Geoscientific Instrumentation Methods and Data Systems Discussions

Geosci. Instrum. Method. Data Syst. Discuss., referee comment RC2
https://doi.org/10.5194/gi-2021-11-RC2, 2021

Comment on gi-2021-11

Anonymous Referee #2

Referee comment on "Evaluation of multivariate time series clustering for imputation of air pollution data" by Wedad Alahamade et al., Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2021-11-RC2, 2021

Summary

The paper focuses on the technical problem of imputing (or spatially interpolating?) missing air pollutant concentration data in a multivariate setting and, more precisely, on the solution of this technical problem by applying methods involving multivariate time series (MVTS) clustering. It builds upon the work by Alahamade et al. (2021, in review) by evaluating (some of) the methods originally proposed therein. For that, hourly real-world data for four main air pollutants (specifically, $PM_{2.5}$, PM_{10}, O_3 and NO_2), as well as several graphical and statistical tools, are utilized. The data have been recorded from year 2015 until year 2018 at 167 stations representing six different environmental types (specifically, rural, urban, suburban background, roadside and industrial), thereby allowing comparisons across these types, other than the comparisons allowed across the four examined air pollutants. The evaluation framework assumes that each air pollutant is missing entirely and imputes it, separately for each station. Moreover, it involves the definition of a training period (i.e., from 2015 to 2017) and a testing period (i.e., 2018), the application of the MVTS clustering and time series imputation methods (resulting to six compared models in total, with three of them using the clustering outcomes, two of them using geographical distances, and an ensemble one using the five previous methods), the computation of prediction evaluation metrics (i.e., the factor of two (FAC2), Mean Bias (MB), Normalised Mean Bias (NMB), Root Mean Squared Error (RMSE), Coefficient of correlation (R) and Index of Agreement (IOA)), the design of Taylor's diagrams, and the conditional quantile analysis. Further, the Daily Air Quality Index (DAQI) is computed for the non-missing values of each original time series and for each imputed time series (corresponding to an original time series), and the agreement between the "observed DAQI" values and the "imputed DAQI" values is investigated. The DAQI is widely used to assess and monitor air pollution levels in the United Kingdom, and is computed based on the available data for five major air pollutants (specifically, O_3, NO_2, PM_{10}, $PM_{2.5}$ and SO_2); if data from one to four air pollutants are not available, the index is computed based on data for the remaining air pollutant(s). It is concluded that the ensemble imputation method (which uses the clustering outcomes produced by the MVTS clustering method) performs well.

General comments

Overall, I believe that the paper is meaningful, interesting and very well-written. Nonetheless, some clarifications and, perhaps, some extra work are also required at the moment.

More precisely, two major comments are provided in this report (see below) that should be carefully addressed, to my view, so that possible terminology-related confusion or misunderstandings are avoided, and further because the paper aims, among others, at showing how different graphical and statistical model evaluation functions enable the selection of the imputation (or spatial interpolation?) model that produces the most plausible imputations (or spatial interpolations?) (see lines 10−12). Because of these two major comments, I recommend major revisions.

A few minor comments are also provided in this report.

Specific major comments

1) According to Van Buuren (2018, Chapter 2.6), "imputation is not prediction" and "RMSE is not informative for evaluating imputation methods". In fact, innovations are set to zero in mean-value or median-value (i.e., non-probabilistic) prediction, while imputation creates a random noise to reflect the uncertainty of the missing values. In this view, Section 5.1.1 and Table 1 provide information that, in the best case, does not mean much by itself, and could even be misleading, unless relevant discussions are provided in the paper. Perhaps, however, the solved problem is not a time series imputation problem (at least, not in the sense explained in Van Buuren 2018, Chapter 2.6), but a spatial interpolation problem or a mean (or median) imputation problem. Related clarifications and extensive discussions should be provided, to my view.

2) Section 5.1.1 seem to be examining the imputations from a different perspective with respect to Section 5.1.3 (and perhaps also with respect to Section 5.1.2), where the conditional quantile analysis is presented. I wonder if the investigations of these two Sections are equally important for assessing the provided modelling solutions. It seems that they can support the assessment of models for different applications.

Specific minor comments

1) The case study conducted by Alahamade et al. (2021, in review) could be briefly described in the manuscript (in terms of its utilized data, evaluation procedures, and more), as this companion work is not available as a preprint. To my understanding (based

on lines 74 and 75), this specific case study has focused on univariate time series, while the present work focuses on multivariate time series, is this correct?

2) Further, in the manuscript the reader is referred to Alahamade et al. (2021, in review) for the full description of the assessed methods. Perhaps, an adapted reproduction of this full description could be added in the supplement (or in an appendix). To my view, this would make the paper complete.

3) Examples of imputed versus observed time series (with missing values) could be presented in the manuscript.

4) Figures could become more reader friendly. More precisely, all the text labels, axis labels and legends in the Figures could become larger, as currently it is quite hard for someone to read them. Also, the main figure titles could be removed, as the information reported there can also be found in the figure captions.

5) All the software packages used for this work should be cited in the manuscript.

6) Lastly, some few typos exist throughout the paper, and could be eliminated during revisions.

References

Van Buuren S (2018) Flexible imputation of missing data, second edition. Chapman and Hall/CRC, Boca Raton. doi:10.1201/9780429492259. Freely available online at: https://stefvanbuuren.name/fimd