Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2017-1126-RC1, 2018

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Associativity Analysis of SO_2 and NO_2 for Alberta Monitoring Data Using KZ Filtering and Hierarchical Clustering" *by* Joana Soares et al.

Anonymous Referee #1

Received and published: 6 February 2018

Review of acp-2017-1126. Associativity Analysis of SO2 and NO2 for Alberta Monitoring Data Using KZ Filtering and Hierarchical Clustering

The papers is well written, scientifically sound, and clearly reflects the large amount of work behind the analysis as well as the authors' knowledge of the relevant scientific literature. The presentation is clear and overall reads really well. In general, the application of data mining techniques to air quality monitoring and modeling is still at an 'embryonic stage' and deserves encouragement.

I advice the editor to accept the manuscript for publication. I only have some minor suggestions on my notes for the authors to consider including in the final version.

General and specific comments:

1) Given the scientific significance and the potentiality of this work, I believe it deserves more visibility. I think the authors are underselling their work. For instance, the title seems to suggest a study with highly technical details which can discourage non-expert readers, whilst could be more general to attract more audience. Consider avoiding the use of KZ in the title, it is just a moving average filter.

2) It's not clear to me the average behind figure 9. It shows the correlation map of each grid cell with any other cell? Does it imply an average over R? or it is the time or spatial series correlation being investigated? Please clarify in the text

3) it is not clear how redundancy is defined: Overlapping variance, coefficient of determination above a certain threshold, . . .;

4) based on this study, can the authors comment on the minimum exposure period (length of time series) for the clustering analysis to be reliable

5) page 4 ,line 17. Consider Vardoulakis et al. 2011. Atmospheric Environment 45 (2011) 5069-5078

6) page 8, line3. Not only 'dissimilarity metrics' but also agglomeration method and definition of correlation coefficient are quite sensitive parameters

7) Can the Euclidean distance be used to spot systematic detection error?

8) page 14 line 16. Can the authors comment on the spatial continuity of the solution? Is it a requirement or the area can contain holes and/or be even detached?

9) Page 15, line37. Solazzo and Galmarini misspelled.

10) Page 15, line37. A source of dissimilarity was found to be the reporting time not harmonised across European countries. Data reporting at the beginning or at the end of the hour can make a significant difference

I invite the authors to comment on the following:

I think we are still far away from using clustering for operational use. Clustering is known to provide some qualitative insight, but it is quantitatively weak as it depends on many parameters. Indeed, a fundamental challenge of clustering is the high sensitivity to the options controlling the underlying algorithms, such as the agglomerative method, the distance metric, the number of clusters, and the cut-off distance are aspects that need to be determined case by case. In particular, the cut-off (the threshold similarity above which clusters are to be considered disjointed) determines the dimension of the sub-space of non-redundant information and is decided by visual inspection of the dendrogram. Supervised clustering (e.g. k-Means) initiated with the results of unsupervised clustering might be more robust.

The application of associativity analysis for detecting potential redundancy in the context of regulatory air quality monitoring might have some pitfalls (most of which are anyway mentioned by the authors in the text, but I think deserve more words). For example, the potential duplicate of information obeys some policy precautionary principle and might reveal useful in some instances (double checking, reduce missing records, cross validation, etc). Further, redundancy should be determined with some long-term climatology and should also serve future decision making in the sense that what might be redundant based on the past ten years of data might not be in the next ten years. In this sense the adoption of models for future scenarios might help.

I think that, more than the estimation of redundancy, the main strengths of the methodology are the potential for classification and the estimation of the area of representativeness (AoR). Indeed I would have framed the whole work in the context of classification. For example, can the methodology assist in the classification of monitoring station based on area-type or site-type? Do the authors expect the diurnal signal to be the associated over long distance? In siting a new station, its area-type can be defined by looking at how the signal of existing stations compares with the signal of the new station? I would invite the authors to add some further considerations about the

potentiality of the methodology devised, also in light that some reflections are already part of the paper, for example the clustering of long term signals.

Concerning the AoR, the authors (or at least some of them) have already experience with the topic, and I have been surprised that it was not expanded in the text, especially since model results are available. The maps in figure 9 indeed show some AoR! The authors mentioned it at the beginning of page 3 but then drop it. For example, some discussion about AoR would fit nicely in section 4.1. Again, in light of better exploiting the large amount of work done, I would invite the authors to consider adding some further words about the potentiality of the analysis for determining the AoR.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2017-1126, 2018.