual ‘democratisation’ of GI science and spatial data sets within society has led to a developing concern with the ways in which GIS experts can translate their theoretical and quantitative concerns and knowledge more effectively into the public realm. Quite apart from the promotion of the subject, it does behove GI experts to think through more fully about ways in which to inform and engage the public (Goodchild, 2006; Elwood, 2006). The development of the Internet as both a process and platform is also timely in terms of the presentation of health information. A number of public agencies including government health departments, specialist agencies such as the CDC in the US and even international bodies such as the World Health Organisation, have all been engaged in the promotion and dissemination of spatial health information in the last decade or so. In addition, this is an area where the potential of the visualisation features of GIS are particularly suitable. Bell et al. (2006) noted the increasing production and consumption of traditional flat maps but suggest that there is a need for more interactive and on-line health information which allows the public to carry out their own queries and analysis. Given concerns around cancer and other data sets, especially in relation to public (mis)interpretation, the clarity and communicative potential of such maps must be carefully thought through. They identify some examples, both from the work of the CDC but also within individual US states, which identify interactive formats, incorporating spatial statistical summaries and displays (see Figure 1 below) to better inform the public.



Select Cancer Type: Colorectal

Site: All Colorectal

Gender: All

Race: All Races

Age: All ages

Stage: All stages

Period: 1994-2002

Classification: Quantile

Classes: 5

| County | | Rate | Count |
|---|---|---|---|
| Adams | | 60.4 | 518 |
| Allegheny | | 61.4 | 9390 |
| Armstrong | | 58.9 | 506 |
| Beaver | | 56.4 | 1247 |
| Bedford | | 52.3 | 287 |
| Berks | | 60.1 | 2287 |
| Blair | | 67.4 | 1012 |
| Bradford | | 54.3 | 363 |
| Bucks | | 61.3 | 3174 |
| Butler | | 61.5 | 1042 |
| Cambria | | 65.9 | 1291 |
| Cameron | | 76.7 | 57 |
| Carbon | | 65.0 | 472 |
| Centre | | 52.1 | 502 |
| Chester | | 56.2 | 2015 |
| Clarion | | 64.5 | 273 |

Figure 1: Example of Interactive Health Statistics Website (Bell et. al., 2006)

Another area where digital health information has developed is evidenced by the gradual normalisation of electronic patient records (EPR), especially in countries such as France, where full patient data is held on a small credit card allowing for a much greater mobility and sharing of digital patient information. In terms of disseminating the tools and base data to set up a Health GIS capacity, the work of the WHO and its Health Mapper system, though flawed, is also an example of work in the developing world to tackle critical and ongoing issues around the management of infectious and increasingly, degenerative diseases. A good regional example of data sharing and disease monitoring is the work of Singhasivanon et al. (1999) on the management and spatial modelling of drug-resistant malaria in the Mekong Delta. Other ongoing examples are the work of the Health Observatories in the UK and the creation in Ireland of a new online product, jointly development by government, academia and private agencies called Health Atlas Ireland (Pringle et al., 2007). While ongoing problems remain with such initiatives, especially in the release of detailed spatial data sets for public consumption and getting the balance right between technical knowledge and public understanding, these are hopeful developments, especially in relation to the democratisation of public health information and the greater partnership involved between GIS experts, health professionals and lay users of health services.

### Summary

The above summaries are precisely that, a brief survey of ongoing GI science research in a number of sub-themes within the broader heading of health and social care planning. While the theoretical and quantitative aspects may be more fully appreciated by reading the cited references, it is hoped that a flavour of the kinds of problems and ways in which those problems have been tackled to date is incorporated above. While the division amongst the different application areas has created an applied structure, it is useful to summarise them in relation to the wider themes of data, analysis and visualisation. Clearly as the value of spatial data has become better understood, the collection and tagging of health data has become more and more effective and detailed. While the quality and level of this spatial ‘tagging’ does vary from country to country and from region to region, there is a much clearer understanding that good analysis requires good data. From an analytical perspective, the ongoing development of spatial statistics and spatial data analysis whether through Bayesian models, improved scan statistics along with increasingly sophisticated methods of data mining have all improved both the precision and quality of data analysis. Notwithstanding this, there are some persistent spatial problems, primarily around sampling, the spatial configurations of the data sets and comparability issues which continued to be problematic for analysis. Finally, the improvements in data and analysis are gradually feeding through into the public arena. While traditional maps and other cartographic output continue to be widely used, there is a development, driven by the ubiquity of the Internet, in other more interactive and innovative ways of displaying spatial data which are leading to better health and social care planning. Issues remain however in the areas of access, interpretation and public spatial literacy.

## APPLICATIONS IN HEALTH AND SOCIAL CARE PLANNING

This section will deal with the work of the authors and their own histories coming out of the UK and Irish systems. A number of surveys and studies carried out by the study team, do, we feel, exist as a useful body of work around spatial analysis and GI Science applications in the health and social care arena and we have been selective in choosing three examples. The chosen studies are driven by concerns with data, analysis and visualisation but also focus in on a core applied problem, with the intention of identifying an outcome or solution as well as identifying future directions for that particular line of enquiry. The first, and longest, example is based on work presented at the 15th European Colloquium on Theoretical and Quantitative Geography, held in September 2007 at Montreux in Switzerland. This work explored the modelling of cross-border hospital access on the island of Ireland. The other two examples are more briefly discussed and relate to: a) a UK based research project on the application of GWR techniques to the modelling of limiting long-term illness in Northern England and, b) a social care application in Southern England. From these empirical works, we have developed a range of evidence that also informs the wider literature. Many of the issues we have found in our research reflect wider concerns and we will end the chapter with some suggestions around how these issues might be tackled.

### Cross-Border Access to Hospital Services on the island of Ireland

With recent changes in the security situation in Northern Ireland, the two governments of the Republic of Ireland and Northern Ireland became interested in modelling economic and social structures across the whole island and health has been one of the key areas explored (Gleeson et al., 2008). Informally, there has been cross-border movement in the utilisation of health care and a recently published study by Jamieson & Butler (2007) identified considerable potential for cross-border collaboration in hospital services, particularly in the vicinity of the border. The pilot study presented here took a geographical or spatial approach to measuring accessibility to acute hospitals and examined how the current configurations both north and south could be expressed in terms of an accessibility score. It also quantitatively investigated another of Jamieson and Butler's themes, namely the relative equity of hospital catchments both North and South as expressed by beds per patient. In the latter half of the 20th century, the introduction in the North of the NHS model created one structure while a theoretical national public hospital system also existed in the South, though characterised by a more complex public-private mix with a stronger role for private health insurance. Northern Ireland had a population of 1.69 million according to the 2001 Census and this was estimated to have increased to approximately 1.74 million by 2006. In the Republic of Ireland Census data put the population at 3.92 million in 2002 and 4.23 million in 2006. Within both jurisdictions there was a range of hospital sizes, expressed in both the number of specialisms and the total bed count, with the latter being the sole measure which was meaningfully used in the model. In all 9 hospitals (Trusts) in the North and 40 hospitals in the South were included in the modelling.

From a spatial modelling perspective, there are a number of important considerations contained in policy discussions, which need to be considered. The re-organisation and improved planning of hospital services is by definition, the core concern of policy in both jurisdictions. Clearly, geographical tensions always exist in any decisions on where to locate services. These will reflect tensions between urban and rural areas, between densely and lightly populated areas and between local, regional and national imperatives. Few decisions made around either additions to, or cuts in, services provision escape the contentious question of exactly *where* these adjustments take place. Ongoing discussions noted in Jamieson & Butler (2007) and Murphy & Killen (2007) around the location of new hospitals (both regional and service-specific), exemplify all of these issues and further emphasise the importance of geography and the need to have spatially-informed decision making.

One area where policy was arguably lacking was in some of the evidence bases that were used for health care planning, particularly those with spatial dimensions (Charlton, Fotheringham & Brunsdon, 2001). It was possible to access annual data on the nature and level of hospital service provision in terms of bed counts, occupancy rates, specialisms and day patient activity. These were associated with individual hospitals but were aggregated up to regional or national level as well. It was also possible to get information on utilisation of services though the spatially-tagged data on this was better in the North than the South with postcodes found only in the former. Both of these sets of data have been studied and analysed but rarely had the locational/spatial aspects of both been put together in a holistic way. Additionally, geographical aspects such as density of population and the impact of distance had rarely been factored in to strategic planning (Murphy & Killen, 2007).

Perhaps the primary value of a GIS based approach is its ability to collate large volumes of information and to produce not one answer but a number of different answers to inform a number of different planning scenarios.

The aim of the project was to test the use of a spatial approach to examine specific aspects of accessibility associated with existing hospital provision. The specific objectives of the study are a) to use GIS to model spatial accessibility to acute hospitals in both Northern Ireland and the Republic of Ireland; b) to model for two different time periods to identify how changes in bed provision and local populations had impacted on accessibility; c) to provide a spatial measure of supply equity in the form of beds per patient and finally d) to explore how changes, even over a very short time period, impacted spatially on improvements or reductions in modelled supply.

The modelling was driven by the three core geographical considerations mentioned previously around: a) the distribution of potential patients (potential need/demand), the configuration of hospitals north and south (potential supply) and the transport network (accessibility based on travel time). Based on the literature on spatial accessibility noted above, three core datasets were identified as being essential. These included a) demographic data at electoral divisions (ED) and output area (OA) levels (drawn from the census), b) point datasets for individual hospitals with associated data on size, status and levels of provision and c) data related to the road networks in both countries. A number of issues arose in relation to spatial scale, compatibilities of classifications and the timing of data collection but a robust initial model was still produced (Pringle et al., 2000; Kitchin, Bartley & Gleeson, 2007). Once all the component parts were in place the data was fed in as a set of spatial data layers into Arc-Info, a proprietary GIS package, using a primarily raster approach. For the travel time each road segment had a unique ID and classification code for the road type which in turn allowed for the modelling of average speeds and hence travel time. The demographic data for population (and sub-populations) for each ED/OA was initially stored within a set of vector polygons but these data were transformed to be attached to an ED/OA centroid for development within the raster model. Finally the locations of all the hospitals were digitised into the system, initially in vector format.

Given that the aims were to produce a working accessibility 'score' as well as to define nominal catchments, the model started by assuming the creation of nominal non-overlapping catchments for each hospital. Once these catchments were defined and mapped, it was possible to use the background demographic data to compute the number of residents in each catchment. Given that we also knew how many beds were available at all the hospitals, we could then compute the national ratio of beds per head of population. To develop this, it was possible to then compute the expected number of beds if local supply followed the national rate and then calculating the ratio of the actual number of beds to the expected number of beds gives us the local bed rate as a location quotient. The second piece of modelling was more complex and primarily carried out within the GIS to combine the road network, travel speeds and the specific locations of the hospitals with small area local population counts to produce an effective ‘cost-distance’ surface which provided us with an accessibility score. The final technical stage was to remodel the accessibility scores with the Border included and excluded to examine its spatial effect on hospital activity.

The initial modelling focused on the period 2001/02. There was an estimated combined island population of 5.59 million in this period. The total number of beds modelled into the system at this time was 14,129 and using the two statistics together and all-island rate of 0.00257 beds per person was identified. Multiplying each modelled catchment’s population by this rate would yield the expected number of beds, which could be compared with the actual number. The initial map (Figure 2) identified a strong clustering of high accessibility values around the urban centres. This was driven very strongly by the location of the hospitals and by extension the higher densities of population close to those hospitals. The areas with the lowest accessibilities consist of two types, firstly the lakes which had been given artificially high values in the modelling and secondly, mountainous areas which were typically areas of low population density and amounts. However, the general low levels of inaccessibility associated with the western seaboard and the upland parts of the north should also be noted as identifying areas with genuinely low accessibility.

The second phase of the modelling looked at the period 2005/06 using updated hospital, road and demographic data. There were strong caveats with the demographic data due to the lack of up-to-date small area data for Northern Ireland and as a result this data was modelled from district level estimates. The accessibility modelling identified for 2005-06 provided very similar results for the earlier period with the same urban clusters being identified as areas with good accessibility and the more remote and rural areas being identified as poorly served. It was difficult to get a very strong sense of change from the spatial accessibility maps as the periods were particularly close and the increases in bed provision were matched by an increase in population across both jurisdictions. From a policy evidence point of view this is a useful finding in that the input of additional beds in this period (15,008 in 05-06 as opposed to 14,129 in 01-02) was countered by an almost identical 6.2% increase in population.

What was more valuable from a comparative point of view in this period was to look at change in a more disaggregated way by looking at provision at a regional scale. This was achieved by looking at the local hospital regions and their modelled bed rates. As noted in the methodology section, for each hospital catchment a form of location quotient for that hospital was calculated which compared actual local provision to the expected provision if national averages were applied.

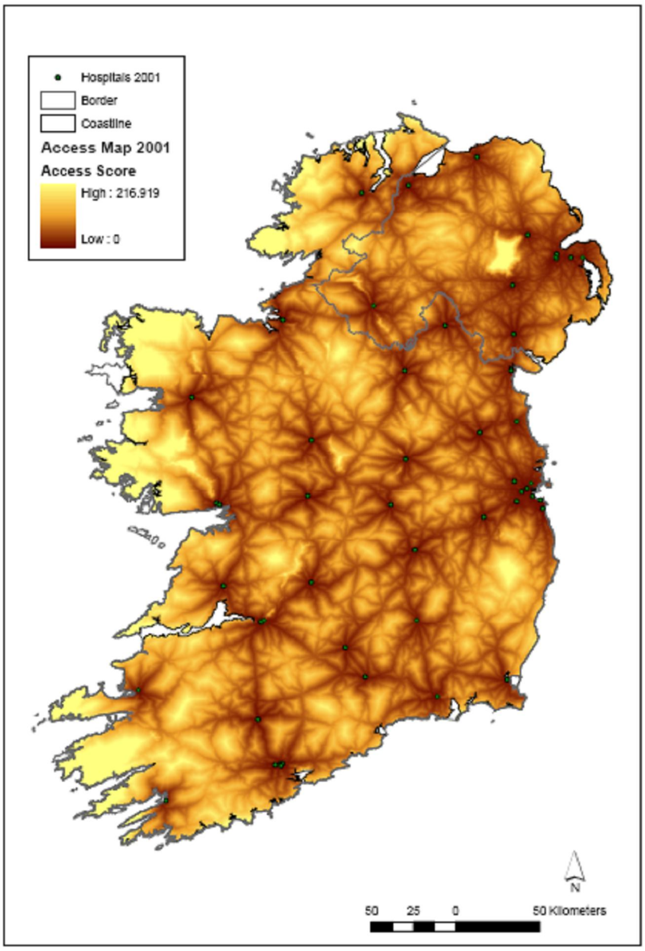


Figure 2: Modelled Accessibility to Hospitals on the Island of Ireland

This allowed for the calculation of location quotients for both periods. When the two time periods were compared (Figure 3), it was possible to tease out more fully changes at a local level. A number of areas showed a reduction or a loss in their location quotients, most definitively in Galway but will small losses recorded across the Midlands and South Coast and in Donegal and South Down as well. While a global pattern of change was hard to establish, this local level change was explained by areas where bed numbers had been reduced or had stayed static against an increasing population. In turn much of the North saw slight increases in the their location quotients as did parts of central and mid-Leinster and even some more remote parts of Mayo and West Cork. Policy makers could find this data, caveats notwithstanding, useful in a number of ways. Spatial approaches such as this identify more exactly where change is taking place. In addition, it should be noted that a reduction in the location quotient for an area like Galway, while in absolute terms suggests a diminution of service provision, could also be seen in relative terms as a reduction in over-supply which in turn brought the area more into line with the national average.

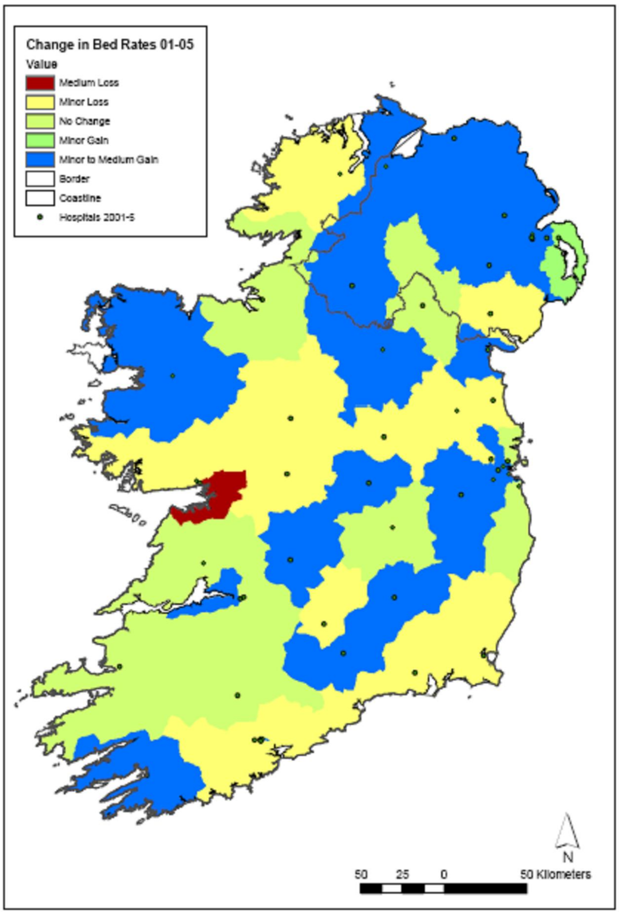


Figure 3: Variation in Service Provision: Change in Regional Bed Supply, 2002-2006, Ireland

As a final part of the modelling the impact of the border was identified and spatially modelled as two scenarios, one with and one without the border. This allowed the impact of a non-border scenario to be modelled and compared with a separate model. This identified the location of areas disadvantaged in terms of accessibility by the presence of the border, as well as the extent, expressed in excess travel time zones, of that disadvantage (Figure 4). Within the GIS it was possible to model in the border as a fixed feature and by subtracting the ‘no-border’ and border accessibility surfaces, it was possible to calculate a time disadvantage grid. This grid was then classified into time bands, vectorised and intersected with the population data to obtain the proportions in each band. While 52% of the populations in these areas were disadvantaged by the presence of the border by less than five minutes, a full 26% of the residents were disadvantaged by fifteen minutes or more. As Figure 4 demonstrated, the GIS was able to not only calculate these inequities but identify exactly where the zones were. This also identified another very useful policy function for a spatial modelling approach.

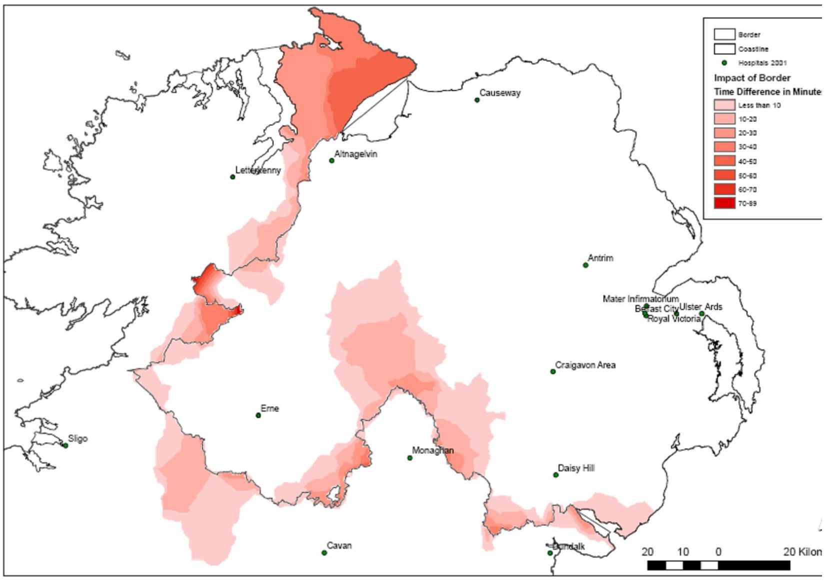


Figure 4: Time Variation in Access related to the presence of the Border

A developed version of the model would be improved by analysis of population data at small area level along with specialist health service data by specialty (utilisation rates, staffing numbers, hospital throughput). However the primary aim of the research was to identify the potential of GIS in terms of ‘scenario modelling’ and allowed for both a spatial and a numerical analysis of the impacts of the existence of the border. A predictive version of the model for say 2011 or even 2015 which included an adjustment to the bed sizes based on planned capacity changes would also be relatively easy to do. A third quantitative approach would be to model individual services which would incorporate the production of service specific accessibility surfaces, which might also be weighted by utilisation data.

### Using GWR to model Limiting Long-term Illness in Northern England

One of the most common technical issues is modelling health care applications in such a way that local, rather than global associations are incorporated into those models. Fotheringham, Charlton and Brunsdon (1998) looked at this problem in relation to the development of what they term geographically-weighted regression (GWR). Though GWR has been further developed in the intervening decade, this application is a health-specific one. Taking as a starting point earlier work around estimation methods (EM) which try to tackle the problem, the basic premise is to see how regression techniques can be improved by shifting from global to local relationships (Jones & Casetti, 1992). Typically, when a regression approach is used to establish a relationship between two variables for a number of spatial units, that relationship, usually expressed as a parameter estimate in the regression equation, is assumed to be the same (*stationary*) across all units. Yet clearly if one was applying this method across a region or country, one would assume that there would be some spatial variation (*spatial nonstationarity*) in that relationship in different parts of the country or region. Such spatial nonstationarity is not accounted for in the model. The methods used in EM tackle this problem by looking at trends in parameter estimates over space. However GWR develops this more fully by assuming that within the regression equation there is a continuous surface of parameter values, and we can obtain estimates of the values of the parameter at various locations of this surface. A number of statistical problems affect such an approach, particularly in the model specification, and are related to the choice of the spatial weighting function (to find the optimal bandwidth for the kernel type approaches used in the model) and also to finding a balance between sample bias and variance.

These issues were applied to a health-related data set, namely levels of limiting-long-term illness (LLTI), in Northern England, to test and map the spatial variation in the local parameter estimates created by GWR. In essence the paper analysed: a) how the GWR parameter estimates (showing spatial non-stationarity) compared with the earlier EM method and, b) how they performed when tested against a number of independent variables used to predict LLTI. The variation between the two different methods is shown in Figure 5 below with a focus on the parameters used in modelling the relationship between LLTI and unemployment.

The spatial distributions of the EM (Quadratic) showed an artificial tendency to the north-west of the map, whereas the GWR version much more closely matched empirical patterns of unemployment on the ground.

Another particularly interesting result was that for population density, another of the independent variables. One would assume that given that LLTI rates are higher in urban areas; these would in turn be associated with greater population densities. Yet the mapped GWR parameters for this variable showed up as very high in more rural part of Durham, one of the counties in the study. However, given the fact that high levels of LLTI in the UK have traditionally been associated with former coal-mining districts, it was found that those parts of rural Durham were actually where most of the former miners lived. This proved the effectiveness of GWR in exposing previously disguised local variations in explanatory statistical relationships. This and other forms of spatial statistics have to date been proved effective, yet have been under-used in other health related studies and continue to provide potential value in the planning of public and chronic health interventions.

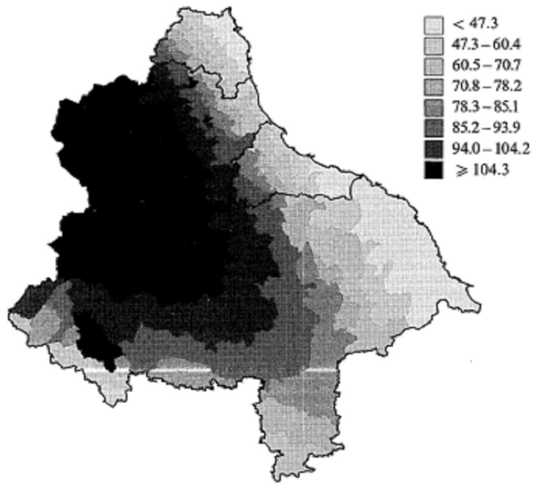


**Figure 9.** Quadratic expansion of the unemployment parameter.

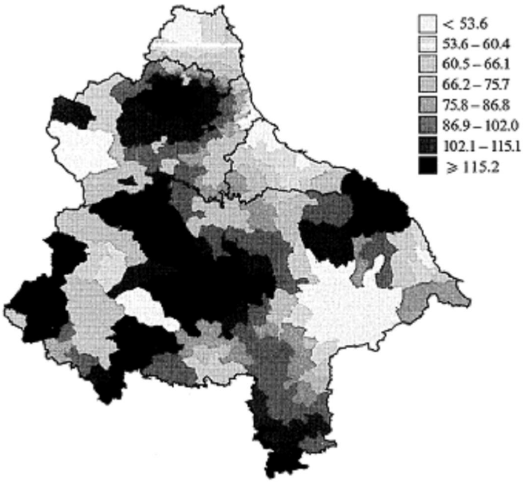


**Figure 14.** Geographically weighted regression distribution of unemployment parameter.

Figure 5: Spatial Variations between EM and GWR Methods (Fotheringham, Charlton & Brunsdon, 1998)

### Modelling Social Care Planning for Informal Carers in Southern England

One of the few examples where GIS was applied in a social care setting, Foley (2002) looked at the potential of GIS to assist in the planning of services to and for informal carers in Sussex, England. While a considerable lack of knowledge of the potential of GIS in this area was identified in discussion with key informants in the services involved, three useful case studies were developed to show its potential. The first case study used GIS to help model the financial implications of different levels of service provision within a spatially-costed model of potential demand. In essence, a cost per spatial unit (ward) for three different levels of service, daily, weekly and monthly, was calculated for respite and short-term care services across the county of East Sussex. This was weighted by existing and potential future demand and helped the local social services department identify exactly where costs might accrue in respect of the different models of services provision proposed.

The second related to identifying the optimal location for a new special school as access to an existing school had been closed off. This involved the mapping of existing utilisation of services and the demographic modelling of a specific sub-population of children (under-fives) who might use the school in future. Two potential sites were chosen and while the final choice was outside of the scope of the project, the narrowing down to two sites was a big step forward for the commissioning agency.

Finally, the work also modelled the impact of changes to service boundaries and the impact on potential catchments for a new voluntary service located on the border of two different authorities and crucially helped the service identify that it had sufficient capacity within the new authority to justify its location. The final section of the work also looked at some of the barriers and opportunities for the use of GIS in social care planning more generally, some of which are persistent and will be discussed below.

## CONCLUSION: SETTING A RESEARCH AGENDA

Having examined the wider literature and the specific recent projects carried out by the authors, it appears that the application of theoretical and quantitative aspects of GI science has reached a certain level of maturation, but is also ready for a new injection of energy. As an application area, health and social care planning are emblematic of this stage. Within the broad areas of data, analysis and visualisation, we have identified a range of ideas, many of which have been around for twenty years or more, such as MAUP, modelling local relationships and the meaningful measurement of health inequalities, which still exercise the minds of theoretical quantitative geographers. The focus on the spatial is in part what separates GI scientists from mathematicians and statisticians, which is also a source of complexity. Many algorithms and models, which work well in the abstract, become more problematic when projected onto Cartesian or three-dimensional space. Spatial behaviours, especially when modelled within the arenas of health and social care planning, have an additional applied aspect which is also a source of complexity. While theoretical aspects of supply, demand, need and access can all be modelled spatially, the planning aspect, essential in any applied use of these ideas, is subject to the input of humans, who have an innate ability to contradict any known theoretical construct. From a social care perspective, professional knowledge of these factors may be better fed into any future GI modelling.

This remains a core challenge to any work in the health and social care planning arena; to make the theoretical and quantitative work make sense in a real world environment. It should also be noted that GI scientists do not work in a vacuum and collaborate in both theoretical and applied ways, with a range of social and health professionals and biomedical experts. In such interdisciplinary settings, this potential is not always met as established practices and approaches often sustain barriers. This is not to say that such models cannot have value on their own. Indeed they often help planners to rethink the ways in which they model their own data. For example, the MAUP notion, familiar to most GI scientists, still comes as a surprise to some social and health care planners. Yet there is rich evidence that the demonstration of spatial analysis and the outputs of that analysis have helped focus planners and administrators attention much more thoroughly on the spatial potential of their own data sets (Foley, 2002; Smith, Higgs & Gould, 2005). Indeed the simple ability to map a health data set has often been sufficient to stimulate the interest of health service managers (Gordon & Womersley, 1997). As a balancing effect, the established work of health and medical professionals, especially in clinical and applied epidemiological settings, has sometimes been underused by GI scientists, and stronger collaboration and better communication must be considered. Finally, it should be noted that many of the algorithms and quantitative tools used by GI scientists in other arenas such as economics, engineering and the biological sciences may also have a wider applicability within health care, especially in the epidemiological and service-planning aspects of the subject (Longley et al. 2005; Wilson & Fotheringham, 2008). These have been barely looked at to date and represent a useful future direction.

With the gradual development of web-based products and the increasing ease with which locational data can be created, there are a number of pointers towards opportunities for the GI science community to embed their ideas into health and social care planning. As was noted above, the increasing publication and dissemination of health data on the Internet has created on the one hand, an expectation and on the other a requirement, to engage the public in this way. The success of Google Earth and the increasingly common applications being developed which link this to public mapping, mash-ups, wiki-environments and location-based services points to ways in which spatial data and visualisation are developing almost beyond theory. Information on health and social care facilities for example, is becoming one of a range of service information which is provided on such services. This is led by developments in the US, with many sites set up to provide answers to spatial queries such as ‘where is my nearest hospital?’ and ‘what residential care homes exist within 30 miles of my home?’. These types of sites are less well developed, though likely to be no less popular, in the more publicly-funded systems in Europe and other parts of the global North. The development of government sponsored web sites, both from within the health arena, but also from wider statistical sources such as the Office for National Statistics in the UK and EUROSTAT at an EU level, have also begun to create expectations for quality health data. In addition, the development of easy-to-adapt technologies such as Google Maps and open source software provides a real opportunity to promote the use of GIS in health and social care planning in the poorer countries of the world.

So what are the possible barriers to accessing this potential wider and deeper supply of spatial data? As noted above with the realisation of the potential of spatial information, data holders, public and private, are experiencing a demand for that information which has made them sit up and take notice. In an information society, where knowledge is power, ownership of spatial data, especially data at a detailed or even individual scale, has become more contested in the last decade. At the same time as the usability of data has improved due to better spatial tagging (often at address or even detailed co-ordinate level) and the development of EPRs, access to that data can remain problematic. This is less of an issue in countries such as Sweden, Finland and Norway, where traditions of public access to private information are well-developed, but in other parts of the EU, data holders often retreat behind issues of privacy and data protection. Indeed recent ‘scares’ around the loss of CDs containing personal information are likely to harden rather than soften that position (Fresco, 2008). While INSPIRE, the recent EU Directive on spatial data should theoretically embed frameworks to make a range of spatial data sets more publicly available, health and social care data do not appear prominently in the initial data appendices (European Union, 2007). A number of successful examples of accessing health data seem to have features in common. Working with and on behalf of health and social care agencies is important as are signed agreements around access. Proofs of concept, whereby sample data provided can be shown to be: a) safe and b) useful (in terms of its output being usable for planning purposes) may also help to build up a culture of trust and openness.

While technical advances in spatial modelling are and will continue to develop, there are a number of ways in which some of that technical work could perhaps be better integrated. While research on accessibility and utilisation often go hand in hand, they are often treated as separate issues in theoretical and modelling terms. This is partly to do with the fact that conflating the two can sometimes lead to counter-intuitive results. In addition, work on the measurement of accessibility is dominated by the supply end, so that much of the modelling looks at population-weighted travel times to generate accessibility surfaces. At the same time, location–allocation approaches are interested in the ways in which the volumes of supply impact on the location of services. Yet it seems as if the logical and strong theoretical connections between these two are not as well developed as they could be. The development of supply-weighted (as opposed to demand-weighted) accessibility and travel time maps seems a logical next step. Damiani et al.'s (2003) attempt to model choice into such an equation is a useful first step but it feels like more could be done. It is unlikely that the problems associated with MAUP will ever be overcome, especially in the area like health and social care planning. Given the fact the organisation of service planning in most jurisdictions is still wedded to fixed boundaries, catchments and administratively-oriented real units, this will remain a problem. Perhaps one approach might be to see if advances in spatial data at an individual level can be used as 'input', thereby providing a greater flexibility for developing different forms of areal units of 'output'. By aggregating better-quality raw data into models and using them to develop more applied, practical but still 'data secure' outputs, a lessening of the grip of MAUP and traditional reporting units might be possible. One interesting example of this from Ireland was the work of McCafferty & Canny (2005) in mapping benefit claimants into previously unused areal units such as housing estates and Traveller (Gypsy) halting sites.

While the development in secondary health care is stronger, there is the feeling that much promise remains in primary and social care settings. There is a suggestion that a gap still remains between the methods/technologies and their use in applied settings and part of the aim of the literature review is to see whether this gap is expressed through issues of spatial literacy, resourcing, data constraints or even human factors such as participation (Elwood, 2006). The increasing use of primary care data in interesting forms of analysis and visualisation coming out of research centres like CASA in London points to ways forward for this type of work (Batty & Longley, 2003). Examples include using geo-demographic data to model potential demand and the use of new survey sources to model interventions in a variety of public health and health promotion application such as teenage pregnancy and social deprivation (Longley and Singleton, 2008; Petersen et al., 2009). The championing by the WHO of the use of GIS in planning health care services in countries as diverse as Ethiopia, Cambodia and Niger points to its global potential. Indeed in many poor countries with poorly developed health surveillance infrastructures the arrival of location-based technologies has been a huge boon to spatially-tagged health data (Kloos et al., 2007). Finally, the challenges of climate change, and particularly its environmental health dimensions, point to ways in which GIS can be applied to more effective planning at a global level.

## REFERENCES

**Bailey, T.C.; Gatrell, A.C.,** 1995, Interactive Spatial Data Analysis. Addison Wesley Longman: Harlow

**Batty, M.; Longley, P.**, 1999, Advanced Spatial Analysis: The CASA Book of GIS. ESRI Press: Redlands

**Beere, P.; Brabyn, L.**, 2006, Providing the evidence: Geographic accessibility of maternity units in New Zealand. In: New Zealand Geographer, 62: 135-142

**Bell, B.S.; Hoskins, R.; Pickle, R.W.; Wartenberg, D.**, 2006, Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. In: International Journal of Health Geographics, 5, 49

**Bell, N.; Schuurman, N.; Hayes, M.**, 2007, Using GIS-based methods of multi-criteria analysis to construct socio-economic deprivation indices. In: International Journal of Health Geographics, 6, 17

**Brabyn, L.; Skelly, C.**, 2002, Modelling population access to New Zealand public hospitals. In: International Journal of Health Geography, 1, 1: 3-12

**Braga, M.; Cislaghi, C.; Luppi, G.; Tasco, C.**, 1998, A Multipurpose, Interactive Mortality Atlas of Italy. In: Gatrell, A.C.; Löytönen, M. (Eds), GIS and Health. London : Taylor and Francis: 125-138

**Bullen, N.; Jones, K.; Moon, G.**, 1996, Defining localities for health planning: A GIS approach. In: Social Science & Medicine, 42, 6: 801-816

**Charlton, M.E.**, 2008, Quantitative Methods and Geographical Information Systems. In: Wilson, J.P.; Fotheringham, A.F. (Eds), The Handbook of Geographic Information Science. Oxford: Blackwell: 379-394

**Charlton, M.E.; Fotheringham, A.S; Brunsdon, C.**, 2001, Analysing Access to Hospital Facilities with GIS. In: Clarke, G.; Madden, M. (Eds), Regional Science in Business. Berlin: Springer: 283-304

**Christie, S.; Fone, D.**, 2003, Equity of access to tertiary hospitals in Wales: a travel time analysis. In: Journal of Public Health Medicine, 25, 4: 344-350

**Cliff, A.D.; Haggett, P.; Ord, J.K.**, 1986, Spatial Aspects of Influenza Epidemics. Pion: London

**Cliff, A.D.; Haggett, P.**, 1988, Atlas of Disease Distributions. Blackwell: Oxford

**Couclelis, H.**, 2003, The certainty of uncertainty: GIS and the limits of geographic knowledge. In: Transactions in GIS, 7: 165-175

**Crabbe, H.; Hamilton, R.; Machin, N.**, 2000, Using GIS and Dispersion Modelling Tools to Assess the Effect of the Environment on Health. In: Transaction in GIS, 4, 3: 235-244

**Croner, C.; Sperling, J.**, 1996, Geographic Information Systems (GIS): New perspectives in understanding human health and environmental relationships. In: Statistics in Medicine, 15: 1961-1977

**Curtis, S.**, 2004, Health and Inequalities: Geographical Perspectives. Sage: London

**Damiani, M.; Propper, C.; Dixon, J.**, 2005, Mapping choice in the NHS: cross sectional study of routinely collected data. In: British Medical Journal, 330, 7486: 284-288

**Dicken, P.; Lloyd, P.E.**, 1990, Location in Space: Theoretical Perspectives in Economic Geography. Harper and Row: New York

**Driedger, S.M.; Kothari, A.; Morrison, J.; Sawada, M.; Crighton, E.; Graham, I.**, 2007, Using participatory design to develop (public) health decision support systems within GIS. In: International Journal of Health Geographics, 6, 53

**Elwood, S.**, 2006, Critical Issues in Participatory GIS: Deconstruction, Reconstructions and New Research Directions. In: Transactions in GIS, 10, 5: 693-708

**Foley, R.**, 2002, Assessing the applicability of GIS in a health and social care setting: planning services for informal carers in East Sussex, England. In: Social Science & Medicine, 55, 1: 79-96

**Fotheringham, A.S.; Charlton, M.E.; Brunsdon, C.**, 1998, Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. In: Environment and Planning A, 30, 11: 1905-1927

**Fotheringham, A.S.; Brunsdon, C.; Charlton, M.E.**, 2000, Quantitative Geography. Perspectives on Spatial Data Analysis. Sage: London

**Fotheringham, A.C.; Rogerson, P.**, 1994, Spatial Analysis and GIS. Taylor and Francis: London

**Fresco, A.**, 2008, Data-loss fiasco caused by 'woefully inadequate' system. In: The Times Online, http://www.timesonline.co.uk /tol/news/uk/crime/article4211711.ece, Accessed October 30th

**Gatrell, A.; Löytönen, M.** (Eds), 1998, GIS and Health. Taylor and Francis: London

**Gatrell, A.C.; Senior, M.L.**, 1999, Health and healthcare applications. In: Longley, P.A., Maguire,_D.J., Goodchild, M.F.; Rhind, D.W. (Eds), Geographical Information Systems. Chichester: John Wiley: 925-938

**Gatrell, A.**, 2002, Geographies of Health. An Introduction. Blackwell: Oxford

**Gatrell A.C.; Rigby, J.E.**, 2004, Spatial perspectives in public health. In: Goodchild, M.F.; Janelle, D.G. (Eds), Spatially Integrated Social Science: Examples in Best Practice. New York: Oxford University Press: 366-380

**Gleeson, J.; Kitchin, R.; Bartley, B.; Driscoll, J.; Foley, R.; Fotheringham, A.S.; Lloyd, C.**, 2008, An Atlas of the Island of Ireland: Mapping Social and Economic Change. NIRSA-AIRO: Maynooth

**Goodard, M.; Smith, P.**, 2001, Equity of access to health care services: Theory and evidence from the UK. In: Social Science & Medicine, 53, 9: 1149-1162

**Goodchild, M.**, 2006, GIScience Ten Years after *Ground Truth*. In: Transactions in GIS, 10, 5: 687-692

**Gordon, A.; Womersley, J.**, 1997, The use of mapping in public health and planning health services. In: Journal of Public Health Medicine, 19, 2: 139-47

**Gordon, D.; Lloyd, E.; Senior, M.; Rigby, J.E.; Shaw, M.; Ben Shlomo, Y.**, 2001, Targeting Poor Health: Report of the Welsh Assembly's National Steering Group on the Allocation of NHS Resources. National Assembly for Wales: Cardiff

**Gould, P.; Wallace, R.**, 1994, Spatial structures and scientific paradoxes in the AIDS pandemic. In: Geografiska Annaler, 76B: 105-116

**Guagliardo, M.**, 2004, Spatial accessibility of primary care: concepts, methods and challenges. In: International Journal of Health Geographics, 3, 3

**Haase, T.; Pratchske, J.**, 2005, Deprivation and its spatial articulation in the Republic of Ireland: new measures of deprivation based on the Census of Population 1991, 1996 and 2002. ADM: Dublin

**Higgs, G.; Smith, D.P.; Gould, M.**, 2005, Findings from a survey on GIS use in the UK National Health Service: organisational challenges and opportunities. In: Health Policy, 72, 1: 105-117

**Hodgson, M.J.**, 1988, An hierarchical location-allocation model for primary health care delivery in a developing area. In: Social Science and Medicine, 26, 153-161

**Horner, A.A.; Taylor, A-M.**, 1979, Grasping the nettle – location strategies for Irish hospitals. In: Administration, 27, 3: 348-370

**Howe, G.M.**, 1972, Man, Environment and Disease in Britain. Barnes and Noble: New York

**Jacquez, G.M.**, 2008, Spatial Cluster Analysis. In: Wilson, J.P.; Fotheringham, A.F. (Eds), The Handbook of Geographic Information Science. Oxford: Blackwell: 395-416

**Jamieson, J.; Butler, M.**, 2007, Removing the Barriers: An Initial Report on the Potential for Greater Cross-Border Co-operation in Hospital Services in Ireland. Centre for Cross-Border Studies: Armagh

**Jarman, B.**, 1984, Underprivileged areas: validation and distribution of scores. In: British Medical Journal, 286: 1705-1709

**Johnson, G.D.**, 2004, Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. In: International Journal of Health Geographics, 3, 29

**Jones, A.P.; Bentham, G.**, 1995, Emergency medical service accessibility and outcome from road traffic accidents. In: Public Health, 109: 169-177

**Jones, A.P.; Jørgensen, S.H.**, 2003, The use of multilevel models for the prediction of road accident outcomes. In: Accident Analysis & Prevention 35, 1: 59-69

**Jones, C.M.; Taylor, G.O.; Whittle, J.G.; Evans, D.; Trotter, D.P.**, 1997, Water fluoridation, tooth decay in 5 year olds, and social deprivation measured by the Jarman score: Analysis of data from British dental surveys. In: British Medical Journal, 315: 514-517

**Jones III, J.P.; Casetti, E.**, 1992, Applications of the Expansion Method. Routledge: London

**Jordan, H.; Roderick, P.; Martin, D.; Barnett, S.**, 2004, Distance, rurality and the need for care: access to health services in South West England. In: International Journal of Health Geographics, 3, 21

**Joseph, A.E.; Phillips, D.R.**, 1984, Accessibility and utilisation: Geographical perspectives on health care delivery. Harper & Row: New York

**Kelly, A.; Teljeur, C.**, 2004, A new national deprivation index for health and health services research. SAHRU: Dublin

**Khan, A.A.; Bhardwaj, S.M.**, 1994, Access to Healthcare: A Conceptual framework and its relevance to Health Care Planning. In: Evaluation & the Health Professions, 17, 1: 60-76

**Kistemann, T.; Friederike Dangendorf, F.; Exner, M.**, 2001, A Geographical Information System (GIS) as a tool for microbial risk assessment in catchment areas of drinking water reservoirs. In: International Journal of Hygiene and Environmental Health, 203, 3: 225-233

**Kitchin, R.; Bartley, B.; Gleeson, J.; Cowman, M.; Fotheringham, A.S.; Lloyd, C.**, 2007, Joined-up thinking across the Irish border: Making the data more compatible. In: Journal of Cross-Border Studies in Ireland, 2: 22-33

**Kloos, H., Assefa, Y., Adugna, A., Mulatu, M.S.; Mariam, D.H.**, 2007, Utilisation of antiretroviral treatment in Ethiopia between February and December 2006: spatial, temporal and demographic patterns. In: International Journal of Health Geographics, 6, 45

**Kumar, N.**, 2004, Changing geographic access to and locational efficiency of health services in two Indian districts between 1981 and 1996. In: Social Science & Medicine, 58: 2045-2067

**Langford, I. H.**, 1994, Using empirical Bayes estimates in the geographical analysis of disease risk. In: Area, 26: 142-149

**Lin, G.; Allen, D.; Penning, M.**, 2002, Examining distance effects on hospitalizations using GIS: a study of three health regions in British Columbia, Canada. In: Environment and Planning A, 34: 2037-2053

**Longley, P.A.; Goodchild, M.F.; Maguire, D.J.; Rhind, D.W.**, 2005, Geographic Information Systems and Science, 2nd Edition. Wiley and Sons: Chichester

**Longley, P.A.; Singelton, A.D.**, 2008, Social Deprivation and Digital Exclusion in England. In: CASA Working Paper, 145. London: UCL Centre for Applied Spatial Analysis

**Love, D.; Lindquist, P.**, 1995, The Geographical Accessibility of Hospitals to the Aged: A Geographic Information Systems Analysis within Illinois. In: Health Services Research, 29, 6: 629-651

**Lovett, A.; Haynes, R.; Sünnenberg, G.; Gale, S.**, 2002, Car travel time and accessibility by bus to general practitioner services: a study using patient registers and GIS. In: Social Science & Medicine, 55, 1: 97-111

**Luo, W.; Wang, F.**, 2003, Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region. In: Environment and Planning B, 30: 865-884

**Martin, D.J.**, 1996, Geographic Information Systems: Socioeconomic Applications. 2nd Edition. Routledge: London

**Martin, D.**, 1998, Optimizing census geography: the separation of collection and output geographies. In: International Journal of Geographical Information Science, 12, 7: 673-685

**Martin, D.; Wrigley, H.; Barnett, S.; Roderick, P.**, 2002, Increasing the sophistication of access measurement in a rural healthcare study. In: Health and Place, 8, 1: 3-13

**McCafferty, D.; Canny, A.**, 2005, Analysing Poverty at the Local Scale in Limerick. Presentation to Mapping Poverty Conference, Combat Poverty Agency, NUI Maynooth, Ireland, September 8th

**McLafferty, S.**, 2003, GIS and Health Care. In: Annual Review of Public Health, 24: 25-42

**Meade, M.; Earickson, R.**, 2000, Medical Geography. 2nd Edition. Guilford: New York

**Melnick, A.**, 2002, Introduction to Geographic Information Systems in Public Health. Jones & Bartlett Publishers: Boston

**Milligan, C.; Fyfe, N.**, 2005, Making Space for Volunteers: exploring the links between voluntary organizations, volunteering and citizenship. In: Journal of Urban Studies, 42, 3: 417-433

**Morrissey, K.; Clarke, G.; Hynes, S.; O'Donoghue, C.,** 2009, Accessibility Modelling. In: Bavaud, F.; Mager, C. (Eds), Handbook of Theoretical and Quantitative Geography. Coll. Workshop, 2. Lausanne: UNIL-FGSE: 311-334

**Moss, M.P.; Schell, M.; Goins, R.T.**, 2006, Using GIS in a first national mapping of functional disability among older American Indians and Alaska natives from the 2000 census. In: International Journal of Health Geographics, 5: 37

**Murphy, E.; Killen, J.**, 2007, Transportation accessibility issues and the location of a national facility: the case for a new paediatric hospital to serve the Republic of Ireland. In: Irish Geography, 40, 1: 1-16

**Noble, M.; Wright, G.; Dibben, C.; Smith, G.A.N.; McLennan, D.; Anttila, C.; Barnes, H.; Mokhtar, C.; Noble, S.; Avenell, D.; Gardner, J.; Covizzi, I.; Lloyd, M.**, 2004, Indices of Deprivation 2004. Report to the Office of the Deputy Prime Minister. Neighbourhood Renewal Unit: London

**Oliver, A.; Mossialos, E.**, 2004, Equity of access to health care: outlining the foundations for action. In: Journal of Epidemiology and Community Health, 58: 655-658

**Oppong, J.R.; Hodgson, M.J,** 1994, Improving Spatial Accessibility to Health Care Facilities in Suhum District, Ghana. In: Professional Geographer, 46, 2: 199-209

**Petersen, J.; Atkinson, P.; Petrie, S.; Gibin, M.; Ashby, D.; Longley, P.**, 2009, Teenage pregnancy—New tools to support local health campaigns. In: Health and Place, forthcoming

**Philibert, M.D.; Pampalon, R.; Hamel, D.; Thouez, J-P.; Loiselle, C.G.**, 2007, Material and social deprivation and health and social services utilisation in Québec: A local-scale evaluation system. In: Social Science and Medicine, 64, 8: 1651-1664

**Singhasivanon, P.; Kidson, C.; Supavej, S.** (Eds), 1999, Mekong Malaria. Malaria, multi-drug resistance and economic development in the Greater Mekong sub-region of South-East Asia, incorporating geographical information systems databases. In: The Southeast Asian Journal of Tropical Medicine and Public Health, 30, Supp. 4: 1-101

**Pringle, D.; Cook, S.; Poole, M.; Moore, A.**, 2000, Cross-border deprivation analysis: a summary guide. Oak Tree Press: Dublin

**Pringle, D.; Johnson, H.; Cullen, C.; Boyle, E.; Doyle, D.; Brazil, J.; McKeown, P.; Staines, A.; Beaton, D.; McIntyre, M.; Hennessy, C.; O'Kiersey, C.**, 2007, Health Atlas Ireland: An open-source mapping, database and statistical system. GeoComputation 2007. National University of Ireland, Maynooth, 3rd-5th September

**Rigby, J.E.; Skelly, C.; Tate, N.J.; Brabyn, L.; King, R.; Borman, B.**, 2005, Inequalities in geographical access to healthcare services: using GIS to inform policy planning and implementation. Paper presented at the 10th International Symposium on Health Information Management Research, SEERC, Greece

**Robitaille, É.; Herjean, P.**, 2008, An analysis of the accessibility of video lottery terminals: the case of Montréal. In: International Journal of Health Geographics, 7: 2

**Rushton, G.**, 1998, Improving the Geographic Basis of Health Surveillance using GIS. In: Gatrell, A.C.; Löytönen, M. (Eds) GIS and Health. Taylor and Francis, London: 63-80

**Rushton, G.**, 2000, GIS to Improve Public Health. In: Transactions in GIS, 4, 1: 1-4

**Sabel, C.; Gatrell, A.C.; Löytönen, M.; Maasilta, P.; Jokelainen, M.**, 2000, Modelling exposure opportunities: Estimating relative risk for motor neurone disease in Finland. In: Social Science and Medicine, 50: 1121-1137

**Schuurman, N.; Fielder, R.S.; Grzybowski, S.; Grund, D.**, 2006, Defining rational hospital catchments for non-urban areas based on travel time. In: International Journal of Health Geographics, 5, 43

**Senior, M.; Rigby, J.E.**, 2001, Unavoidable costs of rurality and remoteness in NHS resource allocation: applying the Scottish evidence to Wales. NHS Wales Resource Allocation Review: Cardiff

**Shuttleworth, I.; Lloyd, C.**, 2005, Analysing commuting using local regression techniques: scale, sensitivity and geographical patterning. In: Environment and Planning A, 37: 81-103

**Teljeur, C.; Barry, J.; Kelly, A.**, 2004, The Potential Impact on Travel Times of Closure and Redistribution of A&E Units in Ireland. In: Irish Medical Journal, 97, 6

**Tomintz, M.; Clarke, G.; Rigby, J.**, 2008, The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services. In: Area, 40, 3: 341-353

**Townsend, P.; Davidson, N.**, 1982, Inequalities in Health: The Black Report. Penguin: Harmondsworth

**Vigneron, E.**, 1997, Santé, société inégalités géographiques en France. In: Actualité et dossier en santé publique, 19: XII-XVI

**Weiner, D.; Harris, T.M.**, 2008, Participatory Geographic Information Science. In: Wilson, J.P.; Fotheringham, A.F. (Eds), The Handbook of Geographic Information Science. Oxford: Blackwell: 466-480

**Wilson, J. P.; Fotheringham, A.S.** (Eds), 2008, The Handbook of Geographic Information Science. Blackwell: Oxford

**Yuan, M.**, 2008, Adding Time into Geographical Information System Databases. In Wilson, J.P.; Fotheringham, A.F. (Eds), The Handbook of Geographic Information Science. Oxford: Blackwell: 169-184

**Zerger, A.; Smith, D.I.**, 2003, Impediments to using GIS for real-time disaster decision support. In: Computers, Environment and Urban Systems, 27, 2: 123-141

**Zhang, X.; Christoffel, K.K.; Mason, M.; Liu, L.**, 2006, Identification of contrastive and comparable school neighborhoods for childhood obesity and physical activity research. In: International Journal of Health Geographics, 5:14

## AUTHORS INFORMATION

**Ronan FOLEY**
ronan.foley@nuim.ie
National Centre for GeoComputation, NUI Maynooth
Department of Geography, NUI Maynooth, Co. Kildare, Ireland

**Martin C. CHARLTON**
martin.charlton@nuim.ie
National Centre for GeoComputation, NUI Maynooth, Co. Kildare, Ireland

**A. Stewart FOTHERINGHAM**
stewart.fotheringham@nuim.ie
National Centre for GeoComputation, NUI Maynooth, Co. Kildare, Ireland

# SPATIAL ANALYSIS OF SOCIAL FACTS

## A TENTATIVE THEORETICAL FRAMEWORK DERIVED FROM TOBLER'S FIRST LAW OF GEOGRAPHY AND BLAU'S MULTILEVEL STRUCTURAL THEORY OF SOCIETY

**Claude GRASLAND**

Université Paris Diderot, UMR 8504 Géographie-cités, France

## ABSTRACT

This document presents an attempt to build a theoretical framework for the spatial analysis of social facts, derived from Tobler's first law of geography ('Everything is related to everything else, but near things are more related than distant things') and Blau's theory of macro sociology and multilevel structural analysis. At the individual level four basic times of position and interaction are defined (geographical/sociological and discrete/continuous). It is then necessary to discuss the effects of scale aggregation and time dynamics on the elementary levels of position and interaction. This part is illustrated by examples about airflows between world cities in 2000 and euro coin diffusion across borders between 2002 and 2007.

## KEYWORDS

Theory, Social science, Geography, Sociology, Spatial analysis, Social morphology, Durkheim, Simmel, Blau, Tobler

## INTRODUCTION

*In friendly tribute to Waldo Tobler and in memory of Peter M. Blau (1918-2002)*

This document presents an attempt to build a theoretical framework for the spatial analysis of social facts, derived from Tobler's first law of geography ('Everything is related to everything else, but near things are more related than distant things') and Blau's theory of macro sociology and multilevel structural analysis.

Waldo Tobler's first law of geography is defined by "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This law was formulated in 1969 and many authors claim it is no longer valid in a global world where distance in general – and physical distance in particular – is less and less important for the understanding of economic, political and social dynamics. Following

Tobler, I argue that focusing on a "space of flows" rather than on a "space of places" as proposed by many authors in their research on Global City (Beaverstock et al. 2000; Sassen, 2002; Castells, 1996) in no way means that physical distance has to be excluded from spatial interaction models. Moreover, I argue that the distinction between "places" and "flows" is a kind of nonsense if we adopt the perspective that distance is not independent from flows.

The Association of American geographers recently organized a special session on Tobler's first law in 2001 and published the results in AAG in 2004, with six contributions (Barnes, 2004; Goodchild, 2004; Miller, 2004; Philips, 2004; Smith, 2004; Sui, 2004) and a reply of Waldo Tobler himself (Tobler, 2004). Some of the authors are critical and consider Tobler's first law as nothing more than an "unlawful relation and verbal inflation" (Smith, 2004) that can be explained by a specific step in the history of sciences (Barnes, 2004) Other authors are very positive but do not really discuss the law and rather use it as a justification of their own scientific activity as geographers, in the field of Geographical Information System (Goodchild, 2004) or Physical Geography (Philips, 2004) for instance. But only two authors (Miller, 2004; Sui, 2004) examine how Tobler's first law could be used as a way to build bridges between Geography and other Social Sciences, in particular sociology.

Based on my previous research in particular (Grasland, 1997), I suggest here that Tobler's first law provides the theoretical basis for a specific research field called spatial analysis of social facts, which could take place at the border between Geography and Sociology. As noticed by many human geography specialists (Rushton, 1993; Harvey, 1969) spatial analysis was for too long confined to the study of objects and paid insufficient attention to the behavior of individuals and groups. The study of the spatial behavior of individuals and groups is a subject that has been largely overlooked in the other social sciences, though it is at the very core of the specific contribution that geography could potentially make to the social sciences generally: "Social science deals with behavior and good theory is about behavior. Since all behavior occurs in spatial contexts, what then distinguishes the work of human geographers from that of social scientists more generally? (…) As all geographers know – though, curiously, many refuse to acknowledge – the spatial arrangement of things usually affect behaviors. The core of spatial analysis to the behavioral geographers is the analysis of behavior in its spatial context" (Rushton, 1993).

- In the first section of this chapter, I present an analysis of Tobler's first law of geography in connection with other works published by Tobler at the same period of his career, in particular the inversion of gravity model that allows linking positions, distances and interactions. I demonstrate that the inversion of gravity model modifies completely the point of view on gravity model as distance can be an output rather than an input of the analysis. What is important in gravity model is not its capacity to predict flows but its capacity to link structures and interactions. As a personal complement to Tobler, I develop the importance of barriers that are related to discrete attributes of spatial position (e.g. belonging to a political unit) and can sometimes be more relevant for explanation than quantitative attributes related to continuous attributes of location like coordinates.

- In the second section, I present Blau's macro sociological theory of social structure, which is a theoretical attempt to link social positions and social interactions through the concept of distances and opportunities. The formal analogy between the approach of Blau and Tobler makes it possible to derive a generalized framework of analysis of social facts, at the border of Geography and Sociology. I demonstrate how the axioms and theorems proposed by Blau for the analysis of social integration can be easily transposed to geographical parameters of position that can be either quantitative (distance) or qualitative (belonging) and are formally equivalent to the distinction made by Blau between graduate social parameters (e.g. income) and nominal social parameters (e.g. religion).

- In the third section, I focus on my personal contribution about the elaboration of a theoretical framework for the analysis of social facts that integrates into a wider perspective the previous discoveries made by Blau and Tobler. I examine in particular the difficulties to combine parameters of social and geographical position in the same explanatory model and I discuss how the analysis of a particular situation has to be integrated into a wider perspective in terms of scales of aggregation (emergence, determination) and time dynamics (memory, anticipation). These theoretical considerations are illustrated by recent research results on the diffusion of euro coins in Europe, which is considered as a proxy of global mobility. This example proves that the proposed theoretical framework can be applied on concrete situations and that both sociological and geographical factors can be nicely combined in the explanation of international mobility.

## FROM TOBLER'S FIRST LAW TO GRAVITY MODELS AND THEIR INVERSE

### Tobler's first law of geography

Looking at the contributions published by AAG on Tobler's first law of geography, I was very surprised to observe that most contributors focused on the famous isolated sentence ("Everything is connected to everything, but near things are more related to each other than distant things") without exploring its relations with the article where it was published, the articles published at the same time period and the following works by Tobler. This is not so surprising if we consider that, except some very famous papers like "Push-Pull migrations laws" (Dorigo & Tobler, 1983) or "Geographical filters and their inverse" (Tobler, 1969), many publications by Tobler are difficult to read as they were unpublished preliminary versions or purely grey literature. When I organized a seminar on Tobler's work with C. Cauvin in 2002, it was very difficult to realize a complete compilation of his paper and the 4 volumes that we were able to summarize still did not cover all of his bibliography. Invited to this seminar, Waldo Tobler filled some of our gaps in its bibliography but not all. And we were finally very surprised that such an initiative of collecting the complete works of Waldo Tobler had not been engaged in the USA before we did it in France.

Tobler's first law was published in the paper entitled "A computer Movie Simulating Urban Growth in the Detroit Region" in the journal of Economic Geography in 1970. In this relatively short 7-page paper, the main concern is the elaboration of a model of population forecast by computer simulation, which is very similar to present-day cellular automata procedures and other dynamic model like SIMPOP (Sanders, 2007). The first law of geography appears in its complete form only at the third page but the first part ("everything is related to everything") is formulated much earlier in the second paragraph of the introduction, with a clear position in favor of a positivist and structural approach of social life: "As a premise, I make the assumption that everything is related to everything else. Superficially considered this would suggest a model of infinite complexity; a corollary inference is that social systems are difficult because they contain many variables; numerous people confuse the number of variables with the degree of complexity. Because of closure, however, models with infinite numbers of variables are in fact sometimes more tractable than models with a finite but large number of variables" (Tobler, 1970). Tobler clearly considers that a scientific approach implies to write simple models able to capture a large part of the reality, rather than trying to cover every dimension. Moreover, the simplest models are the most general and are therefore likely to be transposed from one field of research to another.

About his model, Tobler comments: “The model recognizes that people die, are born and migrate. It does not explain why people die, are born and migrate. Some would insist that I should incorporate more behavioral notions (…). My attitude, rather, is that since I have not explained birth, death or migration, the model might apply to any phenomenon which has these characteristics, e.g. people, plants, animal, machines (which are built, moved, destroyed), or ideas. The level of generality seems inversely related to the specificity of the model” (Tobler, 1970).

My own practice of Tobler’s work has led me to the personal feeling that the key to Tobler’s first law cannot be found in this paper which is rather a claim in favor of scientific geography but not really a demonstration of the law, at least as the second part is concerned (near things are more related to each other than distant things). It is rather in two other papers published at the same period that we can find a clearer demonstration of the interest of a symmetric approach of positions and interactions through distance, which for me is Tobler’s major contribution to theoretical geography. To support my assumption that linking position and process is the core of Tobler’s contribution and the key of interpretation of his first law, it is also important to remind that the paper “Geographical Filter and their Inverses” was published in 1969, one year before the first law, and focused precisely on the theoretical problem of discovering processes through the analysis of forms: “If one assumes that geographical processes operate at various scales, then a filtering by scales could separate processes. The Fourier interpretation of scale is the wavelength, or, equivalently, the form of the spread function. Large-scale processes can thus be separated from small-scale processes as a preliminary step in geographical analyses. Some examples follow.” (Tobler, 1969). This was really a research program that was defined here and the publications of the following years were therefore clear attempts to propose empirical validations in various fields.

In the paper published in 1970 and entitled “Geobotanical Distance between New Zealand and Neighboring Islands” (Tobler et al., 1970), Tobler presented for the first time an empirical application of the inversion of gravity model and showed how it is possible to derive positions from interactions and, more generally, to link the analysis of flows and dissimilarities. In this work realized with natural scientists, Tobler appears as first author (against alphabetical order) and this fact clearly shows that he realized the core of the demonstration. Briefly said, the paper demonstrates how it is possible to propose a simulation model of diffusion of botanic species between the islands of New Zealand and to estimate the effect of distance on this diffusion. Even if we ignore the initial location of species, it is possible to make assumptions on it through an inversion of gravity model. Of course, the model does not fit very well with reality as it is based on Euclidean distance, but further complexity can be added and anyway it is sufficient to demonstrate that near things are more related to each other than distant things. Contrary to common opinion, Tobler does not believe that the concept of distance can be reduced to its basic geometrical component. But he claims that the simplest model should be firstly applied before to propose more sophisticated versions based on residuals. This is very clear in this paper where the authors write: “Model-building is useful not only because it may allow predictions but also because it identifies areas for further research by making assumptions explicit (…) Although the existing model cannot explain all distributions, one-third of the floristic variation between New Zealand and neighboring islands is explained by island location and size alone, demonstrating the importance of spatial arrangement in plant geography”.

One year later, in 1971, Tobler published with S. Wineburg in Nature what we consider his most fascinating realization. Entitled “A Cappadocian Speculation” (Tobler & Wineburg, 1970), this very short paper proceeded to a secondary analysis of archeological data and tried to determine the location and the name of unknown cities mentioned on the cuneiform tablets found by Hrozny near the village of Kültepe in 1925. The author notices immediately in the introduction that “This is theoretical geography in the sense of Bunge, which is conditional on several assumptions.” Contrary to previous researchers that had tried to analyze the relation between cities mentioned on the tablets through historical approach or linguistic approach, Tobler focused immediately on geographical distances that he proposed to estimate through an inversion of the gravity model assumption: “On a purely random basis, one would expect the names of large towns to occur more frequently than the names of small towns. The total expectation is thus that the interaction between places depends on the size of the places and the separation between the places. This rather obvious result has been verified in a large number of societies and for many phenomena. Specifically, we expect the interaction to increase as the places get bigger, and to decrease as they are farther apart. Many functions satisfy such a requirement. For social interaction the most common formulation is the so-called gravity model:

$Iij = k\ Pi\ Pj\ /\ d^{2}ij$

Where Iij is the interaction between places i and j; k is a constant depending on the phenomena; P, is the population of i; P is the population of j; and d is the distance between places i and j" (Tobler & Wineburg, 1970). Of course, we ignore what the size of the cities is and the flows between them, but we can propose an estimation of these variables: the size of cities is estimated by the frequency of quotations on the tablets and the flows are derived from the common citation of names. It is now possible to extract the distance by transforming the previous equation in:

$dij = (k.\ Pi\ Pj / Iij)^{1/2}$

But distance is only an intermediate step of the research as the real goal is to link the names of the unknown cities mentioned in the tablets with archeological spots that were discovered in the region. To fulfill this objective, Tobler proposed to use a trilateration method in order to derive the positions of cities from their distance and to produce a map of hypothetical position as names of a few cities were known and related to precise locations in space. The final result was the famous map presented in Figure 1.

One again, Tobler used Euclidean distance as the basic reference of the analysis but he claimed that this was not an obligation and that the model proposed could be extended to different situations: "Distance may be in hours, dollars, or kilometers; populations may be in income, numbers of people, numbers of telephones and so on; and the interaction may be in numbers of letters exchanged, number of marriages, similarity of artifacts or cultural traits and so on" (Tobler & Wineburg, 1970).

In our opinion, Tobler's most important discovery in 1970-71 was not the efficiency of the gravity model (it was established before), neither the importance of simplicity in scientific model (it was a very common thought in this period). The crucial discovery of Tobler was the fact that positions and interactions can be analyzed in a symmetrical way through the concepts of movements and accessibility. From this point of view, the common distinction proposed by M. Castells (1996) and followers (Beaverstock et al., 2000; Derudder et al., 2007; Sassen, 2002; Taylor, 2006) between the so-called 'space of places' and 'space of flows' is a kind of nonsense. What really matters is to define what the relevant geometries for a common description of both positions and interactions are. A real criticism of Tobler probably relies more on the choice of his favorite geometry (Euclidean, continuous) than on the ignorance of the importance of flows. But we have seen that Tobler always claimed that there was no reason to limit the analysis of relations between flows and structure to the case of Euclidean distance. From this point of view, Tobler's ideas are fully compatible with the network society and we have no reason to reject Tobler's first law on the basis of more and more complex forms of accessibility taking the form of networks. We can compare the position of the main airports according to Euclidean distance and to airflows distances (derived from an inverse gravity model) (Grasland, 2007).

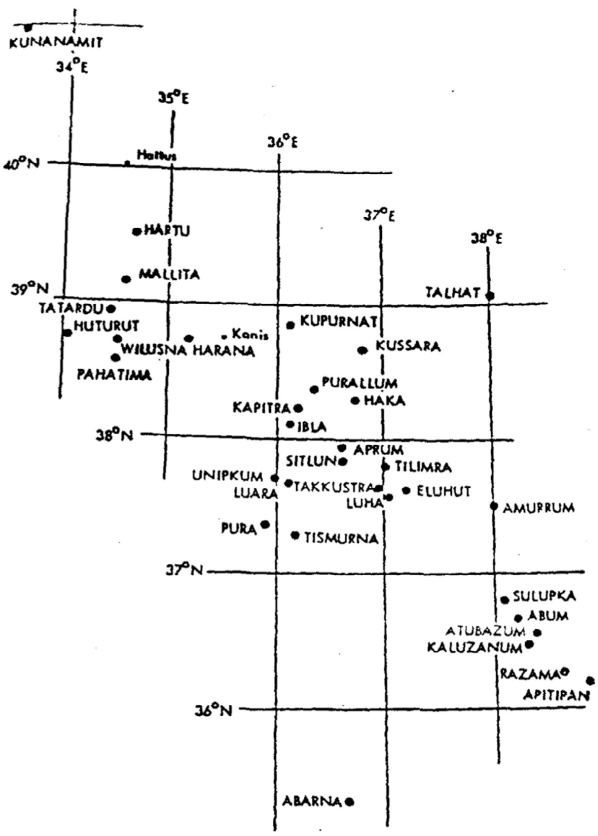


Figure 1: Predicted location of 33 pre-hitite towns by Tobler & Wineburg (1971)

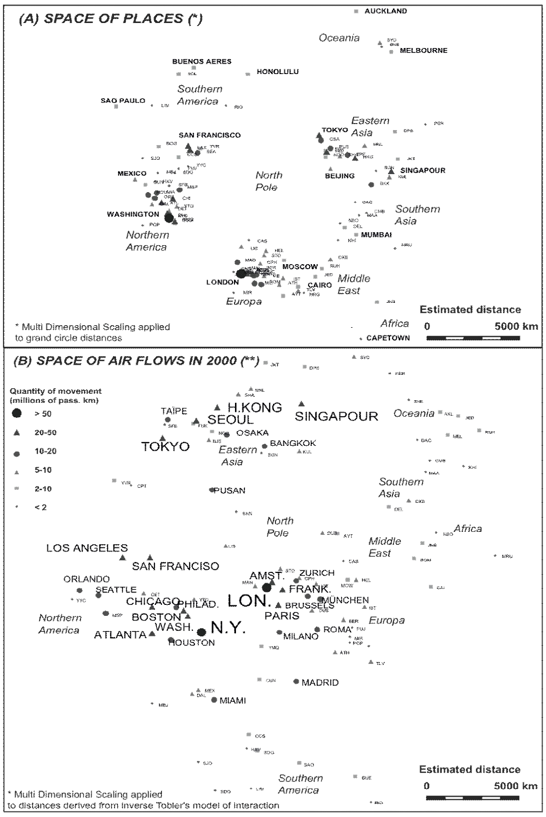


Figure 2: Predicted positions of 100 world cities according to Euclidean distance and air flows in 2000

What I want to illustrate with this two figures is the fact that the interesting scientific question is not to produce an artificial opposition between places and flows but, on the contrary, to propose various representations of places according to various types of flows by means of various types of distances. The map of Euclidean distances between cities is interesting for a given set of flows that are related to physical proximity. For example, the explosion of a nuclear plant power like Chernobyl has much to do with Euclidean distance, even if winds can introduce some anisotropy. If a nuclear problem appears in Toronto, we will obviously use the map (a) based on Euclidean distance to estimate the danger for neighboring cities and consider that Chicago is more likely to be concerned than Paris. But if we are interested in the propagation of a disease by means of air flows, it is certainly more relevant to use a distance based on the frequency of air flows as the one presented on map (b). In this map, the distances are derived directly from the flows through the inverse of a double constraint model. The proximity of cities reveals here the existence of small worlds of interconnected cities that are more connected than expected according to a random network (Guimera et al., 2005). We are exactly in the situation of the “Cappadocian Speculation” but the tablets have been replaced by the number of flights between airports, which could be analyzed by an archeologist of the future, making assumption on the location of cities according to the discovery of a manuscript describing the air connections in 2000.

### The justification of spatial and territorial interaction models in a social sciences perspective

Starting from Tobler’s perspective of joint analysis of positions and flows, the status of the gravity model appears completely different as it no longer is a deterministic model but a tool for the analysis of society that can be analyzed both in classical and inverse forms. To support this point, it is first necessary to remind some basic evidences on the theoretical justification of the decrease of interactions with distance and to complete what Tobler considers as obvious in his publications. At this point, we will introduce a more personal touch by considering not only continuous measure of distance (measured in time, kilometer, cost, energy) but also discrete forms of distance related by common belonging to a geographical area. In our publications and academic teaching courses, we propose to use the term “*spatial interaction*” for general effects of distance measured in a mathematical continuous form and to specify “*territorial interaction*” when we want to measure the discontinuous effect of a partition of space.

One remarkably persistent misconception is the claim that applying Newton’s law of universal gravitation to human movement is the main justification for the use of spatial interaction models. While it is undeniable that it has played a major part in their historical development, from Ravenstein and Reilly to Stewart and Zipf, the analogy should serve primarily as the starting-point rather than the end point of any theoretical inquiry into the role of geographical positions in the development of relations between individuals or groups. This is hardly the place for a comprehensive historical survey of the various stages in the development of models of social interaction. Still, it is worth noting that, in parallel with the refinement of instruments of formalization and modelization of flows at a macroscopic level, several theoretical currents have sought to provide an account of the concrete actions and processes that determine the spatial mobility at the level of individuals or small groups and which explain the emergence of regulations at the higher aggregate level constituted by places. At the same time, the empirical applications carried out at different levels and in different territorial or institutional contexts have tended to make more complex the basic hypotheses of the gravitational model and to refine it significantly with the addition of other variables besides the simple effect of Euclidean distance (linguistic barriers, political borders, infrastructure networks, time or cost of relation, etc.).

In my view, there are three broad categories of hypotheses that are likely to provide the theoretical foundation for models of spatial interaction (effect of distance) and models of territorial interaction (effect of belonging to the same grid). The following outline of these categories will be based on the specific case of migrations.

**Justification by the economic theory**

Firstly, the economic account of mobility starts with the hypothesis of a relation between the probability of relation and the cost of moving. Irrespective of the specific kind of cost incurred (monetary, psychological, temporal), the migrant is assumed to be a rational being seeking to optimize the relation between the benefits provided by mobility and the cost of mobility. If the opportunities of mobility situated at different points in space offer the same benefits for the migrant, he/she will merely seek to minimize the function of use constituted by the cost of moving. In theory, every migrant ought therefore to choose the destination involving the lowest cost, i.e. the closest destination. But because the migrant is in competition with other agents, there is an imbalance between offer and demand of mobility for some destinations. Because of such competition, it may therefore have to opt for a destination that involves a higher cost or abandon all hopes of moving if the cost involved is higher than the predicted benefits. The regulation between offer and demand should produce a balance resulting in a decrease of the probability of a relation according to the cost of moving. Thus formalized, the economic explanation does not prejudice the nature of the cost of moving and its relation to the geographical situation of individuals. An additional hypothesis is therefore required to change the economic determinant (the cost of moving) into a geographical determinant.

In the case of models of spatial interaction, I put forward the hypothesis that the costs of moving borne by the migrant are proportional to a given distance that is presumed to include a range of pressures bearing on the decision of mobility. Thus, in the case of a change of address, the choice of a remote destination implies a monetary cost (the cost of moving), a psychological cost (the rupture of social relations with individuals in the original location), a relational cost (the time involved in moving from the original location to the destination, if the migrant decides to make periodic returns), etc. Most of the distances that determine the overall cost of relation are variously correlated with Euclidean distance, which is the main justification for the efficiency of predictions obtained by means of the gravitational models, even the simplest ones.

The same hypotheses are equally applicable to a model of territorial interaction where the cost is dependent not on the number of kilometers effectively covered but on clearing limits of territorial grids. This is especially apparent in the case of trade flows where crossing a political border entails tax payments (monetary cost) and a waiting period that may vary at the border (temporal cost). But the same principle applies to the mobility of individuals. A person moving to another area or country is often forced to carry out a number of administrative formalities that entail a range of costs (temporal, physical, and financial), not to mention the psychological cost if the territorial grid in question is a particular territory or an inhabited space to which the individual is particularly attached.

My empirical research on internal migration in multinational countries (Belgium, Czechoslovakia, and Cameroon) demonstrated many times that territorial interaction effects are sometimes complementary to spatial interaction effects (Bopda et al., 2000). And in any case, they both should be simultaneously introduced in the model in order to control their relative contribution to explanation. Barriers are not residuals of gravity models based on distance but alternative forms of analysis of reality that imply the existence of discontinuities in social and spatial organizations.

**Justification by means of the theory of information flow**

The economic account of geographical interaction is the most relevant one since it does not reduce the notions of cost and utility to a mere monetary equivalence and since it also includes in its definition other factors such as time, attractiveness, habit, etc. A number of recent studies in spatial economics have tried to formalize subjective distances by showing that they are the product of a learning process and that they are thereby subject to more or less important intervals in relation to the distances that a homo economicus needs to consider.

Hägerstrand’s theory of the field of spatial information usefully complements the economic account by focusing on the information received by or made available to agents that is relevant to their opportunities of destination. The traditional economic paradigm presupposes that every agent is a rational and well-informed being, and assumes that an agent is fully aware of the cost and benefits of every relation that he/she is likely to forge. In practice the paradigm remains unverified. Any material flow is accompanied and preceded by counter-flows of information that enable agents to make a rational choice between several destinations though in a context of limited information. Hägerstrand's central hypothesis is that the probability of material relation between two individuals or groups depends on the quantity of information circulating between these individuals or groups. For example, the theory helps to provide an account of the process of self-maintenance of long-distance migratory channels. The settlement of the first migrants generates a flow of information between the point of origin and the destination, which facilitates the arrival of new migrants, who in turn will generate a still greater flow of information, and so on.

In the case of models of spatial interaction, it is the hypothesis of a deterioration of information in relation to the given distance that helps to account for the decrease in probabilities of relation as a result of distance. Individuals seeking to change address are far better informed about their migratory opportunities in their immediate environment than they would be in a distant environment. Within their immediate environment, they may benefit from local relays (parents, friends, newspapers, classified advertisements...) that facilitate their decision. In a more distant environment, they will incur a variety of costs in order to acquire information about their destination. Faced with two equivalent opportunities (such as two jobs with the same income), the agent will often choose the closer option because more information will be available and a number of uncertainties will presumably be removed, thus enabling the agent to act with full knowledge of the facts.

To a certain extent, Hägerstrand's theory of information flow accounts for the decrease in probabilities of relations with distance. But it applies just as much (if not more so) to models of territorial interaction. Indeed, Hägerstrand (1952, 1953) was the first to articulate the notion of a barrier that prevents the propagation of innovations and information flow. Obstacles to relations, whether physical (mountain range, lakes, rivers), political (borders, administrative limits) or social (linguistic barriers, socio-cultural discontinuities), often result in a decrease of information exchanges between inhabitants located on either side of such barriers. The effects of such barriers on information flow are numerous. Examples of barriers that prevent the complete circulation of messages between two territorial entities include super-absorbing barriers, which destroy the message and the sender, absorbing barriers, which merely destroy the message, and reflexive barriers, which return the message to its point of origin without destroying it. Cases of hermetic barrier are comparatively rare. Permeable barriers that allow for a transmission of some of the messages according to their permeability are far more frequent.

In short, the effect of barriers is both the concentration of information within territorial grids and the correlative weakness of information exchanges between territorial grids. In practice, this means that at an equal distance, an agent will have at their disposal more information pertaining to the opportunities of relation with the inhabitants of their own territorial grid than with those located in different territorial grids. The agent will therefore tend to privilege such relations, even if a higher cost is thereby incurred. Furthermore, if a territorial limit is an obstacle to the acquisition of information (for example as a result of linguistic barriers), the migrant will devote more energy to the acquisition of information related to opportunities of mobility in their new environment. The economic paradigm and the paradigm of information flow thus tend to complement and strengthen one another in the justification of the hypothesis pertaining to models of territorial interaction.

**Justification by the sociological theory of interposed or grouped opportunities**

Among the numerous criticisms leveled at the application of models of spatial interaction to the study of social relations, one of the most significant objections concerns the nature of the distance used to describe probabilities of relations. Most models of spatial interaction use a continuous measure of distance designed to reflect the impact of the cost of moving on the decision to develop a relation. The use of a function of decline of the probability of relation is the same for all the inhabitants or, at the very least, there is an average behavior that describes mobility continuously in relation to distance.

However, the hypothesis is only valid if the opportunities for relations are distributed evenly in space, i.e. if every agent has at their disposal an equal number of short, medium and long-range relations. This is not generally the case, which means that individuals situated in areas offering a low potential for relations are often forced either to incur greater costs or to reduce their mobility. Nevertheless, it is obvious that when mobility constitutes a response to a minimum need for relations, it is the first solution rather than the second that will tend to be adopted.

The solution provided by the sociologist S. Stouffer (1940, 1960) in order to bypass this difficulty is to alter the role of distance and to use it merely as an ordinal criterion allowing an agent to classify opportunities for relations. Generally speaking, the agent classifies its potential destinations in relation to the benefit/cost ratio, and then examines them sequentially. The probability of forging a long-distance relation does not therefore depend on the absolute value of this distance but on the number of opportunities in a more immediate environment. As the number of interposed opportunities (situated at closer range) increases, the possibility that the migrant will already have obtained satisfaction and decided against examining distant opportunities becomes ever more probable.

Stouffer’s solution is illustrated by the example provided in Figure 3. Two individuals situated at points A and B purchase their bread in either one of two bakeries situated in 1 and 2. If we were to apply gravitational logic strictly, we would have to say that the individual located at point B has a greater probability of using bakery 1 than the individual located at point A, because it is located twice as close to the bakery. But the individual located at point B has at its disposal a closer opportunity for a relation and if it decides to minimize the cost of moving, it will opt for bakery 2. By contrast, the individual located at point A has no alternative at its disposal; bakery 1 is the closest bakery to its house (no interposed opportunities). Unless it chooses not to eat bread, it will therefore agree to incur a higher cost of moving. Thus, contrary to predictions made on the gravitational model, the person situated at point A is more likely to use bakery 1 than the person located at point B, if the two agents adopt a strategy that aims to minimize distance. In fact, if ordinal distance is used instead of direct distance (ranks of opportunities), the behavior of the agents is in keeping with the logic of the spatial interaction model.

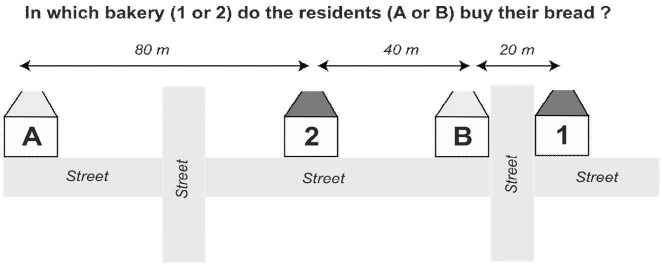

In which bakery (1 or 2) do the residents (A or B) buy their bread ?
80 m
40 m
20 m
A
2
B
1
Street
Street
Street
Street
Street

Figure 3: The case of the two bakeries

Thus Stouffer's model helps to improve models of spatial interaction and to make them more realistic by demystifying the role of distance. The decrease in probabilities of relation as a result of distance does not necessarily correspond to a mechanical effect but may be interpreted by means of a psychological hypothesis of behavior. Indeed, Stouffer's critique mainly pertains to gravitational modeling (simple constraint of preservation of all the flows) but it does not target the double-constraint models of the kind devised by Wilson or Tobler. The latter incorporate competition for destinations, thus helping to predict situations of the kind illustrated in Figure 3. Indeed, if we assume that every inhabitant uses a single bakery and that every bakery has just one customer (a double constraint of preservation of the total flows that are emitted and received), the most likely situation from the point of view of the minimization of mobility costs is that resident A uses bakery 1 and resident B bakery 2.

Stouffer's model therefore validates the hypotheses of spatial interaction models and undermines the objection that they project mechanistic models onto the spatial behavior of individuals. However, what remains to be shown is the validity of the hypothesis of the agents' behavior that underlies Stouffer's theory. Models of spatial interaction posit that agents organize opportunities of relation according to distance, which is presumed to reflect the cost of forging relations. This amounts to having recourse to the economic paradigm, which we previously demonstrated was equally applicable to models of territorial interaction. It also amounts to assuming that the agent is fully informed of the entire range of opportunities of relation and that its classification takes account of the full range of relevant information.

However, as we saw previously with Hägerstrand's model, the agent needs not necessarily to have at his/her disposal the exhaustive range of relevant information, and so the organization of opportunities that serves as a basis for his/her choice may be founded on partial or incomplete information. All other things being equal in relation to distance, the agent may have the use of more complete information related to the opportunities for relations situated within his/her own territorial grid than information pertaining to the opportunities for relations in different territorial grids. The agent's classification may therefore be a function not merely of distance but also of territorial affiliation, which may cause him/her to adopt a behavior that is both spatial and territorial. The migrant may also deliberately choose to prioritize territorial affiliation as a criterion of choice of destination.

It is thus possible to conceive of forms of embedded classification in which a migrant decides (1) to migrate if possible within their affiliation grid and (2) to opt for the closest destination, all other things being equal in relation to affiliation.

In the example of the two bakeries, it is conceivable that it is the children who purchase the bread and that their parents wish them to cross as few roads as possible. The decision rule established by the parents is thus an embedded choice that consists (1) in crossing as few roads as possible (territorial effect) and (2) in choosing the nearest bakery, all other things being equal in terms of the number of streets which have to be crossed (spatial effect). In this case, the inhabitants of residence B will choose bakery 1, which, despite being further, has the advantage of being located in the same block of houses (territorial grid) and does not entail having to cross a street (territorial limit). For the inhabitants of residence A, whoever purchases the bread will have to cross at least one street, and it may even be necessary to go to bakery 2, since the inhabitants of residence B, who are nearer, will more readily make use of the opportunities (bread loaves) located in bakery 1.

Other studies, particularly those by Fotheringham & O'Kelly (1989) have sought to enrich Stouffer's theory of intermediary opportunities with the concept of grouped opportunities. The models of spatial interaction based on the impact of masses and distance do not always take account of the spatial autocorrelation of such masses, i.e. the fact that there are areas at a higher level than the grid under study in which opportunities are either grouped or dispersed. Yet the migrant moving to a distant destination does not merely take account of the opportunities on offer at the destination but also considers those that it will be able to reach around the destination at the cost of a limited displacement. The consideration of groups of opportunities therefore presupposes an alteration of the basic hypotheses pertaining to the two-dimensional (rather than one-dimensional) character of geographical space. The most promising solution seems to reside in the elaboration of embedded models that study interactions between locations at different aggregate levels, thus allowing for the inclusion of concentrations of opportunities for relations at different levels.

The statuses of the theoretical hypotheses upon which the models of geographical interaction are based extend well beyond the fact of a simple analogy with Newtonian physics. And it seems difficult to overlook, at least as a hypothesis, how they might influence the constitution of social networks at different levels. In conclusion, I would insist on the fact that the division of spatial analysis between analysis of location and analysis of interaction is certainly useful for pedagogical reasons (I follow generally the division proposed in the two volumes of the manual published by Pumain & Saint-Julien (1997, 2000) for teaching at the undergraduate level) but it can be also an obstacle for deep understanding of linkage between positions, distances and interactions[1].

## LINKING TOBLER'S FIRST LAW OF GEOGRAPHY WITH BLAUS'S MULTILEVEL STRUCTURAL ANALYSIS

The paradigm of multilevel structural analysis and macro sociological theory of social structure elaborated by the American sociologist P.M. Blau (1918-2002) provides in our opinion a useful theoretical framework for the analysis of the impact of territorial grids and of geographical proximities more generally on the constitution of relations or networks at different levels of social reality (Blau, 1977b, 1993, 1994). While it can hardly be said to overlook the individual determinants of social interaction, Blau's work focuses more specifically on effects of structure and context, i.e. the mid-level analysis of social forms of association or integration (Lizardo, 2006; Scott & Calhoun, 2004). In this respect, Blau's perspective is an extension of Durkheim's work on social integration and social morphology but introduces also Simmel's ideas on crosscutting circle. Although Blau does not explicitly envisage the role of the geographical proximity of individuals as a factor of social integration (Ethington, 1997; Park, 1924), given some readjustments his theory and its central concepts are sufficiently broad to be applicable not only to the study of the impact of sociological positions but also to the study of geographical positions.

Blau's work is therefore especially pertinent in the present context because it is based on a distinction between quantitative and qualitative variables of sociological positions, which is analogous to the distinction used in my own work between distance and barrier effects (Grasland, 1997) for the description of geographical positions or interactions. This study will therefore presently examine how Blau's theory might be extended by considering, in addition to the two kinds of position commonly used by sociologists, two other positions aimed at describing the geographical properties of the location of individuals or groups that interact within a given society. The present analysis refers mostly to a synthesizing article by Blau (1993). This paper is one of Blau's last publications (at the age of 75) and can be considered as a summary of the part of its work devoted to structural analysis of society. It is more or less equivalent to Chapter 1 of his final book published in 1994 (Blau, 1994). We refer also to earlier papers where Blau elaborated the first version of a macro

[1] I generally try to introduce an integrated approach at the graduate level.

sociological theory of social structure (Blau, 1957, 1977a).

### Four elementary types of geographical and sociological positions

Blau's structural analysis begins by defining a multidimensional space by way of describing the social position of individuals according to a range of properties known as structural parameters. According to (Blau, 1993), structural parameters belong to either one of two elementary kinds that may combine to form more complex combinations.

(1) **Nominal parameters** pertain to the qualitative attributes of individuals and define social categories such as ethnic group, religious affiliation, professional activity, etc. The degree of differentiation of nominal positions is a reflection of the heterogeneity defined by Blau in the simplest sense of the term as the probability for two members of a given population to belong to different groups.

(2) **Graduate parameters** refer to quantitative attributes, i.e. continuous distributions of differences of resources and statuses such as income, education, etc. that serve to define social strata. The degree of differentiation of graduate parameters reflects inequality, defined most simply as the coefficient of variation (or Gini index) of differences between levels.

Applied to the study of geographical positions, Blau's distinction overlaps with the distinction made in my own work and discussed in the previous section between territorial parameters and spatial parameters of location.

(3) **Territorial parameters** pertain to qualitative variables of the location of individuals such as their affiliation with different territorial grids, which operate an exhaustive partition of space and society. Every individual belongs to one or several territorial grid; the status of these parameters is formally equivalent to the status of the nominal parameters defined by Blau. In its paper of 1977 on 'Macro sociological theory' Blau includes 'places' as a particular case of nominal parameter and proposes some specific theorems related to the effect of place location and distance. But he does not think it useful to distinguish this parameter from other categorical variables like sex or religion, even if he recognizes that they produce specific effects on social integration and organization of heterogeneity and inequalities.

(4) **Spatial parameters** refer to quantitative variables of location such as a system of coordinates including latitude (X) and longitude (Y). This two-dimensional space may be characterized by a metric structure with the usual properties of mathematical distance (identity, symmetry, triangular

inequality). Note the significant difference with Blau's level variables; insofar as spatial parameters commonly operate in pairs (a system of coordinates within a metric structure) and, in most cases, they are not oriented. Admittedly there may be cases where the spatial configuration of positions is one-dimensional (such as location on a highway or on a hill) or oriented (distance from the town-center). But these are particular situations and the more general case, which the remainder of this study will examine, is the two-dimensional space characterized by a Euclidean metric structure devoid of any particular orientation. It is also important to notice that some geographical distances cannot be directly derived from metric systems of coordinates and are related to network organization (e.g. accessibility by road in time). But it is not necessary, in our opinion, to introduce a third category for this situation as it can be considered that it is a network location producing quantitative measures of distances (time, cost…) that is the formal property of interest here. We have also seen in a previous section that, following Tobler's first law of geography, we can reverse the gravity model and transpose network distances into metric structures by multidimensional scaling (eventually with some loss of information).

| | | **MATHEMATICAL PROPERTIES** | | |
|---|---|---|---|---|
| | | **NOMINAL / QUALITATIVE**<br>(-> Heterogeneity) | **Area of uncertainty** | **GRADUATE / QUANTITATIVE**<br>(-> Inequality) |
| **SOCIETAL PROPERTIES** | **SOCIOLOGICAL POSITIONS**<br>(-> Social distances) | **Social category**<br>*(ex. religion)* | *(ex. Cast, social class or professional category)* | **Social level**<br>*(ex. income)* |
| | **Area of uncertainty** | *(ex. nationality)* | *(ex. ethnical Group)* | *(ex. distance from a center)* |
| | **GEOGRAPHICAL POSITIONS**<br>(-> Geographical distances) | **Territorial location**<br>*(ex. administrative units)* | *(ex. position in urban network)* | **Spatial location**<br>*(ex. latitude, longitude)* |

Figure 4: Four types of sociological and geographical positions with areas of uncertainty

Parameters of sociological position will be used to refer to the attributes of individuals that are theoretically independent of their location on the earth's surface or that are not at the very least directly altered in the short term by the movement of an individual in space. By contrast, parameters of geographical position refer to the attributes of individuals that serve to describe their present location on the earth's surface and that undergo complete or partial modifications every time an individual moves. The distinction is not a straightforward one and will require further refinements in due course with the more explicit consideration of the time scales that underlie the suggested definition. Indeed, the past geographical position of an individual (such as their birthplace) is often liable to become a parameter of sociological position, i.e. a lasting attribute of an individual (ethnic identity, nationality). But this particular point would require a more extensive account well beyond the limits of the present study.

The four types of positions thus defined primarily constitute a formal conceptual framework capable of clarifying a whole range of models and hypotheses. But they only define a partial schema (what, for instance, of economic or historical positions) containing many areas of uncertainty (Figure 4). While a number of attributes may be easily assigned to one of the four categories of propositions outlined above, other attributes are more difficult to assign unambiguously. A whole range of categorical (qualitative) sociological variables refers implicitly or explicitly to a hierarchy (casts, social or professional categories) and may also be interpreted as level variables. Similarly, in geography, the notion of neighborhood may refer either to a common affiliation with the same territorial grid (a person living in the same neighborhood or commune) or to a criterion of geometrical proximity or temporary accessibility (a person living less than 5 km or 10 minutes away). The distinction between sociological and geographical positions is no clearer. Thus, nationality leans towards the realm of sociology (an attribute that is independent of spatial location) in the case of rights of kinship, while it will tend to function as a geographical concept in the case of territorial rights. The distance from a symbolic place of social life (church, monument, sanctuary, town center) often relates to both types of sociological and geographical categories. Note also that ethnic identity may be viewed as implying all these categories, which it incorporates to varying degrees... An excellent discussion of the danger of fuzzy ethnic categories as compared to objective census categories (like administrative province of birth) is presented in Bopda's analysis (2003) of the role of the political capital Yaoundé in the building of Cameroon. A comparative analysis of internal migration and regional integration in multi-national states (Belgium, Cameroon, Czechoslovakia) (Bopda & Grasland, 1994; Bopda et al., 2000) also reveals how difficult it can be to isolate the respective effects of physical distance, administrative borders, linguistic regions and inherited historical divisions. Though imperfect, the four provisional categories outlined above may help to refine and widen the content of Blau's theory.

The focus will now turn to the principal axioms, theorems and corollaries of the theory and to an examination of the implications of the theory for the study of sociological and geographical positions.

**Positions, distances and reductions of relation opportunities**

The fundamental hypothesis of Blau's structural theory is that positions constitute structures that have a weak determining effect on the relations forged by individuals (Blau, 1993). This weak determinism implies that the structure in question may not necessarily exert a direct pressure on individual choices (by ruling out certain relations or making others compulsory), but rather exerts an indirect pressure based on the restriction of opportunities for relations resulting from differences in position and the correlative increase in sociological or geographical 'distances' between individuals (Park, 1924). This point is crucial for the connection we propose to establish with Tobler's first law of geography. As Tobler, Blau is not directly interested in the analysis of structures but in the analysis of the relation between structures and interaction through the intermediate but crucial concepts of distances and opportunities. Therefore they are not subject to the classical criticism of structuralism as they do not try to explain structure by structure but rather try to link macroscopic emerging social and geographical organization with interactions that take place between social units at local level (Boudon, 1984; Degenne & Forsé, 1994). However, this in no way means mean that this approach refers only to individual determinations, on the contrary.

The precise nature of the psychological or sociological mechanisms determining this restriction of opportunities for relations between geographically or sociologically 'distant' individuals varies greatly: internalization of norms, lack of information, preferences, material or immaterial cost, budget-time restriction, etc. But, strictly speaking, such mechanisms do not constitute the object of structural analysis, which focuses less on the individual determinants of relations than their aggregated result. The object therefore is to examine how the structure of positions (distance(s) between the full range of individuals or groups that constitute a society) determines the structure of interactions (relation(s) between individuals or groups constituting a given society) according to a specifically probabilistic approach. However, the primary focus of Blau's structural theory is not to isolate the impact of a specific position (a specific distance), but rather to study the global consequences of the compatibility or incompatibility of the various positions that individuals occupy simultaneously. Therefore the primary focus of his analysis is the internal structure of the multidimensional space of positions. Blau formulates the hypothesis that the internal structure to a certain extent determines tendencies for integration or anomy within a given society. His theory is thus situated in a direct line of descent from both Durkheim's social morphology (Durkheim, 1895, 1897) and Simmel's geometry of social life (1925, 1894, 1896). Simmel draws a distinction between societies founded on concentric social circles where the affiliations of individuals are based on successive inclusions within ever-widening groups, and societies founded on crosscutting social circles in which individuals participate simultaneously in different categories of affiliation that do not include the same members. For Simmel, the former case corresponds to traditional societies, whereas the latter is characteristic of modern societies displaying a high degree of social complexity.

Note that Simmel's distinction concerning social circles is directly applicable to the case of territorial grids. The administrative divisions within a given state may thus abide by a strict interlocking, with each grid containing a finite number of N-1 grids corresponding to Simmel's concentric social circles. But it may also be the case that the state uses specific grids for every control issue or for every problem that it has to resolve. In this case, there are no embeddings but rather a partial overlapping between different grids.

Nonetheless Blau's primary focus is on crosscutting social circles because they are more frequent in the societies which he studies, notably the large American metropolis in which individuals tend to belong simultaneously to ethnic, religious or professional groups defining systems of positions that operate more or less independently of one another and which seldom overlap (Blau & Schwartz, 1984; Schwartz, 1990). Still his theory is just as applicable to concentric social circles, which Blau analyzes indirectly through his study of the percolation of heterogeneity within the various organizational strata of social life. The theory is based on two propositions that may be viewed as axioms and that constitute the premise for a series of theorems (1977b, 1993, 1994):

- **Axiom 1:** the probability of a relation between two individual depends on the number of opportunities for contact between them.
- **Axiom 2:** the proximity of two individuals within the multidimensional space of social positions increases the number of opportunities for contact.

From the point of view of a geographer, the two axioms clearly define a system of models of individual interaction in which the hypotheses correspond exactly to the gravity models of spatial interaction used in geography at a more global level. Models of spatial interaction may be viewed as an aggregated articulation of Blau's theory in the specific case where the position of an individual is described solely in terms of their spatial location. The number of relations between two locations is proportional to the product of their respective populations (a consequence of axiom 1 at the level of population aggregates); the intensity of relations between two locations decreases in accordance with the distance that separates them (reduction of axiom 2 to the case of the geographical position of the individuals constituting the population aggregates). It is hardly necessary to add that these two axioms are more or less equivalent to Tobler's first law of geography as the axiom 1 implies that '*every individual is related to every individual'* with more or less opportunities and axiom 2 can be written as '*near individuals are more related than distant individuals'*.

### Blau's first theorem

The first theorem inferred by Blau from his two basic axioms is that the probability of inter-group relations is a direct function of the heterogeneity of the population in question. On the face of it, the first theorem has very little use given that it merely appears to articulate the notion that the macroscopic diversity of a given population has an impact on the diversity of relations potentially forged between the members of the population. However, the underlying idea is that the processes of segregation that may disrupt the development of relations between individual members of different groups are even less likely to appear as social groups are numerous and of similar size. The idea also rests on the assumption that individuals wish to have a minimal number of opportunities of forging relations to make their choice and that, when they are unable to find a satisfactory number of such opportunities within their own group, they are more willing to consider opportunities for relations with other groups. More generally, the value given to an opportunity for a relation is presumably not a parameter that operates independently of the social context in which relations are forged. What this implies is that the microscopic rule for the evaluation of opportunities for relations between groups depends in some cases on the macroscopic structure of the society in question and of the geographical context of the locations where such relations are forged.

Blau's first theorem pertaining to the impact of qualitative position variables is directly applicable to the domain of political geography, where it has often been used to describe the impact of the number and population of administrative units composing a given state. As the number of constitutive elements of a state increases and remains of roughly equal size, so the degree of autonomy of the latter is potentially reduced in relation to central power. The disintegration of the administrative grid multiplies the opportunities for relations between two individuals located in different territorial grids and reduces in the same proportions the probability that social relations are structured or concentrated within the same territorial unit. Therefore it reduces in theory the risk of one of the territorial grids being erected into an autonomous political entity liable to oppose the power of the central state and to claim its independence as explained by C. Raffestin (1980) and empirically validated by M.C. Maurel (1982a, 1982b, 1984) and V. Rey (1984, 1987) in the case of socialist countries. More generally, the multiplication of territorial grids increases the relative weight of relations that may be subject to the control of central power (such as a declaration of change of address). The degree of territorial fragmentation, defined as the probability that two inhabitants are located in different territorial grids, is thus formally equivalent to the social heterogeneity defined by Blau concerning categorical variables. In the context of a political theory of the kind devised by Ratzel (1897), this degree of territorial fragmentation operates as a measure of global territorial integration, which measures the degree of coalescence of territorial segments and is equivalent to the social integration defined by Durkheim as the degree of coalescence of social segments (1895, 1897).

### Blau's second theorem

Blau's second theorem states that the development of inequalities within a given population increases the probability of relations between partners at different levels. An ostensibly paradoxical suggestion, the theorem is based on the same observation as the first theorem, namely that the diversity of social levels (a quantitative attribute) implies an excessive restriction of opportunities for relations if partners choose only to foster relations with individuals at a similar level. However, Blau's theory (1977b, 1993, 1994) is incomplete because it does not examine the highly diverse forms of social inequality, particularly in view of the existence of discontinuities in the distribution of social levels (Figure 5).

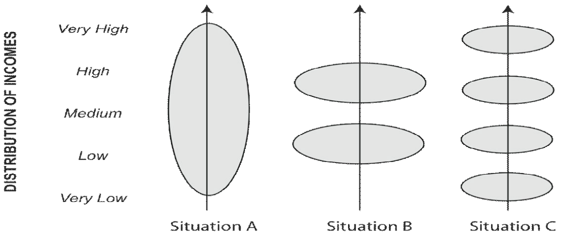


Figure 5: Three theoretical situations of distribution of social levels according to a given criterion (income, education, etc.)

According to this theorem, a society that includes two moderately different middle classes (situation B) is less unfair but probably far more closed than a society consisting of a continuum of individuals at different levels (situation A). In situation B, every individual will have many opportunities for relations with individuals at the same level, whereas in situation A an individual will need to develop relations with other individuals at different levels if they are to have a sufficient number of opportunities. However, Blau's argument is not altogether convincing since we might easily conceive of a situation C in which levels are highly differentiated (variation coefficient comparable to the variation coefficient of situation A) but where there are discontinuities in terms of distribution (as in situation B). Differences of level may not therefore necessarily result in an increase in relations between partners at different levels. The result is a situation that is comparable to situation B if the number of opportunities is adequate in every group separated by discontinuities of the higher and lower groups. It would appear therefore that Blau's second theorem applies specifically to unimodal strata distributions and is more debatable in the case of multimodal distributions. The theorem is still worth applying at a social level. Note also that the more a given population is concentrated in space around a limited number of poles, the more likely it is that long-distance relations will be forged between the inhabitants of this space.

By analogy with the situation of the various strata described above, in situation A the probability of two spatially-distant individuals actually meeting is probably smaller than in situation B. In the first case, every individual situated at a given location will have enough opportunities if they gradually widen their circle of spatial relations. With the exception of individuals in marginal positions (within the space under study), the rule describing the average distance of contact is isotropic. In the second case, every individual is located within the field of a common density gradient that incites them to move to the point of maximum density by way of finding the maximum number of opportunities for relations. The spatial structure therefore potentially favors relations between distant individuals who may nonetheless be subject to a common field of attraction (the town-centre in the case of a town). The most challenging case to interpret is Situation C because the existence of multiple cores provides a large number of opportunities for close-range contacts. On the other hand, individuals who wish to have a number of opportunities higher than the size of the population core of which they are a member are forced to establish relations at a greater distance than in the two previous cases because of the presence of interstitial voids (discontinuities). The result is a situation where there are strong analogies between the theories articulated in geography pertaining to the organization of urban networks (‘Rhenan’ and ‘Parisian’ models) and the distinction drawn in sociology by Granovetter (Philips, 2004) between strong ties that create heavily redundant relation networks (near relations) and weak ties enabling the development of the previous networks (near relations). At a more global level, we have demonstrated in different papers that the spatial distribution of population and wealth can be measured in terms of potential at different scales of neighborhood, providing a good description of macroscopic economic inequalities and flows of migrants and investments that are related to the gradients of this structure (Harvey, 1969; Hägerstrand, 1953).

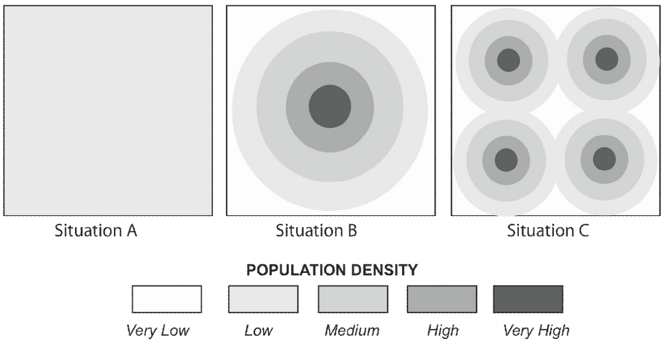


Figure 6: Three theoretical situations of distribution of spatial location of individuals

**Blau’s third theorem**

Blau’s third theorem postulates that the number of relations between groups varies directly as a function of the degree to which differences in social position function independently (Blau, 1977b, 1993, 1994). Conversely, the degree to which relations are confined within the same groups varies directly as the degree to which differences between positions are correlated. This central theorem is premised on the assumption that the multiplicity of independent factors determining opportunities of relation weakens their effect. An individual may wish to foster a relation with another individual from the same ethnic background, with a similar level of income, belonging to the same religion, or sharing the same political views. But if these criteria are weakly correlated within the society in question, the individual will necessarily have to sacrifice one of them unless it accepts a significantly reduced number of opportunities. The individual will therefore have to forge relations with individuals belonging to other groups defined by a range of different criteria. The third theorem is still only valid if we take account of the effects of the independence of the various social positions at the level of the entire population. As with the first theorem, we need to start from the idea that the norms governing relations between different social groups are not individual determinants that function independently of the global context of society. Therefore the systemic effect of the impossibility of applying the entire range of norms governing relations between particular positions (effects of specific barriers) is the emergence of a new global norm based on the weakening of all the norms described above. In other words, a series of independently and randomly distributed discontinuities eventually re-establishes a form of continuity within society.

The same reasoning applies to the study of geographical positions, where it can be shown that a series of discrete obstacles (barriers) eventually defines a continuous distribution of relations (distance) by virtue of their combination. The process may be construed metaphorically as one that involves the closure of a homogeneous plain and by predicting its effect on the relations between individuals situated within the plain (Figure 7). The cost of displacement is essentially a function of the number of obstacles that need to be overcome, with the cost of the trajectories between two obstacles presumed to be minimal. Clearing the obstacle reduces the opportunity for a relation by a factor of two.

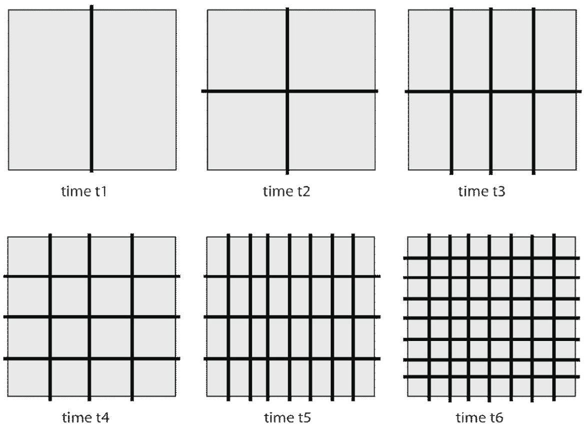


Figure 7: The multiplication of discontinuities re-introduces continuity

At the beginning of the process of closure (t1), the first barrier causes a major discontinuity within the space in question; relations tend presumably to be concentrated within each territorial grid. But the emergence of a barrier that is orthogonal to the previous barrier (t2) causes a reduction in the number of opportunities for relations within each new territorial grid and proportionally increases the opportunities for relations with the inhabitants of contiguous territorial grids. As new barriers that are orthogonal to the previous barriers are established (t3, t4, t5), the weight of opportunities for infra-grid relations becomes minimal and the opportunities for relations with the immediately contiguous territorial grids are reduced in turn. At the end of the process (t6), the opportunities for relations may be entirely described with the help of the measurement of rectilinear distance, which takes no account of the presence of different barriers and depends only on the spatial position of individuals. There has therefore been a gradual shift from a logic of territorial relation (the interaction between individuals depends on their affiliation with different spatial partitions) to a logic of spatial relation (the interaction between two individuals depends on a decreasing and almost continuous function of the distance between them). In mathematical terms, the demonstration can be made as follows:

The rectilinear distance (or Manhattan distance) between two position (xi,yi) and (xj,yj) is defined by:

$$DR_{ij} = \left|x_i - x_j\right| + \left|y_i - y_j\right|$$

It is immediately apparent that the first term is proportional to the number of barriers crossed latitudinally and the second term is proportional to the number of barriers crossed longitudinally. When there is a high quantity of randomly distributed barriers, the distance is therefore proportional to the total number of barriers crossed on the path between i and j

$$DR_{ij} \approx a\sum_{k=1}^{n} B_{ijk}$$

with

$\mathrm{B}_{\mathrm{ijk}} = 1$ if the path cross the barrier

and

$\mathrm{B}_{\mathrm{ijk}} = 0$ if the path do not cross the barrier

If every barrier $B_k$ produces a division by $\lambda_k$ of the opportunities for relations, the function that gives the opportunity of relation between two locations i and j can be derived as follows:

$$O_{ij} = \exp\left[\sum_{k=1}^{n} \ln\left(\frac{1}{\lambda_k}\right) B_{ijk}\right]$$

Assuming that all the barriers have an equivalent effect ($\lambda_1$ =... $\lambda_k$...= $\lambda_N$ = $\gamma$), the result is an estimate of the opportunities for relations that depends only on an exponential function of the rectilinear distance:

$$O_{ij} = \exp\left[\sum_{k=1}^{n} \ln\left(\frac{1}{\lambda_k}\right) B_{ijk}\right] \approx \exp\left[\ln\left(\frac{1}{\gamma}\right) . \left(\frac{1}{a}\right) . DR_{ij}\right] = \exp(-\alpha DR_{ij})$$

$$with \;\; \alpha = \frac{-\ln(\gamma)}{a}$$

This result ought to be applicable to the more general case of Euclidean distance if the direction of the barrier lines is random and does not automatically follow the trajectory of the longitudes and latitudes. But I have not been able to provide the formal mathematical demonstration of this hypothesis.

#### Blau's theorems in relation to the percolation of social heterogeneity

However, the most significant consequences of Blau's structural theory for geography and the study of territorial grids pertain less to the first three theorems (described in different ways) than to their consequences for the analysis of social forms at different aggregate and observational levels.

Blau argues that the integrating effects of social heterogeneity, social inequality and the absence of correlation between social positions are all the more marked since they appear at every organizational level of social life (1977b, 1993, 1994). Because differences that appear at the level of individuals necessarily have an impact at the level of groups or social classes, the main problem is to measure the highest level of the percolation of differentiations within the different levels that constitute any given society. If for instance the object is to study the potential for relations between two groups of students (white/black, men/women) in a given university, then we will need to take account not merely of the relative frequency of each group within the institution, but also of these relative frequencies respectively in each of the disciplines, subject areas and seminar groups within the university. Every level corresponds to specific opportunities for relations (on Campus, in the secretary's office of a given department, in the lecture hall, within classrooms) that reduce opportunities for relations. Heterogeneity at one level (50% of students from every group in the entire university) may indeed conceal a perfect homogeneity across all or part of the lower levels (segregation according to subject matter, seminar groups, etc.).

Blau's analysis of heterogeneity percolation is applicable to a strictly sociological hierarchy of different organizational levels. Schwartz (1990) applied directly Blau's theoretical framework when he studied the percolation of social heterogeneity with a survey of 29,000 Californian students in 1360 classrooms, 362 grade levels, 245 institutions and 18 school districts. The social position of students was described by means of three categorical indicators (sex, ethnic group, parents' profession) and two indicators of level defined by the teachers (school results and behavior in the same classroom). The dependent variable was an inquiry into those individuals whom they 'like to do things with "each other" always'. The results tended to confirm the theory (with the exception of sex ): the higher the percolation of heterogeneity at every level, the higher the frequency of positive appreciations pertaining to the opportunity of forging contacts with students occupying different positions.

But the analysis of percolation may also focus on the hierarchy of spatial or territorial levels. Indeed, a hierarchy of sociological levels is only seldom compounded to varying degrees by a hierarchy of geographical levels, because of the spatial constraints inherent in any form of social life. This argument probably constitutes the strongest defense for the simultaneous integration of sociological positions and geographical positions in the study of social forms and social relations. One of the best demonstrations of such a common approach is provided by a book on residential segregation and school segregation published in 2009 by a geographer (J.C. François) and a sociologist (F. Poupeau). But as explained by J.C. François, such collaboration was the result of more than 10 years of negotiations between researchers having initially very different practices, vocabulary and disciplinary challenges…

In the context of a structural and multiscalar analysis of social relations, the use of variables describing individuals' geographical position is formally possible without making any adjustments to the basic hypotheses. Still, its significance will depend on the possibility of demonstrating at a theoretical level that, for the kind of relation under study, there are spatial or territorial determinants of individual or collective action that operate independently of strictly sociological determinants.

## A FRAMEWORK FOR SPATIAL ANALYSIS OF SOCIAL FACTS AND AN APPLICATION TO THE CASE OF EURO COINS DIFFUSION

The advantage of a theoretical framework linking sociological and geographical interaction (societal models) is that it provides a global formalization of the potential effect of geographical and sociological positions on the development of opportunities for relations between individuals or groups at different observational or organizational levels. The problem is that it tends to focus too exclusively on the determination of relations by structures. It therefore tends to underestimate the importance of two fundamental factors that intervene over time and which have been highlighted (among others) by sociologists working on social networks, by geographers researching urban networks and demographers studying migrations. The individual determinants of the relation take no account of the processes of mediation, absorption or indirect relations that may occur within a given social system. Two individuals occupying distant geographical or sociological positions may come into contact as a result of a third individual acting as an intermediary or of a series of intermediaries that combine to form a ‘path’. The same reasoning applies to the level of groups, where a distinction between strong ties and weak ties is in order (Philips, 2004). A limited number of individuals can thus guarantee the circulation of information between large groups, provided they occupy a strategic position within the system of sociological or geographical positions. This articulation is often facilitated by a hierarchical organization in which every group delegates representatives to a higher-level group, thus guaranteeing a connection between different social segments.

By contrast, we can say following Stouffer (1940, 1960) that two individuals or groups occupying near positions within a geographical or social space can nonetheless forge very weak ties because of the interposition between them of a significant number of intermediary opportunities for relations. The inhabitants of two oases a hundred kilometers apart in a desert probably have more contact than the inhabitants of two suburbs located on opposites sides of the same town or city. Indeed the inhabitants of the two oases have no other option but to cross an expanse of desert if they wish to extend their network of acquaintances; the material obstacle of physical distance will be all the more easily reduced. Conversely, in the case of suburbs, physical distance is not an important obstacle, but the concentration of opportunities for relations at close range (town centre) will tend to absorb a large part of the relations that could potentially be established between inhabitants located on either side of the city. Both examples show that the opportunity for relations between two individuals depends not merely on their respective positions but also on the positions of the other individuals with whom they could potentially forge relations. The estimation of the potential of relations between pairs of individuals (between pairs of positions) is therefore one step in the modelization of interactions that could be established and which are subject to a range of systemic constraints and regulations. As demonstrated by L. Sanders (2007), neither macroscopic, mesoscopic or microscopic models can solve the contradiction and it is only by a combination of levels that it is possible to produce efficient models of social and spatial interactions, irrespective to the mathematical and computing tools used to do it.

Assuming that the entire range of potential interactions between positions has been taken into account at different scales (macro, meso, micro) there still remains the issue of measuring the impact of effective interactions on the formation of opportunities of relations at a given moment. For social systems are not purely Markovian systems in which the state of positions at any given moment helps to predict the relations that could be forged at later times. Assuming that the relations forged between t0 and t1 were entirely determined by the initial structure of the system at time t0, its subsequent evolution will depend both on the new structure at time t1 (altered by the interactions) and the memory of earlier relations that the system is likely to acquire. The relations effectively forged within a given system may modify an earlier structure of positions (social mobility, geographical mobility). But in themselves they also constitute a structure that impacts heavily on the development of future relations. The fact that two individuals i and j enter into a relation with the same individual k increases the probability of them meeting, even if their respective positions have not been modified. In the same way, the fact that two groups A and B that did not wish to enter into contact are nonetheless compelled to forge a relation (for example as a result of their geographical proximity) will change the rules of mutual appreciation, either by reinforcing mutual hostility (conflict) or by lessening such feelings (integration). The relation can modify not merely the structure of positions but also the rules of mutual appreciation entailed by differences of position.

Our contribution is not to solve such big questions (scales, dynamics) but to describe in an approximate way how it could be done in order to link individual positions (discussed in the previous section) with upper levels of organization and changes through time. To illustrate this abstract formalization, we present some examples based on our research on euro coin diffusion with F. Guerin Pace. This example is not chosen by chance but because it can be easily related to Tobler's first law (remember that Tobler was very interested in banknotes diffusion in USA and used it for a theoretical demonstration of analysis of movement (Tobler, 1981) and also to Blau's macro sociological theory and multilevel structural analysis (as euro coin diffusion can be analyzed at very different scales and can reveal very complex combination of factors, both geographical and sociological).

### The structure of a particular observational or organizational level

The internal structure of sociological, economic or geographical systems (individuals, companies and places respectively) invariably contains two main sub-systems describing respectively the elements of the system (the structure) and the interactions between these elements (relations). The positions correspond to different groups of attributes of the elements, according to whether they are quantitative or qualitative, geographical or sociological, etc. Every type of position may of course correspond to several variables and not just a single variable as in the examples cited above. In cases where the elements of the system are not individuals that are deemed to be identical, it is often necessary to draw on another category of indicators that describes the weight or size of the social aggregates (population), their activities (added value) or the space in which they are distributed (surface area).

The differences of position across the full range of criteria define opportunities for relations between pairs of elements. These opportunities correspond to a rate of potential interaction ranging between 0 and 1 when the elements are individuals of equal weight. By contrast, they correspond to a potential volume of relation in cases involving aggregates of different sizes. These opportunities for relations correspond to an offer of relation that is the product of elementary constraints connected with the differences of position and which result from a whole range of economic, psychological and informational factors. The offer of relation is then subject to a number of systemic constraints that result from the total energy available within the system and to its distribution between the different individuals located therein. Every individual may wish to establish a minimum number of relations, but they may not be able to forge an infinite number. Individuals occupying a central position will therefore probably have fewer relations than might have been predicted on the basis of their total number of opportunities for relations.

The relations effectively observed within the system therefore correspond to the effective realization of opportunities for relations relayed to positions under different constraints. They can be analyzed by means of different grids according to the hypotheses made about the position determinants that have the greatest impact. Territorial interaction may correspond for instance to the fact that individuals belonging to the same territorial grid have a higher number of exchanges than individuals belonging to different territorial grids. Spatial interaction corresponds to the fact that the probability of relation decreases with the geographical distance between individuals (measured in time, cost or kilometers). Social heterogeneity implies that individuals with the same affiliation to a given social category generally have more intense relations than individuals belonging to different groups. Social inequality implies that two individuals with identical social positions will forge relations more easily than individuals with different social positions. But these laws of interaction can in some instances be reversed, especially when individuals developing a relation do so on the basis of complementarities rather than affinity or identity.

Furthermore, it is important to bear in mind that the different forms of interaction distinguished here operate simultaneously and that combined effects of varying complexity may appear, as described in Blau's theorems.

Whatever the precise operative mode of the sub-system of interaction, it is at any rate liable to generate two kinds of systemic modification. First of all, relations may cause mobility, i.e. a modification of sociological or geographical positions of individuals who have entered into a relation. The first effect of such mobility is to alter the position of individuals who have entered into direct contact and thus to increase their future opportunities for relations. But mobility can also generate a modification of the opportunities for relations related to the development of networks that facilitate or hinder opportunities for relations between individuals, all other things being equal in relation to their positions (Wasserman & Faust, 1994). A second consequence of relations relates to the memory of past relations, which bring about an alteration of the rules governing the elementary and systemic constraints of the entire system. Two individuals who have forged an initial relation may for instance enter into contact more easily in the future (self-maintenance of relations). Two hostile groups who have entered into contact may see their hostility increase or attenuate. It is therefore not merely the state of the system but also its parameters of operation that are liable to change over time. The resulting framework for analysis can be presented in Figure 8.

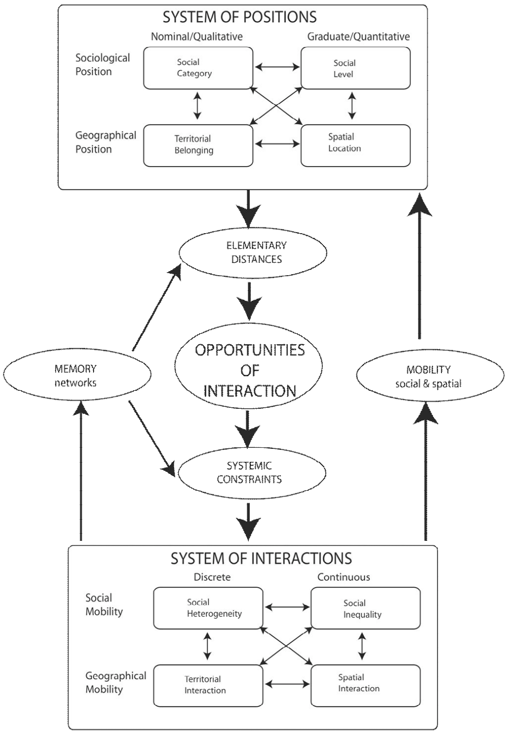


Figure 8: relation between positions and interactions through distances and opportunities

The analysis of euro coin diffusion that we decided to engage with F. Guerin in 2002 is a perfect example of the dilemma that researchers face when trying to analyze complex interactions. We demonstrated very early that the mobility of euro coins at medium and long distances was mainly the product of individual mobility (contrary to banknotes which may be transported over long distance, coins are redistributed on a local scale by the regional bank agencies). Conceiving of euro coin circulation as epidemic disease diffusion (Gould, 1992) could therefore be considered as a global output of the whole geographical moves of people and provide a resulting picture of very different types of movement in space. We decided to focus our analysis on representative surveys of the French population, with samples of 2000 persons chosen randomly according to age, sex, social status, settlement type and of course geographical location. Regular surveys were conducted in France between 2002 and 2008, 3/year at the beginning and 1/year in the most recent period. As these surveys were very expensive, it was not possible to launch a complete survey on the whole countries of the Euro zone, but we were able at least one time to conduct equivalent surveys in Belgium, Germany and Luxembourg, making it possible to benchmark results on both sides of the border.

The important point with such surveys at the individual scale is that the dependent variable (i.e. number of coins of foreign origin in the moneybag) could be explained both by structural attributes (total number of coins, value of coins), by geographical position (place of residence where the survey was realized) and by the person's social attributes (sex, age, activity, income, etc.). It was therefore potentially possible to cross all explanatory factors in the explanation of the number of foreign coins present in the moneybag of individuals and to induce from these results different hypothesis on capital mobility (sociological perspective) or global rules of mobility in the Euro Zone (geographical perspective). But in practice, it appeared very difficult to immediately proceed to such a complex analysis and the first publications were rather analytical, focusing on one scale (individual or territorial units) and one type of attribute (geographical or sociological).

As an example of geographically aggregated model, we can present here a very simple example: a modelization of the diffusion of German coins in France according to distance from the shared border. Individual data was aggregated into spatial administrative units (*départements*) and a topological distance was introduced (shortest path to the border according to contiguity of *départements*). Due to the fact that the survey was representative only at the level of French macro-regions (ZEAT), the choice of *départements* as level of analysis was not fully representative of the territory (Figure 9) and results should be interpreted only after aggregation by classes of topological distance.

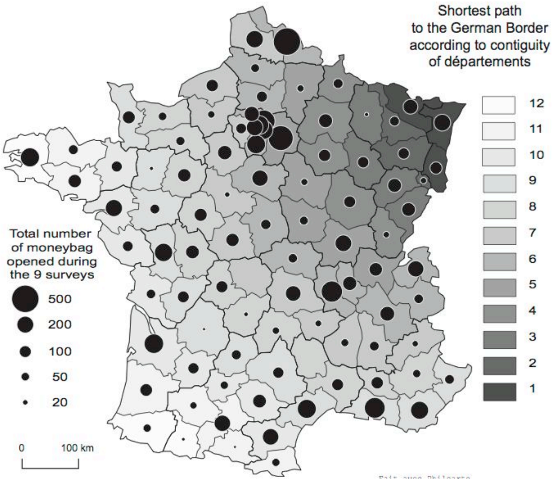
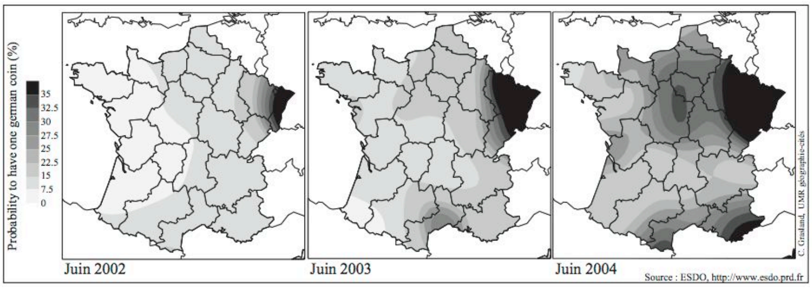


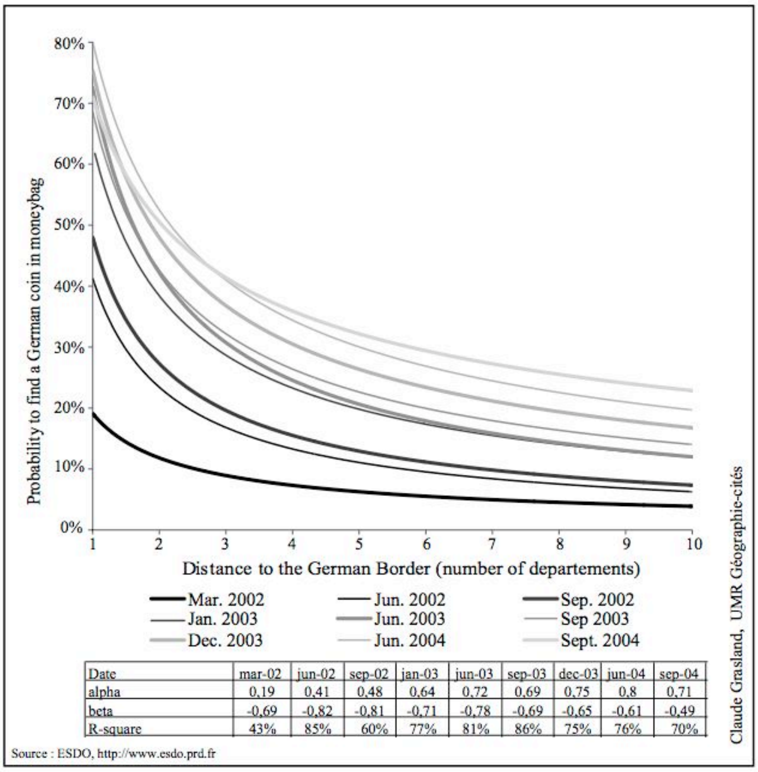
Figure 9: Distance to German border and sample size of euro surveys (2002-2004)

The dependent variable was the proportion of people having at least one German coin in their moneybag, which we considered initially as a better indicator than the proportion of German coins in the total coins of an aggregate of individuals. Indeed, one single person could have a lot of German coins and produce an artificially high proportion for the group that would not reflect the social factor of interest ("how many people have been directly or indirectly in relation with Germany by a social network of coin transportation?").


Probability to have one german coin (%)
35
32.5
30
27.5
25
22.5
15
7.5
0
Juin 2002
Juin 2003
Juin 2004
C. Grasland, UMR géographie-cités

Source : ESDO, http://www.esdo.prd.fr

Figure 10: Map of German coins diffusion in France (2002-2004)



| Date | mar-02 | jun-02 | sep-02 | jan-03 | jun-03 | sep-03 | dec-03 | jun-04 | sep-04 |
|---|---|---|---|---|---|---|---|---|---|
| alpha | 0,19 | 0,41 | 0,48 | 0,64 | 0,72 | 0,69 | 0,75 | 0,8 | 0,71 |
| beta | -0,69 | -0,82 | -0,81 | -0,71 | -0,78 | -0,69 | -0,65 | -0,61 | -0,49 |
| R-square | 43% | 85% | 60% | 77% | 81% | 86% | 75% | 76% | 70% |

Source : ESDO, http://www.esdo.prd.fr

Claude Grasland, UMR Géographie-cités

Figure 11: Modelization of German coins diffusion in France by gravity model (2002-2004)

Looking at the maps (Figure 10), it is clear that the initial diffusion of German coins in France followed a classical diffusion model by proximity even if some anomalies appeared (like peaks in Paris and along the Mediterranean coast). We therefore decided to apply a gravity model of the form p(D)=α. D $^{\beta}$ (with a negative friction parameter β). The probability p for an individual to have one German coin in its moneybag is expressed as a function of the topological distance to the German border D according to a negative Pareto function. As the minimum distance is 1 (*département*s along the border), no problem appears with the distance zero. The parameter α indicates the maximum intensity of diffusion at a very short distance from the border and the parameter β is the distance decay function i.e. the gradient of decrease in the probability to have a German coin (Figure 11).

Even if the size of the samples introduced some noise in the analysis, the model apparently nicely supported classical hypothesis on spatial diffusion, with an early concentration of coins along the border (high value of α reached in June 2003), followed by progressive reduction of the gradient when German coins spread all over the French territory (decrease of β from -0.8/-0.7 to -0.6/-0.5). In other words, exactly what is normally expected by the manual of spatial analysis about spatial diffusion of innovation (Pumain & Saint-Julien, 2000; Saint-Julien, 1985).

This was a nice confirmation of Tobler's first law of geography, but at the same time some doubts appeared because, out of the German case, the situation was not so clear and did not followed the gravity assumptions. In particular, we were very surprised by the fact that Spanish coins were most frequently found in French moneybags than German ones, even after we had removed a bias related to the diffusion of 50cts euro coins by French national bank at the beginning of the process. Belgian coins were also more frequently found in France than German coins. These exceptions could not be explained by simple effects of physical distance and size of the countries as we demonstrated in our presentation at the 13th ECTQG in Lucca in 2003 (Grasland & Guerin-Pace, 2003).

A first way to improve our results was to use more sophisticated modelization of geographical position, and to combine for example continuous effects of distance with discrete effects of barriers. When it was possible to obtain data on diffusion of foreign coins in Belgium, we tried immediately to analyze if the probability to have a French coin depended on distance only or was also related to location in Flemish and Walloons parts of the country (with the specific case of Brussels). The improvement of the model was not only related to the choice of a more complex effect of distance but also to the use of a more disaggregated model at the individual level, making it possible to take into account the effect of the size of the moneybag which follows a very specific statistical distribution (Nuno et al., 2005), not necessarily constant across social groups and which can therefore modify the probability to have foreign coins in one's pocket.

Combining size of moneybag, distance to border and belonging to a linguistic region of Belgium (excluding Brussels), a Poisson regression model of spatial interaction provided a clear demonstration that both spatial interaction (decrease with distance) and territorial interaction (difference of level of diffusion according to linguistic area) accounted for the explanation of the diffusion of French coins in Belgium (Berroir et al., 2006).

Once again, it was nice confirmation of geographical hypothesis and, in my personal case, of the necessary combination of spatial and territorial parameters. But the social dimension was missing. The addition of social parameters (income, age, gender) in the Poisson regression model apparently did not significantly improve the model of diffusion of foreign coins in Belgium, but the sample size was in fact not sufficient to cross so many explanatory variables. In the very early steps of the diffusion process (March and June 2002) we had been able to capture at least one social effect. In the border area of N.E France the probability of having one foreign coin (German, mainly) was significantly higher for active people working in low-level activity groups than for those with higher status. The reverse was observed for internal parts of the territory far from the border. We propose an interpretation of this result related to the mobility of workers across the border in relation with their professional activity. In cross-border areas, people working in Germany or Luxembourg are generally manual workers while people with higher qualifications generally work more in France. In the internal part of the territory, we observe on the contrary a classical situation of higher mobility of people with high level professional activity, inducing regular trips abroad both for work and for leisure. This fascinating result seemed to open the door for an integration of sociological and geographical factors (Grasland et al., 2002a, 2002b). But unfortunately it was not confirmed by further study where this social differentiation of coins apparently disappeared and was not visible in 2003 and following years.

The difficulties that we faced were in fact related to the fact that all of our analysis focused on one level of aggregation (territorial units, individuals) and did not analyze the systemic constraints that were involved in the dynamics of the procedure of diffusion through time. The complexity of the phenomena was related to the fact that different processes of space-time mobility were acting together but with very different characteristics. They could not be captured separately and the resulting picture could not isolate particular effects like tourism contribution to mixture of coins without controlling also the effects of monthly or daily mobility of workers (Grasland & Guerin-Pace, 2004; Grasland et al., 2005). Looking back at the picture of diffusion of German coins between 2002 and 2004 (Figure 10), it is immediately visible that the peak observed along the Mediterranean coast is not related to the same process of diffusion as the gradient observed along the border. And the mystery of domination of Spanish coins can be solved if we enlarge the model to the whole European territory and notice the effect of the intermediary position of France for tourist of northern Europe going on holiday to Spain. British people who are normally not concerned by euros will contribute to the diffusion of French coins in Spain and Spanish coins in France when they go on for holidays by car.

**The articulation of observational and organizational levels**

The external structure of social or spatial systems (Figure 12) also needs to be taken into account theoretically in order to understand its dynamics. The interactions observed between elements correspond to a precise moment (T) and a specific level of organization (L) of spatial or social reality. Yet the reality that is amenable to observation at this level and at this given time is not independent of entrances and exits from other systems occupying different positions in time and within the hierarchy of organizational levels (Durand-Dastès, 1999; Sanders, 2007).

By remaining confined to relations of immediate proximity, the state of the system (L, T) obviously depends first of all on its previous state (L, T-1), which has generated a number of legacies. But it also depends on its future state (L, T+1) to the extent that agents' decisions take account of a range of anticipations pertaining to the future evolution of the system. As for organizational levels, the state of the system (L, T) is the emerging effect of interactions forged between elements at a lower level (L, T-1), but it is also subject to a range of constraints and determinations related to the impact of systems of interaction at a higher level (L, T+1). However, the successive analysis of different organizational levels is not the only possible approach. Increasingly, studies in the field seek to apprehend simultaneously several levels with new statistical instruments of information, especially in the context of multilevel models.

In the case of the analysis of euro coin diffusion, it is only very recently that we were finally able for the first time to demonstrate the existence of social effects related to age, qualification and social status. As explained before, the sample used for Belgium was not sufficient to cross a sufficient number of parameters and the effect of geographical proximity was so strong that other effects were difficult to perceive. But in the case of France, we had the possibility to compile more than 14 surveys. The problem was that, as we had seen in Belgium, the spatial effects of distance were so important in the explanation that it was difficult to depict other social effects when both were introduced in the same model.

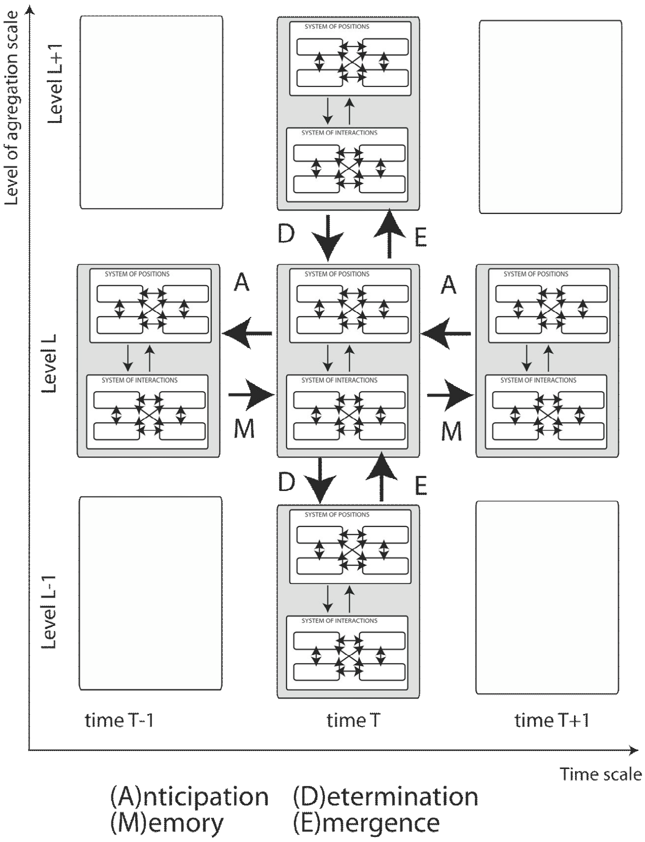


Figure 12: Scale and time levels of analysis

Yet again it was Waldo Tobler who provided indirectly the solution through his distinction between flows and movement (Dorigo & Tobler, 1983). We had tried to analyze the probability to have at least one foreign coin or the number of foreign coins in the moneybag. But we had not taken into account the distance of this foreign coin to its origin and the physical quantity of energy necessary to transport these coins from their origin (abroad) to their current location in France. It is for example important to consider that the discovery of a German coin in a moneybag in Strasbourg (border with Germany) is not exceptional and can be related to a very short distance (less than 5 km). But the discovery of a German coin in Toulouse is something more interesting because, whatever the exact travel, the minimum distance necessary to carry this coin from Germany is at least 800 km and probably more. Considering this idea of movement, we proposed a new variable that evaluates the minimum quantity of movement (energy of transportation) necessary to obtain a given distribution of coins in a moneybag. To do this, we determined the different foreign countries represented in a given moneybag and computed the sum of minimum distance between the place of survey and the nearest border of these countries. French coins were counted as 0 and only foreign coins contributed to the definition of the quantity of movement (minimum number of kilometers) present in the moneybag of a given individual.

In the example presented in Table 1 two moneybags are considered. Moneybag A is made of 15 coins with 9 French, 3 German, 2 Belgian and 1 Spanish coins, which is a rather common distribution. In our previous analysis, this moneybag was considered as equivalent whatever the location of the survey. But now we propose to take into account this location and we will therefore provide different evaluations of this moneybag if it is observed in Strasbourg (coins average distance to origin is equal to 76 km), in Paris (129 km) or Nice (189 km). In the second example, the moneybag involves only 8 coins but with a very exceptional distribution including 5 French coins, 1 from Ireland, 1 from Greece and 1 from Finland. In this case, we can consider that the average distance of origin of coins is very high, whatever the location in France (between 488 and 556 km).

Many improvements can be added to this measure of movement, taking into account the size of the country of origin, the possible clustering effects in coin transportation and, last but not least, the value of coins as we have demonstrated very early that the smallest the value, the lowest the probability of long distance circulation (Grasland et al., 2002). Nevertheless, we will demonstrate that a simple modelization of the raw number of kilometers involve in a moneybag makes it possible to obtain very interesting results on social inequalities in terms of international mobility in the Euro area.

| | **Minimal distance** | | | **Moneybag A** | | | **Moneybag B** | | |
|---|---|---|---|---|---|---|---|---|---|
| | (km, grand circle) | | | (3 AL, 2 BE, 1 ES, 9 FR) | | | (1 FI, 1 IR, 1 GR, 5 FR) | | |
| Country | Stras. | Paris | Nice | Stras. | Paris | Nice | Stras. | Paris | Nice |
| AUT | *191* | *555* | *394* | 0 | 0 | 0 | 0 | 0 | 0 |
| BEL | *166* | *179* | *633* | 332 | 358 | 1267 | 0 | 0 | 0 |
| ALL | *15* | *299* | *402* | 46 | 897 | 1207 | 0 | 0 | 0 |
| ESP | *766* | *680* | *359* | 766 | 680 | 359 | 0 | 0 | 0 |
| FIN | *1493* | *1665* | *1982* | 0 | 0 | 0 | 1493 | 1665 | 1982 |
| FRA | *0* | *0* | *0* | 0 | 0 | 0 | 0 | 0 | 0 |
| GRE | *1349* | *1679* | *1107* | 0 | 0 | 0 | 1349 | 1679 | 1107 |
| IRL | *1060* | *715* | *1359* | 0 | 0 | 0 | 1060 | 715 | 1359 |
| ITA | *247* | *477* | *25* | 0 | 0 | 0 | 0 | 0 | 0 |
| LUX | *131* | *262* | *619* | 0 | 0 | 0 | 0 | 0 | 0 |
| NLD | *265* | *276* | *765* | 0 | 0 | 0 | 0 | 0 | 0 |
| POR | *1322* | *1031* | *1117* | 0 | 0 | 0 | 0 | 0 | 0 |
| **PKM: sum of distances** | | | | **1145** | **1935** | **2834** | **3901** | **4059** | **4448** |
| **KM: average distance of coins** | | | | **76** | **129** | **189** | **488** | **507** | **556** |

Table 1: Estimation of the internationalization of a moneybag by minimum quantity of movement

The model used in this analysis is a Poisson regression model describing the total number of kilometers involved in the moneybag of an individual (*PKM*) according to the size of moneybag (*Nbcoins*), the date of survey (*Time*) and various social parameters describing age and sex structure (*AgeSex*), level of standing of the household based on a synthesis of various parameters (*Standing*), level of education of the individual (*Education*) and finally field of activity with a simple distinction between inactivity, manual workers and others (*Activity*). A scale parameter (*Scale*) is introduced and automatically estimated in order to avoid biases introduced by the unit of measurement of distance. Without this scale parameter, the significance of parameters would be biased and depend on the choice of units in kilometers, miles, etc.

$$KM_i = \; scale \times \exp\begin{bmatrix} a_0 \\ +a_1 \log(Nbcoins) \\ +a_{2,1}Time_1 + .... + a_{2,T-1}Time_{T-1} \\ +a_{3,1}AgeSex_1 + .... + a_{3,A-1}AgeSex_{A-1} \\ +a_{4,1}S\tan ding_1 + .... + a_{4,S-1}S\tan ding_{S-1}\,. \\ +a_{5,1}Education_1 + .... + a_{5,E-1}Education_{E-1} \\ +a_{6,1}Activity_1 + .... + a_{6,W-1}Activity_{W-1} \end{bmatrix} + \varepsilon_i$$

In the case of qualitative variables with N modalities, only (N-1) are introduced in the model as dummy variables and the remaining one defines the effect of the intercept parameter (a0). In our application, the intercept therefore refers to a position which is chosen as reference and is, in this particular case, *the moneybag opened in December 2007 and belonging to a man more than 65 years old with high socio-professional status and high level of education, retired* (the characteristics of W. Tobler, if he had been living in France in 2007…). We could have chosen any other position as reference without changing the result of the model and it is only for empirical comments of the results that this reference is useful.

Before we analyze the detailed results for each position (Table 2), we would like to underline the main discovery which is that all parameters of social position introduced in the model are statistically very significant which means that all contribute to introducing significant differences in the internationalization of moneybags, all things being equal with the number of coins, the time period and the other social parameters introduced as competitor for explanation. More precise results are the following ones:

**The number of coins is less than proportional to the number of kilometers (0.93 <1)** which means that individuals with big moneybag are generally less likely to have a high proportion of coins coming from long distance than people with small moneybag. The size of moneybag is indeed an interesting topic that can be described by a mathematical model (Nuno et al., 2005) and can reveals interesting variations from one country to another or between social groups.

**The internationalization of moneybags increases through time but with some seasonal fluctuations**. The value of the time parameter indeed increased from -1.999 to 0 between March 2002 and December 2007, but the trend is not linear and some positive residuals appears generally in September (due to the effect of great holyday migration) and negative residuals in December or January (when fewer contacts are observed).

**Gender differences are observed in favor of men, but with variations according to age**. Whatever the age group considered, the degree of internationalization of a moneybag is higher for men than for women, suggesting that men are more in contact with international networks or internationalized areas than women. At the same time, we notice that all things being equal with gender, the internationalization of a moneybag is highest for young adults (probably students) and then decreases until 45 years old. A second peak is observed around 60 years old (probably young retired people), followed by a decrease for the older people. As a summary, the highest probability to have an internationalized moneybag is observed for young men (+31% as compared to the reference) and the lowest probability is observed for middle-age women (-11%).

**Social inequalities between households are correlated with international mobility**. Even if the standing variable is not very clearly defined by the Institute that conducted the survey (a mixture of parameters including income, property, etc…), the analysis clearly demonstrate that the internationalization is lower for people with lowest standing (-6.3% for the lowest level as compared to the highest). But there are no differences with the middle group that has in fact the highest level of internationalization (+1.1%). The level of education displays exactly the same conclusion with clear differences for people with the lowest level (-6.0%) and a small advantage for the middle group (+1.1%). In both cases, we can suspect that the observed effect could be partly ecological and related to the place where people are living (and where they have more or fewer probabilities to find foreign coins). It is not necessarily their own mobility that is responsible for the internationalization of their moneybag.

**There are strong differences between blue collars and white collars but not between working and not working people**. The categories are of course very fuzzy and subject to criticism, but it is nevertheless very clear that blue collars are less likely to be in contact with international coins (-14.1%) than the others. On the other hand, white collars are not characterized by higher probability of internationalization as compared to non-working people, despite the facts that we controlled the effect of age and sex.

| **Parameter** | | | **Parameter estimate** | | **Significativity** | | |
|---|---|---|---|---|---|---|---|
| **name** | **category** | **DF** | **value** | **effect** | **Standard error** | **Khi2** | **Pr > Khi 2** |
| **Intercept** | | 1 | 4,811 | | 0,073 | 4391,17 | <.0001*** |
| **Nbcoins** | | 1 | 0,936 | | 0,017 | 2911,97 | <.0001*** |
| **time** | 2002_03 | 1 | -1,999 | -86,5% | 0,117 | 289,86 | <.0001*** |
| **time** | 2002_06 | 1 | -1,715 | -82,0% | 0,075 | 527,39 | <.0001*** |
| **time** | 2002_09 | 1 | -1,095 | -66,5% | 0,059 | 339,07 | <.0001*** |
| **time** | 2003_01 | 1 | -1,155 | -68,5% | 0,062 | 343,02 | <.0001*** |
| **time** | 2003_06 | 1 | -1,155 | -68,5% | 0,062 | 345,29 | <.0001*** |
| **time** | 2003_09 | 1 | -0,776 | -54,0% | 0,055 | 201,91 | <.0001*** |
| **time** | 2003_12 | 1 | -0,723 | -51,5% | 0,051 | 200,80 | <.0001*** |
| **time** | 2004_06 | 1 | -0,539 | -41,7% | 0,050 | 115,46 | <.0001*** |
| **time** | 2004_09 | 1 | -0,310 | -26,7% | 0,047 | 43,85 | <.0001*** |
| **time** | 2004_12 | 1 | -0,369 | -30,9% | 0,045 | 66,40 | <.0001*** |
| **time** | 2005_06 | 1 | -0,273 | -23,9% | 0,047 | 34,38 | <.0001*** |
| **time** | 2005_12 | 1 | -0,130 | -12,2% | 0,044 | 8,86 | 0.0029*** |
| **time** | 2006_06 | 1 | -0,074 | -7,1% | 0,045 | 2,63 | 0.1046- |
| **time** | ***2007_01*** | ***0*** | ***0,000*** | **-** | ***0,000*** | **-** | **.** |
| **Age structure** | W15-24 | 1 | 0,042 | 4,3% | 0,058 | 0,53 | 0.4678- |
| **Age structure** | W25-34 | 1 | 0,025 | 2,6% | 0,051 | 0,24 | 0.6215- |
| **Age structure** | W35-44 | 1 | -0,026 | -2,6% | 0,052 | 0,25 | 0.6145- |
| **Age structure** | W45-54 | 1 | -0,117 | -11,0% | 0,055 | 4,46 | 0.0346* |
| **Age structure** | W55-64 | 1 | 0,049 | 5,0% | 0,053 | 0,84 | 0.3582- |
| **Age structure** | W65 + | 1 | -0,031 | -3,1% | 0,050 | 0,38 | 0.5356- |
| **Age structure** | M15-24 | 1 | 0,271 | 31,2% | 0,060 | 20,64 | <.0001*** |
| **Age structure** | M25-34 | 1 | 0,116 | 12,3% | 0,059 | 3,95 | 0.0470* |
| **Age structure** | M35-44 | 1 | 0,076 | 7,9% | 0,059 | 1,67 | 0.1964- |
| **Age structure** | M45-54 | 1 | 0,089 | 9,3% | 0,058 | 2,41 | 0.1208- |
| **Age structure** | M55-64 | 1 | 0,117 | 12,4% | 0,056 | 4,37 | 0.0366* |
| **Age structure** | ***M65 +*** | ***0*** | ***0,000*** | **-** | ***0,000*** | **-** | **.** |
| **Standing** | Low | 1 | -0,066 | -6,3% | 0,029 | 5,02 | 0.0251* |
| **Standing** | Medium | 1 | 0,011 | 1,1% | 0,026 | 0,16 | 0.6882 |
| **Standing** | ***High*** | ***0*** | ***0,000*** | **-** | ***0,000*** | **-** | **.** |
| **Diplom** | Low | 1 | -0,062 | -6,0% | 0,031 | 4,17 | 0.0411* |
| **Diplom** | Medium | 1 | 0,011 | 1,1% | 0,036 | 0,09 | 0.7659 |
| **Diplom** | ***High*** | ***0*** | ***0,000*** | **-** | ***0,000*** | **-** | **.** |
| **Activity** | White collars | 1 | 0,037 | 3,8% | 0,028 | 1,75 | 0.1863 |
| **Activity** | Blue collars | 1 | -0,152 | -14,1% | 0,043 | 12,58 | 0.0004*** |
| **Activity** | ***Inactive*** | ***0*** | ***0,000*** | **-** | ***0,000*** | **-** | **.** |
| **Scale parameter** | | 0 | 41,087 | | 0,000 | | |

Source of data: Euro Spatial Diffusion Observatory – http://www.esdo.prd.fr

Table 2: A poisson regression model of social diffusion of international foreign coins in France (2002-2007)

These fascinating results - presented here for the first time - have many consequences in both empirical and theoretical fields. For example, in political sciences, they could provide a better understanding of the attitude of French citizens toward the European Union, according to their social and spatial position.

The categories that rejected the Lisbon Treaty in France are more or less exactly the one that have been described here as the less in contact with internationalized networks. We will discuss it in further publications, but we hope that this example will at less convince the reader of the interest of the theoretical framework that we have elaborated here, thanks to Waldo Tobler and Peter Blau.

## CONCLUSION: SPATIAL ANALYSIS AND SOCIAL MORPHOLOGY

In conclusion, I would like to enlarge my point of view to the more general question of the borders between academic disciplines, with a particular focus on Geography and Sociology. In my view, a great opportunity of progress for a unitary Social Science appeared at the end of the 19th century when E. Durkheim elaborated the concept of social morphology. But unfortunately, this opportunity was not realized at this moment of history where each discipline was trying to build its own identity. One century later, I suggest that it could be time to rediscover this idea of Social Morphology, which is very near to many practices of Spatial Analysis.

Spatial analysis may be defined as the body of statistical, mathematical, cartographic and IT methods that aim to describe, measure, analyze or modelize spatial configurations and the evolution and movements of objects or events (men, animals, plants, cultures, volcanoes, temperatures, altitudes, roads, activities, etc.). As a realm of theoretical research, spatial analysis provides measuring instruments that apply to abstract geometrical objects (points, surfaces, lines, networks) and which are characterized by qualitative and quantitative properties that are liable to change over time. The instruments of spatial analysis thus define general forms of distribution and measurements that are independent of the content of the particular objects they are designed to describe. On the face of it, the abstraction of spatial forms enabled by the instruments of spatial analysis does not pertain to any particular disciplinary sphere. But some geographers, notably W. Bunge (1966), have construed spatial analysis as the methodological core of a general theoretical geography (Systematic Geography) that aims to describe the totality of forms and processes that occur on the earth's surface, whether they pertain to the natural world or to the social world. Bunge argues that spatial analysis is not merely a descriptive instrument adapted to categories of objects or heterogeneous phenomena, but that it helps to discover general laws in their distribution at the earth's surface. On the basis of previous research, Bunge has shown for instance that the process governing the successive courses of the Mississippi was not merely analogous to the process that helps to describe the successive itineraries of the Seattle-Tacoma highway, but that both processes are subject to the same account by strictly spatial determinants. According to Bunge, similar arguments apply to the growth of punctual objects such as towns or volcanoes that grow in successive accumulations around a centre, with their density gradient governed by mathematical and to a certain extent predictable rules. Close in this respect to the position later defended by H. Reymond (1981), Bunge sees in the problem of spacing the core of geographical method and the only approach ever likely to provide the foundations of a genuinely scientific domain of inquiry.

According to Durkheim, social morphology is the study of the material substrate of social life, i.e. the human milieu, and of the social structures that function simultaneously as the causes and the consequences of the macroscopic evolution of human societies. In The Rules of Sociological Method, social morphology is defined briefly as “the number and nature of elementary parts that constitute any given society, the way in which such elements are organized, their degree of coalescence, the distribution of population throughout the territory in question, the number and nature of communication routes, the forms of habitats, etc.” (Durkheim, 1895). Durkheim proposed a more precise definition in 1899 when the ‘miscellaneous’ section of L’Année Sociologique was renamed ‘social morphology’: “Social life is based on a substrate determined in size and form. It is constituted by the mass of individuals that compose it, the way in which they occupy its territory, and by the nature and configuration of the range of determinants affecting collective relations. The social substrate varies according to the size and density of the population, the degree to which the population is predominantly urban or rural, the way in which towns and houses are built, the size of the territory effectively occupied by the population, the degree to which the territory is delimited by borders, the nature of its communication routes, etc. On the other hand, the constitution of the substrate directly or indirectly affects all social phenomena, in the same way that mediate and immediate psychological phenomena are related to the state of the brain. We thus have a broad range of issues that are clearly of interest to sociologists and which, since they refer to a common object of inquiry, must pertain to a common realm of scientific inquiry. Such is the core of the science that I propose to call social morphology” (Mucchielli & Robic, 1996).

A number of studies in the epistemology and history of geography and sociology have observed that the idea of creating a multidisciplinary field of research that brings together sociologists, geographers, demographers and economists was not without ulterior motives on the part of Durkheim and his disciples (Mucchielli & Robic, 1996; Rhein, 1982). In many ways, social morphology was conceived by Durkheim as a ‘war machine’ designed to counter the almighty power of history, a means for sociology of asserting itself (by means of a critique of geography, which had been unable to integrate its questions) and an attempt to occupy a promising field of research before other subjects could stake a claim to the very same disciplinary terrain. Though other disciplines responded positively to Durkheim's appeal (which was not the case of geographers in the tradition of Vidal), sociologists could rightfully claim to be the first occupants and thus cement their hold over the research carried out in other disciplines. A century later, the significance of these quarrels has nonetheless considerably weakened. A more productive approach would be to construe Durkheim's proposition literally by reflecting on the usefulness of resurrecting a multidisciplinary field of research centered on social morphology, at the cost of making some minor adjustments to Durkheim's original definition.

According to our theoretical framework, social morphology could be defined as the study of the impact of the position of individuals or groups on the constitution of bonds that may develop between them within a given society. The notion of position is sufficiently broad to include the contributions of sociology and geography within the domain of social morphology. But it also implies other disciplines in the social sciences capable of making a significant contribution of their own (economics, history, and demography). Still, even if social morphology is provisionally restricted to the study of geographical and sociological positions, it is clear that the definition outlined here overlaps to a significant degree with Durkheim's own definition. The notion of geographical position includes phenomena of concentration, dispersion, and accessibility, as well as the role of territorial grids, whether political (borders) or administrative (regions, communes). In theory the concept of sociological position helps to define the elements that constitute any given society, i.e. 'the number and nature of its elementary parts', in Durkheim's earliest definition. An inquiry into the specific contribution of geographical analysis to Durkheim's social morphology outlined over a century ago therefore raises the question of the current connections between geography and sociology. It also raises the question of the identity of geography within the human and social sciences generally (Rhein, 2002).

## ACKNOWLEDGEMENTS

This paper is an individual proposal, but grounded and embedded in a collective community of friends, researchers and students with which I have discussed so many times the issue presented here during the last 20 years. I would first like to thanks my friends **Athanase Bopda** and **Jean-Marc Vincent** who have always encouraged me to delve deeper into the theoretical analysis of the subject, from both sociological and mathematical points of view. Then, I would mention **Violette Rey** and **Thérèse Saint-Julien** who are also friends but were previously director of my PhD (1991) and Habilitation à Diriger des Recherches (1998) and – as far as I know – suffered a lot when they had to correct my manuscripts. I have a collective debt to the members of the Equipe Paris of UMR Géographie-cités. I developed with them my research, until the present provisional conclusion: **Nadine Cattan**, for the research on spatial interaction models and barriers effects; **Renaud Le Goix** and **Jean-Christophe François**, for the seminary named 'Morphologie Spatiale et Morphologie Sociale' that we built together in 1999-2001 and where we invited **Linton Freeman** and **Waldo Tobler; Lena Sanders** and **François Durand-Dastès** who pushed me toward system and model analysis and obliged me to read more international literature before elaborating my own assumptions; **France Guerin-Pace**, for the building of the survey on euro coins; and last but not least, **Helene Mathian** for her incredible support in statistical and spatial analysis. Out of my research team, the reflection was also elaborated with many colleagues in France and other countries, in particular **Jean-Michel Decroly** for the research on spatial discontinuities, **Michel Poulain** for barriers effects, **Stanislas Rehak** for the first analysis of migration models at the individual scale, along the Czech Slovak border. I mention specifically **Denise Pumain** and **Marie-Claire Robic** as directors of the journal l'Espace Geographique who pushed me for the first time to try to publish a theoretical paper on the results of my Habilitation à Diriger des Recherches in 2002. After an excellent critical work of the reviewers (in particular **Catherine Rhein**), the paper was accepted but it was my own decision to delay it as I was finally not convinced to have sufficiently deepened the subject. I will not forget all **my undergraduate and graduate students at the Universities Paris 1 and Paris 7** who helped me during 20 years to precise my personal view on spatial analysis of social facts: when you don't know something, teach it. Finally, I would like to thanks the editors of the present Handbook, **François Bavaud** and **Christophe Mager**, who proposed me to write this chapter and supported me during the last 2 years by friendly but efficient pressure to do the job.

All people mentioned here contributed to the elaboration of the present work, but of course I assume alone the criticism that could be addressed to it.

## REFERENCES

**Barnes, T.J.**, 2004, A Paper Related to Everything but More Related to Local Things. In: Annals of the Association of American Geographers, 94, 2: 278-283

**Beaverstock, J.V.; Smith, R.H.; Taylor, P.J.,** 2000, World City Network: A New Metageography? In: Annals of the Association of American Geographers, 90, 1: 123-134

**Berroir, S.; Grasland, C.; Guerin-Pace, F.; Hamez, G.,** 2006, La diffusion spatiale des pièces euro étrangères en Belgique et en France. In: BELGEO (Revue Belge de Géographie), 5: 793-822

**Blau, P.M.,** 1956, Social Mobility and interpersonal relations. In: American Sociological Review, XXI: 290-295

**Blau, P.M.,** 1957, Formal organisation: dimensions of analysis. In: American Journal of Sociology, LXIII: 58-69.

**Blau, P.M.,** 1977a**,** A macro sociological theory of social structure. In: American Journal of Sociology, 83, 1: 26-54

**Blau, P.M.,** 1977b, Inequality and heterogeneity. Free Press: New York

**Blau, P.M.; Schwartz, J.E.,** 1984, Crosscutting Social Circles. Academic Press: Orlando

**Blau, P.M.**, 1986, Exchange and power in social life. Transaction Publishers: New Jersey

**Blau, P.M.,** 1993, Multilevel Structural Analysis. In: Social Networks, 15: 201-215

**Blau, P. M.,** 1994, Structural Contexts of Opportunities. University of Chicago Press: Chicago

**Bopda, A.,** 2003, Yaoundé et le défi camerounais de l'intégration: a quoi sert une capitale d'Afrique tropicale. Editions du CNRS: Paris

**Bopda, A.; Grasland, C.,** 1994, Migrations, régionalisations et régionalismes au Cameroun. In: Espace, Population, Sociétés, 4: 109-129

**Bopda, A.; Grasland, C.; Poulain, M.,** 2000, Évaluation comparative de l'influence des limites linguistiques sur les comportements migratoires. Applications au cas de la Belgique, du Cameroun et de la Tchécoslovaquie. In: Actes du 6e Colloque international de l'AIDELF, Régimes démographiques et territoires: les frontières en question, 22-26 September. La Rochelle, Paris: Presses Universitaires de France: 107-124

**Boudon, R.,** 1970, L'analyse mathématique des faits sociaux. Plon: Paris

**Boudon, R.,** 1984, La place du désordre. PUF-Quadrige: Paris

**Brunet, R.,** 2001, Le Déchiffrement du monde. Théorie et pratique de la géographie. Belin: Paris

**Bunge, W.,** 1966, Theoretical Geography. Lund Studies: Berlingska Boktryckeriet

**Castells, M.,** 1996, The Rise of the Network Society, The Information Age: Economy, Society and Culture, Vol. I. Blackwell: Cambridge, MA-Oxford, UK

**Cattan, N.**, 2004, Le monde au prisme des réseaux aériens. In: Flux, 58: 32-43

**Cattan, N.; Grasland, C**., 1996, Air Traffic Fields of Western European Cities. In: Pumain, D.; Saint-Julien, T. (Eds), Urban Networks in Europe. John Libey Eurotext / INED: 115-128

**Degenne, A.; Forsé, M**., 1994, Les réseaux sociaux. Masson: Paris

**Derudder, B.; Witlox, F.; Taylor, P.J.,** 2007, Les villes dans les réseaux mondiaux: une nouvelle méthodologie pour cartographier la position relationnelle des villes. In: Revue d'Economie Régionale et Urbaine, 2: 179-200

**Dorigo, G.; Tobler, W.R.,** 1983, Push-Pull Migration Laws. In: Annals of the Association of American Geographers, 73, 1: 1-17

**Durand-Dastès, F.,** 1999, Systèmes et modèles (compilation of publications). (CD ROM) CNRS-UMR 8504

**Durand-Dastès, F.; Grataloup, C.; Levallois, A.,** 1992, Le rôle des flux dans l'organisation des ensembles spatiaux. In: L'information géographique, 1: 35-42

**Durkheim, E.**, 1895, Règles de la méthode sociologique. Quadrige-PUF: Paris. Reprinted in 1993

**Durkheim, E.,** 1897, Le Suicide. Quadrige-PUF: Paris. Reprinted in 1995

**Ethington, P.J.**, 1997, The intellectual construction of 'social distance': toward a recovery of Georg Simmel's Social Geometry. In: Cybergéo, 30. http://www.cybergeo.presse.fr

**Fotheringham, A.S.; O'Kelly, M.E.,** 1989, Spatial Interaction Models: Formulations and Applications. Kluwer Academic Publishers: London

**Freeman, L.,** 1999, Morphologie spatiale et morphologie sociale. Paper presented at the MS2 Conference, CNRS, Paris, France

**Goodchild, M.F.,** 2004, The Validity and Usefulness of Laws in Geographic Information Science and Geography. In: Annals of the Association of American Geographers, 94, 2: 300-303

**Gould, P.,** 1992, Epidémiologie et maladie. In: Bailly, A.; Ferras, R.; Pumain, D. (Eds), Encyclopédie de Géographie. Paris: Economica: 949-969

**Grasland, C.,** 1997, Contribution à l'analyse géographique des maillages territoriaux, Habilitation à diriger des recherches, Université Paris 1, Paris, France

**Graland, C.,** 2007, Global City revisited by Tobler's first law of Geography. In: Proceedings of the 15th European Colloquium on Theoretical and Quantitative Geography, September 7-11, Montreux, Switzerland. Université de Lausanne: Lausanne: 175-180

**Grasland, C.; Guerin-Pace, F.; Garnier, B.; Tostain, A.,** 2002a, L'euro gagne du terrain. In: Pour la Science, 381: 10-11

**Grasland, C.; Guerin-Pace, F.; Garnier, B.,** 2002b, La circulation des euros, reflet de la mobilité des hommes. Population & Sociétés, 384. INED: Paris

**Grasland, C.; Guerin-Pace, F.,** 2003, A simulation of Euro coins diffusion. Paper presented at the 13th European Colloquium on Theoretical and Quantitative Geography, Lucca, Italy, 5-9 September

**Grasland, C.; Guerin-Pace, F.,** 2004, Mobilité européenne, tourisme et diffusion des pièces euros étrangères en France. In: Revue d'Economie Régionale et Urbaine, 5: 793-822

**Grasland, C.; Guerin-Pace, F.; Terrier, C.,** 2005, Interaction spatiale et réseaux sociaux. Modélisation des déterminants de la diffusion internationale des pièces euro (2002-2004). IXe Journées de Méthodologie Statistique, Actes JMS2005. INSEE: Paris. http://jms.insee.fr

**Granovetter, M.S**., 1973, The strength of weak ties. In: American Journal of Sociology, 78: 1360-1380

**Guimera, R.; Mossa, S.; Turtschi, A.; Amaral, L.A.N.,** 2005, The Worldwide air transportation network: Anomalous, centrality, community structure and citie's global role. In: PNAS, 102, 22: 7794-7799

**Harvey, D.,** 1969, Explanation in Geography. Arnold: London

**Hägerstrand, T.,** 1952, The propagation of innovation waves. In: Lund Studies in Geography, B, 4

**Hägerstrand, T.,** 1953, Innovation diffusion as a spatial process. Translation by Pred, A., 1967. University of Chicago Press: Chicago

**Lizardo, O.,** 2006, Peter Blau. In: Ritzer, G. (Ed.), Encyclopedia of Sociology. Blackwell: New York

**Maurel, M.C.,** 1982a, Bureaucratie et contrôle territorial. Le maillage rural en URSS et en Pologne. In: Hérodote, 25: 49-75

**Maurel, M.C.,** 1982b, Territoire et stratégies soviétiques. Economica: Paris

**Maurel, M.C.,** 1984, Pour une géopolitique du territoire: le maillage politico-administratif. In: Hérodote, 33-34: 131-143

**Miller, H.J.,** 2004, Tobler 's First Law and Spatial Analysis. In: Annals of the Association of American Geographers, 94, 2: 284-289

**Mucchielli, L.; Robic, M.-C.,** 1996, La morphologie sociale selon Durkheim: entre sociologie et géographie. In: Muchhielli, L.; Bolandi M. (Eds), La sociologie et sa méthode. Les règles de Durkheim un siècle après. Paris: L'Harmattan: 101-136

**Nuno, J.-C.; Grasland, C.; Blasco, F.; Guérin-Pace, F.; Olarrea, J.; Luque, N.,** 2005, How many coins are you carrying in your pocket? In: Physica A, 354: 432-436

**Park, R.,** 1924, The concept of social distance as applied to the study of racial attitudes and racial relations. In: Journal of Applied Sociology, 8, 6: 339-344

**Philips, J.D.,** 2004, Doing justice to the law. In: Annals of the Association of American Geographers, 94, 2: 290-293

**Pini, G**., 1992, L'interaction spatiale. In: Bailly, A.; Ferras, R.; Pumain, D. (Eds), Encyclopédie de Géographie: 557-576

**Piron, M.**, 1993, Changer d'échelle: une méthode pour l'analyse des systèmes d'échelles. In: Espace Géographique, 2: 147-165

**Poupeau, F.; François, J.C.,** 2009, Le sens du placement. Ségrégation résidentielle et ségrégation scolaire. Editions Raisons d'agir: Paris

**Pumain, D.; Saint-Julien, T.,** 1997, Analyse spatiale t.1: localisations dans l'espace. Armand Colin, Cursus – Géographie: Paris

**Pumain, D.; Saint-Julien, T.,** 2000, Les interactions spatiales. Armand Colin, Cursus – Géographie: Paris

**Raffestin, C.,** 1978, Les construits en géographie humaine: notions et concepts. In: Géopoint: 55-73

**Raffestin, C.**, 1980, Pour une géographie du pouvoir. Litec: Paris

**Ratzel, F.,** 1897, La Géographie Politique - les concepts fondamentaux. Paris: Fayard. First published in 1897

**Rey, V.,** 1984, Sur la pertinence géographique du système national. In: Géopoint: 91-95

**Rey, V.,** 1987, Intégration territoriale et crise spatiale dans les pays d'Europe de l'Est. In: Bulletin de la Société Languedocienne de Géographie, 21, 1-2: 21-28

**Reymond, H.,** 1981, Une problématique pour la géographie: plaidoyer pour une chorotaxie expérimentale. In: Isnard, H.; Racine, J.B.; Reymond, H. (Eds), Problématiques de la géographie. Paris: PUF: 163-249

**Rhein, C., 1982,** La géographie: discipline scolaire et/ou science sociale. In: Revue Française de Sociologie, XXIII, 2: 223-251

**Rhein, C.,** 2002, Intégration sociale, intégration spatiale. In: L'espace géographique, 3: 193-207

**Rushton, G**., 1993, Human behavior in spatial analysis. In: Urban Geography, 5: 447-456

**Sack, R.D.,** 1983, Human territoriality: a theory. In: Annals of the Association of American Geographers, 73, 1: 55-74

**Saint-Julien, T.,** 1985, La diffusion spatiale des innovations. Reclus-Mode d'emploi: Montpellier

**Sanders, L.,** 2007, Models in Spatial Analysis. ISTE: London

**Sassen, S.,** 2002, Global Networks / Linked Cities. Routledge N.Y.: New York

**Schwartz, J.E**., 1990, Penetrating differenciation. In: Calhoun, C.; Meyer, M.W.; Scott, W.R. (Eds), Structures of Power and Constraint. Cambridge: Cambridge University Press: 353-374

**Scott, W.R.; Calhoun, C.,** 2004, Biography of Peter Michael Blau (1918-2002). National Academy of Sciences, Biography & Autobiography: Washington

**Sen, A.; Smith, E.T.**, 1995, Gravity models of spatial interaction behavior. Springer Verlag: Berlin

**Simmel, G.**, 1908, Soziologie. Duncker and Humboldt: Leipzig. First edition in 1908

**Simmel, G.,** 1894, Le problème de la sociologie. Sociologie et Epistémologie. PUF: Paris. Reprinted in 1981

**Simmel, G.**, 1896, Comment les formes sociales se maintiennent. Sociologie et Epistémologie. PUF: Paris. Reprinted in 1981

**Smith, J.M.,** 2004, Unlawful Relations and Verbal Inflation. In: Annals of the Association of American Geographers, 94, 2: 294-299

**Stouffer, S.A**., 1940, Intervening opportunities: a theory relating mobility and distance. In: American Sociological Review, V: 845-867

**Stouffer, S.A.,** 1960, Intervening opportunities and competing migrants. In: Journal of Regional Science, 2: 1-26

**Stouffer, S.A.,** 1962, Social reserach to test ideas. Glencoe, Ill. The Free Press: New York

**Sui, D.Z.**, 2004, Tobler's First Law of Geography: A Big Idea for a Small World? In: Annals of the Association of American Geographers, 94, 2: 269-277

**Taylor, P.J.,** 2005, New political geographies: Global civil society and global governance through world city networks'. In: Political Geography, 24: 703-730

**Taylor, P.J.; Derudder, B.; Saey, P.; Witlox, F.,** 2006, Cities in Globalization: Practices, policies and theories. Routledge: London

**Tobler, W.R.,** 1969, Geographical filters and their inverses. In: Geographical Analysis, 1: 234-253

**Tobler, W.R.,** 2004, On the First Law of Geography: A Reply. In: Annals of the Association of American Geographers, 94, 2: 290-293

**Tobler, W.R.**, 1970, A computer Movie Simulating Urban Growth in the Detroit Region. In: Economic Geography, 46, Supplement: Proceedings. In: IGU Commission on Quantitative Geography: 234-240

**Tobler, W.R.; Mielke, H.W.; Detwyler, T.R.**, 1970, Geobotanical Distance between New Zealand and Neighboring Islands. In: BioScience, 20, 9: 537-542

**Tobler, W.; Wineburg, S.,** 1971, A Cappadocian Speculation. In: Nature, 5297, 231: 39-41

**Tobler, W.R**., 1981, A model of geographical movement. In: Geographical Analysis, 13, 1: 1-20

**Wasserman. S.; Faust. K.,** 1994, Social network analysis. Methods and applications. Cambridge University Press: New York

## AUTHOR INFORMATION

**Claude GRASLAND**
Claude.grasland@parisgeo.cnrs.fr
Université Paris Diderot Paris 7
UMR 8504 Géographie-cités, CNRS, Université Paris 1, Université Paris Diderot, Paris, France

# MACHINE LEARNING MODELS FOR GEOSPATIAL DATA

**Mikhail KANEVSKI[*], Loris FORESTI[*], Christian KAISER[**], Alexei POZDNOUKHOV[*], Vadim TIMONIN[*] and Devis TUIA[*]**

[*]Institute of Geomatics and Analysis of Risk, University of Lausanne, Lausanne, Switzerland, [**]Institute of Geography, University of Lausanne, Lausanne, Switzerland

## ABSTRACT

This chapter presents an introduction to machine learning models/algorithms and their potential applications to geospatial data. The main attention is paid to widely used models which are based on artificial neural networks (multilayer perceptron, general regression neural networks, self-organizing maps) and statistical learning theory (support vector machines). The main ideas of spatial classification, spatial predictions/mapping including automatic algorithms, nonlinear dimensionality reduction and visualization of high dimensional multivariate socio-economic data, treatment and classification of remote sensing images by applying machine learning are illustrated using real data case studies.

## KEYWORDS

Machine learning algorithms, Geospatial data, Mapping and classification, Dimensionality reduction, Remote sensing

## INTRODUCTION

Machine learning (ML), in a general framework, can be considered as a subfield of artificial intelligence that is concerned with the design, development, and application of algorithms and techniques that allow computers to learn from data. Machine learning has a close connection with statistics (especially nonparametric and computational statistics) and theoretical computer science. Since the middle of twentieth century machine learning has evolved from the imitation of a simple neuron and artificial neural networks to a solid interdisciplinary field of basic and applied research having an important influence on many topics: pattern recognition, bio-computing, speech recognition, financial applications, analysis and modeling of high dimensional and multivariate geo- and environmental spatio-temporal data, etc. (Agarwal & Skupin, 2008; Cherkassky & Mulier, 2007; Hastie et al., 2009; Izenman 2008; Kanevski, 2008; Openshaw & Openshaw, 1997; Vapnik, 1998).

In recent years there has been an explosive growth in the development of *adaptive and data-driven* approaches. Among successful and widely used models of ML artificial neural networks (ANN) of different architectures and support vector machines have attracted great attention. Both have demonstrated important and successful applications for geospatial data modeling tasks: spatial predictions (classification and mapping); natural hazards and environmental risk assessments; renewable resources estimates; analysis, modeling and visualization of multivariate socio-economic data; environmental time series predictions; hydroinformatics; treatment and classification of remote sensing images, assimilation of data and science based models; etc. (see references below).

The key feature of the ML models/algorithms is that they learn from data and can be used in cases when the modeled phenomena is not very well described, which is the case in many applications of geospatial data. Machine learning models are adaptive tools, which at present are widely used to solve prediction, characterization, optimization and many other problems.

There exist many kinds of ANN to be used for different problems and cases. Among the most common in geo- and environmental sciences let us mention multilayer perceptron (MLP), radial basis function (RBF) networks, general regression neural networks (GRNN), probabilistic neural networks, Kohonen networks (self-organizing maps, SOM) (Agarwal & Skupin, 2008; Cherkassky & Mulier, 2007; Hastie et al., 2009; Izenman 2008; Openshaw & Openshaw, 1997; Haykin, 2009).

Let us remind that other approaches of geospatial data analysis and treatment (not data-driven) can be considered as model dependent ones. In this case an expert develops a model (e.g., models of spatial correlation) which is then used for the modeling and prediction purposes.

For example, traditional geostatistics can be considered as a well-known model-dependent approach for spatial data and one based on variography, which deals with the analysis and modeling of spatial correlations (Chiles & Delfiner, 1999).

At present, one of the most efficient data-driven approaches is based on statistical learning theory (SLT) (Vapnik 1998). The theory is based on the Structural Risk Minimisation (SRM) principle and has a solid statistical background. When applying SRM one tries not only to reduce training error – to fit the available data with a model, but also to reduce the complexity of the model and to reduce generalization (prediction) error. Many nonlinear learning procedures recently developed in neural networks and statistics can be understood and interpreted in terms of the structural risk minimization inductive principle. A methodology based on SRM is called Support Vector Machines (SVM). At present SLT is still under intensive development and SVM have been finding new areas of application. Many classical models, for example principal component analysis (PCA), were generalized to kernel PCA using so-called “kernel trick” (see details below).

SVM develop robust and non linear data models with excellent generalization abilities that are very important both for monitoring and forecasting. SVM use only support vectors (part of the measurement data points) to derive decision boundaries. They open a way to sampling optimization, estimation of noise in data, quantification of data redundancy, etc. More detailed presentation of SVM application for spatially distributed environmental data is given in Kanevski & Maignan (2004), Kanevski (2008) and Kanevski et al. (2009).

In general, geospatial data are not only data in a geographical low dimensional (2d, 3d) space but rather data embedded into a high dimensional geo-feature spaces, which consist of geographical coordinates and features generated from, for example, digital elevation models, science based models, remote sensing images, etc. In such cases an important problem is the “curse of dimensionality” – an exponential growth (with the dimension of space) of measurement data necessary to fill the space. Therefore dimensionality reduction methods have gained great popularity. Recently, many efficient nonlinear dimensionality reduction techniques were proposed and are efficiently used in many real life applications (Lee & Verlaysen, 2007; Guyon et al., 2005).

In terms of patterns/structures the following main problems closely related to the main problems of learning from data can be considered:

- pattern recognition/pattern detection,
- pattern modeling,
- pattern predictions/pattern completions.

The problem of pattern recognition is closely related to the contemporary exploratory data analysis when the main topic is to find/detect structured information in data without making restrictive hypotheses about data distributions. An important question is how to construct the criteria able to detect and separate “useful” structured information from the noise. For example, geostatistics widely uses the variography – spatial anisotropic correlations analysis in order to detect and to characterize spatial patterns/structures.

Pattern modeling is a task of developing models capable to correctly model structured information taking into account available data, expert knowledge, and science-based models, when these are available. Finally, pattern prediction is a process of forecasting/prediction in space and in time at the points where there are no measurements.

Other traditional topics where machine learning has contributed are problems of optimization and control, calibration of science-based (meteorological, physical) and empirical (cellular automata, multi agent systems) models, modeling/imitation of processes and events, etc. Some recent developments in ML applications for environmental modeling are considered in Cherkassky et al. (2006, 2007). More geographical applications - cellular automata calibration and land-use modeling are presented in Pijanowski et al. (2002) and Almeida et al. (2008). As a universal tool ML can be efficiently integrated with Geographic Information Systems in order to build modeling and data treatment modules.

Geospatial data processed by ML can be of different types: categorical/discrete data – classes (e.g., land use, soil types, geological units), continuous data (concentration of pollution, temperature, river levels, wind fields), and distributions – probability density distribution functions. Different models and tools were developed in machine learning to treat these classes of data and corresponding problems.

Let us consider briefly more formal definitions of the problem of “learning from data”.

### Learning from data. Setting of the learning problem

ML models are based on the statistical treatment of data. The main hypothesis is that data are generated from some unknown processes which can be characterized by the joint probability distribution function p(x,y) describing connection between input(=x) and output(=y). Input-output modeling can be considered as a general problem of mapping.

Statistical interpretation is quite general and covers also deterministic approaches of data description and modeling. Learning machine or machine learning model tries to model the relationship between input and output in order to recognize and to model patterns. Then, the model developed can be used for the prediction purposes.

The general framework of learning from data can be visualized in the following way (Vapnik, 1998; Cherkassky & Mulier, 2007):

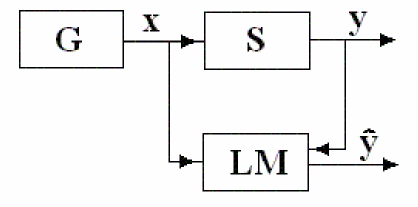


Figure 1: The flow-chart of a generic problem of machine learning from data

In Figure 1: *G* is a generator of i.i.d. (independent and identically distributed) data/vectors $x \in R^n$ drawn from a fixed but unknown probability distribution function *F(x);* (*S*) is a supervisor who returns an output value *y* to every input vector *x*, according to a conditional distribution *F(y|x)*, also fixed but unknown; (*LM*) is a learning machine capable of implementing a set of functions $f(x,\alpha)$, $\Lambda \in \alpha$, where $\Lambda$ is a set of parameters.

In general the learning process can be considered as a process of inferring from data and expert knowledge. Usually this process is a two step procedure (Figure 2): induction – development of a general probabilistic model and then deduction – using a developed model to make predictions at specific points. In many cases this two-step process can be replaced by a direct, so-called transduction process or "from particular to particular". The latter can be more efficient in real-life applications when there usually are not enough data in order to develop a generic inductive model and then use it during the deduction process to make predictions.

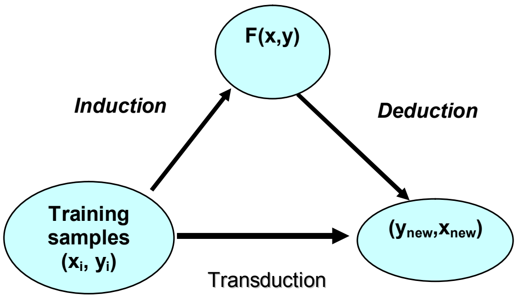


Figure 2: Learning processes: induction, deduction, transduction

In almost all real-life case studies an introduction of statistical model to data is non-trivial because usually only one realization of the phenomena is available: spatial data on pollution, time series of monitoring data, soil and land-use data etc. Statistical treatment of data can be introduced in this case as well under some hypotheses and assumptions. Therefore there are important hypotheses and assumptions that have to be checked (usually it is not a trivial task!) and accepted in order to make statistical (machine learning) inference based on one realization of the phenomena under study: iideness of data, ergodicity (loosely speaking, the convergence of the averaging over space to the averaging over realizations), spatial or temporal stationarity, i.e. absence of trends when important parameters of the model do not change in space/time. In case of geospatial data spatial clustering is an important topic which complicates both treatment of data (representativity of data) and interpretation of the results. The problem of non-stationarity (spatial or temporal) partly can be overcome by using locally adaptive models. Like in statistics these topics are also important in machine learning data treatment and modeling.

### Learning from data. Basic learning algorithms and techniques

Taking into account the definition of learning machine the procedure of learning in general can be considered as a process of the selection of a set of functions *f(x,α)* and tuning/fitting of the corresponding parameters *(α).* Most of machine learning algorithms use wide and flexible libraries of functions capable to model real life data (ML are universal modeling tools).

Common ML models include the following types of learning widely used in practice:

- **supervised learning** – where the algorithm generates a function that maps inputs to desired outputs. In this case output data are available at some (many) input data points. This information is used to train machine in a supervised manner by minimizing a well-defined criterion describing the discrepancy between data and modeling results. Usually this kind of problems falls into the optimization problems in high-dimensional spaces composed of machine's parameters.

- **unsupervised learning** – which models a set of inputs: labeled examples are not available. This is a typical problem of data clustering, data classification, and manifold learning.

- **reinforcement learning** – where the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.

- In recent years a **semi-supervised learning,** which combines both labeled and unlabeled examples to generate an appropriate function or classifier (in fact a "mixture" between supervised and unsupervised techniques) has gained much attention (Chapelle et al., 2006).

The performance and computational analysis of machine learning algorithms is a branch of statistics known as computational learning theory.

The generic procedure of learning can be visualized as flow-charts (Figures 3 and 4).

Supervised learning algorithms can be synthesized in the following diagram:

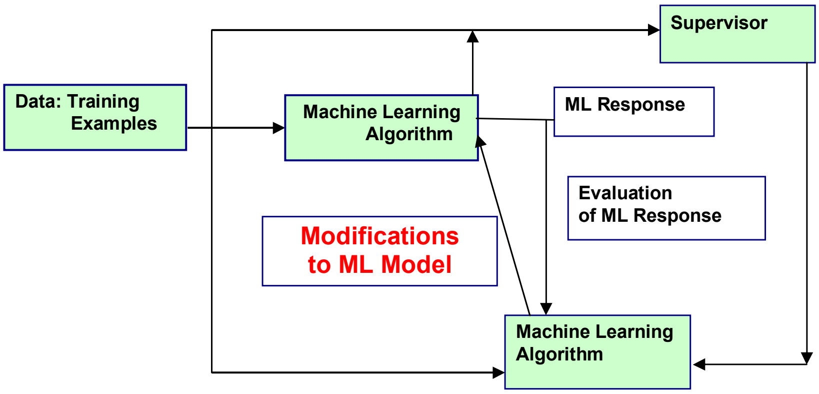

Supervisor
Data: Training Examples
Machine Learning Algorithm
ML Response
Evaluation of ML Response
Modifications to ML Model
Machine Learning Algorithm

Figure 3: Supervised learning

Unsupervised learning algorithms are a modification of the supervised approach:

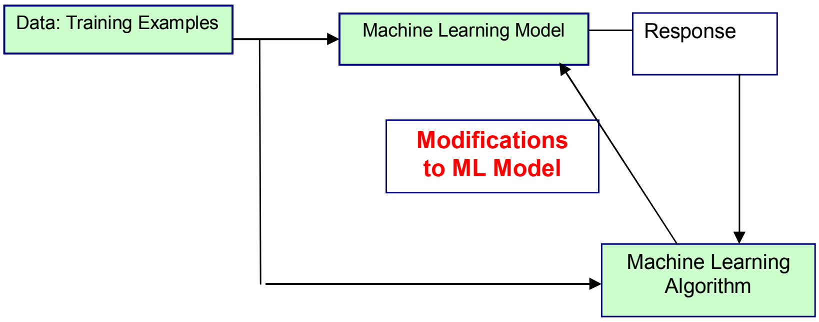


Figure 4: Unsupervised learning

Let us remind that an important property of almost all of the machines is that they are universal tools, loosely speaking they are able to learn any data with a desired precision (Haykin, 2009; Hastie et al., 2009; Cherkassky & Mulier, 2007).

The process of learning usually is quantified by applying a principle of risk minimization.

**Risk minimization**

An important fundamental question is how to describe the quality of learning, i.e. the description of the similarity/dissimilarity between data and a ML model. This quantification is important both during learning and prediction phases.

In order to choose the best available model to the supervisor's response, one measures the **LOSS** or discrepancy L(*y*,f(*x*,$\alpha$)) between the response *y* of the supervisor to a given input *x* and the response f(*x*,$\alpha$) provided by the loss measure which describes the dissimilarity between desired outputs and ML model. Finally, the problem of learning is to minimize the difference between desired and modeled outputs. In case of unsupervised problem the task is to group similar objects into separate classes.

Consider the expected value of the loss, given by the risk function described by the following formula:

$$R(\alpha) = \int L(y, f(x,\alpha))dF(x,y) \qquad (1)$$

where *L* is a loss function and *F* is a joint (input-output) distribution probability function.

The goal is to find the function f($x$,$\alpha_0$) which minimizes the risk in the situation where the joint probability distribution function (pdf) is unknown and the only available information is contained in the training set.

Loss function for the classification problem:

$$L(y,f(x,\alpha))=\begin{cases}0 & \text{if} \quad y=f(x,\alpha)\\ 1 & \text{if} \quad y\neq f(x,\alpha)\end{cases} \tag{2}$$

For the regression/mapping problem (modeling of conditional mean value) the risk is a well known mean-square-error:

$$L(y,f(x,\alpha))=(y-f(x,\alpha))^2 \tag{3}$$

Finally, consider the problem of density estimation from the set of densities p(x,$\alpha$), $\Lambda\in\alpha$. For this problem the following loss-function is usually considered:

$$L(p(x,\alpha))=-\log p(x,\alpha) \tag{4}$$

The criteria presented above are very general. Unfortunately the joint input-output distribution function is not known. Moreover, only a finite number of data measurements (N training data) is available. Therefore most training algorithms for learning machines implement Empirical Risk Minimisation (ERM), i.e. they minimize the empirical error

$$R_{emp}(\alpha)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i,\alpha)) \tag{5}$$

Minimization of empirical risk does not consider the capacity of the learning machine which can result in over-fitting, i.e. using a learning machine with too much capacity for a particular problem. The problem of over-fitting in this case is treated by using different regularization techniques (Bishop, 1995; Bishop 2007; Haykin, 2009): weight decay, early stopping, noise injection, etc.

Structural risk minimization (SRM) was introduced by Vapnik & Chervonenkis in 1974. It is an inductive principle for model selection used for learning from finite training data sets (Vapnik, 1998; Cherkassky & Mulier, 2007). It describes a general model of capacity control and provides a trade-off between hypothesis space complexity (the Vapnik-Chervonenkis dimension of approximating functions) and the quality of fitting the training data (empirical error) (Vapnik, 1998; Cherkassky &

Mulier, 2007). The SRM principle is illustrated in Figure 5. X-axis corresponds to the complexity of the model and Y- axis to the error. According to the SRM principle prediction error is a sum of training error (empirical risk minimization) and complexity term which takes into account penalization of too complex models. In this way a bound on prediction error can be derived which gives an upper limit. This limit does not depend on the distribution of data and therefore is rather pessimistic (too high). In reality, for particular data limits are lower and can be estimated by splitting the data or by using cross-validation technique.

An optimal solution corresponds to the minimum on the curve describing the testing error (= sum of training error and complexity term). The solution with low complexity (oversmoothing/underfitting) is biased - not complex enough to explain the structure in the data. The solution with higher that necessary complexity over-fits data, i.e. noise in data is also modeled. The intermediate complexity solution models structure/patterns and ignores noise, thus following Occam’s razor principle: “The more simple explanation of the phenomena is more likely to be correct” or "entities must not be multiplied beyond necessity" (the law of parsimony).

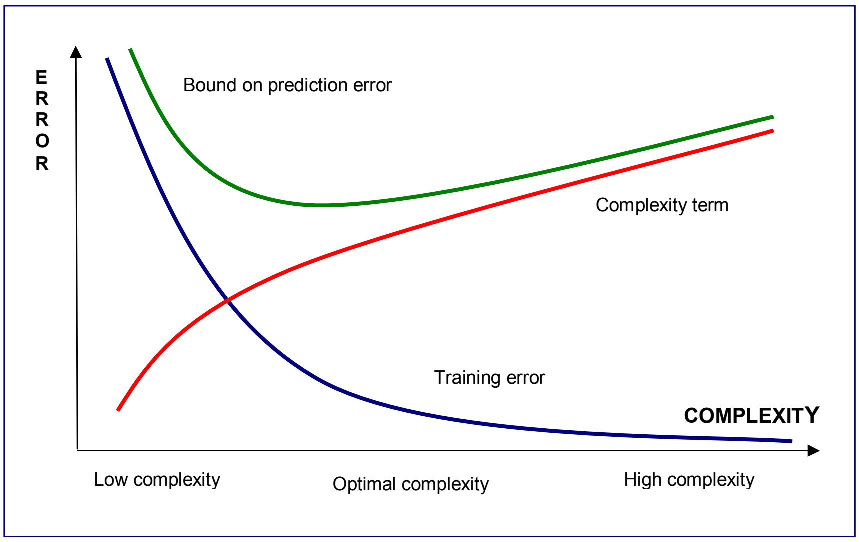


Figure 5: Structural risk minimization principle (Vapnik, 1998)

An important question of data ML modeling is that the main objective is to develop a good model that is a model which has good generalization properties – prediction of new outputs at unknown input points which were not used for modeling purposes. It means that the developed ML model has learned only structured information and ignored a noise present in data. In this case over-fitting is avoided.

Bias-variance dilemma is a step towards the understanding and quantification of this problem.

**Bias-Variance Dilemma**

Let us consider a regression/mapping problem as an example. In general, we can assume that data can be decomposed into unknown function and noise:

$$\begin{aligned}&Y(X)=f(X,\alpha)+\varepsilon\\&where\\&E(\varepsilon)=0,\\&Var(\varepsilon)=\sigma_\varepsilon^2\end{aligned} \tag{7}$$

where *Y(X)* are measured data, *f*(*X*,α) is an unknown function and *ε* is a noise present in the data.

The following expression for the expected prediction error of a regression at an input point *X*=*x₀* using squared-error loss can be derived (Hastie et al., 2009):

$$\begin{aligned}&Err(x_0)=E[(Y-\hat{f}(x_0,\alpha))^2 \mid X=x_0]=\\&\sigma_\varepsilon^2+[E\hat{f}(x_0,\alpha)-f(x_0,\alpha)]^2+E[\hat{f}(x_0,\alpha)-E\hat{f}(x_0,\alpha)]^2=\\&\sigma_\varepsilon^2+Bias^2(\hat{f}(x_0,\alpha))+Var(\hat{f}(x_0,\alpha))=\\&IrreducibleError+Bias^2+Variance\end{aligned}$$

One of the most serious problems that arises in connection with learning by neural networks is over-fitting of the provided training examples. This means that the learned function fits very closely the training data and yet does not generalize well, that is it cannot model sufficiently well unseen data from the same phenomena.

Solution: Balance the statistical bias and statistical variance when doing neural network learning in order to achieve the smallest average generalization error.

The following two processes are important in making a decision about the quality of the model and its generalization ability (Hastie et al., 2009):

A. Model Selection is a process of “estimating the performance of different models in order to choose the (approximate) best one”.

B. Model Assessment is a process of ”having chosen a final model, estimating its prediction error (generalization error) on new data”.

There are many different techniques developed in statistics and the ML community for model selection and model assessment. If we are in a data-rich situation, the best solution is to split data into three data subsets which will be used for 1) training (training data), 2) model selection (test data) and 3) model assessment (validation data). Typical splitting is: 60% is used for training, 20% for testing and 20% for validation purposes.

How can geospatial data be split? Usually random splitting of data is applied. In case of geographically distributed data, especially when spatial clustering is important, declustering procedures can be used as well in order to split data into representative subsets (Kanevski & Maignan, 2004).

Let us give some general comments on ML applications to geospatial data:

- A machine learning model is only as good as the training data. The results depend on the quality and quantity of data.

- Poor training data inevitably leads to unreliable and unpredictable neural networks and support vector machines. Control of the quality of hyper-parameters tuning by using a testing data set is necessary.

- Exploratory Data Analysis and data pre-processing including visualization tools are extremely important! Pre-processing of data can include data transformation as well. Despite the fact that ML are nonlinear tools intelligent data pre-processing can improve learning and make it faster.

- Analysis of the residuals is necessary and helps in understanding the quality of modeling and results.

- If possible, prior to training, add some noise or other randomness to the training examples. This helps to account for noise and natural variability in real data, and tends to produce a more reliable network. This is a well known technique of noise injection which in many cases prevents an over-fitting of data and is equivalent to regularization (Grandvalet et al., 1997; Kanevski et al., 2009).

- Criteria of early stopping are widely used to avoid over-fitting of data when multilayer perceptron is used for modeling purposes. Complex model is iteratively trained (one training pass using all data is called an epoch) until the testing error starts to grow (Figure 6). The model with minimum testing error is fixed as an optimal model and then the validation set is used to estimate the generalization properties of the model developed (see above). If noise in data can be estimated independently, for example, using variography (nugget) or so-called delta- or gamma-tests (Pi & Peterson, 1994; Jones, 2004) splitting of data into training-validation is not necessary because the training procedure can be stopped at the level of the noise. As usual the goal of training is to extract patterns and not a noise.

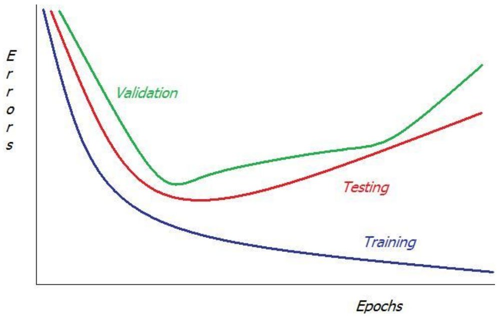


Figure 6: Illustration of an early stopping principle

In the following section unsupervised and supervised learning algorithms are presented and applied to real data case studies.

## UNSUPERVISED LEARNING

Unsupervised techniques are aimed at the exploratory analysis of data. They provide insight into the structures and dependencies hidden in the datasets. This is achieved by finding a simplified representation of a dataset handful for visualization, feature extraction, or descriptive analysis purposes. The main problems encountered here are the problems of clustering and dimensionality reduction (also called embedding).

Most of the methods require defining a distance measure between data samples. Both clustering and dimensionality reduction results depend heavily on this choice, as this is the step when the inherent similarities in data are encoded.

## CLUSTERING

Clustering can be defined as partitioning the dataset into subsets of typical entries such that the samples in each subset share some common characteristics. The commonness is implied by the pre-defined similarities between data samples, and usually requires one to define a problem-specific distance measure in the input space of features. The general types of similarity measures that are used to compare data samples allow distinguishing the groups of typical approaches in unsupervised methods.

There is a vast body on literature on clustering problems, comprehensively reviewed in Jane et al. (1999). Another criteria allowing to distinguish between different groups of the clustering methods include the type of outputs provided by the method (hard assignation to a cluster or fuzzy measures like cluster membership probability), deterministic or stochastic nature of the method (whether the algorithm converges to the same solution in the same initial setting), and the agglomerative or divisive approach to clustering.

### K-means algorithm

Probably the most popular clustering algorithm is known as k-means. Given the dataset $\{x_1, \ldots x_N\}$, it operates as follows:

- K centers $\mu_k$ are initialized randomly in the input space;
- Each sample $x_i$ is assigned to the closest centre with respect to some (usually Euclidean) distance. It is considered as belonging to the cluster $C_k$;
- New centers are computed as $\mu_k^{new} = \frac{1}{N_k}\sum_{i \in C_k}^{N_k} x_i$ .

The last two steps are iterated until convergence. As the centers are updated using all the data at once, this version of the algorithm is known as batch k-means, opposed to the online or stochastic k-means when the update is done by randomly iterating through the samples of the dataset. The stochastic k-means can be faster and less sensitive to the initialization of the centers. K-means methods are aimed at minimization of the intra-cluster variance and demonstrate good performance if the data face distinctive “clouds” and there is no significant correlation between the input features (Figure 7).

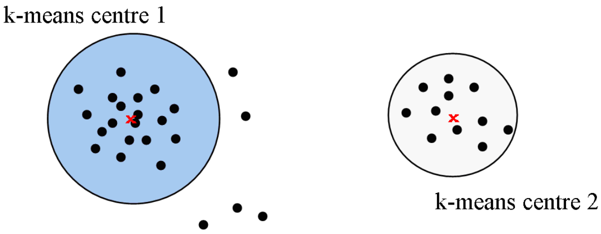


Figure 7: K-means algorithm searches for well-shaped cloud-like clusters in data. It performs poorly if data have correlated (linearly or non-linearly) inputs

### Self-organizing maps

Self-organizing maps (also known as Kohonen maps (Kohonen, 2000)) are a popular method extending the functionality of k-means. SOM places the centers (also called units or neurons) in the data space aiming to "cover" all data points to fit to the topology of the dataset optimally and present it in a two-dimensional map. That is, the centers are not drawn and fitted independently as in the case of k-means, but organized in a two-dimensional map. There exist two common designs of this map: rectangular (every unit has four neighbors) and hexagonal (every unit has six neighbors), shown in Figure 8.

The map is initialized either by random placement of the centers of the map in the data space or by sampling them evenly in the input data subspace spanned by the two largest principal component eigenvectors. This method can increase the training speed significantly as the initial weights may already give a good approximation of the structures in the data. SOM is sensitive to the initialization and may require several runs.

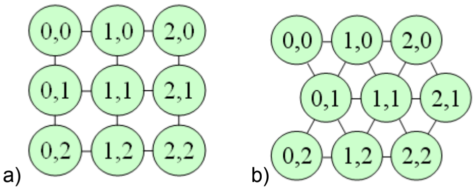


Figure 8: Two different SOM structures with cells in the map space: rectangular (a), and hexagonal (b). The nearest cells (4 for rectangular and 6 for hexagonal, except borders and corners) are connected by the edges

The training process for updating the positions of centers is an online iterative procedure. Samples are presented to the SOM one by one, and the positions of centers are updated. The key point here is that not only the closest centre is updated, but also the centers in its closest (on the map) surrounding. The closeness on the map is defined with the neighborhood function which is responsible for the “cooperation” of centers and their self-organization. This neighborhood influence is then gradually decreased through every iteration epoch. The last epoch of SOM adaptation with no cooperation between centers is the same as in the case of simple k-means.

The last step in SOM training is a post-processing of the obtained map. It is usually done by analyzing the so-called U-matrix, that is the matrix of the distances (in the input space) between the neighboring nodes of the trained SOM structure. With this matrix, one can detect that a group of SOM centers appears far from the rest of the map thus revealing that the data samples captured by these centers form a distinctive cluster. This step can be performed manually but usually an automated procedure such as k-means or hierarchical clustering is used. More detailed description of SOM models and their application for geospatial data can be found in Kohonen (2000), Agarwal & Skupin (2008) and Kanevski et al. (2009).

**Spectral clustering**

The method of spectral clustering originates from a graph-based perspective of the clustering problem (see e.g. Hagen & Kahng, 1992; Ng et al., 2001; Shi & Malik, 2000). Spectral clustering considers the data samples as the nodes of a weighted graph and applies the methods of spectral graph theory to study its inner structure (Figure 9). To describe the affinity between the nodes of the graph, the vertices connecting the *i*-th and *j*-th nodes (the data samples $x_i$ and $x_j$) are entitled with the weights $w_{ij}$ that form the matrix W. To separate the graph into clusters one has to find the “cut” that minimizes the sum of weights that would need to be removed in order to split it. If the weights are attributed according to the distance measure between the data samples $x_i$ and $x_j$, it leads to the clustering of data in the input space. Two common approaches are the n-nearest neighbor one, where $w_{ij}=1$ iff the i-th and j-th samples are amongst the n-nearest neighbors of each other, and the one which uses the Gaussian RBF function of the distance between samples as the value for the weight. The last case implies a parameter to be defined by a user, that is, the width of the Gaussian. The way one attributes the weights can also account for some problem-specific knowledge.

To approach the problem of finding the cut of the graph one needs to analyze the matrix known as graph Laplacian. It is defined as L=W-D where D is the diagonal matrix with element formed by the column-wise sums of elements of W, $d_{ii}=\Sigma_j[w_{ij}]$. The normalized graph Laplacian $L=D^{-1/2}W\ D^{-1/2}$ is often used as well. The foundations of spectral clustering lie in the fact that the eigenspace of the (normalized) graph Laplacian has a particular well-defined structure related to the number of the connected components of the graph. Or, intuitively, if the Laplacian matrix is essentially block-diagonal, it can be easily detected from its eigenspace. While there are many possible approaches that implement this idea, the most popular formulation of the spectral clustering is as follows (Ng et al., 2001):

- Form the affinity matrix W and compute the normalized graph Laplacian;
- Solve the eigenvalue problem of finding the {λ, f} such that Lf=λf;
- To detect k clusters: find the k eigenvectors with largest eigenvalues of L and stack them in columns to form the N-by-k matrix U, normalize it.
- Perform ordinary k-means on the columns of U. Assign the obtained cluster labels to the original N data samples.

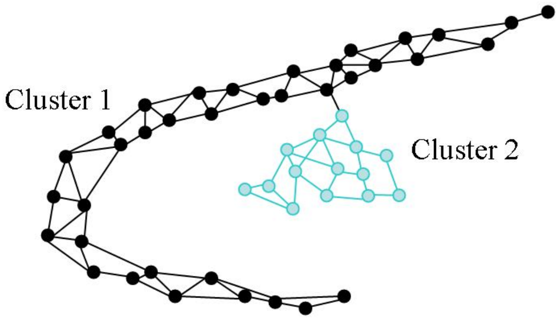


Figure 9: Spectral clustering builds a graph connecting data samples in a local neighborhood and analyses its structure in a search for weakly connected components. Simulated data

## CLUSTERS IN SOCIO-ECONOMIC DATA

To illustrate the use of clustering methods, let us explore the socio-economic data on the region of Lausanne, Switzerland. The dataset was obtained from the population census and includes about 250 different entries defining the social, cultural (mother tongue, nationality, etc.) and economic (employment rate, household type, etc.) characteristics of the population. The data are spatially aggregated from the regular grid cells of 100x100 meters covering the populated area of the region. There are a total of 3359 samples. Population density, which is one of the most important input features, is presented in Figure 10.

Clustering methods may be exploited to provide useful insight into data and to identify structures in the space of socio-economic parameters. Visualization of the obtained clusters on the map can shed some light on the spatial structure of the region. The results of the described methods obtained on the dataset assuming a priori 6 clusters are shown in Figures 10-13. These maps can give rise to interpretations and insight into the spatial agglomeration of the different social groups populating the region at study.

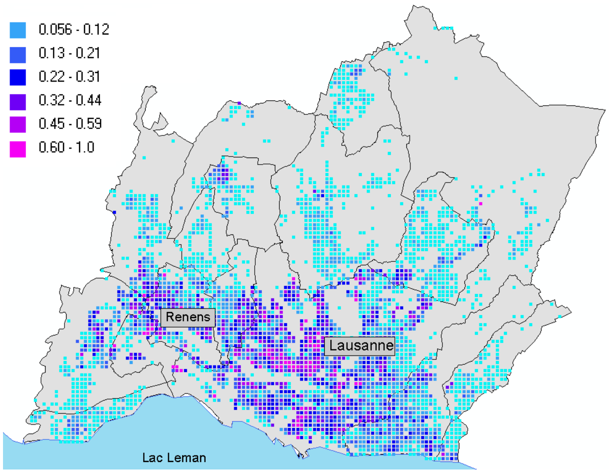


Figure 10: Population density in the region of Lausanne. The values are normalized to the maximum value of density in the region

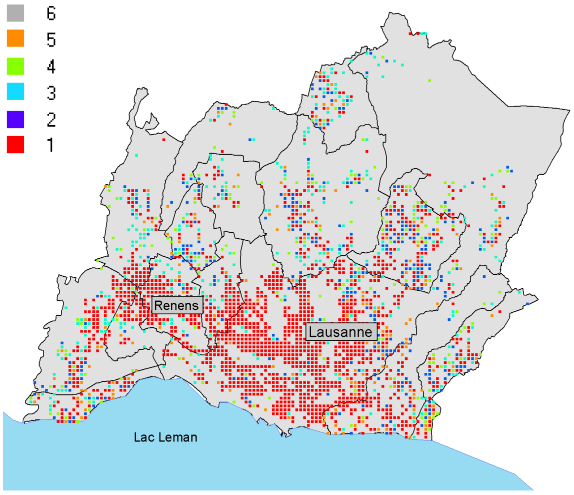


Figure 11: Clusters obtained with k-means method

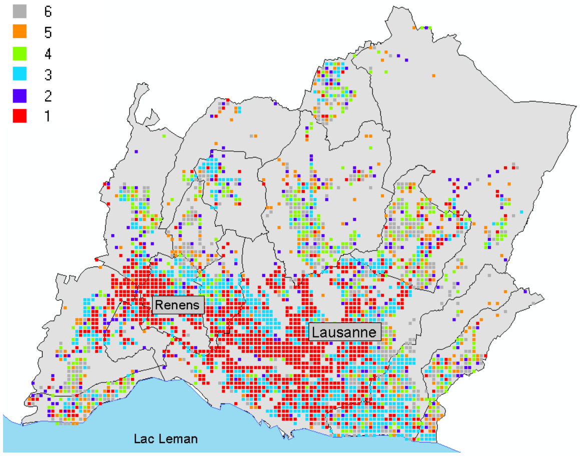


Figure 12: Clusters obtained with SOM

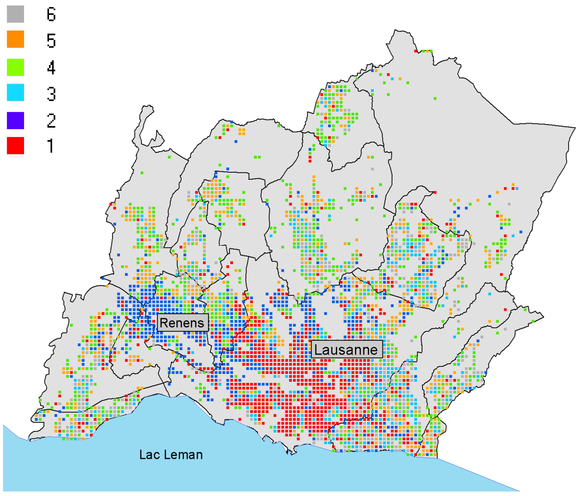


Figure 13: Clusters obtained with spectral clustering

## DIMENSIONALITY REDUCTION

Dimensionality reduction is usually involved while aiming at two goals: first, to produce the low-dimensional representation of data to visualize them, and, second, to extract low numbers of features for further analysis. Here one distinguishes the feature selection and feature extraction, where the latter rather than selecting already existing variables, tries to select a linear or nonlinear combination of the input variables which suits best the problem at hand. An example of a linear method is a well-known principal component analysis. There is a popular and rapidly growing domain of modern nonlinear dimensionality reduction methods known as manifold learning (Lee & Verleysen 2007).

### Principal Component Analysis

The method of principal component analysis (PCA) is a well established statistical method (Pearson, 1901). It searches a linear combination of the input variables (or a linear transformation of the input space) that accounts for maximum variability observed in data (Figure 14). It is provided with the following algorithm:

- Standardize the data and compute the covariance matrix $S = XX^T$
- Compute the eigenvectors and eigenvalues of S, sort them by decreasing values, arrange the eigenvectors in matrix U
- The transformed data set is F=XU

PCA enables to reduce the dimensionality of the data that include linearly correlated inputs. For visualization, the first two principal components span the projection plane providing the most informative (in the sense of variability) viewpoint on the original dataset. The amount of accounted variance can be computed to assist in the choice of the number of components.

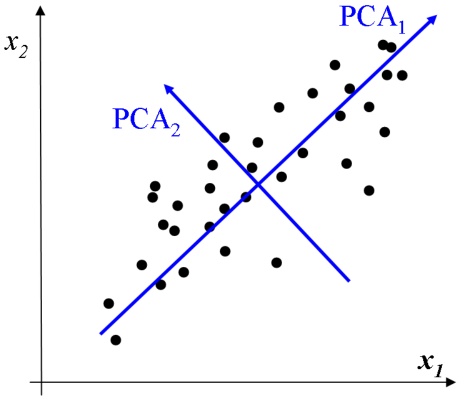


Figure 14: Principle Components form a new orthogonal coordinate system, with the first components spanning along the directions of maximum variance of the data

### Laplacian Eigenmaps

The large variance along a straight line is not necessarily the main characteristic of interest with respect to analyzing the complex structures in the data. A linear projection cannot help unfolding the non-linear structures as the one of the “Swiss roll”, or even a simpler one as shown in Figure 15. It is the local relationships between data samples that may help discover these complex structures. Laplacian eigenmaps (Belkin & Niyogi, 2003) are one particular approach of manifold learning aimed at preserving the local neighborhood relations between the data samples. Here we briefly name the other methods of descriptive manifold learning: locally linear embedding (Roweis & Saul, 2000), ISOMAP (Tenenbaum & De Silva, 2000), maximum variance unfolding (Weinberger & Saul, 2005).

Laplacian eigenmaps operate with the above-mentioned affinity matrix W and its graph Laplacian L and can be summarized as follows:

- Form the affinity matrix W and compute the graph Laplacian L=D-W;
- Solve the eigenproblem Lv=λDv;

- Present data in projections on v starting with the smallest eigenvalues.

This representation keeps the data samples proximate in the input space (according to the affinity matrix) close in the embedded coordinates. Hence the step of constructing the affinity matrix is very important as it should encode the similarities that one desires to keep on the low-dimensional map. Spectral clustering (described above) is essentially a related method that only includes one additional step of clustering the obtained representation using a conventional k-means.

There are some problems encountered when dealing with non-linear dimensionality reduction methods of manifold learning. The first one is the so-called out-of-sample problem. As the embedding is based on the graph built on the training samples, it is not very straightforward to apply the obtained embedding for a new data sample. Then, it is based on the eigenvectors problem and requires matrix computations which are difficult for very large datasets.

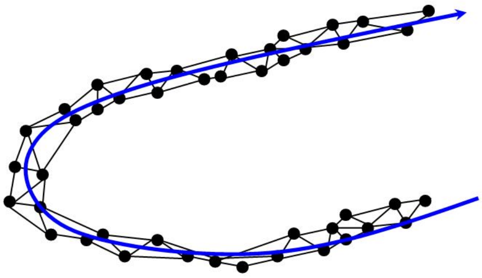


Figure 15: Graph-based non-linear dimensionality reduction methods, including the Laplacian eigenmaps, seek for a “direction” in the graph that carries maximum variance. As the directions found are based on the graph structure, it is not straightforward to represent new samples in the obtained embedding

## DIMENSIONALITY REDUCTION OF SOCIO-ECONOMIC DATA

The approaches for linear and non-linear dimensionality reduction are illustrated using the socio-economic data briefly described above. The methods were applied to the 250-dimensional dataset of 3359 samples. Analysis of variance of PCA components showed that 95% of variance is accounted with more than a hundred components. The first 4 components (accounting for 55% of variance) are shown in Figures 16, 17.

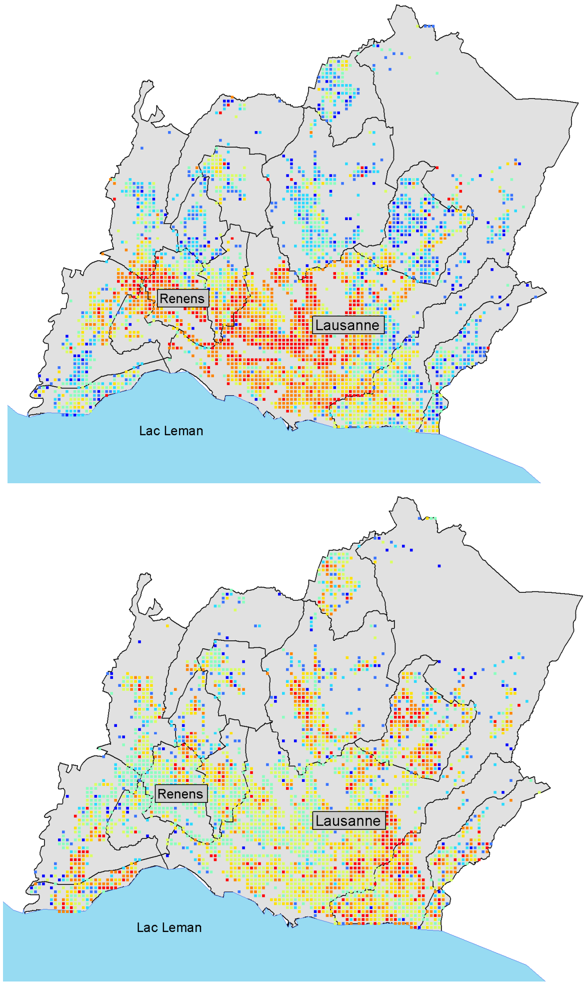


Figure 16. First and second PCA components. The first one clearly follows population density

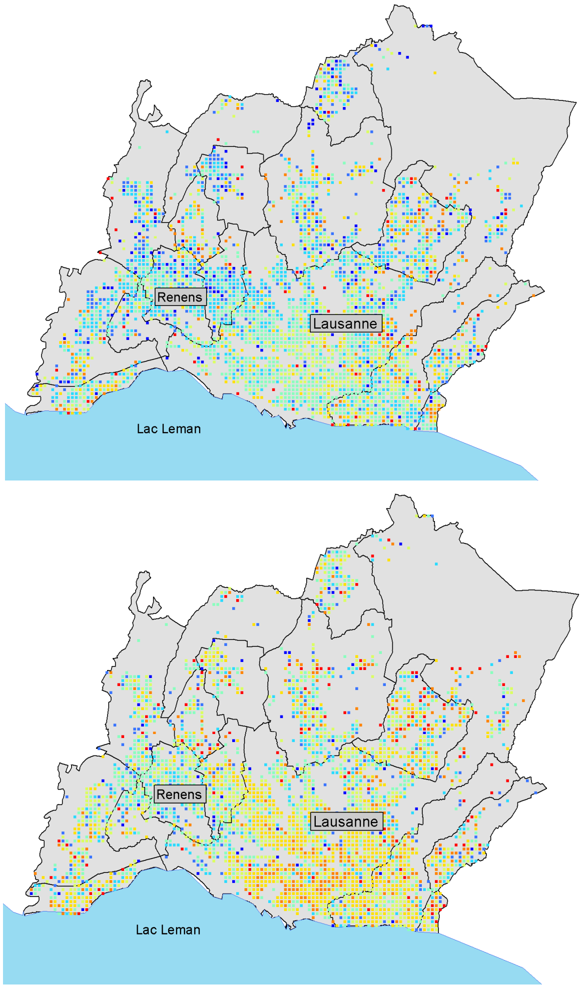


Figure 17. Third and fourth PCA components. The interpretation of these is not straightforward, though one can notice that the two neighbouring towns (Renens and Lausanne) differ

The method of Laplacian Eigenmaps was applied to the same dataset. The first 4 components are shown in Figures 18, 19. For this method, the interpretation of the amount of "accounted variance" is not evident. Some approaches can be based around the analysis of eigenvalues. However, the components revealed clearly the socio-economic diversity of the regions, highlighting more detailed and distinctive features.

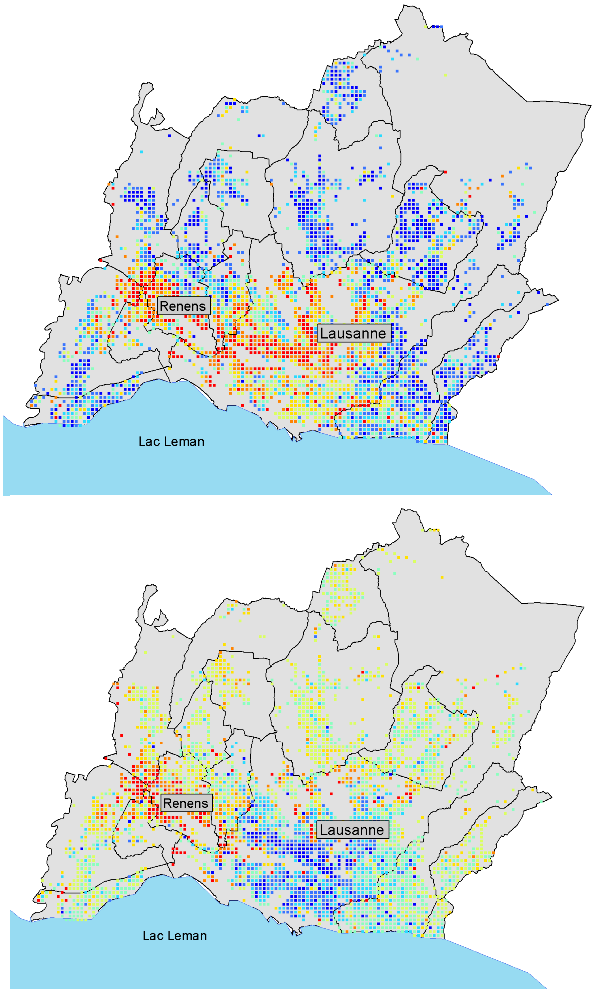


Figure 18. Spatial representation of the first and second components obtained with Laplacian eigenmaps

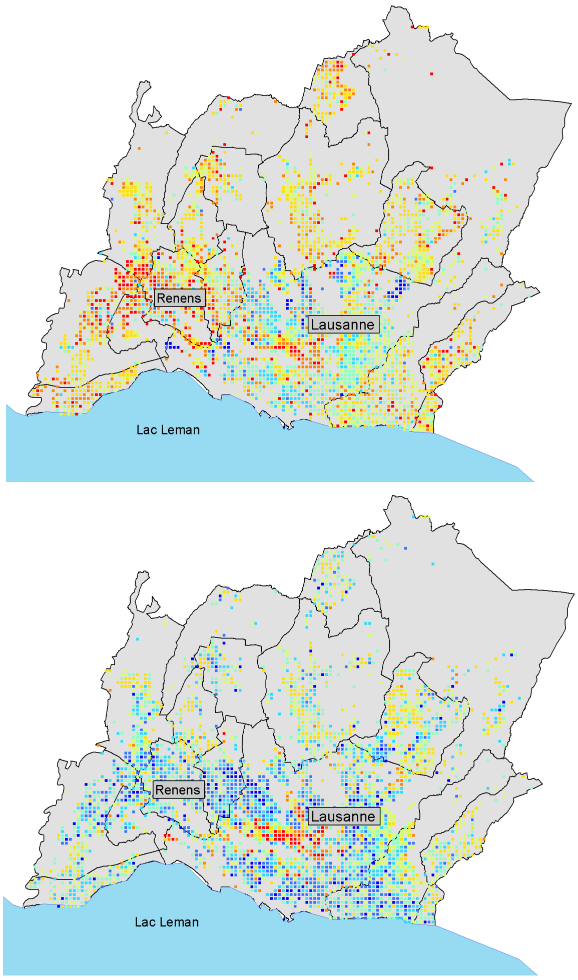


Figure 19. Spatial representation of the third and fourth components obtained with Laplacian eigenmaps

## SUPERVISED LEARNING

In supervised learning, the machine takes advantage of knowledge about the outputs to develop a predictive model. Contrary to the unsupervised methods discussed in the previous section, the algorithm is trained over a set of samples **x** = {$x_1$, $x_2$, …, $x_n$} with associated known outputs **y** = {$y_1$, $y_2$, .., $y_n$}.

In a regression problem, outputs are continuous values and in a classification problem the outputs are discrete values corresponding to classes.

The aim of a supervised method is to build a predictor that

1. Minimizes the error of prediction over known examples, i.e. the couples (*x,y*). The predictor minimizes a certain loss function depending on the task. Empirical or structural risk principles are applied.
2. In general the model must be capable of predicting new unknown examples, i.e. should have good generalization properties.

In this chapter, supervised models are considered in detail both for classification, through the Support Vector Machine (SVM) and for regression, through Multilayer perceptron (MLP) and the General Regression Neural Network (GRNN).

### Supervised learning for classification

As introduced above, the outputs of classification are discrete values corresponding to a given class of interest. For instance, in the example shown below, the outputs are land-use classes such as forest, roads or buildings. From now on, we will refer to known outputs as **labels**.

The principle of supervised classification is demonstrated in Figure 20:

a) A set of features (=variables) is selected to form the input space. In the figure the input space is a 3-dimensional space defined by the three bands of an image in the Red, Green and Blue regions of the visible spectrum. For each pixel that has been labeled we have a couple $(x_i,y_i) = (x^R_i, x^G_i, x^B_i, y_i)$. Unlabeled pixels only possess the values of $x_i$.

b) The labeled pixels can be visualized in the feature space. The coordinates for each pixel correspond to the values on each band. By assigning a different marker to each class, it is possible to visualize the separability of the classes.

c) The aim of the supervised model is to find the best separation between the classes. In this example, most of the labeled pixels are separated correctly, only three pixels are misclassified. Most of the time, misclassification is a necessary trade-off to maintain a low complexity. For instance, if the classification is performed on an aerial photography using RGB values only, green roofs will have a spectral signature which is identical to meadows. Therefore, confusion between the classes has to be admitted.

d) Once the separation between the classes has been defined, all the unlabeled pixels can be classified according to these boundaries by computing their position in the feature space. If they fall in one of the green areas of Figure 20.c, they will be classified as 'green' as well, and so on. Once all the pixels of the image have been classified, a classification map is provided.

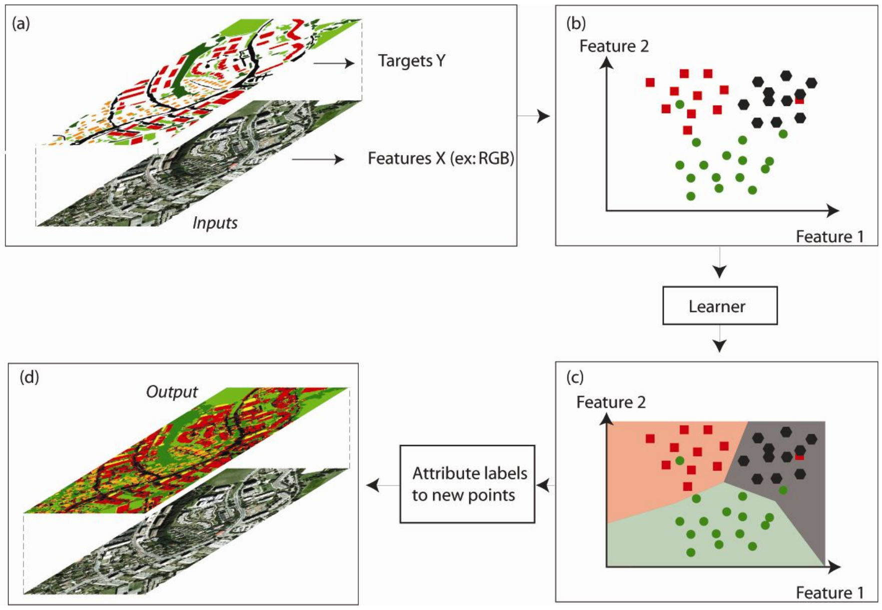


Figure 20: Principle of supervised learning for remote sensing data. (a) Each pixel is associated to features (ex: the spectral bands) AND known labels Y. (b) Knowledge about class membership is used in the feature space to train a learner and (c) define a decision boundary. (d) Finally, the unknown pixels are classified with respect to the decision boundary found and a classification map is provided

There are many algorithms including machine learning that offer solution models to classification tasks: k nearest neighbors (k-NN can be considered as a benchmark model), decision trees, probabilistic neural networks, multilayer perceptron, radial basis functions, support vector machines (Duda et al., 2001; Bishop 2006; Vapnik 1998). SVM have demonstrated excellent efficiency on classification tasks in different fields from remote sensing images to biocomputing and finance.

In the following section basic concepts and ideas concerning support vector machines first introduced in Boser et al. (1992) are presented in detail.

**Support Vector Machines**

First, let us consider the most simple two-class classification problem in a two-dimensional space (Figure 21.a) when data are linearly separable.

The task is to develop a decision boundary which will discriminate the circles $\{y^{C}_{i} = -1\}$ from the squares $\{y^{S}_{i} = 1\}$. Several planes can be drawn to separate correctly these two classes (Figure 21.b). According to Statistical Learning Theory and SRM principle the best solution is (Figure 21.c) a decision boundary (line) which separates two classes with a maximal margin (Vapnik, 1998). In a multidimensional space (where multiple variables are considered simultaneously), the line is replaced by a hyper-plane. By finding the margin with maximal width, we find the best way to linearly separate the labeled samples. This solution has the best generalization properties.

Since only the points lying on the margin are necessary to define the separating hyper-plane, all the other labeled points are not considered by the model.

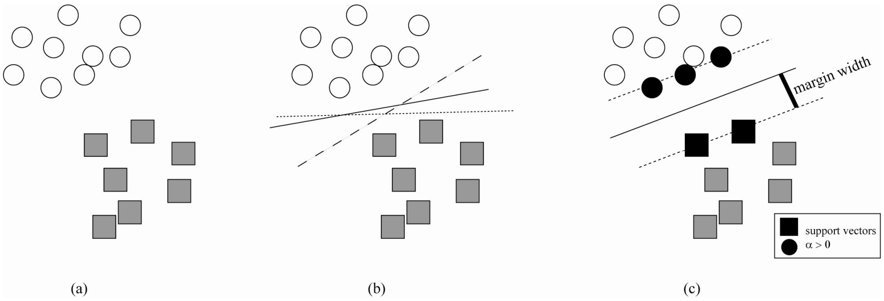


Figure 21: Linear classifiers for a two-class problem. (a) the problem; (b) several linear classifiers separating the two classes; (c) the SVM

Mathematically, this results in the following classifier for the prediction of an unlabeled pixel *q*:

$$f(x) = sign(\sum_i \alpha_. y_i \langle x_i, q \rangle + b) \tag{8}$$

where $\alpha_i$ are coefficients that are i) nonzero and ii) equal to 1 when the labeled sample lies on the margin and $\langle x_i, q \rangle$ is a dot product defining the similarity between the unlabeled pixels and the labeled samples and *b* is the bias. Since the $\alpha_i = 0$ for each sample not lying on the margin, class membership of an unseen point is assessed by the similarity (~ distance) between the labeled pixels and the samples that lie on the margin only. These samples are called the **support vectors**.

Recalling the example of Figure 21, since circles have negative label ($y^C_i = -1$) and positive squares ($y^S_i = 1$), if the new point *q* is globally closer to the support vectors of the class "circles", solution of Eq. (8) will result in a negative value and *q* will be labeled as a circle.

The SVM presented so far can solve only linear separable problems. Slack variables can be considered to allow small errors (see, for instance, Cristianini, 2000), but the algorithm will fail if the data are not linearly separable (Figure 22.a). In order to handle linearly non-separable problems, we might use the so-called kernel trick. If a problem is not linearly separable in the input space, it may be linearly separable in a higher dimensional space H (Figure 22.b). If such a space exists, we can map the labeled samples in the new space and then apply a linear classifier. A linear classification in a higher dimensional feature space (Figure 22.c) will correspond to a nonlinear classification in the input space (Figure 22.d).

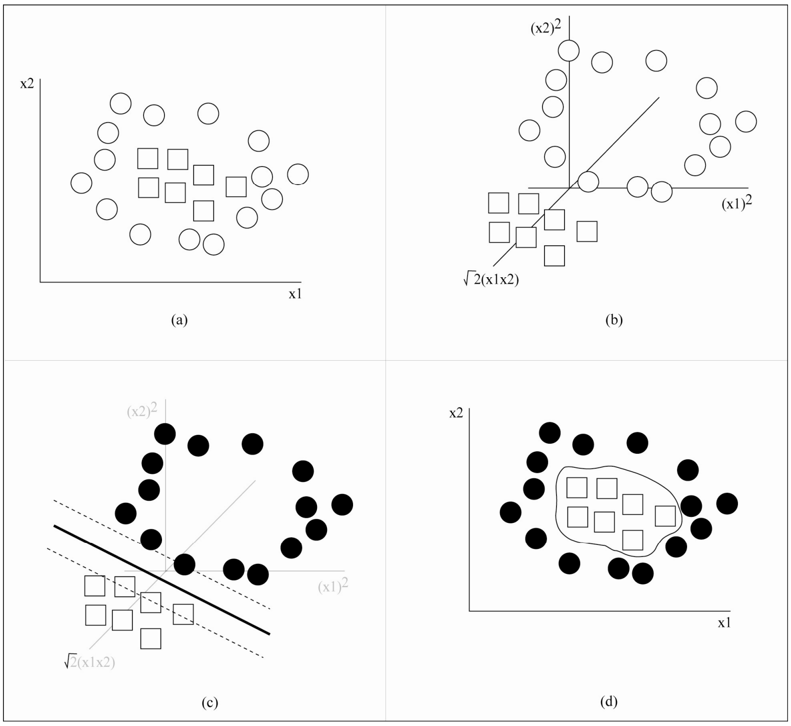


Figure 22: the kernel trick. (a) a linearly non-separable problem in the input space; (b) mapping in a higher dimensional space H; (c) the decision function in H is linear; (d) in the input space it is not

The mapping (computation of the new coordinates) in the new space of all the labeled samples can be solved analytically. For instance, a two-dimensional sample **x** = $\{x_1, x_2\}$ mapped into a 3-dimensional space by a quadratic transform can take the coordinates $\phi(x_i) = \{x_1^2, \sqrt{2}x_1x_2, x_2^2\}$. But by looking at Eq. (8) again, we can notice that the explicit mapping of $x_i$ is not required, rather only the similarity between $x_i$ and $q$. Therefore, there is no need to compute the entire mapping of **x** in H, but only the distances between the mapped **x** and **q**. Such distances can be represented by kernel functions. For instance, a polynomial function of degree 2 encodes a nonlinear similarity $K(x_i,q) = (\langle x_i, q\rangle)^2 = x_1^2q_1^2 + 2x_1x_2q_1q_2 + x_2^2q_2^2 = \langle\phi(x_i), \phi(q)\rangle$. Therefore, K returns the value of the dot product between the mapped samples! Using the kernel trick, the nonlinear SVM solution becomes:

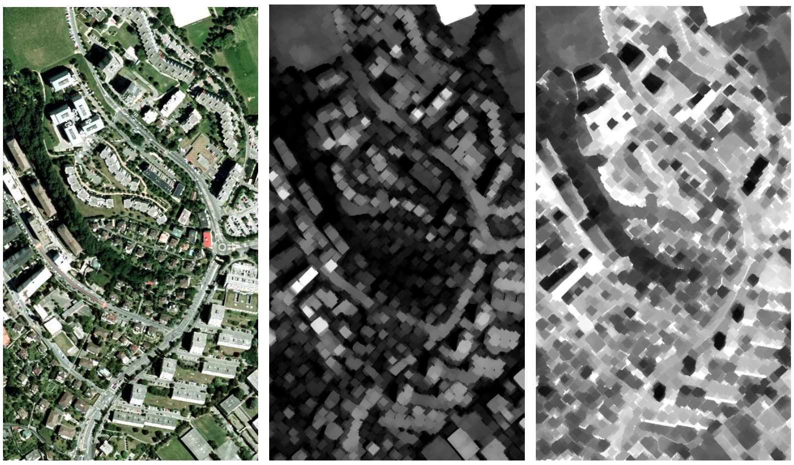
$$f(x) = sign(\sum_i \alpha_{.} y_i K(x_i, q) + b) \tag{9}$$

Summing up, we can compute the SVM solution in a higher dimensional space (where the problem is linearly separable) without knowing explicitly the position of the samples in that space, but only by assessing the distance between them using a kernel function K.

### *3.1.2 Application to aerial photography for land use characterization*

In this section, we will train a SVM for the classification of land use in a neighborhood of the city of Lausanne, Switzerland (Figure 23.a). In particular, we would like to use an aerial photography to discriminate different types of habitat, in particular individual versus collective habitat. This is a very challenging problem, because the spectral information is rather poor (each pixel is only considered by its color coordinates in the RGB space) and roof colors are mixed for the same type of habitat. Moreover, asphalt objects such as roads or parking lots can be easily confused with roofs.

In the first experiment, that we will call MS (Multi-spectral), we use the spectral bands $\boldsymbol{x}^{MS} = (\boldsymbol{x}^{R}, \boldsymbol{x}^{G}, \boldsymbol{x}^{B})$. Therefore in this case pixels are classified by their colors only.



(a) (b) (c)

Figure 23: data considered. (a) aerial photography of the NW of Lausanne; (b) feature , $\boldsymbol{x}^{O15}$ ; (c) feature $\boldsymbol{x}^{C15}$

In order to resolve the confusion discussed above, information about the spatial neighborhood of the pixels has been extracted. This way, size and shape of the objects are taken into account and, for instance, thickness of the roads will help to discriminate road pixels from roof pixels.

Several ways of extracting spatial features have been proposed, going from the simple computation of local means to more advanced feature extraction based on mathematical morphology or texture. In this example, multi-scale morphological filters (Soille, 2004) have been applied to the image. These filters produce features showing the objects that are brighter or darker than their spatial surroundings. In this example, we have added to the three spectral bands 14 morphological features, accounting for morphological opening and closing using diamond-shaped structuring elements of size going from 9 to 21 pixels. Morphological opening filter objects that are brighter than the neighborhood defined by the structured element; morphological closing filter objects that are darker. Two features, obtained by applying opening and closing with a 15 pixels structuring element, are reported in Figure 23.b and 23.c. To know more about these filters for greyscale images, please refer to (Pesaresi, 2001; Benediktsson, 2003). The filters have been applied to the first principal component (see section 2) extracted from the RGB image. In the second experiment, that we will call MM (Mathematical Morphology), we use the spectral bands of the MS experiment plus the morphological features in a stacked vector only $\boldsymbol{x}^{MM} = (\boldsymbol{x}^{R}, \boldsymbol{x}^{G}, \boldsymbol{x}^{B}, \boldsymbol{x}^{O9}, \boldsymbol{x}^{O11}, \ldots, \boldsymbol{x}^{O21}, \boldsymbol{x}^{C9}, \boldsymbol{x}^{C11},\ldots, \boldsymbol{x}^{C21})$, where $\boldsymbol{x}^{O9}$ stands for an opening with a structuring elements of 9 pixels diameter.

As introduced above, the dataset considered (Figure 23.a) is a neighborhood in the North-West of Lausanne and shows mixed residential structures, with individual houses surrounded by apartment blocks and towers, both showing higher urban density. The image was taken in 2004 and has a 1m spatial resolution. Six classes of interest were selected by visual inspection of the image *y = {Trees, Meadows, Individual habitat, Collective habitat, Roads, Shadows}* and 157’723 labeled pixels were highlighted for the analysis. 5'000 randomly selected pixels were used for the training of the model, 10'000 for the tuning of the free parameters and the remaining 142’723 pixels were used to evaluate the performance of the model on unseen data. A multi-class SVM developed using the Torch 3 library (Collobert, 2004) was used for the experiments. Gaussian kernel was used in all the experiments.

In order to compare model performance on new data, the confusion matrices of the predictions of the 142723 test pixels for both the MS (Table 1) and the MM (Table 2) experiments were analyzed. Confusion matrices show the results in terms of predicted pixels (columns) versus ground truth pixels (rows); pixels on the diagonal are correctly classified. The percentage of pixels correctly classified by the SVM is given in the last column, while the last row shows the percentage of pixels correctly classified with respect to the total number of pixels predicted for that class. The visual inspection of the classified images was also carried out to detect improvements in the classification of specific objects. Let's remind that the dimension of the input vector is 3 for the MS model and 17 for the MM model.

For the MS experiment, the global accuracy is 72.7%. Considering accuracies per class, the class *Shadow* is completely neglected by this model and mainly classified as *Trees* because of the color similarity. That means that the model minimizing both the loss function and the generalization term is a model accounting only for 4 classes. That can be explained by the small number of pixels composing the class *Shadow*. The classes which are best classified are *Trees* and *Meadows*. *Roads* are often classified as *Individual habitat* and *Collective habitat*. The color of some roofs is the main reason of such confusion. The same problem is encountered when classifying *Individual* and *Collective habitats*. Here, the similarity in the color of the roofs causes several misclassifications.

| | | **Model output** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MS | Trees | Meadows | I. habitat | C. habitat | Roads | Shadows | Accuracy |
| **Reference** | Trees | **22019** | 2364 | 5 | 105 | 279 | 0 | 89% |
| | Meadows | 1037 | **33984** | 126 | 410 | 169 | 0 | 95 % |
| | I. habitat | 55 | 425 | **15157** | 1047 | 9660 | 0 | 58 % |
| | C. habitat | 288 | 765 | 1348 | **6214** | 2847 | 0 | 54 % |
| | Roads | 370 | 1521 | 9606 | 2572 | **26375** | 0 | 65 % |
| | Shadows | 3931 | 3 | 0 | 0 | 46 | **1** | 0 % |
| | Accuracy | 79 % | 87 % | 58 % | 60 % | 67 % | 0 % | 72.7 % |

Table 1: confusion matrix for MS

The confusion matrix of MM experiment is shown in Table 2. The global accuracy is of 86.1%. Inclusion of morphological features, i.e. including information about the neighboring pixels, has increased performances of the model by 13.4%. The class *Shadow* is not neglected anymore and it is classified at about 90% of accuracy. This is due to the different shape (mainly squared) and size (forest areas are larger than shadows) of shadow areas. The greatest improvement is visible on the classes *Roads*, *Individual* and *Collective Habitats*. For them, the average increase in accuracy is about 17-18%. The classifier can better discriminate these objects, because the morphological features provide information about the size of the object in which the pixel is located.

| | | **Model output** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MM | Trees | Grass | I. habitat | C. habitat | Roads | Shadows | Accuracy |
| **Reference** | Trees | **23715** | 532 | 12 | 82 | 188 | 243 | 96% |
| | Grass | 436 | **34281** | 149 | 440 | 420 | 0 | 96% |
| | I. habitat | 50 | 233 | **19688** | 985 | 5384 | 4 | 75% |
| | C. habitat | 264 | 394 | 1354 | **8278** | 1151 | 21 | 72% |
| | Roads | 244 | 789 | 5223 | 764 | **33413** | 11 | 83% |
| | Shadows | 472 | 0 | 0 | 2 | 4 | **3503** | 88% |
| | Accuracy | 94% | 95% | 75% | 78% | 82% | 93% | 86.1% |

Table 2: confusion matrix for MM

Figure 24 shows the classification maps. For the MS model, two buildings with the roof of the same color are often misclassified (see marker 1a: the collective building is classified as an individual house). This problem is greatly solved by the MM SVM (marker 1b). As mentioned above, the MM model takes into account the information about the structure of the objects and it is therefore able to detect their size and shape. Marker 2a highlights the absence of the class Shadow by the MS model, correctly handled by the MM model (2b). The roofs of the Collective habitat are sometimes made of concrete and are therefore confused with roads by MS (3a). On the contrary, the MM model can better discriminate these objects, even if some confusion is still visible (3b). Green roofs, which are classified as Meadows by MS (4a) are better handled by MM (4b). Finally, the MM model proposes a classification which is less contaminated by high spatial frequencies of the initial image. Therefore, homogeneous surfaces such as *Meadows* and *Trees* (the forest) are more homogeneous (markers 5a and 5b).

The classified results reported have great value when analyzing the urban structure of a city (in our case the spatial distribution of collective and individual habitats). The problem of urban sprawl is strictly related to the question of urban density. Remote sensing images can be used to discriminate automatically between different types of habitat. Machine learning algorithms, used with features that are discriminative for the problem to solve, allow achieving reliable results and provide effective maps for the visualization of urban density.

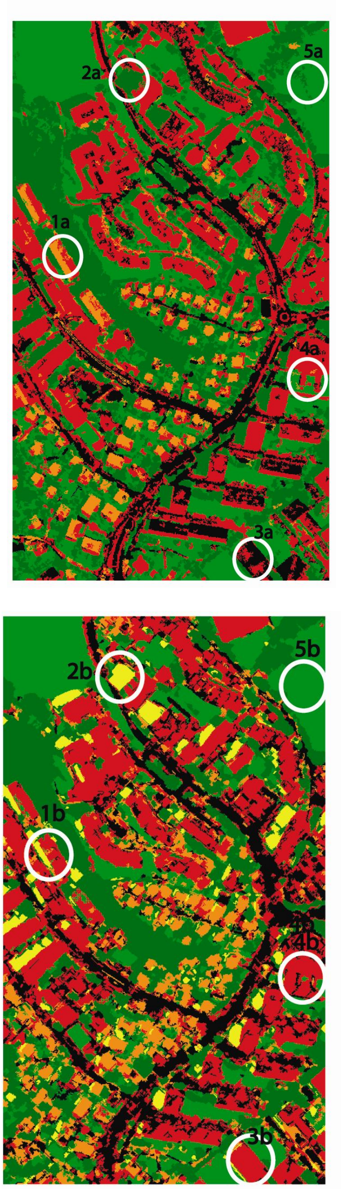


Figure 24: Classified image with RGB (top); classified image with RGB and MM (bottom); Trees = dark green, Meadows = light green, Individual habitat = orange, Collective habitat = red, Roads = black and Shadows = yellow

## Supervised learning for regression

### *Multilayer perceptron*

Multilayer perceptron (MLP) is a “workhorse” of artificial neural networks. It can estimate a dependence function without giving explicitly a mathematical model of how outputs depend on inputs (Haykin, 2009; Bishop, 1995, 2006). It means that there is no prior model of a studied phenomenon whose algorithm should tune to data. MLP “learns from an experience” (or from data).

The following important mathematical result concerning MLP should be mentioned: “MLP with two layers of neurons and non-constant non-decreasing activation function at each hidden neuron can approximate any piecewise continuous function from a closed bounded subset of Euclidean N-dimensional space to Euclidean J-dimensional space with any pre-specified accuracy, provided that sufficiently many neurones are used in the single hidden layer” (Hornik et al., 1989; Cybenko, 1989). This theorem establishes that for any mapping problem, which includes regression and classification, properly trained MLP can find a solution.

The basic unit for information processing, as considered in biological neuroscience, is a neuron. An artificial neuron has inputs that are analogous to dendrites in a biological neuron. It combines these inputs, usually by simple weighted summation, to form an internal activation level. The higher the activation level, the stronger the signal that it will send out to other neurons in the network. The links are called synaptic connections and the bias is known as the activation threshold.

An artificial neuron is a mathematical model that simulates a biological neuron. The simplified model of an artificial neuron is presented in Figure 25. An MLP is a model that simulates a biological neural network, that is, a structured set of interconnected neurons.

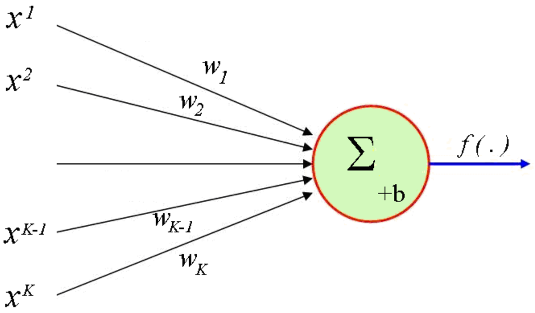


Figure 25: Simple model of artificial neuron

Mathematically, the neuron is the following computational unit

$$Z = f(\sum_{i=1}^{K} w_i x^i + b) \tag{10}$$

which takes the input features $x^i$ (components of some input vector $\boldsymbol{x}$), makes the summation with weights $w_i$, adds a bias $b$ and passes it with a transfer function $f(\cdot)$.

Examples of the activation (transfer) functions are the following S-shaped (sigmoid) functions:

*Logistic* (Figure 26, left):

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{11}$$

*Hyperbolic tangent* (Figure 26, right):

$$f(x) = \tanh(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}} \tag{12}$$

The choice of transfer function is a technical issue due to the stated theoretical properties, and other S-shaped functions are acceptable.

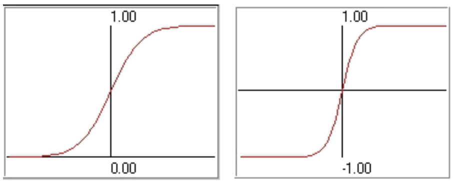


Figure 26: Transfer functions: logistic (left) and hyperbolic tangent (right)

A graphical presentation of MLP structure is given in Figure 27 with a network consisting of 3 inputs, 2 hidden layers with 7 neurons in each and 2 outputs. The network, which solves practical non-linear problems, has hidden (intermediate) layers between the input and output layers.

The power and capabilities of multilayer perceptron stem from the non-linearity used within nodes. An MLP can learn with a supervised learning rule using the backpropagation algorithm. The backward error propagation algorithm (backpropagation) for ANN learning/training caused a breakthrough in the application of multilayer perceptron (Haykin 2009). The backpropagation algorithm gave rise to the iterative gradient algorithms designed to minimize the quadratic error cost function between the actual output of the neural network and the desired output. The error is computed during the forward pass of information flow through the network.

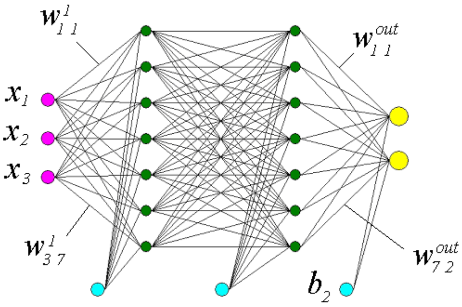


Figure 27: Feed-forward neural network: Multilayer perceptron with 3 input neurons, 7 hidden neurons in the first hidden layer, 7 hidden neurons in the second hidden layer and 2 output neurons (symbolic definition of the net 3-7-7-2). Blue circles are bias neurons (with constant value 1)

*Backpropagation algorithm*

Let us consider briefly the backpropagation algorithm. Although often referred to as a training algorithm, the backpropagation is essentially a method to compute the gradients of the error function with respect to the network weights. These gradients can then be used in any gradient-based optimization algorithm, either of the first- or second-order, online or batch. As usually, for the regression problem the error to be minimized is considered to be a mean squared error (MSE). This error is easily computed, has proved itself in practice and, as shown later, its partial derivatives with respect to individual weights can be computed explicitly.

The outputs of MLP trained with a MSE error function can be interpreted as the conditional mean of the target data, i.e. the regression of a dependent variable (output) conditioned on independent variables (input) (Bishop, 1995; 2006). To simplify the notations, we consider below the model with a single output *t*. It can be easily extended to several outputs by considering the mean squared error averaged over them. For an inputs-output pair (***x**, t*) the error is simply:

$$E_{\mathrm{MSE}}(\boldsymbol{w}) = \frac{1}{2}\left[t - F(\boldsymbol{x}, \boldsymbol{w})\right]^2 \tag{13}$$

Notice that here we are interested in MSE as a function of a set of weights ***w***, since these are the values to be optimized in order to reduce the network error on the training samples.

The basic *backpropagation algorithm* consists of the following steps:

1) Initialization of weights. Usually it is recommended to set all weights and node offsets (biases) to small random values. In many practical studies, the use of simulated annealing to select starting values is more efficient (see Masters, 1993).

2) A pair (input ***x***; desired output *t*) is presented to the network. The actual output of the ANN is computed and the outputs of all the neuron nodes are stored. This completes the forward pass.

3) The derivatives of $E_{\mathrm{MSE}}$ for a single pair (***x**, t*) are computed with respect to the weights in each layer, starting at the output layer with the backward move to the inputs. The derivatives provide information on how much the error depends on the particular weight in the vicinity of a current model, and will be used to optimize its value in order to reduce the error, at least locally. This completes a backward pass.

The key point here is to compute the derivatives of the transfer functions of the neuron nodes with respect to its arguments. Here the smart choice of an activation function comes into play. Since for the logistic function and hyper-tangent, they can be computed through the values of the functions themselves, and the latter are computed and stored after a forward pass, the algorithm simplifies and fastens significantly.

For example, for the logistic function:

$$f = \frac{1}{1+\mathrm{e}^{-x}}, \qquad \frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2} = \frac{1+\mathrm{e}^{-x}-1}{(1+\mathrm{e}^{-x})^2} = (\tfrac{1}{f}-1)f^2 = f(1-f) \tag{14}$$

*Gradient-based MSE minimization*

To minimize the MSE error for the training set, let us construct an iterative procedure of gradient descent. The weights ***w*** are updated iteratively by a gradient rule, with $n$ denoting the iteration number:

$$w_{ij}^{m}(n+1) = w_{ij}^{m}(n) - \eta \frac{\partial E_{\mathrm{MSE}}}{\partial w_{ij}^{m}}(n) \tag{15}$$

that is,

$$w_{ij}^{m}(n+1) = w_{ij}^{m}(n) + \eta \delta_i^m Z_j^{m-1} \tag{16}$$

where $\eta$ is called the rate of learning ($0 < \eta \leq 1$).

A variation of the gradient-based minimization that is often used deals with the addition of a momentum term to the equation for updating weights. In this case one has:

$$w_{ij}^{m}(n+1) = w_{ij}^{m}(n) + \eta \delta_i^m Z_j^{m-1} + \alpha \Delta w_{ij}^{m}(n) \tag{17}$$

where $0 \leq \alpha < 1$ is called a momentum parameter, and $\Delta w_{ij}^{m}(n)$ is an increment of the weight $w_{ij}^{m}$ at the previous iteration.

The effect of the momentum term is to magnify the learning rate for flat regions of the error surface where gradients are more or less constant (or, strictly speaking, were constant at the last iteration). In steep regions of weight space, momentum focuses the movement in a downhill direction by dampening oscillations caused by alternating the sign of the gradient.

Note that it is difficult to define an optimal momentum parameter in advance. In practice, there are more complex algorithms that try to optimize it automatically during the training procedure.

*Batch vs. on-line learning*

There is an important practical question: whether to estimate gradients from a single training example (leading to the so-called stochastic or on-line updating) and make an optimization step at every presented training sample, or first to compute gradients for all training examples (for the epoch) and update the weights once with an averaged gradient (the batch mode). Both approaches are widely used and different recommendations on their efficiency can be found in the literature. In the online case, the order in which the training patterns are presented may affect the direction of the search on the error surface. Some authors (Masters, 1993) prefer using the entire training set for each epoch, because this favors stability in convergence to the optimal weights.

First, all the training samples are presented to the network and the average gradient is computed, that is, the vector containing all the derivatives of MSE and whose dimensionality is equal to the number of weights in the network:

$$\nabla E_{\mathrm{MSE}}(\boldsymbol{w}) = \left\langle \frac{\partial E_{\mathrm{MSE}}}{\partial w_{ij}^{m}} \right\rangle \tag{18}$$

where the average is taken over all the training samples.

The optimization step to modify the vector of weights ***w*** in the batch mode then becomes:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta \nabla E_{\mathrm{MSE}}\left(\boldsymbol{w}(n)\right) \tag{19}$$

The practical aspects of learning the weights (i.e. the optimization algorithm) was an important issue in the development of neural network models. Some popular MLP training approaches deal with the combination of conjugate gradient methods in order to find the local minimum of the error surface, and simulated annealing and/or genetic algorithms in order to escape from the local minima.

Some recent research trends, motivated by the huge size of data sets to be processed, are coming back to the on-line learning scheme, bringing some stochastic elements into the learning process (Bottou, 2003). Interestingly, this sometimes allows both processing of large data and helps in avoiding over-fitting.

*Multiple local minima*

Error surfaces of neural network models are discussed in many textbooks and research papers (see e.g. Hecht-Nielsen, 1990; Haykin, 2009). Because of combinatory permutations of the weights that leave the network input-output function unchanged, these functions typically have a large number of local minima. Therefore the error surfaces can be highly degenerated and have numerous “troughs”. Error surfaces may have a multitude of areas with shallow slopes in multiple dimensions simultaneously. Typically this occurs because particular combinations of weights cause the weighted sums of one or more hidden layer (with sigmoid outputs) to be large in magnitude. When this occurs, the output of that sum (and therefore the value of $E_{MSE}(w)$ is insensitive to small changes in weights, since these simply move the weighted sum value back and forth along one of the shallow tails of the sigmoid function. It has been experimentally established that local minima do actually exist. However, in many problems, convergence to a non-global minimum is acceptable if the error is nevertheless fairly low. The presence of multiple minima does not necessarily presents difficulties in training nets, and a few simple heuristics can often overcome such problems.

**Case study: sediments contamination of lac Léman, Switzerland**

In this real case study, multivariate data on sediments contamination of lac Léman (Geneva lake, Switzerland) are used. The data show the contamination of the lake sediments by heavy metals, from a total of 200 measurements.

The data were divided into two datasets: for training (168 points) and testing (32 points). The structure of MLP was 2-10-10-1 (two inputs - X,Y coordinates, two hidden layers with ten neurons in each hidden layer, and one output – the level of contamination). On the first step of the training procedure, simulated annealing was used to initialize the weights. Then second-order training algorithm Levenberg-Marquardt was used during the main step of the training. The training was stopped as the testing error was increasing. So the model with the minimum test error was selected as a resulting model. The described procedure was repeated 5 times and the model with the lowest test error was adopted as a result.

In Figures 28 (top) the measurements of the level of sediment contamination by zinc (Zn) are presented. In Figure 28 (bottom) the result of MLP mapping is shown. Similarly, the same maps for the sediment contamination by titan (Ti) are given in Figures 29.

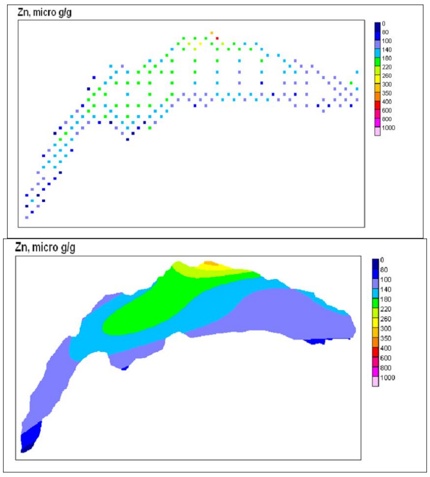


Figure 28: Sediments contamination of lac Léman, Zn: measured (training) data (top), MLP mapping (bottom)

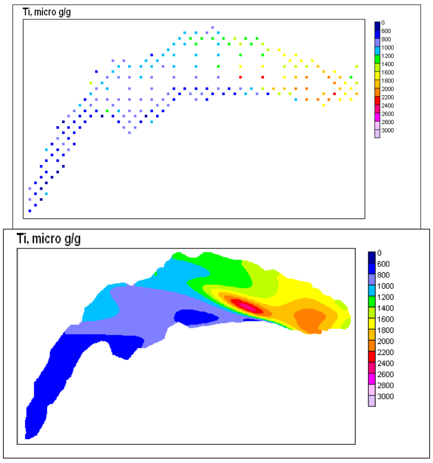


Figure 29: Sediments contamination of lac Léman, Ti: measured (training) data (top), MLP mapping (bottom)

Both MLP have correctly reconstructed spatial structures of pollution patterns. As it was mentioned above, in general, the original data set has to be divided into three parts: training, testing, and validation. The validation subset can be used to estimate the generalization abilities of the model. In the present subsection the main attention was paid only to training of MLP and not to the validation.

### Supervised learning for automatic mapping of geospatial data

MLP is really a powerful tool for performing regression tasks on data of different complexity. But it has one important drawback: difficult and long training procedure.

Automatic modeling technique requires a method which has two important features. First, parameters of the training (tuning) of the algorithm should be selected automatically, and not by the user. Second, the result of mapping should be unique and not dependent on any initial conditions of the training algorithm. One of the possibilities satisfying these conditions is based on General Regression Neural Networks (GRNN) (Kanevski & Maignan, 2004).

*General Regression Neural Network*

GRNN is another name for a well-known statistical nonparametric method - Nadaraya-Watson Kernel Regression Estimator. It was proposed independently in 1964 by Nadaraya (1964) and Watson (1964). In 1991 it was reinterpreted by Specht in terms of neural networks (Specht, 1991). This method is based on kernel nonparametric density estimation proposed by (Parzen, 1962). Details on nonparametric statistics can be found in Hardle (1989) and Fan & Gijbels (1997). See also Fotheringham et al. (2000, 2002) for specific method – Geographically Weighted Regression (GWR), developed in quantitative geography. In fact, the latter is closely related to the local polynomial modeling using a kernel-based approach.

Omitting the details of the mathematical background, let us present the final formula for the regression estimation of *Z(x)* using available measurements $Z_i$:

$$Z(\mathrm{x}) = \frac{\sum_{i=1}^{N} Z_i K\left(\frac{\mathrm{x}-\mathrm{x_i}}{\sigma}\right)}{\sum_{i=1}^{N} K\left(\frac{\mathrm{x}-\mathrm{x_i}}{\sigma}\right)} \qquad i = 1, 2 \ldots, N \tag{20}$$

where *N* is a number of training points, $Z_i$ is a function value of the *i*-th training point having coordinate $x_i$.

The core of this method is a kernel *K*(·). It depends on two parameters: distance to the predicted point and model dependent parameter *σ* – which is a positive number called bandwidth or simply a width of the kernel. Note that $x_i$, in fact, is a centre of the *i*-th kernel.

Different kinds of kernels can be selected from the kernels' library (Hardle, 1989; Fan & Gijbels, 1977). Gaussian is the most widely used kernel

$$K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{\sigma}\right)=\frac{1}{(2\pi\sigma^2)^{p/2}}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right) \qquad i=1,2\dots,N \tag{21}$$

where *p* is a dimension of the input vector *x*.

Gaussian kernel is a standard kernel of GRNN as well. Finally, the GRNN prediction using Gaussian-type of kernel is given by the following formula:

$$Z(\boldsymbol{x})=\frac{\sum_{i=1}^{N} Z_i \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right)}{\sum_{i=1}^{N}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right)} \tag{22}$$

Note that GRNN is a linear estimator (prediction depends on weights linearly), but weights are estimated non-linearly according to the non-linear kernel.

The model described above is the simplest GRNN algorithms. One of the useful improvements is to use multidimensional kernels instead of one-dimensional ones. In a more general setting parameter *σ* may be presented by a covariance matrix. This matrix is a squared symmetrical with dimension *p* by *p* and with the number of parameters equals to *p*(*p*+1)/2.

A model with an anisotropic kernel is much more flexible to model real-world data. It is especially useful in case of complex multidimensional data. For example, for 2D spatial mapping we can use the following parameterization $\sigma$=($\sigma_x$, $\sigma_y$, $\sigma_{xy}$).

Kernel bandwidth is the only hyper-parameter of the GRNN model. There is an optimal value of kernel bandwidth: application of kernels with larger than optimal value of $\sigma$ leads to over-smoothing of data (biased solution); and smaller than optimal value of $\sigma$ produces over-fitting of data.

### GRNN training using cross-validation

As it was mentioned earlier, the only adaptive (free) parameter in the GRNN model is the kernel bandwidth. For its estimation, cross-validation procedure may be applied. In order to find an optimal value of kernel bandwidth usually a grid search is used. It is necessary to define an interval of $\sigma$ values and the number of steps. Then the cross-validation procedure is performed for all $\sigma$ values from the selected interval.

The final result (optimal $\sigma$ value) corresponds to the model with the smallest cross-validation error. The interval and the number of steps have to be consistent in order to catch the expected optimal (with minimum of the error) value. Reliable limits are the minimum distance between points and size of the area under study. In fact, really effective interval is much smaller and can be defined in accordance with the monitoring network structure and/or by using prior expert's knowledge about studied phenomenon.

In case of general anisotropic GRNN model an optimization procedure is performed in a $p(p+1)/2$-dimensional cube in order to find corresponding optimal $\sigma$-values.

### Case study: mapping of transportation data, Switzerland

The Swiss Federal Population Census of the year 2000 includes a matrix of commuter flux between all the 2896 Swiss communes, with the number of commuters for several means of transportation. The commuters which live and work in the same commune are also available. Only the main mean of transportation has been taken into account for each commuter. In Figures 30 (left) the measurements of the number of inhabitants using motorcycles or scooters to reach the job is presented. In Figure 30 (right) the result of GRNN mapping is presented. Similarly, the same maps for habitants using buses or trams are presented in Figures 31.

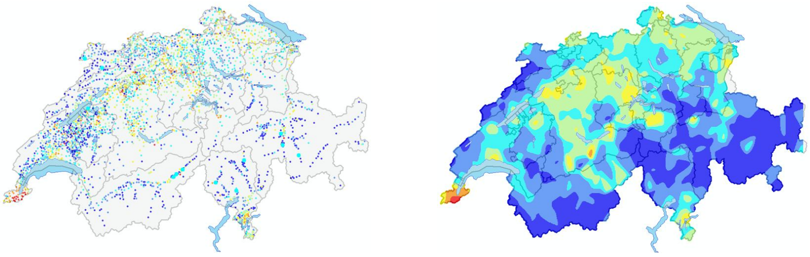


Figure 30: Number of habitants in communes using moto or scooter to reach the job (normalized by population): initial data (left), GRNN mapping (right)

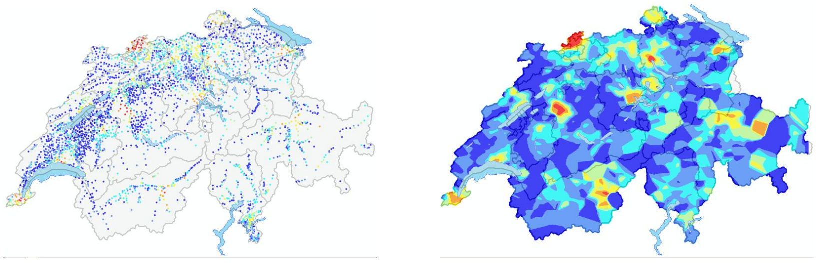


Figure 31: Number of habitants in communes using bus or tram to reach the job (normalized by population): initial data (left), GRNN mapping (right)

Training of GRNN were carried out using cross-validation techniques.

An important generalization of GRNN models can be done by taking into account their similarity with nonparametric statistics: estimation of higher moments, calculation of confidence and prediction intervals, locally adaptive modeling, etc.

## CONCLUSIONS

At present machine learning models/algorithms play a great role in many fields dealing with data analysis modeling and visualization. They are flexible, adaptive, nonlinear, and universal modeling tools based on solid mathematical and statistical background. Despite of some external simplicity, especially taking into account the availability of easy to use software tools, their correct use and interpretation of the results obtained need deep expert knowledge in the corresponding fields. They seem to be indispensable tools when multivariate data are embedded in high dimensional geo-feature spaces and corresponding phenomena are nonlinear, multi-scale and contaminated by noise which is quite a typical situation in real-life applications.

In the present chapter some basic principles of ML models and their efficient applications for geospatial data were considered. After a short general introduction unsupervised and supervised learning techniques were presented with some details and by using real case studies. The future research is foreseen in the directions of applying active learning techniques in order to improve efficiency and quality of data processing using supervised techniques, semi-supervised and manifold learning taking into account unlabeled data and real life constraints. An important research topic is to quantify the uncertainties of the results obtained which will help in decision-making processes. Finally, the problem of spatio-temporal data and science-based models assimilation/integration is still a hot topic of research.

## ACKNOWLEDGEMENTS

The research was supported in part by the Swiss National Science Foundation “GeoKernels. Phase 2” (200020-121835) and “ClusterVille” 100012-113506. The authors thank the CIPEL organization for providing data on lac Léman (Geneva lake).

## REFERENCES

**Almeida, C.; Gleriani, J.; Castejon, E.; Soares-Filho, B**., 2008, Using neural networks and cellular automata fro modeling intra-urban land-use dynamics. In: International Journal of Geographical Information Science, 22: 943-963

**Agarwal, P.; Skupin, A**., 2008, Self-organising Maps: Applications in Geographic Information Sciences. Wiley: New York

**Aha, D.W.** (Ed.), 1997, Lazy Learning. Kluwer Academic: Dordrecht

**Belkin, M.; Niyogi, P.,** 2003, Laplacian eigenmaps for dimensionality reduction and data representation. In: Neural Computation, 15, 6: 1373–1396

**Benediktsson, J.A.**, 2003, Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. In: IEEE Transactions on Geoscience and Remote Sensing, 41: 1940-1949

**Bishop, C.M.,** 2006, Patter Recognition and Machine Learning. Springer: Singapour

**Bishop, C.M.,** 1995, Neural Networks for Pattern Recognition. Oxford University Press: New York

**Boser, B.; Guyon, I.; Vapnik, V.**, 1992, A training algorithm for optimal margin classifiers. In: 5th ACM Workshop on Computational Learning Theory

**Bottou, L.,** 2003, Stochastic Learning, Advanced Lectures on Machine Learning. In: Bousquet, O.; von Luxburg, U. (Eds), Lecture Notes in Artificial Intelligence. Berlin: Springer: 146-168

**Chapelle, O.; Shoelkopf, B.; Zien, A.** (Eds), 2006, Semi-Supervised Learning (Adaptive Computation and Machine Learning). MIT Press: Cambridge MA

**Cherkassky, V.; Hsieh, W.; Krasnopolsky, V.; Solomatine, D.; Valdes, J**., 2007, Special Issue: Computational intelligence in earth and environmental sciences. In: Neural Networks, 20, 4: 433-558

**Cherkassky, V.; Krasnopolsky, V.; Solomatine, D.; Valdes, J**. (Eds), 2006, Special Issue: Earth Sciences and Environmental Applications of Computational Intelligence. Introduction. In: Neural Networks, 19, 2: 111-250

**Cherkassky, V.; Mulier, F.**, 2007, Learning from Data. Concepts, Theory, and Methods. Second edition. Edition.Wiley-Interscience: New York

**Chiles, J.P.; Delfiner P.,** 1999, Geostatistics. Modeling Spatial Uncertainty. Wiley: New York

**Collobert, R.; Bengio, S.; Mariéthoz, J.,** 2002, Torch: a modular machine learning software library. Tech Report IDIAP: Martigny

**Cristianini, N.; Shawe-Taylor, J.**, 2000, An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press: Cambridge

**Cybenko, G.V.,** 1989, Approximation by Superpositions of a Sigmoidal function. In: Mathematics of Control, Signals and Systems, 2, 4: 303-314

**Dubois, G.,** 2005, Automatic mapping algorithms for routine and emergency data. European Commission, JRC Ispra, EUR 21595

**Fan, J.; Gijbels, I.,** 1997, Applied Local Polynomial Modeling and Its Applications. In: Monographs on Statistics and Applied Probability 66. London: Chapman and Hall

**Fotheringham, A.; Brunsdon, Ch.; Charlton, M.,** 2000, Quantitative Geography. Perspectives on Spatial Data Analysis. SAGE Publications: London

**Fotheringham, A.; Brunsdon, Ch.; Charlton, M.,** 2002, Geographically Weighted Regression. The Analysis of Spatially Varying Relationships. John Wiley & Sons: West Sussex

**Hagen, L.; Kahng, A.,** 1992, New spectral methods for ratio cut partitioning and clustering. In: IEEE Trans. on Computer Aided-Design, 11, 9:1074-1085

**Grandvalet, Y.; Canu, S.; Boucheron, S.,** 1997, Noise injection: theoretical prospects. In: Neural computation, 9: 1093-1108

**Guyon, I.; Gunn, S.; Nikravesh, N.; Zadeh, L.,** 2006, Feature Extraction: Foundations and Applications. Springer: New York

**Hardle, W.,** 1989, Applied Nonparametric Regression. Cambridge University Press: Cambridge

**Hastie, T.; Tibshirani, R.; Friedman, J.**, 2009, The Elements of Statistical Learning; Data Mining, Inference, and Prediction. Second edition. Springer Verlag: New York

**Haykin, S.,** 2009, Neural Networks and Learning Machines. Third Edition. Prentice-Hall, Inc.: New York

**Hecht-Nielsen, R.,** 1990, Neurocomputing. Addison-Wesley: Reading

**Hewitson, B.; Crane, R.**, 1994, Neural Nets: Applications in Geography. Kindle Edition

**Hornik, K.; Stinchcombe, M.; White, H.,** 1989, Multilayer feedforward networks are universal approximations. In: Neural Networks, 2: 359-366

**Izenman, A.,** 2008, Modern Multivariate Statistical techniques. Regression, Classification, and Manifold Learning. Springer: New York

**Jain, A.K.; Murty, M.N.; Flynn, P.J.,** 1999, Data clustering: a review. In: ACM Computing Surveys, 31, 3: 264-323

**Jones, A.,** 2004, New tools in non-linear modeling and prediction. In: Comput. Managm. Sci., 1**:** 109-149

**Kanevski, M.; Maignan, M.,** 2004, Analysis and Modeling of Spatial Environmental Data. EPFL Press: Lausanne

**Kanevski, M.; Arutyunyan, R.; Bolshov, L.; Demyanov, V.; Maignan, M.,** 1996, Artificial neural networks and spatial estimations of Chernobyl fallout. In: Geoinformatics, 7, 1-2: 5-11

**Kanevski, M.,** 1999, Spatial Predictions of Soil Contamination Using General Regression Neural Networks. In: Systems Research and Information Systems, 8, 4: 241-256

**Kanevski, M**. (Ed.), 2008, Advanced Mapping of Spatial Environmental Data. iSTE-Wiley: London

**Kanevski, M.; Pozdnoukhov, A.; Timonin, V.,** 2009, Machine Learning for Spatial Environmental Data. Theory, Applications and Software. EPFL Press: Lausanne

**Kohonen, T.,** 2000, Self-organising maps. 3rd Edition. Springer: New York

**Lee, J.; Verleysen, M.,** 2007, Nonlinear Dimensionality Reduction. Springer: New York

**Masters, T.,** 1993, Practical Neural Network Recipes in C++. Academic Press: New York

**Nadaraya, E.A.,** 1964, On estimating regression. In: Theory of Probability and its Applications, 9: 141-142

**Ng, A.Y.; Jordan, M.; Weiss, Y.,** 2001, On spectral clustering: analysis and an algorithm. In: Glen Th. et al. (Eds), Advances in Neural Information Processing Systems, 14: 849-856

**Openshaw, S.; Openshaw, C.,** 1997**,** Artificial Intelligence in Geography. Wiley: London

**Parzen, E.,** 1962, On estimation of a probability density function and mode. In: Annals of Mathematical Statistics, 33: 1065-1076

**Pearson, K.,** 1901, On Lines and Planes of Closest Fit to Systems of Points in Space. In: Philosophical Magazine 2, 6: 559–572

**Pi, H.; Peterson, C.,** 1994, Finding embedding dimension and variable dependencies in time series. In: Neural computation, 6: 509-520

**Pesaresi, M.; Benediktsson, J.A.**, 2001, A new approach for the morphological segmentation of high-resolution satellite images. In: IEEE Transactions on Geoscience and Remote Sensing, 392: 309-320

**Pijanowski, B.; Brown, D.; Shellito, B.; Manik G.,** 2002, Using neural networks and GIS to forecast land use changes: a Land Transformation Model. In: Computers, Environment and Urban Systems, 26: 553-575

**Rosenblatt, M.,** 1956, Remarks on some nonparametric estimates of a density function. In: Annals of Mathematical Statistics, 27: 832-837

**Roweis, S.T.; Saul, L.K.,** 2000, Nonlinear dimensionality reduction by locally linear embedding. In: Science, 290, 5500: 2323-2326

**Shi, J.; Malik, J.,** 2000, Normalized cuts and image segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 8: 888-905

**Soille, P.**, 2004, Morphological Image Analysis. Springer-Verlag: Berlin

**Specht, D.E.,** 1991, A General Regression Neural Network. In: IEEE Transactions on Neural Networks, 2: 568-576

**Tenenbaum, J.B.; de Silva, V.; Langford, J.C.,** 2000, A global geometric framework for nonlinear dimensionality reduction. In: Science, 290, 5500: 2319-2323

**Timonin, V.; Savelieva, E.,** 2005, Spatial Prediction of Radioactivity Using General Regression Neural Network. In: Applied GIS, 1, 2: 19-01 to 19-14. DOI: 10.2104/ag050019

**Vapnik, V.**, 1998, Statistical Learning Theory. Springer-Verlag: Berlin

**Watson, G.S.,** 1964, Smooth regression analysis. In: Sankhya: The Indian Journal of Statistics, Series A, 26: 359-372

**Weinberger, K.Q.; Saul, L.K.,** 2005, Nonlinear dimensionality reduction by semi-definite programming and kernel matrix factorization. Paper presented at the Tenth International Workshop on AI and Statistics (AISTATS-05)

## AUTHORS INFORMATION

**Mikhail KANEVSKI**
Mikhail.Kanevski@unil.ch
IGAR, University of Lausannee
Amphipole, 1015 Lausanne
Switzerland

**Loris FORESTI**
Loris.Foresti@unil.ch
IGAR, University of
Lausannee, Amphipole,
1015 Lausanne, Switzerland

**Christian KAISER**
Christian.Kaiser@unil.ch
IGUL, University of Lausannee
Antropole, 1015 Lausanne
Switzerland

**Alexei POZDNOUKHOV**
Alexei.Pozdnoukhov@unil.ch
IGAR, University of Lausannee
Amphipole, 1015 Lausanne
Switzerland

**Vadim TIMONIN**
Vadim.Timonin@unil.ch
IGAR, University of
Lausannee
Amphipole, 1015 Lausanne
Switzerland

**Devis TUIA**
Devis.Tuia@unil.ch
IGAR, University of Lausannee
Amphipole, 1015 Lausanne
Switzerland

# SOCIAL ACCOUNTING MATRICES: THE DEVELOPMENT AND APPLICATION OF SAMS AT THE LOCAL LEVEL

**Eveline S. van LEEUWEN and Peter NIJKAMP**

Department of spatial economics, VU University Amsterdam, the Netherlands

## ABSTRACT

This contribution aims to highlight the importance of Social Accounting Matrices (SAMs) for the study of regional-economic interactions. After a conceptual review of SAMs, the attention is focused on the empirical meaning of SAMs for economic impact assessment. The potential of SAMs is illustrated by an extensive pedagogical treatment of this tool on the basis of several town-hinterland interactions in 5 different European countries.

## KEYWORDS

Social Accounting Matrices, Multiplier analysis, Towns, Hinterland, Interregional effects

## INTRODUCTION

The Social Accounting Matrix (SAM) has a respectable history in economic research. It comprises a comprehensive, disaggregated, consistent and complete data collection that captures the interdependences that exist within a socio-economic system (Isard et al., 1998). If, for example, households had to pay less tax, they could spend more money on fresh food or beverages. They might then go to a supermarket and spend a larger share of their income there. As a result, the supermarket -or the retail sector in general- would have to obtain more products from the food production sector, which would raise its demand for agricultural products. Because of this increasing demand, more labour input would be needed which would further increase the income of certain households, which again could spend more money. This kind of interdependency between sectors and households can very well be captured within a SAM.

SAMs were initially developed because of growing dissatisfaction with the distributional effects of conventional growth policies, especially in developing countries (see, e.g., Adelman & Robinson, 1978; Pyatt & Round, 1977). In these countries income redistribution is often an important concern. Therefore, researchers in the late 1970s were eager to learn more about the processes and mechanisms dealing with the production of goods and services as well as with the associated income formation and income distribution. Traditionally, input-output models (developed by Leontief already in 1951) were used to analyse production linkages in an economy. Input-output analysis is an established technique in quantitative economic research. It belongs to the family of impact assessment methods and aims to map out the direct and indirect consequences of an initial impulse into an economic system across all economic sectors. It is essentially a method that depicts the system-wide effects of an exogenous change in a relevant economic system (van Leeuwen et al., 2005).

Input-output models are based on the idea that any output requires a corresponding input. Such input may comprise raw materials and services, all coming from other sectors but also labour from households or certain amenities provided by the government. The output consists of a sectoral variety of products and services. Most input-output models are structured to trace changes in the flows of capital and labour between industries in response to a change in final demand; they are demand-driven. However, a conventional input-output model does not take into account the link between increased output, the factorial and household income distribution and increased consumption. Therefore, a new kind of model had to be developed. SAMs combine data on production and income generation, as can be found in input-output tables, together with data about incomes received by different institutions and on the spending of these incomes. Therefore, a SAM allows us not only to analyse (regional) production linkages but also to focus on production-income and income-expenditure relations in a given area, so that distributional effects of a change in final demand can be analysed.

Nowadays, a natural extension of a SAM, a static framework with fixed prices, is a computable general equilibrium (CGE) model, which can be considered dynamic with endogenized prices (Isard et al. 1998). CGE models use a SAM as the base-year but they include also a number of behavioural and structural relationships to describe the behaviour of various actors over time. The CGE approach permits prices of inputs to vary with respect to changes in output prices and, thus, allows the behaviour of economic agents to be captured (van den Bergh & Hofkes, 1999). Notwithstanding the advantages of CGE models, we will address SAMs in this chapter. An important reason for this is that SAMs are able to handle a very disaggregated sector structure. In our empirical analysis, the economic linkages between town and hinterland actors will be described. It is generally thought that towns are important concentration points of economic activities in rural areas, thereby having the capacity to act as a focal point of trade and services for the hinterland. Our analysis will focus on the current economic structure of towns and hinterland, and on existing linkages between those areas using 30 European SAMs describing the local town and hinterland economy (see Appendix A.I for a list of the towns, their population and number of jobs).

In this chapter, we will first describe the SAM framework, including examples of existing SAM-based studies, the structure of a SAM, and its advantages and disadvantages. Next, we will focus on regional SAMs, including the development of a SAM at town level. Then, multiplier analysis will be explained followed by a section with empirical results, describing interregional SAM multipliers at the town-hinterland level. Finally, conclusions are formulated.

## THE SAM FRAMEWORK

### Examples of SAM-based studies

The SAM methodology has been used extensively to analyse a variety of different questions at different levels of geographical aggregation (Isard et al., 1998).

In developing countries, it has been used widely to explore issues such as income distribution (Adelman & Robinson, 1978), the role of the public sector (Pleskovic & Trevino, 1985), and the impact of inter-sectoral linkages on (rural) poverty alleviation (Thorbecke, 1995; Khan, 1999).

In developed countries, SAMs at the national level have been used to analyse the effect of different taxation or subsidy schemes on income distribution (e.g. Roland-Holst & Sancho, 1992; Psaltopoulos et al., 2006). In addition, at present, much emphasis is put on environmental flows, instead of monetary flows. These SAMs are able to integrate, for example, physical water circular flows or emissions into the atmosphere by greenhouse gases (GE), together with the economic flows sourced from the National Accounting (see e.g. Morilla et al., 2007). Another example is the study of Sánchez-Chóliza et al. (2007). Their objective was to assess the environmental impact of the lifestyle enjoyed by the population of Spain; and to estimate the total and per capita pollution associated with household activity. The use of a SAM model facilitated the understanding of how the pollution associated with household activity and consumption patterns “circulates” throughout the map of an economy. The SAM accounts were expressed in terms of different kinds of pollution, such as waste water, $NO_x$, or $CO_2$.

Furthermore, examples can be found of applications at regional or town level. Most of them deal with towns in developing countries (see e.g. Adelman et al., 1988; Parikh & Thorbecke, 1996). Lewis (1991) describes a SAM application on town level of the Kenyan town Kutus. The SAM encompasses both the town of around 5,000 inhabitants, and the 8 km zone around it (hinterland) with a population of 42,000. The SAM was used to test the governmental assumption of agriculturally-driven regional economies and to evaluate non-agricultural production sector activities in the Kutus region. According to the Lewis's multiplier analysis, agricultural activities were indeed very important for the stimulation of regional output and income.

The SAMs used in the present chapter are also developed for spatially disaggregated levels, such as town-hinterland interactions. They are able to make a distinction between the town, a hinterland zone, and the rest of the world (ROW); they are typically interregional SAMs. They will be used to explore the relative economic importance of towns and hinterlands and to distinguish which sectors can be identified as key sectors.

### Structure of a SAM table

A SAM can be described as a general equilibrium data system of income and expenditure accounts, linking production activities, factors of production, and institutions in an economy (Courtney et al., 2007).

Figure 1 shows the economic flows and interrelations captured by a SAM. The industrial production generates value added which is used to pay for primary inputs. These primary inputs consist of profits, wages, and payments to the government. Next, these incomes or receipts, generated in production, are handed over to households or the government. After a redistribution process, incomes are either used for (final) consumption or they are saved. The final consumption leads to new production by industries, and the whole process starts again.

From Figure 1 it becomes clear that input-output tables, which only focus on production linkages, ignore the effects arising from other linkages, as exist, for example, between households' income and the production sectors (final demand).

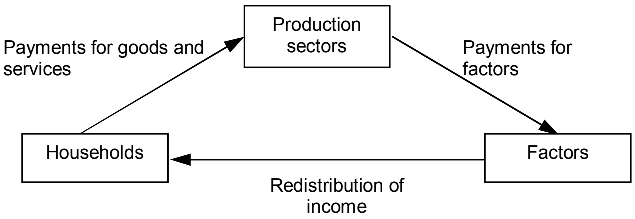


Figure 1: The direction of income flows between the three main types of accounts in a SAM. Source: based on Roberts (2005)

Similar to an input-output table, a SAM presents a series of accounts together in one matrix. It contains a complete list of accounts describing income, expenditure, transfers and production flows (Cohen, 1989). In input-output models, usually only the production accounts are endogenous (implying that changes in the level of expenditures directly follow a change in income), and the factor and household accounts are exogenous (implying that expenditures are set independently of income changes). In a SAM, the production factors, as well as the households' accounts, are endogenous. The exogenous or independent accounts can consist of payments to, and receipts from, the government, actors outside the research area, and investments, value added or savings. Table 1 shows the elements of a (general) SAM.

The first account is the production accounts which are rather similar to an input-output table. The columns of the production accounts describe how firms buy raw materials and intermediate goods from other firms ($A_{11}$). Furthermore, a SAM includes information about the costs of hiring factor services ($A_{21}$) to produce commodities ($Y'_1$). The exogenous part of the first column includes expenditures in the ROW, and value-added, part of which is paid to the government. The rows, which show the receipts, describe the sales to domestic intermediate industries ($A_{11}$), to final consumption of households ($A_{13}$), and to exports to the ROW ($X_1$). The sales to firms or households in the ROW form the exogenous accounts.

| From<br>To | | Endogenous accounts | | | Exogenous accounts | Total |
|---|---|---|---|---|---|---|
| | | Production | Factors | Households | | |
| Endogenous accounts | Production | $A_{11}$ | | $A_{13}$ | $X_1$ | $Y_1$ |
| | Factors | $A_{21}$ | | | $X_2$ | $Y_2$ |
| | Households | | $A_{32}$ | | $X_3$ | $Y_3$ |
| Exogenous accounts | | Residual balance* | | | | |
| Total | | $Y'_1$ | $Y'_2$ | Y'3 | | |

Table 1: The elements of a SAM table. * Used to meet the assumption that Y1= Y'1. Source: Based on Cohen (1989)

The factor accounts include labour and capital accounts. The rows show received payments in the form of wages ($A_{21}$). Factor revenues, such as labour income and part of the profits, are distributed to households ($A_{32}$), after paying the corresponding taxes to the government.

The exogenous part of the factor accounts includes payments to households in the ROW from town or hinterland industries, as well as wage payments of ROW industries to local households.

Finally, the households' accounts include the factor incomes described above ($A_{32}$), as well as household expenditures on the local market ($A_{13}$). The exogenous part ($X_3$) describes direct taxes and the savings from households, as well as their consumption in the ROW.

A SAM is balanced when savings is equal to receipts minus expenditure for all actors (firms and households). Because we do not have information about savings, we use the residual balance to make sure that the row and column sums are the same.

**Advantages and disadvantages of SAM analysis**

A SAM is an analytical and predictive tool to represent and forecast system-wise effects of changes in exogenous factors. A great advantage of a SAM is its ability to capture a wide variety of developments in a (macro-) economy, as it links production, factor and income accounts. A large share of economic interactions takes place within the household sector and a SAM disaggregates the cells involving 'returns for labour' and the household sector into smaller groups (such as different income groups) to show the effect of the different behaviour of these groups. Furthermore, it is a relatively efficient way of presenting data; the presentation of data in a SAM immediately shows the origin and destination of the various flows included. Another advantage is its usefulness as a tool to reconcile different data sources and fill in the gaps. This enables the reliability of existing data to be improved and inconsistencies in data sets of different nature and origin to be revealed (Alarcon et al., 1991).

Most of the disadvantages of a SAM are similar to the disadvantages of input-output tables and concern the production activities accounts. Important, and sometimes restrictive, assumptions made in the input-output model, as well as in the SAM, are that all firms in a given industry employ a constant production technology (usually assumed to be the national average of input, output and labour for that industry), and produce identical products. Because the tables are produced only for a certain period, the model can become irrelevant as a forecasting tool when production techniques change. Other disadvantages are that the model assumes that there are no economies or diseconomies of scale in production or factor substitution, and that they do not incorporate the existence of supply constraints. In a rather static situation, these ceteris paribus conditions are a perfectly acceptable position which has demonstrated its great relevance in a long (spatial-) economic research tradition. However, in a highly dynamic context, with complex space-time system interactions, stable solution trajectories are less likely to occur (Nijkamp, 2007). Finally, the production accounts are essentially based on a linear production technology; doubling the level of agricultural production will in turn double the inputs, the number of jobs, etc. This reveals something of the inflexibility of the model. Thus, the model is entirely demand-driven, implying that bottlenecks in the supply of inputs, or increasing efficiency effects are largely ignored (van Leeuwen et al., 2005).

There are also some practical problems in the development of a (local) SAM. The statistical estimation of a new matrix is very labour-intensive and expensive. This is mainly because much of the information is gathered with help of micro-survey questionnaires. A related problem with this method is that interviewees, firms, or households, are not able to give perfect answers. Sometimes they do not understand the question, or they do not want to tell the truth, and therefore -as a result of a response bias- the results are not always perfect. However, a SAM is still a very useful tool in that it shows effects throughout the whole economy, linking the different accounts.

## A REGIONAL SAM

### From a national to a regional model

The construction of a SAM always involves the integration of data from different data sets. Data required for production accounts often come from input-output tables (which are more widely available) and the distribution flows to institutions come from national income and expenditure accounts. Therefore, the majority of studies using SAMs concern the economies of single countries. Although an economic unit does not necessarily have to be a country, the national borders do provide a natural and artificial boundary for defining a macroeconomic unit (Round, 1988). Often information is available at the national level, which makes it a lot easier to develop a national input-output table or SAM. However, many economic processes on a regional level are very different from those at the national level. Regional, spatial or institutional differences can bring about important economic differences. Smaller regions, for example, are more dependent on trade with other areas, both for the sales of outputs (export) and for the purchase of inputs (import) (Miller & Blair, 1985). Therefore, it can be necessary to develop a regional SAM.

There are several ways to regionalize a national input-output table or a SAM. According to Isard et al. (1998), the more disaggregated a SAM needs to be, the more extensive are the data requirements. They state that the best way to build a regional SAM is to start with the regionalization of the production activities' account using a national input-output table. The simplest way is to use a ‘non-survey method’. Another way is to use the GRIT method: Generating Regionalized Input-output Tables. The GRIT method, developed by Jensen et al. (1979), has the advantage that it combines non-survey methods with survey methods. The GRIT system is designed to produce regional tables that are consistent in accounting terms with each other and with the national table. Therefore it uses location quotients, which describe the regional importance of an industry compared with its national importance, by using output-ratios. However, the developer is able to determine the extent of interference with the statistical processes by introducing primary (e.g. from questionnaires) or other superior regional data. In the next section we will describe the development of SAMs at the local, town-hinterland level.

### Interregional SAMs at town-hinterland level

*Data collection*

For our study, we used data that was collected as part of a trans-national project, the European Union research project ‘Marketowns’[1]. This project focused on the role of small and medium-sized towns as growth poles in regional economic development. For this purpose, it was necessary to measure the flow of goods, services and labour between firms and households in a sample of 30 small and medium-sized rural towns in five EU countries. The participating countries reflect the varied conditions of the existing and enlarged European Union, viz. France, Poland, Portugal, the Netherlands and the UK. In each of the participating countries, six small and medium-sized towns were selected with reference to a set of relevant, predefined criteria: for instance, the condition that no other town with more than 3,000 inhabitants should be located in a hinterland with a radius of approximately 7 km.

In order to compare the nature and strength of linkages throughout the wider economy, we defined three different zones for each town: town, hinterland and the rest of the world (ROW). These were designed to facilitate comparisons between the different areas. As a result, the study area from which households and firms were sampled comprised the town and the hinterland, a 7 km radius around it.

Primary data were collected using self-completion survey techniques to measure the spatial economic behaviour of households and firms. The

[1] The information contained in this chapter is drawn from the MARKETOWNS project funded by the European Commission under the Fifth Framework Programme for Research and Technology Development, Contract QLRT - 2000-01923. The project involves the collaboration of the University of Reading (UK), the University of Plymouth (UK), the Joint Research Unit INRA-ENESAD (France), Agricultural Economics Research Institute LEI (The Netherlands), the Polish Academy of Sciences (Poland) and the University of Trás-os-Montes and Alto Douro (Portugal).

household questionnaire focused on spatial patterns of consumer purchases by distinguishing between different categories of goods and services and expenditure patterns across the three pre-defined geographical zones. Furthermore the place of work was identified. The firm questionnaire dealt with spatial patterns of input and output transactions, including labour costs. Surveys were carried out between September 2002 and May 2003 (Terluin et al., 2003).

*Developing an interregional SAM*

A specific classification and disaggregation of a SAM depends on the questions which the SAM methodology is expected to answer. In this case, the aim is to focus on the spatial interdependency of town and hinterland actors (see also Mayfield et al., 2005). This means that a bi-regional SAM has to be developed, describing both the town and its hinterland, which results in four systems of endogenous accounts (see also Appendix A.II):

- Linkages within the town;
- Linkages within the hinterland;
- Flows from town to hinterland;
- Flows from hinterland to town.

For the generation of the interregional SAM, the most important data are the national input-output table and secondary data, such as number of firms or number of jobs, obtained from government institutions, as well as primary data from (local) surveys (see Figure 2). Once this information has been collected, the next step is to develop a regional input-output table by the use of the earlier described GRIT method. The GRIT method uses location quotients, which describe the regional importance of an industry compared with its national importance, by using output-ratios. Together with additional secondary data on commuting patterns and on production values, value added, employment level, savings, investments, imports, and exports, a regional input-output table describing the town and a table describing the hinterland can be generated.

As mentioned earlier, the most structural difference between a (regional) input-output table and a (regional) SAM is the information on household expenditures, wages, employment, etc. Therefore, secondary data, together with information from the surveys on household groups and firm groups, need to be added and combined with the two regional input-output tables. After the regional SAM has been generated, expert opinions[2] can be requested to verify the cell values of the matrix.

Although the development of the SAMs should take place with great care, it is important to keep in mind that the local focus of the models that have been built results in its own limitations. One of the major problems is the relatively small proportion of the total inputs and outputs from firm production that is retained within the local economy, resulting in small coefficients, making them more liable to statistical error. Another limitation is that the secondary data collected in the five countries (especially in Portugal and Poland) is not exactly the same (sometimes there even was no data available at all) (Mayfield et al., 2005), resulting in various creative solutions.

However, finally, 30 SAMs were developed (see Appendix A.1 for a list of towns), each consisting of 17 production accounts, 4 production factor accounts, 4 household accounts and an exogenous ROW account (see Appendix A.III). Together, they form a very interesting and unique database, especially because they enable us to perform a thorough analysis and comparison of the economic structure of a set of towns located in five different European countries.

[2] In this case, local stakeholders (policy makers and persons who are acquainted with the local economy) were asked to verify the results.

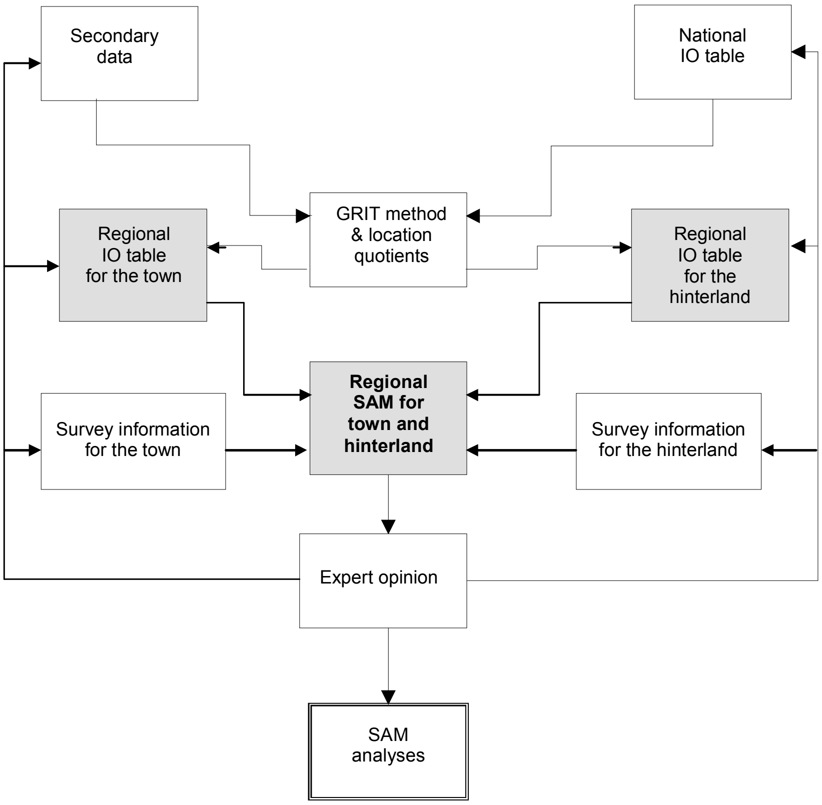

Secondary data
National IO table
GRIT method & location quotients
Regional IO table for the town
Regional IO table for the hinterland
Regional SAM for town and hinterland
Survey information for the town
Survey information for the hinterland
Expert opinion
SAM analyses

Figure 2: Procedure to construct interregional SAMs (Mayfield et al., 2005)

## MULTIPLIER ANALYSIS

### Introduction

SAMs, as I-O tables, can be used to construct multipliers based on the estimated re-circulation of spending within the region: recipients use some of their income for consumption spending, which then results in further income and employment. This 'multiplier effect' appears at three levels. First, the *direct effect* of (production) changes: for example, an increase in retail demand because of a growing population will directly increase the output of the retail industry. *Indirect effects* result from various rounds of the re-spending of, for example, retail receipts in linked industries, such as wholesale or the food sector. This will have an indirect effect on these industries. The third level of effects is the *induced effect.* This effect only occurs when the household accounts are endogenous (which means that it responds to a change in income) as in a SAM. The induced effects include changes in economic activity resulting from household spending of income earned directly or indirectly. These households can, for example, be employees of supermarkets, who spend their income in the local economy (van Leeuwen et al., 2005). The three most frequently used types of multipliers are those that estimate the effects on: (1) outputs of the industries; (2) income earned by households because of new outputs; and (3) employment generated because of the new outputs. In this section, we look at the composition of those multipliers and identify important sectors for the town and hinterland economy.

**Variations in multiplier values in the literature**

The values of the multipliers can differ because of different factors. The size of the multipliers depends, first of all, on the choice of the exogenous and endogenous variables which, in turn, depend on the problem studied (Cohen, 1999). Furthermore, the size depends on the overall size and economic diversity of the region's economy. Regions with large, diversified economies which produce many goods and services will have high multipliers, as households and businesses can find most of the goods and services they need in their own region. Smaller regions, such as cities or towns, will need to import more products and labour (imports can be considered as leakage), resulting in lower multipliers. Regions that serve as central places for the surrounding area will have higher multipliers than more isolated areas. Besides this, the level of economic development is important. Economic theory predicts a higher share of government and more foreign trade at higher levels of economic development, leading to an expected lower output multiplier at a higher development level (Cohen, 1999). Furthermore, the nature of the specific industries concerned can have a significant effect. Multipliers vary across different industries of the economy based on the mix of labour and other inputs and the tendency of each industry to buy goods and services from within the region (less leakage to other regions) (van Leeuwen et al., 2005). The value of SAM multipliers is higher compared with input-output multipliers because, besides capturing effects from production activities, they also include effects on factor and household incomes. The range of values of SAM output multipliers on a national scale lies between 2.1 and 4.5 (see Vogel, 1994; Blane, 1991; Cohen, 1999; Archarya, 2007). As expected, SAM output multipliers at a local or regional scale are usually lower, and have values between 1.3 and 2.3 (see Roberts, 1998; Cohen, 1996; Psaltopoulos et al., 2006). The income multipliers are generally lower compared with output multipliers: at a local scale the values typically range between 1.2 and 1.6.

## INTERREGIONAL SAM MULTIPLIERS AT TOWN-HINTERLAND LEVEL

As described earlier, the town-hinterland SAMs include four systems of endogenous accounts: town-town, hinterland-hinterland, town-hinterland, and hinterland-town flows. The total SAM multiplier is a product of three matrixes: M1, M2 and M3 (see Mayfield et al., 2005). First of all, M1 is the intraregional multiplier matrix, depicting the linkage effects between endogenous accounts wholly within the actors' 'own region' (town or hinterland). Secondly, M2 can be interpreted as the multipliers for all the cross-flows between the town and hinterland. It captures the effects from the town on the hinterland, and vice versa. Thirdly, M3 indicates the 'closed loop' multiplier matrix. This matrix shows the effect that a shock in the town (or hinterland) has on itself through the endogenously defined linkages within the hinterland (or town). Table 2 shows the M1 and M2 multipliers for a shock in the production sector, factor accounts, or household income.

| | | Production | Factor | Household | Production | Factor | Household |
|---|---|---|---|---|---|---|---|
| | | Town | | | Hinterland | | |
| Production | Town | $M1_{town}$<br>(output) | $M1_{town}$<br>(factor) | $M1_{town}$<br>(income) | $M2_{hinterland}$<br>(output) | $M2_{hinterland}$<br>(factor) | $M2_{hinterland}$<br>(income) |
| Factor | | | | | | | |
| Household | | | | | | | |
| | | Town | | | Hinterland | | |
| Production | Hinterland | $M2_{town}$<br>(output) | $M2_{town}$<br>(factor) | $M2_{town}$<br>(income) | $M1_{hinterland}$<br>(output) | $M1_{hinterland}$<br>(factor) | $M1_{hinterland}$<br>(income) |
| Factor | | | | | | | |
| Household | | | | | | | |

Table 2: M1 and M2 output multipliers for town and hinterland (shock to production, factors, or household income)

Evidently, this methodology results in a great number of (sub-) multipliers (output, factor, and income) as well as the possibilities to show linkages between town and hinterland. Our aim is to use the interregional SAMs to find out, for towns in 5 European countries, what important sectors in both town and hinterland economies are, and how strong are the linkages between production and households and between town and hinterland.

### SAM income multipliers

SAM household income multipliers (summation of the $M_1$, $M_2$ and $M_3$ effects) reflect the impact on the regional economy of a shock into household incomes. In the interregional SAMs, the households are divided into four income groups (25 per cent groups). Income group 1 receives the least income, income group 4 the most.

The exact amount of income per household group differs between the five countries because the division is based on the average level of income in a specific country.

Table 3 shows the average SAM income multipliers per country, and Table 4 the average value for the 4 different income groups per country. From the literature, we know that the values of income multipliers are generally lower compared with output multipliers. For England, France and the Netherlands, Table 3 shows values in line with values found in the literature (between 1.2 and 1.6). However, in Portugal and Poland, the values are higher, even 2.11 for the low incomes in the Polish towns. From the data, it appears that particularly the Polish and Portuguese (town) households buy a large amount of necessities in the local economy. Furthermore, more than two-thirds of these households have a job in the zone of residence, which means that they also profit from the induced effects. The reason why there is a higher income multiplier for the town households is that, in all countries, these households buy more products and services locally.

| | England | France | Netherlands | Poland | Portugal | Average |
|---|---|---|---|---|---|---|
| Town | 1.30 | 1.44 | 1.39 | 1.58 | 1.71 | 1.48 |
| Hinterland | 1.28 | 1.35 | 1.35 | 1.48 | 1.69 | 1.43 |

Table 3: Average household income multipliers in town and hinterland for 5 European countries

| | England | France | Netherlands | Poland | Portugal | Average |
|---|---|---|---|---|---|---|
| Town | | | | | | |
| Income group 1 | 1.40 | 1.63 | 1.57 | 2.11 | 1.78 | 1.70 |
| Income group 2 | 1.34 | 1.50 | 1.44 | 1.56 | 1.97 | 1.56 |
| Income group 3 | 1.28 | 1.30 | 1.30 | 1.23 | 1.70 | 1.36 |
| Income group 4 | 1.18 | 1.31 | 1.22 | 1.41 | 1.39 | 1.30 |
| Hinterland | | | | | | |
| Income group 1 | 1.40 | 1.37 | 1.59 | 1.83 | 1.83 | 1.60 |
| Income group 2 | 1.30 | 1.47 | 1.36 | 1.76 | 1.79 | 1.53 |
| Income group 3 | 1.26 | 1.31 | 1.27 | 1.24 | 1.77 | 1.37 |
| Income group 4 | 1.18 | 1.27 | 1.17 | 1.07 | 1.37 | 1.21 |

Table 4: SAM Household income multipliers in town and hinterland for 5 European countries

Interestingly, from Table 4 it appears that, in all countries, both in town and hinterland, the lower the income, the higher the multiplier effect. Households with high income more often have a job outside the local area (outside the town-hinterland area). Furthermore, it appears that richer households are less likely to shop in town or hinterland.

Summarising, it appears that especially in Poland and Portugal the household income multipliers are relatively high. This is mainly because of strong effects on the production output. Furthermore, we can conclude that the higher the level of income of households, the lower the SAM income multiplier. Finally we found that, in general, most of the impact of a shock to the income of households, living either in town or hinterland, goes to town production output. The only exception is the Netherlands where a shock to the income of hinterland households also results in a strong effect on the production output in the hinterland.

### SAM output multipliers

SAM output multipliers show the adjustment in the towns' and hinterlands' total output that would be associated with a change of one unit of output from a certain sector. When, for example, the final demand for manufacturing products increases in towns, this results in an effect in the production sectors in towns, as well as in the production sectors in the hinterlands. But these are not the only effects: there will also be an effect in labour factors, as well as in household incomes in town and hinterland. All these effects together, plus the 'closed loop' effect[3] sum up to the 'industry SAM output multiplier'.

## AGGREGATED OUTPUT MULTIPLIERS

For each town, the output multiplier of 17 sectors in the town and hinterland has been derived. Table 5 shows the average SAM output multiplier values of the aggregated agricultural, manufacturing and service related sectors per country (average of 6 towns). First of all, we can see that the hinterland multipliers have higher values than the town multipliers. In many areas, the total economic output in town is larger than in the hinterland. Furthermore, it appears that local inputs are more important for hinterland firms (higher indirect effects). The service multipliers have relatively high values; only in England, is the output multiplier for the manufacturing sector in town higher than the service multiplier. The explanation for this is that in England (and to a lesser extent in France), the share of exogenous accounts in the total output of the manufacturing sectors is lower than for the service sectors[4]. Especially in the Netherlands and Portugal, the multiplier for the service sector is relatively high (both in the town and the hinterland). The most important reason for this is the stronger effect on factor income and household income in the Netherlands; in Portugal, a stronger effect on the intermediary deliveries also plays a role.

[3] For example, the effect of hinterland households who receive more income because of a shock to the town and who spend this extra income in a shop in town.

[4] In the other three countries, the share of exogenous accounts (which includes payments to the ROW) in the service sectors is lower compared with those in the manufacturing sectors, resulting in higher service multipliers. However, in general, the share of exogenous accounts is very high in England and France (around 82 per cent) compared with the other three countries (70 per cent in the Netherlands and Poland and 65 per cent in Portugal).

| | England | France | Netherlands | Poland | Portugal | Average** |
|---|---|---|---|---|---|---|
| Town | | | | | | |
| Agriculture* | - | - | - | - | - | - |
| Manufacturing | 1.39 | 1.36 | 1.29 | 1.26 | 1.20 | 1.30 |
| Services | 1.32 | 1.41 | 1.56 | 1.45 | 1.51 | 1.45 |
| Hinterland | | | | | | |
| Agriculture*** | 1.25 | 1.28 | 1.52 | 1.94 | 1.65 | 1.53 |
| Manufacturing | 1.42 | 1.30 | 1.35 | 1.35 | 1.52 | 1.39 |
| Services | 1.44 | 1.44 | 1.57 | 1.50 | 1.66 | 1.52 |

Table 5: Aggregated SAM output multipliers for five European countries. * Agriculture is not part of the town economy. ** Average of the five country multipliers. *** Without forestry and fishing

In Poland and Portugal, the agriculture multipliers are relatively high. Especially in Poland this sector is still important; it produces 31 per cent of the total output of the Polish hinterland compared with around 12 per cent in the other four countries (see van Leeuwen, 2008). However, also in Portugal and the Netherlands, the agriculture multipliers are larger than the manufacturing multipliers. This can be explained by the relatively large share of local inputs.

The hinterland multipliers are generally higher and more heterogeneous compared with the town multipliers. This holds especially for Poland and Portugal. In Poland the effect on factor income is stronger in the hinterland. In Portugal, the main reason for higher multipliers in the hinterland is the stronger interregional effect on production activities located in the town.

### SECTORS WITH HIGH MULTIPLIER VALUES

After this general exploration of multiplier values in the five countries, we can focus more on sectors with relatively high multipliers, for the town and hinterland economy. Table 6 shows the disaggregated SAM output multipliers for 13 sectors in the towns (in a town the agricultural sectors are usually not present) and 17 sectors in the hinterland.

| | England | France | Netherlands | Poland | Portugal | Average |
|---|---|---|---|---|---|---|
| Town | | | | | | |
| Arable farming* | - | - | - | - | - | - |
| Dairy and intensive farming | - | - | - | - | - | - |
| Horticulture | - | - | - | - | - | - |
| Mixed farming | - | - | - | - | - | - |
| Forestry and fishing | 1.39 | 1.00 | 1.00 | 1.23 | 1.47 | 1.22 |
| Coal, oil and gas, metal ore, electricity | 1.48 | 1.37 | 1.00 | 1.10 | 1.01 | 1.19 |
| Food, drink and tobacco | 1.34 | 1.44 | 1.22 | 1.22 | 1.06 | 1.26 |
| Textiles, leather, wood, furniture | 1.48 | 1.36 | 1.36 | 1.24 | 1.21 | 1.33 |
| Chemicals, rubber, plastics, glass | 1.43 | 1.44 | 1.19 | 1.38 | 1.10 | 1.31 |
| Metals, machinery, electrical, computing, transport equipments | 1.46 | 1.31 | 1.33 | 1.25 | 1.25 | 1.32 |
| Construction | 1.13 | 1.27 | 1.64 | 1.38 | 1.57 | 1.40 |
| Transport Services | 1.36 | 1.66 | 1.55 | 1.55 | 1.31 | 1.48 |
| wholesale/retail | 1.10 | 1.34 | 1.58 | 1.51 | 1.27 | 1.36 |
| Hotels and catering | 1.19 | 1.62 | 1.89 | 1.55 | 1.78 | 1.61 |
| Banking and financial services | 2.12 | 1.61 | 1.43 | 1.19 | 1.24 | 1.52 |
| Other Business services | 1.02 | 1.07 | 1.31 | 1.28 | 1.56 | 1.25 |
| public administration, education, health, other services | 1.14 | 1.16 | 1.61 | 1.65 | 1.92 | 1.50 |
| Hinterland | | | | | | |
| Arable farming | 1.27 | 1.13 | 1.48 | 1.76 | 1.89 | 1.51 |
| Dairy and intensive farming | 1.22 | 1.17 | 1.73 | 1.81 | 1.74 | 1.53 |
| Horticulture | 1.11 | 1.64 | 1.45 | 1.83 | 1.89 | 1.58 |
| Mixed farming | 1.40 | 1.19 | 1.41 | 2.37 | 1.07 | 1.49 |
| Forestry and fishing | 1.08 | 1.00 | 1.18 | 1.36 | 1.62 | 1.25 |
| Coal, oil and gas, metal ore, electricity | 1.44 | 1.29 | 1.06 | 1.13 | 1.33 | 1.25 |
| Food, drink and tobacco | 1.43 | 1.42 | 1.44 | 1.62 | 1.36 | 1.45 |
| Textiles, leather, wood, furniture | 1.43 | 1.40 | 1.58 | 1.36 | 1.68 | 1.49 |
| Chemicals, rubber, plastics, glass | 1.47 | 1.16 | 1.36 | 1.17 | 1.45 | 1.32 |
| Metals, machinery, electrical, computing, transport equipments | 1.55 | 1.22 | 1.29 | 1.46 | 1.45 | 1.39 |
| Construction | 1.17 | 1.32 | 1.40 | 1.35 | 1.87 | 1.42 |
| Transport Services | 1.29 | 1.60 | 1.47 | 1.67 | 1.79 | 1.57 |
| wholesale/retail | 1.31 | 1.31 | 1.51 | 1.39 | 1.30 | 1.36 |
| Hotels and catering | 1.15 | 1.54 | 1.81 | 1.57 | 1.86 | 1.59 |
| Banking and financial services | 2.44 | 1.65 | 1.44 | 1.60 | 1.28 | 1.68 |
| Other Business services | 1.20 | 1.39 | 1.47 | 1.30 | 1.85 | 1.44 |
| public administration, education, health, other services | 1.27 | 1.14 | 1.72 | 1.49 | 1.88 | 1.50 |

Table 6: SAM output multipliers for sectors in town and hinterland in 5 European countries. * Agriculture is not part of the town economy

First we focus on multipliers of sectors located in town. From Table 6, it becomes clear that in the English towns the industry sectors have higher multiplier values than in the other countries; an exception is the construction sector. However, the banking and financial service sector has the highest multiplier, with a value of 2.12. This is due to strong linkages with 'other services' and public administration. Also an important part of the multiplier is related to wages paid to managers and (non) skilled non-manual employees.

The French towns show a rather similar picture; with relatively high industry multipliers (compared to the other countries). However, here the transport sector has the highest multiplier of the towns, together with hotels and catering and (again) the banking and financial services sector.

In the Netherlands, the situation is slightly different. Among the industry sectors, the construction sector seems to be especially important, with the highest multiplier for this sector of all five countries. This also holds for the hotels and catering service sector, with a high multiplier of 1.89. This last sector has a relatively strong impact on the income of local households. Other important service sectors in the Dutch towns are wholesale and retail, as well as public administration. In Poland and Portugal the key-sectors are rather similar to those of the Dutch towns: the construction sector has the highest multiplier value of the industry sectors and the hotels and catering and the public administration sector are important service sectors.

On average, in the towns of the five European countries, the construction sector is the key-industry sector and the hotels and catering the key-service sector; the sectors with the greatest output impact on the local economy from an exogenous shock. These sectors, together with the agricultural sectors and public administration, also have the strongest impact on local household income.

Secondly, we look at the hinterland sectors. We already saw that the agriculture sector in general can have high multiplier values, especially in Poland and Portugal and to a lesser extent in the Netherlands (see Table 2). According to Table 3, in Poland the agricultural sector with the highest multiplier is mixed farming with a value of 2.37. In Portugal, the horticulture and arable farming have the highest multipliers. This is because of strong effects on both factor income as well as on the household income accounts. Horticulture is also in France, by far, the most important agriculture sector, with a multiplier of 1.64. Also here, strong linkages exist with factor income as well as with deliveries of the arable farming sector. In the Netherlands, the agricultural sector with the highest SAM multiplier is the dairy and intensive farming sector. In these hinterland areas, the intermediate deliveries affect the multiplier, deliveries to the own sector, the energy sector, for machinery or public administration.

Furthermore, in the hinterland, not the construction sector but the 'textiles, leather, wood and furniture' sector has on average the highest multiplier, especially in the Netherlands and Portugal. Another important sector is the ‘food, drink and tobacco’ sector, with an especially high value in Poland. Also in France, this is an important industry sector.

The service sector with on average the highest SAM output multiplier is the banking and financial services sector; especially in England and France this is the key-service sector.  Another important sector in the hinterland as well, is the hotel and catering sector.

To summarise, we found that in general, in all five countries, both in town and hinterland, the sectors with the highest multiplier values are the service-related sectors. In England and France this is the banking and financial service sector, in Poland (and also in France) the transport sector, and in the Netherlands and Portugal this is the hotel and catering, as well as the public administration sector.

## INTER-REGIONAL EFFECTS

Apart from level of redistributive effects (size of the multiplier), multipliers can also show the interdependencies between town and hinterland. The inter-regional effects, obtained through the $M_2$ multiplier, show the linkages between town and hinterland and how strong they are; is the town more dependent on the hinterland, or the hinterland on the town?

Table 7 shows which part of the (redistributive) multiplier effect descends in the other region because of a shock in production activities in the town or hinterland, thus showing the level of interdependency. The inter-regional effects are calculated by dividing the effect of a shock in output in the ‘other’ region by the total effect from the shock minus the initial shock. When the total output multiplier is 1.58 and the effect in the ‘other’ region 0.19, the intra-regional effect is 0.19/0.58*100.

It appears that, on average, the inter-regional effect in the hinterland of a shock in the town is around 22 percent (both for industry as for services). The largest effects appear in the textiles and in the banking sectors in the Netherlands (both 36 per cent); the smallest effects appear in England and Poland. In these two countries the towns are less dependent on the hinterland.

When focusing on the effects in town from a shock in the hinterland, we see that the shares are almost always higher, only in the Netherlands, the hotel and the construction sector in the hinterland are less connected to the town than vice versa. For the hinterland, the linkage with the town is between 30 and 39 per cent. This implies that, in general, the hinterland is more connected to the towns and thus more dependent on intermediary deliveries as well as on labour form the towns.

In the agricultural sectors the linkages seem to be slightly weaker; however, in the Netherlands the horticulture sector depends for 56 percent on the town, both for intermediate deliveries and for labour. In the industrial sectors, on average the strongest inter-relationships exist; 39 percent of the multiplier effect. In France the strongest linkage exists in the construction sector. These construction firms in the hinterland are especially dependent on the intermediary deliveries of the towns. Finally, the interregional effects in the service sector are highest in Portugal, especially in the hotel sector. Also in Poland and England the interregional relationship is strong in this sector. In Poland and Portugal, almost all necessary goods and services are bought in town; apparently these are not available in the hinterland. In England also part of the factor income is paid to town households. The linkage between town and hinterland in the banking sector seems to be very strong in the Netherlands. Besides this dependency on intermediary deliveries, also labour is acquired from the towns.

| | England | France | Netherlands | Poland | Portugal | Average |
|---|---|---|---|---|---|---|
| Effect on hinterland from a shock in town | | | % | | | |
| Agriculture* | - | - | - | - | - | - |
| Dairy and intensive | - | - | - | - | - | - |
| Horticulture | - | - | - | - | - | - |
| Industry | 13 | 24 | 26 | 18 | 27 | 22 |
| Textiles | 15 | 20 | 36 | 21 | 28 | 24 |
| Construction | 12 | 17 | 28 | 18 | 25 | 20 |
| Services | 18 | 23 | 25 | 17 | 20 | 21 |
| Hotels | 17 | 16 | 24 | 9 | 23 | 18 |
| Banking | 16 | 27 | 36 | 18 | 20 | 24 |
| Effect on town from a shock in the hinterland | | | % | | | |
| Agriculture | 24 | 36 | 38 | 28 | 24 | 30 |
| Dairy and intensive | 29 | 40 | 32 | 25 | 29 | 31 |
| Horticulture | 16 | 36 | 56 | 41 | 25 | 35 |
| Industry | 18 | 29 | 54 | 47 | 49 | 39 |
| Textiles | 15 | 23 | 60 | 44 | 39 | 36 |
| Construction | 39 | 44 | 20 | 33 | 41 | 36 |
| Services | 30 | 42 | 29 | 37 | 39 | 35 |
| Hotels | 42 | 26 | 16 | 43 | 42 | 34 |
| Banking | 23 | 40 | 49 | 32 | 31 | 35 |

Table 7: Inter-regional effects of aggregated and key-sectors: effect in the hinterland from a shock in output in town and vice versa.* Agriculture is not part of the town economy

To summarise, it appears that in general the interrelationships between hinterland and town are stronger than vice versa. Although the differences between the sectors are very small, on average the strongest links are found in the industry sectors, both in town and hinterland. Furthermore, the strongest links are found in the Netherlands, because of intermediary deliveries, but mostly because of labour. In Poland and Portugal the hinterland is especially dependent on intermediary deliveries from the towns. In England we find the weakest links.

*SAM employment multipliers*

The employment multipliers indicate the additional employment generated in the regional employment due to an initial employment increase in a particular sector. The employment multipliers are derived from a combination of output multipliers and direct employment coefficients (employment per sector output, see Mayfield et al., 2005: 57).

Table 8 shows the aggregated multipliers for the agricultural, industrial and service related sectors in the five countries. The multiplier values are far more homogeneous compared to the output and household income multipliers: they range between 1.10 and 1.36. On average, the employment multipliers for the town and hinterland sectors seem to be equal in size. Furthermore it appears that, in both areas, the industrial sectors generate the largest effects in employment when a new job is added. In the agricultural sectors the effect is relatively small.

| | England | France | Netherlands | Poland | Portugal | Average** |
|---|---|---|---|---|---|---|
| Town | | | | | | |
| Agriculture* | - | - | - | - | - | - |
| Industry | 1.32 | 1.33 | 1.27 | 1.14 | 1.16 | 1.24 |
| Services | 1.31 | 1.17 | 1.12 | 1.11 | 1.21 | 1.19 |
| Hinterland | | | | | | |
| Agriculture*** | 1.11 | 1.07 | 1.18 | 1.11 | 1.10 | 1.11 |
| Industry | 1.36 | 1.26 | 1.18 | 1.12 | 1.23 | 1.23 |
| Services | 1.32 | 1.20 | 1.09 | 1.10 | 1.24 | 1.19 |

Table 8: Aggregated SAM employment multipliers for 5 European countries * Agriculture is not part of the town economy ** Average of the five country multipliers *** Without forestry and fishing

Although both the output and the income multipliers for the English and French towns are relatively small, the employment multipliers are rather large. Moreover, particularly the Polish employment multipliers are relatively small. This can be explained by the fact that both in England and France, the local number of jobs (per household) is rather small, thus an increase of 1 job has a stronger effect. On the other hand, in Poland, the number of jobs is rather large.

## IMPORTANT EMPLOYMENT SECTORS

Table 9 shows the disaggregated employment multipliers of the sectors with the largest employment multipliers, as well as the sectors that have high output multipliers. Interestingly, some of the sectors have both high output and employment multipliers. This holds for dairy and intensive farming and for banking and financial services. However, other sectors with high output multipliers, such as horticulture, construction and hotel and catering, do not have high employment multipliers, possibly because these sectors are already labour intensive. In all countries, the food, drink and tobacco sector, as well as the banking and financial services sector have relatively high employment multipliers.

| | England | France | Netherlands | Poland | Portugal | Average |
|---|---|---|---|---|---|---|
| Town | | | | | | |
| Dairy and intensive farming* | - | - | - | - | - | - |
| Mixed farming | - | - | - | - | - | - |
| Horticulture | - | - | - | - | - | - |
| Food, drink and tobacco | 1.41 | 1.34 | 1.29 | 1.18 | 1.09 | 1.26 |
| Metals, machinery, electrical, computing, transport equipments | 1.32 | 1.36 | 1.29 | 1.10 | 1.37 | 1.29 |
| Construction | 1.09 | 1.11 | 1.22 | 1.19 | 1.17 | 1.16 |
| Transport Services | 1.48 | 1.34 | 1.22 | 1.10 | 1.21 | 1.27 |
| Hotels and catering | 1.09 | 1.08 | 1.12 | 1.15 | 1.12 | 1.11 |
| Banking and financial services | 2.09 | 1.37 | 1.28 | 1.12 | 1.19 | 1.41 |
| Hinterland | | | | | | |
| Dairy and intensive farming | 1.03 | 1.10 | 1.28 | 1.12 | 1.15 | 1.14 |
| Mixed farming | 1.28 | 1.06 | 1.21 | 1.06 | 1.04 | 1.13 |
| Horticulture | 1.06 | 1.07 | 1.09 | 1.06 | 1.14 | 1.08 |
| Food, drink and tobacco | 1.59 | 1.54 | 1.43 | 1.14 | 1.25 | 1.39 |
| Metals, machinery, electrical, computing, transport equipments | 1.41 | 1.13 | 1.12 | 1.06 | 1.28 | 1.20 |
| Construction | 1.13 | 1.09 | 1.10 | 1.07 | 1.21 | 1.12 |
| Transport Services | 1.17 | 1.29 | 1.11 | 1.11 | 1.39 | 1.21 |
| Hotels and catering | 1.05 | 1.16 | 1.05 | 1.13 | 1.14 | 1.11 |
| Banking and financial services | 2.21 | 1.49 | 1.18 | 1.10 | 1.09 | 1.41 |

Table 9: Disaggregated SAM employment multipliers of the key-output sectors and the sectors with the highest employment multiplier. * Agriculture is not part of the town economy

## INTER-REGIONAL EFFECTS

The level of inter-regional effects of a shock in employment is quite comparable to the effects of a shock in production output; in town the interregional effects are around 20 percent, in the hinterland around 35 percent (see Table 10).

| | England | France | Netherlands | Poland | Portugal | Average |
|---|---|---|---|---|---|---|
| Effect on the hinterland from a shock in town | | | % | | | % |
| Agriculture* | - | - | - | - | - | - |
| Dairy and intensive farming | - | - | - | - | - | - |
| Mixed farming | - | - | - | - | - | - |
| Industry | 10 | 18 | 26 | 16 | 21 | 18 |
| Food, drink and tobacco | 19 | 30 | 26 | 18 | 32 | 23 |
| Metals, machinery… | 3 | 21 | 44 | 19 | 19 | 22 |
| Services | 13 | 15 | 31 | 12 | 14 | 18 |
| Transport Services | 13 | 10 | 22 | 11 | 10 | 14 |
| Banking and financial services | 11 | 18 | 49 | 7 | 14 | 21 |
| Effect on town from a shock in the hinterland | | | % | | | % |
| Agriculture | 12 | 20 | 34 | 41 | 43 | 26 |
| Dairy and intensive farming | 20 | 32 | 32 | 27 | 42 | 28 |
| Mixed farming | 29 | 27 | 33 | 39 | 22 | 32 |
| Industry | 9 | 27 | 63 | 57 | 59 | 39 |
| Food, drink and tobacco | 2 | 16 | 68 | 49 | 49 | 34 |
| Metals, machinery… | 4 | 39 | 49 | 70 | 57 | 40 |
| Services | 20 | 38 | 44 | 55 | 61 | 39 |
| Transport Services | 18 | 16 | 44 | 59 | 37 | 34 |
| Banking and financial services | 12 | 30 | 59 | 49 | 54 | 38 |

Table 10: Inter-regional effects of aggregated and key-sectors: effect in the hinterland from shock in employment in town and vice versa.* Agriculture is not part of the town economy

Although the English and French employment multipliers are relatively high compared to the other countries, still the linkages between town and hinterland are relatively small. In the Netherlands, these effects are much stronger: especially the town-hinterland linkages are strong compared to the other countries; with almost half of the impact of a new job in the banking and financial service sector in town affecting the hinterland. However, the strongest linkages appear between hinterland and town, both in the service and industry sectors in most countries.

## CONCLUSIONS

This chapter has focused on the different possible applications of SAMs. After a conceptual exposition, various results derived from 30 interregional (town and hinterland) SAMs in 5 European countries were presented. The aim was to find out in which countries strong linkages, and thus high multiplier values, appear, what are sectors with large redistributive effects on town and hinterland economies and to what extent are town and hinterland linked.

As well as analytical results, the SAM analysis also generates multipliers which can be used as a more predictive tool. Multipliers show the system-wide direct and indirect effect of the recirculation of spending within the region; recipients use some of their income for consumption spending, which then results in further income and employment, and so forth.

Obviously, households are also part of the macro-economy. In Poland and Portugal, the income multipliers are significantly higher than in the other three countries. This is because Polish and Portuguese (town) households buy a large amount of necessities in the local economy. Furthermore, more than two-thirds of these households have a job in the local area, which means that they also profit from the induced effects.

In all countries, we found a higher multiplier for town-households than for hinterland households. The explanation for this is that, in all countries, these town households buy more products and services locally. Furthermore, it appears that both in town and hinterland, the lower the income, the higher the multiplier.

We also found that, in general, the highest output (measuring the effect of extra demand in output) and income (measuring the effect of increasing income) multipliers are found in Poland and Portugal. In these countries, strong linkages exist between local production activities, as well as between households and local production. This is an indication that in less developed countries rural areas are still relatively isolated, leading to smaller leakages in rural economies. In England and France, the multipliers are relatively low, and in the Netherlands in-between. In all five countries, the service-related sectors generate the highest output multipliers. Only in the English towns (not in the hinterland) are the manufacturing multipliers higher, and in the Polish hinterland the agriculture multipliers.

Furthermore, the hinterland multipliers are in general higher than the town multipliers. An important reason for this is the stronger linkage between hinterland and town than vice versa: the hinterland firms obtain a relatively larger part of their inputs from the towns. This implies that investments (or subsidies) in hinterland activities, preferably in service-related activities, lead to relatively large local effects.

We also find that there are significant national differences. In England, and to a lesser extent in France, the linkage between town and hinterland is weaker, as well as the production-income linkage; these firms have more employees from outside the local area. In the Netherlands, the linkages between town and hinterland are much stronger but the towns are relatively less important.

However, both town and hinterland are especially important for the provision of labour.

We may conclude that SAMs are a powerful tool in (spatial-) economic research, as they are able to map out the complexities of intersectoral and interregional interactions in a manageable format, based on a strict economic methodology. Their wide-spread use illustrates that the use of SAMs greatly enhances an understanding of impacts of shocks or interactions in (multi) regional systems, and hence may be seen as important vehicles for a solid policy analysis.

## REFERENCES

**Acharya, S.,** 2007, Flow Structure in Nepal and the Benefit to the Poor. In: Economics Bulletin, 15, 17: 1-14

**Adelman, I.; Robinson, S.,** 1978, Income Distribution Policy in Developing Countries: A Case Study of Korea. Oxford University Press: Oxford

**Adelman, I.; Taylor, J.E.; Vogel, S.,** 1988, Life in a Mexican Village: A SAM Perspective. In: Journal of Development Studies, 25: 5-24

**Alarcon, J.; van Heemst, J.; Keuning, S.; de Ruiter, W.; Vos, R.,** 1991, The Social Accounting Framework for Development, concepts, construction and applications. Avebury: Aldershot

**Bergh, J.C.J.M.; Hofkes, M.W.,** 1999, Economic models of sustainable development. In: van den Bergh, J.C.J.M. (Ed.), Handbook of Environmental and Resource Economics, Edward Elgar: Cheltenham: 1108-1122

**Blane, L.D.,** 1991, An Inquiry into Kenyan Small Town Development Policy. In: Economic Geography, 6, 2: 147-153

**Cohen, S.I.,** 1989, Analysis of social accounting multipliers over time: the case of the Netherlands. In: Socio-Economic Planning Sciences, 23: 291-302

**Cohen, S.I.,** 1996, Urban Growth and Circular Flow in a SAM-framework: The Case of The Netherlands. In: Socio-Economic Planning Sciences, 30, 1: 1-14

**Cohen, S.I.,** 1999, European structural patterns applied to transition economies: assessment for Poland and Hungary. In: Socio-Economic Planning Sciences, 33, 3: 205-219

**Courtney, P.; Mayfield, L.; Tranter, R.; Jones, P.; Errington, A.,** 2007, Small towns as ‘sub-poles’ in English rural development: Investigating rural-urban linkages using sub-regional social accounting matrices. In: Geoforum, 38, 6: 1219-1232

**Isard, W.; Azis. I.J.; Drennan, M.P.; Miller, R.E.; Saltzman, S.; Thorbecke, E.,** 1998, Methods of Interregional and Regional Analysis. Ashgate: Aldershot

**Jensen, R.C.; Mandeville, T.D.; Karunaratne, N.D.,** 1979, Regional Economic Planning: Generation of Regional Input-Output Analysis. Croom Helm: London

**Khan, H.A.,** 1999, Sectoral Growth and Poverty Alleviation: A Multiplier Decomposition Technique Applied to South Africa. In: World Development, 27, 3: 521-530

**Leontief, W.,** 1951, The Structure of the American Economy. Oxford University Press: New York

**Lewis, B.D.,** 1991, An Inquiry into Kenyan Small Town Development Policy. In: Economic Geography, 67, 2: 147-153

**Mayfield, L.; Courtney, P.; Tranter, R.; Jones, P.,** 2005, The role of small and medium-sized towns in rural development - Final Report. Centre for Agricultural Strategy. Reading

**Miller, R.E.; Blair, P.D.,** 1985, Input-output analysis: Foundations and Extensions. Prentice Hall: Englewood Cliffs

**Morilla, C.R.; Diaz-Salazar, G.L.; Cardenete, M.A.,** 2007, Economic and environmental efficiency using a social accounting matrix. In: Ecological Economics, 60, 4: 774-786

**Nijkamp, P.,** 2007, Ceteris paribus, spatial complexity and spatial equilibrium. An interpretative perspective. In: Regional Science and Urban Economics, 37, 4: 509-516

**Parikh, A.; Thorbecke, E.,** 1996, Impact of Rural Industrialization on Village Life and Economy: A Social Accounting Matrix Approach. In: Economic Development and Cultural Change, 44, 2: 351-377

**Pleskovic, B.; Trevino, G.,** 1985, The use of a social accounting matrix framework for public sector analysis: The case study of Mexico, Monograph no. 17. International Centre for Public Enterprises in Developing Countries: Ljubljana

**Psaltopoulos, D.; Balamou, E.; Thomson, K.J.,** 2006, Rural-Urban Impacts of CAP Measures in Greece: An Inter-regional SAM approach. In: Journal of Agricultural Economics, 57, 3: 441-458

**Pyatt, G.; Round, J.I.,** 1977, Social accounting matrices for development planning. In: Review of Income and Wealth, Series 23, 4: 339-364

**Roberts, D.**, 1998, Rural-urban interdependencies: analysis using an inter-regional SAM model. In: European Review of Agricultural Economics, 25: 506-527

**Roberts, D.**, 2005, The role of households in sustaining rural economies: a structural path analysis. In: European Review of Agricultural Economics, 32: 393-420

**Roland-Holst, D.; Sancho, F.**, 1992, Relative income determination in the United States. In: The Review of Income and Wealth, 38: 311-327

**Round, J.I.**, 1988, Incorporating the International, Regional, and Spatial Dimension into a SAM: Some Methods And Applications. In: Marrigan, F.; McGregor, P.G. (Eds), Recent Advances in Regional Economy Modelling, London papers in regional science, 19: 24-45

**Sánchez-Chóliz, J.; Duarte, R.; Mainara, A.**, 2007, Environmental impact of household activity in Spain. In: Ecological Economics, 62, 2: 308-318

**Terluin, I.J.; van Leeuwen, M.; Pilkes, J.**, 2003, Economic linkages between town and hinterland: a comparative analysis of six small and medium- sized towns in the Netherlands. Agricultural Economics Research Institute LEI: The Hague

**Thorbecke, E.**, 1995, Intersectoral Linkages and Their Impact on Rural Poverty Alleviation: A Social Accounting Matrix Approach. United Nations Development Organisation (UNIDO): Vienna

**Van Leeuwen, E.S.; Nijkamp, P.; Rietveld, P.**, 2005, Regional Input-Output Analysis. In: Kempf-Leonard. K. (Ed.), Encyclopedia of Social Measurement, Elsevier: 317-323

**Van Leeuwen, E.S.**, 2008, Towns Today. Contemporary Functions of Small and Medium-sized Towns in the Rural Economy, dissertation. VU University Amsterdam, the Netherlands

**Vogel, S.J.**, 1994, Structural Changes in Agriculture: Production Linkages and Agricultural Demand-Led Industrialization. In: Oxford Economic Papers, New Series, 46, 1: 136-156

## AUTHORS INFORMATION

**Eveline S. van LEEUWEN**
Eleeuwen@feweb.vu.nl
Department of Spatial Economics
VU University Amsterdam, the Netherlands

**Peter NIJKAMP**
Pnijkamp@feweb.vu.nl
Department of Spatial Economics
VU University Amsterdam, the Netherlands

Appendix A.I: List of European towns included in the analysis together with information about the population in town, hinterland and the total area, as well as the number of jobs in the total area.

| Country | Town | Population (# inhabitants) | | | # Jobs |
|---|---|---|---|---|---|
| | Location | Town | Hinterland | Total | Total |
| England | Leominster | 7,316 | 6,147 | 13,463 | 4,964 |
| | Swanage | 8,571 | 3,667 | 12,238 | 3,741 |
| | Towcester | 6,771 | 13,949 | 20,720 | 6,507 |
| | Tiverton | 9,257 | 8,582 | 17,838 | 6,781 |
| | Burnham-on-Sea | 16,344 | 14,909 | 31,253 | 8,925 |
| | Saffron Walden | 10,331 | 22,564 | 32,895 | 23,168 |
| France | Brioude | 3,131 | 1,969 | 5,100 | 5,771 |
| | Prades | 3,632 | 1,352 | 4,984 | 3,170 |
| | Magny-en-Vexin | 2,296 | 1,994 | 4,290 | 3,573 |
| | Mayenne | 6,548 | 2,428 | 8,976 | 11,177 |
| | Douarnenez | 7,302 | 1,615 | 8,917 | 7,601 |
| | Ballancourt-sur-Essonne | 6,169 | 10,900 | 17,069 | 10,990 |
| The Netherlands | Dalfsen | 6,570 | 16,895 | 23,465 | 6,791 |
| | Bolsward | 9,378 | 18,555 | 27,933 | 9,184 |
| | Oudewater | 7,745 | 51,705 | 59,450 | 30,576 |
| | Schagen | 17,214 | 24,116 | 41,330 | 13,198 |
| | Nunspeet | 19,215 | 27,410 | 46,625 | 17,630 |
| | Gemert | 14,815 | 41,245 | 56,060 | 17,119 |
| Poland | Glogówek | 6,251 | 12,975 | 19,226 | 5,214 |
| | Duzniki | 5,471 | 1,846 | 7,317 | 4,728 |
| | Ożarów | 7,144 | 16,956 | 24,100 | 7,694 |
| | Jędrzejów | 16,667 | 9,076 | 25,743 | 16,354 |
| | Ultsroń | 14,585 | 6,632 | 21,217 | 10,554 |
| | Lask | 20,587 | 11,104 | 31,691 | 8,018 |
| Portugal | Mirandela | 11,186 | 14,633 | 25,819 | 9,148 |
| | Tavira | 12576 | 12,421 | 24,997 | 10,221 |
| | Lixa | 5490 | 52,105 | 57,595 | 27,790 |
| | Vila Real | 32,644 | 17,313 | 49,957 | 20,511 |
| | Silves | 18,836 | 14,994 | 33,830 | 14,945 |
| | Esposende | 10,401 | 22,924 | 33,325 | 15,531 |

Appendix A.II: Format of inter-regional Marketowns SAM (53 x 53) / part 1

| | **Town** | | | **Hinterland** | | | | |
|---|---|---|---|---|---|---|---|---|
| **Town** | Production | Production labour income | Households | Production | Production labour income | Households | Exogenous accounts | Total |
| Production | A1<br>Town inter-industry matrix | B1 | C1<br>Town household expenditures on town goods and services (g&s) | D1<br>Exports from town sector output to hinterland | E1 | F1<br>Hinterland household expenditures on town (g&s) | G1<br>Export from town sector output to ROW, ROW household consumption on town g&s | H1<br>Total output value of town production |
| Production labour income | A2<br>Wage payments by town sector output to town labour income | B2 | C2 | D2<br>Wage payments by hinterland sector output to town labour income | E2 | F2 | G2<br>Wage payments by ROW sector output to town labour income | H2<br>Total factor payments to town |
| Households | A3 | B3<br>Payments to town households from town sector output | C3 | D3 | E3<br>Payments to town households from hinterland sector output | F3 | G3<br>Government transfers to town households | H3<br>Total town households income |

Appendix A.II: Format of inter-regional Marketowns SAM (53 x 53) / part 2

| **Hinterland** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Production | A4<br>Export from hinterland sector output to town | B4 | C4<br>Town household expenditures on hinterland (g&s) | D4<br>Hinterland inter-industry matrix | E4 | F4<br>Hinterland household expenditures on hinterland (g&s) | G4<br>Export from hinterland sector output to ROW, ROW household consumption on hinterland g&s | H4<br>Total output value of hinterland production |
| Production labour income | A5<br>Wage payments by town sector output to hinterland labour income | B5 | C5 | D5<br>Wage payments by hinterland sector output to hinterland labour income | E5 | F5 | G5<br>Wage payments by ROW sector output to hinterland labour income | H5<br>Total factor payments to hinterland |
| Households | A6 | B6<br>Payments to hinterland households from town sector output | C6 | D6 | E6<br>Payments to hinterland households from hinterland sector output | F6 | G6<br>Government transfers to hinterland households | H6<br>Total hinterland household income |
| Exogenous accounts | A7<br>Indirect taxes, VAT, subsidies, imports from ROW of town sector output | B7<br>Payments to households in ROW from town sector output | C7<br>Savings/ direct taxes of town households | D7<br>Indirect taxes, subsidies, imports from ROW of hinterland sector output | E7<br>Payments to households in ROW from hinterland sector output | F7<br>Savings / taxes of hinterland households | G7 | H7 |
| Total | A8<br>Total input value of town sector output | B8<br>Total factor payments of town | C8<br>Total town household expenditure | D8<br>Total input value of hinterland sector output | E8<br>Total factor payments of hinterland | F8<br>Total hinterland household expenditure | G8 | H8 |

Source: Mayfield et al., 2005.

## Appendix A.III: List of accounts in inter-local SAMs of Marketowns

**Production account:**

1. Arable farming
2. Dairy farming
3. Arable farming Intensive farming
4. Horticulture-open ground
5. Horticulture-glass
6. Forestry and fishery
7. Mining of coal, oil and gas
8. Other mining (sand, clay, salt etc)
9. Chemical products
10. Food manufacturing
11. Textile, leather
12. Wood, furniture
13. Paper, offset printing
14. Rubber, plastic, glass
15. Metals, machines
16. Electric apparatus, computers, optical equipment
17. Transport equipment
18. Electricity, water
19. Construction
20. Wholesalers
21. Retailers
22. Hotels, restaurants and catering
23. Transport services
24. Bank, finance and insurance services
25. Real estate, other business services
26. Public administration, education, health, recreation, culture
27. Personal services

**Production factor account:**

1. Labour income management/professional
2. Labour income skilled/partly or unskilled non-manual
3. Labour income skilled manual
4. Labour income partly or unskilled manual

**Households account:**

1. 1st 25%-income group
2. 2nd 25%-income group
3. 3rd 25%-income group
4. 4th 25%-income group

**Exogenous account:**

1. Sum of *rest of world account* (imports/exports), *government account* (taxes/subsidies) and *capital account* (savings/investments).

# INTEGRATING MORPHOLOGY IN URBAN SIMULATION THROUGH RETICULAR AUTOMATA

## SCHELLING'S MODEL OF SEGREGATION AS USED WITH REMUS

**Diego MORENO[*], Dominique BADARIOTTI[**] and Arnaud BANOS[***]**

[*]Laboratoire Société Environnement Territoire, [**]Laboratoire Image Ville et Environnement, [***]Laboratoire Géographie-cités, France

The representation of space through graph formalism permits to integrate the anisotropy of urban space in cellular automaton models. In order to analyze the influence of a graph structure in simulation dynamics we study here a segregation model implemented in a graph-based cellular automaton. The Remus model calculates neighborhood graphs in a city and integrates them to the segregation model proposed by Thomas Schelling.

In the first part we present a state of the art of graph-based cellular automatons and of morpho-dynamic research in urban studies. We present also the Remus methodology as used to create neighborhood graphs based on network accessibility between buildings as a way to integrate complex urban forms in simulation. In the second part we analyze the structure of the neighborhood graphs calculated by Remus for the town of Pau. Finally, in the third part, we study the impact of this structure on the simulation of spatial segregation as described in Schelling's model.

This application demonstrates the possibilities offered by graph formalism. On the one hand it permits the analysis of urban forms through graph indicators. On the other it allows one to integrate the heterogeneity of urban space in simulation. This approach is obviously an interesting way to explore the relationship between form and dynamics in urban studies.

**KEYWORDS**

Urban morphology, Graph theory, Cellular automata, Schelling's segregation model

## INTRODUCTION

The aim of urban cellular automaton models is to explain the dynamics that control the behavior of urban systems. They allow one to study the way these dynamics shape towns and cities, the influence of function on form. But what about the influence of form on function?

The purpose of this chapter is to study the effect of urban morphology on the dynamics of a reticular automaton, the Remus model. Firstly, we present the state of the art of cellular automata, the possibilities offered by graph formalism and the thematic and methodological context that oriented the conception of the Remus model. Afterwards we analyze the urban structure resulting from Remus. Finally, Schelling's classical segregation model is used to test the influence of the urban structure on the automaton behavior.

## FROM CELLULAR TO GRAPH-BASED CELLULAR AUTOMATA

The Remus model concerns cellular automata, urban modeling and graph formalism to represent urban forms. We present here the background of urban morpho-dynamic modeling, cellular automata, graph theory and urban morphology applications and, finally, a definition of graph-based cellular automata.

### Urban form, mobility and morpho-dynamic modeling

Urban morpho-dynamics developments concern the old-age question of the form and function of a town or a city. Nowadays this relation between form and function has become topical again. In the conception of the Remus model, form and function are related through intra-urban mobility.

The relation between mobility and urban forms is highlighted in urban planning researches (Newmann & Kenworthy, 1989; Dupuy, 1995; Dubois-Taine & Chalas, 1997; Wiel, 2002). These studies explain how mobility has contributed to shaping the morphology of cities. The increase in mobility has contributed to urban sprawl and conditioned urban patterns. Some researchers have also inquired about the role that morphology might play on urban mobility (Vernez-Moudon et al., 1997; Gourdon, 2001; Genre-Grandpierre, 2001).

Urban planning has often modified city forms to promote mobility. This relation can be formalized through the accessibility concept.

Accessibility represents an interaction potential which can be used to characterize urban morphology regarding mobility. Geographers, but also economists, have been interested in studying accessibility and proximity within the city (Burmeister & Colletis-Wahl, 1997; Meunier, 1999). Several approaches have been developed to study this urban “hybrid space”, made of physical space separation based on geographical distances, and of virtual space separation based on an economic and social logic (Muhammad et al., 2008). The concept of accessibility has been used in different ways and its measure can be grouped into three types : infrastructure-based measures, activity-based measures and utility-based measures (Geurs & van Eck, 2001, 2003).

But the accessibility concept also permits to integrate urban morphology in modeling approaches. Inequalities in accessibility in urban space have a strong influence on the social behavior of urban systems. Social and mobility fields are closely connected because mobility is generated by the need of people to have social interaction within the city. In an attempt to relate urban forms to social logics, Bill Hillier and Suzan Hanson developed a formal concept called “space syntax” (Hillier & Hanson, 1984). In this approach urban space is conceived of as an emergence of forms controlled by global random processes and regulated by local town-planning rules. The method uses graph formalism to describe the interactions between the morphologic and the social elements of a city. In doing so, it helps integrate the reticular dimension of urban space, rarely considered in urban models.

In most morphological analyses urban space is considered as anisotropic, and proximities between elements are generally represented on a topological (contiguity, connectedness) or a metric (Euclidean) basis. But an understanding of network structures that shape cities is necessary to describe the relation between morphology and urban processes.

Mobility in a street network can be compared to a fluid flow in a constrained milieu. This idea is actually renewed in physics and engineering by the “constructal” approach of Adrian Bejan (Bejan, 1997; Bejan, 2007). In a “constructal” flow, the general frame of the flow network is designed by the constraints of the environment and the fluid dynamics. For example, river basin morphology is obtained by the combination between the physical characteristics of the flow and the obstacles from the relief. Bejan’s constructal theory has been applied to city plans to explain their morphology or to optimize them (Beja & Ledezma, 2007; Reis, 2008).

The confluence of these methods inspired the conception of the Remus model. The use of graph formalism to represent social interactions within the city is inspired from Space Syntax, but also from recent developments on social networks. The idea that there is a strong relationship between urban flows and urban form influenced the choice of network accessibility as a criterion to define the proximity of spatial units.

Fractal analysis has been introduced by Benoît Mandelbrot in his famous article on the length of the coast of Britain (Mandelbrot, 1967). The first applications of this new geometric concept were mainly developed to characterize natural forms (Mandelbrot, 1975, 1977, 1982), but they have also been used in various other fields like mathematics (Bélair, 1987), physics (Schroeder, 1991), computer sciences (Pickover, 1990) or esthetics (Peitgen & Richter, 1986). The computation and study of theoretical fractals and the measure of observed fractals were the two main approaches developed in these early years. Afterwards, the application of fractal computations and measures to urban form studies was developed by geographers and town-planners (Batty & Longley, 1986; Frankhauser, 1990, 1991): it constitutes a different approach in relating form and function. The computation of the fractal dimension of built-up elements through scaling behavior showed the auto-similar characteristic of urban forms on different scales. This fractality may be interpreted as the result of a maximization of the accessibility to local amenities accompanying urban growth (Frankhauser, 1994). These computations can be done for a whole city or for its different morphologic units, with its districts or quarters, all dating back to different periods and resulting from different processes (Badariotti, 2005).

Remus aims also to develop morphologic indicators to characterize urban morphology on different scales. This method permits to reveal the auto-similar characteristics of the urban network structure, just as the fractal dimension measures the auto-similarity of built-up elements of the city.

However, the main particularity of the method exposed here is the integration of the network accessibility criterion in urban simulation. The first urban simulation models (Berry & Lobley, 1964; Forrester, 1969) applied the system theory (Bertalanffy, 1950) to urban modeling. But dynamic systems modeling in the 70's and the more recent developments in complex systems theory were more focused on dynamics than in morphology, considering their lack of tools to analyze urban forms. Even if the influence of transportation networks was early introduced in urban cellular automata (White & Engelen, 1997; Engelen et al., 2002; Barredo et al., 2003), it was not used to define their spatial structure.

In the Remus model, network accessibility is included in the neighborhood definition of cellular automata. This methodology is explained further below.

### From cellular to graph-based cellular automata

A cellular automaton is a mathematical object embodying a lattice of contiguous cells. Cell state changes are controlled by a transition function, which depends on the states of the neighboring cells and on the state of the cell itself. Thus the dynamics of the system are the result of the dynamics of cell states.

The history of Cellular Automata is related to the history of computation. Turing's and Von Neumann's works on self-reproducing machines (Turing, 1950; Von Neumann & Burks, 1966) set down the basis of computing. The self-reproducing Von Neumann's machine was in fact a cellular automaton. In the Hedlund's synthesis of the works on symbolic dynamics (Hedlund, 1969), the mathematical formalism of cellular automata was described for the first time. Cellular automata models were subsequently applied to many disciplines, like linguistics, economics, transportation and urban planning. They are important tools for modeling complex spatial systems, because they permit to simulate the emergence of forms in space as a non-linear response to the aggregation of individual dynamics. This property helps to describe social phenomena based on the combination of individual behaviors.

In geography, the principle of reciprocal influence between a given position and its neighborhood existed in the first geographical models, like the diffusion models developed in the 1950s. The idea of a cell representation of space was employed in Chapin & Weiss (1968) and Lathrop & Hamburg (1965) models of land use change. But in these models, the neighborhood effect is just one of many factors affecting cells.

Tobler (1979) proposed to use cellular automata to study geographical processes. The idea of a Cellular Geography, based on the influence of each spatial unit's land use upon its immediate neighbors, was further expanded by Codd (1968), Albin (1975), Nakajima (1977) and Couclelis (1985, 1988, 1989). Their contributions paved the way for the development of urban cellular automaton models in the 1990s: White & Engelen (1993), Batty & Xie (1994, 1997), Xie (1996), Clarke et al. (1997), Phipps & Langlois (1997), Papini et al. (1998) are only a few of those who applied cellular automaton models to the study of urban change.

White & Engelen (1993) introduced the effect of road-network and infrastructural elements as constraints in their model. This notion of constrained cellular automata was employed in subsequent models (Engelen et al., 2002; Barredo et al., 2003) to integrate the role of accessibility, but also the role of suitability and planning in the transition functions.

Although these models tried to take into account the impact that infrastructure elements have on urban evolution, the effect of networks and transportation modes is difficult to define with cellular structures. Even if the integration of spatial constraints has introduced some space heterogeneity in urban cellular automaton models, their main spatial structure is based on assumptions about isotropy and stationarity of space. Those assumptions may be relaxed to fit more complex relationships in the city. All the same, O'Sullivan (2001a) showed that re-examining the formulation of cellular automata was necessary because of the significant sensitivity that spatial processes have to even small changes in underlying spatial structures.

Couclelis (1997) opened the discussion of the relation between proximal space, Cellular Automata, and a formalization of those models through Geo-algebra (Couclelis, 1997; Takeyama, 1996; Takeyama & Couclelis, 1997). This discussion constituted an attempt to solve the problem of the relation between structure and processes in geographical models, and resulted in the formalization of a geocomputational theory (Couclelis, 1998). Then the proposition emerged to integrate complex neighboring relationships between cells in cellular automata based on graph formalism. O'Sullivan's thesis (2000) proposed a graph-based cellular automaton and explored the effect of non-stationary neighborhoods on the behavior of the system.

### Graph theory and urban applications

At this point it is necessary to consider briefly graph theory and its developments. A graph is defined as a finite set of vertices (or nodes) and a finite set of edges (or links) that connect pairs of vertices.

Several graph indicators exist, but the indicators that are usually used to describe graph structures are the main degree, the path length and the clustering coefficient (Watts & Strogatz, 1998; Jiang, 2007).

- The degree of a vertex is the number of edges that are incident to it. The degree of a vertex $v_i$ is denoted *deg($v_i$).* It can be used to measure the centrality of a vertex in a graph. The average degree of a graph having *n* vertices, $D(G) = \Sigma_i\ deg(v_i) / n$, can be used to measure the density of links in a graph for a fixed number of vertices.

- A path is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence. The path length between a pair of vertices is the smallest number of edges necessary to go from a vertex to another. Thus the average path length indicates if the graph structure connects well distant nodes between them or not.

- The clustering coefficient of a vertex is the ratio between the number of edges incident to it (its degree) and the number of possible edges in a given neighborhood. This measure reflects the probability for two neighbors to be connected. The average clustering coefficient can thus be used to describe the connection between closer nodes.

These measures are often used to characterize different types of networks. Six types of graphs have a particular interest for our analysis.

- Tree graphs, defined as graphs in which any pair of nodes is connected by exactly one path, without cycles. This type of graphs is used to describe hierarchical structures.

- Random graphs, representing ER (Erdös-Rényi) networks, defined by Erdös & Rényi (1959). These graphs are constructed by starting with a fixed number of vertices and connecting them through a random process. The existence of an edge is controlled by a probability *p*.

- Regular networks, in which each node possesses the same number of incident edges.

- Scale-free networks, or BA (Barabási-Albert) networks, defined by Barabási & Albert (1999). These graphs have a large degree distribution that follows a decreasing power law. These graphs can be constructed from an existing graph by adding new vertices with a preferential connection to the existent well-connected vertices.

- Small-world networks, or WS (Watts-Strogatz) networks, first described by Watts & Strogatz (1998). These graphs are highly clustered, like regular graphs, yet have small path lengths, like random graphs. In these graphs each node is well connected locally to its neighborhood and there are some shortcuts that connect distant nodes, like social networks.

- Urban networks, or NG (Newman-Gastner) networks, characterized by Newman (2003a, 2003b) and Gastner (Gastner & Newman, 2004). These networks have been defined by the geometrical analysis of the properties of actual urban networks. The number of nodes is not known, the structure is more or less cellular, as in a lattice, and the maximum degree of any vertex is quite small, no more than 6 in general.

Jiang (2007) demonstrated that urban dual street networks, perceived as graphs whose edges represent street segments and vertices represent segment intersections, usually have scale-free or small-world properties.

**Graph-based Cellular Automata: a new space formalization**

Spatial relationships in cellular automata can be addressed advantageously via an approach based on mathematical graphs. Graph formalism can be used to describe the complexity of proximity relationships, and can be integrated in cellular automaton models for urban simulation purposes (O'Sullivan, 2000, 2001a, 2001b). The theoretical framework of graphs offers a set of methods and tools, beyond the accepted formulation, that facilitates the exploration of the relationships between urban forms and their functions.

From a formal point of view (Takeyama & Couclelis, 1997) a geographic cellular automaton is defined by the quadruplet (*U, N, S, F*), with:

- $U = \{U_1, U_2, \ldots, U_i, \ldots, U_n\}$, a finite set of spatial units (one notices that this finite collection of spatial units may be situated in an unbounded space, a torus, with the West units being related to those of the East, and the North units to those of the South), most often laid out in a regular grid;

- $N = \{N_1, N_2, \ldots, N_i, \ldots, N_n\}$, the set of the spatial units' neighbors, as defined by a topologic criterion (contiguity) or a geometric criterion (distance between centroids). Each neighborhood $N_i$ is a list of all the spatial units deemed to be neighbors of a unit *i* on the basis of a chosen criterion: $N_i = \{U_j, U_k, \ldots, U_m\}$;

- *S*, the cells' possible states, as defined by qualitative or quantitative variables, either discrete or continuous;
- *F*, local transitional functions, probabilistic or deterministic, by which a cell may evolve stepwise over time, depending both on cell state and neighbors states: $S_i^{t+1} = f(S_i^t; S_{Ni}^t)$

With this architecture, a structural component *{U,N}* and a dynamic component *{S,F}* can be usefully distinguished in a typical cellular automata:

$$CA = ( \{ U , N \} , \{ S , F \} ) \qquad (1)$$

The spatial units and their neighbors thus constitute the skeleton upon which the dynamics will be able to play themselves out at the heart of the automata. The first couple of arguments are of particular interest in geography, since they deal with the way in which the underlying spatial structures are defined. Despite that fundamental specificity, geographic automata have nevertheless been most often developed in reference to the formal definition according to which "cellular automata are automata distributed on vertices of a periodic network, i.e. a discrete geometric structure that is preserved by certain operations of translation and rotation" (Weisbuch, 1989). The majority of those referring to it (Nakajima, 1977; Xie, 1996; Batty & Xie, 1994; White & Engelen, 1997; Langlois & Phipps, 1997; Phipps & Langlois, 1997; Engelen et al., 2002; Barredo et al., 2003) adopt that perspective, the limitations of which we shall demonstrate.

Table 1 shows standard representations of the structural component *{U,N}* of a geographic cellular automaton. Space is divided into square cells, of uniform size, arranged evenly over the domain. The neighborhood of each cell is thus stated, for its topologic version, according to two definitions of adjacency:

- Von Neumann's adjacency, whereby two cells are neighbors if they share at least one common boundary;
- Moore's adjacency, in which two cells are neighbors if they share at least one common vertex.

That representation, as intuitive as it is, still has a limited scope. A slightly different interpretation of the structural couple *{U,N}* helpfully overcomes this limitation. In fact, like the display in Table 1, it is possible to view the spatial units $U_i$ as vertices of a mathematical graph, such that the neighboring relationships between vertices can be depicted as the edges of that graph, designated *G(U, N)*.

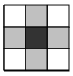
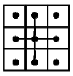
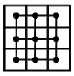
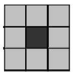
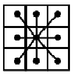
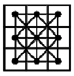
|  | *Classic schema* | *Neighboring Graph for one cell* | *Entire Neighboring Graph* |
|---|---|---|---|
| ***Von Neumann's Neighborhood*** |  |  |  |
| ***Moore's Neighborhood*** |  |  |  |

Table 1: A representation of a cellular automata and related neighboring graphs (Badariotti et al., 2007)

Such a portrayal is at the root of the neighboring graph, comprising all the neighboring relationships (Table 1). From this perspective, the couple *{U,N}* may be rendered as a graph *G=(U,N)*, altering in one stroke the definition of a geographic cellular automaton of equation (1):

$$CA = ( G(U,N) , \{ S , F \} ) \qquad (2)$$

Moreover, one can note that the neighboring graph thus obtained is not necessarily planar, since two edges do not necessarily intersect at a node. Such a remark may seem trivial. However, its implications are not always grasped by geographers, as was rightfully emphasized by O'Sullivan (O'Sullivan, 2000, 2001a, 2001b), for the graph representation immediately highlights the fundamental hypothesis of stationarity on which the majority of geographic cellular automata rest, namely that:

- Spatial units are arranged uniformly across the domain;
- Neighbor relationships are strictly identical for all spatial units (provided units located at the limits of a finite domain are disregarded);
- Transition functions are themselves most often stationary, applying equally to all units of the structure.

In order to take into account the spatial relationship between units, O'Sullivan proposed irregular graph-based cellular automata (O'Sullivan, 2000, 2001a, 2001b). These models go beyond stationarity by focusing on the structural component of cellular automata, which lacked adequate realism as regards urban environments.

Graph-based cellular automata allow one to give a better explanation of the impact of morphology on the behavior and evolution of urban systems.

At this point, we have introduced the common background knowledge that it is useful to position our model, as well as its results, in the thematic urban field, and the current state of art in graph-based automata.

About cellular automata, we have in particular seen that it is an important tool for modeling complex spatial systems, because it permits to simulate the emergence of forms in space like a non-linear response to the aggregation of individual dynamics. This property is useful to describe social phenomena based on the effect of combination of individual behaviors. In fact, the dynamics of each cell state depends on its neighborhood and the definition of the neighborhood is related to the topological relationships between cells. Therefore, the dynamics of the whole system is related to the pattern structure of the lattice and this structure is the result of cell formalism used in cellular automata.

## INTEGRATING MORPHOLOGY IN A RETICULAR AUTOMATON: THE REMUS MODEL

In spite of important contributions that the community of geographers have made to mathematical and computer formalizations of cellular automata for the purpose of better accounting for the specificities of geographic spaces and processes (Ménard et al., 2004), an essential characteristic is still lacking. Usually, neighborhood relationships between cells in geographical automata are most often determined on the one hand by topological (contiguity) or geometric (Euclidean distance) criteria, and on the other hand, on a stationary basis: two options that depict poorly, if at all, the spatial differentiations that arise from the anisotropy of geographical space.

### The Remus methodology

It is necessary to define a method to calculate neighboring graphs to represent anisotropic spaces in graph-based automata. The method used in the Remus model consists in calculating network accessibility in urban environments to build these neighboring graphs.

Inspired by the work of Chua & Yang (1988), Schonfish (1997) and O'Sullivan (O'Sullivan, 2000, 2001a, 2001a), Badariotti et al. (2006) and Banos (2009) developed a methodology to construct graph-based automata integrating intra-urban accessibility. These automata treat the built elements as the basic urban units, and replace strict contiguity with a neighborhood distance. Two spatial units are neighbors if the Euclidean distance between their centroids is less than a specified threshold distance: $d(U_i(x_i,y_i); U_j(x_j,y_j)) \leq d_{threshold}$

This being done, the layout of the spatial units becomes irregular, with spatial discontinuities between buildings imposing themselves as significant structural elements. The graph of the underlying relationships is thus strongly influenced by this non-stationarity, in that minor changes in the neighboring distance $d_{threshold}$ may have broad overall repercussions, due to the sudden connection of partial sub-graphs that are not related at lower values of $d_{threshold}$ (Figure 1).

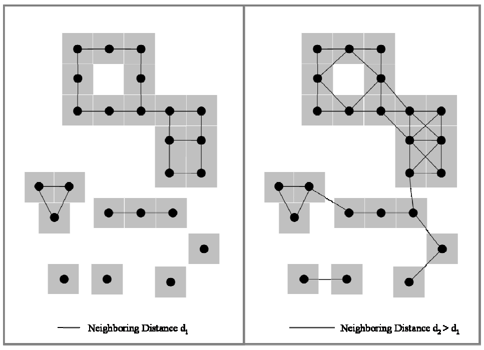


Figure 1: The introduction of heterogeneous spatial structures and its effect on the neighboring graph at various thresholds of Euclidean distance (Badariotti et al., 2007)

To work like this on the basis of Euclidean distances means, however, that one must accept, in addition to the technical advantages that it accrues, the hypothesis that space is fundamentally isotropic: for a given $d_{threshold}$, the neighboring graph depends solely on the spatial layout of units $U_i$. Yet it is evident that the presence of transportation networks with differentiated flow capacities greatly affects this hypothesis. Thus, thinking in terms of network accessibility for a given mode of transportation means introducing into the very structure of the cellular automata the fundamental anisotropy of geographic space.

The many possible relationships between spatial entities, and the spatial and temporal variability of these relationships stemming from the nature and evolution of urban movements and modes of transportation, compel to better specify spatial relationships. In particular it is possible to define for each cell a specific neighborhood capable of evolving over time and which permits to take into account all the complexity of spatial relationships in the city. This dynamic of relationships enables the cells to communicate well beyond their immediate adjacencies for the purpose of fulfilling given functions, much the same as neurons in the brain do. Thus these neighboring relationships are fundamentally functional, not simply geometric ones.

The size and shape of buildings influence the way the geometry of urban pockets forms and how the roadway pattern develops, especially in adjacent built-up areas. When we look at a cadastral view of developed urban areas, we can discern the circulation channels flowing from their buildings. Such traffic channels are as important to the functioning of the city as the buildings themselves are: it is thereby helpful to study the shape of the built-up areas in conjunction with the shape of the network of urban roads. For this reason the model start from a representation of the city that takes into account its morphology from the standpoints both of its buildings (a buildings database) and of its road network (a public thoroughfares database).

For this approach, the construction of the relationship graph involves:

- First, constructing an *urban graph G(V,E)*, the vertices of which include the buildings and the nodes of the road network, and the edges of which comprise the thoroughfares and road-to-building connectors (Figure 2a). The urban graph *G(V,E)* is one that can be termed “Euclidean” or “physical”, in that it corresponds to a “geographic” graph that can be displayed as a diagram or map, and is based on Euclidean spatial distances. This graph is not planar because edges crossing at different levels (e.g. shafts, tunnels, bridges) may or may not entail an intersection.

- Executing shortest path algorithms on the urban graph in order to calculate minimal travel time, by a given mode of transportation, between each pair of vertices $\{V_i; V_j\}$ of the graph *G(V,E)*. This process results in the creation of a *functional graph*, complete and non-planar $G'(U \,/\, U \subset V\,,\, K)$, comprised of vertices, representing the buildings, and edges representing the shortest travel times over the network between all possible pairs of buildings, for a specified mode of transportation (Figure 2b). The graph *G'(U,K)* may be called a functional graph, as it displays the functional linkages that exist on the basis of time-distance between buildings. Since reciprocal (but non symmetric) time-distances inherently exist for any pair of points, this depiction corresponds mathematically to a complete graph, because any two such points are always adjacent (Kaufmann, 1968).

- Constructing a neighboring graph $G_t''(U\ ,\ N_t\ /\ N_t \subset K)$, a sub-graph of the functional graph, in which the vertices represent the buildings and the edges represent the neighboring relationships falling within a given threshold of travel time $t_{threshold}$, such that $t(U_i; U_j) < t_{threshold}$ (Figure 2c).

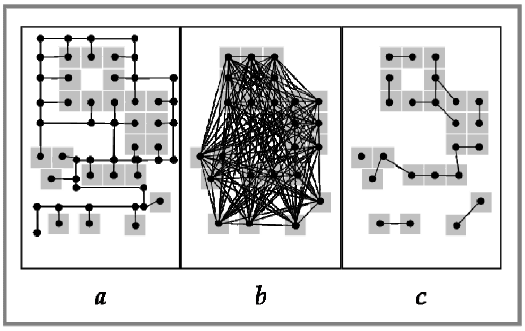


Figure 2: (a) Urban Graph *G(V,E)*; (b) Functional Graph *G'(U,K)*; and (c) Neighboring Graph via the network *Gt"(U,N)* (Badariotti et al., 2007)

The very structure of the neighboring graph that is thus generated is significant, since it does not depend solely on the arrangement of the spatial units, but equally on their spatial relationships via road access. Each neighboring graph thus reflects a specific mode of transportation and a specific threshold time, and is the product of a series of abstracting procedures whose aim is to create a display of the structure of intra-urban accessibility (Figure 3). The neighboring graph thus allows irregular neighborhood relationships to be defined in a dynamic way, reflecting urban dynamics.

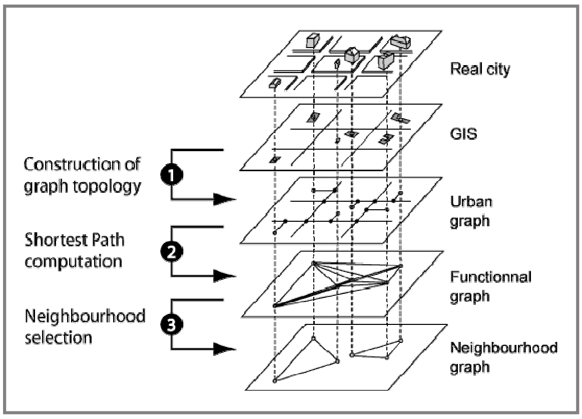


Figure 3: Stages in the construction of neighboring graphs (Badariotti et al., 2007)

**Application to the development of the Remus model**

The Remus model, designed in collaboration with researchers in computer science at LIUPPA (Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour), allows various graphs to be generated (urban, functional and neighboring graph) so as to establish the foundations for an anisotropic and non-stationary cellular automata that is ultimately capable of taking a more realistic account of spatial relationships when simulating urban dynamics.

Generating the urban graph entails importing geometric data from a Geographic Information System (GIS). Such data, generally available from cadastral and transportation network databases, do not necessarily include the topology needed to generate a graph.

Because of this fact, it is necessary to modify the geometry of the road network beforehand for it to adhere, to adhere to the topology of the urban graph. To do this, Remus creates intersections at the junctions of the road segments, grouping the various arcs that comprise a segment of road between intersections and generating a road graph that displays the intersections as well as the vertices, and the road segments as well as the component edges.

Next, Remus incorporates into the graph the built-up data, generating the buildings' access to the road network. To do this, the software:

- determines the road segments closest to each building's centroid (the search-distance around each building can be calibrated);
- eliminates road segments that are inaccessible due to obstacles (e.g., other buildings);
- elects from the remaining segments the one segment with the shortest geometric distance from the building's centroid; and
- generates the access routes between the buildings and the road segments.

Together these two operations generate the urban graph, representing the entire collection of buildings and their best inter-connections via the road network.

During a second stage, the urban graph makes it possible to create the functional graph, reflecting varied threshold distances. This stage involves a preliminary assignment of values to the urban graph, the transit times for each segment being derived from its geometric length and average speed. The average speed may be estimated indirectly from various fields in the roadway database, such as the type of road, or its width, or authorized speed limit.

For this stage, a path-search strategy was applied with the help of the Floyd-Warshall algorithm (Floyd, 1962). This algorithm calculates the shortest possible path between any pair of vertices. However, in light of the complexity of the algorithm and in order to reduce its computation time, a simplification was adopted to help prune the number of possible paths. Thus the urban graph has two levels: the sub-graph of building-to-road connectors *G(B,C)* (Figure 4a) and the sub-graph of roads *G(I,T)* (Figure 4b). The Floyd algorithm is initially applied solely to the road sub-graph; then, secondarily, the best paths are extended by joining the building-to-road connector of every building to a roadway intersection.

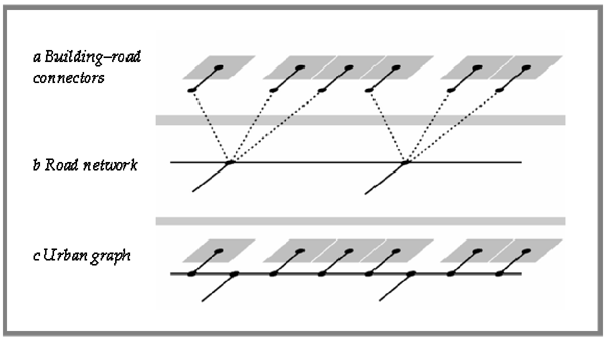


Figure 4: Diagram of the structure of an urban graph: a) Sub-graph of connectors linking the buildings to roads G(B,C); b) Sub-graph of the road network G(I,T); c) Urban graph G(V,E) composed of elements of the sub-graphs (Badariotti et al., 2007)

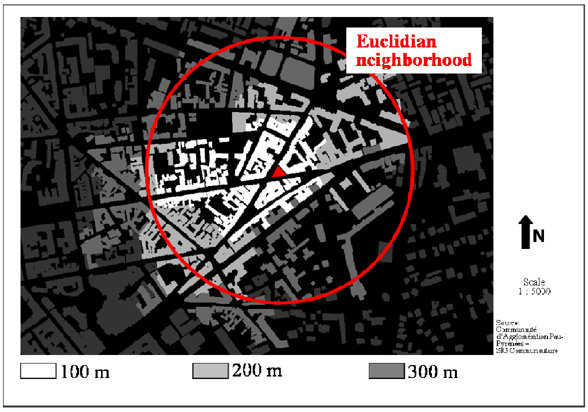
This innovative two-tiered strategy for exploring routes allows one to process very large databases, such as cadastral databases, by reducing the number of vertices traversed and thus drastically decreasing the number of iterations of the algorithm (Table 2).

| Graph | Number of Vertices | Number of Edges | Temporal Complexity (in millions of iterations) |
|---|---|---|---|
| Theoretical Urban Graph *G(V,E)* | 86,336 | 98,362 | 643 540 333 |
| Urban Road Graph *G(I,T)* | 7,590 | 19,616 | 437 245 |
| Functional Graph *G'(U,K)* | 39,373 | 775,096,878 | 61 037 329 |

Table 2: Comparison of the complexity of various graphs used for greater metropolitan Pau (Department of Pyrénées Atlantiques, France)

At the end of the shortest-path operation, a matrix is generated of the en-route times between every pair of buildings. This matrix corresponds to the functional graph *G'(U,K)* in which the vertices *U* depict the building units, and the value assigned to each edge *K* is the distance-time of the shortest path between buildings via the urban road network.

In order to study the structure of the functional graph thus defined, Remus allows for visual exploration of each building's neighborhood via the road network, regrouping the buildings hierarchically in terms of their accessibility at variable distance-times (Map 1).



Source:
Communauté
d'Agglomération Pau-
Pyrénées –
SIG Communautaire

Map 1: Using the REMUS program for visual exploration: neighborhoods within 100, 200, and 300 meters of a selected building in greater metropolitan Pau-Pyrénées

Hierarchically ordering the paths' distance-times makes it possible to select each building's neighbors for a given distance-time threshold. Finally the neighboring graph is generated for a selected time-distance threshold (Figure 5). A localized visualization is thus obtained, showing the proximities of buildings, not as the crow flies, but via the road network.

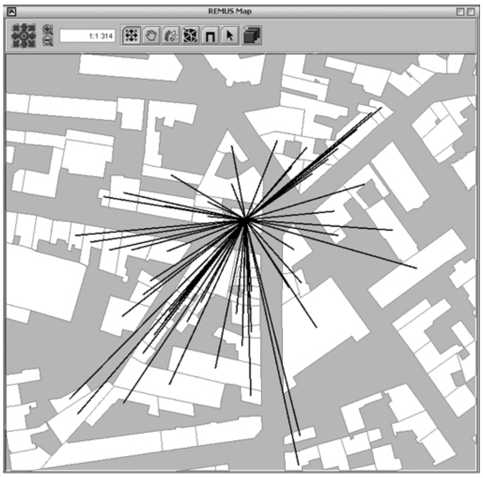


Figure 5: Example of a neighboring graph generated by the Remus program, showing neighbors within 100 meters of a selected building

## APPLICATION OF THE REMUS MODEL TO THE CITY OF PAU

It is now time to propose an application of the Remus formalization in order to study neighborhood graph structures.

In a first global approach, the Remus model was applied to the Communauté d'Agglomération de Pau-Pyrénées to test the concrete applicability of the model on a real space situation and to explore its possibilities to calculate morphological indexes. Pau-Pyrénées has been chosen because of the variety of patterns existing in it. This city is situated in the French south-west and it had the typical evolution of European cities. But the main urban growth occurred in the second half of the twentieth century, in a lapse of time which is very short in the French context. The result is a small town with heterogeneous patterns, which is adequate for our analysis.

The analysis of the general neighborhood graph structure has been made for the whole space of the town to characterize the distribution of the distances between buildings. Other indexes, whose interest for the characterization of different morphologies has been demonstrated, have been tested on specific quarters with a specific morphology: they include a neighborhood average degree analysis on each quarter, an analysis of the scaling behavior of their distance structure and a specific degree distribution analysis.

### Distribution of the distances between buildings in the entire network space

In order to analyze the generated neighborhood graphs, the average degree *D(G)* was calculated for different threshold distances, as:

$$D(G) = \Sigma_i \, deg(v_i) \, / \, n$$

At low threshold distances, just a few buildings are connected to a neighbor and the *D(G)* value tends to zero. As the threshold distance increases, more buildings connect to their neighbors and the average degree increases too. The results were compared with neighborhood graphs based on Euclidean metrics, for random and real distribution of spatial units (buildings). Chart 1 shows the differences in integration of the new neighbors when the threshold distance increases. The irregular growth of the two real-based curves, compared to the random curve, shows clearly that the urban structure is anisotropic.

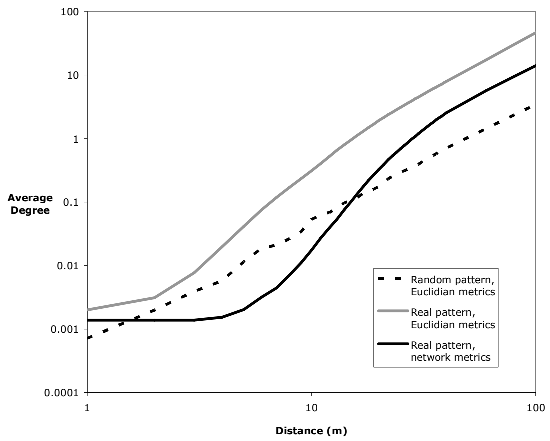


Chart 1: Comparison of neighborhood graphs average degree: a) real spatial distribution of buildings and Euclidean-defined neighborhoods; b) real spatial distribution of buildings and network-defined neighborhoods; c) random distribution and Euclidean-defined neighborhoods

More precisely, we can notice that the growth of the real-based curves is not always regular: at some local threshold points the curves change their direction. So it is obvious that the actual distribution of buildings and the introduction of network accessibility in urban metrics induce an important anisotropy in urban space at local scales.

This anisotropy is due to the urban pattern structure, which permits neighbor integration at three different rates:

- Below a limit of 5 meters, neighbor integration is insignificant because of the distance between building entries which makes neighbor links rare. This limit is determined by the size of spatial units and constrains neighborhood structure at low distances.

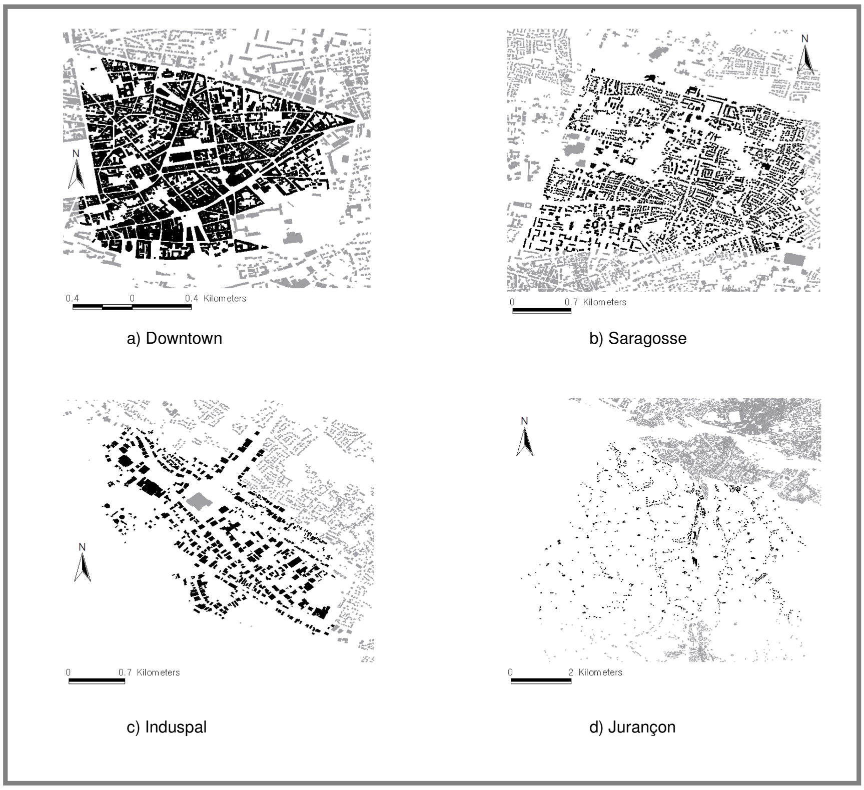
- Between 5 and 20 meters, neighbors are integrated faster when distance increases. The urban pattern permits sudden connections around crossroads that increase the number of neighbors easily.

- Around 20 meters, neighbors integration slows down with distance increase. This threshold corresponds to other types of patterns in which the distance between buildings is greater.

Chart 1 suggests that the urban spatial distribution permits buildings to be distant at a local level and to stay connected at a global level. The observed thresholds can be compared to transition phases and critical thresholds in network percolation (Erdös & Rényi, 1960; Aizenman & Barsky, 1987; Janson et al., 2000). The analysis of percolation threshold in networks is useful to study epidemic dynamics (Barthélemy et al., 2004; Eubank et al., 2004; Keeling, 2005; Serrano & Boguñá, 2006; Meyers, 2007). Urban patterns could be characterized by the way in which they integrate neighbors, in the same way that phase transitions characterize thermodynamic systems.

**Neighborhood graph average degree analysis**

In order to confirm these statements, different urban patterns were analyzed in the Communauté d'Agglomération de Pau-Pyrénées (Figure 6): a) the Downtown pattern corresponding to the medieval city, characterized by contiguous buildings and tight streets; b) the Saragosse neighborhood, the product of a functional urbanism; c) Induspal, an industrial and commercial facilities area; and d) the Jurançon suburbs in the southern sloping area.



a) Downtown

b) Saragosse

c) Induspal

d) Jurançon

Figure 6: The four neighborhoods maps

The average degree of neighborhood graphs was calculated for different distances and for these different patterns, and the results are shown in Chart 2. The neighbor integration is very different from one pattern to another. Downtown the number of neighbors is greater than in the other zones, because of the density heritage from periods when mobility was technologically restricted. The Saragosse area integrates neighbors in an intermediate way, because of its loose pattern, which is the result of a vertical urbanism that liberated space for public zones. The Induspal and Jurançon areas have a similar behavior in terms of neighbor integration, but their patterns are dissimilar. In Induspal the spatial distribution of buildings is homogeneous and greater distances between buildings are explained by space needed for industrial activities, but at Jurançon the spatial distribution is concentrated and a large distance between residential houses is required for the inhabitants to benefit from the natural environment offered by the sunny hills in the south. In these two final cases, the distance between buildings explains neighbor integration, even if urban patterns significantly differ.

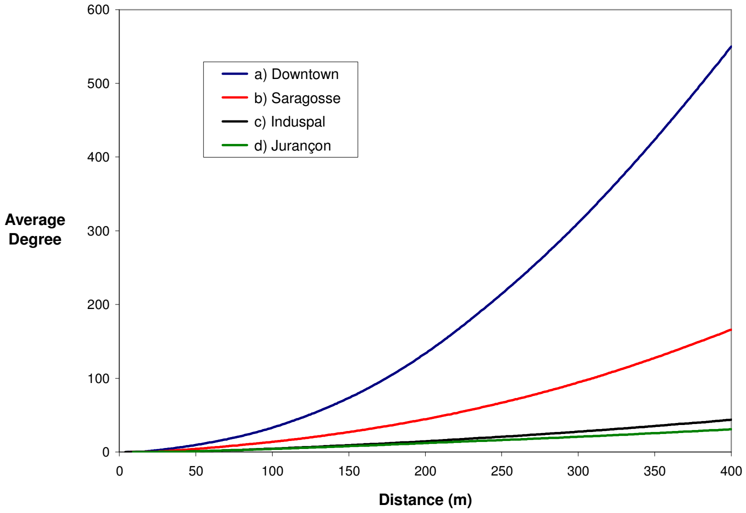


Chart 2: Comparison of neighborhood graphs average degree for four different urban patterns

Thus, variations in neighbor integration graphics correspond to heterogeneous composition in urban patterns, and they permit to characterize different structures through the analysis of their specific response to distance variations.

### Scaling behavior of distance structure

Furthermore, we computed the fractal dimension and the scaling behavior within our network model for the four patterns selected. This was done because the fractal dimension computation is a recognized technique to characterize and to discriminate morphologies: it will give us another indicator to compare the morphodynamics of our panel of urban quarters.

The computation is based on a radial analysis approach, putting into relation the degree of the resulting graph, representing the number of units connected ($N(\varepsilon)$), with the distance steps, representing the size of the gauge $\varepsilon$, as in the following formula:

$$D = log\ (N(\varepsilon))\ /\ log\ (\varepsilon)$$

Which gives, transposed in a graph formalism context:

$$D = log\ (degree)\ /\ log\ (distance)$$

The radial analysis is a common method used to calculate the fractal dimension of a pattern. It refers to a specific point, known as the counting center, and gives the law of distribution of the occupied sites around this point (Caglioni & Rabino, 2004: 4). A circle, or a square, is drawn around the point and the radius or size is gradually increased: at each step it is then possible to calculate the number of occupied points ($N(\varepsilon)$ = degree) and to put it in relation with the size considered ($\varepsilon$ = distance). Reported on a log-log plot, this relation appears as a straight line, where the slope is the *D* dimension.

As *D* may be computed at each step with this method (the denomination "local" *D* is then used), we can also compute the local value of the slope, from step to step, which is similar to computing the local *D* at each step. These "local" *Ds* are represented and give us the curve of the "scaling behavior", which means the local variations of *D*. The scaling behavior curve shows the local variations in the *D* value, and helps visualize the changes in the spatial organization of the analyzed morphology.

The results (Chart 3) show a general trend, like a square root symbol, which is also observed in the scaling behavior curves that are usually computed in urban analysis (Frankhauser, 1994). Although the initial morphological data is described in a bidimensional space, space has been transformed by the computation of space-time distances. This computation induces considerable changes in the geometry of space, which leads to local deformations and overlapping for instance. Consequently, the triangular inequality cannot always be respected and space may be tridimensional instead of bidimensional in certain places. For this reason, the computed fractal dimension may in some cases exceed 2.

We notice that the fractal dimension is always above 2 for Downtown and Saragosse, which means that these quarters are complex and strongly interconnected. The value is above 2 because these two quarters have a "superspace" structure in the graph representation: we no longer are in a regular 2D Euclidean space representation but almost in a 3D one. On the contrary, Induspal and Jurançon are less complex: their fractal dimension is always below 2, which corresponds to a regular measure for an Euclidean urban structure (Badariotti, 2005). This means that these quarters have a space structure in the graph representation which is not very different from a Euclidean space. In all cases, the very smooth slowdown of the curve shows that the general structure of each quarter is quite homogeneous and submitted to the same fractal law.

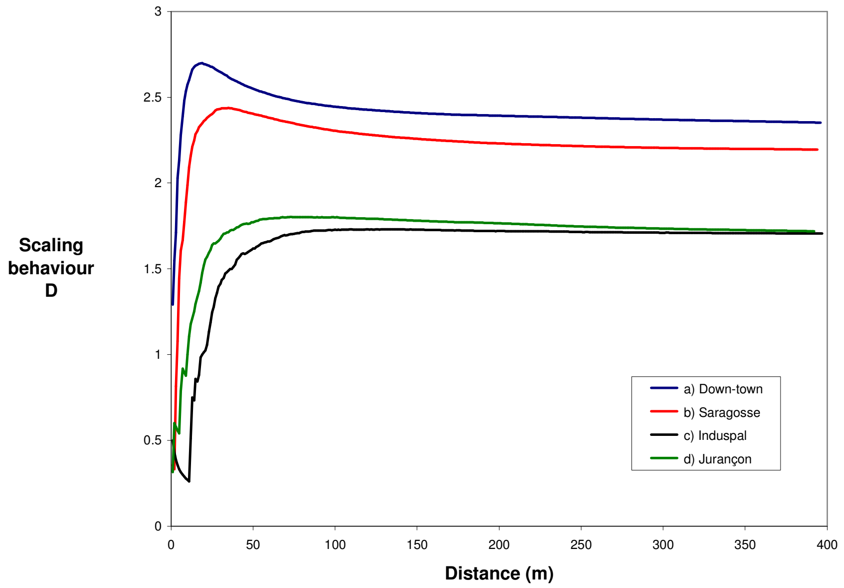
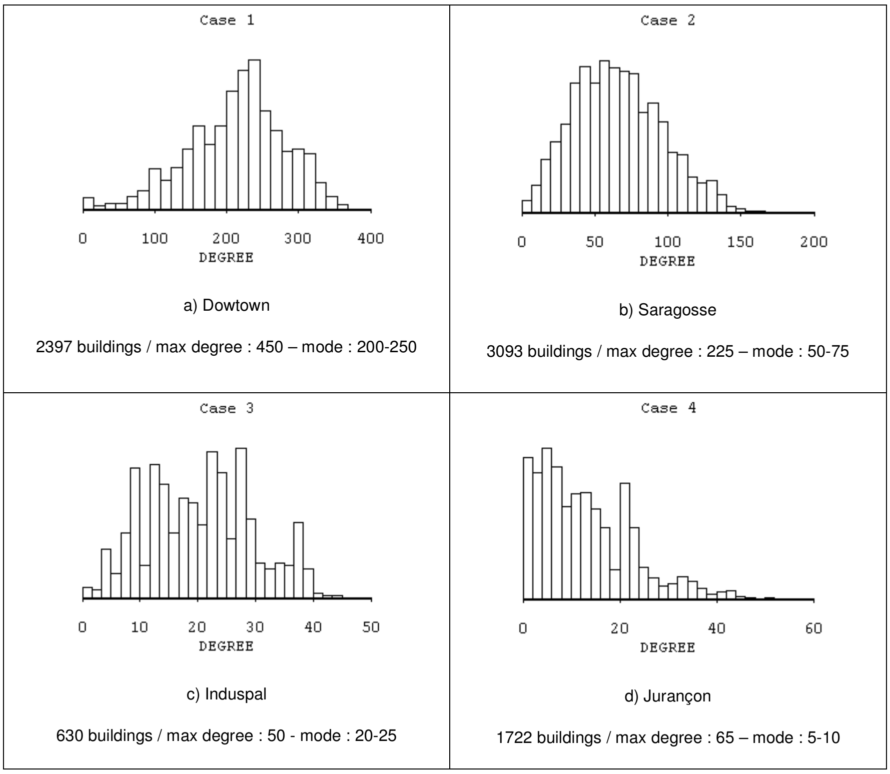


Chart 3: Comparison of the scaling behavior of four different urban patterns

### Degree distribution analysis

In order to study the neighboring graphs corresponding to the four types of pattern with great precision, we decided to compare the statistical distribution of their degree for a threshold distance of 250 m (Figure 7). We can observe that the degree distribution is quite different from one case to another. The statistical law varies from a sub-gaussian distribution (Downtown) to a Poisson distribution (Jurançon): each network presents a high variation in the distribution of the degrees. Two of them are clearly uni-modal but at Induspal and Jurançon we observe a second mode. The level of the modal value is also very different, ranking from 5-10 for Jurançon to 200-250 for Downtown.



a) Dowtown

2397 buildings / max degree : 450 – mode : 200-250

b) Saragosse

3093 buildings / max degree : 225 – mode : 50-75

c) Induspal

630 buildings / max degree : 50 - mode : 20-25

d) Jurançon

1722 buildings / max degree : 65 – mode : 5-10

Figure 7: Degree distribution of buildings in the four zones for a distance of 250 m

We then notice that the observed networks do not really correspond to the NG model, nor to the urban network model, because of the observed high degree level. They may neither be interpreted as an ER model, or random network model, because of the distribution of degrees: in the ER model, each degree is equi-probable, so the distribution would be very different from the observed one.

The observed networks are in fact more compatible with the WS model, or small-world model, characterized by a locally highly interconnected network. In fact, each distribution shows a mode, but this mode is more detached from the other values Downtown: this zone is also the place where the degrees are much higher. The BA model, or scale-free model, which has a large distribution of degrees decreasing slowly as in a hierarchical law, may also be invoked for Jurançon. The other two quarters show intermediate solutions between the two models.

Differences between the four cases can also be observed on the rank-size curves of the degree distribution of each zone (Figure 12). We first notice that the total number of buildings and the maximum degree of each zone are very different from one another. The network structures of these zones show great differences that make their comparison difficult.

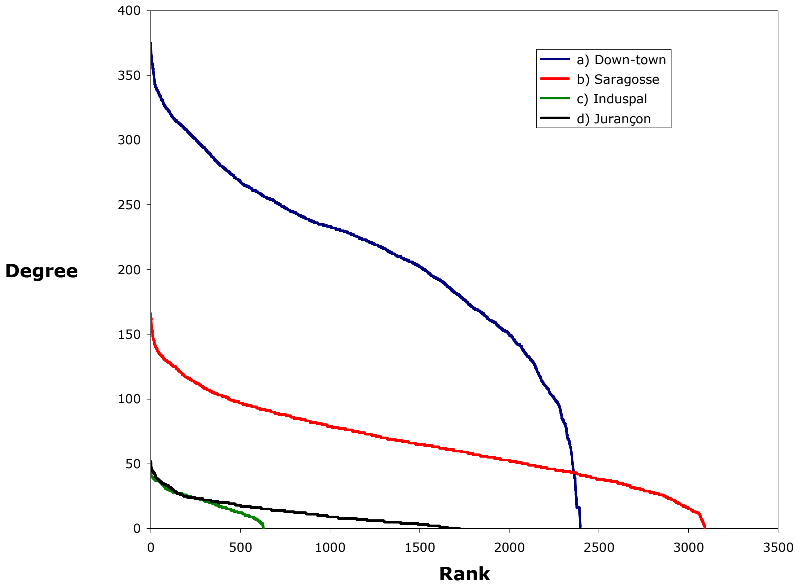


Chart 4: Rank-size distribution of the degrees of four specific quarters

The application of the Remus model to the city of Pau demonstrates the interest of this model. It is clear that the graph formalism used permits the computation of original morphologic indexes that are equally coherent for similar-looking quarters, and discriminate different-looking quarters.

The comparison between random patterns and real patterns, and the computation of Euclidean distances versus graph distances highlight the specific structure of the city's morphology. The idea of a fractal structure of towns or cities remains in the analysis of urban proximities at different distances, which shows great differences according to the quarter under consideration. The degree distribution analysis makes it evident that there is an order in the town, with a rank-sized organization, beyond the complexity of the morphology. So Remus has shown its ability to compute classical and less classical morphologic indexes, differentiating the town's various morphologies.

Remus is not a compilation of techniques. The idea is that urban forms are complex and they have a strong influence on urban dynamics. In this context, Remus appears as a new tool that helps characterize the morphologic complexity of cities from a structural point of view: it can then be used as a new tool to analyze the structure of the city and its effects on dynamics.

## IMPACT OF THE REMUS MODEL ON CELLULAR AUTOMATA PROXIMITY-BASED PROCESSES: APPLICATION TO SCHELLING'S SEGREGATION MODEL

The graph-based neighborhood structure thus constitutes the structure of our cellular automaton model. One problem then: how can we assess the functionality of the Remus model, in particular its way of modeling proximities within the town, to study the impact of morphology on neighborhood processes?

### The Schelling's spatial proximity segregation model

We propose to use a very classical process, the Schelling's segregation model founded on proximity behaviors, and to implement it within Remus.

Indeed, Schelling (1969, 1971, 2006) illustrated the emergency of a highly organized phenomenon from the combination of individual behaviors, using the segregation example. In his "spatial proximity model" (Schelling, 1969), a population is divided into two groups, and the individuals' membership is permanent and recognizable. This membership can be identified by a color or a symbol. Everybody has a particular location at any moment, everybody observes his neighbors' colors or symbols and everybody is capable of moving if he is dissatisfied with the color mixture of his neighbourhood.

Firstly, agents (identified by plusses and zeros) are randomly distributed along a line (Figure 8a). The neighborhood of each agent includes the four nearest neighbors on both sides of him. Dissatisfied individuals move to the nearest point that meets their minimum demand (the nearest point at which half their neighbors will be like themselves at the time they arrive there) between two agents. The first agent to move is the first dissatisfied one, from left to right. Then the agents satisfaction is calculated again. The second agent to move is the next dissatisfied agent situated at the right of the first agent. The process is repeated until the last agent moves. The model reaches stability in one or two steps and the final configuration shows an alternation of homogeneous clusters, irrespective of the initial configuration.

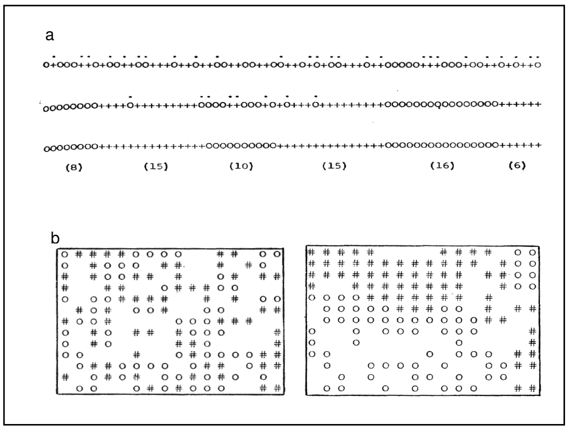


Figure 8: Schelling's dynamic segregation model: a) the linear model: the plusses and zeros represent people scattered in a line, a dot is put over each individual whose neighborhood does not meet his demand, each line represents a step in the model, the bottom line represents the final stable and clustered configuration; b) the two-dimensional model; left: sharps and zeros are randomly distributed among a rectangular space of 208 cells, 70 empty places make the individuals' movements easier, as in the linear model there is an equal number of sharps and zeros; right: the universal demand is that no fewer than half of one's neighbors be of the same sign, the discontent moves to the nearest satisfactory vacant square, the final configuration shows a segregated pattern (Schelling, 1971)

Schelling identified five elements that can be varied in the linear model: neighborhood size, demanded percentage of one's own symbol or color in the neighborhood, ratio of plusses to zeros in the total population, rules governing movement, and original configuration. Varying these parameters always results in segregated patterns. The main variation observed in the outcome of the model is the size of the clusters which grows when the numbers of group members differ.

Similar results are obtained in the two-dimensional model (Figure 8b). It begins with a number *n* of individuals, of which *n/2* agents are sharps and *n/2* are zeros situated randomly on a 13-rows and 16- columns "checkerboard". There are *m* free places, with the ratio *m/n* comprised between 25% and 30%, in order to allow "freedom of movement without making the board too empty" (Schelling, 1971: 154). This board configuration can be described as a regular lattice graph. Initially, the neighborhood of each agent has been defined as their 8 surrounding squares, a Moore neighborhood, as described in Table 1. Each agent has the same behavioral rule as in the linear model: if the proportion of agents of a distinct color in the neighborhood is higher than the tolerance threshold of 50%, this agent will move towards the nearest free place satisfying his demand.

It is obvious from that rule that if the value of threshold is low (agents of low tolerance, with a high demand for similar neighbors), the number of dissatisfied agents is high, and the rate at which the agents move leads to a very structured spatial organization, marked by a more or less clear separation between two types of agents. What is more surprising and makes this little model successful is the following paradox: even if the value of threshold $\lambda$ is high (for instance, 70% of neighbors of another color accepted), the dynamics of the model will still tend towards a spatial separation between agent categories; in the end, the agents will no longer have 70% of neighbors of another color, but something nearer to 30%.

Schelling varied other parameters in the two-dimensional model: unequal demands for the two groups of agents, unequal numbers of agents, size of the neighborhood, different types of agent preferences. He observed that when the number of agents in the two groups differ, the minority group tends to segregate even more than the other one. In a similar way, when tolerance differs in the two groups, the less tolerant group, tends to segregate even more than the other group. When the neighborhood size increases, the tendency to segregate is attenuated for moderate demands and near-equal number groups. Still a "congregationist" preference, like a group demand based on the number of similar individuals in the neighborhood, leads to segregation. Finally, Schelling tested an integrationist strategy in his theoretical model: the two groups prefer a ratio of like colors. The model stabilization is hindered by this strategy, a larger number of individuals are dissatisfied, and the final pattern is more complex than that of the segregated one.

The segregation model described corresponds to one of three types of models that Schelling presents in his article. The two other types are less commented in literature: a "bounded-neighborhood model" and a "tipping model". In these models all the agents share the same neighborhood and boundaries, instead of each neighborhood being defined by reference to its own location. These models are non spatial. Schelling thus analyzes the influence of tolerance on segregation, obviating the main parameter controlling segregation on the "spatial proximity model": the spatial structure of agents' neighborhoods.

The main result of this model, which has inspired many other studies, is that segregation is an emergency of the combination of individual demands, even if this demand is not segregationist. The question is: if agents don't want segregation, why does segregation occur? Segregation seems to be the consequence of the individuals' neighborhood preferences, and it should be related to the characteristics of neighborhood.

Being aware of the extreme simplicity of his model, Schelling recognized that "the results (of this model) are only suggestive, because few of us live in square cells on a checkerboard" (Schelling, 2006: 151). But he didn't reduce the limits of his model to the spatial structure. He accepted that his model was too abstract and artificial to be applied to the real world, because it makes no provision for other factors which have a strong influence on residential choices, like family incomes, family sizes, and the cost of space. "Quantitative measures, of course, refer exclusively to an artificial checkerboard and are unlikely to have any quantitative analogue in the living world" (Schelling, 1971: 158). The model is supposed to be suggestive of the segregating process, and it could be used to compare the different factors that have some influence upon it by studying their influence on the model behavior: "Comparisons among them, however, such as the effect of reducing or enlarging the size of a minority, may be capable of some extension to that world" (Schelling, 1971: 158).

Does the Schelling model simulate segregation? For Schelling the model is only suggestive of segregation, or aggregation in certain cases, but it could not represent real processes. One could argue against this model that it represents a nucleation process due to the asymmetry of migration rules (satisfied individuals don't move), that Schelling overestimated the value of the agents' tolerance in his analysis, and that he doesn't include in it the other factors that are responsible for the segregation (Forsé & Parodi, 2006). But the real questions are: What is the purpose of modelling? Does a model have to represent real dynamics, or is it not just a tool to compare the parameters that may have some influence on these dynamics? Does Schelling study segregation or the self-organizing process in general? Does the ontological definition of a model require to validate its results with reality?

Our position is that the segregation model proposed by Schelling can be used as an example of self-organizing phenomena, but it can't be used to explain the complexity of segregation in a city. However, the understanding of the model behavior permits to study its sensibility to different parameters that can be included in more realistic models. In our case, the model has allowed us to characterize urban patterns, through the understanding of their influence on its dynamics.

### Variants of the spatial proximity model

The spatial proximity model has been extended many times in the literature (Aydinonat, 2007). The model variants can be divided into three types: extension of the utility function that controls the agents' satisfaction, changes in agent movement rules, and variants in the space configuration. Recently, the extension of the utility function to asymmetric peaked functions (Pancs & Vriend, 2007; Barr & Tassier, 2008; O'Sullivan, 2009), and the incorporation of economic variables, using stochastic evolutionary game theory techniques (Young, 1998; Young, 2001; Zhang, 2004a, 2004b; Möbius, 2000; Bøg, 2006; Dokumaci & Sandholm, 2006) have been used to try and explain the segregation in the Schelling model from a formal analytical perspective. The results confirmed Schelling's conclusions. Variants in the way the agents move on the checkerboard, like the introduction of inertia in position preferences or changes in the order of moves (Pancs & Vriend, 2003), lead to the same segregated patterns that exist in the original model outcome.

Variants in the spatial structure of the model are less frequent. Pancs & Vriend (2003) tested circular and torus configurations to avoid border influence on model dynamics. These changes didn't affect the model dynamics. Laurie & Jaggi (2003) extended Schelling's analysis of the influence of neighborhood size on the dynamics of the model. They identified three regimes: an unstable regime which leads invariably to segregation, a stable regime, and an intermediate regime with complex behavior. The increase in the neighborhood size accentuates the dynamics of the model: it stabilizes the model when the agents' threshold demand is low, but impedes stabilization when the threshold tolerance is high. Flache & Hegselmann (2001) tested the influence of irregular grids on cellular automata. Even if they focused their analysis on cooperation and migration models, they have shown that these models are globally robust to variation in the grid structure but are sensitive to local variations of neighbors within a grid. This behavior has not been demonstrated with the Schelling spatial proximity segregation model.

### Implementation of spatial proximity model in Remus

Irregular neighborhoods could have a strong influence on model dynamics. So, to implement Schelling's model on a more realistic urban pattern (like Remus) with a view to exploring the behavior of the Remus model in a complex structure could be interesting. In particular, we wonder if morphology does affect Schelling's model dynamics and what kind of morphologies emerge if we take into account the more realistic way individuals have of living close together, integrating morphology. In other terms: how does Schelling's model work if we are not in a Moore neighborhood? This approach may help reveal the sensibility of segregation processes to an urban structure, defined both by its morphology and network accessibility.

In order to integrate the network structure in Schelling's model, it is possible to replace the Moore neighborhood of each cell by a set of buildings defined by the neighborhood graph calculated with Remus for a specific distance threshold. Then, Schelling's transitional function can be adapted to the reticular structure: an agent occupying a building moves to another building if the proportion of neighboring buildings occupied by agents who are different from it exceeds a certain limit. The number of neighboring buildings is not fixed in the Remus structure. Thus the limit is defined by the ratio between neighboring buildings occupied by different agents and the total number of neighboring buildings. In Figure 9 we show the simulation interface in Remus model. The two agent groups are represented by blue and red colors, and empty buildings are represented in white. The movement of all the unsatisfied agents occurs at the same time at each step. The choice of their destination is made randomly among the empty buildings.

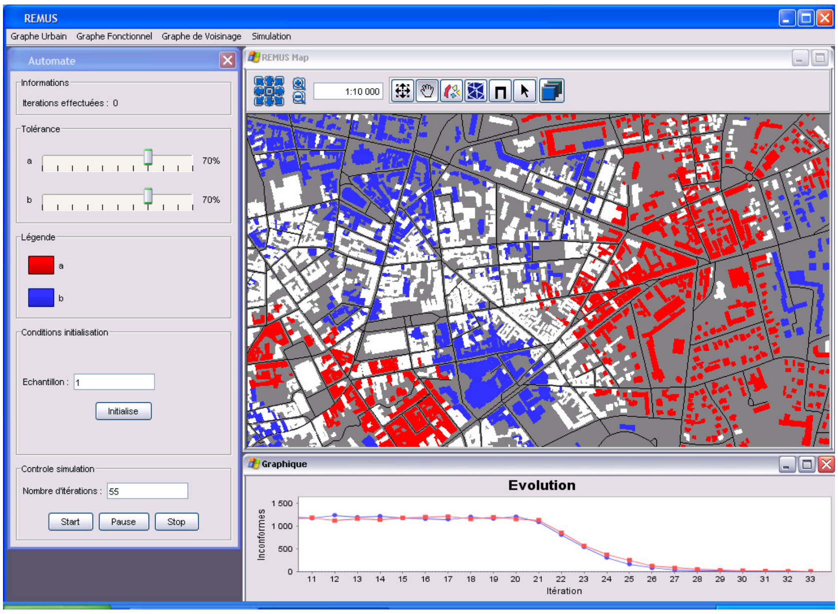

Informations
Iterations effectuées : 0
Tolérance
a 70%
b 70%
Légende
🟥 a
🟦 b
Conditions initialisation
Echantillon : 1
Controle simulation
Nombre d'itérations : 55

Figure 9: The Schelling model implemented in Remus

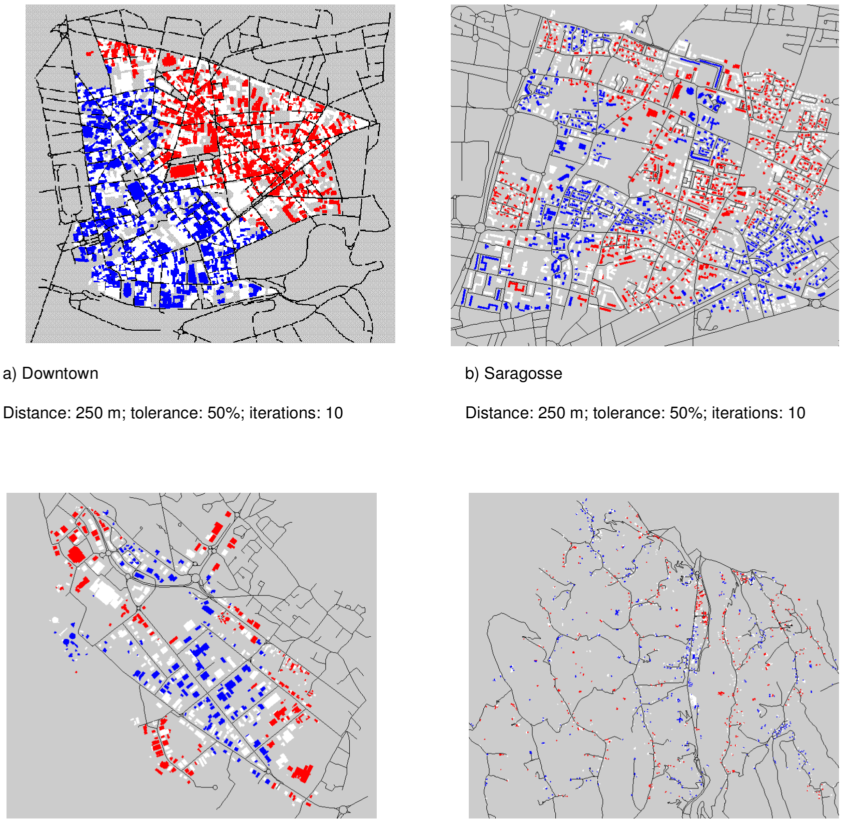
This generalization of the Schelling model permits to compare different types of patterns. The number of iterations needed to stabilize the system depends on the individuals' tolerance to unlike neighbors, as Schelling demonstrated. In terms of dynamics, differences in the pattern structure don't seem to affect the way the model reaches stability. But how does the structure affect spatial segregation? Is the behavior of the inhabitants the same in every part of the city? Is it possible to use the segregated patterns to characterize the structural differences revealed through the neighborhood graph analysis?

### Application of Schelling's model to specific urban zones and discussion

In order to answer these questions, we have tested Schelling's segregation model using the patterns corresponding to the four areas of Pau-Pyrénées (Figure 10).

The results showed a slight influence of the type of pattern on the way the urban system reaches stability. In the zones where the buildings are more distant from one another, like Induspal or Jurançon, stability is rapidly obtained. In the downtown areas with tight urban patterns, the stability is slightly more difficult to obtain. These results are due to the differences in the number of neighbors which are lower in loose patterns for the same distance threshold. These observations confirm Schelling's observation (Schelling, 1971) and Laurie and Jaggi's analysis (Laurie & Jaggi, 2003) about the effect of increasing neighborhood size in the spatial proximity model: when the neighborhood is increased, as the number of neighbors increases, the segregation is attenuated.

However, differences appear in the final configurations of the population's spatial distribution. In Schelling's classical model, the position of clusters was a non-linear response to the dynamics of the model. In the case of urban network proximities calculated by Remus, the urban structure seems to have a strong influence on the spatial distribution of population clustering. The main crossroads, corresponding to the best-connected nodes in urban graph, could represent attractors for clustering: buildings in the crossroads proximity are more connected, and seem to aggregate in big clusters.



a) Downtown

Distance: 250 m; tolerance: 50%; iterations: 10

b) Saragosse

Distance: 250 m; tolerance: 50%; iterations: 10

c) Induspal

Distance: 250 m; tolerance: 50%; iterations: 8

d) Jurançon

Distance: 250 m; tolerance: 50%; iterations: 7

Figure 10: Schelling model tested for different urban patterns in Remus

This result seems to be contradictory with the previous conclusions. Why well-connected buildings segregate more than less connected ones if an increase in the number of neighbors attenuates segregation? In fact, segregation dynamics may be controlled by poorly connected zones separating well-connected ones. Poorly connected zones are unstable, because buildings have fewer neighbors and equilibrium is difficult to obtain. However, they isolate the well-connected zones that reach stable local equilibriums by the formation of a big cluster. This behavior is illustrated in Figure 11. In this graph there is a central node that possesses two neighbors and that divides the graph in two cliques. Inside the cliques, each node has 3 or even 4 neighbors. The stability in the cliques is difficult to obtain when they have an equal number of agents of each color. But when a majority is imposed, segregation is rapidly obtained and the agents quickly separate into two distinct color cliques.

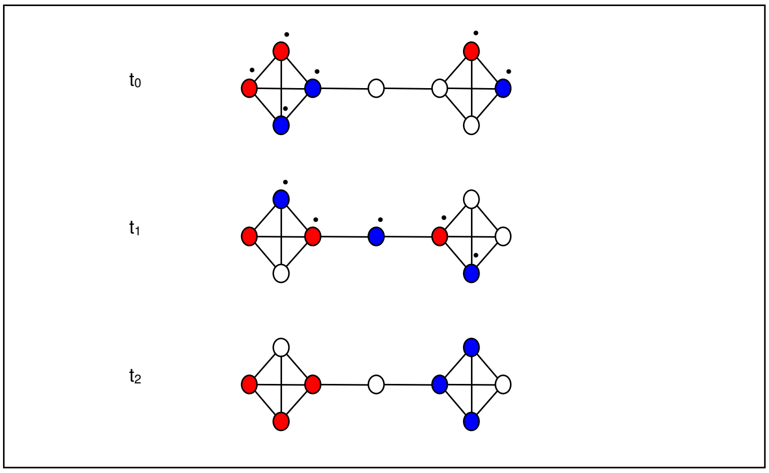


Figure 11: Schelling's segregation model applied to an irregular graph with two cliques: the circles represent nodes occupied by red agents, blue agents, or empty (white nodes); dots over nodes indicate dissatisfied agents; at $t_0$ all the agents are dissatisfied, a stable configuration is impossible with an equal number of agents in each clique; at $t_1$ a red agent is satisfied and a majority is obtained in the left clique, the middle node is occupied but unstable; at $t_2$ stability is reached by the separation of the agents in each clique

This hypothesis has been recently analyzed by Banos (2009). He has proved that this "clique effect" disrupts the model behavior in irregular graphs, because cliques constitute cluster attractors in segregation dynamics. Scale-free and fractal networks are more influenced by this effect because of the hierarchical structure of their degree distribution. These results can be confronted to the conclusions of Fagiolo et al. (2007) about Schelling's model behavior when applied to different types of networks. The latter measured segregation in lattice, random, regular, small-world and scale-free networks. They conclude that the structural network properties do not seem to have a strong impact on segregation levels obtained in the long-run by the system. Moreover, they have demonstrated that an increase in the average degree (the average number of neighbors) will correspond to lower levels of segregation

(Fagiolo et al., 2007). These results extend Schelling's (Schelling, 1971) and the conclusions of Laurie & Jaggi (2003) about the influence of neighborhood size on segregation in network structures.

For them, the relative impact of one agent move is inversely related to the number of links: in a low-degree network an agent move is more likely to affect the satisfaction level of other agents, while in high-degree networks changes are "absorbed" by the agents which easily find a configuration that makes them happy without the need to form clusters.

These statements are based on measures of the average segregation in theoretical graphs. In our actual neighborhood graph analysis, we examine the local variations of segregation and we observe that well-connected sub-graphs have more segregated levels than the less connected zones. But at the whole graph level, the segregation levels are equivalent to those that other spatial structures have. Thus the localization of homogeneous areas that appear characteristically in the Schelling model could depend on the local network structure of the urban space, and more precisely on the combination of well-connected and less-connected zones.

This aspect could be interesting for urban planners and should be explored by geographers. Does the spatial distribution of nodes and degrees affect the spatial distribution of segregated clusters in the city? In other words: does intra-urban accessibility constrain the geography of segregation? To examine this point, we propose to compare patterns of segregation obtained with Schelling's model with the maps showing the spatial distribution of the degrees in the four selected zones (Figure 12). Comparing both distributions (the degree distribution and the segregation distribution), we observe that there is a certain rate of overlapping between these two realities: segregated clusters seem to appear more frequently around places that share the same degree level.

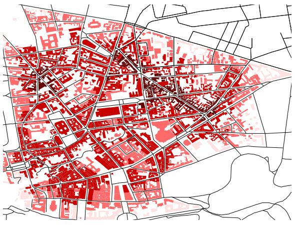
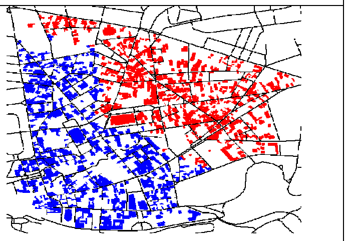
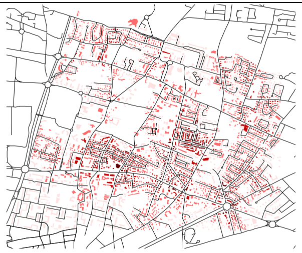
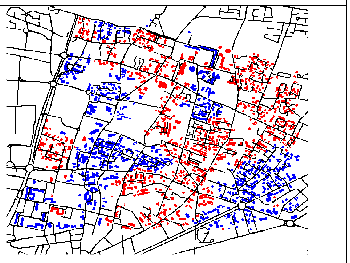
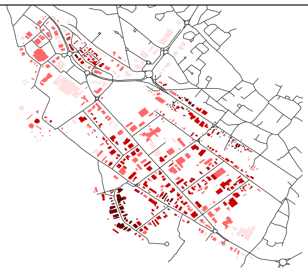
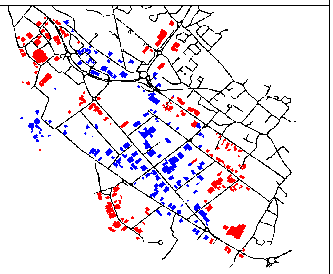
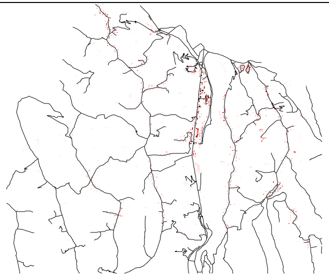
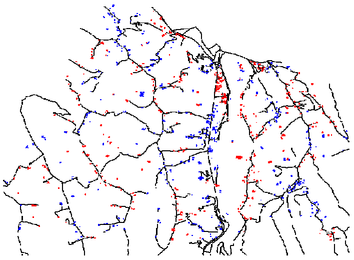
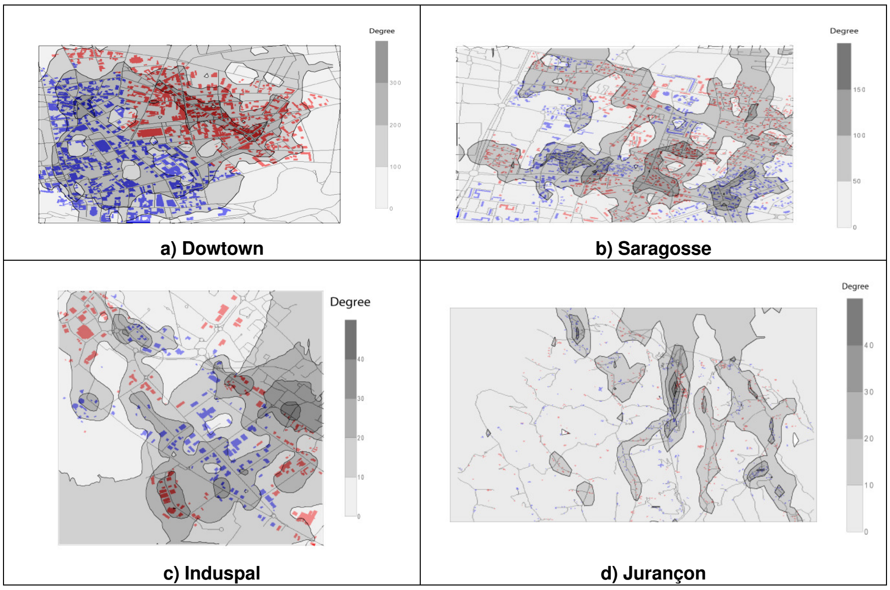
|  | Degree distribution | Pattern of segregation |
|---|---|---|
| a) Downtown |  |  |
| b) Saragosse |  |  |
| c) Induspal |  |  |
| d) Jurançon |  |  |

Figure 12: Degree distribution and patterns of segregation in the selected zones of Pau-Pyrénées

To get a better idea of this overlapping, we interpolated the degree distribution and superimposed it on the map of Schelling's segregated patterns (Figure 13). An attentive observation shows us that in most cases, clusters appear in places with a homogeneous degree level. In particular, we can see on Figure 12 and on Figure 13 that the best-connected places are not really segmented by segregation. They may belong to population *A* or population *B*, but they are very seldom fragmented between the two populations. The zones with a high-degree level seem more to correspond to the cores of the clusters. The less-connected zones seem more mixed, because they separate large clusters. In these zones there is a higher proportion of empty buildings too.



**a) Dowtown**

**b) Saragosse**

**c) Induspal**

**d) Jurançon**

Figure 13: Segregated clusters versus interpolated degree distribution

It is then more evident that the spatial structure plays a role, not really in the dynamics of the social process of segregation in the whole town, but in the segregation patterns that result from these dynamics. There is a certain relationship between the spatial structure of the quarters, and the spatial segregation that we observe, which should be studied more accurately with the help of some quantitative indicators. In particular, it would be interesting to examine the local behavior of the system, comparing statistically the degree of each node with segregation indicators calculated from the stabilized configuration of the output of the simulation.

That approach may be more conclusive about the relationship between connectivity and segregation in Schelling's model.

## CONCLUSION

Studying urban morphology may be undertaken by analyzing the cadastral matrices that index them and the spatial relationships between buildings. This depiction may be viewed from a new angle by using a geographic cellular automata portraying the links of proximity between urban buildings in both Euclidean space and, concurrently, in network space. This new formulation offers several methodological advantages that have to be explored.

The analysis of neighboring graphs shows that urban patterns integrate neighborhood relationships in a differential way when network distance varies. It is possible to determine where groups of sites are most closely clustered, which in effect amounts to defining neighborhood communities among buildings. These neighborhood communities permit to identify potential "small worlds" in the town (Albert & Barabasi, 2002; Newmann, 2003a) and thereby identify urban sub-parts with preferential ties.

On the one hand, the integration of this neighboring graph structure to the Schelling model reflects the potential which reticular automaton models have for depicting urban structures through the analysis of model behavior. On the other hand, the analysis of the effect of urban patterns on the spatial proximity model through Remus contributes to improving the comprehension of social dynamics. Spatial segregation in a network controlled by Schelling's rules is encouraged by a hierarchical distribution of degrees, as in scale-free graphs. The alternation of well-connected and less connected sub-graphs seems to favor clustering around central nodes. Hence future works will focus on the use of such a statistical indicator as spatial autocorrelation between similar buildings in Schelling's model simulation, in order to analyze the behavior of the system when varying metrics, distance and tolerance thresholds in neighborhood definition.

Small-world graphs could encourage diversity, because of the high number of connections for all nodes. This kind of graph structure exists in social networks, but it isn't achievable in urban networks, in which links are not potential but physical. The huge number of connections implies a strong mobility increase, making transport systems more complex and consequently causing their collapse. On the contrary random graphs have a homogeneous node-degree distribution.

But this kind of graphs can have isolated nodes or sub-graphs, and inaccessibility is not acceptable in urban environments.

The structure revealed by the neighbor integration graph of the Communauté d'Agglomération de Pau-Pyrénées could possess a particular distribution of degrees permitting to limit spatial segregation. The urban pattern constrains the number of links at a local level but allows a rapid integration of connections at a larger level ensuring accessibility in the town. This type of pattern can't avoid segregation, as Schelling conceived it. These structures can only limit its extension. However, they could reflect the social equity that underlies the emergence of urban morphology.

This approach can also be used to define proximal zones in a town – kinds of "governmental pockets" – that are useful for public authorities to improve town governance. The hypothesis formulated in this perspective is that spatial proximity via the road network can constitute a reasonable approximation of underlying social proximity.

This new generation of geographic models allows one to represent more accurately the proximity relations in space. Geoalgebra and graph-based cellular automaton models permit to integrate the fundamental anisotropy of space in cellular automaton models. The introduction of accessibility is a key perspective to continue exploring the city morphology and its effects on the internal dynamics of the city.

## GLOSSARY

"Euclidean" or "physical" graphs:

- *G(V,E)*: urban graph (all vertices, all edges).
- *G(B,C)*: sub-graph of connectors between buildings and road segments (buildings, building-to-roadway connectors).
- *G(I,T)*: sub-graph of roads (intersections, thoroughfares).

"Functional" distance-time graphs:

- *G'(U,K)*: functional graph (building units, linkages measured in distance-times).
- *Gt"(U,N)*: neighboring graph (building units, neighboring links according to a specified distance-time threshold and specified mode of transportation).

## REFERENCES

**Aizenman, M.; Barsky, D.,** 1987, Sharpness of the phase transition in percolation models. In: Communications in Mathematical Physics, 108: 489–526

**Albert, R.; Barabási, A.-L.,** 2002, Statistical Mechanics of Complex Networks. In: Review of Modern Physics, 74: 47-97

**Albin, P.S.,** 1975, The Analysis of Complex Socioeconomic Systems. Lexington Books: Lexington

**Aydinonat, N.E.,** 2007, Models, conjectures and exploration: An analysis of Schelling's checkerboard model of residential segregation. In: Journal of Economic Methodology, 14, 4: 429-454

**Badariotti, D.,** 2005, Des fractales pour l'urbanisme ? In: Cahiers de géographie du Québec, 49, 137, sept.: 133-156

**Badariotti, D.; Banos, A.; Moreno, D.,** 2006, Modélisation de la structure spatiale urbaine par un automate cellulaire non stationnaire: le modèle Remus. In: International Conference on Spatial Analysis and GEOmatics, Research & Development: 17

**Badariotti, D.; Banos, A.; Moreno, D.,** 2007, Conception d'un automate cellulaire non stationnaire à base de graphe pour modéliser la structure spatiale urbaine: le modèle Remus. In: Cybergeo, 403. http://www.cybergeo.eu/index10993.html

**Banos, A.,** 2009, Influence des 'effets de clique' dans la dynamique du modèle de ségrégation de Schelling. 16èmes journées de Rochebrune, rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels

**Barabási, A.-L.; Albert, R.,** 1999, Emergence of scaling in random networks. In: Science, 286, 5439: 509-512

**Barr, J.M.; Tassier, T.,** 2008, Segregation and Strategic Neighborhood Interaction. In: Eastern Economic Journal, 34, 4: 480-503

**Barredo, J.I.; Kasanko, M.; McCormick, N.; Lavalle, C.,** 2003, Modelling Dynamic Spatial Processes: Simulation of Urban Future Scenarios through Cellular Automata. In: Landscape and Urban Planning, 64: 145-160

**Barthélemy, M.; Barrat, A.; Pastor-Satorras, R.; Vespignani, A.,** 2004, Velocity and hierarchical spread of epidemic outbreaks in complex networks. In: Physical Review Letters, 92, 178701

**Batty, M.; Longley, P.A.,** 1986, Fractal-based description of urban form. In: Environment and Planning B: Planning and Design, 14, 2: 123-134

**Batty, M.; Xie, Y.,** 1994, From Cells to Cities. In: Environment and Planning B: Planning and Design, 21: 31-48

**Batty M.; Xie Y.,** 1997, Possible Urban Automata. In: Environment and Planning B: Planning and Design, 24, 2: 175-192

**Bejan, A.,** 1997, The Constructal law of structure formation in natural systems with internal flows. In: ASME AES, 37: 257-264

**Bejan, A.,** 2007, Constructal theory of social dynamics. Gilbert W: Merkx

**Bejan, A.; Ledezma, G.A.,** 1998, Streets tree networks and urban growth: optimal geometry for quickest access between a finite-size volume and one point. In: Physica A, 255: 211-217

**Bélair, J.,** 1987, Sur le calcul de la dimension fractale. In: Annales de sciences mathématiques, 1, 11: 7-23

**Berry, B.; Lobley. J.,** 1964, Cities as systems within systems of cities. In: Papers of the regional science association, 13: 147-163

**von Bertalanffy, L.,** 1950, An outline of general system theory. In: British journal of philosophie of science, 1: 139-164

**Bøg, M.,** 2006, Is segregation robust? Unpublished manuscript, Stockholm School of Economics. http://mpra.ub.uni-muenchen.de/8774/

**Burmeister, A.; Colletis-Wahl, K.,** 1997, Les interactions production-transport-espace. In: Revue d'économie régionale et urbaine, 3: 363-386.

**Caglioni, M.; Rabino, G.,** 2004, Contribution to fractal analysis of cities: a study of metropolitan aera of Milan. In: Cybergeo, 269. http://www.cybergeo.eu/index3634.html

**Chua, L.; Yang, L.,** 1988, Cellular Neural Networks: Theory and Applications. In: IEEE Transactions on Circuits and Systems: 1257-1290

**Clarke, K.C.; Hoppen, S.; Gaydos, L.,** 1997, A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay Area. In: Environment and Planning B: Planning and Design, 24: 247-261

**Chapin, F.S.; Weiss, S.F.,** 1968, A probabilistic model for residental growth. In: Transportation Research, 2: 375-390

**Codd, E.,** 1968, Cellular Automata. Academic Press: New York

**Couclelis, H.,** 1985, Cellular worlds: a framework for modeling micro-macro dynamics. In: Environment and Planning A, 17: 585-596

**Couclelis, H.,** 1988, Of mice and men: what rodent populations can teach us about complex spatial dynamics. In: Environment Planning A, 20: 99-109

**Couclelis, H.,** 1989, Macrostructure and microbehavior in a metropolitan area. In: Environment and Planning B: Planning and Design, 16: 141-154

**Couclelis, H.,** 1997, From cellular automata to urban models: new principles for model development and implementation. In: Environment and Planning B: Planning and Design, 24: 165-174

**Couclelis, H.,** 1998, Geocomputation and space. In: Environment and Planning B: Planning and Design, 25th anniversary, 25: 41-47

**Dokumaci, E.; Sandholm, W.H.,** 2006, Schelling Redux: An Evolutionary Dynamic Model of Segregation. Working Paper. University of Wisconsin

**Dubois-Taine, G.; Chalas, Y.** (Eds), 1997, La ville émergente. Ed. de L'aube : Paris

**Dupuy, G.,** 1995, L'auto et la ville. Flammarion : Paris

**Engelen, G.; White, R.; Uljee, I.,** 2002, The MURBANDY and MOLAN Models for Dublin, submitted to European Commission Joint Research Center, Ispra, Italy. ERA-Maptec: Dublin

**Erdös, P.; Rényi, A.,** 1959, On random graphs. In: Publicationes Mathematicae, Debrecen, 6: 290-297

**Erdõs, P.; Rényi, A.,** 1960, On the evolution of random graphs. In: Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5: 17-61

**Eubank, S. et al.,** 2004, Modelling disease outbreaks in realistic urban social networks. In: Nature, 429: 180-184

**Fagiolo, G.; Valente, M.; Vriend, N.J.,** 2007, Segregation in networks. In: Journal of Economic Behavior & Organization, 64: 316-336

**Flache A.; Hegselmann, R.,** 2001, Do Irregular Grids make a Difference? Relaxing the Spatial Regularity Assumption in Cellular Models of Social Dynamics. In: Journal of Artificial Societies and Social Simulation, 4, 4. http://www.soc.surrey.ac.uk/JASSS/4/4/6.html

**Forrester, J.W.,** 1969, Urban dynamics. MIT Press: Cambridge Mass

**Forsé, M.; Parodi, M.,** 2006, La ségrégation spatiale selon Schelling : la perversité est ailleurs. In: Document de travail de l'Observatoire Français des Conjonctures Economiques, 5

**Floyd, R.W.,** 1962, Algorithm 97: Shortest Path. In: Communications of the ACM, 5, 6

**Frankhauser, P.,** 1990, Fractal structures in urban systems. In: Methods of Operations Research, 60: 97-108

**Frankhauser, P.,** 1991, Fractal geometry of urban patterns and their morphogenesis, Discrete Dynamics. In: Nature and society, 2, 2: 127-145

**Frankhauser, P.,** 1994, La fractalité de structures urbaines. Economica : Paris

**Gastner, M.T.; Newman, M.E.,** 2004, The spatial structure of networks urban networks. ArXiv:cond-mat/0407680

**Genre-Grandpierre, C.,** 2001, Laisser une chance aux modes non mécanisés par l'aménagement des réseaux routiers. In: Proceedings of the 5e rencontres de Théoquant. http://thema.univ-fcompte.fr/theoq/pdf/2001/genre.pdf

**Gourdon, J.-L.,** 2001, La rue: essai sur l'économie de la forme urbaine. Charles Mayer : Paris

**Geurs, K.T.; van Eck, J.R.R.,** 2001, Accessibility Measures: Review and Applications. Rijksinstituut voor Volksgezondheid en Milieu (National Institute of Public Health and the Environment, RIVM) and Urban Research Centre, Utrecht University, Bilthoven/Utrecht, Netherlands

**Geurs, K.T.; van Eck, J.R.R.,** 2003, Evaluation of accessibility impacts of land-use scenarios: the implications of job competition, land-use, and infrastructure developments for the Netherlands. In: Environment and Planning B: Planning and Design, 30, 1: 69-87

**Hedlund, G.A.,** 1969, Endomorphisms and automorphisms of the shift dynamical system. In: Math. Systems Theory, 3, 4: 320-375

**Hillier, B.; Hanson, J.,** 1984, The social logic of space. Cambridge University Press: Cambridge

**Janson, S.; Luczak, T.; Rucinski, A.,** 2000, The Phase Transition. In: Random Graphs. Wiley, New York: 103-138

**Jiang, B.,** 2007, A topological pattern of urban street networks: Universality and peculiarity. In: Physica A, 387: 647-655

**Kaufmann, A.,** 1968, Des points et des flèches… La théorie des graphes. Dunod : Paris

**Keeling, M.J.,** 2005, The implications of network structure for epidemic dynamics. In: Theoretical Population Biology, 67, February: 1-8

**Langlois, A.; Phipps M.,** 1997, Automates cellulaires: application à la simulation urbaine. Hermes: Paris

**Lathrop, G.T.; Hamburg, J.R.,** 1965, An Opportunity-Accessibility Model for Allocating Regional Growth. In: Journal of the American Institute of Planners, 31: 95-103

**Laurie, A.J.; Jaggi N.K.,** 2003, Role of 'Vision' in Neighbourhood Racial Segregation: A Variant of the Schelling Segregation Model. In: Urban Studies, 40, 13: 2687-2704

**Mandelbrot, B.B.,** 1967, How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. In: Science, New Series, 156: 636-638

**Mandelbrot, B.B.,** 1975, Les objets fractals : forme, hasard, dimension. Flammarion: Paris

**Mandelbrot, B.B.,** 1977, The fractals – form, chance and dimension. W. H. Freeman and co: San Francisco

**Mandelbrot, B.B.,** 1982, The fractal geometry of nature. W. H. Freeman and co: San Francisco

**Ménard, A.; Filotas, E.; Marceau, D.,** 2004, Automates cellulaires et complexité : perspectives géographiques. In: Documents de l'IAG, http://www.iag.asso.fr/

**Meunier, C.,** 1999, Infrastructures de transport et développement. L'apport de l'économie des réseaux. In: Les cahiers scientifiques du transport, 36, 1999: 69-85.

**Meyers, L.A.,** 2007, Contact network epidemiology: bond percolation applied to infectious disease prediction and control. In: Bull. Amer. Math. Soc., 44: 63-86

**Möbius, M.M.,** 2000, The formation of ghettos as a local interaction phenomenon. Unpublished manuscript, MIT

**Muhammad, S.; Ottens, H.F.L.; De Jong, T.,** 2008, Modeling the impact of telecommuting on future urbanization. In: the Netherlands. Tijdschrift voor Economische en Sociale Geografie (Journal of Economic & Social Geography), 99, 2: 160-177

**Nakajima, T.,** 1977, Application de la théorie de l'automate à la simulation de l'évolution de l'espace urbain. In: Congrès sur la Méthodologie de l'Aménagement et du Développement, Montréal. Association Canadienne-Française pour l'Avancement des Sciences et Comité de Coordination des Centres de Recherche en Aménagement, Développement et Planification (CRADEP): 154-160

**Newman, M.E.J.,** 2003a, Random graphs as models of networks. In: Handbook of graphs and networks. Wiley-VCH: Weinheim: 35-68

**Newman, M.E.J.,** 2003b, The Structure and Function of Complex Networks. In: SIAM Review, 45: 167-256

**Newman, P.; Kenworthy, J.R.,** 1989, Gasoline consumption and cities: a comparison of US cities with a global survey. In: Journal of the American Planning Association, 55: 24-37

**O'Sullivan, D.,** 2000, Graph-based cellular automata models of urban spatial processes, PhD Thesis. University College London: London

**O'Sullivan, D.,** 2001a, Exploring spatial process dynamics using irregular cellular automaton models. In: Geographical Analysis, 33: 1-18

**O'Sullivan, D.,** 2001b, Graph-cellular automata: a generalised discrete urban and regional model. In: Environment and Planning B : Planning and Design, 28: 687-705

**O'Sullivan, D.,** 2009, Schelling's model revisited: Residential sorting with competitive bidding for land. In: Regional Science and Urban Economics, 39: 397-408

**Pancs, R.; Vriend, N.J.,** 2003, Schelling's spatial proximity model of segregation revisited. Dept. of Economics Working Paper, Queen Mary, University of London, 487

**Pancs, R.; Vriend, N.J.,** 2007, Schelling's spatial proximity model of segregation revisited. In: Journal of Public Economics, 9: 1-24

**Papini, L.; Rabino, G.A.; Colonna, A.; Di Stefano, V.; Lombardo, S.,** 1998, Learning Cellular Automata in a Real World: The Case Study of the Rome Metropolitan Area. In: Bandini, S.; Serra, R.; Suggi Liverani, F. (Eds), Cellular Automata: Research Towards Industry, ACRI'96. Proceedings of the Third Conference on Cellular Automata for Research and Industry. London: Springer-Verlag: 165-183

**Peitgen, H.O.; Richter, P.,** 1986, The beauty of fractals. Springer Verlag: Berlin

**Phipps, M.; Langlois, A.,** 1997, Spatial Dynamics, Cellular Automata, and Parallel Processing Computers. In: Environment and Planning B: Planning and Design, 24, 2: 687-705

**Pickover, C.,** 1990, Computers, Pattern, Chaos, and Beauty. St. Martin's Press: New York

**Reis, H.,** 2008, Constructal view of the scaling laws of street networks - the dynamics behind geometry. In: Physica A, 387: 617-622

**Schelling, T.C.,** 1969, Models of Segregation. In: American Economic Review, Papers and Proceedings, 59, 2: 488-493

**Schelling, T.C.,** 1971, Dynamic Models of Segregation. In: Journal of Mathematical Sociology, 1: 143-86

**Schelling, T.C.,** 1978, Micromotives and Macrobehavior. W.W. Norton and Co: New York

**Schonfisch, B.,** 1997, Anisotropy in Cellular Automata. In: BioSystems, 41: 29-41

**Schroeder, M.,** 1991, Fractals, Chaos, Power Laws. W.H. Freeman & Co: New York

**Serrano, M.; Boguñá, M.,** 2006, Percolation and epidemic thresholds in clustered networks. In: Phys. Rev. Lett., 97, 088701

**Takeyama, M.,** 1996, Geo-algebra: a mathematical approach to integrating spatial modeling and GIS. Unpublished doctoral dissertation thesis, Department of Geography, University of California at Santa Barbara, USA

**Takeyama, M.; Couclelis, H.,** 1997, Map dynamics: Integrating Cellular Automata and GIS through Geo-Algebra. In: International Journal of Geographical Information Science, 11, 1: 73-91

**Tobler, W.,** 1979, Cellular Geography. In: Gale, S.; Ollson, G. (Eds), Philosophy in Geography. Dordrecht: D. Reidel: 379-386

**Turing, A.M.,** 1950, Computing machinery and intelligence. In: Mind, 59: 433-460

**Vernez-Moudon, A.; Hess, P.; Snyder, M.C.; Stanilov, K.,** 1997, Effects of site design on pedestrian travel in mixed-use, medium density environments, Report for the Washington State Transportation Center, Federal Highway Administration

**Von Neumann, J.; Burks, A.,** 1996, Theory of Self-Reproducing Automata. University of Illinois Press: Champaign

**Watts, D.J.; Strogatz, S.H.,** 1998, Collective dynamics of 'small-worlds' networks. In: Nature, 393: 440-442

**Weisbuch, G.,** 1989, Dynamique des Systèmes Complexes; une introduction aux réseaux d'automates. InterEditions / CNRS : Paris

**White, R.; Engelen, G.,** 1993, Cellular Automata and Fractal Urban Form: a Cellular Modelling Approach to the Evolution of Urban Land use Patterns. In: Environment and Planning A, 25: 1175- 1199

**White, R.; Engelen, G.,** 1997, Cellular Automata as the Basis of Integrated Dynamic Regional Modeling. In: Environment and Planning B: Planning and Design, 24, 2: 235-246

**Wiel, M.,** 2002, Villes et automobiles. Descartes et Compagnie, coll. Les Urbanités: Paris

**Xie, Y.C.,** 1996, A Generalized Model for Cellular Urban Dynamics. In: Geographical Analysis, 28, 4: 350-373.

**Young, H.P.,** 1998, Individual Strategy and Social Structure. Princeton University Press: Princeton

**Young, H.P.,** 2001, The dynamics of conformity. In: Durlauf, S.N.M Young, H.P. (Eds), Social Dynamics. Washington/Cambridge: Brookings Institution Press/MIT Press: 133-153

**Zhang, J.,** 2004a, A dynamic model of residential segregation. In: Journal of Mathematical Sociology, 28: 147-170

**Zhang, J.,** 2004b, Residential segregation in an all-integrationist world. In: Journal of Economic Behavior and Organization, 24: 533-550

## AUTHORS INFORMATION

**Diego MORENO**
diego.moreno.sierra@gmail.com
Laboratoire Société Environnement Territoire – UMR 5603 CNRS
Université de Pau et des Pays de l'Adour, France

**Dominique BADARIOTTI**
dominique.badariotti@live-cnrs.unistra.fr
Laboratoire Image et Ville – ERL 7230 CNRS
Université de Strasbourg, France

**Arnaud BANOS**
arnaud.banos@parisgeo.cnrs.fr
Laboratoire Géographie-Cités – UMR 8504, France

# ACCESSIBILITY ANALYSIS: AN OVERVIEW AND A HEALTH SERVICE APPLICATION

**Karyn MORRISSEY[*,**,***], Graham CLARKE[***], Stephen HYNES[*,**] and Cathal O'DONOGHUE[*,**]**

[*]Department of Economics, National University of Ireland, Galway, [**]Rural Economic Research Centre, Teagasc, Athenry, Republic of Ireland, [***]School of Geography, University of Leeds, Leeds, UK

## ABSTRACT

The aim of this chapter is to provide a comprehensive overview of accessibility analysis and its usefulness in an applied setting. Specifically, the aim of this chapter is to demonstrate the effectiveness of accessibility measures, in this case a spatial interaction model, in examining whether acute hospital services are optimally located, given the spatial distribution of long term illness (LTI) at the small area level. We do this by using a spatial microsimulation model to simulate micro-level LTI data at the small area, electoral division (ED) level. The simulated data on LTI are then inputted into a spatial interaction model to highlight areas with low acute hospital accessibility given their health status. The policy implications of these results are discussed in relation to both the health care literature and current Irish health care policy.

## KEYWORDS

Accessibility Analysis, Spatial Microsimulation, Health Service Provision, Long Term Illness, Ireland

## INTRODUCTION

Recent years have seen a renewed interest in a more integrated planning approach for service provision. Previously, government investment to improve access to public services had been prioritised either on the basis of the spatial distribution of services or on the availability of transport services. However, ease of access to a variety of services, such as retail, health and recreational services is increasingly recognised as an integral part of daily life and that by increasing individual level access to services other issues such as social exclusion, physical isolation and deprivation can be ameliorated. Thus, a more systematic approach to measuring accessibility would allow scarce public funding to be targeted more effectively at tackling those problems. As a result, the debate on accessibility now centres on a range of issues including:

- Identifying the accessibility needs of people and the impact of poor accessibility on their daily life
- The transport policy interventions which are required to improve rural accessibility
- The development of practical accessibility measures that can be used to aid expenditure decisions and prioritise initiatives

The aim of this chapter is to provide a review of accessibility analysis and to demonstrate its effectiveness in an applied setting, by modelling both demand for, and supply of acute hospital services in Ireland. In particular, interest is focused on estimating the accessibility of those individuals with high health care demands.

The first part of this chapter provides an overview of accessibility analysis to date and outlines the various methods available to examine access levels to services. A secondary aim in this chapter is to create a spatial micro-level dataset for the whole of Ireland, using spatial microsimulation. This dataset will contain a variety of demographic, socio-economic and health attributes for each individual in Ireland at the small area level. Finally, we integrate both methodologies to estimate access to acute hospitals in Ireland for those with highest demand for these services. Although the following analysis is possible at a more aggregate level, a micro-level approach is adopted here to ensure that the full range of demand (health status) and supply (service provision) determinants are available for each individual.

The rest of the chapter is structured as follows: section «Accessibility Analysis» presents an overview of accessibility indicators including both ratio-based indicators of access and simple model-based indicators derived from spatial interaction models. Section «Performance Indicators» provides an overview of performance indicators. Again, a range of access indicators from simple count-based indicators of access to model-based access indicators are outlined. Section «Creating a Spatial Profile of Health Status» gives an overview of the spatial microsimulation process and the data used to create SMILE, (Simulation Model for the Irish Local Economy) a spatial microsimulation model for Ireland. Section «Creating a Spatial Profile of Health Status» continues by outlining the calibration technique, alignment, used to calibrate the long term illness variable produced by SMILE. Section «Combining Spatial Interaction Models & Spatial Microsimulation» provides an outline of how a spatial microsimulation model and a spatial interaction model may be integrated to produce access scores for individuals with a long term illness at the small area level for acute hospital services. Section «Results» provides the results of this analysis. Finally, concluding comments in terms of policy recommendations are presented in section «Conclusions».

## ACCESSIBILITY ANALYSIS

There are numerous measures of accessibility and their application is widespread across different disciplines in both academia and policy analysis. These accessibility measures include; simple statistical ratios, service density estimation, facility coverage measures, minimum distance measures, travel cost measures (McLafferty, 2002; Talon, 2003) and more sophisticated model based measures based on; gravity/spatial interaction models (Fotheringham & O'Kelly, 1989; Clarke & Wilson, 1994). Measurement selection is one of the most important aspects of accessibility analysis and given that there are so many measurements one may use, one must be careful to choose the right measurement for the particular research question. This section will look at the different measures which may be used to analyse accessibility and the relative advantages and disadvantages of each measure of accessibility to service providers.

### Indicators of Accessibility

The most simple and least computationally intensive means of measuring accessibility are those based on container measures (the count of facilities within a given geographical area) or coverage measures (the number of facilities within a given distance from a point of origin). A simple container approach to accessibility analysis is the regional availability approach. This approach is usually calculated as the ratio of population to a service provider in a fixed area or the ratio of a certain age-group/socio-economic class to a provider in a fixed area. A clear advantage of using a container method to estimate accessibility to a given service is that it is easy to use and may take into account non-spatial attributes, such as age and socio-economic status. However, given the absence of formally defined and geographically self-contained catchment areas for many services, particularly health services, container measures of accessibility are unable to fully capture health service accessibility in an area (Congdon, 2001).

However, given the usefulness of container measures as a general measure of accessibility, Luo (2004) and Talbot (2000) both use a modified container measure in health care analysis; A floating catchment area approach (FCA) within a GIS is used to measure accessibility to physicians in Illinois and the distribution of low birth weight rates in New

York State. Using a GIS, spatially disaggregated population data and geo-coded service provider data are combined. Instead of using a predefined catchment area, as in the simple container measure of accessibility, buffers zones of varying sizes are created around service providers. The methodology involves creating a buffer zone centred on the centroid of the administrative boundary in use. Incorporating population and supply data (for example, the number of physicians within the predefined buffer zone) within the buffer zone, one can calculate the ratio of demand to supply. These steps are then repeated for the rest of the administrative boundaries of interest i.e. the catchments 'float over space. These buffer zone sizes are meant to represent a reasonable distance to which an individual would travel for varying levels of health care, and as such better represent the access individuals have to a service within a certain distance.

Although the FCA method of accessibility analysis overcomes some difficulties associated with cross-boundary flows (by extending the radius of the buffer outside the output boundary) what one considers a 'reasonable travel distance' is hugely subjective and depends on an individual's age, income, mobility status and location (especially urban versus rural) and the level of service they receive at the service point. For example, McLafferty (2003) points out that individuals are more willing to travel longer distances for specialised medical care as opposed to general primary care. Similarly, Birkin & Clarke (1991) point out that in a retail context larger retail units are more attractive and as such individuals are willing to travel much further to access the services they provide.

An enhancement to the FCA was introduced by Luo & Wang (2003) which is based on merging two GIS-based accessibility measures – FCA measures and gravity based measures - into one framework. The FCA method is enhanced by defining the service area of physicians based on a time threshold and incorporating a gravity-based method of accessibility to ensure that a nearby physician is more attractive than a remote physician. The addition of the gravity-based component, which includes an 'attractiveness' parameter for the service, means that the relative attractiveness of each health service, based on either size of facility or services offered, etc, may be incorporated into the model.

An alternative to the container methods of estimating access is Kernel Distance Estimation (KDE). KDE depicts the density of point events as spatially continuous variables, with 'peaks' on the map representing high service density and 'valleys' representing low service density (McLafferty & Grady, 2004). KDE is a well established means of estimating access to service providers in urban areas, where areas within a city/town with low service densities may be pinpointed. In contrast, KDE may not be an effective means of measuring access to services in rural areas where the density of service providers is much lower than urban areas for various social and economic reasons. Thus, access measurements for services in rural areas are more realistically measured in terms of distance rather than density.

More computationally intense measures of accessibility include minimum distance and travel cost measures. The minimum distance approach to accessibility analysis measures access as the distance from a certain point to the nearest facility, for example the nearest pharmacy to an individual resident. On the other hand, the travel cost approach calculates the distance (cost) between an origin and every destination possible within a certain area (Talon, 2003), such as the distance from an individual to every single pharmacy in an area. The measurement of distance between two spatial locations can be based on several different metrics, most notably; Euclidean distance and network distance. Access measures based on Euclidean or straight line distance are straightforward to calculate, but may not represent actual travel behaviour accurately (McLafferty, 2003; Talon, 2003). Indeed, as a measure of geographical access, Euclidean distance is flawed as it fails to incorporate the ease, cost and time of travel, and access to transportation (Martin et al., 2002). A second approach is to calculate distance/travel cost along an existing road network. To examine accessibility along road networks, many previous studies have used GIS to calculate network distances (Kalogirou & Foley, 2006; Alvanides & Gilmore, 2007) and travel times based on road type and quality (Martin et al., 2002).

However, these standard accessibility indicators are only a measure of the opportunities that are available to residents of a zone. They do not take account of which facilities are actually used by residents in each zone and are therefore unable to predict the level of interaction that occurs between residences zones and service locations? For rural service planning, a major question is what factors determine the magnitude of spatial interaction between residential areas and some facility location. To answer this question more sophisticated models of accessibility are required. In the next section, indicators based on gravity models or spatial interaction models will be introduced.

**Model-based Indicators**

Model based approaches to accessibility analysis include gravity models and spatial interaction models. Model-based approaches quantify accessibility to services according to resident location, service location, the 'attractiveness' of the service and the frictional effect of distance decay. There are four main factors that continually seem to influence the level of spatial interaction (Birkin & Clarke, 1991); the area's population size and the attributes of an area, service provider competition within an area, the distance or time taken to reach a location, and the relative 'attractiveness' of a service provider. Model-based indicators are considered to be the most realistic indicators of accessibility, given that they are based on the movements of consumers and take into account all of the factors above. There are two main types of model-based accessibility measurements: gravity models and spatial interaction models.

Gravity models are frequently used in urban and regional planning and are so-called because of their analogy to Newton's theory of gravity – two objects attract each other in direct proportion to their masses and the inverse of the distance separating them. In the same way, two zones in a city will attract interactions or movements between them given their relative size to each other and the distance between them. Knox (1978) states that as gravity models are specifically designed to handle distance-decay effects they are easily modified to measure accessibility between an origin and a destination point.

Hansen (1959) provides one of the simplest measures of accessibility (based on accessibility to employment):

$$A_i = \sum_j E_j \exp\left(-\beta c_{ij}\right) \tag{1}$$

where $A_i$ is accessibility in zone $i$ to employment $E_j$ in zone $j$, $c_{ij}$ is the cost of travelling from $i$ to $j$ and $\beta$ is a distance decay parameter. This simple model of employment accessibility embodies the crucial idea that accessibility is related to both distance (and cost) and the scale of opportunities at different locations. An alternative formula for accessibility is Schneider & Symons (1971) gravity model, outlined in Bertuglia et al. (1994):

$$AO_i = \sum_j \frac{Sj}{t_{ij}^{\beta}} \tag{2}$$

where:

$AO_i$ - is an index of access opportunity

$S$ - is the size of some facility $j$

$t$ - is the time taken to travel from $i$ to $j$ and

$\beta$ - is the distance decay parameter.

This measure can be weighted according to some other criteria, such as the rate of car ownership or income level to determine an individual's (or a community's) accessibility to a given service facility. Knox (1978) further modified the gravity model by taking the Schneider and Symons model and incorporating parameters based on travel speeds (by car and by public transport) to transform the $AO_i$ measure to:

$$TA_i = C_i * \Sigma_j (\frac{A_j}{S_a}) + (100 - C_i) * \Sigma_j (\frac{A_j}{S_t}) \qquad (3)$$

Where

$TA_i$ - is the new index of accessibility for zone $i$

$A$ – is the access score in zones $i$ and $j$

$C_i$ - is the percentage of car-owning households in zone $i$, and

$S_a$ and $S_t$ - are the average times taken to travel a distance by car and public transport respectively.

By including a time based measure of access to various destinations, Knox (1978) states that the 'index provides a reasonably sensitive yet robust indicator of the accessibility of different localities to a given spatial distribution of facilities'. However, as we will see in the next section, although gravity models are a reasonable measure of access to services, they are unable to incorporate the kind of behaviour that is necessary for a full accessibility analysis.

Spatial interaction models (SIM) build on the gravity model concept and involve determining through demand, supply and interaction information the attributes that promote flows of people and goods between different locations. There are four types of SIMs currently in use. These models include:

- Destination-Constrained SIMs which assume that the attributes of the supply point are known, i.e. the location of various service providers (supply points).

- Origin-Constrained SIMs which assume that the attributes of the demand point are known, i.e. the location of households (demand points).
- Doubly Constrained models which assume that both the supply or demand point attributes are known.
- Unconstrained models which assume that neither demand nor supply attributes are known.

Most accessibility analysis has traditionally used origin-constrained SIMs. An example of an origin-constrained spatial interaction model is the following model used to estimate shopping flows (Clarke et al., 2002)

$$S_{ij}^{m} = O_i^m A_i^m W_j^{\alpha^m} \exp(-\beta^m d_{ij}) \tag{4}$$

where:

$S_{ij}^{m}$ - is expenditure by household type *m* in residence zone *i* at destination *j*

$O_i^m$ - is the level of consumer expenditure of household type *m* in zone *i* at destination *j*

$A_i^m$ – is a balancing factor that ensures that:

$$\Sigma_j S_{ij}^{m} = O_i^m \tag{5}$$

where $A_i^m$ is calculated as:

$$A_i^m = \frac{1}{\Sigma_j W_j^{\alpha^m} \exp(-\beta^m d_{ij})} \tag{6}$$

$W_j$ – is the attractiveness of the service facility to residents in ED *j*

$\beta^m$ – is the distance decay parameter

$\alpha^m$ – is a parameter reflecting the attractiveness of the destination by household type *m*

$d_{ij}$ – is the distance metric from the origin, *i* to the destination, *j*

The demand side (zone *i*) is usually represented as households aggregated to the smallest geographical output level for which there is data. The supply element of the spatial interaction model represents service locations. Set against the important effect of distance decay is the 'attractiveness' of a particular facility. Indeed, the role of attractiveness is one of counterbalancing the disutility of distance. The attractiveness of an opportunity should be measured based on what characteristics of a potential destination are important to the consumer (Liu & Zhu, 2004). For example it can be measured as the number of retail outlets at a destination, or the physical size of a retail outlet at a destination. In reality the attractiveness of a facility often translates as its physical size (Birkin & Clarke, 1991). Birkin & Clarke show that the distance that people are willing to travel to retail centres varies according to the size of the outlet (in West Yorkshire). The distances travelled were compared with the number of outlets in the city centre. Unsurprisingly, it was found that distance travelled increased with the size of the retail outlet.

The interaction component of the SIM – the distance between the origin and destination - may be calculated in a number of ways, such as Euclidean distance, travel time, or cost analysis. Once a SIM has been developed and calibrated it is possible to predict flows from residential areas to individual facilities.

## PERFORMANCE INDICATORS

Similar to the indicators outlined earlier, performance indicators were historically defined in relation to individual spatial areas which were unconnected to other spatial areas. Obviously this definition of space is fundamentally flawed as people interact with a variety of different locations during the day. As such, a number of performance indicators have been developed that take into account inter-area flows, thus providing a more realistic account of individual interactions with service facilities. The development of accessibility indicators from simple provision indicators to model-based accessibility indicators will be outlined below.

### Simple Provision Indicators

The simplest performance indicator with regard to service provision is a non-model based indicator, based on population/service size ratios. For example, in relation to retail accessibility, one could take retail grocery square feet per household in a geographically defined area. However, the fundamental drawback of this indictor is its inability to account for inter-area flows of consumers. Thus, although many areas may not have a retail store in its zone, this may be because there is a large retail outlet in its neighbouring area. As such, the residents in these zones may not be underserved or have poor access to the facility, especially when using small census tracts. Thus, simple performance indicators are limited in their applicability to real-life scenarios.

The development of GIS has greatly aided accessibility analysis and there are now a number of methods of calculating accessibility based indicators within GIS. Following Clarke et al. (2002) two will be outlined here. Clarke et al. (2002) outline four different procedures for analysing food provision in Cardiff. The first component of this analysis involved using a GIS to map a circle of 500m around each retail grocery outlet (500m is considered to be an appropriate measure of walking distance). Inhabitants living outside these radii may form potential candidates for low accessibility (for this analysis, individuals living in so called food deserts). Next they restricted the search for food deserts to areas with high deprivation, in the belief that food deserts were only a problem if individuals had poor mobility and incomes. However, although this method of accessibility analysis takes into account deprivation and mobility, the full range of shopping opportunities is not taken into account. This analysis assumes (as is often not the case) that individuals use the facility closest to them.

The second performance indicator outlined by Clarke et al. (2002), is a Hansen type accessibility indicator as introduced in section «Accessibility Analysis». This accessibility measure reflects the opportunities to use services that are available at all destinations and is related to the distance (from the origin to the destination) and scale of provision (size etc, of the service facility) and may be calculated as follows:

$$\alpha_i = \sum_j W_j \exp(-\beta^m d_{ij}) \tag{7}$$

This indicator is an improvement on simple accessibility measures, as it allows for all retail facilities to be considered (thus no area has zero provision). Areas with a high number of local services will be given the highest access score. Furthermore, this indicator can be disaggregated by person type, therefore highlighting differences in access levels across age, socio-economic, income, type, etc. However, Clarke et al., point out that this indicator is still only a measure of service opportunities and it is unable to take into account the actual facilities *used* by individuals in an area.

**Model-based Performance Indicators**

Model-based performance indicators are based on the predicted levels of interaction as calculated by a spatial interaction model (SIM). Thus, these models quantify accessibility according to where individuals shop as predicted by the SIM and as such provide a more realistic representation of access based on the movements of consumers rather than simply on the geographical distribution of service outlets in a zone. In this section, two model-based performance indicators are introduced. These indicators were first introduced by Clarke & Wilson in Bertuglia et al. (1994).

The first type of performance indicator relates to individuals and households. These indicators are based on residential location and relate to the ways in which the individual is served by facilities. As such these indicators can be used to estimate the effectiveness of service provision to individuals. The second type of performance indicator relates to service providers and the specific services they offer. Thus, these indicators measure the efficiency of provision by service providers. By identifying spatial variations in effectiveness and efficiency of provision, performance indicators allow the targeting of resources to increase the efficiency of a service facility or to increase the effectiveness of service provision to households. Using both indicators to analyse spatial interaction provides an understanding of how households and facilities are interdependent. Thus, these indicators allow planners to examine whether the problem is in the effectiveness of delivery to residential location or whether the problem is in the efficiency of provision at facility locations.

The following two model-based indicators measure the effectiveness of service delivery to residential areas. As such they measure the *aggregate level of provision and the level of provision per household* respectively. These indicators are important because the effectiveness of a provider in delivering its services is based on its size and location.

The aggregate level of provision in an area is given as:

$$w_i^m = \Sigma_j \frac{S_{ij}^m}{S_{\bullet j}^m} W_j \tag{8}$$

The above equation for estimating the aggregate level of provision in an area is calculated by dividing each SIM output (equation 5) by the sum of all outputs for each zone *j,* where • indicates summation across all zones *i.* This is then multiplied by the attractiveness of outlets in zone *j*. The sum of all these values for residence zone *i* provides the aggregate provision for each zone *i.* This indicator ensures that even if an area does not have a service facility, the area will not have a zero accessibility flow (like the Hansen-type indicators outlined above).

In order to identify areas with low service provision it is necessary to relate this level of provision to the number of households in that area. For example, if service provision for a particular area is low, but population is also low, then the area may not be classified as a problem zone. On the other hand, if an area has relatively low provision and population is high then the model outputs will establish this as a problem area. Also, because of the nature of this performance indicator, it is possible that an area with high provision and high population can appear to be relatively poorly served because the indicator is a measure of the share of a facility that a residence area has (Clarke et al., 2002). Relating this aggregate provision indicator to population in an area will allow the identification of areas where a significant number of households suffer poor accessibility to a particular service. Level of provision per household is an indicator that divides the aggregate level of provision by the number of households in the residence zone, as follows:

$$v_i^m = \frac{w_i^m}{H_i^m} \tag{9}$$

Similarly, a catchment population indicator can be calculated as follows:

$$C_i = \sum_i \frac{S_{ij}}{S_i} * P_i \tag{10}$$

Where $S_i$ is expenditure in area *i*, $S_{ij}$ is expenditure from area *i*, in area *j* and $P_i$ is the population in area *i*. In equation (7), a typical term on the right-hand side involves taking the proportion of provision at *j* which is used by residents of *i* and then summing to obtain a measure of total provision for residents of *i*. Clarke and Wilson state that similarly, equation (8) represents partitions of the residential population which are combined to form a catchment population for the centre (or outlet) at *j*. One can also calculate a version of this in terms of demand, which we can call $\Delta_j$:

$$\Delta_j = \sum j \left( \frac{S_{ij}}{S_i} \right) \times W_j \tag{11}$$

Where $S_i$ is expenditure in area *i*, $S_{ij}$ is expenditure from area *i*, in area *j*. Other typical indicators are $\frac{W_i}{P_i}$ for effectiveness and $\frac{W_i}{C_j}$ for efficiency, where $W_i$ is the aggregate level of provision in zone *i*, $P_j$ is the population in zone *j* and $C_j$ is the catchment of zone *j*. These indicators are then used when the population is disaggregated by type m (social class, car owner, etc) and provision is disaggregated by type of good, *g*. Thus, in relation to policy and planning, the key when assessing accessibility to services in rural areas is to ensure that the spatial distributions of provision and demand are such that both sets of effectiveness and efficiency indicators achieve appropriate targets.

The goal of this section was to provide an overview of accessibility analysis and the methodologies that may be employed to estimate levels of access to services. As outlined above the second aim of this chapter is to create a spatially disaggregated micro-dataset for the whole of Ireland using spatial microsimulation. This dataset may be used for input into the SIMs. The next section provides an overview of SMILE (Simulation Model of the Irish Local Economy) a spatial microsimulation model.

## CREATING A SPATIAL PROFILE OF HEALTH STATUS

Detailed spatial profiles of an individual’s health status would be extremely valuable to both government and non-government organisations in their efforts to target scarce health care resources at the individuals who need them most. However, the construction of detailed spatial profiles of individual health status has been hampered by the non-availability of health data at the small area level. One problem is that household surveys (such as the Living in Ireland (LII) survey), which may include data on individuals’ health is that they are aspatial in nature. Census data, although spatially disaggregated, generally does not include any health data. Thus, if we could merge the data in the LLI with the ED census level data we would have a much richer dataset that would allow us to investigate an individual’s health at a very fine level of spatial resolution. We use spatial microsimulation techniques to accomplish this.

Spatial microsimulation is a means of synthetically creating large-scale micro-datasets at different geographical scales. The development and application of spatial microsimulation models offers considerable scope and potential to analyse the individual composition of an area so that specific policies may be directed to areas with the greatest need for that policy. SMILE is a static spatial microsimulation model that uses a combinational optimisation technique, simulated annealing to match the 2000 LII survey to the 2002 Small Area Population Statistics (SAPS). This provides a micro-level synthetic dataset for the whole population of Ireland (for a full discussion on the datasets and algorithm used to create the statistical match, see Morrissey et al., 2008). The dataset created by SMILE contains the usual demographic, socio-economic, labour force and income variables for both individuals and family units. However, unlike other microsimulation models, SMILE also contains a farm (Hynes et al., 2009) and health component (Morrissey et al., 2008). Table 1 contains the most important health status drivers found in the matched SMILE dataset, of which the most important for this analysis is the long term illness (LTI) variable.

| | LII | Census | Status |
|---|---|---|---|
| **Demographic Profile** | | | |
| Sex | Yes | Cross-tabulated with Age | Matched Variable |
| Martial Status | Yes | Yes | Non-Matched Variable |
| Household Size | Yes | Yes | Matched Variable |
| Number of Children | Yes | Yes | Non-Matched Variable |
| Household Income | Yes | No | Non-Matched Variable |
| **Socio-Economic Profile** | | | |
| Household Income | Yes | No | Non-Matched Variable |
| Education Level | Yes | Yes | Matched Variable |
| Employment Status | Yes | Yes | Non-Matched Variable |
| Occupation | Yes | Yes | Non-Matched Variable |
| **Health Status Profile** | | | |
| Health Status (5 category dummy variable) | Yes | No | Non-Matched Variable |
| Long Term Illness | Yes | No | Non-Matched Variable |
| Physically/Emotionally Hampered | Yes | No | Non-Matched Variable |
| **Utilisation Variables** | | | |
| GP utilisation | Yes | No | Non-Matched Variable |
| Hospital In-patient Utilisation | Yes | No | Non-Matched Variable |
| Nights spent as an in-patient | Yes | No | Non-Matched Variable |
| Medical Specialist Consultant | Yes | No | Non-Matched Variable |
| **Coverage Status** | | | |
| Private Health Insurance Cover | Yes | No | Non-Matched Variable |
| Medical Card Cover | Yes | No | Non-Matched Variable |
| **Risk Factors** | | | |
| Smoker or not | Yes | No | Non-Matched Variable |

Table 1: The main Demographic, Socio-economic and Health variables in SMILE

As mentioned above, spatial microsimulation is a method used to create spatially disaggregated microdata that previously did not exist. Thus, validation of the newly created data is essential, but difficult. SMILE contains a number of internal validation methods, namely z-scores, $z^2$-scores and total absolute measure scores (Hynes et al., 2009). However, a serious weakness of internal validation is that it fails to compare the fitted model with ‘fresh’ external data. External validation arises when the newly created variables in SMILE are compared to similar data that is not used in the original match. With regard to our data, the simulated ‘long term illness’ variable is validated against the 2000, weighted LII Survey. When weighted, the LII is a representative dataset for the whole of Ireland. Validating at the NUTS3 spatial level, it was found that the rates of individuals with a LTI were underestimated in the simulated SMILE dataset compared to the weighted LII survey (Table 2).

The discrepancy between SMILE and the weighted LII dataset at the regional level is not surprising. As reported in Table 1, SMILE’s health component variables are a result of matching the health variables in the LII survey with the SAPS dataset. As with all microdata surveys, there are issues of data quality with the LII dataset (Stuggard, 1996). Thus, given the internal bias contained in SMILE (and according to Baekgaard (2002), all microsimulation models), it was decided to calibrate the LTI variable through a process called alignment (Caldwell & Keister, 1996; Morrison, 2006).

| Regions | Weighted LII | SMILE |
|---|---|---|
| Border | 17% | 20% |
| Dublin | 13% | 22% |
| Mid-East | 15% | 21% |
| Midlands | 15% | 21.5% |
| Mid-West | 16% | 21% |
| South-East | 15% | 20% |
| South-West | 17% | 23% |
| West | 18% | 23% |

Table 2: Comparison of NUTS3 breakdown of LTI suffers as reported by SMILE and the Weighted LII Survey

The objective of calibrating a spatial microsimulation model is to ensure that the simulated output matches exogenous totals at varying levels of spatial disaggregation (Baekgaard, 2002). It may be argued that the need to calibrate simulated data exposes the weaknesses of the simulation process. If the simulation process does not reach its desired target is the model not defective? However in response, the output from a microsimulation model is only as reliable as the original datasets that were used to create the synthetic dataset (Baekgaard, 2002). As outlined above, datasets are prone to a number of errors which arise due to sampling error, data collection error and data processing error. Thus, it may be argued that due to limitations in the original datasets – missing observations, erroneous coding, respondent bias etc, that ex-post calibration is necessary.

Alignment is a technique that may be used to help calibrate the data produced from a spatial microsimulation model. The alignment process may be broken down into a 5 step procedure (Morrissey et al., 2009):

- A logit model is used to estimate the probability of suffering from a LTI for each individual in the dataset.
- From these estimates, a residual is created (a stochastic term)
- An exogenous alignment total is specified (in this case, the total number of individuals with a long term illness in the weighted LII survey)
- Each residual is ranked from highest to lowest
- If the created ranked residual of each person is greater than the logit model estimates for that person then the status of individuals with the highest probability of having a LTI is changed until the exogenously specified totals of suffers with an LTI for each region is complete.

On completion, the alignment process provides a fully calibrated LTI variable which may be used to create a spatial profile of LTI suffers at the small area level for Ireland. Please see Morrissey et al., 2009 for a full discussion of the alignment procedure. This dataset, when combined with a spatial interaction model, may be used to estimate levels of access to acute hospitals for individuals with high health care needs.

## COMBINING SPATIAL INTERACTION MODELS & SPATIAL MICROSIMULATION

In this paper the SIM described in section «Performance Indicators» is used to measure access scores from each ED to their nearest acute hospital. For the purpose of this paper the attractiveness parameter for each acute hospital, $W_j$ is the number of consultants in each acute hospital (a measure of how easy it is to be examined quickly), the number of beds in each hospital (a measure of the size of each hospital), whether the hospital has an A&E facility, a maternity facility, and a psychiatric facility. Each hospital is given one point for each bed and consultant specialist they have, twenty points if they have an A&E department and 10 points if they have a maternity and psychiatric unit. These 'points' are then added together to form the $W_j$ term of the SIM.

The demand variable, $O_i$ is the number of individuals in each ED with a LTI as determined by SMILE's aligning process. The distance variable, $d_{ij}$ is the distance from each ED centroid (*i*), to acute hospital *(j)*. $d_{ij}$ was calculated using time cost analysis (using the current road data for Ireland) in ArcGIS. Using travel time, rather than distance ensures that an up-ward bias is not created for urban areas. Although individuals living in urban areas may be closer to hospitals in terms of distance alone, travel time in built-up, congested urban areas is usually longer over the same distance than in rural areas. The use of travel-time rather than distance in the SIM ensures that each ED is assigned a more realistic travel cost parameter. Using the travel cost function in ArcGIS, the national legal speed limit was applied to each road classification. This ensures that the travel time along each road segment is calculated in a more realistic manner, thus providing a more realistic travel time for each ED to an acute hospital. Combining the data on individuals with a LTI contained in SMILE and the access score from the SIM highlights areas with low acute hospital accessibility given their health service needs.

## RESULTS

As outlined in section «Creating a Spatial Profile of Health Status», an alignment procedure was used to calibrate the LTI variable contained in SMILE. Figure 1 presents the mean percentage of individuals with ill-health in the original, SMILE dataset. As one can see, there is no obvious pattern to the number of individuals with an illness across space. This is hugely surprising given the demographic and socio-economic distribution of Ireland (the West and the North-West of Ireland have disproportionately more individuals over 65) and the link between individual's age and health status.

In contrast, Figure 2 presents the aligned mean percentage of individuals with ill-health, now reported by SMILE. As one can see, there is a definite West versus East divide with regard to ill-health. The West coast of the country from North to South has a much higher rate of ill-health than its Eastern counterpart. This spatial patterning of ill-health is much more consistent than the original matched pattern given with the demographic and socio-economic profile of the two coasts.

From the alignment of the LTI variable we find that similar to research in the UK (Mitchell et al., 2002) an individual's spatial location may be a strong indicator of personal health. Studies have found individuals with similar risk profiles cluster in the same location and therefore increase (or decrease) the risk of mortality and specific morbidities across space (Dorling, 1997; Mitchell et al., 2002).

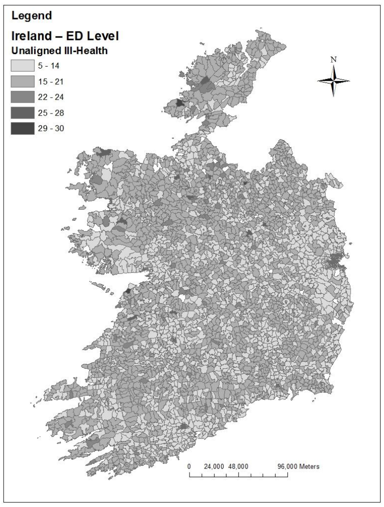


Figure 1: Original Simulated LTI at the ED level (Source: SMILE)

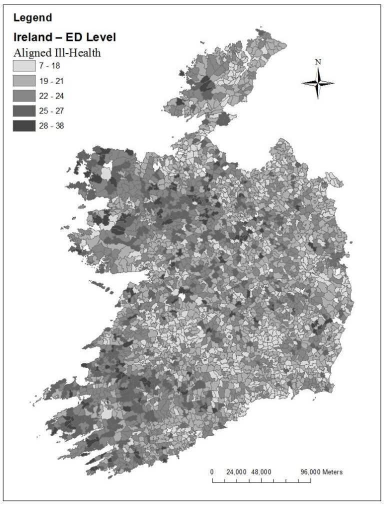


Figure 2: Map of the Average Calibrated LTI at the ED level (Source: SMILE)

Using the two performance indicators specified in section «Model-based Performance Indicators», the access scores for EDs in relation to acute hospital services were calculated for Ireland. The access score for each

ED in Ireland ranged from 0.0001064 (indicating very poor access) in the north west of the country to 3515.3 (indicating very good access) in the north east of the county, while the average access score across the 238 EDs was 72.59. The accessibility scores for each ED with regard to acute hospital services are presented in Figure 3.

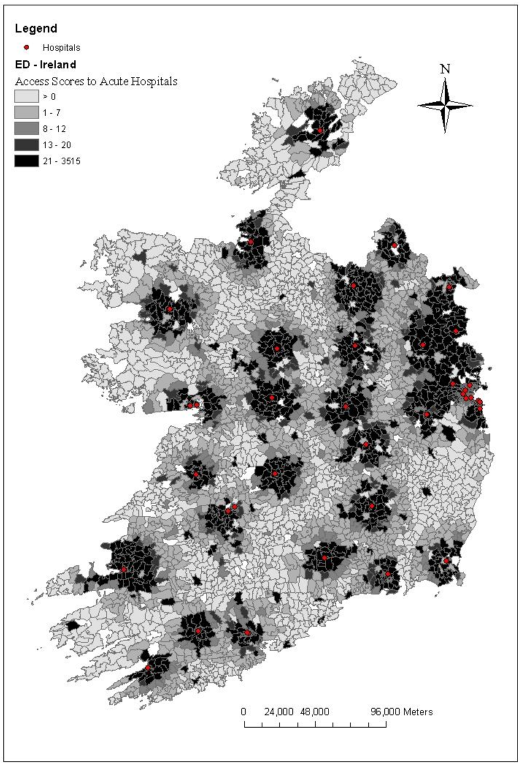


Figure 3: Access Scores for each ED to an acute hospital service in Ireland as calculated by the Accessibility Indicators

It is interesting to note that although the Dublin area has by far the highest number of hospitals, using travel time rather than distance as the travel cost parameter meant that Dublin city EDs did not necessarily have high access scores. Indeed, the average access score for the Dublin area was 202.16, compared to the highest access score (3515.3) in the north east of the country.

Finally, we compare the access indicators scores from the spatial interaction model for Ireland to the weighted mean of residents with a LTI for each ED. A negative relationship between EDs with high rates of residents with LTI and low access to acute hospital services is confirmed in Figure 4. Figure 4 presents a bar graph comparing the weighted mean probability of residents with a LTI for each ED with the mean accessibility scores for each ED.

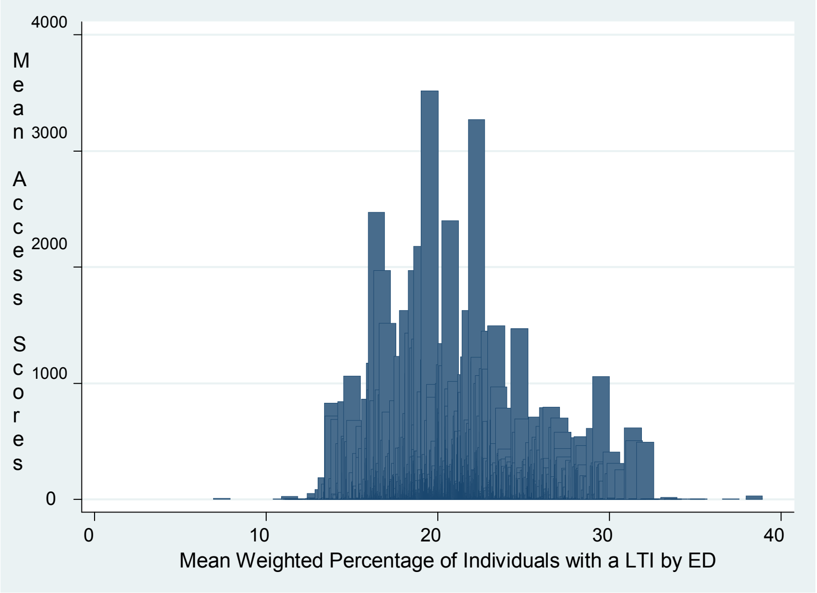


Figure 4: Comparison of the Mean Weighted Percentage of Individuals with a LTI for each ED with the mean accessibility scores for each ED

As one can see from the bar graph, EDs with the highest percentages with a LTI (approximately 32% to 40% of their residents have a LTI) actually have the lowest accessibility to acute hospital services. Conversely, one can see from the bar graph that the EDs with the lowest percentage of individuals with a LTI (between 10% and 24%) have the highest accessibility to acute hospital services.

## CONCLUSIONS

The primary aim of this chapter has been to present an overview of accessibility analysis to date and outline the various methods available to examine access levels to services. There are numerous measures of accessibility and to this end section «Accessibility Analysis» and section «Performance Indicators» present a diverse range of accessibility measurements from simple statistical ratios to more sophisticated model based measures using gravity and spatial interaction models.

A secondary aim of this chapter was to present a methodology and a subsequent case study of such indicators in operation for examining access to acute hospital services for those with an LTI. To address such a research question it was first necessary to create a spatial micro-level dataset for the whole of Ireland, using spatial microsimulation. As such, section «Creating a Spatial Profile of Health Status» gives an outline of SMILE, a spatial microsimulation for Ireland. Using a process called alignment SMILE's LTI variable is calibrated to match exogenous totals from the nationally weighted LII. This process provides us with the first spatial map of LTI for Ireland. Such a map may be of great benefit to health service policy-makers and planners for future planning and resource allocation.

Furthermore, the alignment of the LTI variable across space also provides the first clear indication that individuals' with similar health status cluster in similar locations in Ireland. These results are consistent with previous work in the UK (Mitchell et al., 2002; Dorling, 1997) and therefore provide further evidence that an individuals' 'place' or location may have a strong influence on health status.

Finally, we integrate both a SIM and the micro-data from SMILE to estimate access to acute hospitals in Ireland for those with highest demand for these services. As with Morrissey et al. (2008) and their analysis of GP services for County Galway, in Ireland, our analysis found that despite increases in health spending (Wiley, 1998) inequalities in both health status and access to health services not only exist, but also show huge is a polarisation between urban and rural communities. As such, these results are perhaps unsurprising in the Irish context. Indeed, Tudor Hart on examining the relationship between health care need and health care provision in Wales found that the 'availability of good medical care tended to vary inversely with the need for it in the population served' (Hart, 1971). Our results point to the continued existence of this inverse relationship between health care need and health care provision.

However, there are limitations to this analysis. For example, in Ireland, as in the UK, individuals that attend an acute hospital are often referred by their GP. Therefore further analysis of access to acute hospitals should incorporate the trip from the individual's residence to the GP surgery to the acute hospital. However, currently no data in Ireland exists as to the percentage of individuals who are referred by GPs or self-attendees to acute hospitals. Furthermore, this study has determined that LTI exist in clusters across Ireland. However, it is also necessary to determine why these clusters exist. Is it due to clusters of individuals with similar characteristics, i.e. individuals in these areas have a high risk profile, or is it due to some exogenous, contextual effect? Therefore, further research in this area will involve determining the underlying influences that drive these clusters of ill-health.

Limitations aside, this chapter demonstrates that accessibility analysis (in this case a SIM), when combined with a spatial microsimulation model, can indicate areas with high health care demand and poor health care access. These results in turn can be used to target health care resources in areas with the highest demand and lowest access and therefore optimise government intervention and public resources in a more effective manner.

## REFERENCES

**Alvanides S.; Gilmore, W.,** 2007, Location-allocation for predicting patient flows from the closure of the obstetrics and neonatal Services. In: Applied GIS, 3,1

**Baekgaard, H**., 2002, Micro-macro Linkages & the Alignment of Transition Processes. In: Technical Paper, 25. National Centre for Social and Economic Modelling, University of Canberra

**Clarke, G.P.; Wilson, A.G.**, 1994, A new geography of performance indicators for urban planning. In: Bertugli, C.S. (Ed.), Modelling the City: Performance, policy and Planning. Routledge: London

**Birkin, M.; Clarke G.**, 1991, Spatial Interaction in Geography. In: Geography Review, 4: 16-24

**Caldwell, S.; Keister, L**., 1996, Wealth in America: family stock ownership and accumulation, 1960-1995. In: Clarke, G. (Ed.), Microsimulation for Urban and Regional Policy Analysis. London: Pion

**Clarke, G.P.,** (Ed.), 1996, Microsimulation for Urban and Regional Policy Analysis. Pion: London

**Clarke, G.; Eyre, H.; Guy, C**., 2002, Deriving Indicators of Access to Food Retail Provision in British Cities: Studies of Cardiff, Leeds and Bradford. In: Urban Studies, 39, 11: 2041-2060

**Congdon, P**., 2001, The development of gravity models for hospital patient flows under system change: a Bayesian modelling approach. In: Health Care Management Science, 4: 289-304

**Fotheringham, A.S.; O'Kelly, M.E.,** 1989, Spatial interaction models: formulations and applications. Kluwer Academic Publishers: Boston London

**Hansen, W.G.**, 1996, How Accessibility Shapes Land Use. In: Journal of American Institute of Planners, 25: 73-76

**Hart, T.**, 1971, The Inverse Care Law. In: The Lancet, 1, 7696: 405-412

**Hynes, S.; Morrissey, K.; O'Donoghue, C.; Clarke, G.**, 2009, Building a Static Farm Level Spatial Microsimulation Model for Rural Development and Agricultural Policy Analysis in Ireland. In: International Journal of Agricultural Resources, Governance and Ecology, 8: 282-299

**Kalogirou, S.; Foley, R.,** 2006, Health, Place and Hanly: Modelling Accessibility to Hospitals in Ireland. In: Irish Geography, 39, 1: 52-68

**Knox, P.**, 1978, The interurban ecology of primary medical care: patterns of accessibility and their policy implications. In: Environment and Planning A, 10, 4: 415-435

**Luo, W.,** 2004, Using a GIS-based floating catchment method to assess areas with shortage of physicians. In: Health & Place, 10: 1-11

**Liu, S.; Zhu, X.**, An Integrated GIS Approach to Accessibility Analysis Transactions in GIS, 8, 1; 45-62

**Martin, D.; Wrigley, H.; Barnett, S.; Roderick, P.,** 2002, Increasing the sophistication of access measurement in a rural healthcare study. In: Health and Place, 8: 3-13

**McLafferty, S.L.**, 2003, GIS and Health Care. In: Annual Review Public Health, 24: 25-42

**Mitchell, R.; Dorling, D.; Shaw, M.**, 2002, Population production and modelling mortality – an application of geographic information systems in health inequalities research. In: Health and Place, 8, 1:15-24

**Morrison, R.**, 2006, Make it so: Event Alignment in Dynamic Microsimulation. DYNACAN Team, Ottawa, Canada

**Morrissey, K.; Hynes, S.; Clarke, G.P.; O'Donoghue, C.; Ballas, D.,** 2008, Examining Access to GP Services in Rural Ireland using Microsimulation Analysis. In: Area, 40, 3: 354-464

**Morrissey, K.; Clarke, G.; O'Donoghue, C.,** 2009, The Spatial Pattern of Health Service Utilisation in Ireland. Working Paper 09WPRE03. RERC, NUI, Galway

**Stuggard, N.**, 1996, Reconciliation of UK Household Income Statistics with the National Accounts. Paper presented at the Expert Group on Household Income Statistics, Canberra, Australia

**Talon, E.**, 2003, Neighborhoods as service providers: a methodology for evaluating pedestrian access. In: Environment and Planning B: Planning and Design, 30: 181-200

**Wiley, M.,** 1998, Health Expenditure Trends in Ireland: Past, Present and Future. In: Leahy, A.; Wiley M. (Eds.), The Irish Health System in the 21st Century. Oak Tree Press: Ireland

## AUTHORS INFORMATION

**Karyn MORRISSEY**
karyn.morrissey@teagasc.ie
National University of Ireland, Galway, Department of Economics; Rural Economic Research Centre, Teagasc, Athenry, Republic of Ireland; University of Leeds, School of Geography, Woodhouse Lane, Leeds LS2 9JT, UK

**Cathal O'DONOGHUE**
cathal.odonoghue@nuigalway.ie
National University of Ireland, Galway, Department of Economics; Rural Economic Research Centre, Teagasc, Athenry, Republic of Ireland

**Graham CLARKE**
g.p.clarke@leeds.ac.uk
University of Leeds, School of Geography, Woodhouse Lane, Leeds LS2 9JT, UK

**Stephen HYNES**
stephen.hynes@nuigalway.ie
National University of Ireland, Galway, Department of Economics; Rural Economic Research Centre, Teagasc, Athenry, Republic of Ireland

# CHOICE SET GENERATION IN SPATIAL CONTEXTS

## A RETROSPECTIVE REVIEW

**Francesca PAGLIARA**[*] **and Harry J.P. TIMMERMANS**[**]

[*]University of Naples Federico II, Italy, [**]Eindhoven University of Technology, the Netherlands

### ABSTRACT

The importance of properly specifying choice sets to avoid biased parameters is well recognized in the literature and it is particularly relevant for spatial choice models where alternatives are generally numerous and somewhat artificially defined (i.e. traffic zones). The choice set refers to the set of discrete alternatives considered by an individual in the decision-making process which is a subset of the universal choice set that consists of all alternatives available to the decision-maker. The objective of this paper is to review the various approaches to choice set formation present in the literature. Strengths and weaknesses of each of these approaches are discussed.

### KEYWORDS

Choice set, Spatial choice, Deterministic approach, Behavioural approach, Dominance variables

### INTRODUCTION

The importance of properly specifying choice sets to avoid biased model parameters has been well recognized in the literature (Pellegrini et al., 1997; Dai, 1998). This topic is particularly relevant for spatial choice models where alternatives are generally numerous and somewhat artificially defined (i.e. traffic zones). The choice set refers to the set of discrete alternatives considered by an individual in the decision-making process which is a subset of the universal choice set that consists of all alternatives available to the decision-maker (Nedugadi, 1987). A number of approaches to define choice sets or develop choice set generation procedures have been pursued, mostly in the marketing and transportation literature. Consequently, the majority of empirical tests of these approaches are in the non-spatial domain.

The term choice set has been defined in different ways and should be differentiated from other related concepts. For example, the awareness or knowledge set consists of the subset of choice alternatives in the universal set that a decision-maker knows exists and could therefore potentially be selected (Schocker et al., 1991). It is from the awareness set that the consideration set evolves. A consideration set can be viewed as consisting of those alternatives salient or accessible on a particular choice occasion. While an individual may have knowledge of a large number of alternatives, it is likely that only a few of these will come to mind when a choice has to be made. A choice set then is the set of alternatives that is very seriously considered just before a final choice is made. Thus, these related concepts and definitions assume some kind of hierarchical decision making process during which the set of alternatives considered is gradually reduced. Before discussing the various approaches to modelling choice set generation and composition, and place these in a wider context, it may be relevant to reflect on these definitions. First, these definitions imply subtle differences between these concepts, and as we will see later in this paper, virtually none of the modelling approaches has been really successful in making such subtle differences, especially because most rely on ad hoc, rather arbitrary data collection. In fact, actual procedures of choice set construction are often not specified to elicit choice sets in this strict sense. Secondly, one should realize that the implicitly or explicitly assumed hierarchical process may not be the most relevant one for some types of spatial choice problems. For example, residential choice processes often tend to be more sequential. Individual and household often have no knowledge about available alternatives in the market and thus become involved in active search. Screening information sources (websites, newspaper, etc), they typically encounter a property that they like, which triggers a process of inspection and negotiation. If positive, this may lead to a bid and perhaps successful buy; if not, the search process continues. Similarly, for shopping centre choice, individuals only know a subset of stores and shopping centres in their region, but shopping behaviour can be better described as consisting of a context-dependent repertoire of routine shopping strategies developed through a learning process in interaction with the environment. In other words, we argue that for some choice problems, the concept of a choice set may be a plausible process model, but for many other choice problems, it is better to view the concept as an ad hoc construction of the researcher that does not necessarily reflect any behavioural realism.

The usual approaches to choice set formation in practice are (Ortuzar & Willumsen, 2001):

- The use of heuristic or deterministic choice set generation rules which allow one to exclude certain alternatives, validated using data from the sample.

- The use of choice set information directly from the sample, simply by asking respondents about their perception of available options.

The use of random choice sets, whereby choice probabilities are considered to be the result of a two-stage process: firstly, a choice-set generating process, in which the probability distribution function across all possible choice sets is defined; and secondly, conditional on the specific choice set, the definition of the choice probability of each alternative in the choice set (Lerman, 1984; Richardson, 1982).

Because choice sets are not directly observable, choice set formation is usually simplified by using ad hoc assumptions about the size, composition and interpersonal variability of the choice set of alternatives. The traditional approaches typically are based on a restricted list of deterministic criteria selected by the analyst. For instance, Miller and O'Kelly (1983) assumed that the choice set of a certain individual is composed of all destinations actually chosen by all individuals from the same geographic area. Theoretically, this operational decision is highly problematic because it not only denies that choice sets of the same individual may vary by purpose, time of day and other context variables, but it also disregards any heterogeneity as all individuals are unrealistically assumed to share an identical choice set. Gautschi (1981) restricted the set of feasible locations for major non-grocery shopping in suburban San Francisco to four major retail centers. The criterion used is that these centers generate the bulk of the non-grocery retail business in the study area. Smaller centers are discarded as potential destinations for computational reasons. This is obviously also a rather ad hoc and unrealistic assumption. The method suggested by Black (1984) represents another example. It sets the perimeter of the choice set and all destinations within this perimeter are assumed to belong to the choice set. This approach suffers from the same flaws.

The implications of an invalid measurement of choice sets depend on the situation. Most spatial choice approaches assume that the odds of choosing a particular alternative are independent of the composition of the choice (IIA property). If indeed this assumption is valid and choice set composition has no effect on pairwise choice probabilities, McFadden (1978) proved that random choice sets generate unbiased estimates. If the IIA property does not hold, choice behaviour does depend on choice set composition (which seems realistic for many spatial choice problems). In that case, unbiased estimates can only be obtained if choice set composition is systematically included in the estimation, but such an ideal is very difficult if not impossible to establish when using observed choice behaviour for model estimation. Regardless of this situation, however, the choice probabilities will always depend on the right operational definitions of individual choice sets. If the choice set defined by the analyst includes alternatives actually never evaluated by the decision-maker, the choice model assigns non-negative probabilities to all alternatives in the choice set, including those that are not in the true choice set. Choice set composition will affect predicted market shares as the latent demand is allocated to the alternatives belonging to an individuals' choice set. Thus, predictions of market shares of choice alternatives will be wrong if individual choice sets are misspecified. If the composition of the choice set also influences individual preferences and choice behaviour, the parameters of the estimated utility or preference function will also be biased.

The objective of this paper is to review different choice set generation approaches since the last contribution goes back to the work of Thill (1992). Because operational decisions tend to depend on the underlying modelling approach (i.e. spatial interaction vs. discrete choice vs. activity-based models), these different approaches will be described as well.

This paper is organized into five sections. In the second section the different modelling approaches are discussed, spanning from spatial interaction models to discrete choice and activity-based models. The third section discusses the alternative approaches to delineate choice sets mechanisms. In the fourth section the computational process models are discussed. Conclusions and further research are presented in the last section.

## THEORETICAL APPROACHES

### Spatial interaction models

Spatial interaction models are aggregate models in the sense that both space and activities are grouped into discrete categories. Instead of analyzing particular points in space, zones containing a large number of activities are defined. Moreover, it is assumed that all individual members of a group have similar characteristics. These aggregates interact, generating flows of different kinds, such as trips, migrations, movements of commodities, etc. Spatial interaction models assume that the probability of interaction between two spatial entities (typically aggregate in nature) is proportional to some measure of size/attractiveness of these entities and negatively proportional to some function of spatial separation (distance, travel time, etc). Originally, the family of spatial interaction models (Wilson, 1970) has been applied in many different fields of application. Later, Fotheringham's Competing Destinations (CD) model (Fotheringham, 1983; 1988) became more popular. It differs from the traditional origin-constrained spatial interaction model in that it assumes that the choice of a destination is also influenced by its accessibility to a set of alternative, secondary destinations. In contrast to the conventional retail model which defines flows $S_{od}$ in terms of trip origins $O_o$, floorspace $W_d$ and home-based travel costs $c_{\text{od}}$, the CD model is expressed as:

$$S_{od} = \frac{O_o W_d^{\alpha} A_d^{\gamma} \, e\,xp(\beta c_{od})}{\sum_{d'} O_o W_d^{\alpha} A_d^{\gamma} \, e\,xp(\beta c_{od'})} \tag{1}$$

with the potential measure $A_d$ defined for competing destinations as:

$$A_d^{\gamma} = \sum \quad A^{\gamma}{}_d = \sum_{k \neq d} W_d c^{\delta}{}_{dk} m_{dk} \tag{2}$$

where $c_{dk}$ represents the travel costs between the primary destination $d$ and a nearby destination $k$. The binary variable $m_{dk}$ is equal to one when destination $k$ lies within a pre-specified range of destination $d$ and 0 otherwise. The introduction of the competing destinations potential terms distinguishes the model from the conventional spatial interaction model. A positive value of the scaling index $\gamma$ demonstrates the presence of agglomeration economies, whereas a negative value indicates the presence of competition.

As equation (1) indicated, the original model was based on an extension of the spatial interaction model. Borgers & Timmermans (1985, 1987, 1988) also developed a non-IIA model of spatial choice, but based on utility theory. One variant contained not only proximity among destinations, but also their similarity. In terms of choice sets, it shares the same problems as discussed for the competing destinations model.

### Discrete choice models

The criticism that spatial interaction models lack an underlying theory of choice behaviour, led to the development of discrete choice models. In general, these models are based on observations of individual choice behaviour and therefore this approach can be considered disaggregate. The main assumption underlying these models is that individuals choose rationally between the options available to them, subject to a number of constraints such as income. For instance, when an individual decides to buy a house, a number of options, varying in price, size, type and location will be available and the individual will face a number of restrictions, mainly in the scope of houses on offer and the available budget. Individuals are assumed to choose from the options available to them the one that yields the highest utility.

Discrete choice models are mathematical functions predicting an individual's choice based on the utility or relative attractiveness of alternatives. Random utility models have been frequently used in transportation research (Ben-Akiva & Lerman, 1985) and predict the probability of choosing a destination among all those available for individuals of category *i* travelling from zone *o*, for purpose *s* (possibly in period *h*):

$$p^i[d/osh] = Prob\ [U^i_{d/osh} > U^i_{d'/osh,}\ d' \in I^i_d,\ d' \neq d]] \qquad (3)$$

where $U^i_{d/osh}$ represents the sum of the systematic utility $V^i_{d/osh}$, and a random residual $\varepsilon^i_{d/osh}$, which is assumed to be independently and identically distributed (iid) according to a Gumbel distribution:

$$U^i_{d\ osh} = V^i_{d\ osh} + \varepsilon^i_{d\ osh} \quad \forall d \in I_d{}^i \qquad (4)$$

$$V^i_{d\ osh} = \sum_k \beta_k X^i{}_{kj} \qquad (5)$$

The systematic utility typically is a linear function in the coefficients $\beta_k$ of the attributes $X^i_{kj}$ and $\beta_k$ are parameters to be calibrated (Lerman & Louviere, 1978).

In a spatial context, it is generally assumed that the zones in the study area represent elementary destination choice alternatives. In reality, the destination where one chooses to carry out an activity is not a traffic zone but rather a specific location or locations (i.e. an office or a shopping centre) within a traffic zone. These specific locations are the elementary destination alternatives. Therefore, a traffic zone should be modelled as a compound alternative that results from aggregating its elementary destination alternatives. Different functional forms can be derived depending on whether the elementary alternatives are taken to be the traffic zone or the specific destination locations. Therefore, the two cases will be discussed separately.

**a. Choice alternative: a zone**

When the alternative is assumed to be a generic traffic zone, a Multinomial Logit functional form is generally used:

$$p[d/osh]=\frac{exp(V_{d/osh}/\theta_d)}{\sum_{d'}exp(V_{d'/osh}/\theta_d)} \tag{6}$$

where $V_{d/osh}=E[U_{d/osh}]$ is the systematic utility derived from destination zone $d$ and $\theta_{\mathrm{d}}$ is a scaling parameter. For simplicity of notation the superscript $i$ has been omitted. Examples of destination Multinomial Logit models can be found in Ben-Akiva et al., 1985; Ben-Akiva & Lerman, 1985; Cascetta, 2001; Koppelman & Sethi, 2000; Ortuzar & Willumsen, 2001; Train, 2003.

**b. Choice alternative: elementary destination within a zone**

In the case of destination choice models, the alternative is represented by the elementary destination within the traffic zone. As the real interest of the analyst is to reproduce the distribution of trips between traffic zones and not between elementary destinations, an aggregation procedure of elementary destinations into traffic zones is needed. Therefore, the traffic zone d is a compound alternative made up by the aggregation of $M_d$ elementary alternatives. To calculate $p[d/osh]$, a Nested Logit model is usually used (Williams, 1977). In this case equation (6) becomes:

$$p[d/osh]=\frac{exp(V_{d/osh}/\theta_d+\delta_d Y_d)}{\sum_{d'}exp(V_{d'/osh}/\theta_d+\delta_d Y_{d'})}=\frac{exp[(V_{d/osh}+s_d)/\theta_d]}{\sum_{d'}exp[(V_{d'/osh}+s_{d'})/\theta_d]} \tag{7}$$

where:

$V_{d/osh}=E[U_{d/osh}]$ is the systematic utility of the traffic zone $d$, common to all elementary destinations in $d$;

$\theta_d$ is the parameter of the Gumbel distribution of Ud/osh;

$s_d=$ is the inclusive utility relative to elementary destination choice;

$V_{r/d}=E[U_{r/d}]$ is the systematic utility of the elementary destination $r$ conditioned to traffic zone $d$;

$\theta_r$ is the parameter of the Gumbel distribution of $U_{r/d}$;$\delta_d=\theta_r/\theta_d\leq 1$

From the above another equivalent specification of $s_d$ may be derived:

$$s_d = \overline{V}_d + \theta_r \ln M_d + \theta_r \ln \frac{1}{M_d} \sum_{r'=1,\ldots,M_d} \exp[(V_{r'/d} - \overline{V}_d)/\theta_r] \tag{8}$$

where $\overline{V}_d$ is the mean systematic utility of the elementary destinations and it is equal to:

$$\overline{V}_d = \frac{1}{M_d} \sum_{r'=1,\ldots,M_d} V_{r'/d}$$

This expression is particularly advantageous when the attractiveness of each elementary destination cannot be determined. In fact equation (8), except for the last term on the right side (which represents a heterogeneity term), can be easily calculated if the number $M_d$ of the elementary destinations is known.

For certain trip purposes it is possible to assume that the number of elementary destinations, $M_d$, within each traffic zone d can be measured (e.g., the number of shops for the shopping purpose). However, for several trip purposes the precise identification of the elementary destinations and their number is not possible. In this case, the size variable $M_d$ may be replaced by a size function expressing the actual (unknown) number of elementary destinations as a function of other variables (such as population, employment in different sectors, number of firms of different types, etc.):

$$M_d = \Sigma_k \beta_k Z_{kd} \tag{9}$$

Although the multinomial logit models still are the work horse in discrete choice analysis, some other interesting specifications have been suggested. Because the focus of this paper is on choice sets and issues in choice set generation are invariant to model specification in this stream of model, we will only discuss a few alternative specifications. An interesting approach is the one proposed by Gaudry & Dagenais (1979). Their Dogit model is flexible in that it permits the choice among specific pairs of alternatives to be consistent with the independence from irrelevant alternatives axiom, but not for others.

The handful of empirical studies that have used captivity models for travel mode choice problems have all stressed the appropriateness of the approach for this kind of choice situation (Kitamura & Lam, 1984; Swait & Ben-Akiva, 1986). An application to a spatial context is due to Swait & Ben-Akiva in a shopping scenario. A polycentric urban area is considered in which all possible cases seem a priori well suited for a captivity model whereby the full set of destinations is split into as many choice sets as there are suburban malls plus one (for the downtown area). It is also assumed that one has sufficient empirical evidence showing that, by and large, people shop either downtown or at one of the suburban malls. In this context, the captivity concept is still valid. It only makes sense to restrict the individual's universal choice set to the destinations in these two sets, which define an appropriate partition. Suppose next that a number of shopping destinations are located out of the major retail centres, for instance in strip developments. Because of the geographic dispersion of destinations, individuals are no longer likely to be captive to deterministic sets of destinations. In this context, the a priori specification of these sets would rely on the analyst's judgment. In turn, this would recreate the very conditions of arbitrariness that the modelling of choice-set generation processes was designed to prevent. The appropriateness of a captivity structure in the context of shopping destination choices should be evaluated on a case-by-case basis (Thill, 1992).

Horowitz & Louviere (1990) noticed that choice set formation models, along with conventional discrete choice models, assume implicitly that choice-set generation is independent of the preferences of the decision-maker. In other words, the random components of the choice and choice-generation models are believed to be independent. As a rule, this position is untenable because it lacks a firm behavioural base. One expects intuitively that unobserved destination attributes (e.g. the friendliness of the sales personnel in the context of shopping behaviour) and unobserved individual characteristics (e.g. reluctance to travel) affect both the formation of a set of alternatives and the actual selection from this set.

Traditional travel demand models represent the trips that comprise a journey (a sequence of trips starting and ending at home) assuming that the decisions (choices) made for each trip are independent of those made for other trips in the same journey. It was also noted that these assumptions are reasonable when the journey is a "round trip" with a single destination and two symmetric trips.

However, human activities have become increasingly complex, especially in urban areas. One reflection of this in the domain of transportation is an increasing number of journeys that connect multiple and disparate activities in different locations, i.e. journeys consisting of sequences of trips that influence each other in complex ways. A number of demand models have been proposed in the literature to address the sequence, or chain, of trips making up a journey.

Some of these models represent the activities carried out (i.e. the different purposes of the journey) together with the trips that link them.

The mathematical models proposed to represent trip or activity chains do not have a standard structure as, for example, trip demand models do. This is due both to the relatively recent interest in these models (so there are fewer examples of them), and to the greater complexity of the phenomenon to be represented.

However, the most commonly used modelling structure is based on the concept of a primary activity (destination) for a particular journey. In other words, it is assumed that each journey is associated with a primary activity (or purpose), and that this activity is conducted in a particular place, known as the primary destination. Experimental studies suggest that the activity that the user perceives as primary for a particular journey is determined by relatively few criteria. These include:

- hierarchical level of purpose (in decreasing order, workplace or study, services and professional business, other purposes);
- duration of the activity (the primary activity is that which, within the highest hierarchical level, takes the most time);
- distance from zone of residence (the primary activity, given the same hierarchical level and duration, is that which is carried out in the place furthest from the residence).

Adopting this definition, a system of demand models for trip sequences (journeys) can be specified with a partial share structure analogous to the standard four-step model.

There has been much research aimed at improving the specification of the random utility (or discrete choice) model (Walker & Ben-Akiva, 2002), which models a decision-maker’s choice among a set of mutually exclusive alternatives. A guiding philosophy in these developments is that such enhancements lead to a more behaviourally realistic representation of the choice process, and consequently a better understanding of simpler models structures. A number of important extensions have been presented in the literature, including:

- *Flexible Disturbances* in order to allow for a rich covariance structure and enable estimation of unobserved heterogeneity through, for example, random parameters.

- *Latent Variables* in order to provide a richer explanation of behaviour by explicitly representing the formation and effects of latent constructs such as attitudes and perceptions.
- *Latent Classes* in order to capture latent segmentation in terms of, for example, taste parameters, choice sets, and decision protocols, and:
- Combining *Revealed Preferences and Stated preferences* in order to draw on the advantages of the two types of data, thereby reducing bias and improving efficiency of the parameter estimates.

**Behavioural models**

Although discrete choice models are often viewed as behavioural models because their mathematical expression can be derived from random utility theory, some critical remarks may be in order. First, the fact that the equation can be derived from random utility theory does not make the model behavioural as the same mathematical expression can also be derived from Boltzman's laws of temperature, and principles of bounded rationality. Thus, alternative, also non-behavioural theories will lead to the same outcomes, and fundamentally different theories lead to the same mathematical specification. The problem is that most discrete choice models are applied to the outcomes of decision processes. Researchers interpret the model in terms of an a priori chosen theory, but do not provide additional evidence of construct validity. Second, although the theory is formulated at the individual level, some applications of multinomial logit models are still based on aggregate data.

The problem that most multinomial logit models are based on observed outcomes (revealed preferences) implies the estimated parameters of the utility/preference function only reflect preferences if the environmental choice allows individuals to express their preferences. If choice sets are restricted, for example because of unavailable alternatives or disequilibrium in the market, estimated parameters will reflect some unknown combination of preferences and constraints. To avoid this problem, some alternatives have been suggested. For example, Rushton's revealed preference scaling model (1969) tried to avoid some limitations of aggregate models by creating choice sets, defined in terms of distance and size categories. As this was still arbitrary, Timmermans (1979) suggested including only those alternatives that fall within a reasonable distance circle, explicitly measured for each respondent. Of course, there may be other reasons why an individual would not consider an alternative, beyond distance/travel time, but at least the definition of a choice set is not an arbitrary decision that applies to all individuals.

In this stream of research, several authors, mostly in analytical research, have measured mental maps, action spaces, familiarity sets and similar concepts, some of which might be taken as good representations of choice sets (see Timmermans & Golledge, 1990, for an overview). These action spaces turned out to be strongly related to size, distance and relative location of the alternatives. Thus, although it would be possible to first predict the probability that an alternative belongs to the action space/choice set, and then predict the conditional probability that it is chosen, such models would likely use the same set of variables for the two processes, which would not be a convincing property of such an approach. Alternatively, one could directly use the choice sets, explicitly measured on the basis of the answers to specific questions of respondents. In that case, one would have to generalize the results from the sample to the population.

#### Activity-based approach

Both spatial interaction and discrete choice models typically assume that all trips are made from home. It is well-known however that an increasing share of interactions involves multi-purpose, multi-stop trip chains, embedded in activity-travel patterns. Different types of activity-based models have been suggested in the literature (Timmermans et al., 2002). Closest to the original theoretical framework are the so-called constraints-based models. The main purpose of these models is to assess the feasibility of an activity agenda or identify the number of possible ways any given activity agenda can be implemented, given space-time constraints. Compared to spatial interaction and discrete choice models, these dynamic constraints are generated endogenously. However, the constraints-based models do not include any mechanisms about how individuals cope with such constraints and may dynamically adjust their activity-travel patterns when faced with constraints. Moreover, they typically assume that individuals know all choice alternatives. They lack any behavioural mechanisms and process model. In recent years, several GIS-based approaches have been proposed and implemented that incorporate such spatio-temporal constraints in the definition of destination choice sets for engaging in discretionary activities (Kwan, 1997; Kwan & Hong, 1998).

Other activity-based models are based on the principle of utility-maximizing behaviour. In some sense, they can be viewed as elaborations of the original single-facet discrete choice models, incorporating more choice facets and hence complexity. An example is Bowman & Ben-Akiva (2001). Their model identifies different primary activities (Subsistence (work or school) on tour; Subsistence (work or school) at home; Maintenance (shopping, personal business, etc.) on tour; Maintenance at home; Discretionary (social, recreation, entertainment, etc.) on tour and Discretionary (social, recreation, entertainment, etc.) at home. Activities are combined into tours. There are eight possible types for work/school tour and four possible types for maintenance and discretionary tours. A day activity pattern model predicts the number and purposes of secondary tours, involving six alternatives (No secondary tours; One secondary tour for work or maintenance; Two or more secondary tours for work or maintenance; One secondary tour for discretionary purpose; Two or more secondary tours for discretionary purpose and two or more secondary tours: at least one for work or maintenance and at least one for discretionary purpose). Since not all of the tour types apply to all of the primary activity types, there are 19 possible combinations of primary activity/tour types. Each of the six secondary tour alternatives are possible for all primary activity/tour types, so the model has a total of 19*6= 114 alternatives.

These utility-maximizing models are estimated based on the relative occurrence of sub-patterns in the data. Choice set formation and composition are thus not explicitly considered. Because these models lack any concept of a behavioural process, dynamic choice sets also are not considered in the estimation phase. Ad hoc mechanisms need to be introduced to simulate activity-travel patterns in forecasting and impact assessments.

Albatross (Arentze & Timmermans, 2000, 2004, 2005) is a rule-based model founded on an explicit process representation. It has in common with constraints-models the concept of space-time prisms, but goes beyond these models in that activity schedules are generated and not assumed given. It means that durations are always determined such that the activity fits into the time available. Likewise, activity types, modes and locations are determined such that it will be possible to arrive in time to the next fixed activity. In this way space-time prism constraints are used to exclude certain activity types, modes and locations from the respective choice sets. This is what guarantees that the prism constraints are satisfied. The destination-choice set for a given activity is embedded in the individual’s space-time prism and can be extracted easily by selective sorting of destinations in the prism. Knowledge of the operating constraints can be used to delineate the portion of the space-time geography that can be chosen. This approach to choice-set generation provides an operational framework that encompasses a strong behavioural basis. However, it should be noted that the typical input to modelling dynamic choice sets, does not differentiate between individuals and does not include any explicit measurement of individuals' awareness, consideration or choice sets. In that sense, true choice sets are likely a subset of these dynamic choice sets.

## APPROACHES TO DELINEATE CHOICE SETS

Model estimation requires that the models are specified (i.e. the functional form and the variables are defined), calibrated (i.e. the unknown coefficients are estimated) and validated (i.e. the ability to reproduce the available data is tested). In case of discrete choice and activity-based models, these operations are performed on the basis of disaggregate information for a sample of individuals. Once the models have been specified and calibrated, they can be applied to derive estimates of present demand and/or to derive forecasts of future demand. In order to have correct forecasts, it is fundamental to generate a correct choice set, even if the model validly represents choice behaviour.

Manski (1977) suggested a stochastic, exhaustive and explicit approach to choice set formation. He centred the attention on the choice set formation process suggesting on the two-stage approach:

$$P^i_d = \sum_{C \in G} P^i(d / C) \quad P^i(C / G) \tag{10}$$

where:

$P^i_d$ is the probability that individual *i* chooses alternative *d* *(d∈I); I* is the universal set of alternatives *i*;

$P^i(d/C)$ is the probability that individual *i* chooses *d* given choice set *C*;

$P^i(C/G)$ is the probability that the choice set of individual *i* is *C*.

In this case *G* is the set of all non-empty subsets of *I*.

The main problem that is associated with Manski's approach is that the number of elements to which the sum is extended increases exponentially with the number of alternatives. Therefore, the application of this method becomes very difficult in choice contexts implying a large number of alternatives.

Swait & Ben-Akiva (1987) incorporated random constraints in discrete models of choice set generation. If the value of an attribute, $X^i_{dk}$, is larger than a certain threshold $T^i_{xk}$, the individual will reject alternative d. Thus, in this model an individual will consider an alternative as part of his/her choice set if and only if all its attributes are within their respective thresholds. Earlier, Timmermans et al. (1986) had suggested a similar approach in the context of residential choice behavior, albeit as a representation of a hybrid compensatory/non-compensatory decision making process. Swait & Ben-Akiva assume that thresholds are independent, so the probability that d is included in the choice set of individual i is given by:

$$P^{i}{}_{d\in C} = \prod_{k=1}^{K_d} Pr\,ob\left(X^{i}{}_{dk} \leqslant T^{i} x_k\right) \tag{11}$$

The probability that alternative *d* does not belong to choice set *C* is given by $P^{i}{}_{d\in C} = 1 - P^{i}{}_{d\in C}$ and the probability that the choice set is empty is equal to the probability that no alternative belongs to it. The choice set formation is modelled:

$$P^{i}(C/G)\frac{1}{1-P^{i}{}_{C=0}} = \prod_{d=1}^{n_i}\left\{P^{id_{dC}}{}_{d\in C}\cdot P^{i(1-d_{dC})}{}_{d\notin C}\right\} \tag{12}$$

where $d_{dC}$ is 1 if d is an element of choice set *C* and 0 otherwise.

The main problem of the approach is the high degree of computational complexity associated with a large number of alternatives (more than 6). Moreover, in practice, the distinction between choice set generation and decision seems rather arbitrary in lack of any process data. In that case, one will only observe the outcome of the decision making process. Moreover, the application of the approach will likely involve the use of the same set of attributes, one to delineate the choice set and one to estimate the utility functions, which does not seem very appealing. Heterogeneity is also not taken into account.

Morikawa (1995) suggested an appropriate method when *I* consists of too many alternatives, whereby the set formation process is modelled by pair-wise comparisons of alternatives, reducing computational effort. Specifically, he proposes a two stage model, and each stage employs a different decision making protocol. The first stage models the choice set formation process. A choice set is defined by the alternatives which satisfy a set of constraints. The second stage is modelled by using ordinary discrete choice models. Therefore, an alternative is included in the choice set if and only if all the latent conditioning measures of the alternative satisfy the criteria. This type of choice set formation process is modelled by pairwise comparison of alternatives, which dramatically reduces the computational load of the stochastic choice set models, by estimating the choice set formation model and the discrete choice model simultaneously using only the information on actual choice. While computationally appealing, the approach shares the theoretical and behavioural disadvantages mentioned above.

Ben-Akiva & Boccara (1995) considered only a reduced sample of feasible alternatives where each available alternative can be selected randomly or by using an ordered availability assumption where the probability of each alternative to belong to the choice set is determined as a function of the difference between some socio-demographic characteristic of the household at their current location (such as income or crime rate) and those of each considered alternative.

Cascetta & Papola (2001) provide a solution to the choice set approach through the specification of a dominance measure. While pure dominance is rare in spatial choice problems, different degrees of dominance are accounted for, making the model some mixture of a conventional utility function and a measure of dominance. More specifically, the probability that an alternative d belongs to the decision-maker's choice set *C*, *P(d∈C),* is jointly simulated with its probability of being chosen within this set, P(d/C), by adding the logarithm of *P(d∈C)* in the utility function of alternative *d*, $U_d$. Specifically, in the IAP (Implicit Availability Perception) random utility model the individual choice set is implicitly simulated by adding to the utility of each alternative d the logarithm of its "availability/perception" $AP_d$:

$$U_d^* = U_d + \ln AP_d \tag{13}$$

Therefore, according to the r.u. theory:

$$p[d] = prob[U_{d'} + \ln AP_{d'} \leqslant U_d + \ln AP_d \quad \forall d' \in C] \tag{14}$$

with some assumptions ($U^*_d$ independently and identically Gumbel distributed, $E[\ln AP_d] = \ln E[AP_d] = \ln p(d \in C))$ equation (16) becomes:

$$p[d] = \frac{exp(V_d)[X_d] \quad p(d \in C)[Y_d]}{\Sigma_h exp(V_h)[X_h] \quad p(h \in C)[Y_h]} \tag{15}$$

which is formally analogous to the formulation of Fotheringham (1983).

A simplification consists in reproducing the availability/perception of an alternative *d* by introducing the availability/perception attributes $Y_d$ directly in the systematic utility of that alternative.

In particular, the authors argue that the perception of a zone can be approximated through dominance attributes (Cascetta & Papola, 2005), which can be inserted in the utility function as availability/perception attributes. A dominance value can be assigned to each alternative. They indicate that the new variable can be defined in several ways: it can be a Boolean variable, it can be a variable taking values between 0 and 1; it can be the number of times an alternative is dominated by the others. A dominance ranking of the alternatives can be provided as well, in which the rank occupied by the alternatives is defined by the number of alternatives dominating the alternative itself. The lowest rank (1) will be occupied by the alternative with the greatest number of superior alternatives and, in turn, the model will give it a lower probability of belonging to the set of alternatives. The highest ranks will be occupied by the alternatives with few superior alternatives and, therefore, by those better perceived by the individual. The rank itself can be considered as a dominance attribute in the utility specification.

The proposed methodology aims to bypass the question of excluding a priori any alternative from the choice set to avoid model misspecification. Although the term 'perception' is used, in reality the utility function is extended with an ad hoc proxy variable. Although reference is made to random utility theory, proxy variables defined by the analysts are used, implying that the specification does not seem to have an immediate behavioural basis.

Choice sets are formed by random sampling using the dominance criteria as weights (Cascetta et al., 2007). To provide a direct comparison, the proposed approach uses the dominance variables as an inverse sampling weight. The sampling is performed with replacement, but an alternative is only included once in the choice set. Different choice sets were considered with a different number of alternatives. This approach mobilised by this procedure is equally ad hoc as the approaches discussed previously.

Swait (2001) incorporated limits to the value of the attributes in the formulation of the decision problem (Fig. 1). According to this approach, each attribute k has a lower cut-off or threshold $\tau^i_{Lk}$ and an upper threshold $\tau^i_{Uk}$ and the utility function is expressed as:

$$U^i_d = \theta X^i_d + \theta^L \delta^{Ldi}(\tau^i_L - X^i_d) + \theta^U \delta^{Udq}(X^i_d - \tau^i_U) + \varepsilon^i_d \qquad (16)$$

where $\delta^{Ljq}$ is a diagonal matrix with the element *kk* = 1 if $X^i_{dk} < \tau^i_{Lk}$ and 0 otherwise. Similarly, the *kk* element in $\delta^{Udi} = 1$ if $X^i_{dk} > \tau^i_{Uk}$ and 0 otherwise.

The cut-offs are obtained from a survey (individuals are asked directly) and the parameters $\theta^U$ and $\theta^L$ are estimated. The EBA model is a limiting case where for an attribute *k* (with negative marginal utility) the lower threshold does not exist and $\theta^U$ is minus infinity. This model can be estimated using standard software. In the SP experiment analyzed by Swait (2001) individuals identified the attribute thresholds that were treated as exogenous variables. This approach has the advantage that cut-offs point are explicitly measured. Heterogeneity is not taken into account.

In the approach by Gillbride & Allemby (2004), if $T^i$ is a vector of random thresholds, the alternatives included in the choice set are identified with an indicator function, $\mathrm{I}(X^i_d, T^i)$ that is equal to 1 if the decision rule is satisfied and 0 otherwise.

The choice model can be written in the general form:

$$P^i_d = \mathrm{P}(V^i_d + \varepsilon^i_d \geq V^i_{d'} + \varepsilon^i_{d'} \text{ for all } d \text{ such that } \mathrm{I}(X^i_d, T^i) = 1) \qquad (17)$$

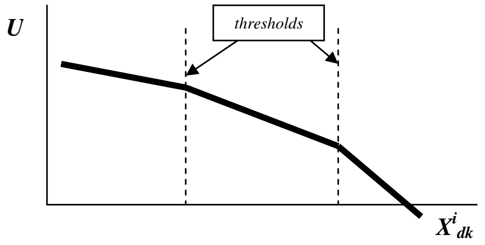


Figure 1: Utility variation incorporating limits in attribute values

In the case of a conjunctive rule, an alternative is acceptable if all its relevant attributes are acceptable. Then, the indicator function is $\prod_k \mathrm{I}(X_{dk}{}^i \leq T_k{}^i) = 1$. Likewise, in case of a disjunctive decision rule, if at least one of the attributes levels is acceptable, then $\sum_k \mathrm{I}(X^i_{dk} \leq T^i_k) = 1$.

Exclusion rules defined by travel time, distance, or site quality have behavioral appeal, yet are fundamentally limited by their one dimensional scope. To remedy this shortcoming while maintaining the concept that trips require costly inputs to yield utility generating outputs, Scrogin et al.

(2004) developed and tested an efficient site exclusion rule by estimating individual specific trip production functions using stochastic frontier models. Measuring the proximity of sites to the efficient frontiers, choice set composition, efficiency and probability of site choice, and non-market benefits were evaluated to the prevalent deterministic rule with recreational site choice data.

Cantillo & Ortuzar (2006) proposed a model considering correlation among thresholds. Also, if thresholds exist they can be random, different among individuals and be a function of socioeconomic features and choice conditions. Their formulation allows estimating both the model taste parameters and those of the threshold probability distribution. If thresholds are distributed according to a joint density function *Ω(δ)* with mean E[$T^i$] = *T* and covariance matrix Var[$T^i$] = ∑, then the probability $P^i_{d\in C} = P\{X^i_d \leq T^i\}$ can be written as:

$$P^{i}{}_{d\in C} = P\left\{X^{i}{}_{d} \leq T^{i}\right\} = \int_{X^{i}{}_{kd}}^{\infty} \int_{X^{i}{}_{2d}}^{\infty} \int_{X^{i}{}_{1d}}^{\infty} \Omega(\delta) d\delta_1 d\delta_2 \ldots . d\delta_3 \tag{18}$$

This approach is an important step forward in the sense that it allows for correlation among thresholds, and an explicit estimation of the threshold distribution parameters, reflecting heterogeneity.

The assumption usually made for road networks is that, before making a trip, the user chooses a sequence of road segments to follow to the destination. This can be represented as a path — a sequence of nodes and links on the graph that represents the road system. Only elementary (loopless) paths are considered, and thus their number is finite.

Definition of the paths considered as choice alternatives, i.e. definition of the choice set, is particularly important since the topological complexity of the network could generate an unrealistically large number of paths between a single O-D pair. The set of feasible paths $K_{odm}$ connecting the centroids *od* over the mode m network should in principle be defined through an explicit choice set model (Cascetta, 2008). In practice, however, two types of heuristic approaches are used.

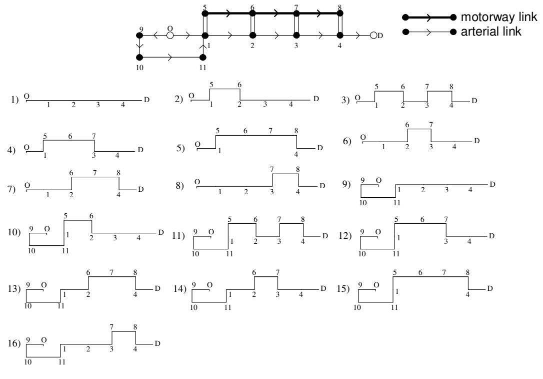
The exhaustive approach considers all elementary paths on the network. This approach may generate a large number of paths that share many links and so are correlated in their perceived (dis)utilities. Furthermore, given the computational complexity of explicitly enumerating all the paths in a network, this operation is usually carried out implicitly (implicit path enumeration) by algorithms that simultaneously calculate path choice probabilities and assign flows.

The selective approach, on the other hand, applies heuristic behavioural rules to identify only a subset of the elementary paths. For example, a path may not include more than one entrance and one exit from the same motorway, may not go further away from the destination, may not have a generalized cost exceeding the minimum path cost by more than a given amount, etc. Various criteria for the selection of feasible paths have been proposed in the literature. They relate to different application contexts (urban/interurban networks) and to different algorithms for generating paths and calculating choice probabilities and link flows. Examples of selection criteria are given in Table 1, where $Z_{o,I}$ and $Z_{o,j}$ represent the minimum cost to reach node *i* and node *j* from origin *o*.

| SELECTION CRITERIA | SPECIFICATION |
|---|---|
| Topological | A path is feasible (Dial efficient) if each link moves away from the origin and/or moves towards the destination, see section 7.3.1a<br>$k \in K_{od}$ if $Z_{o,i} < Z_{o,j}$ $\forall(i,j) \in k$ |
| Comparison of costs | Paths with a generalized cost not exceeding the minimum cost by more than a factor of $\alpha$<br>$k \in K_{od}$ if $g_k \leq (1+\alpha)\ g_{min}$ |
| Progressive | The *n* paths with the lowest generalised costs |
| Multi-attribute | Minimum paths with respect to various attributes (usually performance variables such as travel time, monetary cost, motorway distance, etc.) |
| Behavioural | Paths excluding behaviourally unrealistic link sequences (e.g. repeated entrances and exits for the same motorway) |
| Distinctive | Paths overlapping for no more than a given percentage of their length |

Table 1: Criteria for path feasibility on road networks

In general the selective approach requires explicit path enumeration between each O-D pair, and usually applies a combination of criteria. Fig. 2 depicts an example of the complete set of elementary paths and a selective set for an origin-destination pair.



Paths 1-16: complete set of elementary paths
Paths 1-2 + 4-8: paths selected by applying behavioral (eliminating paths 3 and 11) and topological (eliminating paths 9 to 16, which go further away from the destination) criteria

Figure 2: Examples of exhaustive and selective set of feasible paths

For more sophisticated feasible path generation models, the criteria to be used must be “calibrated” like other parameters in the model. Calibration can be carried out by comparing the paths generated by the model with the paths perceived (or chosen) by a sample of users, adjusting the model to maximize the coverage of the latter by the former.

Some experimental results suggest that a good level of coverage of the paths used by users, at least for interurban networks, can be achieved by generating the first n paths for some intuitively reasonable criterion (e.g. minimum time, minimum monetary cost, maximum motorway use, etc.)

The selective approach guarantees better control of the feasibility of the generated paths while allowing the use of performance attributes that are not additive over links. These advantages are obtained at the expense of greater computational complexity.

Conversely, implicit path enumeration methods are computationally more efficient and are typically used in the assignment models implemented in commercial software. However, it should be emphasized that there has been no systematic analysis of the computational complexity and memory requirements of the two methods. Recent literature suggests a growing tendency towards explicit path enumeration models in applications, perhaps because of the increasing power of the computing resources that are routinely available.

To predict route choice from among the routes that a particular traveller considers, the objective is to identify all the routes that any traveller might consider. Specifically, algorithmic rules have been identified for generating the observed routes to avoid introducing biases in the estimation procedure, and to have useful algorithms for navigation systems. Such algorithms should be able to reflect drivers’ knowledge of the transportation network and their perceptions of travel times and other network variables. Further, there is no benefit to enumerating routes that no traveller would consider. That is, computational effort is one criterion by which to evaluate potential path generation algorithms. The effectiveness of different path generation techniques is defined in terms of the generated routes’ coverage of the observed routes. Ideally, a generated route would match the observed route link-for-link; in this case, the algorithm has replicated the survey route. Other routes may not be replicated, so the distance that the generated route shares in common with the survey route is considered. Thus, the overlap typically is expressed as a percentage of the survey route distance. Finally, coverage is defined as the percentage of observations for which an algorithm or set of algorithms has generated a route that meets a particular threshold for overlap. There are many dimensions in which a path generation algorithm may be designed. A well-known method, known as the K-shortest Path algorithm, generates the first “k” shortest paths for a given origin-destination pair. Two popular heuristics may be classified as link penalty and link elimination methods. Both techniques proceed iteratively after identifying a shortest path. In a link penalty heuristic (see for example De La Barra et al., 1993), the impedance of all links on the shortest path is gradually increased. In a link elimination technique, links on the shortest paths are removed from the network in sequence to generate new routes.

The labelling approach of Ben-Akiva et al. (1984) exploits the availability of multiple link attributes, such as travel time, distance and functional class to formulate different “generalized cost” functions that produce alternative routes. These routes may be labelled according to the criteria such as “minimize time,” “minimize distance,” “maximize use of expressways,” etc. that yielded it. Simulation methods produce alternative feasible paths by drawing impedances from different probability distributions. The distribution type (for example, Gaussian, Gumbel, Poisson), distribution parameters, number of draws and the seed of the pseudo-random number generator are design variables.

The paper by Nielsen et al. (2002) presents the formulation, estimation and calibration of a large-scale stochastic multi-class road traffic assignment model for the Copenhagen Region. The model considers several classes of passenger cars (different trip purposes), vans and trucks, each with its own utility function on which route choices are based. In addition, buses are assigned according to their schedules. The different classes and types of vehicles influence all the speed-flow relationships on links and comprehensive sub-models for intersections and roundabouts describe queues and geometric delays. A combined stochastic assignment model for route, including willingness to pay and tunnel choice has been considered.

The advantages of the stochastic assignment model compared to the logit model are:

- the tunnel is a small part of the transportation infrastructure in Copenhagen. Therefore the choice of travelling through the tunnel is closely interrelated with the choice of route. The choice set is a huge number of possible routes, and can therefore not easily be represented by a discrete choice model;

- by dividing the travellers into tunnel users and non-tunnel-users before assignment, it would not be theoretically and practically possible to achieve consistency between the utility functions in the route choice models and the model for choice of travel through tunnel or not. It would furthermore be difficult to model user equilibrium considering congestion effects, if two separate assignments were carried out;
- in practice it would be difficult to force the route choice model to assign routes through the tunnel, for the group of travellers determined in the logit model as tunnel users.

The bridges that are the alternatives to the tunnel cannot be 'closed' in the assignment procedure, as some of the tunnel users also need to use one of the bridges. Therefore an iterative procedure would be needed in the assignment algorithm to reject 'forbidden' routes, until routes were forced to use the tunnel. This would result in an awkward software program with a large increase in calculation time. It would also result in a model that would be specific for this single infrastructure project.

On the other hand, a stochastic assignment model has been used for wider examinations of willingness to pay, for instance toll roads, road pricing or payment for other new infrastructure projects.

On the basis of its methodological and practical advantages, it was decided to proceed with the stochastic assignment model. Technically this is the first time a model of this type has been estimated on basis of SP data and put into practice on a large traffic network, although the model had been used earlier without this estimation.

The paper by Bekhor et al. (2006) focuses on the problem of estimating a route choice model for a large network. The approach was that to first to generate a choice set and use this choice set to estimate model parameters. The proposed choice set generation method falls in the class of deterministic methods. The advantage of such method is that it can be applied to any urban network with existent resources, since it is based on successive shortest path calculations using travel time and distance variables.

The modelling framework proposed in the paper by Frejinger & Bierlaire (2007) is to identify a subnetwork, i.e. a simplification of the road network only containing easy identifiable and behaviourally relevant roads. In practice, the subnetwork can easily be identified based on the route network hierarchy. The importance and the originality of the approach lie in the possibility to capture the most important correlation without considerably increasing the model complexity. This makes it suitable for a wide spectrum of applications, namely involving realistic large-scale networks. Estimation results are reported of a factor analytic specification of a mixture of Multinomial Logit model, where the correlation among paths is captured by error components.

### COMPUTATIONAL PROCESS MODELS

These models are based on some process assumptions which are used to simulate behaviour. Realising that that choice sets are formed by receiving information by moving around in a particular area and from different sources (social network, active information search, advertising), through learning and adaptation, in principle such a process can be simulated. Consider a newcomer to an urban area. In the beginning, this traveller will have very limited information about locations and their attributes where activities can be conducted, about the transportation networks and their properties (e.g., driving times) and about the institutional context. However by implementing their activities and by travelling, people will learn about the properties of the system, increase their knowledge and in quasi-stationary environments reduce the uncertainty. When travelling, people will see new locations and may decide to explore these, which perhaps may lead to changes in choice sets and adapted activity-travel patterns or a wider repertoire of context-dependent patterns. Arentze & Timmermans (2005a, b) captured these notions in a simulation model, focusing on location choice. Choice sets are dynamically changed because travellers learn new alternatives, some of which are added to their choice set. After some time, when particular locations have not been visited for a long time, the given choice is abandoned. Note that heterogeneity is included because these are individual-level processes. Han et al. (2007) elaborated this approach and also include the effect of social networks. Individual choice sets may change because other members of the social network had positive experiences with a particular destination. Thus, both expectation levels, needs and choice sets dynamically shift and become similar, as a function of the similarity of the members. They showed that under certain conditions (stationary environment), the choice sets reach an equilibrium state.

Although this approach is appealing in the sense that it is based on a rich conceptual framework, it is difficult to validate it. For example, the initial conditions will not be known and one typically also lacks process information. The best thing one can possibly do is test such process models against aggregate system-level counts. However, to the extent that such models are successful, the choice set formation does not represent a specific challenge as it is embedded in the large simulation which attempts to simulate how activity-travel patterns evolve in time and space and are adjusted to changes in the urban environment and other exogenous dynamics.

Developments have recently taken place in the literature on traffic assignment and route choice where the issue of choice set formation has been considered and relevant approaches developed.

## CONCLUSIONS AND FURTHER PERSPECTIVES

The topic of choice set formation and dynamics has been on the active research agenda in various disciplines for a long time. The topic deserves attention because it is not only interesting in its own right, but also to ultimately improve predictions of spatial choice behaviour or even to avoid biased, misleading estimates.

In this overview, we have discussed how the incorporation of choice sets has gained in sophistication over the years. In each of a multitude of modelling traditions, the issue of choice set formation and dynamics has received attention and much progress has been made.

This overview however also clearly indicates that modellers continue struggling with this topic. Even though the inclusion of latent stochastic thresholds and the simultaneously estimation of thresholds and utility functions represents an important step forward in discrete choice analysis, forecasting results still depends on the researchers' specification of the choice set. What seems lacking is a convincing process model that probably needs to be developed with a particular type of spatial choice behaviour in mind. Recreation choice processes differ from tourist choice, which in turn differ from shopping choice and residential choice. Such process models should conceptualize which information is used to learn about the environment (travel, social networks, dedicated information sources, (location-based) advertising, etc), when it will be consulted, when it will lead to exploration and what mechanisms are used to keep an alternative in one's choice set, but also when it will be deleted from the choice set. While it would not be too demanding to formulate such processes, data requirements for calibrating and validating such models are.

## REFERENCES

**Arentze**, **T.A**.; **Timmermans, H.J.P.,** 2000, Albatross: A learning based transportation oriented simulation system. European Institute of Retailing and Services Studies: Eindhoven

**Arentze, T.A.; Timmermans, H.J.P.,** 2004, A learning-based transportation oriented simulation system. In: Transportation Research Part B, 38, 2004: 613-633

**Arentze, T.A.; Timmermans, H.J.P.,** 2005, Albatross 2.0. EIRASS: Eindhoven

**Arentze, T.A.; Timmermans, H.J.P.,** 2005a, Information gain, novelty seeking and travel: A model of dynamic activity-travel behavior under conditions of uncertainty. In: Transportation Research A, 39: 125-145

**Arentze, T.A.; Timmermans, H.J.P.,** 2005b, Representing mental maps and cognitive learning in micro-simulation models of activity-travel choice dynamics. In: Transportation, 32: 321-340

**Bekhor, S.; Ben-Akiva, M.; Ramming, M.S.,** 2006, Evaluation of choice set generation algorithms for route choice models. In: Annals of Operations Research, 144 : 235-247

**Ben-Akiva, M.; Bergman, M.J.; Daly, A.J.; Ramaswamy, R.,** 1984, Modelling Inter Urban Route Choice Behaviour. In: Volmuller, J.; Hamerslag, R. (Eds), Proceedings of the 9th International Symposium on Transportation and Traffic Theory. Utrecht: VNU Press: 299-330

**Ben-Akiva, M**.; **Gunn, G.H.; Silman, L**., 1985, Disaggregate trip distribution models. Paper presented at the Japanese Society of Civil Engineers, Tokyo, Japan

**Ben-Akiva M.; Lerman, S**., 1985, Discrete choice analysis: theory and application to travel demand. MIT Press: Cambridge, Mass

**Ben-Akiva, M.; Boccara, B.**, 1995, Discrete choice models with latent choice sets. In: International Journal of Research in Marketing, 12: 9-24

**Black, W.C.**, 1984, Choice set definition in patronage modeling. In: Journal of Retailing, 60: 63-85

**Borgers, A.W.J.; Timmermans, H.J.P.,** 1987, Choice model specification, substitution and spatial structure effects: A simulation experiment. In: Regional Science and Urban Economics, 17: 29-47

**Borgers, A.W.J.; Timmermans, H.J.P.,** 1988, A context-sensitive model of spatial choice behaviour. In: Golledge, R.G.; Timmermans, H.J.P. (Eds), Behavioural Modelling Approaches in Geography and Urban Planning. Croom Helm, Beckenham: Kent: 159-179

**Bowman, J.L.; Ben-Akiva, M.,** 2001, Activity-based disaggregate travel demand model system with activity schedules. In: Transportation Research A, 35: 1-28

**Cantillo, V.; Ortuzar, J. de D**., 2006, Implications of thresholds in discrete choice modelling. In: Transport Reviews, 26: 667-691

**Cascetta, E.**, 2001, Transportation Systems Engineering: Theory and Methods. Kluwer Academic Publisher: Boston

**Cascetta, E.**, 2008, Transportation Systems Analysis: Models and Applications. Springer: New York

**Cascetta, E.; Papola, A.,** 2001, Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand. In: Transportation Research C, 9: 249-263

**Cascetta, E.; Papola, A.,** 2005, Dominance among alternatives in random utility models: a general framework and an application to destination choice. Paper presented at the European Transport Conference, October, Strasbourg, France

**Cascetta, E.; Pagliara, F.; Axhausen, K.W.,** 2007, The use of dominance variables in spatial choice set generation. Paper presented at the 11th World Conference on Transport Research, 24-28 June, Berkeley, USA

**Dai, J.,** 1998, Calibration and test of a discrete choice model with endogenous choice sets. In: Geographical Analysis, 30: 95-118

**De la Barra, T.; Perez, B.; Anez, J.,** 1993, Multidimensional Path Search and Assignment. In: Proceedings of the 21st PTRC Summer Meeting: 307–319

**Fotheringham, A.S.,** 1983, A new set of spatial interaction models: the theory of competing destination. In: Environment and Planning A, 15: 15-36

**Fotheringham, A.S.,** 1988, Consumer store choice and choice set definition. In: Marketing Science, 7: 299-310

**Frejinger, E.; Bierlaire, M.,** 2007, Capturing correlation with subnetworks in route choice models. In: Transportation Research B, 41: 363-378

**Gaudry, M.J.I.; Dagenais, M.G.,** 1979, The Dogit model. In: Transportation Research B, 13: 105-111

**Gautschi, D.A.,** 1981, Specification of patronage models for retail center choice. In: Journal of Marketing Research, 18 : 162-174

**Gillbride, T.; Allemby, G.**, 2004, A choice model with conjunctive, disjunctive and compensatory screening rules. In: Marketing Science, 23: 391-406

**Han, Q.; Arentze, T.A.; Timmermans, H.J.P; Janssens, D.; Wets, G.,** 2007, Modelling the dynamic formation of activity location choice-sets. Paper presented at the 11th World Conference on Transport Research, 24-28 June, Berkeley, USA

**Horowitz, J.L.; Louviere, J.J.,** 1990, The external validity of choice models based on laboratory choice experiments. In: Fisher, M.M.; Nijkamp, P.; Papageorgoiu, Y.Y. (Eds), Spatial choices and processes. Amsterdam: North-Holland: 247-263

**Kitamura, R.; Lam, T.N.**, 1984, A model of constrained binary choice. In: Volmuller, J.; Hamerslag, R. (Eds), Proceedings of the 9th International Symposium on Transportation and Traffic Theory. Utrecht: VNU Science Press: 49-65

**Koppelman, F.; Sethi, V.,** 2000, Closed form discrete-choice models. In: Hensher, D.A.; Button, K.J. (Eds), Handbook of transport modelling, Volume 1 of Handbook in Transport. Oxford: Pergamon Press: 211-222

**Kwan, M.P.,** 1997, GISICAS: An activity-based travel decision support system using a GIS-interfaced computational process model. In: Ettema, D.F.; Timmermans, H.J.P. (Eds), Activity-based approaches to activity analysis. Oxford: Pergamon Press: 263-282

**Kwan, M.P.; Hong, X.D.,** 1998. Network-based constraints-oriented choice set formation using GIS. In: Geographical Systems, 5: 139-162

**Lerman, S.R.; Louviere, J.J.,** The use of functional measurement to identify the form of utility functions in travel demand models. In: Transportation Research Record, 673: 78-85

**Lerman, S.R.,** 1984, Recent advances in disaggregate demand modelling. In: Florian M. (Ed.), Transportation Planning Models, Amsterdam: North-Holland: 10-17

**Manski, C.**, 1977, The structure of random utility models. In: Theory and Decision, 8: 229-254

**McFadden, D.,** 1978, Modeling the choice of residential location. In: Karlqvist, A.; Lundqvist, L.; Snickars, F.; Weibull, J.W. (Eds), Spatial interaction theory and residential location. Amsterdam: North-Holland: 75-96

**Miller, E.J.; O'Kelly, M.E.,** 1983, Estimating shopping destination models from travel diary data. In: Professional Geographer, 35: 440-449

**Morikawa, T.,** 1995, Correcting state dependence and serial correlation in the RP/SP combined estimation method. In: Transportation, 21: 153-165

**Nedugadi, P.,** 1987, Formation and use of a consideration set: implications for marketing and research on consumer choice. PhD dissertation, University of Florida, Gainesville, USA

**Nielsen, O.; Daly, A.; Frederiksen, R.,** 2002, A stochastic route choice model for car travelers in the Copenhagen region. In: Networks and Spatial Economics, 2: 327-346

**Ortuzar, J. de D.; Willumsen, L.G.,** 2001, Modelling Transport. John Wiley & Sons: New York

**Pellegrini, P.A.; Fotheringham, S.; Lin, G.,** 1997, An empirical evaluation of parameter sensitivity to choice set definition in shopping destination choice models. In: Papers in Regional Science, 76: 257-284

**Richardson, A.J.,** 1982, Search models and choice set generation. In: Transportation Research A, 16: 403-419

**Rushton, G.,** 1969, Analysis of spatial behaviour by revealed space preferences. In: Annals of the Association of American Geographers, 59: 391-400

**Scrogin, D.; Hofler, R.; Boyle K.; Milon, J.W.,** 2004, An efficiency approach to choice set generation. Working paper, University of Central Florida

**Shocker, A.D.; Ben-Akiva, M.; Boccara, B.; Nedugadi, P.,** 1991, Consideration set influences on consumer decision-making and choice: issues, models and suggestions. In: Marketing Letters, 2: 181-197

**Swait, J.; Ben-Akiva, M.,** 1987, Incorporating random constraints in discrete models of choice set generation. In: Transportation Research B, 21: 91-102

**Swait, J.,** 2001, Choice set generation with the generalized extreme value family of discrete choice models. In: Transportation Research B, 35: 643-666

**Thill, J.C.,** 1992, Choice set formation for destination choice modelling. In: Progress in Human Geography, 16: 361-382

**Timmermans, H.J.P.,** 1979, A spatial preference model of regional shopping behaviour. In: Tijdschrift voor Economische en Sociale Geografie, 70: 45-48

**Timmermans**, **H.J.P**.**; Borgers, A.W.J**., 1985, Choice set constraints and spatial decision-making processes. In : Sistemi urbani, 5 : 211-220

**Timmermans, H.J.P.; Borgers, A.W.J.; Veldhuisen, K.J.,** 1986, A hybrid compensatory-noncompensatory model of residential preference structures. In: Journal of Housing and Environmental Research, 1: 227-233

**Timmermans, H.J.P.; Golledge, R.G.,** 1990, Applications of behavioural research on spatial problems II: preference and choice. In: Progress in Human Geography, 14: 311-354

**Timmermans, H.J.P.; Arentze, T.A.; Joh, C.-H.,** 2002, Analyzing space-time behavior: new approaches to old problems. In: Progress in Human Geography, 26: 175-190

**Train, K.**, 2003, Discrete choice methods with simulation. Cambridge University Press: Cambridge

**Walker, J.; Ben-Akiva, M.,** 2002, Generalized Random Utility Model. In: Mathematical Social Sciences, 3: 303-343

**Williams, H.C.W.L.**, 1977, On the formation of travel demand models and economic evaluation measures of user benefit. In: Environment and Planning A, 9: 285-344

**Williams, H; Ortuzar, J. de D.**, 1982, Behavioral theories of dispersion and the mis-specification of travel demand models. In: Transportation Research B, 16: 167-219

**Wilson, A.G.**, 1970, Entropy in urban and regional modelling. Pion Press: London

## AUTHORS INFORMATION

**Francesca PAGLIARA**
fpagliar@unina.it
Department of Transportation Engineering
University of Naples Federico II - Italy

**Harry J. P. TIMMERMANS**
H.J.P.Timmermans@tue.nl
University of Technology
Eindhoven - Holland

# CELLULAR AUTOMATA IN URBAN SIMULATION: BASIC NOTIONS AND RECENT DEVELOPMENTS

**Nuno PINTO[*], António ANTUNES[*] and Josep ROCA[**]**

[*]Department of Civil Engineering, University of Coimbra, Portugal, [**]Center for Land Policy and Valuation, Technical University of Catalonia, Spain

## ABSTRACT

Cellular Automata (CA) is a spatial simulation technique that has been the subject of intensive research for the last two decades. This technique draws its theoretical origins in the 1940s with the research effort made by von Neumann and Ulam for devising mathematical rules for the evolution of biological systems. The intrinsic spatial character of CA suggested their introduction to quantitative geography by Tobler in the 1970s. In this chapter, we firstly present a concise literature survey on CA and their use in geography. The mathematical formulation of CA is presented, as well as their main applications to urban geography and urban studies. The discussion over important CA relaxations is introduced. In the second part of the chapter we present a series of recent developments regarding the use of geographic CA. Two main issues constitute the core of these developments: the choice of the modeling scale and the use of irregular cells. The development of a CA model for simulating change in small urban areas is presented. The use of irregular cells in opposition to the classic regular, pixel-based cells is also discussed. Finally, a reflection is made about future trends in a multi-scale CA for modeling urban and regional growth.

## KEYWORDS

Cellular automata, Land use, Urban change, Small urban areas, Multi-scale

## INTRODUCTION

The use of mathematical tools to model a wide range of spatial problems has been classified for the last decades as an important approach to scientific planning (Batty, 1994). The growth of urban areas is one of the issues which have concentrated a large research effort, and a series of new modeling techniques were introduced and developed throughout the last decades. Before the democratization of the use of computers, back in the 1970s, modeling approaches were usually applied to problems which demanded little information and, consequently, small computational effort. Computers made possible the use of more disaggregate information, shifting models from a large-scale perspective to microsimulation. New techniques were sought and Tobler (1979) proposed the application of a cellular approach for modeling geographic phenomena, introducing the use of cellular automata to geography. During the next two decades, a great effort was made to develop CA-based models: Couclelis (1985), White & Engelen (1993), and Batty (2005) worked on their theoretical issues regarding CA application to urban studies, Batty & Xie (1997) and Clarke et al. (1997) worked on the application of important evolutions of CA to real world problems, Semboloni (2000) studied urban infrastructure development, O'Sullivan (2001b) used an integrate approach based on CA and on graph theory to study gentrification, Semboloni (1997) and Ward, Murray & Phinn (2003) developed multi-scale urban models based on CA, and Silva & Clarke (2002) and Barredo et al. (2003) made applications of previously developed CA models to large metropolitan areas.

This chapter is dedicated to introducing the concept of CA and its main applications to quantitative approaches in geography and urban studies. The chapter has three main parts. The first part focuses on basic notions and on the introduction of CA to geography. The concept and its first applications in geography are discussed in detail, after which some applications are presented. The second part of the chapter is dedicated to an application of CA to small urban areas, using irregular cells; the model and the results obtained for a set of test instances and to a real world case study are presented in detail. In the third part of the chapter we identify some trends for future research in CA, pointing out some issues that need to be addressed.

## BASIC NOTIONS

Cellular automata (CA) were first introduced in the 1940s by John von Neumann, the founder of game theory, and Stanislaw Ulam, who made intensive research in the field of Monte Carlo simulation. Both researchers were dedicated to studying self-reproduction and to modeling biological life, trying to devise a mathematical formula that could reduce the forces governing reproduction to logical rules (Torrens, 2000). The concept of automatic computation lies on the very foundations of CA, as von Neumann was strongly influenced by Alan Turing's Universal Machine, a theoretical concept of a machine capable of universal computing. "A system may be regarded as a universal computer if, given a suitable initial program, it is capable of implementing any finite algorithm through its evolution over time, i.e., that it is capable of producing a working copy as complicated as itself, and the means to make further copies" (Torrens, 2000). This machine would be able to take new actions in a given moment of time based both on a set of parameters and on its knowledge about the external world obtained from an input system of any kind. An automaton can be considered as a Turing-like machine that operates over a tessellated cell space. The term automaton refers to a self-operating machine that "processes information, proceeding logically, inexorably performing its next action after applying data received from outside itself in light of instructions programmed within itself" (Levy, 1992, cited by Torrens, 2000). However, there is a large difference between CA and Turing's machine: CA are parallel processors, with a series of rules working at the same time while Turing machines are serial processors that can only handle one process after another (Torrens, 2000).

**The Concept of Cellular Automata**

According to the first formulation of a 2D cellular automata, due to von Neumann, each automaton cell processes information and performs action considering both a set of transition rules and data received from the environment. Formally, each cell *A* is defined by a set of cell states $S = \{S_1, S_2, \dots, S_N\}$ and a set of transition rules *T*

$$A \leftarrow (S, T) \qquad (1)$$

Transition rules define an automaton state $S_{t+1}$ at time step *t+1* depending on its state at time step *t*, $S_t$ ($S_t, S_{t+1} \in S$) and an input information $I_t$:

$$T : (S_t, I_t) \rightarrow S_{t+1} \qquad (2)$$

A grid of automata becomes cellular automata when the input is influenced by the states of neighboring cells. This definition of neighborhood is basically the set of cells that influences the state of the cell under consideration.

$$A \leftarrow (S, T, R) \qquad (3)$$

where *R* denotes automata neighborhood *A*, and establishes the boundary for drawing the input information *I* necessary for the application of transition rules *T*. Classic neighborhoods such as the von Neumann neighborhood (Figure 1a) and the Moore neighborhood (Figure 1b) are commonly used in 2D cellular automata. The first one is the set of four adjacent cells (usually) corresponding to the main cardinal points. The second one is the entire set of eight adjacent cells. Variations and conjugations of these two types of neighborhood have also been used (e.g., the one depicted in Figure 1c).

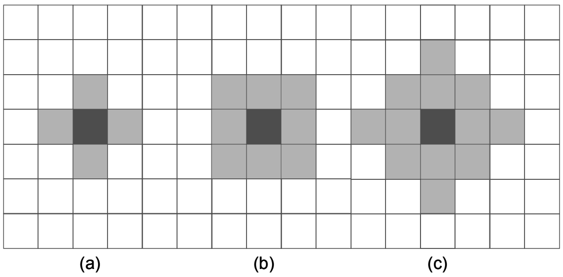


Figure 1: Classic von Neumann's, Moore's and mixed neighborhoods for 2D CA

There are two main moments in the history of CA after the introduction of complex systems theory: the discussion over John Conway's Game of Life in 1970 and the work of Stephen Wolfram in the 1980s (Wolfram, 1984).

The Game of Life was aimed to design a simple set of rules to study microscopic spatial dynamics of population (Benenson & Torrens, 2004). This 2D CA was based on a set of two cell states, alive (1) or dead (0), and on three plus one transition rules: survival (Figure 2a), death (Figure 2b) and birth (Figure 2c). The cell remained dead if none of these rules were applied (Figure 2d). The survival rule states that a living cell survives if it is surrounded by two or three living cells; the death rule states that a cell can die of loneliness (if it only has one neighbor) or because it is overcrowded; the birth rule states that a new cell is born if it has exactly three neighbor living cells; and finally, no cell can be born if it only has one neighbor or if it has more than five neighbors.

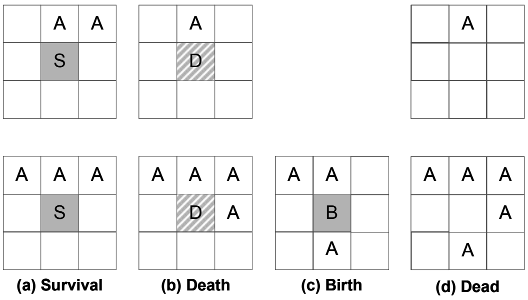


Figure 2: Conway's Game of Life set of rules

From this simple set of rules, the Game of Life could generate extremely complex patterns of growth through iterative simulation. Furthermore, it could be easily replicated following the concept of the Turing Machine. Its main achievement was to formulate a simple interdisciplinary tool for representing complex spatial systems and for modeling their dynamics (Benenson & Torrens, 2004). Conway pursued a configuration that could generate moving configurations of stable patterns. Conway's challenge on the pages of Scientific American for the research of such a configuration was a defining moment on CA history. Robert Gosper and his team at the Massachusetts Institute of Technology implemented a version of Life-CA that was capable of generating a machine that could reproduce copies of itself that were as complicated in their structure (Torrens, 2000). Considering the Game of Life as the first widely known CA, and relating its classification as a game to represent world evolution, it is interesting to consider that, perhaps, CA is like the world in its ability of representing the most complex forms of evolution from simple, well-understood interaction rules (Couclelis, 1997). It is worth stating the procedures of the Game of Life as they embody the key elements of CA (Batty et al., 1997).

Following the introduction of CA as a suitable tool for studying complex system dynamics made possible through the discussion over the Game of Life, research on the limits of system's spatial patterns began. The mathematician Stephen Wolfram made extensive research on CA exploring the final configurations towards which they could evolve to from simple sets of parameters and rules. Using a simple 1D CA, where the neighborhood is the set formed by the cell and its adjacent neighbors (a set of 3 cells), Wolfram studied limiting patterns for 256 different transition rules (Wolfram, 1983, 1984). Based on numeric experiments, Wolfram demonstrated that the final configuration of CA does not depend on the initial state of cells but is defined by transition rules. Wolfram classified CA according to their dynamic behavior and the patterns generated by them, identifying the four main classes described and depicted in Table 1.

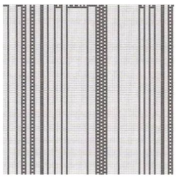
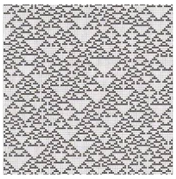
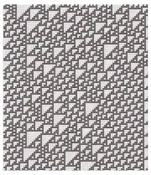
| Class | I | II | III | IV |
|---|---|---|---|---|
| **CA dynamics evolves towards** | Spatially stable pattern – each cell reaches the stable value of "0" and "1" | Sequence of stable or periodic structures – each cell changes its states according to the fixed finite sequence of "0"s and "1"s | Chaotic aperiodic behavior – the sequence of cell state is not periodic, but the spatial patterns repast themselves in time | Complicated localized structures, which are sometimes long-lived and are more complex than those of Class III |
| **Type of system dynamics** | Limit points | Limit cycle | Chaotic attractors | Attractors unspecified |
| **Graphic Output** | |  |  |  |

Table 1: Wolfram's classification for 1D CA behavior

Several other studies tried to devise new classifications by extending the study of Wolfram's classes to larger neighborhoods and different sets of rules and parameters. Wolfram's classification has the disadvantage of making it impossible to classify a given CA based on the transition rule alone, not knowing what the spatial pattern would be (Benenson & Torrens, 2004). However, Wolfram's classification remains the most popular one.

It is also important to mention the classification introduced by C. Langton which is very interesting for geography and urban studies (Benenson & Torrens, 2004). He introduces the concept of inactive cells, i.e. cells that cannot change state during CA evolution. Although a series of new classifications were proposed, several studies on higher-dimensional CA showed that they were similar to Wolfram's classification for 1D CA (Benenson & Torrens, 2004).

### Cellular Automata and Urban Studies

Most geographical theories are static, assuming that the interactions of agents take place in a generic market that remains in a state of equilibrium; this assumption is far from being reasonable, as all cities are continually undergoing complex changes. This fact makes imperative the use of dynamic models based on the processes that occur in the territory (White & Engelen, 1993). The great appeal of CA relates to the fact that many classes of system dynamics can be simulated through it. Another important feature of CA is their ability to give equal weight to the importance of space, time, and system attributes (Batty et al., 1997). The natural ability that CA have to represent systems with complex spatial/temporal behaviors from a small set of simple rules and states made this technique very interesting for geographers and urban researchers. CA are intrinsically spatial and they are used to model a wide range of phenomena due to their ability to represent spatial processes, from forest fires to epidemics, from traffic simulation to regional-scale urbanization, polycentricism, gentrification, historical urbanization, urban growth, form and location (White & Engelen, 1993; Torrens & O'Sullivan, 2001). CA-based modeling also allows the integration of socio-economic and natural systems models in a detailed and realistic way (White & Engelen, 1997).

There are three main classes of urban CA models that are a direct result of the exploration of modifications of the conventional CA: (1) models designed to explore spatial complexity; (2) models designed to research themes of economical, sociological and geographical areas; and (3) models designed to produce operational tools for planning (Torrens, 2000). In this section we present three pioneer studies that explored the formulation and the spatial complexity of CA from a theoretical standpoint.

From the 1960s until the mid 1980's geographers were more focused on the study of comprehensive regional models (Benenson & Torrens, 2004). The criticisms formulated by Douglass Lee (1973) in his "*Requiem…*" established a transition point after which the modeling community started to question the necessity for large scale models and shifted to small scale problems. Problem complexity, even when a small set of spatial relationships was used for modeling purposes, and the lack of experimental data for model calibration generated a search for simpler approaches that could produce more reliable simulations (Klosterman, 1994). CA presented an opportunity to deal with these new modeling requirements. Tobler and Couclelis had set the ground for this cellular approach, in the 1970s and the 1980s, when it really had a major development along with computer graphics, fractal theory, chaos, and complexity.

Waldo Tobler presented in 1970 his study on population growth simulation based on a cellular model (Tobler, 1970). While addressing the interdependence of population growth in a given location and its dependence on the population growth everywhere else in the world, he stated the first law of geography: *everything is related to everything else, but near things are more related than distant things*. This is an important concept that supports the link between CA and geography because of the importance of the concept of spatial neighborhood. Instead of using trend equations with limited sets of coefficients that could surrogate the real behavior of growth in every location, he considered the influence on population growth limited to a neighboring set of locations. Later in 1979, Tobler published another important study where CA are explicitly considered on the study of spatial phenomena (Tobler, 1979). Tobler made a classification of cell models for land use simulation, taking into consideration their dependence on spatial and temporal dynamics (Figure 3). Model type I is an independent model, where the land use for a given location $g_{i,j}$ at time $t+\Delta t$ is not related with the situation at time $t$. Model type II is classified as functionally dependent because land use at a given cell $g_{i,j}$ at time $t+\Delta t$ depends on the land use at that location at time $t$. Model type III is named a historical model as land use for a given location $g_{i,j}$ at time $t+\Delta t$ depends on a series of previous states at previous time periods. Model type IV is classified as multivariate because it depends on several variables other than land use at that location *i,j*. Finally, model type V is classified as the geographical model because land use at a given location $g_{i,j}$ at time $t+\Delta t$ is dependent on the land uses in a given neighborhood of that location at time *t*.

Tobler developed his study using a model type V which he classified as a dynamic one, because land use at a given location $g_{ij}$ at time step $t+\Delta t$ is a function *F* of land use at that location at time *t* and of a measure of the influence of all land uses located at the neighboring cells, $n_{ij}$, as expressed in Equation 4:

$$g_{ij}^{\ t+\Delta t} = F(g_{ij}^{\ t}, n_{ij}) \qquad (4)$$

The neighborhood *n* is based on the traditional von Neumann neighborhood and is defined as a geographical domain of influence. He considers that, because of the different notions of neighborhood different residents have, it should be possible to have dynamic neighborhoods varying on size, shape and orientation.

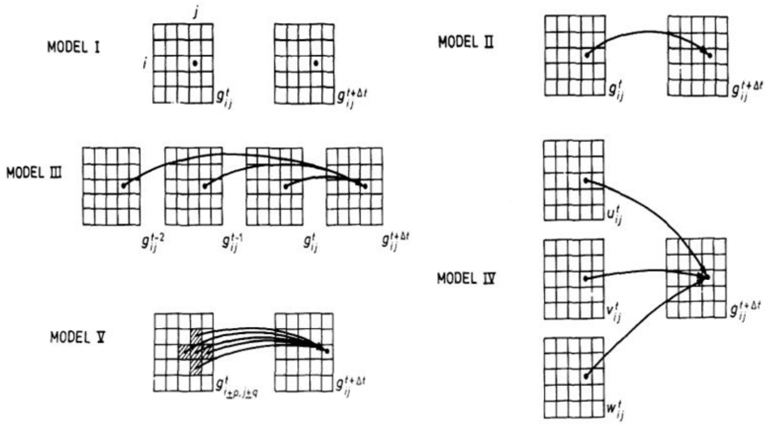


Figure 3: Classification of models of land use change (Tobler, 1979)

Transition rules *F* are defined closely to conventional CA definition. Tobler is more interested in studying geographic features of transition rules such as spatial isotropy and spatial stationarity. By spatial isotropy it is meant that the positioning of neighbors does not have any influence in the transition rule. By spatial stationarity it is meant that the same environment, the same neighborhood, results in the same consequences; that is, the rules do not depend on the location of the cell. These conclusions are, of course, very generic and considerably distant from real world geographic behaviors (Tobler, 1979). Tobler's groundbreaking work was followed by a series of other studies that explored CA and its application to geography (Couclelis, 1985; White & Engelen, 1993).

In the mid 1980s, Helen Couclelis continued the work started by Tobler in the research of CA for urban modeling purposes (Couclelis, 1985). Her work has two main lines: the first one relates to the conceptual and theoretical linkage with the theory of complex systems; the second one relates to the exploration of possible uses in urban planning (White & Engelen, 1993). She establishes a parallelism between the set of transition rules of the Game of Life and urban changing phenomena. A 'live' cell could be interpreted as an urban zone that exceeds some kind of threshold in terms of a given urban function. Transition rule 1 (survival) would then guarantee that this cell would maintain this urban function if two or three neighboring cells also exceeded that threshold. Transition rule 2 (death) would make this cell loose that urban function if four or more neighboring cells also exceeded that threshold because the cell would suffocate or, on the other hand, if there was only one neighboring cell exceeding the threshold (then the cell would die from loneliness). Transition rule 3 (birth) would settle this urban function in a cell if there were exactly three neighboring cells that exceeded the threshold. A cell would remain dead if none of the precedent conditions were verified (rule 4, dead). However, the inherent simplicity of this formulation makes it inappropriate to simulate real world spatial phenomena. Couclelis identified a series of shortcomings on the ability of cellular models to simulate urban phenomena. Issues regarding space dimension and boundary problems are pointed out as the first limitations for modeling cellular worlds. The imposition of universal transition rules to an infinite regular cell space collides with an empirical interpretation of the method. A second limitation is related with the regularity of neighborhoods. In order to be representative of real world phenomena, neighborhoods should be different in shape and size for different cells. Cell regularity and spatial homogeneity within each cell are also issues that make classic CA formulation far from being geographically representative. Finally, it is also important to notice the assumptions of space and time invariance for transition rules as well as the system closure to external perturbations (other than stochastic behaviors) (Couclelis, 1985). The solution relies on the consideration of relaxations that would enhance the ability of a cell-space approach to simulate real world phenomena without loosing both its identity and its simplicity. Couclelis reformulated the structure of the Game of Life to obtain a simple geographic CA formulation, a generic structure that could be applied to a series of spatial problems (see Couclelis (1985)). She also stated that this structure can be easily relaxed in order to improve simulation. Couclelis' formulation is independent from cell shape and size, and even from the existence of a cell space. The cell space concept can be generalized to the point where it describes any discrete time/space model representing components and their interactions (Zeigler, 1976 cited by Couclelis, 1985). It is likely that any model of interest to geographers, whether aggregate or disaggregate, continuous or discrete, deterministic or stochastic, quantitative or categorical, can be expressed in the same language as the cell-space concept (Couclelis, 1985).

Another important concept was introduced by Roger White and Guy Engelen: the concept of constrained CA (White & Engelen, 1993). Conventional CA are modeled with the explicit intention of being as general as possible. Because of this they have two main characteristics: (1) they are defined on a homogeneous cell space, overriding the variability of cell characteristics as they exclusively relate cells with their state values regardless of their location on the cell grid; (2) they are unconstrained, so that the number of cells in each state is determined endogenously by the application of transition rules to the current configuration of cells (White & Engelen, 1997). In their initial work, White & Engelen (1993) made some important assumptions that can be considered an approximation of CA to more likely urban behaviors. The first assumption is related to the consideration of a small number of hierarchically ordered cell states. There are four cell states, vacant (the lowest state), housing, industrial and commercial (the highest one). A cell in the vacant state can change to any other state but the inverse is not possible (thus the city can only grow, which is unlikely to happen). A cell in the industrial state can only change to a commercial state. Another assumption is related to the use of neighborhoods larger than the traditional Moore or von Neumann ones. The interaction between cells is then dependent of a larger area of influence, breaking up with the formal concept of local neighborhood and introducing what is commonly named as action-at-a-distance. On a classic CA model changes of state must be locally influenced, discarding any action-at-a-distance (Batty et al., 1997). In this case, these relationships (depicted in Figure 4) are assumed non-monotonic and may present positive values (meaning an attractive relationship) or negative ones (a repulsive relationship). Finally, White & Engelen considered transition rules that are not based on a simple probability of change but on a composite measure of a transition potential. This transition potential is defined as a weighted sum that simulates the behavioral propensities of the actors who determine land use, as expressed in Equation 5:

$$P_{ij} = S\left(1 + \sum_{h,k,d} m_{kd} I_{hd}\right) \tag{5}$$

where $P_{ij}$ is the transition potential from state $i$ to state $j$, $m_{kd}$ is the calibration parameter applied to a cell in state $k$ at distance $d$, $h$ is the index of cells within a given distance, $I_{hd}$ is equal to 1 if the state of cell $h$ is $k$ and 0 otherwise, and $S$ is a stochastic perturbation. The stochastic perturbation is calculated through the following expression:

$$S = 1 + [-\log(R)]^{\alpha} \tag{6}$$

where $R$ is a uniform random distribution in the interval ]0,1[ and $\alpha$ is a control parameter for the adjustment of the size of the perturbation. This term has a highly skewed distribution so that most values are near one and much larger values occur only infrequently. The weighted sum in the transition potential is multiplied by *S* in order to simulate the stochastic behavior of agents in each component of the transition potential.

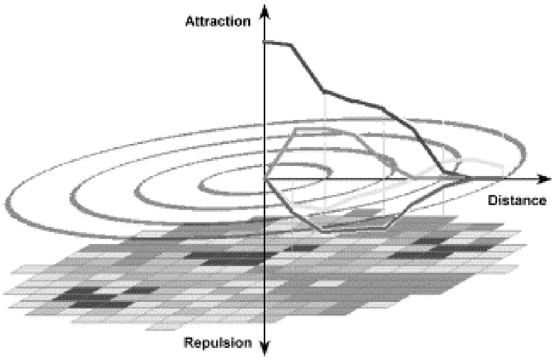


Figure 4: Neighborhood interactions (Benenson & Torrens, 2004)

Another important evolution from the early CA models is related to the introduction of land use demand as an exogenous factor. In classic CA, the model evolves over time from a small set of cells to increasingly more complex systems through transition rules. The number of changeable cells is determined only by the change process itself. White & Engelen (1993) introduced an exogenous parameter that aims to simulate land use demand: there is a predetermined number of cells for each state that is considered in the transition procedure. If the total number of cells is not attained for a given state, there is a probability that converts cells in lower cell states to the given cell state until the demand is fully satisfied. The transition potential is calculated for every cell at every time step of the simulation and is ordered in a descending manner. The set of cells with higher potential is chosen for transition considering land use demand. This approach has been developed and improved, and was used in recent studies (White & Engelen, 2000; Barredo et al., 2003; Pinto, 2006). Land use usually depends on three main types of factors: the inherent qualities of land itself, the effects of neighboring land uses, and the aggregate level of demand for each land use (White & Engelen, 1997).

Another innovation introduced by White & Engelen (1993) is the consideration of fractal measures to assess model performance and simulation results. Several fractal measures are calculated for the set of theoretical cities generated by the model and then they are compared with known fractal measures from real world cities such as Cincinnati, Atlanta, and Houston.

The fractal dimensions obtained for the theoretical cities were in line with the values observed for the set of real world cities, which indicates CA's ability to produce realistic simulations of urban growth.

It is also important to refer the work of Michael Batty on CA and urban studies. Batty is the author of a series of studies where CA is studied both from a theoretical and an operational point of view. His recently published book Cities and Complexity provides a framework for understanding cities as complex entities and presents CA (as well as agent-based simulation and fractal theories) as powerful tools for simulating complexity (Batty, 2005). In his long work on urban simulation, Batty studied CA as a tool for exploring city complexity, mostly by formulating and implementing a series of CA models aimed to explore theoretical issues regarding the technique. He developed, with Yichun Xie, models for what they called "areal automata" to simulate urban structure, and for "linear automata" to simulate the growth of road networks with hierarchical structure (Batty & Xie, 1997). They then generalized these two types of models and developed an urban growth CA model based on cells and networks. These models were based on simple formulations close to the probabilistic formulation of conventional CA.  The models also considered few land uses which shows their exploratory character. Later, a more sophisticated simulator was developed based on the work of Xie. The dynamic urban evolutionary modeling (DUEM) was aimed to capture urban complexity and urban dynamics. It was applied both to theoretical urban structures and to real urban areas, such as Amherst, New York (Xie, 1996). Batty also pioneered the coupling of CA and GIS as a means of developing visualization, which is increasingly important as a simulation tool (Batty et al., 1999b; Batty, 2005).

A series of more recent applications of urban CA are also important to the history of CA and urban modeling. However, it is important to first discuss the main evolutions of CA that are essentially based on relaxations of its four main components. The vast majority of these applications (if not all) introduced important innovations on urban CA pursuing the goal of improving their ability for simulating complex urban phenomena.

### Main Relaxations and Evolutions

It was already mentioned that the simplicity of CA is one of their great attractions for urban modeling. From a simple set of rules operating over a simple cell structure it is possible to achieve complex forms from simple structures. But the classic formulation of CA is limited in its own formal definition, imposing the necessity of relaxing some of its basic components. It can be said that the notion "cellular automata" was used in geography, from the beginning, in a very broad sense, and not as a rigid formal scheme (Benenson & Torrens, 2004). The formal framework of CA defined by their most famous researchers as Ulam, von Neumann, Conway or Wolfram, based on a very simple formulation, is far from being able to represent real cities; many, if not all, urban CA bear little similarity with classic CA. One can question whether urban CA still are evolutions of classic CA or whether they are just cellular-based models (Torrens & O'Sullivan, 2001).

There are four major adaptations of the classic CA concept: (1) most applications to urban systems relax the local neighborhood interaction to incorporate action-at-a-distance (considering this long range interaction as a global consequence of local spatial diffusion, thus reinforcing the application of classic CA); (2) it is hard to identify a scale for urban systems where everything is reducible to one activity in one cell; (3) the need of CA to meet plausible values of change rates; and (4) the use of GIS and map algebra (Batty et al., 1997).

Helen Couclelis pointed out two main characteristics that CA-based models must have: interactivity and realism (Couclelis, 1997). Interactivity is an essential property that CA-based models must comprise. Being CA models of complexity, in which a small change in the initial conditions or transition rules may produce major changes in the results, they must allow the evaluation of small changes in model conditions to make sensitivity analyses possible. CA models should also incorporate good visualization techniques, not only at the graphic level but also in the statistical characterization of the results. The integration with GIS was pointed out as the next big step, because of the natural affinity between CA and raster images, and because of the potential of graphical visualization provided by GIS. However, the more complex is the CA structure, the more difficult it becomes to produce visualization. Another important feature of CA-based models is realism. Couclelis argues that no model based on the classic assumptions of CA – homogeneity, uniformity, universality – can claim good performance when applied to real world problems (Couclelis, 1997). There are two main dimensions for model realism: realism with respect to data and realism with respect to model structure. The linkage between CA and GIS can be, once again, very profitable in order to guarantee data realism. Several CA models are already supported by GIS (Takeyama & Couclelis, 1997; Batty et al., 1999a; Li & Yeh, 2000). Structure realism is achieved through the adaptation of the classic CA structure to the specific needs of simulating complex urban behaviors. CA ability to generate complex patterns from simple cell configurations and through small sets of rules is one of its main attractions, as it has been already stated. But, in order to improve the feasibility of simulated urban landscapes, it is necessary to adapt the formalism of CA to the complexity of urban and social phenomena, adjusting the rigidity of classic CA components to the perception of the real world. Couclelis (1997) identifies a series of relaxations that can be implemented in order to enhance the ability of CA models to correctly acquire the essential behaviors of complex urban dynamics. Figure 5 depicts a series of relaxations that can be implemented for all the four main components of CA.

The majority of CA models referred to so far are based on regular square cells forming an orthogonal cell lattice. The main reason for the use of these simple spatial structures relates to land use data availability from remote sensing maps. In order to enhance spatial representativeness, it is possible to forget this formalism and to use irregular cells as the spatial unit for CA. Tobler refers that there are some analytical advantages in considering the irregular spatial division of political jurisdictions (Tobler, 1979). However, Tobler states that the basic difficulty, of topological nature, relies on the fact that these irregular cells do not all have the same number of adjacent cells, thus their neighborhoods can not be defined by any simple notational scheme. There are very few studies based on irregular cells. The model developed by Vandergue et al. (2000) is cited by Ménard & Marceau (2005) as the only CA model that uses census tracts as cells. This approach has a great potential for linking urban form and reliable data, and is on the basis of the CA model presented in the next section of this chapter. Semboloni (2000) used Voronoi polygons to model urban growth considering cell division. O'Sullivan (2001a, 2001b) developed another CA-based model coupling CA and graph theory. Benenson & Torrens (2004) refer to the existence of theoretical studies that used triangular and hexagonal cells (Gerling, 1990; Eloranta, 1997).

Cell state sets can also be adapted to simulate different land uses in different parts of the simulated area. This modification has some implications not only in the definition of neighborhoods but also in the definition of the set of transition rules.

The concept of neighborhood can also be modified. There is a problem of representativeness in the first place. The importance of the neighborhood is that it defines the geographical domain of influence (Tobler, 1979).

The fact that there are different concepts of neighborhood must be taken into account: the sense of neighborhood of residents of urban areas and residents of rural areas is different, therefore it may be useful to consider the shape, size and form of a neighborhood as a function of the location of its central cell, giving special attention to neighborhoods in boundary areas (Tobler, 1979).

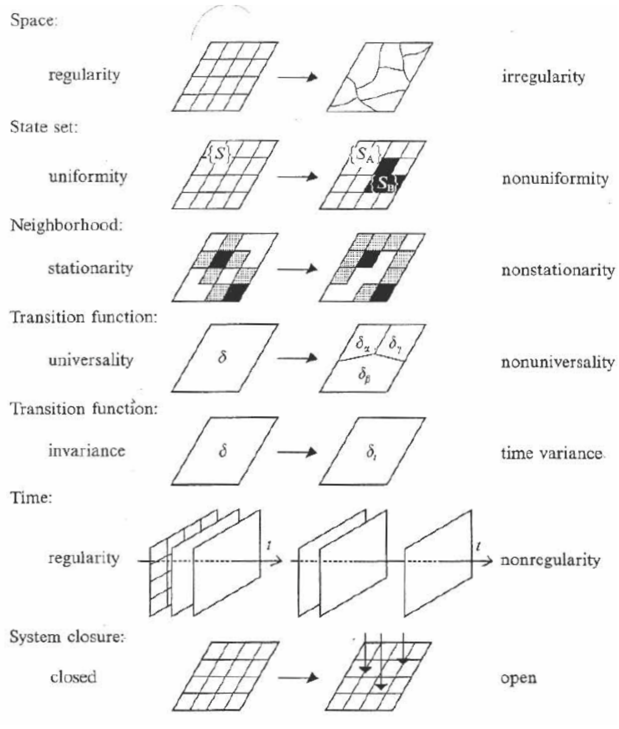


Figure 5: Possible CA relaxations (Couclelis, 1997)

Neighborhoods such as von Neumann's and Moore's are particularly suited to physical phenomena; human interactions are more likely to be explained by wider areas of influence (White & Engelen, 2000). Physical and topographic constraints can also shape the sense of neighborhood, particularly in larger urban areas. Neighborhood structure may also vary in time. There are some studies in the field of tissue and leave growth that considered the creation of new cells within a given neighborhood (Lindenmayer, 1968). Semboloni (2000) also assumed that cells could be divided, thus creating new neighborhoods. Ménard & Marceau (2005) produced an interesting study on spatial scale sensitiveness. They gathered information about the neighborhoods used in several CA models from 1993 to 2003. They defined CA's spatial scale as a set of three components: spatial extent, cell size and neighborhood configuration. They focused their study on the analysis of different cell sizes and neighborhood configurations for a simple two state problem. Small variations in cell size can produce significant variations in results when a given scale threshold is exceeded. The authors state that CA models are not sensitive to variations of neighborhood configuration. Spatial scale sensitivity affects CA models were the cell actually represents a portion of the geographic space. Ménard & Marceau (2005) also stated that, although scale issues are supposed to be considered as a design choice that will not drastically vary during the simulation process, modelers must pay more attention to the subject, because some of the features that influences neighborhood are effectively dependent on scale choice.

Transition rules also underwent major developments from their initial formulation. The main component of CA is the transition rule (Torrens, 2000). The first and most significant evolution regarded the incorporation of stochastic perturbations in transition rules. Deterministic behaviors are not suited for the simulation of complex urban phenomena because they can be considered a result of complex interactions between agents. For this reason, many CA models have incorporated stochastic perturbations in their transition rules (White & Engelen, 1993; Clarke et al., 1997; Barredo et al., 2003). Transition rules can also be considered non-static, varying through time. Their universality, an assumption of the classic CA formulation, can also be relaxed, as different areas and land uses present different behavior changes. There may be also a distinction regarding the way transition rules operate over time. They can be applied sequentially, updating cell states one after another, or in parallel, updating all cells at the same time (Benenson & Torrens, 2004). Von Neumann's self-reproducing CA presented an asynchronous behavior with the cells being updated considering the previous updates generated by the automata. Conway's and Wolfram's approaches were based on synchronous behaviors, with the entire cell set being updated at the same time. There are two main processes of asynchronous updating. The cell set can be updated at a moment in time after a predetermined order according to some cell characteristic. Another method is based on a probability of change in a given moment that is a function of the time the cell has to wait for changing.

Finally, formal CA can be considered as a black box that processes input data towards an output without any interference from the outside world, that is, they can be considered as a closed system. Urban systems are too far from being close; in fact, an urban system is one of the most opened systems that can be found. Therefore, the assumption that an urban system modeled by CA can experience exogenous interference – say stakeholders' actions or political decisions – enhances their ability to simulate urban complexity and, consequently, their representativeness. Many models deal with this relaxation by introducing stochastic perturbations that only occur when certain exogenously defined thresholds are overcome (Clarke et al., 1997).

Relaxations are needed to improve CA's ability to simulate complex urban phenomena. The use of real, disaggregate data made possible the relaxation of any one of the assumptions of classic CA, allowing the models to better fit the complexity of cities (Couclelis, 1997). However, excessive relaxation of traditional CA assumptions may increase exponentially the difficulties to understand the outcomes of a model, in a comeback to widely criticized large-scale simulations (Couclelis, 1997).

## APPLICATIONS OF CA IN URBAN CHANGE PROBLEMS

One of the most widespread CA models is SLEUTH, developed by Keith Clarke to reproduce and predict urban growth (Clarke et al., 1997; Candau, 2000; Clarke, 2002; Silva & Clarke, 2002). Its name is an acronym for Slope, Land use, Exclusion, Urban, Transportation and Hill Shade. SLEUTH has two main modules (Clarke, 2002): first, an urban growth model; second, an embedded land use model that uses information from the urban growth model. This model requires six GIS-based inputs for visualization, used in image format: urbanization, land use, transportation, areas excluded from urbanization, slopes, and hill shading. Urban areas are required for four different time periods for calibration purposes. Urbanization is the product of an urban seed file and at least two road maps that interact with a slope layer to allow the generation of new urban centers. It considers a regular grid space, a neighborhood of eight cells and only two cell states: urban and non-urban. It operates according to five sequential transition rules: (1) diffusion, (2) breed, (3) spread, (4) slope resistance, and (5) road gravity. The growth rate depends on five different factors. The diffusion factor determines the overall dispersion of the distribution of single grid cells and of the movement of new settlements outward through the road network. The breed factor determines how likely a newly generated settlement is to begin its own growth cycle. The spread factor controls how much outward "organic growth" expansion takes place within the system. The slope resistance factor influences the likelihood of settlement existence on steeper slopes. The road gravity is a factor of attraction of new settlements onto the existing road system if new areas fall within a given distance of a road (Clarke, 2002). The operation of these factors leads to four different types of urbanization. First, spontaneous growth: any non-urban cell can be urbanized according to a probability inversely proportional to cell slope. Second, generation of new diffusion centers: each spontaneously urbanized cell can become a new spreading center if it has a given number of neighboring urban cells and reaches a probabilistic threshold defined as a model parameter. Third, diffusion at the edges of urbanized areas: there is a fixed probability (another parameter of the model) that allows an edge cell to become urbanized given a certain number of neighboring urban cells. Fourth, road-influenced diffusion: a new spreading center is chosen given its distance to the road network. It can be dislocated along that road in a randomly selected direction for a given distance (another parameter of the model) for a new location where it can "paste" development. Two randomly chosen neighboring cells would then change state. An important improvement of this model is the consideration of self-modification rules aimed to modify the model's behavior over time giving it the ability of identifying different periods of intensive growth or of little or no growth.

After the first phase of urban growth modeling, SLEUTH will assign land uses to the new urban areas obtained from the first phase of the model. This assignment is made through an embedded model named Deltatron. The main assumption is to consider each cell as a Deltatron, an urban entity that is associated to only one cell state (or land use) (Benenson & Torrens, 2004). These entities evolve during simulation in a way similar to the one described for urban growth, through four stages. Firstly, cells are selected at random as candidate locations for land use change on the basis of how much urban growth has taken place. Each newly urbanized cell is assumed to induce a potential change in land use and, as a result, determines whether the selected cell will keep the same land use or change to another one; this is made through the consideration of a probability that depends on historical change and cell slopes. If a transition occurs and the cell is not a Deltatron already, a new Deltatron is created. Cluster dynamics are defined as an aggregation process of these new Deltatrons and the associated land use transition. The newly transitioned cell acts now as the land use aggregation center. This process behaves closely to the "organic" growth described above. Age is also considered to characterize Deltatrons: as time goes by, Deltatrons get older and are eliminated when their age reaches a given threshold. SLEUTH gets credit for introducing the concept of evolving transition rules with the aim of improving the model's ability to retrieve past behaviors. It is also a straightforward model that can be easily calibrated to different urban and regional areas with a small set of calibration parameters. There are already an important number of applications not only in the United States but also in Europe, Africa and South America.

Another innovative approach was proposed by Li & Yeh (2001). They formulated a CA-based model that uses artificial neural networks (ANN) to simulate urban change. It has been applied to the Pearl River Delta, China. The set of parameters is obtained automatically by the ANN method, which is made more robust because it uses a back-propagation training procedure. The authors consider that the traditional CA formulation has serious problems in obtaining consistent parameters values. The method is believed to be suited for dealing with complex interactions between dependent variables without user involvement, a major issue in urban modeling. The model is capable of analyzing relevant data, and eliminating noise and other redundant data. Another important development of this model is related to transition rules: they are obtained from the ANN training process and not from user's definition. The model only needs to be fed with training data. It is also capable of dealing with uncertainty by simulating alternative developments through network training, integrating potential interventions exogenous to the change process which are very difficult to be deducted from historical data. Basic model structure is depicted in Figure 6.

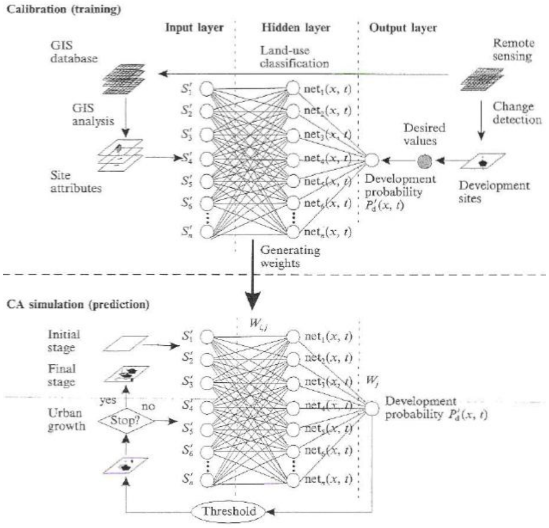


Figure 6: CA model based on artificial neural networks (Li & Yeh, 2001)

O'Sullivan (2001b) proposed a model that couples CA with graph theory to enable research into relationships between spatial structure (represented by graphs) and urban dynamics (simulated by CA), as depicted in Figure 7. The model is based on a partition that needs not be space-filling, and overlapping spatial elements might be used. In order to study urban phenomena, the entities forming the basis of urban morphology (buildings, blocks, streets, census blocks, and administrative/planning zones) are a natural set of spatial elements to be used as graph vertices.

Edges in the graph represent some sort of relationship between vertices, so that any relationships relevant to the model being developed might be used. The neighborhood of each vertex consists of a set of adjacent vertices. Cell states may be defined in a suitable way, considering land use classes as large as necessary to correctly simulate reality. The graph/CA model was applied to study gentrification in small local neighborhoods in central London. Data availability and quality were considered problems for using this approach to micro-scale modeling. The dynamic behavior of complex models was pointed out as an important shortcoming to the application of micro-scale modeling to gentrification problems.

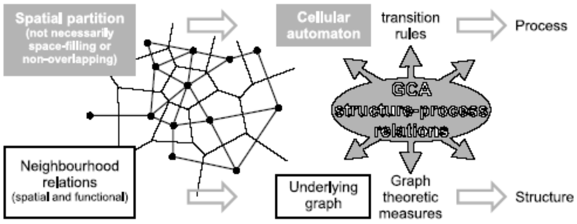


Figure 7: Graph-CA model concept (O'Sullivan, 2001b)

Barredo et al. (2003) developed a constrained CA model named MOLAND, which is an integrated platform that includes a series of macro-scale models for simulating the natural, social, and economic sub-systems that are the main drivers of land use evolution at this scale. These models are coupled with a classic CA model for land use allocation at a local scale, following the approach of White & Engelen (1993). MOLAND uses the concept of transition potential introduced by these authors to build a model based on a regular cell grid with a radial neighborhood of 172 cells. They also considered a neighborhood effect materialized through a user-defined weighting factor matrix, establishing quantitative values for attraction and repulsion, similar to the concept depicted in Figure 4. The model was applied to a series of regions in Europe and also Africa, particularly for supporting risk assessment analyses. One of the case studies was Dublin, for which a thirty-year long historic period (1968 to 1998) was used for calibrating the simulation. Using fractal measures and contingency matrices as goodness of fit functions, the model was able to achieve interesting results. Simulated fractal measures for each land use were quite similar to the real values, and the values for the statistical indicators derived from the contingency matrices can be considered satisfactory.

## CELLULAR AUTOMATA APPLICATION FOR SMALL URBAN AREAS

Nuno Pinto (Pinto, 2006; Pinto & Antunes, 2007) proposed a CA model for simulating growth of small-size urban areas based on the use of irregular cells. These two issues – the use of CA for simulating small urban areas and the use of irregular cells – have barely been addressed so far. There is a great potential on using irregular cells because they generate cell lattices that are highly representative of urban structures. The cell space was modeled considering the intersection between census blocks (for different reference years) and urban areas, defined by settlement boundaries. The model also deals with land use demand from an innovative standpoint. Land use demand depends both on population growth and on the variation of construction density. These two indicators were considered by using population densities (which also vary from one reference year to another) as a measure of land use demand. The model distributes the increase of population over the territory (thus generating the growth of urban areas) rather than sorting out a series of cells for state change based only in a given probability. State transition depends on a measure of transition potential, considered as the calibrated sum of three different components: cell accessibility, land use suitability, and neighborhood interactions between different land uses.

### Model Structure

The cell space is formed by a lattice of irregular cells. Irregularity was modeled through the use of Voronoi polygons for simulating real-world irregular census blocks. On the one hand, census blocks are highly representative of urban structure and, on the other hand, they hold structured and reliable information on demographics and construction. The model works with a set of six aggregate cell states (or land use classes): (1) urban low density (UL), (2) urban high density (UH), (3) industry (I), (4) non-urbanized urban areas (XU), (5) non-urbanized industrial areas (XI), and (6) areas where construction is highly restricted (R). Besides cell state, a set of other cell attributes are defined: cell accessibility, cell suitability, and neighborhood effect. Accessibility is assessed through a calibrated measure of the distance from the cell to the main functional centers of the territory: municipality, civil parish, and main industrial district. Cell suitability is linked to zoning and is assumed to be a binary value that takes the value of 1 if the cell is suitable for a given land use and 0 otherwise. Neighborhood was considered variable in space, discarding the use of local, therefore small, neighborhoods: its dimension was assumed as a model parameter. Another component of the model is the neighborhood effect. It was considered as an aggregate value of the interaction between one cell state (or land use) of a given cell and another cell state of another cell within a certain distance. These linear relationships decay to 0 (no interaction) as the distance between two land use locations (or cells) increases. This value varies from -1 (total repulsion) to 1 (total attraction) and is zero (0) if they do not interact. Transition rules are applied through the consideration of a transition potential. The potential function is a calibrated value of accessibility, land use suitability, and neighborhood effect:

$$P_{i,s}{}^{t} = \left(\chi_P \times A_i + v_P \times S_{i,s}{}^{t} + \theta_P \times N_{i,s}{}^{t}\right) \times \xi; \; i \in C; \; s \in CS \qquad (7)$$

where, for each cell *i* from cell set *C* and for each state *s* from cell state set *CS*, $P_{i,s}{}^{t}$ is the transition potential for cell state *s* of cell *i* at time step *t*, $A_i$ is the accessibility measure for cell *i* (constant during the entire simulation period), $S_{i,s}{}^{t}$ is the suitability value for cell state *s* of cell *i* at time step *t*, $N_{i,s}{}^{t}$ is the neighborhood effect for cell state *s* of cell *i* at time step *t* considering its neighborhood $V_i$, $\chi_P$ is the calibration parameter for accessibility, $v_P$ is the calibration parameter for suitabilities and $\theta_P$ is the calibration parameter for the neighborhood effect. $\xi$ is the stochastic perturbation (similar to the one described by Equation 4). The model deals with land use demand in a perspective different from usual CA models. Land use demand is proportional to the increase in population, as well as to the variation of construction density: the accounting of new public facilities and spaces offered by newly built areas, in result of more demanding planning regulations, usually leads to a decrease on construction density throughout the years. The modeling time step was of ten years, equal to the length of the period covered by the historical data available. The choice is based on two main issues: census data availability and land use dynamics. The assessment of the model performance was made using contingency matrices and corresponding *kappa* index (Couto, 2003). The comparison of the simulation map with the reference map through a measure of similarity is a good comparison measure because it is oriented for the analysis of the entire territory as a distributed structure and not only as a centralized urban layout. But there are urban land uses (cell state R in the present classification) that were not considered in the changing dynamics. The consideration of the entire set of cell states for the calculation of the *kappa* index would produce a distortion on its significance. To avoid this distortion, a modification of the *kappa* measure was considered, named $k_{Mod}$, where only the cell states that take part in the urban change dynamics are considered.

### Calibration with Particle Swarm

The calibration of a model can be achieved through two main approaches. The first approach is based on the use of a sensitivity analysis procedure for each parameter, in which the value for the parameter is changed considering the other parameters fixed. The second approach is based on the use of optimization to search for the set of calibration parameters that optimize a fitness function chosen for the model. Sensitivity analysis is based on a group of procedures that take place after the simulation; it becomes difficult to apply as the number of parameters increases. The main reason for using an optimization approach is to ensure an extensive search of the parameters, leading to the best possible values for the parameters given the fitness function for the CA model. For the problem at hand, the optimization method chosen was Particle Swarm (hereafter referred to as PS). This new optimization algorithm has been giving promising results for complex optimization problems. The concept of this optimization technique is to have a swarm of particles, each one representing a point of the multi-dimensional space of solutions. Each particle is an independent set of calibration parameters which correspond to an independent CA run. This swarm will "fly" throughout the solutions space, having the location of its particles updated by a simple set of position and velocity equations that slightly change the sets of CA calibration parameters. The process evolves during a pre-determined number of iterations and converges to a quasi-optimum solution for the objective function, the set of CA calibration parameters that better fits simulation with reality. In the present case, the objective function is to maximize the similarity between simulation and reference maps, measured by $k_{Mod}$. It is an optimization paradigm that simulates the ability of human societies to process knowledge (Kennedy, 1997). The member's movement in a swarm is the result of the individual effort to maintain an optimum distance between himself and his neighbors (Parsopoulos & Vrahatis, 2002). The individuals' successes influence their searches and those of their peers (Kennedy, 1997). Interaction in these swarms enhances progress towards a solution, as members (or particles) benefit from their own knowledge and from their neighbors' knowledge. In this technique the aim is not the survival of the fittest but the joint effort of the swarm in finding the best solution.

### Application to a Set of Test Instances

The model was applied to a series of theoretical test instances to assess its behavior for a variable set of problems. These test instances were designed to reproduce real-world spatial structures similar to small Portuguese municipalities. A set of twenty test instances was created simulating small municipalities not only in size but also in the total number of cells (two of which are depicted in Figure 8).

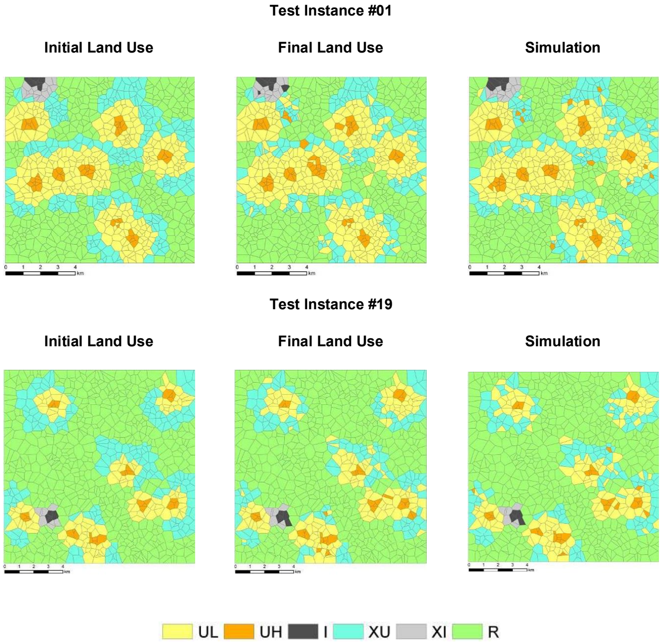
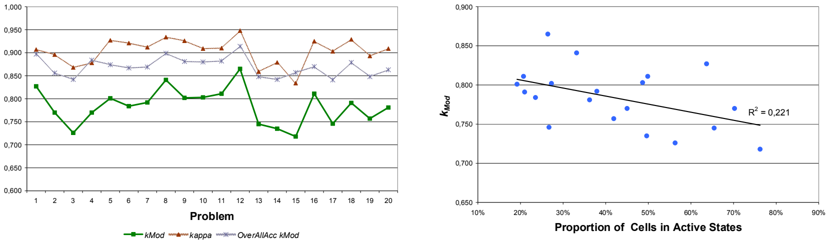


Figure 8: Two examples of test instances

The model was able to reach a remarkable level of agreement between simulation and reference maps. Global $k_{Mod}$ results for the entire set of instances are depicted in Figure 9(a). These results can be considered as very good for a simulation process: 50 percent of the problems achieved a $k_{Mod}$ around 0.80 or higher and 75 percent of them exceeded 0.75. As it was explained before, $k_{Mod}$ is a measure of agreement between simulation and reference maps which does not take into account inactive cell states. Figure 9(a) also presents the variation of the absolute *kappa* measure for the set of test problems. For 65 percent of the problems, the agreement exceeded 0.90 and 95 percent exceeded 0.85. These values are commonly accepted as very good agreement between simulated and reference situations (Barredo et al., 2003). Note that the difference between the values of *kappa* and $k_{Mod}$ is often higher than 0.10, justifying the choice of the use of $k_{Mod}$. The model was also efficient at simulating the total area consumed for each cell state. The average ratio between total simulation area and total reference area for each cell state was of about 1.0 percent for UL cell state, -5.0 percent for UH cell state, and -1.0 percent for XU cell state. For industrial land uses this value increased due to the fact that only one cell out of two or three cells have changed state. A tenuous relationship with a Pearson coefficient of 0.47 can be established between the proportion of active cells and the performance of the model: the smaller this proportion is the better the model performs Figure 9(b).

The agreement produced by simulation was also assessed by the number of cells whose change was matched in simulation and in reality. For 80 percent of the test problems, matching proportions of 35 percent or higher were achieved; for 20 percent of the problems the value of 50 percent of matches was exceeded. These results can also be considered satisfactory, because from visual observation it is clear that the model chose a cell close to the cells that had really changed state.



(a) (b)

Figure 9: (a) Global $k_{Mod}$ and *kappa* results for the set of test problems and (b) Relationship between the proportion of active cells and model performance

### Application to a Real-World Case Study

The results observed from the application to the set of test instances gave us good insights into the use of CA models for simulating urban growth in small urban areas. The model was then applied to a real-world case study – the municipality of Condeixa-a-Nova, Portugal – chosen among a set of small Portuguese municipalities that presented high population and built area growth rates over the past two decades.

The results achieved for the case study were worse than those achieved for the test instances: the $k_{Mod}$ was only 0.621 (Figure 10a). However, it may be also considered a promising result because, on the one hand, this lower value is in line with the trend deduced for the set of test problems (the larger the problem is, the worse the agreement is), and because, on the other hand, it is close to commonly accepted thresholds, for example for remote sensing agreements.

The model also produced good results for two ten-year simulation runs that were used to test its ability to generate plausible future land use scenarios. Starting from the real land use configuration in 2001 and considering a constant growth scenario for both population and employment, the model simulated a densification process around the main urban areas, a behavior that can be considered in line with the expected growth under current planning regulations (Figure 10b).

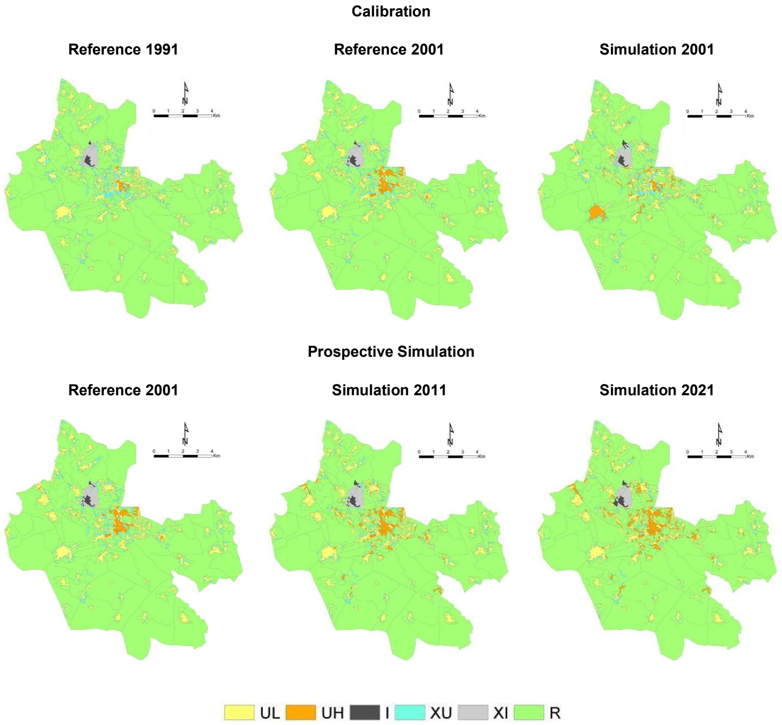


Figure 10: Application to Condeixa-a-Nova, Portugal – (a) Calibration; (b) Prospective Simulation

## FUTURE TRENDS IN CELLULAR AUTOMATA

Cellular automata are an important subject of interest in urban simulation. The concept of CA became even more important with the great developments that multi-agent systems (MAS) experienced in the past years, by considering cells as spatially located agents with well defined behaviors. Cells are non-moving agents that are properly suited for simulating a series of phenomena that are located in space. The consideration of cells as agents requires more research on some of the main concepts of CA, such as scale, cell form, neighborhood, and transition rules. Some hints are pointed out below for the pursuit of these new research paths.

### Multi-Scale Cellular Automata

Different urban phenomena can be observed and modeled considering, on the one hand, cities as parts of larger regional systems and, on the other hand, cities themselves, where the driving forces of urban change depend on local variables. The issue of scale is currently under debate and some models already face the problem of inter-scale interaction (Kocabas & Dragicevic, 2006; Benenson, 2007; White, 2007). The use of both regional and local scales is considered useful to correctly simulate a wide set of complex urban phenomena that occur on different spatial scales. Considering that these demographic and economic relationships depend largely on neighborhood conditions, and that large scale cells can be observed with a considerable amount of reliable data (counties, municipalities, cities), it is clear that a CA model may be used to simulate urban evolution at a regional, therefore macro-scale. At this scale, the assessment of aggregate land use demand – for housing, industrial, or tertiary land uses – through population and employment growth can be more representative of urban growth than the disaggregate amount determined for each land use, which is traditionally used by common CA models. In opposition, at a city/neighborhood scale – the local scale – land use growth outcomes from the distribution of each land use (considered on a disaggregate level). In fact, land use demand can be estimated as the amount of land necessary for each land use considering population and employment growth. Then, a set of disaggregate land uses can be assigned to different cells in order to meet land use demand. This multi-scale approach is the consideration of two levels of simulation (as depicted in Figure 11): the regional level, for which the aim is to assess land use demand from population and employment evolution through time; and the local level, for which the aim is to assign different land uses to different locations (cells or parts of cells) in order to meet land use demand.

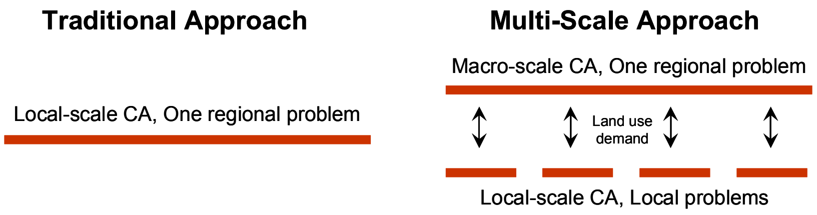

Traditional Approach
Local-scale CA, One regional problem
Multi-Scale Approach
Macro-scale CA, One regional problem
Land use demand
Local-scale CA, Local problems

Figure 11: Multi-scale CA structure versus traditional CA approach

### The Use of Irregular Cells

Common CA models operate with identical square cells that are usually obtained from remote sense maps (White & Engelen, 1993; Clarke et al., 1997; Barredo et al., 2003; Li & Liu, 2006). This approach was the natural evolution from the 2D mathematical CA models and is useful for modelers because it allows the collection of reliable historical data on land use, irrespective of the ability of this framework for fitting real-world processes (Benenson, 2007). However, these cells are not directly related to urban form because of their regularity. The use of irregular cells has implied the idea of capturing the natural irregularity that can be observed on the structure of a large majority of urban areas around the world. The number of CA models that use irregular cells is small (Semboloni, 1997; Vandergue et al., 2000; O'Sullivan, 2001b; Pinto & Antunes, 2007; Stevens et al., 2007). The potential of this approach derives from the possibility of combining urban form and reliable data (Pinto & Antunes, 2007). At a regional scale, municipalities or other intermediate administrative units will in principle be appropriate cells; at a local scale cells should be as close to urban structure as possible. Census blocks can be a good option (they proved to be feasible for the CA application to small urban areas presented before) since they are normally drawn considering the urban structure, and contain extensive and reliable information. There is indeed a large amount of processed data of various types both at a regional and a local scale. The possibility of crossing reliable data and spatial structure is the main gain obtained from using irregular cells.

### Dynamic Neighborhoods

Neighborhood is a critical issue for every type of spatial models, and in particular for CA. Neighborhood is commonly (if not exclusively) considered by the strict concept inherited from the mathematical formulation of CA. This concept is based only on the consideration of a set of physical neighbors to one cell: these neighboring cells can be those which are directly connected to the cell considered, or they can be the group of cells that are within a given range from that cell (White & Engelen, 1993; Benenson & Torrens, 2004; Ménard & Marceau, 2005). But this mathematical perspective is far from being representative of how cities work. The concept of neighborhood must be able to reproduce how agents interact, considering both spatial and functional levels of interaction. At a regional level, neighborhood is influenced by both these levels of interaction. A municipality is strongly influenced not only by its directly connected neighbors – in terms of employment and trade – but also by the main functional centers of the region, where administrative and economic decision centers are located. At a local level, this dual influence is also observed. The choice of location for a given land use is influenced both by the surroundings (a good example is the relationship between residential and industrial land uses) and by the distance to the main functional and employment centers in the city. Neighborhood must shift from the concept of a limited area which remains constant in time to a larger and possibly disconnected part of the territory, varying in time. A cell can have a discontinuous neighborhood that includes its immediate neighbors at a certain extent and some distant functional centers (regional capital, main regional cities) as depicted in Figure 12. This neighborhood can vary during the years, enlarging or decreasing proportionally to the potential of attraction of the cell. This concept of neighborhood implies the consideration of spatial and functional interaction between cells.

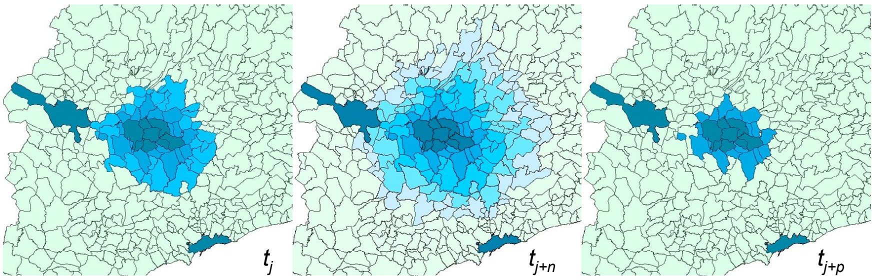


Figure 12: Dynamic neighborhoods

The main goal is to enhance neighborhood representativeness at both regional and local scales. There are physical and ecological barriers that limit or break the shape of a neighborhood. At a regional scale, a mountain, a river, or a delta can act as a physical barrier for many types of interactions. The same occurs at a local level with topographic features, heavy urban infrastructures (road and railways), and city form acting as barriers for defining the perception of neighborhood.

### Adjusting Transition Rules to Different Scales of Interactions

Transition rules are also an important component of CA. They define the way a CA model evolves through time and they are intimately related to the way the model is able to correctly acquire behaviors that are the drivers of the simulation. Classic CA uses probabilistic sets of transition rules (Wolfram, 1984). Although this probabilistic approach has been used in some theoretical urban CA models, transition rules are commonly designed to reproduce complex urban behavior (White & Engelen, 1993; Clarke et al., 1997; Barredo et al., 2003). These rules can also be used to reproduce planning regulations that constrain land use change and demand (Pinto & Antunes, 2007). A multi-scale approach requires some attention on the definition of transition rules for both scales. At a regional level of analysis, it is important to notice that the goal is to simulate the macro interactions that are observed within a region. These interactions can be assessed using macro-scale indicators of population, employment, and commuting flows, for example. Cell states should relate to these indicators and can be defined as a degree of urbanization taken in aggregate or disaggregate form. A binary state of urbanization, occupied or non-occupied, can be used; an alternative approach can use a disaggregated degree of urbanization, in percentage of the available land. It is also possible to consider the main land uses (residential, industrial) in an aggregate or disaggregate way. At a local scale, the goal is to simulate the distribution and growth of land use using a more disaggregate classification. These phenomena depend on a series of factors related with land suitabilities, accessibility, land use demand, and land management policies, among many others. At this scale, transition rules will relate to measures of urban potential for each land use, considering all the interactions between all the land uses, observing planning constraints that can be derived from planning regulations. Cell states are the traditional set of disaggregate cell states that are commonly used in classic CA models.

### Considering Policy Testing in CA

Policy testing is widely recognized as an important feature that must be taken into account in the development of urban simulation. The use of simulation by itself is often considered redundant because it tends to produce results that are sometimes considered futile by planners and decision makers. Simulation must be able to incorporate the practical needs of planners and of the planning process. There are different degrees of implementation of simulation-based methodologies across the world, which relates with different contexts of planning, regulation, and available information. Simulation is oriented toward understanding and reproducing real phenomena in a controlled environment, aiming to explain how these phenomena work and how they can be dealt with. However, these features do not provide planning with the necessary tools for dealing with issues that are strongly influenced by uncertainty. It is necessary to enhance the ability of models in order to use the parameters that characterize complex behaviors, allowing the consideration of policy testing on the generation of simulated scenarios and evolutions. The consideration of policy testing in CA is made using transition rules to reproduce policy constraints. This process is based on the identification of candidate policies to be tested by the model, by considering transition rules that can incorporate the parameters that control thoses policies. Its implementation must be an ongoing process that is expected to last even after the end of the model development stage: every time a new policy is set to be tested it should be possible to change the set of transition rules in order to meet specific simulation needs. This type of simulation is strongly dependent on the historical data gathered for a given problem. Every simulation technique can be considered captive of this data, reducing its ability of forecasting an uncertain future. This is probably the most important shortcoming that is pointed out to modeling by common planners. Therefore, it is important to develop modeling tools that are able to correctly identify historical trends and to break them under given conditions. Major urban transformations are usually the result of single decisions localized in time as the large urban renovation operations or the organization of important sporting events (for example the Olympic Games). These types of transformations are almost impossible to capture by any model because they cannot be derived from historical data. CA work with a set of transition rules that are calibrated considering the historic evolution of urbanization. The model must be able to generate new transition rules after observing a trend that can not be explained by past evolution. Another possibility is the calibration of transformation thresholds. An illustrative example is the change of a cell to land uses that are forbidden by planning regulations. Once those thresholds are crossed, a given cell should be allowed to change to cell states that were forbidden before.

### Land Suitability Indicators for Cell Differentiation

Land suitability is also an important issue to attend. Different land uses demand different land suitabilities for choosing a given location. Although this is not a classic CA component, it assumes great importance when we are dealing with constrained models. A model must be able to correctly differentiate two neighboring cells in terms of their ability for allocate different land uses, which is a crucial issue for improving representativeness. At a regional level, the analysis is made from an aggregate point of view and land suitability is seen as a measure of comparison for general environmental quality or environmental protection policies. It is important to capture the influence of general environmental and physical characteristics in location choice, both for residential and for non-residential land uses. There are municipalities that have more demand for residential land use due not only to their environmental characteristics but also to the distance to large industrial districts or polluted areas. At a local level, the comparison between land suitabilities for every land use is decisive for assessing demand for different and/or competitive land uses. Land suitability is an important constraint to the occupation of a given land plot by a given land use. Therefore, at this level of analysis it is imperative to develop a robust set of land suitability indicators that include physical features and, possibly, measures of levels of service for basic urban infrastructures as water supply or solid waste collection.

### Accessibility Measures for CA Models

Accessibility is strongly linked to land use. Any attempt to simulate complex behaviors that occur within an urban system must take into account accessibility, as it may be considered as one of the most important drivers of urban growth. Accessibility is also strongly dependent on the scale of analysis. In order to correctly simulate accessibility conditions and evolution throughout time, different transportation modes must be considered and their scopes of influence must be attended. Air and maritime transportation modes have a regional scale of influence, as the number of network nodes located in a given region is always only a few. On the contrary, road and rail transportation modes have both a regional scope (regional highways and roads and railroads systems) and a local one (city road network, subway rail systems). Future research will probably focus on the development and testing of different measures of multi-modal accessibility that can be used as an input for a CA model at different scales of simulation.

## CONCLUDING REMARKS

Cellular automata (CA) models are still in the front line of urban simulation techniques. After more than two decades of intensive research on geographic CA, following the pioneer work of Helen Couclelis (Couclelis, 1985), CA has already a solid scientific background that projects its use into the future of simulation. Their intrinsic spatiality guarantees that the concept has an important role to play not only as a modeling tool by itself, but also coupled with other simulation concepts and techniques, such as multi-agents modeling (MAS). Furthermore, this spatiality can prove useful for developing decision support tools for assisting common planning processes, because CA is able to easily operate over land use maps, which are one of the most wanted tools for planners and decision-makers. This feature fosters the use of CA-based simulation within the GIS framework. But to guarantee that CA continues to appear as an interesting concept for both planners and modelers to develop modeling-based planning approaches, it is crucial that the concept evolves towards a more realistic level of simulation by opposition to the more strict mathematical formulation traditionally followed. The research on CA's components is fundamental to achieve this goal. Cells, neighborhood and transition rules still have an important margin for research and development. There are several issues that depend on innovative computational approaches. Calibration, a key issue in modeling, can also be improved through the use of more sophisticated algorithms. New challenges are posed to CA and open new perspectives on their use in simulation. The coupled use of CA with MAS is probably the most obvious evolution of CA, creating the new concept of Geographical Automata Systems (GAS) proposed by Benenson & Torrens (2004). This integration provides a very powerful tool for modeling the physical and socio-economic phenomena that are the main drivers of urbanization, by linking spatial and non-spatial agents, enhancing our capacity for capturing and understanding complex behaviors. This creates new possibilities in forecasting future evolutions of urban growth by ensuring higher degrees of realism and detail. CA's simplicity and spatiality ensure them, as a standalone concept or coupled with other modeling techniques, a promising future.

## REFERENCES

**Barredo, J.; Kasanko, M.; McCormick, N.; Lavalle, C.,** 2003, Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata. In: Landscape and Urban Planning, 64: 145-160

**Batty, M.,** 1994, A chronicle of scientific planning - The Anglo-American modelling experience. In: Journal of the American Planning Association, 60, 1: 7-16

**Batty, M.,** 2005, Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. MIT Press: Cambridge, MA-London

**Batty, M.; Couclelis, H.; Eichen, M.,** 1997, Editorial: Urban systems as cellular automata. In: Environment and Planning B: Planning and Design, 24: 159-164

**Batty, M.; Xie, Y.,** 1997, Possible urban automata. In: Environment and Planning B: Planning and Design, 24: 175-192

**Batty, M.; Xie, Y.; Sun, Z.,** 1999a, Modeling urban dynamics through GIS-based cellular automata. In: Computers, Environment and Urban Systems, 23, 3: 205-233

**Batty, M.; Xie, Y.; Sun, Z.,** 1999b, Modelling urban dynamics through GIS-based cellular automata. In: Computers, Environment and Urban Systems, 23: 205-233

**Benenson, I.,** 2007, Warning! The scale of land-use CA is changing! In: Computers, Environment and Urban Systems, 31, 2: 107-113

**Benenson, I.; Torrens, P.M.,** 2004, Geosimulation - Automata-based modeling of urban phenomena. John Wiley & Sons Ltd: Chechester

**Candau, J.,** 2000, Calibrating a cellular automaton model of urban growth in a timely manner. Paper presented at the 4th International Conference on Integrating GIS and Environmental Modeling, GIS/EM4), Banff, Alberta, Canada, September 2-8

**Clarke, K.,** 2002, Land Use Change Modeling Using SLEUTH. Advanced Training Workshop on Land Use and Land Cover Change Study, Taiwan, National Central University/National Taiwan University/ START, December 9-20th

**Clarke, K.; Hoppen, S.; Gaydos, L.,** 1997, A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. In: Environment and Planning B: Planning and Design, 24: 247-261

**Couclelis, H.,** 1985, Cellular worlds: a framework for modelling micro-macro dynamics. In: Environment and Planning A, 17: 585-596

**Couclelis, H.,** 1997, From cellular automata to urban models: new principles for model development and implementation. In: Environment and Planning B: Planning and Design, 24: 165-174

**Couto, P.,** 2003, Assessing the accuracy of spatial simulation models. In: Ecological Modelling, 167: 181-198

**Eloranta, K.,** 1997, Critical growth phenomena in cellular automata. In: Physica D: Nonlinear Phenomena Lattice Dynamics, 103, 1-4: 478-484

**Gerling, R.W.,** 1990, Classification of triangular and honeycomb cellular automata. In: Physica A: Statistical and Theoretical Physics, 162, 2: 196-209

**Kennedy, J.,** 1997, The particle swarm: Social adaptation of knowledge. In: IEEE International Conference on Evolutionary Computation: 303-308

**Klosterman, R.,** 1994, Large-Scale urban models - Retrospect and prospect. In: Journal of the American Planning Association, 60, 1: 3-6

**Kocabas, V.; Dragicevic, S.,** 2006, Assessing cellular automata model behaviour using a sensitivity analysis approach. In: Computers, Environment and Urban Systems, 30, 6: 921-953

**Lee, D.,** 1973, Requiem for large-scale models. In: Journal of the American Planning Association, 39, 3: 163-178

**Levy, S.,** 1992, Artificial life: the quest for a new creation. Pantheon Books: New York

**Li, X.; Liu, X.,** 2006, An extended cellular automaton using case-based reasoning for simulating urban development in a large complex region. In: International Journal of Geographical Information Science, 20, 10: 1109 - 1136

**Li, X.; Yeh, A.,** 2001, Calibration of cellular automata by using neural networks for the simulation of complex urban systems. In: Environment and Planning A, 33: 1445-1462

**Li, X.; Yeh, A.G.O.,** 2000, Modelling sustainable urban development by the integration of constrained cellular automata and GIS. In: International Journal of Geographical Information Science, 14, 2: 131-152

**Lindenmayer, A.,** 1968, Mathematical models for cellular interaction in development, Parts I and II. In: Journal of Theoretical Biology, 18: 280-315

**Ménard, A.; Marceau, D.J.,** 2005, Exploration of spatial scale sensitivity in geographical cellular automata. In: Environment and Planning B: Planning and Design, 32: 693-714

**O'Sullivan, D.,** 2001a, Exploring Spatial Process Dynamics Using Irregular Cellular Automaton Models. In: Geographical Analysis, 33, 1: 1-18

**O'Sullivan, D.,** 2001b, Graph-based cellular automaton models of urban spatial processes. Thesis. Bartlett School of Architecture and Planning, University College, London, UK

**Parsopoulos, K.E.; Vrahatis, M.N.,** 2002, Recent approaches to global optimization problems through Particle Swarm Optimization. In: Natural Computing, 1: 235-306

**Pinto, N.; Antunes, A.,** 2007, A cellular automata model for the study of small urban areas. In: Proceeedings of the 15th European Colloquium on Theoretical Quantitative Geography, Montreux. University of Lausanne: Lausanne: 303-309

**Pinto, N.,** 2006, A microsimulation approach for modelling the growth of small urban areas. MSc Dissertation. Department of Civil Engineering, School of Engineering of the University of Porto, Porto, Portugal

**Semboloni, F.,** 1997, An urban and regional model based on cellular automata. In: Environment and Planning B-Planning & Design, 24, 4: 589-612

**Semboloni, F.,** 2000, The growth of an urban cluster into a dynamic self-modifying spatial pattern. In: Environment and Planning B-Planning & Design, 27, 4: 549-564

**Silva, E.; Clarke, K.C.,** 2002, Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. In: Computers, Environment and Urban Systems, 26: 525-552

**Stevens, D.; Dragicevic, S.; Rothley, K.,** 2007, iCity: A GIS-CA modelling tool for urban planning and decision making. In: Environmental Modelling & Software, 22, 6: 761-773

**Takeyama, M.; Couclelis, H.,** 1997, Map dynamics: integrating cellular automata and GIS through Geo-Algebra. In: International Journal of Geographical Information Science, 11,1: 73-91

**Tobler, W.,** 1970, A Computer Movie Simulating Urban Growth in the Detroit Region. In: Economic Geography, 46, 2: 234-240

**Tobler, W.,** 1979, Cellular geography. In: Gale, S.; Olsson, G. (Eds), Philosophy in Geography. Boston: D. Reidel: 379-386

**Torrens, P.,** 2000, How cellular models of urban systems work, 1. Theory. CASA Working Papers Series, UCL, London

**Torrens, P.; O'Sullivan, D.,** 2001, Editorial: Cellular automata and urban simulation: where do we go from here? In: Environment and Planning B: Planning and Design, 28: 163-168

**Vandergue, D.; Treuil, J.-P.; Drogoul, D.,** 2000, Modelling urban phenomena with cellular automata. In: Ballot, G.; Weisbuch G. (Eds), Applications of Simulation to Social Science. Stanmore, Middlesex: Hermes Science Publishing: 127-140

**Ward, D.; Murray, A.; Phinn, S.,** 2003, Integrating spatial optimization and cellular automata for evaluating urban change. In: The Annals of Regional Science, 37: 131-148

**White, R.,** 2007, Multi-scale modelling with variable grid CA. In: Proceeedings of the 15th European Colloquium on Theoretical Quantitative Geography, Montreux. University of Lausanne: 443-449

**White, R.; Engelen, G.,** 1993, Cellular Automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. In: Environment and Planning A, 25: 1175-1199

**White, R.; Engelen, G.,** 1997, Cellular automata as the basis of integrated dynamic regional modelling. In: Environment and Planning B: Planning and Design, 24: 235-246

**White, R.; Engelen, G.,** 2000, High-resolution integrated modelling of the spatial dynamics of urban and regional systems. In: Computers, Environment and Urban Systems, 24: 383-400

**Wolfram, S.,** 1983, Statistical mechanics of cellular automata. In: Reviews of Modern Physics, 55, 3: 601-644

**Wolfram, S.,** 1984, Computation theory of cellular automata In: Communications in Mathematical Physics, 96, 1: 15-57

**Xie, Y.,** 1996, A generalized model for cellular urban dynamics In: Geographical Analysis, 284: 350-373

**Zeigler, B.P.,** 1976, Theory of Modeling and Simulation. Wiley: New York

## AUTHORS INFORMATION

**Nuno N. PINTO**
npinto@dec.uc.pt
Department of Civil Engineering
University of Coimbra
Portugal

**António P. ANTUNES**
antunes@dec.uc.pt
Department of Civil Engineering
University of Coimbra
Portugal

**Josep ROCA**
josep.roca@upc.edu
Center for Land Policy and Valuation
Technical University of Catalonia
Spain

# QUALITATIVE METHODS IN GEOGRAPHY AND PLANNING

**Giovanni A. RABINO and Matteo CAGLIONI**

Department of Architecture and Planning, Polytechnic of Milan, Italy

## ABSTRACT

In human sciences including geography, qualitative methods (such as interviewing, content analysis, qualitative data analysis, etc.) have a long tradition, that often has been considered alternative - or, at least, - independent from quantitative methods. But, recently, the awareness of the fallacy of this dichotomy has appeared, giving space to a complementarity of methods or to an unified vision, that is the perspective adopted in this contribution. The paper therefore begins with an analysis of reasons for this change, associated, not by chance, to an increase in use of qualitative methods. Operationally, this is due to a shift in computer use: from simply “number crunching” to performing several tasks (data handling, executing logical algorithms, etc.). More generally, this technical change is induced by the “complex systems” epistemological perspective, where “the qualitative” is re-valued as a signature of complexity.

It is shown that, in order to deal in a rigorous way with qualitative information, we can try many tricks. In some cases, the large memories and the power of computation of the machine are enough for facing problems with qualitative data, otherwise non treatable; in other cases, this goal is reached by the possibility of “visualization” of information or “cooperative solution” (through communication) allowed by the computer. Moreover, in some cases the qualitative is managed by means of algorithms (carrying out a sequence of “logical” statements); in other cases the problem is “hardened” (made quantitative), then the output is “relaxed” (reduced to a probabilistic result).

Even if the set of methods is rapidly increasing, we look for a taxonomy of the bulk of tools, according to potential use in geography and planning. Considering, for instance, characterization of geographical objects, spatial statistical analysis, territorial modeling and simulation, etc., every qualitative method (e.g. conceptual maps, ontologies, folksonomies, cladistics, qualitative regression, cross impact analysis, expert systems, etc.) is classified, with a short description of its main features.

For a better understanding of meaning, procedures and potentialities of qualitative methods, a brief account of four case-studies is given. A software for building "ontology" is used to map a system of interacting agents; "textual statistics" are applied for analyzing the perception of landscape on the basis of an "open answer" (qualitative) survey; the "cross impact analysis" is the qualitative modeling tool built for the evaluation of the consequences of a land-use project; in a qualitative "multi-criteria evaluation" it is grounded the site selection for location of a big infrastructure.

As a conclusion, the perspectives of qualitative methods are discussed, in the light of two emerging phenomena: respectively, the incoming "extreme data" availability age; and the next "highly automatic" computer-based research activity epoch.

**KEYWORDS**

New qualitative methods, Epistemological foundations and methodological principles, Applications: ontology, textual statistics, cross impact analysis, multicriteria evaluation

## INTRODUCTION: THE REASONS OF A FAST GROWING NEW METHODOLOGICAL APPROACH

Even though geography and urban planning are rich in quantitative data (from coordinates, to distances, to other dimensions – in intensity and/or extension – of punctual, areal or spatial interactive phenomena), they are mostly constituted by qualitative information.

We also immediately remark that, in the following work, by qualitative data we mean:

- Lexical information only (i.e. terms of a taxonomy: lion, tiger, jaguar, puma, …);
- Ranks and sorting (1st, 2nd, 3rd, 4th, ...; bad, sufficient, good, optimal, ….; D, C, B, A, A+, A++, …). *Note*: often evaluations are expressed through a numerical mark (i.e. 10). In these cases we have to pay attention to the evaluation method, and determine if the number is only an agreement to express a rank (in this case, the number can be

substitute with another agreement, like a word or character), or it's a real quantitative data (then 10 is the double of 5);

- Binary information (yes/no; true/false; 0/1; 1/2; single/many; present/absent;…), which can be seen like a limit case of the rank.

Dealing with these kinds of data, there are two approaches.

In "natural sciences" (and then also in those fields of geography more related to this view), qualitative data usually have got a limited consideration and ill repute. These are two of the capital sins qualitative data are accused of:

- They evade a "rigorous" treatment (we can't use axiomatic, univocal, and also simple definitions; we can't make explicit the relationships in analytical or geometrical formulas);

- They are questionable, that is they aren't objective in a verifiable way (that is, to a certain extent, until debate of the philosophical problem of the "*qualia*" – see Levin (2001) – that is, in other words, the problem about if what I perceive and say to be "red", it's the same for another person; or if you define as "red" something else, a different perception).

On the other hand, in "human sciences" (and then also in those fields of geography more related to this view), qualitative methods (such as conversational interviewing, focus groups, contents analysis, semantic analysis, etc.; see for instance Denzin & Lincon, 2005), that use data of the kinds listed above, are judged to be a fundamental research approach, because:

- They are fruitful in analyzing cause-effect relations, in generating hypotheses, etc.;

- Only they allow the study of facts which are emergent, deviant or inaccessible to quantitative measures (e.g. gangs, prostitution) and help to understand the reasons for the "outliers" in a given phenomenon.

At the roots of this dichotomy lie two (epistemological) ways of understanding reality, positivism (quantitative approach) and interpretativism (qualitative approach). Presented along the time in different manners (understanding vs explanation, subjectivism vs objectivism, discovery vs justification, non standard research vs standard research,...), these two views are in fact all related to the basic sharing made by the "classic science": Descartes' distinction between *rex extensa* and *res cogitans*, which is not only a separation between subject and object, but also between qualitative and quantitative.

Deeply describing this historical debate is outside the scope of this chapter (see, for instance, Marshall, 2005). What is important here it's to say that this state of art is rapidly changing under the impact of three factors:

- **A better comprehension of the nature and structure of the "knowledge"**. Two aspects have to be taken into account. The first one is the complexity of the cognitive forms (explicit and tacit knowledge, unconscious and by default knowledge, etc.) and the relationships which connect themselves (either as individual knowledge, or social knowledge). The second one is the so called "naturalization" of the knowledge (the recognition of its conditioning by our physical-biological substrate), of which a consequence is the elimination of the strong correlation between "subjective", meant as relative to the subject, and the "subjective" meant as questionable. Also the individual knowledge can be taken back to a stronger and more objective treatment. More generally expressed, it puts together the dichotomy subject/object (qualitative/quantitative) in an articulated continuum of situations;

- **The possibility**, through new technologies like ICT (information and communication technologies: computers, telecommunication, sensors, etc.) **to manage information (and therefore knowledge) in innovative, richer, and more powerful ways with respect to the traditional media** (i.e. documents on paper sheets). To go more in detail, four elements characterize this transformation:

  1. ability to treat (acquire, store, elaborate, recover) a huge quantity of data, and more widening than what it was possible before. The GIS are a paradigmatic example of this.
  2. an increasing ability to reason, either to support human reasoning, or as an autonomous ability of the machine. Crucial for this is the idea of execution – more or less supervised by man – of an algorithm: a sequence of instructions (declarative: i.e. X=5; operative: i.e. GO TO; logical: i.e. IF... THEN…; or algebraic: i.e. X=A+B) which permit, in mathematical fields, from analytical equation computation, to achieve logical deductions and symbolic reasoning; but also permit, in other fields, for example to perform drawings, manage database, manipulate sounds, etc. It's also opportune here to observe that the previous "logical elaboration" joins qualitative methods, which we will see later, with other new

tools for spatial analysis (i.e. Cellular Automata, where the cell state – lexically defined – is modified on the base of logic rules prevalently qualitative; or Multi Agent Systems, where behaviors and characteristics of the agent are often defined only in a qualitative way);

3. A (becoming more) “distributed” organization of knowledge. With this we mean two things: the storing (rapid filing and rapid recovery as well) of the information on different supports (for number and type), also spatially distant and in different form (hyper-media); and the decentralized processing of the knowledge, with easy possibility (also at distance) of collaboration/competition either for the instruments, or for the people;
4. A (growing) predominance of the “visualization” as an expressive modality of knowledge (we can think about passing from text in DOS to icons in WINDOWS). And it’s true that images can transmit knowledge in a superior way (for velocity in communication, type of knowledge, and richness of the contents) with respect to traditional text. However we have to know that the cognition itself also changes, as the consequence of visualization, with not always positive aspects. For example, the “simultaneous” visual knowledge is more hetero-direct than the “sequential” textual one is (where we can control the reading and reflection times). To better understand this point, see Simone (2000).

- Finally, as a consequence, **the emergence of a new “science of complexity”**, with the following tendencies:

  1. at the epistemological level, to define the conceptual and theoretical bases of the complex systems, that is those systems characterized in the structure by a huge number of elements with non-linear relations (differing from classical science with few elements and linear relationships), and in the emergent proprieties from holism, auto-organization, and counter-intuitiveness (in contrast with reductionism, causality, and logic);
  2. at the methodological level, to individuate the approaches to study these systems, adopting on one hand new perspectives on knowledge, and on the other hand the potentialities offered by new ICT.

In this new context, the qualitative data is the expression of what is known about the complex system, a knowledge richer than a knowledge based on simple and linear approximations (which can be expressed in a quantitative way) of this complex system. Even though this knowledge is, as often happens, strongly connected to characteristics of the subject who knows, it doesn't evade objective or objectivable aspects. It is a knowledge for which we have instruments for a strict treatment (that is precise and reproducible).

In conclusion, traditional approaches take advantage of the new emerging perspectives (that is more than an integration between qualitative and quantitative researches, but it is a really unified vision):

- quantitative methods escape the "rhetoric of objectivity", accepting aspects of "subjectivity" (social and historical conditions, individual creativity, serendipity, etc.) pointed out by "relativistic" epistemologists;
- qualitative methods pass over the criticism of excessive arbitrariness (i.e.: not to be scientific);
- both methods benefit of a large set of new tools (or improvement of traditional ones), more rigorous and powerful.

There is nothing to be amazed about if qualitative methods are con Figured today, in "quantitative geography", as an Eldorado to discover and take advantage of.

## THE LOGICS UNDERLYING QUALITATIVE METHODS

Without claiming to be complete – systematic studies in this field don't exist – in this paragraph we discuss about the adopted strategies in order to treat qualitative data in a rigorous way.

It's useful to remember that **the most traditional way is the one of data "numberization"** (in order to express data through numbers, congruent with the scale – lexical, ordinal, binary – of the qualitative information) followed by an elaboration "as if" the data was quantitative. We have to pay attention to two things interpreting the results: to know that something – in knowledge – is lost in this procedure (qualitative data is richer than quantitative data); to be careful in the cogent character of the results (as consequence of the possible effect of what we have lost).

Compared to the traditional way, the new methods use, often in a differently composite way, the four possibilities offered by new ICT (great amount of data, distributive knowledge, etc.).

To understand strategies, let us see the problem in the following way (Figure 1).

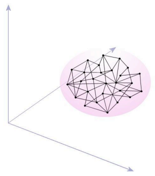


Figure 1: Qualitative data in the knowledge space

In multi-dimensional space of the knowledge, qualitative information is a particular domain (ellipsoid in the figure), which is also a system (also if not explicated) of complex elements and relationships.

The fist strategy (which is the one used above!) is to proceed with **visualization of the systemic structure of the information**. In other words, it consists in mapping the mental model which we have about the phenomenon under exam, recognizing variables and relationships (of different typologies) – see Figure 2.

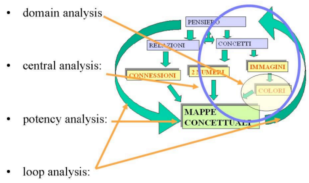


Figure 2: Visualization of a concept (in this case it’s the concept of visualization itself)

Even though visualization is useful by itself, as a tool to clarify information and for a less ambiguous communication (especially for maps of great dimensions), and also as base for other rigorous analyses (i.e. central variable, conceptual loops, etc.), it is often an indispensable starting point for other methodologies. The formal ontologies, for example, can be considered as a specification of the maps, where the concept are defined in a more rigorous way, through a set of well defined relationships (i.e. “is a part of”, “is a kind of”, “is a synonym of”) which allow an automated logical elaboration.

Ontologies are also a good example of the second tendency in the treatment of the qualitative information: **the explication**, often with the aid of large memories and logical-computational abilities of the computers, **of the complex quantitative system behind the information itself**. This has also been seen in the current use, for example, of the growing diffusion of a numerical indicator for the “perceived warmth” (a complex function of temperature, humidity, wind, etc.). This procedure is particularly used in the system of indicators where often a fuzzy concept (i.e. quality of life) is identified as function (more or less simple) for a set of quantitative variables (income, number of cars, etc.).

The objectification of perceptions (then qualitative data, referred to their own subject) profits, in a growing way, of the great possibilities offered by “the net”. Involvement (more or less voluntarily) of a large group of users, on statistic base, can give shared definitions, participative taxonomies (folksonomies), scales of preference. The Figure 3 shows an example of these scales, where also the visualization is applied to communicate the results with more immediacy.

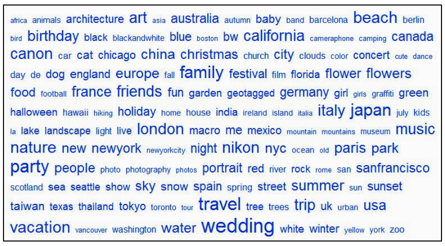

africa animals architecture art asia australia autumn baby band barcelona beach berlin
bird birthday black blackandwhite blue boston bw california cameraphone camping canada
canon car cat chicago china christmas church city clouds color concert cute dance
day de dog england europe fall family festival film florida flower flowers
food football france friends fun garden geotagged germany girl girls graffiti green
halloween hawaii hiking holiday home house india ireland island italia italy japan july kids
la lake landscape light live london macro me mexico mountain mountains museum music
nature new newyork newyorkcity night nikon nyc ocean old paris park
party people photo photography photos portrait red river rock rome san sanfrancisco
scotland sea seattle show sky snow spain spring street summer sun sunset
taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban usa
vacation vancouver washington water wedding white winter yellow york zoo

Figure 3: “Word cloud” (or “tag cloud”) of the most photographed things

Another strategy to elaborate qualitative information is the one that **proceeds** – where it’s possible – in **the cognitive space** (see Figure 1) **as an analogy of the usual geometrical and mathematical spaces**. So, in a literary text, the lemmas (words or groups of words) can be assimilated to elements; and any relationship among lemmas (i.e. the co-occurrence in a phrase) can be assimilated to a proximity or distance among elements. A statistical analysis of the element frequencies and of the distances can rigorously define, besides the syntactical structure of the text, the semantic relevancies and the logical correlations. It’s possible to point out that the matrix of distances can be used for more sophisticated elaborations (i.e. with classification algorithms, with neural network, etc.); and all this can be extended to figurative (images and videos) and audible texts.

Going to consider the specific case of qualitative sorting, where more generally we have to combine (in an opportune way, respecting a particular finality) different sortings relative to a set of objects (you can think about rank of nations relative to different Olympic disciplines). It seems recognizable that three main strategies stand out (in different variants and extensions, sometimes applied together):

- renouncement of a certain amount of information (i.e. calculating totals, passing to binary scale, etc.). We obtain quantitative data that we can elaborate accordingly the following way;
- utilization of information in order to formulate a different problem, but quantitative (i.e. passing – in a convenient way – from a matrix of the original data to a matrix of distances between elements). The result is then elaborated, taking it back to the solution of the initial problem;
- a sophisticated procedure (see Rietveld, 1989) logically based on two principles:

  1. qualitative sortings separate all quantitative sorting (among considered objects) in two sets: the congruent one and the not congruent one with the qualitative sorting;
  2. for the Laplace's indifference principle, all quantitative congruent sortings are possible, and also equiprobable (if not, logically, we would have a different information on the probability)

  and it is operatively based on the wide and fast calculation ability of the computer:

  1. calculating solutions for all quantitative data congruent with the sorting (actually, more exactly, we consider a wide sample of data, uniformly distributed on the range of admissibility);
  2. statistically analyzing the solutions (and then to obtaining results in probabilistic terms).

In order to conclude this paragraph we have to make an observation on treatment of fuzzy data. There is a wide literature about that (see, for example, Dubois & Prade, 1980) also regarding territorial problems (zoning, multi-attribute evaluation, etc.); and which is impossible to summarize in this paper. Though we can remember that, in the fuzzy set theory, the concept of membership function $\mu_A(x)$ is crucial. In the interval [0,1] it indicates the degree of belonging of the element *x* to the set A (i.e. x = man with x age, A = set of the old men). Membership functions, besides for the elements, can be defined also for relationships (i.e. x = distance; A = far away). Moreover analytical models (i.e. regressions) and algorithms (i.e. classifications) can be defined in terms of fuzzy variables and relationships.

Numerical computation of these formulas, in general, goes through a parameterization (of the membership function) for each variable and relationship, like:

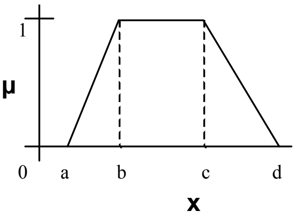


Figure 4: Trapeziform approximation of the membership function

In fact, opportunely using the four numerical value a, b, c, d (and the corresponding values of μ), it's possible perform the right counts and obtain the fuzzy result (approximated).

## AN OVERVIEW OF METHODS AND FIELDS OF APPLICATION

In the previous paragraph we showed a set of strategies to deal with qualitative information. In rigorous qualitative methods, they are used individually (as in conceptual maps) or – more often – jointly (as in cross impact analysis). In this paragraph we propose a list of methods, which, on the base of our personal experience, seem particularly useful for geographical analysis and urban planning. The reader should know though, that the fast growth of this field will provide us with new tools, much more innovative. For each method we present the work guideline (if it hasn't already been done in the previous paragraph) and the most important applicative potentialities. But the reader should consider that many tools are multipurpose as far as their uses are concerned, and that the creativity of the researcher could individuate new application fields.

**Conceptual maps** (Novak & Gowin, 1984; Cañas, Novak & González, 2004) are the main visualization tool (see above). From Einstein to Mendeleiev, or Kekulè, many great scientists have underlined that (mental) visualization of their problem helped them in their discovery. For us, (video/graphic) visualization helps us to work with networks of concepts of a great amount of elements; and help us to remember them. And all this helps the reasoning: clarifying ideas, reviewing the map, etc., until it can be used as knowledge support in order to build other new tools (i.e. simulation models).

A second important value is the communication. It can be used in different participative processes, making it easier to show different points of view of the different stakeholders, or to build a shared knowledge. In order to build the conceptual maps there are different softwares, we can indicate in particular Compendium and CmapTools (this one characterized by a wide online community). Finally we can report also different maps, derived or similar to conceptual ones: **semantic networks**, **frames**, **topic maps**, **causal maps**, etc.

In this list there are also (and they are very important) the **formal ontologies** (Guarino, 1998), a “formal and explicit specification of a shared conceptualization” (see previous paragraph). Following this definition, ontologies present three kinds of applications:

- automatic reasoning, which often is a classification procedure: once an ontology defined (i.e. about different kinds of urban settlements) the machine proceeds by putting the specific instances on the basis of their description (i.e. Charlottesville as “zoomburb”). This is a potentiality not yet well explored in geography;

- guarantee of the interoperability on different systems. They can be information systems, like GIS, and ontology guards against the use of the same term (i.e. street) with different meanings in different countries. But they can also be different systems like technical dictionaries (i.e. of Urban Planning) in different languages, where the ontology allows to translate in a correct way the terms, which are apparently similar (city plan, master plan, etc.), but actually different (for some aspects).

- building meta-model of a model (definition: meta-model is the whole of the information which define the model itself). It’s an increasing necessity not only for a clear comprehension of the model, and not only for the correct implementation of the model in a programming language (or to translate from a language to another one), but also because in the wide diffused object-oriented modeling (i.e. multi agent systems) the analytical formulation is missing, which guarantee a unique definition and reproducibility of the model (two conditions for scientific models). The meta-model can be used for this (and to build that, it’s necessary for opportune languages to be created ad hoc, i.e UML – unified modeling language –).

In conclusion it’s possible to mention already available ontologies (i.e. WordNet), or make a list of different programming languages to build software which manage ontologies (i.e. Cyc, OWL, etc.), but it’s also possible to remember the main software (Protégé) and another one dedicated to urban phenomena (Towntology), which is interesting because it can also use images.

Together with regression analyses, the main tool for interpreting relationships between variables, classification methods (with clustering techniques, which in territorial field define zone systems) are the most used tools in geography, because they are the best tools to put into evidence order forms (classes, classification trees, zones, etc.) within complicated sets of data. **Cladistic** (Kitching et al., 1998), like other similar techniques, extends the application field to qualitative information. What differentiates cladistic from others is the fact that, inside it, the profile (set of information) of an object to be classified, or an ideal-typical profile, is taken as reference point to classify all other object. Taken a set of objects (i.e. municipalities of France, just to remember that it's possible to operate with very huge data) and taken qualitative profiles (i.e. delicatessen: red wine, goat cheese, mustard, etc.) for each municipality, all these techniques generally start with determination of any similarity index among profiles (i.e. number of common deli). The matrix of similarity indices (among the municipalities, in the previous example) is treated with one of the methods usually used for quantitative taxonomy.

To elaborate a cladistic it's possible to use, for example, WinClada software; but nowadays classification algorithms can be found in many qualitative analysis packages (see textual statistic below). Out of curiosity, from personal experience, it's important to indicate that habit to taxonomy on quantitative data is so deeply rooted, that it's necessary to make some efforts to come back to think about "qualitative" comparison (and discover new ways for mining unexploited information).

We have considered above, in an implicit way, that the qualitative profiles of a set of objects were available. If we move our attention on the building of these profiles using key-words, labels (TAG), we are in the field of **folksomony** (Sterling, 2005), union of the words folk and taxonomy. As these terms say, they are the result of a collaborative building of object classifications through freely chosen tags (an action to be referred to the internet world). We must remark that:

- about their peculiarity with respect to traditional classifications, they always have to be referred to the "conceptual models" of the specific communities that create and use them (and these conceptual models are also in rapid evolution);

- about their application fields, performing individual preference aggregation, they indicate the collective ones (social bookmarks); and, allowing to recover information through their classification, they give the possibility to share different textual data in different media (texts, images, videos, sounds, etc.);
- about elaboration potentialities, they allow different successive treatment of the profiles with the described modalities (i.e. cladistics, textual statistics, multi-criteria methodologies).

In geography and in urban planning, utility in preference aggregation and media sharing is already evident on its own, but these potentialities should be much more explored and exploited. About the software, it's possible to find most of them on internet web sites (i.e. social bookmarking: Del.icio.us, Technorati, etc.; images and photo sharing: Flickr; video sharing: YouTube; sound sharing: Freesound; news: Digg; etc.).

Using the terminology just explained, **textual statistics** (Lebart & Salem, 1994; McKee, 2003) can be described as the procedure that, using a large number of Tags (lemmas, made of single words or groups of words), counts frequencies of the occurrence and co-occurrences in the text. Co-occurrences can be defined differently: as contiguity of lemmas, co-presence in a phrase, separation within a max number of words, etc. The analysis of the text structure, or automatic mining of significances, can be performed directly on occurrences and co-occurrences; or operating with classification techniques (i.e. see cladistics). To reach this goal the MDS is also used (multidimensional scaling) (Borg & Groenen, 2005). At the contrary of the simple procedure to acquire distances between objects, given the coordinates (and chosen a metric), this algorithm (quite complex) gets coordinates given the distances between objects. Assuming that n objects can be placed, in general, in a n-1 dimensional space, the procedure – in order to allow the user to interpret the outgoings (through visualization) – "compresses" (projecting with some criteria) the obtained coordinates in space with fewer dimensions (usually, mono, bi, three-dimensional). Many deep meanings of the text are analyzed through reciprocal positions of the lemmas. Among a lot of statistical packages, which perform textual statistics (i.e. SPAD, Hamlet, etc.) we indicate Weka (an open source software).

In geography and urban planning, two are open fields: (innovative) elaboration, starting from the huge amount of existing texts, and treatment of open choice questionnaires (a kind of answer very difficult to analyze before, but – differently from the closed one – it is deprived of a conditioning from surveyor).

Finally we report the **text mining**, generalization of textual statistics, either in the terms of more ways to elaborate texts (information retrieval and information extraction) or in terms of other available elaboration techniques (decision trees, etc.).

The so called **multi-criteria methods** are the generalization to qualitative data of the linear regression. We take $y_i = a_1 x_{1i} + a_2 x_{2i} +$ etc. $+ a_N x_{Ni}$ , where $y_i$ is the dependent variable, $x_{ni}$ is one of N independent variables ($1 \leq n \leq N$), $a_n$ the weight and i the i-th case study in the regression ($1 \leq i \leq M$). If the independent variables and/or the weights are, for example, ranks, the dependent variable $y_i$ will also be a rank. Multi-criteria methods are overall applied as evaluation tools (global comparison of the alternatives, each one evaluated following different criteria, most often qualitative); and they are among the qualitative techniques, with old tradition, for which a wide literature exists. Noted that algorithms exist either for the solution of the quantitative equation or for its calibration (estimation of weights $a_n$, that is to determine preferences with respect to different criteria in evaluations), we invite to refer to that literature (to start i.e. Nijkamp et al., 1985; Roy et al., 2001).

In order to build “scenarios”, a fundamental activity in planning, it’s possible to use the **cross impact analysis** (Gordon & Hayward, 1968), a beautiful example of extension and use of the potentialities of the conceptual map. Once we have built (in a collaborative/competitive way among stakeholders) the map of the causal-effect chains among the factors, which determinate the possible (future) scenarios of a system, being in the field of subjective evaluations, it’s often possible to define a probability (or possibility) in the realization of the causal-effect relationship, for each relation taken singularly, that is the direct impacts (in other words complex synergies among factors cannot be estimated). We call $F^I_{ij}$ the impact matrix (of the factor i on the factor j). Operating with a logic (similar to the one explained above) which we can call “numerical simulation + Laplace’s principle”, this procedure allows to “cross” the impact, resulting in a new matrix $F^F_{ij}$ which includes the inter-dependences and that is depurated from incongruities and subjective judgments. As software for the cross impact analysis we prefer “The Time Machine” (freeware).

The list of the presented methods would not be exhaustive if we would not have mentioned a set of other tools about Artificial Intelligence (Nilsson, 1998) (i.e. **expert systems**, **logic programming**, **case based reasoning**, etc.). They are similar – despite their specificities – by the fact that qualitative knowledge (opportunely coded) is elaborated following well defined rules and chains of rules.

## SELECTED EXAMPLES OF APPLICATION

In order to make more concrete the previous discussion, in this paragraph we present some examples of applications made by authors. For this purpose, for each case, the presentation (obviously brief) will be centered on the problem, on the input data, and on the results (just a subset of results) obtained with the considered qualitative method.

### Ontology of a participative location process

The studied phenomenon (Rabino & Caglioni, 2009) can be described in the following way:

- there are four actors: public administration and (for example) three municipalities involved in localization of an obnoxious or dangerous facility;
- all the actors propose a set of criteria. With these criteria, the alternatives proposed by a technician are evaluated. In this way, we obtain values relative to each alternative for each criterion, measured with different unities;
- in order to make the values all homogeneous, the actors choose all together some utility functions (different for each criterion), adapted to translate all values in scores between 0 and 1, which represent the satisfaction degree of the couple alternative-criterion;
- afterwards, each actor sets, independently from others, the weight for each criterion, based on its own preferences; and this allows to get a rank among all alternatives for each stakeholder;
- then we have a negotiation phase, where the ranks are compared in order to find one and only one winning alternative. The weights are varied, in order to reach a definitive agreement among the actors, and together mitigation and compensation actions are proposed, for those actors who are penalized by the final chose;
- when the agreement is settled, the realization of the work is given to technicians.

This descriptive ontology of the phenomenon can be freed from ambiguities of the natural language through the building of its “formal ontology”. To do this, UML (unified modeling language) diagrams are used (Rhem, 2005). This is a tool to specify the model of the system that we are studying, which allows to properly build the programming code, and eventually to transfer it from a code to another one without interpretation errors.

In Figure 5 we show a diagram about the system’s behavior, the use case diagram. This diagram shows the actors (users of the system), the use cases (scenarios of how they use the system), and their relationships. (Note: the actor is the entity which interacts with a use case, activating the sequence of actions described by the use case and, eventually, getting some precise answers from the system. It could be a person or another system).

In Figure 6 we show a diagram about the structure of the system, the class diagram. By class we mean a category or a group of objects (with this term we also include living beings) which have similar attributes and similar behaviors. The class diagrams show the classes and the relationships between classes (using a standard notation).

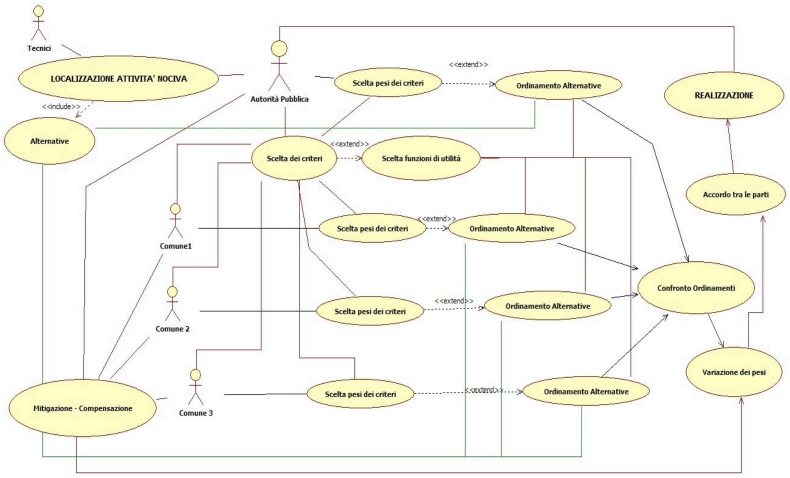


Figure 5: Use case diagram of the location process

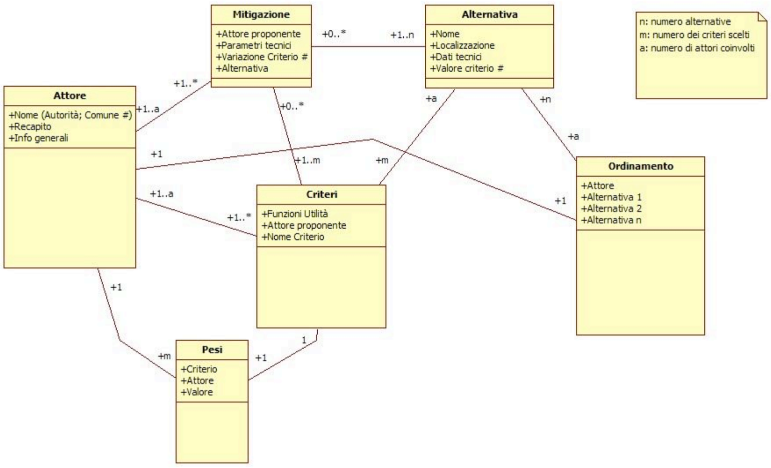


Figure 6: Class diagram of the location process

## Point out the collective imaginary with textual statistics

The problem we are studying, in the landscape research field, is the one that recognizes how the alpine landscape appears in the collective imaginary, in terms of significant elements and feelings aroused.

This study (Rabino & Scarlatti, 2006) was led through an open choice questionnaire, with two questions:

1. If I say “alpine landscape”, what image comes into your mind? Describe it in 5/10 lines

2. What are your feelings and sensations when you think about this image? Describe it in 5/10 lines

The questionnaire had been give to a population sample, stratified by sex, age (young, adult, old), life context (metropolis – centre and hinterland – big, medium, little municipality) and distance from alpine environment (municipality on mountain, on pre-Alps, in plain).

Among the elaborated results, we show several “maps”, obtained with multidimensional scaling (see above), of the relative positions of more significant observed lemmas (connected with the concerned population).

In Figures 7 and 8, we show the mono-dimensional maps respectively about images and emotions for males and females.

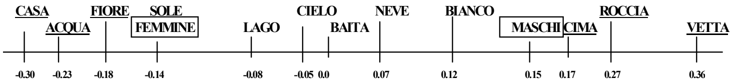


Figure 7: Graph which represent the mental image of males and females

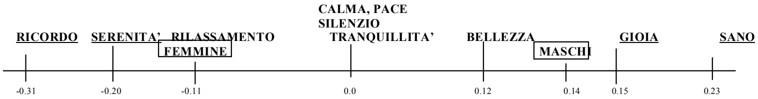


Figure 8: Graph which represent the feelings/emotions of males and females

The contribution to (rigorous) semantic analysis of questionnaires is highlighted by several considerations which can be deducted from those graphs, and that we report immediately below.

The image of landscape, relative to females, is characterized firstly by the words house, water, and flower, and secondly by mountain, grass, river, sun, and white. To this we can oppose the male image of alpine landscape, characterized by the terms rock, peak, and chalet, top, valley. Moreover, females enrich their description with emotions characterized by words like relax, serenity, remembering, but also joy, loneliness, fear, while the males “feel” the mountain in the terms of beauty, health, and cold. The female imaginary seems to privilege, also in the alpine landscape, the elements which can be considered more related to their world: house and water, essential for living. Females have a sunny image of landscape, full of light, depending on the season: mountain, river, grass, sun recall prevalently the summer, while white recalls a winter image of alpine landscape. The word “house” reinforces the concept that, also in the imaginary, the most part of the landscapes, which seem natural, are highly humanized. And also the repeated term “grass” recalls the territory domesticated by man.

In Figure 9, we present another example, a 2D graph, which represents the mental image of the interviewed persons, in relation with the distance to the alpine environment.

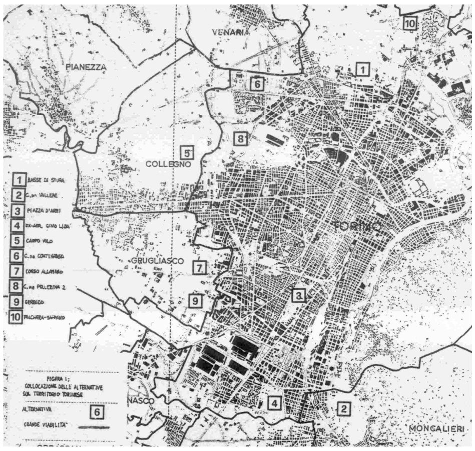


Figure 10: Map of the potential site for the stadium

The other alternatives (3. Piazza D'Armi; 4. Ex Aeroporto G. Lisa; 5. Ex Campovolo; 6. Cascina Continassa; 7. Corso Allamano; 8. Cascina Pellerina; 9. Gerbido; 10. Falchera-Mappano) can be ranked (in a qualitative way, for the difficulties in quantification and/or lack of time for measurements) considering seven criteria agreed upon by all the stakeholders and considered as relevant in the location process:

1. Area extension (for the possibility to add related structures and also new ones in the future); 2. Accessibility (for the user); 3. Integration of the stadium with the public green system; 4. Environmental impact of the infrastructure (with respect to the urban context); 5. Disturbance (to residents caused by the fruition of the stadium); 6. Availability of the site (that is, easiness to purchase); 7. Constrains (of different types: land-use planning, historical values, etc.).

Figure 11 shows the evaluations. These are qualitative even if they are represented with numbers; and it's useful to notice that the higher the values are, the better; and that they are also tying.

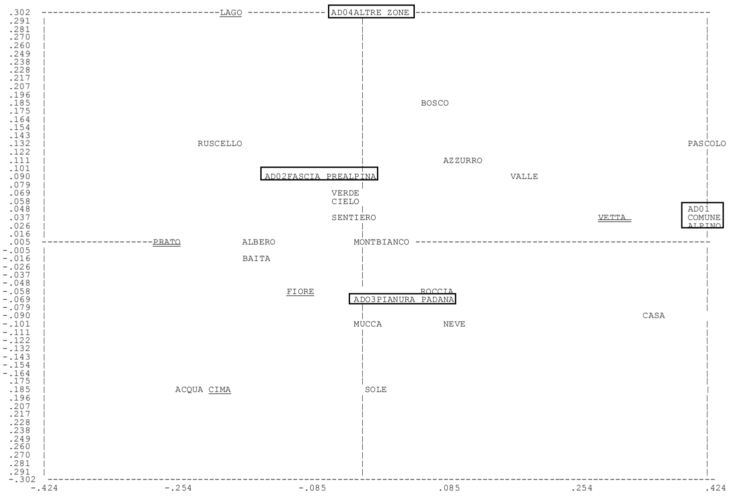


Figure 9: Graph of the mental image of the interviewed people,
in relation with the distance with the alpine environment

A different vision emerges between who resides in the mountain, and live in the mountain everyday, and who resides in the city, either in a big metropolis or a small municipality. It’s the difference between the place for living, and for working, and the place for holidays, where everyday worries don’t take place, where one can relax and contemplate and where entertainment and amusement play a larger part. In fact, people who live in the mountains prefer to use the word “pasturage”, while those living in a plain use the word “grass”. Moreover observing the graph of emotions and feelings (not added to this paper), a word used only by people who live in the mountain is the term “freedom”. It indicates how freedom is important for those people; and it is probably for this reason that they live in the mountains.

### Multi-criteria evaluation of sites available for a big infrastructure

The problem was the choice of a site to build a new stadium in the city of Turin (A.A.V.V., 1987).

After a “research on-the-field”, technicians (Public Administration) proposed to decide on a selection among 10 solutions (see Figure 10), but the alternative 1 (Basse di Stura) and alternative 2 (Vallere) have been rejected because of their naturalistic interest and because of the fog in the winter period (when football matches usually took place).

| | Extension | Accessibility | Green | Env. Impact | Disturb | Availability | Constrains |
|---|---|---|---|---|---|---|---|
| 4. ExAeroporto G.L. | 2 | 1 | 0 | 1 | 1 | 2 | 5 |
| 9. Gerbido | 7 | 2 | 1 | 1 | 2 | 0 | 4 |
| 5.ExCampoVolo | 6 | 1 | 1 | 1 | 2 | 1 | 3 |
| 8. C.Pellerina | 3 | 3 | 2 | 0 | 2 | 2 | 3 |
| 6. C.Continassa | 4 | 3 | 1 | 1 | 2 | 2 | 2 |
| 10. Falchera | 5 | 2 | 1 | 1 | 2 | 0 | 3 |
| 3. PiazzaD'Armi | 1 | 2 | 0 | 1 | 0 | 2 | 4 |
| 7.Corso Allamano | 1 | 1 | 0 | 1 | 2 | 1 | 5 |

Figure 11: Evaluation of the alternatives for location with respect to different criteria

In order to obtain an "overall" evaluation, faced with the difficulties of the public decision makers (but also of the technicians) to express the relative importance of the various criteria, different points of view are explored (point of view = particular set of weights for the criteria), emerged during the discussion and regrouped in four cases: highly innovative, moderately innovative with economic caution; moderately innovative more sensitive to ecological themes; conservative.

The Figure 12 shows the table of weights (precisely, the rank of the weights) for the four points of view (an higher value means higher preference).

| | Highly Innovative | Mod.Innov.Econ | Mod.Innov.Ecol | Conservative |
|---|---|---|---|---|
| Extension | 7 | 7 | 4 | 4 |
| Accessibility | 5 | 6 | 5 | 7 |
| Green | 6 | 3 | 6 | 3 |
| Env. Impact | 4 | 2 | 7 | 2 |
| Disturbance | 3 | 1 | 3 | 1 |
| Availability | 2 | 5 | 1 | 5 |
| Constrains | 1 | 4 | 2 | 6 |

Figure 12: Weights of criteria for different points of view

Multi-criteria evaluation has been conduced with alternatives methods (weighted sum, regime, concordance/discordance, etc.) obviously for the 4 points of view.

As an example (Figure 13) we report the result (appraisal scores) for the weighted sum.

|  | 4.Ex AerGL | 9.Ger bido | 5.ExCampo Volo | 8.Pelle Rina | 6.Continassa | 10.Falchera. | 3.Piazza D'Armi | 7.Corso Allamano |
|---|---|---|---|---|---|---|---|---|
| Highly Innovat. | 0.0739 | **0.1884** | 0.1483 | 0.1626 | 0.1528 | 0.1476 | 0.0647 | 0.0617 |
| Mod.Innov. Econ | 0.1030 | **0.1670** | 0.1303 | 0.1487 | 0.1525 | 0.1232 | 0.0995 | 0.0756 |
| Mod.Innov. Ecol | 0.0894 | 0.1541 | 0.1386 | 0.1367 | **0.1583** | 0.1450 | 0.0888 | 0.0891 |
| Conservat. | 0.1118 | 0.1317 | 0.1050 | **0.1660** | 0.1563 | 0.1121 | 0.1232 | 0.0939 |

Figure 13: Appraisal scores for the weighted sum method

Here, but more generally, we can see a clear predominance of the alternatives: 9. Gerbido; 8. Pellerina; 6. Continassa. This happens with all kinds of methods, and without any substantial differences among the points of view.

The result so formulated is useful to remember, in conclusion, that evaluation isn't a decision, but it's a tool for decision support. Positively, in this case, choosing Cascina Continassa, Public Administration was in agreement with the evaluation process made in a rigorous way, even thought qualitative.

**The cross impact analysis to foresee an urban requalification project**

Inspired and above all supported by Collegno municipality, in the Turin hinterland, this study (Murano, Spaziante & Rabino, 2007) aimed to precise the contents of the project "ImmaginiAmo Collegno" (in italian, a joke with words imagine and "I love [amo] Collegno"), centred on three important actions (to change the functionality of a large group of historical buildings; to recuperate a wide industrial dismissed area; to set up the main road) (see Figure 14), but it is also associated with other interventions, and is therefore able to imply wide consequences for the whole socio-economical and urban organization of the municipality.

The goal was to build future scenarios, taking into account the multiplicity of the aspects (actions and consequences), their difficult estimation and quantification, and the multiplicity of the involved stakeholders (with their own competences, interests, points of view).

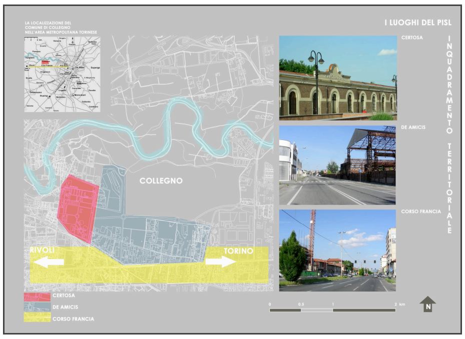

LA LOCALIZZAZIONE DEL COMUNE DI COLLEGNO NELL'AREA METROPOLITANA TORINESE
I LUOGHI DEL PISL
CERTOSA
DE AMICIS
CORSO FRANCIA
INQUADRAMENTO TERRITORIALE
COLLEGNO
RIVOLI
TORINO
0
0.5
1
2 km
N

Figure 14: Location and constitutive elements of the project ImmaginiAmo Collegno

To reach this goal, the utility of the conceptual maps (Figure 15a and 15b) is obvious (first step of the cross impact analysis) as tool to build, in a shared way, the model of the system (which consists in elements and relationships between themselves) and to estimate the strength of these relationships (measure of the direct impacts).

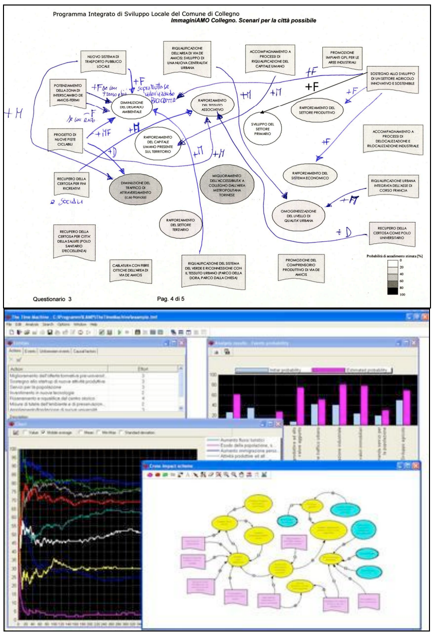

Programma Integrato di Sviluppo Locale del Comune di Collegno
ImmaginiAMO Collegno. Scenari per la città possibile
NUOVO SISTEMA DI TRASPORTO PUBBLICO LOCALE
RIQUALIFICAZIONE DELL'AREA DI VIA DE AMICIS: SVILUPPO DI UNA NUOVA CENTRALITA' URBANA
ACCOMPAGNAMENTO A PROCESSI DI RIQUALIFICAZIONE DEL CAPITALE UMANO
PROMOZIONE IMPIANTI GPL PER LE AREE INDUSTRIALI
SOSTEGNO ALLO SVILUPPO DI UN SETTORE AGRICOLO INNOVATIVO E SOSTENIBILE
POTENZIAMENTO DELLA ZONA DI INTERSCAMBIO DE AMICIS-FERMI
DIMINUZIONE DEL DEGRADO AMBIENTALE
RAFFORZAMENTO DEL TESSUTO ASSOCIATIVO
RAFFORZAMENTO DEL SETTORE PRODUTTIVO
SVILUPPO DEL SETTORE PRIMARIO
PROGETTO DI NUOVE PISTE CICLABILI
RAFFORZAMENTO DEL CAPITALE UMANO PRESENTE SUL TERRITORIO
ACCOMPAGNAMENTO A PROCESSI DI DELOCALIZZAZIONE E RILOCALIZZAZIONE INDUSTRIALE
RECUPERO DELLA CERTOSA PER FINI RICREATIVI
DIMINUZIONE DEL TRAFFICO DI ATTRAVERSAMENTO (c.so Francia)
MIGLIORAMENTO DELL'ACCESSIBILITA' A COLLEGNO DALL'AREA METROPOLITANA TORINESE
RAFFORZAMENTO DEL SISTEMA ECONOMICO
RIQUALIFICAZIONE URBANA INTEGRATA DELL'ASSE DI CORSO FRANCIA
OMOGENEIZZAZIONE DEL LIVELLO DI QUALITA' URBANA
RAFFORZAMENTO DEL SETTORE TERZIARIO
RECUPERO DELLA CERTOSA COME POLO UNIVERSITARIO
RECUPERO DELLA CERTOSA PER CITTA' DELLA SALUTE (POLO SANITARIO D'ECCELLENZA)
CABLATURA CON FIBRE OTTICHE DELL'AREA DI VIA DE AMICIS
RIQUALIFICAZIONE DEL SISTEMA DEL VERDE E RICONNESSIONE CON IL TESSUTO URBANO (PARCO DELLA DORA, PARCO DALLA CHIESA)
PROMOZIONE DEL COMPRENSORIO PRODUTTIVO DI VIA DE AMICIS
Probabilità di accadimento stimata [%]
0
20
40
60
80
100
Questionario 3
Pag. 4 di 5
+F
+F se con treno e bici
+F soprattutto se valorizzando BICICLETTA
-F se con auto
+M
+MF
+M
+D
e sociali
+M
+M
+M
+M
+F
+F
+F
+F
+M
+D
+F

Figure 15: a) map made by a stakeholder, which shows his opinion about the impacts and their strength;
b) collective map imported in the software “The Time Machine” (Blecic, Cecchini & Trunfio, 2005)

Afterwards, going on – starting from this information and following the methodology shortly presented above – to estimate the cross impacts, we can obtain a base scenario (Figure 16a), where it's possible to observe the achievement of several targets (more exactly, the values achieved by some particularly significant indices), without any projectual intervention (trend scenario).

We can see, for example, that in a common perception, the processes in action are leading to an increment of the accessibility of Turin (the pole of the metropolitan area).

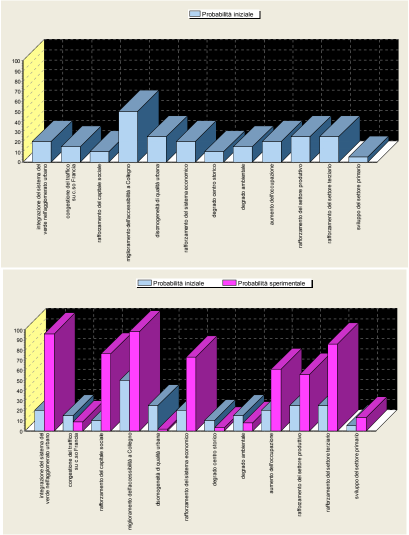


Figure 16: a) trend scenario; b) "all active" scenario

Besides this scenario, it's possible to elaborate alternative scenarios with a mix of different actions (in Figure 16b, for example, "all active" scenario, which is the one that activates all the actions proposed). These outgoings can be used to support the debate, the evaluation process and the participative decision.

To complete, we have to say that the "all active" scenario reported above privileges the accomplishment of the objectives: accessibility, economical development, and public green system; but to do this, the scenario increases congestion, environmental degradation and social segregation.

## LOOKING AHEAD TO FUTURE APPLICATIONS

Lately two institutions, with authority in their field, Microsoft (Emmort, 2005) and Computer Science Corporation (Lutzak, 2005) have published two reports about the evolution of information and computer science for the next ten years.

Concerning the data, the close future is characterized by extreme availability of information. Contrarily to the past (and nowadays), despite an increasing respect of individual privacy, a great number of sensors of different types will give us a huge set of data, in fast sequences over long periods (as an example, we can consider the wide information contained, already now, in “Google Earth”). Great opportunities appear, but we shall face new problems: firstly how to deal in an adequate way with the wide information (the following point about the software is proposed as an answer); secondly, how to resolve specific problems of the information on a large scale (How to verify reliability? How to control coherence? How to prevent falsification, with the scope of disinformation?).

Particularly, about the qualitative/quantitative aspect, we can add that, even though the great amount of today’s qualitative and subjective data will surely find more rigorous substitutes, the growth will concern above all the qualitative data, mainly because it is connected with a plurality of media.

About the elaboration of information, we will go towards “scientist and planning computers”, which, from little obtuse assistants, will become “creative co-protagonists of the scientists (geographers, planners, architects), fading away the man-machine frontier” (free but coherent citation of the report of Emmort). Going on with the citation, “nothing will remain of the old scientists (geographers, etc.). We need to change the training and to adopt new criteria (also ethic) in further studies and researches”.

For us, in our more limited perspective, this means forming ourselves to rigorous elaboration of the qualitative information; but being aware of the fact that this is an aspect of the beautiful tendency towards the reconciliation of two cultures: human sciences and natural (plus artificial) sciences.

## REFERENCES

**A.A.V.V.,** 1987, Scheda di valutazione definitiva delle aree per la localizzazione del nuovo stadio di Torino. Comune di Torino, Torino, mimeo: 1-25

**Blecic I.; Cecchini A.; Trunfio, G.A.,** 2005, Costruire scenari futuri per la pianificazione territoriale strategica: metodologie e strumenti. In: Cecchini A.; Plaisant, A. (Eds), Analisi e Modelli per la Pianificazione. Teoria e pratica: lo stato dell'arte. Milano: Franco Angeli: 1-13

**Borg, I.; Groenen, P.**, 2005, Modern Multidimensional Scaling: theory and applications. Springer-Verlag: New York

**Cañas, A.J.; Novak, J.D.; González, F.M.** (Eds), 2004, Concept maps: Theory, methodology, technology. Paper presented at the first international conference on concept mapping. Universidad Pública de Navarra, Pamplona, Spain

**Denzin, N.K.; Lincon, Y.S.** (Eds), 2005, The Sage Handbook of Qualitative Research. Sage Publication Ltd: Thousand Oaks (CA)

**Dubois, D.; Prade, H.**, 1980, Fuzzy Sets and Systems: Theory and Applications. Academic Press: New York

**Emmort, S.**, 2005, Towards 2020 Science. http://research.microsoft.com/towards 2020science

**Gordon, T.J.; Hayward, H.,** 1968, Initial experiments with the cross-impact method of forecasting. In: Futures 1, 2: 100-116

**Guarino, N.,** 1998, Formal ontology in information systems. In: Proceedings of FOIS'98, Trento, Italy, 6-8 June. Amsterdam, IOS Press: 3-15

**Kitching, I.J.; Forey, P.L.; Humphries, C.J.; Williams, D.M.,** 1998, Cladistics. Oxford University Press: Oxford

**Lebart, L.; Salem, A.**, 1994, Statistique textuelle. Dunod: Paris

**Levin, J.**, 2001, Qualia. In: Wilson, R.A.; Keil, F.C. (Eds), The MIT Encyclopedia of The Cognitive Sciences. Cambridge (MA): MIT Press, Cambridge (MA): 693-694

**Lutzak, E.,** 2005, Extreme Data: Rethinking the "I". In: IT. Leading Edge Forum, Computer Science Corporation. http://csc.com

**Marshall, G.,** 2005, Qualitative versus quantitative debate. In: Scott, J.; Marshall, G. (Eds), A Dictionary of Sociology. Oxford: Oxford University Press

**McKee, A.,** 2003, Textual Analysis: A Beginner's Guide. Sage Publications Ltd: Thousand Oaks (CA)

**Murano, C.; Spaziante, A.; Rabino, G.A.,** 2007, L'analyse des impacts croisés dans un processus de projet participatif. Paper presented at the 8th "Rencontres Theo Quant ", Besançon, France

**Nijkamp, P.; Leitner, H.; Wrigley, N.,** 1985, Measuring the unmeasurable. Martinus Nijhoff: Dordrecht

**Nilsson, N.,** 1988, Artificial Intelligence: A New Synthesis. Morgan Kaufmann Publisher: San Francisco

**Novak, J.D.; Gowin, D.B.,** 1984, Learning How to Learn. Cambridge University Press: Cambridge (MA)

**Rabino, G.A.; Caglioni, M.**, 2009, Modelling, metamodelling and UML language in spatial analysis. Paper presented at the V INPUT Conference, Lecco, Italy, forthcoming

**Rabino, G.A.; Scarlatti, F.**, 2006, Statistiche testuali, mappe concettuali, reti bayesiane: applicazioni nella valutazione del paesaggio. In: Moroni, S.; Patassini, D. (Eds), Problemi valutativi nel governo del territorio e dell'ambiente. Milano: Franco Angeli: 180-203

**Rhem, A.J.,** 2005, UML for Developing Knowledge Management Systems. Auerbach: London

**Rietveld, P.**, 1989, Using ordinal information in decision making under uncertainty. In: Syst. Anal. Model. Simul., 6, 9: 659-672

**Roy, B.; Bouyssou, D.; Jacquet-Lagrèze, E.; Perny, P.; Slowinski, R.; Vanderpooten, D.; Vincke, P.,** 2001, Aiding decisions with multiple criteria: essays in honor of Bernard Roy. Springer: London

**Simone, R.**, 2000, La terza fase. Laterza: Bari

**Sterling, B.**, 2005, What's the best way to tag, bag, and sort data? Give it to the unorganized masses. In: WIRED, 13 april: 1-5

## AUTHORS INFORMATION

**Giovanni A. RABINO**
giovanni.rabino@polimi.it
Politecnico di Milano
Dipartimento di Architettura e Pianificazione,
Milan, Italy

**Matteo CAGLIONI**
caglioni.matteo@gmail.com
Politecnico di Milano
Dipartimento di Architettura e Pianificazione,
Milan, Italy

# A SMALL WORLD PERSPECTIVE ON URBAN SYSTEMS

**Céline ROZENBLAT[*] and Guy MELANÇON[**]**

[*]IGUL, University of Lausanne, Switzerland, [**]LaBRI, University of Bordeaux - INRIA, France

## ABSTRACT

The theory of small world network as initiated by Watts and Strogatz 1998 has drawn new insights on spatial analysis as well as to systems theory. Its concepts and methods are particularly relevant to geography where spatial interaction is mainstream, and where interactions can be described and studied using large volume of exchanges or similarity matrices.

Networks indeed organize through direct links or by indirect paths, inducing topological proximities simultaneously involving spatial, social, cultural or organizational dimensions. Network synergies build over similarities and are fed by complementarities between or inside cities, the two effects potentially amplifying each other according to the "preferential attachment" hypothesis that has been explored in a number of different scientific fields (Barabàsi & Albert, 1999; Barabàsi, 2002; Newmann et al., 2006). In fact, according to Barabási & Albert (1999), the high level of hierarchy observed in "scale-free networks" results from "preferential attachment" which characterizes the development of networks: new connections appear preferentially close nodes already having the largest number of connections. In this way, the improvement in the network accessibility of the new connection will probably be greater. But at the same time, network regions gathering dense and numerous weak links (Granovetter, 1973, 1985) or network entities acting as bridges between several components (Burt, 2004) offer a higher capacity for urban communities to benefit from opportunities and create future synergies. Several methodologies have been suggested on how such denser and more coherent regions (also called communities or clusters) in term of links can be identified (Watts & Strogatz, 1998; Watts, 1999; Barabasi & Albert, 2000; Barabasi, 2003; Auber et al. 2003, Newmann et al., 2006).

These communities not only possess a high level of dependency between their member entities, but also show a low level of “vulnerability” allowing for numerous redundancies (Burt, 2000, 2005).

The SPANGEO project 2005-2008 (SPAtial Networks in GEOgraphy), gathering a team of geographers and computer scientists, has conducted empirical studies to survey concepts and measures developed in other related fields such as physics, sociology or communication science. The relevancy and potential interpretation of weighted or non weighted measures on edges and nodes were examined and analyzed at different scales (intra-urban, inter-urban or both). New classification and clustering schemes based on the relative local density of subgraphs were developed. The article describes how these notions and methods bring a contribution on a conceptual level, in terms of measures, delineations, explanatory analysis and visualization of geographical phenomena.

## KEYWORDS

Urban Systems, Networks, Graphs, Small Worlds, Multi-level approach, Multi geographical scale

## INTRODUCTION

The increasing complexity of spatial organizations, in particular of cities, challenges the study of spatial distribution. Places tend to be seen more and more as nodes in specialized networks and their future is undoubtedly strongly dependant on their position in these networks (Castells, 1996). Geographical distances and accessibility always matter in spatial dynamics, but in parallel, other distances appear as relevant: social, economical, cultural, organizational distances can also interfere in spatial dynamics. As a consequence, Geography as a science of society in space, must work as defining and measuring how spatial accessibility dynamics influence other kinds of distances, and how processes emerge from spatial or territorial dimensions. From this perspective, in a social science work division, geography could be defined as the science dedicated to the identification of spatial patterns and their implications at specific geographical scale levels.

These geographical levels are themselves evolving according to systems of social processes. For example, globalization processes are developing as transport and communication speed is increasing (and as costs are decreasing), but also according to organizational transformations of society, deriving from actors themselves, like companies and subsidiaries, social groups and institutional arrangements. These types of interactions and their dynamics, explain why globalization has been developed at a worldwide scale in terms of stock exchange, while "global value chains" of production are developing, most of the time, at a continental scale leveraging from "regional" international trade arrangements (Doz et al., 2001; Dicken et al., 2001).

The hypothesis we just laid down suggests that all transformations strongly depend on internal and external component relations, calling for a study of the relative positions of places in networks and at different geographical scales. Geographical interactions can be studied through multi-dimension networks, which can be considered as graphs. Until now, most geographical network studies were based either on Gravity models (Wilson, 1967) or on general indexes of graph connectivity or accessibility (Kansky, 1963). Graph theory and more particularly Watts' theory of small world networks (Watts, 1999), can feed this trend of research and improve spatial analysis as a complement to systems theory. Networks indeed organize through direct links or by indirect paths, inducing topological proximities simultaneously involving spatial, social, cultural or organizational dimensions. Network synergies, either based on similarities or on complementarities between or inside cities, create specialized and more or less stable proximities between these entities (Powell, 1990). Interrelations between geographical scales of organization also allow understanding equilibrium or disequilibrium of territories emerging at different geographical scales. The concepts and methods of the small world theory are particularly relevant to geography where spatial interaction is mainstream, and where interactions can be described and studied using large volumes of exchanges or similarity matrices. As far as we know, this type of empirical approach combining a conceptual approach of "small world theory" and dedicated tools has not been developed in Geography.

Incidentally, the multi-level analysis and the visualization of graphs bring about new questions. How do networks link similar actors reinforcing their social proximities as well as spatial specializations and segregations? In this sense, how do networks (not only transport and communication networks but also social and economic networks) increase anisotropy of space? How far are networks from random processes, creating densely connected groups of nodes separated from others like "Small world" networks (Watts, Strogatz, 1998; Watts, 1999)? What are the properties of these networks and what consequences can we infer on geographical places especially for cities?

In this presentation, we'll stress on the advantages to analyze and visualize networks through their network properties (section "Small world theory and scale-free networks"). Then, we will underline topological positions in a static state as well as from a dynamic perspective (section "Visualizing topological proximities in geographical networks") and transfer them onto properties of city systems (section "Network properties as city systems' properties"). The final purpose discusses levels of organization, taking into account empirical results on hierarchical clustering (section "Multi-level clustering approaches").

## SMALL WORLD THEORY AND SCALE-FREE NETWORKS

This chapter is devoted to the introduction of small world theory and scale-free networks. Laying down these formal notions in a separate section will allow us to freely build the discussion around examples studied by members of the SPANGEO project without having to pause for mathematical digressions.

Graph theory, developed especially by Erdös and his school since the 1950s was based on the concept of "random graph", making the strong assumption that all relationships are likely to occur with equal probability (Erdös & Renyi, 1959, 1960). Put simply a random graph is one where edges are drawn at random playing heads or tails – that is there are no region in the graph where edges have a higher probability to appear, leading to uniformly distributed connections among nodes.

Since the late 1990s, many contributions have proposed alternatives to this model, showing the importance of "small worlds" in the organization of societies and their dynamics, where networks are not seen as homogeneous (Newman, 2000; Newmann et al., 2006). That is, small worlds are built from groups of tightly connected *communities*, themselves connected through privileged nodes. The small world phenomena experimentally identified by Milgram (1967) in social networks was observed in many other fields, giving this notion a taste of universality. In biology, for instance, networks of interacting proteins show a small world character, where communities of proteins are embody cell functions (Vespignani, 2003). The internet is another well known and extensively studied example (Adamic, 1999). Two main properties typify these networks:

- The average path length from one individual to another is small and compares to the average path length of random graphs (with the same number of nodes and edges);
- They have a strong propensity to create sub-groups (also called communities or clusters). This is the strongest characteristics of small world networks, which is the result of several micro processes like transitivity, homophily, and range of ties.

Different statistics can be used to assess of the presence of communities in a graph and confirm its small-worldness. The well known *clustering coefficient* introduced by Watts relates to transitivity and roughly measures the probability that neighbors of a node are connected. The work by Watts & Strogatz (1998) has given birth to several extensions and variations of this measure. Ancient work also provides ingredients to investigate networks. The Jaccard measure (Jaccard, 1901), for instance, assigns a value to an edge depending on the number of common neighbors its incident nodes has; the topological overlap matrix index (Ravazs et al., 2002) is similar to the Jaccard index and was introduced in the context of bioinformatics. As mentioned earlier, small worlds organize into subgroups and are connected through privileged nodes acting as *bridges*; Burt (2000, 2005) refers to this phenomenon as *structural holes*. Many authors have designed metrics aiming at the identification of these bridges, the most used certainly being the *betweenness centrality* defined by Freeman (1977) (see also Brandes, 2001). The *strength metric* defined in Auber et al. (2003) also aims at the identification of bridges (see Sallaberry & Melançon (2008) for a short survey on edge metrics).

In parallel to Watts and Strogatz's theory of small worlds, some authors have observed and investigated another property of networks, referring to them as being scale-free. These networks organize into a hierarchical order with respect to the degree of nodes and evolve along a basic process. According to Barabási & Albert (1999) (see also Barabasi, 2002) who popularized these "scale free networks", this order hierarchy is due to the *"preferential attachment"* that characterizes the development of networks: new links preferentially link to nodes already having the largest number of links, favoring the hierarchy. In reality, we can find previously such explanations of the Zipf law proposed by Simon (1955), referring to the *"Yule process"* (Yule, 1925).

The process thus induces the distribution degree of nodes to follow a power-law, that is the frequency of nodes of degree $k$ is approximately given by $p(k) \sim c\,\exp(-\gamma k)$. In other words, a sharp inequality exists between strongly central entities and others who cling to these centralities[1].

The absence of scale of spatial networks makes sense, as the forthcoming sections shall explain. However, it makes things much harder for their visual exploration and (automatic) analysis. Most often, networks show both characters simultaneously. That is, networks do contain small communities of interacting entities, while the overall organization of the network is dominated by a scale-free process. The high degree nodes thus, make it hard to identify communities or bridges and hide the small world structure. Hybrid approaches have to be designed in order to reveal the structure and properties of these complex networks.

## VISUALIZING TOPOLOGICAL PROXIMITIES IN GEOGRAPHICAL NETWORKS

Visualizations are common tools for geographers, using cartography to stress geographical organizations and patterns. Nevertheless, as far as geographical distance is not the only one factor constraining the system and its dynamics, a topological representation can be much more useful to highlight organizational structures.

A basic geographical example is the worldwide container traffic network, where the visualization allows a good overview of the overall structure (Figure 1) (Ducruet et al., 2010). On the one hand, the representation ignores the geographical coordinates of ports. On the other hand, their location in the graph is determined by applying a layout based on topological proximities for maritime exchanges. Therefore, the geographical situation is preserved due to the continual movement from one port to another.

Fundamentally, two main sub-systems exist: Asia-Pacific, and Europe-Atlantic. These sub-systems are connected through a limited number of nodes: Panama and Suez canals. Removing those two global pivots would thus split the world system in two parts, although one may notice some other links (e.g. Magellan Strait and Cape of Good Hope) but those remain very limited compared to the main trunk lines concentrated at the two canals. Also, Port ranges appear as the Scandinavian range in the top left of the figure 1, close to American ranges.

[1] The interested reader is referred to Bornholdt & Schuster (2003).

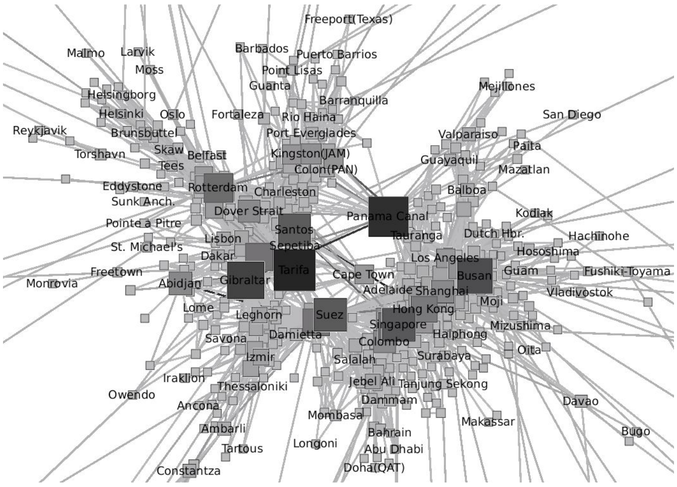


Figure 1: Worldwide container traffic Network 2006

Another example maintaining some geographical positions in the topologic network representation is the topic of commuters (Figure 2). The example of Switzerland shows especially the linguistic proximities between municipalities with 3 visible isolated zones: Swiss French speaking at left, Italian speaking at right, and Swiss German in the center, cut into two parts between a system around Bern (the political federal capital), and another around Zürich (the economic capital).

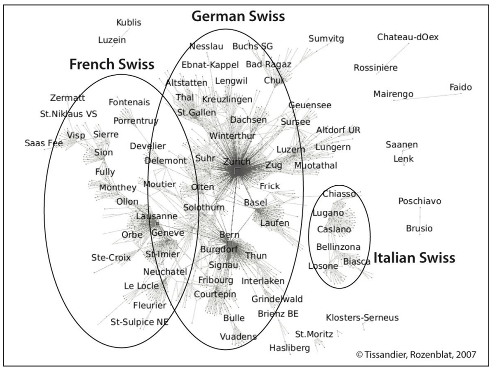


Figure 2: Commuters between Swiss Municipalities in 2000

## NETWORK PROPERTIES AS CITY SYSTEMS' PROPERTIES

### Urban systems as small worlds

The small world property has a clear interpretation for urban systems in general. The short average distance between entities can be seen as a consequence of redundant paths through different routes between city networks, strengthening direct and/or indirect interaction. This mutual interaction occurs mainly through "clusters" which can form inside cities or towns, heavily connected by groups.

Empirical evidences of these networks are founded on some basic properties reinforcing these strengths (Monge & Contractor, 2003) which, in transposition, could make sense for geographical network organizations:

- Transitivity: This elementary property was early shown by sociologists (Weimann, 1980): If an individual A is linked to B, and B to C, then the likelihood is high that A should be linked to C.

  In geographical perspectives, we can transpose this property to interaction and interdependency effects. If a place A is closely related and interdependent to B, and B to C then A is interdependent to C. For example, commuter exchanges between municipalities create this kind of chain dependencies (Figure 3). If a factory employing several thousands of people in A suddenly closes, it will affect residential population of B sending commuters in A, but in consequence, it also will affect C sending commuters to B. Sometimes, this transitive relation is materialized by real exchanges (like commuters), but even if there's not such direct materialized exchanges, interdependencies occur as well through indirect effects. So some local chains create tangled networks allowing to very integrated local territory systems (Tissandier et al., 2010).

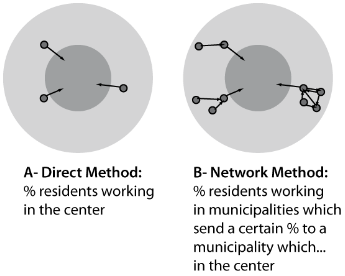


Figure 3: Urban Areas delineation using transitivity property

At the inter-urban scale, interdependencies also make sense in mutual exchanges (symmetric or not), increasing proximities in terms of space (local or regional economical networks) or in terms of specialization (worldwide clusters linked through multinational firms' global value chains or through research centers). These new proximities stimulate growth of exchanges, strengthening closeness of some specific groups of cities or territories in spatial, economic, social or cultural proximities.

- Homophily: Networks between actors, territories or cities with similar attributes have larger probability of occurring. These exogenous criteria interact actually with networks because exchanges could also transform places as they transform individuals with imitation behaviors and diffusions of many types of ideas, concepts and technologies.
- Range: in parallel to homophily, diversity of elements of places and cities linked by networks seems essential to the reproduction and renewal of territorial systems. This diversity could be seen in terms of exogenous characteristics (for example diversity of activity sectors) as well as endogenous: the diversity of links to other places.

**Urban systems as scale-free networks - Hierarchies of individuals**

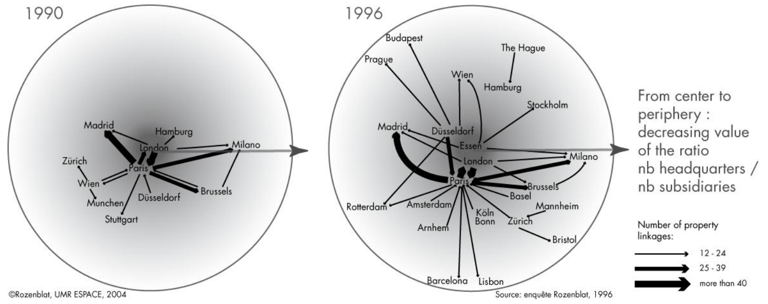
There are evidences in urban systems that innovations spread preferentially from the largest cities, already concentrating multiple networks, to the smallest ones (Batty, 2005; Pumain, 2006). At a micro level of individual behavior, these processes are in a large part due to the search of security rather than maximization, in a context of partial information in game theory (Luce & Raiffa, 1957). The "bounded rationality" mainly developed by Simon (1957, 1972, 1999) explains, in a context of random probability of meeting opportunities, why individuals choose the first satisfactory opportunity, rather than the best one. These trends lead to a strengthening of major cities, where opportunities are more diversified and where interactions are maximized, decreasing transaction costs (Coase, 1937). This process enables some specific properties of centralities, equivalence or structural equivalence positions for places or for groups of places according to positions of located individuals.

- Centralities: inequalities between places in terms of relative position in networks are obviously dependent on accessibility or reachability into some sets of places, in a certain context and at a certain period (Bretagnolle, 1999). Many kinds of indexes bring close views on these central positions, like ***Degree*** (number of links), ***Reachability*** (distance to all other points), ***Betweenness Centrality*** (number of shortest paths passing by a point), or ***Generalized Strahler*** index (Delest & Aubert, 2003; Auber et al., 2004) measuring how many trees could be built from a node : every measure can be weighted or not (Wasserman & Faust, 1994; Newmann, 2003; Brandes & Erlebach, 2005). They are all very useful to compare node positions in a graph, or even positions of the same node in different graphs (multiplex graph).

- Dependencies and power: when social or economic networks are oriented, either reflecting ownership between enterprises, decision-making, or hierarchy. The balance between in-degree ($Deg^{in}_i$) and out Degree ($Deg^{out}_i$) can be a measure of such dependencies or power. Thus, an index of power $P_i$ can be calculated for each city (Rozenblat, 2004):

  $$P_i = Deg^{out}{}_i - Deg^{in}{}_i$$

  For example, this methodology can easily underline cities of power in multinational firms’ networks according to the opposition between headquarters and subsidiaries (Figure 4). The ties are oriented and the centrality in the graph is calculated by the share of the number of Headquarters (representing the out Degree) by the number of subsidiaries (in Degree). Between 1990 and 1996, the graph has spread all over Europe in particular areas like south and Eastern Europe. London, with a high share of headquarter/subsidiaries, is the most central with many American Headquarters controlling subsidiaries in the continent. But Paris seems much more connected to other European big cities.

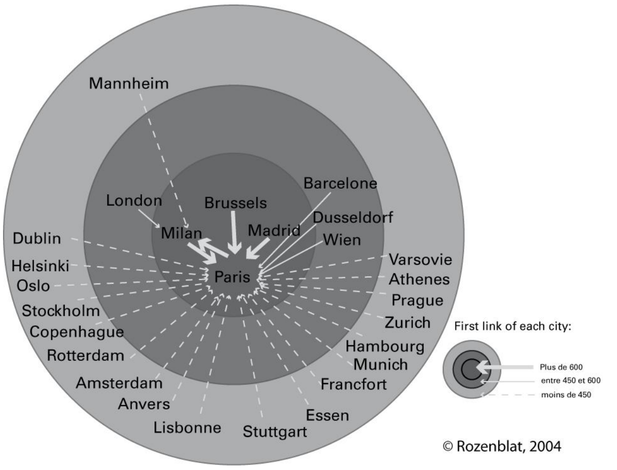


©Rozenblat, UMR ESPACE, 2004

Source: enquête Rozenblat, 1996

Figure 4: Hierarchy of European cities according to their control of Multinational firms 1990-1996

Also, one can highlight dual positions where some cities are only linked to another. According to Nystuen & Dacey's methodology (1961), one can build some conceptual hierarchy with the highest link of each city connecting to a bigger city. All these links create some "regional systems". So, hierarchy can also emerge from non-oriented ties according to weighted dependencies. Figure 5 shows this construction applied to European city connections through complete graphs built inside each firm of a sample of 100 groups in 1996 (Figure 5). The first link of the majority of European cities is then oriented towards Paris, the most connected city according to this kind of firms' graph.



© Rozenblat, 2004

Figure 5: Hierarchy of European cities according to their positions in Multinational firms in 1996

- Equivalence and Structural equivalence: Organizational structure is viewed as a pattern of relations among positions. White et al. (1976) and Burt (1982) have developed positional theories of structural equivalence According to Monge & Contractor, this theory "argues that people maintain attitudes, values, and beliefs consistent with their organizational positions irrespective of the amount of communication that they have with others in their organizational networks" (2003: 19). In a geographical perspective, we could argue that places like cities are formed by tangled and overlapped social, economic and institutional networks which put them in equivalent positions as well. For example, cities of the same country are equivalent faced with their political national capital in many ways from hierarchical administration to hierarchical economy and culture. On the other hand, national capitals are in such structural equivalence faced with their respective national urban systems even if some national networks are more hierarchical than others.

The confrontation of weighted degree and betweenness centrality of each city in every level allows to underline *Bridge cities* (Barrat et al., 2005; Guimera et al., 2005). In fact, degree and betweenness centrality are very correlated even if they have different natures: local (for degree) and global (for BC). But a low degree with a high betweenness centrality shows especially the role of bridges between clusters (Figure 6-A). We can discuss the threshold to choose in order to select this kind of bridges (Figure 6-B). On the other hand, we can enlarge this approach testing all kinds of confrontations of local and global measures.

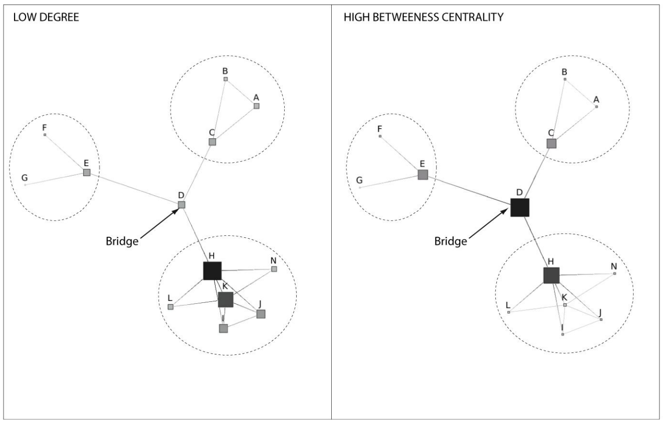


Figure 6: Identification of bridges

Individuals or organizations acting as bridges (or connectors) can override the "structural holes" strengthening the entire "social capital", but in the first place their own capital (Burt, 2000, 2005). These bridges have a very strong strategic centrality even if they have a small number of links (Guimera et al., 2005). A city can form a "bridge" between several cities when it is located at the interface between different groups or clusters, or between different units (as a relay between its national cities and abroad, between continental and the world).

In dynamic terms, such urban placements require high flexibility in networks and strong adaptability of cities and reactivity for their actors. In contrast, too strong a hierarchy between nodes freezes the system, but leaves individuals excluded from all networks. The limitations of networks between what is integrated and what is not are highly constitutive of their operation and evolution (Barabasi, 2003).

### Hierarchies of groups

If we assume that interaction implies interdependencies, then groups of cities which exchange more together than with others are more dependent on each other. They constitute groups that can be defined in different ways (we'll see below). But whatever method of classification is used, these groups imply some relative closure, due in part to transitivity property seen previously. For other cases, they constitute some star around a central node, linked itself to outside (a Hub). So, a kind of "cohesion" measuring the relative density of the group's graph can be interpreted as strength of interdependencies. As in classical classifications, some nodes (like cities) contribute more than others to the cohesion of their groups: this is what we'll call the *"contribution"* of each point to its group. Besides, each city can also have links going out to the group. Then, we can also define a second measure, specifying from the point of view of each node, its participation to several groups: this is what we'll call "*participation"* (Guimera et al., 2005). Contribution and participation indexes can be measured as individual shares for one node to one group (or the opposite).

For contribution of a node i to a group g:

$$Contr_i^g = \frac{Deg_i^g}{\sum_{i=1}^{n} Deg_i^g}$$

Where $Deg_i^g$ is the degree of the node i (number of links) in a group g.

For the participation of the node i to a group g:

$$Part_i^g = \frac{Deg_i^g}{Deg_i}$$

Where $Deg_i^g$ is the degree of the node i inside the group g and $Deg_i$ is the total degree of i.

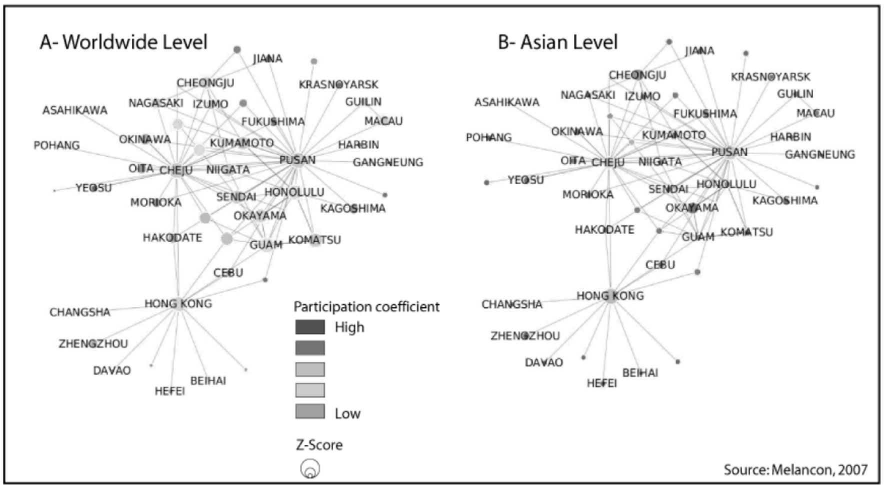
These kinds of indexes can identify hubs in networks. Their role is often concentrated in or between particular communities. Extending Guimera et al. (2005) who introduced z-score and the "*participation coefficient*" of nodes, we take edge weights into account. The z-score of a node measures how much of the node's connection is devoted to its own cluster, while its participating coefficient measures how much its connections cover all other clusters. Because we cluster the graph into a hierarchy of subgraphs, we can moreover measure Guimera et al. (2005) indices at every level and study how the indices vary as we go down the hierarchy of clusters. Compared to our indexes the Guimera et al. (2005) indices are:

$$z - score_i = Part_i^g$$

where g is the main group of i.

$$Guimera - Participation_i = 1 - \sum_{g=1}^{p} Part_i^{g^2}$$

The figure below provides a good illustration (Figure 7). Nodes have been sized according to their z-score (internal connections) and colored according to their participation coefficient (external connections). For example, Hong-Kong has got a higher role at the worldwide level than at the Asian level.



Source: Melancon, 2007

Figure 7: Identification of Hubs by Contribution indexes

For a more global view of each node's role, an entropy index H could take into account the balance between contributions of a node to different groups. It measures the share of the maximal entropy: LOG(n). For Entropy of n nodes i contribution to the group g:

$$HContr_g = \sum\nolimits_{i=1}^{n} \frac{Deg_i^g}{\sum\nolimits_{i=1}^{n} Deg_i^g} \times LOG\left(\frac{Deg_i^g}{\sum\nolimits_{i=1}^{n} Deg_i^g}\right) \Bigg/ LOG(n)$$

Symmetrically, an entropy index H could take into account the balance between participations of different nodes to a group. It measures the share of the maximal entropy: LOG(p). For entropy of the node i participation to different p groups g:

$$HPart_i = \sum\nolimits_{g=1}^{p} \frac{Deg_i^g}{Deg_i} \times LOG\left(\frac{Deg_i^g}{Deg_i}\right) \Bigg/ LOG(p)$$

The signification of these entropy indexes could be interpreted as is usual for entropy: the degree of information contained into each group or information for each node (Shannon, 1948; Lin, 1999). The higher the information included inside a node or a group, the higher is the resilience and the capability to renew. So, the maximum entropy corresponds to a great variety of links for each node (balanced participation to many groups) or at the level of clusters, various contributions of nodes forming each group (and not only one in the case of star patterns). This variety contains a high disorder and could be interpreted as the systems' high capability to maintain and renew (Bailey, 1990; Burt, 2005). It illustrates for a particular node, the "*strength of Weak Ties*" of Granovetter (1973) where a node has got more opportunity of a large number of weak links relying on differentiated clusters formed by larger links. In those clusters, vicinities' economic links can develop processes of agglomeration economies or network economies. However weak links between clusters represent diversity, which also produces saving networks, which allows cities both to operate and to renew themselves.

This coupling between strong and weak links allows the reproduction of the urban system and its transformations (Guimera et al., 2005; Uzzi, 1997). We could identify such types of clusters between cities from different kinds of networks (transport, communication, migrations, multinational companies etc.), and we could specify for each city the intensity of its membership in a cluster, assess to what extent these clusters correspond to territorial logic (of continents, country, urban areas), or logic of economic specialization (cities specialized in the same fields).

Besides diversity, specialization and "closure" are common patterns in networks, according to the general form of "scale free" laws (see before). Then the $\beta$ index of the "scale free function" can also constitute a norm to measure high concentration (Barabasi & Albert, 1996; Batty, 2005; Paulus et al., 2006). Then the measures of scale free contributions inside each group could be adopted, but the difficulty remains in defining these groups and their relative positions.

## MULTI-LEVEL CLUSTERING APPROACHES

Clusters approaches are easy if clusters are disjoint. But clusters can also be formed at different levels, which can be nested, tangled or overlapped. Therefore, a multi-level approach is useful to perceive all the dimensions of the processes, to compare them and to identify their complementary roles in the system dynamics.

The “small world” networks approach identifies those parts of networks — such as “clusters” — which are the densest and most coherent (Watts & Strogatz, 1998; Watts, 1999; Barabasi & Albert, 2000; Barabasi, 2003; Newmann et al., 2006). Such parts not only demonstrate a high level of dependency between their individual components, but also a low level of “vulnerability” in networks that allow for numerous redundancies (Burt, 2000, 2005). Thus, both through direct links and by more indirect paths, networks organize new topological “proximities” constituted simultaneously of spatial, social, cultural or organizational proximities. In 2004, Newman and Girvan posited a measure of clustering on non-value graphs predicated on the maximization of intra-group connections (Q value). We adapted it to weighted graphs.

G a weighted graph

C= {C1, C2, ..., Cp} a partition of this graph into clusters C1, C2,..., Cp

w(e) is the weight of the edge e (Passenger Traffic in the case of the air worldwide passenger network, for example).

The weighted quality measure $Q_{val}$ is defined as follows:

$Q_{val} = \sum i\ Q_{val}(Ci),\ 1 \leq i \leq p.$

and

$Q_{val}(Ci) = (\sum_{e \text{ in } E(Ci)} w(e) / \sum_{e \text{ in } E} w(e)\ ) - (\sum_{e \text{ in } E(Ci,*)} w(e) / \sum_{e \text{ in } E} w(e)\ )^2$

The application of this method to the international air-traffic network — whether it is non-valued (Guimera & al., 2005), or valued, nevertheless demonstrates that groups of cities organize themselves geographically by continents (Figure 8).

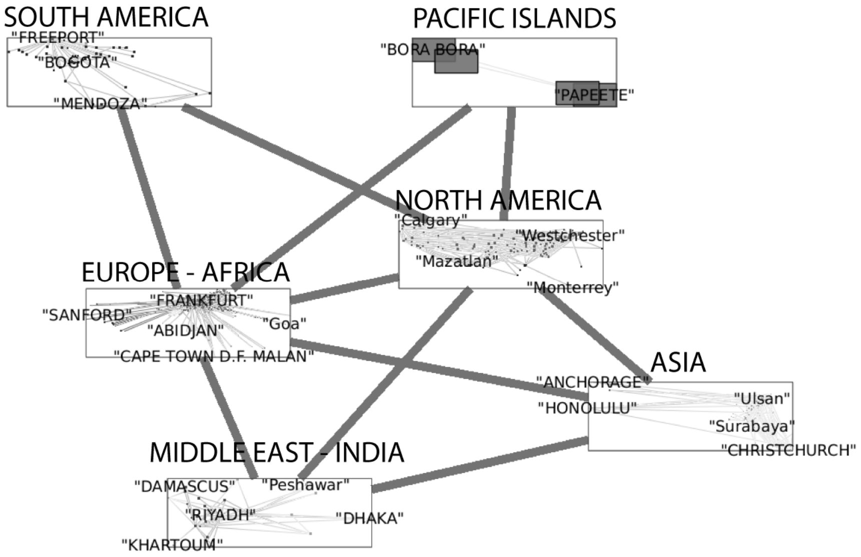


Figure 8: Classification of Worldwide air traffic according to Weighted Quality Mesure (Qval)

The clusters are seen as communities and are captured with the use of edge metrics conveying topological properties of the underlying communities. The metrics we used are from Auber et al. (2003) adapted to weighted edges – traffic flows (see also Amiel, Melançon & Rozenblat, 2005). A Strength index is calculated for each edge (corresponding to the density of ties around this edge) (Figure 9). Then filtering out edges with a low value induces a graph whose connected components form "communities". This procedure is iterated into each community to define sub-communities and so on at various levels.

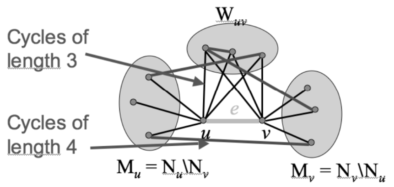


Figure 9: Measure of an edge strength index

For this operation, one calculates a ratio based on the number of cycles of length 3 involving an edge e:

$$\gamma_3(e) = \frac{|\mathrm{W}_{uv}|}{|\mathrm{M}_u| + |\mathrm{M}_u| + |\mathrm{W}_{uv}|}$$

A similar ratio can be defined for cycles of length 4:

$$\gamma_4(e) = \mathrm{s}(\mathrm{M}_u, \mathrm{W}_{uv}) + \mathrm{s}(\mathrm{M}_v, \mathrm{W}_{uv}) + \mathrm{s}(\mathrm{M}_u, \mathrm{M}_v) + \mathrm{s}(\mathrm{W}_{uv}, \mathrm{W}_{uv})$$

Then the strength index is the sum of both: $\Sigma(e) = \gamma_3(e) + \gamma_4(e)$

To ignore edges with low strength indexes, we used a Quality index (Delest et al., 2006). This metric completes Burt's network constraint, measuring how much an actor is constrained by his/her direct neighborhood (Burt, 2000, 2005). In both cases, it is expected that edges with a low value act as bridges between tighter communities.

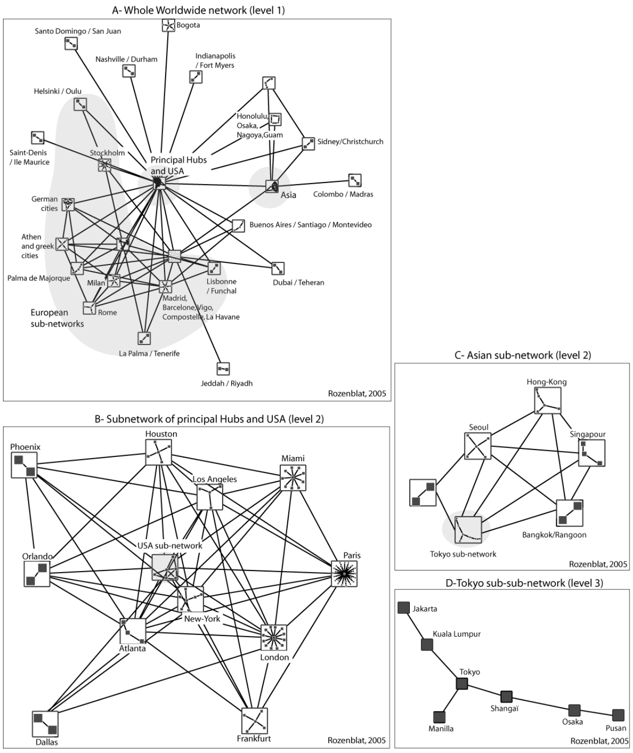


Figure 10: Classification of Worldwide air traffic according to Strength index

As can be expected, communities become more and more cohesive as levels go down (Figure 10). In this approach, the principal hubs are all organized in the same class (Figure 10.b). While geographical proximities persist, they are also involved in local hierarchies, either around the national capitals of Europe (Fan, 2006) or they are established by airline companies, as it is the case in Asia (Figure 10.c and 10.d).

Because it allows as much weight to the density of the graphs as to the strength of connections, this method appears more suited to represent a collection of weak connections forming multi-level groups of cities organized according to their proximities both geographical and functional. The levels are distinguished by the degree of cohesion that they contribute to the network. They may be constituted with respect to geographical imperatives (continents), or to organizational (hubs) or economic imperatives (airline companies). A further improvement to this approach to clustering would presently consist in defining not strictly circumscribed groups but rather a "fuzzy" belonging of each city to various groups, which would allow for the interactions of a given city to be deployed over a wide variety of levels at once. Results could help measure how network levels are evolving in a multi-dimensional space comprising the companies' alliances and competitions, the network economies given by the hub system, the global laws of deregulation, restrictions for environment and the transportation demand.

## CONCLUSION

The developing organization of cities into multi-level networks seems to occur not only in the field of transport systems but in organizational systems as well (Dicken & al., 2001). While it is undoubtedly true that agent networks transcend the limitations of geographical scales (Castells, 1996), it is also the case — as Whitley has demonstrated (1998) — that the organizational levels of the "global commodity chain" are in the process of crystallizing from local to worldwide scales. In the context of this realization, multinational corporations do not develop within territories that are straightaway homogenous and open as argued by Ohmae (1990); rather they transversally weave a system of territories — overlapping or not as the case may be — in which everyone produces their own rules and regulations. The structure of the "transnational field" continues to be strongly influenced by the "international field" and so continues to depend on "the exchange rates, labor and fiscal regulation, those 'externalities' of which the company can take advantage of, and the size and quality of the market." (Dollfus, 2001: 104-105). Considered from this perspective, the continental level is seen to increase its cohesive role by virtue of the creation of free-trade zones, which tends to reinforce continental systems (Ohmae, 1995; Yeung, 2002; Rugman, 2001; Rozenblat, 2004; Dicken, 2007), but it is not the only one in a multi-dimensional point of view.

In parallel to this, intense economic specialization creates groups of cities which are more and more interrelated (a tendency that is reinforced by policies of “poles of excellence” or “poles of competitiveness”). The most frequently quoted example of this specialization is the financial “global city” linking New York, London and Tokyo (Sassen, 1991). The financial capitals of the world as an ensemble attach themselves to each other either directly or indirectly by constituting specialized subgroups or/and geographical sub-groups.

Between these groups of highly interlinked cities, “relay” cities constitute “obligatory pathways” between the groups that can be depicted as “bridges” (Figure 6). A telling example of this is constituted by the national capitals which host the national head-offices of the foreign subsidiaries of multinational corporations — head-offices which, in their turn, invest in businesses or establishments elsewhere than in the host country. (The same model also functions at a continental level articulated around continental centers (Pumain & Rozenblat, 1993; Rugman, 2001; Alderson & Beckfield, 2004; Rozenblat, 2004; Rozenblat & Pumain, 2007). Employee meetings are often located at these “intermediate” head-offices. At the level of cities, capitals perform the function of “bridgehead” for all the other national cities: for capital cities dramatize themselves by the expedient, for example, of providing a preferential location for representational offices (Brussels in Europe). Through this role of “display relay” for all the other cities of a given country, the capital thus concentrates a large proportion of relevant relationships both upstream and downstream: international and national. This position of “relay” or “bridge”, affords the capital better access to the whole network as well as increased control over information transfers (Rozenblat & Pumain, 2007). In a further dimension, the network positions of different cities which share similar attributes of “dependency” with regard to “relay” cities, display “equivalent” characteristics. Within each “sub-group” a marked hierarchy of cities’ levels of centrality in the network is maintained, while other cities remain altogether peripheral, being only attached indirectly to the network by virtue of their connection to participating cities. The recent reinforcement of the centrality of European national capitals is due to this process (Rozenblat & Cicille, 2003; Grasland, 2006; Rozenblat et al., 2008). To this end, one can wonder about the accumulation of social capital of cities in certain industries or businesses: to what extent networks of certain activities, such as chemicals, can cause the emergence of other networks in relatively remote areas? To do this, we can refer to studies focusing on linkages between activities within certain cities (Hall & Pain, 2006).

The clusters and hierarchies create dynamics that can support, or move against the sustainable development of urban systems, which can be defined as the ability to renew and maintain geographical and social networks' diversification. Thus, in some types of network, "Clustered" components disconnected from the rest of the network can easily show that some groups of cities or clusters can be found relatively isolated from the rest of the urban system, without being able to overcome the decline of their core business. This process of fragmentation is detrimental to the isolated cluster that can not participate in the rest of the network, but also for the entire system that loses its diversity. In contrast, one can assume that the cities of the same country do not develop all the same networks of expertise. This contributes to maintain a diversity of national urban networks, provided that those mutual links between cities maintain a certain level of relations to enjoy this diversity.

It is of essential importance that our knowledge of all the properties of these networks of cities should be deepened in a multi-level perspective. Identifying, qualifying and measuring each level could lead to a better understanding of globalization, because at each level, there are so many processes that allow the emergence of urban properties in which "the whole becomes not only more than but very different from the sum of its parts" (Anderson (1972) in Lane, 2006).

## ACKNOWLEDGEMENTS

The authors would like to thank the editors of the Handbook to allow them publishing a synthetic view of their collaboration. This collaboration has been concretized in the SPANGEO project, managed by the two authors and joining Geography with Computer Science. It was supported from 2005 to 2008 by the French national agency of research. The collaboration of the authors has been largely enriched and developed by all the members of the SPANGEO project who collaborated in a very motivating and enthusiastic way. In this respect, the paper has a big debt to all the group and the authors would like to thank them warmely: Charles Albert-Bohan, David Auber, Romain Bourqui, Marie-Noelle Comin, Maylis Delest, César Ducruet, Bérengère Gautier, Jean-François Gleyze, Pierre-Yves Koenig, Alain L'Hostis, Patrick Mary, Morgan Mathiaut, Eric Mermet, Trung-Tien Phan-Quang, Francesca Rota, Patrice Tissandier, Sandra Vinciguerra.

## REFERENCES

**Adamic, L.A.**, 1999, The Small World Web. In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science 1696. New York: Springer-Verlag: 443-452

**Alderson, A.S.; Beckfield, J.**, 2004, Power and position in the World City System. In: American Journal of Sociology, 109, 4: 811-851

**Amiel, M.; Mélançon, G.; Rozenblat, C.**, 2005, Réseaux multi-niveaux: l'exemple des échanges aériens mondiaux. In: Mappemonde, 79-3. http://mappemonde.mgm.fr/num7/index.html

**Auber, D.; Chiricota, Y.; Jourdan, F.; Melançon G**., 2003, Multiscale navigation of Small World Networks. In: IEEE Symposium on Information Visualisation. Seattle, GA, USA: IEEE Computer Science Press: 75-81

**Auber, D.; Delest, M.; Chiricota, Y.**, 2004, Strahler based Graph Clustering using Convolution. In: Proceedings of the Eighth International Conference on Information Visualisation (IV'04), IEEE Computer Society: 44-51

**Bagler, G.**, 2008, Analysis of the Airport Network of India as a complex weighted network. In: Physica A, 387: 2972-2980

**Bailey, K.D.**, 1990, Social Entropy Theory. State University of New York Press: New York

**Barabási, A.; Albert, R.**, 1999, Emergence of scaling in random networks. In: Science, 286: 509-512

**Barabási, A.**, 2002, The New Science of Networks. Perseus: Cambridge, MA

**Barrat, A.; Barthélémy, M.; Vespignani, A.**, 2005, The effects of spatial constraints on the evolution of weighted complex networks. In: J. Stat. Mech., P05003

**Batty, M.**, 2005, Cities and Complexity: understanding Cities with Cellular Automata, Agent-Based Models and Fractals. The MIT Press: Cambrige, MA

**Bornholdt, S.; Schuster, G.** (Eds), 2003, Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH: London

**Brandes, U.**, 2001, A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical Sociology 25, 2: 163-177

**Brandes, U.; Erlebach, T.**, 2005, Handbook of Network Analysis: Methodological foundations. Springer: Berlin

**Bretagnolle, A.; Pumain, D.; Rozenblat, C.,** 1999, Space-time Contraction and the Dynamics of Urban Systems. In: Cybergeo: Revue européenne de géographie. http://www.cybergeo.eu/index373.html

**Burt, R.**, 2000, The Network Structure of social Capital. In: Sutton, R.; Staw, B. (Eds), Research in organizational Behavior. JAI Press: Greenwich, CT

**Burt, R.**, 2005, Brokerage and Closure: an introduction of social Capital. Oxford University Press: Oxford

**Button, K.J.; Taylor, S.Y.**, 2000, International air transportation and economic development. In: Journal of Air Transport Management, 6: 209-222

**Castells, M.**, 1996, The rise of the Network Society. Blackwell: Oxford

**Coase, R.**, 1937, The nature of the firm. In: Economica, 4: 386-405

**Delest, M.; Auber, D.**, 2003, A Clustering Algorithm for Huge Trees. In: Advances in Applied Mathematics, 31: 46-60

**Delest, M.; Fedou, J.M.; Melancon, G.**, 2006, A Quality Measure for Multi-Level Community Structure. Paper presented at the SYNASC 8th International Conference, September, Timisoara, Romania

**Dicken, P.**, 2007, Global Shift: Mapping the Changing Contours of the World Economy. Sage: London, 5th edition

**Dicken, P.; Kelly, P.F.; Olds, Kr.; Wai-Chung Yeung, H.**, 2001, Chain networks, territories and scales: toward a relational framework for analysing the global economy. In: Global Networks, 1, 2: 89-112

**Ducruet, C.; Rozenblat, C., Zaidi, F.,** 2010, Ports in a world maritime system: a multi-level analysis. In: Journal of Transport Geography (forthcoming)

**Erdős, P.; Rényi, A.**, 1959, On Random Graphs, I. In: Publicationes Mathematicae, 6: 290-297

**Erdös, P.; Rényi, A.**, 1960, The Evolution of Random Graphs. In: Magyar Tud. Akad. Mat. Kutató Int. Közl., 5: 17-61

**Fan, T.**, 2006, Improvements in intra-European inter-city flight connectivity:1996–2004. In: Journal of Transport Geography, 14: 273-286

**Freeman, L.C.**, 1977, A set of measures of centrality based upon betweenness. In: Sociometry, 40: 35-41

**Grabher, G.**, 2006, Trading routes, bypasses, and risky intersections: mapping travels of networks between economic sociology and economic geography. In: Progress in Human Geography, 30, 2: 163-189

**Granovetter, M.**, 1973, The strength of Weak Ties. In: American journal of sociology, 78, 6:1360-1380

**Guimera, R.; Mossa, S.; Turtschi, A.; Amaral, L.A.N.**, 2005, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. In: PNAS, 102, 22: 7794-7799

**Hall P.; Pain, C.**, 2006, The polycentric metropolis: learning from Mega-City Regions in Europe. Earthscan: London

**Kansky, K.**, 1963, Structure of transportation networks: relationships between network geography and regional characteristics. University of Chicago, Department of Geography, Research Papers 84

**Lane, D.**, 2006, Hierarchy, Complexity, Society. In: Pumain, D. (Ed.), Hierarchy in Natural and Social Sciences. Springer: New York: 81-119

**Lin, S.K.**, 1999, Diversity and Entropy. In: Entropy (Journal), 1, 1: 1-3

**Luce, R.D.; Raiffa, H.**, 1957, Games and Decisions: Introduction and Critical Survey. Wiley & Sons: New York (see Chapter 3, section 2)

**Melançon, G.**, 2006, Just how dense are dense graphs in the real world? A methodological note. Paper presented at the BELIV Workshop within AVI Conference, Venice Italy

**Milgram, St.**, 1967, The Small World. In: Psychology Today, 2: 60-67

**Monge, P.; Contractor, N.**, 2003, Theories of communication networks. Oxford University Press: Oxford

**Newman, M.**, 2000, Models of the Small World. In: Journal Statistical Physics 101: 819- 841

**Newman, M.**, 2003, The structure and function of complex networks. In: SIAM Review 45: 167-256, arXiv:cond-mat/0303516v1

**Newman, M.; Girvan, M.**, 2004, Finding and evaluating community structure in networks. In: Physical Review E, 69, 026113

**Newman, M.; Watts, D.; Barabási, A.-L.** (Eds), 2006, The Structure and Dynamics of Networks. Princeton University Press: Princeton

**Nystuen, J.D.; Dacey, M.F.**, 1961, A graph theory interpretation of nodal regions. In: Regional science Association, Papers and Proceedings, 7: 29-42

**Ohmae, K.**, 1990, The Borderless World. Harper & Row: New York

**Ohmae, K.**, 1995, The End of the Nation State-How Region States Harness the Prosperity of the Global Economy. Free Press, McMillan Inc.: New York

**Paulus, F.; Pumain, D.; Vacchiani-Marcuzzo, C.; Lobo, J.**, 2006, An evolutionary theory for interpreting urban scaling laws. In: Cybergéo: Revue européenne de géographie. http://www.cybergeo.eu/index2519.html

**Powell, W.W.**, 1990, Neither Market nor Hiearchy: Network Forms of organization. In: Research in Organizational Research, 12: 295-336

**Pumain, D.**, 2006, Alternative explanations of hierarchical differentiation in Urban Systems. In: Pumain, D. (Ed.), Hierarchy in Natural and Social Sciences. Berlin: Springer : 169-222

**Radicchi, F.; Castellano, C.; Cecconi, F.; Loreto, V.; Parisi, D.**, 2004, Defining and identifying communities in networks. In: Proceedings of the National Academy of Science USA, 101: 2658-2663

**Ravasz, E.; Somera, A.L.: Mongru, D.A.; Oltvai, Z.N.; Barabasi, A.**, 2002, Hierarchical Organization of Modularity in Metabolic Networks. In: Science Magazine, 297: 1551-1555

**Rozenblat, C.**, 2004, Tissus de villes: réseaux et systèmes urbains en Europe. Thèse d'Habilitation à Diriger des Recherches

**Rozenblat, C.**, 2007, Villes et réseaux petits-mondes. In: Da Cunha, A.; Matthey, L. (Eds), La ville et l'urbain: des savoirs émergents. Lausanne: Presses Polytechniques romandes: 81-105

**Rozenblat, C.; Mélançon, G.; Amiel, M.; Auber, D.; Discazeaux, C.; L'Hostis, A.; Langlois, P.; Larribe, S.**, 2006, Worldwide Multi-Level Networks of cities emerging from air traffic. Paper presented at the IGU Conference "Cities of Tomorrow", Santiago de Compostela, July 31 to August 7. http://doc.rero.ch/search.py?recid=5893&ln=fr

**Rozenblat, C.; Pumain, D.**, 1993, The location of multinational firms in the European urban system. In: Urban Studies, 10: 1691-1709

**Rugman, A.M.**, 2001, The end of Globalization: why global strategy is a myth & how to profit from realities of regional markets. AMACOM: New York

**Sallaberry, A.; Melançon, G.**, 2008, Edge Metrics for Visual Graph Analytics: A Comparative Study. In: Proceedings of the Information Visualization Conference, London, UK, IEEE Computer Society: 610-615

**Sassen, S.**, 1991, The Global City: New York, London, Tokyo. Princeton University Press: Princeton

**Shannon, Cl.E.**, 1948, A mathematical theory of communication. In: Bell System Technical Journal, 27: 379-423; 623-656

**Simmel, G.**, 1896, Superiority and subordination as subject-matter of sociology. In: American Journal of Sociology, 2, 2: 167-189

**Simon, H.A.**, 1955, On a class of skew distribution functions. In: Biometrika, 42: 425-440

**Simon, H.A.**, 1972, Theory of bounded rationality. In: McGuire, C.B.; Radner, R. (Eds), Decision and organization. New York: Amsterdam: North Holland publishing company: 161-176

**Simon, H.A.**, 1999, Bounded rationality in social science: Today and Tomorrow. In: Mind & Society. Fondazione Rosselli: 25-39

**SPANGEO** Project (2005-2008):
http://s4.parisgeo.cnrs.fr/spangeo/spangeo11.htm

**Tissandier, P.; Phan-Quang, T.; Archambault, D.,** 2010, Defining polycentric Urban Areas through commuting cohesions. In: Rozenblat, C.; Mélançon, G. (Dir.), Multilevel Analysis and visualization of Geographical Networks. Springer (in press)

**Uzzi, Br.**, 1997, Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. In: Administrative Science Quarterly, 42: 35-67

**Vespignani, A.**, 2003, Evolution Thinks Modular. In: Nature, 35, 2: 118-119

**Wasserman, S.; Faust, K.**, 1994, Social network analysis. Cambridge University Press: Cambridge

**Watts, D.J.**, 1999, Small Worlds. Princeton University Press: Princeton

**Watts, D.J.; Strogatz, S.H.**, 1998, Collective Dynamics of Small-World Networks. In: Nature, 393: 440-442

**White, H.C.; Boorman, S.A.; Breiger, R.L.**, 1976, Social structure from multiple networks. I. Blockmodels of roles and positions. In: American Journal of Sociology, 81, 4: 730-780

**Whitley, R.**, 1998, Internationalization and varieties of capitalism: the limited effects of cross-national coordination of the economic activities on the nature of business systems. In: Review of International Political Economy, 5: 445-381

**Wilson, A.G.**, 1967, A Statistical Theory of Spatial Distribution Models. In: Transportation Research, 1: 253-269

**Yeung, H.W.C**., 2002, The limits to globalization theory: a geographic perspective on global economic change. In: Economic Geography, 78: 285-305

**Yule, G.U.**, 1925, A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S. In: Philosophical Transactions of the Royal Society of London, Ser. B, 213: 21-87

## AUTHORS INFORMATION

**Céline ROZENBLAT**
celine.rozenblat@unil.ch
Université de Lausanne
Institut de Géographie –
Faculté des Géosciences
Bâtiment Anthropole - Bureau 4064
Quartier Dorigny
CH - 1015 Lausanne

**Guy MELANÇON**
melancon@labri.fr
INRIA Bordeaux -- Sud-Ouest
CNRS UMR 5800 LaBRI
Campus Université Bordeaux I
351 Cours de la libération
33405 Talence Cedex
France

In 2007, the Institute of Geography of the University of Lausanne (IGUL) organized the Fifteenth European Colloquium for Quantitative and Theoretical Geography (ECTQG07), held for the first time in Switzerland (Montreux). The Colloquium also constituted the annual Workshop of the Faculty of Geosciences and Environment (FGSE), which is situated at the cross-section between natural and social sciences, and focuses on planetary changes of natural or anthropogenic origin, particularly in alpine and metropolitan contexts.

The high-quality and variety of the contributions at the Colloquium reflected the broad range of issues and methods investigated in Europe, as well as the vigor and scope of contemporary research in quantitative and theoretical geography. To deepen the questions raised, several European research teams were asked to write articles presenting and evaluating, for various geographical fields, the progress achieved, the methods used, and the scientific prospects. Six themes were selected, by alphabetical order:

- **Cellular automata, multi-agent systems and cooperative phenomena**
- **Epistemological issues in geography**
- **Geographical flows and networks**
- **Health geography and epidemiology**
- **Resource management and risk analysis**
- **Urban economic geography and spatial economic analysis**

ISBN-13: 978-2-9403-6808-2