Surface-Water Quality Monitoring

Subjects: Environmental Sciences
Contributor: Lorena Etcheverry

The monitoring of surface-water quality followed by water-quality modeling and analysis are essential for generating effective strategies in surface-water-resource management. However, worldwide, particularly in developing countries, water-quality studies are limited due to the lack of a complete and reliable dataset of surface-water-quality variables.

Keywords: data scarcity ; water quality ; missing data ; univariate imputation ; multivariate imputation ; machine learning ; hydroinformatics

1. Introduction

Monitoring, modeling and management represent the three foundations for building an effective pollution-control strategy [1]. They strictly depend on each other: there is no management without modeling and no modeling without exhaustive monitoring. Therefore, any problem related to data collection is then reflected in the performance of the modeling and management phases. Consequently, it is crucial first to acknowledge what improvement would result if all the available data could be well exploited [2].

The issue of missing data frequently occurs in environmental fields due to sensor failures, weak or inexistent strategy for coordinating monitoring campaigns, a change in the measurement site, in data collectors or to the equipment over time, budget issues [3][4]. Such water-quality data problem is particularly significant in developing countries where monitoring stations and monitoring frequency is scarce, and the percentage of missing data is exceptionally high [5].

It is possible to deal with missing data in two different ways: deletion or imputation [6]. Deletion consists of removing the observations or the features characterized by missing values, while imputation involves reconstructing missing data. Deletion is typically the default method adopted since it is rapid and straightforward [7]. However, in several fields, there are many examples in which such a technique presented some restrictions. It reduces the dataset size and may lead to biased results and a loss of critical information, mainly when a high percentage of missing values characterizes the dataset. Among the most straightforward imputation techniques, there are mean imputation and linear interpolation (which rely only on the available time-series data to perform the imputation), arithmetic, and weighted averaging. However, these techniques have shown poor performance when the dataset is characterized by a significant length of the missing sequence [5].

Another common approach used to fill in missing data, which is part of the univariate imputation methods, is to use information from the neighboring monitoring stations. The inverse distance weight (IDW) is a technique that has been successfully adopted for environmental datasets, particularly for meteorological variables [8][9][10][11].

In the last decade, progressively more advanced techniques have been adopted to reconstruct environmental time series [12][13]. Among them, machine-learning techniques that can handle multivariate inputs are the most widely used. Aguilera et al. [5] adopted three different methods (spatio-temporal kriging, multiple imputations by chained equations through predictive mean matching and random forest) to reconstruct daily precipitation time series characterized by extreme missingness (>90%). They found that spatio-temporal kriging simulates rainfall distribution under missing chronological patterns more reliably than the other two techniques. Sattari et al. [14] provided an in-depth comparison of ten different statistical and machine-learning models to impute monthly precipitation data. Computational results showed that arithmetic averaging, multiple linear regressors and non-linear iterative partial least squares perform best among the classical statistical methods. The multiple imputation technique performed best when rainfall data from more than one dependent station were considered. In addition, Barrios et al. [10] compared the performance of five models for filling monthly precipitation records, finding that artificial neural network, multiple linear regression and IDW showed the best performance.

Most of the imputation works presented in the scientific literature refer to meteorological variables and, sometimes, to hydrologic variables like streamflow [15]. To our knowledge, there are few works related to the imputation of water-quality data. Tabari and Talaee [16] employed artificial neural networks to successfully recover missing values of 13 water-quality parameters at five monitoring stations in the South of Iran. Srebotnjak et al. [17] adopted hot-deck imputation to improve a country-level water quality index, calculated by considering dissolved oxygen, electrical conductivity, *pH*, total phosphorus and total nitrogen. Ratolojanahary et al. [7] assessed for the first time the problem of high omission rate (even higher than 80%) in a water-quality dataset by adopting four machine-learning models (random forest, boosted regression trees, k-nearest neighbors and support vector regression). However, there is no comprehensive evaluation of different types of imputation models in the context of water-quality data characterized by a high percentage of incompleteness.

2. Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach. Take Uruguay as An Example

Effective water-resource management requires the analysis of a large number of water-quality information over space and time. However, in many parts of the world, particularly in developing countries, the monitoring of water-quality variables is usually characterized by few monitoring stations over the territory, where observations are recorded with a low frequency and are characterized by an important percentage of missing data. Therefore, in this study, we evaluated the performance of several statistical and machine-learning techniques (univariate and multivariate) in imputing a water-quality dataset characterized by eight water quality variables measured at six monitoring stations. Particularly, we aimed to augment the water-quality dataset, from bi-monthly to monthly frequency. The percentage of missing values ranges between 50% and 70% (high missingness percentage), and the water-quality variables are characterized by a high temporal and spatial distribution. The study area considered was one of the most critical Uruguayan watersheds, Santa Lucía Chico, since it provides water to more than 60% of the national population. This was an interesting study area to analyze since it is a mixed lotic and lentic system and the six monitoring stations are located along the mainstream (SLC01, SLC02 and PS01) and in the reservoir (PS03, PS04 and PS02). In this way, it was appealing to assess the performance of several models in these two different surface-water bodies.

There are few related works on the imputation of water-quality data, and they are relatively recent. In 2012, Srebotnjak et al. [17] showed that hot-deck imputation can improve geographical coverage of a country-level water quality index, calculated considering dissolved oxygen, electrical conductivity, *pH*, total phosphorus and total nitrogen. This water-quality index is a composite indicator to track water quality over time and space, easily interpretable since it varies from 0 to 100. Still, it does not allow a detailed analysis of each water-quality variable used to calculate it. Therefore, this type of index does not allow us to answer scientific questions such as which compounds are significant indicators for specific land use categories or the spatio-temporal behavior of a particular problematic compound in a particular area of study. To overcome these limitations, we decided to directly impute each water-quality variable and not a global index, which allows us to use the imputed data for more advanced analyses.

In 2015, Tabari and Talaee [16] obtained acceptable results (RMSE ranges between 0.016 and 4475) in imputing a large dataset of water-quality information (13 variables) measured, with a monthly frequency, at five monitoring sites along the Maroon River (Southwest of Iran). It should be noted that this study has already adopted the concept of helper variables to improve the imputation process based on the correlations among water-quality variables. The correlation between *EC* and *Turb* that we used in our analysis is confirmed in this study. In Tabari and Talaee [16], the results were insufficient for *EC*, *Turb* and total dissolved solids (*TDS*) at all monitoring stations, showing RMSE values between 100 and higher than 4000. They employed only two artificial neural networks as imputation models: multilayer perceptron and radial bias function. In our study, we improved such results using more imputation techniques and founding that SVR model shows better performance for *EC* and *Turb*.

In 2019, Ratolojanahary et al. [7] tackled for the very first time the problem of high rate missingness (higher than 80%) in a water-quality dataset of a drinking water well employing four machine learning models (RF, KNNR, SVR and boosted regression trees, similar to our AB). Their outcomes showed that SVR provides the best performance (notably in terms of average prediction error). However, this study does not introduce the temporal dimension into the imputation process, and, therefore, temporal variability of water-quality parameters is not considered a challenge. Spatial variability is also not addressed, as the authors analyzed water well. These aspects are included in our study. Furthermore, we confirm that the performance of SVR is better than AB and KNNR in the imputation of water quality data.

It is also important to note that our work pioneered the use of IDW for water-quality data imputation, and this method performed the best among all the methods analyzed. Some recent works proposed using IDW to interpolate water quality

in scenarios where spatial variability may be negligible, as in the case of lakes [18] or where temporal variability is low, as in the case of groundwater [19].

Some of the correlations found in our work were also reported in previous studies in the same study area [20][21][22][23]: a robust correlation among nitrogen compounds, in its dissolved and particle-bound form; a strong inverse correlation between by T_w and *DO*.

3. Conclusions

We tackled the challenge of data imputation in a multivariate water-quality dataset characterized by a high percentage of missing data (between 50% and 70%). In particular, the variables T_w, *EC*, *pH*, *DO*, *TN*, NO_2^-, NO_3^- and *Turb* of six monitoring stations located along the Santa Lucía Chico river (Uruguay) were considered for this study. Adopting a multi-model approach was crucial since the best model for imputing any water-quality variable does not exist. The statistical and machine-learning models implemented were IDW, RFR, RR, BR, AB, HR, SVR and KNNR.

The imputation outcomes were overall adequate. More than 76% of the imputed data can be considered “satisfactory” (NSE > 0.45). This was validated by calculating PBIAS (>96% of the imputed data is “satisfactory”) and KGE (all the imputations are considered “good”). It is interesting to notice that the performance is always remarkable at the three monitoring stations located in the Paso Severino reservoir, while they may be “unsatisfactory” at some monitoring stations located along the Santa Lucía Chico river (upstream the reservoir). Among the implemented models, IDW was chosen as the best model 17 times since it is the only model that considers the temporal and spatial variability that characterizes the variables under study.

We pave the path to future water-quality research in the watershed under study (e.g., implementation of reliable modeling tools, water-quality prediction and scenario analysis). Hopefully, the results obtained will help water managers and researchers worldwide make the most of existing water-quality data to improve modeling and generate effective pollution-control strategies.

Our current results are promising, but we believe that it is possible to improve the present methodology by integrating physical knowledge that considers the spatial information of the available water-quality data. Our future work intends to transform the current approach, based on machine learning, into a hybrid method where the data-driven techniques incorporate physical aspects during their training.

References

1. Whitehead, P.; Dolk, M.; Peters, R.; Leckie, H. Water Quality Modelling, Monitoring, and Management. In Water Science, Policy, and Management; Dadson, S.J., Garrick, D.E., Penning-Rowsell, E.C., Hall, J.W., Hope, R., Hughes, J., Eds.; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2019.
2. Gorgoglione, A.; Castro, A.; Chreties, C.; Etcheverry, L. Overcoming Data Scarcity in Earth Science. Data 2020, 5, 5.
3. Teegavarapu, R.S.V.; Aly, A.; Pathak, C.S.; Ahlquist, J.; Fuelberg, H.; Hood, J. Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: Use of optimal weighting parameters and nearest neighbour-based corrections. Int. J. Climatol. 2018, 38, 776–793.
4. Mital, U.; Dwivedi, D.; Brown, J.B.; Faybishenko, B.; Painter, S.L.; Steefel, C.I. Sequential imputation of missing spatio-temporal precipitation data using random forests. Front. Water 2020, 2, 20.
5. Aguilera, H.; Guardiola-Albert, C.; Serrano-Hidalgo, C. Estimating extremely large amounts of missing precipitation data. J. Hydroinformatics 2020, 22, 578–592.
6. Buhi, E. Out of sight, not out of mind: Strategies for handling missing data. Am. J. Health Behav. 2008, 32, 83–92.
7. Ratolojanahary, R.; Ngouna, R.H.; Medjaher, K.; Junca-Bourié, J.; Dauriac, F.; Sebilo, M. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. Expert Syst. Appl. 2019, 131, 299–307.
8. Lo Presti, R.; Barca, E.; Passarella, G. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environ. Monit. Assess. 2010, 160, 1–22.
9. Chen, F.W.; Liu, C.W. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. Paddy Water Environ. 2012, 10, 209–222.

10. Barrios, A.; Trincado, G.; Garreaud, R. Alternative approaches for estimating missing climate data: Application to monthly precipitation records in South-Central Chile. For. Ecosyst. 2018, 5, 28.
11. Gong, G.; Mattevada, S.; O'Bryant, S.E. Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas. Environ. Res. 2014, 130, 59–69.
12. Aissia, M.-A.B.; Chebana, F.; Ouarda, T. Multivariate missing data in hydrology–Review and applications. Adv. Water Resour. 2017, 110, 299–309.
13. Chivers, B.D.; Wallbank, J.; Cole, S.C.; Sebek, O.; Stanley, S.; Fry, M.; Leontidis, G. Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. J. Hydrol. 2020, 588, 125126.
14. Sattari, M.-T.; Rezazadeh-Joudi, A.; Kusiak, A. Assessment of different methods for estimation of missing data in precipitation studies. Hydrol. Res. 2017, 48, 1032–1044.
15. Oriani, F.; Borghi, A.; Straubhaar, J.; Mariethoz, G.; Renard, P. Missing data simulation inside flow rate time series using multiple-point statistics. Environ. Model. Softw. 2016, 86, 264–276.
16. Tabari, H.; Talaee, P.H. Recontrsuction of river water quality missing data using artificial neural networks. Water Qual. Res. J. Can. 2015, 50, 4.
17. Srebotnjak, T.; Carr, G.; de Sherbinin, A.; Rickwood, C. A global Water Quality Index and hot-deck imputation of missing data. Ecol. Indic. 2012, 17, 108–119.
18. Jácome, G.; Valarezo, C.; Yoo, C. Assessment of water quality monitoring for the optimal sensor placement in lake Yahuarcocha using pattern recognition techniques and geographical information systems. Environ. Monit. Assess. 2018, 190, 259.
19. Kanga, I.S.; Naimi, M.; Chikhaoui, M. Groundwater quality assessment using water quality index and geographic information system based in Sebou River Basin in the North-West region of Morocco. Int. J. Energ. Water Res. 2020, 4, 347–355.
20. Gorgoglione, A.; Gregorio, J.; Ríos, A.; Alonso, J.; Chreties, C.; Fossati, M. Influence of land use/land cover on surface-water quality of Santa Lucía river, Uruguay. Sustainability 2020, 12, 4692.
21. Goyenola, G.; Meerhoff, M.; Teixeira-de Mello, F.; González-Bergonzoni, I.; Graeber, D.; Fosalba, C.; Vidal, N.; Mazzeo, N.; Ovesen, N.B.; Jeppesen, E.; et al. Phosphorus dynamics in lowland streams as a response to climatic, hydrological and agricultural land use gradients. Hydrol. Earth Syst. Sci. Discuss. 2015, 12, 3349–3390.
22. Gorgoglione, A.; Alonso, J.; Chreties, C.; Fossati, M. Assessing temporal and spatial patterns of surface-water quality with a multivariate approach: A case study in Uruguay. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Changchun, China, 21–23 August 2020; Volume 612, p. 012002.
23. Barreto, P.; Dogliotti, S.; Perdomo, C. Surface water quality of intensive farming areas within the Santa Lucia River basin of Uruguay. Air Soil Water Res. 2017, 10, 1178622117715446.