An empirical comparison of some outlier detection methods with longitudinal data

Marcello D'Orazio[1]

Abstract

This paper investigates the problem of detecting outliers in longitudinal data. It compares well-known methods in official statistics with some proposals from the data mining or machine learning field, which are based on the distance between observations or binary partitioning trees. The comparison is done by applying them to panel survey data related to different types of statistical units. Traditional methods are relatively simple and enable the direct identification of potential outliers; however, they require specific assumptions. Recent methods provide just a score whose magnitude is directly related to the chance of having an outlier. All the methods require setting a number of tuning parameters; however, the most recent methods show higher flexibility and are sometimes more effective than traditional ones. Additionally, these methods can be applied in the multidimensional case.

Keywords: Boxplot, isolation forest, nearest neighbour distance, outlier detection, panel data.

DOI: https://doi.org/10.1481/ISTATRIVISTASTATISTICAUFFICIALE_1-2-3.2023.06.

[1] Marcello D’Orazio (madorazi@istat.it), Italian National Institute of Statistics - Istat.

The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The author would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

Data collected in sample surveys as well as data in administrative registers often contain errors that, if not corrected, may affect the accuracy of the final estimates. For these reasons, National Statistical Institutes (NSIs) have always invested considerable resources in verifying the incoming data to detect actual or potential errors. This process is known as *data editing* (or *statistical data editing;* de Waal *et al.*, 2011; sometimes also referred to as *input data validation*) and aims also at identifying *missing values* that are then replaced by *imputed values*[2]; the *imputation* process also permits replacing the values identified as erroneous (deleted and imputed). The editing and imputation sub-process can make use of a variety of statistical methods depending on many factors: the type of data source; the data collection mode; the nature of variables (continuous, categorical, or mixed-type) and the relationship existing between them; the nature of errors and their potential impact on the final estimates, etc.

This paper focusses on the subset of data editing methods tailored to *outlier detection*; "an outlier is a data value that lies in the tail of the statistical distribution of a set of data values" (UNECE, 2000)[3]. The underlying idea is that "outliers in the distribution of uncorrected (raw) data are more likely to be incorrect". For instance, when observing a single continuous variable (household income, firm production, harvested area in a farm, etc.), an outlier can be the outcome of a *measurement error*, *i.e*. the observed value is not the true value (and the true value is not expected to be in the tail of the distribution). An outlier can also be a non-erroneous "extreme" value that, although it has a significant influence on the final estimates, may deserve a "special" treatment in the analysis[4].

This article examines traditional and recent approaches to outlier detection in longitudinal data, where a continuous variable is observed on the same set

[2] "Data editing and imputation" is a sub-process in the "Process" phase of the Generic Statistical Business Process Model (GSBPM; https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1).

[3] "An outlier is an observation which is not fitted well by a model for the majority of the data. For instance, an outlier may lie in the tail of the statistical distribution or 'far away from the centre' of the data" (Memobust Glossary; https://ec.europa.eu/eurostat/cros/content/glossary_en).

[4] In sample surveys, in theory, it is possible to distinguish between a *representative outlier* (*i.e*. a unit in the sample that represents other units in the target population that have similar values) and a *non-representative outlier* (when in the target population no other units are showing similar values; https://ec.europa.eu/eurostat/cros/content/glossary_en).

of units at different time points. In the case of panel surveys, official statistics can observe households, firms, or agricultural holdings. The following Section briefly describes well-known outlier detection methods. It illustrates some nonparametric approaches suggested in the field of data mining or machine learning, which have great potential when applied to this specific setting and, more generally, in official statistics. Section 3 compares the outcomes of the application of the reviewed methods to panel data related to different types of statistical units, often investigated in surveys conducted by NSIs. Finally, Section 4 drafts some concluding remarks and future areas of work.

2. Some outlier detection methods for longitudinal data

When a given quantitative non-negative variable *Y* is observed repeatedly over time on the same set of units, we can expect a high correlation between subsequent measurements; this feature represents useful information in setting up an efficient outlier detection procedure. It becomes even more relevant when the objective is to estimate the change over time of a population parameter related to *Y*.

Formally, let $(y_{t1}, y_{t2}, \dots, y_{tn})$ be the values of *Y* observed at time *t* on a set of *n* units, being $y_{ti} \geq 0$; the ratio $r_i = y_{t_2 i}/y_{t_1 i}$ denotes the "individual change" from time t_1 to time t_2 for unit *i*, being $t_1 > t_2$. Data editing literature suggests various methods to check whether the individual change (r_i) is too large or too low (*e.g*. the theme "Editing for Longitudinal Data" in Eurostat, 2014); a very popular one was suggested by Hidiroglou and Berthelot (1986), and its characteristics are summarised in the following 2.1.

2.1 Hidiroglou-Berthelot method for outlier detection

Hidiroglou and Berthelot (1986) suggested examining the empirical distribution of the ratios $r_i = y_{t_2 i}/y_{t_1 i}$ $(i = 1,2,\dots,m;$ being $m \leq n,$ after discarding 0s and missing values, if any, in both $y_{t_2 i}$ and $y_{t_1 2}$). In particular, they first transform the ratios in the following manner:

$$s_i = \begin{cases} 1 - \frac{r_M}{r_i}, & \text{if} \quad 0 < r_i < r_M \\ \\ \frac{r_i}{r_M} - 1, & \text{if} \qquad r_i \geq r_M \end{cases} \tag{1}$$

such that s_i=0 when a ratio is equal to the median of ratios $(r_i = r_M)$; then, to account for the magnitude of data and give more "importance" to units involving high values of the *Y* variable, suggest to derive the following *score*:

$$E_i = s_i\left[\max\left(y_{t_1 i}, y_{t_2 i}\right)\right]^U \tag{2}$$

where *U* can range from 0 to 1 $(0 \leq U \leq 1)$ and controls the role of the magnitude in determining the importance associated with the centred ratios, a common choice consists of setting *U* = 0.5.

Identification of potential outliers relies on the assumption that the scores are approximately distributed according to a Gaussian distribution; in practice, the parameters of the Gaussian distribution are estimated using robust methods and units outside the interval:

$$[E_M - C \times d_{Q1}, E_M + C \times d_{Q3}] \tag{3}$$

are identified as potential outliers. In expression (3):

$$d_{Q1} = max\left(E_M - E_{Q1}, |A \times E_M|\right) \quad d_{Q3} = max\left(E_{Q3} - E_M, |A \times E_M|\right) \tag{4}$$

being $\boldsymbol{E_{Q1}}$, $\boldsymbol{E_M}$, $\boldsymbol{E_{Q3}}$ and the quartiles of the *E* scores. The constant *A* is a positive small quantity (suggested *A* = 0.05) introduced to overcome cases of $E_M = E_{Q3}$ or $E_M = E_{Q1}$ that may occur when the ratios are too concentrated around their median. The parameter *C* determines how far from the median the bounds should be; commonly suggested values are $C = 4$ or $C = 7$, but larger values can be considered, depending on the tails of the distribution of the *E* scores. In practice, the bounds (3) allow for a slight skewness in the distribution of the *E* scores.

Recently, Hidiroglou and Emond (2018) suggested replacing $\boldsymbol{E_{Q1}}$**and** $\boldsymbol{E_{Q3}}$ with, respectively, the percentiles $\boldsymbol{E_{P10}}$ **and** $\boldsymbol{E_{P90}}$ when a large proportion of units (>1/4) share the same value of the ratio, since in this case the "standard" method would detect too many observations as potential outliers.

Practically, the decision about the values of the "tuning" constants *U* and *C* is not straightforward and requires a graphical investigation of the distribution of scores, as well as different attempts with alternative values of both the constants. A helpful practical suggestion is to start inspecting the (suspicious) ratios by sorting them in decreasing order with respect to the absolute value of the score $(|E_i|)$. Hidiroglou and Emond (2018) also suggested an additional graphical inspection procedure.

In the R environment (R Core Team, 2022), the Hidiroglou Berthelot (HB) procedure is implemented by the function `HBmethod()` available in the package *univOutl* (D'Orazio, 2022), which also includes graphical facilities for inspecting the scores, in line with Hidiroglou and Emond's recommendations (2018). In addition, *univOutl* has facilities to identify outliers in univariate cases with methods based on robust location and scale estimates of the parameters of the Gaussian distribution.

2.2 Nonparametric methods

The nonparametric methods for outlier detection are very popular because they do not introduce an explicit assumption on the underlying distribution. For the sake of simplicity, when describing these methods, it is assumed that the problem of detecting outliers arises when a generic continuous variable *X* is observed on a set of *m* observations. The following Subsections describe some well-known approaches using boxplots, as well as recent proposals in the field of data mining and machine learning, particularly *distance and tree-based methods*.

2.2.1 Outlier detection with boxplots

Drawing a *boxplot* (*box-and-whisker* plot) is a popular approach to outlier detection:

$$[Q1_x - c \times IQR_x;\ \ Q3_x + c \times IQR_x] \tag{5}$$

where IQR_x is the inter-quartile range $(IQR_x = Q3_x - Q1_x)$ and, usually, $c = 1.5,\ 2 \text{ or } 3$; units outside the bounds (*whiskers*) are considered outliers.

To account for moderate skewness, Hubert and Vandervieren (2008) suggested an "adjusted" boxplot:

$$[Q1_x - 1.5 \times exp(aM) \times IQR_x;\ \ Q3_x + 1.5 \times exp(bM) \times IQR_x] \tag{6}$$

being *M* the *medcouple* measure of skewness ($-1 \leq M \leq 1$; Brys *et al*., 2004) that when greater than 0 indicates positive skewness and requires setting *a* = -4 and *b* = 3 in expression (6); on the contrary, with negative skewness (*M* < 0) it is suggested to set *a* = and *b* = 4. The authors claim that the adjusted boxplot fences in (6) work with moderate skewness, *i.e*. $-0.6 \leq M \leq 0.6$. Unfortunately, the adjusted boxplot permits only to set *c* = 1.5, and it is not possible to use alternative values.

D'Orazio (2022) implemented standard and adjusted boxplots in the function `boxB()` of the R package *univOutl*. Additionally, the function `HBmethod()` enables the application of the adjusted boxplot to the *E* scores.

2.2.2 Distance-based outlier detection

The idea of using distance measures in outlier detection is a direct consequence of the fact that we search for observations that are far from the centre of the data. In practice, distance-based outlier detection methods search for an observation that has very few other observations close to it; the fewer observations close to a unit, the higher the chance that it is an outlier. Knorr and Ng (1998) suggest identifying an outlier as an observation that has fewer than *k* observations at a distance less than or equal to a threshold δ. This approach requires deciding: (i) how to measure the distance, (ii) the distance threshold δ, and (iii) the *k* parameter. The first two choices are strictly related and relatively simple in the univariate setting (but not in the multidimensional case), as different distance functions may leave the set of nearest neighbours of a given unit unchanged.

Knorr and Ng's approach (1998) does not provide a ranking for the potential outliers. To overcome this difficulty, Ramaswamy *et al*. (2000) suggest identifying the potential outliers by calculating the *k* nearest neighbour (*k*-NN) distance; in practice, if $\boldsymbol{d_i^{(k)}}$is the distance of the ith from its *k*-nearest neighbour, the units showing the largest values of $\boldsymbol{d_i^{(k)}}$ are potential outliers. This simple approach can be computationally demanding in the presence of many observations and variables; however, some algorithms simplify the search (Hautamäki *et al*., 2004). The problem becomes much simpler in the univariate case, where the initial ordering of the units reduces the computational effort.

The *k*-NN distance proposed by Ramaswamy *et al*. (2000) is a very popular approach, and many variants are available. A well-known extension assigns to each unit a "weight" consisting of the sum of its distance from the corresponding *k* nearest neighbour observations (Angiulli and Pizzuti, 2002):

$$\boldsymbol{\omega_i^{(k)} = \sum_{j=1}^{k} d_i^{(j)}} \tag{7}$$

Hautamäki *et al.* (2004) suggest using the average. Common choices for the parameter *k* are 5 or 10; however, the literature does not provide a rule of thumb. Campos *et al.* (2016) note that the sum of distances makes the scores less sensitive to the value of *k*. Obviously, if *k* is too large, then the weight may become quite large even for non-outlying observations, since, as shown by expression (7), the final score will be the result of a sum of a larger number of terms.

In general, distance-based outlier detection methods are strictly related to outliers' detection based on kernel density estimation techniques; this paper will not address such methods, but it is worth noting that when *k*-NN is applied to density estimation problems, a possible rule of thumb consists in setting $k = ceiling(\sqrt{m})$ and, more in general, $k{\sim}m^{4/5}$.

The main drawback with *k*-NN methods is that they do not directly identify the potential outliers, like boxplots or the HB method; instead, they provide a summary score (distance or "weight") whose magnitude indicates the chance of being an outlier; the larger the score, the higher is the chance that a given observation is an outlier. To identify a possible threshold such that observations with a score above the threshold are identified as potential outliers, Hautamäki *et al*. (2004) suggest:

$$u_0 = \varepsilon \times max\left[u_{(i+1)}^{(k)} - u_{(i)}^{(k)}\right],\ i = 1,2,\dots,m-1 \tag{8}$$

where $u_i^{(k)} = \omega_i^{(k)}$ or $u_i^{(k)} = d_i^{(k)}$, $u_{(i+1)}^{(k)} \geq u_{(i)}^{(k)}$ and ε is a user-defined constant $0 < \varepsilon < 1$. This rule, introduces an additional parameter to set (ε); practically, a graphical inspection of the ordered scores $u_i^{(k)}$ can be more effective: once sorted them increasingly, good candidate thresholds the values corresponding to "jumps" in the plot (abnormal increase in the score).

The approach by Knorr and Ng (1998) is closely related to the DBSCAN (*Density-based spatial clustering applications with noise*) clustering algorithm (Ester *et al*., 1996), where the observations not "reachable" by any other observation are identified as *noisy* observations (outliers). The "reachability" depends on a distance threshold *δ*; in practice, two observations *i* and *j* are *directly reachable* if their distance is less than, or equal to, *δ* $(d_{i,j} \leq \delta)$. At the same time, they are only *reachable* if there is a path of three or more observations to go from *i* to *j*, where each couple of units in the path is directly reachable. The DBSCAN algorithm requires setting also a value for *g* that is needed to identify the *core observations*, *i.e.* observations that have at least $k = g - 1$ distinct units at a distance smaller than or equal to *δ*. To identify a value for *δ*, it is suggested to plot the *k*-NN distances in increasing order and set *δ* equal to the distance where the plot shows a jump.

In R, some distance-based methods for outlier detection are implemented in the package *DDoutlier* (Madsen, 2018), although *k*-NN distance is calculated in many other R packages; the package *dbscan* (Hahsler *et al*., 2019 and

2022) implements the DBSCAN clustering algorithm but has also facilities to calculate the *k*-NN distance efficiently.

2.2.3 Outlier detection with isolation forest

The *isolation forest* is an unsupervised decision-tree-based algorithm that consists of fitting an ensemble of *isolation trees* (Liu *et al*., 2008 and 2012). The underlying idea is that outlying observations have a higher chance of being separated from the others in one branch of the partitioning tree, with relatively few splits. In the univariate case an arbitrary threshold $\boldsymbol{x_o}$ is selected at random within the range of X $([\boldsymbol{min}(x_i), \boldsymbol{max}(x_i)])$ and all the observations are divided into two groups according to whether they show higher or lower values than $\boldsymbol{x_o}$. This randomised splitting process is applied recursively (*i.e.* divide the units into two groups then repeat the process in each group, and so on) until no further split is possible or until meeting some other criteria. The final outcome is an isolation tree where the more observations show similar *X* values, the longer (more splits) it will take to separate them into small groups (or alone) compared to less occurring *X* values; for this reason, the *isolation depth* (number of splits needed to isolate a unit) can be considered as a tool for detecting outliers.

Since a high variability would characterise the isolation depth estimated in a single isolation tree, its reduction can be achieved by building an ensemble of isolation trees – the isolation forest – and then deriving the final score by averaging over the fitted trees. As in random forests, each single isolation tree is fit on a bootstrap sample of $\boldsymbol{q}$ $(\boldsymbol{q} < \boldsymbol{m})$ observations randomly selected. In the Liu *et al*. proposal (2008 and 2012), the partitioning stops when a node has only one observation, or all units in a node have the same values (in some cases, it is also introduced a maximum value for the tree height, *e.g.* $\boldsymbol{l_{max} = ceiling(log_2(q))}$).

Formally, if $\boldsymbol{h(x_i)}$ is the *path-length or depths*, *i.e*. the number of splits to reach the *i*-th observation in a fitted tree, Liu *et al.* (2008, 2012) suggest to associate to each observation the following score:

$$u_i = 2^{-\frac{E[h(x_i)]}{c(q)}} \tag{9}$$

where $E[h(x_i)]$ is the average path length across the ensemble of the fitted trees and

$$c(q) = 2 \times H(q-1) - 2\frac{q-1}{q} \tag{10}$$

being $H(\cdot)$ the harmonic number. The Authors demonstrate that the resulting score ranges from 0 to 1 $(0 < u_i \leq 1)$ being a monotonic function of $h(x_i)$ and, in particular

- when $E[h(x_i)] \rightarrow q-1$ then $u_i \rightarrow 0$;
- when $E[h(x_i)] \rightarrow c(q)$ then $u_i \rightarrow 0.5$;
- when $E[h(x_i)] \rightarrow 0$ then $u_i \rightarrow 1$.

Practically, scores close to 1 indicate observations with a very short average path length that tend to be isolated earlier than the other ones and therefore denote outlying observations. As a consequence, setting a threshold score u_0 will return as outliers all the units having a score $u_i > u_0$. Generally, it is suggested to consider $u_0 > 0.5$, but a graphical inspection of the ordered scores can be beneficial in deciding u_0.

The isolation forest is very efficient and can handle multi-modal distributions. It requires setting two tuning parameters, the subsample size *q* and the number of trees to fit. In the first case, Liu *et al*. (2008, 2012) claim that even a small subsample size (*q = 256*) can work with massive datasets, while at least 100 trees should be fitted; this latter number should be increased when the achieved scores are on average quite below 0.5, as this may point out a problem of unreliable estimation of the average path length. It is worth noting that the standard method proposed by Liu *et al*. (2008, 2012) is also developed to handle outlier detection in a multidimensional framework, where the creation of each tree requires a recursive random selection of one of the available variables and its corresponding random splitting point. In the univariate case, with a single variable, there is just the random selection of the splitting point, and consequently, there is no need to grow a large number of trees. It is worth noting that, in the multivariate case, the standard algorithm (Liu *et al.*, 2008, 2012) essentially consists of an ensemble of results related to the application of the isolation forest independently to each variable. To compensate for this drawback, Hariri *et al*. (2021) proposed an *extended isolation forest* that, in the branching step, considers jointly two or more variables; for instance, when two variables are randomly selected, then the algorithm partitions repeatedly the units according to a regression line whose intercept and slope are randomly generated each time.

In R, the standard isolation forest is implemented in the *solitude* package (Srikanth, 2021), while the *isotree* package (Cortes, 2022) implements the "base" isolation forest algorithm, as well as some of its variants.

3. Application of the chosen methods to some data from panel surveys

This section examines the performance of the methods presented in the previous Section when applied to various datasets related to panel or pseudo-panel surveys, as described in Table 3.1.

Table 3.1 – Datasets used in the experiments

Dataset/survey	Number of units	Type of units	Description
RDPerfComp	509	Firms	R&D performing US manufacturing; yearly observations from 1982 to 1989 of the following variables: production, labour and capital (a).
RiceFarms	171	Farms	The Indonesian rice farm dataset comprises 171 farms that produce rice, which were observed six times. Several variables are available, including hectares of cultivated area, gross output of rice in kilograms, and net output, among others (b).
Wages	595	Individuals	A panel of 595 individuals from 1976 to 1982, taken from the Panel Study of Income Dynamics (PSID); many variables available (see footnote 2)
Survey on Household Income and Wealth (SHIW)	3,804	Households	Subset of panel households observed in 2014 and 2016; many variables available: net income, consumption, wealth, etc. (c).

Source: Author's processing

(a) https://www.nuffield.ox.ac.uk/users/bond/index.html. See also the R package pder https://CRAN.R-project.org/package=pder.

(b) R package plm https://cran.r-project.org/package=plm.

(c) Bank of Italy, Survey on Household Income and Wealth, years 2014 and 2016. Public use anonymised microdata distributed for research purposes https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/distribuzione-microdati/index.html?com.dotmarketing.htmlpage.language=1.

In practice, in each dataset, the HB procedure is applied to the chosen variable and the resulting $\boldsymbol{E_i}$ scores, calculated using expression (2), become the input of the outlier detection techniques listed in the first column of Table 3.2, whose corresponding parameters/tuning constants are given in the second column of the table. All analyses were carried out in the R environment. Columns 3 and 4 in Table 3.2 provide details related to the chosen R package, function, and corresponding arguments (arguments not explicitly mentioned are set equal to their default values)[5].

[5] The used R code can be found in the GitHub repository: https://github.com/marcellodo/univOutl.

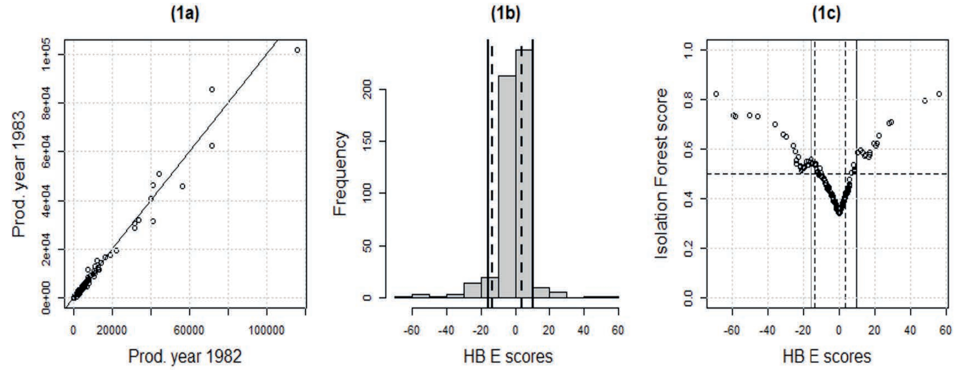
Table 3.2 – Methods, R functions and corresponding tuning parameters

Method	Parameters	R function and package	Arguments in the R function
Hidiroglou-Berthelot ("HB")	Quartiles (deciles for Wages dataset); $U = 0.5$; $C = 7$; $A = 0.05$	`HBmethod()` in *univOutl*	`U=0.5, A=0.05, C=7, pct=0.25 (pct=0.10` for Wages dataset)
Skewness-adjusted boxplot ("SABP") (see eqn. (6))		`boxB()` in *univOutl*	`k=1.5, method='adjbox'`
Isolation Forest ("IF")	No subsampling; 500 trees	`isolation.forest()` in isotree	`ntrees=500`
DBSCAN	Three runs with different values of g (6, 11,16) and different thresholds for δ (decided after graphical inspection of the sorted (g-1)-NN distances calculated on the $\boldsymbol{E_i}$)	`dbscan()` in *dbscan*	`minPts=6, minPts=11, minPts=16` `eps` set equal to the decided for each combination of `minPts` and the various datasets
k-NN outlier detection ("k-NN-dist")	Three runs with different values of k (5, 10,15)	`kNNdist()` in *dbscan*	k`=5, k=10, k=15`
k-NN weights ("k-NN-weight"; see expression (7))	Three runs with different values of k (5, 10,15)	`kNNdist()` in *dbscan*	`k=5, k=10, k=15, all=TRUE`

Source: Author's processing

The variable examined in the RDPerfComp dataset is the firms' production in 1983 compared to 1982. Figure 3.1 reports the observed scatterplot (1a); plot (1b) shows the distribution of the HB scores ($\boldsymbol{E_i}$) and the vertical continuous lines indicate the HB bounds provided by (3). In contrast, the dashed lines represent the fences of the skewness-adjusted boxplot, as provided by equation (6). The histogram shows a moderate negative skewness ($M = – 0.2338$), and the SABP fences identify a higher number of potential outliers if compared to HB (with $C = 7$ and $A = 0.5$).

Figure 3.1 - Scatterplot of the data related to firms' production (1a), distribution of the HB scores (1b), and relation between HB and IF scores (1c)

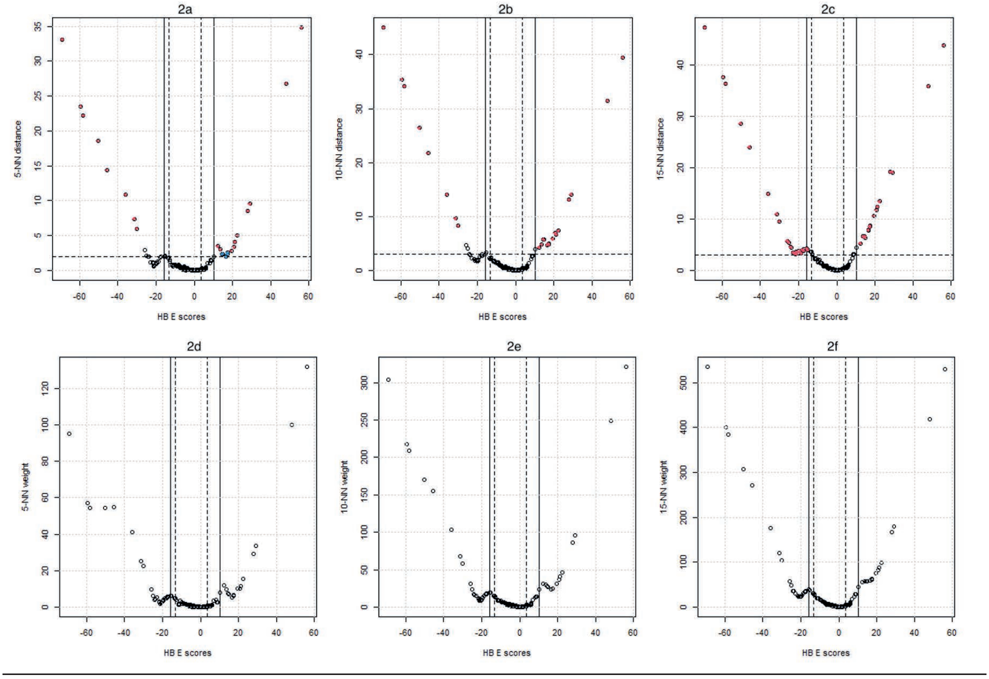


Source: Author's processing

The right-side plot (1c) reports the scatterplot of the IF scores (u_i) vs.E_i. Many observations identified as outliers by the HB method have an IF score slightly above 0.5, while the observations with the maximum observed IF score (slightly beyond 0.8, with a maximum achievable score of 1) are relatively few.

Figure 3.2 reports the scatterplots of scores provided by *k*-NN-dist (2a-2c) and *k*-NN-weight (2d-2f) compared to the input HB scores (E_i). Scatterplots (2a-2c) also show the findings of DBSCAN with respectively $\delta = 2$, $\delta = 3$ and $\delta = 3$. In these plots, the red-colour points indicate the noisy points (outliers), while the blue-colour ones form a separate cluster of observations, far from most of the observations that, however, are not identified as outliers. In this example, the outliers returned by DBSCAN are always fewer than those provided by the standard HB method. More generally, 5-NN and 10-NN are more effective than 15-NN distance in identifying potential outliers (units with higher distance).

Figure 3.2 - Firms' production data, relationship between HB and scores provided by the *k*-NN methods



Source: Author's processing

Scatterplot (2d), (2e), and (2f) compare the E_i with the sum of the *k*-NN distances (*k*-NN-weight). If compared to the “standard” *k*-NN-dist, the sum of the distances (weight), as expected, seems less sensitive to the value of *k* (Campos *et al*., 2016) and helps more in detecting the potential outliers (units with the highest weight $\omega_i^{(k)}$), in particular when $k = 5$ and $k = 10$

Table 3.3 shows the estimated Kendall’s *tau* correlation coefficient between the various scores obtained at the end of the different procedures for outlier detection (for HB, it is considered the absolute value $|E_i|$). Correlations are relatively high, indicating a high concordance between rankings of the scores produced by the different methods. IF scores are highly correlated with $|E_i|$. Concordance between $|E_i|$ and the scores provided by the application of *k*-NN methods increases with increasing values of *k*.

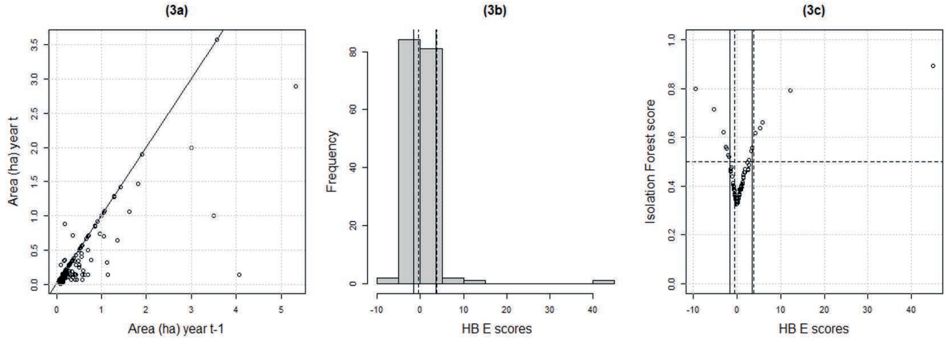
Table 3.3 – Kendall’s correlation between the scores assigned to the firms

	IF	5-NN-dist	10-NN-dist	15-NN-dist	5-NN-weight	10-NN-weight	15-NN-weight
|E|	0.8984	0.7672	0.8269	0.8689	0.7435	0.8094	0.8471
IF		0.8091	0.8665	0.9009	0.7845	0.8544	0.8920
5-NN-dist			0.829	0.7974	0.9045	0.8931	0.8587
10-NN-dist				0.8913	0.8050	0.9198	0.9414
15-NN-dist					0.7717	0.8596	0.9144
5-NN-weight						0.8712	0.8296
10-NN-weight							0.9400

Source: Author’s processing

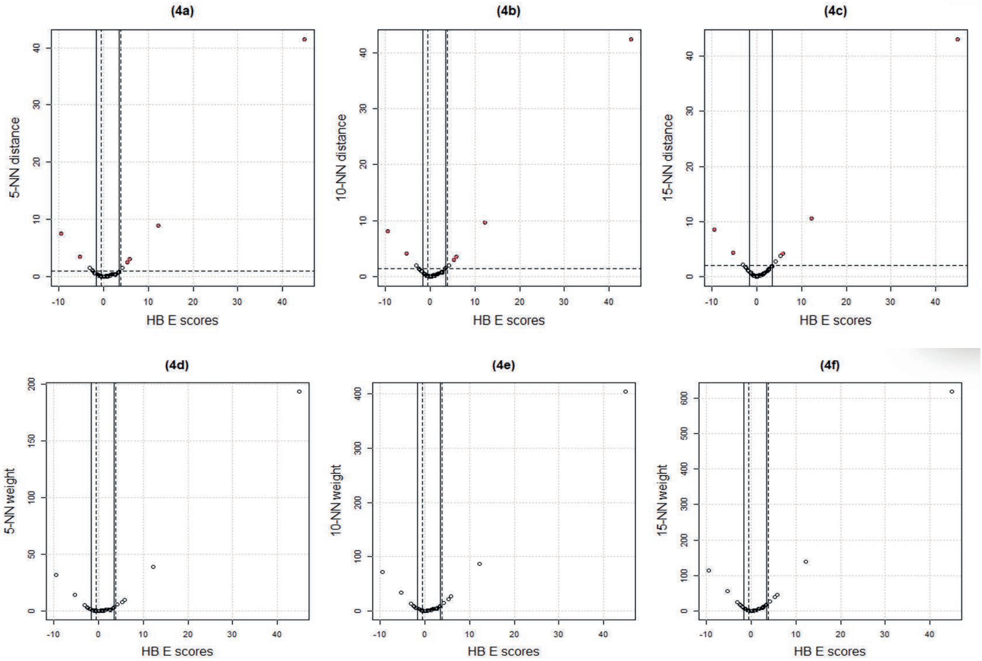
Figures 3.3 and 3.4 summarise the results of the different outlier detection procedures when applied to the area harvested for rice production in the observed farms (5th time occasion vs. the 4th) listed in the RiceFarm dataset. Histogram (3b) indicates a moderate positive skewness (M=0.3697). The fences of the SABP are close to the bounds of HB intervals, particularly on the right tail of the distribution; as expected, SABP appears to better account for moderate positive skewness. The identified outlying farms are relatively few and, in general, show an IF score greater than 0.6 (with few exceptions, located in the left tail). DBSCAN with the chosen distance thresholds ($\delta = 1$, $\delta = 1.5$ and $\delta = 2$, respectively; decided by after graphical inspection of the sorted *k*-NN distances) identifies quite a few outliers, slightly less than those identified by HB or SABP. Plots related to *k*-NN methods (4a-4c and 4d-4f) show that there are a few outlying farms with scores (*k*-NN distance or *k*-NN weight) that are not close to those of the majority of farms.

Figure 3.3 - Scatterplot of the area for rice production (3a), distribution of the HB scores (3b), and relation between HB and IF scores (3c)



Source: Author's processing

Figure 3.4 - Rice-growing area in farms, relationship between HB and scores provided by the k-NN methods



Source: Author's processing

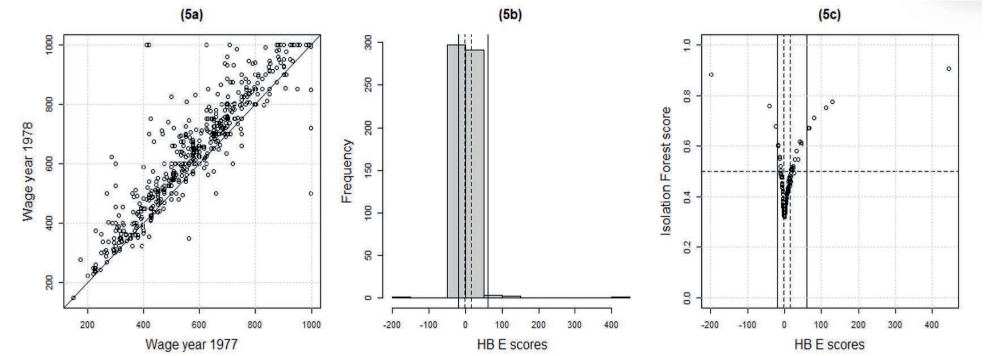
Table 3.4 shows that, also in this case, the IF score is the one with higher correlation (measured in terms of Kendall's tau) with the absolute value of the HB scores ($|E_i|$). Rankings based on IF scores tend to agree more with those provided by 10-NN and 15-NN methods. In general, correlations are all relatively high.

Table 3.4 - Kendall's correlation between the scores assigned to the farms producing rice

	IF	5-NN-dist	10-NN-dist	15-NN-dist	5-NN-weight	10-NN-weight	15-NN-weight
|E|	0.8627	0.7320	0.8285	0.8306	0.7064	0.8205	0.8401
IF		0.8084	0.9080	0.9125	0.7798	0.9054	0.9299
5-NN-dist			0.8071	0.7934	0.8978	0.8775	0.8306
10-NN-dist				0.9154	0.7624	0.9105	0.9480
15-NN-dist					0.7595	0.8890	0.9449
5-NN-weight						0.8333	0.7910
10-NN-weight							0.9392

Source: Author's processing

Figure 3.5 - Scatterplot of the individuals' wages (5a), distribution of the HB scores (5b), and relation between HB and IF scores (5c)

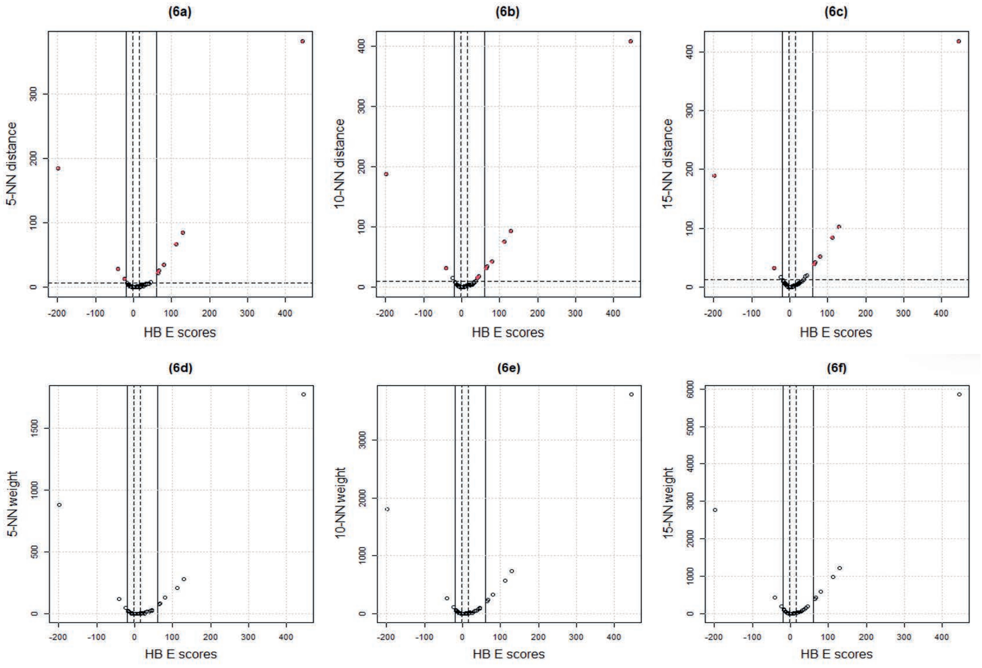


Source: Author's processing

Figures 3.5 and 3.6 show the results obtained by applying the investigated outlier detection methods when analysing the change in individuals' wages from 1977 to 1978, reported in the Wages dataset. Plot (5a) shows that there is an increase in wage for a large subset of individuals. The distribution of the HB E_i scores is positively skewed (M=0.3162), leading to identification of relatively few outliers; in this case, since there is a high concentration of the E_i around the median, in expression (4) it was decided to replace E_{Q1} and

E_{Q3} with respectively E_{P10} and E_{P90}, as suggested by Hidiroglou and Emond (2018). This is the reason for the large discrepancy between the HB bounds and those provided by the SABP.

Figure 3.6 - Wages data, relationship between HB and scores provided by the k-NN methods



Source: Author’s processing

Outliers identified by the HB method are individuals with an IF score of 0.7 or greater. DBSCAN returns the same outliers identified by HB, except g = 6 (k = 5) (scatterplot 6b) where some additional individuals are identified as outliers. In general, scores provided by the methods based on k-NN show clearly identifiable potential outliers that generally correspond to those identified by the HB procedure.

On average, the estimated correlations reported in Table 3.5 are lower than those calculated with other datasets, indicating that in this case, the rankings provided by the scores do not fully agree. As in other cases, the IF scores are those with higher correlation with the starting $|E_i|$.

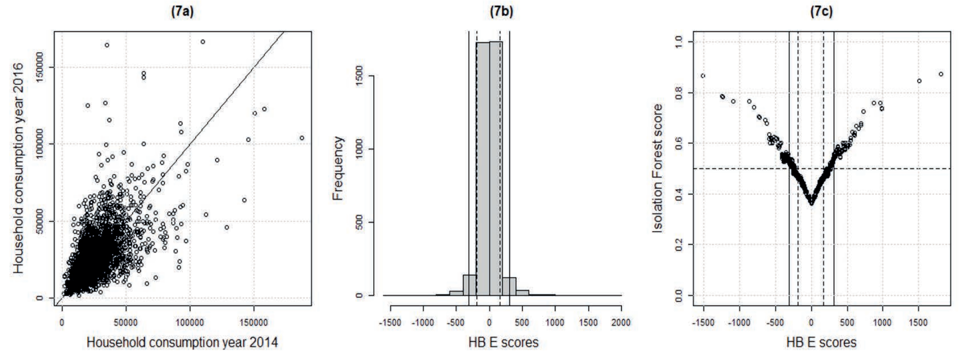
Table 3.5 – Kendall's correlation between the scores assigned to the individual wages

	IF	5-NN-dist	10-NN-dist	15-NN-dist	5-NN-weight	10-NN-weight	15-NN-weight
|*E*|	0.8101	0.5775	0.7045	0.7698	0.5420	0.6654	0.7254
IF		0.6376	0.7695	0.8371	0.6068	0.7426	0.8061
5-NN-dist			0.7287	0.6678	0.8617	0.8377	0.7629
10-NN-dist				0.8396	0.6712	0.8727	0.9200
15-NN-dist					0.6225	0.7861	0.8803
5-NN-weight						0.7836	0.7088
10-NN-weight							0.8976

Source: Author's processing

Figure 3.7 and 3.8 summarise the analyses done on household consumption observed in 2014 and 2016 for the panel component of the Survey on Household Income and Wealth (SHIW) carried out by the Bank of Italy (the survey is biennial). The histogram shows $\boldsymbol{E_i}$ having a symmetric distribution ($M = -\,0.024$) and the HB bounds ($C = 7$ and $A = 0.5$) determine a relatively high number of outlying households. Similarly, the number of households with IF scores greater than 0.6 is non-negligible, but this subset reduces significantly if the threshold is set to $\boldsymbol{u_0} = 0.7$.

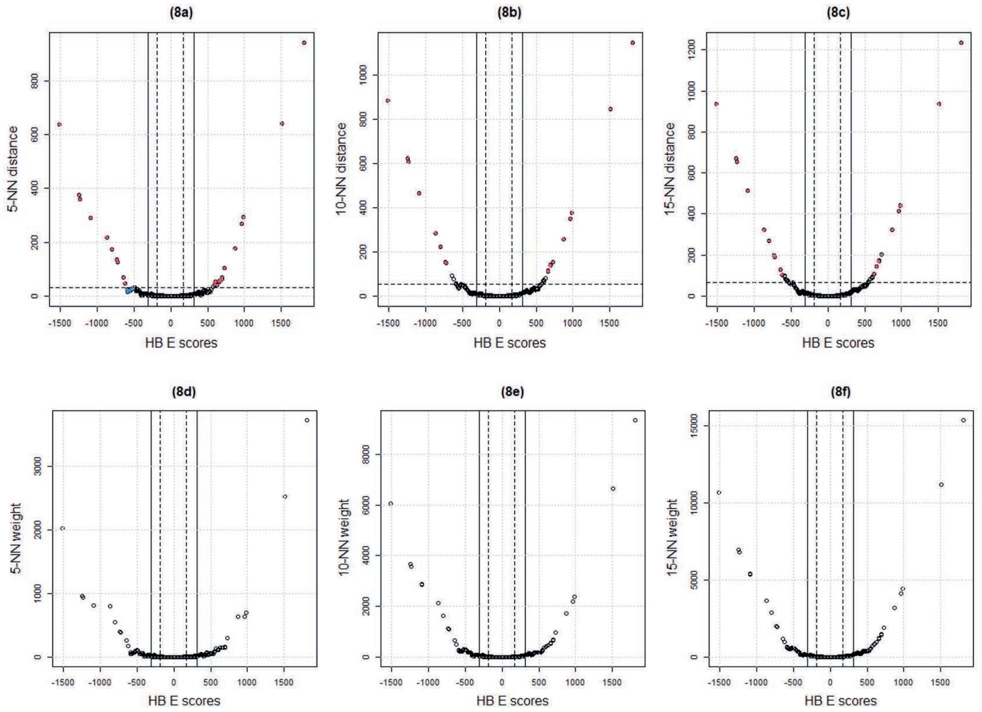
Figure 3.7 - Scatterplot of the household consumption (7a), distribution of the HB scores (7b), and relation between HB and IF scores (7c)



Source: Author's processing

DCSCAN clustering procedure (plots 8a-8c; with $\delta = 30$, $\delta = 55$ and $\delta = 65$, respectively; decided by after graphical inspection of the sorted *k*-NN distances) returns a number of outliers much smaller if compared to the HB method. More generally, all the plots related to *k*-NN distances or *k*-NN weight (8d, 8f) return final scores that facilitate the identification of the outlying observations.

FFigure 3.8 - Household consumption data, relationship between HB and scores provided by the k-NN methods



Source: Author’s processing

Kendall’s correlations in Table 3.6 show the same tendency highlighted in other situations.

Table 3.6 - Kendall’s correlation between the scores assigned to the individual household consumption

	IF	5-NN-dist	10-NN-dist	15-NN-dist	5-NN-weight	10-NN-weight	15-NN-weight
|*E*|	0.9340	0.6702	0.7519	0.7896	0.6493	0.7268	0.7668
IF		0.7048	0.7877	0.8232	0.6833	0.7663	0.8083
5-NN-dist			0.7751	0.7417	0.8737	0.8677	0.8158
10-NN-dist				0.8610	0.7390	0.8820	0.9180
15-NN-dist					0.7129	0.8204	0.8965
5-NN-weight						0.8406	0.7845
10-NN-weight							0.9169

Source: Author’s processing

It is worth noting that all scatterplots comparing the HB and IF scores show a "V" shaped diagram except Figure 3.1 (1c) (firms' production), where the E scores show an asymmetric distribution with moderate negative skewness ($M = –0.2338$) but quite "long" tails. The rule of thumb, which identifies units with an IF score greater than 0.5 in SHIW and firms' datasets as potential outliers, returns a relatively high fraction of potential outliers compared to others. This outcome suggests that such a rule should be applied carefully, rather than automatically.

When comparing the HB E scores with those provided by k-NN and "k-NN weight", the scatterplots show a kind of "U" shaped curve with some irregularities depending on the asymmetry in the distribution of the E scores; an exception is again demonstrated by the firms' production data (Figure 3.2). In general, all these scatterplots exhibit some differences when passing from $k = 5$ to $k = 10$. In comparison, shapes remain almost the same for $k = 10$ and to $k = 15$ (obviously, the magnitude of the distance-based scores increases by increasing the values of k), indicating that increasing the value of k too much may not be helpful. DBSCAN is closely related to k-NN since g = k + *1*, and the analysis of the k-NN distances is required to identify a threshold (parameter δ); it is not a simple task and we opted for a subjective choice guided by a graphical inspection instead of using expression (8) which would require setting the additional tuning constant ε; it is worth noting that for each of the considered datasets the obtained results remain almost stable when varying the combination of the tuning parameters (g and δ); more in general, it seems that this approach returns a relatively small number of observations having however a high chance of being outliers.

4 Conclusions

This paper compares traditional and recent approaches to detecting outliers with longitudinal data, a relatively simple situation that can be practically addressed by applying univariate outlier detection methods. The traditional approaches considered in this study, the HB method and the boxplot, are also popular in official statistics because they have the advantage of permitting a direct identification of the potential outliers (units outside the estimated bounds). The HB method requires setting a series of tuning parameters depending also on the observed distribution of the scores ($\boldsymbol{E_i}$) derived by transforming the initial ratios ($\boldsymbol{r_i = y_{t_2 i}/y_{t_1 i}}$); the method assumes an approximate Gaussian distribution for $\boldsymbol{E_i}$, allowing for slight skewness, but choosing the tuning parameters (to derive the $\boldsymbol{E_i}$ and the final bounds) may require more attempts. Skewness-adjusted boxplot does not explicitly assume a distribution for $\boldsymbol{E_i}$ (apart from that of working with a unimodal unknown distribution) and allows for a moderate skewness; on the contrary, it is not flexible enough as the bounds become too narrow with empirical distributions showing very long tails.

In the wide range of nonparametric methods for outlier detection developed in the fields of data mining and machine learning, we believe that those based on k-NN distances and isolation forests can be efficient and able to handle panel survey data collected in NSIs. These methods offer more flexibility than traditional ones, as they can adapt to different empirical distributions. They ultimately assign a score to each observation, where the larger the score, the higher the chance of being an outlier. This is also their major drawback because it's up to the practitioner to set a threshold such that units with a score beyond it are identified as potential outliers. Only DBSCAN ends up with a clear identification of outliers, but the price to pay is that of setting a threshold for the distance, in addition to the value of *g*. In the case studies considered in this comparison, with the chosen combination of input parameters, this approach generally returns a smaller number of potential outliers compared to traditional techniques and k-NN. For these reasons, DBSCAN seems preferable to k-NN methods; also because it permits the capture of "non-standard" distributions of the $\boldsymbol{E_i}$.

Setting the starting tuning parameters is simpler in the case of the isolation forest, where the practitioner should decide on the size of the bootstrap sample and the number of trees to grow, guided by some rule of thumb mentioned in the literature. The isolation forest has the additional advantage of producing scores ranging in the [0,1] interval, whose midpoint (0.5) represents a good initial candidate for setting a threshold.

In general, the great advantage of “new” nonparametric methods is that they are designed to work also in the multidimensional setting, in contrast to the HB and the boxplot. This is an appealing feature in official statistics, where the data sources often include many variables collected on the same set of units. Additional investigation is, however, required to better understand the pros and cons of these relatively “new” nonparametric methods.

References

Angiulli, F., and C. Pizzuti. 2002. “Fast Outlier Detection in High Dimensional Spaces”. In Elomaa, T., H. Mannila, and H. Toivonen (Eds.). *Principles of Data Mining and Knowledge Discovery. 6th European Conference, PKDD 2002, Helsinki, Finland, August 19–23, 2002, Proceedings*: 15-27. Berlin, Heidelberg, Germany: Springer. https://doi.org/10.1007/3-540-45681-3_2.

Brys, G., M. Hubert, and A. Struyf. 2004. “A Robust Measure of Skewness”. *Journal of Computational and Graphical Statistics*, Volume 13, Issue 4: 996-1017. https://www.jstor.org/stable/27594089.

Campos, G.O., A. Zimek, J. Sander, R.J.G.B. Campello, B. Micenková, E. Schubert, I. Assent, and M.E. Houle. 2016. “On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study”. *Data Mining and Knowledge Discovery*, Volume 30: 891-927. https://doi.org/10.1007/s10618-015-0444-8.

Cortes, D. 2022. *isotree: Isolation-Based Outlier Detection. R package version 0.5.15*. https://CRAN.R-project.org/package=isotree.

de Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ, U.S.: John Wiley & Sons.

D’Orazio, M. 2022. *univOutl: Detection of Univariate Outliers. R package version 0.3*. https://CRAN.R-project.org/package=univOutl.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. 1996. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In Simoudis, E., J. Han, and U. Fayyad (Eds.). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*: 226–231. Washington, D.C., U.S.: AAAI Press.

Eurostat. 2014. *Memobust Handbook on Methodology of Modern Business Statistics*. Luxembourg: Eurostat. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en.

Hautamäki, V., I. Kärkkäinen, and P. Fränti. 2004. “Outlier Detection Using k-Nearest Neighbour Graph”. In *Proceedings of the 17th International Conference on Pattern Recognition - ICPR 2004*: 430-433. New York, NY, U.S.: IEEE.

Hahsler, M., and M. Piekenbrock. 2022. *dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R package version 1.1-10.* https://CRAN.R-project.org/package=dbscan.

Hahsler, M., M. Piekenbrock, and D. Doran. 2019. "dbscan: Fast Density-Based Clustering with R". *Journal of Statistical Software*, Volume 91, Issue 1: 1-30. https://doi.org/10.18637/jss.v091.i01.

Hariri, S., M.C. Kind, and R.J. Brunner. 2021. "Extended Isolation Forest". In *EEE Transactions on Knowledge and Data Engineering*, Volume 33, N. 4: 1479-1489. https://doi.org/10.1109/TKDE.2019.2947676.

Hidiroglou, M.A., and J.-M. Berthelot. 1986. "Statistical editing and Imputation for Periodic Business Surveys". *Survey Methodology*, Volume 12, N. 1: 73-83. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1986001/article/14442-eng.pdf?st=6rMUAGoq.

Hidiroglou, M.A., and N. Emond. 2018. "Modifying the Hidiroglou-Berthelot (HB) method". *Unpublished Note*. Ottawa, Ontario, Canada: Statistics Canada, Business Survey Methods Division.

Hubert, M., and E. Vandervieren. 2008. "An Adjusted Boxplot for Skewed Distributions". *Computational Statistics & Data Analysis*, Volume 52, Issue 12: 5186-5201. https://doi.org/10.1016/j.csda.2007.11.008.

Italian National Institute of Statistics - Istat, Statistics Netherlands - CBS, and Swiss Federal Statistical Office - SFSO. 2007. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Luxembourg: Eurostat, EDIMBUS Project. http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf.

Knorr, E.M., and R.T. Ng. 1998. "Algorithms for Mining Distance-Based Outliers in Large Datasets". In Proceedings of the 24th *International Conference on Very Large Databases (VLDB '98):* 392-403. San Francisco, CA, U.S.: Morgan Kaufmann Publishers Inc. https://www.vldb.org/conf/1998/p392.pdf.

Liu, F.T., K.M. Ting, and Z.-H. Zhou. 2012. "Isolation-based anomaly detection". *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), Volume 6, Issue 1 (Article 3): 1-39. https://doi.org/10.1145/2133360.2133363.

Liu, F.T., K.M. Ting, and Z.-H. Zhou. 2008. "Isolation forest". In *Proceedings of the Eighth IEEE International Conference on Data Mining* (*ICDM*): 413–422. New York, NY, U.S.: IEEE. https://doi.org/10.1109/ICDM.2008.17.

Madsen, J.H. 2018. *DDoutlier: Distance & Density-Based Outlier Detection. R package version 0.1.0*. https://CRAN.R-project.org/package=DDoutlier.

R Core Team. 2022. R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ramaswamy, S., R. Rastogi, and K. Shim. 2000. "Efficient Algorithms for Mining Outliers from Large Data Sets". *ACM SIGMOD Record*, Volume 29, Issue 2: 427-438. https://doi.org/10.1145/335191.335437.

Srikanth, K. 2021. *solitude: An Implementation of Isolation Forest. R package version 1.1.3*. https://CRAN.R-project.org/package=solitude.

United Nations Statistical Commission, and Economic Commission For Europe - UNECE. 2000. *Glossary of Terms on Statistical Data Editing*. Geneva, Switzerland: United Nations. https://unece.org/DAM/stats/publications/editing/SDEGlossary.pdf.