data

MDPI

Editorial

Overcoming Data Scarcity in Earth Science

Angela Gorgoglione [1,*], Alberto Castro [2], Christian Chreties [1] and Lorena Etcheverry [2]

[1] Department of Fluid Mechanics and Environmental Engineering (IMFIA), School of Engineering, Universidad de la República, Montevideo 11300, Uruguay; chreties@fing.edu.uy
[2] Department of Computer Science (InCo), School of Engineering, Universidad de la República, Montevideo 11300, Uruguay; acastro@fing.edu.uy (A.C.); lorenae@fing.edu.uy (L.E.)
* Correspondence: agorgoglione@fing.edu.uy

Received: 26 December 2019; Accepted: 30 December 2019; Published: 1 January 2020

Abstract: The Data Scarcity problem is repeatedly encountered in environmental research. This may induce an inadequate representation of the response's complexity in any environmental system to any input/change (natural and human-induced). In such a case, before getting engaged with new expensive studies to gather and analyze additional data, it is reasonable first to understand what enhancement in estimates of system performance would result if all the available data could be well exploited. The purpose of this Special Issue, "Overcoming Data Scarcity in Earth Science" in the *Data* journal, is to draw attention to the body of knowledge that leads at improving the capacity of exploiting the available data to better represent, understand, predict, and manage the behavior of environmental systems at meaningful space-time scales. This Special Issue contains six publications (three research articles, one review, and two data descriptors) covering a wide range of environmental fields: geophysics, meteorology/climatology, ecology, water quality, and hydrology.

Keywords: earth-science data; data scarcity; missing data; data quality; data imputation; statistical methods; machine learning; environmental modeling; environmental observations

1. Introduction

Environmental modeling deals with the representation of processes that occur in the real world in space and time. Based on differential equations, dynamic models mostly describe the processes that transform the environment through time. The spatial interactions and topological rules are mostly managed by geographic information systems (GIS) [1]. These mathematical models heavily rely on the data collected by direct field observations. However, a functional and complete dataset of any environmental variable is difficult to collect because of two main reasons: (i) the low reliability in the measurements (e.g., due to issues related to the equipment location or occurrences of equipment malfunctions); and (ii) the high cost of the monitoring campaigns [2,3]. The lack of an adequate amount of Earth-science data may induce an unsatisfactory and not reliable representation of the response's complexity of an environmental system to any input/change, both natural and human-induced. In this case, before undertaking expensive studies to collect and analyze additional environmental data, it is reasonable to first understand what improvement in estimates of system performance would result if all the available data could be well exploited [4].

Missing data imputation is a crucial task in cases where it is fundamental to use all available data and not neglect records with missing values [5]. Since the 1980s, many techniques to impute missing data have been proposed [6,7]. Generally speaking, the methods for filling in an incomplete dataset can be divided into two main categories: single imputation and multiple imputations [6]. Single imputation, i.e., filling in precisely one value for each missing one, intuitively has many appealing features, e.g., standard complete-data methods can be applied directly, and the substantial effort

required to create imputations needs to be carried out only once. Multiple-imputation is a method of generating multiple simulated values for each missing item to reflect appropriately the uncertainty related to missing data [8].

A well-known and computationally simple method for the imputation of missing data is the mean substitution. However, it can disrupt the inherent structure of the data considerably, leading to significant errors in the covariance/correlation matrix and thereby degrading the performance of the model based on this data set [9]. A slightly better approach is to impute the missing elements from an ANOVA model [8]. More advanced imputation methods have been developed, and several methods and algorithms are now available.

The purpose of this Editorial is twofold: (i) combine and address the contributions of this Special Issue to use them as a basis in this area of science; (ii) encourage communication among the various disciplines by identifying and grouping complementary research solutions.

2. Summary

The main goal of the Special Issue "Overcoming Data Scarcity in Earth Science" in the *Data* journal, was to emphasize the body of knowledge that aims at enhancing the capacity of exploiting the available data to better characterize, understand, predict, and manage the behavior of environmental systems at all practical scales. This Special Issue contains six publications (three research articles, one review, and two data descriptors) covering a wide range of environmental disciplines: hydrology [10], water quality [11], meteorology/climatology [12,13], ecology [14], and geophysics [15].

2.1. Hydrology

In their article, Abraham et al. presented an application of machine learning for classifying soil into hydrologic groups [10]. Based on several soil characteristics such as the value of saturated hydraulic conductivity, and percentages of sand, silt, and clay, the authors trained machine learning models to classify soil into four hydrologic groups (Group A: soils with high infiltration rate and low runoff; Group B: soils with a moderate infiltration rate; Group C: soils with a slow infiltration rate; Group D: a very slow infiltration rate and high runoff potential). Afterward, they compared the results of the classification obtained using four different algorithms, (i) k-Nearest Neighbors (kNN), (ii) Support Vector Machine (SVM) with Gaussian Kernel, (iii) Decision Trees, (iv) Classification Bagged Ensembles and TreeBagger (Random Forest), with those obtained using estimation based on soil texture. Overall, kNN, Decision Tree, and TreeBagger performed better then SVM-Gaussian Kernel and Classification Bagged Ensemble. Among the four hydrologic groups, the authors noticed that group B had the highest rate of false positives.

2.2. Water Quality

Zavareh and Maggioni proposed an approach to analyzing water quality data based on rough set theory (RST) [11]. They collected six water quality indicators (temperature, pH, dissolved oxygen, turbidity, specific conductivity, and nitrate concentration) at the outlet of the catchment that contains the George Mason University campus in Fairfax (VA, United States) over three years (October 2015–December 2017). They evaluated the efficiency of using RST to estimate one water quality indicator based on other given (known) indicators. The authors stated that RST does not require any prior information on the dataset and represents a powerful tool able to deal with uncertainty and vagueness in the sample. Overall, RST was proven capable of finding primary indicators and discovering decision-making rules. RST-based decision-making rules can be a remarkable aid for analysts and planners for their decision-making process.

2.3. Meteorology/Climatology

In their work, Cazes Boezio and Ortelli evaluated the use of data-assimilation techniques from field measurements into initial conditions of atmospheric numerical simulations to obtain wind estimates in

Uruguay (South America), at heights of 100 m above the ground and lower [12]. The wind was assessed with hourly frequency in a regular grid that covers the entire country. The field data to be assimilated was measured with anemometers placed 100 m above the ground in local wind farms. The data was assimilated into initial conditions for the Weather Research and Forecast regional model (WRF) of the National Center of Atmospheric Research (NCAR) using the module for data assimilation included in this model, the WRF-DA module. The authors stated that in addition to its direct use in the numerical prediction process, the results of data assimilation can be considered as "pseudo-observations" of atmospheric variables in regular grids.

In his data-descriptor publication, Mistry introduced a new high-resolution global gridded dataset of climate-extreme indices (CEIs) based on sub-daily precipitation and temperature data from the Global Land Data Assimilation System (GLDAS) [13]. This dataset, called "CEI_0p25_1970_2016", includes 71 annual (monthly in some cases) CEIs at 0.25° × 0.25° gridded resolution, covering 47 years over the period 1970–2016. The author stated that CEI_0p25_1970_2016 fills gaps in existing CEI datasets by encompassing more indices and by being the only comprehensive global gridded CEI data available at high spatial resolution. The data of individual indices are freely downloadable in the commonly used Network Common Data Form 4 (NetCDF4) format. Potential applications of CEI_0p25_1970_2016 include the evaluation of sectoral impacts (e.g., hydrology, agriculture, energy, health), as well as the identification of spatial and temporal patterns that show similar historical of high/low temperature and precipitation extremes.

2.4. Ecology

In their thorough review, Pascoe et al. identified and discussed how the currently available environmental Earth data are lacking concerning their applications in species distribution modeling, mainly when predicting the potential distribution of invasive arthropods that vector pathogens (IAVPs) at significant space-time scales [14]. The authors examined the issues related to the interpolation of weather-station data, and the lack of microclimatic data, which is significant to the environment experienced by IAVPs. Furthermore, they provided some suggestions for filling these data gaps. The optimal resolution of environmental data relevant to IAVP ecology will likely vary according to the species under consideration, but they assumed that this resolution would typically be less than 1 m and hourly. The authors encourage modelers and ecologists to take a proactive approach in collecting small resolution data using data loggers, crowdsourcing, unmanned aerial vehicles or controlled environmental studies. They proposed that these proximally-sensed data, as well as remotely-sensed data, be made open access in a user-friendly database.

2.5. Geophysics

In their work, Bataleva et al. developed a sophisticated geophysical station that collects, processes, and store geophysical information, in particular, electrical and magnetic components of the natural electromagnetic field, useful for the study of geodynamic processes occurring in the Earth's crust and upper mantle [15]. This station is located in the territory of the Bishkek Geodynamic Proving Ground, located in the active seismic zone of the Northern Tien Shan (on the border between China and Kyrgyzstan, Central Asia).

3. Statistics

The following tables (from Tables 1–4) represent some statistics about the publications belonging to the Special Issue "Overcoming Data Scarcity in Earth Science" in the *Data* journal.

Table 1. Brief report of the Special Issue.

Submission	Quantity
Received	9
Published after review	6
Rejected	3
Acceptance rate	66.67%
Median publication time	57 days

Table 2. Type of publications belonging to the Special Issue.

Type of Publication	Quantity	Percentage
Article	3	50
Review	1	17
Data descriptor	2	33
Total	6	100

Table 3. Disciplines covered by the publications of the Special Issue.

Discipline	Quantity	Percentage
Hydrology	1	17
Water quality	1	17
Meteorology/climatology	2	33
Ecology	1	17
Geodynamics	1	17
Total	6	100

Table 4. Countries of the authors.

Country	Quantity	Percentage
Czech Republic	1	5
Italy	5	26
Kyrgyzstan	3	16
Netherland	1	5
United States	7	37
Uruguay	2	11
Total	18	100

Author Contributions: Conceptualization, A.G.; writing—original draft preparation, A.G.; writing—review and editing, A.C., C.C., and L.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We gratefully acknowledge the technical and administrative support of the *Data* journal team. We also want to thank the Authors who contributed towards this Special Issue on "Overcoming Data Scarcity in Earth Science", as well as the Reviewers who provided the authors with suggestions and constructive feedback.

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

References

1. Chaulya, S.K.; Prasad, G.M. Chapter 7—Application of cloud computing technology in mining industry. In *Sensing and Monitoring Technologies for Mines and Hazardous Areas*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 351–396.
2. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.L.; Oki, L.R.; Haver, D.L.; Young, T.M. Uncertainty in the parameterization of sediment build-up and wash-off processes in the simulation of water quality in urban areas. *Environ. Model. Softw.* **2019**, *111*, 170–181. [CrossRef]

3. Gorgoglione, A.; Gioia, A.; Iacobellis, V.; Piccinni, A.F.; Ranieri, E. A rationale for pollutograph evaluation in ungauged areas, using daily rainfall patterns: Case studies of the Apulian region in Southern Italy. *Appl. Environ. Soil Sci.* **2016**, *2016*, 9327614. [CrossRef]
4. Gorgoglione, A.; Gioia, A.; Iacobellis, V. A Framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. *Sustainability* **2019**, *11*, 4933. [CrossRef]
5. Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a realbreast cancer problem. *Artif. Intell. Med.* **2010**, *50*, 105–115. [CrossRef] [PubMed]
6. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
7. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; CRC Press: Boca Raton, FL, USA, 2010.
8. Junninen, H.; Niska, H.; Tuppurainen, K.; Ruuskanen, J.; Kolehmainen, M. Methods for imputation of missing values in air quality data sets. *Atmosph. Environ.* **2004**, *38*, 2895–2907. [CrossRef]
9. Tutz, G.; Ramzan, S. Improved methods for the imputation of missing data by nearest neighbor methods. *Comput. Stat. Data Anal.* **2015**, *90*, 84–99. [CrossRef]
10. Abraham, S.; Huynh, C.; Vu, H. Classification of soils into hydrologic groups using machine learning. *Data* **2020**, *5*, 2. [CrossRef]
11. Zavareh, M.; Maggioni, V. Application of rough set theory to water quality analysis: A case study. *Data* **2018**, *3*, 50. [CrossRef]
12. Cazes Boezio, G.; Ortelli, S. Use of the WRF-DA 3D-Var data assimilation system to obtain wind speed estimates in regular grids from measurements at wind farms in Uruguay. *Data* **2019**, *4*, 142. [CrossRef]
13. Mistry, M.N. A high-resolution global gridded historical dataset of climate extreme indices. *Data* **2019**, *4*, 41. [CrossRef]
14. Pascoe, E.L.; Pareeth, S.; Rocchini, D.; Marcantonio, M. A Lack of "environmental earth data" at the microhabitat scale impacts efforts to control invasive arthropods that vector pathogens. *Data* **2019**, *4*, 133. [CrossRef]
15. Bataleva, E.; Rybin, A.; Matiukov, V. System for collecting, processing, visualization, and storage of the MT-Monitoring data. *Data* **2019**, *4*, 99. [CrossRef]