Flexible Techniques to Detect Typical Hidden Errors in Large Longitudinal Datasets

Renato Bruni , Cinzia Daraio , Simone Di Leo *

Posted Date: 1 March 2024

doi: 10.20944/preprints202403.0012.v1

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

 doi:10.20944/preprints202403.0012.v1

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 of the Sapienza University of Rome; 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 same units at different times) requires the development of flexible techniques to automatically detect common errors in such data. Besides obvious and easily identifiable cases, such as missing or out-of-range data, large longitudinal dataset often present problems not easily traceable by the techniques used for generic datasets. In particular, elusive and baffling problems are i) inversion of one or more values from one unit to another; ii) anomalous jumps in the series of values, iii) errors in the timing of the values due to a recalculation operated by the data providers to compensate previous errors. This work 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. The proposed approach identifies suspect erroneous data and is applicable in a variety of contexts. We implement this approach in a relevant database of European Higher Education institutions (ETER) by analyzing Total academic staff, that is one of the most important variables, used in empirical analysis as proxy of size and also considered by policy makers at European level.

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 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 the data quality should consider that the very concept of data quality is not one-dimensional but multidimensional [3,4]. In particular, the following seven dimension are usually identified: accuracy, completeness, consistency, validity, timeliness, uniqueness, and integrity And even though the names of those dimensions may vary in literature, the same key concepts are widely recognized (mainly rooted in the seminal paper [5]).

Due to the relevance of the issue, many authors have proposed methods or guidelines to assess problems on data quality [6–13]. However, few works focus on the problems that specifically regards the case of numerical data describing repeated observations of the same units over a period of time. This type of data are often called *longitudinal* data, or also 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,

 doi:10.20944/preprints202403.0012.v1

one ubiquitous problem affecting almost all data-related applications is the presence of errors in the data. Unfortunately, longitudinal data make no exception in this. Thus, when data containing errors are used for some study, 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* form 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. The easiest identifiable cases are for example missing values or out-of-range values. Due to the spread of this problem, several techniques have been developed in different fields of science to cope with similar situations. There exist several *imputation* techniques for the reconstruction of missing or out-of-range values, see for example [14,15]. Some method for estimating measurement errors in Longitudinal Data are based on latent variable modelling [16,17]. Another technique, called the MultiTrait MultiError approach, is presented in [18] to estimate multiple types of errors concurrently using a combination of experimental design and latent variable modelling.

Other methods are based on data integration, when the data under analysis are also contained or derivable from different sources [19,20].

However, in addition to the "usual" types of errors, large longitudinal dataset may contain peculiar types of errors that are not easily identifiable by the techniques used for finding and/or correcting errors into 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* are inserted in the space of *A*. We call this situation ***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 of too low for unit *A*, the generalist imputation techniques will probably try to reconstruct the correct value based on elaborations involving unit A, and 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, if we discover that such a variable has a high volatility, the situation can also be normal after all. We call this situation ***anomalous jump***. In this case, we need to identify some threshold above which the values should be considered erroneous. This is a very delicate issue, and standard error detection techniques are often 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 are actually belonging 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 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.

This work proposes a statistical-mathematical approach based on a system of indicators that define a rational process to assess and improve the quality of data (as suggested by [21]. In particular, the proposed approach is able to identify suspect erroneous data suffering from the described problems by working at the formal level, regardless of the specific meaning of the data. Therefore, it is applicable in a variety of contexts. Moreover, our approach contains a certain degree of flexibility,

 doi:10.20944/preprints202403.0012.v1

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 variable "Total academic staff". This is one of the most important variables, used in empirical analysis as a proxy of the size of the institutions and also considered by the policy makers at European level.

2. Materials and Methods

As explained in Section 1, in large numerical longitudinal databases we identify the following three main consistency problems that are specific to the case of longitudinal data and are difficult to treat 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. The method consists of several steps for each of the above three problems, detailed in Sections 2.1, 2.2 and 2.3. As materials, we conducted our experiments on the ETER database, described in section 2.4.

2.1. Inversion Problem

To identify inversion problem between two units *A* and *B*, we evaluate two types of conditions that we call here *H*1 and *H*2. The first type (*H*1) consists in 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 *H*1 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...m = U. Unit *i* has values of a variable (or attribute) *v* over several time instants t=1...n=S. Define now $\Delta v_i^{(t,t+1)}$ as the difference (*delta*) between the two values assumed by unit *i* in two consecutive time instants *t*, *t*+1 for variable *v*, that is:

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

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 the $\Delta v_i^{(n,n+1)}$ 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 a for each unit *i* the value DV_i defined as the modulus of the product between the sums of the positive deltas and the sum of negative deltas:

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

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 all collapsing to zero when the intrinsic variability of a unit must be nonnegative. 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*s in the entire dataset considered:

 doi:10.20944/preprints202403.0012.v1

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

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 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 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 all collapsing to zero when this is not acceptable.

H1.f. Now, to compute a reasonably 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 inversion of a value between two consecutive units *A* and *B* by computing *H*1, we need that four conditions are 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 *H*1 is given by the following boolean expression:

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

If $H1_{(A,B)}^t$ is true, then to have a probable swap problem we also need a corresponding condition $H2_{(A,B)}^t$ to be true. The generic condition *H*2 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*:

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

H2.b. We define now 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 taking the values of the subsequent unit *i*+1 (the one with which the values could have been exchanged):

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

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, that means the minimum is $|N_i^t|$, then inversion is probable.

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

 doi:10.20944/preprints202403.0012.v1

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

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

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: $|(-112-10)(107+10)| = 14274$ for unit 1 and $|(-129)(120+5+5)| = 16770$ 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 thresholds T is 32.17 for unit 1 and 25.59 for unit 2.

Now 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.

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

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

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, $F_1{}^{2} = 6$, thus $F_1{}^{2} \neq I_1{}^{2}$ 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 that before, obtained as follows. After the computation of the threshold T_i described in Section 2.1, we execute the following steps.

a. 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.
b. Compute VI_i as the integer upper part of the value LGM_i plus a constant c representing another element of customization of the procedure. This value can be determined either with a priori reasoning or even derived from the data itself.

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

c. Compute GMT_i as the sum of $GM_i + T_i$. In practice, we are summing 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 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.

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

 doi:10.20944/preprints202403.0012.v1

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 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.

As $|\Delta v_3^{(2,3)}| = 280 > 177.15$ and $|\Delta v_3^{(3,4)}| = 290 > 177.15$, we report an anomalous jump flag for the period t=2,3 and for 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.

Table 2. Values and **Δ** of the unit considered for the Anomalous Jump example.

	$\mathbf{v^1}$	$\mathbf{v^2}$	$\mathbf{v^3}$	$\mathbf{v^4}$	$\mathbf{v^5}$	$\mathbf{\Delta^{(1,2)}}$	$\mathbf{\Delta^{(2,3)}}$	$\mathbf{\Delta^{(3,4)}}$	$\mathbf{\Delta^{(4,5)}}$
Unit 3	200	220	500	210	230	-20	-280	290	-20

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 sign are both above the threshold TT_i in modulus:

$$[(\Delta v_i^{(t-1,t)} > 0 \ \wedge \ \Delta v_i^{(t,t+1)} < 0) \vee (\Delta v_i^{(t-1,t)} < 0 \ \wedge \ \Delta v_i^{(t-1,t)} > 0)] \ \wedge \ (|\Delta v_i^{(t-1,t)}| > TT_i \ \wedge \ |\Delta v_i^{(t,t+1)}| > TT_i) \quad (9)$$

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

Example 3. We provide an example of 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.

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

All the described operations are available in the Microsoft Excel file contained in [22]. This file can be used to operate the described checks with any data, by simply pasting them in the sheet "Main Table". Each row must represent a single unit of analysis. The excel file is also adaptable to use units with 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) [23] is a key initiative for understanding the higher education landscape in Europe developed after the successful AQUAMETH project [24,25]. 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 [26].

As of March 2024, ETER includes 41 European countries and provides data from 2011 to 2020, with a total of over 3,500 HEIs. ETER collects a wide range of data on HEIs, including: institutional

 doi:10.20944/preprints202403.0012.v1

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 [27]. However, as described in Section 3, the proposed techniques were able to locate several cases of the specific longitudinal data problems described above.

ETER provides comprehensive documentation of the methodologies used and the data collection processes, ensuring transparency and replicability. ETER contributes to a better understanding of the higher education landscape and is a valuable resource for researchers, policy makers and stakeholders in European higher education. Within ETER, we selected the case of the variable Total academic personnel in headcount (HC) because it is widely used in empirical analysis and by policy makers as a proxy for the size of the universities. Therefore, it is one of the most important variables, and it is of paramount importance to detect any possible errors on that. Total academic personnel in HC, according to the ETER manual, includes

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, and
iv) PhD students employed for teaching assistance or research.

We report our experiments on the largest EU countries present in ETER, i.e., Germany, France, Italy, Spain, Poland and Portugal, for a total of 1587 HEIs, in the time period from 2011 to 2020. Table 4 shows the subdivision by country. Table 5 reports the number of HEIs having complete data for each year.

Table 4. Number of HEIs available in ETER for each country in the period 2011-2020.

	HEIs available in ETER
Italy	219
Germany	424
Spain	84
France	417
Poland	314
Portugal	129

Table 5. Number of HEIs with the variable total academic staff (HC) available in 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

 doi:10.20944/preprints202403.0012.v1

Poland	0	0	0	0	0	0	247	243	241	237	968
Portugal	113	106	94	91	90	95	90	90	89	92	950

3. Results

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 variable Total academic personnel in headcount. 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 is zero, it was replaced with 1 to avoid all collapsing to zero. In the computation of VI_i the constant c was set at 8 by means of an experimental fine tuning. Table 6 reports, for each Country, the total number of flags raised by the described techniques. In particular, we indicate $H1$ 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 on academic staff in the period 2011-2020 (i.e., the column *Total* in Table 5). Tables 7–9 report the years over which the error flags were raised.

As observable, 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 higher for Germany mainly because that 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 jump anomalies in the dataset. This type of problem is strongly conditioned by the data collection method carried out by 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 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).

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. 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

 doi:10.20944/preprints202403.0012.v1

Portugal	2	0	0	0	2	0	0	2	0	1	7

Table 8. Number of anomalous jump flags raised by country and by delta.

	Δ2011-2012	Δ2012-2013	Δ2013-2014	Δ2014-2015	Δ2015-2016	Δ2016-2017	Δ2017-2018	Δ2018-2019	Δ2019-2020	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 9. 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

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. We have identified the following problems: i) inversion of one or more values from one unit to another; ii) anomalous jumps in the series of values, iii) errors in the timing of the values due to a recalculation operated by the data providers to compensate previous errors. This list could even be extended in future studies. We devised techniques to identify the potential errors, based on a system of indicators. These techniques were wanted 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. These techniques have been implemented in a Microsoft Excel spreadsheet, publicly available in [23] 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 on an important example of 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 still found by the proposed techniques. Thus, thanks to the described approach, the data quality of the dataset could be further improved.

 doi:10.20944/preprints202403.0012.v1

5. Conclusions

When dealing with large numerical longitudinal databases, there exist errors specific for this type of datasets that are not recognized by standard error detection and correction techniques. This work 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. 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 Total academic staff. This variable is one of the most important and delicate ones, it is often used in empirical analysis as a proxy of the size of the institutions, and it is also one of the main variables considered by policy makers at the European and national level. Empirical results show the effectiveness of the proposed techniques and the computational viability of the approach. 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.

Author Contributions: Conceptualization, Bruni R. and Daraio C.; methodology, Bruni R.; software, Di Leo S.; writing Bruni R., Daraio C. and Di Leo S. 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 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 discussion.

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

References

1. *Quality framework and guidelines for OECD statistical activities*; OECD Publishing, 2011.
2. Daraio, C., Iazzolino, G., Laise, D., Coniglio, I.M., Di Leo, S. Meta-choices in ranking knowledge-based organizations. *Management Decision* **2022**, *60*, 995–1016. https://doi.org/10.1108/MD-01-2021-0069
3. Ballou, D.P., & Pazer, H.L. Modeling data and process quality in multi-input, multi-output information systems. *Management science* **1985**, *31*, 150–162. https://doi.org/10.1287/mnsc.31.2.150
4. Pipino, L.L., Lee, Y.W., Wang, R.Y. Data quality assessment. *Communications of the ACM* **2002**, *45*, 211–218. https://doi.org/10.1145/505248.506010
5. Wang, R.Y., Strong, D.M. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* **1996**, *12*, 5–33. https://doi.org/10.1080/07421222.1996.11518099
6. Wang, R.Y., Ziad, M., Lee, Y.W. *Data quality*; Springer Science & Business Media, 2006; Volume 23.
7. Sadiq, S. (Ed.) *Handbook of data quality: Research and practice*; Springer Science & Business Media: 2013.
8. Batini, C., Barone, D., Cabitza, F., Grega, S. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems* **2011**, *3*, 60–79. https://doi.org/10.5121/ijdms.2011.3105
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. https://doi.org/10.3390/sym10040099
11. Corrales, D.C., Ledezma, A., & Corrales, J.C. From theory to practice: A data quality framework for classification tasks. *Symmetry* **2018**, *10*, 248. https://doi.org/10.3390/sym10070248
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. https://doi.org/10.3390/sym13081440
13. Daraio, C., Di Leo, S., Scannapieco, M. Accounting for quality in data integration systems: a completeness-aware integration approach. *Scientometrics* **2022**, *127*, 1465–1490. https://doi.org/10.1007/s11192-022-04266-0

 doi:10.20944/preprints202403.0012.v1

14. Bruni, R. Error Correction for Massive Data Sets. *Optimization Methods and Software* **2005**, *20*, 295–314. https://doi.org/10.1080/10556780512331318281.
15. Bruni, R., Daraio, C., Aureli, D. Imputation techniques for the Reconstruction of Missing Interconnected Data from higher Educational Institutions. *Knowledge-Based Systems* **2021**, *212*, 106512. https://doi.org/10.1016/j.knosys.2020.106512
16. Alwin, D. The *margins of error: A study of reliability in survey measurement*; Wiley-Blackwell, 2007.
17. Saris, W., Gallhofer, I. *Design, evaluation, and analysis of questionnaires for survey research*; Wiley-Interscience, 2007.
18. Cernat, A., Oberski, D. Estimating Measurement Error in Longitudinal Data Using the Longitudinal MultiTrait Multi Error Approach. *Structural Equation Modeling: A Multidisciplinary Journal* **2023**, *30*, 592–603. https://doi.org/10.1080/10705511.2022.2145961.
19. Oberski, D. L., Kirchner, A., Eckman, S., Kreuter, F. Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association* **2017**, *112*, 1477–1489. https://doi.org/10.1080/01621459.2017.1302338
20. Pavlopoulos, D., Pankowska, P., Bakker, B., Oberski, D. Modelling error dependence in categorical longitudinal data. In *Measurement error in longitudinal data*; Oxford University Press, 2021. https://doi.org/10.1093/oso/9780198859987.003.0008
21. Batini, C., Cappiello, C., Francalanci, C., Maurino, A. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* **2009**, *41*, 1–52.
22. Bruni, R., Daraio, C.; Di Leo, S. A detection tool for longitudinal data specific errors applied to the case of European universities. *Mendeley Data* **2024**, V1. https://doi.org/10.17632/syyc7t4z54
23. ETER Project Website. Available online: https://www.eter-project.com/#/home (accessed on 23 February 2024).
24. Bonaccorsi, A., Daraio, C. (Eds.) *Universities and strategic knowledge creation: Specialization and performance in Europe*; Edward Elgar Publishing, 2007.
25. Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., ... Eeckaut, P. V. The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy* **2011**, *40*, 148–164. https://doi.org/10.1016/j.respol.2010.10.009
26. Lepori, B., et al. *Establishing a European tertiary education register*; Publications Office of the European Union, 2016; ISBN 978-92-79-52368-7. https://doi.org/10.2766/755061
27. 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. *Journal of Data and Information Science* **2020**, *5*, 129–160. https://doi.org/10.2478/jdis-2020-0029