EGU21-10004, updated on 26 Oct 2021
https://doi.org/10.5194/egusphere-egu21-10004
EGU General Assembly 2021

Missing values and lack of information in water management datasets: an approach based on Bayesian Networks

Rosa F Ropero[1], M Julia Flores[2], and Rafael Rumí[1]
[1]University of Almería, Dpt. Mathematics, Data Analysis Research Group , Spain (rosa.ropero@ual.es)
[2]Computing Systems Department, SIMD I3A, University of Castilla-la Mancha, Campus Univ., Albacete, Spain

Environmental data often present missing values or lack of information that make modelling tasks difficult. Under the framework of SAICMA Research Project, a flood risk management system is modelled for Andalusian Mediterranean catchment using information from the Andalusian Hydrological System. Hourly data were collected from October 2011 to September 2020, and present two issues:

- In Guadarranque River, for the dam level variable there is no data from May to August 2020, probably because of sensor damage.
- No information about river level is collected in the lower part of Guadiaro River, which make difficult to estimate flood risk in the coastal area.

In order to avoid removing dam variable from the entire model (or those missing months), or even reject modelling one river system, this abstract aims to provide modelling solutions based on Bayesian networks (BNs) that overcome this limitation.

Guarranque River. Missing values.

Dataset contains 75687 observations for 6 continuous variables. BNs regression models based on fixed structures (Naïve Bayes, NB, and Tree Augmented Naïve, TAN) were learnt using the complete dataset (until September 2019) with the aim of predicting the dam level variable as accurately as possible. A scenario was carried out with data from October 2019 to March 2020 and compared the prediction made for the target variable with the real data. Results show both NB (rmse: 6.29) and TAN (rmse: 5.74) are able to predict the behaviour of the target variable.

Besides, a BN based on expert's structural learning was learnt with real data and both datasets with imputed values by NB and TAN. Results show models learnt with imputed data (NB: 3.33; TAN: 3.07) improve the error rate of model with respect to real data (4.26).

Guadairo River. Lack of information.

Dataset contains 73636 observations with 14 continuous variables. Since rainfall variables present a high percentage of zero values (over 94%), they were discretised by Equal Frequency method

with 4 intervals. The aim is to predict flooding risk in the coastal area but no data is collected from this area. Thus, an unsupervised classification based on hybrid BNs was performed. Here, target variable classifies all observations into a set of homogeneous groups and gives, for each observation, the probability of belonging to each group. Results show a total of 3 groups:

- Group 0, “Normal situation”: with rainfall values equal to 0, and mean of river level very low.
- Group 1, “Storm situation”: mean rainfall values are over 0.3 mm and all river level variables duplicate the mean with respect to group 0.
- Group 2, “Extreme situation”: Both rainfall and river level means values present the highest values far away from both previous groups.

Even when validation shows this methodology is able to identify extreme events, further work is needed. In this sense, data from autumn-winter season (from October 2020 to March 2021) will be used. Including this new information it would be possible to check if last extreme events (flooding event during December and Filomenastorm during January) are identified.