MUNI | RECETOX

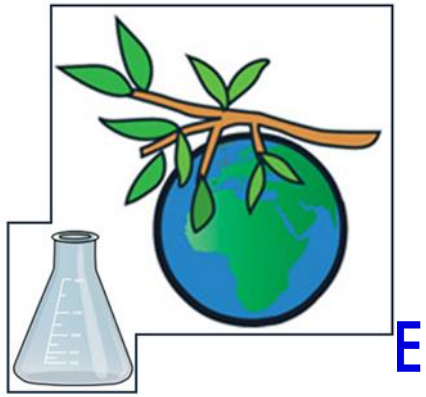
E4221-Modelování a intepretace environmentálních dat

Klára Komprdová

Přehled přednášek a cvičení

- Úvod do statistického modelování, experimentální design, nejistoty modelů
- Prostorové modelyI- prostorová autokorelace
- Prostorové modelyII – interpolační techniky
- Hodnocení časových řad
- Vícerozměrné metody pro identifikaci a klasifikaci znečištění

Environmentální informace, data a studie



ECETOX

Jak na to, když nás zajímá třeba…

Výskyt a hladiny látek v životním prostředí

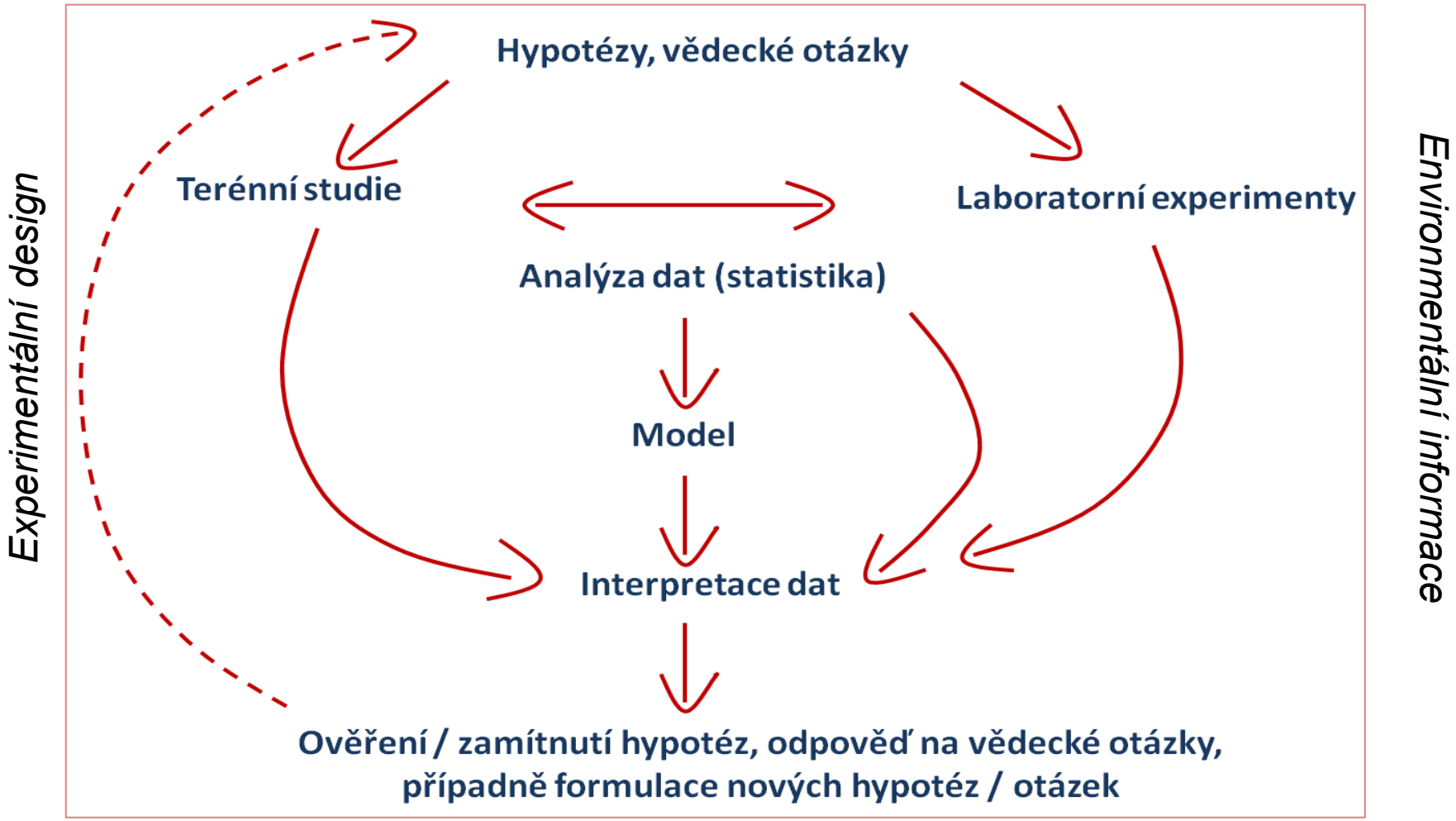
Osud látek v prostředí (např. transport a distribuce)

Monitoring časových a prostorových trendů různých jevů a skutečností

Srovnání modelů s měřenými daty

Rozhodování v oblasti životního prostředí, analýza nákladů a přínosů

Metodický postup u environmentálních studií



Environmentální data

Laboratorními experimenty a terénními studiemi získáváme environmentální data. Jsou to zaznamenané údaje o určitých skutečnostech životního prostředí.

- primární
- agregovaná
- indikátory (ukazatele) životního prostředí
  - kvalitativní indikátory
  - kvantitativní indikátory

Hřebíček a Kubásek, 2011

Environmentální informace

Environmentální informace jsou jakékoli informace v písemné, obrazové, zvukové, elektronické nebo jiné podobě o:

- stavu složek životního prostředí
- faktorech, které ovlivňují nebo mohou ovlivnit stav složek prostředí
- opatřeních, které ovlivňují nebo mohou ovlivnit složky a faktory
- zprávách o provádění právních předpisů o životním prostředí
- analýzách nákladů a přínosů použitých v rámci aplikace opatření
- stavu lidského zdraví a bezpečnosti

Environmentální informační systémy

Existují různé environmentální informační systémy (EIS), které zpracovávají, vyhledávají a prezentují environmentální data.
EIS jsou budovány:

- veřejnou správou na národní úrovni
- veřejnou správou na mezinárodní úrovni
- vědeckými institucemi
- nevládními organizacemi
- podnikatelskou sférou

Příklad environmentální databáze a informačního systému

Global Environmental Assessment Information System (GENASIS)

http://www.genasis.cz

GENASIS poskytuje informace o persistentních organických polutantech

MUNI | RECETOX

Jak hledat informace o životním prostředí v ČR?

Příkladem webového portálu nevládní organizace, která seznamuje se základními informačními zdroji o životním prostředí v ČR, je:

http://arnika.org/jak-a-kde-najit-informace-o-zivotnim-prostredi-cr

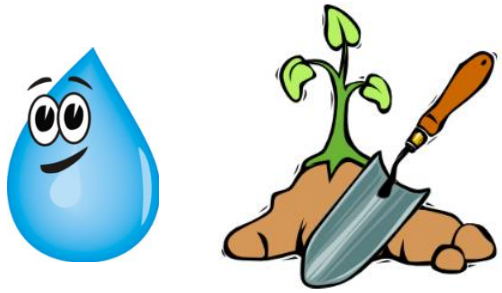
MUNI | RECETOX

Experimentální design

Pokud jsme nenašli požadovaná data a informace, musíme je sami vytvořit

laboratorní studie a experimenty

experimenty a odběry vzorků v terénu



kombinací obého

Experimentální design

Datový soubor, který hodnotíme, by měl být:

- **dostatečně velký** – tj. měl by obsahovat množství vzorků dostatečné pro popis situace, statistické vyhodnocení, spolehlivé modelování apod.
- **nezávislý** – tj. design
- **reprezentativní** – tj. měl by pokrývat celou oblast našeho zájmu; celý rozsah možností, které zkoumáme by měl být objektivní a nic nepreferovat
- **získaný konzistentní metodologií** – tj. měl by zaručit odběr/analýzu vzorků stejnou metodikou nebo srovnatelnými metodikami
- **se signifikantní přesností** – tj. měla by být získána takovými metodami, které jsou výrazně přesnější než variabilita souboru

Hengl (2007)

! PROBLÉM: v reálu tomu tak často není
PROTO je nutné vše dobře plánovat! !

Experimentální design – příklad

Zavádění nové analytické metody v laboratoři pro stanovení různých koncentrací vybraného polutantu v několika environmentálních matricích. Soubor dat by měl splňovat tyto podmínky:

- **dostatečně velký** – soubor různých naspikeovaných koncentrací polutantu v matricích musí dostatečně pokrýt gradient znečištění
- **reprezentativní** – metodu je třeba vyzkoušet na všech matricích, které budou v budoucnu studovány
- **nezávislý** – existuje-li podezření, že metoda má horší výsledky u nízkých koncentrací polutantu, není možné je do studie nezahrnout
- **získaný konzistentní metodologií** – celý analytický postup musí být stále stejný, jak u zavádění metody, tak u její následné rutinní aplikace na reálné vzorky
- **se signifikantní přesností** – limity detekce a kvantifikace musí odpovídat reálným hladinám polutantu v prostředí

Rozdělení modelů

Popisuje budoucí stav systému nebo jeho podmínek?

ANO dynamické modely - závislé na čase - s*pojité, diskrétní*
NE statické modely - nezávislé na čase

Popisují prostorovou strukturu?

ANO prostorově heterogenní - *diskrétní*, *spojité*
NE prostorově homogenní modely

Zahrnuje náhodnou složku?

ANO stochastické modely
NE deterministické modely

Podle čeho vybírat model?

Výběr modelu záleží na zkoumaném problému. Je třeba brát v potaz tyto aspekty:

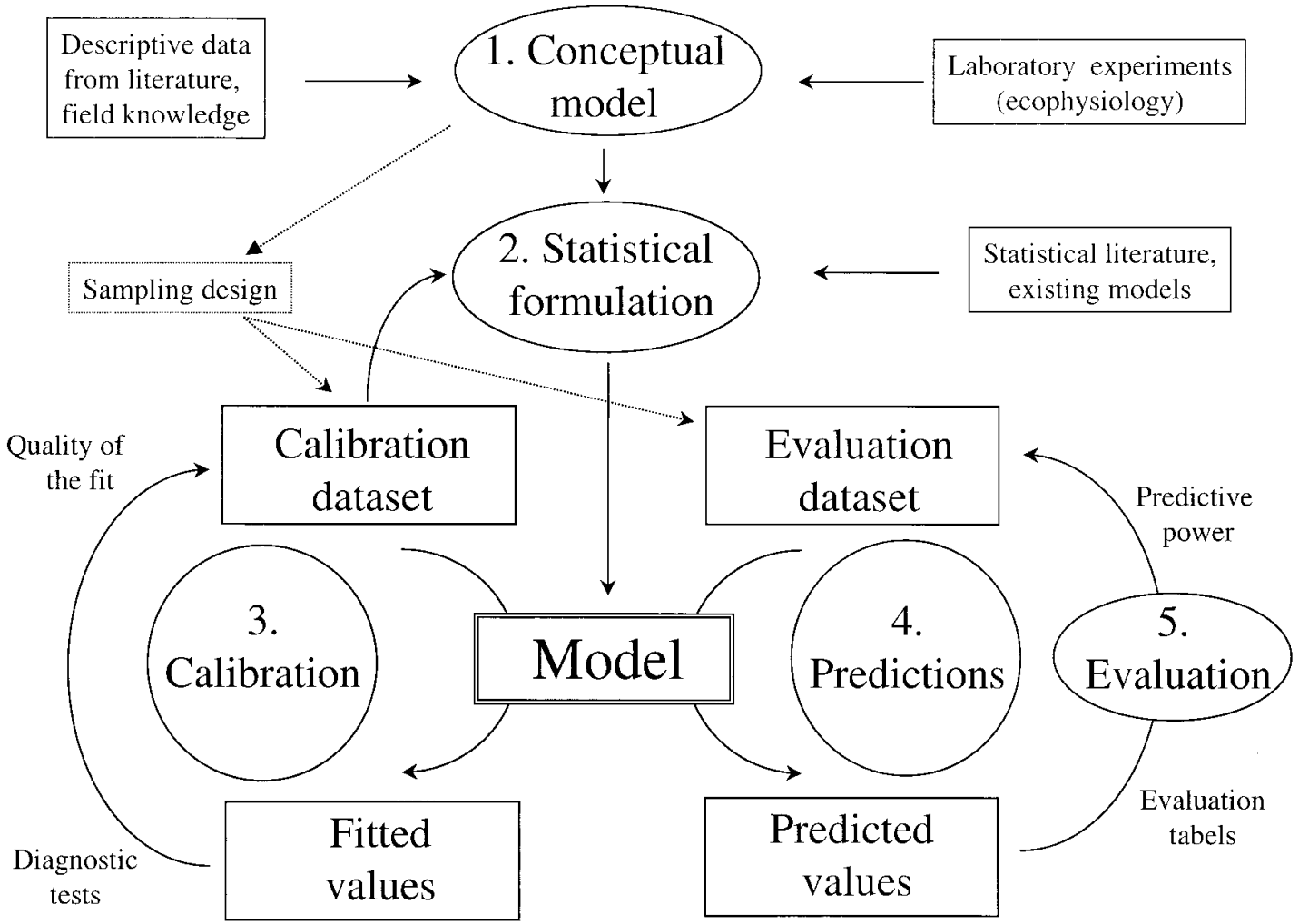
- povaha problému, hypotézy, řešené otázky
- měřítko – např. velikost zkoumaného území
- povaha dat, které jsou k dispozici – např. odlehlé hodnoty
- velikost datového souboru, který je k dispozici - metody vhodné pro malé/velké soubory
- přesnost modelu
- interpretovatelnost modelu

- a řadu dalších

Proces modelování

- design vzorkování a zpracování dat (z literatury, předešlých experimentů)
- terénní sběr dat a laboratorní analýzy
- analýza datového souboru a tvorba modelu
- kalibrace a validace modelu
- interpretace modelu, jeho srovnání s realitou
- použití modelu

Proces modelování



Typy dat

Různé **typy dat** rozlišujeme podle toho, jakých hodnot může daná skupina dat nabývat nebo jaké operace s nimi lze provádět.

- **kvalitativní (kategoriální)**: lze pouze určit, zda jsou dvě „hodnoty“ stejné nebo se liší
  - např. typ půdy
- **semikvantitativní (ordinální):** lze určit rovněž pořadí hodnot
  - např. teplota po stupních
- **kvantitativní (spojité):** lze provádět všechny matematické operace, mohou mít intervalovou nebo poměrovou podobu
  - např. koncentrace látek
- **binární:** lze je považovat za kvantitativní, semikvantitativní i kvalitativní proměnnou
  - výskyt/ nevýskyt látky (informace typu ANO/NE)

Nejistoty modelů

Nejistoty, se kterými se při modelování potýkáme, s nimiž je třeba počítat a které musíme znát, jsou zejména dvou typů:

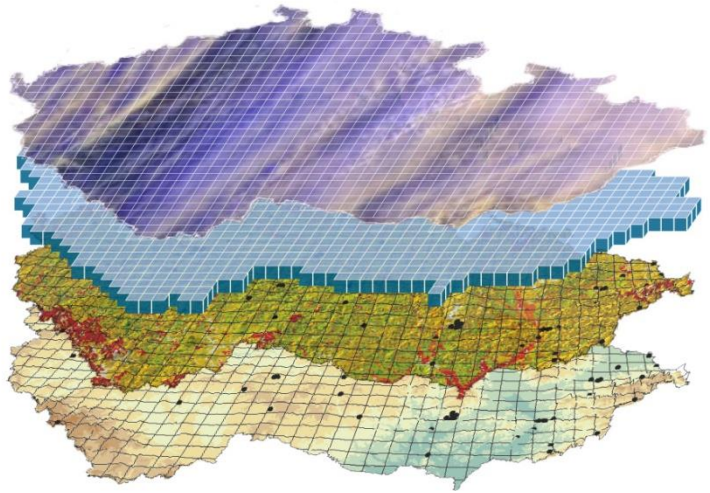
- nejistoty proměnných (plynoucí z chyb při odběru vzorků a analýze v laboratoři, agregace dat, odečítání hodnot z map, designu experimentu apod…), které do modelu vstupují
- nejistoty modelů samotných (konstrukce modelů, zjednodušující předpoklady…)

Prostorové modelování - Jak jsou data rozložená v prostoru?

Prostorové modelování

Prostorová analýza :

- Hledá a popisuje různé vzory v geografickém prostoru
- Snaží se porozumět prostorovým jevům
- Využití geografických informačních systémů



Co nás zajímá?

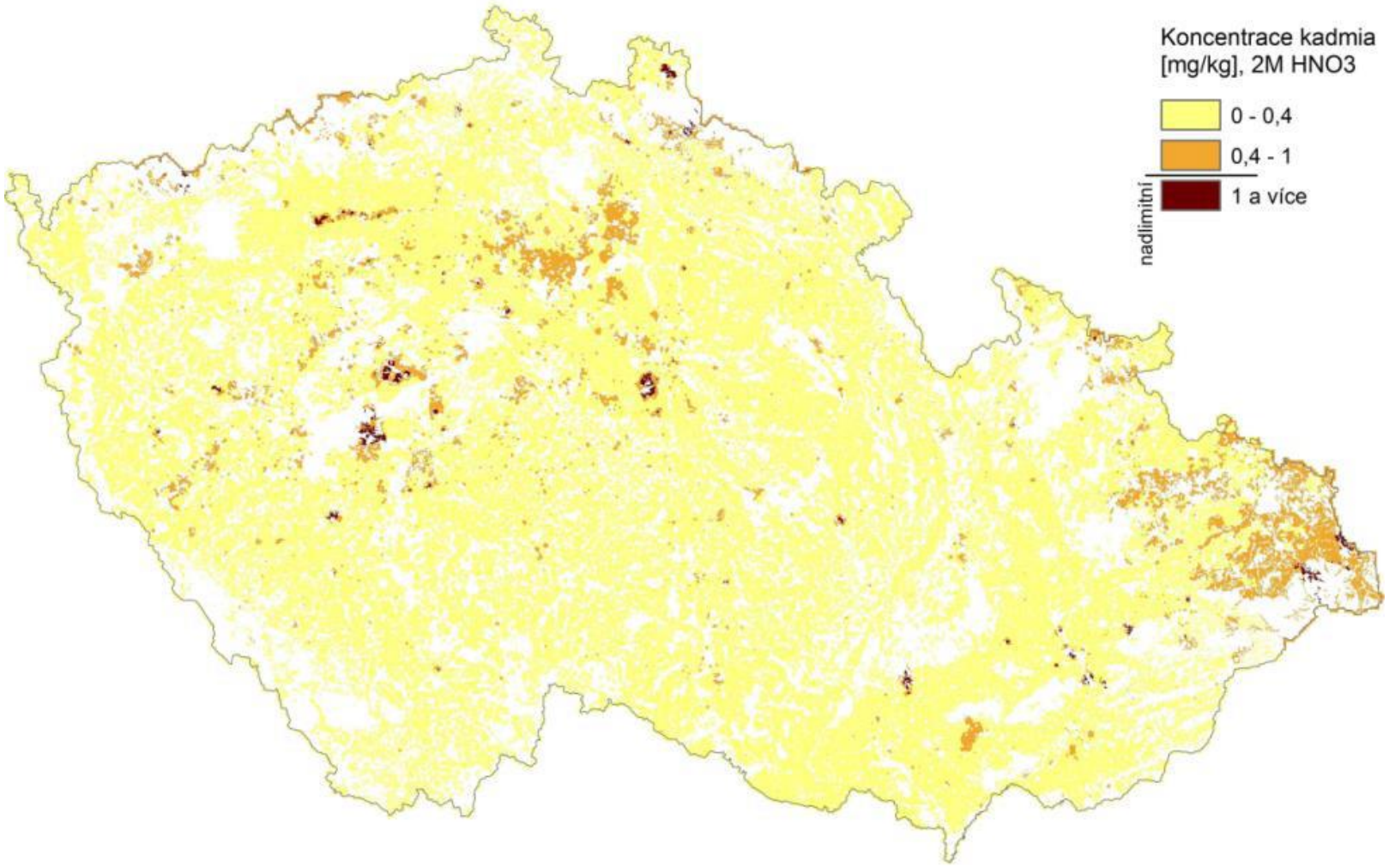
- Jak se pozorování mění v prostoru?
- Co způsobuje tuto změnu v prostoru?
- Kolik pozorování (např. lokalit) potřebujeme, abychom dokázali popsat prostorovou variabilitu?
- Jaká bude hodnota proměnné na novém místě?
- Jaká je nejistota našeho odhadu (predikce)?

T. Hengl (2007) A Practical Guide to Geostatistical Mapping of Environmental Variables

Co všechno můžeme modelovat v prostoru?

- konkrétní hodnoty – (koncentrace, početnosti…)
- pravděpodobnosti – (pst překročení limitu…)
- presence/absence – (přítomnost/nepřítomnost polutantu…)
- nejvíce pravděpodobná entita – (typy půdy, převažující typ znečištění, využití krajiny…)

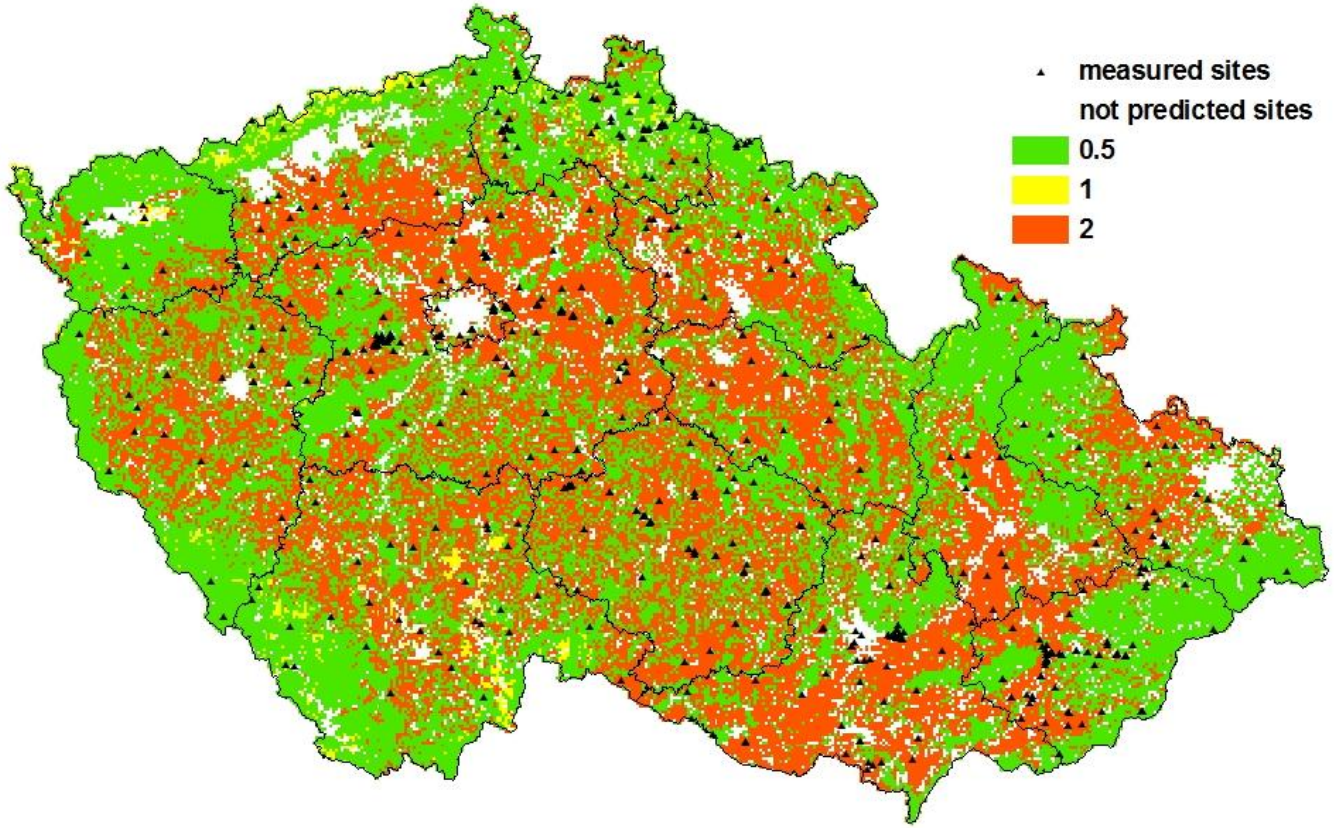
Koncentrační mapa



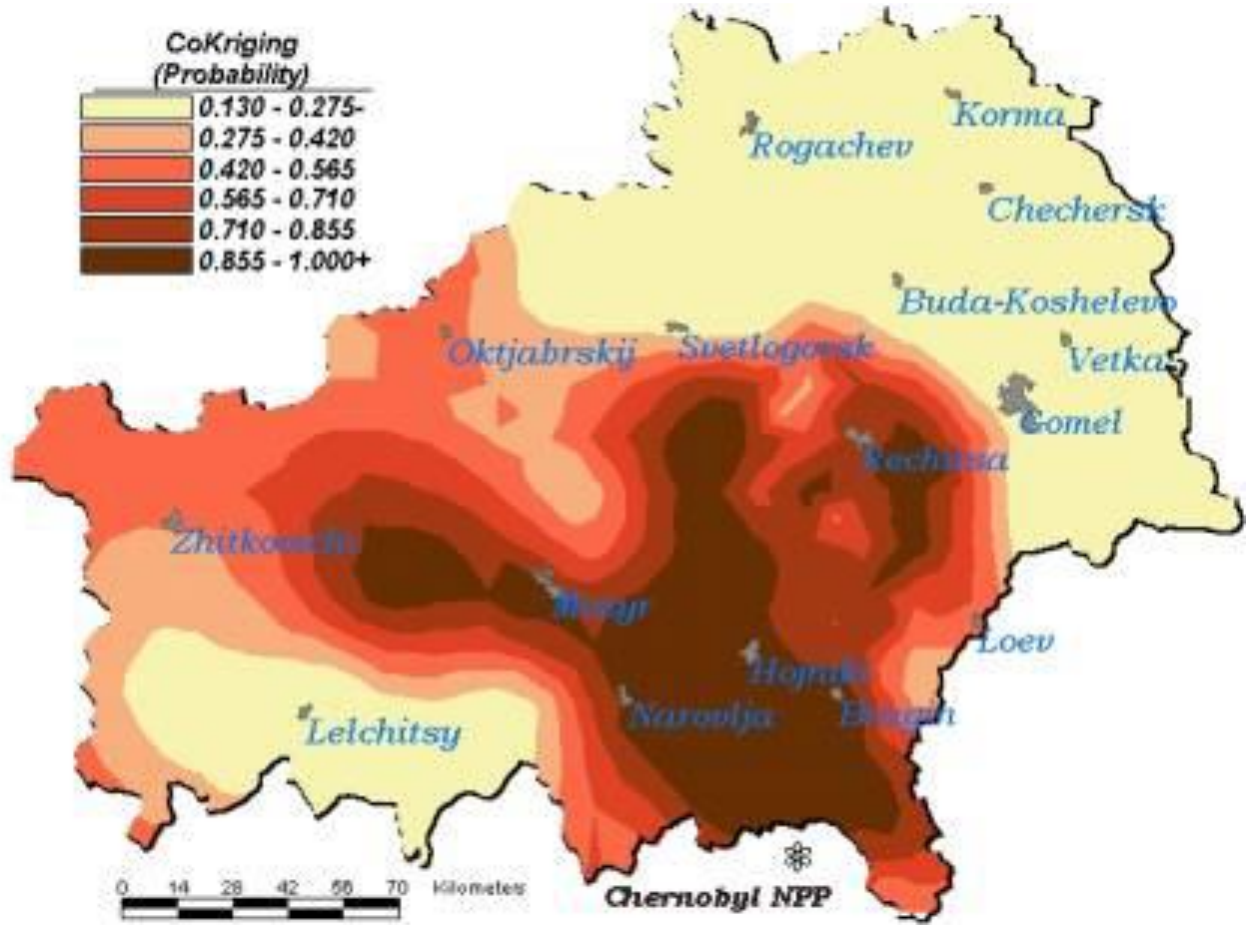
Koncentrace **kadmia** na území ČR s využitím metody **IDW**

MUNI | RECETOX

Mapa zásob DDT (kg/km^2)



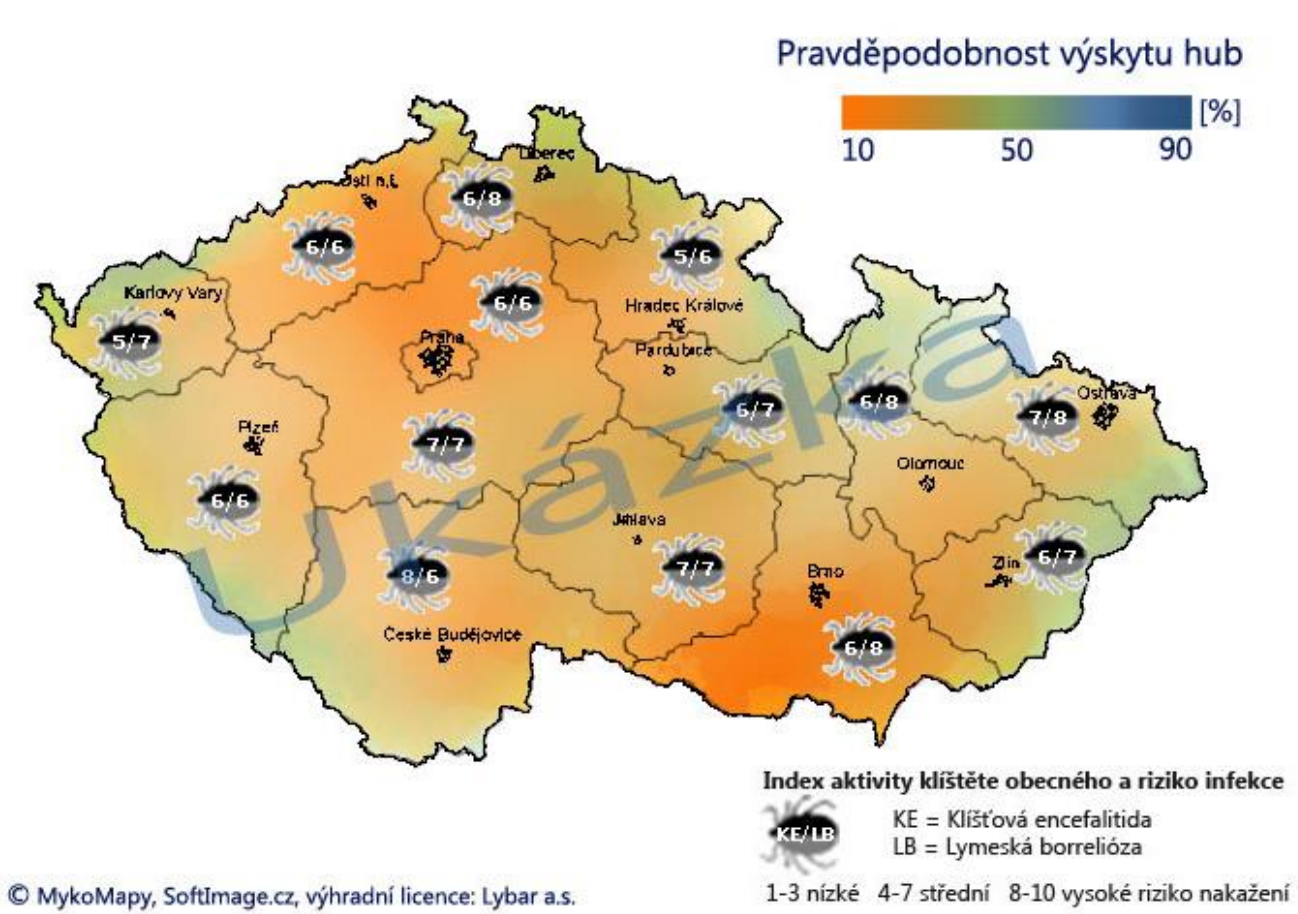
Pravděpodobnostní mapa I



Pravděpodobnost překročení limitní hodnoty 100 Bq/m^2 u 241Americia v půdě v oblasti severně od Černobylu v roce 1992 (Krivoruchko 1999)

MUNI | RECETOX

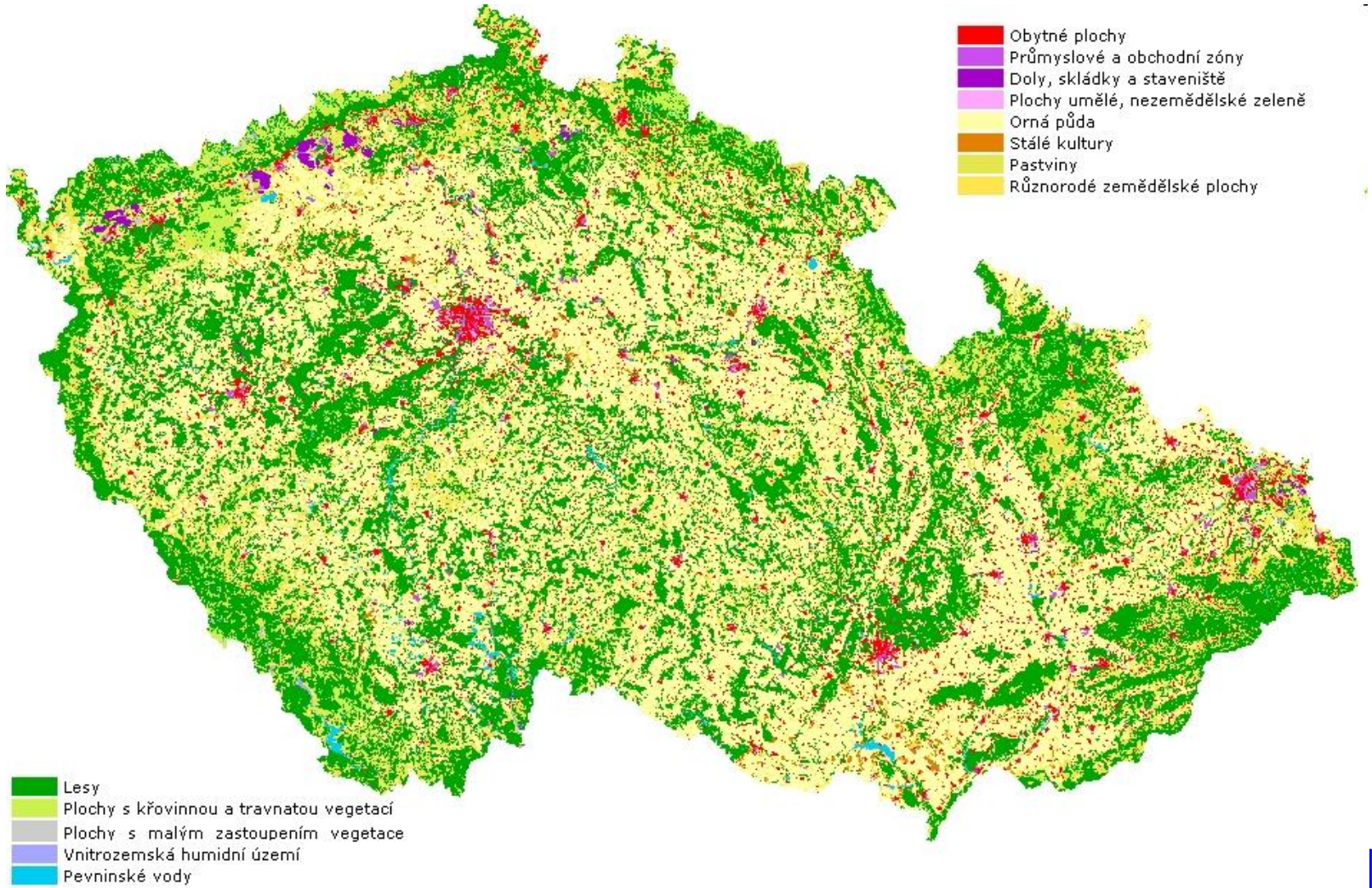
Pravděpodobnostní mapa II



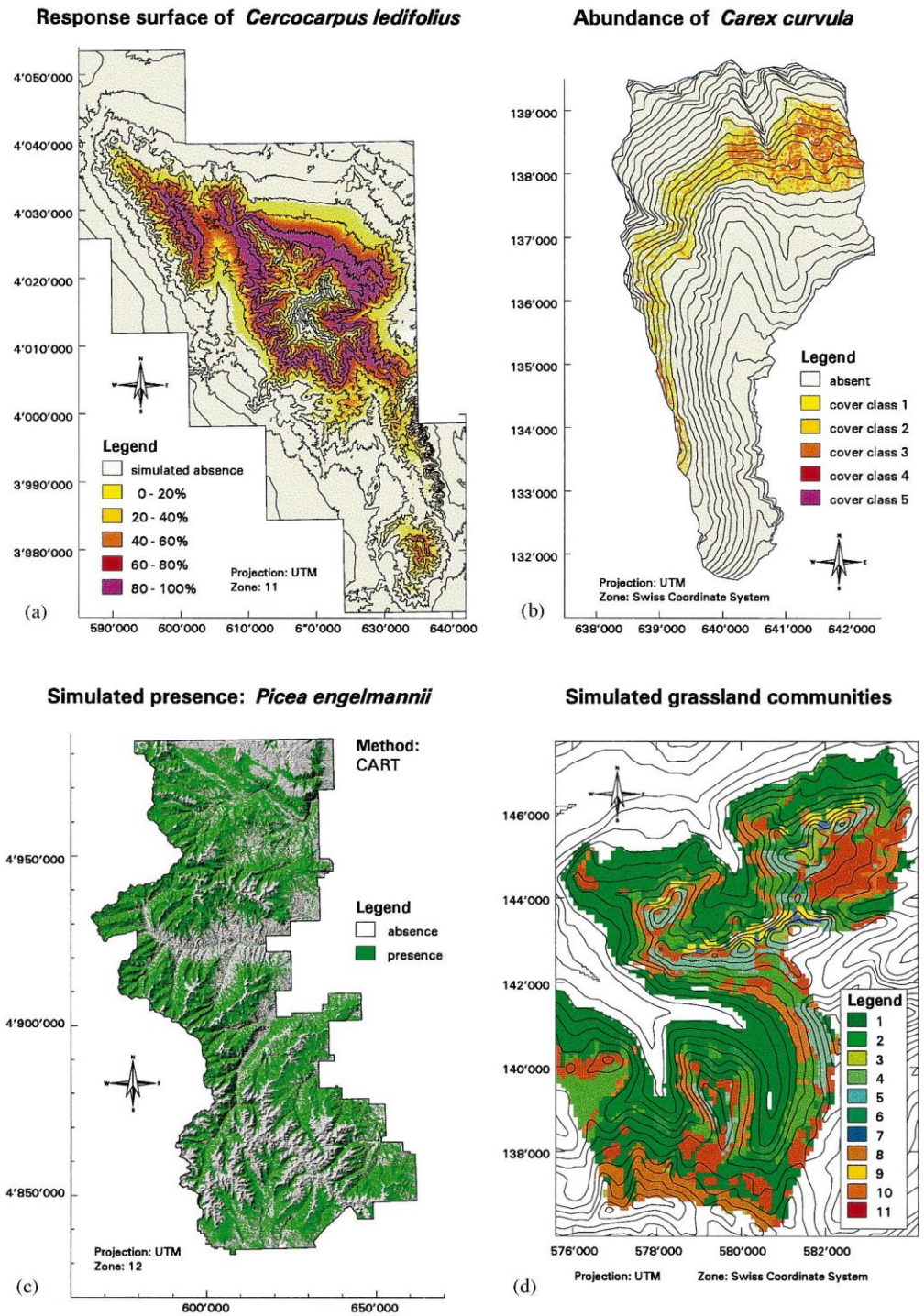
Mykomapa - předpovídá pravděpodobnost růstu hub na území ČR
a současně informuje o možném riziku nakažení nemocemi přenášenými klíšťaty

MUNI | RECETOX

Mapa krajinného pokryvu



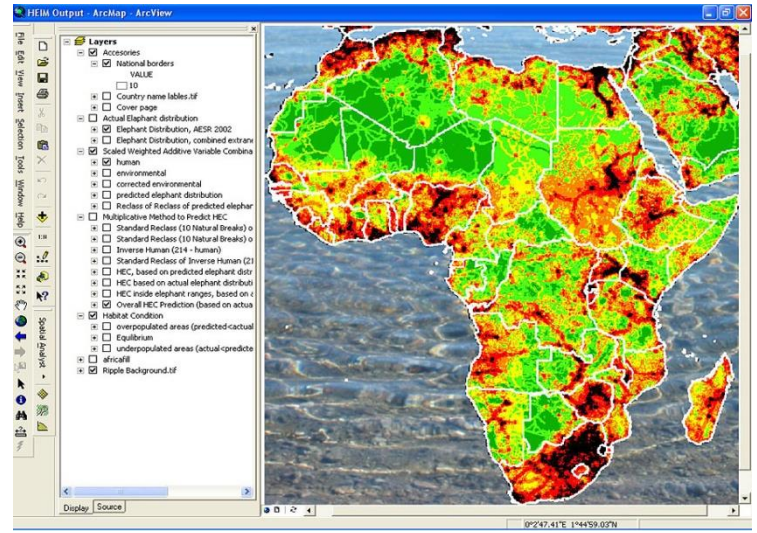
RECETOX

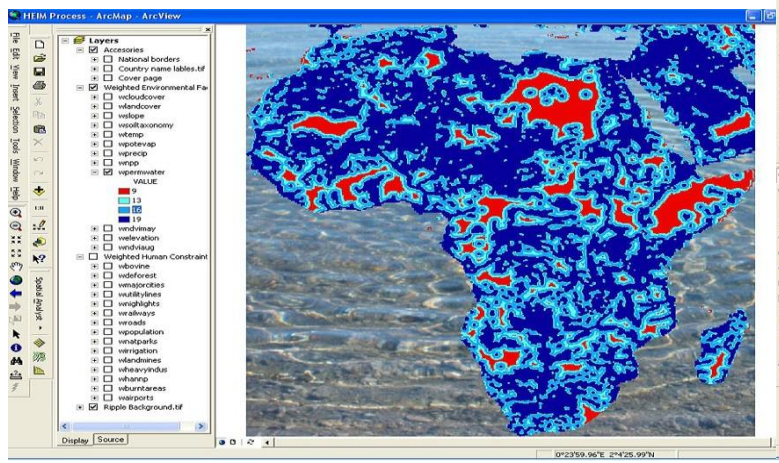


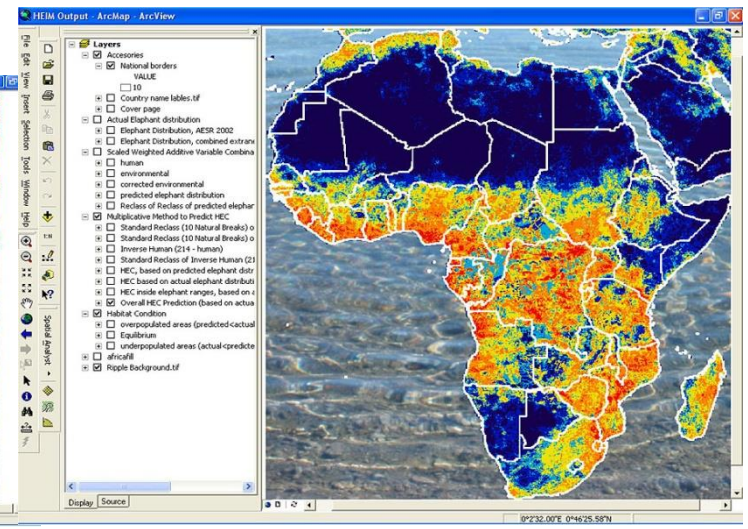
Modelování vegetace
a) pravděpodobnosti
b) abundanční skóre
c) výskyt/nevýskyt
d) vegetační typy

MUNI | RECETOX

Guisan & Zimmermann, 2000

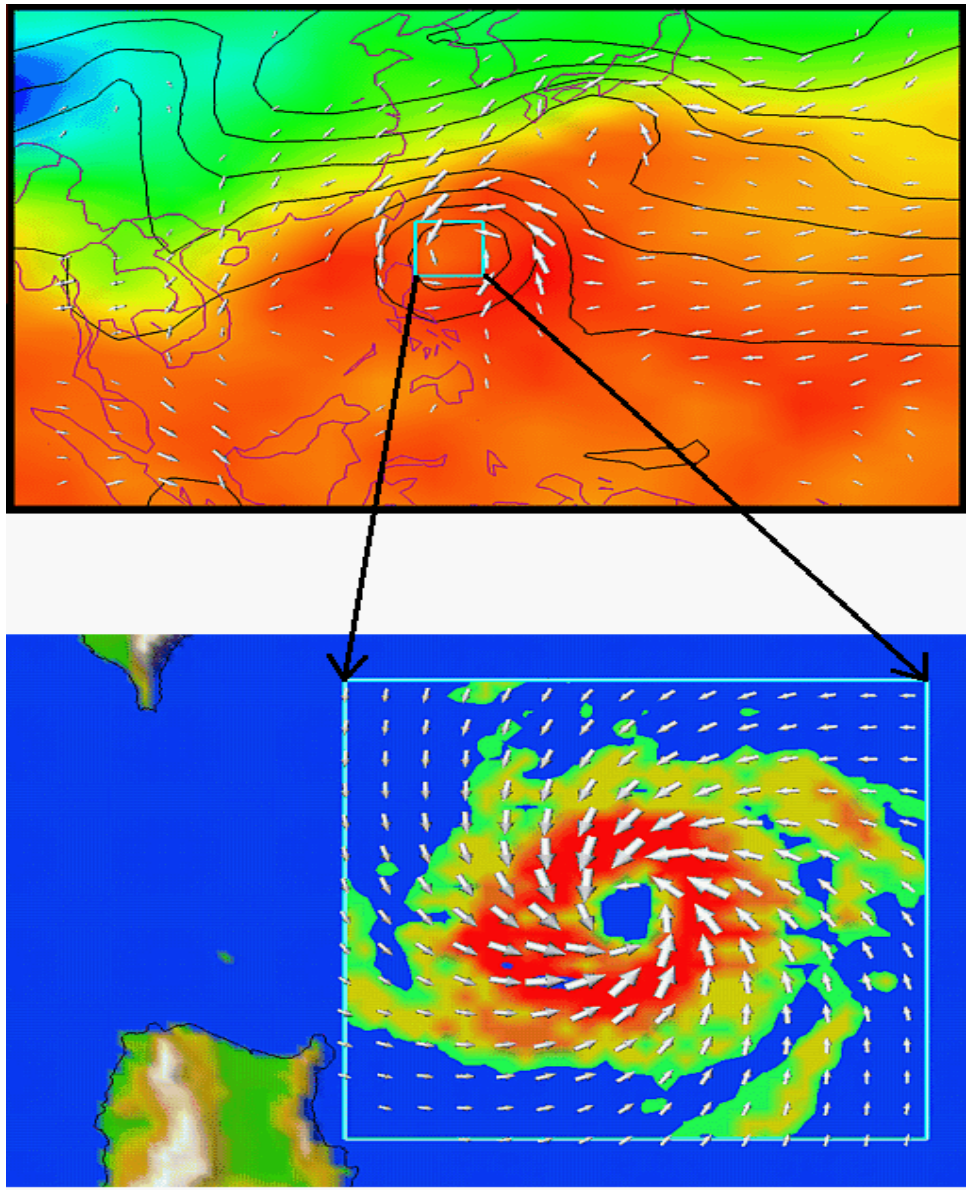






www.esri.com, GIS in Africa

MUNI | RECETOX



http://www.gfdl.noaa.gov/global-warming-and-hurricanes-figures

Prostorová distribuce a plán vzorkování (sampling design)

Kvalitní datový soubor

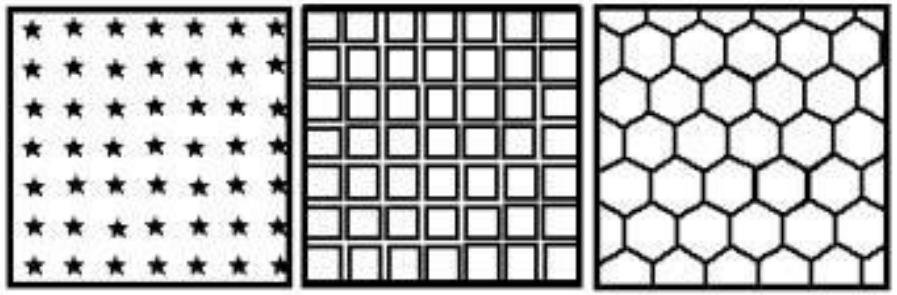
- dostatečně velký
- reprezentativní
- získán konzistentní metodologií
- se signifikantní přesností
- nezávislý

- **Vzorkování**
- jednoduchý náhodný výběr
- systematický výběr
- stratifikovaný náhodný výběr
- preferenční sběr

Testování prostorové distribuce

Náhodný typ distribuce pro 3 typy prvků: body, linie, areály

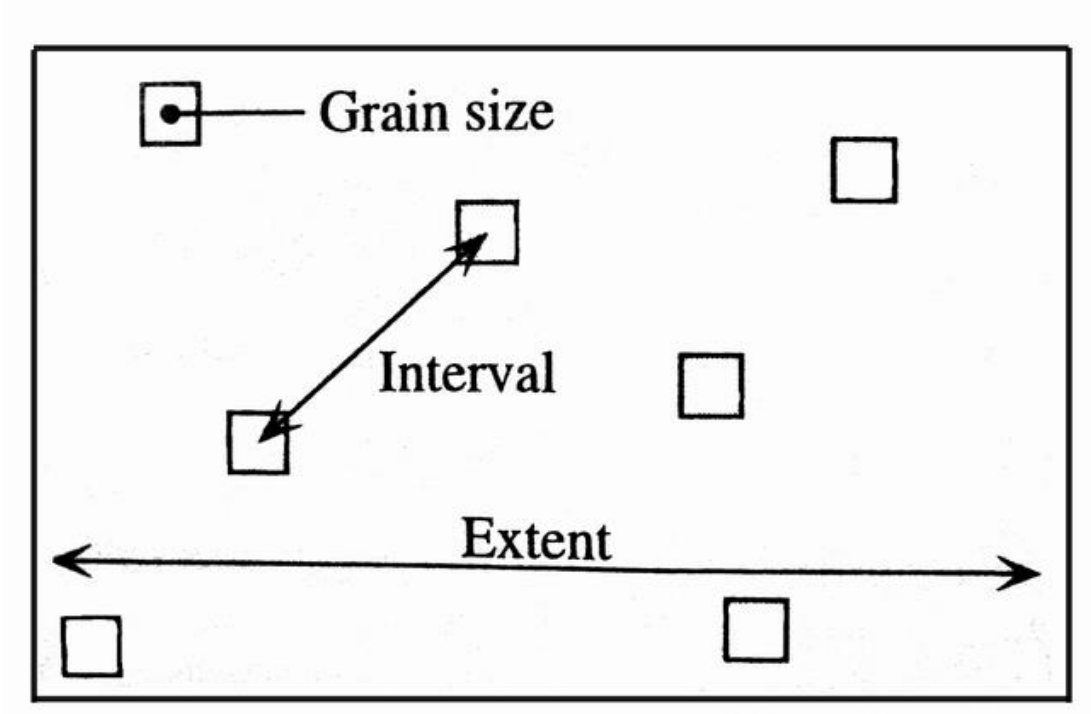
Pravidelný typ distribuce pro 3 typy prvků: body, linie, areály



Shlukový typ distribuce pro 3 typy prvků: body, linie, areály

Komponenty vzorkování

- velikost zrna (grain size) je velikost základní vzorkovací jednotky, může být vyjádřena jako průměr, plocha či objem
- interval (sampling interval) je průměrná vzdálenost mezi sousedícími vzorkovacími jednotkami
- rozsah (extent) – celková délka, plocha nebo objem zahrnutý do studie



(Legendre & Legendre, 1998)

MUNI | RECETOX

Interpolace x Extrapolace

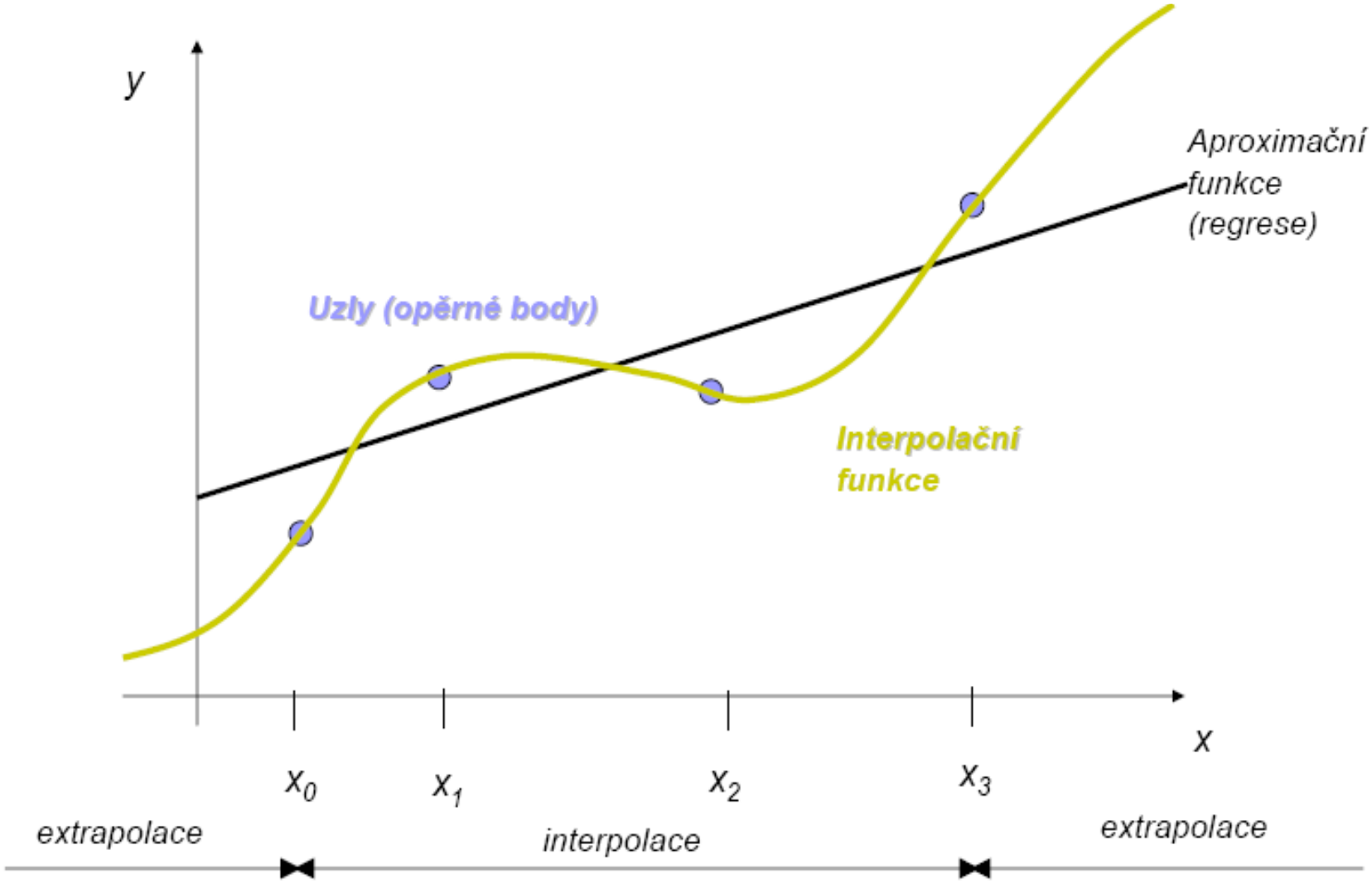
Interpolace – pro „známé“ území (oblast o které máme informace)

- nejsou potřeba žádné další informace o podmínkách daného území
- parametry modelu jsou voleny libovolně či empiricky
- neodhaduje se predikční chyba
- většinou nejsou kladeny žádné statistické předpoklady

Extrapolace – použití modelu na nové území

- potřebujeme další informace o podmínkách daného území
- složitější modely
- odhad chyby predikce
- statistické předpoklady
- sada parametrických i neparametrických metod

Interpolace, aproximace, extrapolace

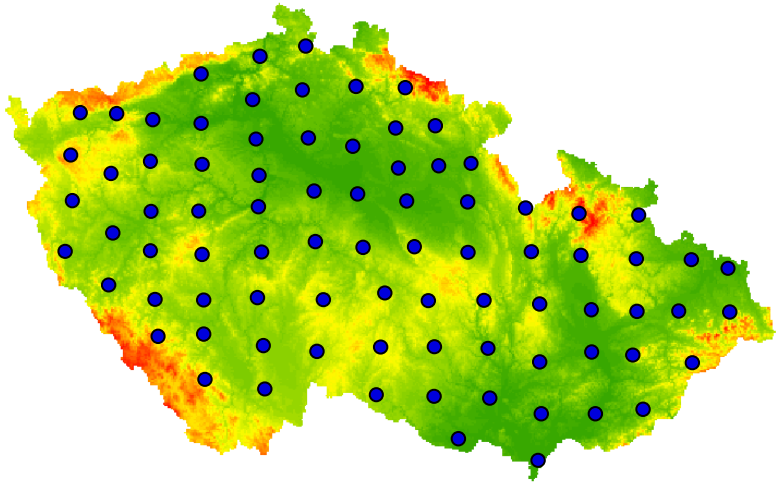


F. Ježek (2006) Interpolace funkcí

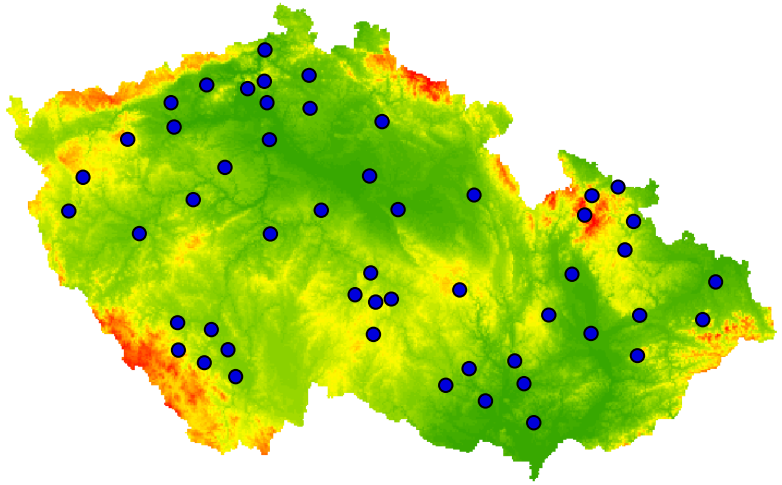
RECETOX

Interpolace x Extrapolace

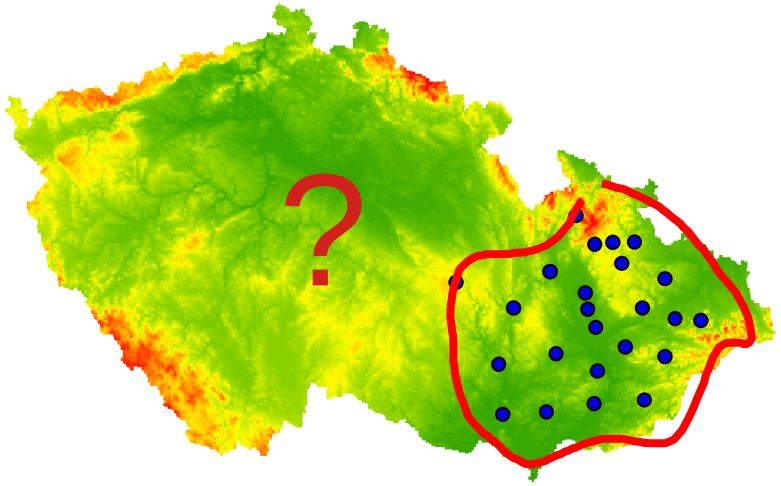
Interpolace – rovnoměrné vzorkování



Interpolace – nerovnoměrné vzorkování



Extrapolace – prediktivní modelování



MUNI | RECETOX

Interpolační metody

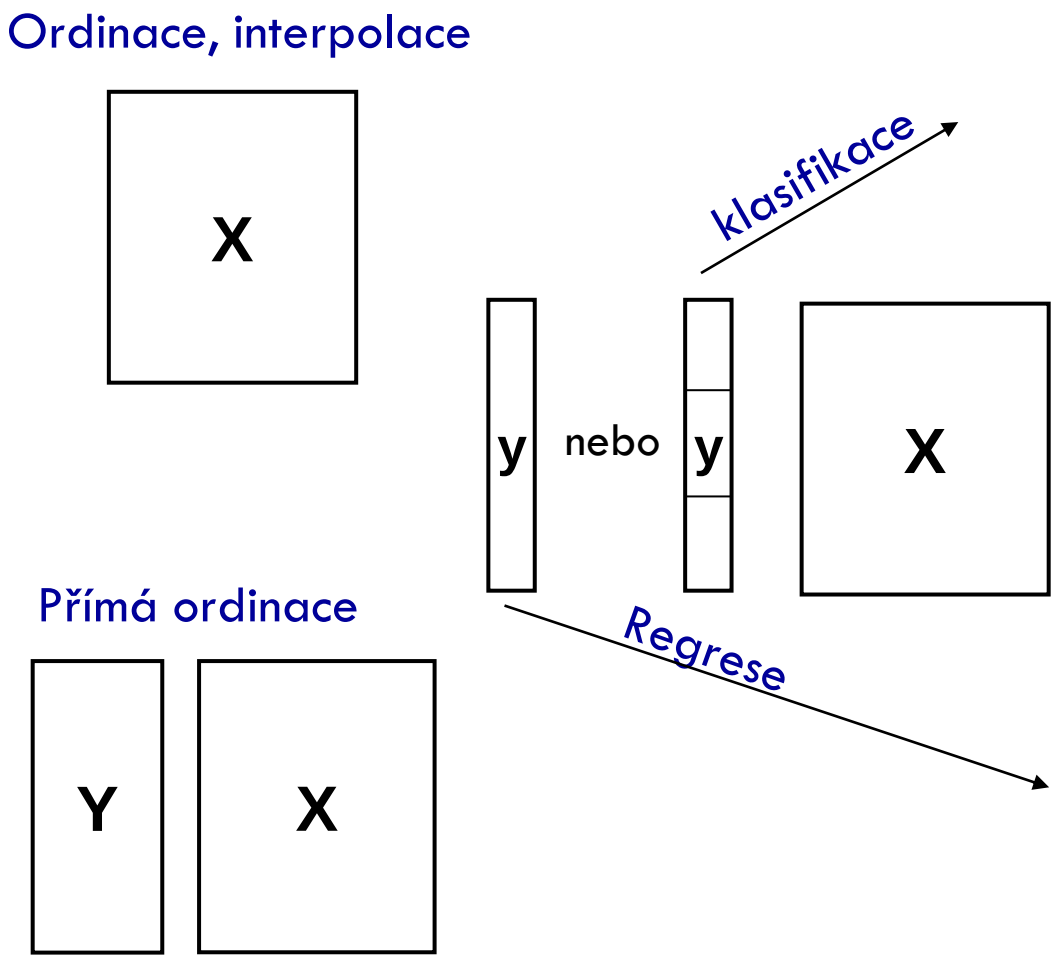
Rozdělení metod:

- *Deterministické* (MECHANICAL/EMPIRICAL MODELS) – (IDW – Inverse distance interpolation, Regression on coordinates, Splines …)
  - parametry modelu jsou voleny libovolně či empiricky
  - neodhaduje se predikční chyba
  - většinou nejsou kladeny žádné statistické předpoklady

- *Geostatistické* (STATISTICAL (PROBABILITY) MODELS) – využívají prostorovou strukturu celého pole, pro celé pole lze spočítat chybu interpolace (různé typy krigingu–obyčejný, univerzální, blokový, cokriging, Bayesian Maximum Entropy)
  - odhad parametrů v modelu objektivně-teorie pravděpodobnosti
  - odhad chyby predikce
  - statistické přepoklady

- *Metody prediktivního modelování*
  - Sada parametrických i neparametrických metod

Pokročilejší modelovací přístupy



Klasifikace

- Metody založené na stromech
- Lineární dikriminační analýza
- Neuronové sítě
- Metoda podpůrných vektorů
- Logistická regrese
- Bayesovský klasifikátor

…

Regrese

- Klasický lineání model
- Lineární zobecněné a aditivní modely
- Nelineární regrese
- Na stromech založené techniky
- Neuronové sítě
- Metoda podpůrných vektorů
- Na stromech založené techniky

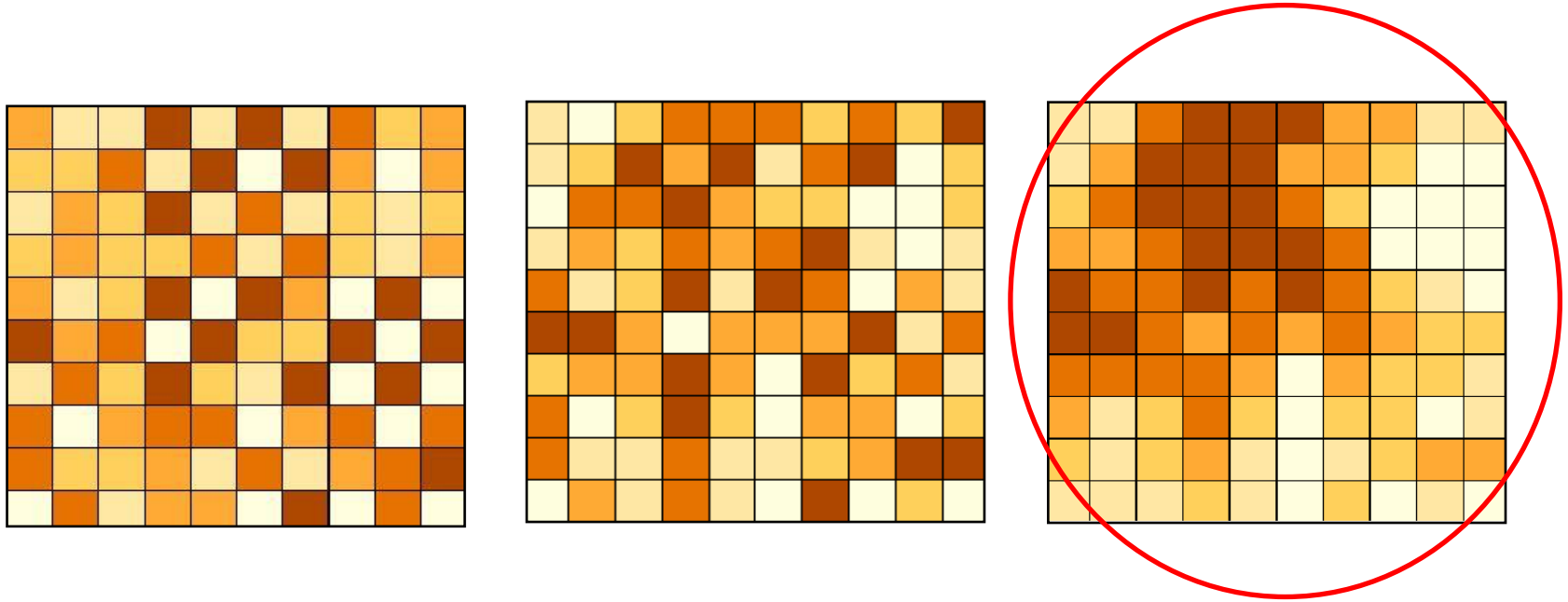
…

Prostorová autokorelace

“everything is related to everything else, but near things are more related than distant things” Waldo Tobler

Prostorová autokorelace

Negativní *Náhodná* *Pozitivní*



Prostorová autokorelace

- existence autokorelace prostorových dat je obvyklá
- způsobuje selhávání některých základních předpokladů statistické analýzy, zejména:
  - nezávislosti jednotlivých pozorování
  - nedostatku předpokladů, týkajících se chyb a reziduí v regresní analýze
- nevhodné použití klasických metod korelační a regresní analýzy u dat, která nesou prostorovou informaci
- byly vyvinuty prostorové modely a metody zohledňující autokorelaci
- řada způsobů pro testování existence prostorové autokorelace

Měření prostorové autokorelace

- existence autokorelace prostorových dat je obvyklá
- před výpočtem prostorových autokorelačních koeficientů je potřeba spočítat matici geografických vzdáleností [D_{hi}] mezi lokalitami
- autokorelační koeficienty jsou spočítány pro jednotlivé vzdálenostní třídy *d*
- váhy w_{hi} (*Kronecker deltas*) kde: w_{hi} = 1 - lokalita *h* a *i* jsou ve vzdálenosti *d*

  w_{hi} = 0 jinak
- pouze páry lokalit (*h*,*i*) ve vzdálenostní třídě *d* jsou použity pro výpočet příslušného koeficientu
- *W* je suma všech vah w_{hi} pro danou vzdálenostní třídu (počet párů použitých k vypočítání koeficientu)

Měření prostorové autokorelace

Statistické měření pro zjištění prostorové autokorelace

Moranův index (I)

$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y)(y_i - y)}{\frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2}$$

Gearyho index (*C*)

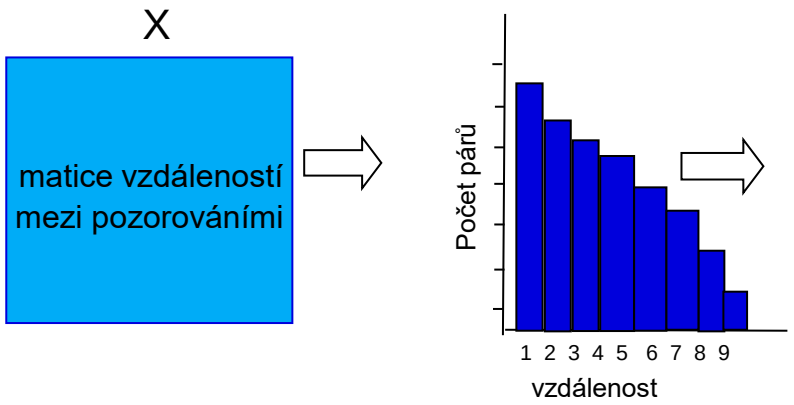
$$c(d) = \frac{\frac{1}{2W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y_i)^2}{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - y)^2}$$

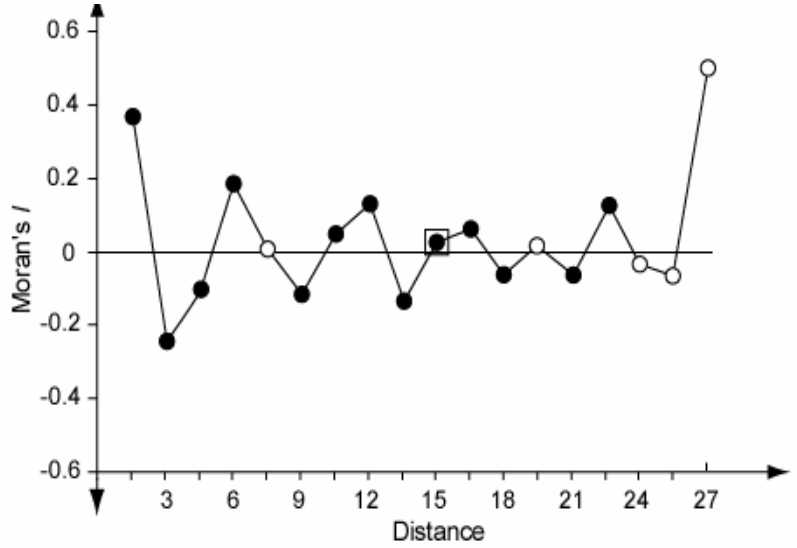
y_h a y_i jsou hodnoty pozorované na místě *h* a *i*, *w* jsou váhy a $\bar{y}$ je průměr hodnot

Moranův index – podobný Pearsonovu korelačnímu koeficientu (-1,1)
Gearyho index – vzdálenostního typu (0, > 1)

Prostorový korelogram

Prostorový korelogram – autokorelační hodnoty x vzdálenosti pozorování





Výpočet indexů pro jednotlivé vzdálenosti

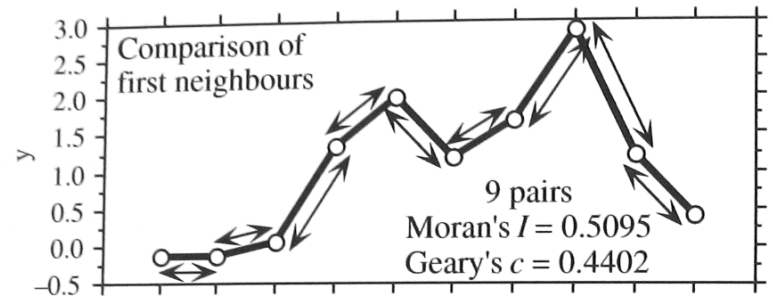
Vzdálenost

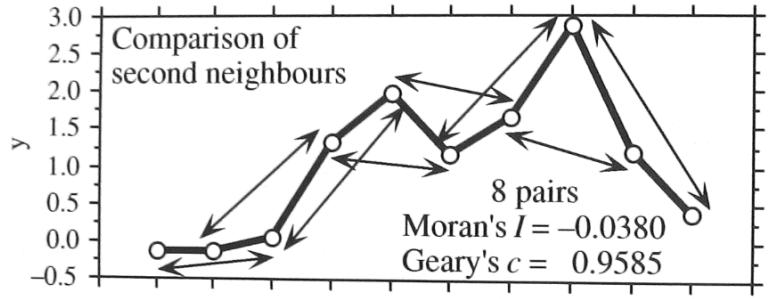
d = 1

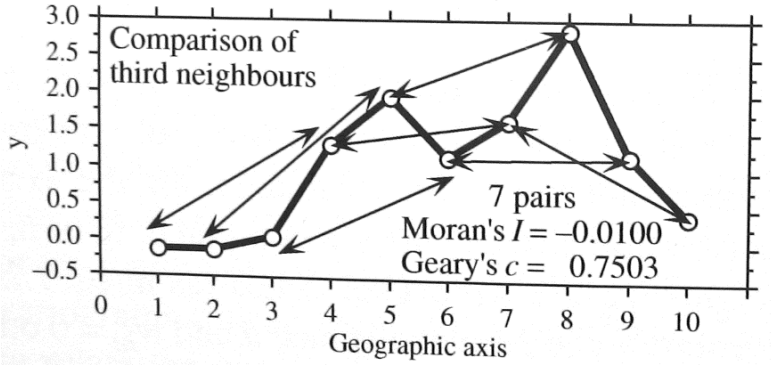
d = 2

d = 3

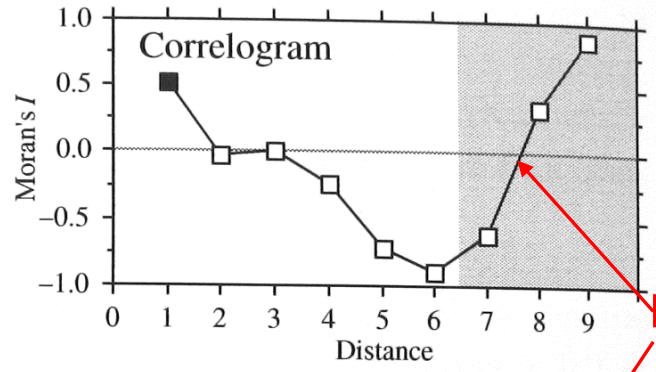
atd.



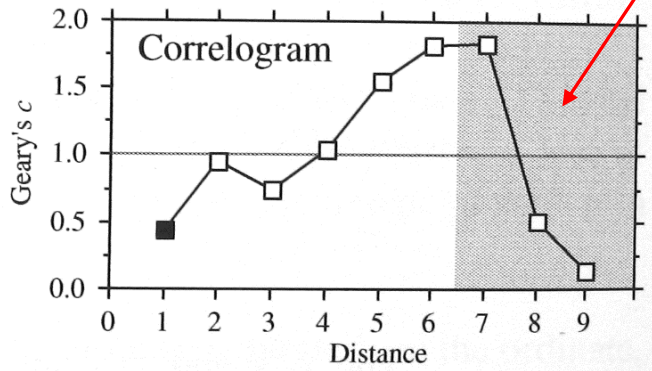


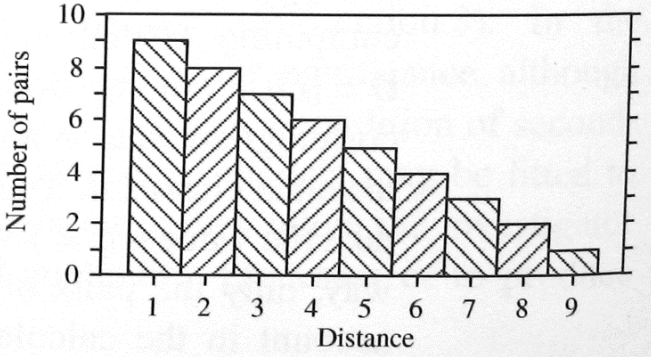


etc.



malé N!

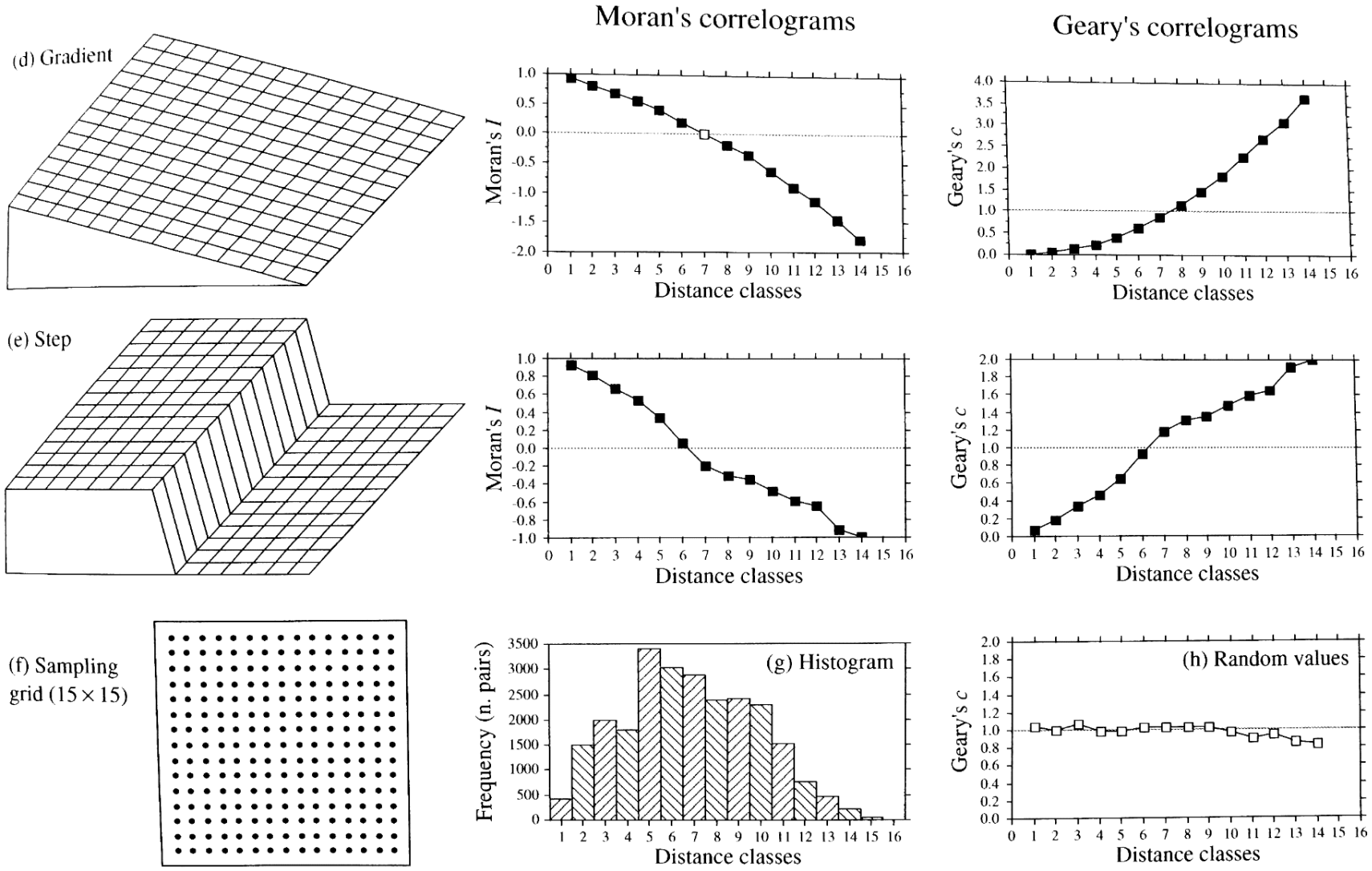




Legendre, 2003

RECETOX

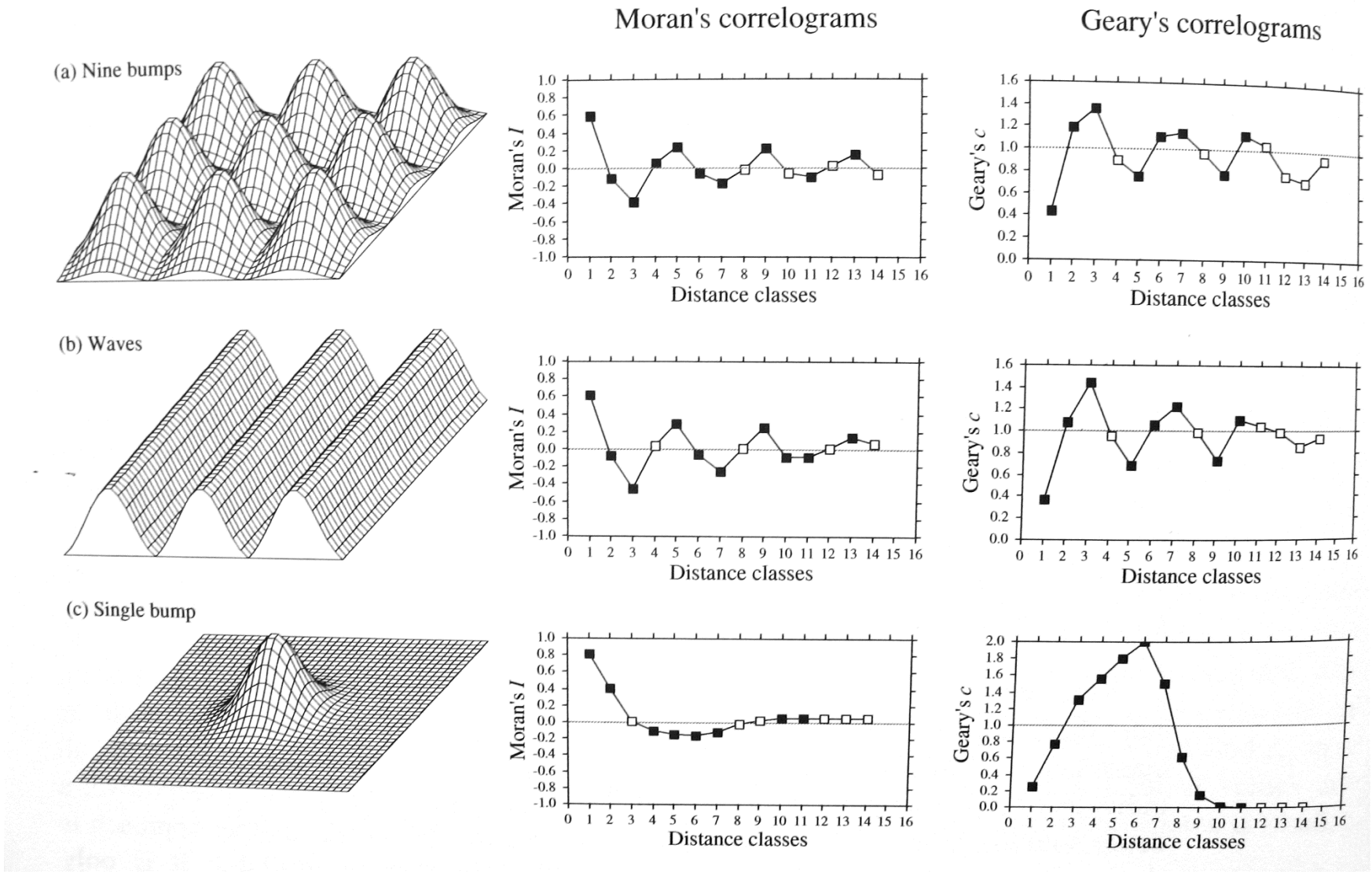
Prostorový korelogram



Legendre, 2003

RECETOX

Prostorový korelogram II



Legendre, 2003

RECETOX

Moranův index (I)-testování

- nulová hodnota znamená náhodnou prostorovou distribuci
- pro testování hypotézy se hodnoty Moranova indexu transformuhí na z-skóre (hodnoty větší než 1.96 nebo menší než -1.96 → prostorová autokorelace je významná na hladině významnosti 5%

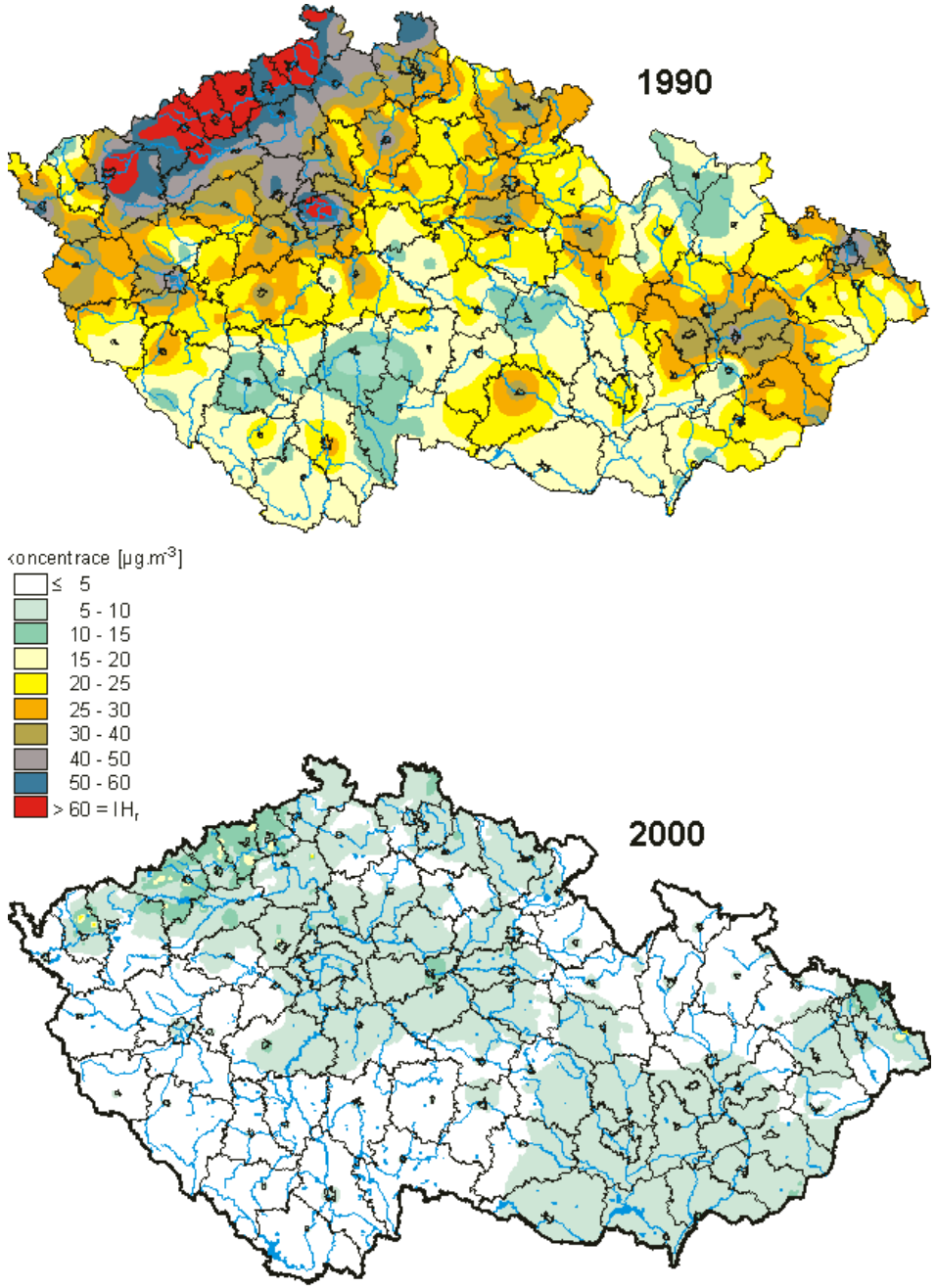
$$z = \frac{x - \bar{x}}{\sigma}$$

- *x* je skóre, které chceme standardizovat a *σ* je směrodatná odchylka

Interpolační metody
IDW, Kriging, Trendová analýza

IDW - Inverse distance weighted – inverzní vážená vzdálenost

- Nejjednodušší neparametrická technika
- Interpolační prostor (povrch) by měl být ovlivněn spíše bližšími body než vzdálenými
- Interpolační prostor je váženým průměrem rozložení bodů a váha přiřazená každému bodu se zmenšuje se vzrůstající vzdáleností od interpolovaného bodu



Obr. 2-97 Pole ročních aritmetických průměrů koncentrací, oxid siřičitý, 1990 a 2000

Příklad použití metody IDW-

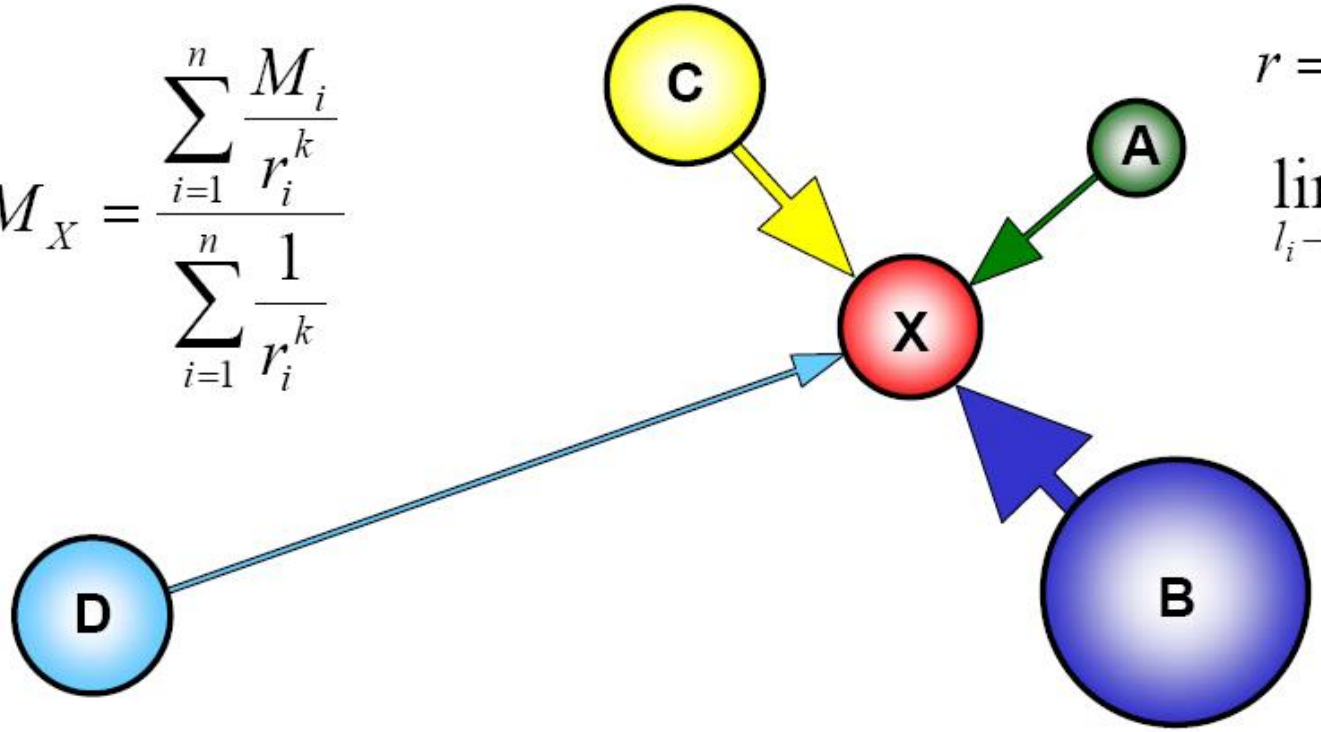
koncentrace SO_2

MUNI | RECETOX

IDW - Inverse distance weighted – inverzní vážená vzdálenost

Velikost příspěvku je přímo úměrná velikosti hodnoty a na druhé straně nepřímo úměrná vzdálenosti.

$$M_X = \frac{\sum_{i=1}^{n} \frac{M_i}{r_i^k}}{\sum_{i=1}^{n} \frac{1}{r_i^k}}$$



$$r = \sqrt{dX^2 + dY^2}$$

$$\lim_{l_i \to 0} M_X = M_i$$

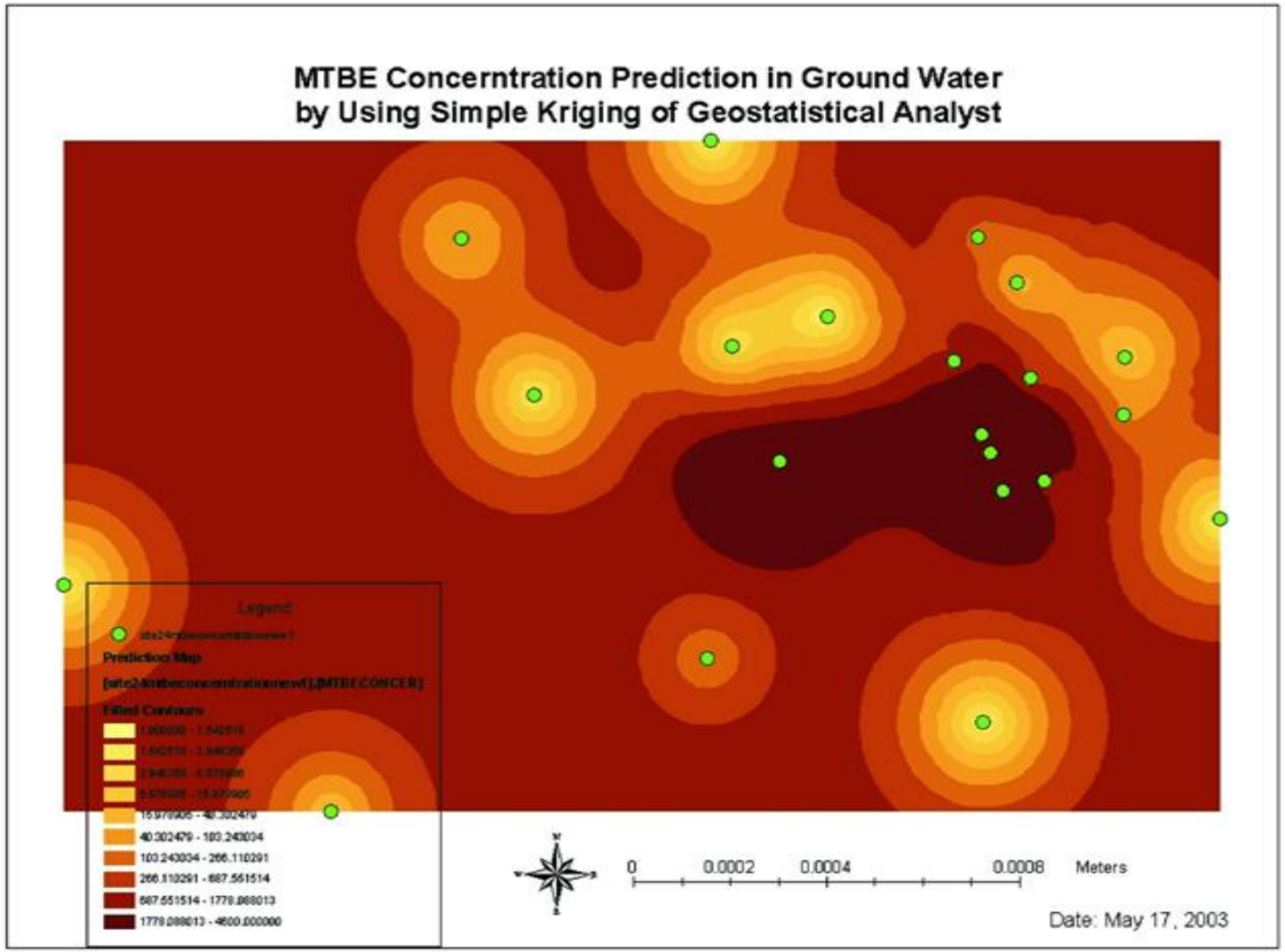
„M_i“je známá hodnota v i-tém místě, „r_i“ vzdálenost i-tého místa od místa X, „k“ je vhodná mocnina vzdálenosti (např. 1 nebo 2) a n je počet bodů.

Kriging

Francouzský matematik Georges Matheron odvodil matematický popis krigingu na základě práce důlního inženýra Daniela Gerharduse Kriga, po němž tuto metodu také roku 1962 nazval

Daniel Gerhardus Krige
26 August 1919

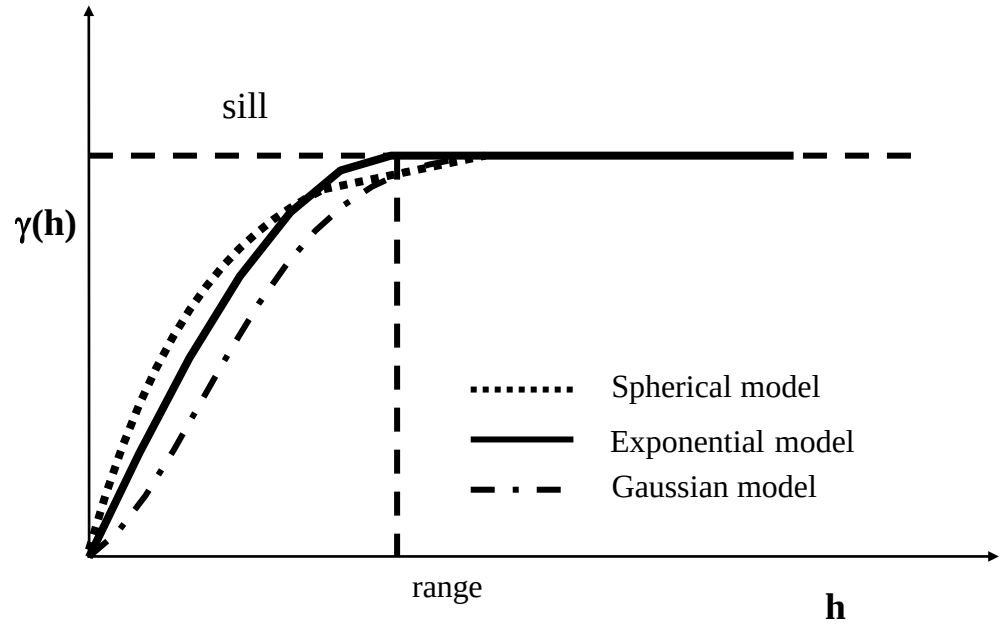
- při hledání zlatých dolů v jižní Africe!



MUNI | RECETOX

Kriging

- Sofistikovanější IDW – jak odhadnout váhy jednotlivých bodů?
  - odhadnout váhy které odrážejí skutečnou prostorovou autokorelační strukturu
  - Semivariance – rozdíly mezi nejbližšími body → teoretický variogram



Variogram

- sumarizuje sílu asociace mezi pozorováními jako funkci vzdálenosti
- experimentální variogram je graf, který ukazuje jak se ½ mocninného rozdílu mezi dvěma hodnotami (semivariance) mění se vzdáleností mezi pozorováními
- očekáváme menší semivarianci v menších vzdálenostech a stabilní semivarianci mezi hodně vzdálenými pozorováními

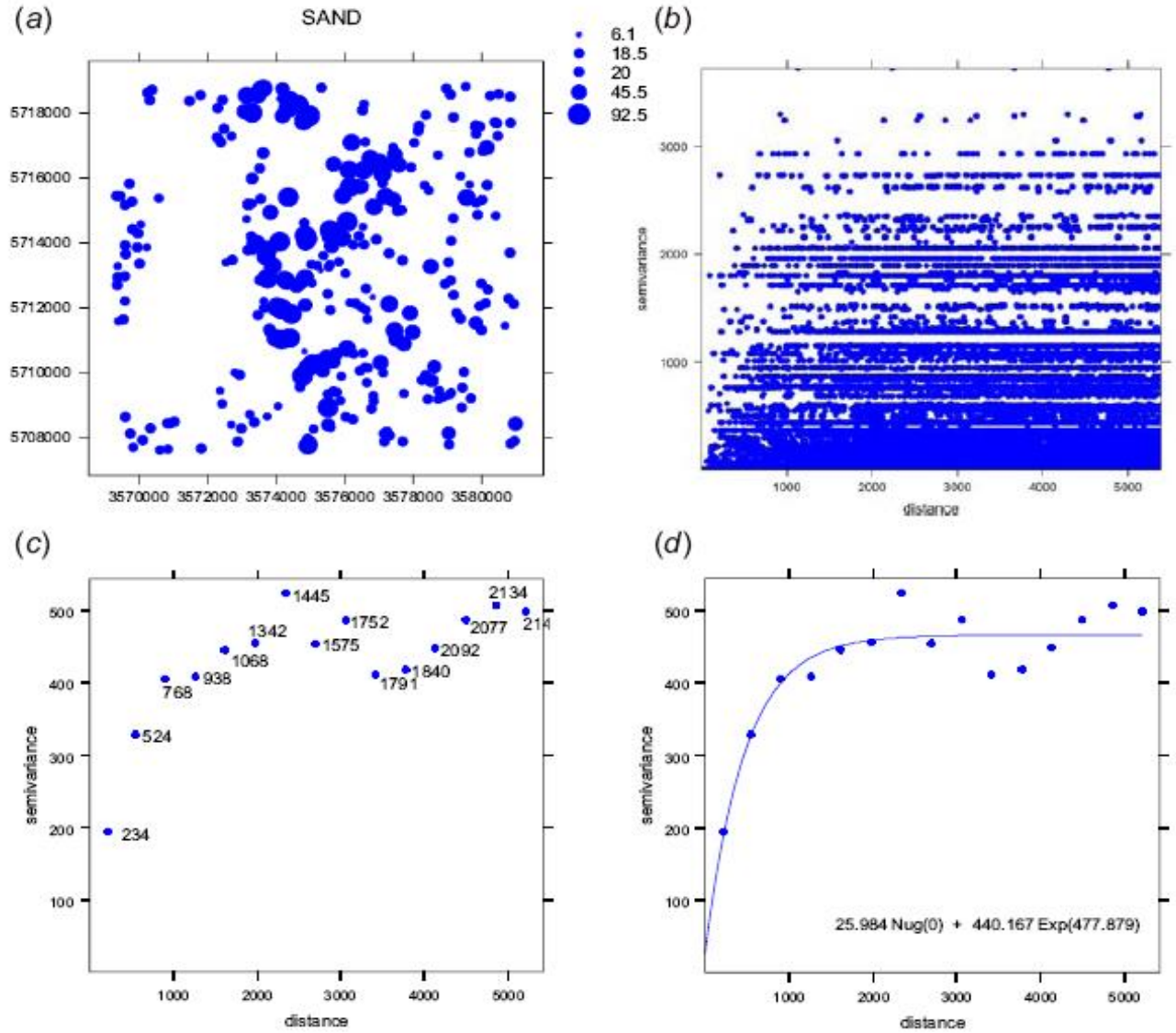
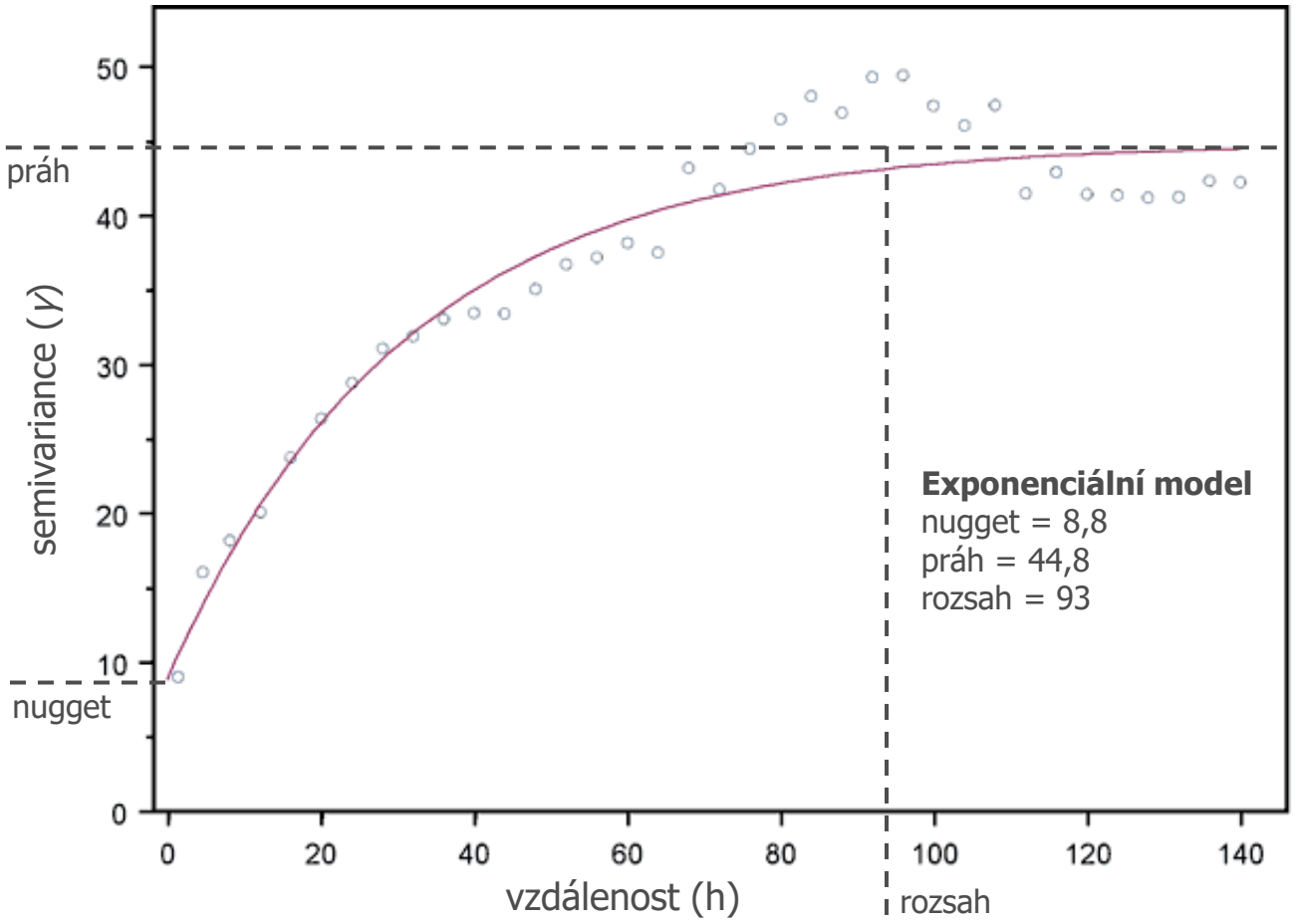


Fig. 1.7: Steps of variogram modelling: (a) location of points (300), (b) variogram cloud showing semivariances for 44850 pairs, (c) semivariances aggregated to lags of about 300 m, and (d) the final variogram model fitted using the default settings in gstat.

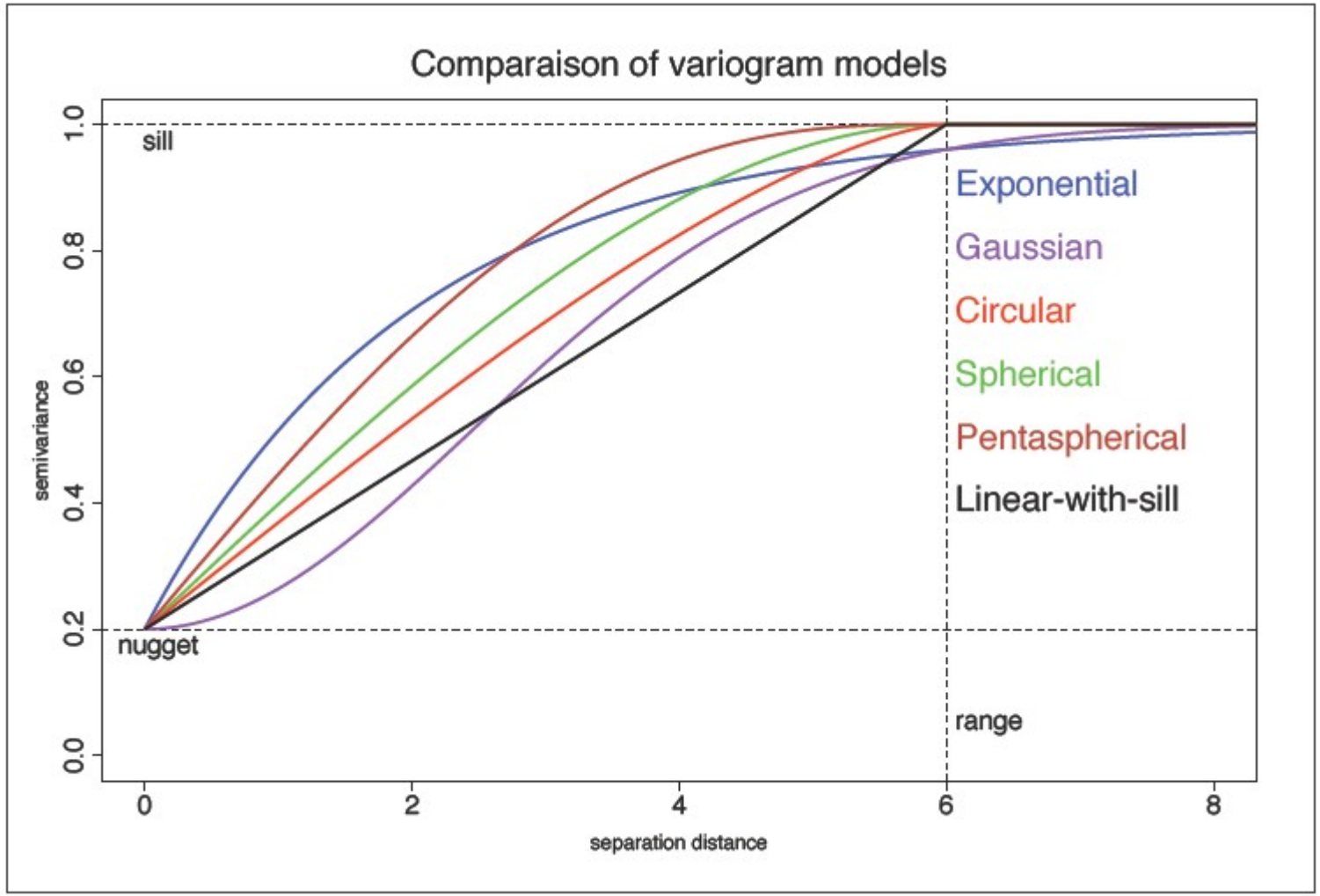
T. Hengl (2007) A Practical Guide to Geostatistical Mapping of Environmental Variables

MUNI | RECETOX

Exponenciální model semivariogramu



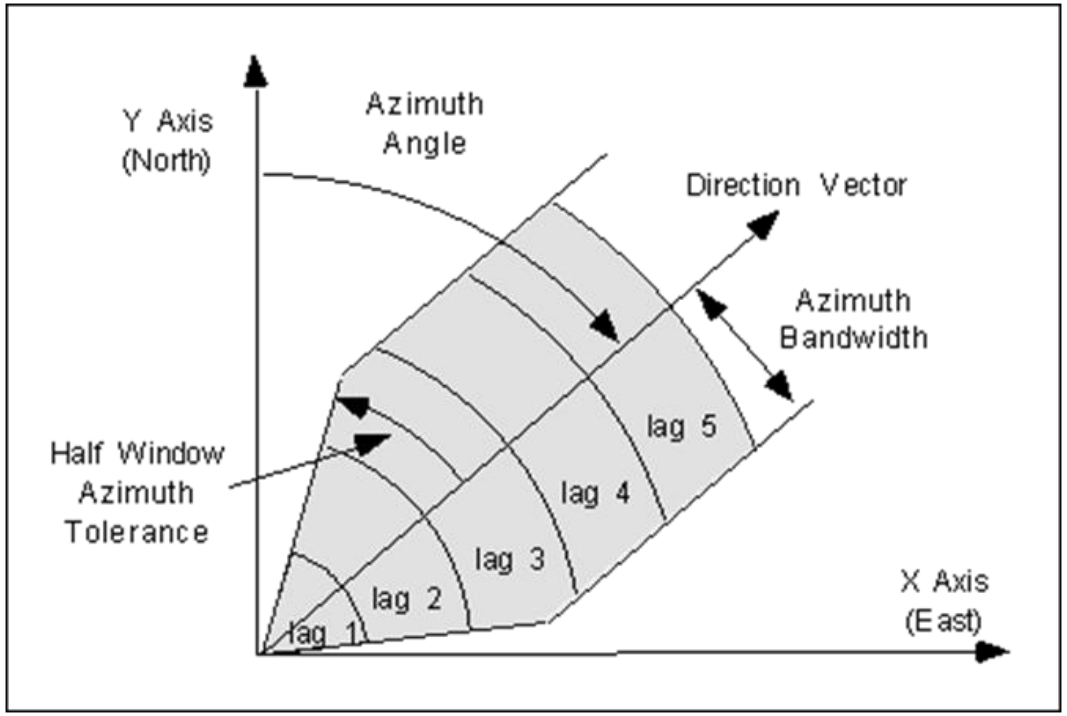
Modely variogramu



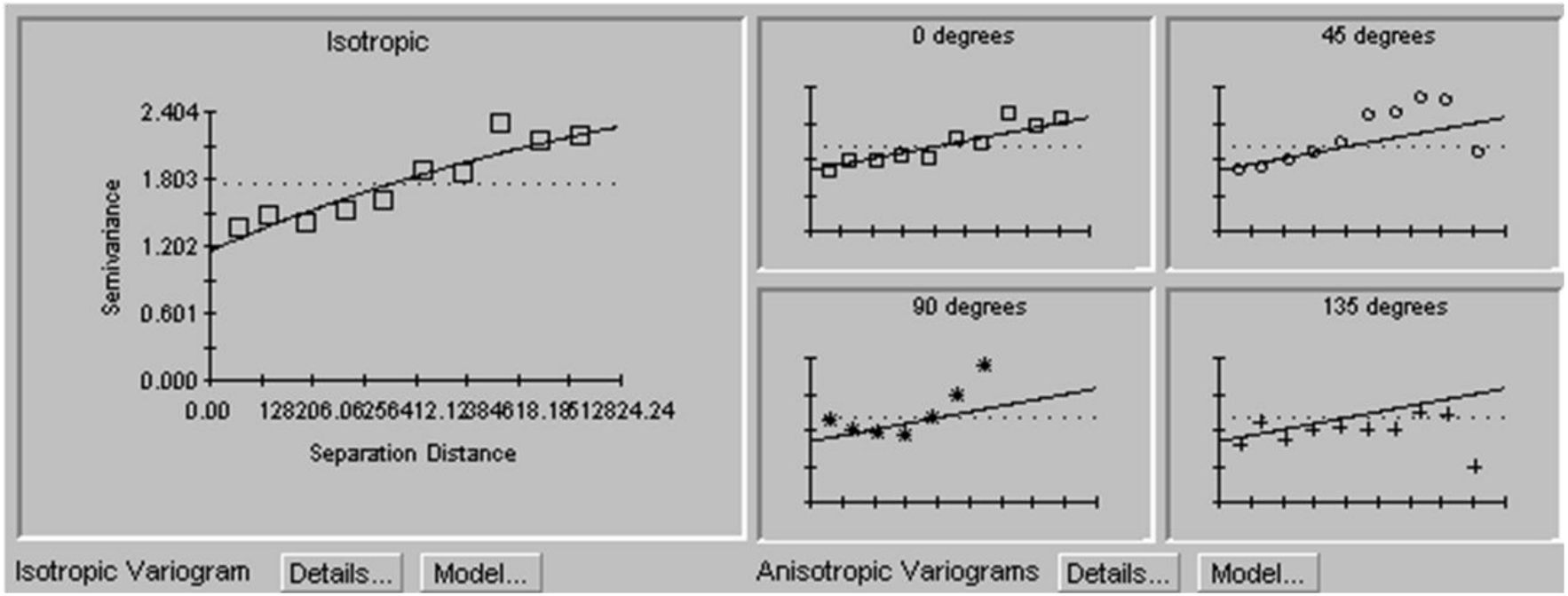
Isotropní x anisotropní variogram

Isotropní – v každém směru stejný variogram

Anisotropní – v různém směru různý variogram



Isotropní x anisotropní variogram



Trend surface analysis – Trendová analýza

- metoda pro vytváření vyhlazených (*smoothed*) map
- odhady proměnných v daných lokalitách jsou získány regresním modelem kalibrované přes celou studovanou plochu
- Vyjádříme proměnnou **y** (odpověď) jako nelineární funkci geografických souřadnic X a Y jednotlivých ploch, kde byly proměnné sledovány
- trend surface analysis je aplikace polynomiální regerese k prostorově uspořádaným datům
- Postup: vycentrujeme (na průměr) ***y***, *Y*, *X* (intercept = 0); vybereme stupeň polynomu; vyřadíme nesignifikantní členy (*backward elimination*), dokud všechny členy polynomiální rovnice nebudou signifikantní; vypočítáme nové odhady ***y***

Model jednoduché lineární regrese -opakování

$$Y = \alpha + \beta X + \varepsilon$$

Náhodná variabilita

Intercept

Sklon, též regresní koeficient
Slope

Závisle proměnná
Odpověď
Dependent v., response

Nezávisle proměnná, prediktor,
Independent v.

Polynomiální regrese

- polynomiální regrese - libovolnou funkci lze nahradit (v omezeném rozsahu hodnot prediktoru) polynomem

- mám představu (třeba z nějaké teorie), jak má závislost vypadat, a věřím, že residuály budou náhodně kolem predikované hodnoty

- tradiční názvy kvadratická regrese, kubická regrese

Polynomiální regrese

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_m X^m + \varepsilon$$

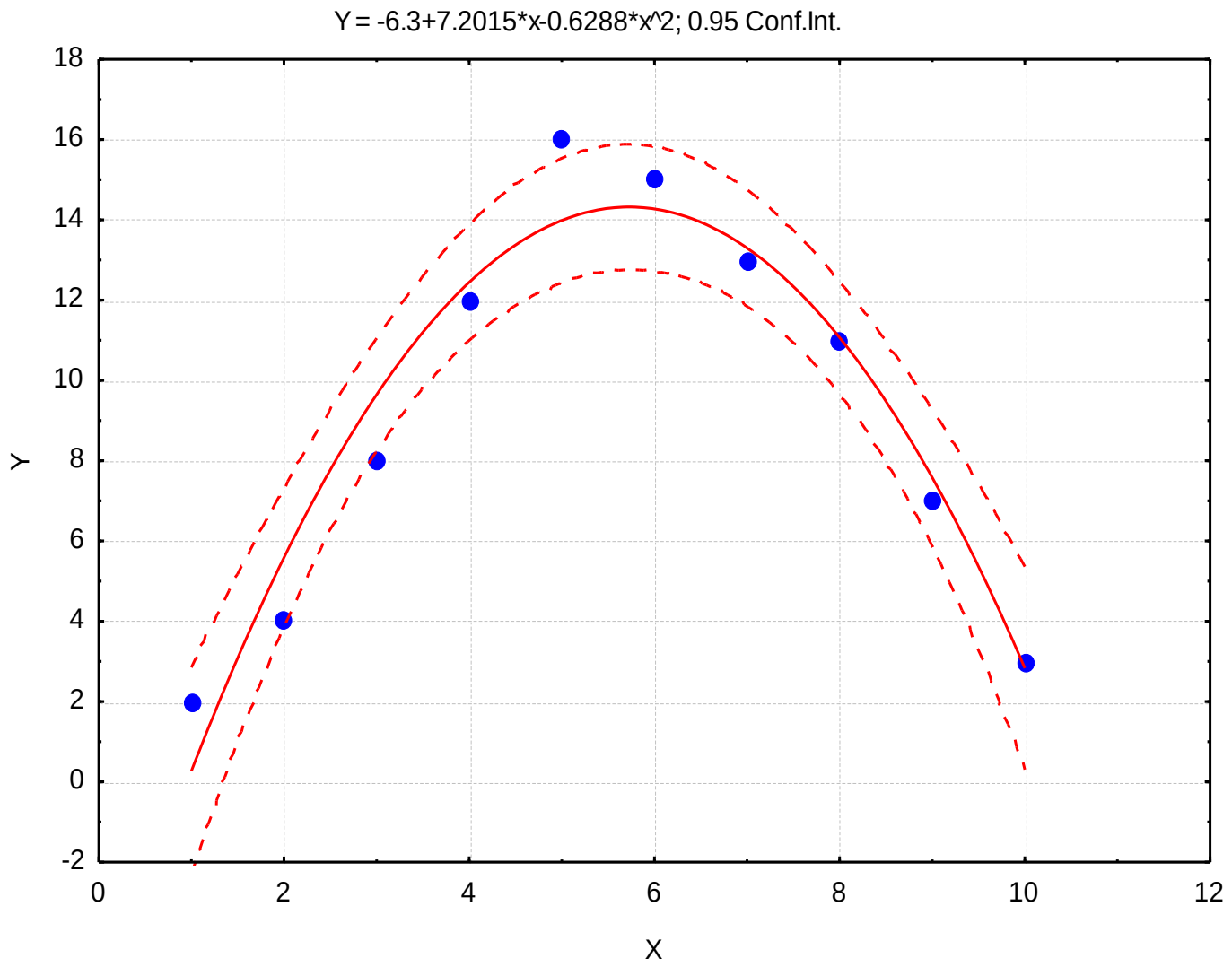
- mnohonásobná lineární regrese, kde prediktory jsou X, X^2, X^3 atd. se počítá stejně (tj. opět kriterium nejmenšího součtu residuálních čtverců, které má opět (normálně) jedno minimum).
- do modelu jsou přidávány pouze proměnné, které snižují residuální chybu modelu:

  dopředný výběr (*forward elimination*) – začínáme s konstantou (interceptem) a postupně se přidávají jednotlivé členy

  zpětný výběr (*backward elimination*) – začínáme se všemi členy, postupně se odebírají ty, které přispívají k nejmenšímu snížení residuální chyby

- obdobný význam má i R^2

Polynomiální regrese



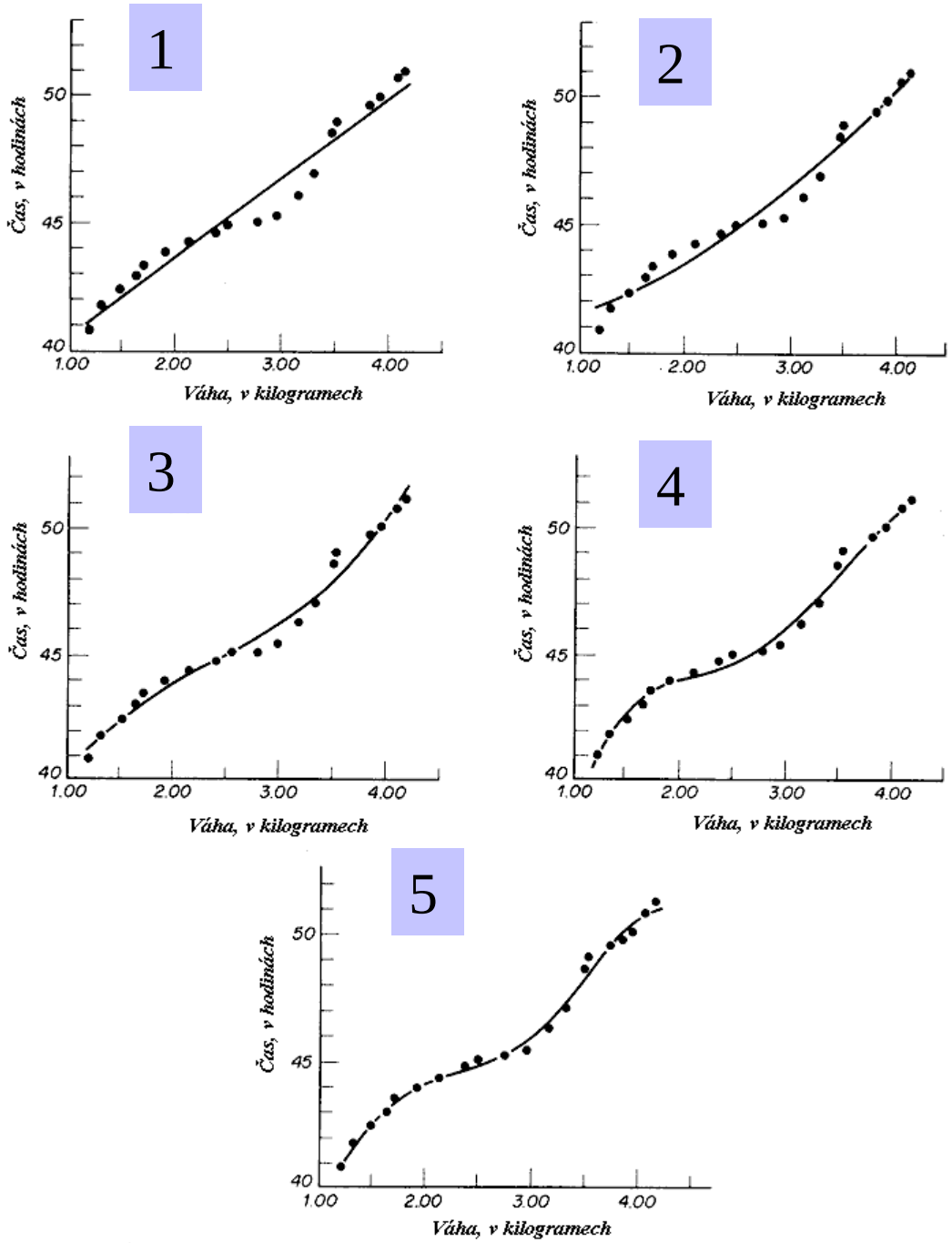
kvadratická regrese může být vysoce průkazná, i když lineární regrese průkazná není

průkaznost kvadratického členu můžeme chápat jako důkaz nelinearity vztahu

MUNI | RECETOX

Se zvyšujícím se stupněm

polynomu stoupá “flexibilita”



Pozor! Zvyšující se složitost nemusí znamenat lepší predikční schopnost

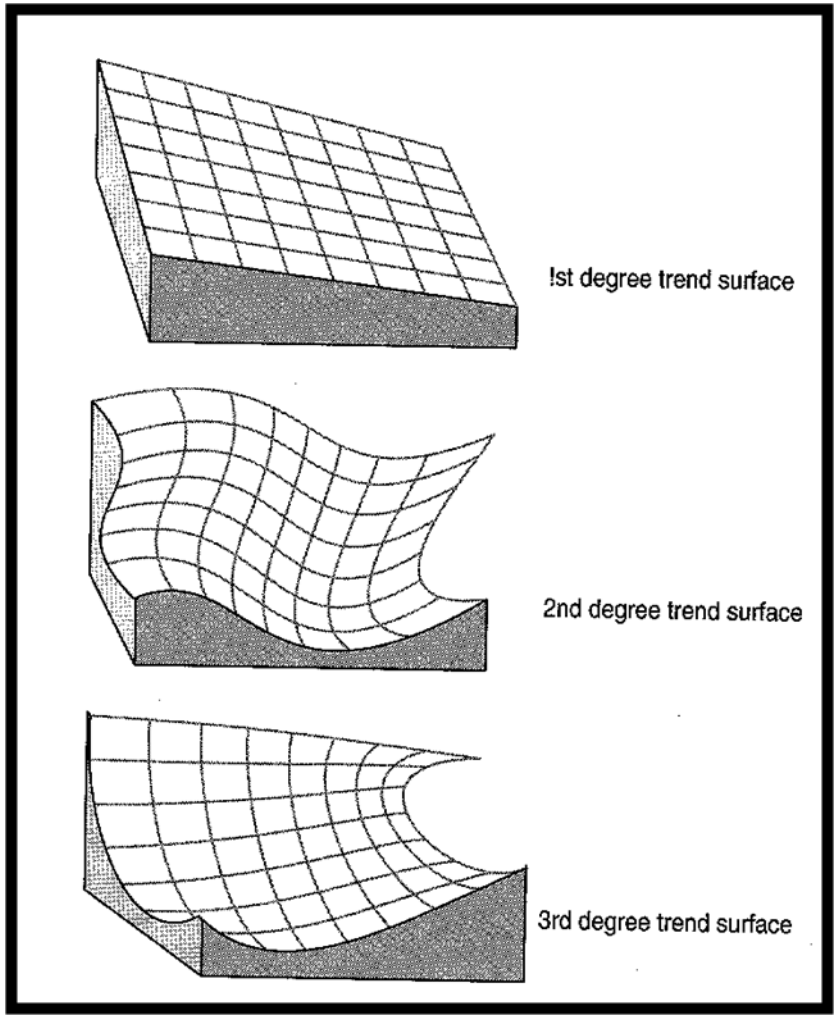
RECETOX

Zpět k trendové analýze...

- většinou polynom max. 3. stupně
- zkoumáme závislost proměnné na prostorové struktuře
- máme představu (z teorie), jak má závislost vypadat
- proměnnou můžeme rozdělit na dvě komponenty – trend a odchylky od trendu (residua)
  - trend je celkový (globální) „*pattern*“ (lineární –klesající, stoupající; kvadratický, kubický)
  - residua reprezentují lokální „*pattern*“

$$y = a + \beta_0 x + \beta_1 y + \beta_2 x^2 + \beta_3 xy + \beta_4 y^2$$

Globální trend

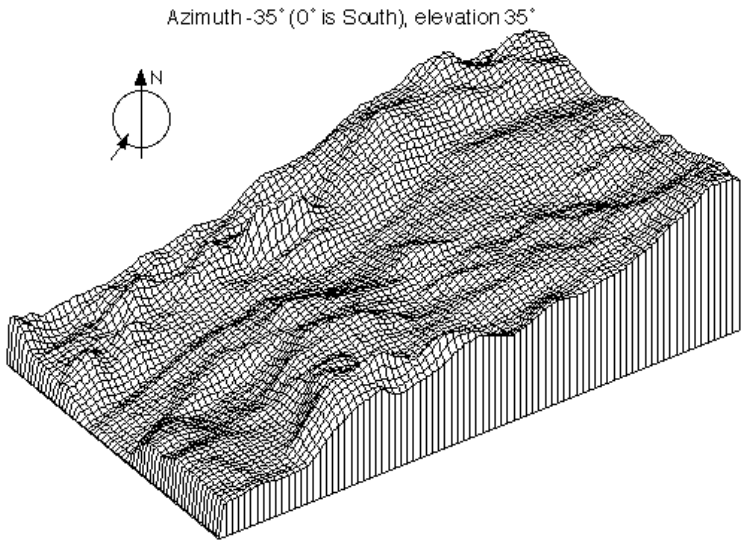


Lineární

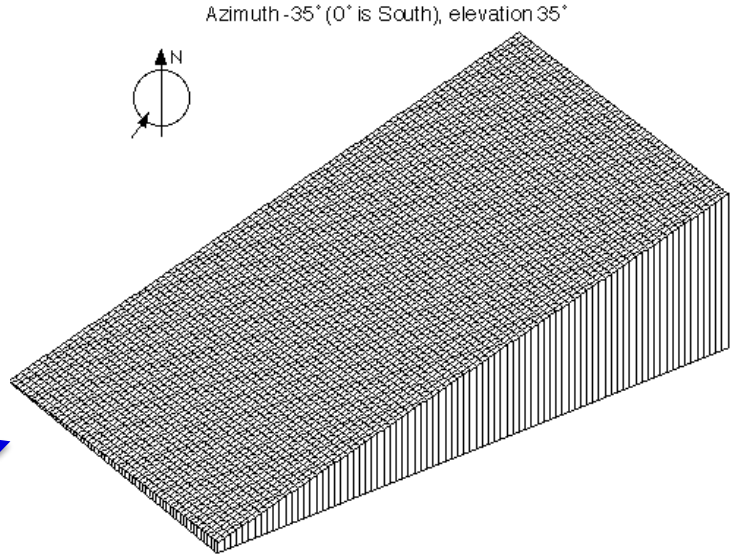
Kvadratický

Kubický

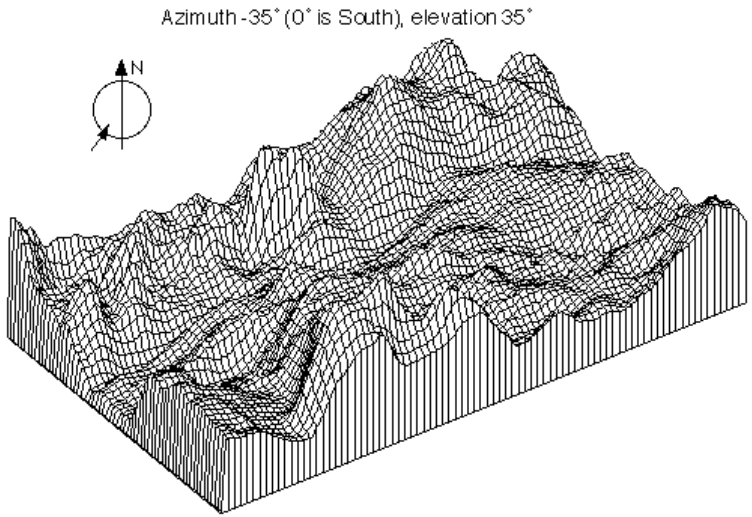
Příklad



Globální gradient + lokální změny



1. stupeň polynomu → gradient od východu na západ

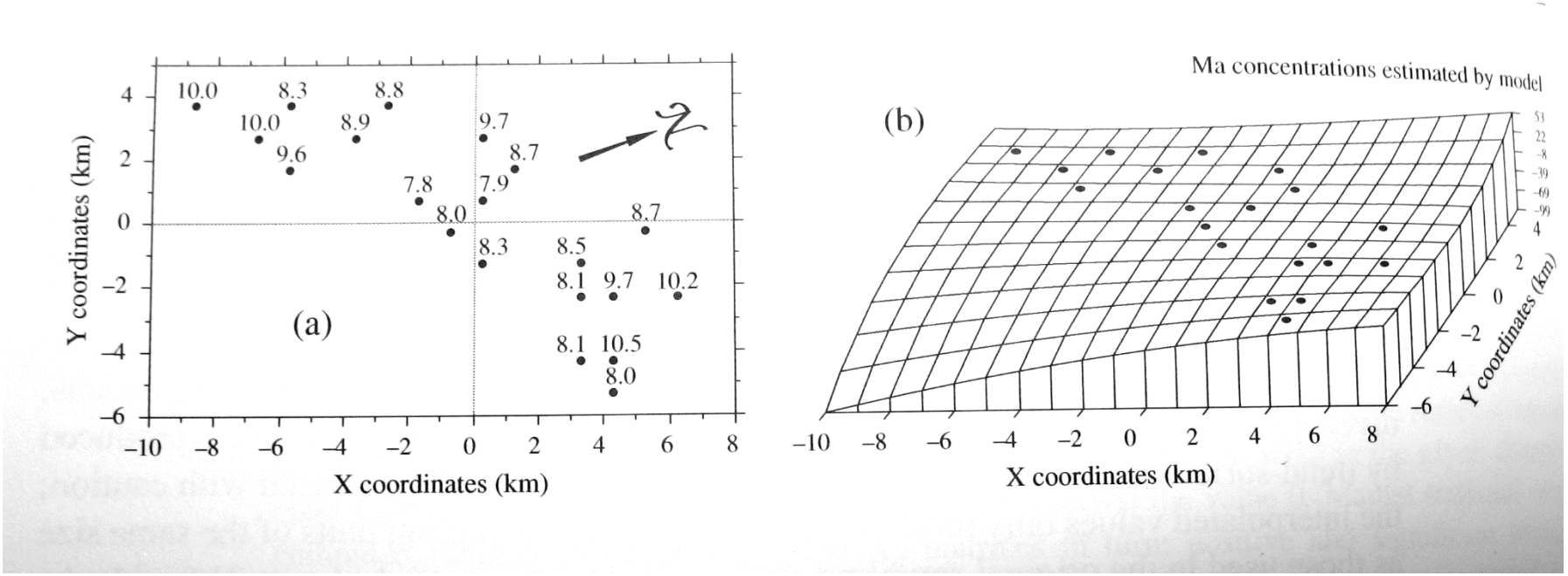


residua → lokální změny

http://www.kgs.ku.edu

Příklad – koncentrace aerobních bakterií

20 vzorkovacích míst



Legendre, 2003

Příklad – koncentrace aerobních bakterií II

- Začínáme s rovnicí 3. řádu
- Rovnice 1. řádu (X, Y, $X*Y$) $R^2 = 0.02$ ($p = 0.52$) - není významný lineární trend
- Rovnice 2. řádu (X^2, Y^2,…) $R^2 = 0.39$ ($p = 0.21$) – stále nevýznamný trend
- Rovnice 3. řádu (X^3, Y^3,…) $R^2 = 0.87$ pro všechny členy- významný trend– některé členy můžeme odstranit – zpětné odstranění
- Finální rovnice: $y = 8.13 – 0.16XY - 0.09Y^2 + 0.04X^2Y + 0.14XY^2 + 0.10Y^3$ ($R^2 = 0.81$, $p = 0.0001$)
- Používáme pouze je-li viditelná jednoduchá závislost!

Shrnutí

- tři techniky pro prostorovou interpolaci:
  - IDW – nejjednodušší, vhodný pro velký počet bodů k „vyhlazení“ plochy → váženo pouze vzdáleností
  - Kriging – několik druhů; není potřeba pravidelné vzorkování; váhy odrážejí prostorovou strukturu → semivariogram – pozor na stat. předpoklady!
  - Trend surface analysis – využívá polynomiální regrese; k odhadu prostorové závislosti využívá souřadnice; pozor na stat. předpoklady!
- Tyto metody se v environm. vědách používají nejčastěji → existují další interpolační metody- někdy příště ☺

- Prostorovou distribuci můžeme předem otestovat pomocí Moranova korelačního indexu (I) a Gearyho vzdálenostního indexu (C); v ArcGIS dostupný pouze Moranův
  - Distribuce: náhodná, shluková, negativní

Časová řada

Co je to časová řada

řada hodnot věcně a prostorově vymezeného ukazatele, která je uspořádaná v čase

$y_t = f(t)$ $\quad\quad$ $t = 1, 2, \ldots, n$

Např. Teplota vody měřená každý den ve stanovenou dobu

Dekompozice časové řady

- Každá časová řada obsahuje tři základní komponenty, které je potřeba odlišit a identifikovat (tzv. dekompozice časové řady):
- Trend (deterministický/stochastický)
- Periodická složka - Sezónní vlivy (seasonals)
- Šum (noise)
- Metody: vyhlazování, regrese, ARMA, ARIMA,sezónní rozklad (...)

Typologie časových řad

- možnost predikce
  - stochastické (*obsahují prvek náhody*)
  - deterministické (*lze přesně předpovědět vývoj*)
- interval sledování
  - krátkodobé *(kratší než 1 rok – měsíční, kvartální)*
  - dlouhodobé *(standardně roční)*
- podle sledované veličiny
  - intervalové *(popisují tokovou veličinu)*
  - okamžikové *(popisují stavovou veličinu)*
  - absolutní *(původní získané hodnoty)*
  - odvozená *(transformované hodnoty, např. indexy)*

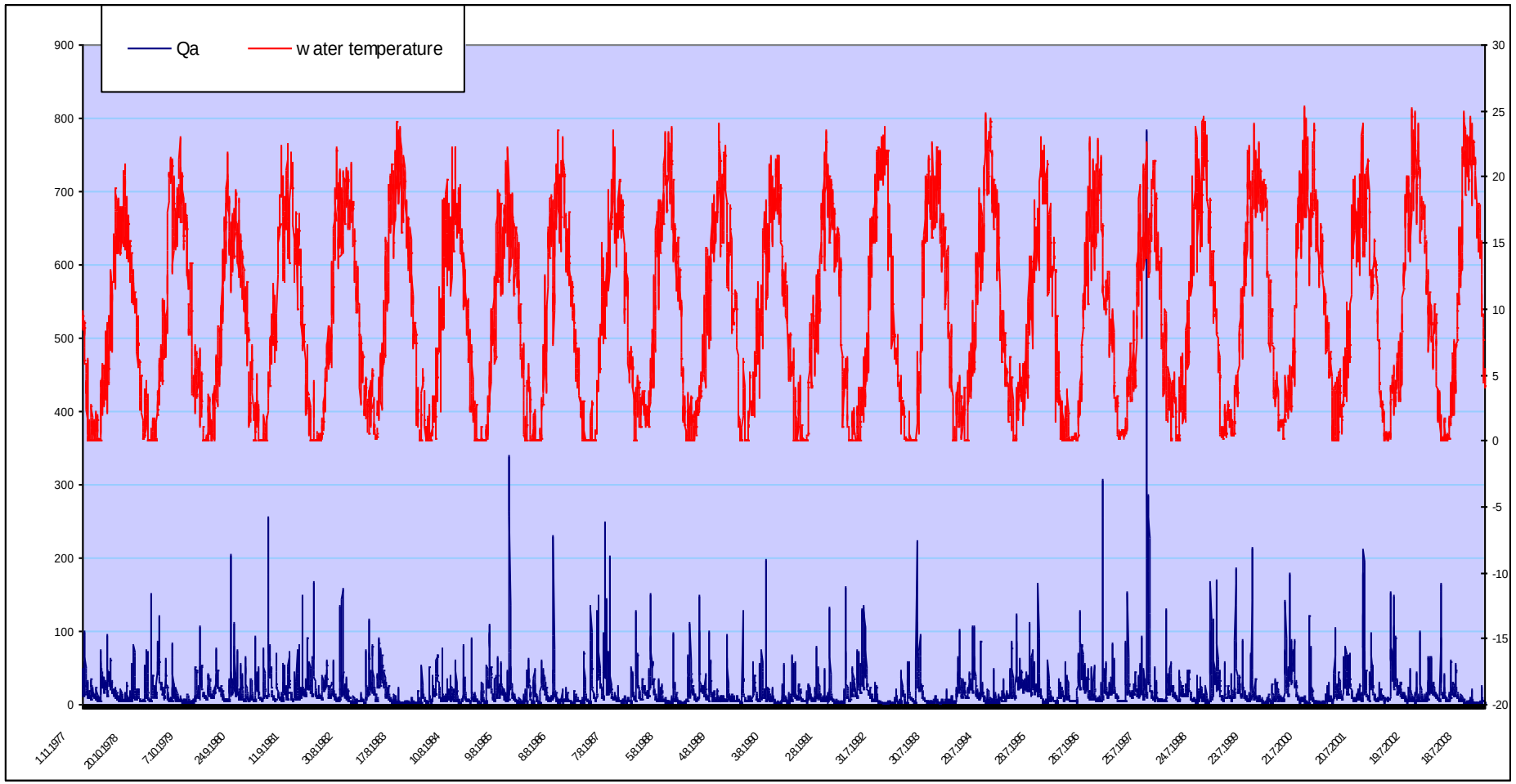
Problémy při analýze časových řad

- **problémy s délkou řady a volbou intervalu pozorování**
  - krátký interval (dlouhá řada) vede k zbytečné redundanci informace
  - dlouhý interval (krátká řada) znamená riziko ztráty informace
- **problémy s kalendářem**
  - různé délky let a měsíců (standardní měsíc 30 nebo 365/12)
  - různé počty pracovních dní v měsíci (vyrovnání)
  - pohyblivé svátky (Velikonoce)
- Stacionární x nestacionární časové řady

Grafy časových řad

grafické vyjádření časové řady:

spojnicový graf



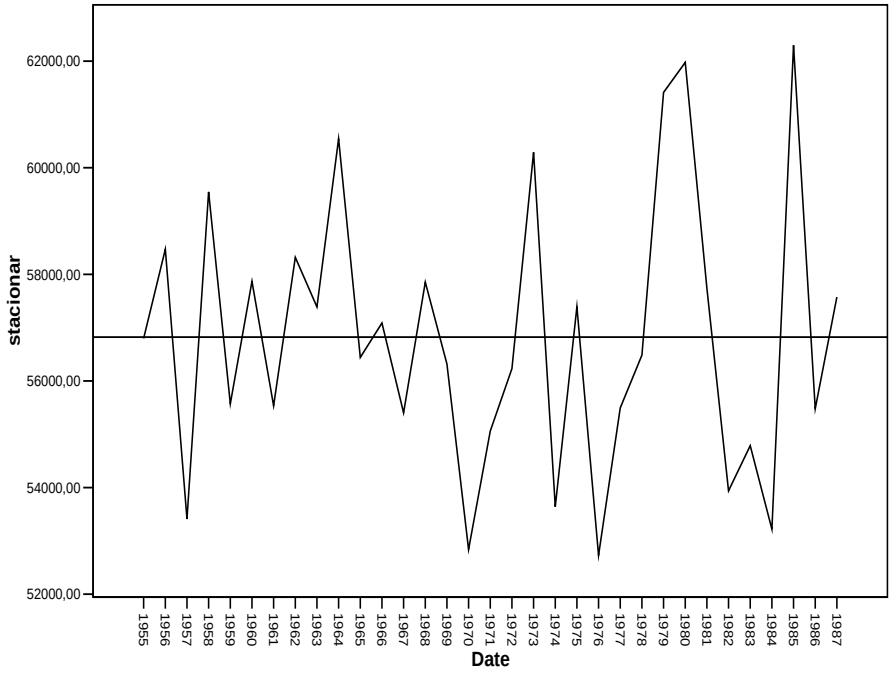
CETOX

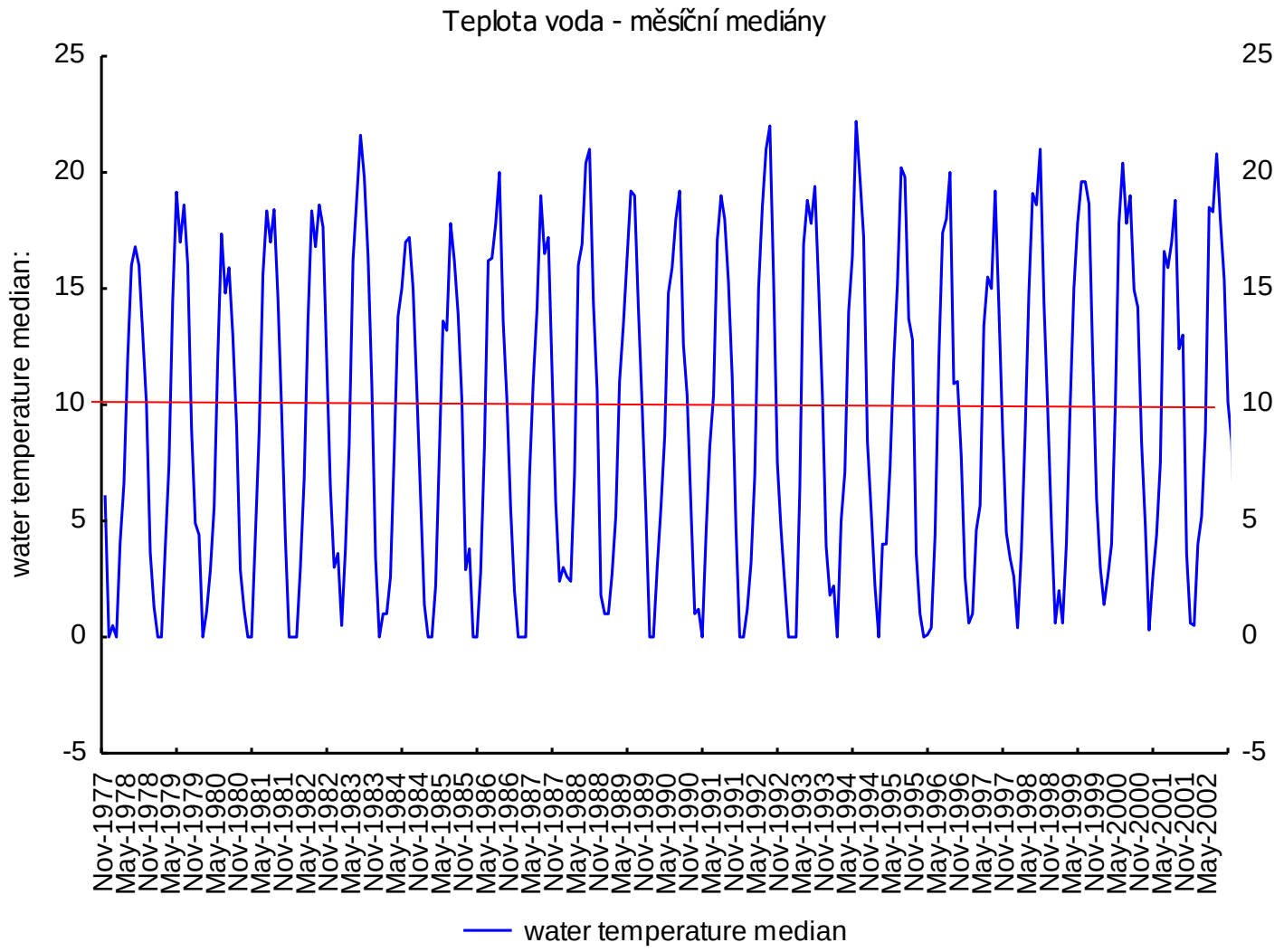
Stacionární řada

- Časovou řadu považujeme za stacionární, pokud splňuje následující podmínky:
  - má konstantní průměr
  - má konstantní variabilitu
- Stacionarita je jednou z nutných podmínek řady metod analýzy časové řady
- Stacionarity lze docílit transformací na řadu diferencí či odečtením trendu

Stacionární časová řada

řada bez zjevného trendu → hodnoty kolísají kolem konstanty

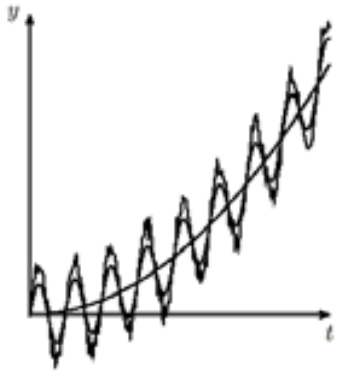




MUNI | RECETOX

Základy analýzy časových řad

Hlavní cíle analýzy časových řad



1. odhalení zákonitostí a příčin dosavadního **vývoje**
2. **prognóza** chování časových řad

Každá řada může obsahovat čtyři základní složky:

- *trend (Tt)*
- *periodická (sezónní) složka (St)*
- *cyklická složka (Ct)*
- *náhodná složka (εt)*

První tři složky tvoří systematickou část řady.

Trendová složka časové řady

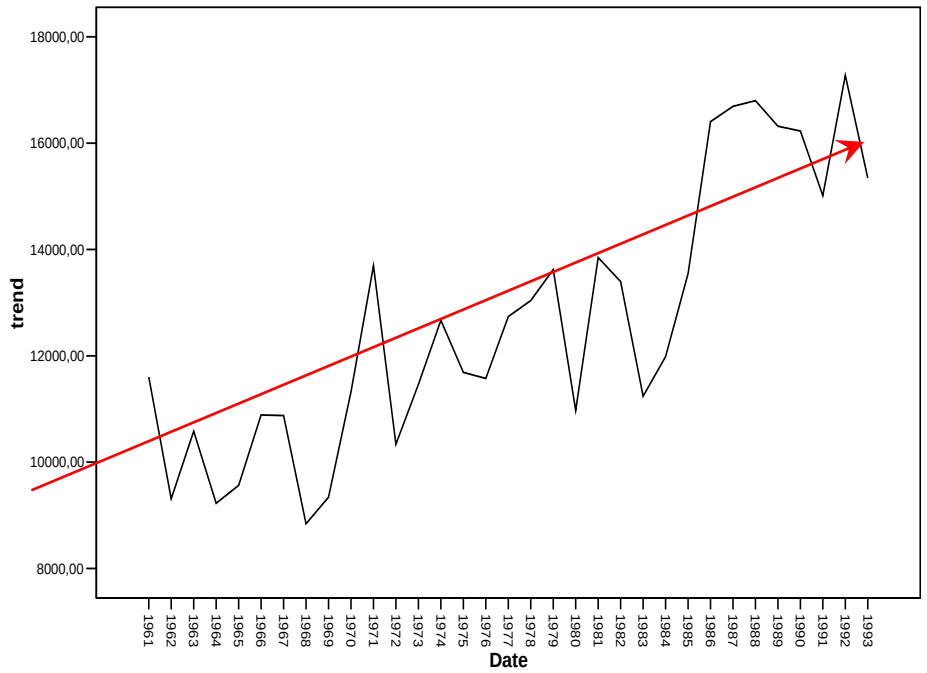
Trend je obecná tendence vývoje zkoumaného jevu za dlouhé období.

- je výsledkem dlouhodobých a stálých procesů (v měřítku posuzované délky časové řady)
- trend může být lineární či nelineární
- trend může být rostoucí, klesající nebo může existovat řada bez trendu

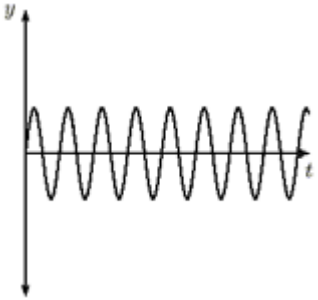
Časové řady bez trendu se označují jako stacionární.

Časová řada s trendem

řada má rostoucí nebo klesající trend



Periodická složka časové řady

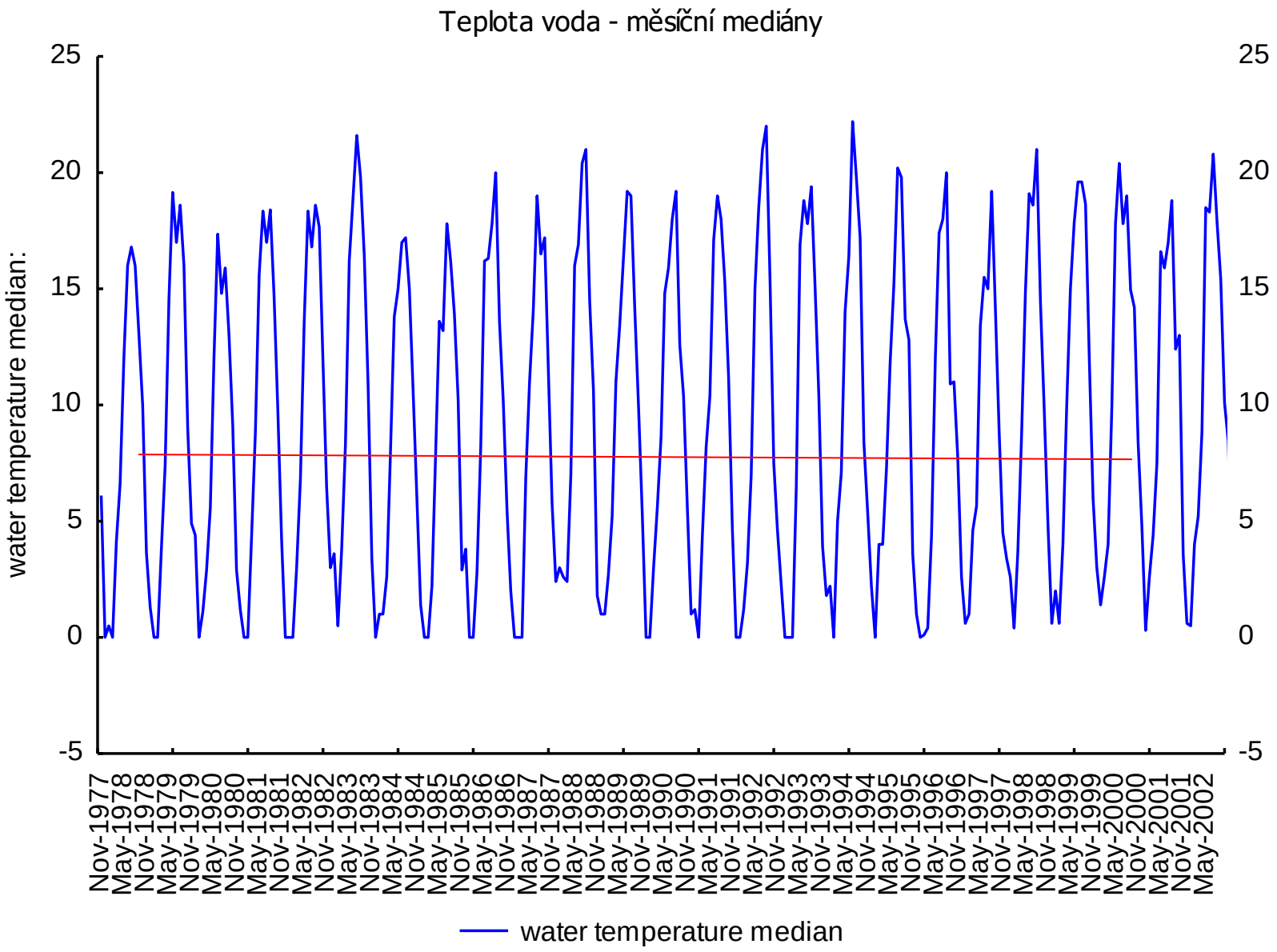


Periodická složka je pravidelně se opakující odchylka od trendové složky s pevnou délkou **periody T**

- perioda této složky je menší než celková velikost sledovaného období

- typickým případem jsou **sezónní kolísání** a nebo řady denních, měsíčních, čtvrtletních ukazatelů

- příčiny sezónnosti jsou různé, většinou však dobře definovatelné



MUNI | RECETOX

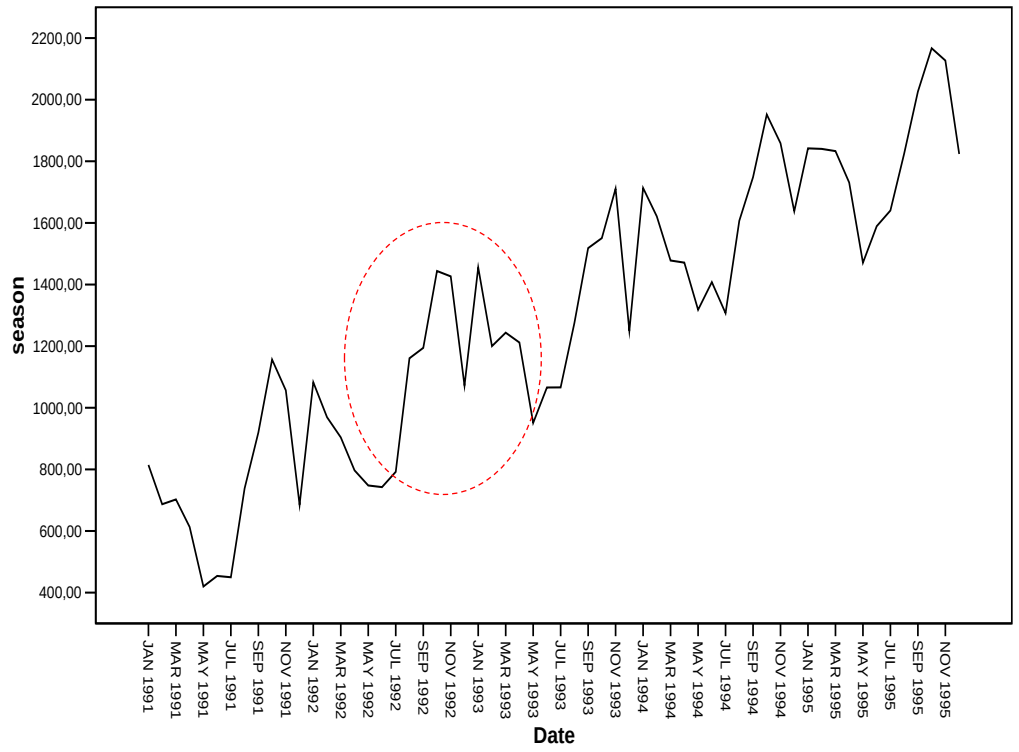
Cyklická složka

Cyklická složka udává kolísání okolo trendu v důsledku dlouhodobého cyklického vývoje

- cyklická složka může vykazovat změny v délce a amplitudě cyklu
- délka cyklu je tedy většinou neznámá (př. demografický trend, kolísání teploty vzduchu)
- délka cyklu je delší než 1 rok, v některých případech se označuje jako „střednědobý trend“
- bývá typickou součástí časových řad meteorologických prvků (př. problém globálního oteplování) či hydrologických jevů

Časová řada se sezónností

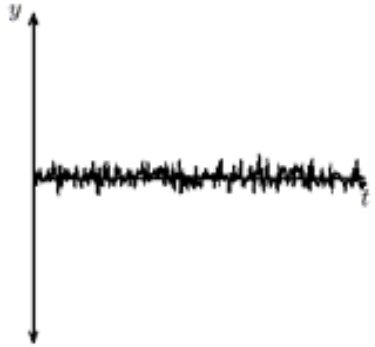
řada má opakující se charakter v rámci jednotlivých sezón



MUNI | RECETOX

Náhodná složka časové řady

Náhodná (stochastická) složka se nedá popsat žádnou funkcí času



- „zbývá" po vyloučení trendu, sezónní a cyklické složky
- jejím zdrojem jsou v **jednotlivostech** nepostižitelné jevy
- lze ji však popsat pravděpodobnostně

Transformace časové řady

Transformace časové řady

Jedná se o úpravu původní časové řady, tak aby

1. splňovala podmínky pro následnou analýzu (např. linearizace, stacionarita atd.)

2. zvýrazňovala dále analyzovanou složku

- přidání konstanty $y = y + C$
- linearizace řady $y = ln(y)$
- odečtení průměru
- standardizace
- odečtení hodnot trendové funkce (…stacionarita)

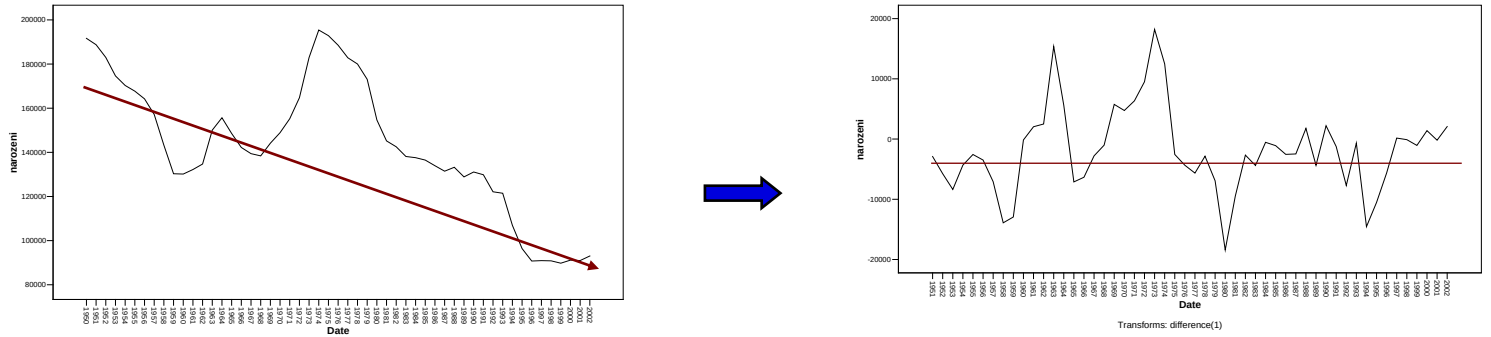
Transformace časové řady

$y_t \rightarrow u_t$

vytvoření nové časové řady

s lepšími parametry pro analýzu

diference $u_t = y_t^{(1)} = y_t - y_{t-1}$ $u_t = \mathrm{DIFF}(y_t, 1)$



lineární trend časové řady se mění na stacionární
trendový filtr – zbaví časovou řadu trendové složky

MUNI | RECETOX

Diference vyšších řádů

1. diference $y_t^{(1)} = y_t - y_{t-1}$ DIFF(y_t, 1)

lineární trend → konstantní (stacionární)

kvadratický trend → lineární

…

2. diference $y_t^{(2)} = y_t^{(1)} - y_{t-1}^{(1)}$ DIFF(y_t, 2)

n. diference $y_t^{(n)} = y_t^{(n-1)} - y_{t-1}^{(n-1)}$ DIFF(y_t, *n*)

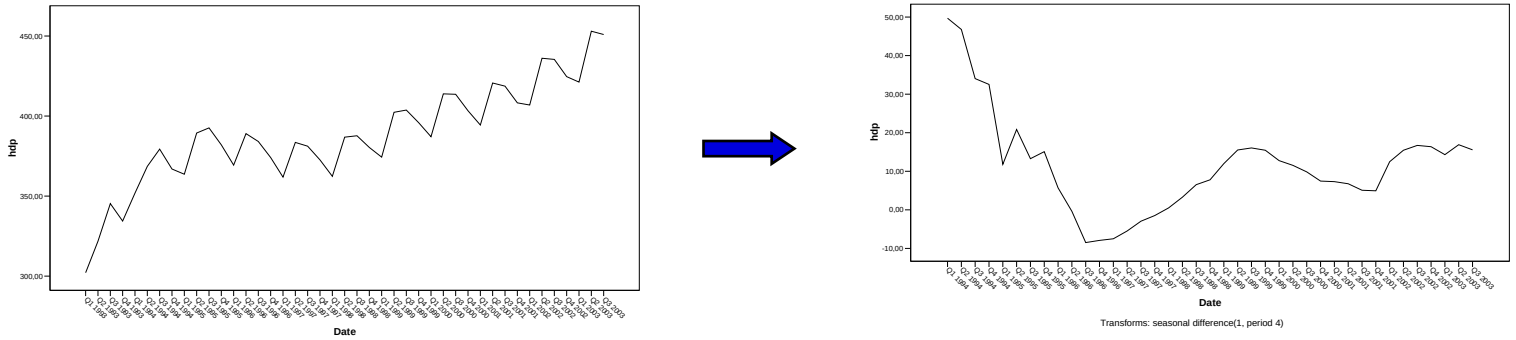
exponenciální trend → exponenciální

(odolný proti diferencování)

Sezónní diference

sezónní diference (délka sezóny = k)

$u_t = y_t^{(s1)} = y_t - y_{t-k}$ $u_t = \text{SDIFF}(y_t, 1)$

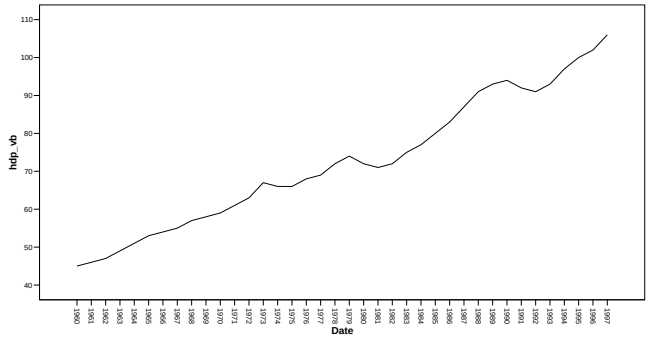


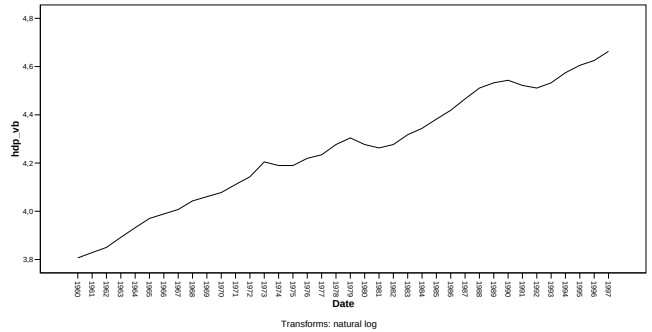
snižuje nebo odstraňuje vliv sezónnosti
sezónní filtr – zbaví časovou řadu sezónnosti

Logaritmická transformace

$u_t = \ln y_t$

exponenciální trend → lineární





logaritmická diference:

$u_t = \ln y_t - \ln y_{t-1} = ln\ (y_t / y_{t-1})$

exponenciální trend → stacionární

MUNI | RECETOX

Doplnění chybějících hodnot

pro další zpracování časové řady je třeba chybějící hodnoty nahradit jejich odhady

globální odhady:

- **SMEAN** (*series mean*) – průměr celé řady
- **TREND** (*linear trend at point*) – lineární trend celé řady

lokální odhady:

- **MEAN** (*mean of nearby points*) – průměr okolních hodnot
- **MEDIAN** (*median of nearby points*) – medián z okolních hodnot
- **LINT** (*linear interpolation*) – lin. interpolace z okolních bodů

Autokorelace

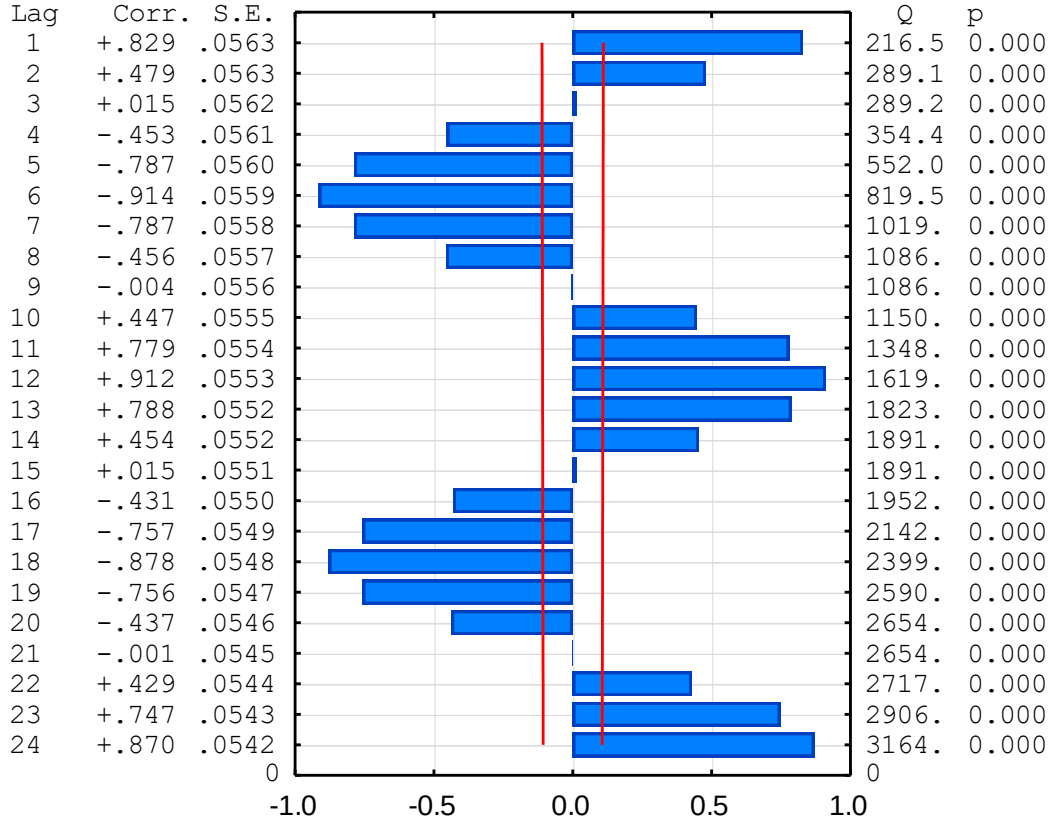
U stacionární časové řady -korelace dvou časově posunutých pozorování (autokorelace), závisí na délce posunu

Autocorrelation Function

water temperature median

(Standard errors are white-noise estimates)

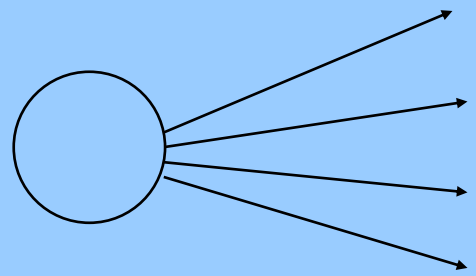
Lag	Corr.	S.E.	Q	p
1	+.829	.0563	216.5	0.000
2	+.479	.0563	289.1	0.000
3	+.015	.0562	289.2	0.000
4	-.453	.0561	354.4	0.000
5	-.787	.0560	552.0	0.000
6	-.914	.0559	819.5	0.000
7	-.787	.0558	1019.	0.000
8	-.456	.0557	1086.	0.000
9	-.004	.0556	1086.	0.000
10	+.447	.0555	1150.	0.000
11	+.779	.0554	1348.	0.000
12	+.912	.0553	1619.	0.000
13	+.788	.0552	1823.	0.000
14	+.454	.0552	1891.	0.000
15	+.015	.0551	1891.	0.000
16	-.431	.0550	1952.	0.000
17	-.757	.0549	2142.	0.000
18	-.878	.0548	2399.	0.000
19	-.756	.0547	2590.	0.000
20	-.437	.0546	2654.	0.000
21	-.001	.0545	2654.	0.000
22	+.429	.0544	2717.	0.000
23	+.747	.0543	2906.	0.000
24	+.870	.0542	3164.	0.000



MUNI | RECETOX

Klasický model časové řady

popis forem (časového) pohybu



T_t – trendová složka

C_t – cyklická složka

S_t – sezónní složka

e_t – reziduální (náhodná) složka

T_t C_t S_t – deterministické složky (*lze modelovat*)

e_t – stochastická složka (*nelze předvídat*)

Aditivní a multiplikativní model

aditivní model

$$y_t = T_t + C_t + S_t + \varepsilon_t$$

trend – vyjádřen regresní funkcí
cyklická a sezónní složka – přírůstky k trendu

multiplikativní model

$$y_t = T_t \cdot C_t \cdot S_t \cdot \varepsilon_t$$

trend – vyjádřen regresní funkcí
cyklická a sezónní složka – indexy

Klasický nesezónní model s konstantními parametry

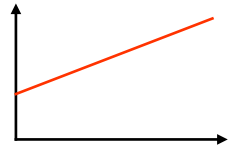
$$y_t = T_t + (C_t) + \varepsilon_t$$

cyklická složka – projevuje se až u dlouhodobých řad, obvykle ji neuvažujeme

hlavní úkol – volba trendové funkce T_t

lineární funkce

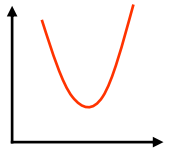
$$T_t = \beta_0 + \beta_1 t$$



konstantní 1. diference

kvadratická funkce

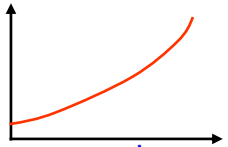
$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2$$



konstantní 2. diference

exponenciální funkce

$$T_t = \beta_0 \cdot \beta_1^t$$



konstantní diference logaritmů

MUNI | RECETOX

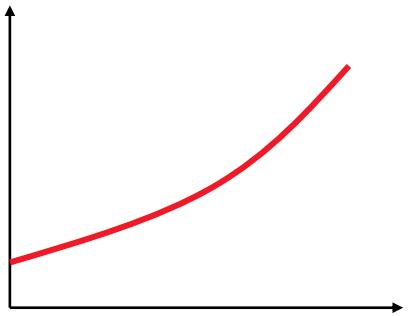
Klasický x adaptivní přístup

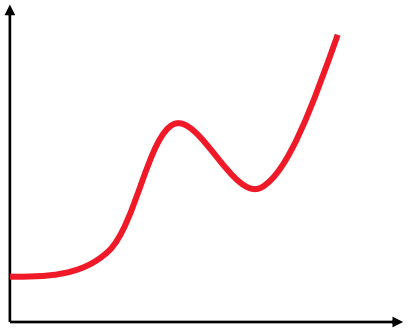
- ***klasický přístup***

  model má stále stejné parametry
  nevyvíjí se v čase

- ***adaptivní přístup***

  model má proměnné parametry
  reaguje na strukturální změny v čase





Modely analýzy časových řad

Časová řada – hodnota ukazatele je funkcí času a náhodné složky

K analýze a popisu časových řad se používá několika základních modelů:

A. Klasický (formální) model
B. Box-Jenkinsova metodologie
C. Lineární dynamické a regresní modely
D. Spektrální analýza

ARIMA

Komplexní lineární model, složený ze tří dílčích částí (nemusí se vždy vyskytovat všechny tři):

- **AR (Autoregresive)** – autoregresivní proces

  lineární kombinace vlivů minulých hodnot

- **I (Integrative)** - náhodná procházka

  odfiltrování nestacionární složky dat

- **MA (Moving Average)** – metoda klouzavých průměrů

  lineární kombinace vlivů minulých chyb (šoků)

Vlastnosti modelů ARIMA

- Mimořádně flexibilní
- Relativně náročné pro výpočet a pro pochopení výsledků (obtížná interpretace parametrů)
- Náročné na kvalitu a počet naměřených dat
- Předpoklad: alespoň 50 měření