A LATENT-CLASS APPROACH TO MISSING VALUE IMPUTATION IN INCOMPLETE MULTIVARIATE WAVE METRIC DATASETS

Marco Picone [1]*, Francesco Lagona [1], Gabriele Nardone [2] and Mauro Bencivenga [2]
[1] GRASPA and DIPES - University Roma Tre, Rome, Italy - marcopic83@libero.it
[2] ISPRA - Marine Service, Rome, Italy

Abstract
We propose a latent-class model where the joint distribution of linear and circular data is specified through a finite mixture of conditionally independent Gamma and von Mises distributions. Missing values are imputed by drawing multiple imputations from the predictive distribution of the missing values given the observed data, at the maximum likelihood estimates. The procedure is illustrated on an incomplete dataset that includes measurements of wind speed and direction and significant wave height and direction, taken by a buoy and two tide gauges of the Italian wave-metric network.
Keywords: Waves, Wind/Font, Sicilian Channel

Environmental multivariate data are disseminated by environmental protection agencies for a variety of purposes, including the estimation of statistical models that detect significant relationships between different environmental measurements. Incomplete datasets, where some of the measurements are missing, pose a serious obstacle in the fulfillment of these purposes. There is an extensive literature about the estimation of statistical models in the presence of missing values. However, these methods often require the expertise of a trained statistician, as they involve both computational and methodological issues that can be challenging, depending on the nature of the mechanism that generate the missing values and the complexity of the model that is exploited for analysis. To reduce the workload of the data analyst, incomplete data should be provided in a way that they can be analyzed by "standard" methods, i.e. methods that require the availability of complete data information. Environmental data could be completed by imputing missing values according to an imputation model. This approach is referred to as single imputation. It is however well known that if the data analyst uses complete-data methods for analyzing the completed dataset by treating imputed values as if they were real data, this generally leads to variance estimates that are too low, confidence intervals which are too narrow, and wrong tests (real significance level above nominal level).

Multiple imputation (MI [1]) has been suggested as a way of overcoming the variance estimation problem that arises under a single-imputation strategy. Under a MI protocol, the data-base constructor (or imputer) and the end user (or data analyst) are thought as distinct entities. The data-base constructor draws a number of imputed values from the predictive distribution of the missing values, given the observed data, computed on the basis of an imputation model. The resulting completed datasets are appended together to provide an augmented dataset to the data analyst, who can exploit standard methods to simultaneously examine these datasets and, appropriately pooling the results, use them to correct for the variability in the imputations, which differs from the variability in the observed data. Directions about the pooling procedure are provided by the imputer and involve simple calculations, which can be carried out by a data-analyst who is not necessarily a trained statistician. Under a MI strategy, imputation is typically carried out by estimating a parametric model from the complete cases and using the predictive distribution of the missing data given the observed data to draw a number of imputations for each missing value in the incomplete dataset. Under a frequentist approach, the imputation model is estimated by maximum likelihood and imputations are drawn from the conditional distribution of the missing values, given the observed data, evaluated at the maximum likelihood estimate of the parameters that have been obtained from the observed data.

Latent-class, mixture models have been proposed as flexible imputation models when incomplete datasets include categorical [2] or mixed categorical and continuos variables [3]. We extend this approach to the case of linear-circular variables, by specifying a multivariate mixture of Gamma and von Mises distributions. The model clusters incomplete data into homogenous groups and exploits this classification to complete records with missing values. Parameters of the mixture model are estimated by maximization of the likelihood from the observed data, by a suitable E-M algorithm that allows for missing values. Imputations are then drawn from the multivariate conditional distribution of the missing values given the observed data, evaluated at the maximum likelihood estimate. We evaluate the performance of this imputation method by means of predictive intervals and nonparametric cross-validation.

Results of the MI procedure are illustrated on an incomplete dataset that includes hourly measurements of wave height and direction, taken in the period 10/13-11/11/2003 by the buoy of Mazzara del Vallo located at about 10 Km from the southern coast of Sicily. The dataset also includes eight-hours moving averages of wind speed and direction, taken from the two nearest tidal stations, respectively located at Porto Empedocle (Sicilian coast, about 100 Km from the buoy) and at Lampedusa Island (about 250 Km from the buoy).

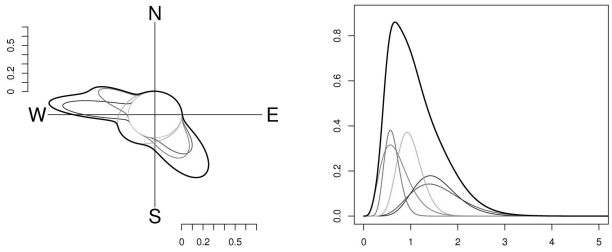

Fig. 1. The 5-components mixture distribution of wave direction (left) and height (right), as estimated by a latent-class multivariate model

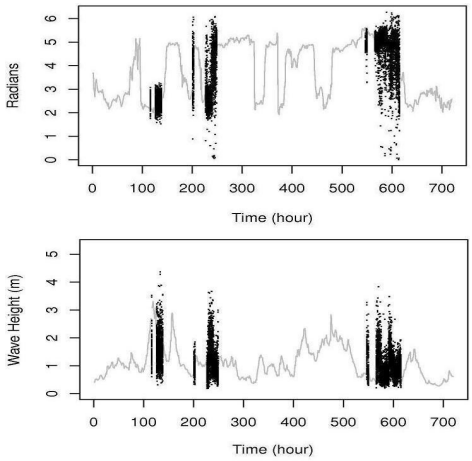

Fig. 2. Imputed values (dots) and observed data (grey line) of wave direction (top) and height (bottom), yielded by a 5-components latent-class model

References
1 - Rubin DB., 1987. Multiple Imputation for Nonresponse in Surveys. New York, John Wiley.
2 - Vermunt JK, Van Ginkel JR, Van der Ark LA, Sijtsma K., 2008. Multiple imputation of categorical data using latent class analysis, *Sociological Methodology*, 33: 369-297.
3 - Hunt L. and Jorgensen, M., 2003. Mixture model clustering for mixed data with missing information, *Comput. Stat. Data Anal.*: 41: 429-440.