Reviewer's report

Title: Using Self-Organizing Maps To Classify Days By Air Quality For Air Pollution Epidemiological Mixtures Research

Version: 1 Date: 11 April 2014

Reviewer: John Molitor

Reviewer's report:

Summary

The authors do a good job of introducing the sub-field of multi-pollutant modeling, and discuss the current reasons for using the multi-pollutant approach and the challenges associated with such inference. A general discussion of supervised versus unsupervised clustering and its application to multi-pollutant modeling ensues.

The authors then introduce the self-organizing map (SOM) and give some history behind the approach. A dataset is then introduced, which contains time-series measurements of daily concentrations for 12 air pollutants for years 2000 to 2007. Some mathematics regarding the SOM algorithm is then provided, followed by a discussion of tuning parameters and the issues related to selection of the appropriate number of classes for the SOM algorithm. Regarding this last point, the preference for using the CH index to choose cluster number is discussed, along with other statistical measures used to analyze SOM output.

The results were then given. The CH statistics chose a difficult to interpret 2-class result, so a 4x4 map was then chosen, revealing two primary types of classes related to primary and secondary pollutants. The data were further analyzed to characterize the kinds of pollutants associated with seasonal trends.

Major Compulsory Revisions

The question as to how to analyze multi-pollutant profiles is an ongoing one, more work is needed, and this approach presents a useful addition to this literature. Nevertheless, I'm disappointed that the authors have not provided an analysis that examines associations between air pollution profiles and health outcomes. I realize that the authors are perhaps setting the stage for further development in this area, nevertheless, it's difficult to assess the quality of this approach without applying it in a manner which addresses what I think is the ultimate goal, associating multi-pollutant profiles with health.

Further, the approach presented does not go much beyond simply using a series of R packages (such as the SOM package) and reporting results. There is little in the way of development of a truly new methodology. The paper does provide a demonstration of the use of these packages, though the dataset presented is lacking in terms providing the kind of substantive insights that would be
informative to, say, and epidemiologist of other health researcher. I do, however, see some advantage in demonstrating the use of a package which, as far as I know, has not been applied to this type of problem.

If I understand things correctly, the SOM procedure creates nodes associated with a multi-pollutant profile. Could these nodes be mapped? One could define all spatial points closest to a center as a region, color in each region with a unique color and provide a legend associating these colors (regions) with a prototypical multi-pollutant profile.

Some kind of geographical map presenting locations of typical multi-pollutant profiles would greatly improve the paper. Unfortunately, many graphs are provided, presumably provided by the R packages, which are unlikely to be informative to applied health or exposure researchers.

In short, this paper is useful, but some additional work is needed to make the output interpretable to applied health and air pollution exposure reserachers.

Discretionary Revisions:

What is the similarity between SOM and Voronoi tessellations?

The approach used in (12) does find an “optimal” clustering obtained for interpretation purposes. However, the main point of this approach is that all inference, including uncertainty estimation related to the optimal clustering, is done via Bayesian model averaging which averages over cluster uncertainty. This kind of analysis is much more in the spirit of Bayesian Dirichlet Process estimation rather than cluster analysis which is usually associated with “hard” versus “fuzzy” clustering.

Regarding the statement, “While unsupervised classifications are considered objective”. Considered by whom? As you point out, they involve many subjective assumptions.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.