INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 31, číslo 4, prosinec 2020

ROZLIŠOVÁNÍ REALIZACÍ NÁHODNÝCH MNOŽIN
DISTINGUISHING BETWEEN REALISATIONS OF RANDOM SETS

Kateřina Helisová[1], **Jakub Staněk**[2]

Adresa: [1]FEL ČVUT v Praze, Technická 2, 166 27 Praha 6; [2]MFF UK, Sokolovská 83, 186 75 Praha 8

E-mail: heliskat@fel.cvut.cz, stanekj@karlin.mff.cuni.cz

Abstrakt: Článek je přehledem dosud zkoumaných a publikovaných metod rozlišování dvou realizací náhodných množin ve smyslu rozhodnutí, zda jsou si realizace (ne)podobné svými základními rysy. Metody se zaměřují na různá pojetí podobnosti realizací, přičemž všechny jsou založené na dvou náhodných výběrech funkcí – po jednom z každé realizace – popisujících specifické rysy komponent. Shoda pravděpodobnostních rozdělení těchto funkcí je pak testována a realizace jsou považovány za podobné, jestliže shoda není zamítnuta. Metody jsou zde stručně popsány a numericky ilustrovány na simulační studii. Na závěr je pak uvedeno srovnání výhod a nevýhod jednotlivých metod.
Klíčová slova: Morfologický skelet, $\mathcal{N}$-vzdálenost, náhodná množina, obálkový test, opěrná funkce, podobnost, realizace, sousedství komponenty, spojitá komponenta.

Abstract: The paper concerns an overview of recently developed methods for comparing two realisations of random sets in the sense of deciding whether the basic features of the realisations are (dis)similar. The methods concern different definitions of similarity while all of them are based on deriving a sample of functions from each realisation, which describe the shapes or even the mutual positions of the realisation components. The equality of probability distributions of the functions is then tested, and the realisations are considered to be similar if the eguality is accepted. The methods are briefly described, their advantages and disadvantages are mentioned, and the conclusions are justified by a simulation study.
Keywords: Morphological skeleton, $\mathcal{N}$-distance, Random set, Envelope test, Support function, Similarity, Realisation, Neighborhood, Connected component.

1. Úvod

Realizaci náhodné množiny si můžeme představit jako geometrické znázornění objektu, jehož tvar je výsledkem nějakého náhodného procesu. Takovým

objektem může být třeba rostlinný porost, výskyt konkrétního nerostu v hornině, shluk buněk v tkáni apod. Matematické a zvláště pak statistické studie náhodných množin, tedy procesů, kterými jsou výše popsané objekty vytvářeny, nám mohou přinést spoustu užitečných informací o chování těchto procesů, a tudíž i o možnostech jejich regulace a praktického využití.

V posledních letech se modelování a statistické analýzy náhodných množin staly velice populárními, a to zvláště díky rozvoji výpočetní techniky umožňující rozsáhlé simulační studie, na nichž jsou tyto analýzy často založeny. Z teoretického hlediska se jimi zabývali např. [12], [13] nebo [20], jiní pak zkoumali jejich aplikace např. v biologii [15], lékařství [8], materiálových vědách [19] a dalších oblastech.

Obvykle se pro realizaci náhodné množiny (v praxi pak většinou binární obrázek, který je aproximací této realizace v daném rozlišení) snažíme nalézt vhodný model. Občas ale znalost konkrétního modelu není nutná, neboť je našim cílem porovnání dvou či více realizací ve smyslu rozhodnutí, zda realizace pocházejí ze stejného procesu, viz např. rozlišování tkáně zasažené buďto zhoubným nádorem nebo mastopatií [16].

A právě takovým porovnáním se zabývají níže popsané metody. Na základě podobnosti dvou realizací rozhodují, zda příslušné dva objekty mohou pocházet ze stejného vytvářecího zdroje (např. rakovinové bujení vs. nezhoubný nádor, zdravý rostlinný porost vs. porost zasažený škůdcem apod.). Jen poznamenejme, že samotné přiřazení realizací ke konkrétním procesům tento článek neřeší, to je úloha klasifikace. Zde se zabýváme pouze otázkou, zda jsou si dvě realizace podobné či nikoliv. Podobnost je zde definovaná jako shoda pravděpodobnostních rozdělení skupiny funkcí, které množinu popisují. Pro různé metody jsou však tyto funkce různé, neboť v různých praktických situacích nás zajímají rozdílné aspekty (např. v případě tkáně nás zajímá hlavně tvar buněk, zatímco v případě rostlinných porostů i vzájemná poloha shluků rostlin), tudíž některé metody mohou dvě realizace považovat za podobné a jiné je mohou rozlišit. Našim cílem tedy není konstrukce testů shody pravděpodobnostních rozdělení náhodných množin, z níž realizace pocházejí, nýbrž testování podobnosti realizací dle odpovídající definice. Rozlišením dvou realizací pak máme na mysli situaci, kdy test podobnost realizací zamítne.

Přestože existují klasické nástroje pro popis náhodných množin jako např. kovarianční funkce [1], kontaktní distribuční funkce [1], funkce na morfologických operacích (dilatace, eroze, otevírání a zavírání množiny) [20], granulometrická funkce [20] atd., jsou situace, kdy tyto charakteristiky nemusí být k rozlišení dvou realizací dostatečné. Všechny navíc mají tu nevýhodu, že pro jednu realizaci vždy obdržíme pouze jeden odhad dané funkce, takže

srovnání realizací pak probíhá více méně optickým porovnáním těchto funkcí, nikoliv formálním statistickým testem shody pravděpodobnostních rozdělení skupiny funkcí, která nám definuje podobnost. Tímto problémem se v posledních letech zabývaly publikace [3], [5], [6] a [7], jejichž přehled, stručný popis a srovnání jejich výhod a nevýhod je náplní tohoto článku.

Článek je organizován následovně. Kapitola 2. uvádí základní definice a již existující teoretické výsledky použité v dále popisovaných metodách, v kapitole 3. popíšeme tyto metody a nakonec je v kapitole 4. porovnáme a výsledky ilustrujeme na simulační studii.

2. Teoretické základy

2.1. Náhodné množiny a mozaiky

Definice 2.1. *Nechť $\mathcal{F}$ je rodina uzavřených množin a $\mathcal{C}$ rodina kompaktních množin v $\mathbb{R}^2$. Nechť (Ω, Σ, P) je pravděpodobnostní prostor. Zobrazení $\mathbf{X} : \Omega \to \mathcal{F}$ se nazývá náhodná uzavřená množina, jestliže pro každou kompaktní množinu $K \in \mathcal{C}$ platí $\{\omega \in \Omega : \mathbf{X} \cap K \neq \varnothing\} \in \Sigma$. Rozdělení $P_{\mathbf{X}}$ náhodné množiny $\mathbf{X}$ je dáno vztahem $P_{\mathbf{X}}(F) = P(\{\omega \in \Omega : \mathbf{X}(\omega) \in F\})$ pro $F \in \mathcal{B}(\mathcal{F})$, kde $\mathcal{B}(\mathcal{F})$ je borelovská σ-algebra na $\mathcal{F}$.*

Poznámka 2.1. *Je-li množina hodnot zobrazení $\mathbf{X}$ složená pouze z konvexních kompaktních množin, což v některých pasážích níže předpokládáme, mluvíme o $\mathbf{X}$ jako o náhodné konvexní kompaktní množině.*

Pro jednoduchost si pro naše účely můžeme představit náhodnou množinu jako dvoudimenzionální geometrický objekt náhodného tvaru. Realizací X náhodné množiny $\mathbf{X}$ pak myslíme konkrétní tvar, tj. konkrétní množinu $X = \mathbf{X}(\omega)$ pro dané ω. Jelikož v praxi obvykle dostáváme realizace ve formě digitálních binárních obrázků, budeme používat pojem „realizace" i pro tyto obrázky. Pro práci s realizacemi pak budeme potřebovat následující pojmy.

Definice 2.2. *Pro množiny $X, Y \in \mathcal{F}$ definujeme Hausdorffovu metriku*

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y}\|x-y\|,\ \sup_{y\in Y}\inf_{x\in X}\|x-y\|\right\},$$

kde $\|\cdot\|$ značí eukleidovskou vzdálenost na $\mathbb{R}^2$.

Pro binární obrázky složené z černých a bílých pixelů, které pro výpočetní účely reprezentujeme maticí jedniček a nul, definujeme následující pojem.

Definice 2.3. *Frobeniova norma matice A o rozměrech $m \times n$ je definována vztahem $\|A\|_F = \left(\sum_{i=1}^{m}\sum_{i=1}^{n}(a_{ij})^2\right)^{1/2}$.*

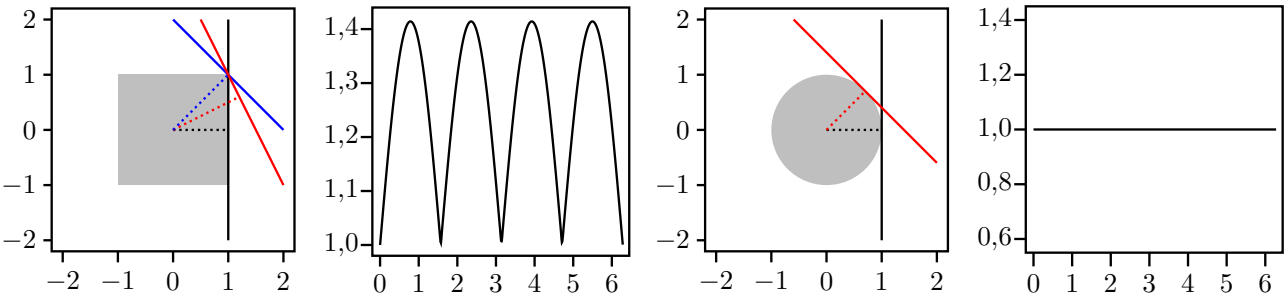
Významná část teorie náhodných množin je věnována náhodným konvexním kompaktním množinám. Jedna z níže prezentovaných metod je přímo založena na aproximaci realizací těmito množinami, přičemž hlavní roli zde hraje následující funkce.

Definice 2.4. *Pro (náhodnou) konvexní kompaktní množinu* $\mathbf{X}$ *definujeme její opěrnou funkci*

$$h_{\mathbf{X}}(u) = \sup_{x \in \mathbf{X}} \langle u, x \rangle, \quad u \in \partial b(0,1),$$

kde $\partial b(0,1)$ *je jednotková sféra v* $\mathbb{R}^2$*, tj. kružnice se středem v počátku a jednotkovým poloměrem.*

Body na jednotkové sféře přitom můžeme nahradit úhly mezi vodorovnou osou a spojnicí počátku s příslušným bodem na kružnici. Pak opěrnou funkci intepretujeme jako vzdálenost počátku od opěrné nadroviny, tj. od přímky, která je kolmá na daný směr a leží v maximální možné vzdálenosti tak, aby s danou množinou měla neprázdný průnik, viz obr. 1. Jinými slovy opěrná funkce popisuje jistý dosah množiny ve všech směrech.



Obrázek 1: Ukázka opěrné funkce pro čtverec s těžištěm v počátku, tedy pro centrovaný geometrický objekt s rovnými hranicemi (levé dva obrázky), a pro kruh se středem v počátku, tedy pro centrovaný geometrický objekt se zaoblenými hranicemi (pravé dva obrázky).

Věta 2.1 (Lavie [10])**.** *Dvě náhodné konvexní kompaktní množiny jsou stejně rozdělené právě tehdy, jsou-li stejná všechna konečně-dimenzionální rozdělení jejich opěrných funkcí.*

Nakonec definujeme speciální mozaiku tvořenou průnikem Voronoiovy mozaiky [1] a sjednocení kruhů se středy v bodech generujících tuto mozaiku.

Definice 2.5. *Mějme* $\{b_1, \ldots, b_n\}$ *konečnou konfiguraci kruhů se středy* $\{x_1, x_2, \ldots, x_n\}$ *a stejnými poloměry. Označme*

$$B_i = \{y \in b_i : \|y - x_i\| \leqslant \|y - x_j\| \text{ pro všechna } j \neq i\}.$$

Systém množin B_i se nazývá Voronoiova mozaika na sjednocení kruhů $\bigcup_{i=1}^{n} b_i$.

Poznámka 2.2. *Buňky B_i jsou množiny bodů z kruhu b_i, které jsou blíže středu kruhu b_i než středu jakéhokoliv jiného kruhu, tedy se jedná o konvexní kompaktní množiny, a tudíž je Voronoiova mozaika sjednocením konvexních kompaktních množin pokrývajícím celé sjednocení kruhů $\bigcup_{i=1}^{n} b_i$.*

2.2. Morfologický skelet množiny

Mějme $X \subseteq \mathbb{R}^2$, $A \subset \mathbb{R}^2$ obsahující počátek, označme $\check{A} = \{-a : a \in A\}$ a $X_a = \{x + a;\ x \in X\}$.

Definice 2.6. *Dilatace, eroze, otevření a zavření množiny X kompaktním prvkem (množinou) A jsou definovány dle [20] následovně:*

$$
\begin{aligned}
D_A(X) &= X \oplus A = \{z \in \mathbb{R}^2 : \check{A}_z \cap X \neq \varnothing\} = \textstyle\bigcup_{a \in A} X_a, \\
E_A(X) &= X \ominus \check{A} = \{z \in \mathbb{R}^2 : A_z \subseteq X\} = \textstyle\bigcap_{a \in A} X_{-a}, \\
O_A(X) &= D_A(E_A(X)) = (X \ominus \check{A}) \oplus A, \\
C_A(X) &= E_A(D_A(X)) = (X \oplus A) \ominus \check{A}.
\end{aligned}
$$

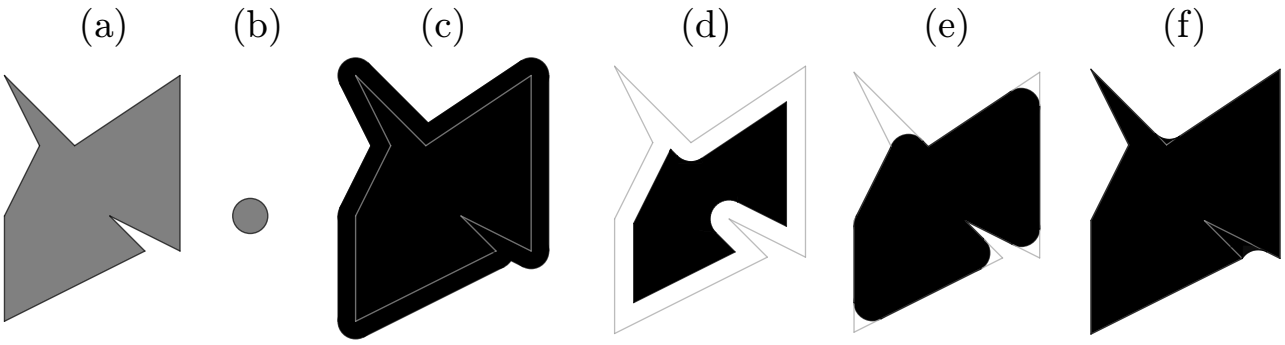
Dilataci si můžeme představit jako zvětšení, jakoby „nafouknutí“, množiny X (pokud je kompaktním prvkem A kruh o poloměru r, což je v aplikacích nejčastější případ, jedná se o zvětšení množiny X o pás šířky r). Eroze je naopak zmenšení množiny X. Otevření si můžeme představit jako proces, kdy množinu nejprve zmenšíme, čímž odstraníme tenké výčnělky, a poté množinu zase zvětšíme. Při operaci zavření množiny naopak nejprve množinu zvětšíme, čímž uzavřeme úzká místa, a poté ji zase zmenšíme. Ukázku těchto operací vidíme na obr. 2.

Definice 2.7. *Kruh rB_x se středem v bodě x a poloměrem r se nazývá maximalní vzhledem k množině X, pokud neexistuje žádný jiný kruh B' obsažený celým povrchem v X a obsahující rB_x, tj.*

$$rB_x \subseteq B' \subseteq X \Rightarrow rB_x = B'.$$

Definice 2.8. *Nechť $I_X^{\max}$ je množina maximálních kruhů vzhledem k X. Morfologický skelet (nebo také kostra; dále jen skelet) $SK(X)$ množiny X je definován jako množina všech středů maximálních kruhů, tj.*

$$SK(X) = \{x;\ rB_x \in I_X^{\max},\ r > 0\}.$$



Obrázek 2: Ukázka morfologických operací z def. 2.6: (a) množina X, (b) kompaktní prvek A, (c) dilatace, (d) eroze, (e) otevření a (f) zavření množiny X prvkem A.

Pro $r > 0$ označme $S_r(X) = \{x \in SK(X) : rB_x \in I_X^{\max}\}$ r-tou skeletovou podmnožinu. Je zřejmé, že pak platí

$$SK(X) = \bigcup_{r>0} S_r(X).$$

Nyní uvažujme X coby binární obrázek. Kruh rB je nahrazen diskrétní aproximací kruhu nB s poloměrem $n \in \mathbb{N}$ s využitím manhattanské vzdálenosti, tj. $nB = \{(x, y) \in \mathbb{Z}^2;\ |x| + |y| \leqslant n\}$. Najděme $N \in \mathbb{N}$ takové, že eroze $E_{NB}(X) \neq \varnothing$ a $E_{(N+1)B}(X) = \varnothing$. Skelet $SK(X)$ je pak definován jako

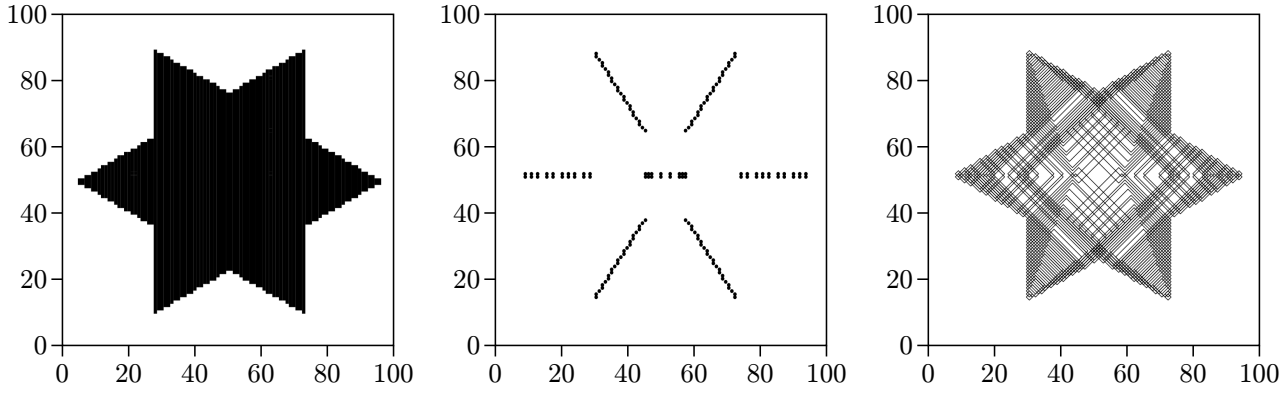
$$SK(X) = \bigcup_{n=0}^{N} S_n(X),$$

kde $S_n(X) = E_{nB}(X) \backslash O_B(E_{nB}(X))$ je analogie n-té skeletové podmnožiny v diskrétním případě.

Algoritmus pro konstrukci skeletu je tedy následující [11]:

1. $n := 0$, $E_1 := X$,
2. $E_2 := E_1 \ominus B$,
3. jestliže $E_2 = \varnothing$, pak $N := n$, $S_n(X) := E_1$ a STOP,
4. $O := E_2 \oplus B$,
5. $S_n(X) := E_1 \backslash O$,
6. $n := n + 1$, $E_1 := E_2$ a běž zpět na bod 2.

Na obr. 3 je ukázka binárního obrázku, jeho skeletu a rekonstrukce původního obrázku pomocí skeletu a příslušných maximálních kruhů, tj. sjednocení všech maximálních kruhů. Poznamenejme, že pro lepší vizualizaci jsme při zobrazení rekonstrukce použili k vykreslení hranic maximálních kruhů jejich spojitou verzi, tj. $\{(x, y) \in \mathbb{R}^2;\ |x| + |y| = n\}$.



Obrázek 3: Ukázka binárního obrázku množiny (vlevo), jejího skeletu (uprostřed) a rekonstrukce původní množiny pomocí skeletu a příslušných maximálních kruhů, tj. sjednocení všech maximálních kruhů (vpravo; pro lepší vizualizaci zobrazeny pouze hranice těchto kruhů).

2.3. Testování shody rozdělení náhodných funkcí

2.3.1. Obálkový test Jako první test shody rozdělení náhodných funkcí jsme použili tzv. obálkový test z [18] (lze jej nalézt také v českém jazyce [17]), který funguje následovně.

Mějme $s+1$ zaměnitelných (exchangeable) náhodných objektů, které jsou popsány funkcionálními charakteristikami $T_i(u)$, $i = 1, ..., s + 1$, $u \in I$ (I je indexová množina). Víme, že $T_2(u), \ldots, T_{s+1}(u)$ pocházejí ze stejného rozdělení, a chceme testovat, zda $T_1(u)$ pochází z téhož rozdělení. Pro každé $u \in I$ označme $R_i^{\uparrow}(u)$, resp. $R_i^{\downarrow}(u)$, pořadí příslušné hodnoty $T_i(u)$ od nejmenšího (pořadí 1) do největšího (pořadí $s + 1$), resp. od největšího (pořadí 1) do nejmenšího (pořadí $s + 1$). Pro každé $u \in I$, pak definujme u-tou hloubku charakteristiky $T_i(u)$ jako $R_i(u) = \min\left(R_i^{\uparrow}(u), R_i^{\downarrow}(u)\right)$.

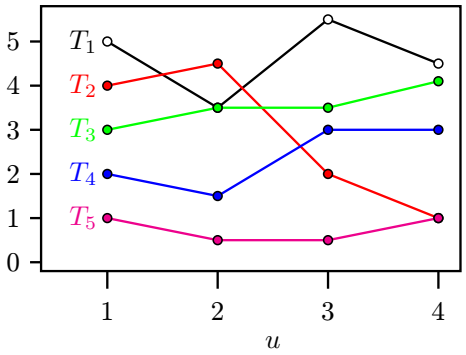
V praxi obvykle pozorujeme $T_i(u)$ na diskrétní indexové množině $I = \{u_1, \ldots, u_n\}$, tj. $T_i(u) = (T_i(u_1), \ldots, T_i(u_n))$. Označme $\widetilde{s} = \lfloor (s + 2)/2 \rfloor$, $\mathbf{N}_i = (N_{i1}, \ldots N_{i\widetilde{s}})$, kde $N_{ik} = \sum_{j=1}^{n} \mathbf{1}(R_i(u_j) = k)$, a definujme

$$\mathbf{N}_i < \mathbf{N}_j \Longleftrightarrow \exists m \leqslant \widetilde{s}\ \forall k < m : N_{ik} = N_{jk}\ \&\ N_{im} > N_{jm}.$$

Pak p-hodnota testu je

$$p = \frac{1}{s+1}\left(1 + \sum_{i=1}^{s+1} \mathbf{1}(\mathbf{N}_i \prec \mathbf{N}_1)\right).$$

Ukázku výpočtu $R_i(u)$ a $\mathbf{N}_i$ lze vidět na obr. 4.



$$\begin{array}{lllll} R_1(1)=1 & R_2(1)=2 & R_3(1)=3 & R_4(1)=2 & R_5(1)=1 \\ R_1(2)=2 & R_2(2)=1 & R_3(2)=2 & R_4(2)=2 & R_5(2)=1 \\ R_1(3)=1 & R_2(3)=2 & R_3(3)=2 & R_4(3)=3 & R_5(3)=1 \\ R_1(4)=1 & R_2(4)=1 & R_3(4)=2 & R_4(4)=3 & R_5(4)=1 \end{array}$$

$$\mathbf{N}_1 = (3,1,0),\ \mathbf{N}_2 = (2,2,0),\ \mathbf{N}_3 = (0,3,1),$$
$$\mathbf{N}_4 = (0,2,2),\ \mathbf{N}_5 = (4,0,0)$$

$$\Rightarrow \mathbf{N}_5 \prec \mathbf{N}_1 \prec \mathbf{N}_2 \prec \mathbf{N}_3 \prec \mathbf{N}_4$$

Obrázek 4: Ukázka výpočtu $R_i(u)$ a $\mathbf{N}_i$ v obálkovém testu.

Jelikož v našem případě budeme chtít testovat shodu rozdělení dvou náhodných funkcí $t^{(1)}$ a $t^{(2)}$, použijeme permutační verzi testu, která funguje následovně. Máme dva vzorky funkcí, konkrétně $t_1^{(1)}(u), \ldots, t_{m_1}^{(1)}(u)$ získané z realizace X a $t_1^{(2)}(u), \ldots, t_{m_2}^{(2)}(u)$ získané z realizace Y. Testovací charakteristika je daná normovaným rozdílem jejich průměrů, tj.

$$T_1(u) = \frac{\bar{t}^{(1)}(u) \ - \ \bar{t}^{(2)}(u)}{\sqrt{\text{var } t^{(1)}(u) \ + \ \text{var } t^{(2)}(u)}}.$$

Dále použijeme permutační Monte Carlo test, viz [2], tj. vezmeme všechny funkce $t_1^{(1)}(u), \ldots, t_{m_1}^{(1)}(u), t_1^{(2)}(u), \ldots, t_{m_2}^{(2)}(u)$ dohromady a vytvoříme s jejich náhodných permutací. Následně každou permutaci rozdělíme na dvě skupiny o rozsahu m_1 a m_2 funkcí a analogicky spočteme $T_i(u)$, $i = 2, \ldots, s+1$, jako normovaný rozdíl průměrů funkcí z i-té permutace. Z hodnot $T_i(u)$ pak spočteme $\mathbf{N}_i$ a z nich následně hledanou p-hodnotu.

2.3.2. Test založený na $\mathcal{N}$-vzdálenosti Druhý test shody rozdělení náhodných funkcí je založen na teorii $\mathcal{N}$-vzdáleností, viz [9], kterou stručně uvádíme níže.

Nechť $\mathcal{X}$ je neprázdná množina. Uvažujme negativně definitní jádro $\mathcal{L} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ [9].

Definice 2.9. *Negativně definitní jádro $\mathcal{L}$ se nazývá* silně negativně definitní jádro*, jestliže pro libovolnou pravděpodobnostní míru μ a libovolnou funkci f :*

$\mathcal{X} \to \mathbb{R}$ *takovou, že* $\int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) = 0$ *a* $\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)f(x)f(y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$ *existuje a je konečný, platí, že* $\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)f(x)f(y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = 0$ *implikuje* $f(x) = 0$ μ*-skoro všude.*

Pro zobrazení $\mathcal{L} : \mathcal{X}\times\mathcal{X} \to \mathbb{C}$ označme $B_{\mathcal{L}}$ množinu všech měr μ takových, že $\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$ existuje.

Věta 2.2 (Klebanov [9])**.** *Nechť* $\mathcal{L}(x,y) = \mathcal{L}(y,x)$*. Pak*

$$\begin{aligned}\mathcal{N}(\mu,\nu) = 2\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y) - \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\ - \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y) \geqslant 0 \end{aligned} \tag{1}$$

platí pro všechny míry $\mu,\nu \in \mathcal{B}_{\mathcal{L}}$ *s rovností pro* $\mu = \nu$ *tehdy a jen tehdy, pokud* $\mathcal{L}$ *je silně negativně definitní jádro.*

Dále v textu budeme výraz $\mathcal{N}(\mu,\nu)$ z (1) nazývat $\mathcal{N}$-vzdáleností měr μ a ν. V [9] můžeme nalézt několik příkladu silně negativně definitního jádra $\mathcal{L}$. Jelikož se zde zaměřujeme na testování rozdělení náhodných funkcí $t^{(1)}$ a $t^{(2)}$, použijeme jádro z [6] zkonstruované speciálně pro náhodné funkce. Konkrétně jestliže vyčíslujeme náhodnou funkci $t^{(1)}$, resp. $t^{(2)}$, v diskrétních proměnných $I = \{u_1,\ldots,u_n\}$, jádro je

$$\mathcal{L}(t^{(1)},t^{(2)}) = \sum_{K\in 2^I}\left(\sum_{u_k\in K}\left(t^{(1)}(u_k) - t^{(2)}(u_k)\right)^2\right)^{1/2},$$

kde 2^I značí množinu všech podmnožin indexové množiny I.

Odhad $\mathcal{N}$-vzdálenosti (náhodných) funkcí $t^{(1)}$ a $t^{(2)}$ založený na náhodných výběrech $t_1^{(1)}(u),\ldots,t_{m_1}^{(1)}(u)$ a $t_1^{(2)}(u),\ldots,t_{m_2}^{(2)}(u)$ je

$$\begin{aligned}\widehat{\mathcal{N}_1} = \frac{2}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathcal{L}(t_i^{(1)},t_j^{(2)}) - \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_1}\mathcal{L}(t_i^{(1)},t_j^{(1)})\\ - \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\mathcal{L}(t_i^{(2)},t_j^{(2)}).\end{aligned} \tag{2}$$

Ten pak hraje roli testové statistiky.

Pro testování hypotézy, že dvě skupiny funkcí $t_1^{(1)}(u),\ldots,t_{m_1}^{(1)}(u)$ a $t_1^{(2)}(u)$, $t_2^{(2)}(u),\ldots,t_{m_2}^{(2)}(u)$ pocházejí ze stejného rozdělení pak opět použijeme Monte

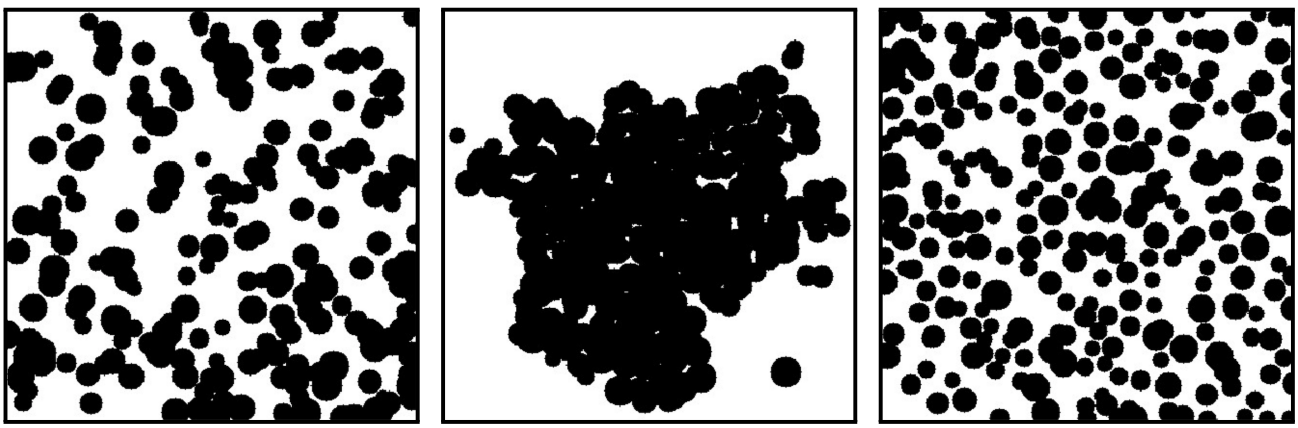
Carlo permutační test, tj. vytvoříme s permutací ze všech funkcí $t_1^{(1)}(u), \ldots, t_{m_1}^{(1)}(u), t_1^{(2)}(u), \ldots, t_{m_2}^{(2)}(u)$, každý zpermutovaný vzorek rozdělíme do dvou skupin o délkách m_1 a m_2 a analogicky k (2) spočteme $\widehat{\mathcal{N}}_i$ pro i-tou permutaci, $i = 2, \ldots, s+1$. Pak p-hodnota testu je

$$p = \frac{\#\{i \in \{2, \ldots, s+1\} : \widehat{\mathcal{N}}_i \geqslant \widehat{\mathcal{N}}_1\} + 1}{s+1}.$$

3. Metody rozlišování realizací náhodných množin

Společným cílem metod, které níže uvádíme, je rozhodnutí, zda dvě realizace X a Y náhodných množin $\mathbf{X}$, resp. $\mathbf{Y}$, jsou si podobné ve smyslu daném pro každou metodu zvlášť. Metody jsou založené na získání skupin funkcí pro každou z realizací X a Y popisující specifické rysy daných realizací a následném testování shody pravděpodobnostních rozdělení těchto funkcí pomocí obálkového testu a testu založeném na $\mathcal{N}$-vzdálenosti. Realizace pak považujeme za podobné, pokud test shodu nezamítá, a nepodobné (rozlišené), pokud je shoda zamítnuta.

Popisované procedury jsou ilustrovány na simulační studii zahrnující tři modely. Prvním je boolský proces, tedy sjednocení náhodných geometrických objektů, jejichž tvary i polohy jsou vzájemně nezávislé (v našem případě se jedná o sjednocení náhodného počtu kruhů s náhodnými poloměry), druhým modelem je shlukový proces a třetím pak proces odpuzujících se komponent, viz [7] pro detaily o parametrech daných modelů a [14] pro detaily o jejich simulacích. Realizace těchto modelů jsou zobrazené na obr. 5.



Obrázek 5: Ukázka realizací boolského modelu (vlevo), shlukového modelu (uprostřed) a modelu odpuzujících se komponent (vpravo), které byly použité pro simulační studii.

Ve studii uvažujeme 200 simulovaných realizací každého modelu v rozlišení 400×400 pixelů, které dále hrají roli jednotek v popisovaných procedurách, a srovnáváme vždy 100 a 100 realizací nejprve stejných a poté různých modelů, tj. máme šest párů modelů ke srovnání. Jelikož výstupy metod jsou p-hodnoty, získáváme tak pro každý srovnávaný pár 100 p-hodnot. Ty jsou zobrazeny formou histogramů. Poznamenejme, že p-hodnoty blízké nule znamenají, že shoda pravděpodobnostních rozdělení příslušných testovacích funkcí je zamítnuta, tudíž realizace považujeme za rozdílné. Proto v simulační studii očekáváme p-hodnoty koncentrované blízko nuly, pokud srovnáváme realizace z různých procesů, zatímco pro realizace stejných procesů by p-hodnoty měly být rovnoměrně rozložené na intervalu $[0, 1]$.

3.1. Aproximace sjednocením konvexních kompaktních množin

Myšlenka této metody, která je detailně popsaná v [6] a [7], spočívá v aproximaci realizací X a Y sjednoceními konvexních kompaktních množin, vyčíslení jejich opěrných funkcí a náhodném výběru těchto funkcí, které pak hrají roli testovacích funkcí $t^{(1)}$ a $t^{(2)}$. Realizace pak považujeme za podobné, pocházejí-li $t^{(1)}$ a $t^{(2)}$ ze stejného rozdělení, což je hypotéza, kterou zde testujeme.

Aproximace realizací funguje následovně:

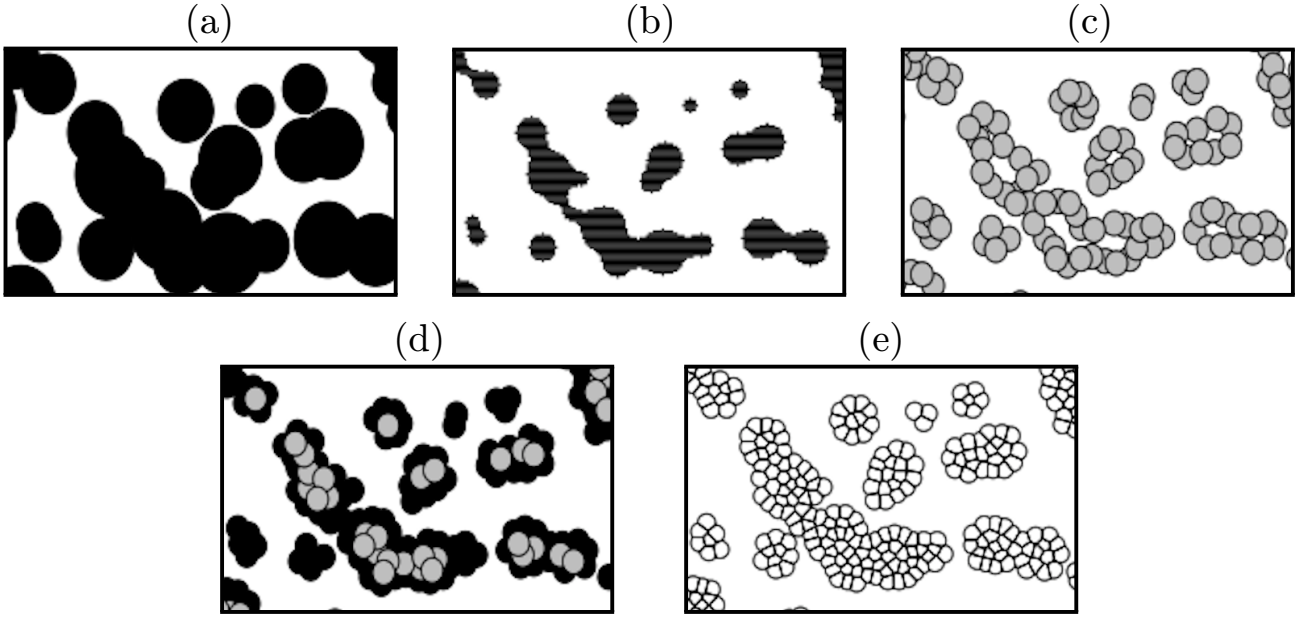
1. Zvolíme hodnotu r a pokryjeme realizaci kruhy o poloměru r použitím Poissonova disc-sampling algoritmu, viz [4], tj.
   (a) vybereme náhodně bod x_1 rovnoměrně uvnitř realizace a zkonstruujeme kruh $b_1(x_1, r)$,
   (b) vybereme náhodně bod x_2 rovnoměrně uvnitř realizace, avšak mimo kruh b_1, a zkonstruujeme kruh $b_2(x_2, r)$,
   (c) pokračujeme tímto způsobem, dokud není pokrytá celá realizace,
2. zkonstruujeme Voronoiovu mozaiku na $\bigcup b_i$.

Je zřejmé, že plocha takto vytvořené aproximace je větší než plocha původní realizace, proto zavádíme dvě korekce v kroku 1., a to:

(i) provedeme erozi realizace kruhem o poloměru r,

(ii) způsobem popsaným výše pokrýváme dilatovanou realizaci, přičemž nejprve vybíráme středy pokrývajících kruhů mezi hraničními pixely a až poté, co je pokrytá celá hranice, pokrýváme náhodně vnitřek,

což je graficky ilustrováno na obr. 6 i s následnou konstrukcí Voronoiovy mozaiky.



Obrázek 6: Pokrývání realizace sjednocením kruhů se stejnými poloměry a následná konstrukce Voronoiovy mozaiky: aproximovaná realizace (a), její eroze (b), pokrývání hranice realizace po erozi (c), pokrývání vnitřku (d) a odpovídající Voronoiova mozaika (e).

Poissonův disc-sampling algoritmus s následnou konstrukcí Voroniovy mozaiky jsme zvolili hlavně pro jeho jednoduchou interpretaci: v případě shlukových procesů má mozaika více vnitřních (tj. mnohoúhelníkových) než vnějších (tj. částečně zaoblených) buněk, zatímco dlouhé tenké komponenty mají naopak více vnějších a méně vnitřních buněk, což je právě rozdíl dobře podchytitelný opěrnými funkcemi, neboť zaobleným částem buněk odpovídají v opěrných funkcích konstantní úseky (v hodnotě poloměru příslušného kruhu), viz levá část obr. 1 na str. 6, zatímco opěrné funkce mnohoúhelníkových buněk vytvářejí v grafech „kopečkovité" úseky, viz pravá část obr. 1 na str. 6. Avšak nejen to. I realizace složené z více menších spíše okrouhlých komponent, resp. z dlouhých tenkých komponent, které obě mají více vnějších buněk, jsou rozeznatelné, neboť zaoblené části v prvním případě jsou delší než v případě druhém, což je opět rozlišeno příslušnými opěrnými funkcemi.

Problémem však je volba optimálního poloměru pro pokrývací kruhy. Ten totiž nesmí být ani moc malý, neboť příslušná Voronoiova mozaika by měla příliš mnoho vnitřních buněk a ty jsou stejně rozdělené, tudíž nenesou informaci o tvaru komponent, avšak nesmí být ani moc velký, abychom při aproximaci zachovali tvar původní realizace a také abychom měli dostatečný

vzorek buněk pro testování. Několik (spíše heuristických) doporučení lze nalézt v [7].

Další slabinou této procedury je fakt, že dvě buňky Voronoiovy mozaiky mající stejný tvar, ale jinou orientaci, mají různé opěrné funkce. Proto v [7] autoři navrhují jisté přerovnání opěrných funkcí, což sice metodu činí efektivnější, ale i přesto dochází k významné ztrátě informace.

Nevýhodou této metody je také náhodnost pokrývání, která s sebou rovněž nese jistou míru nepřesnosti.

3.2. Srovnání skeletů a příslušných maximálních kruhů

Ve druhé proceduře, studované detailněji v [3], definujeme podobnost realizací náhodných množin pomocí funkce popisující přírustky objemu kolem bodů skeletů daných realizací.

Mějme realizaci X stacionární náhodné množiny ve formě binárního obrázku a její skelet $SK(X)$. Pro každý bod $x_i \in SK(X)$ označme r_i poloměr příslušného maximálního kruhu. Poznamenejme však, že dle algoritmus v kapitole 2.2., jeden izolovaný pixel představuje kruh s nulovým poloměrem, avšak pro definici testovací funkce (3) (viz níže), je důležité, aby i jeden pixel představoval kruh s kladným poloměrem. Proto pro další účely stavovíme, že každému bodu $x_i \in S_n(X)$ připadá poloměr odpovídajícího maximálního kruhu $r_i = n + 1$. Testovací funkci v bodě x_i pak definujeme jako

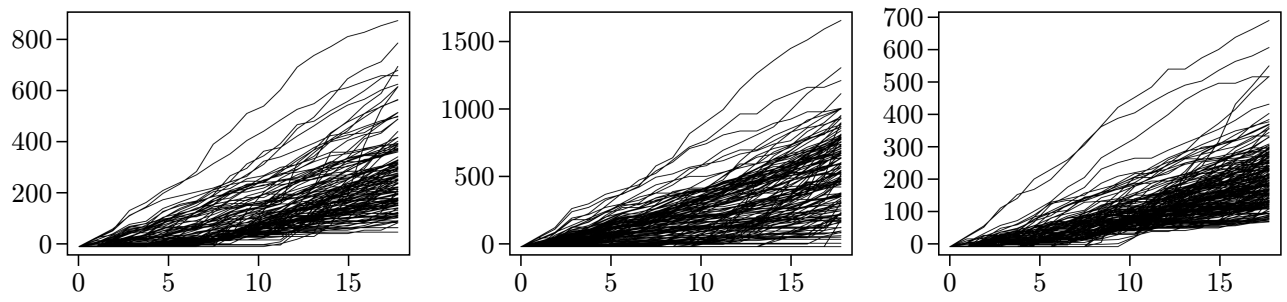
$$t_i(u\ ) = \sum_{j \neq i} r_j \mathbf{1}_{||x_i - x_j|| < u}, \quad u = 1, 2, \ldots, U_{\max}, \quad U_{\max} \in \mathbb{N}. \tag{3}$$

Interpretace je taková, že výraz $r_j \cdot 1$ z (3) je přírustek plochy přidáním j-tého kruhu do celkové plochy množiny X kolem bodu x_i, neboť je to plocha obdélníku $r_j \times 1$ kterým přibližně při seřazení bodů skeletu za sebe j-tý maximální kruh do celkové plochy přispívá.

Bohužel neexistuje obecné pravidlo, jak v (3) stanovit parametr $U_{\max}$. Musí být dostatečně velký, aby testovací funkce podchytily rozprostření masy v realizaci, ale nesmí být příliš velký, neboť pak dochází k následující komplikaci. Testovací funkce t_i a t_j jsou si podobné, pokud x_i a x_j jsou si blízko a $U_{\max}$ je příliš velké (neboť mají spoustu maximálních kruhů společných). Proto zde zavádíme minimální vzdálenost $D_{\min}$ mezi body, které mohou být zahrnuty do náhodného výběru následných testů (tyto body dále nazýváme testovacími body). Je přitom zřejmé, že volba $D_{\min}$ je úzce spjatá s volbou $U_{\max}$, neboť obě tyto hodnoty dohromady ovlivňují, jak moc se maximální kruhy přispívající do hodnot testovacích funkcí překrývají, tj. jak moc jsou závislé. Pro dané $U_{\max}$ by přirozenou volbou $D_{\min}$ bylo $D_{\min} = 2U_{\max}$, neboť

v takovém případě je každá dvojice (x, r) zahrnutá pouze v jedné testovací funkci t_i. Avšak to si nemůžeme dovolit, pokud obrázek není dostatečně velký. Širší diskusi na toto téma lze nalézt v [3].

Stručně řečeno, doporučuje se nejprve vykreslit pro každou realizaci několik testovacích funkcí na delším definičním oboru a na základě grafů rozhodnout, jaký interval pro u je dostatečný k podchycení rysů dané realizace, např. na obr. 7 pozorujeme od hodnoty $u = 10$ dále hustší shluky funkcí pro proces odpuzujících se komponent než pro jiné procesy, nebo mnohem větší funkční hodnoty od $u = 5$ pro shlukový proces než pro ostatní procesy.



Obrázek 7: Ukázka testovacích funkcí z jedné realizace boolského procesu (vlevo), shlukového procesu (uprostřed) a procesu s odpuzujícími se komponentami (vpravo) použitých při rozlišování metodou srovnávání skeletů a příslušných maximálních kruhů.

Chceme-li tedy porovnávat dvě realizace, podíváme se, zda a od jaké hodnoty u se začínají jejich testovací funkce vizuálně lišit. Jestliže se ani pro velká u (velká vzhledem k velikosti obrázku) funkce neliší, vezmeme $U_{\max}$ co největší tak, abychom v souladu s $D_{\min}$ měli na vstupu alespoň pár desítek testovacích bodů. Pokud se od nějaké hodnoty u lišit začínají, hodnotu $U_{\max}$ stanovíme tak, aby $U_{\max} > u$, ale abychom zároveň v souladu s $D_{\min}$ i zde měli na vstupu dostatek testovacích bodů. Pokud nelze stanovit $(D_{\min}, U_{\max})$ splňující uvedené podmínky, není tato metoda k porovnávání daných realizací vhodná.

Poté, co určíme výše zmíněné parametry, porovnáme realizace X a Y tak, že náhodně vybereme testovací body, vyčíslíme příslušné testovací funkce $t_1^{(1)}(u), \ldots, t_{m_1}^{(1)}(u)$ a $t_1^{(2)}(u), \ldots, t_{m_2}^{(2)}(u)$, a stejně jako v předešlé metodě aplikujeme obálkový test a test založený na $\mathcal{N}$-vzdálenosti k otestování hypotézy, že skupiny funkcí $t_1^{(1)}(u), \ldots, t_{m_1}^{(1)}(u)$ a $t_1^{(2)}(u), \ldots, t_{m_2}^{(2)}(u)$ pocházejí ze stejného rozdělení.

Když pomineme komplikace ohledně volby $U_{\max}$ a $D_{\min}$, má tato metoda spoustu výhod. Za prvé je ve srovnání s předešlou metodou až na náhodný vý-

běr testovacích bodů deterministická a jednoznačná, tj. nedochází k náhodné aproximaci a rekonstrukce realizace z jejího skeletu je zcela shodná s původní realizací. Za druhé, testovací funkce popisuje přírustek plochy v okolí příslušného testovacího bodu ve všech směrech, takže nedochází ke ztrátě informace vlivem rotací, což vede k větší přesnosti v případě hledání nepodobností. A vskutku, jak ukazuje simulační studie, nám tato metoda dává nejlepší výsledky při rozlišování na pohled podobných realizací, které však pocházejí z různých modelů (jmenovitě z booleského procesu a procesu odpuzujících se komponent).

3.3. Srovnání komponent a jejich sousedství

Ve třetí metodě, uvedené v publikaci [5], uvažujme realizace X a Y složené z komponent $C_i,\ i \in J$ (indexová množina). Obvykle uvažujeme C_i coby spojité komponenty, ale ve specifických případech můžeme použít i jiné dělení (např. pokud jsou komponenty složené z biologických buněk, které lze ve vstupním binárním obrázku identifikovat, můžeme jako C_i označit tyto jednotlivé buňky). Pro každou C_i definujeme její i-té sousedství N_i jako oblast obsahující celou C_i a k tomu všechny body pozorovacího okna, které jsou blíž C_i než jakékoliv jiné $C_j,\ j \in J\backslash\{i\}$ ve smyslu Hausdorffovy metriky, tj. $N_i = \{y \in W : d_H(\{y\}, C_i) \leqslant d_H(\{y\}, C_j)$ pro všechna $i \neq j\}$.

Hlavní myšlenkou je srovnat v realizacích X a Y, které uvažujeme coby binární obrázky, tedy matice jedniček a nul (odpovídajících pixelům množin, resp. jejich doplňků), symetrické diference jednotlivých komponent $C_1 \Delta C_2 := \{x : x \in C_1, x \notin C_2 \ \text{ nebo } \ x \notin C_1, x \in C_2\}$, kde C_1 a C_2 jsou vycentrovány ve svých těžištích, resp. symetrické diference jednotlivých sousedství $N_1 \Delta N_2$ pro N_1 a N_2 vycentrované do těžišť C_1, resp. C_2, jejichž plochy jsou $||C_1 - C_2||_F$, resp. $||N_1 - N_2||_F$. Srovnáváme přitom každou komponentu z realizace X s každou komponentou z realizace Y. Jelikož funkce

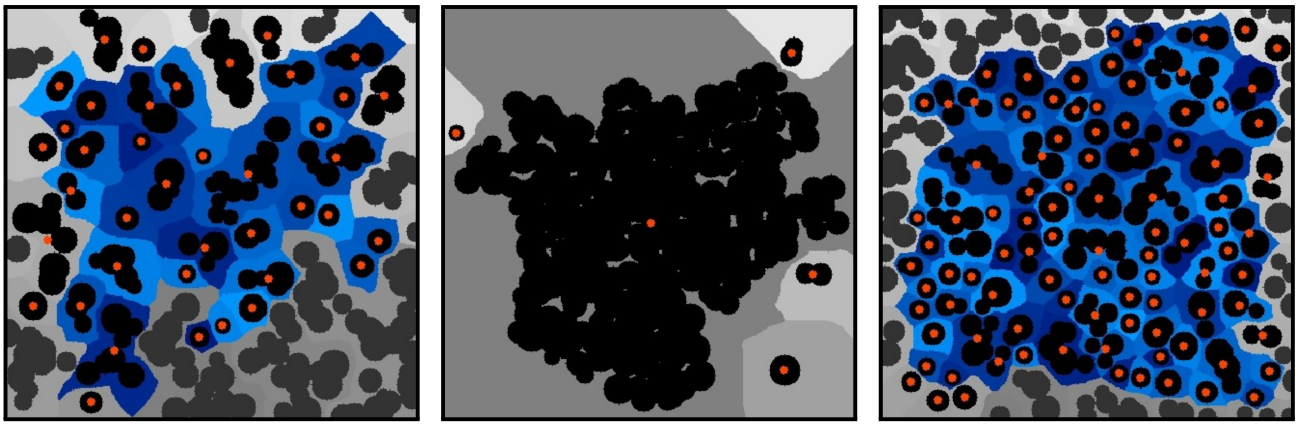
$$\mathcal{L}_C(C_1, C_2) = ||C_1 - C_2||_F, \quad \mathcal{L}_N(N_1, N_2) = ||N_1 - N_2||_F,$$

$$\mathcal{L}_{CN}((C_1, N_1), (C_2, N_2)) = \sqrt{||C_1 - C_2||_F^2 + ||N_1 - N_2||_F^2}$$

tvoří silně negativně definitní jádra, aplikujeme test založený na $\mathcal{N}$-vzdálenostech na tato jádra místo $\mathcal{L}(t^{(1)}, t^{(2)})$ a dle (2) z nich spočteme $\widehat{\mathcal{N}}_1$ a $\widehat{\mathcal{N}}_i$, $i = 2, \ldots, s$. To znamená, že v tomto případě hrají komponenty C_i a sousedství N_i role testovacích funkcí z předešlých kapitol. Označíme-li $C_1^{(1)}, \ldots, C_{m_1}^{(1)}$ a $C_1^{(2)}, \ldots, C_{m_2}^{(2)}$ náhodný výběr komponent z realizací X, resp. Y, a $N_1^{(1)}, \ldots, N_{m_1}^{(1)}$ a $N_1^{(2)}, \ldots, N_{m_2}^{(2)}$ jejich příslušná okolí, pak zde uvažujeme tři verze nulové hypotézy, a to za prvé tvrzení, že $C_1^{(1)}(u), \ldots, C_{m_1}^{(1)}(u)$ a $C_1^{(2)}(u), \ldots,$

$C_{m_2}^{(2)}(u)$ pocházejí ze stejného rozdělení, za druhé tvrzení, že $N_1^{(1)}(u), \ldots, N_{m_1}^{(1)}(u)$ a $N_1^{(2)}(u), \ldots, N_{m_2}^{(2)}(u)$ pocházejí ze stejného rozdělení, a za třetí tvrzení, že dvojice $(C_1, N_1)^{(1)}(u), \ldots, (C_{m_1}, N_{m_1})^{(1)}$ a $(C_1, N_1)^{(2)}, \ldots, (C_{m_2}, N_{m_2})^{(2)}$ pocházejí ze stejného rozdělení.

Poznamenejme, že tato metoda je velice citlivá na okrajové efekty, proto zahrnujeme do studie pouze komponenty, které neprotínají hranici pozorovacího okna (znázorněny černou barvou v obr. 8, zatímco vynechané komponenty jsou znázorněny tmavě šedou barvou). Analogicky při porovnávání sousedství uvažujeme pouze ta, která jsou celá obsažena v pozorovacím okně (znázorněna různými odstíny modré v obr. 8).



Obrázek 8: Uvažované komponenty (vyznačené černě) v boolském procesu (vlevo), shlukovém procesu (uprostřed) a procesu odpuzujících se komponent (vpravo), jejich těžiště (červené body) a uvažovaná sousedství (různé odstíny modré).

4. Srovnání metod

V tab. 1 shrneme uvedené metody, jejich použití, výhody a nevýhody, přičemž první metodu budeme zkráceně nazývat „aproximace", druhou „skelety" a třetí „komponenty a sousedství". Pro verzi aproximace, v níž používáme obálkový test, pak budeme používat značení „AO", pro verzi s použitím $\mathcal{N}$-vzdálenosti „AN", analogicky tyto verze skeletů budeme značit jako „SO", resp. „SN", a pro porovnání komponent, sousedství a komponent se sousedstvím dohromady budeme používat zkratky „K", „S", resp. „KS".

Ze statistického hlediska můžeme vynést závěr, že všechny uvedené metody jsou velice přesné při porovnávání realizací stejných modelů, neboť příslušné histogramy p-hodnot provedených testů vykazují rovnoměrné rozdělení na intervalu $[0, 1]$ (proto zde tyto histogramy neuvádíme). Zajímavější

jsou histogramy p-hodnot v případě srovnávání různých modelů, viz obr. 9. Z nich můžeme vidět, že v metodě aproximace nám test založený na $\mathcal{N}$-vzdálenostech dává výrazně lepší výsledky než obálkový test, což je nejspíš dáno skutečností, že obálkový test srovnává pouze průměry opěrných funkcí, takže nemůže dostatečně podchytit zejména podobnosti různě orientovaných buněk podobného tvaru. Nicméně i test založený na $\mathcal{N}$-vzdálenostech vykazuje dost nepřesností (ve smyslu vysokých p-hodnot). Významně vyšší přesnosti dosahujeme v případě skeletů. Ty nejsou vázány na orientaci, takže obě varianty testů (obálkový i založený na $\mathcal{N}$-vzdálenostech) dávají podobnou přesnost, která je zároveň nejvyšší ze všech uvedených metod.

Pro metodu komponent a sousedství jsme použili pouze test založený na $\mathcal{N}$-vzdálenostech, neboť testovacími charakteristikami nejsou přímo funkce, nýbrž komponenty a příslušná sousedství, pro něž jsme nenalezli vhodnou verzi obálkového testu. Zde z histogramů vidíme, že metoda velice špatně odlišuje shlukový proces od ostatních procesů, což je způsobeno jednak malým množstvím komponent ve shlukovém procesu, jednak jednou velkou dominantní komponentou v každé realizaci tohoto procesu, která významně ovlivní hodnotu symetrické diference, čímž potlačí hodnoty symetrických diferencí ostatních dvojic komponent jak v původních realizacích, tak v následných permutacích.

Nakonec v tab. 2 uvádíme podíly p-hodnot menších nebo rovných 0,001, 0,01, 0,05 a 0,1, které nejsou z histogramů patrné, pro různé modely, jejichž realizace jsou si vizuálně nejpodobnější, jmenovitě pro boolský proces a proces odpuzujících se komponent. Z ní naopak, v porovnání k předešlému, vidíme, že při rozlišování těchto procesů je metoda komponent a sousedství významně silnější než metoda aproximace. Nejsilnější metodou i zde zůstávají skelety.

Závěrem bychom tedy mohli říct, že výběr vhodné metody závisí jednak na účelu a jednak na vzhledu realizace. Záleží-li nám nejen na tvarech, ale i na vzájemné poloze komponent v realizaci, je vhodné použít skelety. Zajímá-li nás pouze tvar komponent, nabízí se aproximace nebo komponenty a sousedství, přičemž druhá jmenovaná metoda přesněji rozlišuje komponenty podobných velikostí a první si lépe poradí v případě extrémně velikých komponent či menšího počtu komponent v realizaci.

Poděkování

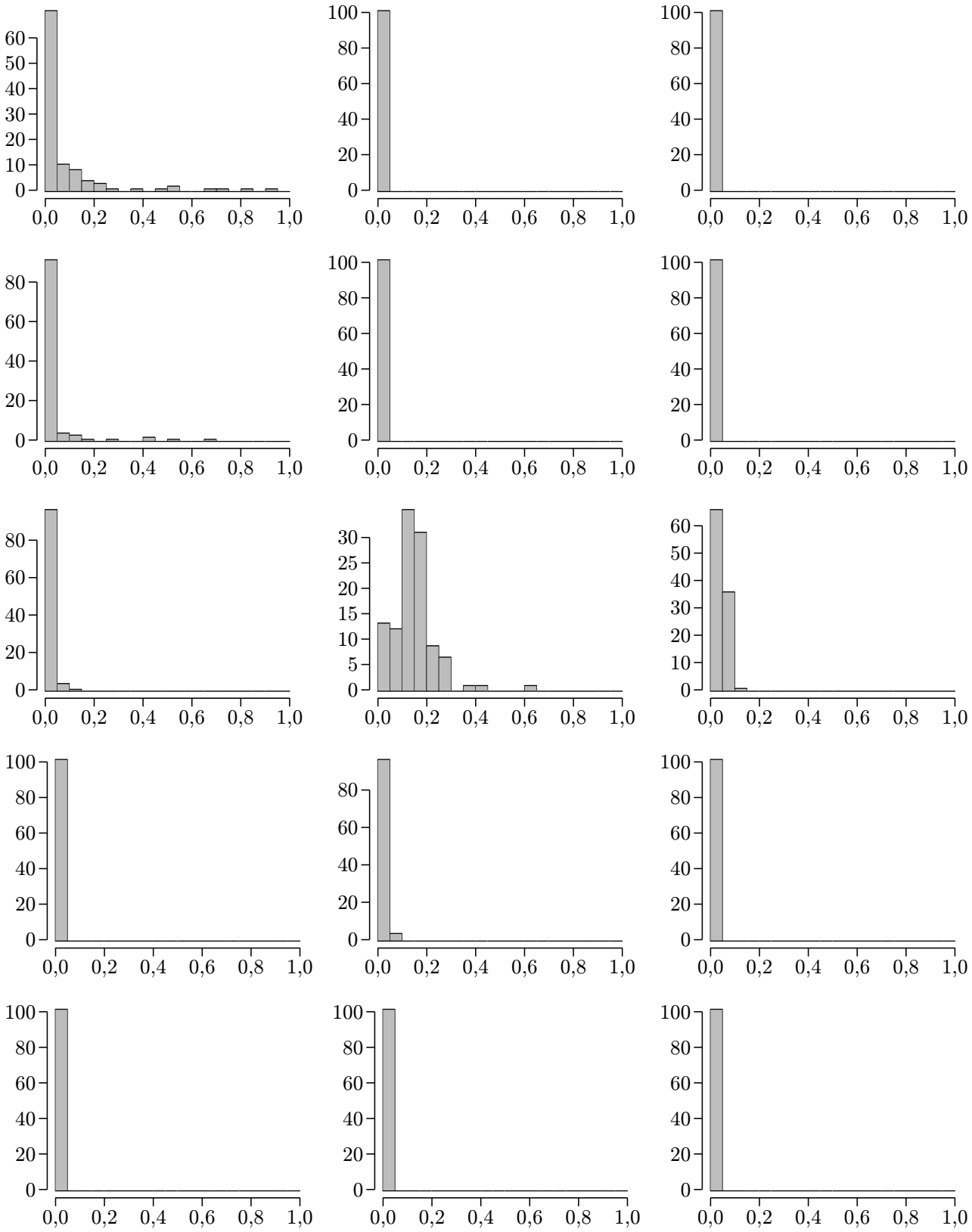
Podpořeno grantem GAČR 19-04412S.

Tabulka 1: Přehled použití, výhod a nevýhod prezentovaných metod.

Použití		
Aproximace	**Skeletony**	**Komponenty a sousedství**
rozlišuje realizace zejména podle délky hranice vzhledem k velikosti a podle tvaru hranice komponent	bere v úvahu jak plochu realizací tak vzdálenosti komponent a jejich částí	rozlišuje realizace podle podobnosti komponent či jiných částí (dle aplikace)

Výhody						
Aproximace		**Skeletony**		**Komponenty a sousedství**		
jednoduchá interpretace testovacích funkcí		vysoká přesnost při rozlišování podobných realizací		flexibilita (lze uvažovat různé, nejen spojité komponenty)		
AO	AN	SO	SN	K	S	KS
rychlejší než AN	přesnější než AO	rychlejší než SN	lehce přesnější než SO	rychlejší než S a KS	berou v úvahu pozice komponent	

Nevýhody						
Aproximace		**Skeletony**		**Komponenty a sousedství**		
náhodná aproximace, volba poloměru pokrývání, vliv orientace buněk mozaiky		málo intuitivní interpretace, nutná volba vstupních parametrů		vyžaduje na vstupu velké množství komponent		
AO	AN	SO	SN	K	S	KS
nejméně přesná ze všech metod	málo přesná vzhledem k časové náročnosti	–	lehce vyšší časová náročnost	spíše nižší přesnost	vzhledem k K výrazně vyšší časová náročnost	



Obrázek 9: Histogramy p-hodnot pro srovnání boolského procesu s procesem s odpuzujícími se komponentami (vlevo), boolského se shlukovým procesem (uprostřed) a shlukového procesu s procesem s odpuzujícími se komponentami (vpravo) metodami AO (první řádek), AN (druhý řádek), K (třetí řádek), S (čtvrtý řádek), KS, SO and SN (stejné histogramy v pátém řádku).

Tabulka 2: Procenta malých p-hodnot pro boolský proces a proces odpuzujících se komponent.

Metoda	$\leqslant 0{,}001$	$\leqslant 0{,}01$	$\leqslant 0{,}05$	$\leqslant 0{,}1$
AO	5 %	5 %	66 %	76 %
AN	50 %	72 %	87 %	91 %
SO	99 %	100 %	100 %	100 %
SN	100 %	100 %	100 %	100 %
K	67 %	83 %	95 %	99 %
KN	97 %	100 %	100 %	100 %
S	99 %	100 %	100 %	100 %

Literatura

[1] Chiu, S. N., Stoyan, D., Kendall, W. S., Mecke, J.: *Stochastic Geometry and its Applications.* John Wiley & Sons, New York, 2013, doi: 10.1002/9781118658222. *cit. 4, 6*

[2] Davison, A. C. Hinkley, D. V.: *Bootstrap Methods and their Application,* Volume 1 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997, doi: 10.1017/CBO9780511802843. *cit. 10*

[3] Debayle, J., Gotovac Đogaš, V., Helisová, K., Staněk, J., Zikmundová, M.: Assessing Similarity of Random Sets via Skeletons. *Methodology and Computing in Applied Probability*, 2020+, doi: 10.1007/s11009-020-09785-y. *cit. 5, 15 a 16*

[4] Ebeida, M. S., Davidson, A. A., Patney, A., Knupp, P. M., Mitchell, S. A., Owens, J. D.: Efficient Maximal Poisson-Disk Sampling. In *SIGGRAPH '11*: ACM SIGGRAPH 2011 papers, Hugues Hoppe Eds., July 2011, doi: 10.1145/1964921.1964944. *cit. 13*

[5] Gotovac Đogaš, V.: Similarity Between Random Sets Consisting of many Components. *Image Analysis and Stereology* **38**, 185–99, 2019, doi: 10.5566/ias.2017. *cit. 5, 17*

[6] Gotovac Đogaš, V., Helisová, K.: Testing Equality of Distributions of Random Convex Compact Sets via Theory of $\mathcal{N}$-Distances. *Methodology and Computing in Applied Probability,* 2020+, doi: 10.1007/s11009-019-09747-z. *cit. 5, 11 a 13*

[7] Gotovac Đogaš, V., Helisová, K., Ugrina, I.: Assessing Dissimilarity of Random Sets Through Convex Compact Approximations, Support Functions and

Envelope Tests. *Image Analysis and Stereology* **35**, 181–193, 2016, doi: 10.5566/ias.1490. *cit. 5, 12, 13 a 15*

[8] Hermann, P., Mrkvička, T., Mattfeldt, T., Minárová, M., Helisová, K., Nicolis, O., Wartner, F., Stehlík, M.: Fractal and Stochastic Geometry Inference for Breast Cancer: A Case Study with Random Fractal Models and Quermass-Interaction Process. *Statistics in Medicine* **34**, 2636–2661, 2015, doi: 10.1002/sim.6497. *cit. 4*

[9] Klebanov, L. B.: *$\mathcal{N}$-distances and Their Applications.* Karolinum Press, Charles University, Prague, 2006. *cit. 10, 11*

[10] Lavie, M.: Characteristic Function for Random Sets and Convergence of Sums of Independent Random Sets. *Acta Math Viet* **25**, 87–99, 2000. *cit. 6*

[11] Maragos, P. A., Schafer, R. W.: Morphological Skeleton Representation and Coding of Binary Images. *IEEE Trans. on Acoustics, Speech, and Signal Processing* **34**, 1228–1244, 1986, doi: 10.1109/TASSP.1986.1164959. *cit. 8*

[12] Matheron, G.: *Random Sets and Integral Geometry.* John Wiley & Sons, New-York, 1975. *cit. 4*

[13] Molchanov, I.: *Theory of Random Sets.* Springer, New York, 2005, doi: 10.1007/1-84628-150-4. *cit. 4*

[14] Møller, J., Helisová, K.: Power Diagrams and Interaction Processes for Unions of Discs. *Advances in Applied Probability* **40**, 321–47, 2008, doi: 10.1239/aap/1214950206. *cit. 12*

[15] Møller, J., Helisová, K.: Likelihood Inference for Unions of Interacting Discs. *Scandinavian Journal of Statistics* **37**, 365–81, 2010, doi: 10.1111/j.1467-9469.2009.00660.x. *cit. 4*

[16] Mrkvička, T., Mattfeldt, T.: Testing Histological Images of Mammary Tissues on Compatibility with the Boolean Model of Random Sets. *Image Analysis and Stereology* **30**, 11–18, 2011, doi: 10.5566/ias.v30.p11-18. *cit. 4*

[17] Mrkvička, T.: Globální obálkové testy aneb jak otestovat vhodnost statistického modelu na základě funkcionální charakteristiky. *Pokroky matematiky, fyziky a astronomie* **62**(1), 17–23. *cit. 9*

[18] Myllymäki, M., Mrkvička, T., Grabarnik, P., Henri Seijo, H., Hahn, U.: Global Envelope Tests for Spatial Processes. *Journal of the Royal Statistical Society*, Series B (Statistical Methodology) **79**, 381–404, 2017, doi: 10.1111/rssb.12172. *cit. 9*

[19] Neumann, M., Staněk, J., Pecho, O. M., Holzer, L., Beneš, V., Schmidt, V.: Stochastic 3D Modeling of Complex Three-phase Microstructures in SOFC-electrodes with Completely Connected Phases. *Computational Materials Science* **118**, 353–64, 2016, doi: 10.1016/j.commatsci.2016.03.013. *cit. 4*

[20] Serra, J.: *Image Analysis and Mathematical Morphology,* Vol. 2: Theoretical Advances. Academic Press, 1982, doi: 10.1002/cyto.990040213. *cit. 4, 7*

ODHAD VARIANČNÍ MATICE VE VYSOKÉ DIMENZI

COVARIANCE MATRIX ESTIMATION IN HIGH-DIMENSIONAL PROBLEMS

Marie Turčičová[1], Kryštof Eben[2]

Adresa: [1]KPMS MFF UK, Sokolovská 49/83, 186 00, Praha 3 – Karlín
[1,2]Ústav informatiky AV ČR, Pod Vodárenskou věží 271/2, 182 07, Praha 8

E-mail: [1]turcicova@cs.cas.cz, [2]eben@cs.cas.cz

Abstrakt: V řadě statistických aplikací, kde je dimenze náhodného vektoru vysoká v porovnání s počtem dostupných měření, je velkým problémem odhad varianční matice. Klasická výběrová varianční matice má v takovém případě řadu nežádoucích vlastností, zejména nízkou hodnost a malou spolehlivost odhadu jednotlivých prvků. Tento článek obsahuje přehled metod, které se v tomto případě k odhadu varianční matice používají. Pozornost je nejdříve věnována výpočetně jednoduchým metodám pracujícím po prvcích, mezi které patří například metoda smrštění (shrinkage), posílení diagonály (tapering) a další. Dále je uveden přehled složitějších přístupů, které používají parametrické modely založené na různých dodatečných předpokladech o vlastnostech náhodného vektoru, zejména normality, kovarianční stacionarity nebo markovské vlastnosti. Parametrické modely se používají jak k popisu poklesu vlastních čísel, tak k přímému modelování varianční matice či její inverze. Parametry příslušných modelů lze odhadovat standardními statistickými postupy.

Klíčová slova: varianční matice, odhad, vysoká dimenze, regularizace.

Abstract: In many statistical applications, where the dimension of a random vector highly exceeds the number of available measurements, the estimation of covariance matrix poses a challenge. The sample covariance matrix has several undesirable properties in this case, specifically low rank and poor accuracy of estimation of its individual elements. This paper provides an overview of methods that are used for covariance matrix estimation in high-dimensional problems. First, we pay attention to computationally simple methods which usually work element-wise, such as shrinkage, tapering, etc. Further, more complex approaches are presented, which employ parametric models based on additional assumptions about the properties of the random vector, especially normality, covariance stationarity and Markov property. Parametric

models are used to describe the decay of eigenvalues or to model the covariance matrix or its inverse. Parameters of the corresponding models can be estimated by standard statistical techniques.

Keywords: covariance matrix, estimator, high-dimension, regularization.

1. Úvod

V tomto článku se budeme zabývat metodami odhadu varianční matice náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_p)^\top$ délky p založenými na náhodném výběru, jehož rozsah N je výrazně menší než p. Nejdůležitější aplikace těchto metod se týkají vícedimenzionálních náhodných polí, zejména v prostorové statistice, kdy například v každém z 500 000 ($= p$) bodů třídimenzionální mřížky o rozměrech $100 \times 100 \times 50$ zjišťujeme hodnotu určité proměnné, zatímco dostupný náhodný výběr má rozsah v řádu desítek. Taková situace je typická pro atmosférické vědy, kde stav atmosféry popisovaný numerickým modelem má vysokou dimenzi, ale výpočetní náročnost modelu nedovoluje pracovat s velkým počtem „scénářů“ a tvořit náhodné výběry o velkém rozsahu, které by dobře popisovaly stochastické vlastnosti modelových polí.

Kromě prostorové statistiky se problém odhadování varianční matice v situaci malého výběru objevuje v mnoha dalších oborech, např. finančnictví, genetice, lékařské statistice, epidemiologii nebo zpracování obrazu. Jednou z typických úloh, kde odhad varianční matice hraje klíčovou roli, je začleňování aktuálních pozorování do dynamického systému (tzv. datová asimilace).

Metody odhadu uvedené v tomto článku jsou určeny zejména pro tuto aplikaci, tak jak se používá ve vědách o atmosféře. Příslušné náhodné pole pak obsahuje např. hodnoty různých meteorologických veličin. Při modelování variančních matic vícerozměrných polí se obvykle body mřížky uspořádají do vektoru, například u dvourozměrného pole se zkonstruuje vektor obsahující postupně všechny sloupce matice mřížky. Vektor $\mathbf{X}$ v dalším textu obvykle značí takto uspořádané náhodné pole. Pak je možné použít různé metody odhadu včetně standardních, je však třeba vzít v úvahu speciální strukturu varianční matice vyplývající z povahy původního problému.

V celém článku budeme předpokládat, že rozdělení vektoru $\mathbf{X}$ má konečné druhé momenty a označíme $\mathrm{var}(\mathbf{X}) \equiv \Sigma$. Standardním odhadem varianční matice Σ založeným na náhodném výběru $\mathbf{X}_1, \ldots, \mathbf{X}_N$ je výběrová varianční matice

$$S = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{X}_k - \overline{\mathbf{X}})(\mathbf{X}_k - \overline{\mathbf{X}})^\top, \tag{1}$$

kde $\overline{\mathbf{X}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{X}_k$. V naší situaci malého počtu pozorování je to odhad nekvalitní a trpí celou řadou nedostatků. Je-li rozdělení $\mathbf{X}$ spojité a $p/N < 1$, ale ne dostatečně blízké nule, je sice matice S skoro jistě invertibilní, ale je obvykle špatně podmíněná, což mimo jiné znamená, že její inverze S^{-1} je nekvalitním odhadem matice Σ^{-1}. V případě $p >> N$ má již matice S nízkou hodnost a výpočet její inverze není vůbec možný. Odhad matice přesnosti Σ^{-1} je přitom v řadě situací klíčový. Ke špatné podmíněnosti matice S se přidává jako další nežádoucí jev výskyt tzv. rušivých kovariancí (angl. spurious covariances), což jsou nereálně vysoké odhady kovariancí vzniklé pouze v důsledku malého rozsahu výběru.

V tomto článku se zaměříme na metody regularizace odhadu varianční matice, kde regularizací rozumíme libovolnou techniku vedoucí k regulární matici, která bude dobrým odhadem skutečné varianční matice. Metody regularizace můžeme dělit na neparametrické a parametrické, dále rozlišovat podle toho, zda probíhají v původním (fyzickém) prostoru nebo v transformovaném prostoru (spektrálním, waveletovém), a konečně také podle toho, zda je regularizována varianční matice nebo její inverze. Níže popsané neparametrické metody se často dají analogicky použít i po transformaci, proto je v sekci 2. uvedeme jen ve fyzickém prostoru. V sekci 3. popíšeme různé parametrické modely založené na dodatečných předpokladech o struktuře odhadované varianční matice, a to jak ve fyzickém prostoru, tak v transformovaném prostoru a pro inverzní varianční matici.

2. Neparametrické metody

2.1. Odhad metodou smrštění (tzv. shrinkage odhad)

Je známo, že deficit v rozsahu výběru má za následek zkreslení struktury vlastních čísel výběrové varianční matice v tom smyslu, že největší (nejmenší) vlastní čísla jsou při odhadu nadhodnocena (podhodnocena). Celá řada metod se zabývá korekcí tohoto fenoménu. Příkladem mohou být odhady metodou smrštění, tzv. *shrinkage* odhady. Asi nejznámější z nich je lineární *shrinkage* odhad definovaný v článku [16], který je dobře podmíněný a vede ke zpřesnění odhadu varianční matice, aniž by se předpokládala jakákoliv speciální struktura. Tento odhad je tvaru

$$S_1 = \rho\nu I \ + (1-\rho)S, \tag{2}$$

kde $0 < \rho < 1$ a $\nu > 0$. Výsledný odhad tedy vzniká smrštěním výběrové varianční matice S k diagonální matici. V článku [16] se optimální odhad

parametrů ρ, ν hledá prostřednictvím minimalizace střední kvadratické ztráty

$$\min_{\nu,\rho} \mathrm{E} \left\| \rho\nu I \ + (1-\rho)S \ - \Sigma \right\|_F^2 ,$$

kde $\|A\|_F = \sqrt{\mathrm{tr}(AA^\top)}$ značí Frobeniovu normu matice A. Při výpočtu minimalizace se využije, že $\mathrm{E}\, S = \Sigma$. Autoři zjistili, že optimální koeficienty lineární kombinace (2) jsou

$$\nu = \frac{1}{p}\,\mathrm{tr}(\Sigma), \qquad\qquad \rho = \frac{\beta^2}{\alpha^2+\beta^2} = \frac{\beta^2}{\delta^2}, \tag{3}$$

kde

$$\beta^2 = \frac{1}{p}\,\mathrm{E}\,\|S-\Sigma\|_F^2 , \qquad \alpha^2 = \frac{1}{p}\,\|\Sigma - \nu I\ \|_F^2 , \qquad \delta^2 = \frac{1}{p}\,\mathrm{E}\,\|S-\nu I\ \|_F^2 ,$$

které bohužel závisejí na neznámé varianční matici Σ. V článku [16] lze ale nalézt pro parametry ν, β, α a δ konzistentní odhady. Jejich dosazením do (2), s využitím vztahu (3) pro ρ, pak získáme pozitivně definitní odhad matice Σ.

Důvod, proč odhad (2) funguje dobře, lze nahlédnout z toho, že výběrová vlastní čísla (tj. vlastní čísla výběrové varianční matice S) jsou více rozptýlena kolem svého průměru než skutečná vlastní čísla Σ. V [16] je dokázáno, že použitím konvexní kombinace (2) dojde ke smrštění výběrových vlastních čísel k jejich průměru, a tím k celkovému zlepšení odhadu. Myšlenka vylepšení odhadu varianční matice prostřednictvím smrštění výběrových vlastních čísel k jejich průměru je podrobněji popsána v přehledovém článku [18].

Vzorec (2) může mít obecnější tvar

$$S_1^* = \rho T + (1-\rho)S, \tag{4}$$

kde T je cílová matice (tzv. *target matrix*), která disponuje kýženými vlastnostmi varianční matice, jako je plná hodnost a pozitivní definitnost. Cílová matice T se často volí diagonální.

Místo přímého odhadu koeficientu ρ ve vzorci (4), popř. (2), lze smrštění dosáhnout také tak, že se dostupný náhodný výběr rozšíří o další simulovaná pozorování, která se nagenerují z rozdělení s varianční maticí T. Z takto vzniklého souboru se pak vypočte výběrová varianční matice (1). Tato procedura má podobný efekt jako smrštění původní výběrové varianční matice S k cílové matici T. Konstrukce nových členů se odvíjí od předpokládaných statistických vlastností náhodného vektoru $\mathbf{X}$.

2.2. Posílení diagonály matice (tzv. tapering)

Efektivní a zároveň jednoduchý způsob jak potlačit rušivé kovariance ve výběrové varianční matici je vynásobit S po prvcích řídkou pozitivně definitní maticí M. Označíme-li výsledný odhad jako S_2, pak

$$S_2 = S \circ M, \tag{5}$$

kde $\circ$ značí Schurův součin. Tato metoda bývá v anglické literatuře označována jako *tapering* (v [21] jako *banding*) a matice S_2 jako *tapered matrix*. Díky řídkosti matice M je řídká i matice S_2, což přináší zejména v datové asimilaci velké výpočetní výhody.

Je-li M pozitivně definitní matice, pak díky větě o Schurově součinu (např. [10], str. 479, Věta 7.5.3) je zajištěno, že S_2 bude také pozitivně definitní.

V kontextu datové asimilace byla regularizace kovariance pomocí Schurova součinu navržena v [11] a [9] a vychází z principů modelování kovariance náhodného pole pomocí kovarianční funkce. Nechť tedy $\mathcal{X}(\boldsymbol{s})$ je spojité náhodné pole definované na omezené doméně $\mathfrak{D} \subset \mathbb{R}^3$ pokrývající část atmosféry Země a $\mathbf{X} = (X_1, \ldots, X_p)^\top$ představuje nějakou diskretizaci $\mathcal{X}$, tj. $X_i = \mathcal{X}(\boldsymbol{s}_i),\ \boldsymbol{s}_i \in \mathfrak{D}$. Funkce $C(\boldsymbol{s}, \boldsymbol{t}) = \mathtt{cov}(\mathcal{X}(\boldsymbol{s}), \mathcal{X}(\boldsymbol{t}))$ se nazývá kovarianční funkce pole $\mathcal{X}$. Platí, že pro každou diskretizaci $\mathbf{X}$ je matice s prvky $C(\boldsymbol{s_i}, \boldsymbol{s_j})$ pozitivně semidefinitní a naopak, je-li pro libovolný výběr bodů $\boldsymbol{s_i}$ kovarianční matice s prvky $C(\boldsymbol{s_i}, \boldsymbol{s_j})$ pozitivně semidefinitní, definuje C kovarianční funkci.

Matice M je v [11] konstruována tak, že je pozitivně definitní a její (i, j)-tý prvek M_{ij} je roven $\varrho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_{\mathbb{R}^3})$, kde $\varrho\colon [0, \infty) \to [0, 1]$ je funkce s kompaktním nosičem a $\|\cdot\|_{\mathbb{R}^3}$ je norma v prostoru $\mathbb{R}^3$. Jinými slovy, ϱ je možné ztotožnit s dobře definovanou korelační funkcí a (5) tedy modeluje kovarianci pole jako Schurův součin výběrové kovariance s korelační maticí danou funkcí ϱ. Zkonstruovat takovou funkci (zejména s kompaktním nosičem) je netriviální problém, kterým se zabývá článek [8], z něhož je v [9] vzata funkce ϱ hladká a monotónně klesající k nule, konkrétně jde o polynom pátého řádu [8, vztah (4.10)]. Graf tvarem připomíná pravou polovinu Gaussovy křivky, která však nabývá nulové hodnoty v konečné vzdálenosti, což zajišťuje řídkost M.

Již pod názvem *tapering* se tato regularizační metoda objevuje v článku [7], kde se matice M určuje z minimalizace vzdálenosti měřené střední čtvercovou chybou

$$\mathrm{MSE}(S_2) = \mathrm{E}\,\|\Sigma - S_2\|_F^2 = \mathrm{E}\left(\mathrm{tr}\left((\Sigma - S \circ M)^2\right)\right). \tag{6}$$

Minimalizace by se měla provádět přes třídu pozitivně definitních matic M, což je výpočetně náročné. Toto omezení se proto ignoruje a výsledná ma-

tice M, pro jejíž prvky lze při daném Σ získat explicitní formuli, se dodatečně převede na pozitivně definitní matici dalšími heuristickými postupy (např. se v jejím spektrálním rozkladu ponechají pouze kladná vlastní čísla a ostatní se položí rovna nějakému $\varepsilon > 0$).

Druhou možností je pak v souladu s články [11] a [9] matici M parametrizovat pomocí validní kovarianční funkce, která bude popisovat dosah korelací, a prostřednictvím minimalizace MSE odhadovat její parametry. Tam je již pozitivní definitnost zaručena, řídkost ale zajištěna není, což dělá tuto metodu poněkud méně výpočetně atraktivní.

2.3. Vynulování zanedbatelných prvků (tzv. tresholding)

Pro úplnost na závěr zmiňme metodu, pro niž se v anglické odborné literatuře používá termín *tresholding*. Myšlenka vychází z toho, že vysokodimenzionální varianční matice, které se vyskytují v praktických aplikacích, jsou obvykle řídké, a obsahují tudíž mnoho nulových prvků. Vylepšený odhad, navržený původně v [3], pak je

$$T_t(S) \equiv \left(s_{ij}\mathbf{1}_{[|s_{ij}|\geqslant t]} : i,j = 1,\ldots,p\right), \tag{7}$$

kde $T_t(S)$ značí operátor tresholdingu aplikovaný na výběrovou varianční matici a $t > 0$ je zvolený práh, pod nímž již prvky pokládáme za zanedbatelné. Autoři [3] doporučují t zvolit podle následujícího postupu: Výběr se náhodně rozdělí na dvě části o velikostech $N_1 = N\left(1 - \frac{1}{\log N}\right)$ a $N_2 = \frac{N}{\log N}$ a vypočtou se výběrové varianční matice S_{N_1} a S_{N_2}. Tento postup se opakuje K-krát a t se zvolí tak, aby minimalizovalo

$$R(t) = \frac{1}{K}\sum_{k=1}^{K} \|T_t(S_{N_1,k}) - S_{N_2,k}\|_F^2 . \tag{8}$$

Velkou předností metod jako tapering a tresholding je jednoduchá implementace. Nevýhodou však může být to, že pozitivní definitnost výsledku často není zajištěna. Pro tresholding je v [3] alespoň ukázána stejnoměrná konzistence v operátorové normě na množině řídkých matic, a to za podmínky, že $\frac{\log p}{N} \to 0$.

3. Parametrické metody

Pokud náhodný vektor $\mathbf{X}$ má specifické vlastnosti, které se promítají do tvaru jeho varianční matice, lze jejich zohledněním ve formě předpokladů odhad

výrazně vylepšit. Například u meteorologických veličin je často možné předpokládat normalitu, kovarianční stacionaritu nebo prostorovou markovskou vlastnost. Každá z těchto vlastností poskytuje možnost vylepšení odhadu varianční matice pomocí specifického parametrického modelu, jehož parametry lze odhadnout standardními statistickými metodami. Přesnost výsledného odhadu a jeho kvalita, která se projeví v dalším použití (např. v datové asimilaci), se pak odvíjí od toho, jak realistické tyto dodatečné předpoklady byly.

U parametrických metod je nezbytné vyhnout se příliš komplexním modelům s velkým počtem parametrů. Proto řada přístupů využívá různé transformace, které vedou k přibližné diagonalitě varianční matice v transformovaném prostoru. Tím dochází k velké redukci počtu parametrů.

3.1. Regularizace ve spektrálním a waveletovém prostoru

Tento přístup je založen na reprezentaci vektoru $\mathbf{X}$ rozvojem

$$\mathbf{X} = \mathrm{E}\,\mathbf{X} + \sum_{i=1}^{p} d_i^{1/2} \xi_i \mathbf{v}_i, \tag{9}$$

kde d_i jsou koeficienty, ξ_i jsou nezávislé náhodné veličiny s nulovou střední hodnotou a jednotkovým rozptylem a $\mathbf{v_i}$ jsou ortonormální vektory v $\mathbb{R}^p$. Pro varianční matici pak máme

$$\Sigma = FDF^\top, \qquad \text{kde } D = \mathrm{diag}(d_1, \ldots, d_p) \tag{10}$$

a matice F má ve sloupcích vektory $\mathbf{v}_i$. Je-li (10) spektrální rozklad varianční matice, dostáváme známý rozklad do hlavních komponent (Karhunen-Loève). Ve velké dimenzi je ale většina výběrových vlastních čísel nulová a odhad teoretického rozvoje (9) je obtížný. Pro regularizaci se proto využívají vhodné deterministicky dané báze (nebo obecněji i framy jako v případě waveletové transformace).

Modelujeme-li pole $\mathbf{X}$ pomocí (9), můžeme matici Σ odhadnout prostřednictvím odhadu diagonální matice $D = F^\top \Sigma F$. Pokud nepřijmeme žádné další předpoklady, je možné podobně jako v předchozí sekci použít některou neparametrickou metodu na výběrovou varianční matici S pole $\mathbf{X}$ převedenou do spektrálního prostoru, tj. na $F^\top S F$. Tato matice přirozeně není přesně diagonální, ale obsahuje (v praxi často malé) nenulové mimodiagonální členy. Nejjednodušším způsobem regularizace odhadu je pak jejich zanedbání [13]. Jedná se vlastně o *tapering* spektrální varianční matice podle vzorce (5) pro $M = I$.

Za předpokladu konkrétního rozdělení vektoru $\mathbf{X}$ lze diagonální prvky $\{d_{ii}\}_{i=1}^{p}$ spektrální varianční matice D odhadnout, například metodou maximální věrohodnosti založenou na náhodném výběru $\mathbf{X}_1, \ldots, \mathbf{X}_N$. V případě normálního rozdělení dostáváme [25]

$$\widehat{d}_{ii} = \frac{1}{N} \sum_{k=1}^{N} X_{ki}^2, \quad i = 1, \ldots, p, \tag{11}$$

kde X_{ki} značí i-tý prvek vektoru $\mathbf{X}_k$, $k = 1, \ldots, N$.

3.1.1. Regularizace ve spektrálním prostoru Doposud jsme nepředpokládali nic o výběru ortogonální matice F. Podle očekávání častou volbou pro F bude Fourierova báze. V meteorologii podobně jako ve fyzice obecně se často vyskytují děje difúzního typu a je přirozené, že se modely popisující difúzi úspěšně používají i pro popis stochastických vlastností některých fyzikálních veličin. Tak se ukazuje souvislost mezi stacionaritou pole, Fourierovou transformací a Laplaceovou rovnicí, a je možné založit na ní regularizační techniky. Pokud např. můžeme považovat vektor $\mathbf{X}$ za pozorování omezeného časového úseku kovariančně stacionárního procesu $\{X_t,\, t \in \mathbb{Z}\}$ pro $t = 1, \ldots, p$, vede diskrétní Fourierova transformace vektoru $\mathbf{X}$ k přibližné diagonalitě[1] varianční matice v transformovaném prostoru [6], tj. můžeme předpokládat model

$$\Sigma = FDF^{\top}, \tag{12}$$

kde matice F provádí Fourierovu transformaci.

V meteorologických aplikacích vektor $\mathbf{X}$ často představuje diskretizaci spojité omezené domény, tedy oblasti pokrývající část atmosféry Země, a index má význam souřadnice v prostoru. Předpokládejme pro jednoduchost, že spojitou doménou je úsečka $[0, 1]$ a vektor $\mathbf{X}$ délky p ji rozděluje na $p + 1$ dílků délky $h = \frac{1}{p+1}$. Nyní můžeme využít faktu, že matice F je tvořena vlastními vektory diskrétního Laplaceova operátoru $L\colon \mathbb{R}^p \to \mathbb{R}^p$, který je v jedné dimenzi identický s operátorem druhé derivace a lze ho reprezentovat maticí

$$L = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & 0 & \ldots & 0 \\ 1 & -2 & 1 & \ldots & 0 \\ & & \ddots & & \\ 0 & \ldots & 1 & -2 & 1 \\ 0 & \ldots & 0 & 1 & -2 \end{pmatrix}. \tag{13}$$

[1] Matice Σ je v transformovaném prostoru přesně diagonální pouze v případě, kdy je proces $\{X_t,\, t \in \mathbb{Z}\}$ periodický s periodou p, tj. $X_t = X_{t+p},\ \forall t \in \mathbb{Z}$.

Pak pro $\mathbf{u} = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p$ je j-tá složka vektoru $L(\mathbf{u}) \equiv L\mathbf{u}$ rovna

$$L(\mathbf{u})_j = \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2},$$

kde $j = 1, \ldots, p$. Pro body u_0 a u_{p+1} (ležící na krajích úsečky $[0, 1]$) se uvažují různé okrajové podmínky. Definice (13) odpovídá tzv. Dirichletově okrajové podmínce, při níž je $u_0 = u_{p+1} = 0$. Vlastní vektory $\mathbf{v}_i$ mají v tomto případě tvar

$$\mathbf{v}_i = \sqrt{\frac{2}{p+1}} \begin{pmatrix} \sin\left(\frac{1}{p+1} i\pi\right) \\ \sin\left(\frac{2}{p+1} i\pi\right) \\ \vdots \\ \sin\left(\frac{p}{p+1} i\pi\right) \end{pmatrix}, \quad i = 1, \ldots, p \tag{14}$$

a dostáváme (jak je možno se přesvědčit) spektrální rozklad $L = F\Lambda F^\top$, kde Λ je diagonální matice s vlastními čísly

$$\lambda_i = -4(p+1)^2 \sin^2\left(\frac{\pi}{2(p+1)} i\right), \quad i = 1, \ldots, p \tag{15}$$

na diagonále.

Máme tedy $L = F\Lambda F^\top$ a $\Sigma = FDF^\top$ a je přirozené modelovat Σ jako funkci Laplaceova operátoru. Je-li totiž $L = F\Lambda F^\top$, pak pro spojitou funkci f lze definovat matici $f(L)$ pomocí spektrálního rozkladu $Ff(\Lambda)F^\top$, kde $f(\Lambda) = \mathrm{diag}(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_p))$.

Vlastní čísla d_{ii} můžeme modelovat jako vhodnou funkci vlastních čísel (15) Laplaceova operátoru, tj. $d_{ii} = f(\lambda_i)$, $i = 1, \ldots, p$. Jelikož $\mathbf{X}$ představuje diskretizaci spojitého náhodného pole $\mathcal{X}$ na kompaktní oblasti $\mathfrak{D}$, lze variační matici Σ pokládat za diskretizaci variančního operátoru $\mathcal{C}$ pole $\mathcal{X}$. Za předpokladu spojitosti kovarianční funkce $\mathrm{cov}(\mathcal{X}(\boldsymbol{s}), \mathcal{X}(\boldsymbol{t}))$, $\boldsymbol{s}, \boldsymbol{t} \in \mathfrak{D}$ pole $\mathcal{X}$ lze pomocí metod známých z numerické matematiky ukázat, že vlastní čísla $\{d_{11}, \ldots, d_{pp}\}$ matice Σ konvergují pro $p \to \infty$ k vlastním číslům variančního operátoru $\mathcal{C}$ [2, 19]. Jelikož každý varianční operátor musí mít konečnou stopu ([14], Věta 2.1), musejí $\{d_{ii}\}_{i=1}^p$ s rostoucím i rychle klesat k nule i pro p konečné.

Předpokládejme tedy $\Sigma = Ff(\Lambda)F^\top$ pro nějakou vhodnou f, a tudíž $d_{ii} = f(\lambda_i)$. Kvůli požadavku na konečnost stopy varianční matice musí funkce f pro $\lambda \to -\infty$ rychle klesat k nule, neboť $\{\lambda_i\}_{i=1}^p$ je klesající posloupnost záporných čísel.

V praxi se užívá zejména polynomiální a exponenciální model:

$$d_{ii} = c(-\lambda_i)^{-\alpha}, \quad \alpha > 1 \quad \text{(polynomiální model)}$$
$$d_{ii} = c\mathrm{e}^{\alpha\lambda_i} \quad \text{(exponenciální model)},$$

kde $i = 1, \ldots, p$. Koeficienty c, α jsou parametry daného modelu. To představuje v daném případě snížení počtu parametrů z p na 2, přesto i takto jednoduchý model může být přijatelný. Za předpokladu konkrétního rozdělení $\mathbf{X}$ lze pak parametry odhadnout například metodou maximální věrohodnosti [25]. Použití takového modelu koriguje strukturu vlastních čísel podobně jako shrinkage (viz začátek sekce 2.1.), vyhlazuje také průběh odhadů $\widehat{d}_{ii}$ a tím přispívá k redukci šumu.

Náš příklad byl pouze ilustrativní. Zajímavé a hluboké vztahy mezi určitým typem stochastické rovnice difúze, náhodnými poli s (prostorovou) markovskou vlastností a kovariancemi z tzv. Matérnovy třídy popisuje článek [15] a též přehledný článek [23]. Použití těchto metod v praxi pak ilustruje [17].

3.1.2. Regularizace ve waveletovém prostoru Reprezentuje-li $\mathbf{X}$ náhodný vektor či náhodné pole v prostoru jako je tomu u meteorologických veličin, je zřejmé, že takové pole může mít jak charakteristiky vlnové povahy tak i lokální vlastnosti vázané na umístění v prostoru. Fourierova transformace postihne frekvenční komponenty vektoru $\mathbf{X}$, ale nezachytí lokální jevy. Nacházíme zde podobně jako v řadě dalších aplikací určitou komplementaritu spektrální a prostorové reprezentace.

Nástrojem ke „kompromisní“ reprezentaci takového náhodného pole je tzv. waveletová transformace [4], která je realizována podobně jako u Fourierovy transformace rozvojem typu (9) a rozkladem varianční matice (10). Matice transformace se skládá z bázových funkcí, tvořených různě umístěnými a škálovanými vlnkami (odtud název *wavelets*). Obecněji je možno v rozvoji (9) použít místo ortonormální báze tzv. frame, který má některé vlastnosti báze, ale může mít i více než p členů. Různé volby framů pak dávají waveletové reprezentace různých typů.

Po transformaci varianční matice do waveletového prostoru lze pak zanedbáním příslušných koeficientů vyřadit bázové funkce s nízkou vypovídací schopností. Často je možné podobně jako u Fourierovy transformace zanedbat mimodiagonální prvky transformované matice, což odpovídá lokálnímu průměrování prostorových korelací [20]. Můžeme také předpokládat vhodnou řídkou strukturu kovarianční matice ve waveletovém prostoru. Po transformaci zpět do původního prostoru pak dostáváme varianční matici očištěnou o šum.

Reprezentace ve waveletovém prostoru často přináší dramatické snížení dimenze problému. Koeficienty bázových funkcí lze pak odhadovat pomocí vhodného parametrického nebo i semiparametrického modelu.

3.1.3. Využití dalších typů rozkladu varianční matice k jejímu odhadu Kromě spektrální a waveletové transformace lze ke konstrukci pozitivně definitního odhadu varianční matice využít i jiné typy rozkladu. Jednoduchý rozklad

$$\Sigma = BRB \tag{16}$$

na matici směrodatných odchylek B a korelační matici R má velký praktický význam, neboť oba faktory rozkladu jsou snadno interpretovatelné. To umožňuje odhadovat B i R zvlášť, což může mít význam zejména v aplikacích, kde je jeden z těchto prvků důležitější.

Druhým typem rozkladu, který se při odhadování s úspěchem používá, je Choleského rozklad

$$\Sigma = CC^\top, \tag{17}$$

kde C je (pro pozitivně definitní Σ) jednoznačně daná dolní trojúhelníková matice s pozitivními prvky na diagonále. V určitých případech může být výhodnější modifikovaný Choleského rozklad

$$\Sigma = LD^2L^\top, \tag{18}$$

kde $L = CD^{-1}$ je dolní trojúhelníková matice a D^2 je diagonální matice, kterou lze opět modelovat zvlášť.

Při těchto typech odhadu lze s výhodou využít znalost (prostorových nebo časových) vazeb mezi složkami vektoru $\mathbf{X}$. Zafixováním veličin ve vhodném pořadí lze totiž docílit pásové struktury Σ a následně řídkosti některých členů ve zvoleném rozkladu.

3.2. Lineární model kovariance

Výše popsané metody nepředpokládaly žádnou speciální strukturu varianční matice nebo její inverze a zabývaly se převážně regularizací výběrové kovarianční matice v původním nebo transformovaném prostoru. Alternativou k těmto metodám je do jisté míry komplementární přístup, který vychází ze znalosti speciální struktury varianční matice nebo její inverze. Taková struktura může být dána prostorovými nebo i frekvenčními charakteristikami výchozího náhodného pole. Často lze strukturu varianční matice dobře postihnout lineárním modelem [1]

$$\Sigma = \alpha_1 U_1 + \ldots + \alpha_q U_q, \tag{19}$$

kde U_i jsou známé, lineárně nezávislé (v prostoru matic), symetrické matice a α_i jsou neznámé parametry takové, že výsledná matice je pozitivně definitní. Tento model je zcela obecný, neboť pro $q = \frac{p(p+1)}{2}$ lze každou varianční matici zapsat ve tvaru (19) s $\alpha_{ij} = \sigma_{ij} = \sigma_{ji}$ a maticemi U_{ij}, které jsou nulové až na jednotky na pozicích (i, j) a (j, i). Teoreticky tak pokrývá libovolnou odhadovací metodu pracující po prvcích, jako je např. *tapering* nebo *banding*. Podrobnější přehled metod vycházejících z lineárního modelu lze najít v [21].

Parametry $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ je možné odhadovat standardními statistickými metodami, jejich použití ale značně komplikuje podmínka, že výsledná lineární kombinace musí být pozitivně definitní matice. Nutné a postačující podmínky pro existenci explicitního maximálně věrohodného odhadu jsou k dispozici v článku [24]. K výpočtu řešení věrohodnostní rovnice je přitom použita iterativní metoda navržená v [1].

Jednou z možností jak zajistit pozitivní definitnost odhadu je použít lineární model na logaritmus varianční matice

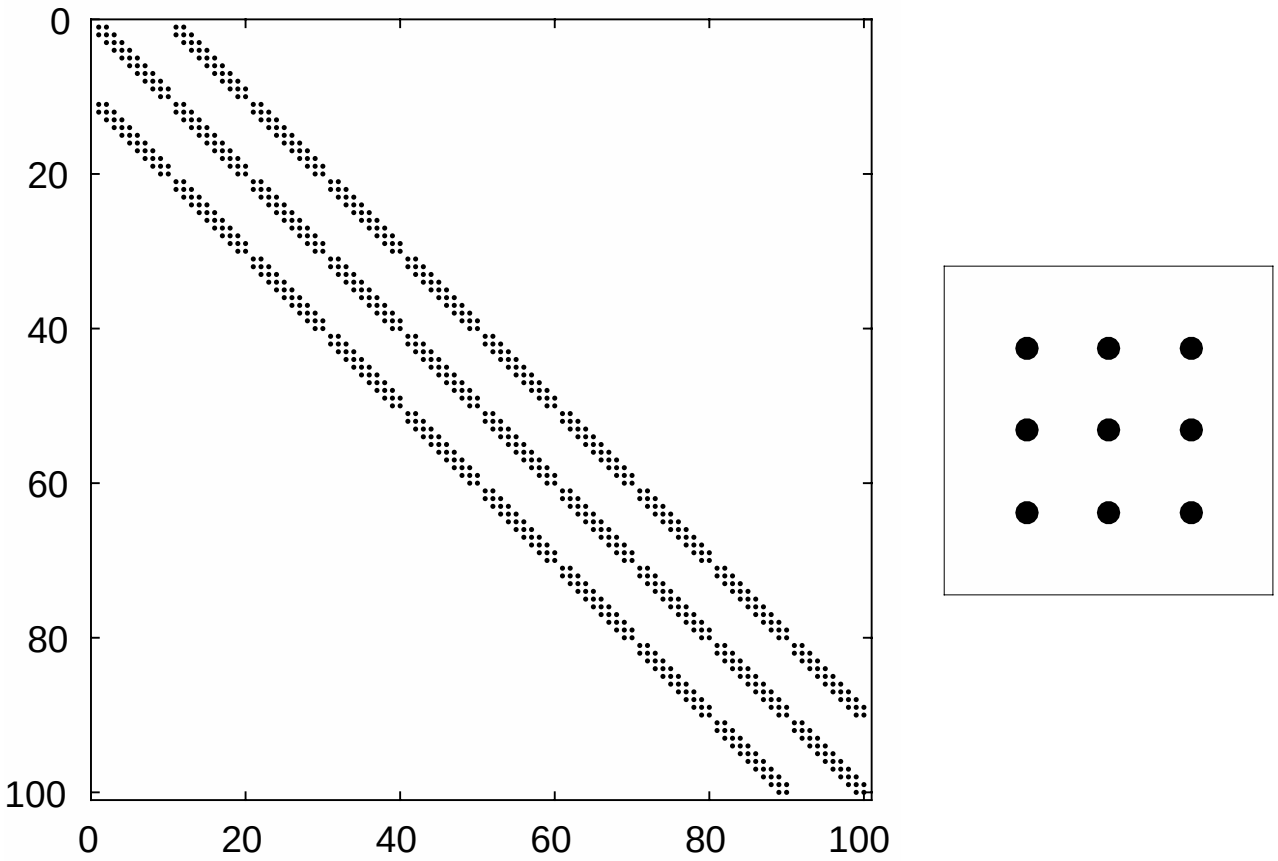
$$\log \Sigma = \alpha_1 U_1 + \ldots + \alpha_q U_q, \tag{20}$$

kde U_i jsou opět známé matice, ale na α_i nejsou kladeny žádné další požadavky. Nevýhodou tohoto přístupu je to, že $\log \Sigma$ nemá žádnou statistickou interpretaci. Autoři [5] uvádějí maximálně věrohodný odhad pro parametry modelu (20), a to včetně jeho asymptotických vlastností.

3.3. Regularizace v inverzním prostoru

Motivací pro regularizaci varianční matice skrze její inverzi, tzv. matici přesnosti, je fakt, že u normálně rozdělených vektorů $\mathbf{X}$ má inverze korelační matice na (i, j)-tém místě prvek úměrný parciálnímu korelačnímu koeficientu mezi X_i a X_j při vyloučení všech ostatních prvků.

Má-li tedy vektor $\mathbf{X}$ normální rozdělení, pak (i, j)-tý prvek matice přesnosti je nulový právě tehdy, když X_i a X_j jsou nezávislé podmíněně na všech zbývajících prvcích [22]. Tento výsledek lze s výhodou využít u polí, která jsou navíc tzv. prostorově markovská. Vektor $\mathbf{X}$ má prostorovou markovskou vlastnost, pokud pro každé $i = 1, \ldots, p$ je podmíněné rozdělení X_i závislé pouze na bodech z jeho okolí. Na ostatních bodech (které jsou daným okolím od bodu X_i odděleny) je X_i podmíněně nezávislé. Odtud vyplývá, že matice přesnosti gaussovského markovského pole je řídká pásová matice. Toho je možno využít pro regularizaci. Obrázek 1 ilustruje nenulové prvky matice přesnosti dvourozměrného markovského pole $\mathbf{X}$, jehož každý bod závisí pouze na 8 nejbližších sousedech.



Obrázek 1: Matice přesnosti dvourozměrného markovského pole o rozměru 10×10 bodů (body jsou uspořádány do sloupce), v němž každý bod závisí na 8 nejbližších sousedech. Vpravo je schéma bodu a jeho 8 nejbližších sousedů.

Metody použité v [15], [17] a [23], o kterých jsme se zmiňovali v odstavci 3.1.1. v souvislosti se stochastickou rovnicí difúze, se týkají gaussovských polí s markovskou vlastností a spadají tedy svojí povahou do modelování v inverzním prostoru. Níže uvedeme ještě další použitelné techniky.

3.3.1. Lineární model pro matici přesnosti Lineární model lze použít jak pro varianční matici tak pro její inverzi, což je pro gaussovské markovské pole velmi výhodné. Uvažujme model

$$\Sigma^{-1} = \beta_1 A_1 + \ldots + \beta_r A_r, \tag{21}$$

kde matice $A_1, \ldots, A_r$ jsou známé řídké matice a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$ jsou neznámé parametry. Vhodnou volbou pro matice $A_1, \ldots, A_r$ mohou být opět matice A_{ij}, které mají na místě (i, j) a (j, i) jedničku a jinde nuly. Obdobně je možno použít symetrické matice obsahující nenulové hodnoty pouze na vybraných subdiagonálách, nebo jejich úsecích.

Parametry $\boldsymbol{\beta}$ lze odhadnout například metodou maximální věrohodnosti [27]. Výsledný odhad ale není dán explicitně a je výsledkem numerické maximalizace věrohodnosti. Explicitní vzorec pro odhad parametrů (viz [26]) lze získat metodou tzv. score-matchingu [12], která, stejně jako metoda maximální věrohodnosti, poskytuje za určitých předpokladů konzistentní odhad.

4. Závěr

Pokrok současné měřicí a výpočetní techniky vede často k situaci, kdy rozsah naměřených či modelových dat je velký, ale možnost opakování experimentu je velmi omezená. Problematika odhadu varianční matice na základě náhodného výběru, jehož rozsah je malý v porovnání s dimenzí jednotlivých vektorů, se tak objevuje v mnoha statistických aplikacích. V tomto článku jsme se zaměřili zejména na metody odhadu používané při datové asimilaci, pro niž je kvalitní odhad varianční matice klíčovým prvkem ovlivňujícím kvalitu následné predikce, například při numerické předpovědi počasí. Metody jsou však vyvíjeny v souvislosti s mnoha dalšími obory, často souběžně v různých vědeckých komunitách, a je zde řada otevřených problémů.

Výčet metod uvedených v tomto článku není jistě vyčerpávající, ale doufáme, že představuje přehled základních přístupů a směrů výzkumu v dané oblasti.

Poděkování

Vznik článku byl částečně podporován z grantu TA ČR č. TL01000238. Děkujeme Pavlu Krčovi, který zachránil text článku před jeho kompletním ztracením.

Literatura

[1] Anderson, T. W. (1973): Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure. *Ann. Stat.* **1**(1), 135–141, doi: 10.1214/aos/1193342389. *cit. 34, 35*

[2] Atkinson, K. (1975): Convergence Rates for Approximate Eigenvalues of Compact Integral Operators. *SIAM J. Numer. Anal.* **12**(2), 213–222. *cit. 32*

[3] Bickel, P. J., Levina, E. (2008): Covariance Regularization by Thresholding. *Ann. Stat.* **36**(6), 2577–2604, doi: 10.1214/08-AOS600. *cit. 29*

[4] Burrus, C. S., Gopinath, R. A., Guo, H. (1998): *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, New Jersey. *cit. 33*

[5] Chiu, T. Y. M., Leonard, T., Tsui, K. W. (1996): The Matrix-Logarithmic Covariance Model. *J. Amer. Statist. Assoc.* **91**, 198–210, doi: 10.2307/2291396. *cit. 35*

[6] Dwivedi, Y., Rao, S. S. (2011): A Test for Second-Order Stationarity of a Time Series Based on the Discrete Fourier Transform. *J. Time Ser. Anal.* **32**, 68–91, doi: 10.1111/j.1467-9892.2010.00685.x. *cit. 31*

[7] Furrer, R., Bengtsson, T. (2007): Estimation of High-Dimensional Prior and Posterior Covariance Matrices in Kalman Filter Variants. *J. Multivar. Anal.* **98**(2), 227–255, doi: 10.5167/uzh-21542. *cit. 28*

[8] Gaspari, G., Cohn, S. E. (1999): Construction of Correlation Functions in Two and Three Dimensions. *Q. J. R. Meteorol. Soc.* **125**(554), 723–757, doi: 10.1002/qj.49712555417. *cit. 28*

[9] Hamill, T. M., Whitaker, J.S., Snyder, Ch. (2001): Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter. *Mon. Weather Rev.* **129**(11), 2776–2790, doi: 10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2. *cit. 28, 29*

[10] Horn, R. A., Johnson, Ch. R. (2013): *Matrix Analysis*. Cambridge University Press, Second Edition. *cit. 28*

[11] Houtekamer, P. L., Mitchell, H. L. (2001): A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. *Mon. Weather Rev.* **129**(1), 123–137, doi: 10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2 *cit. 28, 29*

[12] Hyvärinen, A. (2005): Estimation of Non-Normalized Statistical Models by Score Matching. *J. Mach. Learn. Res.* **6**(24), 695–709, doi: 10.5194/npg-22-485-2015. *cit. 37*

[13] Kasanický I., Mandel J., Vejmelka M. (2015): Spectral Diagonal Ensemble Kalman Filters. *Nonlinear Process. Geophys.* **2**, 115–143, doi: 10.5194/npg-22-485-2015. *cit. 30*

[14] Kuo, H. H. (1975):. *Gaussian Measures in Banach Spaces*. Lecture Notes in Mathematics. Vol. 463, Springer-Verlag, Berlin, doi: 10.1007/BFb0082007. *cit. 32*

[15] Lindgren, F., Rue, H., Lindström, J. (2011): An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(4), 423–498, doi: 10.1111/j.1467-9868.2011.00777.x. *cit. 33, 36*

[16] Ledoit, O., Wolf, M. (2004): A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *J. Multivar. Anal.* **88**(2), 365–411, doi: 10.1016/S0047-259X(03)00096-4. *cit. 26, 27*

[17] Mirouze, I., Weaver, A. T. (2010): Representation of Correlation Functions in Variational Assimilation Using an Implicit Diffusion Operator. *Q. J. R. Meteorol. Soc.* **136**, 1421–1443, doi: 10.1002/qj.643. *cit. 33, 36*

[18] Muirhead, R. J. (1987): Developments in Eigenvalue Estimation. *Adv. Multivariate Statist. Anal.* 277–288, doi: 10.1007/978-94-017-0653-7_14. *cit. 27*

[19] Osborn, J. E. (1975): Spectral Approximation for Compact Operators. *Math. Comput.* **29**(131), 712–725, doi: 10.2307/2005282. *cit. 32*

[20] Pannekoucke O., Berre, L., Desroziers, G. (2007): Filtering Properties of Wavelets for Local Background-Error Correlations. *Q. J. R. Meteorol. Soc.* **133**, 363–379, doi: 10.1002/qj.33. *cit. 33*

[21] Pourahmadi, M. (2011): Covariance Estimation: The GLM and Regularization Perspectives. *Stat. Sci.* **26**(3), 369–387, doi: 10.1214/11-STS358. *cit. 28, 35*

[22] Rue, H., Held, L. (2005): Gaussian Markov Random Fields: Theory And Applications, *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC. *cit. 35*

[23] Simpson, D., Lindgren, F., Rue, H. (2012): Think Continuous: Markovian Gaussian Models in Spatial Statistics. *Spat. Stat.* **1**, 16–29, doi: 10.1016/j.spasta.2012.02.003. *cit. 33, 36*

[24] Szatrowski, T. H. (1980): Necessary and Sufficient Conditions for Explicit Solutions in the Multivariate Normal Estimation Problem for Patterned Means and Covariances. *Ann. Stat.* **8**(4), 802–810, doi: 10.1214/aos/1176345072. *cit. 35*

[25] Turčičová, M., Mandel J., Eben, K. (2019): Multilevel Maximum Likelihood Estimation with Application to Covariance Matrices. *Commun. Stat. – Theory and Methods* **48**(4), 909–925, doi: 10.1080/03610926.2017.1422755. *cit. 31, 33*

[26] Turčičová, M., Mandel, J., Eben, K. (2020): Score Matching Filters for Gaussian Markov Fields with Linear Model of Precision Matrix. *Článek se připravuje. cit. 37*

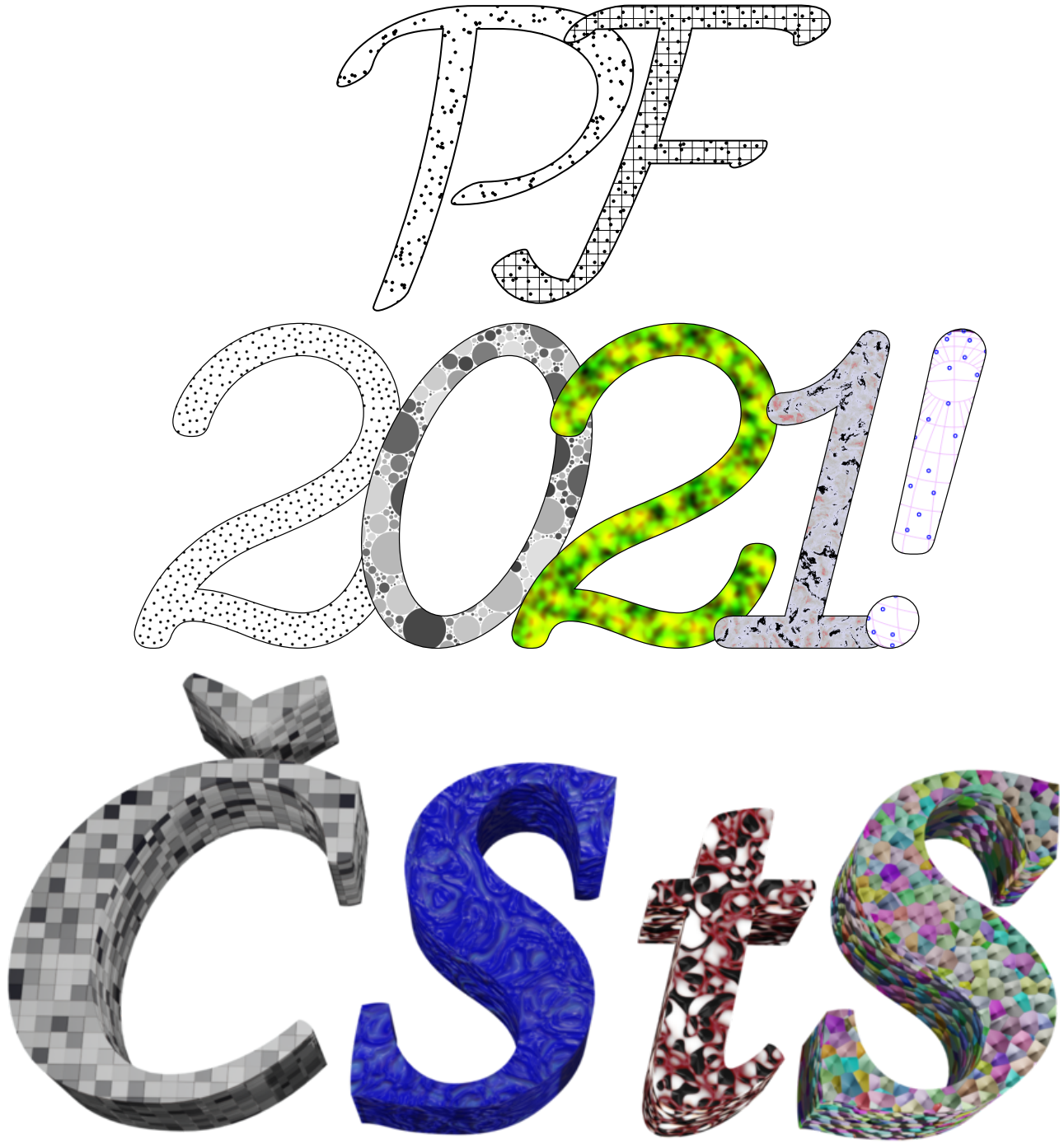
[27] Ueno, G., Tsuchiya, T. (2009): Covariance Regularization in Inverse Space. *Q. J. R. Meteorol. Soc.* **135**(642), 1133–1156, doi: 10.1002/qj.445. *cit. 37*

PF2021! ANEB NENÍ ŠUM JAKO ŠUM
PF2021! OR DIFFERENT TYPES OF NOISE

Pavel Stříž

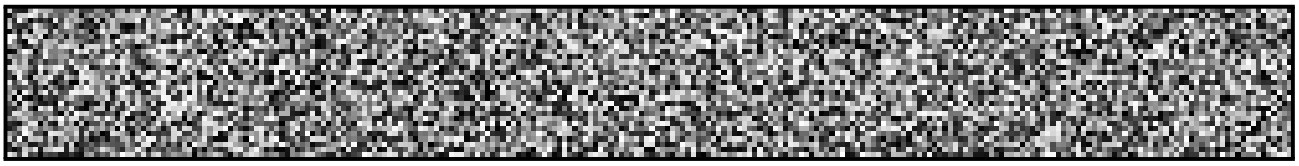
E-mail: pavel@striz.cz

Motto: [...] 12585 33893 43498 [...], str. 165, řádek 08210, sloupce 5–7.
A Million Random Digits with 100,000 Normal Deviates
www.rand.org/pubs/monograph_reports/MR1418.html



1. Úvodem pár vzpomínek

Mé první vzpomínky na práci s náhodností spadají pod Sinclair BASIC na základní škole, kdy jsme např. volili den v týdnu jako jedno z čísel 1 až 7 a dál s tím pracovali. Pár let poté jsem podobné příklady procvičoval v QBASICu s otevřenou knihou *Sbírka úloh z programování* od manželů Töpferových. V průběhu času se člověk dostal k fyzice a šumu u televizí až ke kosmickému záření.



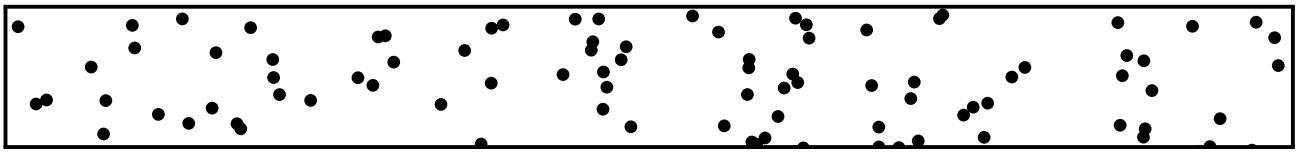
Zaujala mě práce Norberta Wienera, nejen jeho aplikace zpětné vazby, ale i práce nad Brownovým pohybem. To jsou ty známé základoškolské fyzikální pokusy vhození hypermanganu do vody. Dnes je bílý šum (angl. white noise) brán vážně ve všech významnějších oblastech přírodovědeckého bádání, včetně ekonomie, informatiky, statistiky a geometrie.

Z poslední doby se mi do ruky dostala kniha Davida Johnstona *Random Number Generators—Principles and Practices* z roku 2018, kde se to hemží kódy v C a Pythonu. Dle MSC2020 bychom naši článkovou rešerši začali asi v oblastech 60H40 (White noise theory), 65Cxx (Probabilistic methods...) či 11K45 (Pseudo-random numbers; Monte Carlo methods).

2. PF aneb Základ černobíle

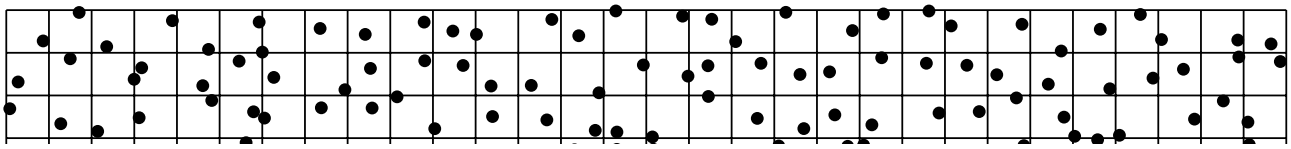
Na první ukázce, písmenku $\mathcal{P}$, vidíme typickou situaci užití pseudonáhodných čísel ve 2D (angl. pseudo-random numbers). Bez ohledu na kvalitu statistických vlastností to není po vizuální stránce příjemné. Jsou tam prázdná oka, kruhy se mohou překrývat. K testování lze doporučit `www.random.org`. Zde je ukázka přes Ti*k*Z.

```
\documentclass{standalone}
\usepackage{tikz}
\tikzset{inner sep=0pt, outer sep=0pt}
\begin{document}
\begin{tikzpicture}
\foreach \x in {1,...,900} {
  \pgfmathparse{rnd*100} \let\malx=\pgfmathresult
  \pgfmathparse{rnd*100} \let\maly=\pgfmathresult
  \node[xshift=\malx mm, yshift=\maly mm, circle, fill, minimum width=1mm]{};
} % end of \foreach \x
\end{tikzpicture}
\end{document}
```



Druhou stranou mince by byla dokonalá mřížka z bodů. Spojení obou nápadů vzniká stratifikovaný výběr (angl. supersampling či jittered grid). Prvně si plochu rozdělíme na menší čtverce a v každém volíme po jednom bodu. Chceme-li bodů víc, zjemníme mřížku. Dostáváme písmenko $\mathcal{F}$. Vizuálně je to lepší, ale stále tam jsou místy body u sebe a občas řeky. To vadí především typografům z čteného textu odstavců. U mřížky se mohou objekty překrývat. Opět vzorek přes Ti*k*Z.

```
\documentclass{standalone}
\usepackage{tikz} \tikzset{inner sep=0pt, outer sep=0pt}
\begin{document}
\begin{tikzpicture}
\draw[step=3.3333mm] (0,0) grid (100mm,100mm);
\foreach \x in {1,...,30} {
  \foreach \y in {1,...,30} {
    \pgfmathparse{3.3333*(rnd-1+\x)} \let\malx=\pgfmathresult
    \pgfmathparse{3.3333*(rnd-1+\y)} \let\maly=\pgfmathresult
    \node[xshift=\malx mm, yshift=\maly mm, circle, fill, minimum
     width=1mm]{};
  } % end of \foreach \y
} % end of \foreach \x
\end{tikzpicture}
\end{document}
```

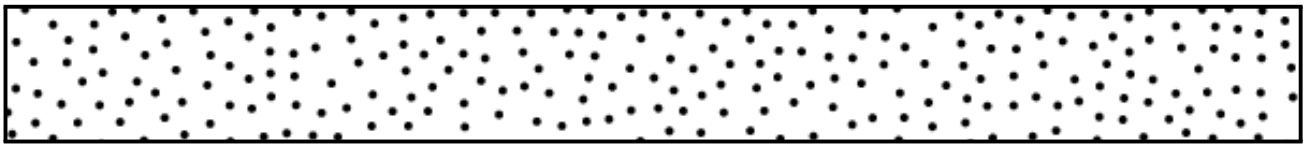


3. 2021 aneb Přes stupně šedi do barvy

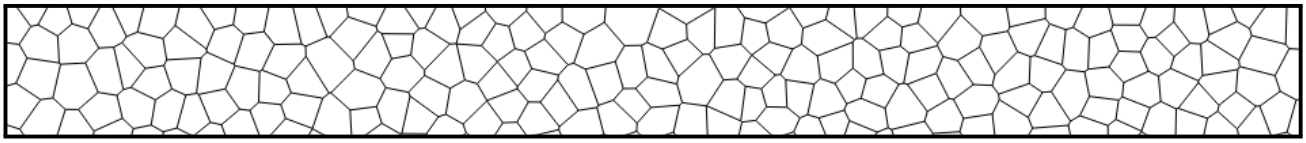
Grafici šli dál.

3.1. Robert Bridson

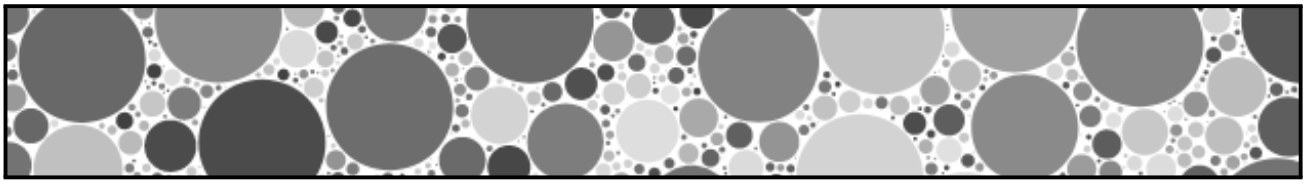
Jason Davies (zde, je znám hlavně díky mrakům slov) či Mike Bostock (zde, zakladatel serverů `bl.ocks.org` a `observablehq.com`) představují Bridsonův algoritmus (2007) generování bodů s geometrickým vztahem, že žádný nový bod nesmí být blíž než určená vzdálenost (angl. Poisson-disc sampling). Nelze již mluvit o pokusu náhodného generování, neb existuje mezi body vztah. Vizuálně je to pro lidské oko příjemné. Ukázka je v první číslici *2.* Davies vykresluje body přes canvas HTML5, Bostock do svg přes `D3js`.



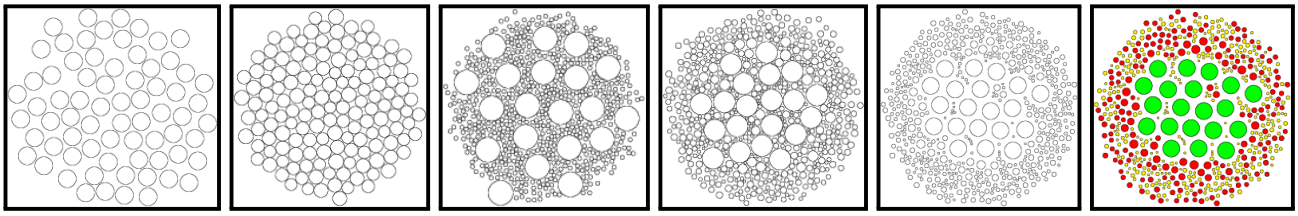
Ze středů lze snadno získat Voronoi diagram.



Zobecnění přináší algoritmus Mitchell's best-candidate, kdy se generuje sada k bodů a vybírá se z nich jen jeden, takový, který je nevzdálenější vůči všem ostatním. To dává možnost např. volit různý poloměr kruhů (číslice *0*).



Zájemci mohou nahlédnout na mé starší pokusy přes Lua, byť tedy užité algoritmy jsou pomalé.

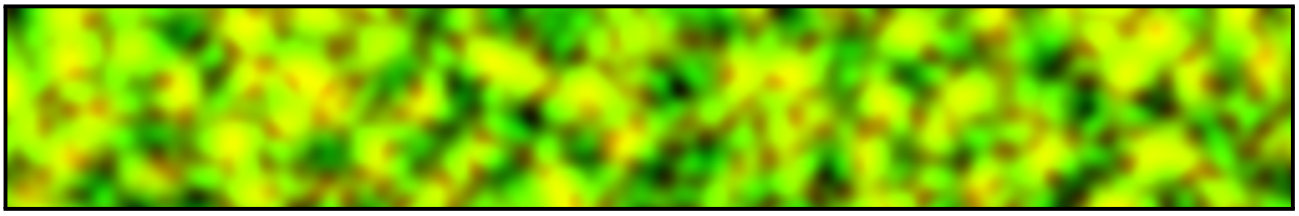


3.2. Ken Perlin

Mezník představuje Perlinův šum (angl. Perlin noise) z roku 1983 (příspěvek Kena Perlina z roku 1985 se jmenuje *An Image Synthesizer*), kdy dochází k interpolaci mezi body. Tím lze snadno vytvářet šum ve 2D a 3D, se zahrnutím času či barvy i ve vyšších rozměrech. Zde je typická ukázka, vypadá to jak výškový model (angl. DEM) známý z geografických inf. systémů.

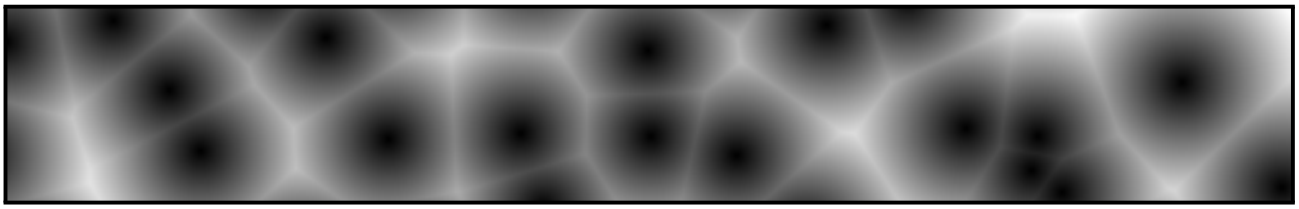
Za pozornost stojí i článek Ken Perlin a Fabrice Neyret: *Flow Noise*. Podobné výsledky dostáváme pomocí algoritmu Simplex noise a volně dostupné varianty OpenSimplex noise. Zde jsou dostupné implementace v Javě, JavaScriptu a nezapomínejme na Rust. Zaujal mě článek *Recursive Wang Tiles for Real-Time Blue Noise*. Pro studenty lze doporučit na YouTube kanál The Coding Train Daniela Shiffmana v jeho oblíbeném nástroji `p5js` a jeho knihu *The Nature of Code*.

Druhá cifra *2* vznikla v JavaScriptu v balíčku `simplex-noise` přes `npm`.

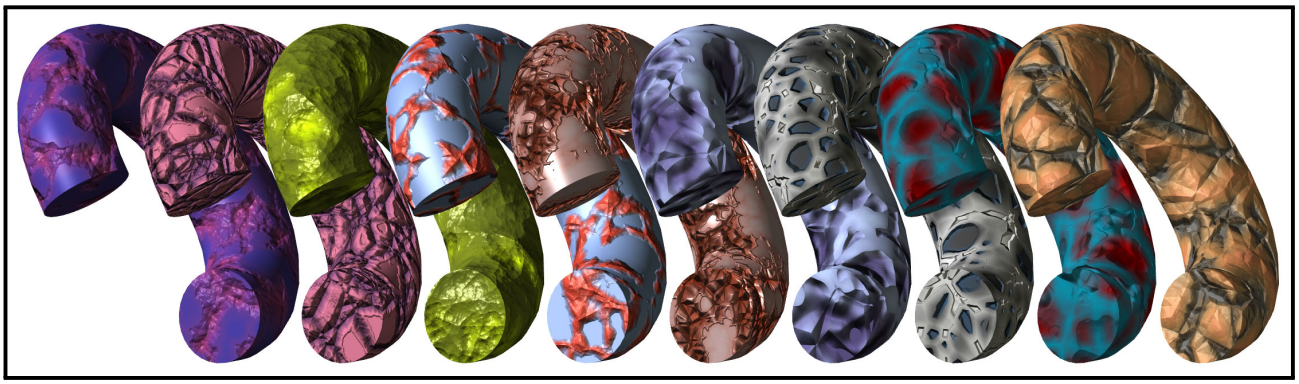


3.3. Steven Worley

Další skok přichází v roce 1996, kdy Steven Worley na konferenci představuje tvorbu procedurální textury.



Zde jsou ukázky z jeho článku *A cellular texture basis function*.

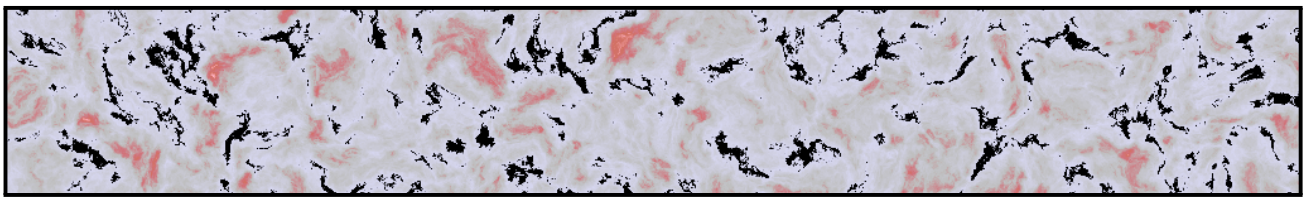


Poslední cifra, *1*, je z `examples/texturegranite.rs` z knihovny `noise` v0.6.0 na `crates.io`, konkrétně soubor `texture_granite_planar.png`. Získáváme malbu skoro jako od Jacksona Pollocka.

Po instalaci:

```
$ curl --proto '=https' --tlsv1.2 -sSf https://sh.rustup.rs | sh
$ echo "export PATH=$HOME/.cargo/bin:$PATH" >>~/.bashrc
$ source $HOME/.cargo/env
$ git clone https://github.com/Razaekel/noise-rs.git
$ cd noise-rs
$ cargo build
```

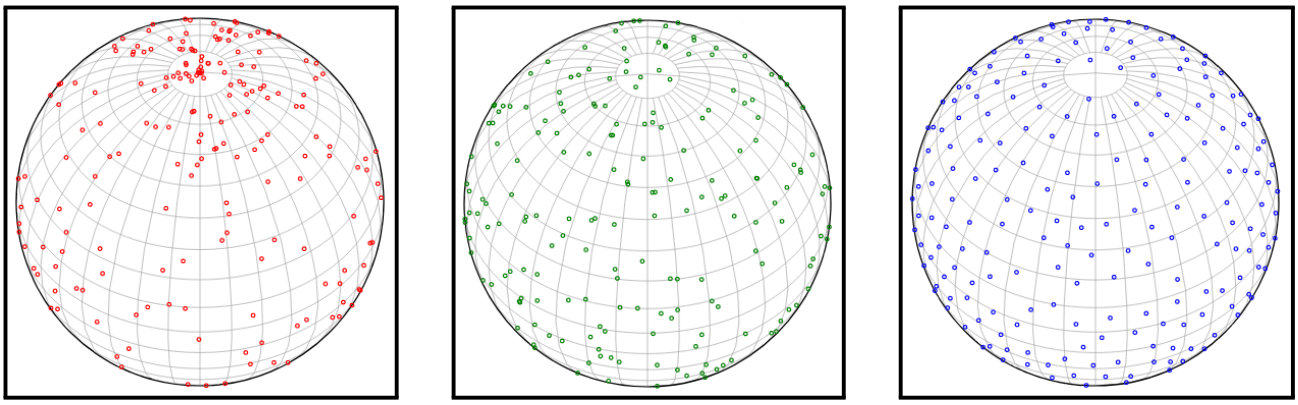
jsem spouštěl `cargo run --example texturegranite`.



Vážnějším zájemcům doporučuji knihu Patricio Gonzales Vivo a Jen Lowe: *The Book of Shaders* se zvýrazněnými ukázkami. Kniha vzniká od roku 2015 a je stále v přípravě a ve vývoji v OpenGL ES. Vážným zájemcům pak doporučuji knihy Ebert, Musgrave, Peachey, Perlin a Worley: *Texturing and Modeling: A Procedural Approach,* 3. vyd. z roku 2002 a z roku 2017 Tanya X. Short a Tarn Adams: *Procedural Generation in Game Design.*

4. Koule na místo vykřičníku

Rust má v ukázkách příklad vzniku textury pasovanou na kouli. Je to předchozí kód, další vygenerovaný soubor `texture_granite_sphere.png` ve složce `example_images`. Rovnovnoměrné rozdělení bodů na kouli je známý a vyřešený problém, viz MathWorld. Jason Davies srovnává intuitivní řešení ve sférické soustavě souřadnic (vlevo), řešení s korekcí (uprostřed) a aplikovaný Mitchellův algoritmus výběru nejlepšího kandidáta (koule vpravo). Charakteristikami se dostáváme na hranici modrého šumu.



5. ČStS aneb Vzorky pomocí procedurálních textur

Situace se komplikuje, pokud zvolíme obecný 3D objekt.

Existuje řada programů, které umí pracovat s texturami. Mezi nejvýraznější svobodné programy patří Blender (zkr. BS). Ten je mezi námi už od roku 1988, jen o 10 let mladší než TEX (1978) a o 5 let starší než R (1993). Některé postupy jsou vlastní, některé inspirované jinými programy na 3D grafiku. Jednou z inspirací byl program Filter Forge, kde se užívá jazyk Lua.

Za pozornost dávám knihu Richarda Egdahla *Texture Magic: Procedural Textures for Blender Cycles,* kterou považuji za představitele Blenderu do verze 2.79b. Důležité je, že Blender je specialista na 3D a položení textur na libovolné 3D objekty je přirozený krok grafiků.

Blender má nadstavbu Sverchok (zkr. SV, rusky сверчок) na parametricky definované 3D objekty. Příchod nadstavby Animation Nodes (zkr. AN) znamená mezník u animování s následnou možností úpravy textů.

Blender udělal obří skok od verze 2.80 s tahem k nástroji Everything Nodes, kdy přes grafické rozhraní (angl. shader nodes) by měly být dostupné téměř všechny nástroje Blenderu, speciálně systém částic.

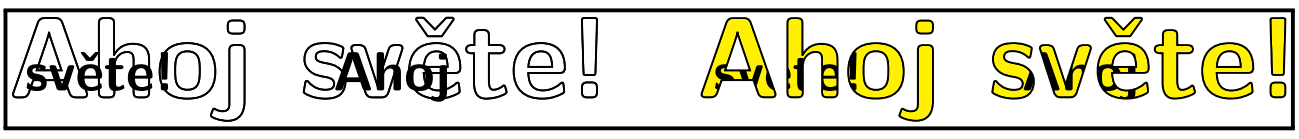
Lze doporučit YouTube kanál LiveNoding od Jimma Gunawaneho.

Znaky v ČStS jsou vytvořené v Blenderu za pomocí procedurálních textur a vyrenderovány jako rastrové obrázky: *Č* přes White Noise Texture, *S* přes Noise Texture, *t* přes Musgrave Texture a *S* pomocí Voronoi Texture.

6. Kompletace novoročenky místo Závěru

Pro TEXisty bude zajímavá první část. Přes Ti*k*Z vkládám pozadí dovnitř znaků. Zde je minimální ukázka „Ahoj světe!“ bez a s vyříznutím. Dávám na sebe pozadí velkých slov, malá slova a obrys velkých znaků.

```
\documentclass{article}
\pagestyle{empty}
\usepackage{tikz}
\begin{document}
\def\ukaz#1#2#3#4{%
\begin{tikzpicture}[text height=2ex, text depth=1ex]
\node{\Huge\bfseries\sffamily
  \pgfsetstrokecolor{black}\pgfsetfillcolor{yellow}%
  \pdfextension literal {#1 Tr}%
  \makebox[0pt][l]{#2}%
  \makebox[0pt][l]{\large
    \pdfextension literal {#3 Tr}%
    \tikz[trim left, baseline=0mm]
      \tikz{\node[xshift=7mm, yshift=1.5mm, text height=2ex, text
     depth=1ex]{\color{black}#4};
      };% end of \tikz
  }% end of \makebox
  \pdfextension literal {1 Tr 0.4 w}%
  \makebox[0pt][l]{#2}%
  };% end of \node
\end{tikzpicture}%
}% end of \ukaz
\ukaz{1}{Ahoj}{0}{světe!}\kern2cm \ukaz{1}{světe!}{0}{Ahoj}\kern28mm
\ukaz{6}{Ahoj}{0}{světe!}\kern2cm \ukaz{6}{světe!}{0}{Ahoj}
\end{document}
```



[...] but with Perlin noise I may pick numbers like this: 2, 3, 4, 3, 4, 5, 6, 5, 4, 5, 6, 7 [...]

Daniel Shiffman @ Perlin Noise and Flow Fields
www.youtube.com/watch?v=sor1nwNIP9A

VYHLÁŠENÍ SOUTĚŽE: DATA ANALYSIS COMPETITION 2021

ANNOUNCEMENT: DATA ANALYSIS COMPETITION 2021

Zdeněk Hlávka

E-mail: dac.iasc@email.cz

The International Association for Statistical Computing (IASC) announces the **Data Analysis Competition 2021.**

IASC is an Association of the International Statistical Institute (ISI) whose objectives are to promote the theory, methods, and practice of statistical computing and to foster interest and knowledge in effective and efficient statistical computing through international contacts among researchers and professionals in statistics, computer science, and related areas at universities, organizations, institutions, governments, and the general public in different countries of the world, to convert data into information and knowledge.

For the 2021 Data Competition, winners will be invited to present their work at the **Data Science, Statistics, and Visualisation Conference (DSSV-2021)** to be held at **Erasmus University, Rotterdam (July 7–9, 2021)** or at some other event organized by IASC (see `https://iasc-isi.org/events-all/`), depending on the current COVID-19 situation. The winners will also be invited to submit a manuscript for possible publication (following peer review) to IASC's **Journal of Data Science, Statistics, and Visualisation** (see `https://jdssv.org`). In addition, IASC will sponsor participation at the virtual **63rd ISI World Statistics Congress (July 12–16, 2021)** and **one year IASC membership** to all authors of submissions selected by the Committee on Data Analysis Competition; see `https://iasc-isi.org/become-a-member/` for an overview of IASC membership benefits.

The theme of the 2021 competition is around the **analysis of quality-of-life related data** and the submission should clearly describe the significance of your findings either for individuals or for the society. The primary data set may come from one or more databases but connecting information from different databases may help to obtain interesting and original conclusions.

There are currently many sources related to the spread of COVID-19, allowing to investigate topics like social inequalities in post-COVID-19 world or to compare effects of governments' actions, but we encourage you to investigate also any other quality-of-life related issues (health, education, envi-

ronment, social inequalities, etc.) In your analysis, you may concentrate on individuals, a single region (e.g., your own country), a continent or even the entire world.

Your entry must be submitted as a poster in PDF format and you must clearly specify the source of your data, i.e., by listing the relevant URLs and the steps required to obtain the data.

The competition is open to everyone who is interested in presenting their poster at **WSC 2021.** You are allowed to work individually or in a small group of up to five participants on your poster. Posters will be judged according to these criteria:

1. Appropriateness of analysis.
2. Novelty of approaches used in the analysis.
3. Clarity of objectives, approaches, data management, displays, and results.
4. Significance of findings.
5. Generalizability of approaches to data sets in other areas.
6. Overall quality of poster.

Not all posters are expected to meet all criteria to the same degree. Your poster may contain links to databases, computer programs, animations, and other external sources and it may (but need not) be accompanied by a short description (maximum five pages). All materials must be submitted in PDF format.

Updates and results of previous editions of the Data Analysis Competition may be found at `https://iasc-isi.org/data-analysis-competition/`.

Deadline for submission
Final submissions (PDF) are due on 30th April 2021.
Submit to: `dac.iasc@email.cz`.

Further information
For all inquiries contact:
Associate Professor Zdeněk Hlávka: `dac.iasc@email.cz`.

Zdeněk Hlávka	Jürgen Symanzik
Chairman of the Organizing Committee	President of the IASC

NEDÁVNO VYDANÉ KNIHY
RECENTLY PUBLISHED BOOKS

Redakce časopisu

Nakladatelství VŠE Oeconomica začátkem září 2020 vydalo v elektronické podobě dvoudílná skripta dostupná zdarma **R snadno a rychle** od česko-slovenského autorského tandemu Jakub Danko a Karel Šafr. Naprostý začátečník se v nich může na pracovní úrovni seznámit se základy jazyka R, vizualizací dat a programováním.

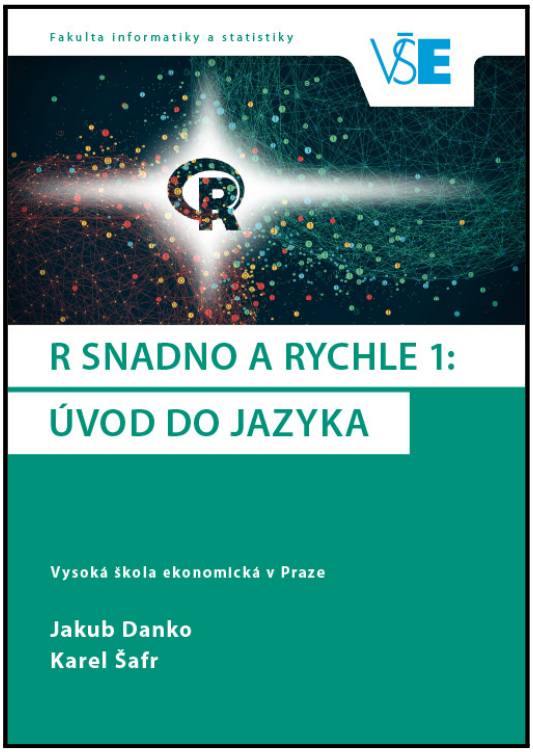
První učební text je věnován základům programování v jazyku R se zaměřením na studenty kvantitativních oborů nejen Fakulty informatiky a statistiky. Publikace využitelná též pro všechny zájemce o analýzu dat v softwaru R. Učební text zahrnuje základy samotného programovacího jazyka, zpracování a přípravu dat pro statistické zpracování, aplikaci základních i pokročilých statistických metod v R a taktéž prezentaci a vizualizaci výstupů.

V druhé publikaci autoři vysvětlují problematiku od instalace programu až po využití pokročilých nástrojů pro analýzu dat. Kniha je určena všem zájemcům o program od začátečníků po středně pokročilé uživatele.

Publikace jsou ke stažení a více informací hledejte na webu nakladatelství Oeconomica:

`https://oeconomica.vse.cz/publikace/r-snadno-a-rychle-1`

`https://oeconomica.vse.cz/publikace/r-snadno-a-rychle-2`



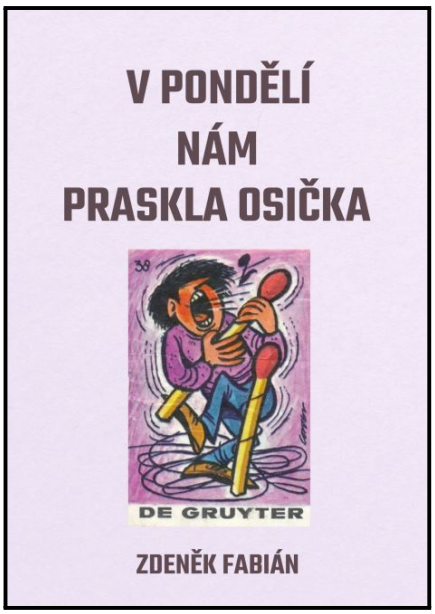
~ ~ ~

Dobrá zpráva pro příznivce tvorby Zdeňka Fabiána, mj. autora hitu **Járo,** viz `statspol.cz/konference/robust/robust-pisen-jaro/`. V nedávné době vyšly dvě jeho nové e-knihy. Knihy jsou nabízeny za Baťovské ceny v četných internetových knihkupectvích. Připomínáme také, že k dostání, tentokrát jak v tištěné podobě (Luxor), tak jako e-kniha (např. Alza) je i Zdeňkova knížka Úlety Matěje Broučka, vydaná v roce 2018.

V pondělí nám praskla osička. Nečekané zrození písničkáře na vedlejší úvazek, humorné písničky z doby normalizace i písničky z doby dnešní, s doprovodnými povídkami a pohádkou vyprávěnou jazykem Rudého práva.

Zápalky nepatří do rukou zvířat. Humorné vyprávění o autorově dětské vášni sbírání krabiček od sirek, kterými jsou dále ilustrovány činy a názory zvířátek z doby normalizace i z té dnešní, které se provizorně říká porevoluční, a která jistě časem dostane trefnější jméno.

Úlety Matěje Broučka. Autor, skryt za pseudonymem RNDr. Matěj Brouček, listuje, řečeno s Šimkem a Grossmannem, ve svém brožovaném životě. Podobně jako slavný jmenovec, i on je cestovatelem. Autor jej vysílá do nepříliš exotických destinací, které jen trochu zcestovatelý čtenář jistě zná, a to ne proto, aby objevoval nové demografické, geologické, geografické, etnografické, pornografické či kulturní výdobytky, ale aby se porval s drobnými lapáliemi, které mohou číhat na služebních cestách, vědeckých konferencích, na dovolené nebo třeba na vojenském cvičení či v lázních. A taky že číhaly. Zatímco příhody Matějova předskokana popisované panem Čechem jsou od A do Z vyfabulované, procento pravdivosti příhod v této knížce je $100 - \varepsilon$, kde ε je číslo menší než malé.



Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je sázen v programu TEX, ve formátu LuaHBTEX s písmy balíku CSfonts.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

Informace pro autory jsou na stránkách společnosti, http://www.statspol.cz/.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.