Testování nezávislosti v prostorových modelech s kótami

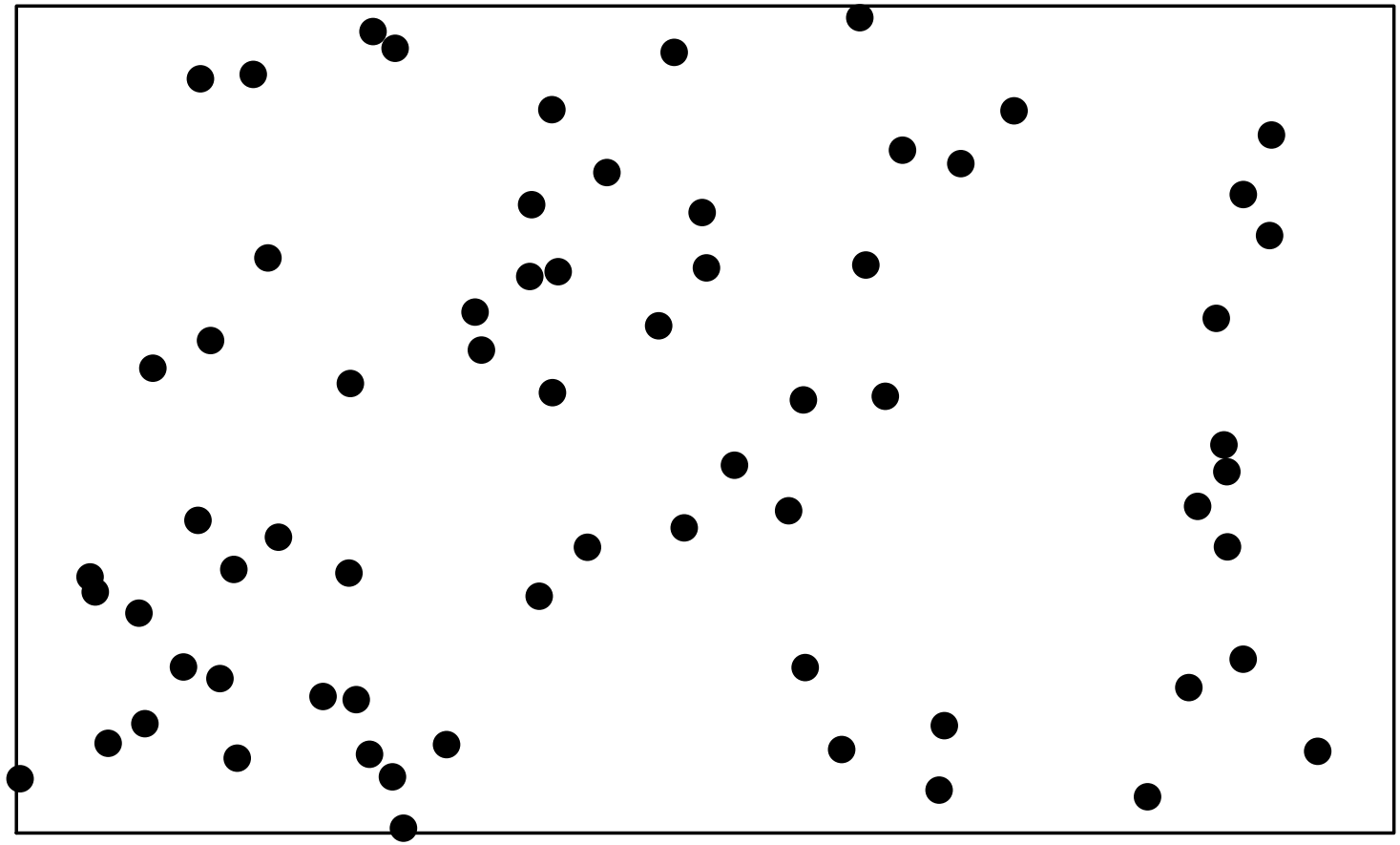
Zbyněk Pawlas

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fyzikální fakulta
Univerzita Karlova

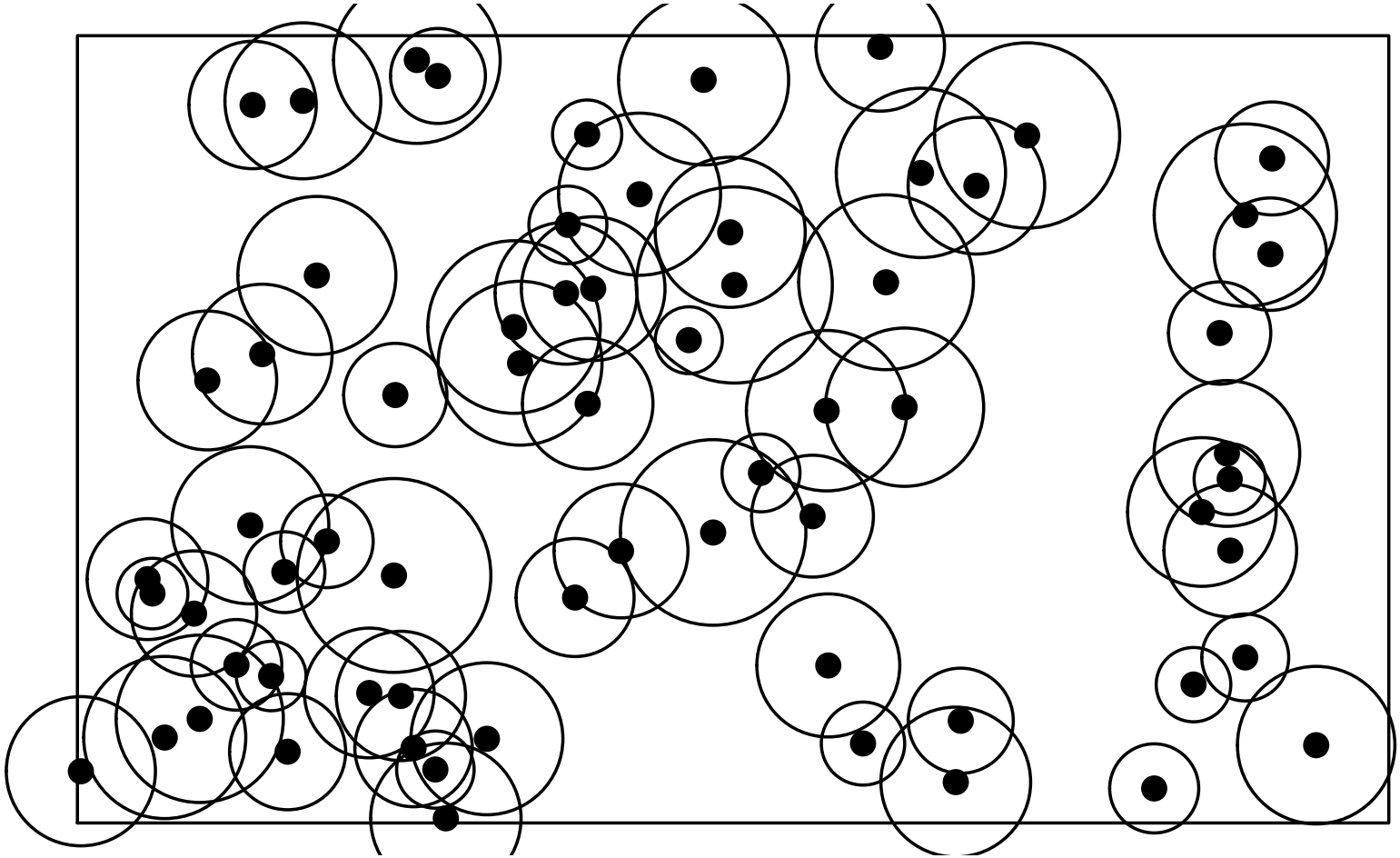
19. letní škola JČMF ROBUST 2016

15. září 2016, Loučná nad Desnou

Realizace bodového procesu

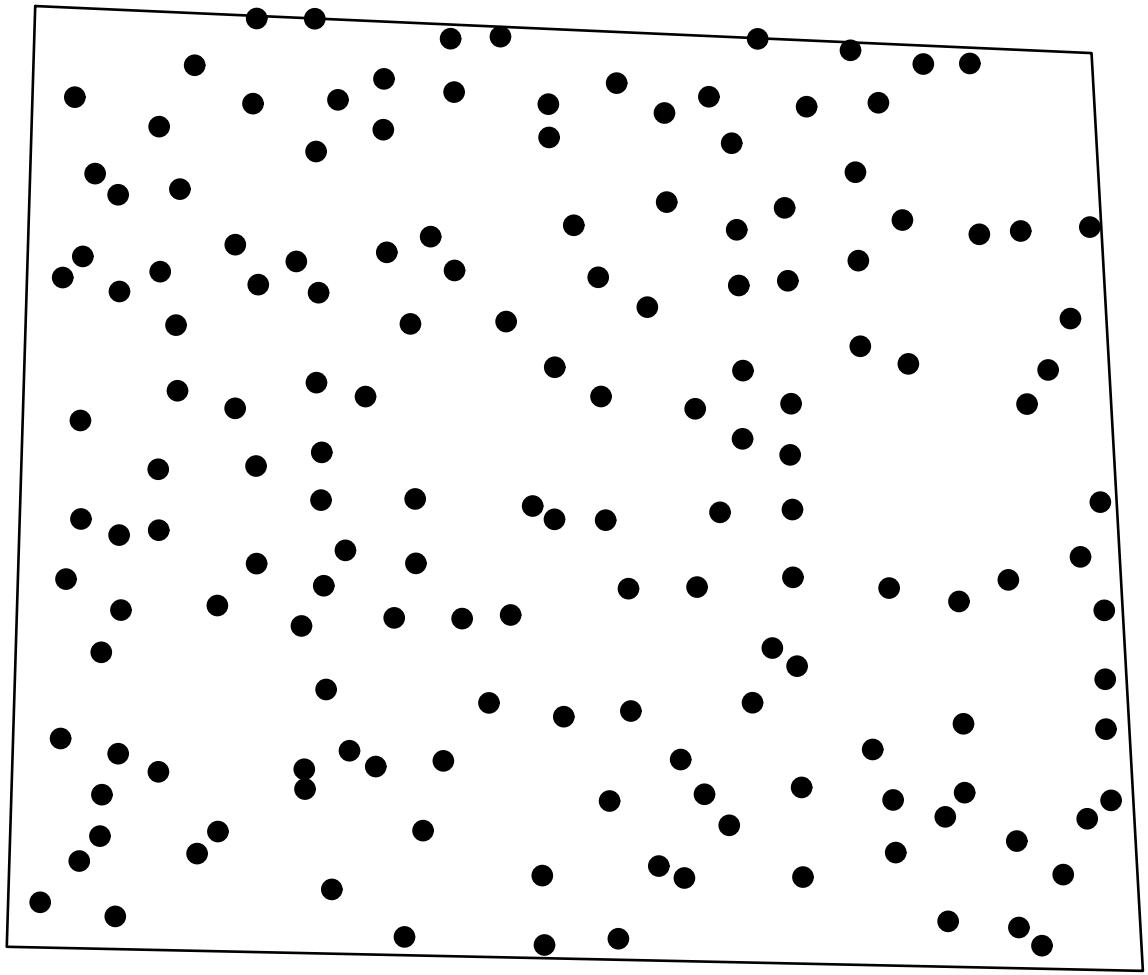


Realizace kótovaného bodového procesu



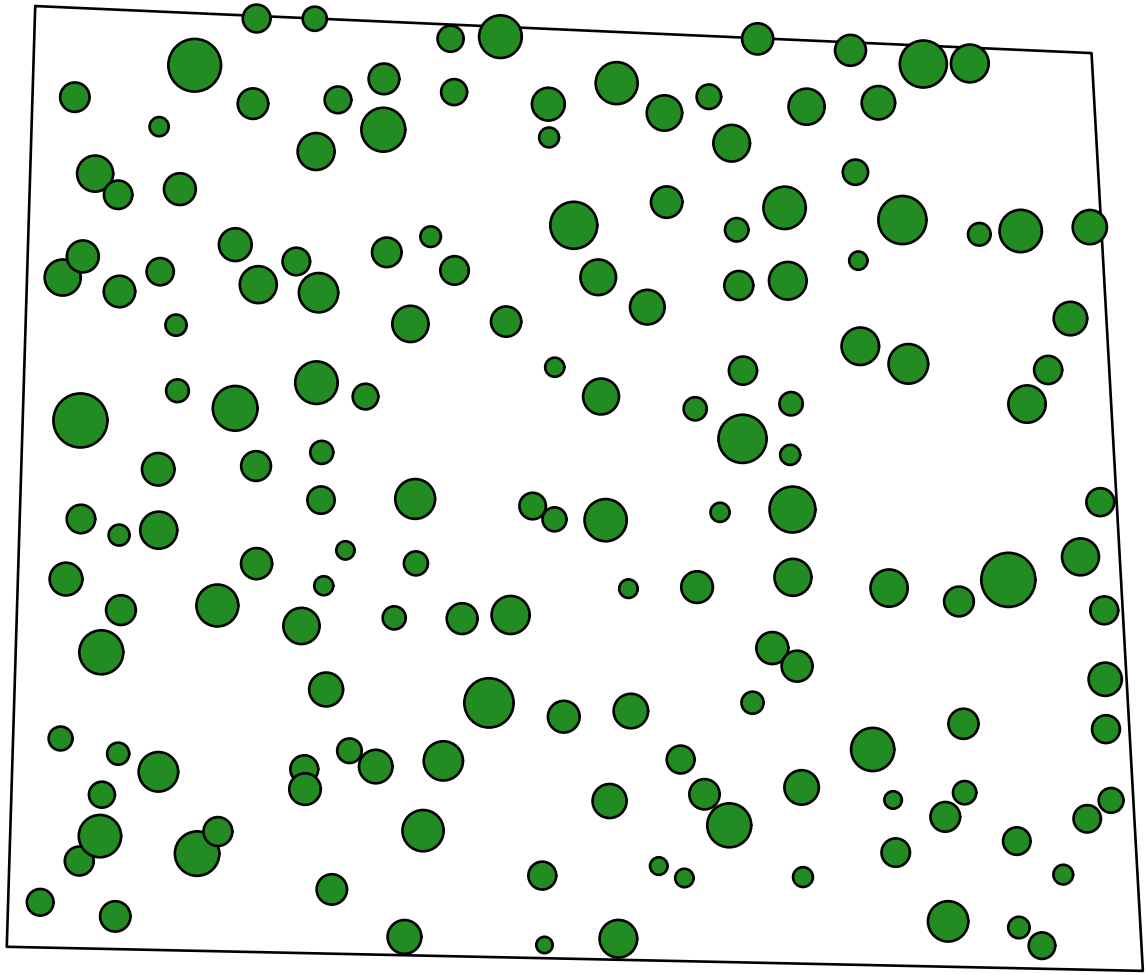
Polohy stromů

ÚHÚL, Brandýs nad Labem



Polohy a výčetní tloušťky

ÚHÚL, Brandýs nad Labem



Pradědy

Kótované bodové procesy

bodový proces

$$\Phi = \sum_{i \geq 1} \delta_{X_i}$$

$\{X_i\}$ náhodné body v $\mathbb{R}^d$
$\Phi(B) < \infty$ pro omezené borelovské $B \subseteq \mathbb{R}^d$

kótovaný bodový proces

$$\Phi_m = \sum_{i \geq 1} \delta_{(X_i, M_i)}$$

$\{M_i\}$ náhodné elementy v prostoru kót $\mathbb{M}$
$\Phi_m(B \times \mathbb{M}) < \infty$ pro omezené borelovské $B \subseteq \mathbb{R}^d$

uvažujeme kvantitativní kóty: $\mathbb{M} = \bar{\mathbb{R}}$

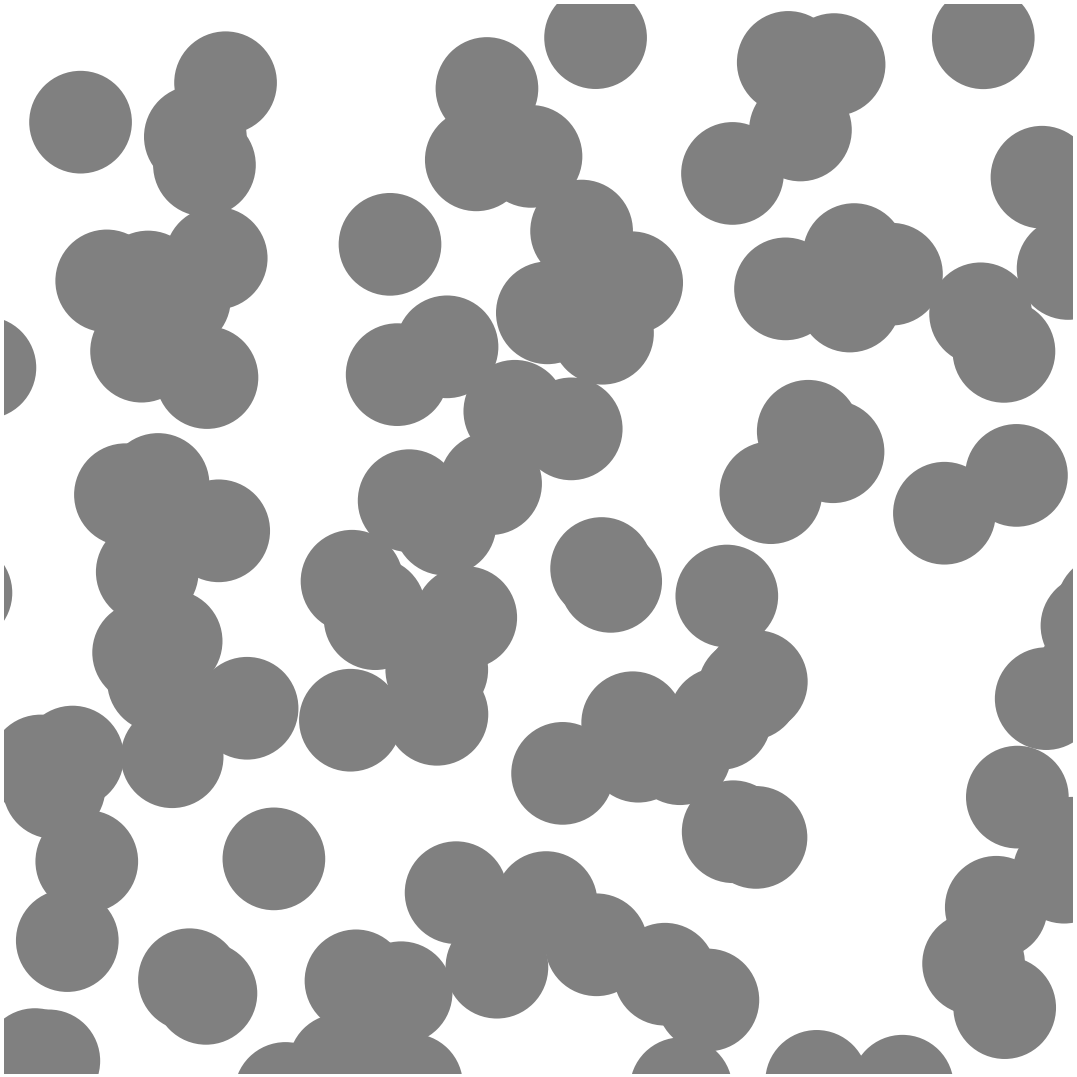
Druhy kótování

$$\Phi_m = \sum_{i \geq 1} \delta_{(X_i, M_i)}$$

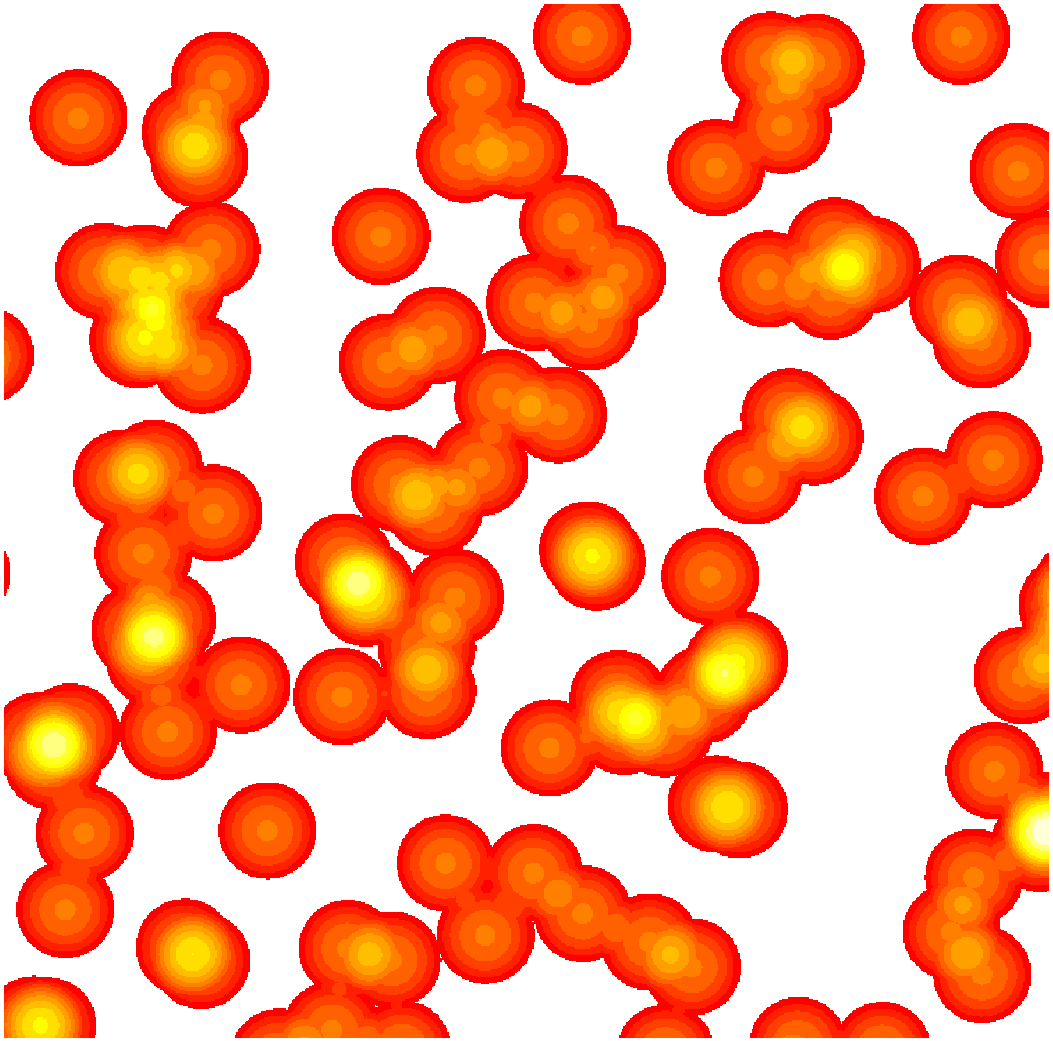
- **nezávislé kótování**
  Φ_m je nezávisle kótovaný bodový proces,
  když $\{X_i\}$ a $\{M_i\}$ jsou nezávislé a $\{M_i\}$ jsou i.i.d.

- **geostatistické kótování**
  Φ_m je geostatisticky kótovaný bodový proces,
  když $\{X_i\}$ a náhodné pole $\{M(x) : x \in \mathbb{R}^d\}$ jsou nezávislé
  a $M_i = M(X_i)$

Realizace náhodné uzavřené množiny



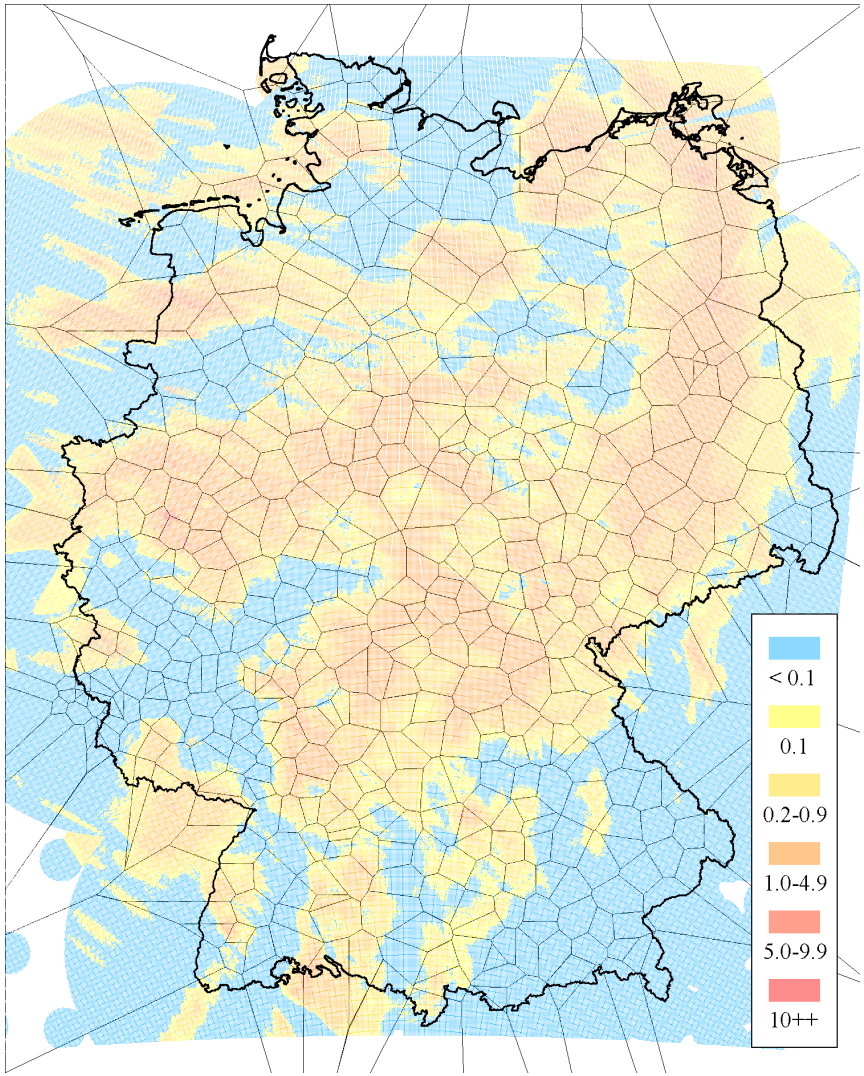
Realizace kótované náhodné uzavřené množiny



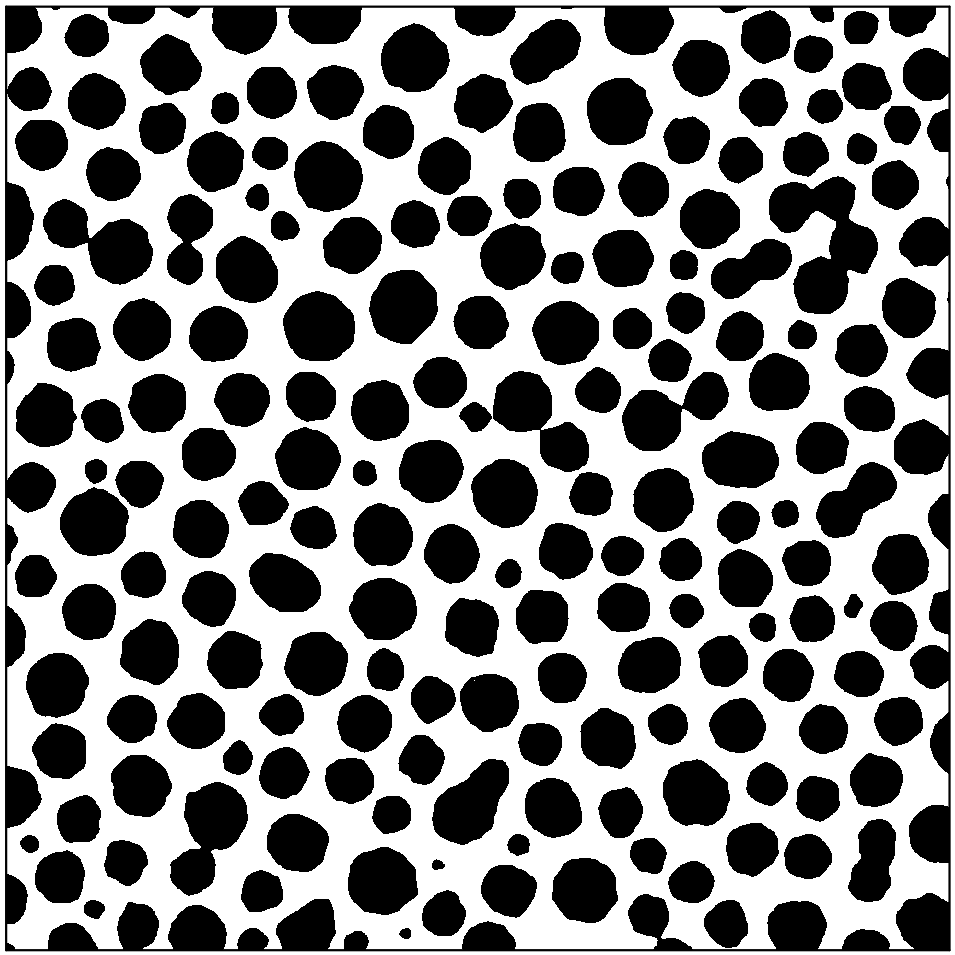
Místa srážek

Radarová data

Deutscher Wetterdienst (DWD)

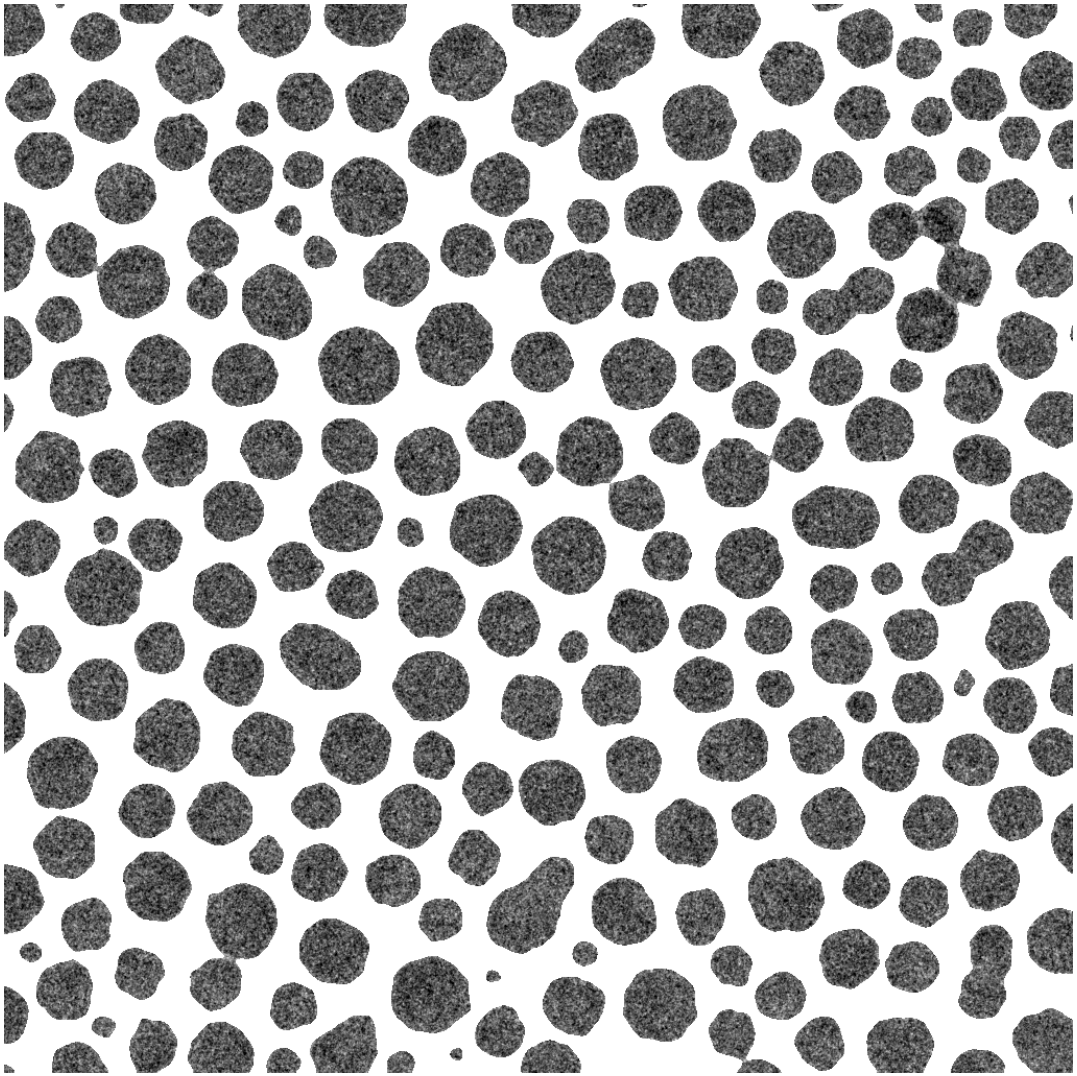


Směs dvou organických polovodičů



Materialová data

Molecular Materials and Nanosystems, Eindhoven University of Technology



Kótované náhodné uzavřené množiny

Ξ náhodná uzavřená množina v $\mathbb{R}^d$

$\Xi_m = (\Xi, M)$ kótovaná náhodná uzavřená množina

$M : \Xi \to \bar{\mathbb{R}}$ shora polospojitá
$\{(x, m) \in \Xi \times \bar{\mathbb{R}} : m \leq M(x)\}$ je náhodná uzavřená množina v $\mathbb{R}^d \times \bar{\mathbb{R}}$

Ξ_m je tzv. geostatisticky kótovaná náhodná uzavřená množina, když $\Xi_m = (\Xi, M|_\Xi)$, kde množina Ξ a náhodné pole $\{M(x) : x \in \mathbb{R}^d\}$ jsou nezávislé

Speciální případ

je-li $\Phi = \sum_{i \geq 1} \delta_{X_i}$ bodový proces, potom $\Xi = \cup_{i \geq 1} X_i$ je náhodná uzavřená množina

je-li $\Phi_m = \sum_{i \geq 1} \delta_{(X_i, M_i)}$ kótovaný bodový proces, $\Xi = \cup_{i \geq 1} X_i$ a $M : \Xi \to \bar{\mathbb{R}}$ taková, že $M(X_i) = M_i$, potom $\Xi_m = (\Xi, M)$ je kótovaná náhodná uzavřená množina

Stacionarita

Kótovaný bodový proces Φ_m je stacionární, pokud

$$\Phi_m = \sum_{i \geq 1} \delta_{(X_i, M_i)} \quad \text{a} \quad \Phi_m + y = \sum_{i \geq 1} \delta_{(X_i + y, M_i)}$$

mají stejné rozdělení pro každé $y \in \mathbb{R}^d$.

$\lambda = \mathbb{E}\Phi_m([0,1]^d \times \bar{\mathbb{R}}) = \mathbb{E}\Phi([0,1]^d)$ je intenzita procesu Φ_m (a také procesu Φ)

$\mathbb{Q}(U) = \frac{\mathbb{E}\Phi_m(B \times U)}{\mathbb{E}\Phi_m(B \times \bar{\mathbb{R}})}$ je stacionární rozdělení kót

Kótovaná náhodná uzavřená množina Ξ_m je stacionární, pokud $\Xi_m = (\Xi, M)$ a $\Xi_m + y = (\Xi + y, M_y)$ mají stejné rozdělení pro každé $y \in \mathbb{R}^d$, kde $M_y(x) = M(x - y)$.

Objemový podíl

Předpokládejme, že

$$0 < \lambda = \mathbb{E}\mathcal{H}^k(\Xi, M) \cap ([0,1]^d \times \bar{\mathbb{R}})) = \mathbb{E}\mathcal{H}^k(\Xi \cap [0,1]^d) < \infty$$

pro nějaké $k \in \{0, 1, \ldots, d\}$.

Tuto konstantu λ nazveme *k-objemový podíl* kótované množiny Ξ_m (a také množiny Ξ).

Pro $k = 0$ jde o intenzitu kótovaného bodového procesu.

$$\mathbb{Q}(U) = \frac{\mathbb{E}\mathcal{H}^k((\Xi, M) \cap (B \times U))}{\mathbb{E}\mathcal{H}^k((\Xi, M) \cap (B \times \bar{\mathbb{R}}))}$$

je *stacionární rozdělení kót* kótované množiny Ξ_m

Test nezávislosti poloh a kót

Cíl: otestovat nulovou hypotézu geostatistického kótování

Pokud jsou polohy bodů (bodový proces) nebo oblasti výskytu (náhodná uzavřená množina) nezávislé na kótách, pak lze k analýze a modelování obou složek přistupovat zvlášť.

Alternativa: kóty závisejí na polohách, typicky na lokální hustotě bodů/množiny v daném místě

Problémy:

- volba vhodné testové statistiky
- rozdělení testové statistiky za nulové hypotézy

K-funkce

testovou statistiku můžeme založit na kumulativní popisné charakteristice druhého řádu, tzv. K-funkci

stacionární náhodná uzavřená množina Ξ s k-objemovým podílem λ:

$$K(r) = \frac{1}{\lambda^2 |A|} \mathbb{E} \int_{\Xi} \int_{\Xi \cap A} \mathbf{1}\{0 < \|x - y\| \leq r\} \, \mathcal{H}^k(\mathrm{d}x) \, \mathcal{H}^k(\mathrm{d}y)$$

stacionární kótovaná náhodná uzavřená množina Ξ_m s k-objemovým podílem λ:

$$K_t(r) = \frac{1}{\lambda^2 c_t |A|} \mathbb{E} \int_{\Xi} \int_{\Xi \cap A} t(M(x), M(y)) \times \mathbf{1}\{0 < \|x - y\| \leq r\} \, \mathcal{H}^k(\mathrm{d}x) \, \mathcal{H}^k(\mathrm{d}y),$$

kde $c_t = \int \int t(m_1, m_2) \, \mathbb{Q}(\mathrm{d}m_1) \mathbb{Q}(\mathrm{d}m_2)$

K-funkce za nulové hypotézy

volme $t(m_1, m_2) = m_1$:

$$K_t(r) = K_{m\cdot}(r) = \frac{1}{\lambda^2 \mu |A|} \mathbb{E} \int_{\Xi \cap A} M(x) \mathcal{H}^k(b(x,r) \cap \Xi) \, \mathcal{H}^k(\mathrm{d}x),$$

kde $\mu = \int m \, \mathbb{Q}(\mathrm{d}m)$ je střední kóta

pro geostatisticky kótované modely platí

$$K_{m\cdot}(r) = K(r)$$

Odhad K-funkce

realizace modelu pozorovaná v omezeném okně $W \subseteq \mathbb{R}^d$

$$\widehat{\lambda K_{m\cdot}(r)} = \frac{\int_{\Xi \cap W} M(x) \mathcal{H}^k(b(x,r) \cap \Xi)\, \mathcal{H}^k(\mathrm{d}x)}{\int_{\Xi \cap W} M(x)\, \mathcal{H}^k(\mathrm{d}x)},$$

je podílově nestranný odhad $\lambda K_{m\cdot}(r)$

testová množina $\{\xi_1, \ldots, \xi_N\}$ ve W

$$\widehat{\lambda K_{m\cdot}(r)} = \frac{\sum_{i=1}^{N} M(\xi_i) \mathbf{1}\{\xi_i \in \Xi\} \mathcal{H}^k(\Xi \cap b(\xi_i, r))}{\sum_{i=1}^{N} M(\xi_i) \mathbf{1}\{\xi_i \in \Xi\}}$$

rozdělení $\widehat{\lambda K_{m\cdot}(r)}$ komplikované

Monte Carlo testy

většina testů v prostorové statistice je založena na simulacích

popisná funkcionální charakteristika $T(r)$

empirická charakteristika $\hat{T}_0(r)$ odhadnutá z dat

porovnání $\hat{T}_0(r)$ s odhady $\hat{T}_1(r), \ldots, \hat{T}_q(r)$ ze simulací z nulového modelu

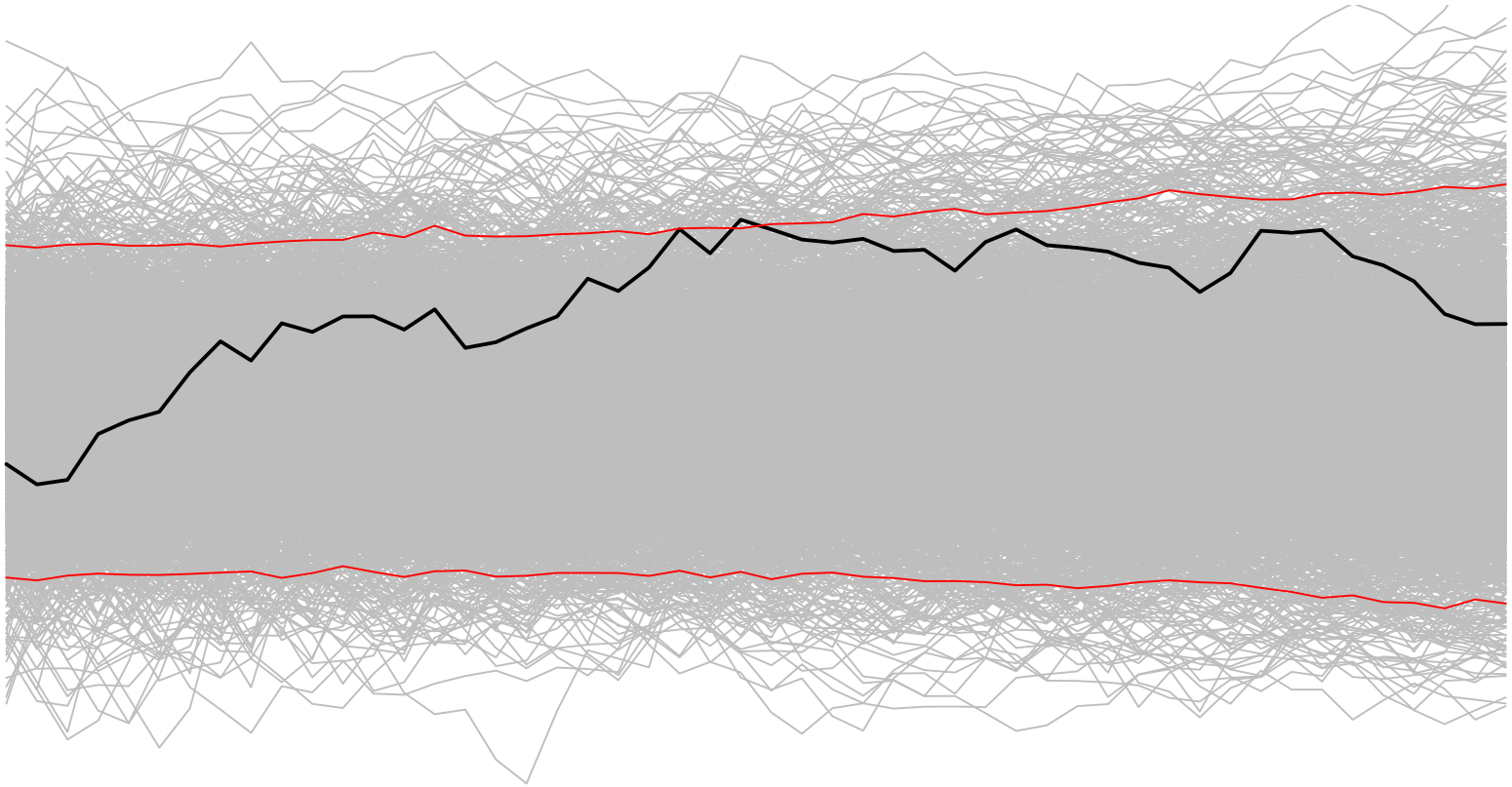
$$T_{\text{low}}(r) = \min^{(k)} \hat{T}_i(r), \quad T_{\text{upp}}(r) = \max^{(k)} \hat{T}_i(r)$$

H_0 zamítneme, když $\hat{T}_0(r) \notin (T_{\text{low}}(r), T_{\text{upp}}(r))$

$k = (q+1)\alpha/2$

bodové obálky, problém mnohonásobného testování

Bodové obálky



Globální pořádkové obálkové testy

M. Myllymäki, T. Mrkvička, P. Grabarnik, H. Seijo, U. Hahn (2016) – JRSS Ser. B

$\hat{T}_0(r), \ldots, \hat{T}_q(r)$ spočteno pro $r_1, \ldots, r_\ell$

pro každé $r \in \{r_1, \ldots, r_\ell\}$:

$$T_{(0)}(r) \leq T_{(1)}(r) \leq \cdots \leq T_{(q)}(r)$$

bodové pořadí

$$R_{(0)}(r) = 1, R_{(q)}(r) = 1, R_{(1)}(r) = 2, R_{(q-1)}(r) = 2, \ldots$$

globální pořadí

$$R_i = \min\{R_i(r), r \in \{r_1, \ldots, r_\ell\}\}$$

Globální pořádkové obálkové testy

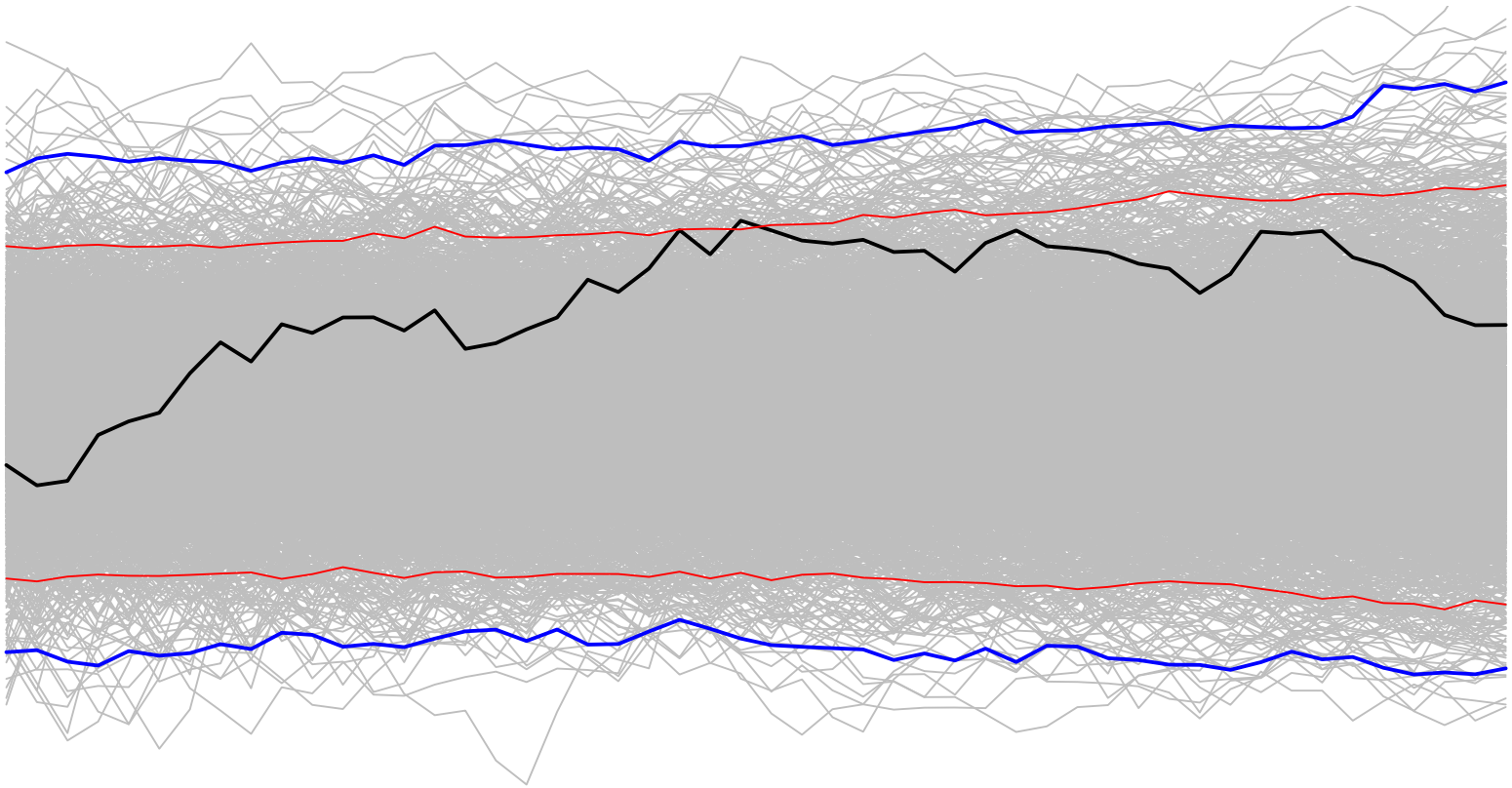
$$p_{\text{low}} = 1 - \frac{1}{q+1} \sum_{i=1}^{q} \mathbf{1}\{R_i \geq R_0\},$$

$$p_{\text{upp}} = 1 - \frac{1}{q+1} \sum_{i=1}^{q} \mathbf{1}\{R_i > R_0\}$$

H_0 zamítneme, když $(p_{\text{upp}} + p_{\text{low}})/2 < \alpha$

konzervativněji: $p_{\text{upp}} < \alpha$

Globální pořádkové obálky



Simulace z nulové hypotézy

potřebujeme vygenerovat q nových výběrů

Problém: Jak generovat simulace z nulové hypotézy?

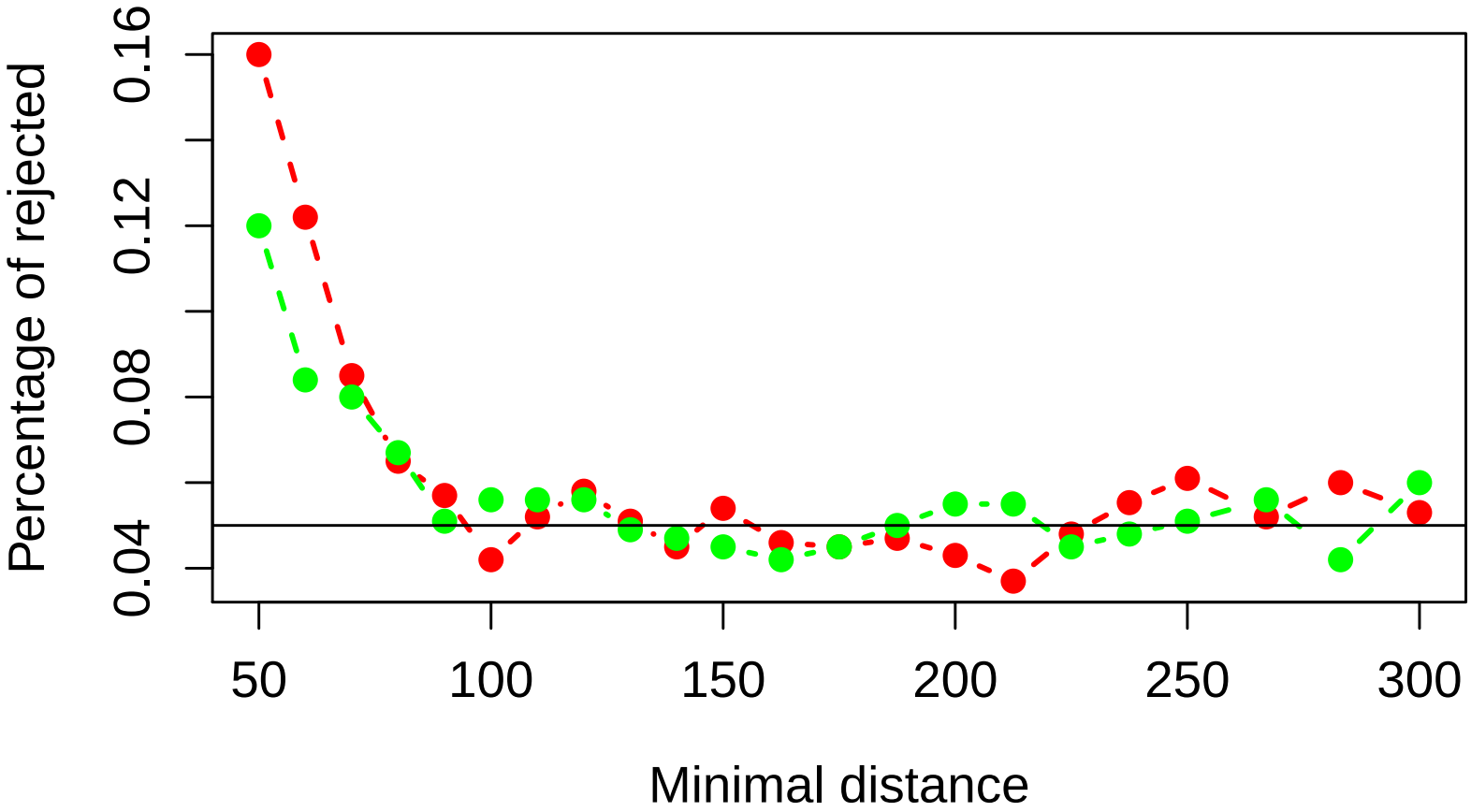
Při neparametrickém přístupu obvykle nějaký resampling.

V našem případě tzv. random reallocation.

odhad $\widehat{K_{m\cdot}(r)}$ na základě $\xi_1, \ldots, \xi_N$
ponecháme Ξ a náhodně zpermutujeme hodnoty
$\Gamma(\xi_1), \ldots, \Gamma(\xi_n)$

POZOR: obecně nedostáváme nezávislé výběry
korelace mezi $\Gamma(\xi_1), \ldots, \Gamma(\xi_n)$ můžou způsobit problémy

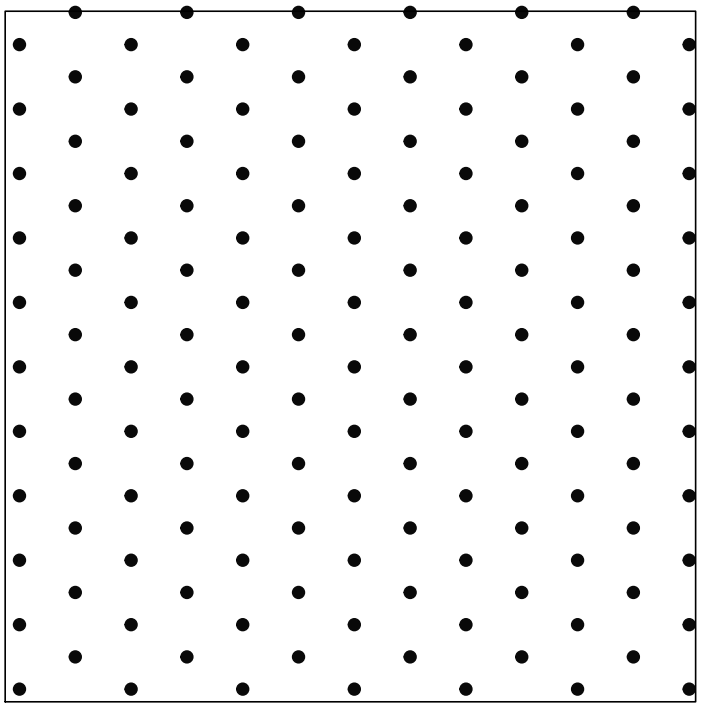
Empirická hladina testu



Volba testových bodů

- potřebujeme $\Gamma(\xi_1), \ldots, \Gamma(\xi_n)$ nekorelované $\Rightarrow$ vzdálenosti mezi $\xi_1, \ldots, \xi_N$ větší než dosah korelací kót
- větší počet testových bodů zvyšuje kvalitu odhadů

vhodná volba: hexagonální mřížka (v rovině)



Výsledky pro reálná data

meteorologická data:

$$p = 0{,}0007; \qquad p = 0{,}001$$

materiálová data:

$$p = 0{,}124; \qquad p = 0{,}727$$

Děkuji za pozornost