Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation

Olle Järv , Henrikki Tenkanen & Tuuli Toivonen

Taylor & Francis
Taylor & Francis Group

Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation

Olle Järv 🅐, Henrikki Tenkanen 🅐 and Tuuli Toivonen 🅐

Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland

ABSTRACT

Novel digital data sources allow us to attain enhanced knowledge about locations and mobilities of people in space and time. Already a fast-growing body of literature demonstrates the applicability and feasibility of mobile phone-based data in social sciences for considering mobile devices as proxies for people. However, the implementation of such data imposes many theoretical and methodological challenges. One major issue is the uneven spatial resolution of mobile phone data due to the spatial configuration of mobile network base stations and its spatial interpolation. To date, different interpolation techniques are applied to transform mobile phone data into other spatial divisions. However, these do not consider the temporality and societal context that shapes the human presence and mobility in space and time. The paper aims, first, to contribute to mobile phone-based research by addressing the need to give more attention to the spatial interpolation of given data, and further by proposing a dasymetric interpolation approach to enhance the spatial accuracy of mobile phone data. Second, it contributes to population modelling research by combining spatial, temporal and volumetric dasymetric mapping and integrating it with mobile phone data. In doing so, the paper presents a generic conceptual framework of a multi-temporal function-based dasymetric (MFD) interpolation method for mobile phone data. Empirical results demonstrate how the proposed interpolation method can improve the spatial accuracy of both night-time and daytime population distributions derived from different mobile phone data sets by taking advantage of ancillary data sources. The proposed interpolation method can be applied for both location- and person-based research, and is a fruitful starting point for improving the spatial interpolation methods for mobile phone data. We share the implementation of our method in GitHub as open access Python code.

1. Introduction

To better understand how societies function and to improve social justice and environmental sustainability, we need to attain enhanced spatio-temporal knowledge about the locations and mobilities of people (Hägerstrand 1970, Sheller and Urry 2006, Cresswell

and Merriman 2011, Kwan 2013). This knowledge has become possible due to the global adoption of mobile information and communication devices in the age of the Big Data, while human interactions through mobile devices create digital footprints of people in space and time (Kitchin 2014). Thus, ubiquitous mobile phones are considered as proxies for people (Asakura and Hato 2004, González *et al*. 2008, Ahas *et al*. 2010).

For example, passively collected time-stamped and location-aware information on mobile phone use, such as call detail records (CDRs), is rapidly enhancing our understanding about individual human mobility (Yuan *et al*. 2012, Järv *et al*. 2014). Such data is applied to understand the relation between human mobility and social networks (Phithakkitnukoon *et al*. 2012), ethnic differences in human activity spaces (Järv *et al*. 2015) and for estimating travel times to public facilities (Wesolowski *et al*. 2015). At the population level, flourishing literature demonstrates how CDR data is used for investigating spatial distributions and temporal dynamics of local and national population (Reades *et al*. 2009, Kang *et al*. 2012, Deville *et al*. 2014), seasonal migration patterns (Silm and Ahas 2010), spatial structures of cities (Louail *et al*. 2014), traffic flow (Wang *et al*. 2012), disease risks (Wesolowski *et al*. 2012) and socio-economic levels of the population (Soto *et al*. 2011, Blumenstock *et al*. 2015).

However, investigation of human presence and mobility based on CDR data imposes theoretical and methodological challenges. For example, one challenging issue is representing the whole population while not everyone is using mobile phones for different reasons (e.g. young children, elderly). Socio-economic biases may exist in a sample while network operators have different segmentation strategies and service costs. Phone users' call activity habits and patterns vary in space and time according to individuals' socio-economic characteristics, preferences, lifestyle, habits and work attributes (Castells *et al*. 2007). Also, authority (legislation) and environmental (physical extent of mobile network) domains may limit where and when mobile phones can be used. These aspects may influence research findings derived from CDR data while creating spatial, temporal and socio-economic biases at both the individual and aggregated levels (Yuan *et al*. 2012, Wesolowski *et al*. 2013, Järv *et al*. 2014, Zhao *et al*. 2016).

One important methodological challenge is the spatial perspective of CDR data since the spatial accuracy of given data depends on the spatial distribution of mobile network base stations, which is not equally distributed in space and changes over time. Furthermore, while more advanced interpolation techniques are applied (e.g. Louail *et al*. 2014, Pei *et al*. 2014), they do not consider the environmental and societal contexts that shape the human presence and mobility in space and time. To obtain more precise spatio-temporal knowledge about human presence and mobility, we need to develop a spatial interpolation technique for CDR data that considers the social structures and human practices.

This paper seeks to contribute to mobile phone-based research in the social sciences by addressing the need to give more attention to the spatial perspective of passively collected mobile phone data (e.g. CDR data) and by overcoming the uneven spatial accuracy problem in base station coverage areas. We suggest applying a dasymetric interpolation approach that is well known in population mapping studies. With this paper we aim: (1) to put forward a generic conceptual framework for a multi-temporal function-based dasymetric (MFD) interpolation method for mobile phone data; and (2) to empirically investigate how and to what extent the proposed method improves a night-time and daytime population distribution derived from mobile phone data. In

doing so, this paper also contributes to dasymetric population modelling by combining spatial, temporal and volumetric mapping techniques while incorporating a function-based population allocation with mobile phone data. To our knowledge, this is the first attempt to comprehensively apply a dasymetric technique for interpolating mobile phone data.

Next, we provide a brief overview on how the spatial perspective is considered in passively collected mobile phone-based research. We then introduce a dasymetric interpolation approach and put forward the conceptual framework of the proposed dasymetric method and demonstrate the empirical results in the case of Tallinn, Estonia. Finally, we discuss the method, obtained outcomes and future steps.

2. Background and literature review

2.1. *The spatial perspective of passively collected mobile phone data*

To a large extent, the spatial accuracy of passively collected mobile phone data, such as CDR data, is determined by the locations of base stations – the geographical location of each mobile phone use is allocated to the base station that provided the network signal. Hence, the spatial accuracy of data corresponds to the coverage areas of a mobile network base station, which is spatially not fixed and varies according to the density of base stations (for some exceptions, see, e.g. Calabrese *et al*. 2013). The simplest way to examine given data from spatial perspective is to use (aggregated) geographic locations of base stations (antennae) as point-based units. Latter approach is widely applied in individual mobility studies (González *et al*. 2008, Yuan *et al*. 2012, Järv *et al*. 2014) and its relation to social phenomena and processes (Phithakkitnukoon *et al*. 2012, Demissie *et al*. 2013, Järv *et al*. 2015, Wesolowski *et al*. 2015). However, this type of data is difficult to integrate and validate with other data sources.

Certainly, CDR data is further assigned to discrete areal polygons using different techniques. Most commonly, a spatial overlay technique is used to attribute given data to predefined spatial (e.g. administrative and statistical) units given the locations of a base station (Ahas *et al*. 2010, Wesolowski *et al*. 2012, Trasarti *et al*. 2015, Williams *et al*. 2015). Conventional choropleth mapping allows for integrating and validating the mobile phone data with other attribute data in a given spatial division. However, choropleth mapping ignores the fact that: (1) the spatial distribution of base stations does not depend on predefined spatial units – an administrative unit may have none or many base stations; and (2) the spatial division of coverage areas and predefined spatial units does not utterly coincide – a base station located at one predefined spatial unit may provide a mobile network signal partly (or entirely) to other predefined units and thus represent CDR data from an area that is larger than where the given base station is actually located.

Another widely applied technique is to interpolate the study area to theoretical coverage areas (polygons) of a base station using a Voronoi tessellation (Delaunay triangulation) technique in Euclidean space (e.g. Wang *et al*. 2012) with some exceptions (see Iqbal *et al*. 2014, Jacobs-Crisioni *et al*. 2014). This way, station-based mobile phone data can be attributed to spatial units indicating base stations' coverage areas. The use of Voronoi tessellation provides more accurate spatial distribution as compared to previous approaches; however, the resulting spatial division is not compatible with

other existing spatial divisions of administration and, hence, does not allow data integration or validation with data in other spatial divisions. Moreover, the two overlay techniques described above depend on the locations of base stations, and any changes in the spatial configuration of a mobile network over time may create biases in both longitudinal and repeated research.

The latter weaknesses can be overcome by applying a straightforward areal weighting spatial interpolation technique; the spatial division of coverage areas and its attributes are simply interpolated by areal intersection to desired spatial units such as statistical grids (Candia *et al*. 2008, Louail *et al*. 2014) or administrative areas (Soto *et al*. 2011, Deville *et al*. 2014). Areal weighting is also applied the other way round; predefined spatial units and its attributes are interpolated into coverage areas of base stations (Kang *et al*. 2012, Doyle *et al*. 2014). Indeed, using statistical grids or administrative areas enables data integration and validation with other data sources in any spatial division. However, this method assumes that mobile phone data as a proxy for people is equally distributed on a plane surface without considering temporality, environmental context or societal structures that shape human presence and mobility in space and time.

Several attempts are made to apply more sophisticated spatial interpolation techniques to enhance the spatial accuracy of mobile phone data within a coverage area of a base station. Ratti *et al*. (2006) applied a neighbour linear interpolation method based on centroids of a Voronoi polygon to represent the probable population in a raster surface. Csáji *et al*. (2013) refined the spatial locations of people within each Voronoi polygon based on the probability density calculation from the maximum likelihood estimation and the signal decay of a base station. Similarly, Pei *et al*. (2014) applied a probabilistic distribution approach to refine the spatial accuracy of people within each Voronoi polygon to grid cells using the inverse distance weighting. Nevertheless, the given sophisticated spatial interpolation techniques still assume space as a plane surface and people in space as a function of the distance from a base station. To the authors' knowledge, only in the supporting information in Deville *et al*. (2014) was there a preliminary attempt made to apply the dasymetric technique by ancillary land-use data to interpolating CDR data.

2.2. Dasymetric interpolation approach and the integration of time

In population research, several spatial interpolation methods are used to go beyond aggregated and predefined census tract and administrative units for finer-scale spatial resolution (Eicher and Brewer 2001, Mennis 2003, Langford 2006). Given the feasibility and reliability of widely used dasymetric mapping, the dasymetric interpolation technique is one of the best approaches for refining the spatial resolution of population data (Wu *et al*. 2005, Mennis and Hultgren 2006). The dasymetric technique for refining population distribution is successfully applied, for example, for studying tourism (Vaz and Campos 2013), accessibility (Langford *et al*. 2008, Shannon 2014), risk exposures and disaster management (Linard and Tatem 2012, Smith *et al*. 2014).

In general, the dasymetric spatial interpolation technique transforms population data from one set of spatial units (source zones) into another spatial division (target zones) using additional ancillary data sources – data that can be related directly or indirectly to

the spatial presence of people for assisting the interpolation. Physical environmental (land use, land cover and zonal data) information is widely used to extract a populated area from a non-populated area and further conduct spatial weighting based on selected attributes (Dobson *et al.* 2000, Ruther *et al.* 2015). Additional ancillary variables considered in dasymetric mapping include the national mailing information (Langford *et al.* 2008), addresses of businesses (Greger 2015), road network proximity and traffic counts (Smith *et al.* 2014), points-of-interests (Bakillah *et al.* 2014), night-imagery (Dobson *et al.* 2000), detailed cadastral maps (Maantay *et al.* 2007) and building functions with vertical and volumetric information (Aubrecht *et al.* 2009, Greger 2015, Biljecki *et al.* 2016).

In recent years, dasymetric interpolation techniques have witnessed fast development given new data sources, improved statistical assessment for estimating accuracy, and advancing multiple areal interpolation with relational data sources, 3D modelling and spatio-temporal interpolation (see, e.g. Nagle *et al.* 2014, Ruther *et al.* 2015, Biljecki *et al.* 2016, Mennis 2016). Interestingly, the concept of the dasymetric interpolation technique is well suited for integrating the temporal perspective (hourly, weekday and seasonal rhythms) for population mapping (e.g. Martin *et al.* 2015, Mennis 2016). Temporally sensitive multidimensional dasymetric interpolation modelling of a population based on census data has already been elaborated upon since 2000 (Dobson *et al.* 2000, Aubrecht *et al.* 2009). Since then, different conceptual frameworks have been proposed for dynamic population distribution modelling while also distinguishing population subgroups, such as local residents, visitors and transit (Bhaduri *et al.* 2007, Aubrecht *et al.* 2014, Martin *et al.* 2015, Mennis 2016). In these studies, the spatial distribution of the population is interpolated as a function of time, while the spatial layer is related to time-dependent ancillary data sources about human presence and activities (e.g. time use survey, activity travel diary).

However, these advanced dasymetric interpolation methods do not consider the verticality of a built environment, and vice versa, spatially advanced building population mapping methods do not consider temporality (Ural *et al.* 2011, Biljecki *et al.* 2016), except a study by Greger (2015). In urbanised societies, the vertical dimension and volumetric data about a built environment are two of the essential environmental attributes that determine the spatial distribution of people (Ahola *et al.* 2007, Biljecki *et al.* 2016).

Thus, we propose a conceptual framework of the MFD interpolation for mobile phone data that also takes into account the verticality of a built environment, similar to Greger (2015). This enables obtaining more accurate spatio-temporal information about human presence and mobility derived from mobile phone data such as CDR, which increases the applicability and reliability of given data.

3. The conceptual framework of an MFD interpolation for mobile phone data

Since mobile phone data as a proxy for people is predominantly estimated by coverage areas of a base station (or antennae) as source zones, we propose a generic MFD interpolation method that first disaggregates the distribution of mobile phone data within each coverage area, and then aggregates mobile phone data to the desired
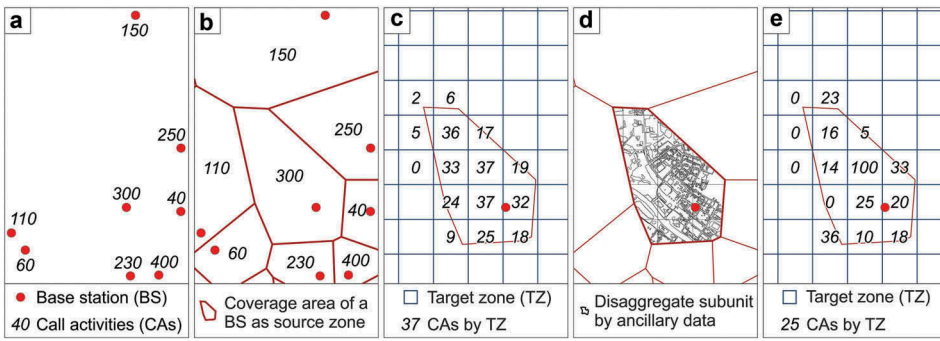
Figure 1. The representation of MFD interpolation method – point-based mobile phone data at base station level (a) – is usually assigned to theoretical coverage areas of a base station as source zones (b) and interpolated to target zones using simple areal weighting method (c). The proposed MFD method disaggregates mobile phone data within each source zone using ancillary data (d) to transform it to desired target zones (e).

spatial division as target zones using different spatially and temporally sensitive ancillary data sources (Figure 1). At minimum, three ancillary data sources are needed – a spatial layer with land-cover data, volume (height) of built environment and time-dependent human activity data.

In general, the MFD interpolation method for mobile phone data has five modelling steps (Figure 2): (I) the preparation of a physical surface layer; (II) the spatial disaggregation of a physical surface layer by both source zones (i.e. coverage areas of a base
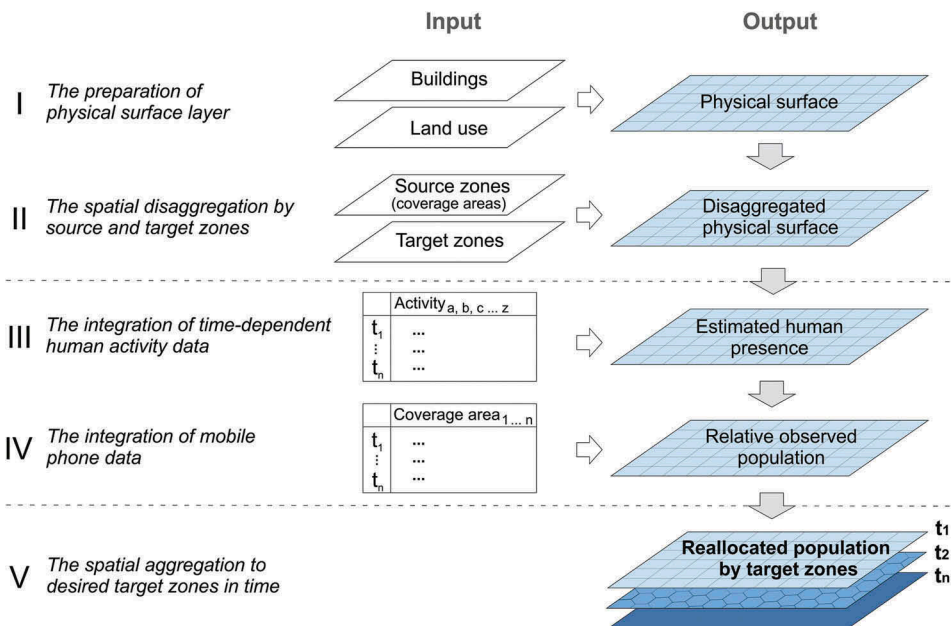

Figure 2. The conceptual framework for the proposed MFD interpolation method for refining mobile phone data as a proxy for people in five modelling steps.

station) and target zone layers; (III) the integration of time-dependent human activity data with a disaggregated physical surface layer; (IV) the integration of mobile phone data with a disaggregated physical surface layer; and (V) the spatial aggregation of disaggregated surface layer to desired target zones as the output of the MFD method. In the MFD method, any desired spatial division can be used as target zones according to any research need – statistical grid cells (as in this case study), census tracts, administrative units, transport analysis zones or any other spatial division. Next, each modelling step is described in the following subsections.

3.1. *The preparation of physical surface layer*

First, the physical surface layer is prepared, where (1) the vertical dimension is incorporated and (2) a functional attribute is assigned, which can be linked to human activity data. The vertical dimension is essential since the verticality of a built environment determines the population density whereas the best indicator is the total floor area of a building (e.g. Biljecki *et al*. 2016). The method is flexible regarding the vertical dimension, and input data can be either ready-made data or estimated (from building footprints and heights) for the buildings' total floor area (m^2) or volume (m^3).

The activity function attribute of each spatial unit enables linking the estimated human presence by activity type in time with the physical surface layer. The method is flexible regarding the detail level of activity function classification as long as it is possible to link with time usage of people by classified human activity data (Eurostat 2009, UNSD 2016) and corresponds to available input data and research needs. We propose to apply six activity function types – residential, work and school, retail and service, transport, restricted area and other areas, since these types have distinguished temporal usage patterns.

The required physical surface layer may be already available in case of detail cadastral data, including the vertical dimension or 3D models on a built environment (Biljecki *et al*. 2016). However, there are several ways to prepare a given layer, for example, by integrating land-use and/or land-cover data with building information; building footprints may be derived from cadastral maps, satellite imagery or volunteered geographic information databases such as OpenStreetMap (OSM) or GoogleMaps. Building height can be derived from areal imagery and remote sensing such as LIDAR measurements (Alahmadi *et al*. 2013). While information about the vertical dimension at building level is not available, more rough estimates may be applied such as the average height of a city block or a neighbourhood.

Building functions related to human activity types in the spatial layer can be further enhanced with additional crowdsourced data such as OSM and online address (business) registers (Greger 2015), and even assign several functionalities to a building if such data is available. Finally, each spatial unit in the physical surface layer includes three attributes: activity function type *a*; surface unit type (building; non-building) *u*; and a vertical dimension (number of floors; height).

3.2. *The spatial disaggregation by source and target zones*

In the second step of the MFD method, two spatial layers are applied in addition to the physical surface layer – the spatial division of mobile phone data as source zones and the spatial division of desired spatial units as target zones (Figure 1). The spatial division of mobile phone data is usually represented by the theoretical coverage areas of a base station using a Voronoi tessellation method. The desired spatial division of target zones is applied according to a research need.

First, a geometric union technique is applied for these three spatial layers to disaggregate the physical surface layer into subunits in order to designate each subunit to a unique source zone *j* and target zone *z*. Second, the area for each disaggregated subunit polygon is calculated, which enables calculating an estimated total floor area *FA* for each spatial subunit *s* by multiplying the area with the vertical dimension regarding the number of floors or height, whereas in both cases the default value is set to 1 for non-building units. This allows for a calculation of a relative floor area *RFA* for each subunit *s* within a base station *j* given the total sum of estimated total floor area *FA* of all subunits within a base station *j* as follows:

$$RFA = \frac{FA_s}{\sum FA_s \in j} \qquad (1)$$

At this point, each disaggregated subunit in the physical surface layer has three additional attributes: source zone ID, target zone ID and a relative floor area *RFA* within a coverage area of a base station to which a given unit belongs.

3.3. The integration of time-dependent human activity data

In the third step of the MFD method, the disaggregated physical surface layer is linked with time-dependent human activity data based on the activity function type of a spatial subunit. The link allows for a more accurate interpolation of mobile phone data between disaggregated spatial subunits within each base station coverage area, while taking into account the probabilities of human activity types.

In general, the probability of people to conduct certain activities at certain spatial subunits at a certain time can be acquired from a time-use survey (Eurostat 2009, UNSD 2016) or activity travel diary data (see Schönfelder and Axhausen 2010). Given data sources allow for an extraction of the average distribution of people by activity type in time that can be associated with spatial subunits based on the activity location type. Any classification of human activity types can be applied, as long as the activity classification is congruent with activity function types of a spatial subunit, such as associating home activities with residential areas. Ideally, the data source is from the case study area. However, the activity classification can represent a more general (e.g. national) level or even from a neighbouring country or elsewhere, as long as the classification represents human activity patterns similar enough for the case study area given the similar social and environmental contexts.

The temporal resolution of human activity data in the MFD method depends on available data sources and research needs. Thus, the basis of the method is the distribution of people by activity type at given time intervals, for example, hourly

intervals. The estimated distribution of people by activity type (e.g. being at home) in a given time unit can be further refined if input data contains more fine-grained information about weekly (working day vs. weekend) and seasonal (e.g. summer vs. winter; holiday weeks and months) variations in mobility and activity patterns of people.

In this modelling step, an estimated human presence *EHP* in time unit *t* is calculated for each disaggregated spatial subunit *s* within a base station *j* given its *RFA* and time-dependent probability for human presence, which is a combination of a daily hour factor *H*, a weekday factor *W* and a seasonal (monthly) factor *M*, according to a human activity function type *a* and spatial unit type *u* (land or building). The total sum of *EHP* of disaggregated spatial subunits within each coverage area of a base station *j* in time unit *t* is 1. *EHP* is calculated as follows:

$$EHP_t = (H_{au}W_{au}M_{au}) \times RFA_{sj} \tag{2}$$

MFD method can be used to calculate *EHP* for the whole population, whereas it can also be calculated separately for each subgroup of a population (e.g. youth, working-age people, and elderly; local vs. non-locals) if the given division is available for both human activity data and mobile phone data. Another possibility is to calculate *EHP* separately for local residents and non-residents (e.g. transit, visitors and tourists). This would provide a more accurate refinement of population distribution within a base station.

It is possible to derive the hourly distribution of people by activity type from time-use survey data. Activity travel diaries can provide weekday and seasonal differences in the probable human presence by activity type (Schönfelder and Axhausen 2010). Seasonal differences can also be derived from longitudinal time-activity studies where personal time use in indoors and outdoors (e.g. Hussein *et al*. 2012) can be linked to spatial unit type (building, land) as an indication for seasonal influence. This combines hourly, weekday and seasonal factors for calculating the distribution of people by activity type as a function of time.

3.4. *The integration of mobile phone data*

In the fourth step of the MFD method, mobile phone data is linked to spatial subunits in the physical spatial layer by base station *j* (source zones) and disaggregated into each spatial subunits according to its *EHP* by given time unit *t* (e.g. hourly intervals as in this study). While mobile phone data is considered as a proxy for population mapping, it is recommended to normalise mobile phone data to represent the relative population distribution within the study area. Thus, the total sum of the relative share of mobile phone data *RMP* in a study area *S* in time unit *t* is always 1. *RMP* for each base station *j* is calculated from all mobile phone data *MP* conducted within a study area *S* in given time unit *t* as follows:

$$RMP = \frac{MP_{tj}}{\sum MP_{tj} \in S} \tag{3}$$

Next, normalised mobile phone data of each base station *j* in time unit *t* is interpolated into disaggregated spatial subunits *s* within a coverage area of a given base station *j* regarding *EHP* coefficients of each spatial subunit *s*. A relative observed population *ROP* for each disaggregated subunit *s* for given time unit *t* is calculated based on an

estimated human presence EHP_t and a relative share of mobile phone data RMP_t in the given time unit t as follows:

$$ROP_t = EHP_t \times RMP_t \qquad (4)$$

3.5. *The spatial aggregation to desired target zones*

In the final step, disaggregated spatial subunits in the physical surface layer are dissolved by desired target zones z (each subunit is already assigned to a unique target zone in the second step of the method). Thus, the share of relative observed population ROP_t of each subunit s within each target zone z in time t is summarised to $ZROP$ as follows:

$$ZROP_t = \sum_{s \in z} ROP_t \qquad (5)$$

As an outcome, the distribution of a relative observed population derived from mobile phone data, which was originally assigned to the coverage areas of a base station, is now more realistically transformed to the desired spatial division of target zones. In the next section, an empirical implementation of the proposed MFD method is described.

4. Implementation of the proposed MFD method

We selected Tallinn, the capital city of Estonia, as our case study area to empirically test the proposed MFD method given the availability of mobile phone data. Tallinn is a medium-sized European city and the biggest city in Estonia, located within a total area of 145 km^2 and with over 400,000 inhabitants (Figure 3). Based on the conceptual framework of the proposed MFD interpolation method for mobile phone data, we developed an open access tool for the Python programming language, which is freely available on GitHub (http://doi.org/10.5281/zenodo.252612).

We investigated mobile phone data from a 1-month study period in March 2015, linked to the 290 base stations that provide the coverage area for a mobile network signal for the case study area. Three different random mobile phone data sets at the base station level are applied: (1) raw CDR data from working days (Monday–Friday) at night (2 AM–6 AM); (2) raw CDR data during working days at daytime (4 PM–5 PM); and (3) the most probable home locations of mobile phone users derived from CDR data using the anchor point method (Ahas *et al.* 2010). Comprehensive details about applied mobile phone data sets are provided in Section S1 in the Supplemental material.

In addition to mobile phone data (source zones), we applied five different ancillary data sets derived from six data sources (Table 1) for implementing the proposed MFD method for mobile phone data for Tallinn, Estonia. We applied two target zone layers to investigate the method for 500 m (0.25 km^2, $n = 664$) and 100 m (0.01 km^2, $n = 16,280$) statistical grid cell layers from Statistics Estonia.

Since ready-made data is not available, we prepared the physical surface layer using four ancillary data sets: (1) land-cover data to classify land-cover parcels by six activity
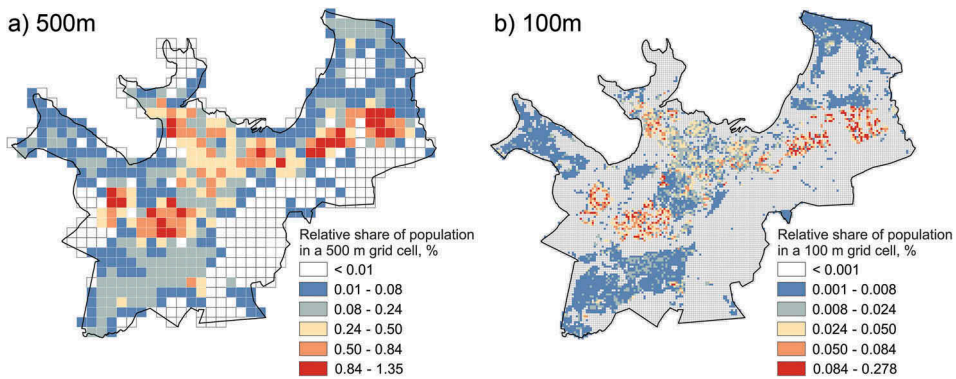
Figure 3. The population registry data representing a night-time population distribution in Tallinn for 500 m (a) and 100 m spatial resolution (b) in 2015.

Table 1. The list of ancillary data sets used in the case study.

Data set	Attributes	Source	Date
Land-cover data	Land-cover parcels with classification (polygon)	The Estonian National Topographic Database (ENTD), the Land Board of Estonia (LBE)	2015
Building data	1. Building footprints (polygon) with binary usage classification (residential or public; other)	The ENTD, the LBE	2015
	2. Building heights based on LIDAR measurements with 1 m accuracy (raster)	The nDSM data, the LBE	2013
	3. Building functionality based on crowdsourcing input on building usage (point) for refining building usage classification	The OpenStreetMap Database (OSM)	2015
Human activity data	The average hourly distribution of people by activity type based on cross-sectional time-use survey study	The Estonian Time Use Survey, Statistics Estonia	2010
Population register data	Residential population data at building level, which is aggregated to both target zones (100 m and 500 m grid cells)	The Population registry data, Tallinn municipality	2015

function types (residential, work, retail and service, other, transport, restricted); (2) building footprint data to integrate buildings with the main usage function (residential and public, non-residential); (3) normalised Digital Surface Model (nDSM) data from LIDAR to calculate building heights for estimating the number of floors; and (4) the Open Street Map database to provide additional information to extract buildings with public, service and retail as the main functions. An overlay method was applied to create a physical surface layer where each disaggregated spatial subunit in the physical surface layer has three attributes: the number of floors, activity function type and spatial unit type of either building or land.

In the second step, a geometric union between the physical surface layer, the Voronoi polygon layer representing theoretical coverage areas and official statistical grid layer for target zones is applied. We received two different disaggregated physical surface layers since we are investigating the results of MFD method at two target zone levels. In the third step, we applied a national time-use survey data and calculated the average hourly distribution of people by activity type to estimate the human presence for spatial units

in the physical surface layer. We also applied a seasonal factor to refine the estimated human presence for each disaggregated spatial subunit. In the fourth step, we applied all three mobile phone data sets and calculated the relative observed population according to MFD. Finally, we created six spatial output layers in the two target zones for three data sets. A detailed description of the implementation of the MFD method in Tallinn is provided in Section S2 in Supplemental material.

To compare the outcome of the MFD method, we created another six spatial output layers where population distribution derived from mobile phone data is interpolated to target zones using the simple areal weighting (AW) interpolation method. We applied population register data at the building level that is aggregated to both target zone layers as the best available baseline to empirically demonstrate how the proposed MFD method improves a night-time (sleeping) population distribution compared to the AW method. To evaluate the accuracy and validity of the proposed MFD method against AW method for refining population distribution derived from mobile phone data, we performed three statistical analysis: (1) linear regression for comparing correlation coefficients and standard error of the estimate (Maantay *et al*. 2007); (2) the mean absolute error; and (3) the coefficient of variation (CV) based on root mean square error of each target zone normalised by the baseline target zone value (Mennis 2016).

5. Results

5.1. *Population distribution by activity function type*

In general, the proposed MFD interpolation method refines a night-time population distribution derived from mobile phone data by activity function type significantly better than areal weighting (AW) when compared to reference data (Figure 4). For the case of a night-time population distribution, some 94% of people are at home according to a national time-use survey. In comparison, the proposed MFD method relocates 86%
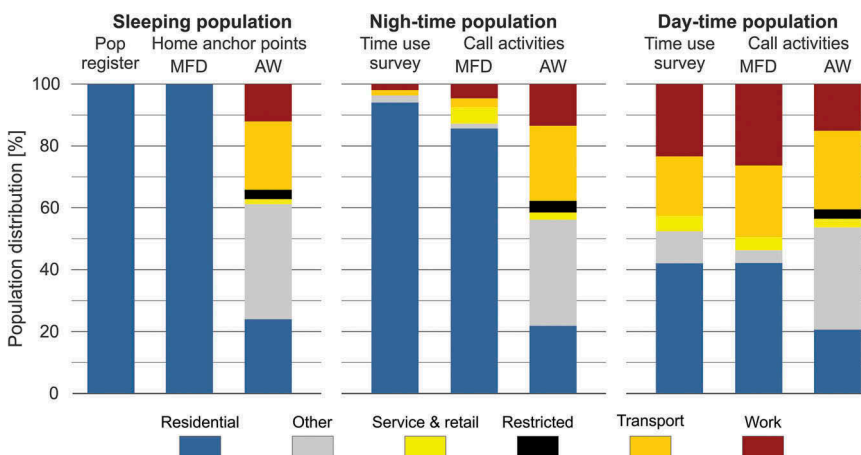

Figure 4. The population distribution by activity function type by MFD and AW interpolation methods regarding three mobile phone data sets in comparison to population register and national time-use survey data.

of a night-time population derived from night-time call activities to residential areas, whereas with AW the given share is only 22%. In the case of a sleeping population distribution at night, we can consider that 100% of population register data indicates a sleeping population in residential areas. In comparison, we applied home anchor points of mobile phone users derived from call activities as an indicator for the sleeping population. The MFD method relocates 100% of home anchor points to residential areas, whereas the given share is only 24% in the case of the AW method.

During the day (as an example of 4 PM to 5 PM), population distribution by activity function type between the two interpolation approaches is less obvious, although population division with the MFD method coincides more with a national time-use survey.

5.2. *Night-time population distribution in target zones*

A visual comparison between two interpolation methods in interpolating the spatial distribution of a night-time population derived from home anchor points (Figure 5) and call activities conducted between 2 AM and 6 AM (Figure S3 in Supplemental material) reveals distinct spatial differences regarding both target zones; the MFD method
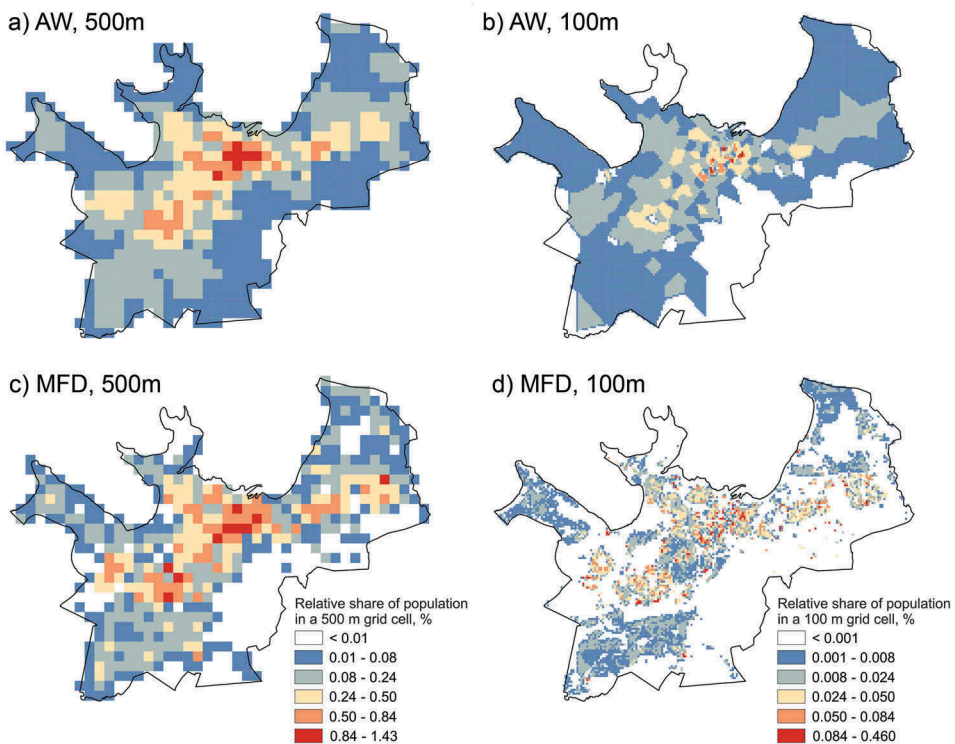

Figure 5. The distribution of home anchor points derived from call activities as an indication for a night-time population by a simple AW and the proposed MFD interpolation methods with both target zone layers.

relocates the population to residential areas compared to the AW method, which distributes people equally in space.

The comparison of population register data as a baseline distribution for a night-time population distribution and population distribution in the case of home anchor points (Figure 6) and call activities (Figure S4 in Supplemental material) demonstrates the differences with the baseline distribution between two interpolation approaches, especially in case of 100 m grid cells. In general, the population distribution using the MFD method has fewer differences with population register data, whereas the influence of the method is more significant in areas where larger coverage areas of a base station occur.

Interestingly, the given difference maps reveal an inherent weakness of mobile phone data. There is a tendency to systematically overestimate the population in the city centre (blue colour) where human communication is significantly more active. However, the aim of the proposed MFD method is not to resolve the given bias, but to refine mobile phone data (e.g. population) within the coverage areas of a base station. Thus, the proposed method relocates the population into more populated target zones and increases the number of target zones with no population as compared to a simple AW method (Figure S5 and Figure S6 in Supplemental material).
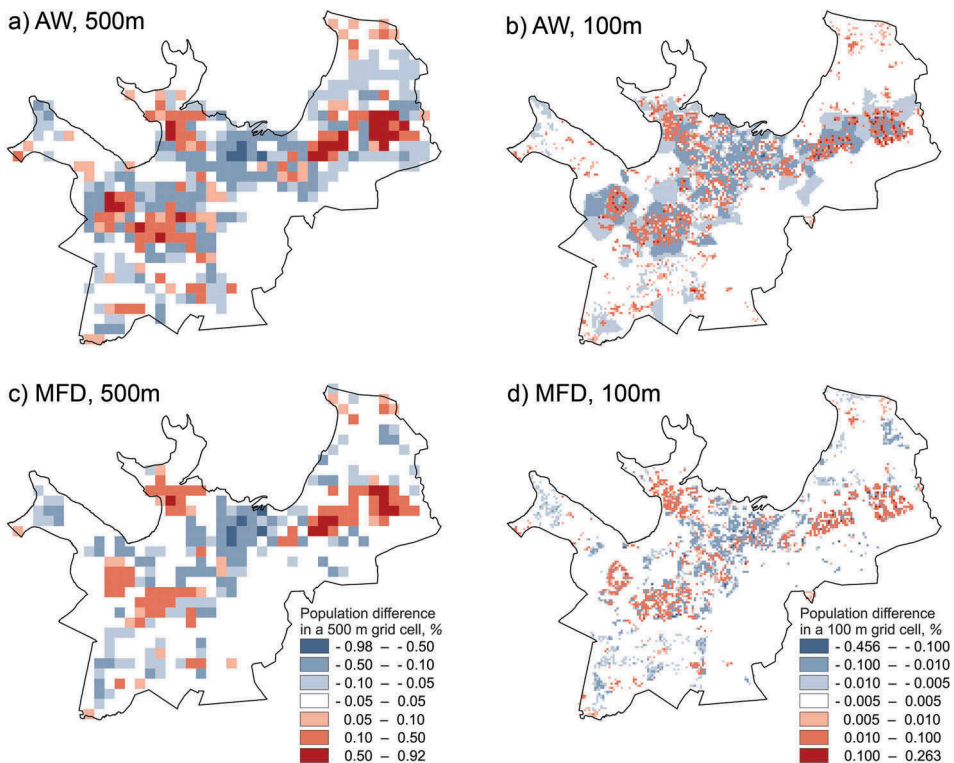

Figure 6. Difference (percentage points) in a night-time population distribution between population register data as the baseline distribution and home anchor points by MFD and AW interpolation methods with both target zone layers. Target zones in red colour indicate underestimation and blue colour overestimation of population, indicated by home anchor points.

Table 2. Summary of evaluating the results of both interpolation methods against population register data regarding two target zone levels and two mobile phone data sets to indicate a night-time population.

Target zone level	Evaluation method		Call activities 2 AM to 6 AM		Home anchor points	
			MFD	AW	MFD	AW
500 m grid cells (n = 664)	Linear regression	Correlation coefficient	0.674	0.511	0.832	0.686
		Std. Error	0.189	0.220	0.142	0.186
	The mean absolute error (MAE)		0.087	0.122	0.072	0.107
	Coefficient of variation (CV) based on the RMSE		0.211	0.242	0.142	0.186
100 m grid cells (n = 16280)	Linear regression	Correlation coefficient	0.472	0.233	0.618	0.347
		Std. Error	0.016	0.018	0.014	0.017
	The mean absolute error (MAE)		0.005	0.008	0.004	0.008
	Coefficient of variation (CV) based on the RMSE		0.019	0.019	0.015	0.017

5.3. *Evaluating MFD method for interpolating population distribution*

Table 2 summarises the statistical comparison of the MFD method and simple AW method for refining mobile phone-based population distribution in comparison to population register data as a baseline distribution for a night-time population. Regardless of the evaluation measure, the MFD method outperforms the AW method; the former has higher correlation coefficients with a lower standard error (Figure S7 and Figure S8 in Supplemental material), lower mean absolute error and coefficient of variation of RMSE, especially for the 500 m × 500 m target zone resolution. The MFD method improves population distribution more for home anchor points since it corresponds better with population register data, as both indicate a sleeping population. The finding supports the visual evidence presented in the previous subsections suggesting that incorporating spatio-temporally dependent ancillary data in interpolating mobile phone data by coverage areas of base station clearly improves the refinement of human presence to target zones as compared to a simple areal weighting interpolation method.

5.4. *Daytime population distribution in target zones*

A visual comparison between two interpolation methods in interpolating the spatial distribution of a daytime population derived from call activities conducted between 4 PM and 5 PM (Figure 7) reveals fewer distinct spatial differences regarding both target zones. During a given time period, call activities are clearly concentrated to the city centre, where a dense base station network already exists, and the proposed MFD method has a less significant effect on refining population distribution as compared to a night-time population.

Regardless of target zone resolution, MFD method has a less significant effect in relocating a daytime population to target zones within the study area (Figure 8) given the strong correlation in population distribution between two interpolation approaches (Figure S9 in Supplemental material). On a smaller scale, however, some significant population relocations occur at grid level. In the case of 500 m target zones, for example, in almost 50% of all target zones, the MFD method relocates ±0.01–0.05% points of the relative observed population (Figure S10 in Supplemental material). This indicates ±40–200 relocated people
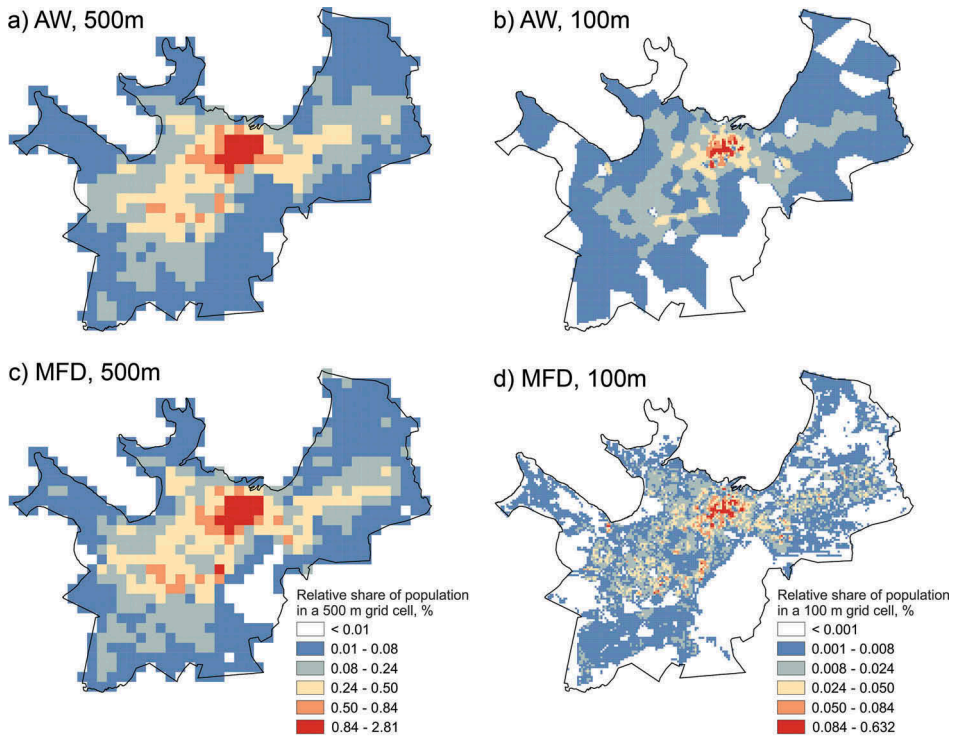
Figure 7. The distribution of call activities (4 PM to 5 PM) as an indication for a daytime population in given time period by simple AW and the proposed MFD interpolation methods with both target zone layers.
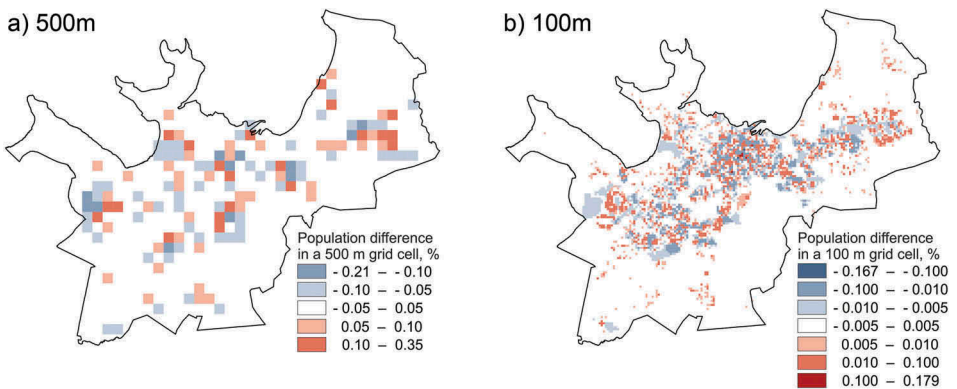

Figure 8. Difference (percentage points) in the distribution of call activities (4 PM – 5 PM) as an indication for a daytime population between the proposed MFD and simple AW interpolation methods with both target zone layers. Target zones in red colour indicate higher population distribution with the proposed MFD interpolation method as compared to simple AW. Grid cells with blue colour indicate the opposite.

in a target zone as compared to the population estimate by the AW method. Furthermore, almost 10% of all target zones have more than ±200 relocated people in a target zone with the MFD method.

6. Discussion and conclusions

The implementation of passively collected mobile phone data such as CDR data to investigate human presence and mobility is mushrooming in social sciences. However, the given data has theoretical and methodological challenges that one needs to acknowledge. This paper gives attention to one of the important challenges that has not been widely discussed to date – the uneven spatial resolution of CDR data due to the uneven spatial configuration of mobile network base stations, and its spatial interpolation. We introduced a dasymetric interpolation approach as one promising way to overcome this. Thus, we proposed a generic conceptual framework of the multi-temporal function-based dasymetric (MFD) interpolation to enhance the spatial resolution of mobile phone data, while taking into account the socio-spatial structure (land use, building volume) and time-dependent human activity data.

The empirical findings demonstrate that the applied MFD interpolation method transforms CDR data as an indication of population distribution to desired target zones (statistical grid cells) more accurately than a simple areal weighting method. The finding was confirmed by evaluating the results of both interpolation methods against population register data as a baseline for a night-time population. In general, the results were consistent with call activity and home anchor point data sets and two different target zone levels with 100 and 500 m resolution. The MFD method had less of an effect for a daytime population interpolation compared to the areal weighting method at the city level; however, it refines the population between target zones significantly on a smaller scale, which is essential, for example, in planning and risk assessment at the micro level.

As the results clearly show, the proposed interpolation method refines the population distribution estimates within a base station coverage area. Yet, it does not resolve the inherent bias of mobile phone data – the fact that it is influenced by the variance of human interactions with mobile devices in space and time, whereas one spatial outcome is the tendency to overestimate the city centre, given its higher probability for human interaction. However, the proposed MFD method can reduce the given biases to some extent when introducing population subgroups to the method, which is similar to Martin *et al.* (2015).

In general, the method can be improved by providing more accurate ancillary data about the socio-spatial structure and human activities or by incorporating additional data sources. For instance, traffic count data refines an estimated human presence for different sections of a road network. Also, more accurate knowledge about coverage areas of a base station would improve the spatial interpolation of CDR data since commonly (also in this case study), a straightforward Voronoi tessellation for theoretical coverage areas of a base station is applied, which does not coincide with the actual coverage areas. Certainly, more detailed research about the sensitive scale issue of *EHP* at different spatial resolutions and the comprehensive validation of accuracy estimates of the method need to be conducted in future.

We did not demonstrate how the proposed MFD method can be applied to examine societal issues. However, the method is generally applicable in the broad field of social sciences since the method: (1) can transform mobile phone data to any desired spatial division of target zones (e.g. grids and census tracts) given the research need; (2) enables

reliable data comparison, integration and validation of mobile phone data with other data sources; (3) enables reliable mobile phone data interpolation for longitudinal and repeated research; and (4) can improve a systematic estimate of population distribution in time and space from a small scale (e.g. neighbourhood level) to a country level (e.g. global level), depending on mobile phone and ancillary data availability. Thus, the method can provide dynamic population distribution mapping to assess object-oriented risks (Smith *et al*. 2014) or to model more realistic spatio-temporal accessibility to services (Tenkanen *et al*. 2016).

We only applied the proposed interpolation method for population modelling at an aggregated level; however, it can also be used for person-based research for refining personal mobility patterns and activity spaces at the individual level. The method can be developed to refine the probable location of an individual within a base station depending on the individual's spatio-temporal activities derived from one's mobile phone usage. For example, individual meaningful locations (e.g. home and work) derived from one's CDR data at the base station level (Ahas *et al*. 2010) allow the model to calculate personal probabilities for a given individual to be associated with a certain type of spatial subunit within certain coverage areas in a given time. Certainly, one must acknowledge and preserve the privacy of phone users regarding the use of mobile phone data (see, e.g. Yin *et al*. 2015).

We conclude that the proposed multi-temporal, function-based dasymetric interpolation method offers a promising approach to increase the usability and reliability of passively collected mobile phone data. Our empirical findings provide a solid base to improve the spatial interpolation methods for mobile phone-based research. Furthermore, this study contributes to developing dasymetric population modelling by incorporating mobile phone data. To facilitate the future development of the proposed approach, we share the implementation of our method in GitHub as open access Python code.

ORCID

Olle Järv 🄳 http://orcid.org/0000-0003-3446-1545
Henrikki Tenkanen 🄳 http://orcid.org/0000-0002-0918-4710
Tuuli Toivonen 🄳 http://orcid.org/0000-002-6625-4922

References

Ahas, R., *et al.*, 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17 (1), 3–27. doi:10.1080/10630731003597306

Ahola, T., *et al.*, 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science*, 21 (8), 935–953. doi:10.1080/13658810701349078

Alahmadi, M., Atkinson, P., and Martin, D., 2013. Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41, 167–176. doi:10.1016/j.compenvurbsys.2013.06.002

Asakura, Y. and Hato, E., 2004. Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, 12 (3–4), 273–291. doi:10.1016/j.trc.2004.07.010

Aubrecht, C., *et al.*, 2009. Integrating earth observation and GIScience for high resolution spatial and functional modeling of urban land use. *Computers, Environment and Urban Systems*, 33 (1), 15–25. doi:10.1016/j.compenvurbsys.2008.09.007

Aubrecht, C., Steinnocher, K., and Huber, H., 2014. DynaPop–population distribution dynamics as basis for social impact evaluation in crisis management. *In*: *Proceedings of the 11th International ISCRAM Conference*. University Park, PA: The Pennsylvania State University, 314–318.

Bakillah, M., *et al.*, 2014. Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, 28 (9), 1940–1963. doi:10.1080/13658816.2014.909045

Bhaduri, B., *et al.*, 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69 (1–2), 103–117. doi:10.1007/s10708-007-9105-9

Biljecki, F., *et al.*, 2016. Population estimation using a 3D city model: a multi-scale country-wide study in the Netherlands. *PLoS One*, 11 (6), e0156808. doi:10.1371/journal.pone.0156808

Blumenstock, J., Cadamuro, G., and On, R., 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350 (6264), 1073–1076. doi:10.1126/science.aac4420

Calabrese, F., *et al.*, 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26, 301–313. doi:10.1016/j.trc.2012.09.009

Candia, J., *et al.*, 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41 (22), 224015. doi:10.1088/1751-8113/41/22/224015

Castells, M., *et al.*, 2007. *Mobile communication and society: a global perspective: a project of the annenberg research network on international communication*. Cambridge, MA: MIT Press.

Cresswell, T. and Merriman, P., 2011. *Geographies of mobilities: practices, spaces, subjects*. Farnham: Ashgate.

Csáji, B.C., *et al.*, 2013. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and Its Applications*, 392 (6), 1459–1473. doi:10.1016/j.physa.2012.11.040

Demissie, M.G., de Almeida Correia, G.H., and Bento, C., 2013. Exploring cellular network handover information for urban mobility analysis. *Journal of Transport Geography*, 31 (0), 164–170. doi:10.1016/j.jtrangeo.2013.06.016

Deville, P., *et al.*, 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111 (45), 15888–15893. doi:10.1073/pnas.1408439111

Dobson, J.E., *et al*., 2000. LandScan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, 66 (7), 849–857.

Doyle, J., *et al*., 2014. Population mobility dynamics estimated from mobile telephony data. *Journal of Urban Technology*, 21 (2), 109–132. doi:10.1080/10630732.2014.888904

Eicher, C.L. and Brewer, C.A., 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science*, 28 (2), 125–138. doi:10.1559/152304001782173727

Eurostat, 2009. *Harmonised European time use surveys: 2008 guidelines*. Luxembourg: European Communities.

González, M.C., Hidalgo, C.A., and Barabási, A.-L., 2008. Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782. doi:10.1038/nature06958

Greger, K., 2015. Spatio-temporal building population estimation for highly urbanized areas using GIS: spatio-temporal building population estimation. *Transactions in GIS*, 19 (1), 129–150. doi:10.1111/tgis.2015.19.issue-1

Hägerstrand, T., 1970. What about people in regional science? *Papers of the Regional Science Association*, 24 (1), 6–21. doi:10.1007/BF01936872

Hussein, T., Paasonen, P., and Kulmala, M., 2012. Activity pattern of a selected group of school occupants and their family members in Helsinki — Finland. *Science of the Total Environment*, 425, 289–292. doi:10.1016/j.scitotenv.2012.03.002

Iqbal, M.S., *et al*., 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40, 63–74. doi:10.1016/j.trc.2014.01.002

Jacobs-Crisioni, C., *et al*., 2014. Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: an exploratory study using mobile phone data. *Environment and Planning A*, 46 (11), 2769–2785. doi:10.1068/a130309p

Järv, O., Ahas, R., and Witlox, F., 2014. Understanding monthly variability in human activity spaces: a twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38 (0), 122–135. doi:10.1016/j.trc.2013.11.003

Järv, O., *et al*., 2015. Ethnic differences in activity spaces as a characteristic of segregation: a study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*, 52 (14), 2680–2698. doi:10.1177/0042098014550459

Kang, C., *et al*., 2012. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19 (4), 3–21. doi:10.1080/10630732.2012.715479

Kitchin, R., 2014. *The data revolution: big data, open data, data infrastructures and their consequences*. London: Sage Publications Ltd. doi:10.4135/9781473909472

Kwan, M.-P., 2013. Beyond space (As We Knew It): toward temporally integrated geographies of segregation, health, and accessibility. *Annals of the Association of American Geographers*, 103 (5), 1078–1086. doi:10.1080/00045608.2013.792177

Langford, M., 2006. Obtaining population estimates in non-census reporting zones: an evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems*, 30 (2), 161–180. doi:10.1016/j.compenvurbsys.2004.07.001

Langford, M., *et al*., 2008. Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems*, 32 (1), 66–80. doi:10.1016/j.compenvurbsys.2007.06.001

Linard, C. and Tatem, A.J., 2012. Large-scale spatial population databases in infectious disease research. *International Journal of Health Geographics*, 11 (1), 1. doi:10.1186/1476-072X-11-7

Louail, T., *et al*., 2014. From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4, 5276. doi:10.1038/srep05276

Maantay, J.A., Maroko, A.R., and Herrmann, C., 2007. Mapping population distribution in the urban environment: the Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science*, 34 (2), 77–102. doi:10.1559/152304007781002190

Martin, D., Cockings, S., and Leung, S., 2015. Developing a flexible framework for spatiotemporal population modeling. *Annals of the Association of American Geographers*, 105 (4), 754–772. doi:10.1080/00045608.2015.1022089

Mennis, J., 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55 (1), 31–42.

Mennis, J., 2016. Dasymetric spatiotemporal interpolation. *The Professional Geographer*, 68 (1), 92–102. doi:10.1080/00330124.2015.1033669

Mennis, J. and Hultgren, T., 2006. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, 33 (3), 179–194. doi:10.1559/152304006779077309

Nagle, N.N., et al., 2014. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104 (1), 80–95. doi:10.1080/00045608.2013.843439

Pei, T., et al., 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28 (9), 1988–2007. doi:10.1080/13658816.2014.913794

Phithakkitnukoon, S., Smoreda, Z., and Olivier, P., 2012. Socio-geography of human mobility: a study using longitudinal mobile phone data. *PLoS ONE*, 7 (6), e39253. doi:10.1371/journal.pone.0039253

Ratti, C., et al., 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33 (5), 727–748. doi:10.1068/b32047

Reades, J., Calabrese, F., and Ratti, C., 2009. Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36 (5), 824–836. doi:10.1068/b34133t

Ruther, M., Leyk, S., and Buttenfield, B.P., 2015. Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. *Giscience & Remote Sensing*, 52 (2), 158–178. doi:10.1080/15481603.2015.1018856

Schönfelder, S. and Axhausen, K.W., 2010. *Urban rhythms and travel behaviour : spatial and temporal phenomena of daily travel*. Farnham: Ashgate.

Shannon, J., 2014. What does SNAP benefit usage tell us about food access in low-income neighborhoods? *Social Science & Medicine*, 107, 89–99. doi:10.1016/j.socscimed.2014.02.021

Sheller, M. and Urry, J., 2006. The new mobilities paradigm. *Environment and Planning A*, 38 (2), 207–226. doi:10.1068/a37268

Silm, S. and Ahas, R., 2010. The seasonal variability of population in Estonian municipalities. *Environment and Planning A*, 42 (10), 2527–2546. doi:10.1068/a43139

Smith, A., Martin, D., and Cockings, S., 2014. Spatio-temporal population modelling for enhanced assessment of urban exposure to flood risk. *Applied Spatial Analysis and Policy*, 9 (2), 145–163. doi:10.1007/s12061-014-9110-6

Soto, V., et al., 2011. Prediction of socioeconomic levels using cell phone records. *In*: J.A. Konstan, et al. eds. *User Modeling, Adaption and Personalization*. UMAP 2011. Lecture Notes in Computer Science, vol 6787. Berlin: Springer, 377–388.

Tenkanen, H., et al., 2016. Health research needs more comprehensive accessibility measures: integrating time and transport modes from open data. *International Journal of Health Geographics*, 15 (1), 1–12. doi:10.1186/s12942-016-0052-x

Trasarti, R., et al., 2015. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*, 39 (3–4), 347–362. doi:10.1016/j.telpol.2013.12.002

UNSD, 2016. United Nations statistics division time use data portal [online]. Available from: unstats.un.org/unsd/gender/timeuse/index.html [Accessed 10 Oct 2016].

Ural, S., Hussain, E., and Shan, J., 2011. Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation*, 13 (6), 841–852. doi:10.1016/j.jag.2011.06.004

Vaz, E. and Campos, A.C., 2013. A multi-dasymetric mapping approach for tourism. *Journal of Spatial and Organizational Dynamics*, 1 (4), 265–277.

Wang, P., et al., 2012. Understanding road usage patterns in urban areas. *Scientific Reports*, 2, 1001.

Wesolowski, A., et al., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface*, 10 (81), 20120986. doi:10.1098/rsif.2012.0986

Wesolowski, A., *et al*., 2012. Quantifying the impact of human mobility on malaria. *Science*, 338 (6104), 267–270. doi:10.1126/science.1223467

Wesolowski, A., *et al*., 2015. Quantifying the impact of accessibility on preventive healthcare in sub-Saharan Africa using mobile phone data. *Epidemiology*, 26 (2), 223–228. doi:10.1097/EDE.0000000000000239

Williams, N.E., *et al*., 2015. Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS ONE*, 10 (7), e0133630. doi:10.1371/journal.pone.0133630

Wu, S., Qiu, X., and Wang, L., 2005. Population estimation methods in GIS and remote sensing: a review. *GIScience & Remote Sensing*, 42 (1), 80–96. doi:10.2747/1548-1603.42.1.80

Yin, L., *et al*., 2015. Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PLoS One*, 10 (10), e0140589. doi:10.1371/journal.pone.0140589

Yuan, Y., Raubal, M., and Liu, Y., 2012. Correlating mobile phone usage and travel behavior – a case study of Harbin, China. *Special Issue: Geoinformatics 2010*, 36 (2), 118–130.

Zhao, Z., *et al*., 2016. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30 (9), 1738–1762. doi:10.1080/13658816.2015.1137298

Supplemental Information for

Enhancing spatial accuracy of mobile phone data

using multi-temporal dasymetric interpolation

CONTENTS

S1. Description of Mobile Phone Data

In general, mobile devices are widely used in Estonia. Some 95% of Estonia's 1.3 million inhabitants used mobile phones in 2008, and the mobile phone subscribers (SIM cards) penetration rate was over 150% in 2014.

The call detail record (CDR) data used in this study is collected by the largest mobile network operator in Estonia with a market share of 40–50% of mobile subscribers in Estonia. CDR data is further processed by a spin-off company Positium LBS, who provided the datasets for this study. The data is recorded in accordance with Estonian legislation for billing purposes by the operator, and not for the purposes of this study. This study is in accordance with European Union legislation on the use of personal data (European Commission 2002). The database contains records of all outgoing call activities initiated by a mobile phone owner. Each CDR includes the unique identification (ID) number of the phone (randomly generated by the operator for every mobile phone), the exact time (HH:MM:SS) and date (YYYY.MM.DD) of call activity, and geographical coordinates of the mobile network antenna that provided the network signal. The precision of the spatial accuracy of the CDR corresponds to the coverage area of the network antenna. The precision of the temporal accuracy of the CDR depends on call activity of people.

We study only mobile phone data that is linked to those 290 base stations that provide the coverage area of mobile network signal for the city of Tallinn (Järv *et al.* 2014; p. 127). Since the actual coverage areas of a base station are unknown, Voronoi tessellation is applied to calculate theoretical coverage areas (polygons) for each base station. The average spatial extent of given theoretical coverage areas is 1.1 km^2, however, coverage areas are not spatially fixed (Figure S1). The spatial extent of a theoretical coverage area at a minimum is 0.01 km^2 and maximum is 48.4 km^2, whereas a lower quartile value is 0.13 km^2, a median 0.36 km^2 and an upper quartile 0.88 km^2. The spatial extent of coverage areas is directly related to demand of mobile phone usage. The smallest coverage areas are located in densely populated city (sub)centres and larger in more sparsely populated areas.
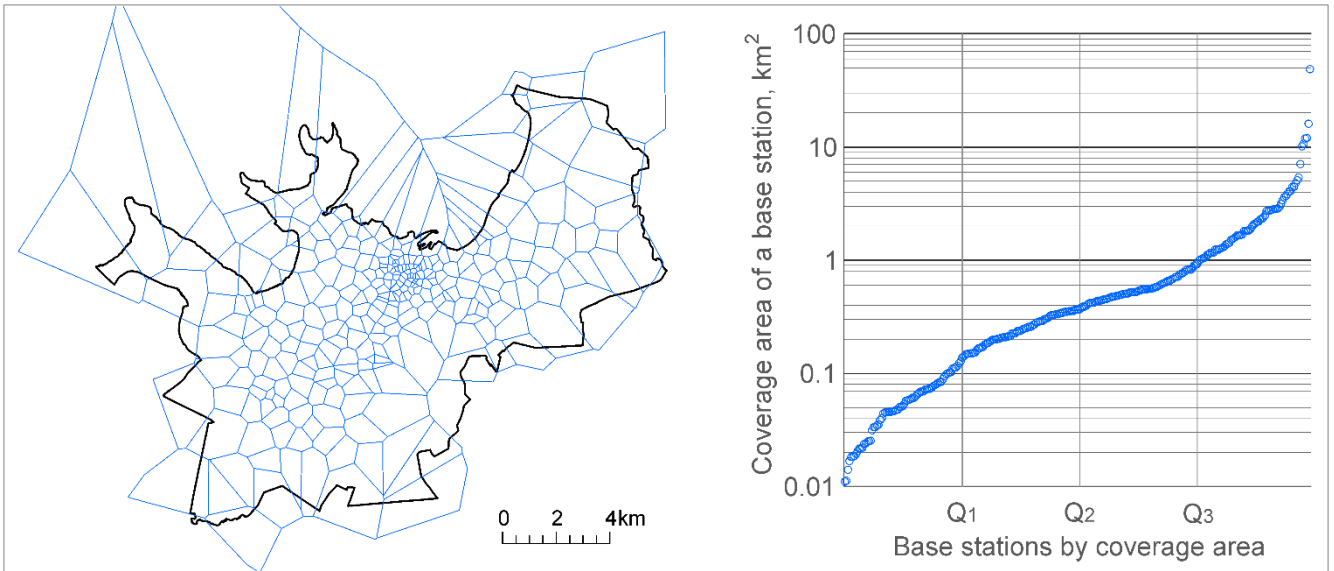
Figure S1. All studied 290 theoretical coverage areas of a base station in Tallinn and the ranking of coverage areas by their spatial extent in a logarithmic scale.

We investigated the proposed multi-temporal function-based dasymetric (MFD) method for mobile phone data with three mobile phone data sets selected from a random sample of 50% from the original databases in March 2015 provided by Positium LBS. Two data sets are extracted from a one-month raw CDR database consisting of a 5.3 million call activities made during working days (Mon.–Fri.) by 58,470 subscribers in the study area. From the latter we select one data set including CDR data from 2 AM to 6 AM to represent a nighttime population and second data set includes CDR data from 4 PM to 5 PM as an example to represent a daytime population. The third data set consists of the most probable home locations of given phone users at a base station level. These data were derived from raw CDR data using the anchor point model (Ahas *et al.* 2010). For each individual in the sample, this model finds the two most frequently used mobile network base stations where mobile phone usage occurs during a month. These base stations are called as anchor points of one's daily life. Furthermore, by taking into account (1) the average time of day and its standard deviations of (outgoing) CAs conducted in each activity location and (2) the spatial relationships between neighbouring activity locations, the model distinguishes home- and work-time locations. In total, the anchor point model extracted 56,778 the most probable home locations that are distributed between selected 290 base stations in the study area.

Certainly, individual mobile phone usage varies significantly in temporal and spatial terms as well as in relation to the overall amount of phone usage. To minimise differences between phone users in terms of total amount of CDRs, we calculated a weighted physical presence for each individual in each base station for each hour of the given time frame based on the number of CDRs generated in each base station from all CDRs conducted during a given hour. Hence, we can better reflect the physical presence of phone users and their time spent between the coverage areas of a base station depending on one's call activity frequency in each base station. For example, to some extent, this solves the problem of the amplification of CDRs conducted by very active phone users who are immobile and located in the city centre (office workers) or who are very mobile throughout the city (e.g. taxi drivers, couriers).

S2. Implementation of MFD method in Tallinn, Estonia

S2.1. The preparation of physical surface layer

To prepare a physical surface layer in the multi-temporal function-based dasymetric (MFD) method for the spatial disaggregation of the study area, first, from The Estonian National Topographic Database (ENTD) a land cover layer is created by aggregating land cover parcels based on six activity function classes, as indicated in Table S1. This enables linking the land cover layer with the classification of building usage and human activity types, similarly to a study by Greger (2015).

Table S1. The classification of activity function type and linkage to land parcels by land cover and buildings by functionality, and the association with human activity types.

Activity function type	Land parcels by land cover type	Building units by functionality type	Human activity type
Residential	residential area	residential, accommodation, prison	at home or accommodation
Work	industrial and business area	industrial, office, public institutions predominantly not for *in situ* services (e.g. police department, hospital, town hall), educational	at work or school
Retail & Service		retail and service (e.g. gas station, cleaner service, local bank branch)	shopping and using services
Other	open area (e.g. park, square), forest, cemetery	public institutions for *in situ* services (e.g. libraries, sport facilities), leisure and entertainment, eatery, churches	other (leisure) activities
Transport	transport networks (e.g. road, railway, parking)		travelling, on the move
Restricted	water, wetland, arable land, restricted area (e.g. wasteland)		*not accessible, no activity*

Second, we prepared the building layer by refining building usage type and estimating total floor area. First, all buildings with a footprint area over 20 m2 are selected from the up-to-date building footprint layer provided by the ENTD. These building units initially have the binary classification of a residential (or public) building and non-residential building. To further refine building classification according to six land cover classes (Table S1), we applied the OSM database to obtain more detailed information about activities and functionalities of buildings. Classification is done semi-automatically since only one building usage type for each building is assigned to indicate the main usage. Hence, from residential buildings assigned by the ENTD all public buildings were distinguished and further assigned with either work/educational or leisure-related usage type. The main usage type for accommodation buildings is considered a residential usage type since customers use it as a temporary home, although it has an important work-related usage type as well. Other (i.e. non-residential building) buildings assigned by the ENTD are considered to have a work-related building usage type. Additionally, the OSM data reveals buildings that have public, retail and service as the main functional type. We acknowledge that it is rather arbitrary to assign one usage type for each building and not to consider multifunctional buildings, however, we believe this ancillary data to be sufficient for highlighting how the dasymetric interpolation approach increases the spatial accuracy of mobile phone data.

Third, we included a vertical dimension for buildings. Since we do not have ready-made available data about the total floor area for each building, we calculated the estimated values based on the building footprint area and building height. We prefer to apply estimates for total floor area (m2) of buildings, which provides a more accurate population estimation than rough building volumes (m3), based on a study by Biljecki et al. (2016). The average height for each building is calculated based on LIDAR measurement points at 1 m accuracy obtained from the normalised Digital Surface Model, which intersects building footprint polygons (see Alahmadi *et al.* 2013). To further estimate the number of floors, the average building height is divided by the floor height set for 3.5 m for residential and 4.5 m for other buildings. The given floor height estimations were set by dividing the measured building height with the actual number of floors based on a random sample of buildings. We further excluded buildings less than 2 m high from the building layer as unsuitable to be inhabited by people. Applied floor heights are rather arbitrary; however, we believe this ancillary data to be sufficient for the case study to highlight how the dasymetric interpolation approach increases the spatial accuracy of mobile phone data.

Fourth, an overlay method was applied to calculate the geometric union of a land cover layer with a building layer, whereas in the case of building polygons, only building attributes remain. For subunits with the surface unit type of land the default value for the number of floors is set to 1. Now, each disaggregated spatial subunit in the physical surface layer has three attributes: the number of floors, activity function type, and spatial unit type of either building or land.

S2.2. The spatial disaggregation by source and target zones

In the second step, a geometric union with the physical surface layer, the layer representing 290 Voronoi polygons as theoretical coverage areas of a base station (source zones) and the official statistical grid cell layer (target zones) is applied to designate a unique base station ID and a unique official statistical grid cell ID for each spatial unit in the physical surface layer. While we apply two statistical grid cell layers with 100 m x 100 m and 500 m x 500 m resolutions, we obtain two physical surface layers with different disaggregation levels. Next, the area for each disaggregated subunit polygon is calculated, which enables a calculation of the estimated total floor *FA* for each spatial subunit by multiplying area with the number of floors. This allows for a calculation of a relative floor area *RFA* for each subunit within a base station from the sum of *FA* within the given base station, as illustrated in Table S2. For example, base station ID 10 covers an area of 70 km^2 with a total sum of *FA* 160 km^2. While a subunit ID 10101 has *FA* 50 km^2, then it has *RFA* 0.29 (50/170=0.29) within the base station ID 10. For comparison, in the case of a simple areal weighting method, all subunits would have *RFA* 0.14 (10/70=0.14).

Table S2. An illustration of calculating a relative floor area (RFA) for each subunit within a base station from the estimated total floor area of the base station (FA).

Subunit ID	Base Station ID (source zone)	Grid cell ID (target zone)	Activity function type	Spatial unit type	Area [km^2]	Number of floors	FA [km^2]	RFA [%]
10101	10	10	Residential	building	10	5	50	0.29
10201	10	20	Residential	land	10	1	10	0.06
10202	10	20	Work	building	10	3	30	0.18
10203	10	20	Work	land	10	1	10	0.06
10301	10	30	Retail&servic	building	10	5	50	0.29
10302	10	30	Restricted	land	10	1	10	0.06
10303	10	30	Transport	land	10	1	10	0.06
Total					70		170	1

S2.3. The integration of time-dependent human activity data

In the third step of the MFD method, the physical surface layer is linked with data on human activity in time. For Tallinn, we applied the time usage of people by activity type on working days (Mon.–Fri.) provided by Statistics Estonia (Figure S2). The data is based on the Estonian Time Use Survey in 2010 that was conducted using the harmonised methodology of Eurostat time use surveys (Eurostat 2009). The average hourly distribution of people by activity type is calculated based on the average time usage of people by activity type for each hour of a day at an aggregated level. In this case study, subpopulation groups are not considered as proposed by the MFD interpolation method, but is calculated for the entire population.
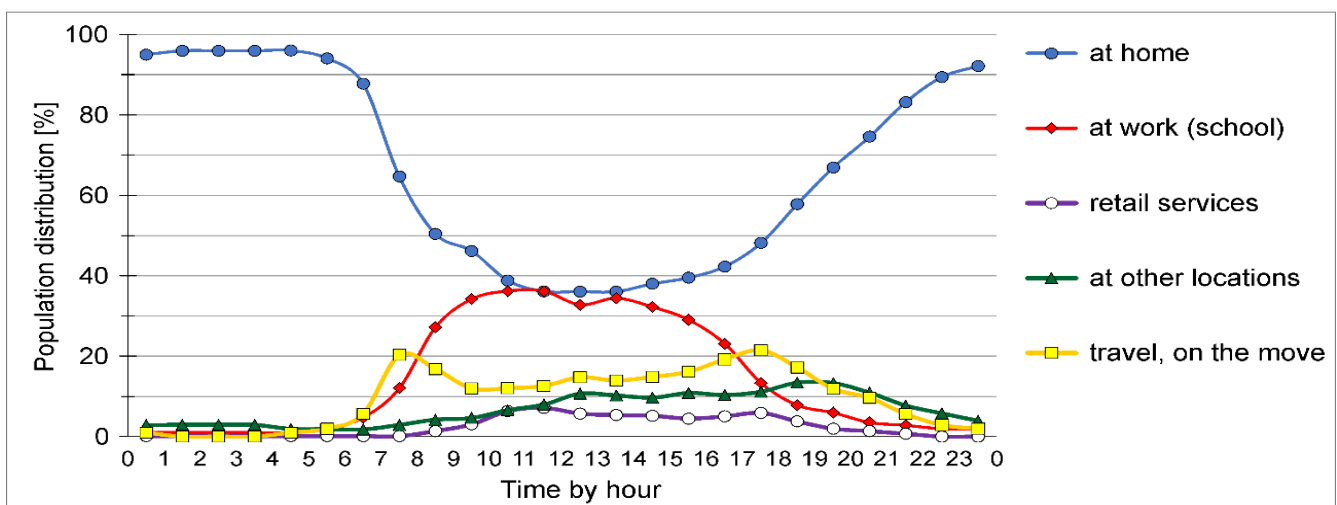

Figure S2. The average hourly distribution of people by human activity type classification on an average working day in Estonia. Source: The Estonian Time Use Survey data in 2010 from Statistics Estonia.

Temporally sensitive human activity information is incorporated to disaggregated spatial subunits in the physical surface layer by linking human activity type with activity function type of subunits (Table S1). To calculate an estimated human presence *EHP* for each disaggregated spatial subunit for every hour of the day, both a daily hour factor *H* and a seasonal factor *M* for estimating human activity is used (Table S3). First, the average hourly distribution of people by activity type is considered as a proxy for *H* for each human activity type. For example, both subunits of ID 10101 and 10201 have an activity location type of a residential building, and hence, for the time period between 4 PM and 5 PM its *H* has the coefficient of 0.42, since at given time period, 42% of Estonians are at home, on average. For spatial subunits with restricted activity function type we consider them inaccessible for people and assume no human presence at any time.

Table S3. An illustration of calculating an estimated human presence (EHP) for each subunit within a theoretical coverage area (Voronoi polygon) of the base station ID 10 in the case of the time period between 4 PM and 5 PM.

Subunit ID	Grid cell ID (target zone)	Activity function type	Spatial unit type	*RFA* [%]	*H* [%]	RFA x H	*M* [%]	$\sum (RFA \times H) \in l$ x M	*EHP* [%]
10101	10	Residential	building	0.29	0.42	0.122	0.90	0.132	0.58
10201	20	Residential	land	0.06	0.42	0.025	0.10	0.015	0.07
10202	20	Work	building	0.18	0.23	0.041	0.90	0.049	0.21
10203	20	Work	land	0.06	0.23	0.014	0.10	0.006	0.02
10301	30	Retail&servic	building	0.29	0.05	0.015	1.00	0.015	0.07
10302	30	Restricted	land	0.06	0.00	0.000	1.00	0.000	0.00
10303	30	Transport	land	0.06	0.19	0.011	1.00	0.011	0.05
Total				1		0.228		0.228	1

Second, the seasonal factor *M* is applied since seasonality influences the division of human presence between indoors and outdoors. Here, *M* is applied as the probability coefficient of people to refine locations of people between disaggregate spatial subunits based on spatial unit type (land; building) within each activity function type and within each base station. A study by Hussein et al. (2012) found that people spend some 90% of time indoors and 10% outdoors in Helsinki, Finland during March. We expect these findings to fit well in our case study given the general similarities of human activity travel behaviour, identical climatological conditions between Helsinki and Tallinn, and both studies are conducted in the same season of the year. Hence, an assumption is made that if an activity function type (e.g. at home) includes both spatial unit types (e.g. residential building; residential land), then *M* has the coefficient of 0.90 for building subunits and 0.10 for land subunits in spring (March). In this study, these coefficients are used for three activity function types (residential; work; and other) and for other activity function types *M* has the coefficient of 1. For example, subunit ID 10101 and 10201 have both residential activity function type, while the former has *M* as 0.90 since it is a building, and the latter 0.1 since it is a land subunit (Table S4).

For example, both subunits of ID 10101 and 10201 have residential activity function type and are located within a Voronoi polygon (theoretical coverage area) of base station ID 10. For the time period between 4 PM and 5 PM *EHP* for subunit ID 10101 is calculated as follows: (((0.29 x 0.42) + 0.025) x 0.9) / 0.228 = 0.58 (Table S4). In other words, for a given hour of the day, some 58% of all people located within the coverage area of the base station ID 10 are estimated to be located within the subunit ID 10101.

S2.4. The integration of mobile phone data

In the fourth step in the MFD method, mobile phone data is linked to the physical spatial layer by base station ID. First, the hourly distribution of mobile phone data as a proxy for people distributed between Voronoi polygons of a base station in given time is normalised to represent the relative population distribution *RMP* of the case study area. Thus, the total sum of *RMP* by base station coverage area (Voronoi polygon) is always 1. Here, *RMP* is calculated for all three mobile phone data sets used in this study (Section S1). For example, if *RMP* is 0.200 for the Voronoi polygon of a base station ID 10, some 20% of all mobile phone data in the case study area is located within a given Voronoi polygon (Table S4). Second, the relative observed population *ROP* for each subunit is

calculated based on *EHP* of given subunit within a base station and *RMP* of a base station to which a given spatial subunit is assigned. For comparison, *ROP* is also calculated based on a simple areal weighting (AW) and a vertical areal weighting (VAW) interpolation method.

Table S4. An example of calculating relative observed population ROP for each subunit within the base station ID 10 regarding simple areal weighting (AW), vertical areal weighting (VAW) and multi-temporal function-based dasymetric (MFD) interpolation method for the time period 4 PM – 5 PM.

Subunit ID	Grid cell ID (target zone)	Activity function type	Spatial unit type	*AW* [%]	*RFA* [%]	*EHP* [%]	*RMP* [%]	*ROP* [%]		
								MFD *(EHP X RMP)*	AW *(AW X RMP)*	VAW *(RFA X RMP)*
10101	10	Residential	building	0.143	0.29	0.58		0.116	0.028	0.058
10201	20	Residential	land	0.143	0.06	0.07		0.014	0.029	0.012
10202	20	Work	building	0.143	0.18	0.21		0.042	0.028	0.036
10203	20	Work	land	0.143	0.06	0.02		0.004	0.029	0.012
10301	30	Retail&servic	building	0.143	0.29	0.07		0.014	0.028	0.058
10302	30	Restricted	land	0.143	0.06	0.00		0.000	0.029	0.012
10303	30	Transport	land	0.143	0.06	0.05		0.010	0.028	0.012
Total				1	1	1	0.200	0.200	0.200	0.200

For example, *ROP* based on *EHP* for the subunit ID 10101 is 0.58 x 0.200 = 0.116, which means that some 11.6% of all people in the case study area derived from mobile phone data is located within given subunit (Table S4). For comparison, with a vertical areal weighting (VAW) method, *ROP* for the same subunit would be 0.29 x 0.200 = 0.058 indicating 5.8% of people to be located in given subunit. With a simple areal weighting (AW) method *ROP* would be 0.029 since each subunit has an equal area of 10 km^2 (10/70 x 0.200=0.029).

S2.5. The spatial aggregation to desired target zones

Finally, each disaggregate subunit now has **ROP** based on *EHP* and is assigned to a unique official statistical grid cell ID, which in this case study is an indication of the target zones layer of the MFD method. By summarising *ROP* of all disaggregate spatial subunits by the official statistical grid cell, the method now provides an outcome of *ROP* by official grid cells as target zones in this case study. Hence, the MFD method has interpolated relative observed population derived from mobile phone data with spatial resolution of Voronoi polygons of a base station (source zones) into official statistical grid cells (target zones).

In the case of the base station ID 10, its 20% of relative observed population from all population is further interpolated into three predefined grid cells whereas, for example, some 11.6% of the population is estimated to be located within the grid cell ID 10 (Table S5). For comparison, with a simple areal weighting (AW) interpolation method, only some 2.8% of the population, and with a vertical areal weighting (VAW) interpolation method, some 5.8% would have been estimated to be located in grid cell ID 10.

For the implementation of the proposed MFD interpolation method for mobile phone data in Tallinn we developed an open access tool for the Python programming language, which is freely available from GitHub (http://doi.org/10.5281/zenodo.252612).

Table S5. An example of interpolating relative observed population (ROP) from a theoretical coverage area of a base station (Voronoi polygon) into predefined grid cells in case of the proposed multi-temporal function-based dasymetric (MFD), simple areal weighting (AW) and vertical areal weighting (VAW) interpolation method.

Grid cell ID (target zone)	Base Station ID (source zone)	*ROP* [%]		
		MFD	AW	VAW
10	10	0.116	0.028	0.058
20	10	0.060	0.086	0.060
30	10	0.024	0.086	0.082
Total		0.200	0.200	0.200

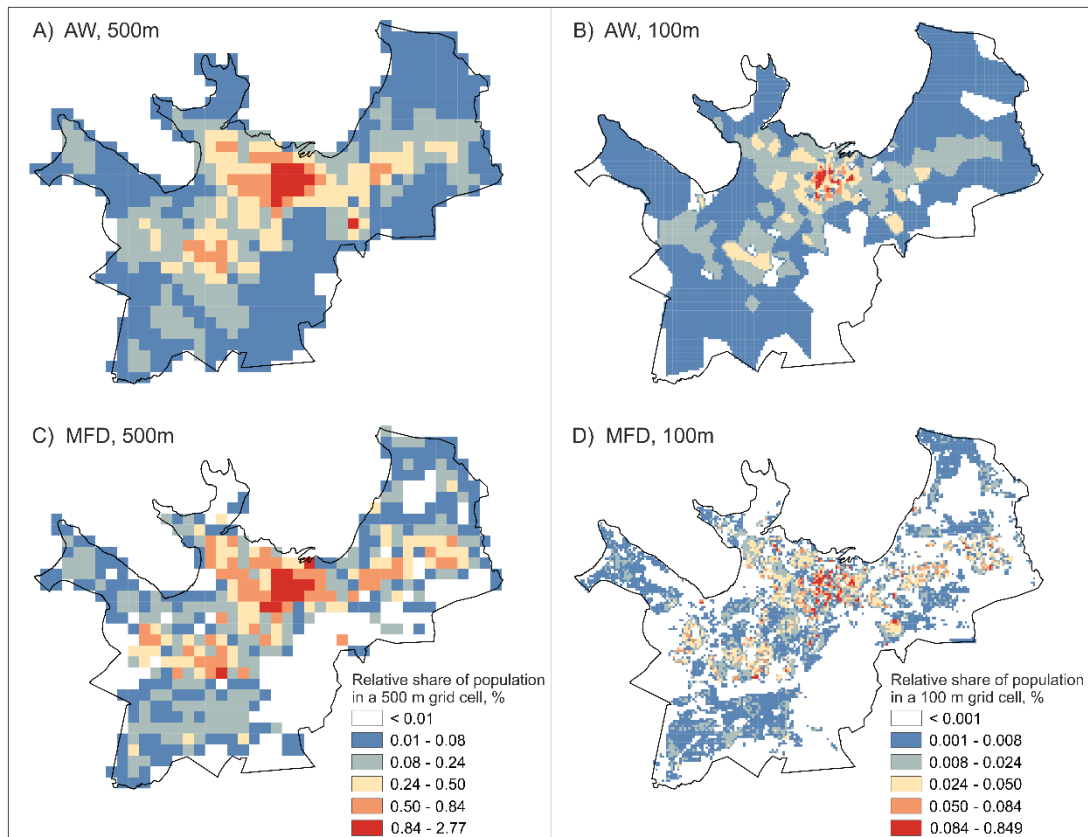## S3. Nighttime population distribution

Figure S3. The distribution of call activities (2 AM – 6 AM) as an indication for a nighttime population by a simple areal weighting (AW) and the proposed MFD interpolation methods with both target zone layers in Tallinn.
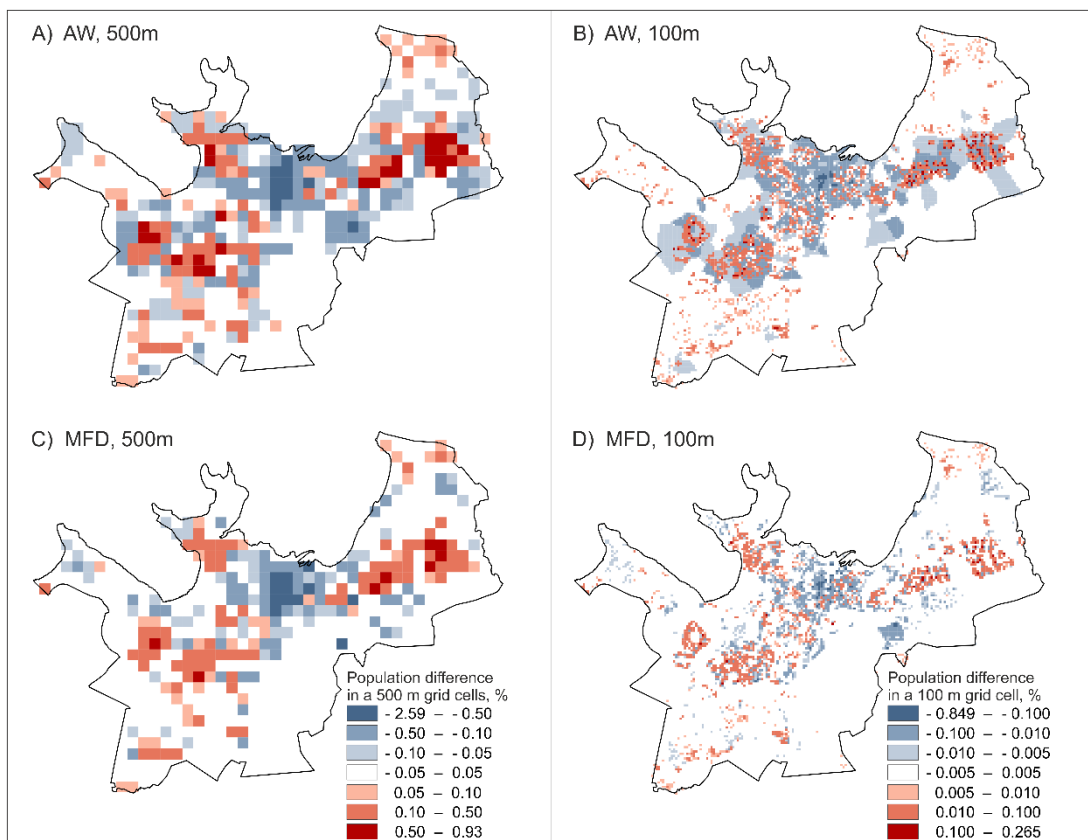

Figure S4. Difference (percentage points) in a nighttime population distribution between population register data as the baseline distribution and call activities (2 AM – 6 AM) by the proposed MFD and a simple areal weighting (AW) interpolation methods with both target zone layers. Target zones in red colour indicate underestimation and blue colour, overestimation of population indicated by call activities.
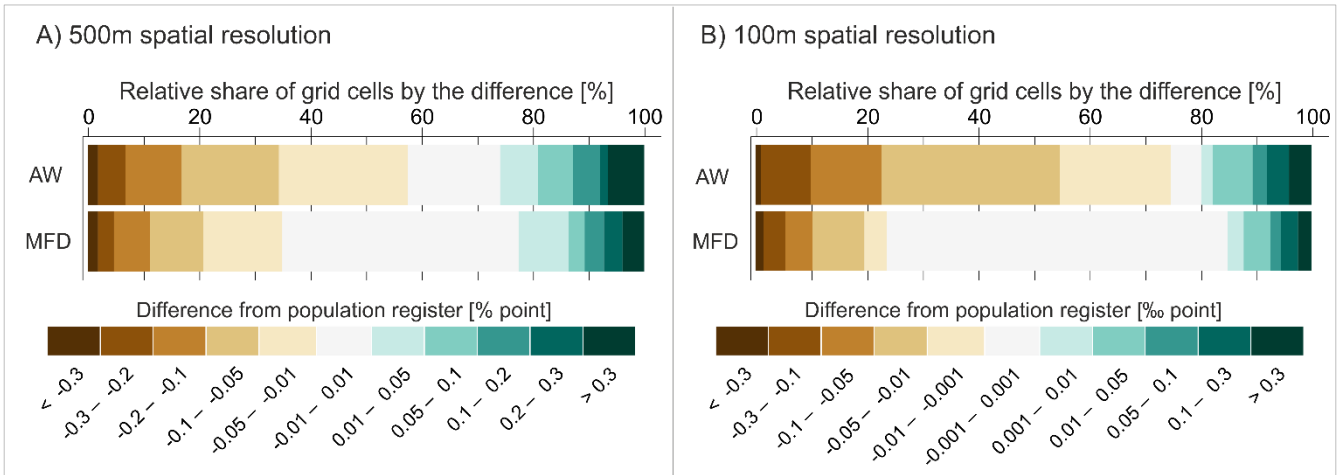
Figure S5. The distribution of target zones by the difference of a sleeping population in a target zone between population register data as the baseline distribution and home anchor points by the proposed MFD and a simple areal weighting (AW) interpolation methods with both target zone layers. Note that differences are indicated in percentage points for 500 m and in per-millage points for 100 m target zone resolution.
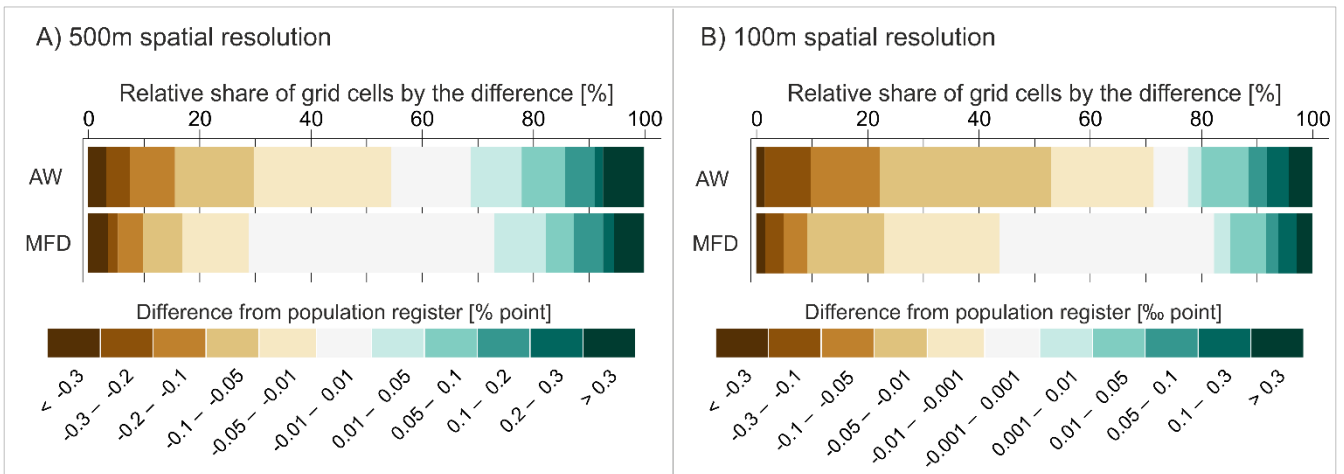

Figure S6. The distribution of target zones by the difference of a nighttime population in a target zone between population register data as the baseline distribution and call activities (2 AM – 6 AM) by the proposed MFD and a simple areal weighting (AW) interpolation methods with both target zone layers. Note that differences are indicated in percentage points for 500 m and in per-millage points for 100 m target zone resolution.
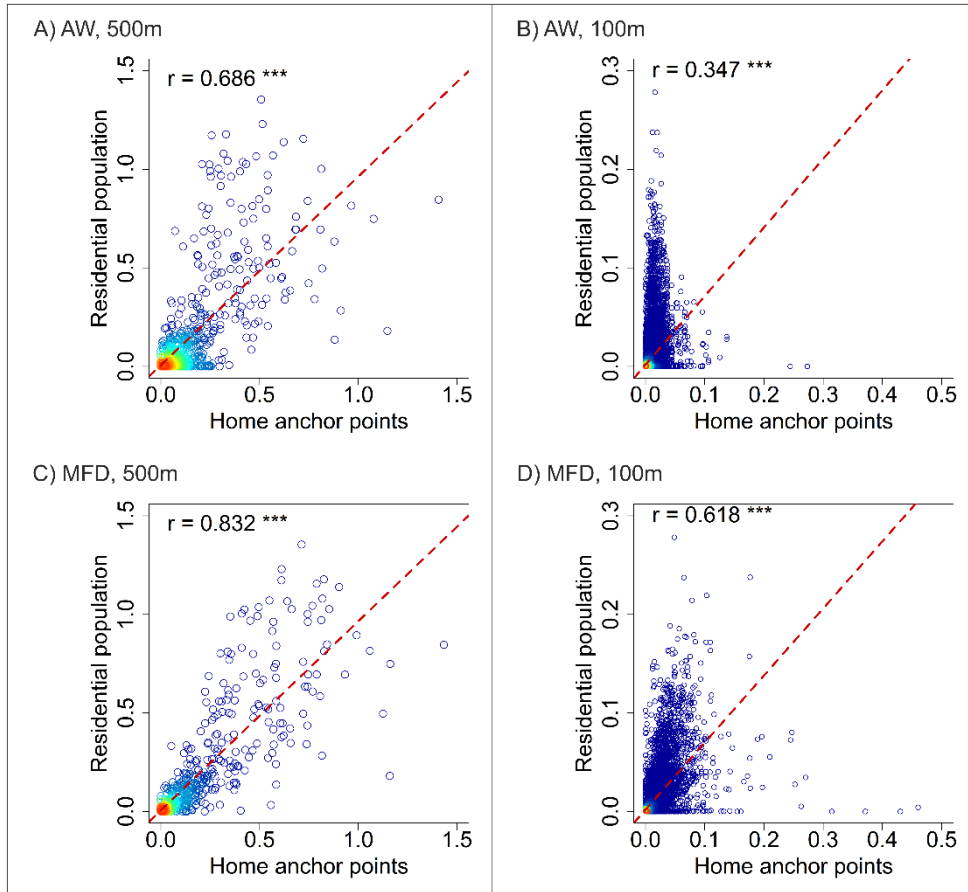
Figure S7. The correlation of a nighttime population distribution between population register data and home anchor points refined by the proposed MFD and a simple areal weighting (AW) interpolation methods with both target zone layers.
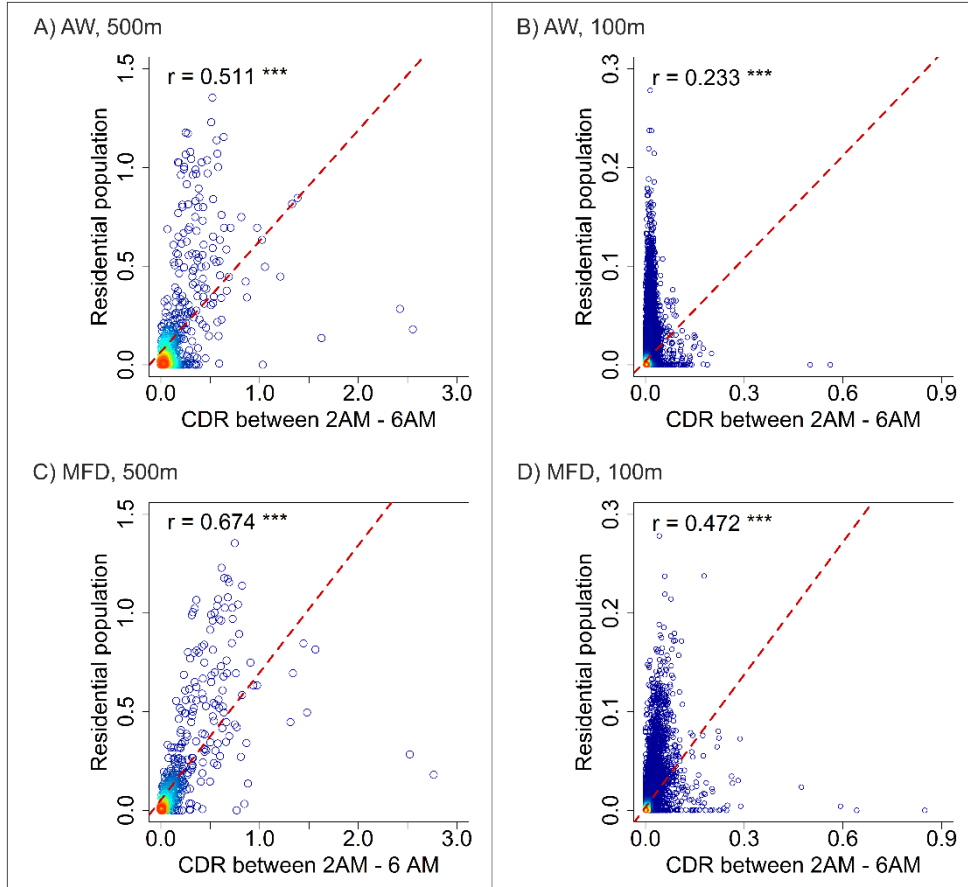

Figure S8. The correlation of a nighttime population distribution between population register data and call activities (2 AM – 6 AM) refined by the proposed MFD and an areal weighting (AW) interpolation methods with two target zone layers.

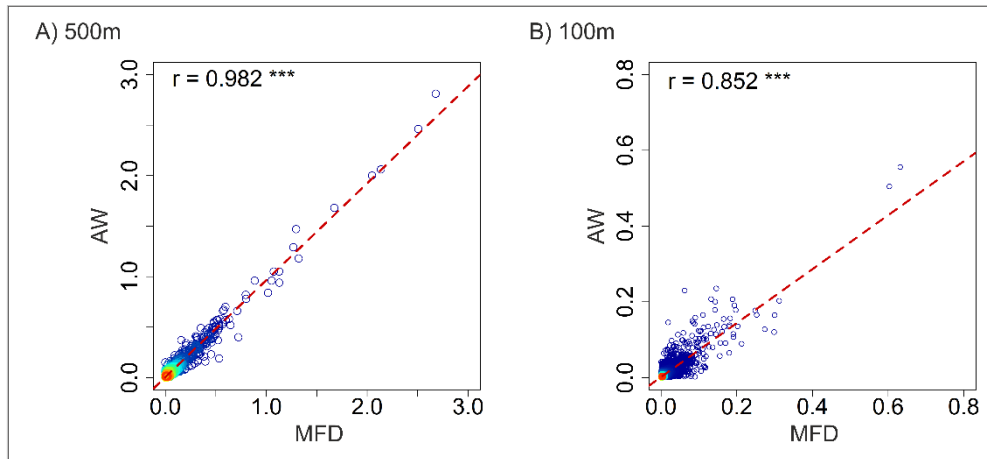## S4. Daytime population distribution

Figure S9. The correlation of a daytime population distribution between population register data and call activities (4 PM – 5 PM) refined by the proposed MFD and a simple areal weighting (AW) interpolation methods with both target zone layers.
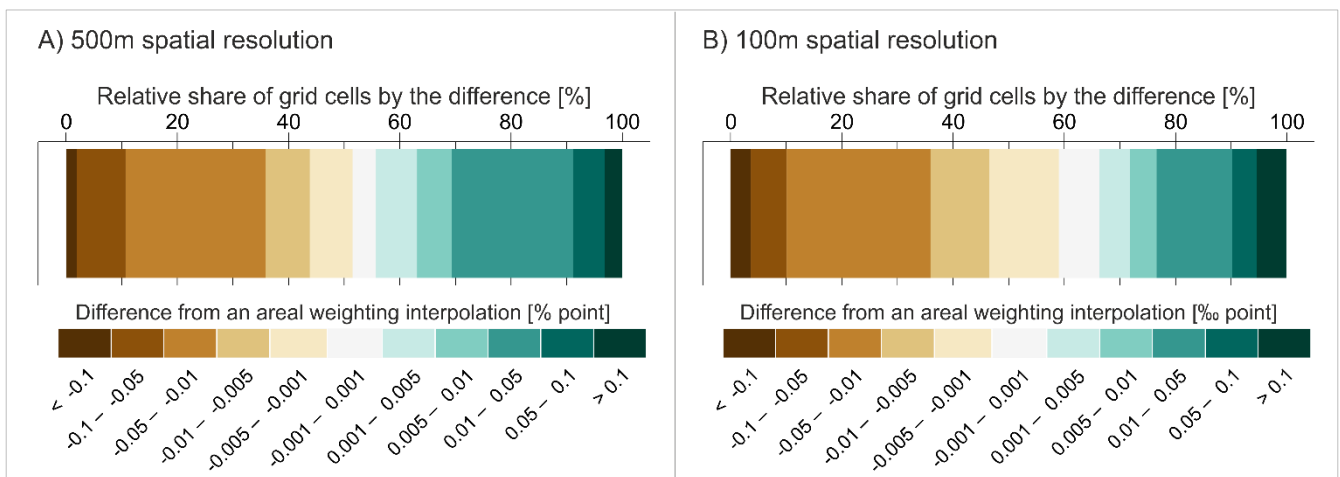

Figure S10. The distribution of target zones by the difference of a daytime population derived from call activities (4 PM – 5 PM) in a target zone between the proposed MFD and a simple areal weighting (AW) interpolation methods with both target zone layers.

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., and Tiru, M., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17 (1), 3–27.

Alahmadi, M., Atkinson, P., and Martin, D., 2013. Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41, 167–176.

Biljecki, F., Arroyo Ohori, K., Ledoux, H., Peters, R., and Stoter, J., 2016. Population Estimation Using a 3D City Model: A Multi-Scale Country-Wide Study in the Netherlands. *PLOS ONE*, 11 (6), e0156808.

European Commission, 2002. Directive 2002/58/EC of the European Parliament and of the Council on privacy and electronic communications. *Official Journal of the European Communities*, 37–47.

Eurostat, 2009. *Harmonised European time use surveys: 2008 guidelines*. Luxembourg: European Communities.

Greger, K., 2015. Spatio-Temporal Building Population Estimation for Highly Urbanized Areas Using GIS: Spatio-Temporal Building Population Estimation. *Transactions in GIS*, 19 (1), 129–150.

Hussein, T., Paasonen, P., and Kulmala, M., 2012. Activity pattern of a selected group of school occupants and their family members in Helsinki — Finland. *Science of The Total Environment*, 425, 289–292.

Järv, O., Ahas, R., and Witlox, F., 2014. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38 (0), 122–135.