INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 37, číslo 2, červen 2026

Obsah

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.

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

ZPRÁVA O ČINNOSTI ČESKÉ STATISTICKÉ SPOLEČNOSTI (ČSTS) V ROCE 2025
CZECH STATISTICAL SOCIETY IN 2025

Martina Litschmannová

E-mail: martina.litschmannova@vsb.cz

Členská základna

K 31. 12. 2025 měla Česká statistická společnost 215 členů. V průběhu roku 2025 přibylo 11 nových členů, naopak 7 členům bylo členství ukončeno. Celkový počet členů se tak v průběhu roku zvýšil o 4.

Během roku zemřel doc. RNDr. Zdeněk Karpíšek, CSc.

K 6. 3. 2026 uhradilo členské příspěvky za rok 2025 152 členů (71 % členské základny).

Akce pořádané či spolupořádané ČStS

1. **Členská schůze ČStS** se uskutečnila dne **31. ledna 2025** v prostorách Českého statistického úřadu v Praze. Zúčastnilo se jí 25 členů společnosti. Byla přednesena zpráva o činnosti za rok 2024 a zpráva o hospodaření za rok 2024. Součástí programu byla diskuse o činnosti společnosti a plánovaných aktivitách.
2. Konference **STAKAN (Statističtí kantoři)** se konala ve dnech **30. května – 1. června 2025 v Pavlově**. Akce byla zaměřena na sdílení zkušeností z výuky statistiky a aplikací statistických metod. Zúčastnilo se jí 34 účastníků, z toho 4 ze Slovenska. Celkem bylo předneseno 14 odborných příspěvků.
3. Ve dnech **1. – 3. července 2025** v Žilině proběhla konference **OSSConf 2025**, Otvorený softvér vo vzdelávani, výskume a IT riešeniách. Akce šíří povědomí o open-source software, datech i hardware. Konference je rozdělená do sekcí, ve kterých zaznělo několik příspěvků na analýzu dat v R a Pythonu a také několik příspěvků z oblasti optimalizace a náročných výpočtů.
4. V roce 2025 se zástupci společnosti aktivně účastnili mezinárodních aktivit. Členové ČStS se zúčastnili konference **Austrian Statistical Days 2025**, která se konala ve dnech **2. – 4. září 2025 v Linci (Rakousko)**. ČStS reprezentoval místopředseda společnosti doc. Mgr. Ondřej Vencálek, Ph.D., spolu s dalšími členy společnosti.

5. ČStS se dále spolupodílela na organizaci semináře **ENERGY DAYS 2025**, který se konal ve dnech **6. – 8. listopadu 2025 v Praze**. Seminář byl zaměřen na problematiku modelování a analýzy energetických systémů.
6. **Miku(k)lášský statistický den** se uskutečnil dne **4. prosince 2025** v Praze. Na akci bylo předneseno 5 odborných příspěvků a zúčastnilo se jí 40 účastníků z akademické i aplikační sféry.

Spolupráce s Českým statistickým úřadem

Český statistický úřad je dlouhodobým partnerem České statistické společnosti a podporuje její činnost zejména poskytnutím sídla společnosti, organizační podporou při pořádání akcí a zajištěním tisku Informačního bulletinu ČStS. Prostory ČSÚ jsou rovněž pravidelně využívány pro konání členských schůzí a dalších akcí společnosti.

Členství v mezinárodních organizacích

Česká statistická společnost je členem federace evropských národních statistických společností **FENStatS** (The Federation of European National Statistical Societies).

Členský příspěvek (200 eur) do organizace FENStatS byl opět uhrazen s podporou Akademie věd ČR prostřednictvím Rady vědeckých společností.

ČStS při jednáních FENStatS zastupuje doc. Ing. Tomáš Hlavsa, Ph.D.

Členství v Radě vědeckých společností ČR

Česká statistická společnost je členem Rady vědeckých společností ČR (RVS) při Akademii věd ČR.

Další činnost

- V roce 2025 bylo vydáno **5 čísel Informačního bulletinu České statistické společnosti**.
- Byla průběžně aktualizována **webová stránka společnosti**, na níž byly zveřejňovány informace o akcích společnosti, konferencích a dalších aktivitách statistické komunity. http://www.statspol.cz/
- Byla průběžně aktualizována databáze členů společnosti.

Akce chystané v roce 2026

- **ROBUST**, 18. – 23. ledna 2026, Srní.
- **Členská schůze ČStS**, 6. března 2026, Praha.
- **SIS-FENStatS 2026**, 22. – 25. června 2026, Řím, Itálie – sekce Statistical Literacy SPE (Portugalsko) + SSS (Slovinsko) + ČStS (Česko) – doc. Ing. Tomáš Hlavsa, Ph.D.
- **OSSConf 2026**, Otvorený softvér vo vzdelávani, výskume a IT riešeniách, 1. – 3. července 2026, Žilina, Slovensko – ČStS je spoluorganizátorem konference. https://ossconf.fri.uniza.sk/
- **AMSE 2026**, 25. – 29. srpna 2026, Martin, Slovensko – ČStS je partnerem konference. https://www.amse-conference.eu/
- **Olomoucké statistické dny 2026**, 10. – 11. září 2026, Olomouc. https://www.statspol.cz/STATDNY2026/
- **Miku(k)lášský den 2026,** ve čtvrtek 3. 12. 2026.
- Další aktivity podporující spolupráci statistické komunity.

V Ostravě dne 6. 3. 2026

Martina Litschmannová
předsedkyně České statistické společnosti

METHODS FOR DISSIMILARITY ASSESSMENT AND CLASSIFICATION OF REALISATIONS OF RANDOM SETS

METODY ROZLIŠOVÁNÍ A KLASIFIKACE REALIZACÍ NÁHODNÝCH MNOŽIN

Bogdan Radović

Address: Department of Mathematics, Faculty of Eletrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6

E-mail: radovbog@fel.cvut.cz

Abstract: The paper provides an overview of newly developed methods for assessing dissimilarity and performing classification of random sets realisations. One of the methods for assessing dissimilarity focuses on the shape of individual components, specifically on the perimeter-to-area ratios and boundary curvature, and employs a permutation test based on $\mathcal{N}$-distances. The second method also examines boundary shape, but additionally incorporates selected topological features of the random sets it aims to compare. In this case, inference is carried out using a permutation version of global envelope test applied to the difference of statistical depths between two samples of random sets. Further, a summary of recently proposed framework for classifying families of sets that arise as realisations of random set models, which employs both supervised and unsupervised approaches, is presented. Each method is briefly outlined, their strengths and limitations discussed, and the conclusions supported by results from a simulation study.

Keywords: Classification, DD-plot, Global envelope test, $\mathcal{N}$-distance, Random set, Realisation, Similarity, Statistical depth, Topological data analysis.

Abstrakt: Článek poskytuje přehled nově vytvořených metod pro rozlišování a klasifikaci realizací náhodných množin. Jedna z metod rozlišování se zaměřuje na tvar jednotlivých komponent, konkrétně na poměr obvodu k ploše a zakřivení hranice, a využívá permutační test založený na $\mathcal{N}$-vzdálenostech. Druhá metoda rovněž zkoumá tvar hranice, avšak navíc zahrnuje vybrané topologické charakteristiky porovnávaných náhodných množin. V tomto případě je inference prováděna pomocí permutační verze globálního obálkového testu aplikovaného na rozdíl statistických hloubek mezi dvěma výběry náhodných množin. Dále je představena metoda pro klasifikaci realizací náhodných množin, který využívá jak přístupy tzv. učení s učitelem, tak bez něj. Každá metoda je stručně popsána, jsou diskutovány její silné stránky i omezení a závěry jsou ilustrovány na výsledcích simulační studie.

Klíčová slova: Globální obálkový test, DD-plot, klasifikace, $\mathcal{N}$-vzdálenost, náhodná množina, podobnost, realizace, statistická hloubka, topologická analýza dat.

1. Introduction

In recent years, modeling and statistical analysis of random sets have become increasingly popular, as they have been proven to be an effective tool for studying the geometry of random objects from a statistical perspective. Advances in computing technology further enabled extensive simulation studies on which these analyses are often based. From a theoretical perspective, they have been investigated, for example, in [2], [19] and many others.

A realisation of a random set may be viewed as a geometric mapping of an object whose shape arises from a random process. Examples of such objects include vegetation configuration in an ecosystem, the presence of specific minerals or gaps within materials, clusters of cells in biological tissue, etc. Statistical investigations of processes generating these objects can yield valuable insights into their behavior, the knowledge of which can yield numerous practical benefits. Different authors studied their application in medicine [15], ecology [21], material science [24] or biology [25].

Usually, for a realisation of a random set, we attempt to find a suitable model. However, knowledge of a specific model is sometimes not necessary, since our goal may be to compare two or more realisations in order to decide whether they are similar (i.e., they originate from the same process) or not.

Classical tools for describing random sets, such as the covariance function, the contact distribution function [2], functions based on morphological operations (dilation, erosion, opening, and closing of a set), the granulometric function [27], and others, may not be sufficient to distinguish between two realisations due to the fact that for a single realisation we obtain only one estimate of the given function, which makes it impossible to perform a formal statistical test of equality of the probability distributions. To tackle this problem, numerous methods for distinguishing between two realisations of random sets have been developed over the past decade, which focused on deriving a sample of functions from each realisation, followed by testing the equality of their distributions (for more details, see [3], [10], [11], [12]). However, these functions differ across methods, which means that some methods may consider two realisations similar, while others may distinguish between them. This is in accordance with the practical setting, as different situations involve different aspects of interest (e.g., when studying tissues, we are more interested in the shape of cells, as it is known that cancerous cells are char-

acterised by their irregular shape and size, while, when studying forests, the shape of tree crowns is regular, but bigger trees affect the size of neighbouring crowns, so we may be more interested in their position).

Continuing in the same fashion, one of the methods described in this article and published originally in [13] develops another approach for distinguishing between realisations of random sets: a sample of functions is derived from a realisation of a random set, which are then used to test the similarity. Furthermore, a different approach for distinguishing between random sets, originally published in [14], is offered based on non-parametric analysis of the distribution of random sets via statistical depths. Finally, in a more general setting, again coming from practical situations, the former method was used in [26] to classify realisations of random sets based on their similarity. This article presents an overview of the abovementioned, already published methods.

The article is organised as follows. Section 2 provides an overview of the relevant definitions and established theoretical results on statistical depths, nonparametric statistical testing using DD-plots (depth-depth plots), curvatures of planar curves, and statistical testing using $\mathcal{N}$-distance theory. Section 3 presents and compares recently developed methods for assessing dissimilarity of two realisations of random sets. Finally, section 4 gives an overview of the recently developed framework for classifying realisation of random sets based on their similarity.

2. Theoretical background

Let (Ω, Σ, P) be a probability space, $\mathcal{F}$ be the family of closed sets in $\mathbb{R}^d$ and $\mathfrak{F} = \sigma\{\mathcal{F}^K : K \text{ is a compact set in } \mathbb{R}^d\}$, where $\mathcal{F}^K = \{F \in \mathcal{F} : F \cap K \neq \emptyset\}$. Then a random closed set $\mathbf{X}$ is a measurable mapping $\mathbf{X} : (\Omega, \Sigma) \mapsto (\mathcal{F}, \mathfrak{F})$.

A set $F \in \mathcal{F}$ belongs to the support of random set $\mathbf{X}$ if $\mathbf{X}$ belongs to any open neighbourhood of F in the Fell topology [7] with positive probability.

Let $\mathbf{X}$ be a random closed set. Its depth function $D(\cdot, \mathbf{X})$ is a function of a closed set F taking values in $[0, 1]$. If F does not belong to the support of $\mathbf{X}$ then $D(F, \mathbf{X}) = 0$.

2.1. Statistical depth for random sets

In this section, several depths for general random sets are introduced together with a way in which they can be estimated given a sample of random sets.

2.1.1. Depths based on representations of closed sets by functions There are several ways to represent a closed set F as a function, the most obvious being its characteristic or indicator function

$$\mathbf{1}_F(x) = \begin{cases} 1, & x \in F, \\ 0, & x \notin F. \end{cases}$$

When using functional band depth [17] together with the indicator function representation of the set, a nice formulation of set depth in terms of random set notation is obtained as

$$D_{band}(F, \mathbf{X}) = P\left(\bigcap_{i=1}^{m} \mathbf{X}_i \subseteq F \subseteq \bigcup_{i=1}^{m} \mathbf{X}_i\right). \tag{1}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_m$ are i.i.d. random sets with the same distribution as $\mathbf{X}$ and $m \in \mathbb{N}$. This depth was introduced in [29].

Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ be a sample, i.e., a realisation of m i.i.d. random sets $\mathbf{X}_1, \ldots, \mathbf{X}_m$ with the same distribution as $\mathbf{X}$. For chosen $n \in \mathbb{N},\ n \leq m$, the estimate of the depth (1) is given by following U-statistics

$$D_{band,n}(F, \mathcal{X}) = \frac{1}{\binom{m}{n}} \sum_{1 \leqslant k_1 < \cdots < k_n \leqslant m} \mathbf{1}\left(\bigcap_{i=1}^{n} X_{k_i} \subseteq F \subseteq \bigcup_{i=1}^{n} X_{k_i}\right). \tag{2}$$

For large m and n, $\binom{m}{n}$ can become extremely large, making calculations of (2) too slow. In these cases, we can choose $s < \binom{m}{n}$, bootstrap s samples $X_1^{(j)}, \ldots, X_n^{(j)},\ j = 1, \ldots, s$ of length n from the sample $\mathcal{X}$ and obtain

$$D_{band,n,s}(F, \mathcal{X}) = \frac{1}{s} \sum_{j=1}^{s} \mathbf{1}\left(\bigcap_{1=1}^{n} X_i^{(j)} \subseteq F \subseteq \bigcup_{i=1}^{n} X_i^{(j)}\right).$$

Another representation of a closed set, commonly used in image analysis, is by the distance function or the signed distance function. Let $W \subset \mathbb{R}^d$ be an observation window. The signed distance function of a set $F \subset W$ is defined by $f_F : W \to \mathbb{R}$ as

$$f_F(x) = \begin{cases} d(x, F), & x \notin F, \\ -d(x, W \setminus F) & x \in F, \end{cases}$$

where $d(x, F)$ stands for an arbitrary distance from a point x to the set F. An arbitrary functional depth can be applied to compute its depth values. These depths will be denoted by D_{sign}.

2.1.2. Simplicial depth A novel approach for defining depths for random sets is introduced, which appears particularly natural in the models involving Minkowski addition (see, e.g., [18]).

Let $\mathbf{X}$ be a random closed set, $\mathbf{X}_1, \ldots, \mathbf{X}_m$ a sequence of i.i.d. random sets with the same distribution as $\mathbf{X}$, sample $\mathcal{X} = \{X_1, \ldots, X_m\}$ its realisation, and F an arbitrary closed set.

Suppose that $F_1, \ldots, F_m$ are arbitrary closed sets. Given that $conv(F_1, \ldots, F_m) = \{p_1F_1 + \ldots + p_mF_m : p_1, \ldots, p_m \geq 0, p_1 + \ldots + p_m = 1\}$, where $+$ stands for Minkowski addition of the sets, simplicial depth can be defined as

$$D_{sim}(F, \mathbf{X}) = P(\{\omega \in \Omega : \exists L, U \ \in conv(\mathbf{X}_1(\omega), \ldots, \mathbf{X}_m(\omega)) \\ \text{such that} \ \ L \subseteq F \subseteq U \ \}).$$

To estimate the depth $D_{sim}(X_i, \mathbf{X})$, $i = 1, \ldots, m$, a bootstrap approach can be used. First, choose $n \leq m$ and $s \in \mathbb{N}$. Then s times take a random sub-sample of $\mathcal{X}$ of length n. In this way s sub-samples $X_1'^{(j)}, \ldots, X_n'^{(j)}$, $j = 1, \ldots, s$ are obtained. Then, the estimate of $D_{sim}(X_i, \mathbf{X})$ is

$$D_{sim}(X_i, \mathcal{X}) := \frac{1}{s} \sum_{j=1}^{s} \mathbf{1}(\exists L, U \in conv(X_1'^{(j)}, \ldots, X_n'^{(j)}) : \ L \subseteq X_i \subseteq U).$$

Since $conv(X_1'^{(j)}, \ldots, X_n'^{(j)})$ consists of an infinite number of points, in practice the above condition cannot be verified for all $L, U \in conv(X_1'^{(j)}, \ldots, X_n'^{(j)})$. However, it is enough to choose $N \in \mathbb{N}$ and verify only for $L, U \in \{p_1X_1'^{(j)} + \ldots + p_nX_n'^{(j)} : p_l = \frac{n_l}{N}, \ n_l \in \mathbb{N}, \ \sum_{l=1}^{n} n_l = N\}$.

2.2. Testing equality in distribution of two samples of random sets using statistical depths

The concept of DD-plot was introduced in [5] to compare multivariate distributions of two samples using statistical depth measures.

Let $\mathcal{X} = \{X_1, \ldots, X_{m_1}\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_{m_2}\}$ be the two samples of random sets that we aim to compare. Given an arbitrary depth function D, the **DD-plot** of the pooled sample is defined as a collection of ordered pairs $DD(\mathcal{X}, \mathcal{Y}) = \{(D(F, \mathcal{X}), D(F, \mathcal{Y})), F \in \mathcal{X} \cup \mathcal{Y}\}$. Hence, the DD-plot consists of $m_1 + m_2$ ordered pairs of numbers between 0 and 1.

Since depths are designed to reflect distributional features of samples, DD-plots constructed from equally distributed samples are expected to be

similar to the identity. Thus, the scatter plot generated by the DD-plot should concentrate around the diagonal line $(0,0)-(1,1)$.

In contrast, when the underlying distributions differ, the DD-plot typically appears to be more dispersed, and it may exhibit irregular patterns. The nature of its deviation from the diagonal can provide insights into the nature of distributional differences (see [5]).

The authors of [6] further propose some statistics that quantify the degree of concentration of the DD-plot around the $(0,0)-(1,1)$ diagonal. However, the test has low power when comparing samples of a random particle model (for more details see [14]) and fails to recognise differences between the distributions of the samples. Thus, it makes sense to develop a new statistical test that would overcome these obstacles.

2.3. Characteristics of shape of a random planar set

Definition 2.1. *Consider a smooth 2D curve $\mathcal{C}$ parametrised by a parameter $\varphi \in [0,\phi] \subset \mathbb{R}$, i.e., $\mathcal{C}(\varphi) = (x(\varphi), y(\varphi))$. The curvature κ of $\mathcal{C}$ in a point $\mathcal{C}(\varphi)$ is defined as*

$$\kappa(\mathcal{C}(\varphi)) = \frac{x'(\varphi)y''(\varphi) - x''(\varphi)y'(\varphi)}{(x'^2(\varphi) + y'^2(\varphi))^{3/2}},$$

if the right hand side is well defined.

Let us assume that the curve $\mathcal{C}$ is continuous, closed (i.e. $\mathcal{C}(0) = \mathcal{C}(\phi)$) and it does not intersect itself (i.e. $\mathcal{C}(\varphi_1) = \mathcal{C}(\varphi_2) \Rightarrow \varphi_1 = \varphi_2$). Consider a connected planar set X whose boundary is given by the curve $\mathcal{C}$. It can be shown [1] that for the curvature $\kappa(z)$ evaluated in a given point $z \in \mathcal{C}$ and for a disc $b(z,r)$ with the center in z and a radius r small enough, it holds that

$$\kappa(z) \approx \frac{3A^*_{b(z,r)}}{r^3} - \frac{3\pi}{2r} = \frac{3\pi}{r}\left(\frac{A^*_{b(z,r)}}{A_{b(z,r)}} - \frac{1}{2}\right), \tag{3}$$

where $A_{b(z,r)}$ is the area of the disc $b(z,r)$ and $A^*_{b(z,r)}$ is the area of $b(z,r) \cap X$.

Consider a connected random set $\mathbf{X}$, i.e. the random set whose realisations are connected. Denote $B_{\mathbf{X}}$ the boundary of $\mathbf{X}$ and $\kappa_{\mathbf{X}}(z)$ the (random) curvature at the point $z \in B_{\mathbf{X}}$. From (3), we can see that for a disc $b(z,r)$ with suitably chosen radius r, it holds that (up to a constant that can be neglected)

$$\kappa_{\mathbf{X}}(z) \propto \frac{A^*_{b(z,r),\mathbf{X}}}{A_{b(z,r)}} = O_{\mathbf{X},b(z,r)},$$

where $A_{b(z,r)}$ is the area of $b(z,r)$ and $A^*_{b(z,r),\mathbf{X}}$ is the area of $b(z,r) \cap \mathbf{X}$. Finally, we define the function

$$\widetilde{\kappa}_{\mathbf{X},r}(u) = |B_{\mathbf{X}}|^{-1} \int_{B_{\mathbf{X}}} \mathbf{1}(O_{\mathbf{X},b(z,r)} \leq u)\, \mathrm{d}z, \quad u \in [0,1],$$

which is an analogy of the distribution function of the curvature at points on the boundary, but it is evaluated for all boundary points, so it describes the distribution for strongly dependent values. The object of our interest is the function, analogous to density function, describing the distribution of the curvature along the boundary

$$t_{\mathbf{X},r}(u) = \widetilde{\kappa}'_{\mathbf{X},r}(u). \tag{4}$$

In the sequel, the function (4), which describes the curvature of the boundary of the set $\mathbf{X}$, is called the C-function, and it will be one of the characteristics used for the inference below.

The second characteristic of the random set $\mathbf{X}$ is the random variable describing the ratio of the perimeter and the area of $\mathbf{X}$. It is denoted as $R_{\mathbf{X}}$ and called the P/A-ratios in the sequel.

In practice, we observe realisations X of the random set $\mathbf{X}$ in the form of binary images, so we need to adjust the definitions of the characteristics defined above to the realisations consisting of black and white pixels. The pixels play the role of units in the sequel. The P/A-ratio is simply given by the number of boundary pixels divided by the number of all pixels of the component. For evaluating the C-function, fix a radius $r \in \mathbb{N}$, denote Pix the set of all pixels of the binary image, $z_1, \ldots, z_n$ all boundary pixels, and for each boundary pixel z_i, define

$$T(z_i) = \frac{\sharp\{p \in Pix : p \in b(z_i,r) \cap X\}}{\sharp\{p \in Pix : p \in b(z_i,r)\}}.$$

Then, the approximation of the function $t_{\mathbf{X},r}(u)$ from (4) is

$$t(u) = \frac{\sharp\{i \in \{1,\ldots,n\} : T(z_i) \in [u - 1/l, u)\}}{n} \quad \text{for } u = \frac{1}{l}, \frac{2}{l}, \ldots, 1, \tag{5}$$

where l is the number of pixels that form the disc $b(.,r)$.

2.4. Testing equality in distribution based on $\mathcal{N}$-distance of probability measures

In this paper, the procedure for testing equality in distribution of random variables and random functions comes from the theory of $\mathcal{N}$-distances, which

is briefly recalled in the following paragraphs. More details concerning this topic can be found in [16].

Let $\mathcal{S}$ be a nonempty set.

Definition 2.2. *A map $\mathcal{L} : \mathcal{S} \times \mathcal{S} \to \mathbb{C}$ is called* negative definite kernel *if for any $n \in \mathbb{N}$, arbitrary $c_1, ..., c_n \in \mathbb{C}$ such that $\sum_{i=1}^{n} c_i = 0$ and arbitrary $x_1, ..., x_n \in \mathcal{S}$ it holds*

$$\sum_{i}^{n} \sum_{j}^{n} \mathcal{L}(x_i, x_j) c_i \overline{c}_j \leq 0.$$

Definition 2.3. *The negative definite kernel $\mathcal{L}$ is called* strongly negative definite kernel *if for an arbitrary probability measure μ and an arbitrary $f : \mathcal{S} \to \mathbb{R}$ such that $\int_{\mathcal{S}} f(x)\,\mathrm{d}\mu(x) = 0$ and the double integral $\int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{L}(x, y) f(x) \cdot f(y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$ exists and is finite, the relation*

$$\int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{L}(x, y) f(x) f(y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = 0$$

implies that $f(x) = 0$ μ-a.e.

For a map $\mathcal{L} : \mathcal{S} \times \mathcal{S} \to \mathbb{C}$, denote $\mathcal{B}_{\mathcal{L}}$ the set of all measures μ such that $\int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{L}(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$ exists.

Theorem 2.1 (Klebanov, 2006)**.** *Let $\mathcal{L}(x, y) = \mathcal{L}(y, x)$. Then*

$$\begin{aligned} \mathcal{N}(\mu, \nu) = 2 \int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{L}(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y) - \int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{L}(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\ - \int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{L}(x, y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y) \geq 0 \end{aligned} \tag{6}$$

holds for all measures $\mu, \nu \in \mathcal{B}_{\mathcal{L}}$ with equality in the case $\mu = \nu$ only, if and only if $\mathcal{L}$ is a strongly negative definite kernel.

In the following text, the term $\mathcal{N}(\mu, \nu)$ from (6) is called the $\mathcal{N}$-distance of the measures μ and ν. Suppose that we have observations $x_1, \ldots, x_{m_1}$ from the distribution μ and $y_1, \ldots, y_{m_2}$ from the distribution ν, then the $\mathcal{N}$-distance of the measures μ and ν is estimated as

$$\widehat{\mathcal{N}}_1 = \frac{2}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathcal{L}(x_i, y_j) - \frac{1}{m_1^2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} \mathcal{L}(x_i, x_j) - \frac{1}{m_2^2} \sum_{i=1}^{m_2} \sum_{j=1}^{m_2} \mathcal{L}(y_i, y_j). \tag{7}$$

Many examples of strongly negative definite kernels $\mathcal{L}$ are introduced in [16] for the case that the observations are real numbers, i.e., realisations of real random variables. One of the examples, used in our paper, is the Euclidean distance

$$\mathcal{L}(x, y) = |x - y|. \tag{8}$$

When the measures μ and ν correspond to distributions of random functions, then we use the kernel introduced in [12], constructed especially for such functions as follows. Consider two functions f_1 and f_2 evaluated in discrete arguments $u_1, \ldots, u_n$, $n \in \mathbb{N}$. Then the strongly negative definite kernel is

$$\mathcal{L}(f_1, f_2) = \sum_{m=1}^{D} \sum_{\{k_1,\ldots,k_m\} \subseteq \{1,\ldots,n\}} \left(\sum_{l=1}^{m} \left(f_1(u_{k_l}) - f_2(u_{k_l}) \right)^2 \right)^{1/2}, \tag{9}$$

where D is a chosen constant specifying the depth of dependence, see [12] for more details.

Then, we use the Monte Carlo permutation test, i.e., we make S permutations of all observed values $x_1, \ldots, x_{m_1}, y_1, \ldots y_{m_2}$, split each permutation into two groups of lengths m_1 and m_2, and, analogously to (7), we calculate $\widehat{\mathcal{N}}_i$ for the i-th permutation, $i = 2, \ldots, S + 1$. Then the p-value of the test is

$$p = \frac{\sharp\{i \in \{2, \ldots, S+1\} : \widehat{\mathcal{N}}_i \geq \widehat{\mathcal{N}}_1\} + 1}{S + 1}.$$

3. Methods for assessing similarity of random sets and their realisations

The primary aim of the methods introduced below is to decide whether two realisations X and Y of random sets $\mathbf{X}$ and $\mathbf{Y}$, respectively, can be regarded as similar, with similarity defined specifically for each method. The first procedure, based on depths, uses a permutation version of the global envelope test [23], where the test function is the difference between statistical depths. The second procedure, referred to as two-step method, is based on deriving, for each realisation X and Y, a collection of functions that describe specific features introduced above. The equality of the corresponding probability distributions of these functions is then tested using a test based on the $\mathcal{N}$-distance. Realisations are considered similar if the null hypothesis of distributional equality is not rejected, and dissimilar (that is, distinguishable) otherwise.

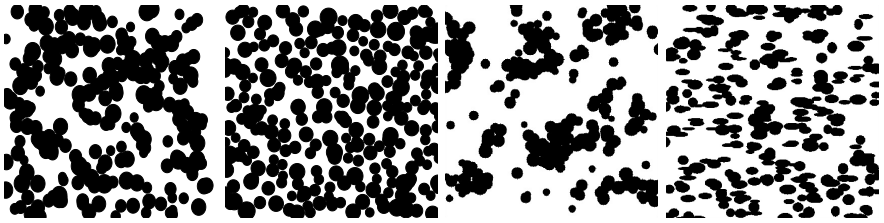


Figure 1: Example of realisation of the Boolean, the repulsive, the cluster and the ellipse model, respectively.

The described procedures are illustrated through a simulation study involving four models. The first model is the Boolean disc model, that is, a union of a random number of discs with random radii. The second model is the Boolean ellipse model, that is, a union of a random number of ellipses of random size. The third model is a cluster process, and the fourth is a repulsive process, both instances of Quermass-interaction process with suitably chosen parameters (for details on their simulation, see [20]; for details on the model parameters see [14]). Realisations of these models are shown in Figure 1.

In the study, for each model 200 realisations of size 400×400 pixels are considered. When we compare realisations coming from the same model, we divide the set of 200 realisations from that model into two sets of 100 realisations, forming 100 pairs. In this way, we obtain 100 p-values for each pair of models compared. Similarly, in the case where we compare different models, we consider 100 realisations of each model. It gives us 100 pairs of compared realisations, and corresponding p-values. p-values close to zero indicate rejection of the equality of the probability distributions of the corresponding test functions, meaning that the realisations are considered to be dissimilar. Therefore, we expect p-values to be concentrated near zero when comparing realisations from different processes, whereas for realisations from the same process, the p-values should be approximately uniformly distributed on the interval (0,1). The percentage of p-values smaller than 0.05 is presented tabularly for both methods.

3.1. Method based on statistical depths

For testing equality in the distribution of the two samples $\mathcal{X} = \{X_1, \ldots, X_{m_1}\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_{m_2}\}$ of random sets, an approach based on the permutation version of the global envelope test is proposed where the test function is the difference between depths. Note that the difference between depth is

proportional to the signed distance from the corresponding point in a DD-plot to a (0,0)–(1,1) line. In more detail, the test function is

$$T_1(k) = D(F_k, \mathcal{X}) - D(F_k, \mathcal{Y}), \quad k = 1, \ldots, m_1 + m_2,$$

where F_k is the k-th set in the joined sample $X_1, \ldots, X_{m_1}, Y_1, \ldots, Y_{m_2}$. The depth D should be chosen in a way that reflects the nature of the problem. Continuing the research from [4], the focus is on the shape of the components, namely the shape of the boundary, and topological features such as connected components and holes. The study in [14] shows that the depth based on the signed distance function (D_{sign}) performs the best when focusing on the shape of the boundary, while holes and connected components were best detected by the band (D_{band}) and simplicial (D_{sim}) depths.

For chosen number of permutations S, we permute the joined sample $X_1, \ldots, X_{m_1}, Y_1, \ldots, Y_{m_2}$, and obtain two samples $\mathcal{X}_i^*$ and $\mathcal{Y}_i^*$, the first sample consisting of the first m_1 sets in the permuted sample and the second consisting of the last m_2 sets in the permuted sample. We calculate $T_i(k) = D(F_k', \mathcal{X}_i^*) - D(F_k', \mathcal{Y}_i^*)$, $i = 2, \ldots, S+1$, where F_k' is the k-th set in the permuted sample. Under the assumption of exchangeability of the joined sample, the distribution of $T_i(k)$, $i = 2, \ldots, S+1$ should remain the same as the distribution of $T_1(k)$. In this way, we obtain $S+1$ objects $T_1(k), \ldots, T_{S+1}(k)$, $k = 1, \ldots, m_1 + m_2$.

For each $k = 1, \ldots, m_1 + m_2$, let $R_i^{\uparrow}(k)$ and $R_i^{\downarrow}(k)$ denote the ranks of the values $T^{(i)}(k)$ from the smallest value with rank 1 to the largest one with rank $S+1$ and from the largest value with rank 1 to the smallest one with rank $S+1$, respectively. For each $k = 1, \ldots, m_1 + m_2$, we define the k-wise ranks of $T_i(k)$ as $R_i(k) = \min\left(R_i^{\uparrow}(k), R_i^{\downarrow}(k)\right)$. The extreme ranks R_i are obtained by $R_i = \min_k R_i(k)$.

In order to avoid the possibility of the ties, we use the area measure refinement of the extreme rank R_i. It is constructed as follows.

Let $T_{[1]}(k) \leq T_{[2]}(r) \leq \cdots \leq T_{[S+1]}(k)$ denote the ordered set of values $T_i(k), i = 1, \ldots, S+1$. The continuous rank of $T_{[i]}(k)$ is

$$c_{[i]}(k) = j + \frac{T_{[i]}(k) - T_{[i-1]}(k)}{T_{[i+1]}(k) - T_{[i-1]}(k)}, \qquad \text{for } i = 2, \ldots, S,$$

and

$$c_{[1]}(k) = \exp\left(-\frac{T_{[2]}(k) - T_{[1]}(k)}{T_{[S+1]}(k) - T_{[1]}(k)}\right),$$

$$c_{[S+1]}(k) = S + 1 - \exp\left(-\frac{T_{[S+1]}(k) - T_{[S]}(k)}{T_{[S]}(k) - T_{[1]}(k)}\right).$$

Let $C_i(r) = S + 1 - c_i(k)$. Then the area rank measure a_i is defined as

$$a_i = \frac{1}{S+1}\left(R_i - \frac{1}{m_1+m_2}\sum_k \left(R_i - C_i(k)\right)\mathbf{1}\left(C_i(k) < R_i\right)\right).$$

Finally, the p-value of the test is the percentage of curves that have more extreme ranks than the observed one, that is,

$$p = \frac{1}{S+1}\left(\sum_{i=1}^{S+1}\mathbf{1}(a_i < a_1)\right).$$

3.2. Two-step method focusing on shape of connected components

Consider connected random sets $\mathbf{X}$ and $\mathbf{Y}$ with the C-functions $t_{\mathbf{X},r}$ and $t_{\mathbf{Y},r}$ and the P/A-ratios $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$, respectively. The similarity of random sets is defined so that two connected random sets $\mathbf{X}$ and $\mathbf{Y}$ are considered to be similar if the distributions of $\lim_{r\to 0} t_{\mathbf{X},r}$ and $\lim_{r\to 0} t_{\mathbf{Y},r}$ as well as the distributions of $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$ are equal. Since realisations usually consist of more than one component, the definition needs to be extended. If we can suppose that the components in each realisation are independent and come from the same distribution, then we can define similarity of two random sets so that they are considered to be similar, if the distribution of their components is similar in the above mentioned meaning.

Consider two samples, namely $\mathcal{X} = (X_1, \ldots, X_{m_1})$ and $\mathcal{Y} = (Y_1, \ldots, Y_{m_2})$, of realisations of connected random sets $\mathbf{X}$ and $\mathbf{Y}$, respectively. We want to test the null hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ are similar. First, we evaluate the P/A-ratios $R_{X_1}, \ldots, R_{X_{m_1}}$, $R_{Y_1}, \ldots, R_{Y_{m_2}}$ of the perimeters and areas of the corresponding realisations. Based on these values, we estimate the $\mathcal{N}$-distance of the ratios of $\mathbf{X}$ and $\mathbf{Y}$ by (7), using (8) where we set $x_i = R_{X_i}$, $i = 1, \ldots, m_1$ and $y_j = R_{X_j}$, $j = 1, \ldots, m_2$. Let us denote it by $\widehat{\mathcal{N}}_1^R$. Then, we evaluate the testing functions $t(u)$ from (5), which describe the boundary curvatures $t_{X_1}(u), \ldots, t_{X_{m_1}}(u)$, $t_{Y_1}(u), \ldots, t_{Y_{m_2}}(u)$, calculate the $\mathcal{N}$-distance of the functions corresponding to $\mathbf{X}$ and $\mathbf{Y}$, respectively, using (9) and (7), and denote this $\mathcal{N}$-distance as $\widehat{\mathcal{N}}_1^t$. The couple $(\widehat{\mathcal{N}}_1^R, \widehat{\mathcal{N}}_1^t)$ is the test statistic. Here, we use the Monte Carlo permutation test described above, i.e. we make S permutations of all realisations $X_1, \ldots, X_{m_1}$ and $Y_1, \ldots, Y_{m_2}$, and split them into two groups of sizes m_1 and m_2, respectively, in order to obtain

$(\widehat{\mathcal{N}}_i^R, \widehat{\mathcal{N}}_i^t)$, $i = 2, \ldots, S+1$, and evaluate the p-value as

$$p = \frac{\sharp\{i \in \{2, \ldots, S+1\} : \widehat{\mathcal{N}}_i^R \geq \widehat{\mathcal{N}}_1^R \wedge \widehat{\mathcal{N}}_i^t \geq \widehat{\mathcal{N}}_1^t\} + 1}{S+1}.$$

3.3. Comparing the methods

	R vs B	R vs Be	R vs C	B vs Be	B vs C	Be vs C
D_{sign}	6	100	12	54	20	54
D_{band}	42	100	44	98	54	94
D_{sim}	26	94	22	82	34	94
2S	100	100	57	100	90	98

Table 1: The percentages of the p-values smaller than 0.05 when comparing different pairs of models, where B stands for Boolean disc model, Be for Boolean ellipse model, C for cluster model, and R for repulsive model. The first row corresponds to the percentages of p-values obtained when using depth based on the signed distance function, the second row when using band depth, the third when using simplicial depth, and the fourth when using two-step method.

From a statistical perspective, both methods perform very well when comparing realisations generated by the same model. This is evidenced by the uniform distribution of the corresponding p-value histograms over the interval (0, 1). More informative are the p-values obtained when comparing different models. The percentage of p-values smaller than 0.05 when comparing pairs of different models is shown in Table 1.

As mentioned in section 3.1, the three depth notions capture different geometric aspects of random sets. The band depth and the simplicial depth are primarily sensitive to size and topological structure. They perform well in detecting holes, disconnected components, and global inclusion differences, because both rely on set containment relations (intersection/union for band depth and Minkowski convex combinations for simplicial depth). Among the depth-based methods, the band depth showed the strongest empirical power in simulation studies and worked particularly well in distinguishing distributions of connected components. Simplicial depth has a solid geometric interpretation, but is computationally demanding and slightly less powerful than band depth in practice. In contrast, depth based on the signed distance function focuses on boundary geometry rather than topological structure. It is particularly sensitive to differences in convexity and local boundary irregularities, making it more effective for detecting shape deformation rather than

Method	**Advantages**	**Disadvantages**
D_{sign}	excellent detection of boundary curvature / convexity	high sensitivity to local boundary noise low power for topological differences properties depend on chosen distance metric high computational complexity (distance maps)
D_{band}	very good detection of holes, connected components and size differences low sensitivity to local boundary noise highest power among depth methods	high computational complexity (combinatorial, bootstrap) low detection of boundary curvature / convexity
D_{sim}	good detection of holes, connected components and size differences low sensitivity to local boundary noise	high computational complexity (convex approximations) low power low detection of boundary curvature / convexity
2S	flexibility (specific objects of interest can be considered) simple interpretation high accuracy	possible dependence of components

Table 2: Advantages and disadvantages of the considered methods, namely the method based on statistical depth when using signed distance function (first row), band depth (second row), or simplicial depth (third row), and the two-step method (fourth row).

structural changes such as holes. However, its properties depend on the chosen distance metric. It may also be more sensitive to noise in image-based representations. Overall, the band depth is most effective for topology-driven differences, simplicial depth offers strong geometric justification with moderate performance, and the signed distance depth excels in detecting boundary shape differences.

In the case of the two-step method (2S), the p-values presented in Table 1 suggest that this method has the greatest power among all the methods presented here and in [4]. The only obstacle shown by the simulation study is when comparing the cluster and repulsive models, which can be explained by the fact that the cluster model contains only a small number of large components, while the rest are repulsive-like, see Figure 1.

Table 2 summarises advantages and disadvantages of the presented methods. To conclude, the selection of a suitable method depends both on the intended objective and on the visual characteristics of the realisations. If the analysis requires accounting not only for component shapes but also for their topology, it may be more comprehensive to employ the method based on statistical depths, while if the main interest is on the shape of the component, the analysis should be based on two-step method.

4. Classification of realisations of random sets

The first step in classifying realisations of random sets is to construct the distances between individual realisations. Since the two-step method showed the greatest power among the existing methods, we decided to use the same distance that was described in section 3.2. Following [13], we focused on two characteristics described in section 2.3, namely P/A-ratio and C-function.

In the method presented below, we take a less restrictive view than in the case of testing the equality of distributions, and simply regard two realisations as more similar, the smaller their $\mathcal{N}$-distance is. Thus, when we refer to two realisations as similar, we mean that the empirical $\mathcal{N}$-distance between them is small, but not necessarily equal to zero.

We consider both supervised and unsupervised classification methods while drawing mainly from [8], [9], [22], and [28]. The common objective is to partition given realisations $X_1, \ldots, X_n$ of random closed sets $\mathbf{X}_1, \ldots, \mathbf{X}_n$ into k classes based on their similarity. This framework also enables us to assign a class label to any newly observed realisation.

For this purpose, the $\mathcal{N}$-distance between two realisations X_i and X_j is computed using the estimator given by (7) with the negative definite kernels (8) and (9). The former kernel is applied when only P/A-ratios are taken into account, while the latter is used when considering only C-functions. If both P/A-ratios and C-functions are considered together, we simply incorporate the value of P/A-ratio as an additional point of the C-function and then use (9).

Finally, let $K = \{1, 2, \ldots, k\}$ denote the set of class labels to be assigned to the realisations. Let $(\mathbf{X}_i, G_i),\ i = 1, ..., n$, be a sample of n independent

pairs, where the random variable G takes values in K. Note that we use (X_i, g_i) to denote the observation of the pair $(\mathbf{X}_i, G_i),\ i = 1, \ldots, n$.

The presented procedures are illustrated through a simulation study involving three models, namely the Boolean disc model, the cluster model, and the repulsive model, described in section 3, and shown in Figure 1. In the study, 200 realisations of size 400 × 400 pixels are considered. Since realisations of the models significantly differ in the number of components, an investigation of possible influence of the number of components on the calculation of the $\mathcal{N}$-distance between two realisations was done, similarly as in [13]. Namely, we calculate the distances using samples of 10, 20 and 'All' components, where 'All' means the number of components in the realisation with a smaller number of components. Further, we studied influnce of the amount of data (i.e., the number of realisations at disposal) on the performance of the classifier, considering samples of 20, 50 or 100 realisations. For supervised classification, data are partitioned into training and test sets in 75:25 ratio, while for unsupervised clustering, we process the entire dataset at once fixing $k = 3$.

4.1. Supervised classification of random sets

The idea of supervised classification is based on the Bayes rule. Given a realisation X of the random set $\mathbf{X}$, we estimate the posterior probabilities

$$p_c(X) = P(G = c|\mathbf{X} = X), \quad c \in K.$$

The realisation X is then assigned to the class with the highest estimated posterior probability.

We can use the kernel-type estimator

$$\widehat{p}_c(X) = \frac{\sum_{i=1}^{n} \mathbf{1}(g_i = c)\mathcal{K}(h^{-1}\mathcal{N}(X, X_i))}{\sum_{i=1}^{n} \mathcal{K}(h^{-1}\mathcal{N}(X, X_i))}, \tag{10}$$

where $\mathcal{K}$ is a kernel with the support $[0, 1]$ (i.e. $\mathcal{K}$ is positive and non-increasing in $[0, 1]$ and $\int_0^1 \mathcal{K} = 1$), and h is a bandwidth (a strictly positive smoothing parameter). It means that the closer X_i is to X, the larger is the value $\mathcal{K}(h^{-1}\mathcal{N}(X, X_i))$. Note that only X_i's with the distance less than h from X are taken into account since for $\mathcal{N}(X_i, X) > h$ it holds that $\mathcal{N}(X_i, X)/h > 1$ and therefore kernel $\mathcal{K}$ with the support $[0, 1]$ assigns zero value to such X_i. Thus, among the realisations X_i's belonging to the c-th class, the closer X_i is to X, the larger is its effect on the c-th estimated posterior probability, and X_i's that are farther than h have no effect at all.

As stated in [8], it is efficient to set the bandwidth h so that only m nearest neighbours of the realisation X are taken into account to calculate the kernel estimator (10). In order to choose the optimal m for each realisation X_{i_0}, denote by $h_{m(X_{i_0})}$ the bandwidth such that $\sharp\{i : \mathcal{N}(X_{i_0}, X_i) < h_{m(X_{i_0})}\} = m$. Further, denote

$$Loss(m, i_0) = \sum_{c=1}^{k} \left(\mathbf{1}(g_{i_0} = c) - p_{c,m}^{(-i_0)}(X_{i_0})\right)^2,$$

where

$$p_{c,m}^{(-i_0)}(X_{i_0}) = \frac{\sum_{i:i\neq i_0} \mathbf{1}(g_i = c)\mathcal{K}(h_{m(X_{i_0})}^{-1}\mathcal{N}(X_{i_0}, X_i))}{\sum_{i:i\neq i_0} \mathcal{K}(h_{m(X_{i_0})}^{-1}\mathcal{N}(X_{i_0}, X_i))}.$$

Then the optimal number of nearest neighbours m_{Loss} for X_{i_0} is

$$m_{Loss}(X_{i_0}) = \arg\min_m Loss(m, i_0)$$

and the corresponding bandwidth is the value $h_{m_{Loss}(X_{i_0})}$. We refer to this method as k-nn.

4.2. Unsupervised classification of random sets

4.2.1. Non-hierarchical clustering When studying unsupervised classification, we start with the well-known k-medoid algorithm described e.g. in [9]. The aim is to divide the realisations $X_1, \ldots, X_n$ to k classes.

The procedure begins by arbitrarily selecting k realisations $X_{i_1}, \ldots, X_{i_k}$ to serve as initial medoids. For each realisation X_i, $i = 1, \ldots, n$, the $\mathcal{N}$-distance from every current medoid is computed, and X_i is assigned to the class represented by the medoid from which it has the smallest $\mathcal{N}$-distance. Within each class, a new medoid is then determined as the realisation whose total sum of $\mathcal{N}$-distances to all other realisations in that class is minimal. Using these updated medoids, the assignment step is repeated. The procedure continues iteratively until the class memberships stabilise and no realisation changes its assigned class. Although the problem of selecting the optimal number of clusters k is widely discussed in the literature, we do not address it here, since in practical applications the number of clusters is typically determined by the specific context of the problem.

4.2.2. Hierarchical clustering Except for non-hierarchical clustering described in the previous section, we take a different approach by applying an

agglomerative hierarchical clustering approach, namely the Ward's method [28] which joins clusters sequentially using the Lance-Williams algorithm with suitably chosen parameters [22].

At the initial step, each realisation forms its own singleton cluster. At every subsequent step, the two clusters with the smallest mutual $\mathcal{N}$-distance are merged, and the distances between clusters are updated accordingly. For a set of realisations $X_1, \ldots, X_n$, the procedure starts by identifying the pair X_i, X_j with the smallest $\mathcal{N}(X_i, X_j)$ among all pairs. These two realisations are then replaced by a new cluster $\widetilde{X} = X_i \cup X_j$. The clustering process continues in the same manner: whenever two clusters $\widetilde{X}_1$ and $\widetilde{X}_2$, containing m_1 and m_2 realisations respectively, are merged, their distance to another cluster $\widetilde{X}_3$ with m_3 realisations is updated according to

$$\mathcal{N}(\widetilde{X}_1 \cup \widetilde{X}_2, \widetilde{X}_3) = \\ = \sqrt{\frac{m_1 + m_3}{m}\mathcal{N}^2(\widetilde{X}_1, \widetilde{X}_3) + \frac{m_2 + m_3}{m}\mathcal{N}^2(\widetilde{X}_2, \widetilde{X}_3) - \frac{m_3}{m}\mathcal{N}^2(\widetilde{X}_1, \widetilde{X}_2)},$$

where $m = m_1 + m_2 + m_3$. The procedure proceeds iteratively until all original realisations are merged into a single cluster. An advantage of this hierarchical approach is that the number of clusters does not need to be specified in advance; for any desired number k, one can simply stop the algorithm at stage in the hierarchical tree at which the original set of realisations is partitioned into k clusters, which corresponds to k steps before the final merge.

4.3. Comparing the methods

Figure 2 presents boxplots of the misclassification rate (defined as a ratio of misclassified realisations), when classifing a set consisting of 100 realisations from each model for different number of components considered. It is clear that the supervised case outperforms the non-supervised case in all settings, regardless of whether $\mathcal{N}$-distance was calculated based on only C-functions, only P/A-ratios, or both characteristics together. When each characteristic is used independently, the maximum misclassification rate is higher; however, it decreases once both characteristics are taken into account simultaneously. This confirms that integrating curvature with perimeter–area information yields a more reliable classification of random set realisations. The simulation study further shows that the proposed similarity-based classifiers are robust across a range of random set models, with their accuracy improving as the number of connected components grows. The reason behind it comes from the fact that the components affect the shape of their neighbouring com-

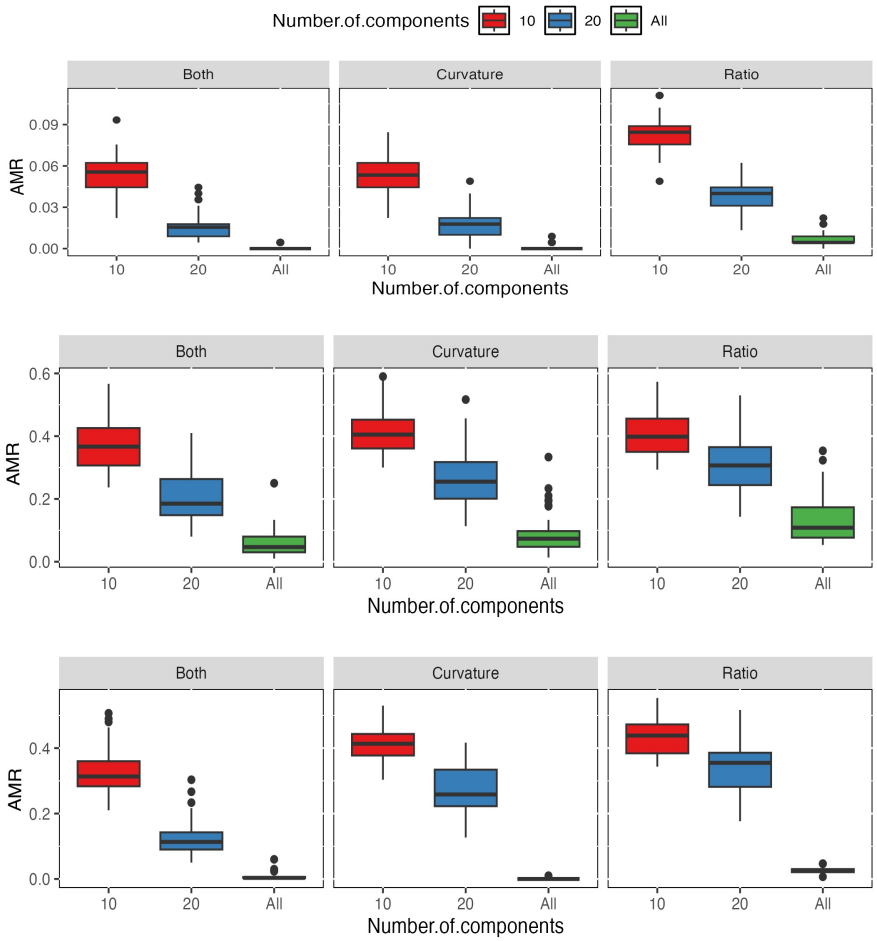


Figure 2: Boxplots of misclassification rate (AMR) for 50 runs of k-nn (top row), k-medoids (middle row) and hierarchical clustering algorithm (bottom row) when considering samples of 100 realisations from each model using both P/A-ratio and C-function, only the C-function 'Curvature' and only the P/A-ratio 'Ratio' for discrimination, respectively. For each setting, misclassification rates for different number of components considered (namely 10, 20 and 'All') are shown.

ponents, meaning that with a higher number of components considered, we will better catch the internal structure of a realisation, which leads to better categorisation. Furthermore, it was shown that the accuracy of classification improves with increasing number of realisations classified in the case of supervised learning, while it was not affected by it in the case of unsupervised learning.

In summary, a general framework for classifying random set realisations is introduced based on functional characteristics combined with $\mathcal{N}$-distances as similarity measures. By integrating supervised and unsupervised strategies, the approach remains applicable to both labelled and unlabelled data. Future research could aim to extend the methodology to higher-dimensional random sets, incorporate additional geometric or topological descriptors, study the performance of classifiers in the unbalanced case, and examine the theoretical properties of the proposed methods such as asymptotic behaviour and computational optimisation for large-scale datasets.

Acknowledgment: Supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS26/031/OHK3/1T/13.

Bibliography

[1] Bullard, J. V., Garboczi, E. J., Carter, W. C., & Fuller, E. R. Jr. (1995). Numerical methods for computing interfacial mean curvature. *Computational Materials Science, Vol. 4: 103–16.* https://doi.org/10.1016/0927-0256(95)00014-H *cit. 11*

[2] Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. (2013). Stochastic geometry and its applications. *John Wiley & Sons, New York. cit. 7*

[3] Debayle, J., Gotovac Đogaš, V., Helisová, K., Staněk, J., & Zikmundová, M. (2021). Assessing similarity of random sets via skeletons. *Methodology and Computing in Applied Probability.* https://doi.org/10.1007/s11009-020-09785-y *cit. 7*

[4] Debayle, J., Gotovac Đogaš, V, Helisová, K. & Staněk, J. (2021). Methods for assessing similarity of random sets. *10th International Symposium on Signal, Image, Video and Communications (ISIVC)*, Saint-Etienne, France, 2021, pp. 1–6. https://doi.org/10.1109/ISIVC49222.2021.9487547 *cit. 16, 20*

[5] Liu R. Y., Parelius., J. M., & Signh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference, *The Annals of*

Statistics, 27(3), pp. 783–858. https://doi.org/10.1214/aos/1018031260 *cit. 10, 11*

[6] Calle-Saldarriaga, A., Laniado H., & Zuluaga F. (2021). Homogeneity Test for Functional Data based on Data-Depth Plots. *Chemometrics and Intelligent Laboratory Systems, 219.* https://doi.org/10.1016/j.chemolab.2021.104420 *cit. 11*

[7] Fell, J. M. G. (1962). A Hausdorff topology for the closed subsets of a locally compact non-Hausdorff space. *Proceedings of the American Mathematical Society, 13(3), 472–476.* https://doi.org/10.1090/s0002-9939-1962-0139135-6 *cit. 8*

[8] Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice.* Springer, New York. *cit. 20, 22*

[9] Gordon, A. D. (1999). *Classification, 2nd Edition.* Chapman and Hall, Boca Raton. *cit. 20, 22*

[10] Gotovac, V., Helisová, K., & Ugrina, I (2016). Assessing dissimilarity of random sets through convex compact approximations, support functions and envelope tests. *Image Analysis and Stereology, 35, 181–93.* https://doi.org/10.5566/ias.1490 *cit. 7*

[11] Gotovac, V. (2019). Similarity between random sets consisting of many components. *Image Analysis and Stereology, 38, 185–99.* https://doi.org/10.5566/ias.2017 *cit. 7*

[12] Gotovac Đogaš, V., & Helisová, K. (2021). Testing equality of distributions of random convex compact sets via theory of $\mathcal{N}$-distances. *Methodology and Computing in Applied Probability, 23, 503–526.* https://doi.org/10.1007/s11009-019-09747-z *cit. 7, 14*

[13] Gotovac Đogaš, V., Helisová, K., Radović, B., Staněk, J., Zikmundová M., & Brejchová K. (2021). Two-step method for assessing similarity of random sets. *Image Analysis and Stereology, 40, 127–140.* https://doi.org/10.5566/ias.2600 *cit. 8, 20 a 21*

[14] Gotovac Dogaš, V. (2024). Depth for samples of sets with applications to testing equality in distribution of two samples of random sets. *Journal of Statistical Computation and Simulation, 94(16), 3507–3532.* https://doi.org/10.1080/00949655.2024.2394898 *cit. 8, 11, 15 a 16*

[15] Hermann, P., Mrkvička, T., Mattfeldt, T., Minárová, M., Helisová, K., Nicolis, O., Wartner, F., & Stehlík, M. (2015). 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.* https://doi.org/10.1002/sim.6497 *cit. 7*

[16] Klebanov, L. B. (2006). *$\mathcal{N}$-distances and their applications.* Karolinum Press, Charles University, Prague. *cit. 13, 14*

[17] López-Pintado, S. & Romo, J. (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association, 104(486), 718–734.* https://doi.org/10.1198/jasa.2009.0108 *cit. 9*

[18] Micheletti, A., Patti, S., & Villa, E. (2005). Crystal growth simulations: a new mathematical model based on the Minkowski sum of sets *In Industry Days 2003-2004 (D.Aquilano et al. Eds), volume 2 of The MIRIAM Project, pp. 130–140. Esculapio, Bologna.* http://www.mat.unimi.it/users/villa/files/Minkowski_simulations.pdf *cit. 10*

[19] Molchanov, I. (2013). *Theory of random sets.* Springer, New York. *cit. 7*

[20] Møller, J. & Helisová, K. (2008). Power diagrams and interaction processes for unions of discs. *Advances in Applied Probability* 40, 321–347. *cit. 15*

[21] Møller, J., Helisová, K. (2010). Likelihood inference for unions of interacting discs *Scandinavian Journal of Statistics, 37, pp. 365–81.* https://doi.org/10.1111/j.1467-9469.2009.00660.x *cit. 7*

[22] Murtagh, F., & Contreras, P. (2011). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2, 86–97.* https://doi.org/10.1002/widm.53 *cit. 20, 23*

[23] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., & Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79, 381–404.* https://doi.org/10.1111/rssb.12172 *cit. 14*

[24] Neumann, M., Staněk, J., Pecho, O. M., Holzer, L., Beneš, V., & Schmidt, V., (2016). Stochastic 3D modeling of complex three-phase microstructures in SOFC-electrodes with completely connected phases, *Computational Materials Science, 118, pp. 353–364* https://doi.org/10.1016/j.commatsci.2016.03.013 *cit. 7*

[25] Perricone, V., et al. (2022). Hexagonal Voronoi pattern detected in the microstructural design of the echinoid skeleton, *J R Soc Interface, 19:193, p. 20220226.* https://dx.doi.org/10.1098/rsif.2022.0226 *cit. 7*

[26] Radović, B., Gotovac Đogaš, V., & Helisová K. (2026+). Classification of realisations of random sets [Preprint]. https://doi.org/10.48550/arXiv.2511.00937 *cit. 8*

[27] Serra, J. (1982) Image Analysis and Mathematical Morphology, Vol. 2: Theoretical Advances. *Academic Press.* https://doi.org/10.1111/j.1365-2818.1988.tb01425.x *cit. 7*

[28] Ward, J. H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58, 236–244.* https://doi.org/10.2307/2282967 *cit. 20, 23*

[29] Whitaker, R.T., Mirzargar, M., Kirby, R. (2013), Contour boxplots: a method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE T Vis Comput Gr, 19, pp. 2713–2722.* https://doi.org/10.1109/TVCG.2013.143 *cit. 9*

GENEROVÁNÍ SYNTETICKÝCH DAT POMOCÍ GENERATIVNÍ AI

GENERATING SYNTHETIC DATA USING GENERATIVE AI

Mikuláš Gangur, Olga Martinčíková-Sojková

Adresa: Katedra ekonomie a kvantitativních metod, Fakulta ekonomická, Západočeská univerzita v Plzni, Plzeň, Česká republika

E-mail: gangur@fek.zcu.cz

Abstrakt: Příspěvek demonstruje použití jazykových modelů umělé inteligence ke generování syntetických dat pro statistické zpracování a úlohy analýzy dat. Výstupem práce s velkým jazykovým modelem (LLM) je požadovaný kód ve vybraném programovém prostředí (Matlabu), který generuje požadovaná syntetická data a je součástí automatického generátoru parametrizovaných úloh. Vyvinutý kód Matlab je zvláště užitečný při generování skutečných kategoriálních datových hodnot a je také schopen implementovat realistické vztahy mezi hodnotami různých kategoriálních atributů. Na příkladech promptů, jejich postupných úprav a jejich výstupů jsou ukázány schopnosti LLM i jejich slabiny a možná řešení k překonání těchto nedostatků.

Klíčová slova: automatické generování statistických úloh, XSL transformace, generování syntetických dat, AI LLM, Moodle, e-learning.

Abstract: The paper demonstrates the use of artificial intelligence language models to generate synthetic data for statistical processing and data analysis tasks. The output of work with a large language model (LLM) is the required code in a selected programming environment (Matlab), which generates the required synthetic data and is part of the automatic generator of parameterized tasks. The developed Matlab code is particularly useful in generating real categorical data values and is also able to implement realistic relationships between the values of different categorical attributes. Examples of prompts, their successive modifications, and their outputs demonstrate the capabilities of LLM as well as their weaknesses and possible solutions to overcome these shortcomings.

Keywords: automatic generation of statistical tasks, XSL transformations, synthetic data generation, AI LLM, Moodle, e-learning.

Úvod

Studijní aktivity, které mají podobu cvičných testů pro sebehodnocení studentů, představují důležitou součást učení prostřednictvím e-learningových kurzů. U těchto testů bez časového omezení není omezen počet opakování a otázka je zadávána v adaptivním režimu s podrobným komentářem (zpětnou vazbou) ke každému problému. Elektronické testování je užitečné pro studenty, aby si byli vědomi svých stylů učení, silných a slabých stránek, a ze strany učitele je k dispozici řada metod a přístupů k výběru těch nejvhodnějších [6]. Systémy pro řízení výuky (LMS) nám pomáhají nejen ve výukovém procesu, ale jsou také velmi užitečné při testování úkolů v různých oblastech [1].

Existuje mnoho aplikací pro automatizované generování testů, včetně například generátoru náhodných testů [5], testu Maestro II [7] nebo systému správy databáze otázek [8]. Tyto aplikace, stejně jako většina systémů pro řízení vzdělávání (LMS), používají testy generované náhodně ze souboru otázek v bance otázek (databázi otázek). Příprava otázek a budování takové banky otázek je obtížný a časově náročný úkol. Automatizace a zjednodušení této práce nejsou ve výše uvedených systémech řešeny.

Pro vytváření banky úloh využíváme univerzální princip automatického generování parametrizovaných úloh a pro tyto účely jsme vyvinuli Automatický generátor parametrizovaných úloh. Použité principy zjednodušují a zefektivňují budování banky otázek, viz [2]. Pomocí tohoto generátoru byly vytvořeny pro různé předměty banky obsahující tisíce unikátních problémů jako pomůcka v různých kurzech z různých oblastí matematiky, mezi které patří i kurzy statistických předmětů. Takto rozsáhlé banky úloh umožňují generování unikátních testů pro každého studenta účastnícího se kurzu.

1. Princip automatického generátoru úloh

Jádrem celého procesu je automatizovaný generátor úloh, viz [2]. Principem generujícího systému je funkční předpis řešeného problému (Řešitel) spolu s generátorem vstupních dat (IDG). Řešitel je funkce s proměnným počtem vstupních parametrů v závislosti na konkrétním problému. Generátor umožňuje automatické generování „vhodných“ vstupních dat na základě určených pravidel. Tato pravidla popisují vztahy ve vstupních datech. Algoritmus IDG se často implementuje jako procedura zpětného sledování řešení problému od očekávaných náhodně generovaných výsledků problému zpět ke vstupním datům. Výstupem IDG je kolekce vstupních dat a výstupem Řešitele je kolekce požadovaných výstupních dat. Následující seznam uka-

zuje datové typy parametrů náhodně generovaných vstupů a výstupů: číslo a znakový řetězec, tabulky, matice, popis, obrázky a datové sobory.

Data vygenerovaná popsaným postupem se používají jako vstup do generátoru úloh spolu se šablonou textu otázky a se strukturou (šablonou) univerzálního výstupního formátu XML. Generátor umožňuje zpracování úlohy vyžadující číselnou odpověď (NUM), úlohy s krátkou odpovědí (SA), dlouhou odpovědí (Popis) a úlohy s výběrem odpovědi (MC). Velmi důležitá je možnost zpracování úlohy s vloženou odpovědí (Cloze). Tento typ otázky může sestávat ze všech výše uvedených typů úloh. Většina vytvořených matematických/přírodovědných úloh vyžaduje uzavřenou otázku s ohledem na strukturovaná řešení (odpovědi).

Generátor otázek vloží vygenerovaná vstupní data do textu úlohy a poté tento text spolu s případnými komentáři vloží do šablony požadované úlohy a také do výstupních výsledků, které Řešitel vypočítá jako odpověď (řešení problému). V případě problému s NUM generátor vloží odpovědi pouze pro jednu otázku; v případě Cloze úlohy vloží odpovědi pro více otázek. Použitý přístup umožňuje určit více odpovědí pro otázku s NUM s danou úrovní validity. Výstupem generátoru otázek je soubor XML v univerzálním navrhovaném formátu, který je možné transformovat např. do formátu vybraného LMS pomocí konkrétní šablony úlohy nebo do formátu LaTeX verzí testu, které se skládají z náhodně vybraných a náhodně generovaných problémů. Podle použitého slovníku lze otázku generovat v různých jazycích. Aktuální verze generátoru je schopna zpracovávat otázky v češtině i angličtině. Celý proces je popsán na obrázku 1.

Aplikace generátoru pro generování matematických/přírodovědných úloh může být vytvořena v libovolném programovém prostředí. Aktuální funkční verze je implementována v Matlabu, který nabízí svou funkcionalitu a nástroje pro efektivní řešení dalších problémů s výše zmíněnou numerickou přesností. V praxi byl generátor použit pro tvorbu středoškolských matematických úloh. Pomocí tohoto generátoru bylo vytvořeno 250 úloh z oblasti kombinatoriky, analytické geometrie, rovnic, nerovnic a funkcí, což vedlo k vygenerování 13 000 unikátních úloh podle těchto vzorů. Finanční a investiční matematika je další oblastí použití automatického generátoru. Bylo připraveno 100 vzorů úloh a vygenerováno 12 000 unikátních úloh. Stejným způsobem byly vytvořeny úlohy pro předměty Základy statistiky, Statistické zpracování dat, popř. Vícerozměrná analýza ekonomických dat. Všechny tyto úlohy tvoří banky úloh implementovaných v jednotlivých e-kurzech v LMS Moodle.

2. Příklad použití generátoru

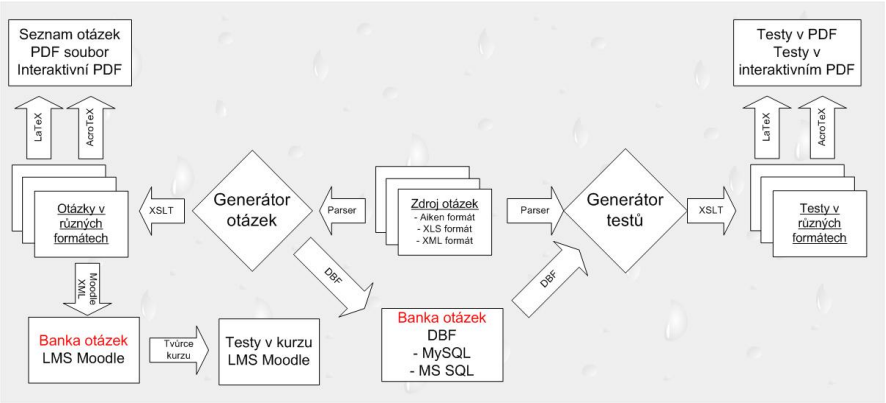
Na obrázcích 2–4 jsou ukázány 3 kroky generování vybrané úlohy. Proměnné pro vstupní parametry jsou označeny symboly `##` na obou stranách. V dané úloze lze nalézt proměnné `##pocet_obligaci##`, `##nominalni_hodnota##` a další, viz obrázek 2.

Výstupem z generujícího procesu je zadání a řešení úlohy ve formátu XML, které je vstupem do XSL transformace, viz obrázek 3.

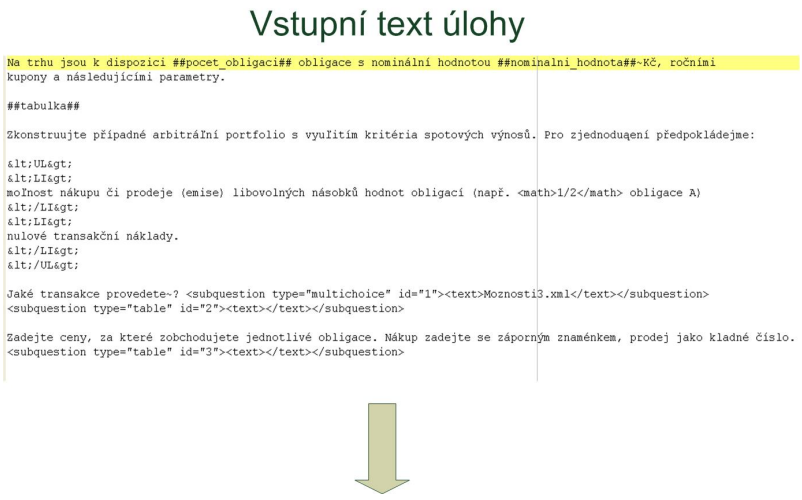
XSL transformace vygeneruje dle zadané XSL šablony zadání a řešení úlohy v požadovaném formátu. Nejvíce používané formáty jsou Moodle XML nebo LaTeX, ze kterého generátor vytvoří PDF soubor. Dalšími implementovanými formáty je AcroTeX pro vytvoření interaktivních PDF souborů, Aiken popř. formát pro převod úloh do struktury cvičení ve skriptech s výsledky na konci skript, viz obrázek 4.

3. Princip generování úloh s datovými soubory

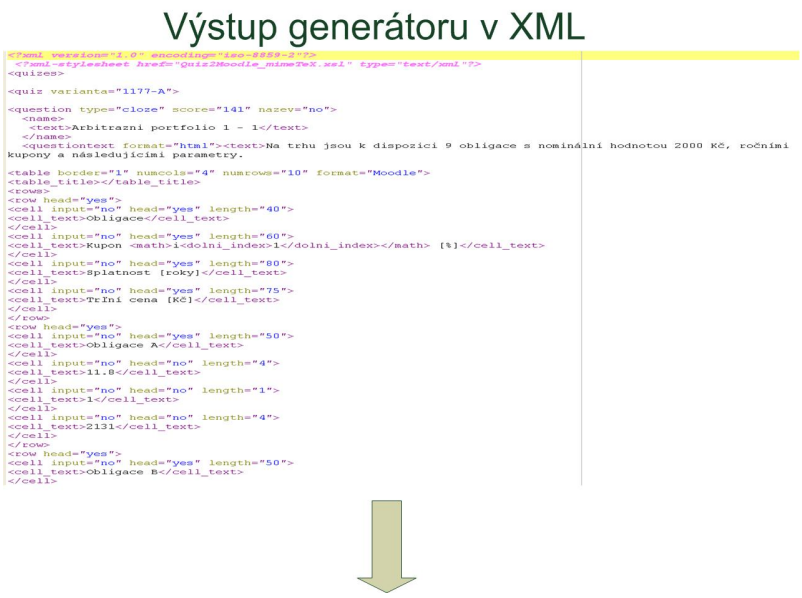
Jedním z výše uvedených datových typů vstupních parametrů, se kterými generátor pracuje, je datový soubor. Při generování dat ve vybraném cílovém formátu Moodle XML máme dvě možnosti řešení problému jako přiloženého datového souboru:



Obrázek 1: Proces generování úlohy a následná transformace do požadovaného formátu



Obrázek 2: Princip generování úlohy – krok 1 – Vstupní vzor

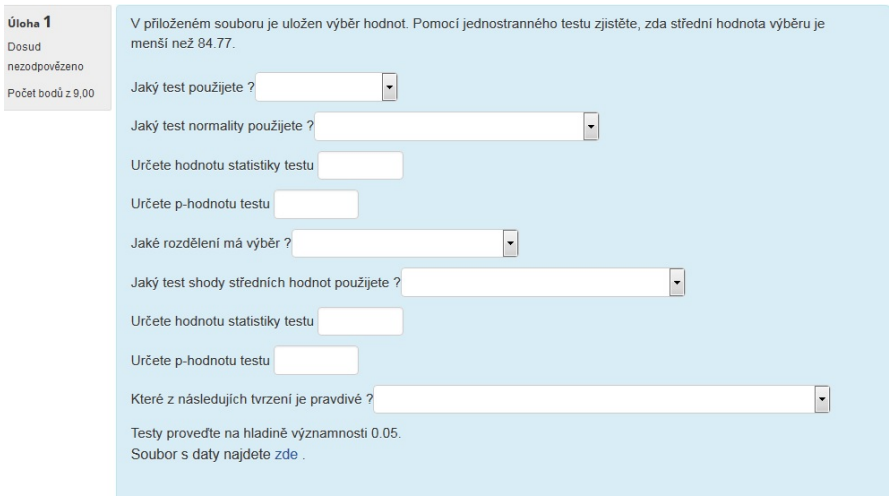


Obrázek 3: Princip generování úlohy – krok 2 – XML soubor

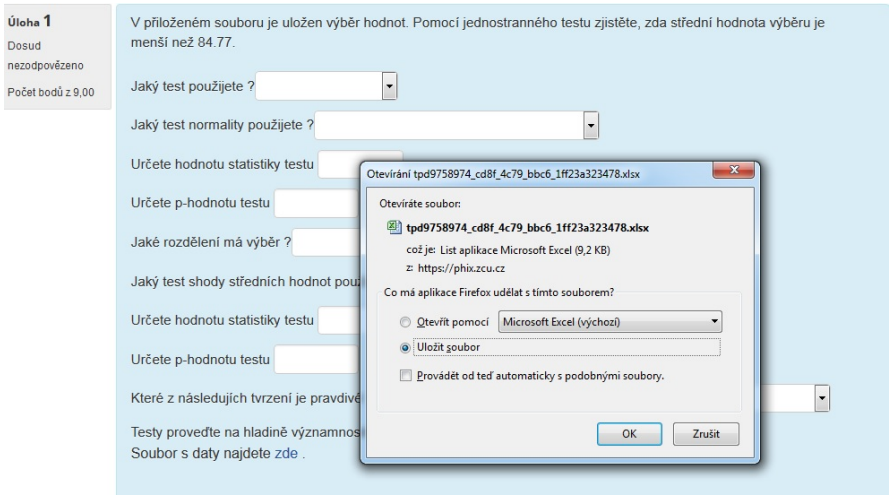
Obrázek 4: Princip generování úlohy – krok 3 – XSL Transformace

```
V přiloženém souboru je uložen výběr hodnot. Pomocí
##test_typ## testu zjistěte, zda střední hodnota výběru je
##vetsi_mensi## než ##hodnota##.

<br />

<subquestion type="multichoice" id="1" file=""><text>Jaký
test použijete~?</text></subquestion>
<subquestion type="multichoice" id="2" file=""><text>Jaký
test normality použijete~?</text></subquestion>
<subquestion type="numerical" id="3"><text>Určete hodnotu
statistiky testu</text></subquestion>
<subquestion type="numerical" id="4"><text>Určete p-hodnotu
testu</text></subquestion>
<subquestion type="multichoice" id="5" file=""><text>Jaké
rozdělení má výběr~?</text></subquestion>
<subquestion type="multichoice" id="6" file=""><text>Jaký
test shody středních hodnot použijete~?</text></subquestion>
<subquestion type="numerical" id="7"><text>Urči hodnotu
statistiky testu</text></subquestion>
<subquestion type="numerical" id="8"><text>Urči p-hodnotu
testu</text></subquestion>
<subquestion type="multichoice" id="9" file=""><text>Které
z následujících tvrzení je pravdivé~?</text></subquestion>

<br />

Testy proveďte na hladině významnosti ##alpha##.
```

Obrázek 5: Šablona zadání statistické úlohy



Obrázek 6: Vygenerované zadání v okně zobrazení úlohy v LMS Moodle



Obrázek 7: Dialogové okno při stažení datového souboru

1. Generovaná data se zobrazí na webové stránce ve formě tabulky bez jakýchkoliv omezení na konci zadání. Data lze označit v zadání a kopírovat je do požadované aplikace a dále je zpracovat.
2. Generovaná data jsou uložena do nezávisle vytvořeného souboru v požadovaném formátu (například txt, csv, xlsx) a tento soubor je následně uložen do repositáře LMS. Poté je vygenerován hypertextový odkaz na tento zdroj a tento odkaz je vložen do textu zadání.

Druhá možnost je popsána v následujícím příkladu. Máme vytvořenu šablonu zadání statistické úlohy, viz obrázek 5.

Obrázek 6 ukazuje vygenerovanou úlohu v okně zobrazení úlohy v LMS Moodle a poté na obrázku 7 je ukázán dialog při stahování souboru se statistickými daty, která uživatel použije k vyplnění řešení úlohy.

Detailnější popis uvedeného řešení lze nalézt v [3].

4. Úlohy s datovými soubory v cloudu

Druhým nejčastěji používaným výstupním formátem úloh je PDF soubor. V tomto případě nelze datový soubor přiložit k zadání jako v LMS Moodle. Řešením je uložení souboru na cloudu do složky s veřejným přístupem a vygenerovat uživatelsky čitelný URL odkaz na tento soubor na cloudu. Stejné řešení volíme i v případě velkých datových souborů, které mohou být společné pro více úloh, viz [4].

S ohledem na životnost vygenerovaného datového souboru je rozdělujeme do dvou skupin.

1. Datové soubory jsou úzce propojeny s daným vygenerovaným úkolem. V případě generování úkolu do banky otázek v Moodlu jsou tyto soubory součástí souboru s popisem celého úkolu. Například u statistických úloh se jedná o relativně menší výběrové soubory s výběry dat, které jsou pro každý úkol jedinečné.
2. Datové soubory jsou společné pro více úkolů. Takové soubory často obsahují velké množství dat a není žádoucí je ukládat do repozitáře společně s definicí každého úkolu. Příkladem mohou být základní soubory s velkým objemem celé populace.

Uvedené rozdělení ovlivňuje i správu těchto souborů při jejich ukládání do cloudu v případě generování úloh do výstupních formátů (např. TEX, LATEX, AcroTEX).

- Dočasné soubory lze v cloudu po určité době zrušit. Tyto soubory jsou generovány pouze pro účely jednorázových písemných testů generovaných ve formátu PDF.

- Pokud je nutné datový soubor uchovávat delší dobu, je možné takový soubor označit štítkem pro trvalý datový soubor (například odkaz na cvičný test v interaktivním formátu PDF lze vložit do e-kurzu v LMS na delší dobu).

V následujícím seznamu jsou uvedené přínosy popsaného řešení:

- Možnost generování datových kolekcí s ohledem na řešený problém a zejména integrace těchto dat do zadání úlohy ve formě připojeného datového souboru je velmi užitečná.
- Pro jednotlivé implementace generátoru je nutné řešit zejména správu souborů v cloudovém řešení v případě generovaných PDF dokumentů s vloženými URL odkazy na datové soubory.
- Řešení následné identifikace a stažení souboru z cloudu je možné pomocí aplikace, která je volána z vygenerovaného URL odkazu s jedním parametrem ve formě názvu datového souboru.

5. Generování syntetických dat

5.1. Motivace

Výchozí předpoklady pro použití syntetických dat a jejich generování jsou uvedeny v následujícím přehledu.

- Problémy s dostupností dat:
  - Reálná data jsou často omezena obavami o soukromí, bezpečnost a dostupnost.
  - Syntetická data mohou tyto problémy vyřešit, zejména v oblastech jako jsou finance, zdravotnictví a vzdělávání.
- Problémy s různou strukturou a velikostí dat. Odlišné podmínky testování znalostí studentů.
- Umělá inteligence a generování syntetických dat:
  - Role umělé inteligence při generování velkých a komplexních datových sad pro vzdělávací a výzkumné účely.
  - Jazykové modely založené na umělé inteligenci (např. ChatGPT, Microsoft Copilot) používané ke generování řízených tabulkových datových sad.

5.2. Metodologie řešení

Klíčové kroky při generování syntetických dat jsou následující:

1. Návrh datové struktury (např. auta, zaměstnanci).
2. Generování náhodných číselných a kategoriálních hodnot.
3. Vytvoření kódu Matlab pro automatizaci generování dat.

Pro každý z uvedených kroků je možné využít AI LLM. Při testování využití LLM byly identifikovány následující problémy.

- Obtížnost s vazbou hodnot proměnných tak, aby v rámci jedné statistické jednotky co nejvíce odpovídaly realitě.
- Potřeba složitých pokynů pro vedení modelů a omezení v dosažení plné realističnosti dat.

Další část ukazuje některé výsledky testování využití AI LLM. AI byla využita zejména pro generování hodnot kategoriálních statistických znaků. Kardinální statistické znaky je možné generovat přímo v kódu Matlabu dle zadaných parametrů jako je rozsah, distribuce apod.

5.3. Výsledky

První návrh řešení předpokládal přímé využití AI LLM při generování úlohy generátorem a ukládání datového souboru do generované úlohy. Problémy se ukázaly zejména při využití počátečních promptů (zero-promptů).

- Generování tabulkových dat (např. atributů automobilů) pomocí jednoduchých výzev vedlo k nejednoznačným výsledkům.
- Vazby mezi jednotlivými atributy neodpovídaly realitě, např. se na výstupu objevila značka auta Škoda s modelem Mustang.
- Výstup často vyžadoval vstup uživatele k opravě chyb (např. nekonzistentní kategoriální hodnoty).

Na základě těchto výsledků jsme přistoupili ke klíčovým úpravám při generování dat i při konstrukci promptu.

- Přechod na vygenerování kódu v Matlabu umožnil lepší kontrolu a reprodukovatelnost.

  Tímto řešením jsme ztratili větší flexibilitu při generování syntetických datových souborů. Na druhou stranu jsme tím umožnili kontrolu vygenerovaného kódu, který může být vložen do kódu generátoru vstupních dat a kontrolovaně generovat syntetická data v průběhu generování úlohy.

- Vylepšení promptu (např. zadání kardinálních a faktorových atributů) zlepšila proces generování dat.

Následují detailní příklady promptů v jednoduché studii generování datové sady automobilů

- **Prompt:** Vytvořte atributy pro tabulku Automobily s datovým typem a vytvořte kód Matlab pro generování 100 příkladů a uložte jej do souboru CSV.
- **Vytvořený kód:** Skript Matlab s náhodnými hodnotami pro atributy jako „Výrobce“, „Rok“, „Cena“ atd.
- **Vylepšení:** Rozšíření promptu pro zpřesnění konkrétních datových struktur zlepšilo reálnost, soudržnost a konzistenci datové sady.

Následuje rozšíření a zpřesnění promptu.

- **Prompt:** Vytvořte datovou strukturu pro automobily s faktorem a kardinálními atributy s minimálně 10 hodnotami pro značky a minimálně 3 modely pro každou značku, včetně náhodných příkladů v kódu Matlab.
- **Vytvořený kód:** Počáteční prompt generoval značky a modely, které spolu nesouvisely (např. Toyota s Mustangem).
- **Vylepšení:** Přepnutí do režimu Exact problém opravilo.

Odpověď na sestavený prompt používá informace a znalosti získané z předcházejících dotazů. Po novém přihlášení se model resetuje na prompt s nulovým počtem pokusů, což způsobuje, že „zapomene“ na předchozí kontext a vygeneruje nesprávné výsledky. Z tohoto důvodu je nutné pracovat na rozšíření promptu a použití tzv. Chain-of-thought promptu.

- **Prompt:** Vytvořte datovou strukturu pro auta se 3 faktory s několika obměnami a 5 kardinálními atributy. Každý faktor a kardinální atribut pojmenujte skutečným názvem. Vygenerujte 10 skutečných značek a 3 reálné odpovídající modely s jedinečnými názvy pro každou značku. Vygenerujte 3 skutečné motorizace pro každý model. Vygenerujte cenu pro každý model. Vygenerujte 100 příkladů v kódu Matlab.
- **Vytvořený kód:** Model splňoval většinu požadavků, ale často selhával při generování hodnot pro některé atributy, jako je barva a cena, s nekonzistentními výsledky při opakovaném spouštění.

5.4. Výzvy a řešení

S ohledem na provedené testy byly identifikovány následující problémy:

- Variabilita výstupů modelu AI.

- Nekonzistence v generování kategoriálních dat (např. nesprávné modely přiřazené značkám).
- Model „zapomíná“ na atributy při opakovaných promptech.

Následuje shrnutí možných řešení popsaných problémů. Tato řešení byla implementována při dalších testech.

- Při využití AI LLM se zaměřit spíše na generování kódu místo na přímé generování dat.
- Použít strukturované a detailní prompty ke sladění výstupů modelu s požadovanou strukturou a vlastnostmi dat (za obvyklý počet slov u „dobrého“ promptu se považuje 500 slov).

6. Závěr

V příspěvku byly popsány základní principy automatického generátoru parametrizovaných úloh doplněné příklady procesu generování úlohy. Jako jeden z datových typů vstupních informací je možné použít i datový soubor. Byly popsány různé možnosti uložení vygenerovaných dat zejména do banky úloh v LMS Moodle nebo jako soubor do uložiště na cloudu. Jádrem příspěvku je popis procesu generování syntetických statistických dat s podporou AI LLM a následná hlavní zjištění z výsledků provedených testů.

- Jazykové modely založené na AI se jeví jako slibné při generování syntetických dat, ale vyžadují pečlivý návrh promptu.
- Přechod na výstup generujícího procesu v podobě programového kódu poskytuje spolehlivější metodu pro generování strukturovaných dat než přímé generování těchto dat s pomocí AI.

Z uvedeného vyplývají následující náměty pro další výzkum. Prozkoumat generování konzistentních numerických dat na základě specifických vlastností nebo rozdělení s pomocí AI LLM. Další zdokonalení technik promptu pro zajištění konzistentnějších výstupů napříč více pokusy. Po zvýšení konzistence výstupů AI je možné se vrátit k původnímu návrhu řešení, tj. přistoupit k testování možností online využití AI LLM při generování syntetických dat v průběhu vytváření statistických úloh založených na analýze vícerozměrných statistických dat.

Literatura

[1] Aleksic-Maslac, K.; Vasic, D.; Korican, M.: Student learning contribution through E-learning dimension at course "management information systems". In *WSEAS Transactions on Information Science and Applications 7 (3)*, WSEAS Publishing, 2010, s. 331–340. *cit. 30*

[2] Gangur, M.: Automatic generation of cloze questions. In *Proceedings of 3rd International Conference on Computer Supported Education Vol. 1*, SciTePress, Portugal, May 2011, ISBN 978-989-8425-49-2, s. 264–269. *cit. 30*

[3] Gangur, M.: Automated generation of statistical tasks. In *DIVAI 2018: 12th International Scientific Conference on Distance Learning in Applied Informatics*, Wolters Kluwer, May 2018, ISBN 978-80-7598-059-5, s. 47–58. *cit. 36*

[4] Gangur, M.: The use of cloud for generated statistics data file storage in statistics tasks. In *DIVAI 2022: 14th International Scientific Conference on Distance Learning in Applied Informatics*, Wolters Kluwer, May 2022, ISBN 978-80-7676-410-1, s. 373–383. *cit. 36*

[5] RTG-PRO: Random Test Generator-PRO. January 2011. https://random-test-generator-pro.software.informer.com/. *cit. 30*

[6] Šimonová, I.; Poulová, P.; Bílek, M.: Learning styles within eLearning: Didactic strategies. In *Proceedings of the 10th WSEAS International Conference on Applied Computer and Applied Computational Science*, 2011, s. 160–164. *cit. 30*

[7] TM-II: Test Maestro II. January 2011. https://test-maestro-ii.apponic.com/. *cit. 30*

[8] Yang, A.; Wu, J.; Wang, L.: Research and design of test question database management system based on the three-tier structure. In *WSEAS Transactions on Systems 7 (12)*, 2008, s. 1473–1483. *cit. 30*

PLÁN AKCIÍ SŠDS V ROKU 2026
PLANNED EVENTS OF SLOVAK STATISTICAL AND DEMOGRAPHIC SOCIETY

Iveta Stankovičová

E-mail: iveta.stankovicova@gmail.com

Akcie SŠDS v roku 2026 – plán

18. 3. 2026 – 23. 1. 2026, podporené cudzie vedecké či odborné
ROBUST 2026, Šumava (Srní), Česká republika
Hlavný organizátor: ČMS JČMF, ČStS

Marec 2026, popularizačné pre študentov
Analytika očami profesionálov, FHI EUBA, Bratislava

23. 4. 2026, vlastné vedecké či odborné
Pohľady na ekonomiku Slovenska 2026
Online, 3 až 4 prednášky prognostikov

Apríl 2026, popularizačné pre študentov
Prehliadka prác mladých štatistikov a demografov, Bratislava
Súťaž prác študentov vysokých škôl

24. 5. 2026 – 26. 5. 2026, vlastné vedecké či odborné
EKOMSTAT 2026, Trenčianske Teplice
Zodpovedný organizátor: Ivan Lichner

18. 6. 2026 – 20. 6. 2026, vlastné vedecké či odborné
23. slovenská štatistická a demografická konferencia, Stará Lesná

26. 8. 2026 – 29. 8. 2026, podporené cudzie vedecké či odborné
AMSE 2026, Martin
Hlavný organizátor: EF UMB v Banskej Bystrici

28. 11. 2026, vlastné vedecké či odborné
Výpočtová štatistika 2026, Bratislava

Priebežne
Regionálne akcie: diskusné popoludnia, prednášky

Vypracovala: Iveta Stankovičová,
predsedníčka SŠDS

SEMINÁŘ HÁJEK 100 LET OD NAROZENÍ HÁJEK STATISTICAL DAYS 2026

Jaromír Antoch

E-mail: antoch@karlin.mff.cuni.cz

Dámy a pánové,

letos uplyne 100 let od narození profesora Jaroslava Hájka, s jehož jménem se převážná většina z Vás setkala.

Naše katedra se proto rozhodla uspořádat při této příležitosti seminář, který se uskuteční ve dnech **4. – 6. června 2026** na **KPMS MFF UK v Praze.** Seminář bude zdarma. Podrobnosti o semináři najdete na adrese

https://www.karlin.mff.cuni.cz/~hajek/

Na seminář jsme pozvali několik zahraničních kolegů, kteří profesora Hájka ať již osobně pamatují, nebo na jeho práce navázali. Vedle toho jsme vyhradili prostor pro vystoupení jednotlivých účastníků. Budeme velmi rádi, pokud mezi nimi butede i vy, ať již aktivně, či pasivně.

Pozvaní řečníci:

- Rudy Beran (UC Davis),
- Marc Hallin (Université libre de Bruxelles),
- Zuzana Prášková (Charles University),
- Aad van der Vaart (TU Delft),
- Noel Veraverbeke (Hasselt University),
- Silvelyn Zwanzig (Uppsala University).

Budeme také velmi rádi, pokud budete tuto informaci šířit mezi svými kolegy jak domácími, tak zahraničními, a studenty.

Kontakt na organizátory: hajek@karlin.mff.cuni.cz.

Za organizátory se na Vás těší Marie Hušková, Jaromír Antoch, Daniel Hlubinka, Miloš Kopa a Michal Pešta.

KONFERENCE USER! 2026 VE VARŠAVĚ
USER! 2026 CONFERENCE WARSAW

Przemysław Biecek

E-mail: useR2026@r-project.org

Join the useR! at Warsaw!

We're excited to announce that useR! Conference in 2026 is being hosted in Warsaw, bringing together data scientists, statisticians, and researchers from across the world.

What makes this edition special is that it is jointly organized by three leading local universities, showcasing the strength of Warsaw's academic community.

Expect inspiring talks, hands-on workshops, and plenty of chances to meet fellow R enthusiasts. Whether you're into research, data science, or just curious about R, this is the place to be.

When: July, 6–9, 2026.

Where: SGH Warsaw School of Economics.

Keynote speakers:

- **Dariia Mykhailyshyna:** R Under Sirens: Research, Students, and Community in Wartime Ukraine.
- **Peter Dalgaard:** Release management and governance structure of the R project.
- **Dianne Cook:** Interactive Graphics for Understanding and Interpreting Nonlinear Model Behaviour in High Dimensions, using R.
- **Jakub Nowosad:** A world still to be mapped: reflections on geocomputation in R.
- **Kari Jordan:** The Work Behind the Work: Sustaining R Through Community.

More information: https://user2026.r-project.org/.

STATISTICKÉ DNY 2026: POZVÁNKA
STATISTICAL DAYS 2026: AN INVITATION

Martina Litschmannová

E-mail: martina.litschmannova@vsb.cz

Česká statistická společnost pořádá i v roce 2026 tradiční Statistické dny, letos ve spolupráci s Katedrou matematické analýzy a aplikací matematiky, PřF UP Olomouc. Konference se věnuje aktuálním tématům spojeným s teorií, aplikacemi a výukou statistiky. Setkání odborníků, pedagogů i studentů se tentokrát uskuteční v hanácké metropoli — Olomouci.

Olomoucké statistické dny 2026

Registrace | **Abstrakt** | **Vložné** | **Příspěvek:** do 10. července 2026

Kde: PřF Univerzity Palackého v Olomouci & Penzion Tõde (Pravda)

Kdy: 10. – 11. září 2026

Pro letošní ročník organizátoři zvolili tři hlavní tematické okruhy:

- Aplikace statistiky v medicíně.
- Aplikace statistiky v průmyslu.
- Inovace výuky statistiky v éře AI.

V programu rádi uvítáme i příspěvky ze sociologie, ekonomie, financí, teorie pravděpodobnosti a matematické statistiky, dále příspěvky k výuce statistiky a statistickému softwaru. Finální podoba programu bude sestavena na základě počtu a tematického zaměření přihlášených příspěvků.

Účastníkům konference bude zajištěno ubytování v Penzionu Tõde, https://www.sestrynp.cz/cs/clanek-penzion, který se nachází v klidné uličce na olomouckých hradbách v historickém centru města. Z jedné strany penzion nabízí výhled do zahrady a přilehlého parku, z druhé do malebné historické uličky. Budova je třípodlažní, zrekonstruovaná, s kapacitou 25 lůžek ve 13 pokojích.

Odborný program bude probíhat v prostorách Přírodovědecké fakulty Univerzity Palackého v Olomouci na třídě 17. listopadu.

Bližší informace na: https://www.statspol.cz/STATDNY2026.

Za organizační výbor se na Vás moc těší Ondřej Vencálek, Kamila Fačevicová, Elizabeth Princová, Jaromír Antoch, Martina Litschmannová a Tomáš Löster.