Expert Systems With Applications 206 (2022) 117809

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automation of cleaning and ensembles for outliers detection in questionnaire data

Vojtěch Uher [a,*], Pavla Dráždilová [a], Jan Platoš [a], Petr Badura [b]

[a] Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VSB - Technical University of Ostrava, 17. listopadu 15/2172, Ostrava, 708 33, Czech Republic

[b] Department of Recreation and Leisure Studies, Faculty of Physical Culture, Palacký University Olomouc, třída Míru 117, Olomouc, 771 11, Czech Republic

ARTICLE INFO

Dataset link: http://hbsc.cz/lockdown2020/

Keywords:
Anomaly detection
Outliers
Questionnaire data
Data cleaning
HBSC

ABSTRACT

This article is focused on the automatic detection of the corrupted or inappropriate responses in questionnaire data using unsupervised outliers detection. The questionnaire surveys are often used in psychology research to collect self-report data and their preprocessing takes a lot of manual effort. Unlike with numerical data where the distance-based outliers prevail, the records in questionnaires have to be assessed from various perspectives that do not relate so much. We identify the most frequent types of errors in questionnaires. For each of them, we suggest different outliers detection methods ranking the records with the usage of normalized scores. Considering the similarity between pairs of outlier scores (some are highly uncorrelated), we propose an ensemble based on the union of outliers detected by different methods. Our outlier detection framework consists of some well-known algorithms but we also propose novel approaches addressing the typical issues of questionnaires. The selected methods are based on distance, entropy, and probability. The experimental section describes the process of assembling the methods and selecting their parameters for the final model detecting significant outliers in the real-world HBSC dataset.

1. Introduction

The anomaly (or outlier) detection reveals observations that seem to be inconsistent with the rest of the data. Hawkins (1980): "An outlier is an observation which deviates so much from the other observations as to arouse suspicious that it was generated by a different mechanism". An outlier may indicate corrupted data such as manual error, coding error, low-quality measurements, system failure, etc. On the other hand, suspicious data can also represent some unknown or infrequent type of data that a dataset does not capture correctly. Outliers need to be detected, assessed, excluded from the analysis, or fixed if possible. They can negatively affect data distribution and weaken the reliability and credibility of the statistical analysis and its results (Shao, Zheng, Gu, Hu, & Qin, 2022; Wilcox, 2019; Yuan & Gomer, 2021).

The objective of this article is to assemble and test a robust framework using an ensemble of unsupervised methods for outliers detection in raw questionnaire data. The data usually comes from questionnaire surveys (Saris & Gallhofer, 2014) that are typically used in social and behavioral research, psychology, health, etc. They cover demographic information, personal characteristics, or opinions of people. Unlike the machine-acquired numerical data used in the industrial and security areas, the questionnaire data has some specific issues coming from the human factors and the whole methodology designing the questionnaires to find out the required information. In questionnaire data, an observation represents the answers of one person (respondent) in a survey. It typically suffers from the following issues: many incomplete questionnaires, empty and intentionally wrong answers (e.g., repetitive patterns, self-contradictory responses), fast responses, questions are discrete categorical un/ordered items with a various number of options, normality or independence of items cannot be generally secured. We specifically examine the data from the Health Behavior in School-aged Children (HBSC) 2020 study (Inchley, Currie, Cosma, & Samdal, 2018; Ng, Cosma, Svacina, Boniel-Nissim, & Badura, 2021) that focuses on health and health behaviors of 11-, 13-, and 15-year-olds in the context of their social environments. We do not have any training set or any prior information about outliers.

Most of the existing approaches solve outlier detection in numerical data, using some proximity-based methods (Agrawal & Agrawal, 2015; Chandola, Banerjee, & Kumar, 2009; Zimek, Schubert, & Kriegel, 2012) (e.g. k-nearest neighbors, local outlier factor). The results are usually validated using supervised learning and labeled data. We aim at unsupervised statistical and machine learning methods for multivariate data with a special focus on the questionnaire-related issues. The goal

* Corresponding author.
E-mail addresses: vojtech.uher@vsb.cz (V. Uher), pavla.drazdilova@vsb.cz (P. Dráždilová), jan.platos@vsb.cz (J. Platoš), petr.badura@upol.cz (P. Badura).

https://doi.org/10.1016/j.eswa.2022.117809
Received 2 March 2022; Received in revised form 23 May 2022; Accepted 6 June 2022
Available online 13 June 2022

is to maximally automate the data cleaning, validation, and outliers detection. These activities are still often done manually (Waure, Poscia, Virdis, Pietro, & Ricciardi, 2015) and they take a lot of time.

The main contributions of this article include:

1. A crucial automatic data cleaning and preprocessing procedure.
2. Selected outlier detection methods optimized for questionnaire data that compute an outlier score reducing the data space onto one final variable.
   (a) Widespread methods (Mahalanobis distance, k-nearest neighbors, local outlier factor).
   (b) Innovated methods (the entropy-based scores dealing with empty answers and repetitive patterns, the correlation-based methods examining the significant dependencies between variables and their corruption, the probability score dealing with rare and improbable answers).
3. A selection of the proper statistical method determining the range of scores representing the finally removed outliers. The scores are assessed according to the distribution of the outlier score variable.
4. An ensemble outlier detection procedure that unites the outliers detected by selected methods assessing the outliers from different perspectives.
5. The case study computed on the HBSC 2020 data evaluating the proposed pipeline and the impact of outliers to data variance and covariance of attributes (Cronbach's alpha (Cronbach, 1951)).

The article is organized as follows. Section 2 summarizes the state-of-the-art literature with a special focus on the categorical and questionnaire data. Section 3 defines our proposed pipeline for preprocessing questionnaire data and ensemble outlier detection. In Section 4, the types of outliers are listed, and the applied outliers detection methods are described in detail. Section 5 tests the defined methods with the HBSC 2020 data. The experiments prove that our framework efficiently addresses the named problems and significantly improves the distribution and the statistical properties of the HBSC 2020 data.

2. Related work

The questionnaires suffer from many errors and difficulties that must be solved before any further analysis of the collected data is performed. The procedure addressing these issues consists of data cleaning, preprocessing, and outlier detection. The cleaning and preprocessing need to be conducted before the outlier detection methods, and it has been described in the literature many times (García, Luengo, & Herrera, 2015; Van den Broeck, Argeseanu Cunningham, Eeckels, & Herbst, 2005; Zhu, Hernandez, Mueller, Dong, & Forman, 2013). We focus on this topic in Section 3.

The outlier detection algorithms can be basically divided into two categories: supervised and unsupervised. The supervised algorithms are based on training a classifier using the labeled training dataset and they include e.g. Support Vector Machine (Utkin, 2014), Artificial Neural Networks (Kieu, Yang, & Jensen, 2018; Naseer et al., 2018), and Decision Tree (Ramachandran & Kishorebabu, 2019). As it is very difficult to obtain labeled data, the unsupervised methods that do not require labeled data are often used to compute a degree of outlierness. They include e.g. statistical (Hubert & Vandervieren, 2008; Tukey et al., 1977), proximity-based (Breunig, Kriegel, Ng, & Sander, 2000; Leys, Klein, Dominicy, & Ley, 2018; Shao et al., 2022), and clustering-based methods (Jiang, Liu, Du, & Sui, 2016; Wang, Wang, & Wilkes, 2012). We focus on the unsupervised methods applicable to multivariate questionnaire data.

Most of the named approaches are designed for numerical data. The questionnaire data usually contains categorical or ordinal discrete variables with a small number of categories. Categorical data is also common in areas such as network intrusion detection (Sari et al., 2015), social networks (Aggarwal, Zhao, & Philip, 2011), industrial processes (Zhu, Ge, Song, & Gao, 2018), sensor faults in sensor networks (Zhang, Meratnia, & Havinga, 2010), credit card fraud (Malini & Pushpa, 2017), etc. The outlier detection methods are based on different approaches of data processing (Akoglu, Tong, Vreeken, & Faloutsos, 2012; Ienco, Pensa, & Meo, 2016). A survey of various outlier detection methods for categorical data is presented by Taha and Hadi (2019).

While the general categorical data has been widely investigated in the literature, articles about outliers detection in questionnaire data are rare (Sakurai et al., 2019; Zijlstra, van der Ark, & Sijtsma, 2011). The selected methods applied to categorical data can be divided into several categories:

- **Probability-based:** (Zijlstra et al., 2011; Zijlstra, Van Der Ark, & Sijtsma, 2007) proposed a method detecting outliers in questionnaire data based on the frequency of answers. The answers are sorted by frequencies so that the infrequent answers have a higher index/rank. The final outlier score of one questionnaire is a sum of the ranks of the contained answers. The greater the score is the more improbable and suspicious answers are contained. This approach is very simple but it can be highly biased if a specific answer predominates.
- **Proximity-based:** This category includes methods based on distance (e.g. k-nearest neighbors Chandola, Banerjee, & Kumar, 2007) or density (e.g. local outlier factor Breunig et al., 2000). The main problem of these methods is to determine the size of neighborhood or estimate the number of samples expected within the neighborhood depending on the distribution of data (Uher, Gajdoš, & Snášel, 2018; Uher, Gajdoš, Snášel, Lai, & Radeckỳ, 2019). Another issue can be the usability of the Euclidean distance in high dimensions.
  A common-neighbor-based distance function was developed by Li, Lee, and Lang (2007) to measure the proximity of a pair of data points in distance-based outlier detection method for high-dimensional categorical data. A weighted density that takes into account the density and uncertainty of each categorical variable is defined by Zhao, Liang, and Cao (2014).
  Multivariate outliers are often detected by variants of the Mahalanobis distance which represents the distance from the distribution of data (Ben-Gal, 2005; Cabana, Lillo, & Laniado, 2019; Leys et al., 2018) and it is non-parametric.
  The principal component analysis (PCA) is used for outliers detection or outliers visualization (Har-Shemesh, Quax, Lansing, & Sloot, 2020; Sakurai et al., 2019; Zhu et al., 2018). PCA is also one of the methods used by Jebreel et al. (2020) to sanitize survey data that relies on combining the classification outcomes of unsupervised machine learning algorithms aimed at detecting wrong answers. Deng and Wang (2018) proposed a modified kernel PCA (KPCA) method with local outlier factor (LOF) to construct multivariate statistical process monitoring methods. Several variants of LOF (Breunig et al., 2000) have been proposed to handle different data types. A variant of LOF combined with information entropy is applied by Xie, Li, Wu, and Zhang (2016) for detecting outliers in medical insurance data. Yu, Qian, Lu, and Zhou (2006) used a mutual-reinforcement-based local outlier detection κ-LOF to handle categorical attributes. The variants of k-nearest neighbors are also often used for outliers detection (Chandola et al., 2007; Chen, Miao, & Zhang, 2010). The outlier score can be computed as a sum of distances between the query object q and its k nearest neighbors. Some other types usually require parameters k and distance λ to assess if neighborhood with radius λ contains at least k objects, or the λ is used as a distance threshold (Chandola et al., 2007; Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002; Knorr & Ng, 1998).

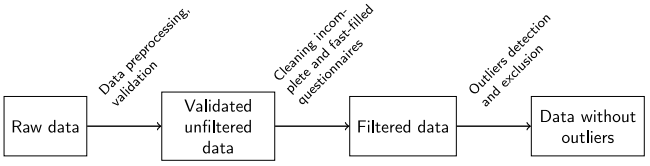


Fig. 1. Flow of data processing.

- **Entropy-based:** Algorithms based on the entropy were investigated in the area of outlier mining for categorical data (He, Deng, Xu, & Huang, 2006; Pacheco, Ali, & Trappenberg, 2019; Yuan, Zhang, & Feng, 2018). The Shannon entropy is often used to search for a set of observations that minimize the Shannon entropy of the observations remaining after the outliers elimination (He et al., 2006). An extension of Shannon information entropy in rough sets is the approximation accuracy entropy (AAE) introduced by Jiang, Zhao, Du, Xue, and Peng (2019). This approach addresses the limits of the proximity-based methods.

Finally, the evaluation of the detected outliers is an important task. But it is difficult to determine the extent to which newly proposed methods are improving compared to established methods in unlabeled data. Campos et al. (2016) conduct an extensive experimental study on the performance of a representative set of methods for unsupervised outlier detection, across a wide variety of datasets prepared for this purpose. As the methods differ and have various distributions of scores, each of them returns slightly different outliers. The ensemble learning generally combines multiple weak classifiers to obtain better overall performance but it mostly aims at supervised methods on labeled data. Several papers (Kriegel, Kroger, Schubert, & Zimek, 2011; Schubert, Wojdanowski, Zimek, & Kriegel, 2012) analyze the rank similarities, correlation of scores, regularization and normalization of scores to make them comparable, and weights estimation for ensemble design. The class labels are helpful for performance evaluation and optimal parameters selection. Most of the named papers rely on the algorithms based on kNN and LOF. There are also unsupervised approaches that compute ensemble outlier score to detect the outliers from multiple perspectives without any knowledge of data. They include e.g. average scoring, maximum scoring, or threshold sum (Aggarwal & Sathe, 2015; Kandanaarachchi, 2021). The main weakness of these ensembles is that they usually lead to average results that blur the outliers specific for standalone methods. The parameters optimization is insecure because the outlier labels are unknown. We overcome this problem with the usage of separate score distribution analysis and unification of identified outliers.

For detailed survey of outliers or anomaly detection methods in various application fields, we refer to articles (Agrawal & Agrawal, 2015; Chandola et al., 2007, 2009; Das, Schneider, & Neill, 2008; Hodge & Austin, 2004; Wang, Bah, & Hammad, 2019).

3. Proposed procedure

The specific issues related to questionnaire data and the design of the final ensembles are discussed here. We present our proposed procedure which starts with raw data, then continues with the specific steps of cleaning and preprocessing of data, sets the requirements for outliers detection, and defines the ensemble method combining the results of multiple algorithms. The pipeline is illustrated in Fig. 1.

3.1. Cleaning and preprocessing of questionnaires

A dataset generally forms a table where columns represent the asked questions (attributes) and rows represent the answers of respondents (individual questionnaires). Our algorithms aim only at the multiple-choice questions (predetermined list of options) with a single answer including rating scales, Likert scales, or matrix questions (batteries) (Harpe, 2015). They can be seen as categorical variables with a small number of options (mostly up to 10). Other types of questions have to be assessed separately and are omitted before the outliers detection methods are applied. We do not set any other a priori conditions for data.

The preserved data has to be validated and cleaned. There are several widely-used approaches for cleaning the questionnaire data (García et al., 2015; Van den Broeck et al., 2005; Zhu et al., 2013) that generally lead to three common issues: **Unrealistic answers** (e.g. wrong type, value out of range), **Incomplete questionnaires** (many missing values), **Response time** (very short times indicate untrustworthy answers). The unrealistic answers can be detected with the utilization of the questionnaire codebook defining unambiguous attribute details. The exclusion of incomplete questionnaires or the fastest respondents can be simply done automatically by setting reasonable thresholds. Our tested preprocessing procedure goes through the following steps:

1. **Selection of questions with a predetermined list of options** - the omission of irrelevant variables such as strings, dates, personal information (birth date, weight, height, etc.), system and browser information (response times, device information, software version, etc.)
2. **"Other" answers** — any open-ended responses in questions with predetermined options are said to be missing (empty).
3. **Elimination of questions with bad distribution** — the omission of dichotomous variables and strongly unilateral variables with more than 80% of answers choosing the same option. Such questions bring a strong bias and undesirably strong correlation between attributes.
4. **Missing values** — all values representing empty answers are replaced by zeros.
5. **Renumbering of option indices** — makes a continuous sequence indexing the options of a question in the original order. The indices start with zero for missing values and continue with other natural numbers. This unifies the indexing and annuls prioritization of any option.
6. **Normalization** — as each question may have a different number of options, their indices recorded as answers in questionnaire data are normalized by variable onto the interval $\langle 0, 1 \rangle$. The zero still represents a missing answer.
7. **Fast response times** — we eliminate all the questionnaires with almost zero response time (if available). Some small threshold has to be chosen.
8. **Questionnaires with many missing values** — we eliminate all the questionnaires with more than 70% of missing values. This step usually also eliminates the records with zero time as the rapid respondents do not fill anything.

The restrictions listed above do not mean that those variables, values, or questionnaires are completely wrong. They can carry some useful information but they are undesirable for automatic outlier detection methods that are used in this paper. Moreover, the questions with a predetermined list of options usually form a major part of questionnaires that takes most of the manual effort.

The incomplete questionnaires often form a huge part of data and their elimination is crucial to get any reasonable perspective on outliers. We generally work with two datasets that are used for comparison:

- **Unfiltered data** — the original dataset after the whole preprocessing except to the points (7. and 8.)
- **Filtered data** — it is the unfiltered data after the elimination of questionnaires with zero times (7.) and more than 70% of missing values (8.)

The preprocessed data can be represented by a matrix $\mathbf{A} = \left[a_{ij}\right] \in \mathbb{R}^{n\times m} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_m\right]$ of questionnaires, where m is the number of questions, n is the number of respondents and $\mathbf{a}_j \in \mathbb{R}^{n\times 1}$ is a column of matrix $\mathbf{A}$ (single variable). For some outliers detection methods, the matrix $\mathbf{A}$ needs to be normalized by columns to the range $\langle 0, 1\rangle$. Let $min(\mathbf{a}_j)$ and $max(\mathbf{a}_j)$ be the minimum and maximum value of a column $\mathbf{a}_j$, the matrix normalized by columns $\tilde{\mathbf{A}} = \left[\tilde{a}_{ij}\right] \in \mathbb{R}^{n\times m}$ is computed as

$$\tilde{\mathbf{A}} = \left[\frac{\mathbf{a}_1 - min(\mathbf{a}_1)}{max(\mathbf{a}_1) - min(\mathbf{a}_1)}, \quad \cdots \quad , \quad \frac{\mathbf{a}_m - min(\mathbf{a}_m)}{max(\mathbf{a}_m) - min(\mathbf{a}_m)}\right]. \tag{1}$$

3.2. Unsupervised outlier scores

Once the questionnaires are preprocessed, the outlier detection algorithms are applied. We identified several common issues that should be detected in questionnaire data and can increase suspicion of outlierness. The issues include predictable patterns of responses, self-contradictory responses, or inconsistency with data distribution. As these defects cannot be revealed by a single approach, multiple outlier detection algorithms are performed, namely, time score, the kth order empirical entropy, correlation-based methods, probability score, Mahalanobis distance, local outlier factor (LOF), and the k-nearest neighbors (kNN).

Each method assigns an outlier score $\omega_i \in \langle 0, 1\rangle$ to each ith questionnaire of the data matrix $\mathbf{A}$ such that $\omega_i = 1$ means the maximal outlier score and $\omega_i = 0$ means a maximally standard questionnaire. Despite the dimension of the problem (number of questions), each outlier score is represented by one variable. If an outlier score is within a different range of values it has to be normalized. All the methods are unsupervised and work without any a priori knowledge. They can process questionnaires with an arbitrary number of questions and a various number of available options per question.

The tested methods are described in detail in Section 4.

3.3. Ensemble outliers detection

Given the outlier scores computed by selected methods, the final set of outliers has to be detected by combining all of them. The standard approaches define an ensemble based on a collective outlier score such as average scoring, or maximum scoring (Aggarwal & Sathe, 2015) which are also tested in the experimental section. These ensembles mostly expect a positive correlation between outlier scores which is typical for methods based on the distance and the nearest neighbors. In our case, the methods are based on different principles and scores do not always correlate positively. The distributions of scores are highly unbalanced and they have positive skewness. Despite the unified scores, it happens that a questionnaire with a very high entropy score has a very low correlation score. The entropy assesses the predictable patterns while correlation-based methods assess if the responses are consistent with detected dependence between questions. This problem will be discussed in detail in Section 5. Using any sum of scores leads to a very average distribution of ensemble score where it is difficult to detect reasonable outliers. Therefore, we avoid computing an ensemble score and we propose a method based on the union of outliers detected by individual algorithms.

The crucial task is to define a procedure that decides which questionnaires will be selected as outliers. Traditional algorithms usually choose K-worst observations or those having scores over some threshold (Aggarwal & Sathe, 2015; Kriegel et al., 2011; Schubert et al., 2012). Both methods are tricky in unsupervised outlier detection as the parameters are user-defined (estimated) and they do not represent the score distribution well. A standard statistical method for outliers identification is the box-plot method (Grubbs, 1969; Tukey et al., 1977). It uses the interquartile range (IQR) which is the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1) of the outlier score. The correct data are expected between the lower and upper whiskers of the box-plot which are computed as $low = Q_1 - 1.5 \cdot IQR$ and $up = Q_3 + 1.5 \cdot IQR$. Everything outside this range is said to be an outlier. The box-plot is applicable for univariate data with approximately normal distribution. The outlier score computed by any of our methods is a single variable; however, its distribution is hardly normal. Our experiments showed that practically all the scores tested in this article have a distribution with positive skewness. Therefore, the *adjusted box-plot (ABP)* method (Hubert & Vandervieren, 2008) is preferred here which is the box-plot method modified for skewed distributions. It was optimized to simulate distributions such as Γ, χ^2, F, Parento, or Lognormal distributions by properly parameterized exponential function based on skewness. The *adjusted box-plot rule* and its lower and upper whiskers $\langle low, up\rangle$ are defined as follows:

$$\begin{gathered}\textbf{if } MC \geq 0 \textbf{ then } \langle Q_1 - 1.5 \cdot IQR \cdot e^{-4MC}, Q_3 + 1.5 \cdot IQR \cdot e^{3MC}\rangle \\ \textbf{if } MC < 0 \textbf{ then } \langle Q_1 - 1.5 \cdot IQR \cdot e^{-3MC}, Q_3 + 1.5 \cdot IQR \cdot e^{4MC}\rangle \\ \textbf{where} \\ MC = med_{x_i \leq Q_2 \leq x_j} h(x_i, x_j) \textbf{ and } h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}.\end{gathered} \tag{2}$$

The MC represents the medcouple of the distribution which is used for skewness computation. The non-negative MC means the right positive skewness and the negative one means the left negative skewness. The Q_2 represents a median of the outlier score. The right skewness means that the distribution has the right tail longer with most of the data concentrated to the left. As the right tail represents the highest outlier scores of the most suspicious entries, we apply only the upper limit up to cut off the worst outliers. The lower limit low is not applied because the low outlier scores represent the standard entries. Therefore, the adjusted box-plot rule allows us to identify the outliers much more precisely in the right tail than the standard box-plot rule.

This rule is applied to all outlier scores computed by selected outlier detection methods. The only exception is the LOF where we use its own methodology based on the outlier factor. It means that each method identifies its own set of outliers. To incorporate multiple methods into a robust ensemble model, a union of individual sets of outliers is computed:

$$O = \bigcup_{m=1}^{l} \{i \in \{1, \ldots, n\}, | \, \omega_{m,i} > up_m\}, \tag{3}$$

where O is a set of outliers indices, l is the number of computed outlier scores (methods), n is the number of questionnaires, $\omega_{m,i}$ is the score of the mth method and the ith questionnaire and up_m is the upper limit of the mth method for outlier detection identified by the adjusted box-plot rule.

4. Outliers detection methods

There are several common perspectives on how to define an outlier in questionnaire data:

(1) **Predictable patterns:** Some cheating respondents often fill the questionnaires using the same answer choice over and over again or they use some repetitive pattern of answers, such as “a,b,c,d”. These answers are also worthless for further exploration.
(2) **Self-contradictory responses:** Some questions can be related or logically dependent. Questionnaire designers even place several similar questions into questionnaire forms to check if a respondent understands the formulation of questions. Self-contradictory answers reveal a lack of understanding or cheating.
(3) **Other inconsistent responses:** Although the data is filtered using all the previous approaches, some inconsistent respondents can persist. These questionnaires have to be assessed by a data analyst. They can be wrong from the perspective of the dataset distribution, or some can represent unusual but accurate replies. Nevertheless, the detection of outlying observations is generally desired.

(4) **Response time:** Although the fastest respondents are excluded at the beginning, the response time can be also used as an outlier score.
(5) **Graphic visualization:** A graphic interpretation of numerical or categorical attributes can reveal some outlying observations that do not fit the data distribution. This is good for visual verification of the data cleaning process.

The methods were selected and designed to be unsupervised, i.e. they are functional without any a priori knowledge of data. Some of them are well-known methods and we modified some others to better adapt them for questionnaire data. A method assigns an outlier score $\omega_i \in \langle 0, 1\rangle$ to each ith questionnaire of the data matrix $\mathbf{A}$ such that $\omega_i = 1$ means the maximal outlier score and $\omega_i = 0$ means a maximally standard questionnaire. If a method naturally returns a different range of values, the scores have to be normalized. The following subsections define the methods in detail. They aim to different aspects of data to provide various perspectives on outliers. We combine methods later in Section 5 to achieve a robust model for searching outliers in questionnaire data.

4.1. Proposed methods addressing issues in questionnaires

This section describes our unsupervised outlier detection methods for questionnaire data based on entropy, correlation of attributes, and probability of answers.

4.1.1. The kth order empirical entropy

Roughly speaking, entropy is a measure of uncertainty and it is often used in text compression algorithms. It shows whether the text contains some repetitive patterns of symbols and this states the lower limit of the compressed representation of data. We propose to compute the entropy to find outliers within a set of questionnaires. As the questionnaire data mostly consists of categorical data, each questionnaire can be transformed into a sequence of answers. It is supposed that cheating respondents try to fill the questionnaires with some predictable patterns such as "a, a, a", "b, b, b", "a, b, c" etc. These patterns lead to very low entropy of the sequence, and thus, it is utilized to detect the suspicious answers for outliers detection. It also detects the questionnaires with many empty answers.

The empirical entropy (Beirlant, Dudewicz, Györfi, & Van der Meulen, 1997; Shannon, 1948) was defined for a discrete variable X with possible outcomes $x_1, \dots, x_l$ representing the symbols with occurrence probabilities $P(x_1), \dots, P(x_l)$. This is also called the zeroth-order empirical entropy and it is based only on the distribution of symbols in a finite string s. The kth order empirical entropy (Manzini, 1999) extends this principle to higher orders. The kth order empirical entropy $H_k(s)$ is a conditional entropy, where the random variable represents all strings of length k and it is computed as

$$H_k(s) = -\sum_{w\in\Sigma^k}\sum_{x\in\Sigma} P(wx)\log P(x|w) = -\sum_{w\in\Sigma^k}\sum_{x\in\Sigma}\frac{n_{wx}}{|s|}\log\frac{n_{wx}}{n_w}, \tag{4}$$

where w is a word consisting of k symbols, $P(wx)$ is occurrence probability of a word wx and $P(x|w)$ is conditional probability that a symbol x follows a word w in a string s. The right side of Eq. (4) represents the calculation of entropy using frequencies n_{wx} of the word wx and n_w of the word w in the string s having length $|s|$.

In the questionnaire data, the string s_i is a sequence $s_i = (a_{i1}, \dots, a_{im})$ of numbers representing the answers to questions of the ith respondent in the matrix A. Each question represented by answer $a_{ij} \in \{0, 1, \dots, l_j\}$ for $j = 1, \dots, m$ allows only a small finite set of l_j numerical values extended by 0 for empty answers, where it is possible that $l_j \neq l_y$ for $j \neq y$ and $j, y \in \langle 0, m\rangle$.

The implementation of $H_k(s)$ is based on the dictionary of unique subsequences of k answers $s_p = (a_p, \dots, a_{p+k-1})$ for $p = 1, \dots, m-k+1$ and their frequencies in the sequence s_i. A subsequence s_p represents a word w in (4) with frequency n_w. The same approach can be used to find subsequences of length $k+1$ for a word wx and frequency n_{wx}.

In this article, kth order empirical entropy is used to compute the kth order empirical entropy of answers for each questionnaire. The lower the entropy is the more predictable and suspicious answers are contained. The questionnaires with low entropy should be excluded from the dataset or detected as outliers, therefore the outlier score is $\omega_i = 1 - H_k(s_i)$. This is a row-based method which means that the outlier score ω_i is computed using just answers of one respondent.

Partial entropy of batteries of questions A typical type of question is a battery of questions that contains a block of questions having the same number of answers. It usually consists of related questions or measures some quality (e.g. 1 — best, 5 — worst), etc. As the battery questions are consolidated together, they usually bring respondents to cheat more often than other types of questions. Therefore, we created another type of entropy-based outliers detection which uses only the batteries. As a dataset consists of many answered questionnaires, it is possible to detect the specific set of options for each question. Then the consolidated ranges of questions with the same possible answers are supposed to belong to the same battery of questions (part).

Given a sequence of answers $s_i = (a_{i1}, \dots, a_{im})$ and a set of all the t battery intervals $B = \{b_1, \dots, b_t\}$, where $b_j = (l_j, r_j)$, $l_j < r_j$ and $l_j, r_j \in \langle 1, m\rangle$, then $e_j = H_k((a_{il_j}, \dots, a_{ir_j}))$ is the kth order entropy (4) of the battery interval b_j. The entropies of t battery intervals (parts) form a vector $\mathbf{e}_i = (e_{i1}, \dots, e_{it})$, whose Euclidean norm $\|\mathbf{e}_i\|$ sets the final outlier score $\omega_i = 1 - \|\mathbf{e}_i\|$ of the given sequence s_i.

This score is computed for all questionnaires and used to detect the outliers. The k should be set to a reasonably small number ($k = 1$ or $k = 2$) because the size of batteries is relatively small.

4.1.2. Correlations

Another family of methods we propose is based on the assumption that there are some dependencies between questions. Outliers are detected by answers that do not suit the dependencies.

Given the normalized matrix $\tilde{\mathbf{A}}$ (1), its correlation matrix $\mathbf{R} = corr(\tilde{\mathbf{A}})$ is computed, where the function $corr$ computes the correlation matrix using the Spearman's correlation coefficient (Sahoo, 2015). Note that the missing answers are omitted for the correlation computation to avoid fake dependencies between questions with many missing answers. The vacancies are filled with zeros for further computing. The matrix $\mathbf{R}$ captures the linear dependencies between all attributes represented by answers to questions. The correlation matrix $\mathbf{R}$ is utilized to detect the significant correlations that reveal some related questions. It is supposed that a strong violation of correlations is suspicious and such questionnaires should be detected as outliers.

Let an $\mathbf{R} = \left[r_{kl}\right] \in \mathbb{R}^{m\times m}$ be a correlation matrix of $\tilde{\mathbf{A}} = \left[\tilde{a}_{ij}\right] \in \mathbb{R}^{n\times m}$. Let a $corrMin$ be a threshold of minimal significant correlation and $dist(\tilde{a}_{ik}, \tilde{a}_{il})$ be a distance for values of two attributes. A τ is a number of the greatest distances that are used to calculate a score for each questionnaire. The basic algorithm can be written as follows:

(1) Set $corrMin$, τ and define function $dist : \mathbb{R}^2 \to \mathbb{R}$.
(2) Find all significant correlations: $S = \{(k, l)\,;\, abs(r_{kl}) \geq corrMin \text{ for } k < l\}$.
(3) For each questionnaire i from $\tilde{\mathbf{A}}$ compute a set of distances between correlated attributes: $D_i = \{d; \forall (k, l) \in S : d = dist(\tilde{a}_{ik}, \tilde{a}_{il})$ **if** $\tilde{a}_{ik} \neq 0 \wedge \tilde{a}_{il} \neq 0$ **else** $0\}$.
(4) Sort each set D_i in descending order.
(5) Compute the score ω_i of a questionnaire i as average of the first τ distances: $\omega_i = \sum_{j=1}^{\tau} D_{ij}/\tau$.
(6) Normalize the scores ω_i for $i = 1, \dots, n$ to the range $\langle 0, 1\rangle$ if necessary.

The score ω_i of the ith questionnaire is represented by its τ worst violations of the assumed dependencies of attributes. The distance function $dist$ can be defined by different types of metrics. The metrics

can be expressed as a distance between a pair of values of the correlated attributes or as a distance between the values and the distribution of the correlated attributes. We tested two basic methods: linear regression (Sahoo, 2015; Witten & Frank, 2002; Zaki, Meira Jr, & Meira, 2014) and difference as they are defined in the following subsections.

Linear regression As the correlation reveals a linear dependency between attributes, it is supposed that its course can be expressed by linear function estimated by linear regression (Sahoo, 2015; Witten & Frank, 2002). The metric is computed as an orthogonal distance between point and line. Given the two correlated columns $\tilde{\mathbf{a}}_k$ and $\tilde{\mathbf{a}}_l$ of the normalized matrix $\tilde{\mathbf{A}}$, the linear regression computes the intercept b_0 and the slope b_1 parameters for the slope-intercept form of the linear equation $y = b_0 + b_1 x$ which can be expressed in the implicit form as $b_0 + b_1 x - y = 0$. The distance from the line for the columns k and l is computed

$$Regress(\tilde{a}_{ik}, \tilde{a}_{il}) = \frac{|b_0 + b_1 \tilde{a}_{ik} - \tilde{a}_{il}|}{\sqrt{b_1^2 + 1}}. \tag{5}$$

Difference The simplest distance can be computed by the difference of attributes values. Given the correlation matrix $\mathbf{R}$ of the normalized matrix $\tilde{\mathbf{A}}$, the distance between the columns k and l for the ith row of $\tilde{\mathbf{A}}$ is computed as

$$Diff(\tilde{a}_{ik}, \tilde{a}_{il}) = \begin{cases} abs(1 - \tilde{a}_{ik} - \tilde{a}_{il}) : & r_{kl} < 0 \\ abs(\tilde{a}_{ik} - \tilde{a}_{il}) : & r_{kl} \geq 0 \end{cases} \tag{6}$$

The difference in Eq. (6) reflects both a positive and negative correlation coefficient. It is a cheaper estimate of linear regression.

4.1.3. Probability score

Another approach is a question-based outlier score presented by Zijlstra et al. (2007) which counts an individual's frequency of unpopular answers. The point is that the frequent answers are not suspicious while many improbable answers per a questionnaire are more suspicious. This method is called $O+$ and it judges the improbable answers by each question/attribute separately.

In short, the probabilities of existing options within a question are computed so that the most probable option has the lowest score $O_{ij} = 0$ and the least probable one has the highest score $O_{ij} = o_j - 1$ for the ith respondent and the jth question having o_j existing options. The final outlier score of one questionnaire is a sum of the ranks of the contained answers. The greater the score is, the more improbable answers are included.

Formally, given the matrix of questionnaires $\mathbf{A}$ and a probability function $P(X_j)$ for $j = 1, \ldots, m$ representing the relative frequency of each option of the jth attribute, then the outlier score O_{ij} is determined using the rank number of $P(X_j = a_{ij})$ denoted $rank\left[P(X_j = a_{ij})\right] \in \{1, \ldots, o_j\}$, such that

$$O_{ij} = o_j - rank\left[P(X_j = a_{ij})\right], \tag{7}$$

where o_j is the number of options of the jth question. The total outlier score O_i+ of the ith questionnaire is

$$O_i+ = \sum_{j=1}^{m} O_{ij}. \tag{8}$$

Zijlstra et al. (2007) assumes that all the m questions have the same number of existing options. However, this condition is not met very often. Thus, we expect various o_j for different j which means that the ranking has to be normalized to get the ranks to the comparable level. Otherwise, the questions with more options would have a greater weight in the total score. The sum in (8) is reformulated as

$$O_i+ = \sum_{j=1}^{m} O_{ij}/(o_j - 1). \tag{9}$$

Although there is some bias due to different o_j, all the m questions are contained in every questionnaire of $\mathbf{A}$, and therefore, the total outlier scores $\omega_i = O_i+$ should be comparable.

Note that the empty answers should be skipped by $O+$ method as they can easily become the most frequent in some questions. This would make the empty answers the most probable and the rest highly suspicious.

4.2. Widespread methods

The selected general outliers detection methods based on response time and distance are briefly summarized here:

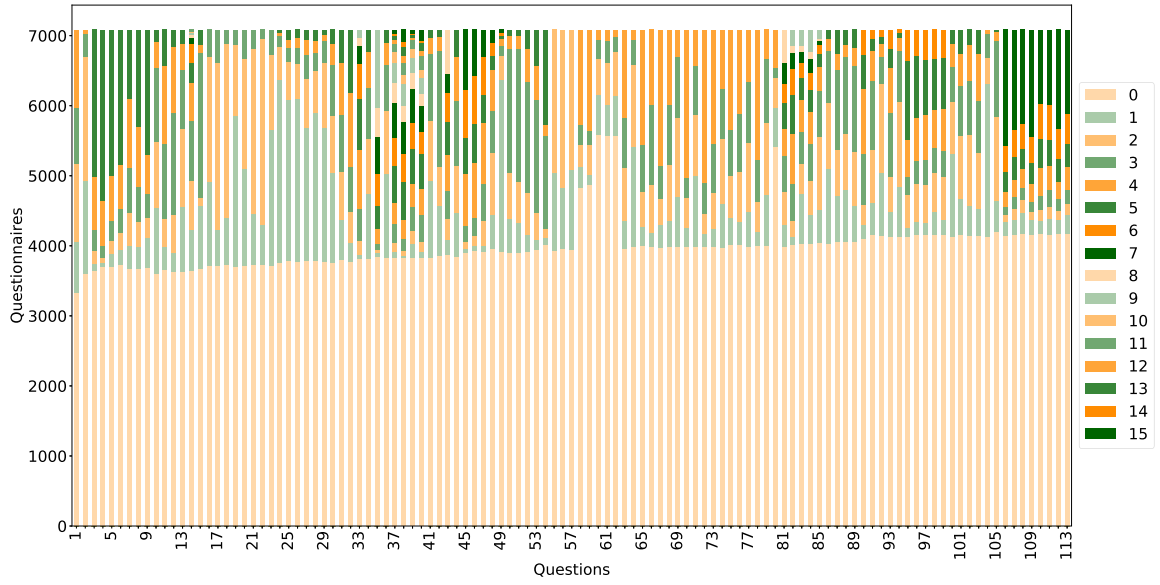
- **Time**: The simplest method is to compute a total time (if available) to reveal how much attention a respondent paid to questions. A very short time probably indicates fake or incomplete answers. The normalized time is used as an outlier score.
- **Mahalanobis distance** (Leys et al., 2018; Mahalanobis, 1936) is a non-parametric method that measures the distance of a point to the center of the dataset using the covariance structure of the data in the multidimensional space.
- **Local outlier factor (LOF)** (Breunig et al., 2000) is a density-based method using a single parameter $MinPts$, which is the number of the nearest neighbors used to define the local neighborhood. Unlike global methods that define a fixed neighborhood for all points, the LOF is able to judge the outliers from the local perspective. The basic idea is that the density around a point is compared with the density around its neighbors. The assumption is that the density around an outlier is considerably different than the density around its neighbors. The outlier factor around 1 means a similar density as neighborhood, a factor smaller than 1 represents an inlier, and a factor greater than 1 is an outlier. Unlike in other methods, we do not provide the outlier score as a single random variable but we apply this LOF rule. The LOF is stable for $MinPts$ between 10 and several hundred (Breunig et al., 2000) depending on the data and the size of clusters. The $MinPts$ should not be greater than the size of the smallest distinguishable cluster. As the LOF uses a spherical neighborhood based on the Euclidean distance, it is inaccurate in high dimensions (curse of dimensionality Zaki et al., 2014). Moreover, the dimensionality increases the time complexity. Therefore, we apply the Principal component analysis (PCA) (Witten & Frank, 2002; Zaki et al., 2014) to reduce the dimension before the LOF application.
- k-**Nearest Neighbors** (k**NN**) (Chandola et al., 2007; Chen et al., 2010) The kNN simply computes the distances to the k closest objects and uses their mean directly as an outlier score. Unlike in LOF, the outliers are identified globally as the scores are compared across the whole dataset.

5. Case study: HBSC data

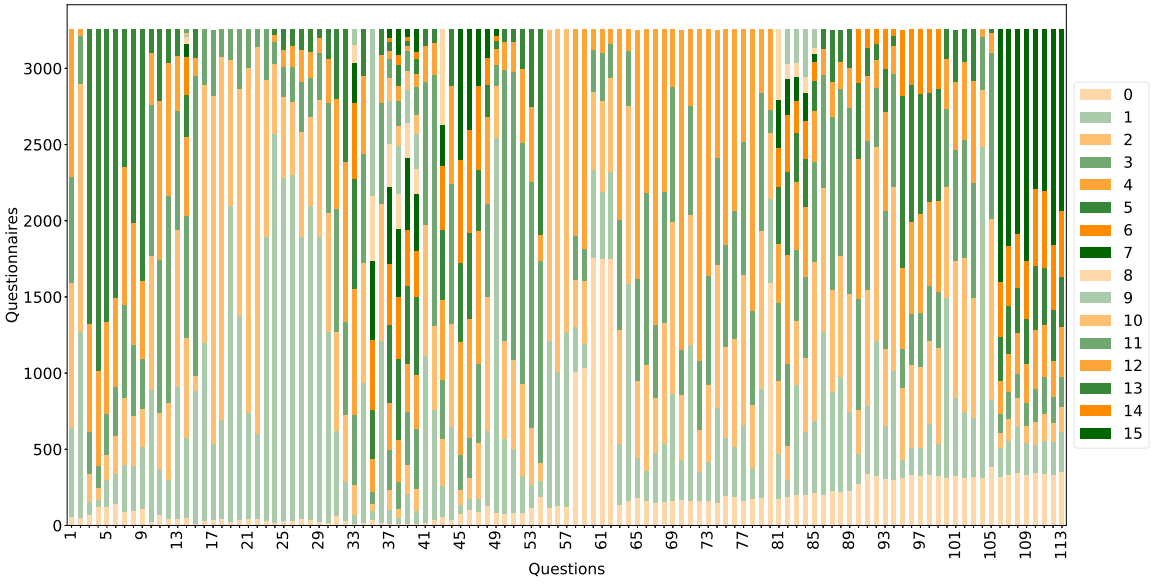
This section evaluates our proposed procedure including questionnaire data cleaning, transformation, and outlier detection tested on a real-world dataset. We bring a case study showing the impacts of the model on the HBSC 2020 data. We got the data available in uncleaned form with all the issues named in this article. Most of the questionnaire survey data is available only in cleaned form, thus we cannot sufficiently test our methods on them. However, the model can be applied to any uncleaned questionnaire data where the outlying and faulty observations are still included.

5.1. HBSC data

The data for the present article comes from a project directly linked to the Health Behavior in School-aged Children (HBSC) study. The HBSC is an international World Health Organization-collaborative questionnaire-based survey, which focuses on health and health behaviors of 11-, 13-, and 15-year-olds in the context of their social environments (family, peers, and school). The design of the HBSC study



(a) Unfiltered data



(b) Filtered data

Fig. 2. Frequency of distinct answers for all 113 questions in HBSC 2020 data represented by stacked bar plots. The number of options per question varies between 3 and 16.

enables not only to check cross-national comparisons (Badura et al., 2021) but also to monitor trends in the health and health behaviors of young people during their transition from childhood to adulthood, as a crucial period for long-term establishment of such behaviors. The data file used in the present article was drawn from a survey conducted in the Czech Republic in June 2020, i.e. during the final stage of the first wave of COVID-19 epidemic. The aim of this data collection was to assess the impact of the lockdown, especially in terms of school closure and ban of sport or other leisure-time activities, on the adolescents' lives, time use and health behaviors (Ng et al., 2021). Overall, 141 schools from 14 administrative regions of the Czech Republic took part of the survey which contains 7082 unique entries gathered between 1st and 30th June 2020.

5.2. Preprocessing of HBSC 2020

The effect of cleaning and preprocessing (Section 3.1) is briefly introduced here to illustrate its importance. The original HBSC 2020 dataset contains 7082 respondents and 232 variables. The variables include the questions, personal data, system/browser information, time data, and some other auxiliary variables. After the application of the cleaning procedure, the dataset was reduced to 113 multiple choice questions with a single answer. The number of possible options varies between 3 and 16 per question. Fig. 2 shows the distribution of answers per each question. Fig. 2(a) represents the unfiltered dataset with all 7082 respondents and Fig. 2(b) represents the filtered data after the elimination of questionnaires with more than 70% of missing values (3255 respondents). The plots show that 54% of the unfiltered dataset is the questionnaires with a huge amount of missing values (>70%).

Table 1
Tested methods and their parameters and types of outliers they reveal.

Name	Description	Outlier types	Parameters
en$k$	kth order empirical entropy	Predictable patterns, missing values	$k \in \{1,2,3,4\}$
pen$k$	Partial en$k$ of batteries	Predictable patterns, missing values	$k \in \{1,2\}$
Maha	Mahalanobis distance	Inconsistent responses	–
$O+$	Probability score	Inconsistent responses	–
dcτ	Correlation (difference)	Self-contradictory responses	$corrMin = 0.7,\ \tau \in \{3,6,12\}$
rcτ	Correlation (linear regression)	Self-contradictory responses	$corrMin = 0.7,\ \tau \in \{3,6,12\}$
LOFd	Local outlier factor	Inconsistent responses	PCA dims $d \in \{3,10\}$, $MinPts = 100$
kNN100	k-Nearest Neighbors	Inconsistent responses	$k = 100$

The missing values are represented by option 0. The difference between Figs. 2(a) and 2(b) is enormous and it would not be possible to read any reasonable information from the data without proper cleaning and outliers elimination. The questionnaire data is heavily disrupted by missing values, human errors, and false and inconsistent answers.

5.3. Model for outliers detection and its parameters

This section describes the utilized outlier detection ensemble that is experimentally verified in the next sections on HBSC 2020 data. Both the unfiltered (7082 respondents) and filtered (3255 respondents) datasets with 113 variables are tested for comparison.

The tested methods and their parameters are summarized in Table 1. It contains all the methods defined previously in Section 4 with parameters that were selected experimentally for the HBSC 2020 dataset. Four different orders of entropy are tested because the entropy for any higher order is too low (score is too high) on 113 variables. Only two orders of entropy are tested with the partial entropy because the batteries of questions are even smaller (say about 10 questions). The data often contains many incomplete questionnaires that were not filled to the end by respondents in time. A large block of missing answers leads to very low entropy and high suspiciousness. However, even incomplete questionnaires with reasonable answers in the opening can be valuable according to our experience. Thus, we judge by the entropy only the initial part of a questionnaire that is filled. The missing answers at the end are simply cut off and those in the middle are preserved. The Mahalanobis distance and $O+$ are nonparametric. In the case of correlation-based methods, we set $corrMin = 0.7$ according to the correlation table which produces 15 significant correlations for both filtered and unfiltered data. The number of the worst distances between correlated variables τ is set to 3, 6, and 12 to test which τ is the most representative. The LOF method is based on a spherical neighborhood which limits its applicability to high-dimensional data. As the data has 113 variables we reduce them to 3 and 10 dimensions by the PCA. According to the discussion in Section 4.2, the $MinPts$ is set to 100 as the data has several thousands of respondents. Small $MinPts$ leads to many outliers detected from the perspective of dense small clusters. Similarly, the kNN method is set to $k = 100$ which corresponds to the LOF method. After several experiments for $k \in \{50, 80, 100\}$ we found out that there is no significant difference, and therefore, $k = 100$ is good for our experiments.

5.4. Experiments

The model described in the previous section is applied to the HBSC 2020 data. This section contains our experiments and the discussion about the properties of outlier detection methods and the effect of data cleaning. Of course, not every method returns the best-expected outliers. Thus, the experiments guided us to select the appropriate combination of methods and parameters for our final ensemble model detecting the outliers in HBSC 2020 data.

5.4.1. Outlier scores distribution

All the variants of methods and parameters described in Table 1 were computed. First, we analyze the distributions of outlier scores and similarities between different methods. The methods with bad distribution are excluded from the model.

Fig. 3 shows standard box-plots of different outlier scores for filtered and unfiltered data. Generally, box-plots with most of the values placed in the smaller half are expected. As the scores assign the highest values to the most inconsistent questionnaires, we expect to find outliers above the upper whisker. Another comparison can be seen in Fig. 4 which illustrates how the scores correlate and how this correlation differs between filtered and unfiltered data. We use Spearman's correlation coefficient which compares the ranks of questionnaires and is more robust in noisy data.

The box-plots of unfiltered data show that there are certain groups of methods with similar logic there. The correlation-based methods correspond better to the distance-based methods and the probability score and they all go against the entropy-based methods and time. This shows that the methods really detect different types of outliers and they complement each other. The same effect can be seen in the correlation matrix in Fig. 4 where there are positive correlations within the named groups and negative correlations between methods from different groups.

The box-plots show that the scores computed on unfiltered data have a terrible distribution. They are strongly affected by incomplete questionnaires which leads to extremely low or extremely high outlier scores. This is because the incomplete records represent a majority of data which deforms the statistical indicators so much that the standard data looks inconsistent. Detection of outliers based on unfiltered data is a random selection of questionnaires.

The distribution of outlier scores computed on the filtered data looks reasonable and it suits the expectation that most of the questionnaires are rated by low scores while the outliers have high scores. The only methods that do not suit this expectation very well are *en3*, *en4* and *time*. The correlation matrix of the filtered data also confirms that the *en3* and *en4* behave on the contrary to other entropies. It seems that a subword of length 3 or 4 is too long to get any reasonable entropy on 113 questions. Therefore, we exclude *en3* and *en4* from the model.

In both experiments, the time factor is included, however; the differences between responding times were so huge that most of the questionnaires lead to very high scores considering the long times in contrast to the short ones. Therefore, we exclude only the questionnaires with zero times during the preprocessing which strongly correspond to empty questionnaires. The time factor is not used in our model for outliers detection.

The method with the least variance of scores is the LOF because it judges the outliers from the local perspective based on the density of the current cluster. It has its own technique for outliers identification (Section 4.2). The kNN also explores the neighborhood of each object/questionnaire but its scores are judged globally in comparison with all the other questionnaires using the ABP rule.

It can be seen that the difference between correlation-based methods with different τ is minimal. Therefore, we use only *dc12* and *rc12* to detect outliers as they have smoother distribution than the scores for lower τ.

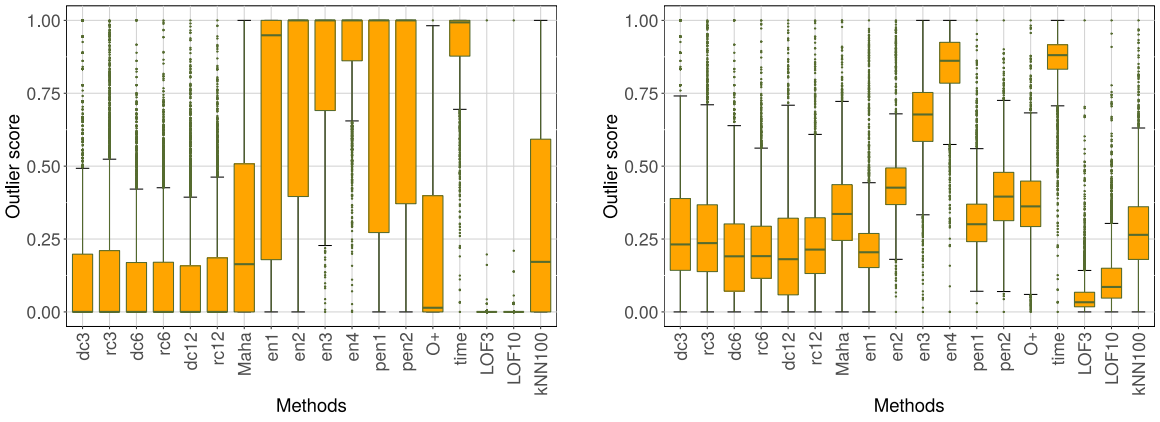


Fig. 3. Comparison between box-plots for outliers detection methods using the unfiltered (left) and filtered (right) HBSC 2020 data.

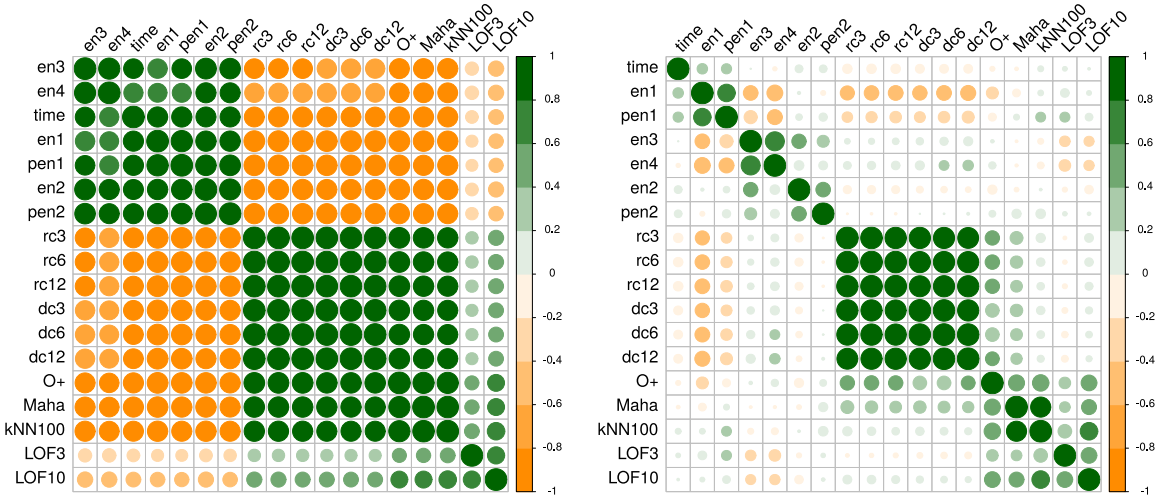


Fig. 4. Comparison between Spearman correlations for methods of outliers detection on the unfiltered (left) and filtered (right) HBSC 2020 data.

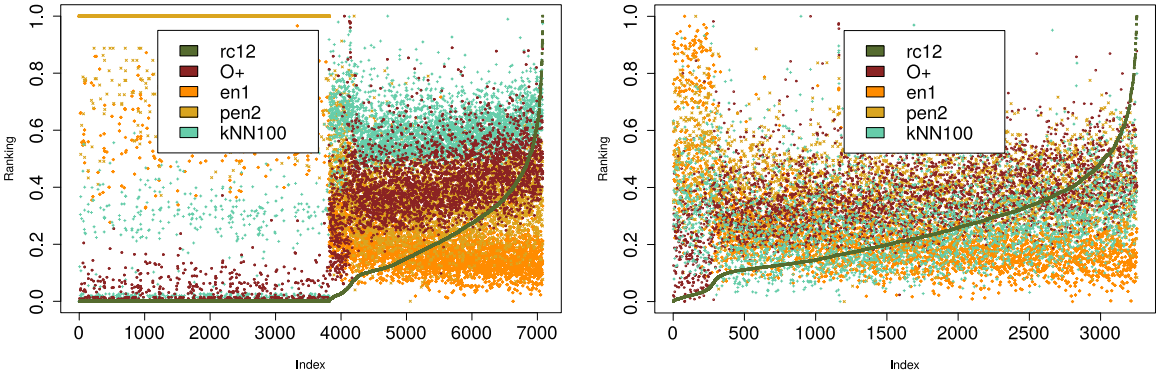


Fig. 5. Comparison among some scoring methods for outliers detection on the unfiltered (left) and filtered (right) HBSC 2020 data.

Fig. 5 presents the distribution of some selected scores on unfiltered and filtered data. The charts are sorted according to the *rc12* score. The figure shows again that the entropy goes against the other scores. The *rc12* gives a very low score to incomplete questionnaires as the difference between correlated attributes is almost zero. Other distance-based methods behave similarly. The *LOF10* is very low for most of the questionnaires in the unfiltered data as there are too many incomplete entries which strongly affect the local density factor. However, the entropy of empty questionnaires is very low, so that, their score is almost one. The *O+* method is closer to *rc12* but it detects almost no outliers in unfiltered data (Fig. 3) because it is based on the frequency of answers. As the empty answers are very frequent, they strongly shift the *O+* distribution to low numbers. Even though we skip the empty answers, their values are missing in the final sum in contrast to the properly filled questionnaires.

In conclusion, the set of outliers (denoted O_{All}) consists of the outliers detected by *en1, en2, pen1, pen2, Maha, O+, dc12, rc12, LOF3, LOF10* and *kNN100* methods. The ensemble is computed using the

Table 2
Intersection table showing the numbers of outliers found by multiple methods. The diagonal represents the numbers found by particular methods. The last row represents unique outliers undetected by other methods.

	en1	en2	pen1	pen2	Maha	*O*+	dc12	rc12	LOF3	LOF10	kNN100
en1	**73**	19	50	14	4	12	0	0	17	9	16
en2	19	**62**	21	16	1	0	0	0	4	2	2
pen1	50	21	**74**	20	4	9	0	0	11	5	12
pen2	14	16	20	**40**	2	3	0	0	6	3	5
Maha	4	0	1	2	**20**	5	0	3	1	3	12
O+	12	0	9	3	5	**28**	0	1	14	5	12
dc12	0	0	0	0	0	0	**2**	2	0	0	0
rc12	0	0	0	0	3	1	2	**19**	0	0	2
LOF3	17	4	11	6	1	14	0	0	**65**	17	11
LOF10	9	2	5	3	3	5	0	0	17	**30**	8
kNN100	16	2	12	5	12	12	0	2	11	8	**28**
Unique	9	23	10	9	5	7	0	14	33	9	2

adjusted box-plot rule and the union of the detected outliers as it was described in Section 3.3.

5.4.2. Outliers detection and their influence

Once the data is preprocessed and the scores are computed, it is necessary to identify the right outliers in the filtered data and assess their influence. We do not want to exclude less or too many questionnaires if it does not bring any benefit.

Table 2 summarizes the numbers of outliers detected by each method, the intersection of found outliers between pairs of methods, and also the numbers of unique outliers found by each method. This table helps us to assess the similarity and significance based on the specific outliers. It shows that most of the outliers are detected by entropy-based methods and *LOF3*. There is also a large intersection between them which suggests that it is not necessary to use all of them.

Next, we test the statistical influence of the excluded outliers based on the separate and ensemble methods. The statistics are summarized in Table 3. We use two statistics: Cronbach's alpha coefficient and the variance of the total score $A+$. The point is to find out if the complete dataset has different statistics from the dataset without the excluded outliers. If the results of the statistical analysis are different the outliers are considered influential. Although each question can have a different number of possible options, the normalization of columns transforms the options onto the values in $\langle 0, 1\rangle$ range. As each questionnaire contains the same questions, the total sum of values per questionnaire can be a representative value reducing the whole data to one variable.

Definition 1 (*Total Score*)**.** Given the matrix $\tilde{\mathbf{A}} = \left[\tilde{a}_{ij}\right] \in \mathbb{R}^{n\times m} = \left[\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_m\right]$ normalized by attributes, the total score $A+$ is computed as

$$A+ = \sum_{j=1}^{m} \tilde{\mathbf{a}}_j.$$

The Cronbach's alpha coefficient (Cronbach, 1951) is a common test score reliability coefficient. It computes the covariance between each pair of attributes and its change after exclusion of outliers signalizes the type and significance of outliers.

Definition 2 (*Cronbach's Alpha*)**.** Let the $\tilde{\mathbf{A}} = \left[\tilde{a}_{ij}\right] \in \mathbb{R}^{n\times m} = \left[\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_m\right]$ be a data matrix normalized by attributes. Let $Cov(\tilde{\mathbf{a}}_j, \tilde{\mathbf{a}}_k)$ denote the sample covariance between the values on attributes j and k, and let σ^2_{A+} be the sample variance of total score $A+$, then

$$\alpha = \frac{m}{m-1} \cdot \frac{\sum\sum_{j\neq k} Cov(\tilde{\mathbf{a}}_j, \tilde{\mathbf{a}}_k)}{\sigma^2_{A+}}.$$

The greater the $\alpha \in \langle 0, 1\rangle$ is, the stronger is the covariance between attributes. The variance of the total score σ^2_{A+} shows how compact the dataset is. The significant reduction of the variance means that the excluded questionnaires are far from the rest of the data. The outlying observations increase the variance of the total score.

Once the statistics are computed, it is necessary to examine if the difference is significant or if it is comparable with a random exclusion of questionnaires. Let the K be the number of outliers found by a method, the procedure is as follows: (1) Do 1000 random omitting of K questionnaires from the original data. (2) Compute a statistic for each of the 1000 versions of data without K randomly excluded questionnaires. (3) The 1000 statistics represent a random variable. Using its distribution, the 2.5th and the 97.5th percentiles are determined. (4) The null hypothesis is that the influence of the K outliers is the same as the influence of any K randomly deleted questionnaires, which means that its statistic is within the range of the 2.5th and the 97.5th percentiles of the distribution. Otherwise, the null hypothesis is rejected and the omission of our K outliers is statistically significant for a significance level equal to 5%.

Tables 2 and 3 shows several interesting observations. The *dc12* finds only two outliers, it is insignificant and it is completely included in the *rc12* method. The *LOF3* method identifies 65 outliers and 33 of them are unique. However, these outliers are not statistically significant. It seems that the PCA reduction to only 3 dimensions is too drastic, a lot of information was lost and the outliers detected with *LOF3* are irrelevant. From the entropy-based methods, we select the *en1* and *pen2*. The *en1* has many common outliers with *pen1* and the *pen1* is insignificant. The *pen2* is added as the better variant of the two partial entropy methods. Table 3 shows that the distance-based methods such as *Maha* or *LOF10* influence the variance σ^2_{A+} more than the α, while the correlation (*rc12*) or probability methods (*O*+) influence more the α. The rest of the methods (especially entropies) influence both significantly. It is a logical consequence of the utilized methods. Omission of questionnaires corrupting the correlation of attributes or questionnaires with too many unusual answers has to increase the covariance between attributes. The omission of distant observations automatically reduces the variance. The entropies reduce both because they omit questionnaires with many empty answers and predictable patterns which are both distant and the patterns also falsely intensify the covariance between attributes.

After considering these experiments, we decided to select the following methods for the final ensemble (denoted O_{Fin}): *en1, pen2, Maha, O+, rc12, LOF10* and *kNN100*. Table 3 also compares the O_{All} with O_{Fin}. Both are highly significant and strongly reduce the statistics. However, while the O_{Fin} deletes only 163 outliers (5.01% of questionnaires) the O_{All} deletes 235 outliers (7.22% of questionnaires). The point of the outliers detection methods is to preserve the maximum of the original data and delete only the worst cases that negatively affect the statistical analysis. Table 3 also contains a comparison with simple widespread ensembles: maximum (M) and average (A) scoring. For each questionnaire, the maximum/average score of final methods (Fin) is computed and the final ensemble score is assessed. First, the questionnaires with the K-greatest ensemble scores are selected as outliers, where the K equals the number of outliers found by the union ensemble O_{Fin}. Therefore, M_{Fin}/A_{Fin} represent 163-greatest

Table 3
Statistical significance of outliers detected by different outlier scores, their union (O_{All}) and union of outliers detected by finally selected methods (O_{Fin}). The O_{Fin} is compared with M_{Fin}/A_{Fin} representing the same number of outliers detected using the maximum/average scoring, and the $\overline{M}_{Fin}/\overline{A}_{Fin}$ representing outliers detected by the ABP using maximum/average scoring. **Columns**: K — number of outliers detected by ABP method; $K\%$ — percentage of suspected observations; $\alpha_{2.5p^{th}}$, $\alpha_{97.5p^{th}}$, $\alpha(K)$, $s_{\alpha(K)}$ — Cronbach's alpha statistics, and $\sigma^2_{2.5p^{th}}$, $\sigma^2_{97.5p^{th}}$, $\sigma^2(K)$, $s_{\sigma^2(K)}$ — variance of the total score σ^2_{A+} (lower/upper percentiles of 1000 random eliminations, value for the K found outliers, and significance); ++: significant value increase over random omission, + : insignificant value increase over random omission, −− : significant value decrease over random omission, -: insignificant value decrease over random omission.

Methods	K	$K\%$	$\alpha_{2.5p^{th}}$	$\alpha_{97.5p^{th}}$	$\alpha(K)$	$s_{\alpha(K)}$	$\sigma^2_{2.5p^{th}}$	$\sigma^2_{97.5p^{th}}$	$\sigma^2(K)$	$s_{\sigma^2(K)}$
en1	73	2.24	.90383	.90656	.88297	−−	77.74972	80.09752	62.66853	−−
en2	62	1.91	.90366	.90639	.87281	−−	77.59456	79.94232	58.03576	−−
pen1	74	2.27	.90375	.90658	.90404	–	77.65353	80.09816	77.79473	–
pen2	40	1.23	.90413	.90612	.89301	−−	77.97641	79.68803	69.18034	−−
Maha	20	0.61	.90443	.90578	.90455	–	78.23596	79.39965	77.80248	−−
$O+$	28	0.86	.90422	.90591	.90669	++	78.05610	79.52942	79.21280	+
dc12	2	0.06	.90491	.90541	.90539	+	78.64709	79.06095	79.05215	+
rc12	19	0.58	.90449	.90576	.90595	++	78.27367	79.38607	79.33506	+
LOF3	65	2.00	.90368	.90639	.90587	+	77.63919	79.95919	78.22690	–
LOF10	30	0.92	.90426	.90594	.90493	–	78.08243	79.55177	78.02416	−−
kNN100	28	0.86	.90428	.90592	.90364	−−	78.12332	79.52470	76.59394	−−
Union-based K										
O_{All}	235	7.22	.90259	.90759	.82791	−−	76.62232	80.98984	41.08061	−−
O_{Fin}	163	5.01	.90294	.90728	.86620	−−	76.94837	80.76940	53.63949	−−
M_{Fin}	163	5.01	.90294	.90728	.83732	−−	76.94837	80.76940	44.08652	−−
A_{Fin}	163	5.01	.90294	.90728	.90552	+	76.94837	80.76940	75.88644	−−
Adjusted box-plot										
$\overline{M}_{Fin}$	44	1.35	.90406	.90617	.89296	−−	77.93390	79.75355	68.95358	−−
$\overline{A}_{Fin}$	33	1.01	.90420	.90600	.90518	–	78.02488	79.59790	77.78858	−−

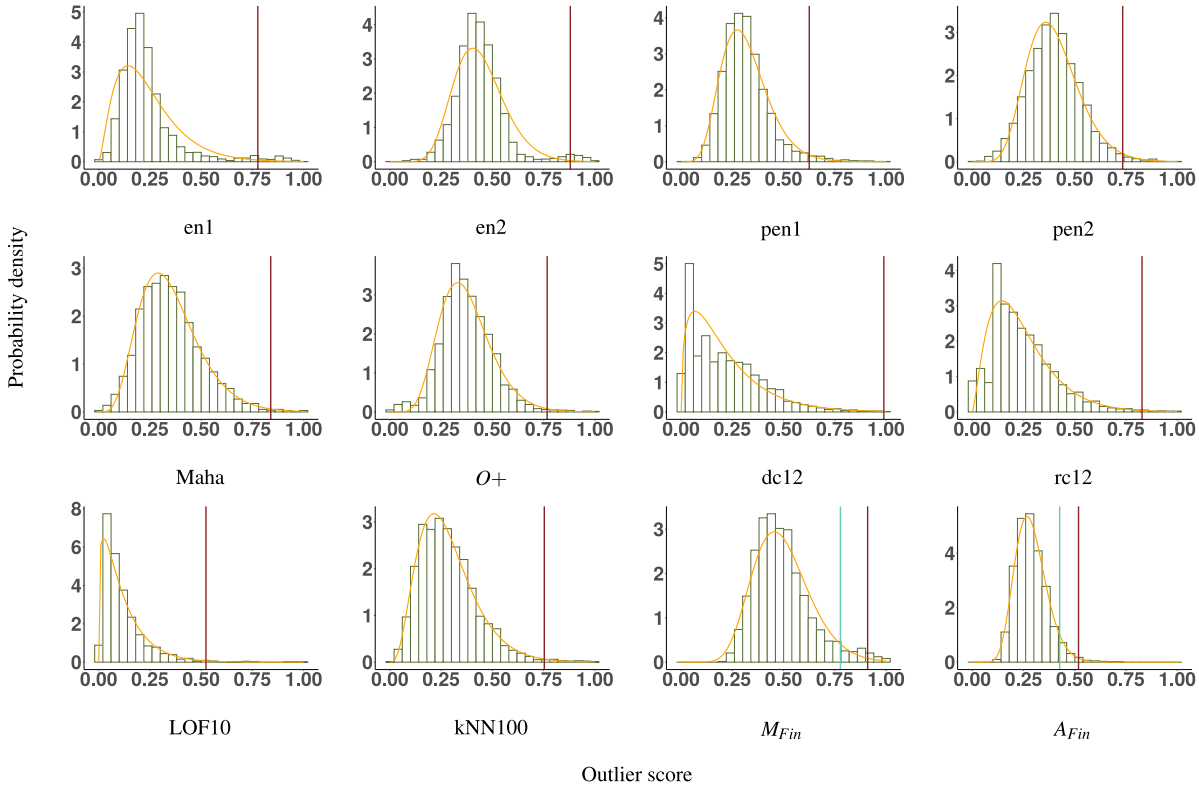


Fig. 6. Histograms and estimated gamma function for distributions of ranking methods on filtered HBSC 2020 data.

maximum/average scores. The table shows that the average scoring leads to weak and insignificant outliers while the maximum scoring returns even more significant outliers than the union ensemble. The problem here is how to define the K without knowledge of O_{Fin}. Next, we used the maximum/average scores and applied the adjusted box-plot (ABP) method to them. The K is individually detected for each ensemble. The outliers detected by ABP are designed $\overline{M}_{Fin}$ and $\overline{A}_{Fin}$. The results show that the ABP-based ensembles detect much fewer outliers than the union ensembles and they are also less significant. The maximum scoring is performing much better than the average scoring again. The maximum and average scoring smooth the differences and specifics of the individual outlier detection methods. As some of the tested methods are unbalanced and strongly uncorrelated the smoothing leads to unrepresentative results biased by scores distributed within larger values. The union ensemble selects more outliers, but it also better reflects the individual methods, regardless of their distributions.

It can be seen in Fig. 6 which illustrates the distribution of the selected scores on the filtered data. All the distributions have positive

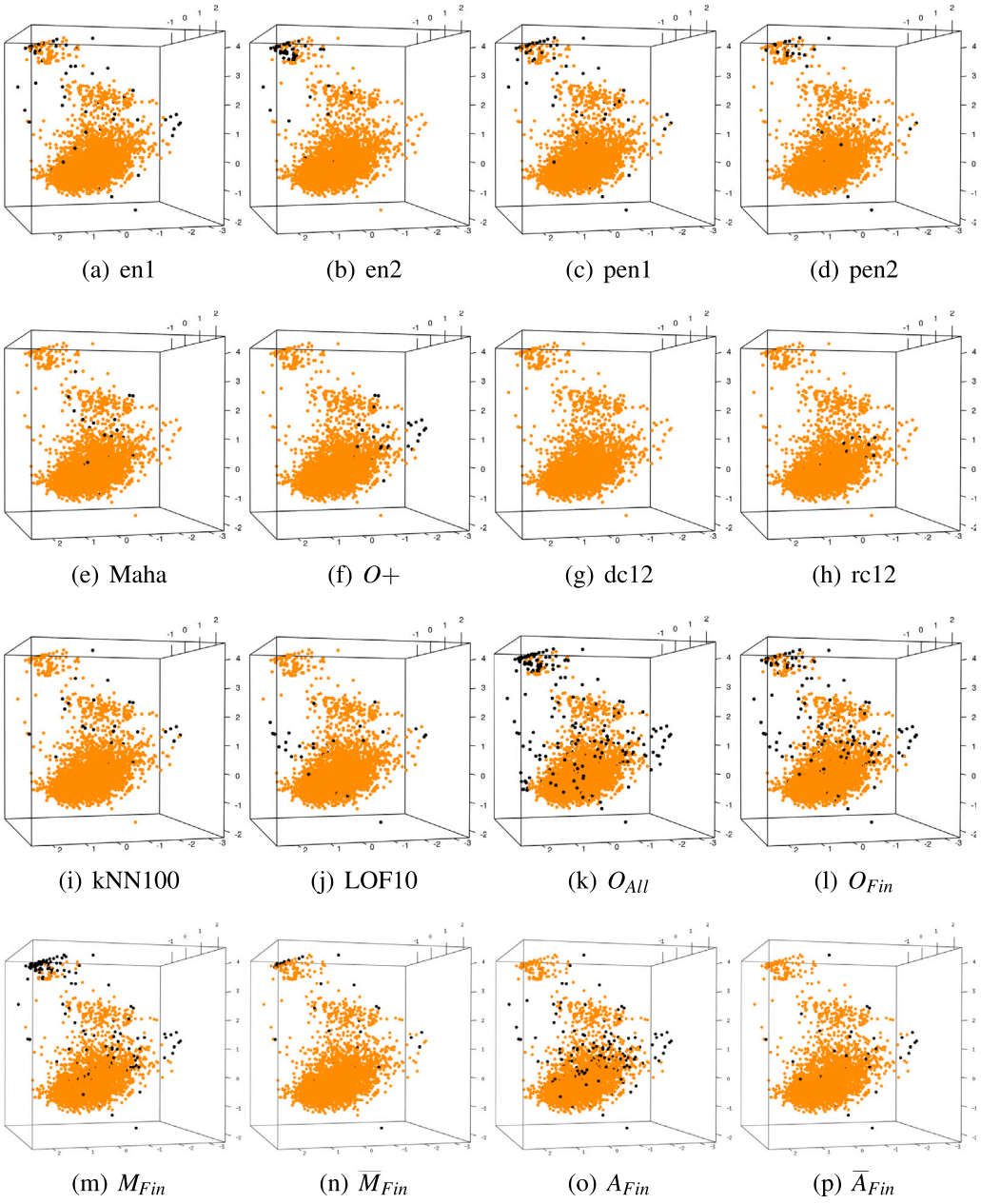


Fig. 7. Visualization of filtered data in 3D by PCA with highlighted outliers (black points) found by selected methods.

skewness. We also add a gamma function which was estimated based on the probability distribution of each score. The gamma function is typically used as a continuous model for distributions with positive skewness and it seems that it fits the scores very well. This also defends our decision to use the adjusted box-plot method instead of the simple box-plot to identify the outliers. Despite normalization onto interval $\langle 0, 1\rangle$, the scores are not directly comparable. The red line indicates the incision from which the outliers are cut off by the ABP. We also add the distributions of M_{Fin} and A_{Fin} and the green line representing the incision of the 163 greatest outliers. The separate application of ABP and the union of outliers secures the corresponding method significance and no prioritization due to higher scores in the distribution.

Fig. 7 illustrates the filtered HBSC 2020 data reduced to 3 dimensions by PCA and the outliers detected by the considered methods. The unions O_{All} and O_{Fin} and the ensembles M_{Fin}, $\overline{M}_{Fin}$, A_{Fin}, and $\overline{A}_{Fin}$ are also visualized. Although the 3D visualizations reduced from the original 113 attributes are strongly lossy, they illustrate the main differences between approaches. The O_{Fin} deletes mostly the points outside the main cluster while the O_{All} deletes some more points inside the main cluster as well. The identification of outliers is an unclear task that includes the specific knowledge of the data and various perspectives. The outliers do not have any exact definition. However, experiments show that our final set of methods covers the most required perspectives on questionnaire data.

5.4.3. Computation time analysis

Table 4 shows execution times of separate outlier detection methods that are summarized in Table 3 and measured on the filtered HBSC2020 data. The table also contains the preprocessing time which is common for all methods and it includes loading the data file, cleaning procedure (missing values, renumbering, normalization), computation of correlation matrix, means and variances of individual variables. The preprocessed statistics are utilized in separate outliers detection methods. The ensembles O_{All} and O_{Fin} are illustrated by the total times here which consist of the measured methods and the preprocessing. The shortest times are reached by simple methods such as kNN, Maha

Table 4
Execution times in seconds (average and standard deviation of 10 measurements) of preprocessing (common for all methods), each standalone method and total times of ensembles. Measured for filtered HBSC2020 data with 3255 respondents.

Method/time (s)	Avg.	St.dev.
Preprocessing	47.6000	2.5123
en1	2.0953	0.1625
en2	2.8078	0.1956
pen1	7.6344	0.5288
pen2	8.1000	0.5557
Maha	0.7234	0.0748
O+	0.2000	0.0161
dc12	2.6703	0.2075
rc12	2.6703	0.2067
kNN100	0.5391	0.0552
LOF3	2.8453	0.3167
LOF10	2.9219	0.0612
O_{All}	80.8078	4.1382
O_{Fin}	64.8500	3.4523

and *O*+. The times of other more sophisticated methods are comparable (< 3 s), only the partial entropy computation takes about 8 s. However, the methods are very different and based on different parameters, perspectives and statistical properties of data, e.g. entropies are row-based while correlations or *O*+ are column-based, kNN, LOF or Maha are proximity-based etc. The execution times only illustrate the relative comparison between methods that form together the final ensemble method. It is clear that the O_{Fin} consisting only of the finally selected methods is faster than the O_{All} calculating with all of them. This article examines the ensemble of diverse outlier detection methods and their effect to data properties. A deep analysis of complexities of standalone methods is beyond the scope of this article.

The program is implemented in Python 3.10.4, Pandas 1.4.1, and Scikit-learn 1.0.2. All experiments run on the following hardware: Intel Core i7-1185G7 @ 3.0 GHz, 32 GB RAM, Windows 10 64-bit

6. Conclusion

This article introduces a robust framework for ensemble outliers detection in raw multivariate questionnaire data. The proposed methods are tested on the HBSC 2020 data focused on the health behaviors of children in the context of social environments. It demonstrates the importance of preprocessing and cleaning data before using methods for outliers detection. We describe the specific outlier types in questionnaires assessed from different perspectives (e.g. incomplete, faked, and self-contradictory responses, or predictable patterns), and we propose various methods addressing them. It is advisable to use multiple methods to form an ensemble computing a final union of the selected statistically significant sets of detected outliers. The article explains how methods differ from the others and which of them correlate. The whole process of outliers detection is standardized and automated with minimal parameterization. The framework is able to work with the skewed or non-normal distribution of data and it uses the dependencies between variables to detect self-contradictory responses violating the detected correlations. We also propose two approaches using the *k*th order entropy for predictable patterns detection. It also works with questionnaires containing a different number of options per question.

The experimental part of the article illustrates the properties of the HBSC 2020 data and our proposed procedure for data cleaning and outliers detection. We recommend excluding 3827 out of 7082 questionnaires during the cleaning phase and 163 outliers suggested by the final set of selected methods. The case study exhaustively examines the data, the influence of the proposed operations and justifies the selection of methods and their parameters. The suggested 163 outliers represent a subset of 5.01% of questionnaires in the prefiltered dataset which is a reasonable portion of data. The experiments also show that some methods detect many suspicious questionnaires but they are not statistically significant, and therefore, they are preserved. The charts (Fig. 4) and the intersection table (Table 2) also verify that the methods return different outliers and some of them have a negative correlation. Therefore, the union of separately detected outliers is preferred over the average and maximum scoring ensembles.

CRediT authorship contribution statement

Vojtěch Uher: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Pavla Dráždilová:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Jan Platoš:** Funding acquisition, Supervision, Validation. **Petr Badura:** Data curation, Funding acquisition, Validation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data statement

Data subject to third party restrictions. The data that support the findings of this study are available from the Faculty of Physical Culture, Palacký University Olomouc. Restrictions apply to the availability of these data, which were used under license for this study. Data are available upon request at http://hbsc.cz/lockdown2020/ with the permission of the Faculty of Physical Culture, Palacký University Olomouc.

Acknowledgments

The work was supported by the Czech Science Foundation [20-25019S], by the Ministry of Education, Youth and Sports, Inter-Excellence, Czech Republic [LTT18020], by the European Regional Development Fund-Project "Effective Use of Social Research Studies for Practice" [No. CZ.02.1.01/0.0/0.0/16_025/0007294], and by SGS, VSB – Technical University of Ostrava, Czech Republic, under the grant "Parallel processing of Big Data IX" [No. SGS2022/12].

References

Aggarwal, C. C., & Sathe, S. (2015). Theoretical foundations and algorithms for outlier ensembles. *SIGKDD Explorations Newsletter*, *17*(1), 24–47.

Aggarwal, C. C., Zhao, Y., & Philip, S. Y. (2011). Outlier detection in graph streams. In *2011 IEEE 27th International conference on data engineering* (pp. 399–409). IEEE.

Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, *60*, 708–713.

Akoglu, L., Tong, H., Vreeken, J., & Faloutsos, C. (2012). Fast and reliable anomaly detection in categorical data. In *Proceedings of the 21st ACM international conference on information and knowledge management*, (pp. 415–424).

Badura, P., Hamrik, Z., Dierckens, M., Gobiņa, I., Malinowska-Cieślik, M., Furstova, J., et al. (2021). After the bell: adolescents' organised leisure-time activities and well-being in the context of social and socioeconomic inequalities. *Journal of Epidemiology and Community Health*.

Beirlant, J., Dudewicz, E. J., Györfi, L., & Van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, *6*(1), 17–39.

Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook* (pp. 131–146). Springer.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104).

Cabana, E., Lillo, R. E., & Laniado, H. (2019). Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators. *Statistical Papers*, 1–27.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., et al. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, *30*(4), 891–927.

Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. *ACM Computing Surveys*, *14*, 15.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58.

Chen, Y., Miao, D., & Zhang, H. (2010). Neighborhood outlier detection. *Expert Systems with Applications*, *37*(12), 8745–8749.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Das, K., Schneider, J., & Neill, D. B. (2008). Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 169–176).

Deng, X., & Wang, L. (2018). Modified kernel principal component analysis using double-weighted local outlier factor and its application to nonlinear process monitoring. *ISA Transactions*, *72*, 218–228.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security* (pp. 77–101). Springer.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining, vol. 72*. Springer.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1–21.

Har-Shemesh, O., Quax, R., Lansing, J. S., & Sloot, P. M. (2020). Questionnaire data analysis using information geometry. *Scientific Reports*, *10*(1), 1–9.

Harpe, S. E. (2015). How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850.

Hawkins, D. M. (1980). *Identification of outliers, vol. 11*. Springer.

He, Z., Deng, S., Xu, X., & Huang, J. Z. (2006). A fast greedy algorithm for outlier mining. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 567–576). Springer.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186–5201.

Ienco, D., Pensa, R. G., & Meo, R. (2016). A semisupervised approach to the detection and characterization of outliers in categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(5), 1017–1029.

Inchley, J., Currie, D., Cosma, A., & Samdal, O. (2018). Health behaviour in school-aged children (HBSC) study protocol: Background, methodology and mandatory items for the 2017/18 survey. *International Report*.

Jebreel, N. M., Haffar, R., Singh, A. K., Sánchez, D., Domingo-Ferrer, J., & Blanco-Justicia, A. (2020). Detecting bad answers in survey data through unsupervised machine learning. In *International conference on privacy in statistical databases* (pp. 309–320). Springer.

Jiang, F., Liu, G., Du, J., & Sui, Y. (2016). Initialization of K-modes clustering using outlier detection techniques. *Information Sciences*, *332*, 167–183.

Jiang, F., Zhao, H., Du, J., Xue, Y., & Peng, Y. (2019). Outlier detection based on approximation accuracy entropy. *International Journal of Machine Learning and Cybernetics*, *10*(9), 2483–2499.

Kandanaarachchi, S. (2021). Unsupervised anomaly detection ensembles using item response theory. URL https://arxiv.org/abs/2106.06243.

Kieu, T., Yang, B., & Jensen, C. S. (2018). Outlier detection for multidimensional time series using deep neural networks. In *2018 19th IEEE international conference on mobile data management (MDM)* (pp. 125–134). IEEE.

Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *VLDB, vol. 98* (pp. 392–403). Citeseer.

Kriegel, H.-P., Kroger, P., Schubert, E., & Zimek, A. (2011). Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM international conference on data mining* (pp. 13–24). SIAM.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156.

Li, S., Lee, R., & Lang, S.-D. (2007). Mining distance-based outliers from categorical data. In *Seventh IEEE international conference on data mining workshops (ICDMW 2007)* (pp. 225–230). IEEE.

Mahalanobis, P. C. (1936). *On the generalized distance in statistics*. National Institute of Science of India.

Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. In *2017 Third international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB)* (pp. 255–258). IEEE.

Manzini, G. (1999). An analysis of the burrows-wheeler transform. In *SODA '99, Proceedings of the tenth annual ACM-SIAM symposium on discrete algorithms* (pp. 669–677). USA: Society for Industrial and Applied Mathematics.

Naseer, S., Saleem, Y., Khalid, S., Bashir, M. K., Han, J., Iqbal, M. M., et al. (2018). Enhanced network anomaly detection based on deep neural networks. *IEEE Access*, *6*, 48231–48246.

Ng, K., Cosma, A., Svacina, K., Boniel-Nissim, M., & Badura, P. (2021). Czech adolescents' remote school and health experiences during the spring 2020 COVID-19 lockdown. *Preventive Medicine Reports*, *22*.

Pacheco, A. G., Ali, A.-R., & Trappenberg, T. (2019). Skin cancer detection based on deep learning and entropy to detect outlier samples. arXiv preprint arXiv: 1909.04525.

Ramachandran, V., & Kishorebabu, V. (2019). A tri-state filter for the removal of salt and pepper noise in mammogram images. *Journal of Medical Systems*, *43*(2), 1–10.

Sahoo, P. (2015). *Probability and mathematical statistics*.

Sakurai, R., Ueki, M., Makino, S., Hozawa, A., Kuriyama, S., Takai-Igarashi, T., et al. (2019). Outlier detection for questionnaire data in biobanks. *International Journal of Epidemiology*, *48*(4), 1305–1315.

Sari, A., et al. (2015). A review of anomaly detection systems in cloud networks and survey of cloud security measures in cloud storage applications. *Journal of Information Security*, *6*(02), 142.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.

Schubert, E., Wojdanowski, R., Zimek, A., & Kriegel, H.-P. (2012). On evaluation of outlier rankings and outlier scores. In *Proceedings of the 2012 SIAM international conference on data mining* (pp. 1047–1058). SIAM.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shao, C., Zheng, S., Gu, C., Hu, Y., & Qin, X. (2022). A novel outlier detection method for monitoring data in dam engineering. *Expert Systems with Applications*, *193*, Article 116476.

Taha, A., & Hadi, A. S. (2019). Anomaly detection methods for categorical data: A review. *ACM Computing Surveys*, *52*(2), 1–35.

Tukey, J. W., et al. (1977). *Exploratory data analysis, vol. 2*. Reading, Mass..

Uher, V., Gajdoš, P., & Snášel, V. (2018). Proposal of effective orthogonal and hexagonal hierarchical structures for disc queries. In *2018 3rd International Conference on Control, Robotics and Cybernetics (CRC)* (pp. 20–26). IEEE.

Uher, V., Gajdoš, P., Snášel, V., Lai, Y.-C., & Radeckỳ, M. (2019). Hierarchical hexagonal clustering and indexing. *Symmetry*, *11*(6), 731.

Utkin, L. V. (2014). A framework for imprecise robust one-class classification models. *International Journal of Machine Learning and Cybernetics*, *5*(3), 379–393.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, *2*(10), Article e267.

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964–108000.

Wang, X., Wang, X. L., & Wilkes, D. M. (2012). A minimum spanning tree-inspired clustering-based outlier detection technique. In *Industrial conference on data mining* (pp. 209–223). Springer.

Waure, C. d., Poscia, A., Virdis, A., Pietro, M. L. D., & Ricciardi, W. (2015). Study population, questionnaire, data management and sample description. *Annali Dell'Istituto Superiore Di Sanita*, *51*, 96–98.

Wilcox, R. R. (2019). Robust regression: Testing global hypotheses about the slopes when there is multicollinearity or heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 355–369.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

Xie, Z., Li, X., Wu, W., & Zhang, X. (2016). An improved outlier detection algorithm to medical insurance. In *International conference on intelligent data engineering and automated learning* (pp. 436–445). Springer.

Yu, J. X., Qian, W., Lu, H., & Zhou, A. (2006). Finding centric local outliers in categorical/numerical spaces. *Knowledge and Information Systems*, *9*(3), 309–338.

Yuan, K.-H., & Gomer, B. (2021). An overview of applied robust methods. *British Journal of Mathematical and Statistical Psychology*, *74*(S1), 199–246.

Yuan, Z., Zhang, X., & Feng, S. (2018). Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures. *Expert Systems with Applications*, *112*, 243–257.

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials*, *12*(2), 159–170.

Zhao, X., Liang, J., & Cao, F. (2014). A simple and effective outlier detection algorithm for categorical data. *International Journal of Machine Learning and Cybernetics*, *5*(3), 469–477.

Zhu, J., Ge, Z., Song, Z., & Gao, F. (2018). Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control*, *46*, 107–133.

Zhu, Y., Hernandez, L. M., Mueller, P., Dong, Y., & Forman, M. R. (2013). Data acquisition and preprocessing in studies on humans: what is not taught in statistics classes? *The American Statistician*, *67*(4), 235–241.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, *36*(2), 186–212.

Zijlstra, W. P., Van Der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, *42*(3), 531–555.

Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(5), 363–387.