Reviewer's report

Title: Spatio-temporal Modeling of Particulate Air Pollution in the Conterminous United States using Geographic and Meteorological Predictors

Version: 2 Date: 12 June 2014

Reviewer: Lianfa Li

Reviewer's report:

This paper proposed generalized additive mixed models to estimate spatially- and temporally-resolved exposure of atmospheric particulate matter. Accurate spatiotemporal estimates of ambient air pollutant concentrations are important since such estimates can potentially improve estimation of health effects of air pollutants. The models of this paper described multiple-levels (regional and fine-scale) spatial and temporal gradients in monthly mass concentrations of fine (PM2.5), inhalable (PM10) and coarse mode particle mass (PM2.5-10) for the conterminous US. The PM2.5 models had an accuracy of 0.77 (CV R2), and the PM10 and PM2.5-10 had a varying accuracy from 0.46-0.58. Regional variations were found in the effects of geographic and meteorological covariates as well as predictive performance of the models. The proposed modeling methods are interesting and useful for estimation of chronic exposure of air pollutants in larger geographic regions like the conterminous US. This paper is acceptable for publications after the authors address the following concerns.

Major Compulsory Revisions

1. The air pollution data in this study came from different sources such as AQS, IMPROVE, CASTNet and SFU etc. (Line 133-142). Did the authors examine whether there was systematic bias between different sources? Although AQS contributed the bulk of the monthly means and sites, such bias could have potential negative influence on the results. How would the authors avoid or decrease it if the bias exists?

2. This paper used the kernel density method to derive several significant spatial covariates such as traffic density (Line 155-157) and density of point-source emission (Line 179-181). For land-use covariates, the proportions of residential and urban land-use were calculated for each location using neighborhoods of 1km and 4 km. To inform the readers, can the authors have a brief introduction to kernel density and its advantages or drawbacks in comparison with other methods such as the buffer-based statistical method? How did this study determine the bandwidth of kernel density and how it influenced on prediction? Why was kernel density not used to extract the land-use covariates?

3. In equation 3, gt,r(si) accounts for residual monthly spatial variability within the region, specified as spatial bivariate thin-plate penalized spline terms with basis dimension. Did the authors try or test other spatial modeling methods such as
varoiogram used in Baysian spatial models? In comparison with these methods, what advantages and drawback (such as over-fitting) does use of spatial location as a bivariate in the model have?

4. For generation of spatial covariates, different buffer distances were given. For example, for residential and urban land-use ((Line 167-170)), just buffer distances (1 and 4 km) were used to derive the traffic density. How did the authors select the buffering distance? Have they tested other buffer distances? Again in derivation of traffic density (Line 155-157), the neighborhood of 100 m was used. What about the other buffer distances?

5. This paper used different cell sizes to extract the grid surfaces such as 30 m cell size for traffic density and 500 m cell size for population density (tract-level). Could such differences in the cell sizes of covariates lead to inconsistency in assigning the corresponding covariates to each subject location?

6. In modeling, the authors first listed the equation 3 and then adopted a two-stage modeling approach to fit equation 3. Are the output of equation 3 and that of the two-stage approach consistent? The authors may use a small sample to test this and compare whether the outcomes from the two approaches are consistent.

7. In selection of the predictive covariates (Line 361-364), did the authors measure multiple collinearity between covariates such as tract-level and county-level population density, elevation and spatial locations (x/y)?

8. For point-source emission, the densities of just 3 km, 7.5 km and 15 km were measured. What about the other distance such as 2km, 5km? How did the authors select the distance?

Minor Essential Revisions

9. In Line 103, it will be better for the reader if the authors have a brief introduction to what the difference between accuracy and precision is and in which aspect they measure the model’s performance?

10. Elevation data (Line 182-183) from the USGS were based on grid or vector? If it was based on grid, what is its resolution?

11. In Line 184-188, the authors used a transformation formula to make the covariates more uniform distribution. It is better that the authors can give the corresponding references or evidence to support this?

Discretionary Revisions

12. In Line 472-473, what are the interactions? It seemed that the effects of two GIS-based time invariant spatial covariates (urban land use within 1 km and elevation) just varied by regions.

13. Tract-level population density was negatively associated with measured PM10 levels. What about its variation by regions?
14. This paper assumes stationary and isotropic spatial variation (Line 543-544). Given the large regions across conterminous US, this assumption may not be true and caution must be taken when modeling under such an assumption.

**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.

**Declaration of competing interests:**

I declare that I have no competing interests.