Data in Brief 34 (2021) 106611

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Information reconstruction in educational institutions data from the European tertiary education registry

Renato Bruni*, Cinzia Daraio, Davide Aureli

University of Rome "Sapienza", Italy

ARTICLE INFO

Article history:
Received 28 October 2020
Revised 25 November 2020
Accepted 26 November 2020
Available online 30 November 2020

Keywords:
Higher education institutions
Data imputation
Reconstructed institutional microdata
Machine learning
European universities

ABSTRACT

Universities and other organizations providing higher level education are collectively called Higher Education Institutions. Their detail data, for instance number of students, number of graduates, etc., constitute the basis for several important analyses of the educational systems. This work provides data of the European Tertiary Education Register (ETER), which describes the Educational Institutions of Europe. These data have been gathered through the National Statistical Authorities of all the Countries participant in the ETER Project. However, they include many scattered missing values. Therefore, we have developed and applied an imputation methodology (see "Imputation Techniques for the Reconstruction of Missing Interconnected Data from Higher Educational Institutions, Bruni et al. [3]) to replace the missing values with feasible values being as similar as possible to the original values that have been lost and are now unknown. Thus, we also provide the imputed version of the same dataset, which allows more in-depth analyses of the European Higher Education Institutions. Both datasets (before and after imputation) are provided in two versions: with or without bibliometric information for the Institutions, so the user can also consider these additional information if interested.

DOI of original article: 10.1016/j.knosys.2020.106512
* Corresponding author.
E-mail address: bruni@diag.uniroma1.it (R. Bruni).

https://doi.org/10.1016/j.dib.2020.106611

Specifications Table

Subject	Information Systems and Management
Specific subject area	Data from Higher Education Institutions, imputed with machine learning techniques. To be used for research on education policies.
Type of data	Tables in the form of MS Excel spreadsheets.
How data were acquired	The data have been gathered through the National Statistical Authorities of the Countries participant in the ETER Project and were downloaded from the ETER Project website: https://www.eter-project.com/#/home. Subsequently, many values that were missing have been imputed with machine learning techniques, see "Imputation Techniques for the Reconstruction of Missing Interconnected Data from Higher Educational Institutions, Bruni et al. [3].
Data format	Raw, Imputed and Analyzed
Parameters for data collection	The datasets has been collected for the years 2011, 2012, 2013, 2014, 2015, 2016,2017 and contain the following data from European Higher Education Institutions: Total students enrolled, Total graduates, Total PhD students, Total PhD graduates, Total academic staff (measured in "Full Time Equivalent" or in "Head Count"), Total non-academic staff (measured as above), Total expenditure, Total revenues.
Description of data collection	The raw data have been gathered through the National Statistical Authorities of the Countries participant in the ETER Project and were downloaded from the ETER Project website: https://www.eter-project.com/#/home. However they contained several missing values. Therefore, we have developed and applied an imputation methodology (see "Imputation Techniques for the Reconstruction of Missing Interconnected Data from Higher Educational Institutions, Bruni et al. [3]) to replace the missing values with feasible values being as similar as possible to the original values that have been lost.
Data source location	Primary data sources: ETER Project website: https://www.eter-project.com/#/home
Data accessibility	With the article
Related research article	R. Bruni, C. Daraio, D. Aureli: Imputation Techniques for the Reconstruction of Missing Interconnected Data from Higher Educational Institutions. Knowledge-Based Systems https://doi.org/10.1016/j.knosys.2020.106512

Value of the Data

- These data are the only available information at the micro level for European Higher Education Institutions.
- Researchers interested in analysing the European Higher Education Institutions can benefit from these datasets, especially from the reconstructed versions of them, to enhance the completeness and the representativeness of their analysis.
- The reconstructed data provided here can be elaborated to provide empirical evidence on European Higher Education Institutions, in order to study and compare the educational systems of the different countries.

1. Data Description

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). Their detail data, also called HEI microdata (e.g., number of students, number of graduates, etc.), constitute the basis for several important analyses of the educational systems [1]. However, this type of detail data are generally much more difficult to be obtained than the corresponding aggregate data.

In particular, 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. It has been gathered and assembled through the course of several research projects, also at the EU level. The latest edition of this project involves the Joint Research center, the Directorate General for Education and Culture of the European Commission, EUROSTAT and the National Statistical Authorities of the participating Countries.

ETER currently 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 this required the cooperation of a large number of entities.

However, also due to the vastness of the project, the current ETER database includes many scattered missing values, and this prevents many important analyses of these data. Therefore, we have developed and applied the imputation methodology described in [3] to replace the missing values with feasible values being as similar as possible to the original values that have been lost. The codes implementing this imputation methodology in Python are available in [4].

Thus, in this work we also provide the imputed version of the most relevant variables included in the ETER dataset, namely: Total students enrolled, Total graduates, Total PhD students, Total PhD graduates, Total academic staff (researchers and professors, measured in "Full Time Equivalent - FTE" or in "Head Count - HC"), Total non-academic staff (technical and administrative staff, again measured in "Full Time Equivalent" or in "Head Count"), Total current expenditure, Total current revenues. This information allows many more analyses of the European HEIs. Both the mentioned datasets (ETER before and after imputation) are provided in two distinct versions: with or without bibliometric information for the Institutions. Hence the user is also enabled to consider these additional information, if interested. The bibliometric data are obtained from the CWTS Web of Science bibliometric database provided in the RISIS European Project (https://risis.eu/).

Each dataset is composed of a single MS Excel file. Hence, the files provided here are:

original_ETER_dataset.xlsx: The ETER dataset before the imputation of the missing values, and without bibliometric information for the Institutions.

imputed_ETER_dataset.xlsx: The ETER dataset after the imputation of the missing values, and without bibliometric information for the Institutions. Note that not all missing values can be imputed, because some specific types of information are particularly scarce in the original dataset, and so the imputation procedure occasionally does not have enough information to work.

original_ETER_dataset_bibliometrics.xlsx: The ETER dataset before the imputation of the missing values, enriched with bibliometric information for the Institutions obtained from the CWTS Web of Science database.

imputed_ETER_dataset_bibliometrics.xlsx: The ETER dataset, enriched with bibliometric information form the CWTS Web of Science database, after the imputation of the missing values.

Moreover, we also provide a version of the imputed datasets where some values that were expressed as "a", meaning "not available" (for example the number of PhD students of an Institution offering no PhD programs) are replaced with numbers (in the example, zeroes) to increase usability in some software packages not accepting symbols. They are called **imputed_ETER_dataset.numbers.xlsx** and **imputed_ETER_dataset_bibliometrics.numbers.xlsx**.

Finally, we also provide a version of the imputed datasets where the records are grouped by Institutions, and not just ordered by year as it happens in raw ETER data. They are particularly useful to visualize more easily the evolution of a specific Institution over the years. They are called **imputed_ETER_dataset.reordered.xlsx** and **imputed_ETER_dataset_bibliometrics.reordered.xlsx**

2. Experimental Design, Materials and Methods

The raw data have been gathered through the National Statistical Authorities of the Countries participant in the ETER Project and were downloaded from the ETER Project website: https://www.eter-project.com/#/home. Those data have already passed a data quality check described in [5]. However, they still contained several missing values, often located in the time series describing the detail of each institution (number of students enrolled in each year, number of graduates in each year, total academic staff in each year, etc.).

Data Imputation is the process of replacing the missing values with feasible values being as similar as possible to the original values that have been lost and are now unknown [2]. This task was evidently required for this dataset. However, this kind of data constitute a particularly difficult case for data Imputation techniques, since they contain multivariate time series. Moreover, all the data of an Institution are interconnected: the number of students is not independent from the number of graduates, or from the academic staff, and so on (see also [3]).

We have developed and applied an imputation methodology specifically designed for this kind of data. In particular, depending on the type of missing pattern in the time series, we used two different approaches. When the time series contains at least two non-missing values, we used a technique based on the available values of the sequence called *trend smoothing imputation*, which is basically a combination of regression and averaging. This technique has been designed to be able to capture the trend and the size from the available data and use it to reconstruct the missing part of the sequence.

On the other hand, when the time series is all missing, or contains only one non-missing value, we used a technique based on the use of the values of other similar institutions, called donors, to reconstruct the missing values. This was done by respecting the size of the institution under imputation and the ratios existing among its interconnected variables. The whole methodology has been described in detail in [3], and the corresponding implementation in Python language is provided in [4].

Ethics Statement

This work did not involve the use of human subjects or animals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

The EU Horizon 2020 RISIS2 Project (grant agreement No. 824091) is gratefully acknowledged.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2020.106611.

References

[1] A. Bonaccorsi, C. Daraio (Eds.) Universities and strategic knowledge creation. Specialization and Performance in Europe. Edward Elgar Publisher, Cheltenham (UK), 2007
[2] R. Bruni, Error correction for massive data sets, Opt. Methods Softw. 20 (2005) 295–314.
[3] R. Bruni, C. Daraio, D. Aureli: Imputation techniques for the reconstruction of missing interconnected data from higher educational institutions, Knowl.-Based Syst. 10.1016/j.knosys.2020.106512
[4] R. Bruni, C. Daraio, D. Aureli: Optimization methods for the imputation of missing values in educational institutions data, co-submitted to MethodsX
[5] 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), doi:10.2478/jdis-2020-0029.