Centrum pro výzkum
toxických látek
v prostředí

Prostorové modelování
-Interpolační techniky
-stanovení prostorové autokorelace

Klára Komprdová

esf evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Interpolace x Extrapolace

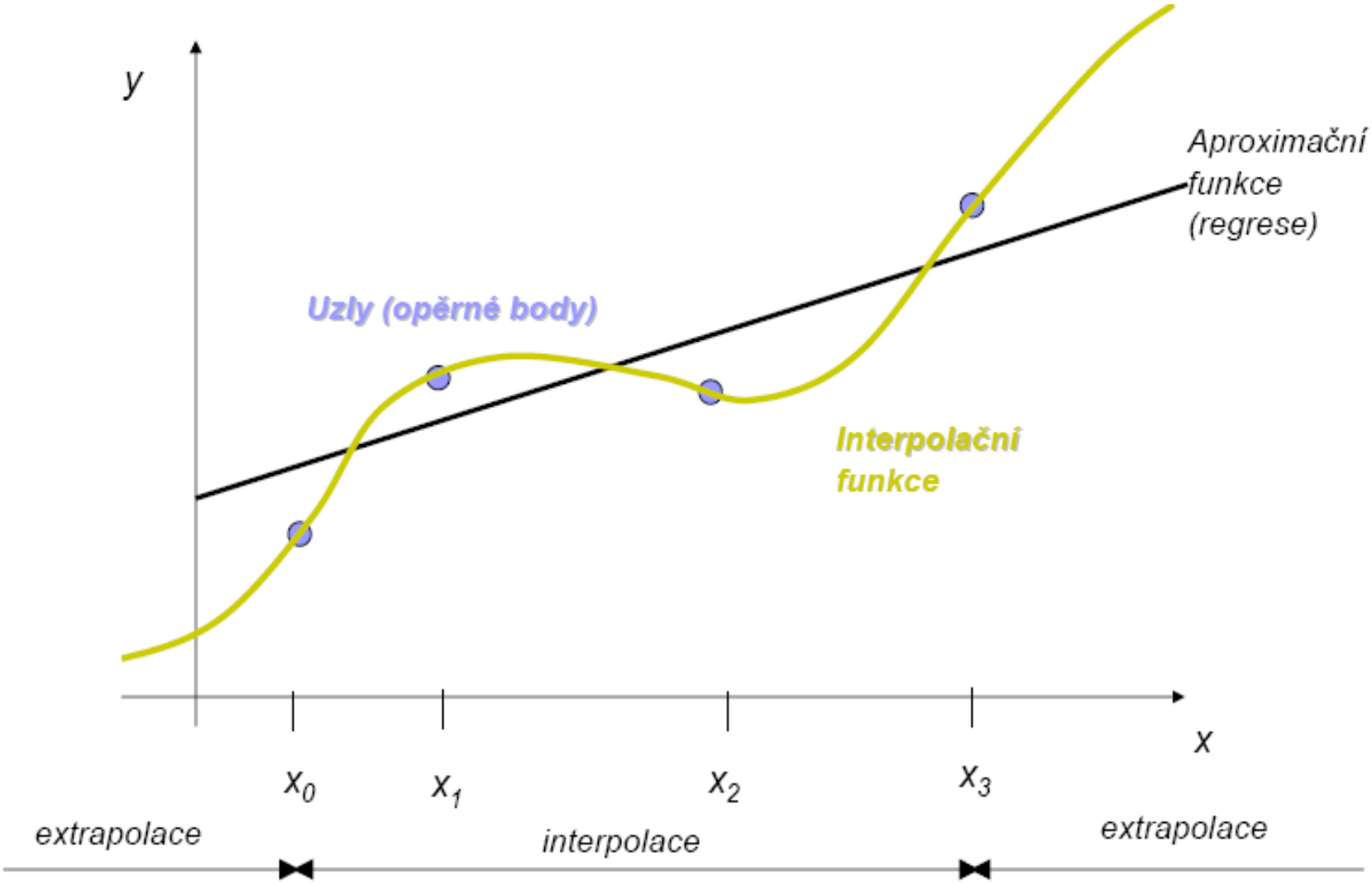
Interpolace – pro „známé“ území (oblast o které máme informace)

- 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

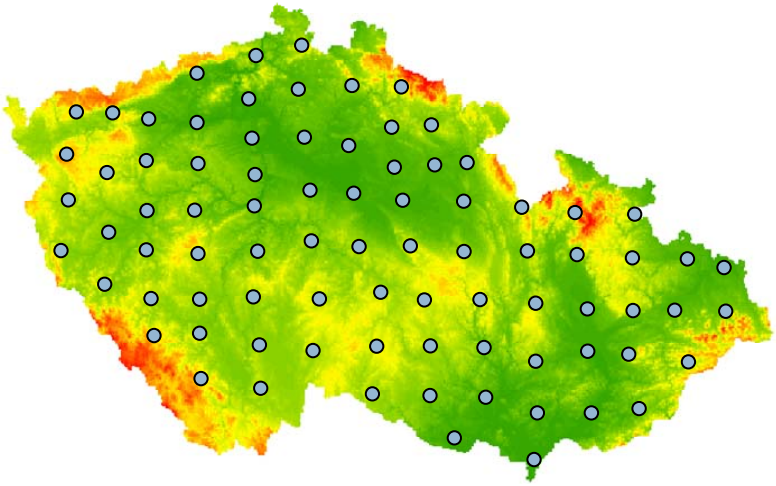
Interpolace, aproximace, extrapolace



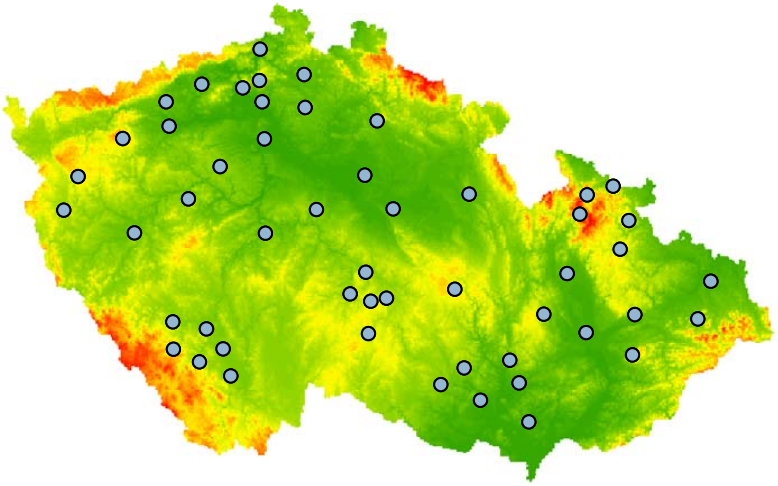
F. Ježek (2006) Interpolace funkcí

Interpolace x Extrapolace

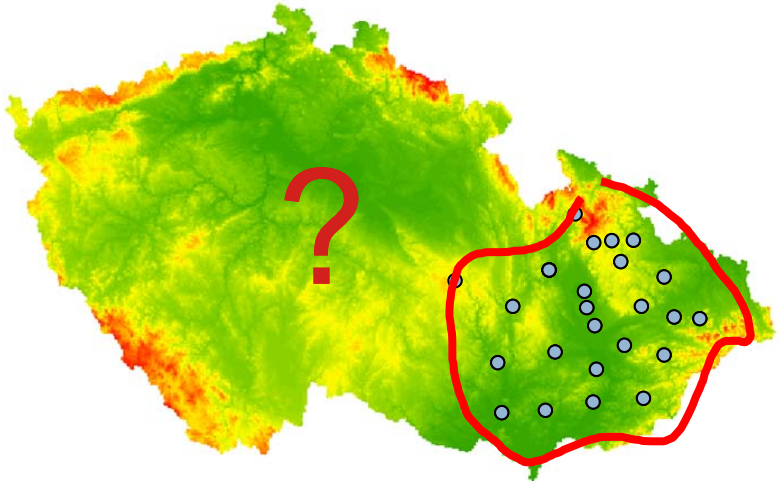
Interpolace – rovnoměrné vzorkování



Interpolace – nerovnoměrné vzorkování



Extrapolace – prediktivní modelování



Interpolační metody

rozdělení metod:

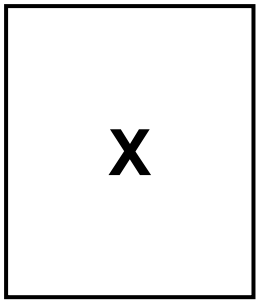
- *Deterministické* (MECHANICAL/EMPIRICAL MODELS) – (IDW –Inverse distance interpolation, Regression on coordinates, Splines …)
  - parametry modelu jsou voleny libovolně či empiricky
  - neodhaduje se predikční chyba
  - většinou nejsou kladeny žádné statistické předpoklady

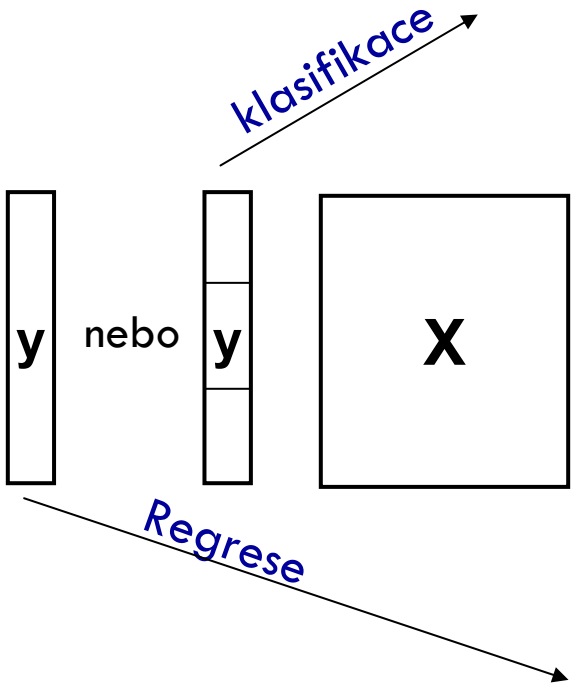
- *Geostatistické* (STATISTICAL (PROBABILITY) MODELS) – využívají prostorovou strukturu celého pole, pro celé pole lze spočítat chybu interpolace (různé typy krigingu– obyčejný, univerzální, blokový, cokriging, Bayesian Maximum Entropy)
  - odhad parametrů v modelu objektivně-teorie pravděpodobnosti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

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

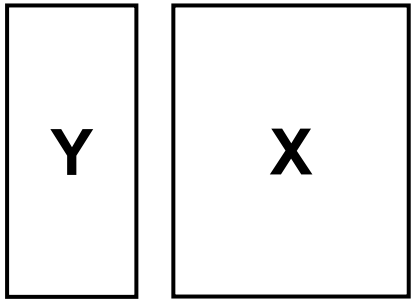
Pokročilejší modelovací přístupy

Ordinace, interpolace





Přímá ordinace



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**

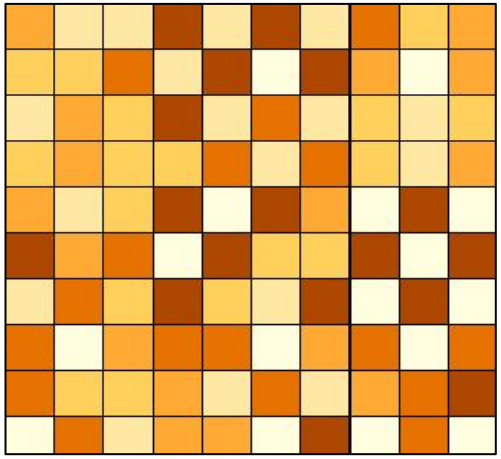
…

Prostorová autokorelace

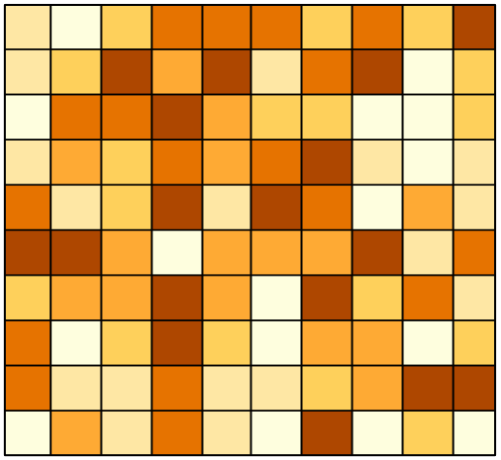
“everything is related to everything else, but near things are more related than distant things” Waldo Tobler

Prostorová autokorelace

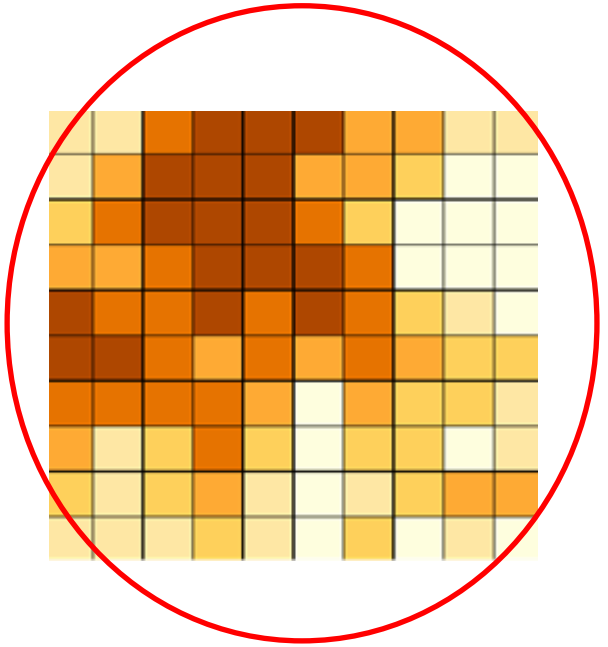
Negativní



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

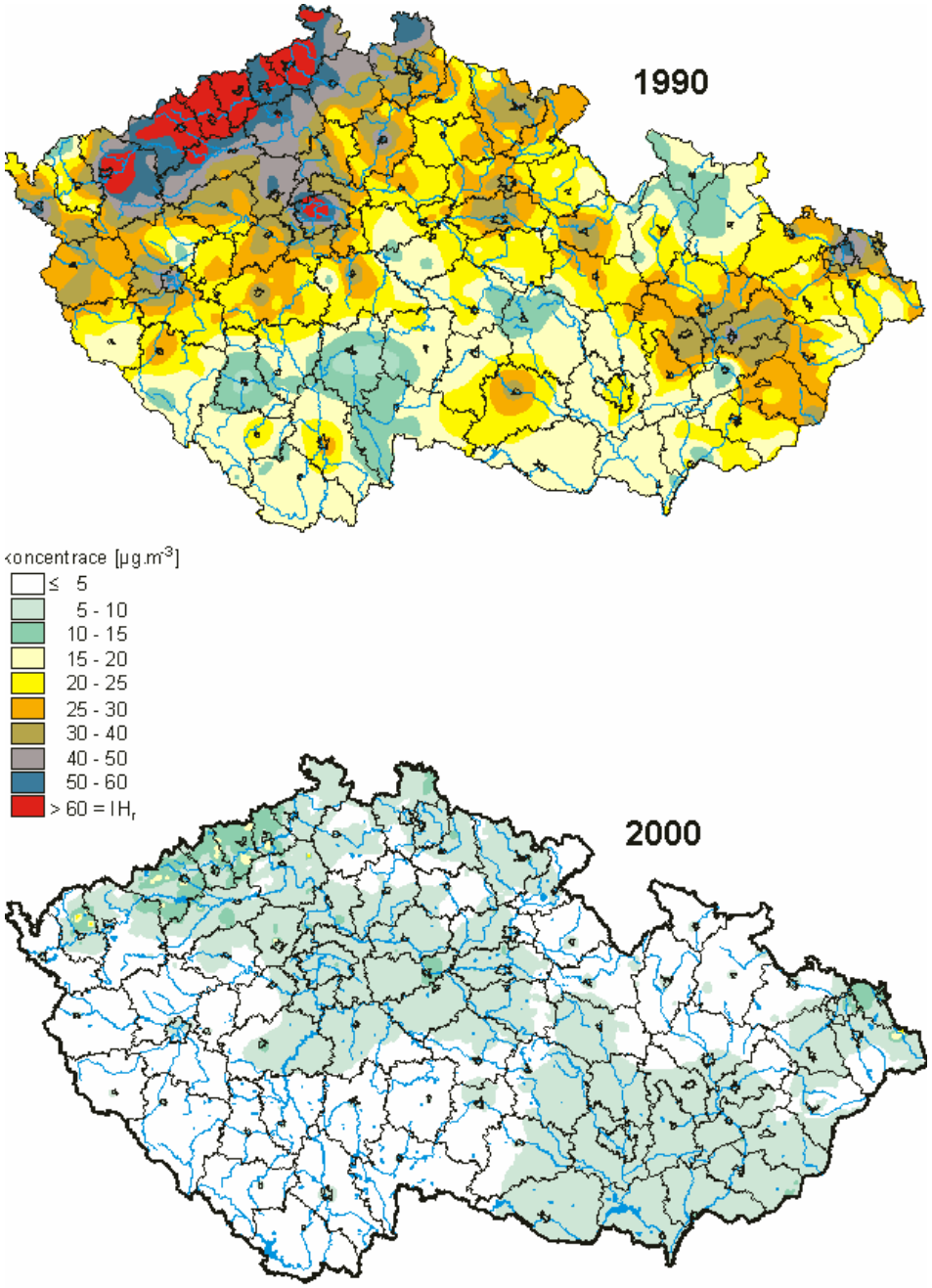
Interpolační metody

- Využívají prostorové autokorelace

Nejznámější kriging a metoda inverzní vzdálenosti

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

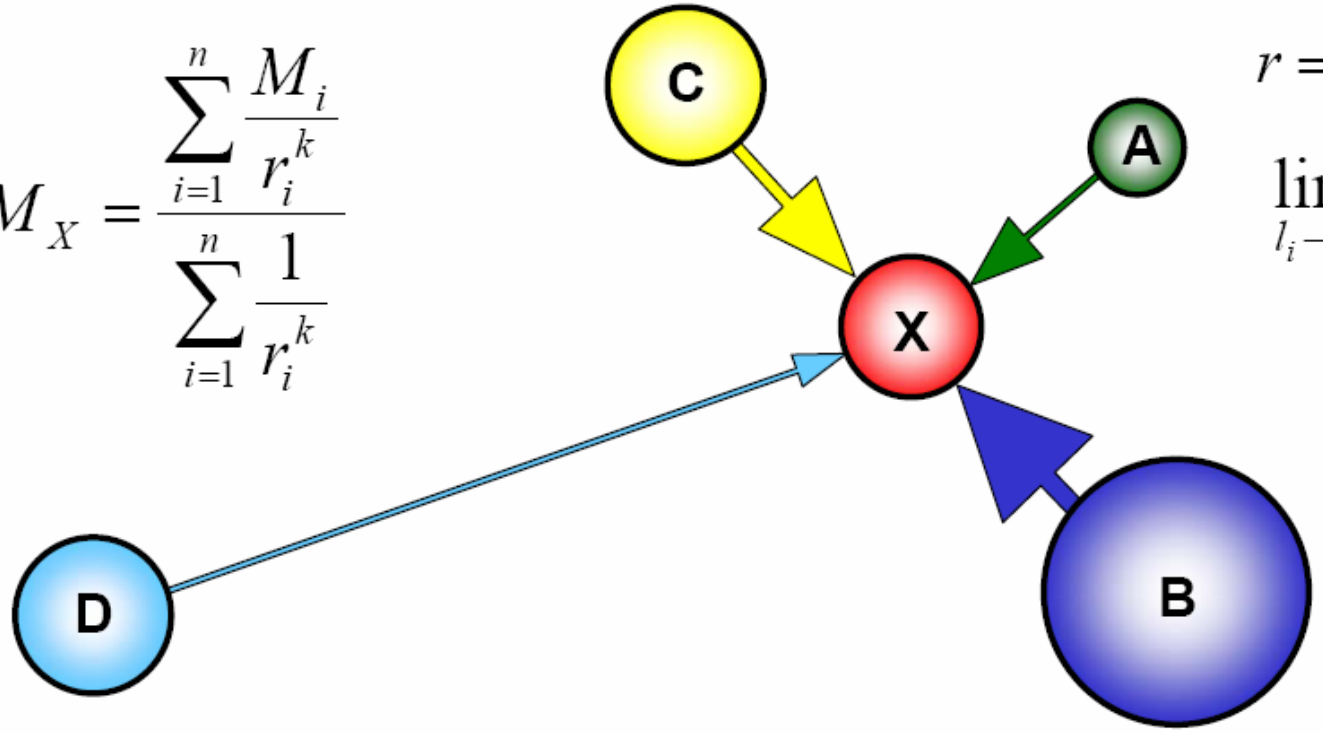
Příklad použití metody IDW- koncentrace SO_2

web ČHMÚ

IDW - Inverse distance weighted – inverzní vážená vzdálenost

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

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

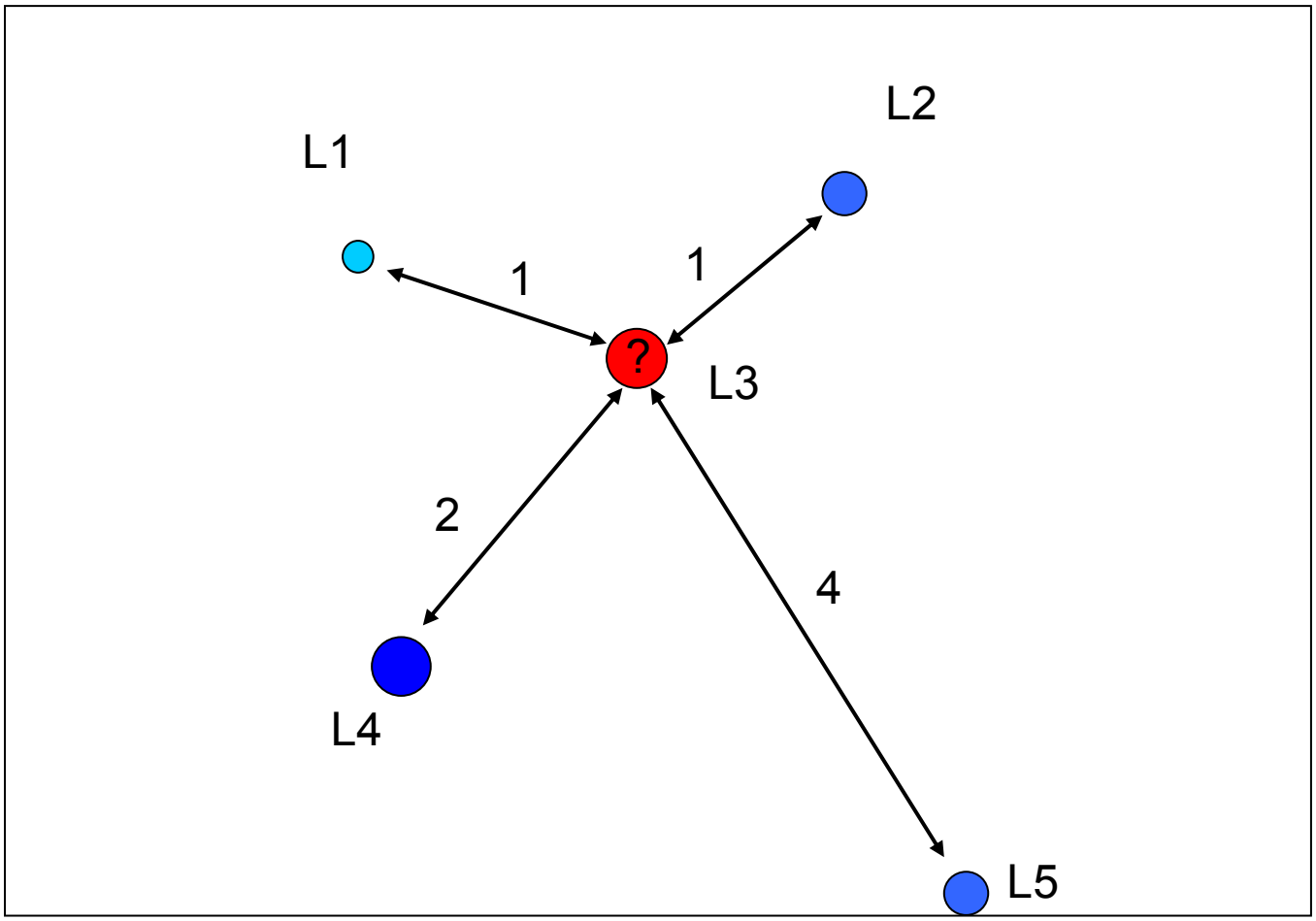


$$r = \sqrt{dX^2 + dY^2}$$

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

„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.

Příklad – IDW – jaká je hodnota na lokalitě L3?



	hodnoty
L1	5
L2	10
L3	?
L4	15
L5	10

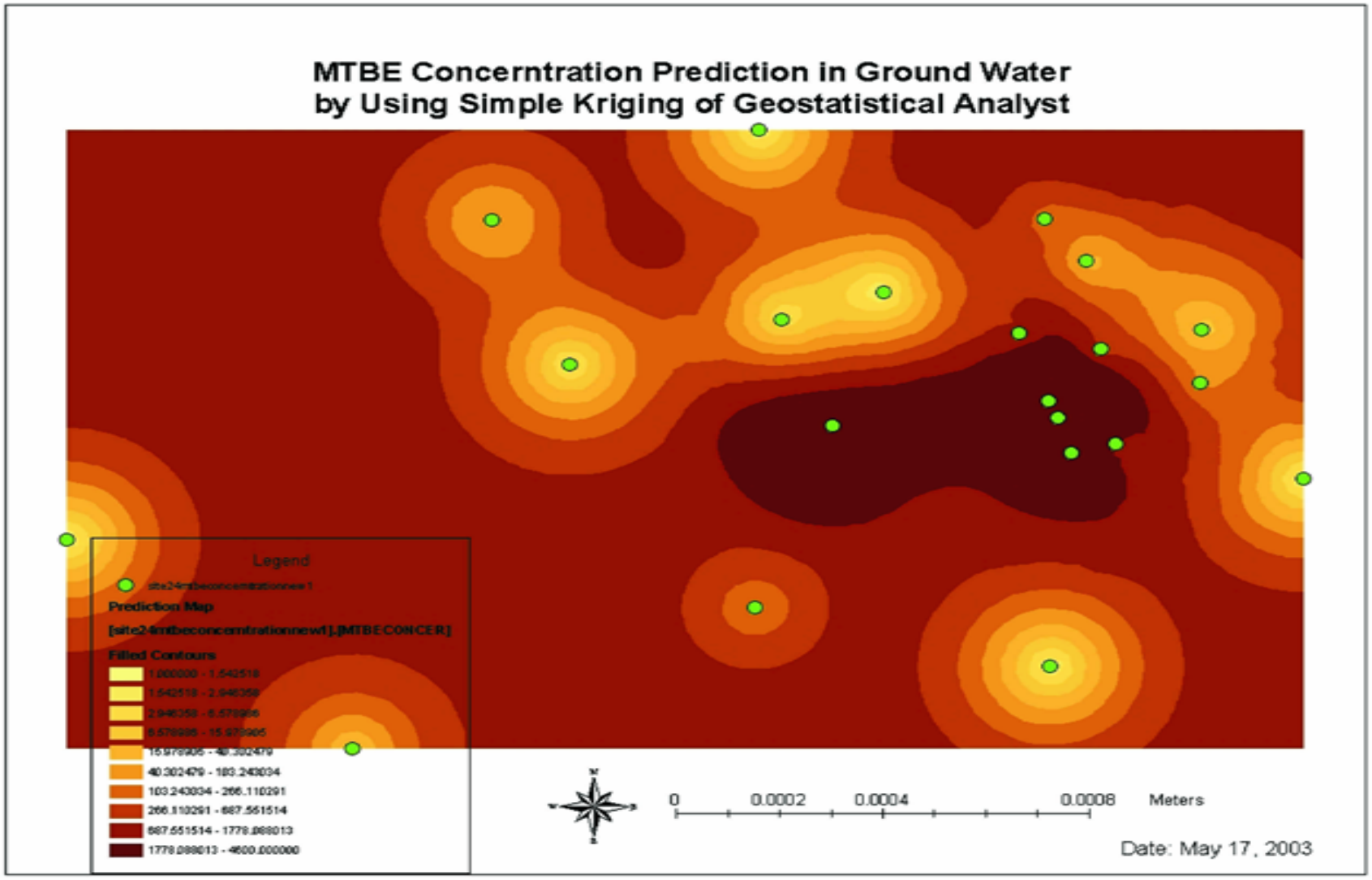
Kriging

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

- při hledání zlatých dolů v jižní Africe!

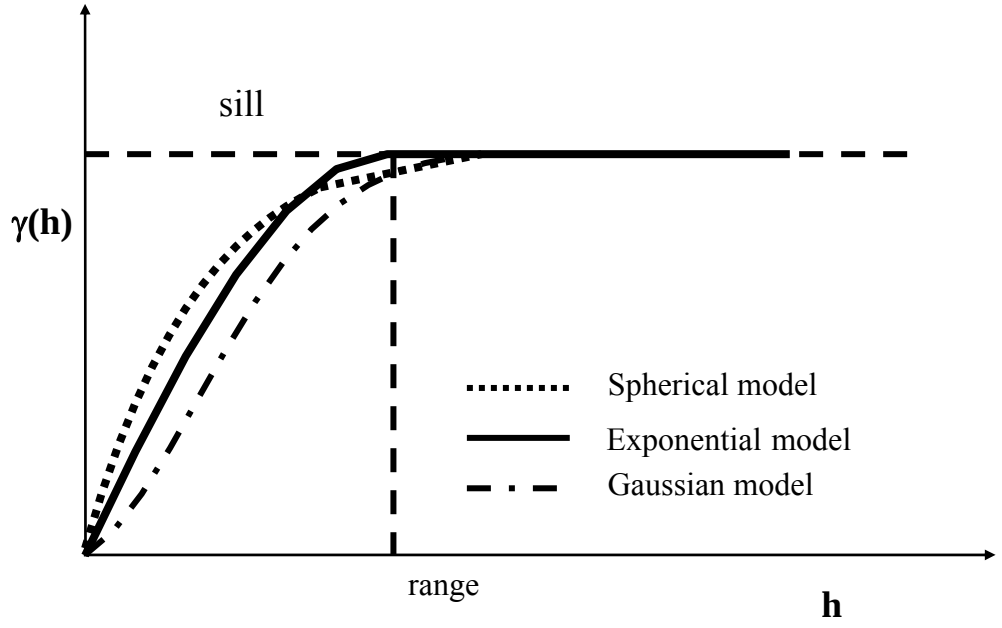
Daniel Gerhardus Krige
26 August 1919

V městě Santa Monica, Kalifornie - byl detekován MTBE (Methyl-3-butyl ether se přidává do benzínu jako antidetonátor) ve spodní vodě v koncentracích ve stovkách ppm (parts per million). Protože MTBE koncentrace byly deset tisíckrát větší než hodnoty doporučené pro pitnou vodu, byly uzavřeny tři z pěti studní, které poskytovaly vodu 40% populace města.



Kriging II

- 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

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{N(h)}^{i=1} \left[Z(x_i) - Z(x_i + h) \right]^2$$

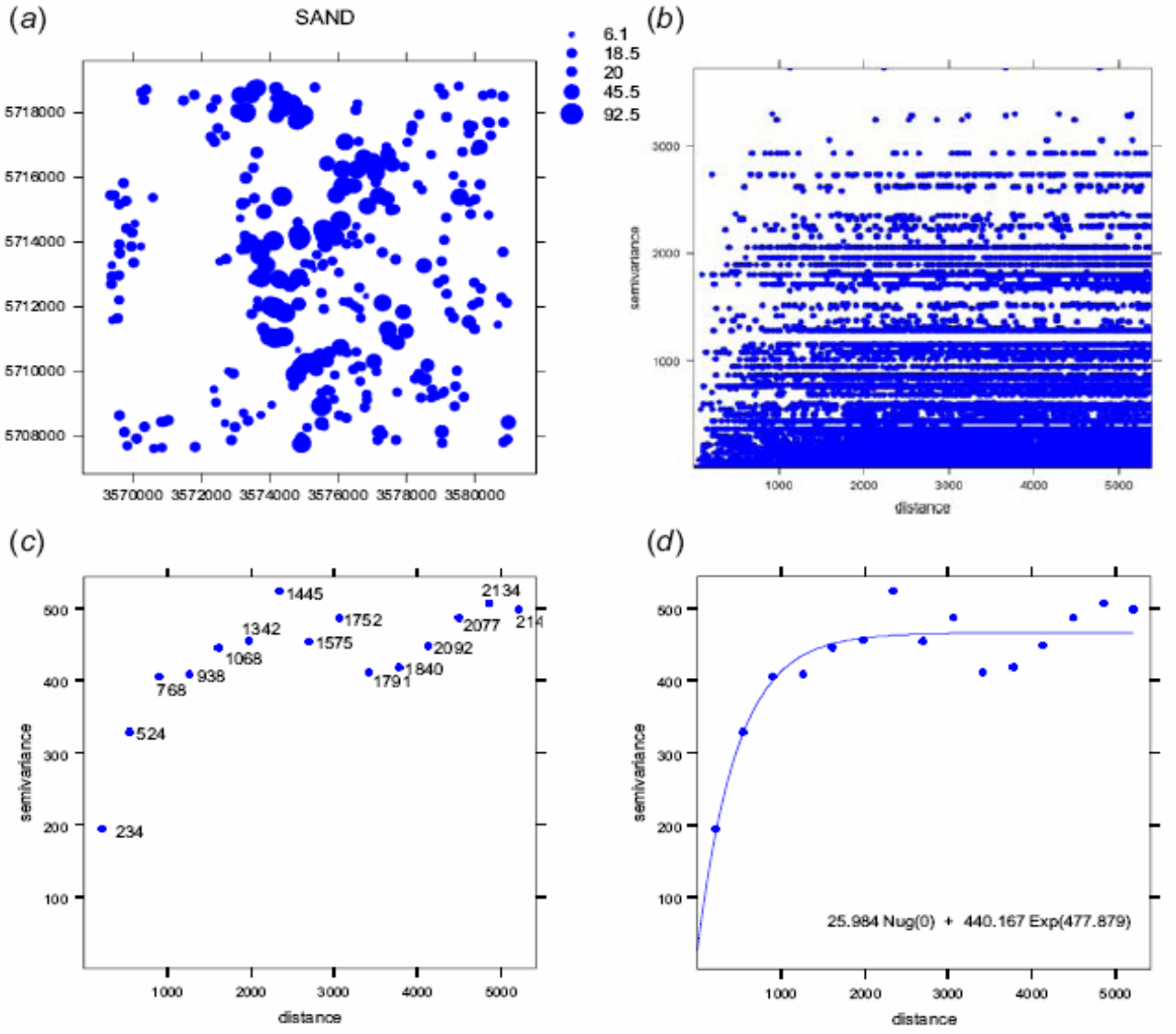
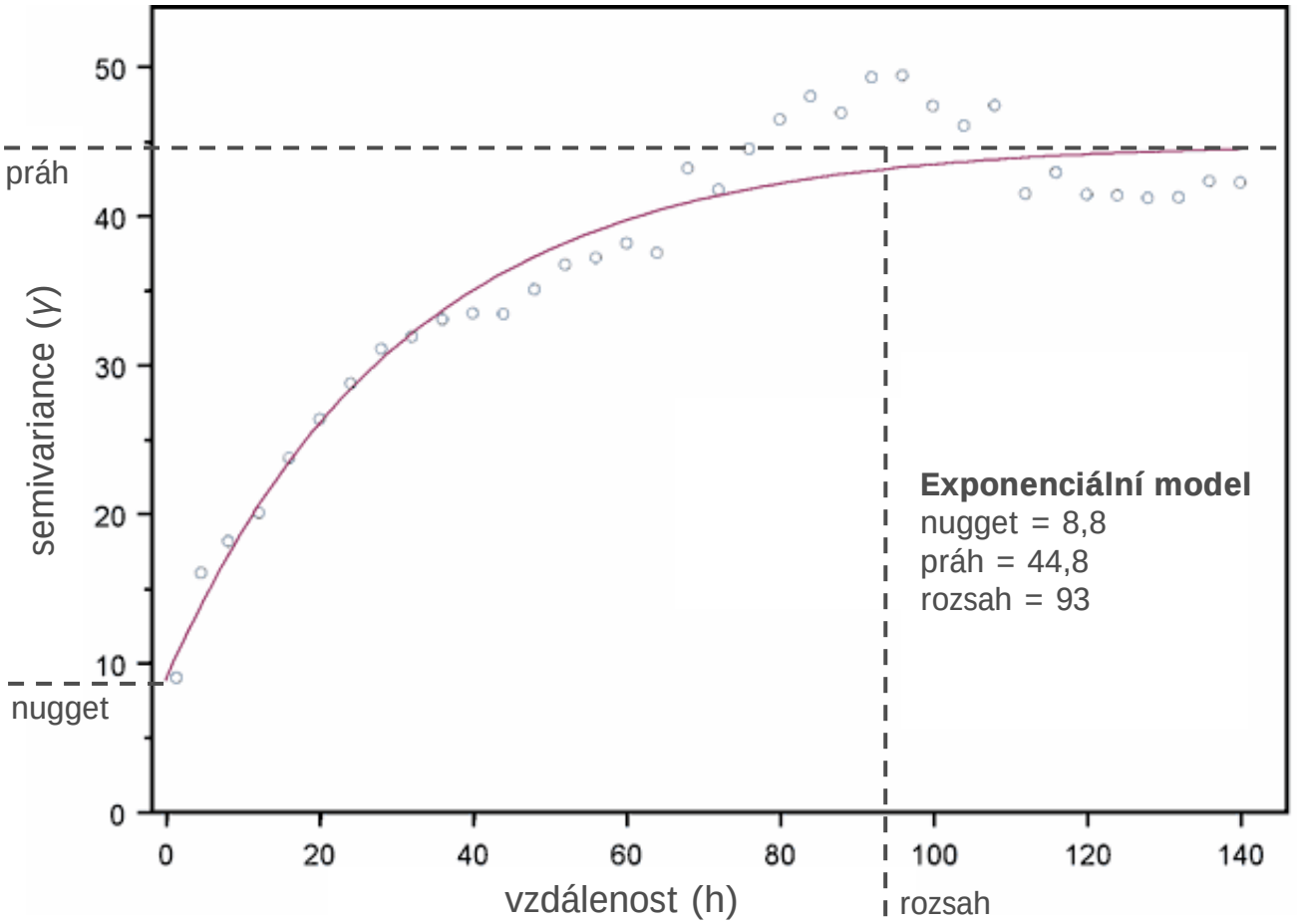
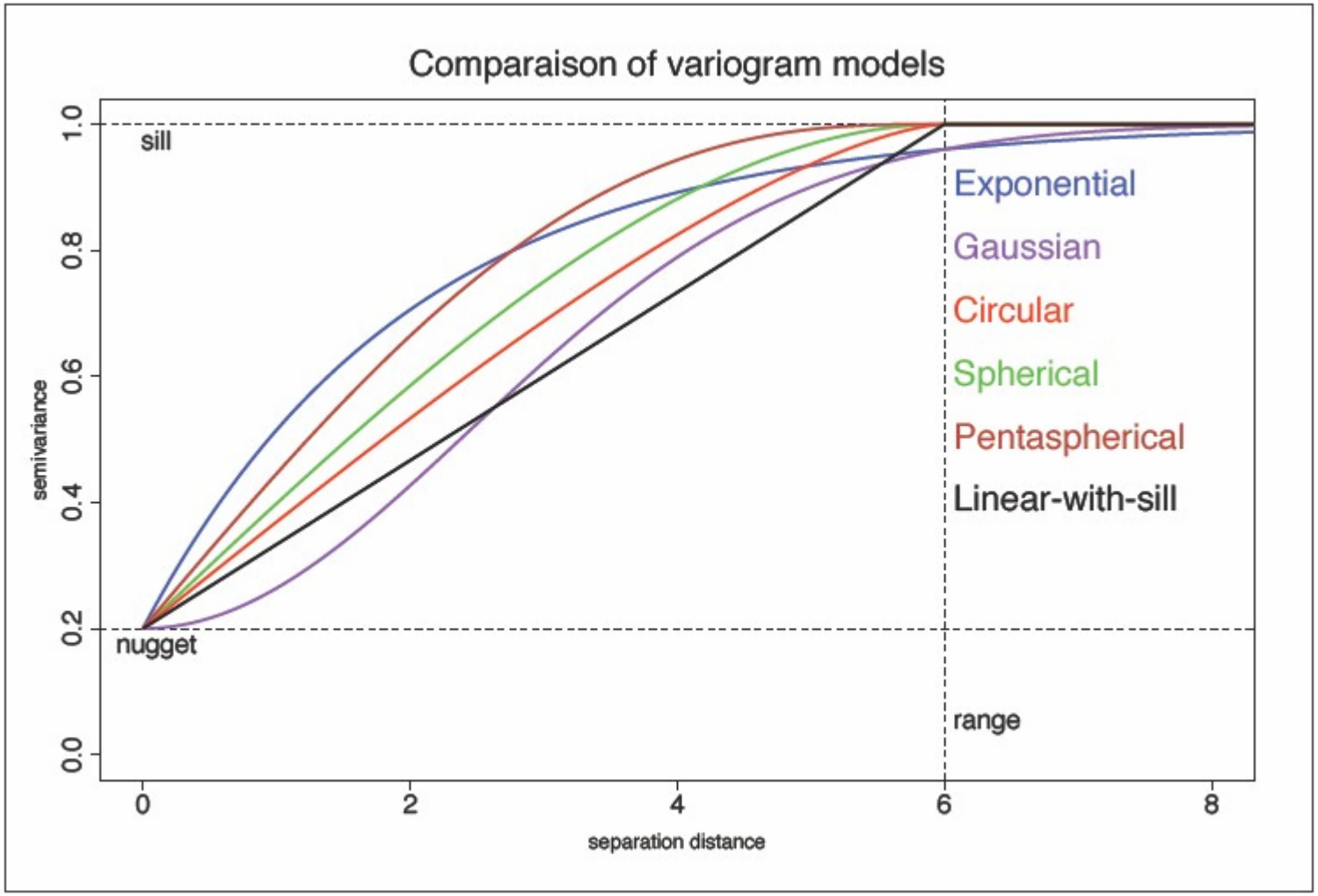


Fig. 1.7: Steps of variogram modelling: (a) location of points (300), (b) variogram cloud showing semivariances for 44850 pairs, (c) semivariances aggregated to lags of about 300 m, and (d) the final variogram model fitted using the default settings in gstat.

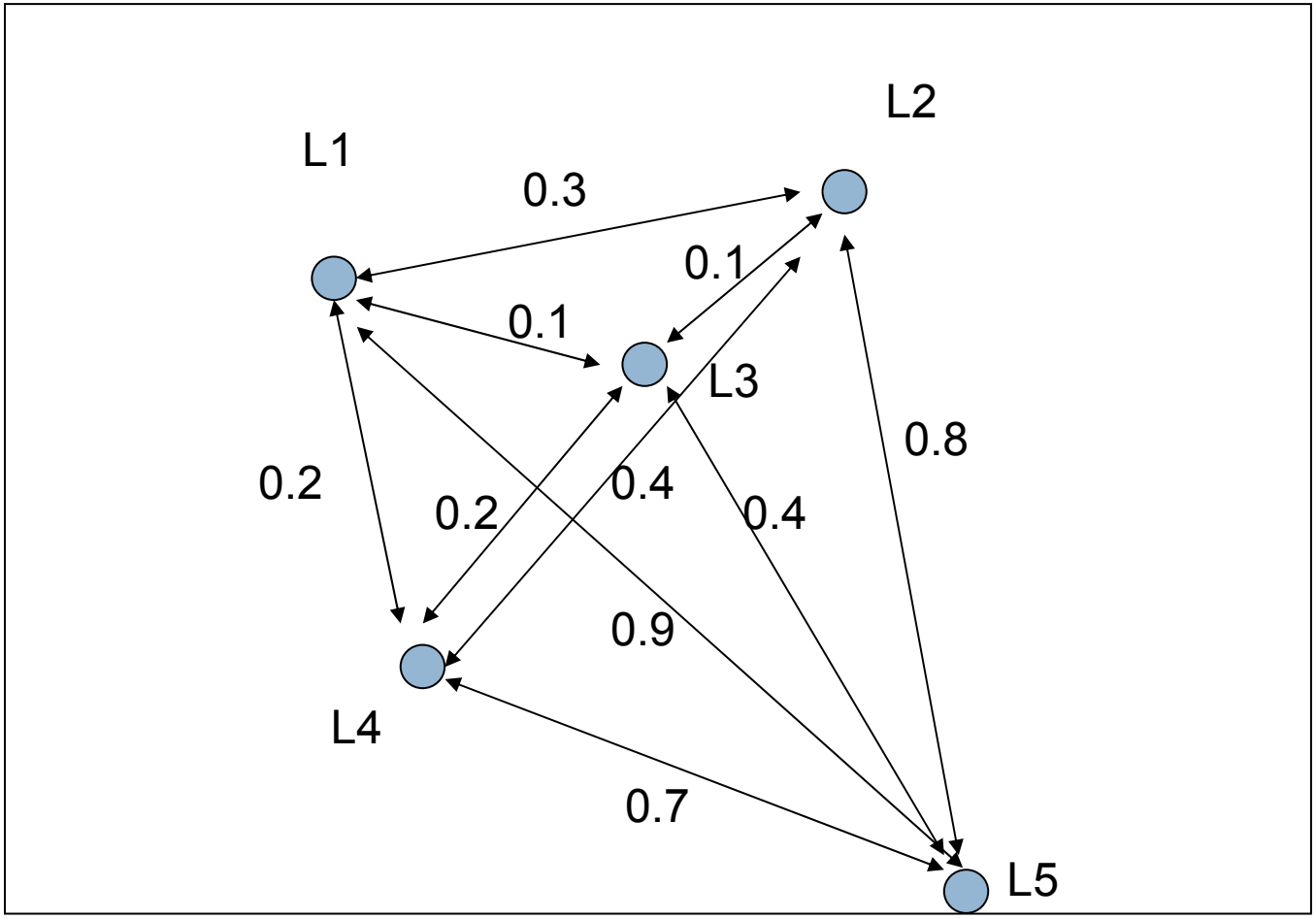
Exponenciální model semivariogramu



Modely variogramu



Příklad - variogram



	hodnoty
L1	5
L2	10
L3	15
L4	15
L5	10

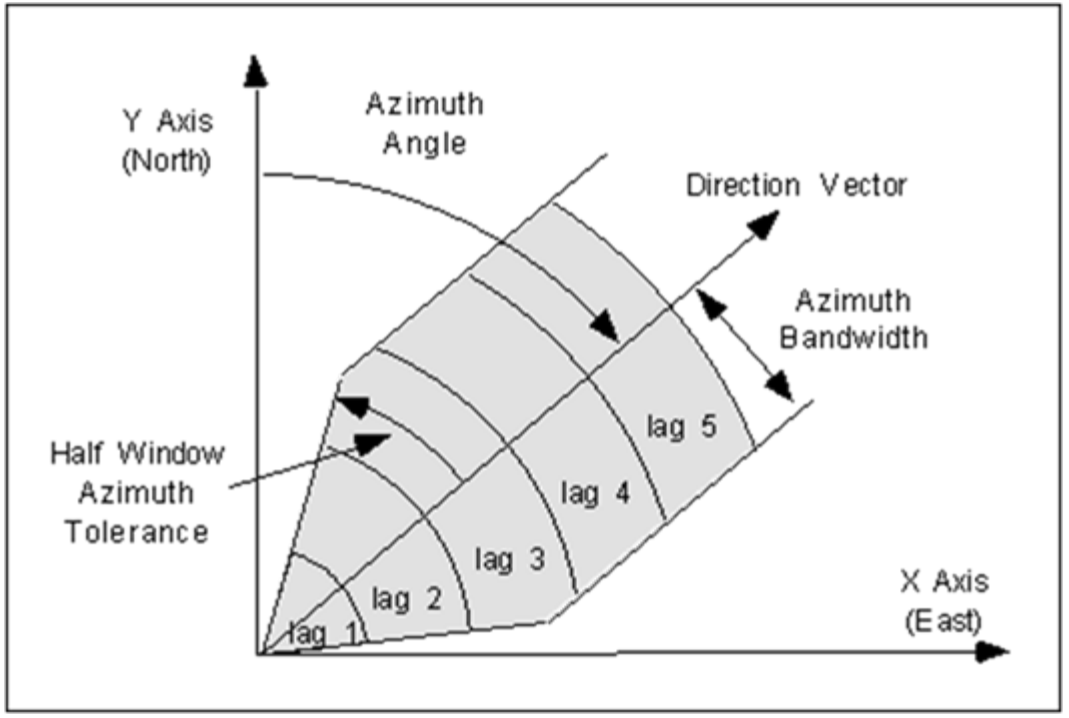
Příklad

Z uvedených hodnot vytvořte semivariogram

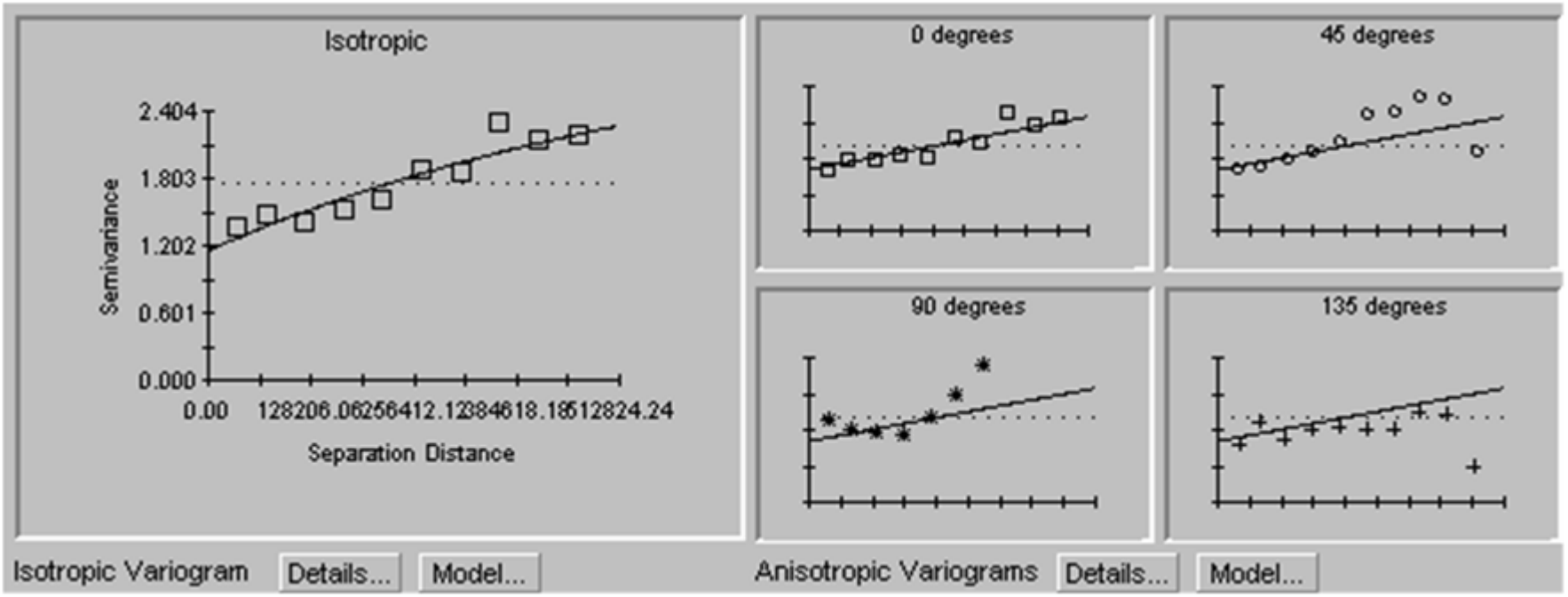
Matice vzdáleností	L1	L2	L3	L4	L5		hodnoty
L1	0	0,3	0,1	0,2	0,9	L1	5
L2	0,3	0	0,1	0,4	0,8	L2	10
L3	0,1	0,1	0	0,2	0,4	L3	15
L4	0,2	0,4	0,2	0	0,7	L4	15
L5	0,9	0,8	0,4	0,7	0	L5	10

Kriging

- 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
- Cíl: vyjádřit 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
- získáme rovnici, která je lineární ve svých parametrech, i když odpověď ***y*** k vysvětlujícím proměnným *x* může být nelineární
- 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

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

Náhodná variabilita

Intercept

Sklon, též regresní koeficient
Slope

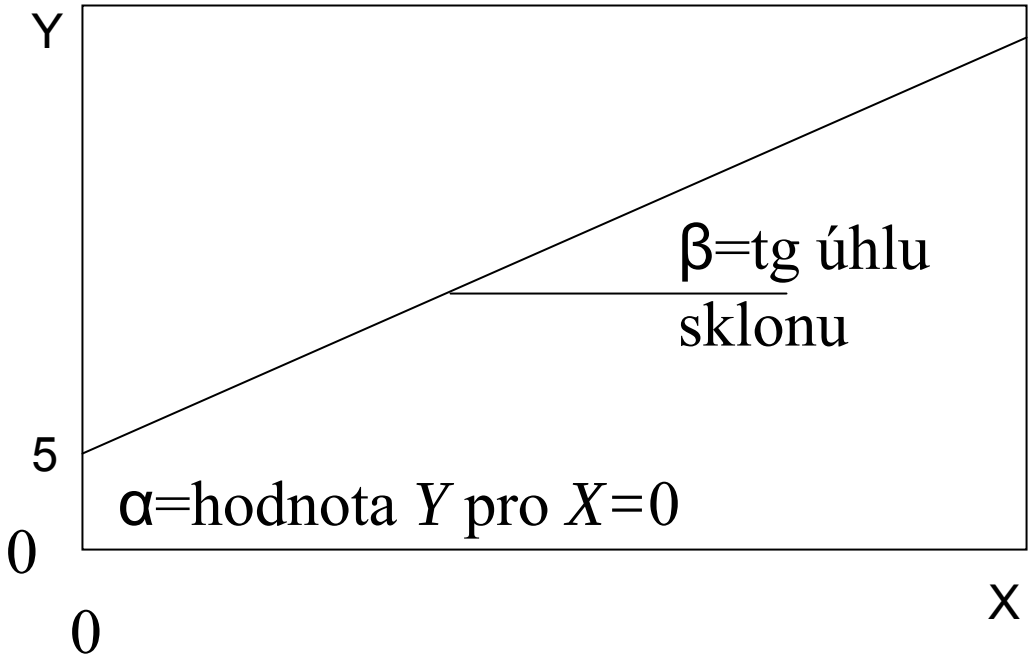
Závisle proměnná
Odpověď
Dependent v., response

Nezávisle proměnná, prediktor,
Independent v.

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

Regresní koeficient = sklon přímky, udává o kolik se změní *Y* při změně *X* o jednotku. Je to tedy hodnota závislá na jednotkách, ve kterých měříme *X* a *Y*. Jde od -∞ do +∞.

Y = 5+0.2*X



Předpokládáme tedy: $$Y = \alpha + \beta X + \varepsilon$$

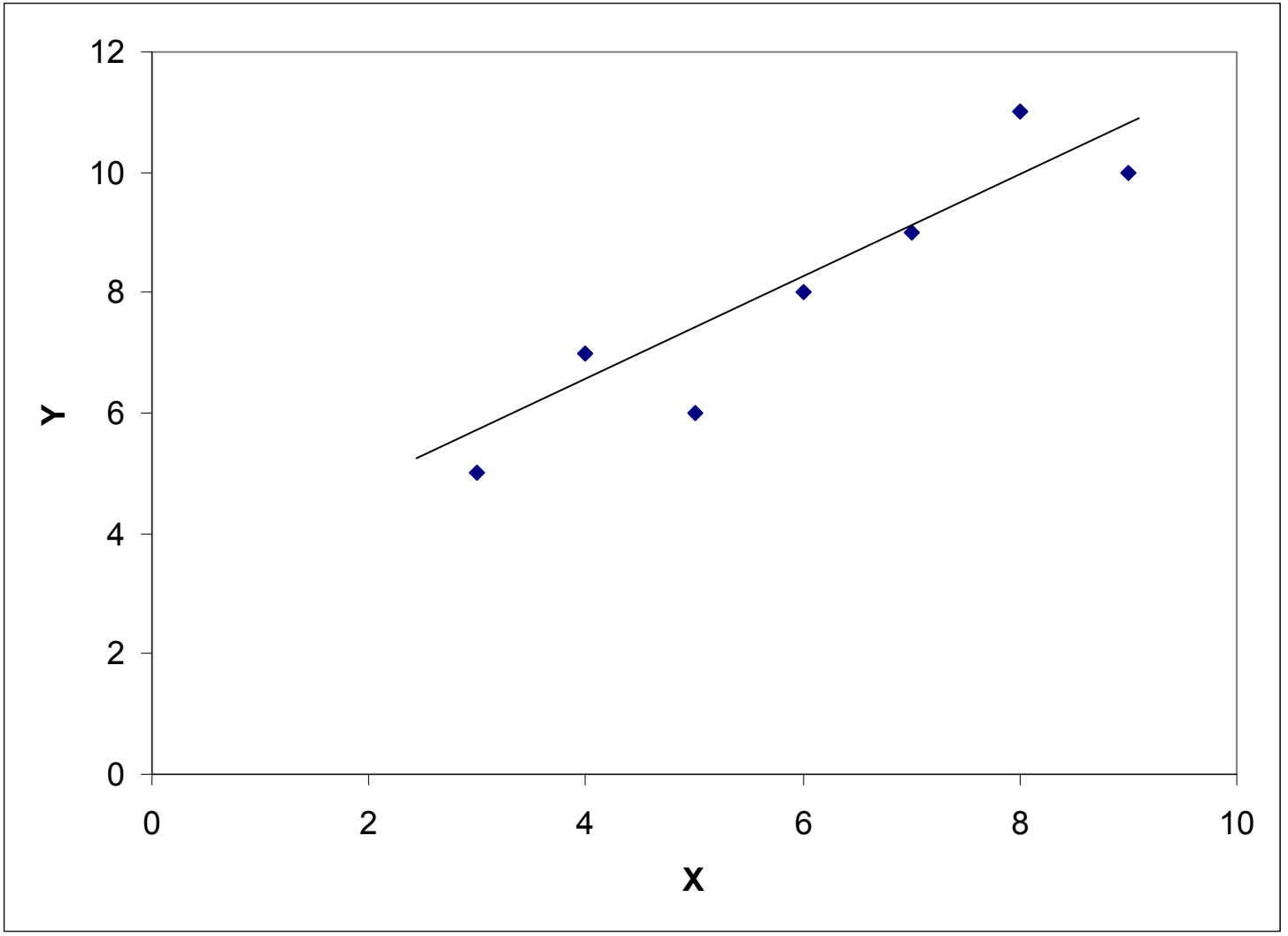
X je změřeno přesně

Y je zatíženo chybou

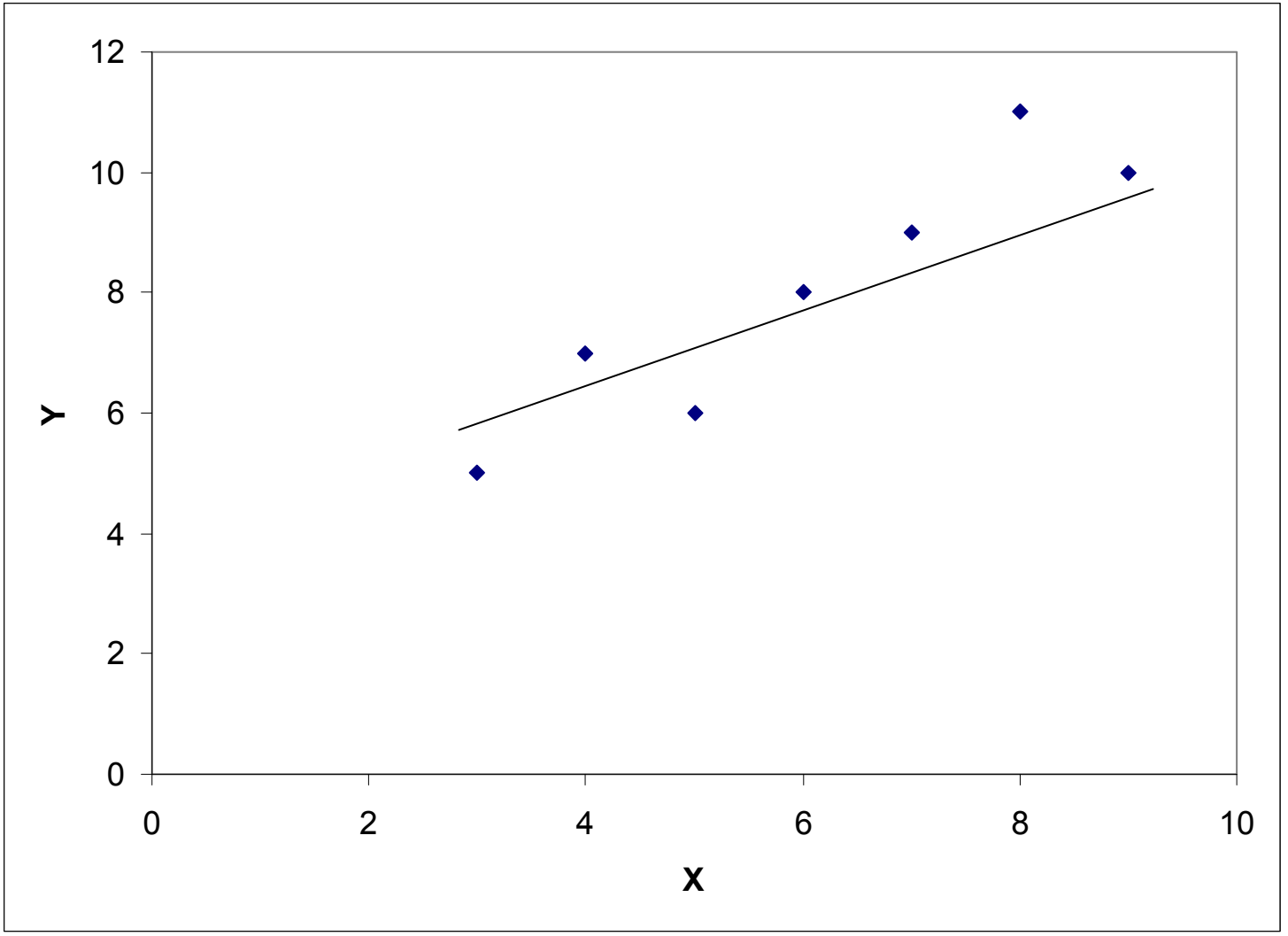
střední hodnota *Y* závisí lineárně na *X*

variance “kolem přímky” je stále stejná (homogenita variance)

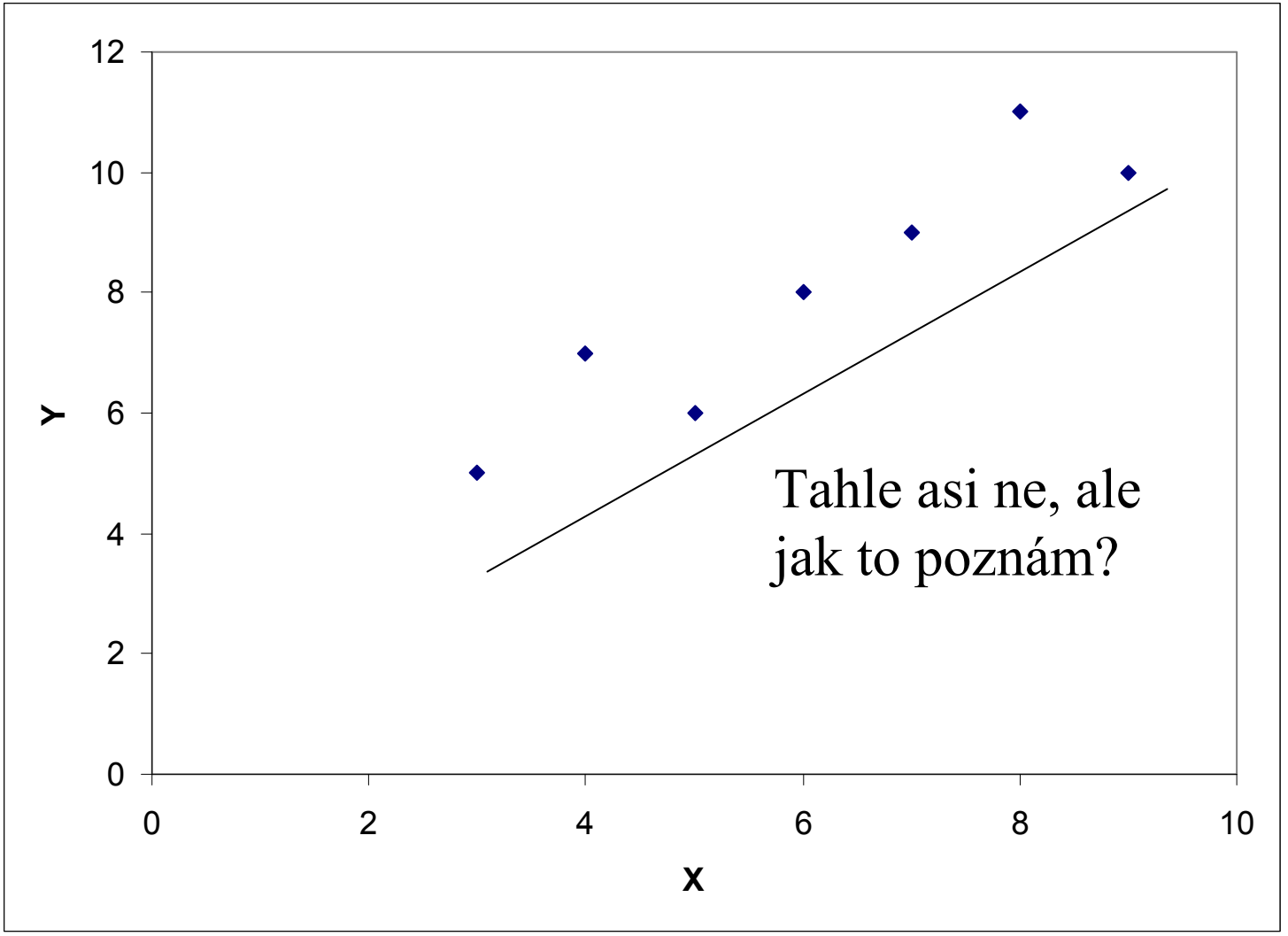
Která přímka je nejlepší?



Která přímka je nejlepší?

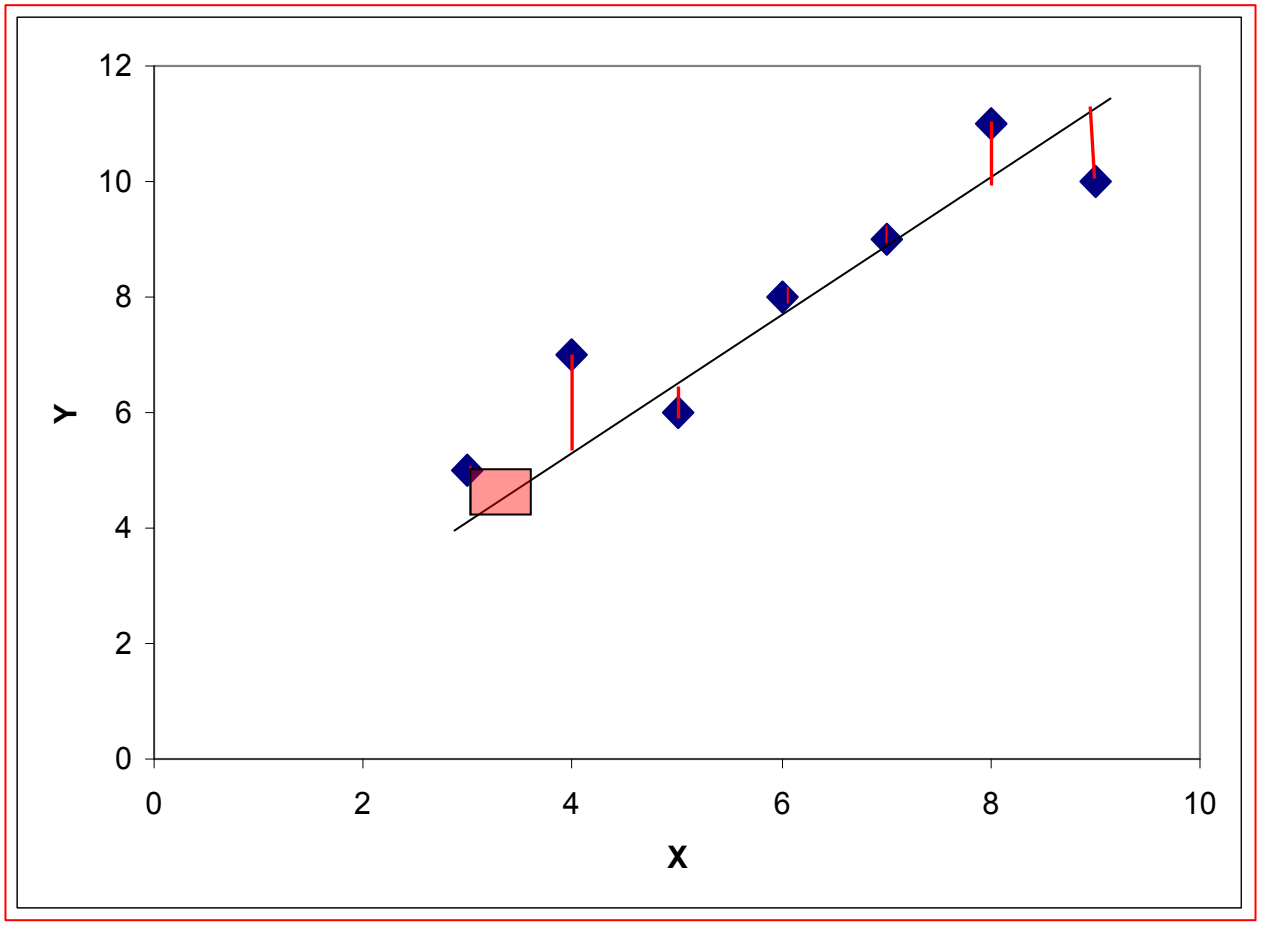


Která přímka je nejlepší?



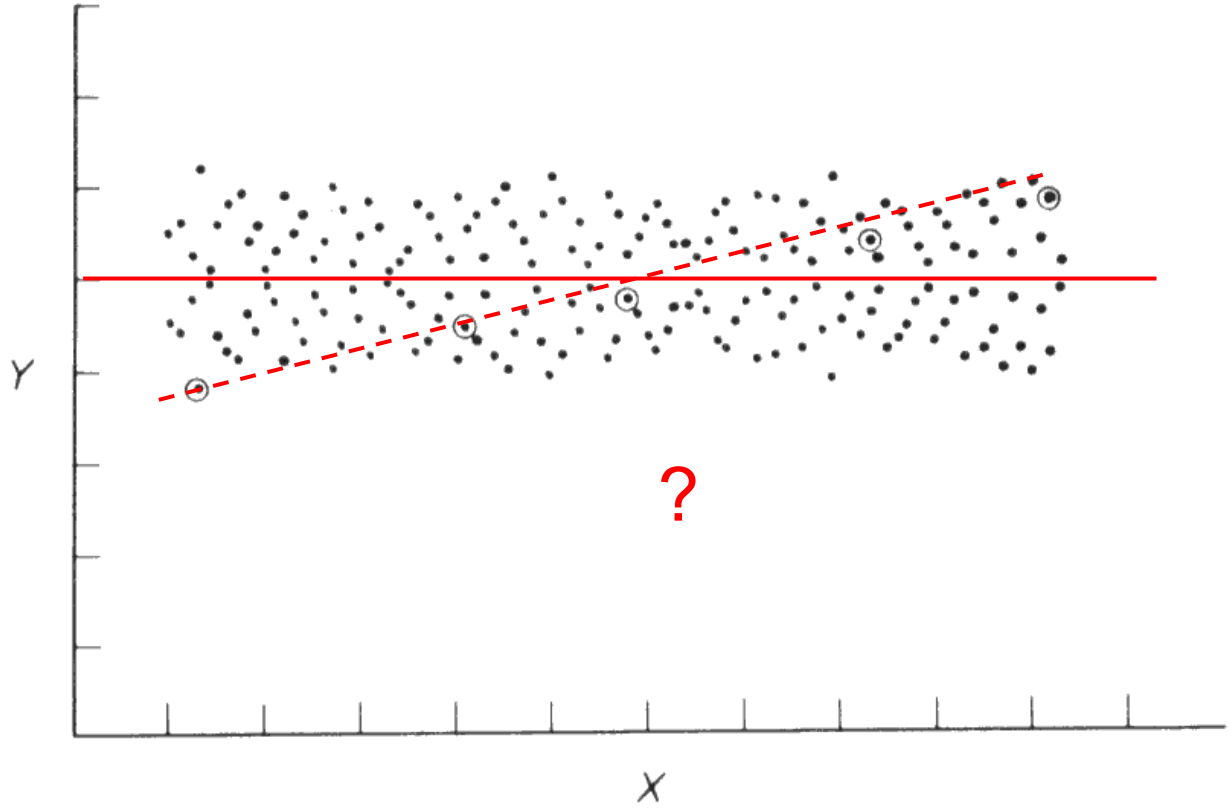
Tj. nejlepší je ta přímka, která má nejmenší součet druhých mocnin (čtverců) residuálů

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$



Svislá - nikoliv kolmá vzdálenost k přímce!!!

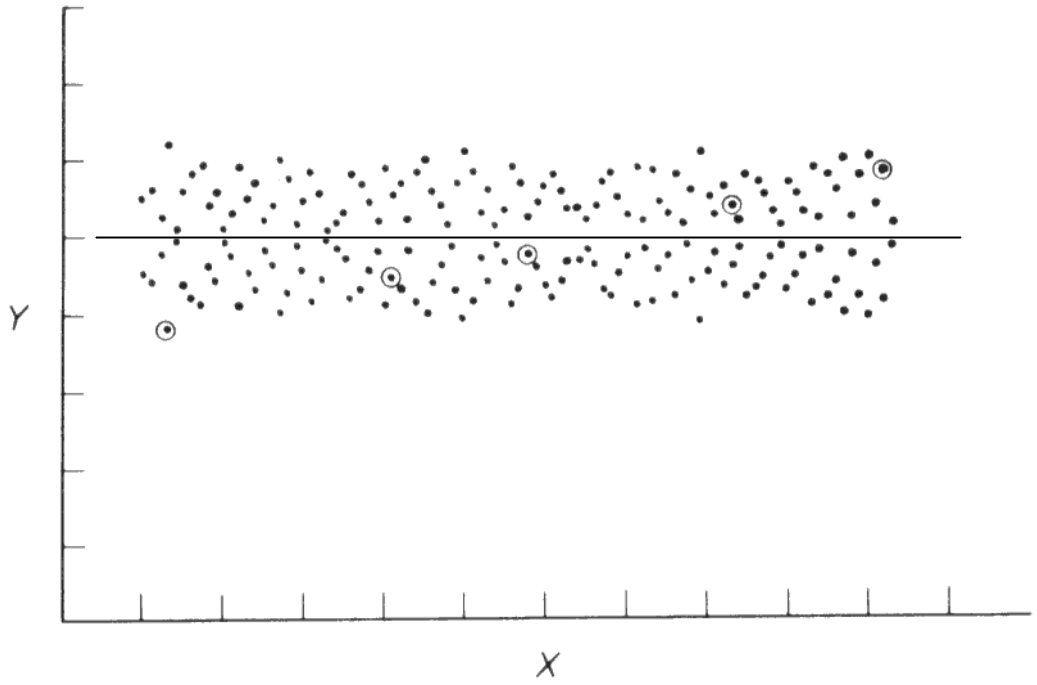
b je výběrovým odhadem skutečné hodnoty *β*



Každý odhad je zatížen nějakou chybou - z variability dat můžeme spočítat střední chybu odhadu *b*

Hypotetický základní soubor dat, s regresním koeficientem β rovným nule. Zakroužkované body mohou být možným výběrem pěti pozorování.

V případě nezávislosti β=0



Dosažená hladina významnosti pro test

H_0: β=0

je pravděpodobnost, že takhle dobrou závislost dostaneme čistě náhodou, pokud jsou proměnné nezávislé

Koeficient determinace - procento vysvětlené variability

$$R^2 = \frac{variabilita\_vysvetlena\_modelem}{celkova\_variabilita\_Y} = 1 - \frac{residualni\_variabilita}{celkova\_variabilita\_Y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = 1 - \frac{SS_e}{SS_{TOT}}$$

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

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

aplikace mnohonásobné lineární regrese, kde prediktory jsou X, X^2, X^3 atd. počítá se stejně (tj. opět kriterium nejmenšího součtu residuálních čtverců, které má opět (normálně) jedno minimum).

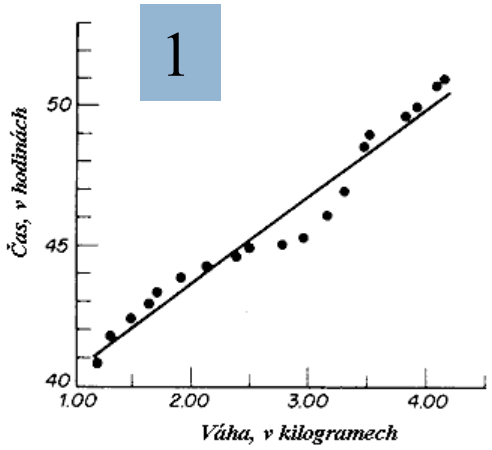
Do modelu jsou přidávány pouze proměnné, které snižují residuální chybu modelu:

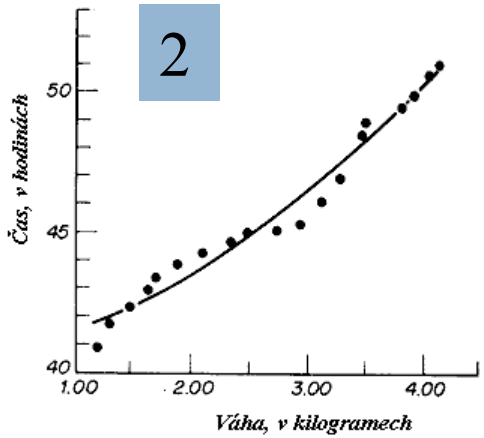
dopředný výběr (*forward elimination*) – začínáme s konstantou (interceptem) a postupně se přidávají jednotlivé členy

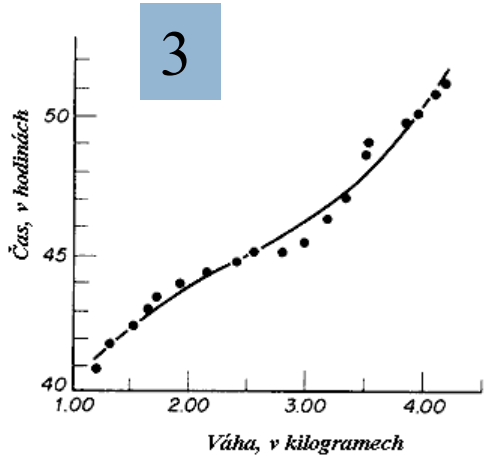
zpětný výběr (*backward elimination*) – začínáme se všemi členy, postupně se odebírají ty, které přispívají k nejmenšímu snížení residuální chyby

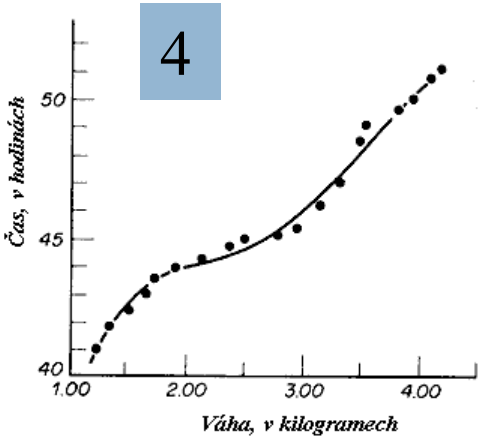
Obdobný význam má i R^2

Se zvyšujícím se stupněm polynomu stoupá “flexibilita”

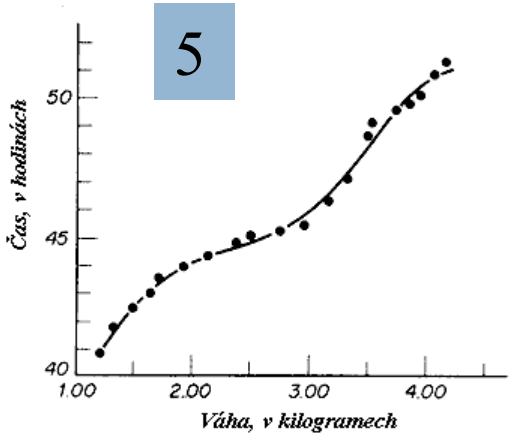




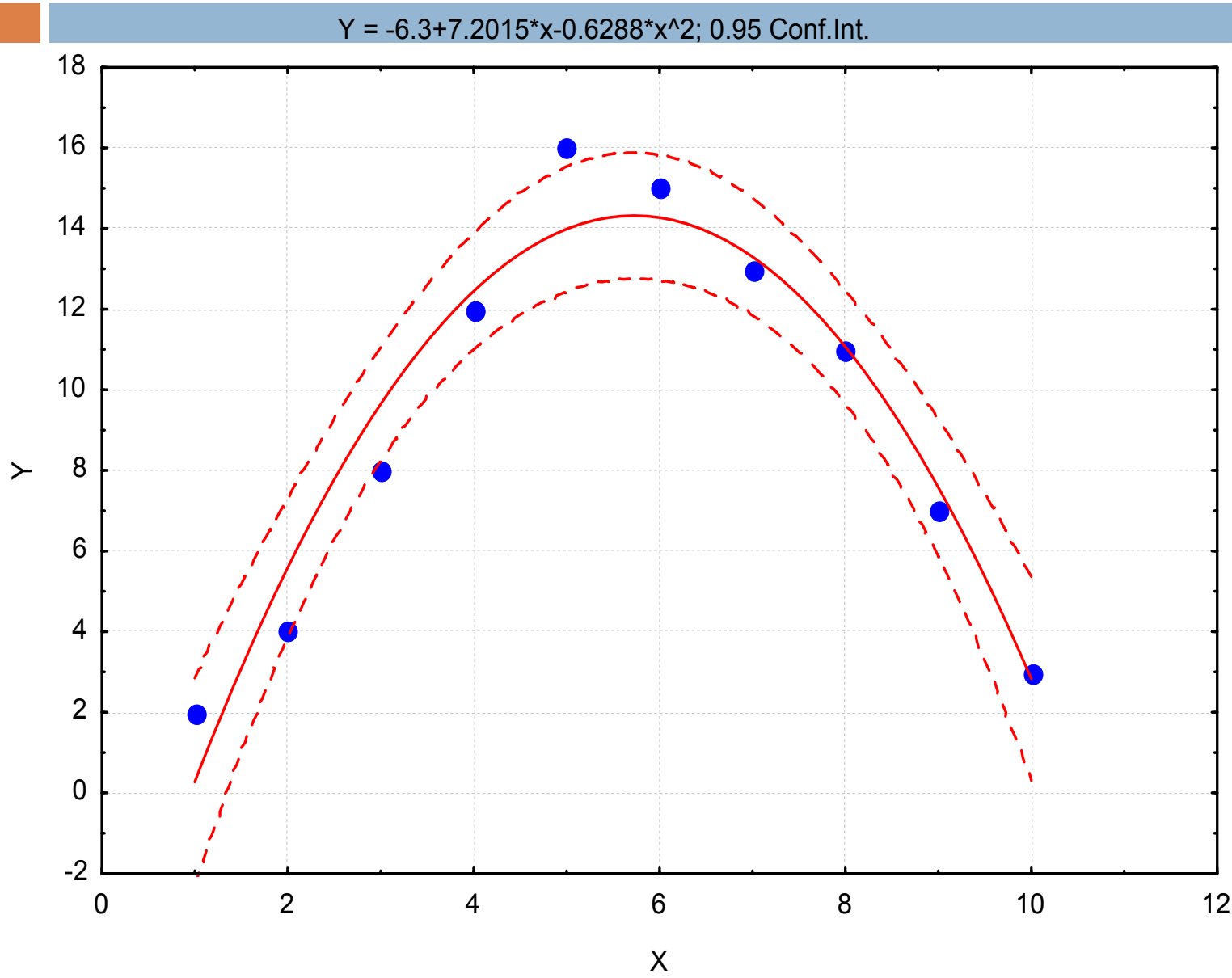




Pozor! Zvyšující se složitost nemusí znamenat lepší predikční schopnost



Polynomiální regrese



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

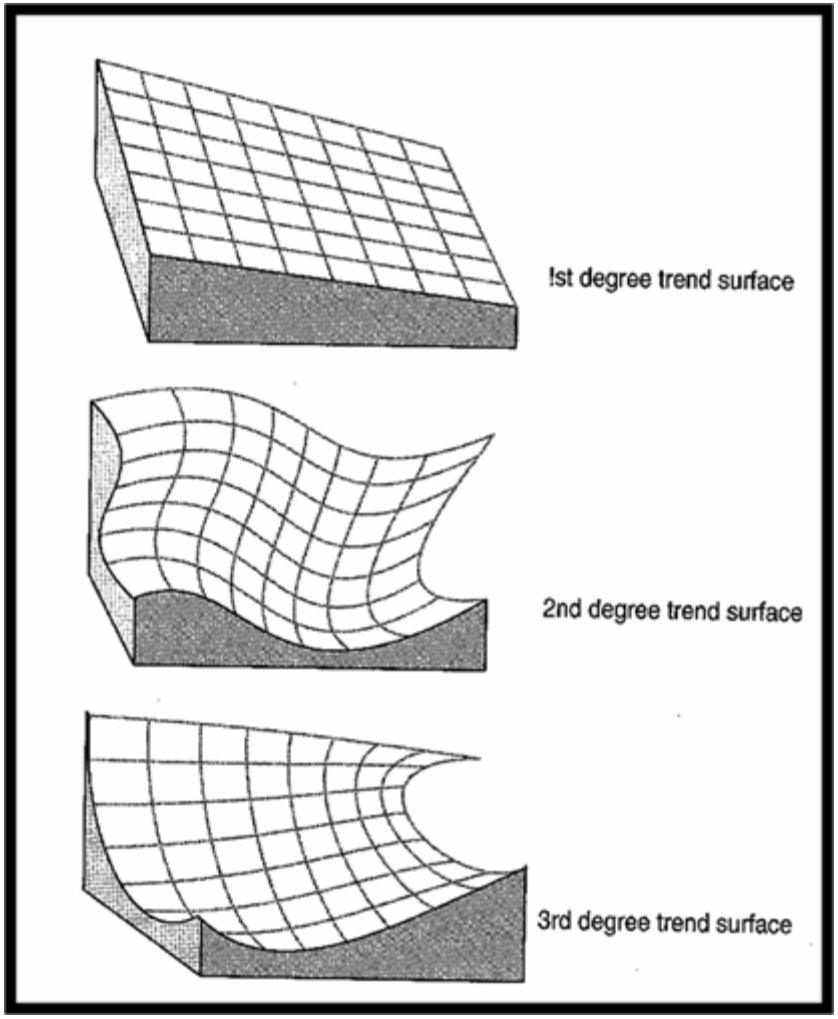
průkaznost kvadratického členu můžeme chápat jako důkaz nelinearity vztahu

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

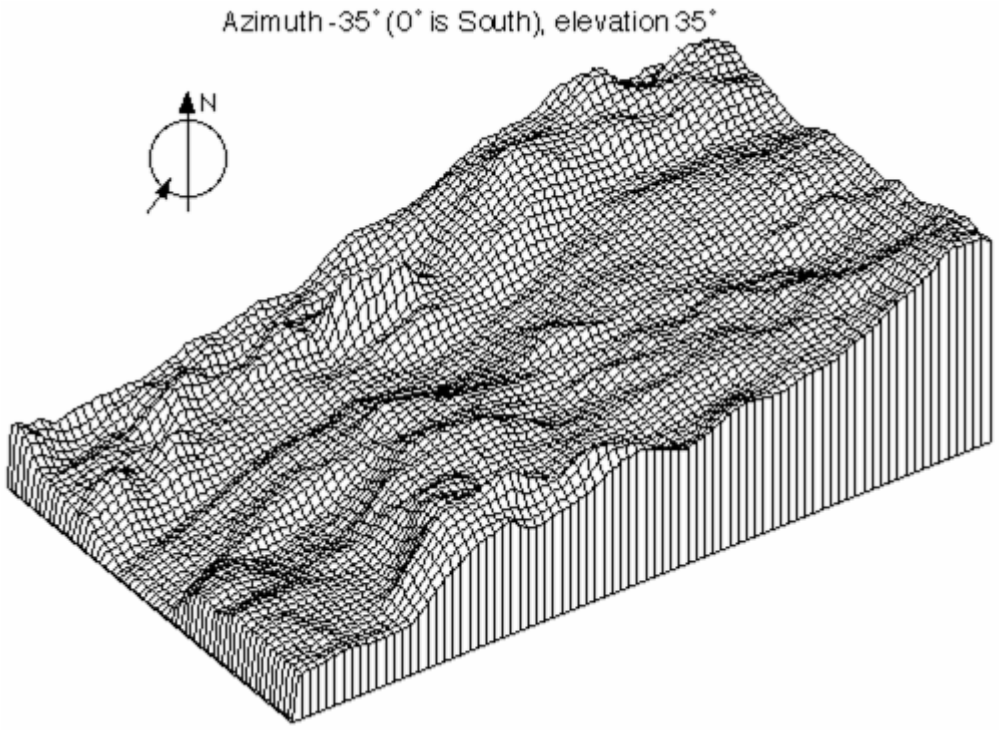


Lineární

Kvadratický

Kubický

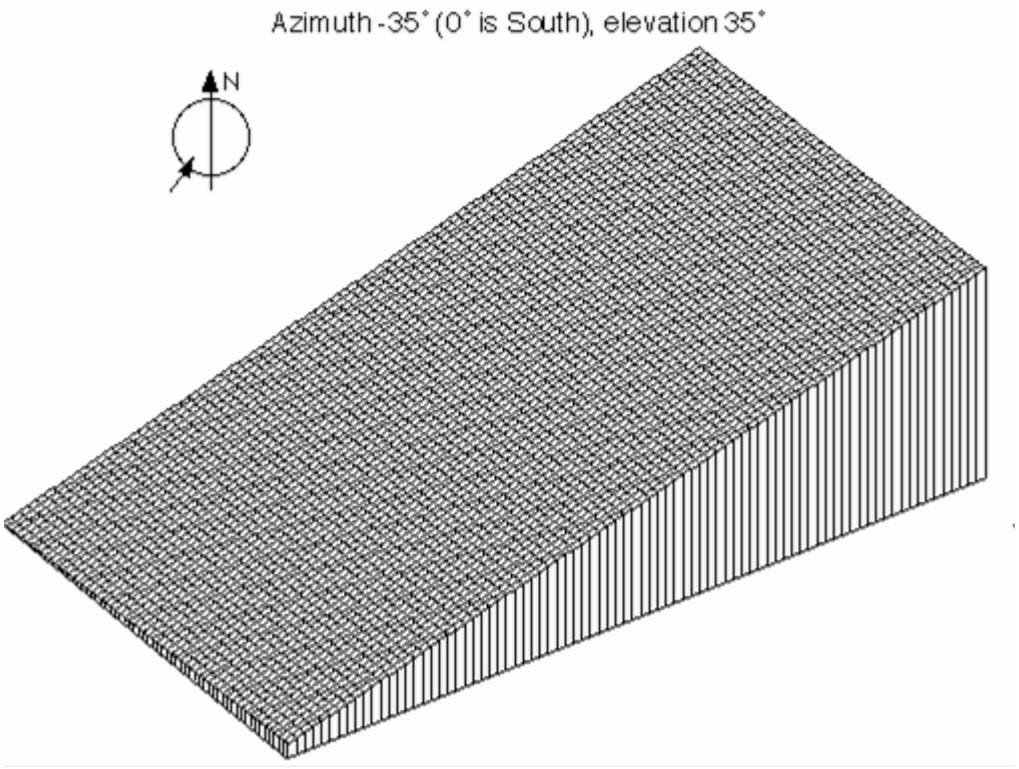
příklad



1. Globální gradient + lokální změny

http://www.kgs.ku.edu

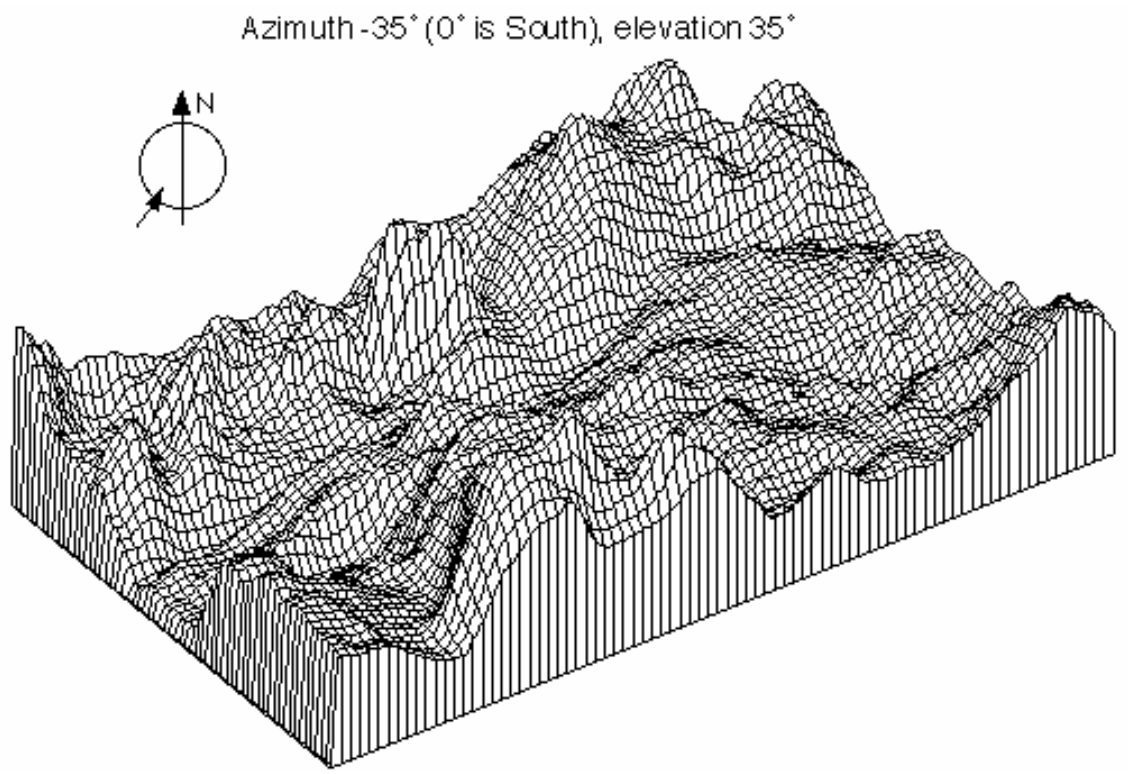
Pouze gradient



1. stupeň polynomu → gradient od východu na západ

http://www.kgs.ku.edu

Pouze lokální změny po odstranění gradientu

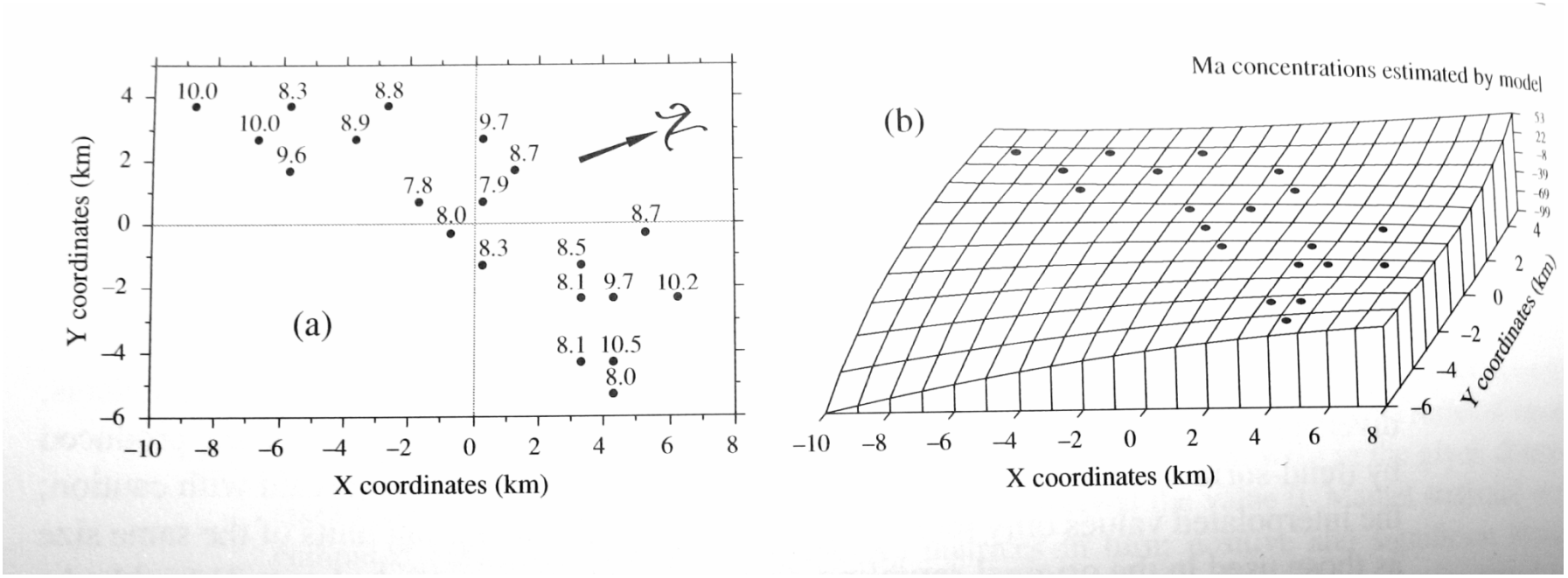


residua → lokální změny

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

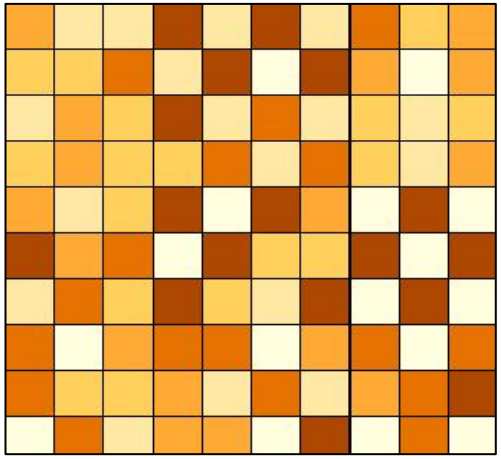
- Finální rovnice: $y = 8.13 - 0.16XY - 0.09Y^2 + 0.04X^2Y + 0.14XY^2 + 0.10Y^3$ ($R^2 = 0.81$, $p = 0.0001$)

- Používáme pouze je-li viditelná jednoduchá závislost!

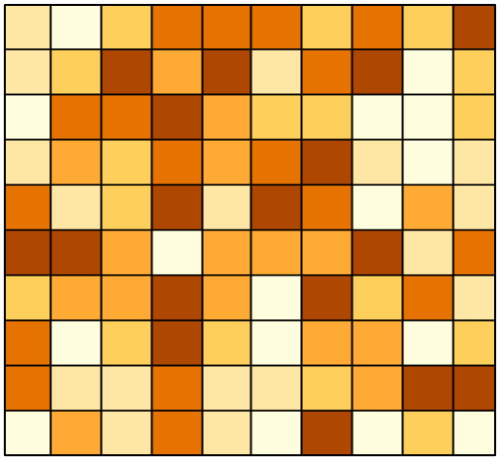
Měření prostorové autokorelace

Prostorová autokorelace

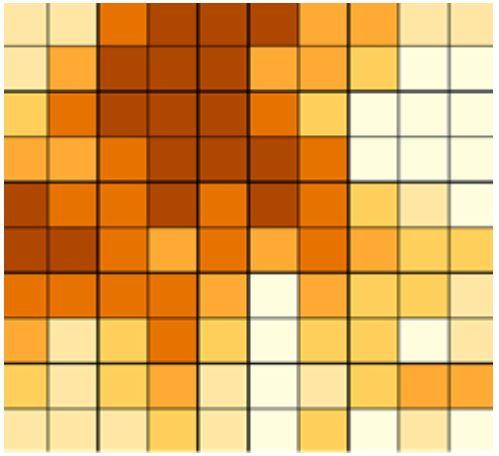
Negativní



Náhodná



Pozitivní



Měření prostorové autokorelace

Statistické měření pro zjištění prostorové autokorelace – Moranův index (I) *a* Gearyho index (*C)*

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

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

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)

Moranův index (I)

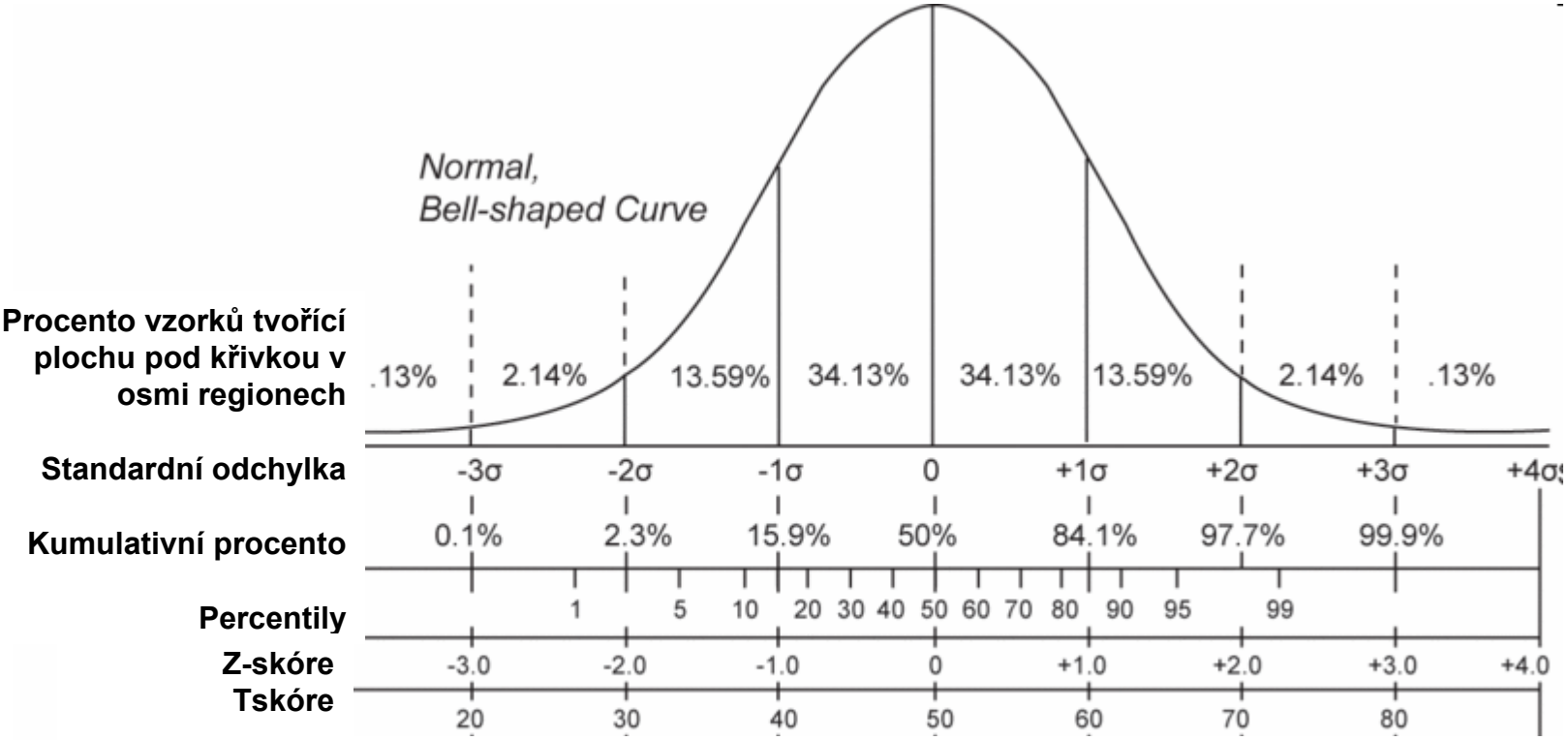
- Nulová hodnota znamená náhodnou prostorovou distribuci
- Pro testování hypotézy se hodnoty Moranova indexu transformuhí na z-skóre (hodnoty větší než 1.96 nebo menší než -1.96 → prostorová autokorelace je významná na hladině významnosti 5%

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

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

- Kriging má smysl provádět, pokud distribuce není náhodná!

Normální rozdělení

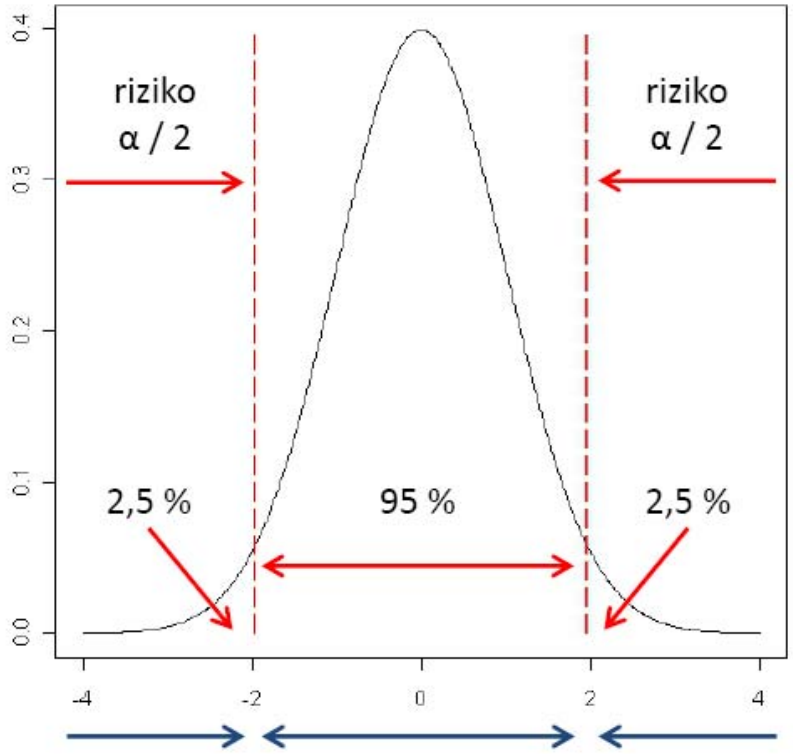


wikipedia

Zamítnutí / nezamítnutí nulové hypotézy

Oboustranný test při α = 0,05

$$H_0 : \theta_1 = \theta_2 \quad H_1 : \theta_1 \neq \theta_2$$



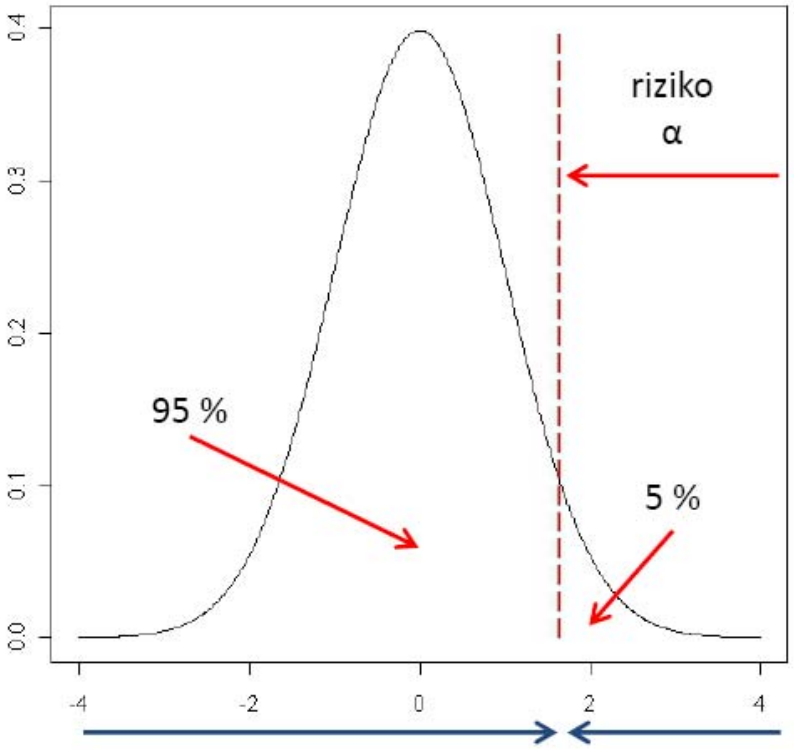
Padne-li testová statistika sem – **zamítáme H_0**

Padne-li testová statistika sem – **nezamítáme H_0**

Padne-li testová statistika sem – **zamítáme H_0**

Jednostranný test při α = 0,05

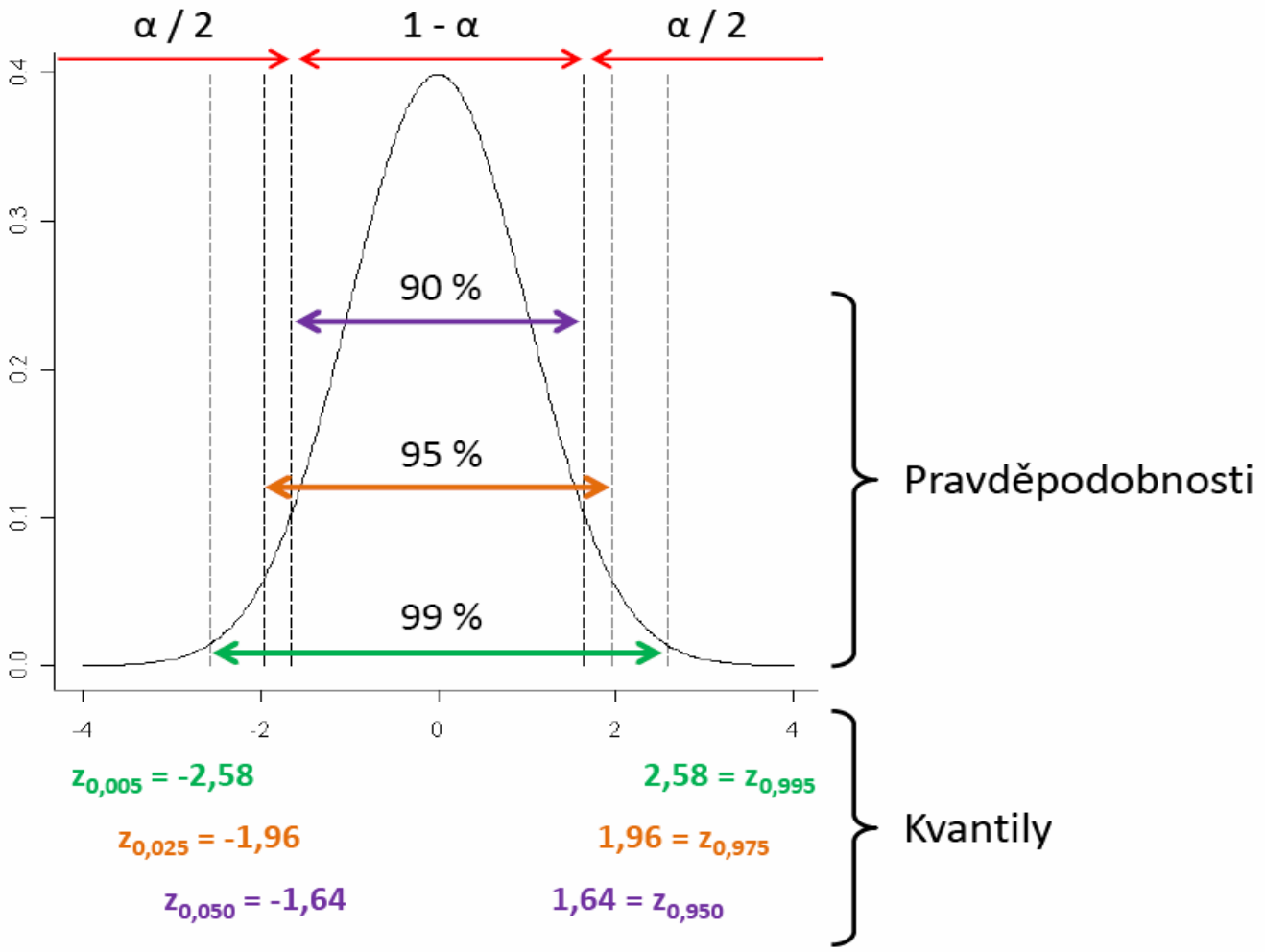
$$H_0 : \theta_1 = \theta_0 \quad H_1 : \theta_1 > \theta_0$$



Padne-li testová statistika sem – **nezamítáme H_0**

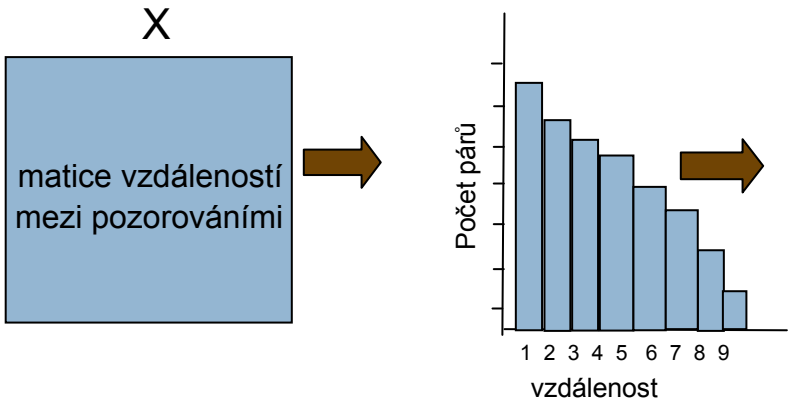
Padne-li testová statistika sem – **zamítáme H_0**

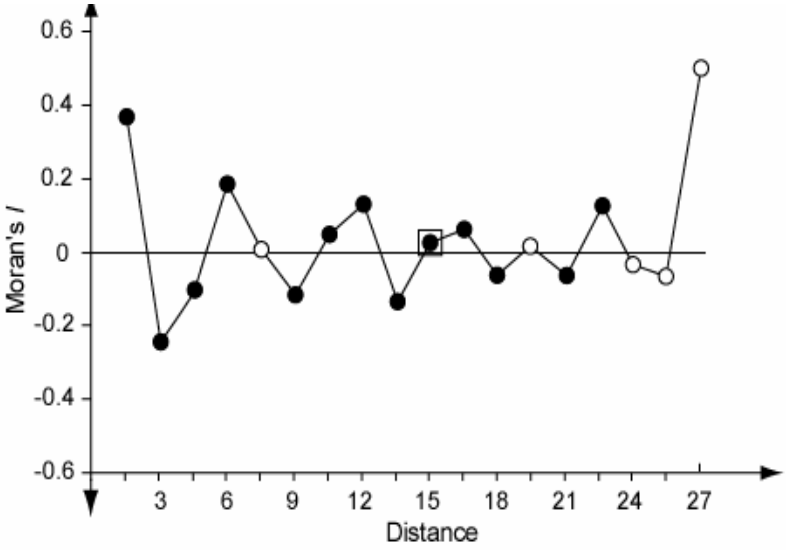
Kvantily standardizovaného normálního rozdělení



Prostorový korelogram

- *Prostorový korelogram* – autokorelační hodnoty x vzdálenosti pozorování





Výpočet indexů pro jednotlivé vzdálenosti

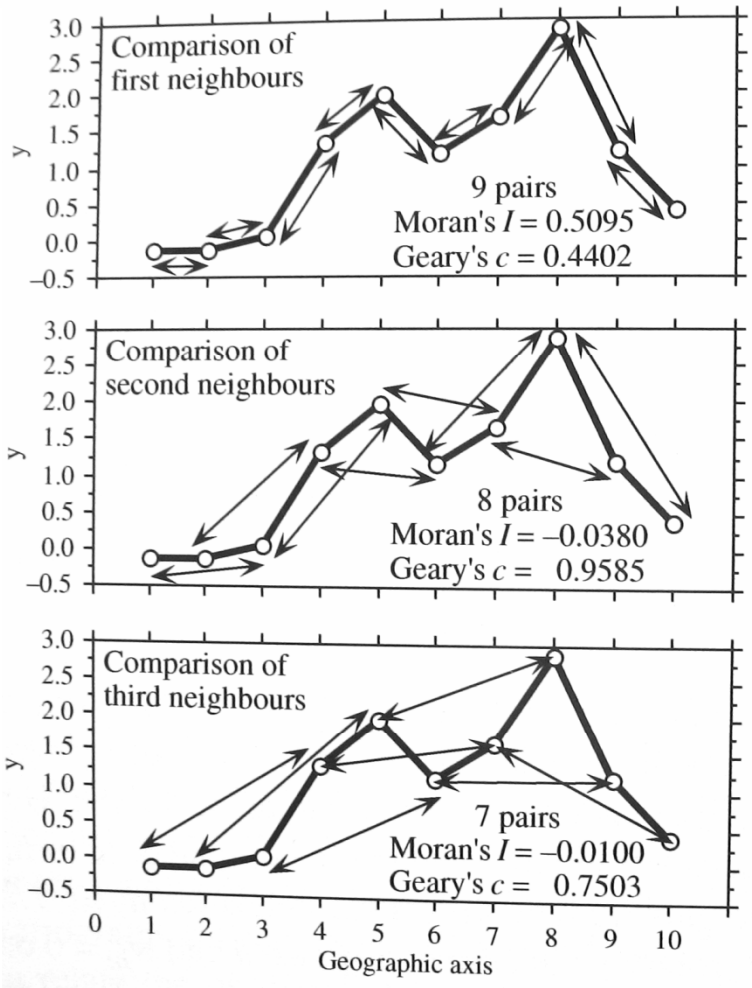
Vzdálenost

d = 1

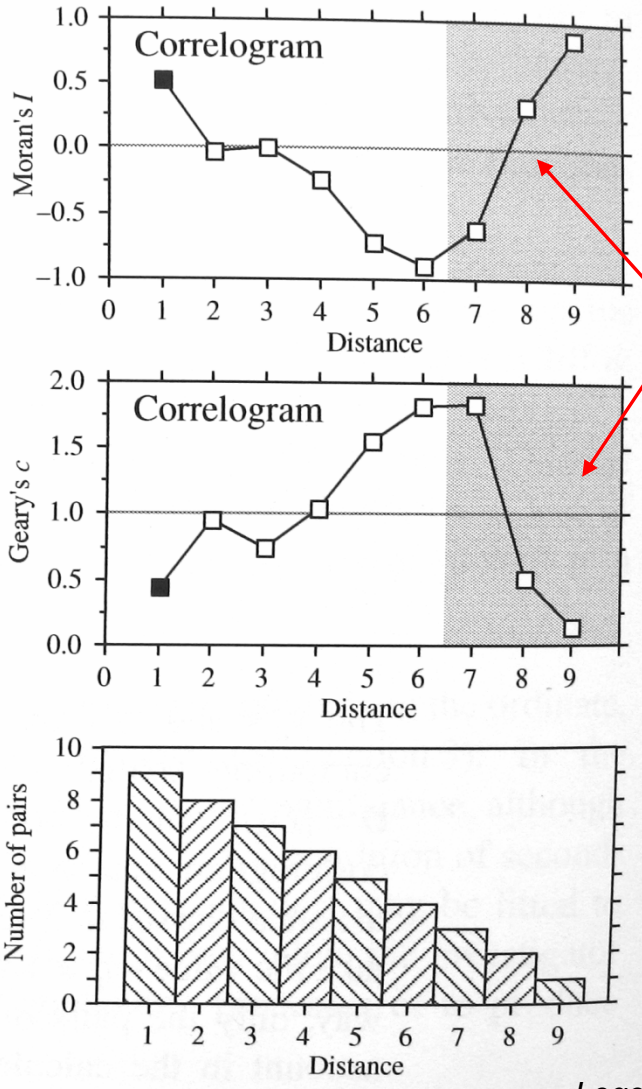
d = 2

d = 3

atd.



etc.



malé N!

Legendre, 2003

Měření prostorové autokorelace

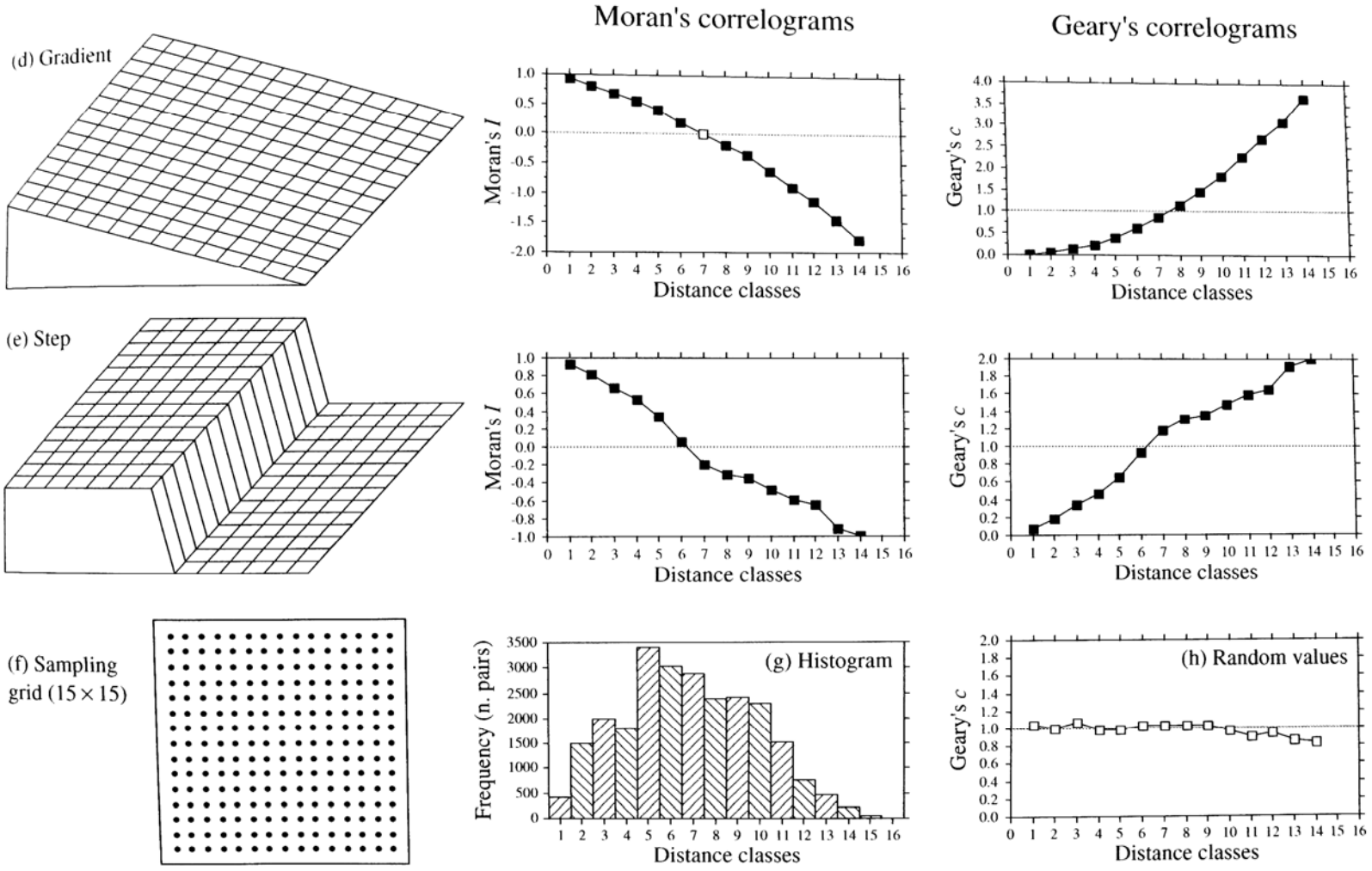
- existence autokorelace prostorových dat je obvyklá
- před výpočtem prostorových autokorelačních koeficientů je potřeba spočítat matici geografických vzdáleností mezi lokalitami $d = [D_{hi}]$
- Autokorelační koeficienty jsou spočítány pro jednotlivé vzdálenostní třídy *d*
- Váhy w_{hi} (*Kronecker 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)

Přesná ukázka výpočtu zde: http://www.gitta.info/DiscrSpatVari/en/html/spat_depend_learningObject8.html

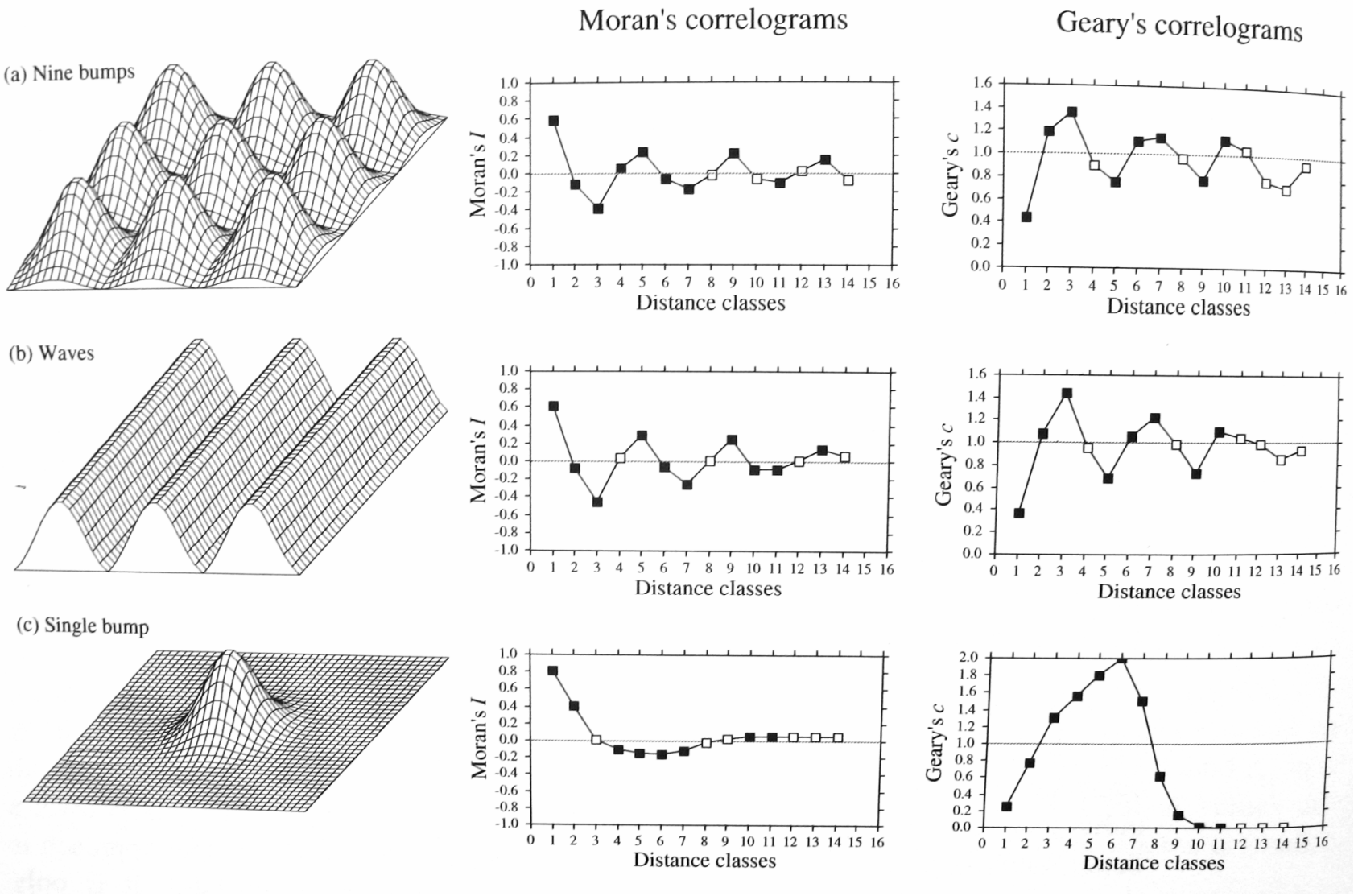
hodnoty	25	20	25	10	10	5	5	1	1
Vzdálenost (km)	1	1	1	2	2	3	3	3	3

Prostorový korelogram



Legendre, 2003

Prostorový korelogram II



Legendre, 2003

Shrnutí

- Ukázali jsme si 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

Cvičení v ArcGIS - úkol

- Každý pro svůj kraj (okres) zjistí prostorovou strukturu koncentrací Hg a Pb v půdě pomocí Moranova Indexu (soubor kovyCR.sta)
- Na základě výsledků prostorové distribuce zvolte vhodné interpolační metody a vytvořte mapu
- Úkol pošlete ve Wordu → všechny výsledky + postup a zvolená nastavení

Literatura

- HENGL, T. A Practical Guide to Geostatistical Mapping of Environmental Variables. Luxemburg: EUR 22904 EN Scientific and Technical Research series, Office for Official Publications of the European Communities, 2007. 143 s. ISBN 978-92-79-0690.

- Legendre P., Legendre L., 1998. Numerical ecology (second ed.). Elsevier, Amsterdam.

- Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659–1673.