An Experimental Comparison of Methods to Handle Missing Values in Environmental Datasets

A. D. Maldonado
University of Almería, ana.d.maldonado@ual.es

P. A. Aguilera
University of Almería, aguilera@ual.es

A. Salmeron
University of Almería, antonio.salmeron@ual.es

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

Part of the Civil Engineering Commons, Data Storage Systems Commons, Environmental Engineering Commons, Hydraulic Engineering Commons, and the Other Civil and Environmental Engineering Commons

https://scholarsarchive.byu.edu/iemssconference/2016/Stream-C/3

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
An Experimental Comparison of Methods to Handle Missing Values in Environmental Datasets

A.D. Maldonado¹, P.A. Aguilera², A. Salmerón¹
¹ Department of Mathematics, University of Almería, Almería, Spain (ana.d.maldonado@ual.es, antonio.salmeron@ual.es)
² Informatics and Environment Laboratory, Department of Biology and Geology, University of Almería, Almería, Spain (aguilera@ual.es)

Abstract: This paper reports on a comparison of different techniques to handle missing data in a real environmental dataset containing missing values. In particular, we handled a dataset related to a surface water quality index. The chosen techniques were regression imputation by linear regression, a model tree and a Bayesian network; multiple imputation by chained equations; and data augmentation by a Bayesian network. The models were tested by analyzing the predictive maps and by comparing the density function of the predicted variable with the observed one. The experimental results showed that the imputation by linear regression and the multiple imputation by chained equations maintained the characteristics of the response variable more than the remaining models.

Keywords: IBMWP; surface waters; incomplete datasets; model comparison

1 INTRODUCTION

Incomplete datasets are fairly common in Environmental Sciences and Ecology. Missing data may occur due to different reasons, including insufficient sampling, loss of samples or failure of measuring instruments (Junninen et al. 2004, Rangeti et al. 2015). The impact of the missing values depends on the data missingness mechanism, which is typically classified into 3 categories: missing completely at random (MCAR), meaning that the probability of an observation to be missing is not related to its own values nor to other variables; missing at random (MAR), in which the probability of an observation to be missing is related to the values of other variables; and missing not at random (MNAR); where the probability of an observation to be missing is related to its own values. In addition, there is a fourth type of missingness, called filtered or censored values, which refers to the structure of the data, regarding data that cannot exist (Conrady and Jouffret 2011, Schafer and Graham 2002).

The methods to deal with missing values range from deletion to sophisticated imputation algorithms. When there are few MCAR values, it is acceptable to delete those observations (listwise deletion). However, deletion when the missing data do not occur at random introduces bias because the observed values do not represent the entire sample (Ellington et al. 2015, Wayman 2003). The opposite of deletion is the imputation of values. The simplest and least recommended imputation technique is the unconditional mean imputation. This technique fills in the blanks with the mean of the marginal distribution of the non-missing values. More advanced is the conditional mean imputation (also known as regression imputation) procedure, which consists in fitting a model using the observed values and afterwards inferring the missing values from the estimated model. Both imputation methods have been largely criticized since the former reduces variability and underestimates standard errors and the latter overestimates test statistics and reduces variance since the imputed data are estimated from a fit model (Soley-Bori 2013, Fichman 2003). Furthermore, from the point of view of Ecology, imputing with the mean is not acceptable, since ecosystems are considered to be heterogeneous; thus, the mean homogenizes the system.

A step further was made when multiple imputation (MI) (Rubin 1976) and data augmentation (DA) techniques were developed. In MI, imputed values are drawn from the distribution of the variable, filling in the blanks $m$ times; then the $m$-imputed datasets are analyzed by standard procedures, and the resultant parameter estimates from each analysis are combined to produce a final result (Nakagawa and Freckleton 2008, Soley-Bori 2013, Wayman 2003). DA procedures are iterative
techniques that use the posterior distribution of the missing variables to replace the blanks until the parameter estimates no longer improve. Examples of DA procedures are Maximum likelihood (ML) estimation or expectation maximization algorithm (EM), among others.

It is worthwhile to highlight that the main goal when dealing with missing data is not to recover those values (by filling in with a substitute value), but to perform proper predictions over the response variable and to maintain the characteristics of their distribution along with the relationships among them (Barceló 2008). However, the problem becomes difficult to tackle when the only variable containing missing values is the response since we are interested in knowing plausible values for this variable.

The goal of this paper is to compare different techniques to handle missing data in Environmental datasets, in particular in a dataset concerning surface waters, where the response variable is a water quality index: the Iberian Bio-monitoring Working Party index (IBMWP). To do so, we applied different methods to the same dataset, where the only variable containing missing values is the response. In particular, we compared the conditional mean technique with methods based on multiple imputation and data augmentation.

2 METHODOLOGY

2.1 Study Area and Data

The study area is Andalusia, the southernmost region in Spain. The main mountain ranges are the Sierra Morena mountain range in the North and the Baetic Systems in the South (comprising the Prebaetic, Subbaetic and Penibaetic Systems), which are separated by the Baetic Depression, the lowest territory in Andalusia (Figure 1). Andalusia is divided into 6 official watersheds, which are subdivided into 62 official sub-watersheds.

A set of thematic maps, available at the Andalusian Environmental Information Network (REDIAM), was used to build our dataset, including climatic, topographic, land-cover and surface waters maps. The coordinate system for all the datasets is based on the European Terrestrial Reference System 1989 (ETRS89). The geographic information system ArcGis (ESR®ArcMap™10.2.2) was used to obtain the variables described in Table 1.

Figure 1. Protected areas and main geomorphological features in Andalusia.

First, the contributing area of a set of sampling points, containing information about the IBMWP and having been taken at different rivers, was delineated in order to determine the sampling units of the study area (Maldonado et al. 2015). However, these sampling units, which represent watersheds, did not cover the entire study area. Therefore, we randomly distributed "missing" sampling points at those locations where no measure of IBMWP had been taken before and calculated their contributing area. We obtained 661 watersheds, 333 of which did not have information regarding IBMWP, either for the non-existence of the sampling point or because the existent sampling point had no information.

Afterwards, the complete set of watersheds was used to delimit the value of the remaining variables. Finally, a dataset composed of 17 variables taking values over 661 observations (watersheds) was obtained. Table 1 summarizes the variables selected to carry out the experiments presented in this paper. Since these variables were measured or calculated in different units, the data were rescaled to interval [0,1] to prevent numerical instability problems.
### Table 1. Summary of variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBMWDP</td>
<td>Iberian Bio-monitoring Working Party index. Data collected from sampling stations belonging to the Biological Quality Network of the EU WFD in rivers in Andalusia.</td>
</tr>
<tr>
<td>Point X (m)</td>
<td>Longitude coordinate of each sampling point.</td>
</tr>
<tr>
<td>Point Y (m)</td>
<td>Latitude coordinate of each sampling point.</td>
</tr>
<tr>
<td>Stream order</td>
<td>Order of the stream in which the sampling point was taken.</td>
</tr>
<tr>
<td>K</td>
<td>Average soil permeability, encoded as 1= low, 2= medium and 3 = high.</td>
</tr>
<tr>
<td>PET (mm)</td>
<td>Average annual potential evapotranspiration for the 30-year period 1971-2000 of each watershed.</td>
</tr>
<tr>
<td>PPT (mm)</td>
<td>Average annual rainfall for the 30-year period 1971-2000 of each watershed.</td>
</tr>
<tr>
<td>T (°C)</td>
<td>Average annual mean temperature for the 30-year period 1971-2000 of each watershed.</td>
</tr>
<tr>
<td>Slope (%)</td>
<td>Average slope of each watershed.</td>
</tr>
<tr>
<td>Z (m a.s.l.)</td>
<td>Average elevation of each watershed.</td>
</tr>
<tr>
<td>Area (km²)</td>
<td>Watershed area. Upslope area that drains water to each sampling point. These watersheds were delineated using ArcGis Hydrology toolset.</td>
</tr>
<tr>
<td>Evenness Shannon Index</td>
<td>Index of landscape structure (Atauri and de Lucio 2001), being calculated from the 172 land-uses integrating the Spanish Land Occupation Information System (SIOSE).</td>
</tr>
<tr>
<td>Artificial (%)</td>
<td>Percentage of land-use that refers to human modified land covers, such as cities, artificial bodies of water or areas under construction. It comprises 77 out of 172 land-use classes in SIOSE.</td>
</tr>
<tr>
<td>Herbaceous crops (%)</td>
<td>Percentage of herbaceous monocultures, not distinguishing between rain-fed and irrigated crops. It comprises 1 land-use class in SIOSE.</td>
</tr>
<tr>
<td>Olive grove (%)</td>
<td>Percentage of olive groves. It comprises 1 land-use class in SIOSE.</td>
</tr>
<tr>
<td>Other crops (%)</td>
<td>Percentage of agricultural land-uses with low representativeness in the study area, including paddy fields, greenhouses, woody crops and associations of crops. It comprises 27 out of 172 land-use classes in SIOSE.</td>
</tr>
<tr>
<td>Natural (%)</td>
<td>Percentage of land-use that refers to natural ecosystems, including non-modified bodies of water, forest, shrub and grassland covers. It comprises 66 out of 172 land-use classes in SIOSE.</td>
</tr>
</tbody>
</table>

### 2.2 Techniques for Handling Missing Data

Let \( Y \) be the response variable, with \( Y_{\text{obs}} \) being the observed values and \( Y_{\text{mis}} \) the missing ones. Let \( X = (X_1, ..., X_n) \) be a set of predictive complete variables, with \( X_{\text{obs}} \) denoting the cases where \( Y = Y_{\text{obs}} \). We have considered the following methods for handling missing data:

- **Conditional mean imputation.** More precisely, we considered three methods in this group: (i) Imputation by multiple linear regression (LR). A multiple linear regression is fit with the observed data, which returns predicted values as imputations. (ii) Imputation by regression tree (M5P). The M5P algorithm (Wang and Witten 1997) implemented in Weka software (Hall et al. 2009) was used to build a regression tree, where each branch determines a region of possible values of the explanatory variables and each leaf contains a linear regression fit to the data corresponding to its branch. This algorithm ignores the missing values of the response variable, which are imputed from the model prediction. The tree was pruned in order to avoid overfitting. (iii) Imputation by Bayesian network (TANcm). A tree augmented naive Bayes (TAN) (Friedman et al. 1997) was built with the observed data and the missing values were imputed using the model prediction. In TAN models, each explanatory variable has one more parent besides the response variable. The relationships between the variables were modeled using Mixtures of Truncated Exponentials (MTEs) functions (Moral et al. 2001).

- **Multiple Imputation by Chained Equations (MICE).** This is a 3-stage technique: imputation, analysis and pooling. We followed the default procedure by the mice R package (van Buuren and Groothuis-Oudshoorn 2011), which imputes univariate missing data using predictive mean matching (PMM). PMM is truncated to the range of \( Y_{\text{obs}} \) since the imputed values are taken from one of the closest \( Y_{\text{obs}} \) in the dataset. More precisely, the PMM algorithm estimates a linear regression of \( Y_{\text{obs}} \) on \( X_{\text{obs}} \) with coefficients \( \beta \). Afterwards, a random draw from the posterior predictive distribution of the coefficients is made, producing new coefficients, \( \beta^* \), which are used to generate predictions for \( Y \) for all cases (\( Y_{\text{obs}} \) and \( Y_{\text{mis}} \)).
Whenever \( Y = Y_{\text{mis}} \), the algorithm identifies a set of cases with \( Y = Y_{\text{obs}} \) whose predictions are close to the predicted value for \( Y_{\text{mis}} \), and randomly chooses one of them to substitute the missing value (Heitjan and Little 1991; Schenker and Taylor 1996). In our experiments, this process was repeated 20 times. The 20 imputed datasets were analyzed using multiple linear regression, obtaining a regression equation for each imputed dataset. Finally, the regression coefficients of the 20 linear models were pooled, i.e. were averaged, into a single multiple linear regression, which was used to predict the values of the response variable.

- Data augmentation by Bayesian network (TAN-DA). Developed by Fernández et al. (2010), it is a variation of the DA algorithm (Tanner and Wong 1987). A TAN model was obtained using an iterative algorithm that generates the missing values of the response variable from its conditional expectation given the explanatory variables. This algorithm iterates until the root mean square error of the model no longer improves.

2.3 Testing Methods

Since real data were not available, comparing the predictions of the missing values with the real measures was not possible. In order to test whether or not the methodologies applied gave plausible results, the distribution of the imputed variable was compared to the distribution of the observed variable by means of density and scatter plots. We also analyzed the complete maps drawn from the predictive models to assess whether or not they are reasonable. The predictive maps were obtained from the evaluation of the sampling points by means of the aforementioned methods. Each point was colored with an intensity varying from dark green for high IBMWP index to dark red for low IBMWP index. For graphical reasons, watersheds contributing to each sampling point were colored instead.

3 RESULTS AND DISCUSSION

Figure 2 shows the predictive maps obtained from the 5 methods along with the observed map. The maps show the estimated value of the IBMWP index in surface waters, with high IBMWP indicating high water quality. The results were reclassified into 5 categories for the sake of legibility: very low [0, 35], low [36, 70], moderate [71, 105], high [106, 140] and very high [141, 280]. These intervals are the same for the 6 maps presented in figure 2. The qualitative description of the interval (low, moderate, high) is only used for reference purposes, since the evaluation of the ecological status of rivers usually considers other characteristics of the watershed (Rico et al. 1992). The maps will be analyzed according to the geomorphological features shown in Figure 1: the Baetic Depression, the Sierra Morena mountain range and the Baetic Systems.

The Baetic Depression is the region with more similarities among the predictive maps. This area has fertile soils and a high agricultural production, mainly comprising rain-fed herbaceous crops in the low-lying plain and irrigated herbaceous crops along the Guadalquivir River. There are 3 protected areas, including the Doñana National Park, which is the largest reserve of bird in Europe. The 5 models predicted watersheds with low to very low IBMWP index in the Baetic Depression. Doñana's low index prediction may be due to the pressures suffered by the park, including the Aznalcollar mine disaster in 1998 (Grimalt et al. 1999), the strawberry farms or the annual pilgrimage of El Rocío, among others. The observed values in this region range from low to very low. Therefore, intensive agriculture and constant pressures on the protected areas explain the predictions made by the 5 models.

The Sierra Morena mountain range shows some differences among the predictive maps, especially in the El Andévalo County, which is located in the westernmost part of this mountain range and has low fertility soils, with grasslands and shrubs being the natural dominant vegetation in the area. In the past, the El Andévalo County based a great deal of its economy on mining and cattle industry, until the former decayed and the latter allowed the emergence of the Spanish Dehesa (Olea and San Miguel-Ayanz 2006). Nowadays, timber industry is becoming more important in this area. According to the LR, MICE and M5P models, the IBMWP index is low in this area, while TANda predicted moderate values, and TANcm ranged from high to very high values for the index. The observed data in this County show watersheds with very low IBMWP index. The economic activity in the region may be the reason for the observed water quality. Therefore, the predictions made by LR, MICE and M5P seem to be more reasonable in this case.
Figure 2. Maps showing the predicted values obtained from each approach, except for (a), which represents the observed data. (b), (d) and (f) show the predicted values based on conditional mean, using M5P, LR and TAN algorithms, respectively; (c) shows the predictions by the pooled multiple linear regression obtained by means of the MICE procedure; (e) shows the predicted values obtained from the TAN-DA algorithm.

The remaining Counties in Sierra Morena differ from the El Andévalo in elevation and in protected areas, with the former being at a higher elevation and having 6 designated protected areas. The difference in elevation influences the land-uses in Sierra Morena, with the exploitation of the Spanish Dehesa playing an essential role in the economy of the area. The northernmost lands are mainly covered by rain-fed crops and Spanish Dehesa. The observed data are scarce and mainly show high to very high values in the East and low to very high values in the West. The good state of preservation of this territory may be due to emigration, with about a 50% of the population having left the countryside by the middle of the 20th century (Rescia 2010). LR and MICE mainly predicted moderate values in the western sector and high values in the eastern one. TANDa mainly predicted high, with some moderate values; TANDc mainly predicted very high and high, with some low values; and M5P mainly predicted values ranging from low to high, with some very low values. Based on the socioeconomic information and the observed data, predictions obtained by TANDa, LR and MICE seem to be more plausible in this case.

The Baetic Systems contain the highest peaks in the study area and 14 protected areas. The observed data show higher water quality in the Prebaetic and Penibaetic Systems than in the Subbaetic System, which could be due to the higher elevation and steepness of the former, discouraging the introduction of agricultural practices as a consequence of inaccessibility. The 5 models predicted high to very high water quality in terms of IBMWP index in those watersheds located at higher elevations, in a protected area and mainly covered by natural vegetation, which matches the observed data. On the other hand, the easternmost area is an arid region, with abrupt orography in
the inner area contrasting with flat areas in the littoral. The observed values are scarce in this particular region, with the index ranging from very low to moderate. LR, MICE and TANda mainly predicted low to moderate values, whereas M5P and TANcm mainly predicted values ranging from very low to very high for the same region. In the Southeast, MICE predicted a very low value for a particular watershed, which corresponds to a highly populated area, whereas LR and TANda predicted low, M5P predicted moderate and TANcm predicted high values for the same watershed. LR and TANdm also predicted high values at some upstream watersheds. The results drawn by LR, MICE and TANda seem to be more feasible; however, reaching a conclusion about this arid region is speculative.

In order to check whether or not the characteristics of the observed variable had been preserved, we compared the distribution of the predicted variables with the observed variable. Figure 3 shows the density functions of the variables. The black curve represents the observed values whereas the colored curves represent the predicted values obtained from each model. The observed density curve is right-skewed, with a mean of 73.64 and a standard deviation of 63.69. The most different density curve was drawn by the M5P model, having a shorter range and several peaks. The density curve by TANcm was bimodal, concentrating the density approximately around the values 50, corresponding to the flatlands of the Baetic Depression, and 130, corresponding to the forested watersheds. LR and MICE showed almost identical density curves, which were similar to the observed density function, but less skewed. TANda showed a density curve similar to LR and MICE, but with 3 peaks, due to the imputation of more high values.

Figure 2. Probability density functions of the observed (black) and the predicted (colored) values.

Figure 4 shows dot-plots for the observed and the predicted values. These plots represent the dispersion of the variable given its order in the dataset. In the "observed data" plot, the black dots represent the observed data; whereas, in the remaining plots, the black dots represent the predicted observed data and the red dots the predicted missing data. It is noticeable that LR, MICE and M5P imputed more low values than TANda or TANcm. In the M5P plot, there is a large concentration of imputed values at 60, which explains the highest peak present in its density function.
4 CONCLUSIONS

In this paper we sought to complete the blanks in the proposed study area. Since the missing values could not be recovered, the models had to be evaluated by expert knowledge, concluding that the predictions by LR and MICE where plausible in the entire study area. The results obtained from TANda were also reasonable, except for the El Andévalo area. M5P and TANcm models were the least reliable. These conclusions were supported by the density curves, with LR and MICE being the models that maintained the characteristics of the response variable the most.

From the aquatic ecology point of view, all 5 models managed to identify the transition between the Baetic Depression and the mountain ranges, with the former being severely impaired and the latter having better surface water quality. The analysis of the maps highlighted that the water quality index, IBMWP, is susceptible to decrease in surface waters running on vast low-lying areas covered by monocultures.

Figure 4. Scatterplot of the observed values (top left) and the predicted values, with the black dots representing the predicted values that were observed and the red dots the predicted values that were missing.

ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness, through the TIN2013-46638-C3-1-P project. A.D. Maldonado is being supported by the Spanish Ministry of Education, Culture and Sport through an FPU research grant, FPU2013/00547.

REFERENCES

A.D. Maldonado et al. / An experimental comparison of methods to handle missing values in environmental datasets


