Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Imputation techniques for the reconstruction of missing interconnected data from higher Educational Institutions

Renato Bruni [a,*], Cinzia Daraio [a], Davide Aureli [b]

[a] Department of Computer Control and Management Engineering, "Sapienza" University of Rome, Rome, Italy
[b] Department of Information Engineering, Electronics and Telecommunications, "Sapienza" University of Rome, Rome, Italy

ARTICLE INFO

Article history:
Received 19 June 2020
Received in revised form 10 September 2020
Accepted 6 October 2020
Available online xxxx

Keywords:
Data imputation
Information reconstruction
Machine learning
Educational Institutions

ABSTRACT

Educational Institutions data constitute the basis for several important analyses on the educational systems; however they often contain not negligible shares of missing values, for several reasons. We consider in this work the relevant case of the European Tertiary Education Register (ETER), describing the Educational Institutions of Europe. The presence of missing values prevents the full exploitation of this database, since several types of analyses that could be performed are currently impracticable. The imputation of artificial data, reconstructed with the aim of being statistically equivalent to the (unknown) missing data, would allow to overcome these problems. A main complication in the imputation of this type of data is given by the correlations that exist among all the variables. We propose several imputation techniques designed to deal with the different types of missing values appearing in these interconnected data. We use these techniques to impute the database. Moreover, we evaluate the accuracy of the proposed approach by artificially introducing missing data, by imputing them, and by comparing imputed and original values. Results show that the information reconstruction does not introduce statistically significant changes in the data and that the imputed values are close enough to the original values.

1. Introduction

Organizations providing *higher level education*, such as traditional universities, universities of applied sciences, polytechnics, community colleges, liberal arts colleges, etc. are collectively called Higher Education Institutions (HEIs). The data describing each individual HEI (for instance number of students, number of graduates, etc.) are called HEI microdata. Their availability is essential to support an empirical-oriented and robust policy making in the dynamic landscape of educational systems [1]. A pioneer research project called Aquameth [2] was the first attempt to gather microdata about European HEIs. Since this first experience, the presence of missing or noisy data and the lack of comparability among data appeared to be the most critical obstacles to an appropriate usage of the collected data [3]. After the Aquameth project, a large scale study called EUMIDA was commissioned by the European Union from 2009 to 2011, and showed the feasibility of a European-level data collection on individual HEIs [1]. Since then, it has been underlined the need to build a register of Higher Education Institutions in Europe.

At present, the European Tertiary Education Register (ETER) is a database collecting information on European HEIs, concerning their basic characteristics and geographical position, number of students, graduates, doctorates, staff, fields of education, income, expenditure and research activities. The main feature of ETER is providing data at the level of individual institutions, being therefore complementary to the educational statistics at the country and regional level provided by EUROSTAT. ETER is a European Commission initiative, and constitutes an Erasmus+ project fully financed by the European Commission. This project is managed by the Joint Research Centre and by the Directorate General for Education and Culture of the European Commission, in cooperation with EUROSTAT and the National Statistical Authorities of the participating Countries. ETER covers EU-27 countries (Austria, Belgium, Bulgaria, Croatia, Republic of Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden), as well as Albania, Iceland, Liechtenstein, Montenegro, Norway, Serbia, Switzerland, Turkey, United Kingdom and the Republic of North Macedonia. At the time of writing, data have been collected from the year 2011 (academic year 2011/2012) until 2016 or, occasionally, 2017.

The data are gathered through the National Statistical Authorities of the different Countries, and not directly collected by the project consortium. Notwithstanding the considerable effort

The code (and data) in this article has been certified as Reproducible by Code Ocean:https://help.codeocean.com/en/articles/1120151-code-ocean-s-verification-process-for-computational-reproducibility. More information on the Reproducibility Badge Initiative isavailable at www.elsevier.com/locate/knosys.

[*] Corresponding author.
E-mail address: bruni@diag.uniroma1.it (R. Bruni).

https://doi.org/10.1016/j.knosys.2020.106512
0950-7051/© 2020 Elsevier B.V. All rights reserved.

in the data collection, the current ETER database includes many scattered missing values in the variables and this creates problems for the usage of those variables. In particular, it clearly does not allow the micro analysis of the institutions containing the missing values, and also prevents the macro analysis (at the aggregate level) of many categories of institutions, whenever they include the incomplete ones. Thus, the main goal of this work is to propose a methodology to reconstruct those missing values for the key variables of the institutions.

Imputation techniques have been studied in many fields to tackle the widespread problem of missing data, see e.g. [4–6]. Consequently, many approaches to the problem have been proposed, based on several different techniques, including: k-nearest neighbors [7] or other forms of proximity search [8], fuzzy clustering [9,10], bagging algorithms [11], ensemble of neural networks [12], autoencoders neural networks [13], denoising autoencoders [14], other deep neural networks [15], Bayesian networks [16,17], integer linear programming [18], pattern sequence forecasting [19], use of a knowledge base [20], similarity rules [21], each of which possibly combined or hybridized with additional techniques. Existing methods to deal with missing data can be organized according to different perspectives. For example, one may discriminate on the basis of the source of the imputed information, which can be: (a) the other variables of the same unit under imputation, taking advantage of some existing relations among the variables; (b) other units similar to the treated one, which are often called *donors*; (c) other sources, external to the dataset under imputation, but storing the same information, e.g. statistical ledgers. Donors are generally complete units [22,23], though even incomplete ones could be used if necessary [24], and their use often requires the solution of optimization problems [25]. Another discrimination sometimes adopted is between statistical methods and machine learning approaches [8].

Currently, there is no evidence of a single "best" imputation method. Rather, it appears that the performance of each method depends in large measure on the dataset characteristics, as also suggested in [26]. The imputation of missing data in time series is a particularly difficult task [27], and many general techniques are not able to satisfactory deal with this case. And the subcase of multivariate time series stays at the core of the most challenging tasks, as observed in [16,28].

Educational data have important peculiarities. To begin with, they contain multivariate time series (number of students, number of graduates, etc.). Furthermore, it can be observed that almost all data of an institution are *interconnected*. The number of graduates is not independent from the number of students, the expenditure is not independent from the staff, just to make some easy examples. Thus, the imputation becomes particularly difficult: imputed values must belong to time series and each of them may impact on the situation of the whole institution. Therefore, imputation techniques for this type of data should be specifically designed. An approach to improve the data quality for the same ETER dataset, not dealing with missing data and so complementing the present work, has been developed in [29].

The proposed imputation methodology combines an interdisciplinary set of tools coming from information management, machine learning and statistics with an investigation of the relations existing among the variables and of the types of missing patterns actually contained in the data. The proposed methodology works at the formal level, with a data driven approach, i.e., learning from the data many aspects of the imputation techniques. Therefore, it could also be used to impute datasets with educational data from other origins, or even datasets with different meaning but sharing the features of time series and interconnection among the variables. We validate our approach by comparing the statistical features of the data before and after imputation. Moreover, we evaluate its reconstruction accuracy by means of the following experiments: (1) we take a set of complete records and artificially introduce in them missing values; (2) we use the proposed imputation techniques to impute those missing values; (3) we compare the imputed values with the original known values, and we study the occurrence of significant differences.

The work is organized as follows. Section 2 explains the variables and the different types of missing values that appear in the current version of ETER. Sections 3 and 4 introduce the proposed imputation techniques, distinguishing between those based on functions of the available values of the same institution and those based on other institutions (donors). Section 5 shows the results of the application of the proposed methodology to reconstruct the selected ETER variables. Section 6 describes the tests of accuracy carried out to further evaluate the performance of the proposed methodology. Section 7 draws conclusions.

2. Distribution of the missing values

ETER database is composed of 3208 units, each representing a single European HEI over a number of years. Each institution I_j contains a number of values s^j_{vk}, where $v \in V$ is the index of the variable (e.g., students), $k \in T$ that of the year (e.g., 2011) and $j \in J$ that of the institution (e.g., Sapienza University). The target of our imputation is constituted of the following 8 variables:

- Total number of students enrolled (called in ETER "Total Students Enrolled at ISCED 5-7")
- Total number of graduates (in ETER "Tot. Stud. Grad. at ISCED 5-7")
- Total number of Ph.D. students (in ETER "Tot. Stud. Enr. at ISCED 8")
- Total number of Ph.D. graduates (in ETER "Tot. Stud. Gr. at ISCED 8")
- Total academic staff (researchers and professors, measured in ETER either in "Full Time Equivalent - FTE" or in "Head Count - HC").
- Total non-academic staff (technical and administrative staff, measured in ETER either in "Full Time Equivalent - FTE" or in "Head Count - HC").
- Total current expenditure (measured in ETER in Euro).
- Total current revenues (measured in ETER in Euro).

These variables are selected because they are very relevant in many types of analyses. Moreover they are the main descriptors of the *size* of an institution. The importance of size in the characterization of the HEIs, and its connection to their overall performance, is well known in the specialized literature, see e.g. [30]. Each of the above variables takes a value for each year of the time horizon, which at present goes from 2011 to 2016 for the majority of the institutions (even if some may have less years, or some may have also 2017). To lighten the notation, when there is no ambiguity, we will denote the values of the *time series* for one generic variable v simply with $v_1, v_2, \ldots, v_t$ (without reference to the index j of the institution), and the set of years indices of the time series simply with $\{1, 2, \ldots, t\} = T$. For example, the values of Total graduates for the years 2011–2016 in an institution called"AAA", which in full notation are

$$\{v^{\mathrm{AAA}}_{\mathrm{graduates2011}},\ v^{\mathrm{AAA}}_{\mathrm{graduates2012}},\ \ldots,\ v^{\mathrm{AAA}}_{\mathrm{graduates2016}}\},$$

may be simply denoted by $\{grad_1, grad_2, grad_3, grad_4, grad_5, grad_6\}$.

A specificity of these data, with respect to other imputation cases, is given by the relations connecting all the above values. Indeed, the different values of each single time series are obviously related. For example, the number of students enrolled in

Table 1
Missing values in current ETER database.

	Institutions without missing in that variable	% on the total of institutions	Total number of missing in that variable
Total Students Enrolled	2206	69%	3075
Total Graduates	1963	61%	3891
Total PhD Students	891	28%	10349
Total PhD Graduates	844	26%	10741
Total Academic staff FTE	1629	51%	7915
Total Academic staff HC	1595	50%	7284
Total Non-academic staff FTE	1455	45%	8807
Total Non-academic staff HC	1341	42%	9145
Total Expenditure	987	31%	10041
Total Revenues	1021	32%	10252

Table 2
Missing values before and after the imputation for the variables Total Students and Total Graduates.

Variable	Total students		Total graduates	
Before or after imputation	Before	After	Before	After
Institutions without missing	2206	3153	1963	3108
% on the total	69%	98%	61%	97%
One isolated internal missing	98	0	157	0
One isolated extreme missing	68	14	259	29
Missing sequence of length 2	392	12	256	17
Missing sequence of length 3	212	3	339	6
Missing sequence of length 4	7	0	22	3
Missing sequence of length 5	37	26	44	27
Missing sequence of length 6	152	1	196	15
Missing sequence of length 7	52	0	66	6
Total number of missing	3075	183	3891	360

2011 is necessarily related to the number of students enrolled in 2012, in 2013, and so on, and in most of the cases the time series exhibits a *trend*, in the sense that, if some values are for instance increasing, it can be expected that the next value will be likely still increasing. However, trend may also change direction during the time series. Moreover, all the above variables are positively correlated. For example, the number of graduates is generally a certain proportion of the number of students, because they were indeed students in the previous year; the staff tends to increase or decrease in some measure with the number students, and so do the expenditures and the revenues.

We will denote the missing value with "$\sqcup$". We are interested in distinguishing the following types of missing values:

- *Isolated Internal Missing*: this type of missing occurs when $v_i = \sqcup$ for $2 \leq i \leq t-1$ and both $v_{i-1} \neq \sqcup$ and $v_{i+1} \neq \sqcup$.
- *Isolated Extreme Missing*: this type of missing occurs when $v_i = \sqcup$ for $i = 1$ or $i = t$ and respectively $v_2 \neq \sqcup$ or $v_{t-1} \neq \sqcup$.
- *Missing Sequence of length L*: this type of missing occurs when $v_i = \cdots = v_{i+L} = \sqcup$ for $i \leq t - L$ with $t - L \geq 1$ and $v_{i+L+1} \neq \sqcup$ (the missing sequence does not cover the whole time series).
- *Full Sequence Missing*: this type of missing occurs when $v_1 = \cdots = v_t = \sqcup$ (the missing sequence covers the whole time series).

The current situation of missing values is reported in Table 1. More details on the types of missing data are in Tables 2–6 in Section 5. Note that the most common type of missing values are those contained in the full sequence missing.

3. Imputation based on the available values of the sequence

This Section describes a category of imputation techniques based on functions of the *available values* in the sequence under imputation (i.e., the values which are not missing), starting with the very simple approaches based on average and linear regression that will be used as building blocks for the proposed Trend Smoothing technique. We also discuss, for each imputation technique, the type of missing for which it should be used.

3.1. Imputation based on weighted average

Very intuitive imputation techniques are based on averages of the values already available in the sequence. This approach relies on the assumption that the values in the time series are related, and not independent [31]. In general, there exist several mathematical types of averages. In our case, it appears reasonable to assume that, generally speaking, the closer the data are on the time scale, the most related their values are. Therefore, we propose to impute a value v_i for an isolated internal missing by using a *weighted arithmetic mean* of the surrounding values which are available, as follows.

$$v_i = \sum_{\substack{h \in T, \\ h \neq i}} w_h v_h, \qquad \text{with} \quad \sum_{\substack{h \in T, \\ h \neq i}} w_h = 1$$

The weights w_h should progressively decrease with the time distance between h and the instant i under imputation. We define yearly decrement d a positive value < 1 such that

$$\begin{array}{ll} w_h = d w_{h+1} & \forall h = 1, \ldots, i-1 \\ w_{h+1} = d w_h & \forall h = i+1, \ldots, t. \end{array}$$

For instance, if $d = 0.5$, $i = 4$ and we denote the value of w_3 by α, we have $w_1 = 1/2\ w_2 = 1/4\,\alpha$, $w_2 = 1/2\,\alpha$, $w_3 = \alpha$, $w_5 = \alpha$, $w_6 = 1/2\,\alpha$, and so on.

However, this condition alone is not enough to find the w_h. Thus, we propose to determine their values by using the following data-driven approach: for each value v_i of each institution j, with $i \in \{T \setminus 1\}$, we compute the variation δ_{ij} with respect to the preceding value v_{i-1} as follows:

$$\delta_{ij} = \frac{(v_i - v_{i-1})}{v_{i-1}}.$$

The set of these variations can now be studied to find its average value $\bar{\delta}$ by using the arithmetic mean, and two extreme values e_1, e_2 such that they contain 90% of the values of the variations. Then, we search for the yearly decrement d that better fits the average variation $\bar{\delta}$. By defining the set D of the indices of the institutions containing isolated missing values, and by assuming, with a slight abuse of notation, that i always represents the index of that missing value, this is obtained by solving the following optimization problem,

$$\min_{d \in (0,1)} \Big(\sum_{j \in D} (v_i^j(d) - \bar{\delta} v_{i-1}^j)^2 + \sum_{\forall i,j} (v_{i+1}^j - \bar{\delta} v_i^j(d))^2\Big),$$

where each $(v_i^j(d) - \bar{\delta} v_{i-1}^j)$ represents the difference between the imputed value $v_i^j(d)$, for which the dependence on d is made

Table 3
Missing values before and after the imputation for the variables Total PhD Students and Total PhD Graduates.

Variable	Total PhD students		Total PhD graduates	
Before or after imputation	Before	After	Before	After
Institutions without missing	891	2977	844	2980
% on the total	28%	93%	26%	93%
One isolated internal missing	86	41	121	44
One isolated extreme missing	118	55	114	51
Missing sequence of length 2	386	26	376	24
Missing sequence of length 3	333	24	151	24
Missing sequence of length 4	78	1	260	0
Missing sequence of length 5	117	28	119	28
Missing sequence of length 6	1014	5	1028	5
Missing sequence of length 7	199	54	214	54
Total number of missing	10349	772	10741	763

Table 4
Missing values before and after the imputation for the variable Total Academic staff FTE and HC.

Variable	Academic staff FTE		Academic staff HC	
Before or after imputation	Before	After	Before	After
Institutions without missing	1629	2921	1595	2977
% on the total	51%	91%	50%	93%
One isolated internal missing	59	17	66	1
One isolated extreme missing	61	81	93	64
Missing sequence of length 2	204	64	387	55
Missing sequence of length 3	119	8	318	19
Missing sequence of length 4	48	6	80	4
Missing sequence of length 5	155	34	102	33
Missing sequence of length 6	650	58	529	58
Missing sequence of length 7	309	21	199	6
Total number of missing	7915	939	7284	803

Table 5
Missing values before and after the imputation for the variable Total Non-academic staff FTE and HC.

Variable	Non-acad. staff FTE		Non-acad. staff HC	
Before or after imputation	Before	After	Before	After
Institutions without missing	1455	2886	1341	2815
% on the total	45%	90%	42%	88%
One isolated internal missing	59	1	105	1
One isolated extreme missing	66	103	122	122
Missing sequence of length 2	241	71	244	71
Missing sequence of length 3	125	11	93	17
Missing sequence of length 4	62	8	67	10
Missing sequence of length 5	88	13	102	16
Missing sequence of length 6	829	95	870	134
Missing sequence of length 7	309	23	309	26
Total number of missing	8807	1107	9145	1422

explicit, and the value $\bar{\delta} v^j_{i-1}$, which is the value obtainable for v^j_i when assuming the average variation $\bar{\delta}$ from the preceding value v^j_{i-1}, and each $(v^j_{i+1} - \bar{\delta} v^j_i(d))$ represents the difference between the following value v^j_{i+1} and the value $\bar{\delta} v^j_i(d)$ obtainable for v^j_{i+1} from the imputed value $v^j_i(d)$ when assuming again the average variation $\bar{\delta}$. Note that v^j_{i-1} and v^j_{i+1} could be such that the average variation $\bar{\delta}$ cannot hold for $v^j_i(d)$, however the minimization of the above squared sums aims at providing the value of d that better shares the differences.

The described Weighted Average Imputation appears feasible to impute isolated missing values, in particular when they are internal. It can be adapted to the external case, by considering only one side of the former case, however it may fail to capture the data trend when there is a distinct increase or decrease of the values over the years. In any case, the Weighted Average approach could be used as a building block to develop more sophisticated imputation techniques for missing sequences.

3.2. Imputation based on linear regression

Another basic technique that can be used to reconstruct values in a time series is *linear regression*. The missing value v_i is approximated with the value given by the straight line interpolating the available values v_h, with $h \in T$, $h \neq i$. This approach relies on the assumption that the values in the time series are not only related but also subject to a temporal evolution, which often exhibits a trend. See [31] for further discussion of the field of application of this approach. This approach does not need to compute weights or other parameters. It appears feasible to impute isolated missing values, in particular extreme ones, because it is able to capture the data trend.

However, when there is a sharp increase or decrease in the available values, it may predict negative values, which are clearly infeasible. For example, if we have the following sequence ⊔, 100, 250, 410, 550, 690, the Linear Regression Imputation would provide -44 for the value of the first period. Clearly, negative values are not acceptable, and even replacing them with 0 would not be a good solution. In similar cases, we propose to smoothen the trend by computing a value $v_1 \in (0, v_2)$. In particular, v_1 can be computed with the exponentiation operation as $(v_2)^c$, with $c \in (0, 1)$. For instance, using $c = 0.5$ gives the square root, which in the example above would produce $v_1 = \sqrt{100} = 10$, which is a more reasonable value.

To select the value of the exponent c, we propose to use again a data-driven approach. When focusing on the case of initial isolated missing, we define the set E of the institutions which would have a negative initial value v_1 if approximated with linear regression, and we compute a new average variation $\bar{\delta}_E$ limited to the institutions in E. Now, we search for the value of c that better fits this new average variation $\bar{\delta}_E$ by solving the following optimization problem,

$$\min_{c\in(0,1)} \sum_{j\in E} (v_2^j - \bar{\delta}_E \; v_1^j(c))^2,$$

where each $(v_2^j - \bar{\delta}_E \; v_1^j(c))$ represents the difference between the available value v_2^j and the value $\bar{\delta}_E \; v_1^j(c)$, which is the value obtainable for v_2^j when assuming $v_1^j(c) = (v_2^j)^c$ and the average variation $\bar{\delta}_E$.

Specularly, when focusing on the case of final isolated missing, we define the set F of the institutions which would have a negative final value v_t if approximated with linear regression, and we compute another average variation $\bar{\delta}_F$ limited to the institutions in F. In this case, we search for the value of c that better fits this new average variation $\bar{\delta}_F$ by solving the following optimization problem,

$$\min_{c\in(0,1)} \sum_{j\in F} (v_t^j - \bar{\delta}_F \; v_{t-1}^j(c))^2,$$

where each $(v_t^j(c) - \bar{\delta}_F \; v_{t-1}^j)$ represents the difference between the imputed value $v_t^j(c)$ and the value $\bar{\delta}_F \; v_{t-1}^j$, which is the value obtainable for v_t^j when assuming $v_t^j(c) = (v_{t-1}^j)^c$ and the average variation $\bar{\delta}_F$.

The described Linear Regression Imputation, possibly integrating the described exponentiation technique, appears feasible to impute isolated missing values, in particular when they are extreme. Moreover, similarly to the approach described in the previous Section, it can be used as a building block to develop more sophisticated imputation techniques for missing sequences.

3.3. Trend smoothing imputation

To be able to capture a trend in the series, but at the same time to be not excessively (mis)lead by it, we propose to combine the two simple approaches described above by means of the following technique. By denoting with WA_i the value given for instant i by the weighted average approach, and by LR_i the value given for instant i by the linear regression approach, the actual value imputed for v_i would be obtained as a combination of the two, as follows.

$$v_i = \frac{a^2}{a^2+1} WA_i + \frac{1}{a^2+1} LR_i$$

Note that the value LR_i is intended to be already smoothened by means of the exponentiation operation explained above. Coefficient a should give more importance to the contribution of the linear regression when the slope of the interpolating straight line is nearly flat, and more importance to the contribution of the weighted average when the same slope is too vertical. This transition from one extreme to the other should be without discontinuities. Therefore, by denoting with m the angular coefficient of the interpolating straight line, with $T' \subset T$ the set of the available values, and with s a constant value usually set at 2, we use

$$a = \frac{s|m|}{\min_{\tau\in T'}\{v_\tau\}}.$$

In other words, we follow the trend given by the available values when it is reasonably increasing or decreasing; we progressively consider the trend less reliable when the available values make it too steep. The absolute value of the angular coefficient is divided by the smaller available value in order to "normalize" it, because an increase for example from 1000 to 1500 is more reasonable than an increase from 1 to 501, which will produce the same value of m.

Table 6
Missing values before and after the imputation for the variables Total Expenditure and Total Revenues.

Variable	Total expenditure		Total revenues	
Before or after imputation	Before	After	Before	After
Institutions without missing	987	3003	1021	3001
% on the total	31%	94%	32%	94%
One isolated internal missing	81	0	103	0
One isolated extreme missing	450	36	359	36
Missing sequence of length 2	429	64	222	63
Missing sequence of length 3	287	15	248	13
Missing sequence of length 4	83	6	88	7
Missing sequence of length 5	75	11	106	13
Missing sequence of length 6	945	71	1051	72
Missing sequence of length 7	202	7	202	7
Total number of missing	10041	763	10252	775

The technique proposed in this Section is called *Trend Smoothing Imputation*, since it does follow the trend, however it progressively smoothen it when needed, without discontinuities. This technique appears feasible for isolated missing, but also for missing sequence of length $L \le t - 2$. In particular, we will use it for this last case in our experiments. The case of missing sequences of length $t-1$ actually contain only one non-missing value, hence no trend exists. Hence, these cases are better assimilated to the cases of full sequence missing.

Finally, in the case of interconnected variables, for example number of students and number of graduates, we may compute an average *ratio* between the values available year by year for that couple of variables, and impose that the imputed values remains "not too far" from that ratio. This request can be practically implemented according to the peculiarities of the specific case.

4. Imputation techniques based on donors

The techniques described in the previous Sections basically rely on some type of prolongation of the values available in the sequence. When the full sequence of values is missing, or even when there is only one available value and the rest of the sequence is missing, those techniques are clearly inapplicable. However, unfortunately, the case of full sequence missing is the most common type of missing. Note, indeed, that a full sequence missing generally corresponds to 6 missing values with the current time horizon.

Therefore, it is important to deal with such type of missing, even if this happens to be the most difficult case. Thus, we need to recur to a different imputation technique, which is also well established in the field [6]. This technique is called *donor imputation*, and is based on the search for a complete (and correct) record (i.e., institution) being as similar as possible to the incomplete records under imputation. Similarity is judged by defining a *distance* function among records, and thus minimum distance corresponds to maximum similarity. When this complete record at minimum distance is found, it is used as a *donor*: the missing values of the incomplete record are replaced by the corresponding values of the complete records. The incomplete record under imputation is also called the *recipient*. A possible variant is to search for a small set of records at minimum distance (e.g., 3) and not just one, and then select the donor randomly among them For this reason, this technique is sometimes called *k*-nearest neighbor imputation. Of course, this approach is easily applied when the set of possible donors is large enough to have donors in the vicinity of each incomplete record that must be treated. A recognized statistical advantage of the donor imputation with respect to other imputation techniques is that it does not create artificial and possibly unlikely values; rather, it takes values already appearing in the dataset, and they are selected with higher probability whenever they are more frequent, see also [18]. Hence, donor imputation tends to preserve the data properties and their frequency distributions without hypothesis on the distribution of the data, which may be questionable [12].

The similarity criterion is fundamental and must be defined on the specific case. In the case of educational institutions, we identify a number of variables under which institutions should be considered similar. They are essentially variables describing the size, the type and the geographical location of the institution. For each of these variables, a difference in the values may give a certain contribution to the total distance or directly lead to the exclusion of the record as possible donor. In the specific case of the ETER database, containing European institutions, we set up the following distance calculation scheme.

1. Variable: "Institution Category standardized", whose value can be 1 = University, 2 = University of Applied Science, 0= Other. For this variable we accept only donors from the same category.
2. Variable: "Distance education institution", which tells whether the institution is essentially telematic or traditional. The difference in this variable gives a contribution to the distance denoted by p_1, which should be set at a value corresponding to a strong penalization.
3. Variable: "Institution Category", which reports the category of the institution with more granularity than the former "Institution Category standardized". The difference in this variable gives again a contribution to the distance of p_1.
4. Variables: "Total Current expenditure", "Total Current revenues", "Total academic staff" (which can be measured either in Full Time Equivalent or in Headcount). These variables basically describe the size of the institution. For these numerical variables, each difference between two values v' and v'' gives a contribution to the distance computed as follows

   $$p_1 \frac{|v' - v''|}{\max\{v', v''\}}.$$

   In other words, the contribution is between p_1 and 0, depending on the absolute difference between the two values normalized by dividing it by the largest value. Hence, the maximum contribution to the total distance given by these three variables is $3p_1$. Size indicators are very important for the selection of a donor; unfortunately these variables are often missing in the ETER database, so their use is limited in practice.
5. Variable: "Country", reporting the country of the institution. We define similarity according to geographical areas reported below.

   - Area 1: Belgium, Liechten., Luxembourg, Netherlands, Switzerland.
   - Area 2: Austria, Germany.
   - Area 3: Greece, Italy, Portugal, Spain.
   - Area 4: Czech Republic, Slovakia, Estonia, Lithuania, Latvia, Hungary, Poland.
   - Area 5: Albania, Bulgaria, Croatia, North Macedonia, Romania, Serbia, Slovenia, Montenegro.
   - Area 6: Finland, Norway, Denmark, Iceland, Sweden.
   - Area 7: Ireland, Malta, United Kingdom.
   - Area 8: France.
   - Area 9: Cyprus, Turkey.

   The same country gives a contribution of 0 to the distance. Different countries in the same area gives a contribution to the distance denoted by p_2, which should be set at a value providing a light penalization. Different areas give a contribution to the distance of p_1 (strong penalization).
6. Variable: "Legal status", reporting whether the institution is public or private. Difference in this variable gives a contribution of p_2 to distance.

Unfortunately, in the case of ETER, the number of possible donors is not large enough to guarantee the presence of suitable donors for each HEI. In particular, some types of institutions or some Countries have very few complete institutions. And even exploiting also the incomplete institutions, used as sets of partial donors for the same recipient, the problem remains unsolved. Hence, to avoid using donors too different from the recipient, which will produce unacceptable data, we need to impose *filters* on acceptable donors, i.e., criteria to recognize and exclude unacceptable donors. For this reason, for some of the incomplete HEIs it is not possible to obtain donors, and so they remain not imputed.

A first filter for donors is implemented by comparing the average values of donor and recipient on a number of size-related variables, such as: Total Current Expenditures; Total Current Revenues; Total Students Enrolled; Total Graduates; Total Academic Staff. When the two corresponding average values of donor and recipient are both available, and their difference is larger than a threshold (e.g., 30%), the donor is considered not acceptable. This filter also uses values of the variable under imputation that may possibly be available in the recipient (e.g., recipient has only 1 value and the rest of the sequence is missing), again by comparing the average values of donor and recipient.

Moreover, some HEIs may be not suitable to donate because, even if they are complete, their values are too uncommon, and it is not advisable to replicate them. We select some of them by computing, for each institution, the ratios between all pairs of variables (e.g., graduates/student, Ph.D. students/ graduates, expenditure/students, etc.) and by excluding those having extreme values of ratios (e.g., top and bottom 2%). Some other are selected by computing the volatility of the sequences, and by excluding those having too volatile values (e.g., top 2%).

Another technique which we introduce to deal with the relative scarcity of donors is *scaling*. When the recipient has a missing sequence containing one non-missing value, for instance v_2, and the sequence of the donor's values is $(w_1, w_2, \dots, w_t)$, we learn from v_2 and w_2 a size ratio $r = v_2/w_2$ between the two institutions, and scale the donor's values by making the recipient sequence become $(rw_1, v_2, rw_3, \dots, rw_t)$. On the other hand, when the recipient has a full sequence missing, scaling can be done by using another recipient's variable strongly related to the variable under imputation. For example, if the recipient has values $(s_1, s_2, \dots, s_t)$ for students and all missing for graduates, and the donor has $(t_1, t_2, \dots, t_t)$ for students and $(g_1, g_2, \dots, g_t)$ for graduates, we learn the sequence of size ratios $(r_1 = s_1/t_1, r_2 = s_2/t_2, \dots, r_t = s_t/t_t)$, and scale the donor's values by making the recipient graduates become $(r_1g_1, r_2g_2, \dots, r_tg_t)$. This allows the imputation of values suitable for the recipient's size even if that differs from the donor's size.

Furthermore, similarly to case of Trend Smoothing imputation, for the couples of variables which are practically linked, such as number of students and number of graduates, we use the trend of the variable available in the recipient to improve the selection of the donor. For example, if we are imputing the whole sequence of number of students and the number of graduates is available in the recipient and is increasing, we accept only donors with increasing number of graduates, or, if that is missing in the donor, with increasing number of students. On the other hand, when the recipient misses both the sequences of number of students and number of graduates, the imputations of both variables should either be done by the same donor or by two partial donors with sequences compatible for size, ratio and trend. Finally, we set a limit on the number of times an institution can be used as donor, to avoid any risk of replicating the same values too often.

5. Results on the real missing values of ETER database

We report in this section the result of the imputation of all the ETER missing values. Therefore, the *original* values (i.e., those lost due to the missing) are not available for a direct check. While the techniques described in Sections 3.1, 3.2, 3.3 always provide a value, the donor imputation described in Section 4 may leave unimputed institutions when no acceptable donor is available for them. More specifically, Tables 2–6 describe the types of missing appearing in the database before and after the imputation operations, respectively for: Total Students and Total Graduates; Total Ph.D. Students and Total Ph.D. Graduates; Total Academic staff FTE and HC; Total Non-academic staff FTE and HC; Total Expenditure and Total Revenues. All the implementation codes of our methodology are available in [32].

As observable, the percentage or institutions without missing goes from being often lower than 50% before imputation to being generally over 90% after imputation. The non imputed cases are all due to the unavailability of acceptable donors for some institutions; they may be short sequences because those institutions often contain less than 6 years in their time horizon. Clearly, they could be imputed by further relaxing the filters used on the donors, however the quality of the imputed values would worsen. Thus, we prefer to pursue this compromise between *coverage* and *quality* of the imputation. The results of our methodology on ETER database are available in [33].

Since the original values are not available, the quality of the imputation cannot be evaluated by matching original and imputed values. Hence, we evaluate it by comparing, for the dataset *before* imputation and *after* imputation:

- the frequency distribution of each variable;
- the ratios between selected couple of variables. This type of measure is particularly informative for interconnected variables.

We do this by using the so called violin plots juxtaposing the frequency distributions of the variables in Figs. 1–10, and the box plots juxtaposing the statistical description of the ratios in Figs. 11–13. We consider informative the following ratios: Graduates/Students; Ph.D. Graduates/Ph.D. Students, Expenditure/Revenues, Expenditure/All Students, Revenues/All Students; Academic Staff FTE/All Students; Academic Staff HC/All Students; Non-academic Staff FTE/All Students; Non-academic Staff HC/All Students. With “All Students” we denote the sum of Students and Ph.D. Students; this quantity have been introduced to evaluate fairly the institutions focused on producing Ph.D., for which a ratio over Students would be misleading.

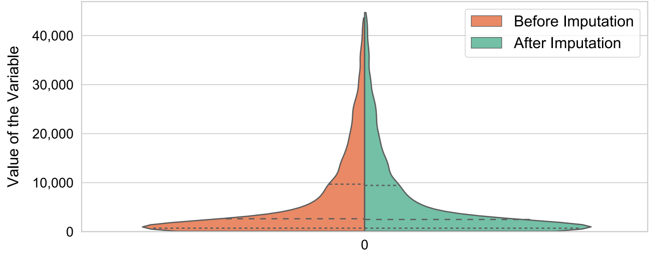

Fig. 1. Violin plot comparing the distributions before and after imputation for Students.

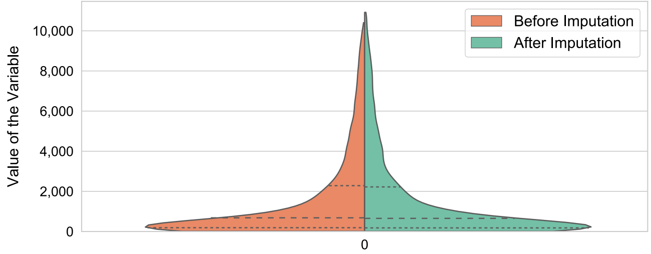

Fig. 2. Violin plot comparing the distributions before and after imputation for Graduates.

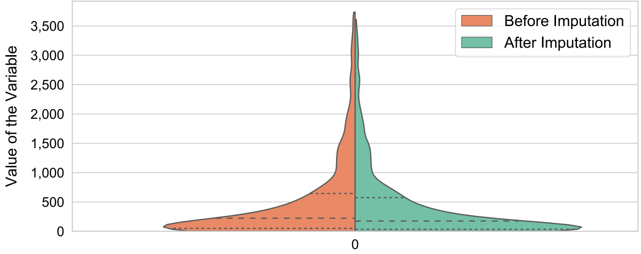

Fig. 3. Violin plot of the distrib. before and after imputation for Ph.D. Students.

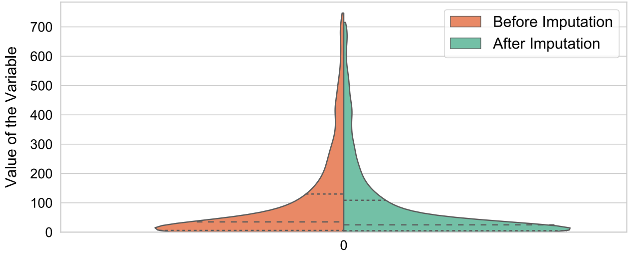

Fig. 4. Violin plot of the distrib. before and after imputation for Ph.D. Graduates.

Note that we chose to compare the ratios of the data before imputation to those of the imputed data only (and not to those of all data after imputation), because, with the second choice, possible differences would be too “diluted” to be evident. Note, moreover, that the missing values are not equally distributed over the institutions, but sometimes more concentrated on small institutions, especially for expenditure and revenues. Thus, when the small institutions are imputed, small values would appear more frequently in the distribution, and this behavior is correct. On the contrary, imputing the small institutions with values similar to those of the larger institutions would not be correct. Therefore, the “ideal” imputation does not necessarily correspond to the exact replication of the data distribution in the left side of the violin plots.

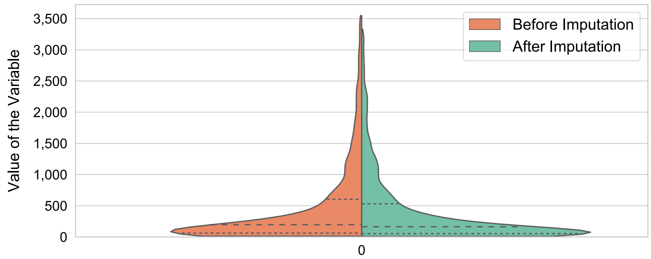

Fig. 5. Violin plot of the distrib. before and after imputation for Academic Staff FTE.

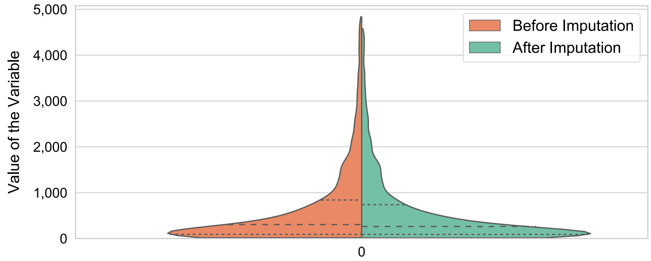

Fig. 6. Violin plot of the distrib. before and after imputation for Academic Staff HC.

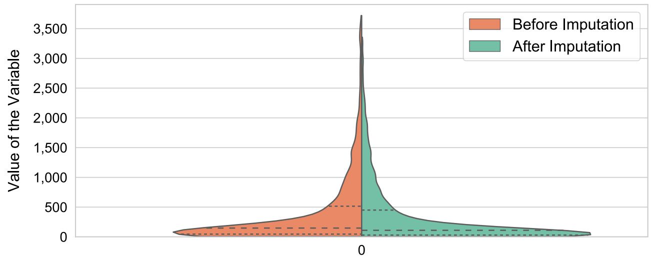

Fig. 7. Violin plot of the distrib. before and after imputation for Non-academic Staff FTE.

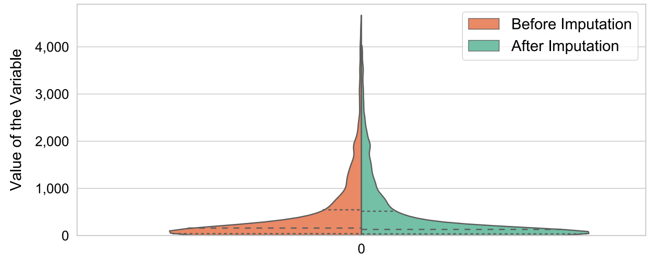

Fig. 8. Violin plot of the distrib. before and after imputation for Non-academic Staff HC.

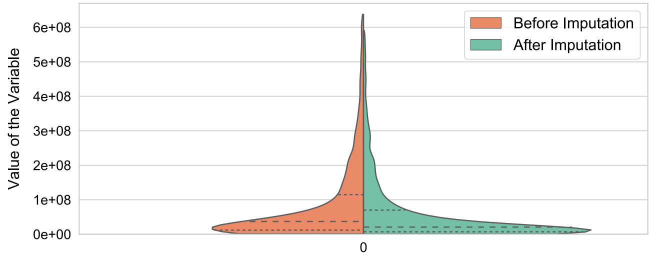

Fig. 9. Violin plot of the distrib. before and after imputation for Total Expenditure.

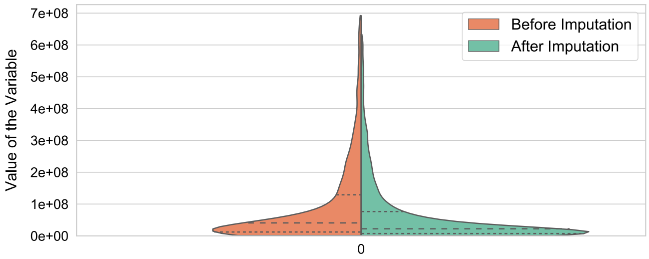

Fig. 10. Violin plot of the distrib. before and after imputation for Total Revenues.

From these analyses, we can observe that: (1) the imputation increased considerably the coverage of the database; some missing are still present due to relative scarcity of donors, however the usability has greatly improved; (2) the distributions of the data have been generally preserved, however in some cases, mainly Expenditure and Revenues (Figs. 9 and 10), the number of small values correctly increased, because the missing were localized mainly on small institutions; (3) the ratios of the imputed data show that they maintain very well the relations between the interconnected variables, so the data quality appears very satisfactory.

6. Performance and accuracy evaluation

To evaluate the accuracy of the proposed methodology, we set up the following experiments using 4 particularly relevant variables: Total Students, Total Graduates, Total Academic Staff FTE, Total Expenditure.

1. We identify a dataset composed by the institutions having all the values for the variable in analysis, and artificially introduce in this variable random missing values according to perturbation schemes described below.
2. For each type of such missing values, we impute the values using the technique developed for that type of missing among the proposed ones.
3. We compare the imputed values with the original values, which in these experiments are known, and we study the occurrence of significant differences.

6.1. Perturbation scheme

We use the following perturbation scheme. For each single variable of the 4 in analysis, we perturb the dataset in 3 alternative manners. We introduce in a first case isolated missing values, in a second case missing sequences of length L, and in a third case full (or almost full) sequences missing, as explained below.

- **Perturbation 1**. Each record is perturbed with one isolated missing value randomly located over the time horizon T.
- **Perturbation 2**. Each record is perturbed with one missing sequence randomly located over the horizon T. The length of the sequence is 2 with probability 0.5, 3 with probability 0.35, and 4 with probability 0.15.
- **Perturbation 3**. The dataset is randomly partitioned in two, taking care of keeping, in each partition, representatives of all countries and types of Institutions. This is obtained by splitting each country and type of institution independently. Then, one partition is perturbed, and the other partition is used as set of possible donors. Hence, the set of possible donors does not include any of the Institutions that undergo perturbation. The perturbation introduces full sequence missing with probability 0.5, and missing sequence of length 5 (i.e., only 1 value is available in the time horizon) randomly located over the time horizon with probability 0.5.

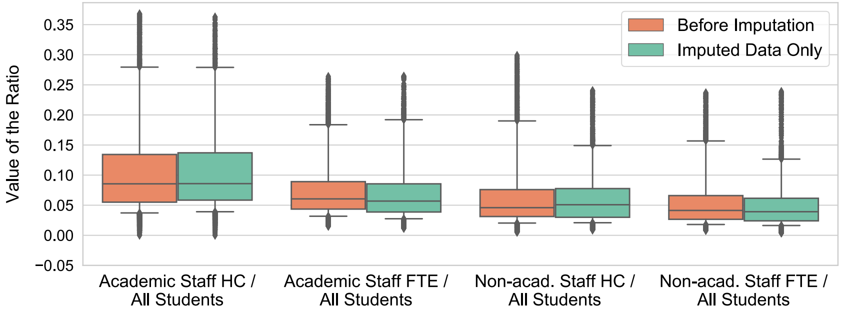

Fig. 11. Box plot comparing the ratios reported near to each plot of the data before imputation to those of imputed data only. All Students means Students + Ph.D. Students.

Therefore, in the first two tests, every record is perturbed and will undergo imputation, while in the third test this happens for half of the records. Note that such a situation is quite worse than the standard situation in real databases. Therefore, the results on the occurrences of significant differences can be seen as a worst-case bound over the results of a real application.

For the first perturbation case, we apply the Weighted Average imputation technique described in Section 3.1 to the internal isolated missing values, and Linear Regression imputation, integrated with the exponentiation technique described in Section 3.2, to the external isolated missing values. Results on this case are reported in Table 7 and Fig. 14. For the second perturbation case, we apply the Trend Smoothing imputation described in Section 3.3, and results are reported in Table 8 and Fig. 15. For the third perturbation case, we apply Donor Imputation with the distance function and filter admissible donors as described in Section 4. Distance parameters p_1 and p_2 are set respectively at 3 and 1. Note that, in this experiment, each institutions received a donor, so no missing values remains after the imputation. Results for this third case are reported in Table 9 and Fig. 16.

6.2. Evaluation of the reconstruction accuracy

This Section aims at evaluating the accuracy in the reconstruction of the imputed values, i.e., how similar to the original value is the imputed value. For each imputed value v_i of each institution j, with $i \in T$, we compute the difference η_{ij} with respect to the original value $v_i^\star$:

$$\eta_{ij} = \frac{(v_i - v_i^\star)}{v_i^\star}$$

An "ideal" imputation would provide very "small" values for η_{ij}, with 0 being the limit. However, 0 is not a realistic target, and we need a scale to determine which values are actually "small". To do so, we use again a data-driven measure and we consider the two extreme values e_1 and e_2, defining a so called Interval of Moderate Variations (IMV), containing 90% of the values of the variations δ_{ij}, as explained in Section 3.1 . Now, we study the frequency of the imputations whose corresponding η_{ij} lay within or outside the $IMV = [e_1, e_2]$. We define "significant" a difference for which $\eta_{ij} \notin [e_1, e_2]$. Moreover, to consider also a data-independent measure, we study the frequency of the imputations whose corresponding η_{ij} lay within or outside a Fixed Interval (FI) defined as $FI = [-10\%, 10\%]$, which is more restrictive than the previous interval in the analyzed cases. The outcome of these analyses is in Tables 7–9.

As a general observation, we note that the imputed values lay quite near to the original ones in the majority of the cases, both in data-driven and in absolute measurements. Since the range of possible value is very wide, this is a very positive result.

Moreover, even when the imputed values are not so near, we hypothesize that the positive and negative errors should statistically compensate each other. To test this hypothesis, we consider "global" descriptors of the data, providing in Figs. 14, 15, 16 the box plots for: (1) the original data suppressed by the perturbation; and (2) the new data imputed by our methodology. Note that this is a very selective analysis, aimed at magnifying possible differences in the data values, and it is possible only in this case, because we actually know the original data suppressed in the perturbation. As observable, our hypothesis appears to be confirmed, since the imputed data appear statistically equivalent to the suppressed data to a considerable extent.

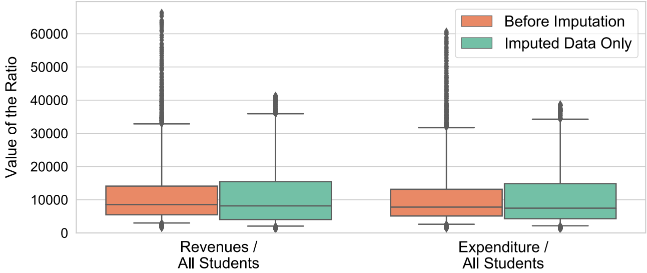

Fig. 12. Box plot comparing the ratios reported near to each plot of the data before imputation to those of imputed data only. All Students means Students + Ph.D. Students.

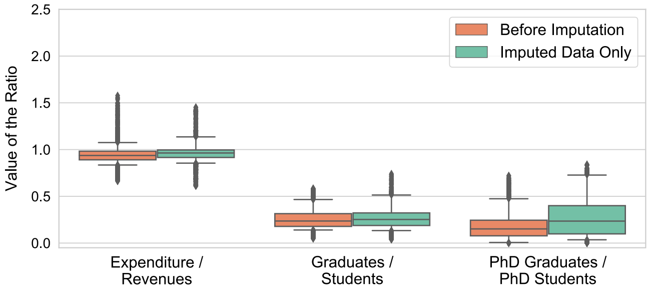

Fig. 13. Box plot comparing the ratios reported near to each plot of the data before imputation to those of imputed data only.

7. Conclusions

Data describing the situation of Educational Institutions are currently used for a wide variety of analyses, even to support important economic and political decisions. However, similar data often contain non negligible shares of missing values, also due to the structure of the gathering process, and this may invalidate the above operations. The missing information should therefore be optimally reconstructed, by imputing data as similar as possible to the unknown original ones. This is a very difficult task. We develop imputation techniques for the reconstruction of partial

Table 7
Analysis of the values imputed to deal with Perturbation 1.

	Total students	Total graduates	Total acad. staff	T. Expenditure
IMV	[−18.93, 29.13]	[−30.76 , 53.28]	[−17.65 , 22.72]	[−10.28 , 21.24]
Imputations laying in *IMV*	97.56%	96.48%	95.93%	90.24%
FI	[-10%, 10%]	[-10%, 10%]	[-10%, 10%]	[-10%, 10%]
Imputations laying in *FI*	88.35%	73.18%	89.70%	83.19%

Table 8
Analysis of the values imputed to deal with Perturbation 2.

	Total students	Total graduates	Total acad. staff	T. Expenditure
IMV	[−18.93, 29.13]	[−30.76 , 53.28]	[−17.65 , 22.72]	[−10.28 , 21.24]
Imputations laying in *IMV*	95.18%	93.74%	93.64%	81.12%
FI	[-10%, 10%]	[-10%, 10%]	[-10%, 10%]	[-10%, 10%]
Imputations laying in *FI*	79.69%	60.31%	80.62%	73.34%

Table 9
Analysis of the values imputed to deal with Perturbation 3.

	Total students	Total graduates	Total acad. staff	T. Expenditure
IMV	[−18.93, 29.13]	[−30.76 , 53.28]	[−17.65 , 22.72]	[−10.28 , 21.24]
Imputations laying in *IMV*	52.85%	64.88%	49.78%	43.64%
FI	[-10%, 10%]	[-10%, 10%]	[-10%, 10%]	[-10%, 10%]
Imputations laying in *FI*	31.94%	24.66%	33.75%	34.43%

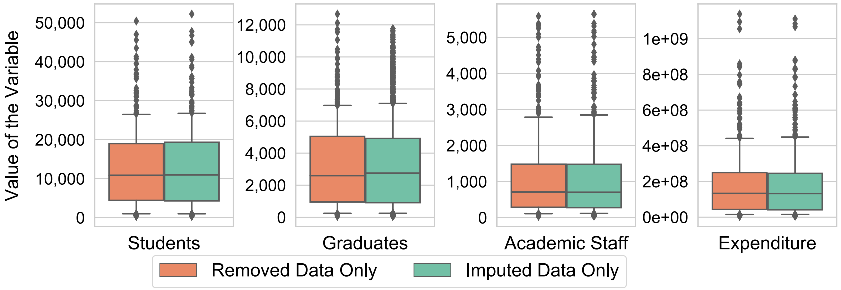

Fig. 14. Box plot of the data removed by Perturbation 1 and the data imputed by the procedure.

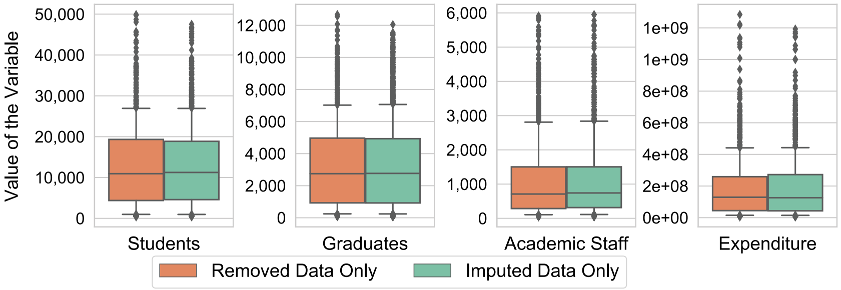

Fig. 15. Box plot of the data removed by Perturbation 2 and the data imputed by the procedure.

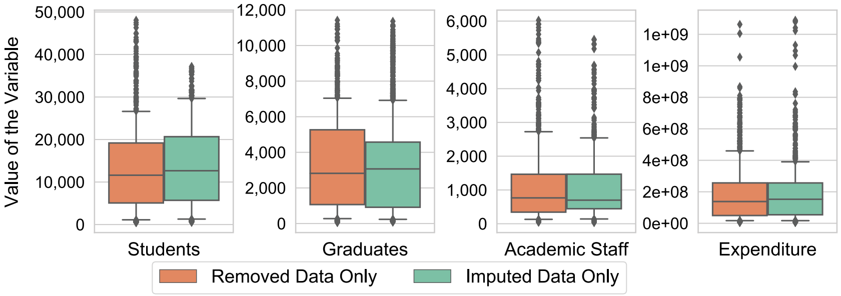

Fig. 16. Box plot of the data removed by Perturbation 3 and the data imputed by the procedure.

numerical sequences based on the combination of weighted average and linear regression, and techniques for the reconstruction of full numerical sequences based on the use of donors. We search for data-driven optimal solutions in the sense that we aim at maximizing the conservation of the global data features. Experiments on real-world data containing real missing values confirm that the imputation process is practically feasible and very useful. Experiments on real-world data artificially perturbed by inserting several types of missing values show that the reconstructed data are satisfactory similar to the original data.

One additional important advantage of the proposed procedure is that it works at the formal level, with a data driven approach. Hence, it could be adapted to impute different datasets containing educational data from other origins, or even datasets with different meaning but sharing the feature of interconnection among the variables. Future work includes the integration in the proposed methodology with the web scraping techniques described in [34,35] by using the Universities' websites, or with the Logic-based techniques described in [36,37] to extract data-supported logic descriptions of the Institutions.

CRediT authorship contribution statement

Renato Bruni: Conceptualization, Methodology, Investigation, Writing. **Cinzia Daraio:** Conceptualization, Writing, Funding acquisition. **Davide Aureli:** Software, Investigation, Data curation.

Declaration of competing interest

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

Acknowledgments

The EU Horizon 2020 RISIS2 Project, grant agreement N. 824091, is gratefully acknowledged.

References

[1] A. Bonaccorsi (Ed.), Knowledge, Diversity and Performance in European Higher Education: A Changing LandScape, Edward Elgar Publishing, Cheltenham (UK), 2014.

[2] C. Daraio, et al., The European university landscape: A micro characterization based on evidence from the Aquameth project, Res. Policy 40 (2011) 148–164.

[3] C. Daraio A. Bonaccorsi (Ed.), Universities and Strategic Knowledge Creation. Specialization and Performance in Europe, Edward Elgar Publisher, Cheltenham (UK), 2007.

[4] S. Van Buuren, Flexible Imputation of Missing Data, CRC press, Boca Raton, FL, USA, 2018.

[5] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed., John Wiley & Sons, New York, 2002.

[6] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.

[7] R. Razavi-Far, B. Cheng, M. Saif, M, Ahmadi similarity-learning information-fusion schemes for missing data imputation, Knowl.-Based Syst. 187 (2020) 104805.

[8] C.-F. Tsai, M.-L. Li, W.-C. Lin, A class center based approach for missing value imputation, Knowl.-Based Syst. 151 (2018) 124–135.

[9] S. Nikfalazar, C. Yeh, S. Bedingfield, H.A. Khorshidi, Missing data imputation using decision trees and fuzzy clustering with iterative learning, Knowl. Inf. Syst. 62 (2020) 2419–2437.

[10] A.M. Sefidian, N. Daneshpour, Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model, Expert Syst. Appl. 115 (2019) 68–94.

[11] A. Andiojaya, H. Demirhan, A bagging algorithm for the imputation of missing values in time series, Expert Syst. Appl. 129 (2019) 10–26.

[12] I. Gheyas, L. Smith, A neural network-based framework for the reconstruction of incomplete data sets, Neurocomputing 73 (2010) 3039–3065.

[13] S. Jyoti Choudhury, N.R. Pal, Imputation of missing data with neural networks for classification, Knowl.-Based Syst. 18215 (2019) Article 104838.

[14] Q. Ma, W.-C. Lee, T.-Y. Fu, Y. Gu, G. Yu, MIDIA: exploring denoising autoencoders for missing data imputation, Data Min. Knowl. Discov. (2020) http://dx.doi.org/10.1007/s10618-020-00706-8.

[15] J. Lin, N. Li, M.A. Alam, Y. Ma, Data-driven missing data imputation in cluster monitoring system based on deep neural network, Appl. Intell. 50 (2020) 860–877.

[16] Q. Lan, X. Xu, H. Ma, G. Li, Multivariable data imputation for the analysis of incomplete credit data, Expert Syst. Appl. 141 (2020) 112926.

[17] C. Ye, H. Wang, W. Lu, J. Li, Effective Bayesian-network-based missing value imputation enhanced by crowdsourcing, Knowl.-Based Syst. 190 (2020) 105199.

[18] R. Bruni, Error correction for massive data sets, Optim. Methods Softw. 20 (2–3) (2005) 295–314.

[19] N. Bokde, M.W. Beck, F. Martínez Álvarez, K. Kulat, A novel imputation methodology for time series based on pattern sequence forecasting, Pattern Recognit. Lett. 116 (2018) 88–96.

[20] Z. Qi, H. Wang, J. Li, H. Gao, FROG: Inference from knowledge base for missing value imputation, Knowl.-Based Syst. 145 (2018) 77–90.

[21] S. Song, Y. Sun, A. Zhang, L. Chen, J. Wang, Enriching data imputation under similarity rule constraints, IEEE Trans. Knowl. Data Eng. 32 (2020) 275–287.

[22] K. Hron, M. Templ, P. Filzmoser, Imputation of missing values for compositional data using classical and robust methods, Comput. Statist. Data Anal. 54 (12) (2010) 3095–3107.

[23] J. Tian, B. Yu, D. Yu, et al., Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering, Appl. Intell. 40 (2014) 376–388.

[24] S. Zhang, Z. Jin, X. Zhu, Missing data imputation by utilizing information within incomplete instances, J. Syst. Softw. 84 (3) (2011) 452–459.

[25] D. Bertsimas, C. Pawlowski, Y.D. Zhuo, From predictive methods to missing data imputation: an optimization approach, J. Mach. Learn. Res. 18 (2017) 7133–7171.

[26] M.L. Yadav, B. Roychoudhury, Handling missing values: A study of popular imputation packages in r, Knowl.-Based Syst. 160 (2018) 104–118.

[27] M. Lepot, J.B. Aubin, F.H. Clemens, Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment, Water 9 (2017) 796.

[28] F. Bashir, H.-L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, Neurocomputing 276 (2018) 23–30.

[29] C. Daraio, R. Bruni, G. Catalano, A. Daraio, G. Matteucci, M. Scannapieco, D. Wagner-Schuster, B. Lepori, European Tertiary education register (ETER): Evolution of the data quality approach, J. Data Inf. Sci. (2020) http://dx.doi.org/10.2478/jdis-2020-0019.

[30] C. Daraio, A. Bonaccorsi, L. Simar, Efficiency and economies of scale and specialization in European universities. a directional distance approach, J. Informetr. 9 (2015) 430–448.

[31] J.D. Hamilton, Time Series Analysis, vol. 2, Princeton university press, Princeton, NJ, 1994, pp. 690–696.

[32] D. Aureli, R. Bruni, C. Daraio, Optimization methods for the imputation of missing values in educational institutions data. MethodsX.

[33] R. Bruni, C. Daraio, D. Aureli, Information Reconstruction in Educational Institutions Data from the European Tertiary Education Registry. Data in Brief.

[34] G. Bianchi, R. Bruni, C. Daraio, A.L. Palma, G. Perani, F. Scalfati, Exploring the potentialities of automatic extraction of university webometric information, in: Proceedings of 17th International Conference on Scientometrics and Informetrics, ISSI 2019, 2019, pp. 2094–2105.

[35] R. Bruni, G. Bianchi, Website categorization: A formal approach and robustness analysis in the case of e-commerce detection, Expert Syst. Appl. 142 (2020) 113001.

[36] R. Bruni, G. Bianchi, Effective classification using binarization and statistical analysis, IEEE Trans. Knowl. Data Eng. 27 (9) (2015) 2349–2361.

[37] R. Bruni, G. Bianchi, C. Dolente, C. Leporelli, Logical analysis of data as a tool for the analysis of probabilistic discrete choice behavior, Comput. Oper. Res. 106 (2019) 191–201.