Evaluation of Multi-variate Time Series Clustering for Imputation of Air Pollution Data

Wedad Alahamade[1,3], Iain Lake[2], Claire E. Reeves[2], and Beatriz De La Iglesia[1]

1 School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

2 School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

3 School of Computing Sciences, Taibah University, Medina 42353, Saudi Arabia

- **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.

Thank you for that very positive evaluation. We appreciate all the hard work in reviewing the paper.

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.

We thank the reviewer for very helpful comments that have made us think about our techniques and how the sit in the context of imputation/prediction/spatial interpolation. After some careful thought we can offer the following arguments which we hope will address some of the reviewer's comments.

According to Robinson et al. (2011, Chapter 4), imputation and interpolation process aim to fill the missing data in order to general a completed dataset. The **imputation** process is defined by Little and Rubin (2002) as any process that replaces missing values with other predicted or observed values. While **interpolation** is a type of estimation that aims to fill the gap or missing values between two known points.

Imputation is more general than interpolation, so we called our proposed approach time series imputation because we used the observed time series to impute missing time series (whole TS) in stations where one pollutant is not measured but other pollutants are. In this process, we are not filling the missing values within the time series but imputing a new TS. Also, it is not prediction since we are not predicting new values like we may do using a regression technique.

Our approach is based on multivariate time series clustering, where we group stations based on their fused temporal similarity of the measured pollutants without considering any spatial information about the stations. Based on the derived results by the clustering process we aggregated the spatial similarity of the stations to the clustering results to develop a model that is able to impute a plausible concentration for unmeasured pollutants.

So our imputation method uses observed data from multiple methods. However, it could be argued that it is close to the spatial interpolation process even though it is not completely

based on spatial information, that is, we did not use any geographical information with the proposed MVTS time series clustering. Adding to that, the main goal of the spatial interpolation is to fill in the gaps (points/locations with unknown measured) using points with known values to cover a certain geographical area. Our goal is to impute unmeasured pollutants (whole TS) in several stations where they are not measured using the fused similarity between stations of other pollutants.

We would also argue that our proposed method reflects the uncertainty because we are not using a prediction of a single new value, but we are using the observed values from other stations (either exact values, or approximate values using the mean) based on the MVTS clustering results and the geographical similarity between stations.

The RMSE is used to evaluated how close the modelled to the observed TS as an initial evaluation measure for time series deviation and it is not used on its own. It is used as an initial step to measure how close the imputed/modelled data is to the real time series in the training set to select the best imputation model that gives the lowest average of errors with the support of other statistical measures that agree with the RMSE such as Coefficient of correlation, index of Agreement, and Mean Bias.

Importantly, in the evaluation process, we don't rely on the RMSE alone to select the best imputation method, but we used several graphical and statistical evaluation functions that represent the uncertainty between modelled and observed TS such as:

- Taylor's diagram analysis is used to evaluate the Correlation Coefficient and the standard deviation to represent the variability between modelled and observed concentrations.

- Conditional quantile plots are used to analyse the spread, distribution, and the uncertainty between the modelled and observed pollutant concentrations. In fact, when we look at the conditional quantile plots we can see that there is a variability in the confidence intervals between modelled and observed data.

The analysis results obtained by the graphical evaluation functions support the results obtained by the RMSE to select the best imputation model.

A new paragraph in page 2 (lines 44-55) discuss these differences. We also mention the uncertainty of our models in a new line in page 7, line 189-190
Note: all changes in the manuscript are highlighted in red to ease checking.

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 quantil 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.

The goal of these sections is to compare and evaluate the performance of the proposed models based on the air quality modelling techniques and using different techniques. So section 5.1.1 presents statistical analysis which the reviewer has in part questioned, but the other forms of evaluation in 5.1.1 and 5.1.3 complement this and offer different perspectives, which we believe is a strength of the paper.

Openair is a tool for air quality data analysis for comparing models against measurements and models against other models.

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?

We thank the reviewer for the advice. Our previous work is now accepted for publication and to appear imminently in the Neurocomputing Journal (Alahamade et al. (2021, to appear)) so we will soon be able to include a new reference and that saves us from having to reproduce unnecessary detail in this paper.

To clarify what is on each paper, on the previous paper (Alahamade et al. (2021, to appear)), we proposed an intermediate fusion approach to cluster stations based on aggregated similarity of the four air pollutants (O3, NO2, PM10, and PM2.5) using the k-means clustering algorithm. We called that multivariate time series (MVTS) clustering. The clustering result is then used with the proposed imputation models to impute missing pollutants (whole TS) in stations.

In the previous paper, the proposed approach is evaluated and compared with the univariate TS clustering where each pollutant is used to derive its own clusters and then imputation is based on that clustering solution, so independent for each pollutant. These two approaches are compared in terms of the quality of the clustering results using the clustering validity indices (CVIs) and the imputation quality using the RMSE and its standard deviation.

For work in the previous paper and this paper, we used the same dataset, that included hourly pollutants concentrations of the four air pollutants. In the current paper, we extend the work by applying the imputation solution using the results of the MVTS clustering to real data and using extensive evaluation methods to demonstrate its effectiveness. This enables us to extend our understanding of pollutant behaviour.

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.

The manuscript in question is now accepted for publication and to appear imminently so we can include a reference which will save putting unnecessary content which can be problematic for the journal (as the previous journal may have some copyright over the material).

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

Thank you for advise which has been followed.

We added an examples for each pollutant, with some discussion, to represent the imputed and the observed TS for small period of time that contains missing observations within the TS. This can now be found in pages number 15-17 (figures in pages 22-23).

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.

Thank you for advise which has been followed.

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

Thank you for advise which has been followed. This can now be found in page number 5.

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

We have done a further round of proofreading to try to catch any typos.

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

Robinson A., Hamann J. (2011) Imputation and Interpolation. In: Forest Analytics with R. Use R. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-7762-5_4

Little, R. J., & Rubin, D. B. (2002). Single imputation methods. Statistical analysis with missing data, 59-74.

Alahamade, W., Lake, I., Reeves, C. E., and De La Iglesia, B.: A Multi-variate Time Series clustering approach based on Intermediate Fusion: A case study in air pollution data imputation, Neurocomputing, accepted, 2021