

Article

Flexible Techniques to Detect Typical Hidden Errors in Large Longitudinal Datasets

Renato Bruni , Cinzia Daraio  and Simone Di Leo *

Department of Computer, Control and Management Engineering, Sapienza University of Rome, 00185 Roma, Italy; bruni@diag.uniroma1.it (R.B.); daraio@diag.uniroma1.it (C.D.)

* Correspondence: dileo@diag.uniroma1.it

Abstract: The increasing availability of longitudinal data (repeated numerical observations of the same units at different times) requires the development of flexible techniques to automatically detect errors in such data. Besides standard types of errors, which can be treated with generic error correction techniques, large longitudinal datasets may present specific problems not easily traceable by the generic techniques. In particular, after applying those generic techniques, time series in the data may contain trends, natural fluctuations and possible surviving errors. To study the data evolution, one main issue is distinguishing those elusive errors from the rest, which should be kept as they are and not flattened or altered. This work responds to this need by identifying some types of elusive errors and by proposing a statistical-mathematical approach to capture their complexity that can be applied after the above generic techniques. The proposed approach is based on a system of indicators and works at the formal level by studying the differences between consecutive values of data series and the symmetries and asymmetries of these differences. It operates regardless of the specific meaning of the data and is thus applicable in a variety of contexts. We implement this approach in a relevant database of European Higher Education institutions (ETER) by analyzing two key variables: “Total academic staff” and “Total number of enrolled students”, which are two of the most important variables, often used in empirical analysis as a proxy for size, and are considered by policymakers at the European level. The results are very promising.



check for updates

Citation: Bruni, R.; Daraio, C.; Di Leo, S. Flexible Techniques to Detect Typical Hidden Errors in Large Longitudinal Datasets. *Symmetry* **2024**, *16*, 529. <https://doi.org/10.3390/sym16050529>

Academic Editors: Sunil Jha, Malgorzata Rataj, Xiaorui Zhang and Shangce Gao

Received: 27 February 2024

Revised: 11 April 2024

Accepted: 22 April 2024

Published: 28 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: big data; information processing; information reconstruction; data quality; longitudinal data sequences

1. Introduction

In the context of an increasingly data-driven economy, data quality is of paramount importance for organizations of all types and sizes and a lack of attention to it can lead to several costs and inefficiencies. According to the quality framework of the Organisation for Economic Cooperation and Development (OECD) [1], data quality is defined as the “fitness for use” with respect to user needs. Data quality can be viewed as an overarching principle that must be kept into account when designing models of metrics [2]. Every technique developed to improve data quality should consider that the very concept of data quality is not one-dimensional but multidimensional [3–5]. In particular, the following seven dimensions are usually identified: accuracy, completeness, consistency, validity, timeliness, uniqueness, and integrity.

Due to the relevance of the issue, many authors have proposed methods or guidelines to assess problems with data quality [6–12]. However, few works focus on the problems that specifically regard the case of numerical data describing repeated observations of the same units over a period of time. This type of data is often called *longitudinal* data or panel data. If we restrict our attention to one single unit over the whole time period, then we obtain a single *time series*. If, on the contrary, we consider all the different units but restrict our attention to one single time instant, then we obtain *cross-sectional* data. In

recent years, longitudinal data have become more and more abundant, and researchers have been exploring the vast possibilities given by their study, typically by using advanced artificial intelligence techniques that are now able to deal with huge datasets. However, one ubiquitous problem affecting almost all data-related applications is the presence of errors in the data. Unfortunately, longitudinal data make no exception to this. Thus, when data containing errors are used for some studies, the results will contain a certain degree of unreliability. Or, in other words, when data contain errors, the problem that we solve is actually *different* from the real problem that was to be solved.

The presence of errors in data may be due to several causes, and consequently, there exist many types of data errors. Easily identifiable cases are, for example, missing values or out-of-range values. For the rest of this work, we assume that the error location is unknown and that the original exact value is not available, since otherwise simple replacement operations would fix the issue. As this type of problem is very widespread, many techniques have been developed in different fields of science to cope with these situations. There exist several *imputation* techniques for the reconstruction of missing or out-of-range values; see for example [13,14]. Some methods for estimating measurement errors in longitudinal data are based on latent variable modeling [15,16]. Another technique, called the MultiTrait MultiError approach, is presented in [17] to estimate multiple types of errors concurrently using a combination of experimental design and latent variable modeling. Survey [18] identifies three main error types in time series: single-point big errors, single-point small errors and continuous errors. Continuous errors are the case where an error occurs at several consecutive time points. A single-point error occurs discontinuously in a time series and only occurs on a single data point at intervals. A big error means that the observed value of the data point is far from the true value, as opposed to a small error. Single-point errors may often be identified by searching for outliers. An outlier is “An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [19]. Following surveys [18,20], methods to detect anomalies/outliers in time series can be grouped as follows: density-based methods like Local Outlier Factor [21]; threshold methods based on statistical descriptions like standard deviations or median absolute deviation; see for example [22]; statistical test methods like Grubbs’ test [23]; machine learning methods based on the use of a training set of outliers, like isolation forest [24]; model-based methods like moving average (MA), autoregression (AR) and their variations (like ARMA or ARIMA); see for example [25]. These methods are effective in cleaning a single time series but suffer from the following limitations: they are generally quite data-demanding (so they do not scale well when the series is not long enough) and they often require statistical assumptions on the data (e.g., they need a Gaussian distribution).

Finally, there exist methods based on data integration, when the data under analysis are also contained or derivable from different sources [26,27], and methods based on checksum to reconstruct information that was originally known but has been damaged. However, these categories are not applicable to our case, because the original exact value is neither known nor available from other sources.

In addition to the “usual” types of errors, large longitudinal datasets may contain peculiar types of errors that are not easily identifiable by the techniques used for finding and/or correcting errors in generic datasets. In the context of numerical longitudinal data obtained from several sources and assembled to form one database, the following situations can lead to some very typical errors:

- (1) When the time series of the different units are written/stored one next to the other, one or more values from one unit A may be erroneously inserted in the space allowed to a contiguous unit B , and vice versa, the corresponding values from B may be inserted in the space of A . We call this situation the *inversion of values between units*. This type of error is often not detectable by general error detection techniques. Moreover, even if the problem is detected, because for example a value v_i is too high or too low for unit A , the generalist imputation techniques will probably try to reconstruct the

correct value based on elaborations involving unit A, ignoring that the correct values are already stored in the database but in the space of record B. Several problems may arise if this type of error is not fully recognized.

- (2) Data contain one or more large “jumps” in the values of the time series corresponding to one unit. For example, given a unit A, imagine that the values of one of its variables are 100, 120, 280, 130, 120, 150. The third value is far from the others, so we may suspect some problem. However, depending on the degree of volatility, the situation can also be normal after all. We call this situation an *anomalous jump*. In this case, we need to identify a threshold above which the values should be considered erroneous. This is a very delicate issue, and standard error detection techniques may be insufficient in this case.
- (3) A time series is composed of values produced by a data provider (for example, an agent or an organization) at every given interval of time (for example, every year). In this case, it may happen that the data provider computes a value v_t for a given time t and later discovers that v_t was incorrect, because some units should have been added to v_t but they were not considered, so v_t should actually be increased by δ_t , or because some units counted in v_t actually belong to the next time interval, so v_t should be decreased by $-\delta_t$. In this case, if it is too late to modify v_t , the data provider often tries to compensate for the error by modifying the next value produced, v_{t+1} , providing $v_{t+1} + \delta_t$ in the first case and $v_{t+1} - \delta_t$ in the second. We call this situation *recalculation operated by the data provider*. Clearly, this type of problem is hardly detectable by general error detection techniques, and again, several problems may arise if this type of error is not fully recognized.

Another limitation of the above-described error correction techniques is that they generally look at the single time series and do not consider the whole database structure in the search for potential errors, so they overlook errors that may arise from its tabular format, like the inversion and the recalculation problems. This work responds to this need by proposing a statistical-mathematical approach based on a system of indicators that define a rational process to assess and improve the quality of the data (as suggested by [28]). In particular, the proposed approach is able to identify suspect erroneous data suffering from the three described problems by working at the formal level, regardless of the specific meaning of the data. Therefore, it is applicable even after a generic error correction step in a variety of contexts. Note that we aim at pinpointing the described error situations without flattening or altering the rest of the data, as some noise removal techniques unfortunately do. Therefore, one delicate issue is distinguishing natural fluctuations from erroneous jumps. We pursue this aim by studying the symmetries and asymmetries of the differences (deltas) between consecutive values of the time series. Our approach incorporates also a certain degree of flexibility, because it is based on a number of mathematical conditions that can be slightly changed to adapt to different cases and take into account different realities.

We implement this approach in a large and relevant database of European Higher Education institutions (ETER) by analyzing the two key variables “Total academic staff” and “Number of enrolled students”. These are two of the most important variables, often used in empirical analysis as a proxy for the size of the institutions, and are considered by the policymakers at the European level.

2. Materials and Methods

As explained in Section 1, there exist several error correction techniques for large datasets. Nonetheless, in large numerical longitudinal databases, we identify the following three main consistency problems that are specific to the case of longitudinal data and are hardly treatable with standard error detection and correction techniques:

1. Inversion of values between units;
2. Anomalous jump;
3. Recalculation operated by the data provider.

The proposed methodology aims to the identification of possible errors by raising check flags (which can later be examined by database managers) on suspect data. Our method can also be applied after other standard error correction procedures, and it consists of several steps for each of the above three problems, detailed in Sections 2.1–2.3. As materials, we conducted our experiments on the ETER database, described in Section 2.4.

2.1. Inversion Problem

To identify the inversion problem between two units A and B , we evaluate two types of conditions that we call here $H1$ and $H2$. The first type ($H1$) consists of assessing, for each possible couple of units A and B , whether there are possible systematic exchanges between the values of A and B over one or more time instants through the evaluation of the differences (called Δ) between each pair of temporally consecutive values of the same variable. In more detail, the generic condition $H1$ is evaluated by executing the following steps.

H1.a. Denote by i the index of the generic unit (a row in the dataset), with $i = 1, \dots, m = U$. The series of the values of a variable (or attribute) v over the time instants $t = 1, \dots, n = S$ for unit i is denoted by (v_i^1, \dots, v_i^n) . Then, define $\Delta v_i^{(t,t+1)}$ as the difference (*delta*) between the two values assumed by unit i in two consecutive time instants $t, t + 1$ for variable v as follows:

$$\Delta v_i^{(t,t+1)} = v_i^t - v_i^{t+1} \quad (1)$$

Those deltas are computed for each period of the dataset and for each unit (and for each variable if there is more than one variable in the dataset). Obviously, for the last period n , $\Delta v_i^{(n,n+1)}$ is not computable. The generic value $\Delta v_i^{(t,t+1)}$ can take on a negative or a positive value. We define as P the set of the indices t for which $\Delta v_i^{(t,t+1)}$ is positive and as N the set of the same indices for which $\Delta v_i^{(t,t+1)}$ is negative.

H1.b. Compute, for each unit i , the value DV_i defined as the modulus of the product between the sum of the positive deltas and the sum of negative deltas as follows:

$$DV_i = \left| \sum_{t \in P} \Delta v_i^{(t,t+1)} \sum_{t \in N} \Delta v_i^{(t,t+1)} \right|. \quad (2)$$

This is somehow a measure of the *intrinsic variability* of the unit i . Indeed, in practical cases, this measures the fact that some units will be “changing” their values more than others. In case any of the $\sum_{t \in P} \Delta v_i^{(t,t+1)}$ or $\sum_{t \in N} \Delta v_i^{(t,t+1)}$ is equal to zero, its value is changed to 1 to avoid it collapsing to zero when the intrinsic variability of a unit must be non-negative. Note that this is one of the customizable aspects, depending on the practical case under study.

H1.c. Compute the DM_i value for each unit i as the ratio between DV_i and the arithmetic mean of all DV_i in the entire dataset considered as follows:

$$DM_i = \frac{DV_i}{\frac{\sum_{i \in U} DV_i}{m}} \quad (3)$$

This value represents a normalization of the above measure of intrinsic variability. The normalization should be conducted over some homogeneous set of units to which unit i belongs. Thus, depending on the context, such a homogeneous set must be identified. For example, in the case presented in Section 3, there is strong heterogeneity in data from different national contexts (i.e., different countries). For this reason, the DV_i is averaged by the mean of DV_i over the country to which the unit belongs.

H1.d. The numerical values of the above DM_i may still vary greatly. To avoid numerical instability, we compress their scale by computing the cubic root, obtaining values called RQ_i representing the compressed normalized intrinsic variability of the unit.

$$RQ_i = \sqrt[3]{DM_i} \quad (4)$$

H1.e. Compute the value GM_i as the geometric mean of all the deltas in the module of unit i . This value represents an evaluation of the size of the unit. If some of the deltas are zero, then they can again be replaced with 1 to avoid them all collapsing to zero when this is not acceptable.

H1.f. Now, to compute a reasonable upper limit on the delta values that unit i could attain, we multiply the compressed normalized intrinsic variability by the measure of the size of the unit, obtaining the following threshold T_i :

$$T_i = GM_i RQ_i \quad (5)$$

H1.g. Now, to finally recognize the situation of the inversion of a value between two consecutive units A and B by computing $H1$, we need four conditions to be verified at the same time: unit A has two consecutive deltas larger (in modulus) than the threshold T_A and with opposite signs (w.l.o.g, the first is positive and the second is negative), and unit B for the same time instants has again two consecutive deltas larger (in modulus) than the threshold T_B but with signs reversed with respect to A (the first is negative and the second is positive). In practice, condition $H1$ is given by the following Boolean expression:

$$H1_{(A,B)}^t: \{[(\Delta v_A^{(t-1,t)} > 0 \wedge \Delta v_A^{(t,t+1)} < 0) \wedge (\Delta v_B^{(t,t+1)} < 0 \wedge \Delta v_B^{(t,t+1)} > 0)] \vee [(\Delta v_A^{(t-1,t)} < 0 \wedge \Delta v_A^{(t,t+1)} > 0) \wedge (\Delta v_B^{(t-1,t)} > 0 \wedge \Delta v_B^{(t,t+1)} < 0)]\} \wedge (|\Delta v_A^{(t-1,t)}| > T_A \wedge |\Delta v_A^{(t,t+1)}| > T_A \wedge |\Delta v_B^{(t-1,t)}| > T_B \wedge |\Delta v_B^{(t,t+1)}| > T_B) \quad (6)$$

If $H1_{(A,B)}^t$ is true, then we also need a corresponding condition $H2_{(A,B)}^t$ to be true to have a probable swap problem. The generic condition $H2$ is evaluated by the following steps.

H2.a. For each unit i , we define I_i^t as the distance of the value v_i^t at time t from the mean value of v over time without the value at time t as follows:

$$I_i^t = v_i^t - (\sum_{k \in S \setminus t} v_i^k) / n - 1 \quad (7)$$

H2.b. We define N_i^t as the distance of the value v_i^t at time t from the mean value of v over time without the value at time t , but this time we take the values of the subsequent unit $i+1$ (the one with which the values could have been exchanged), as follows:

$$N_i^t = v_i^t - (\sum_{k \in S \setminus t} v_{i+1}^k) / n - 1 \quad (8)$$

H2.c. Finally, we define F_i^t as the minimum between the modulus of the two above values: In practice, we are comparing the distance between value v_i^t and all the other values of unit i and between v_i^t and all the other values of unit $i+1$. If v_i^t is closer to the values of unit $i+1$, then the minimum is $|N_i^t|$ and inversion is probable.

$$F_i^t = \min(|I_i^t|, |N_i^t|) \quad (9)$$

Hence, condition $H2$ for units A and B is evaluated as follows:

$$H2_{(A,B)}^t: F_i^t \neq |I_i^t| \tag{10}$$

Conditions $H1$ and $H2$ are computed and checked for every couple of units A and B and every time instant t . If $H1_{(A,B)}^t$ is true and $H2_{(A,B)}^t$ is also true, a possible swapping error flag is raised for units A and B at time instant t ; otherwise, no flag is raised. Note that this error may even affect more than one time instant of the same two units.

Example 1. We provide an example of the check for the inversion problem for two units (called unit 1 and unit 2) on a variable v of a longitudinal dataset with $t = 5$. The data of the units are shown in Table 1. We first compute the deltas for each unit; see Table 1. For instance, unit 1 has $v_1^2 = 18$ and $v_1^3 = 130$, hence $\Delta v_1^{(2,3)} = 18 - 130 = -112$. After this, DV is equal to $|(-112 - 10)(107 + 10)| = 14,274$ for unit 1 and $|(-129)(120 + 5 + 5)| = 16,770$ for unit 2. Subsequently, the value of the geometric mean GM is 33.09 for unit 1 and 24.94 for unit 2; DM is 0.92 for unit 1 and 1.08 for unit 2, and RQ is 0.97 for unit 1 and 1.03 for unit 2. Consequently, the threshold T is 32.17 for unit 1 and 25.59 for unit 2.

Table 1. Values v and Δ of the inversion problem example.

	v^1	v^2	v^3	v^4	v^5	$\Delta^{(1,2)}$	$\Delta^{(2,3)}$	$\Delta^{(3,4)}$	$\Delta^{(4,5)}$
Unit 1	125	18	130	120	130	107	-112	10	-10
Unit 2	21	150	30	25	20	-129	120	5	5

Then, we find $H1_{(1,2)}^t$. Considering that for unit 1, $\Delta v_1^{(1,2)} > 0$ and $\Delta v_1^{(2,3)} < 0$ and, for unit 2, $\Delta v_2^{(1,2)} < 0$ and $\Delta v_2^{(2,3)} > 0$, the first part of the $H1$ condition is verified. Additionally, all those Δv exceed the respective thresholds T . Therefore, $H1_{(1,2)}^2$ is true.

To evaluate $H2_{(1,2)}^2$, we compute I_1^2 and N_1^2 for unit 1 and time 2.

We have value $I_1^2 = 18 - (125 + 18 + 130 + 120 + 130 - 18)/4 = -108.25$.

Value $N_1^2 = 18 - (21 + 150 + 30 + 25 + 20 - 150)/4 = -6$.

Since -6 has the smallest modulus value, $F_1^2 = 6$, $F_1^2 \neq I_1^2$ and $H2_{(1,2)}^2$ is true. As both conditions are true, a probable inversion error flag is reported for the period $t = 2$.

2.2. Anomalous Jump Problem

To identify anomalous jumps, we now compute for each unit i a threshold with tolerance, TT_i , larger than before, obtained as follows. After the computation of the threshold T_i described in Section 2.1, we execute the following steps.

- Calculate the value LGM_i as the natural logarithm of the GM_i value presented in Section 2.1. This logarithm of the size represents a compressed measure of the size of the unit.
- Compute VI_i as the integer upper part of the value LGM_i plus a constant c representing another element of the customization of the procedure. This value can be determined either with a priori reasoning or even derived from the data itself, for example, by defining a training set of anomalous/not anomalous jumps and by choosing the value of c maximizing the detection performance.

$$VI_i = \lceil LGM_i + c \rceil$$

- Compute GMT_i as the sum of $GM_i + T_i$. In practice, we are summing the size and threshold for unit i , obtaining a kind of deformation of the threshold by its size.

- d. Finally, identify the threshold with tolerance TT_i as the largest value between the two size-derived values described above. This is used as an upper bound on the reasonable jumps observed in the values of the unit.

$$TT_i = \max(VI_i, GMT_i).$$

Now, an anomalous jump flag is raised for a unit i in a time $t, t + 1$ for variable v if the module of $\Delta v_i^{(t,t+1)}$ is greater than the threshold TT_i .

Example 2. We provide an example of an anomalous jump problem. Consider a unit (called unit 3) with variable v of a longitudinal dataset with $t = 5$. The data and the deltas of the unit are shown in Table 2. We compute the threshold $T = 101.66$, as already seen in the previous example. Then, we find $LGM = 4.32$, $VI = 13$ and $GMT = 177.15$. By considering $c = 8$ and the mean of deltas = 10, the resulting threshold with tolerance TT value is 177.15.

Table 2. Values and Δ of the unit considered for the anomalous jump example.

	v^1	v^2	v^3	v^4	v^5	$\Delta (1,2)$	$\Delta (2,3)$	$\Delta (3,4)$	$\Delta (4,5)$
Unit 3	200	220	400	210	230	−20	−180	190	−20

As $|\Delta v_3^{(2,3)}| = 180 > 177.15$ and $|\Delta v_3^{(3,4)}| = 190 > 177.15$, we report an anomalous jump flag for the period $t = 2, 3$ and the period $t = 3, 4$. The data manager will have to check the values of $t = 2, t = 3$ and $t = 4$ to understand the reasons for this anomalous jump.

2.3. Recalculation Problem

To identify a recalculation operated by the data provider we use the above threshold with tolerance TT_i . We suspect a recalculation problem on unit i if two contiguous deltas of opposite signs are both above the threshold TT_i in the modulus as follows:

$$[(\Delta v_i^{(t-1,t)} > 0 \wedge \Delta v_i^{(t,t+1)} < 0) \vee (\Delta v_i^{(t-1,t)} < 0 \wedge \Delta v_i^{(t,t+1)} > 0)] \wedge (|\Delta v_i^{(t-1,t)}| > TT_i \wedge |\Delta v_i^{(t,t+1)}| > TT_i) \quad (11)$$

If this condition is true, a possible recalculation flag is raised.

Example 3. We provide an example of a recalculation problem. Consider a unit (called unit 4) with variable v of a longitudinal dataset with $t = 5$. The data and the deltas of the unit are shown in Table 3. Following the steps described above, after computing the threshold $T = 39.40$, we find $LGM = 3.80$, $VI = 12$ and $GMT = 84.24$. The resulting TT value for the unit is 84.24. A flag of possible recalculation error is raised for period $t = 3$ since $\Delta v_4^{(2,3)} > 0$ and $\Delta v_4^{(3,4)} < 0$, while simultaneously $|\Delta v_4^{(2,3)}| = 87 > 84.24$ and $|\Delta v_4^{(3,4)}| = 155 > 84.24$.

Table 3. Values and Δ of the unit considered for the recalculation example.

	v^1	v^2	v^3	v^4	v^5	$\Delta (1,2)$	$\Delta (2,3)$	$\Delta (3,4)$	$\Delta (4,5)$
Unit 4	163	167	80	235	160	−4	87	−155	75

All the described operations are available in the Microsoft Excel file contained in [29]. This file can be used to operate the described checks with any data, by simply pasting them into the sheet “Main Table”. Each row must represent a single unit of analysis. The Excel file is also adaptable to use units with a variable number of time instants. The minimum number of time instants must be inserted in cell MIN OSS in the sheet “Threshold Calculation”.

2.4. Data

The European Tertiary Education Register (ETER) [30] is a key initiative for understanding the higher education landscape in Europe developed after the successful AQUAMETH project [31]. This database provides a reference list of Higher Education Institutions (HEIs) and institutional data on their activities and achievements, including students, graduates, staff and finances. It thus complements national and regional education statistics provided by EUROSTAT [32].

As of March 2024, ETER includes 41 European countries and provides data from 2011 to 2020, with a total of over 3500 HEIs. ETER collects a wide range of data on HEIs, including institutional characteristics (type, size, specialization), student information (enrolment, graduates, mobility), staff (lecturers, researchers, administrative staff), finances (income, expenditure, investment) and research and development activities. ETER complies extensively with statistical regulations and manuals, in particular the UOE Manual on Data Collection on Formal Education and the OECD Frascati Manual on Research and Experimental Development Statistics. This ensures the comparability of data with other international sources. Collaboration with a network of experts and data providers in all participating countries ensures that information is collected from reliable and consistent sources. Established methodologies are used to define variables and indicators, enabling the re-use of collected data for statistical purposes and comparability with other sources. Data undergo rigorous quality control and validation to identify and correct errors or inconsistencies, as described in [33]. However, as described in Section 3, the proposed techniques were able to locate several cases of the specific longitudinal data problems described above.

ETER contributes to a better understanding of the higher education landscape and is a valuable resource for researchers, policymakers and stakeholders in European higher education. Within ETER, we selected the case of the two variables “Total academic personnel” in headcount (HC) and “Total number of enrolled students” because they are widely used in empirical analysis and by policymakers as a proxy for the size of the universities. Therefore, they are two of the most important variables, and it is of paramount importance to detect any possible errors in them.

Total academic personnel in HC include the following:

- (i) The number of academic staff whose primary assignment is instruction, research or public service;
- (ii) Staff who hold an academic rank, like professor, assistant professor, lecturer or an equivalent title;
- (iii) Staff with other titles (like dean, head of department, etc.) if their principal activity is instruction or research;
- (iv) PhD students employed for teaching assistance or research.

Data on students are divided by the level of education of the program to which they are enrolled, using the International Standard Classification of Education (ISCED) in its 2011 version. This version includes the distinction between “Bologna” levels of education (Bachelor, Master and Doctorate). The “Total number of enrolled students” includes students from ISCED 5 (short-cycle tertiary), ISCED 6 (bachelor), ISCED 7 (Master) and does not include ISCED 8 (Doctoral level).

We report our experiments on the largest EU countries present in the ETER, i.e., Germany, France, Italy, Spain, Poland and Portugal, for a total of 1587 HEIs, in the time period from 2011 to 2020. Tables 4 and 5 report the number of HEIs having complete data, respectively, for the variables “Total academic personnel” and “Total number of enrolled students”. The time interval of these two variables is annual.

Table 4. Number of HEIs with the variable total academic staff (HC) available in the ETER for each country and year in the period 2011–2020.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Italy	115	115	115	114	114	114	114	114	114	114	1143
Germany	365	378	383	385	385	383	400	400	396	399	3874
Spain	77	80	80	81	81	80	82	83	83	84	811
France	131	132	130	129	126	0	123	123	119	111	1124
Poland	0	0	0	0	0	0	247	243	241	237	968
Portugal	113	106	94	91	90	95	90	90	89	92	950

Table 5. Number of HEIs with the variable total number of enrolled students available in the ETER for each country and year in the period 2011–2020.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Italy	176	176	176	215	215	216	206	207	207	208	2002
Germany	368	380	384	388	389	391	402	401	401	399	3903
Spain	77	80	80	81	82	82	82	83	83	84	814
France	339	340	368	375	377	0	212	209	206	203	2629
Poland	286	272	280	282	281	274	248	245	243	239	2650
Portugal	113	106	94	91	90	96	90	90	89	92	951

3. Results

3.1. Experiments with the Proposed Techniques

All the computations described in Section 2 have been implemented in Microsoft Excel and run directly from a spreadsheet. Those controls have been applied to the described ETER database, considering the case of the variables Total academic personnel and Total number of enrolled students. All HEIs from Germany, France, Italy, Spain, Poland and Portugal with available values for that variable were considered, for a total of 1587 HEIs. In the computation of DV_i and GM_i , if some factor was zero, it was replaced with 1 to avoid them collapsing to zero. In the computation of VI_i , the constant c was set at 8 by means of experimental fine-tuning. Tables 6 and 7 (respectively, for the two variables Total academic personnel and Total number of enrolled students) report, for each country, the total number of flags raised by the described techniques. In particular, we indicate the $H1$ and $H2$ flags separately and then, when both are true, the number of inversion flags. The values in the brackets show the ratio between the number of flags and the sum of all universities with available data for the considered variable in the period 2011–2020 (i.e., the column *Total* in Tables 4 and 5). After that, Tables 8–10 for the variable Total academic personnel and Tables 11–13 for the variable total number of enrolled students report the years over which the error flags were raised.

Table 6. Total number of flags raised for variable total academic staff by countries.

	# of H1 Flags	# of H2 Flags	# of Inversions Flags	# of Jumps Flags	# of Recalculation Flags
Italy	159 (0.14)	287 (0.25)	40 (0.03)	396 (0.35)	58 (0.05)
Germany	314 (0.08)	398 (0.10)	34 (0.01)	1059 (0.27)	32 (0.01)
Spain	24 (0.03)	81 (0.10)	4 (0.005)	249 (0.31)	21 (0.03)
France	18 (0.02)	20 (0.02)	1 (0.00)	160 (0.14)	5 (0.004)
Poland	79 (0.08)	71 (0.07)	12 (0.01)	9 (0.01)	18 (0.02)
Portugal	50 (0.05)	131 (0.14)	7 (0.01)	236 (0.25)	32 (0.03)

Table 7. Total number of flags raised for variable total enrolled students by countries.

	# of H1 Flags	# of H2 Flags	# of Inversions Flags	# of Jumps Flags	# of Recalculation Flags
Italy	151 (0.08)	192 (0.1)	18 (0.01)	728 (0.36)	111 (0.06)
Germany	150 (0.04)	553 (0.14)	24 (0.01)	1452 (0.37)	99 (0.03)
Spain	35 (0.04)	86 (0.11)	5 (0.01)	289 (0.36)	17 (0.02)
France	96 (0.04)	638 (0.24)	27 (0.01)	795 (0.3)	193 (0.07)
Poland	85 (0.03)	428 (0.16)	12 (0)	871 (0.33)	41 (0.02)
Portugal	23 (0.02)	166 (0.17)	5 (0.01)	256 (0.27)	13 (0.01)

Table 8. Number of inversion flags raised by country and by year (2011–2020).

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Italy	5	1	0	0	2	3	4	4	5	16	40
Germany	9	1	1	0	2	2	1	3	2	13	34
Spain	4	0	0	0	0	0	0	0	0	0	4
France	0	0	0	0	0	0	0	0	0	1	1
Poland	0	0	0	0	0	0	0	9	2	1	12
Portugal	2	0	0	0	2	0	0	2	0	1	7

Table 9. Number of anomalous jump flags raised by country and by delta.

	$\Delta 2011-2012$	$\Delta 2012-2013$	$\Delta 2013-2014$	$\Delta 2014-2015$	$\Delta 2015-2016$	$\Delta 2016-2017$	$\Delta 2017-2018$	$\Delta 2018-2019$	$\Delta 2019-2020$	Total
Italy	42	52	53	42	39	40	38	42	48	396
Germany	132	144	106	111	115	110	112	101	128	1059
Spain	26	21	15	58	31	30	19	26	23	249
France	102	5	6	16	0	0	12	10	9	160
Poland	1	1	1	1	1	1	1	1	1	9
Portugal	31	28	25	30	34	17	20	32	19	236

Table 10. Number of recalculation flags raised by country and by year (2011–2020).

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Italy	N.A.	5	9	13	6	5	5	9	6	N.A.	58
Germany	N.A.	0	2	0	6	8	3	6	7	N.A.	32
Spain	N.A.	3	0	3	4	4	2	2	3	N.A.	21
France	N.A.	2	1	0	0	0	0	1	1	N.A.	5
Poland	N.A.	0	0	0	0	0	0	18	0	N.A.	18
Portugal	N.A.	0	2	0	6	8	3	6	7	N.A.	32

Table 11. Number of inversion flags raised by country and by year (2011–2020).

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Italy	6	1	1	0	0	0	0	0	2	8	18
Germany	16	1	0	2	0	3	1	1	0	0	24
Spain	4	0	0	0	0	0	0	0	0	1	5
France	16	2	4	0	0	0	0	0	0	5	27
Poland	6	0	0	2	1	0	0	0	0	3	12
Portugal	4	0	0	0	1	0	0	0	0	0	5

Table 12. Number of anomalous jump flags raised by country and by delta.

	$\Delta 2011$ – 2012	$\Delta 2012$ – 2013	$\Delta 2013$ – 2014	$\Delta 2014$ – 2015	$\Delta 2015$ – 2016	$\Delta 2016$ – 2017	$\Delta 2017$ – 2018	$\Delta 2018$ – 2019	Total
Italy	83	60	67	86	116	92	75	72	728
Germany	208	190	178	145	142	151	133	139	1452
Spain	32	39	49	39	40	21	18	28	289
France	94	77	75	302	0	0	118	107	795
Poland	124	139	106	83	88	158	54	50	871
Portugal	56	35	18	22	17	17	27	28	256

Table 13. Number of recalculation flags raised by country and by year (2011–2020).

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Italy	N.A.	10	9	6	15	33	19	7	12	N.A.	111
Germany	N.A.	13	11	12	10	17	8	13	15	N.A.	99
Spain	N.A.	4	3	5	1	0	2	1	1	N.A.	17
France	N.A.	19	17	49	0	0	0	103	5	N.A.	193
Poland	N.A.	5	7	7	3	13	4	0	2	N.A.	41
Portugal	N.A.	6	1	0	3	0	1	0	2	N.A.	13

As may be observed, the procedures were able to detect the described problems in every country, notwithstanding the great care taken in obtaining correct data from the different data providers. The values are often higher for Germany mainly because this country has a much larger number of HEIs. If we consider the same values divided by the number of HEIs in the country, we obtain a much more uniform distribution of the errors.

The results show a strong presence of recalculations in the dataset. This type of problem is strongly conditioned by the data collection method carried out by the ETER, which recomputes the values every year and may change from year to year in some of its definitions. Furthermore, one piece of information that unfortunately cannot be evaluated by only looking at the ETER concerns the various reforms of contractual forms that have taken place over the years in the different countries and the role conventions in the institutions (for example, in some countries like Italy, teaching assistants have been phased out as a contractual form).

The running times required by the proposed procedures, implemented in Microsoft Excel, are smaller than a few seconds for each whole country on a standard PC with i7 CPU and 16Gb of RAM running Microsoft Windows 11 OS.

3.2. Comparison with Other Existing Methods

This section contains a comparison of the proposed approach and other four error detection techniques available in the literature:

- (1) Local Outlier Factor (LOF)
- (2) Z-score threshold (Z-Score)
- (3) Interquartile Range threshold (IQR)
- (4) Hampel identifier (HI)

Since the other techniques cannot detect inversion or recalculation problems, our comparison is necessarily limited to the anomalous jump problem and is performed on a sample of Italian HEIs.

The above methods mainly operate by identifying outliers, and we apply them to the deltas to detect “outlier jumps”, roughly corresponding to our concept of an anomalous jump.

LOF compares the Local Readability Density (LRD) of a point to that of its neighbors. An LOF score of approximately 1 indicates that the LRD around the point is comparable to that of its neighbors, so the point is not an outlier. Points that have a substantially

lower LRD than their neighbors are more “far away” from the others. They are considered outliers if they produce a score lying outside an interval I_{LOF} . The minimum number of neighboring points considered was set to 3 to check each value with the two adjacent ones. The extremes of the I_{LOF} interval are computed as the average of the distance between each pair of normalized deltas in a time series (simply the difference of the two normalized values) plus or minus the standard deviation of those distances. If a time series has missing values, this technique cannot work, and the data unit is not checked.

Z-scores quantify how far from the mean an observation is when data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls in. To calculate a Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, $Z = (X - \mu)/\sigma$, where X is an observation, μ is the mean of the population and σ is the standard deviation of the population. The larger the Z-score, the more the value is different from the average, and values above a threshold are declared outliers. We use 2 as the threshold. The main limitation of this method regards the normal distribution assumptions: if data are not normally distributed, this approach might be not accurate.

The Interquartile Range threshold computes an interval I_{IQR} whose width is the difference between first and third quartiles $Q1$ and $Q3$ and whose extremes are given by $Q1 - 1.5(Q3 - Q1)$ and $Q3 + 1.5(Q3 - Q1)$. Deltas lying outside the interval I_{IQR} are considered outliers.

The Hampel Identifier [34] computes the median η of the deltas in a time series and their Median Absolute Deviation (MAD). Then, it computes an interval I_{HI} whose extremes are $\eta - 3MAD$ and $\eta + 3MAD$. Deltas lying outside the interval I_{HI} are considered outliers.

The following Figure 1 reports the results of the anomalous jump detection of our method, described in Section 2.2, and of the above-described four methods for the case of the “Total academic staff” from a sample of the first 26 Italian HEIs in the ETER dataset. The limited size of the sample is due to the fact that the real anomalous jumps, needed to evaluate the performance of each method, were not known in advance and had to be manually identified by experts in the field for the present comparison with time-consuming inspections.

As is observable, the accuracy, defined as the percentage of correct detections over all examined cases (sum of true positive and true negative cases), is the highest for our method (91%), followed by LOF (81%). Similar experiments on the whole Italian situation show that the proposed method almost always finds a larger number of cases, and in the manual controls, performed only on a subset of the alerted cases for obvious reasons of time, the proposed technique appears to have a very good discrimination power. Therefore, the overall outcome indicates that the proposed approach has a detection power of anomalous jumps that is at least comparable, when not superior, to that of each other single technique tested.

To offer another insight, in Figure 2, we focus on a single case, which is the series of the values for the variable total academic staff of a randomly extracted real unit (unit IT0010). Here, the situation can be inspected and judged by a human, and it appears that the steepest jumps are $\Delta 2016-2017$, $\Delta 2018-2019$ and $\Delta 2019-2020$, and they have been judged anomalous by experts in the field.

The results of each method are reported in the same figure, by coloring in azure the unalerted cases and, respectively, in yellow, red, green, purple and orange the deltas detected as anomalous jumps. As is observable, our approach is the only one correctly recognizing all three cases.

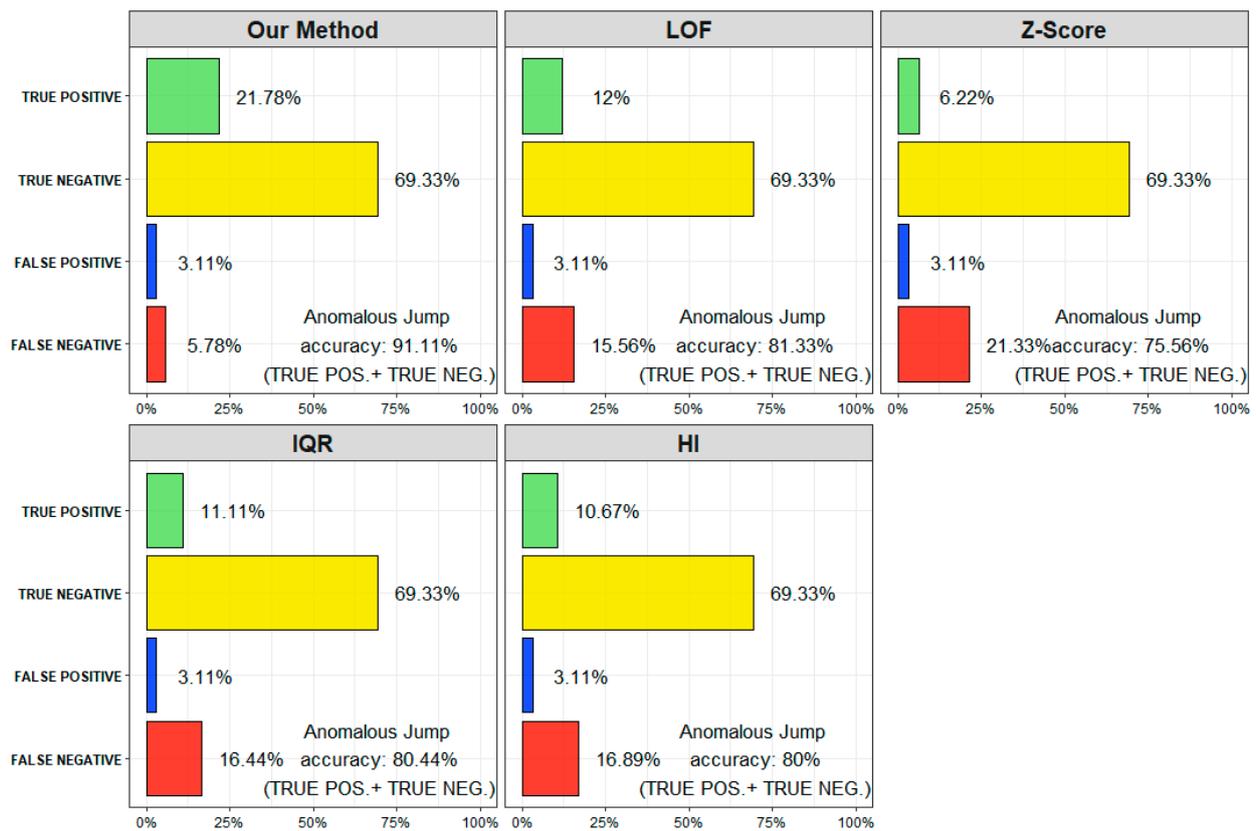


Figure 1. Percentages of anomalous jumps detected by each method.



Figure 2. Anomalous jumps detected by each method on a specific case.

4. Discussion

The issues addressed in this work arise from the analysis of large numerical longitudinal databases. This type of data is becoming more and more accessible, and they are now used for many important analyses. Unfortunately, they may contain errors, like almost any other type of data. In addition, to the generic errors commonly found in other types of data, longitudinal datasets often harbor subtle problems that generic techniques fail to trace. Those elusive types of errors should be automatically identified, so that they can be inspected and possibly removed by a human inspector. This work responds to this need by identifying three types of those errors and by proposing a statistical-mathematical approach to capture their complexity that can be applied after other generic techniques. In particular, we have identified the following types of errors:

- (1) The inversion problem, that is, the swapping of one or more values between two neighboring units;
- (2) The anomalous jump, that is, the presence of a jump between two consecutive values in a time series with the size being out of the ordinary;
- (3) The recalculation problem, which happens when the data provider discovers an error (typically a timing attribution error) on an already published value and operates a recalculation on the next value to compensate for the previous errors.

This list could even be extended in future studies. We devised techniques to identify potential errors of the described types that can be applied after a generic error correction step. We wanted these techniques to possess the following features: be computationally viable even for large datasets; work at the formal level, regardless of the meaning of the data, to be used in several contexts; be flexible to adapt to different situations. The proposed techniques are based on a system of indicators and have been implemented in a Microsoft Excel spreadsheet, publicly available in [29] from the Mendeley Data repository, to favor transparency and replicability of our experiments and to provide an easily accessible tool for anybody interested in using the proposed techniques on other datasets. We applied these techniques to an important example of a large longitudinal database, the ETER database, gathered from the different European countries and obtained by means of several passages. In this case, notwithstanding the great care spent in improving the quality of the data, several cases of the described problems were found by the proposed techniques. Thus, thanks to the described approach, the data quality of the dataset could be further improved.

5. Conclusions

When dealing with large numerical longitudinal databases, there exist errors specific to this type of dataset that are hardly identifiable or not identifiable at all by standard error detection and correction techniques. This work identifies some of these problems and proposes a statistical-mathematical approach based on a system of indicators that is able to capture the complexity of the described problems by working at the formal level, regardless of the specific meaning of the data. In particular, the types of errors analyzed in this work are as follows: (i) the inversion of one or more values from one unit to another; (ii) anomalous jumps in the series of values; (iii) errors in the temporal attribution of the values due to a recalculation operated by the data providers to compensate previous errors. The techniques to detect such errors were implemented in MS Excel and applied to the important database of European Higher Education institutions (ETER) by analyzing two key variables, namely, the total academic staff and the number of enrolled students. Note that these variables are two of the most important and delicate ones, and special care should be devoted to their correction. Each of them is often used in empirical analysis as a proxy for the size of the institutions, and they are also two of the main variables considered by policymakers at the European and national levels.

Empirical results show the effectiveness of the proposed techniques and the computational viability of the approach. Comparison with other existing techniques, which is possible only for the anomalous jump problem, reveals a superior detection power of our approach, whose accuracy on a sample reaches 91%. The implementation of the approach

in Microsoft Excel makes it easy to use for researchers and functionaries working with large longitudinal databases. Moreover, it ensures the replicability of the approach and its applicability in other contexts.

Future work includes the identification of further cases of longitudinal data-specific errors and the development of techniques for their localization. With regard to the ETER dataset, we plan to extend the described techniques to other variables. This work has considerable applications and extensions. For example, it could be suggested to the national data collectors of ETER (typically the National Statistic offices) to use our approach to perform checks, already at the national level, and correct the data before sending them to ETER to maximize the accuracy of the data provided. The Excel tool provided in the work allows an easy implementation of our method on the variables of interest before providing the data.

Author Contributions: Conceptualization, R.B. and C.D.; methodology, R.B.; software, S.D.L.; writing R.B., C.D. and S.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Sapienza research grants RM120172B870E2E2 and RM12117A8A5DBD18.

Data Availability Statement: The European Tertiary Education Register (ETER) is available from the ETER project website: <https://www.eter-project.com/#/home> (accessed on 10 February 2024) The Microsoft Excel file of the implementation of the proposed techniques is available from: Bruni, R., Daraio, C.; Di Leo, S. (2024), "A detection tool for longitudinal data specific errors applied to the case of European universities", Mendeley Data, V1, doi: 10.17632/syyc7t4z54.

Acknowledgments: We thank Benedetto Lepori and Daniel Wagner-Schuster for useful discussions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. OECD. *Quality Framework and Guidelines for OECD Statistical Activities*; OECD Publishing: Paris, France, 2011.
2. Daraio, C.; Iazzolino, G.; Laise, D.; Coniglio, I.M.; Di Leo, S. Meta-choices in ranking knowledge-based organizations. *Manag. Decis.* **2021**, *60*, 995–1016. [CrossRef]
3. Ballou, D.P.; Pazer, H.L. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Manag. Sci.* **1985**, *31*, 150–162. [CrossRef]
4. Pipino, L.L.; Lee, Y.W.; Wang, R.Y. Data quality assessment. *Commun. ACM* **2002**, *45*, 211–218. [CrossRef]
5. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
6. Wang, R.Y.; Ziad, M.; Lee, Y.W. *Data Quality*; Springer Science & Business Media: Berlin, Germany, 2006; Volume 23.
7. Sadiq, S. (Ed.) *Handbook of Data Quality: Research and Practice*; Springer Science & Business Media: Berlin, Germany, 2013.
8. Carlo, B.; Daniele, B.; Federico, C.; Simone, G. A Data Quality Methodology for Heterogeneous Data. *Int. J. Database Manag. Syst.* **2011**, *3*, 60–79. [CrossRef]
9. Batini, C.; Scannapieco, M. *Data and Information Quality*; Springer International Publishing: Cham, Switzerland, 2016.
10. Corrales, D.C.; Corrales, J.C.; Ledezma, A. How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. *Symmetry* **2018**, *10*, 99. [CrossRef]
11. Corrales, D.C.; Ledezma, A.; Corrales, J.C. From Theory to Practice: A Data Quality Framework for Classification Tasks. *Symmetry* **2018**, *10*, 248. [CrossRef]
12. Liu, C.; Peng, G.; Kong, Y.; Li, S.; Chen, S. Data Quality Affecting Big Data Analytics in Smart Factories: Research Themes, Issues and Methods. *Symmetry* **2021**, *13*, 1440. [CrossRef]
13. Bruni, R. Error correction for massive datasets. *Optim. Methods Softw.* **2005**, *20*, 297–316. [CrossRef]
14. Bruni, R.; Daraio, C.; Aureli, D. Imputation techniques for the reconstruction of missing interconnected data from higher Educational Institutions. *Knowl.-Based Syst.* **2020**, *212*, 106512. [CrossRef]
15. Alwin, D. *The Margins of Error: A Study of Reliability in Survey Measurement*; Wiley-Blackwell: Hoboken, NJ, USA, 2007.
16. Saris, W.E.; Gallhofer, I.N. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*; Wiley: Hoboken, NJ, USA, 2007; ISBN 9780470114957.
17. Cernat, A.; Oberski, D. Estimating Measurement Error in Longitudinal Data Using the Longitudinal MultiTrait MultiError Approach. *Struct. Equ. Model. A Multidiscip. J.* **2022**, *30*, 592–603. [CrossRef]
18. Wang, X.; Wang, C. Time Series Data Cleaning: A Survey. *IEEE Access* **2019**, *8*, 1866–1881. [CrossRef]
19. Hawkins, D.M. *Identification of Outliers*; Chapman and Hall: London, UK, 1980; Volume 11.
20. Blázquez-García, A.; Conde, A.; Mori, U.; Lozano, J.A. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Comput. Surv.* **2021**, *54*, 1–33. [CrossRef]

21. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
22. Yang, J.; Rahardja, S.; Fränti, P. Outlier detection: How to threshold outlier scores? In Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing, Sanya, China, 19–21 December 2019; pp. 1–6.
23. Grubbs, F.E. Sample Criteria for Testing Outlying Observations. *Ann. Math. Stat.* **1950**, *21*, 27–58. [[CrossRef](#)]
24. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422.
25. Brockwell, P.J.; Davis, R.A. *Time Series: Theory and Methods*; Springer Science & Business Media: Berlin, Germany, 1991.
26. Oberski, D.L.; Kirchner, A.; Eckman, S.; Kreuter, F. Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models. *J. Am. Stat. Assoc.* **2017**, *112*, 1477–1489. [[CrossRef](#)]
27. Pavlopoulos, D.; Pankowska, P.; Bakker, B.; Oberski, D. Modelling error dependence in categorical longitudinal data. In *Measurement Error in Longitudinal Data*; Oxford University Press: Oxford, UK, 2021. [[CrossRef](#)]
28. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **2009**, *41*, 1–52. [[CrossRef](#)]
29. Bruni, R.; Daraio, C.; Di Leo, S. A Detection Tool for Longitudinal Data Specific Errors Applied to the Case of European Universities. Mendeley Data, V1. 2024. Available online: <https://data.mendeley.com/datasets/syyc7t4z54/1> (accessed on 23 February 2024).
30. ETER Project Website. Available online: <https://www.eter-project.com/#/home> (accessed on 23 February 2024).
31. Daraio, C.; Bonaccorsi, A.; Geuna, A.; Lepori, B.; Bach, L.; Bogetoft, P.; Cardoso, M.F.; Castro-Martinez, E.; Crespi, G.; de Lucio, I.F.; et al. The European university landscape: A micro characterization based on evidence from the Aquameth project. *Res. Policy* **2011**, *40*, 148–164. [[CrossRef](#)]
32. Lepori, B.; Bonaccorsi, A.; Daraio, A.; Daraio, C.; Gunnes, H.; Hovdhaugen, E.; Ploder, M.; Scannapieco, M.; Wagner-Schuster, D. *Establishing a European Tertiary Education Register*; Publications Office of the European Union: Luxembourg, 2016; ISBN 978-92-79-52368-7. [[CrossRef](#)]
33. Daraio, C.; Bruni, R.; Catalano, G.; Daraio, A.; Matteucci, G.; Scannapieco, M.; Wagner-Schuster, D.; Lepori, B. A Tailor-made Data Quality Approach for Higher Educational Data. *J. Data Inf. Sci.* **2020**, *5*, 129–160. [[CrossRef](#)]
34. Hampel, F.R. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **1974**, *69*, 383–393. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.