Reviewer's report

Title: Construction of an environmental quality index for public health research

Version: 1 Date: 28 January 2014

Reviewer: Howard Chang

Reviewer's report:

This manuscript describes the development of a county-level measure that reflects environmental quality. The study has several strengths including the use of an impressive collection of publicly available datasets, inclusion of sociodemographic variables, and its application on a national scale. While considerable details are provided on data acquisition and variable selection, there are a few methodological details missing to ensure reproducibility. Specific comments are given below and these should be considered prior to potential publication.

1. I’m not quite sure what does “ambient” environment mean in the Introduction section. Does this exclude residential or indoor exposure?

2. The authors stated that their choice of variables, while extensive, was based on a literature review of environmental factors associated with infant mortality. A discussion on how the index developed here may or may not be useful for other health outcomes (e.g. cancer, morbidity) will be helpful.

3. The authors should provide more details on how kriging was applied to counties with missing data. First, what spatial dependence structure was used? Was this choice really valid for the entire continental US? Second, how did the authors account for counties having very different geographical sizes? Third, what is the extent of missing data? Are missing data preferentially present in rural areas?

4. It’s not clear how normality of the data was assessed. Surprisingly, it appears that a log-transformation was applied universally to any variable that was not normal. I am particularly concerned about the variables for the sociodemographic domain which are all percentages. Wouldn’t a logit transformation be more suitable? Also it’s not clear if the normality assumption was checked for the composite variables.

5. How much variation does the first PC of each domain explain the variation in the variables? Does this justify using only the first PC? How does this impact interpretation and the usefulness of the domain-specific and the overall index?

6. Some explanations may be needed for why the trends in variables across urbanicity (Table 1’s) are not associated with similar trends in the loadings (Table 2’s).
7. One major limitation of the work is the coarse spatial and temporal scale, which the authors recognize in the Discussion. It will be helpful if the authors can specifically point to which variables and domains might be more susceptible to the effect of aggregation at the county-level (e.g. SES variables that exhibit strong spatial heterogeneity even at the Census tract level), as well as how would these effects impact the overall EQI development and its application.

8. The numerical values with E in Table 1’s may be revised. For example, perhaps report median household value in $10,000 and NO2 in ppm. I'm also not sure if the factor loadings require 4 decimal places. Fewer decimal points might also help convey the differences and seminaries in loadings across urbanicity levels.

**Level of interest:** An article of outstanding merit and interest 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.