Peer Review Report for the Technical Basis for the EPA's Development of Significant Impact Thresholds for $PM_{2.5}$ and Ozone

EPA-454/S-18-001
March 2018

Peer Review Report for the Technical Basis for the EPA's Development of Significant Impact Thresholds for $PM_{2.5}$ and Ozone

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Air Quality Assessment Division
Research Triangle Park, NC

Overview

As part of the OMB review process for the draft guidance document for $PM_{2.5}$ and ozone SILs[1], the EPA agreed to conduct a peer review of the technical basis document (TBD).[2] This summary of the peer review provides the charge questions supplied to the peer reviewers, a summary of the comments received from the reviewers, and overviews of changes made to the TBD and additional analyses conducted in response to reviewer comments.

Peer review process

The peer review was conducted under an EPA contract to the University of North Carolina at Chapel Hill that has been used for technical review purposes similar to this work in the past. The peer review was overseen by Dr. Sarav Arunachalam, Research Associate Professor with the Center for Environmental Modeling for Policy Development. The EPA provided a list of six potential reviewers, from which the contractor obtained agreements from three reviewers to conduct the peer review. The peer reviews were conducted by environmental statisticians on faculty at major U.S. universities. The three reviewers were (bios for each reviewer are provided in Appendix A to this document):

- **Candace Berrett, PhD**; Assistant Professor, Department of Statistics, Brigham Young University
- **Veronica Berrocal, PhD**; John G Searle Assistant Professor of Biostatistics, University of Michigan School of Public Health
- **Bo Li, PhD**; Associate Professor, Department of Statistics, University of Illinois at Urbana-Champaign

Charge questions

The charge questions were developed by EPA in consultation with OMB. The final set of charge questions sent to the reviewers were as follows:

1) Are the relevant technical aspects of the statistical procedure clearly described?
    a. Are input data (EPA's AQS) and their characteristics sufficiently described?
    b. Is it clear what is being estimated?
    c. Is the bootstrap procedure described in sufficient detail to allow reproduction?
2) Are the descriptions of statistical concepts clear and accurate?

[1] Guidance for Comment: Significant Impact Levels for Ozone and Fine Particle in the Prevention of Significant Deterioration Permitting Program, https://www.epa.gov/nsr/draft-guidance-comment-significant-impact-levels-ozone-and-fine-particle-prevention-significant

[2] Technical Basis for the EPA's Development of Significant Impact Levels for $PM_{2.5}$ and Ozone, Office of Air Quality Planning and Standards, RTP, NC, 2017, EPA 454/R-17-002.

    a. Are the descriptions of statistical significance and significance testing clearly and sufficiently described to assist the layperson in understanding the analysis?
    b. Do the examples provided in the TBD illustrate the concepts of statistics sufficiently for the layperson to understand the analysis?
3) Are the assumptions and choices in the analysis clearly described and supported?
    a. Are the assumptions and choices in the analysis sufficiently documented?
    b. Does the document sufficiently describe the sensitivity of results to the choices and assumptions in the analysis? For example, are the technical considerations that support the policy decision to aggregate the variability to a single national value clearly articulated?
4) Are the procedures appropriate for the analytical goals?
    a. Is bootstrapping an appropriate technique to quantify the variability in the air quality design value statistics? Is the bootstrapping analysis a reasonable approach to inform a policy determination of Significant Impact Levels (*i.e.*, threshold levels)?
5) In your assessment, is there need for further analysis or clarification? Do you have suggestions for improving the document?

The peer review occurred parallel to the public comment period, from August 1 through September 30, 2016. The peer reviewers were given approximately 30 days to review the package, which included all three SILs documents (*i.e.*, in addition to the TBD, the policy memo[3] and legal memo[4] were provided to the reviewers). Each individual peer reviewer provided their comments to the UNC contractor, who then anonymized and delivered the reviews to the EPA as PDF documents, similar to how peer review comments would be handled by a scientific journal. The peer review responses are provided in their entirety in Appendix B of this report.

Summary of reviewer responses

The reviewer comments were largely supportive of the TBD and the analysis presented therein.

Reviewer 1

Reviewer 1 offered a few editorial comments but was very supportive of the methods, presentation, and conclusions from the analysis. Their response to charge question 3b was particularly expressive:

"The bootstrap is applied appropriately, and the selection of 50% conference interval to obtain conservative SILs is reasonable. The selection of a single national value is not optimal considering the spatial variability, but taking the consistency of policy into consideration and given the fact that there are no large scale trends in ambient air variability are present, it is not

[4] Legal Support Memorandum, Application of Significant Impact Levels in the Air Quality Demonstration for Prevention of Significant Deterioration Permitting under the Clean Air Act.

unreasonable to have a single national value. Using the median rather the mean provides a more robust SIL for NAAQS."

Reviewer 2

Reviewer 2 offered a few editorial comments but also had several comments related to spatial variability. In particular, in contrast to Reviewer 1, Reviewer 2 felt that the spatial variability and dependence was not sufficiently accounted for.

Reviewer 3

Reviewer 3 offered a few editorial comments but also had several comments related to clarity and specificity, particularly with respect to the statistical terminology. The reviewer also had one technical comment related to the considerations of temporal dependence on the sampling and how this was accounted for in the bootstrap technique.

Summary of responses to peer review and public comments

The EPA made a number of revisions to the TBD, including (1) updating the analysis to include more recent data, (2) editing a number of sections for clarity and accuracy, and (3) conducting new and updated analyses to investigate issues raised by the reviewers.

Updated analysis

The bootstrapping methods for $PM_{2.5}$ data processing for the calculation of the $PM_{2.5}$ design values were also adjusted slightly to better align the methods with standard practice for calculating design values. Specifically:

- The rounding conventions for calculating $PM_{2.5}$ design values were applied in accordance with the EPA's regulations.[5] The original document applied the appropriate truncation conventions for ozone (*i.e.*, truncate to the whole ppb)[6]; however, the rounding rules for $PM_{2.5}$ were not correctly applied (*i.e.*, round design values to the nearest whole $\mu g/m^3$ for the daily NAAQS and the nearest 10^{th} of a $\mu g/m^3$ for the annual NAAQS).

[5] Appendix N to Part 50—Interpretation of the National Ambient Air Quality Standards for $PM_{2.5}$.
[6] Appendix U to Part 50 - Interpretation of the Primary and Secondary National Ambient Air Quality Standards for Ozone.

- The selection of the 98th percentile value for the daily $PM_{2.5}$ value was corrected to use Table 1 provided in the CFR[5] rather than calculating the 98th percentile value based on the number of samples.

These updates had no impact on the recommended annual $PM_{2.5}$ SIL values (0.2 μg/m^3), while the daily $PM_{2.5}$ SIL value increased slightly, from 1.3 μg/m^3 to 1.5 μg/m3, primarily due to the updated 98th percentile selection approach, rather than the application of the rounding concentrations.

Editorial comments

Updates to the TBD were made to address editorial comments from all three peer-reviewers as well as in response to comments received during the public comment period. The majority of these were minor edits so they are not highlighted here but reflected in the revised TBD. However, significant editorial changes were made in section 4.1 in response to both Reviewer 3 and public comments. Section 4.1 was heavily revised with much of the discussion moved to the policy document in order to clarify the difference in the technical analysis and the policy choices made from the available options derived from the technical data. Specifically, we updated section III of the policy document to more clearly describe what information informed the selection of the EPA recommended SIL values and what the policy decision was based upon.

In response to Reviewer 3, the EPA conducted an additional analysis that examined the impact of persistence in ambient concentrations (*i.e.*, concentrations on one day being similar to concentrations on the following or previous day, which could occur due to similar meteorological conditions). The analysis focused on ozone because the EPA believes this pollutant would most likely show the impact of temporal persistence. While the EPA had conducted a form of this analysis during the development of the SILs package, the new analysis more rigorously analyzed the impact of the persistence of pollution events by analyzing the temporal correlation between ambient data at individual sites using standard statistical techniques and aggregated this correlation across the country. In simple terms, the analysis calculates correlation coefficients (using linear regressions) between data from day 1 with data from day 2, between data from day 1 and data from day 3, etc. The correlation between these pairs of days can inform the degree to which concentrations on a particular day can be predicted by concentrations from the previous days and how long pollution events might typically occur. The lag found from the correlation analysis (*i.e.*, 7 days) was used to conduct a block-sampling of the data and a re-run of the bootstrap analysis. The block sampling modified the bootstrap analysis to include the 3 days before and after each randomly selected day, such that blocks of consecutive days were analyzed. This procedure, thus, accounts for any temporal persistence that may be present in the air quality variability. The results at the 50% confidence interval were minimally different from the original, non-parametric analysis that assumed no lag. This additional analysis and the results are documented in section 6 of the appendix to the TBD.

Reviewer 3 also made specific comments on the spatial correlation among sites; in particular, that they did not agree with the EPA's assertion that there is not a significant spatial correlation among sites. Reviewers 1 and 2 also commented on the presence of a correlation between the

spatial variability. Reviewer 1 specifically commented that there were no large scale trends, which was also the EPA's conclusion. In general, the EPA believes that the disagreement by Reviewer 3 is a matter of phrasing in the original TBD. There is a spatial correlation in both ozone and $PM_{2.5}$ in that most areas show relatively small variability and that there is not a strong spatial correlation in the location of sites with high variability. The document was revised to emphasize our intent in describing the spatial correlation. However, we also conducted several analyses to explore the existence of spatial groupings, *i.e.*, to determine if there are natural grouping of monitors with similar levels of variability. Three separate analyses were conducted, as follows:

- A cluster analysis was done using the latitude, longitude, and variability at each site in order to allow spatial variables to form natural groupings with similar levels of air quality variability.

- The NOAA climate regions were used to segregate data into predefined spatial groups based on similar weather patterns. The air quality variability of each climate region was then compared on a region-to-region basis and with the data aggregated to the national level to determine if the subsets were quantitatively different from one another.

Each analysis was conducted separately based on the air quality variability from both the annual and 24-hr $PM_{2.5}$ standards for the 2014-2016 data. The first two analyses were conducted using a "K-means" clustering algorithm and a hierarchical clustering algorithm. The K-means algorithm uses a pre-determined number of clusters and initially randomly assigns all sites to clusters. The difference between the cluster centers and all individuals are calculated, then sites are reassigned to the most similar cluster. The algorithm repeats a set number of times or until a minimum convergence threshold is reached. Hierarchical algorithms do not use a predetermined a number of clusters, but instead start with each site as part of their own cluster. The first step in a hierarchical analysis combines the two most similar clusters (which are just the two most similar sites at the first step). Each subsequent step combines the next closest clusters, until only two clusters are left, which contain all the individual sites.

The results of these additional analyses, which attempted to identify naturual groupings of sites based on similar levels of variability (*e.g.*, sites with consistently high variability), are presented in section 7 of the appendix to the TSD. The three separate analysis described above were conducted with each averaging period, resulting in 14 different sets of clusters. The results across these 14 sets of clusters varied widely. Several analyses did identify a unique region based on a specific clustering technique and averaging period, but these results were not consistent across clustering techniques or averaging periods. For example, the latitude, longitude, and variability analysis (first option in the list above) indicated several unique regions based on the 24-hr standard using the K-means algorithm. However, the K-means algorithm did not identify unique regions for the annual standard. Similarly, for this dataset, the hierarchical analysis identified sites with unique levels of variability for the 24-hr standard, but these sites were not spatially grouped (*e.g.*, the most unique group spanned at least 5 states, ranging from North Carolina to

Maine). Many of the analyses did not identify any unique groupings at all. When the results are considered as a whole, they support the EPA's original position that there are no large scale trends and that a national SIL is reasonable.

Appendix A: Peer reviewer bios

Candace Berrett, PhD; Assistant Professor, Department of Statistics, Brigham Young University
cberrett@stat.byu.edu; 801-422-7055; http://statistics.byu.edu/directory/berrett-candace

- Publications Chair, Section on the Environment, American Statistical Association, 2015-2016
- Program Chair, Environmental Sciences Section, International Society of Bayesian Analysis, 2014-2015
- Berrett, C. and Calder, C. A. (2016) “Bayesian spatial binary classification," Accepted for publication in Spatial Statistics.
- Co-PI, 2014, “Spatial Uncertainty: Data, Modeling, and Communication," National Institutes of Health (NIH).

Veronica Berrocal, PhD; John G Searle Assistant Professor of Biostatistics, University of Michigan School of Public Health

berrocal@umich.edu; 734-763-5965; https://sph.umich.edu/faculty-profiles/berrocal-veronica.html

Relevant Selected Publications:

- Professor of Spatial Statistics and Modern Statistical Methods, University of Michigan
- Young Investigator Award, Section on the Environment, American Statistical Association, 2015
- Chair, Section on Statistics and the Environment, American Statistical Association, 2017
- Associate Editor, **Journal of Agricultural, Biological, and Environmental Statistics, 2015**
- V. J. Berrocal, A. E. Gelfand, and D. M. Holland. (2010). A spatio-temporal downscaler for output from numerical models. Journal of Agricultural, Biological and Environmental Statistics, 15, 176-197.
- V. J. Berrocal, A. E. Gelfand, and D. M. Holland. (2014). Assessing exceedance of ozone standards: A space-time downscaler for fourth highest ozone concentrations. Environmetrics 25(4) · May 2014
- V. J. Berrocal, A. E. Gelfand, D. M. Holland, J. Burke, M. L. Miranda. (2011). On the use of a PM2:5 simulator to explain birthweight. Environmetrics, 22, 553-571.

Bo Li, PhD; Associate Professor, Department of Statistics, University of Illinois at Urbana-Champaign

libo@illinois.edu; 217-333-2167; http://www.stat.illinois.edu/people/faculty/boli.shtml

Relevant Experience and Selected Publications:

- Professor of Spatial Statistics and Analysis of Variance
- Young Investigator Award, Section on the Environment, American Statistical Association, 2011
- Associate Editor, **Journal of Agricultural, Biological, and Environmental Statistics, 2013**
- Li, B., Zhang, X. and Smerdon, J., Comparison between spatio-temporal random processes and application to climate model data (2016), *Environmetrics*, to appear.
- Li, B. and Smerdon, J. E., Defining spatial assessment metrics for evaluation of paleoclimatic field reconstructions of the Common Era (2012) *Environmetrics*, Vol. 23, 394-406.

Appendix B: Peer reviewer comments

Comments from peer reviewer 1

1) Are the relevant technical aspects of the statistical procedure clearly described? -- Yes.
a. Are input data (EPA's AQS) and their characteristics sufficiently described? -- Yes, the data is described clearly.
b. Is it clear what is being estimated? -- Yes, the ozone, annual PM2.5 and 24-hr PM2.5 NAAQS on page 8 is very clear.
c. Is the bootstrap procedure described in sufficient detail to allow reproduction? -- Yes, this is clear.

2) Are the descriptions of statistical concepts clear and accurate? -- Yes.
a. Are the descriptions of statistical significance and significance testing clearly and sufficiently described to assist the layperson in understanding the analysis? -- Although I am not a layperson in statistics, I think the concept is well explained in plain language.
b. Do the examples provided in the TBD illustrate the concepts of statistic sufficiently for the layperson to understand the analysis? -- Yes, they are straightforward to follow.

3) Are the assumptions and choices in the analysis clearly described and supported? -- Yes
a. Are the assumptions and choices in the analysis sufficiently documented? -- Yes, all details are well documented.
b. Does the document sufficiently describe the sensitivity of results to the choices and assumptions in the analysis? For example, are the technical considerations that support the policy decision to aggregate the variability to a single national value clearly articulated? -- Yes, the report carefully studied the spatial variability and the temporal variability for PM2.5 at different sampling frequencies. The bootstrap is applied appropriately, and the selection of 50% conference interval to obtain conservative SILs is reasonable. The selection of a single national value is not optimal considering the spatial variability, but taking the consistency of policy into consideration and given the fact that there are no large scale trends in ambient air variability are present, it is not unreasonable to have a single national value. Using the median rather the mean provides a more robust SIL for NAAQS.

4) Are the procedures appropriate for the analytical goals? -- Yes
a. Is bootstrapping an appropriate technique to quantify the variability in the air quality design value statistics? Is the bootstrapping analysis a reasonable approach to inform a policy determination of Significant Impact Levels (i.e., threshold levels)? -- Yes, the bootstrap is a sound statistical approach. it is very popular due to its flexibility that no parametric model or strong assumptions are required. The bootstrap is applied appropriately to quantify the variability in design values.

5) In your assessment, is there need for further analysis or clarification? Do you have suggestions for improving the document?

I read the document twice. At the first time, I was a little confused with what NAAQS represents in many places. My understanding of NAAQS is that it is a set of standards

or thresholds for different statistics (or called DV here), but then it seems NAAQS is used more often as the statistics defined for NAAQS. For example, the x-axis labels in Figures 11 and 13 used NAAQS as the statistics. Although I finally realized what NAAQS often represents in the document, it might be more clear to explicitly point out that it is the statistics defined for NAAQS rather than the thresholds are actually referred to.

Page 5, The definition of "design value" is also confusing. The definition says it is "a statistic or summary metric based on the most recent one or three years ...". This seems to imply that the design value (DV) is a statistic or summary that is computed based on the sample of monitored data only for new source or modification. It seems to imply that the purpose of computing DV is to evaluate the contribution of source(s). However, later the DV is calculated based on all data measured during 2000-2014 and the results are used to derive SIL which if I understand correctly would serve the thresholds for NAAQS. I would suggest to remove "the most recent" in the definition on page 5 so it reads like "a statistic or summary metric based on one or three years ...".

Page 34 last paragraph, "using only the 1:1 monitors would produce smaller estimates of the variability". This is hard to understand intuitively. Suppose we have continuous observations in time, i.e., a continuous time series. Now we take daily values from this series for 1:1 monitors and also take values every three days for 1:3 monitors, then I expect that the daily values would exhibit no less if not more variability than the values every three days. Is there a better explanation from the characteristics of data collection for the smaller variability with 1:1 monitors? For example, since the 1:3 monitors collect data at different times during the day than the 1:1 monitors, these times may happen to have more variable PM2.5?

Page 11, line -2, ".5" seems redundant.

Page 39, first paragraph, line -5, suggests --> suggest

Comments from peer reviewer 2

Peer review of EPA's draft guidance and supporting documents recommending Significant Impact levels (SILs) for ozone and fine particle pollution that may be used in the Prevention of Significant Deterioration (PSD) permitting program

September 29, 2016

I was charged with examining the EPA's drafts of the guidance, legal, technical, and technical appendix documents regarding SILs for Ozone and $PM_{2.5}$. Overall I found the documents to contain sound and well-explained statistical methodology in order to identify ozone and $PM_{2.5}$ SILs for the US. Below I detail my responses to the charge questions.

1. **Are the relevant technical aspects of the statistical procedure clearly described?**

    a. Are input data (EPA's AQS) and their characteristics sufficiently described?

    Yes. Section 2.1 of the Technical Basis document provides details (e.g., where to access and how collected) about each data set, figures mapping the locations of the monitors, and details about the different types of monitors for each data set.

    b. Is it clear what is being estimated?

    Yes. Section 2.1 explicitly defines the DVs for primary ozone NAAQS, primary annual $PM_{2.5}$ NAAQS, and 24-hr $PM_{2.5}$ NAAQS. Section 1 describes the need for and the explanation of a SIL for each of these pollutants.

    c. Is the bootstrap procedure described in sufficient detail to allow reproduction?

    Yes. Section 2.2.3 describes the purpose of bootstrapping and a detailed procedure of how the bootstrap was implemented for each DV in this analysis. Following this outline, replication would be easily doable.

2. **Are the descriptions of statistical concepts clear and accurate?**

    a. Are the descriptions of statistical significance and significance testing clearly and sufficiently described to assist the layperson in understanding the analysis?

    Yes. Sections 1 and 2.2.1 describe statistical significance and "testing" (as it relates to confidence intervals) and connect these concepts to the SIL. Figure 3 is very useful for showing the difference between a 50% CI and 95% CI.

    b. Do the examples provided in the TBD [sic] illustrate the concepts of statistics sufficiently for the layperson to understand the analysis?

    *Yes. However, for clarification purposes, the hypothetical example on page 13 should start, "Suppose the **observed** annual mean $PM_{2.5}$ concentration…" to distinguish this number from the unobserved population mean, to which the previous paragraphs were referring.*

3. Are the assumptions and choices in analysis clearly described and supported?

a. Are the assumptions and choices in the analysis sufficiently documented?

Yes, the technical document describes all assumptions and modeling choices well.

b. Does the document sufficiently describe the sensitivity of results to the choices and assumptions in the analysis? For example, are the technical considerations that support the policy decision to aggregate the variability to a single number clearly articulated?

Yes, however, see part a.i and a.ii of my response to question 5.

4. Are the procedures appropriate for the analytical goals?

a. Is bootstrapping an appropriate technique to quantify the variability in the air quality design value statistics? Is the bootstrapping analysis a reasonable approach to inform a policy determination of Significant Impact Levels?

Yes. Bootstrapping is a method shown to perform well for quantifying uncertainty for a variety of statistics. That said, I have some concern about its ability to properly quantify the uncertainty for the 24-hr $PM_{2.5}$ DV, particularly for monitoring stations with 1:6 sampling frequency. At these sites, there are not many data points to capture much variability for the 98^{th} percentile. However, these sites are relatively few and the DV is an average across three years, thus reducing potential bias. It's not a red flag, but it is something to consider moving forward with the analysis.

5. **In your assessment, is there need for further analysis or clarification? Do you have suggestions for improving the document?**

This document is well written and clearly defines statistical terms and meets the criteria defined therein. I make one suggestion for revision within the document (listed in item a.iii below; and a few typos are noted at the end of the document). While I don't think there is a need for further analysis at this time, I think future iterations of this analysis should consider two items:

a. Spatial variation.

i. The bootstrap method as implemented in this analysis does not account for the strong spatial dependence (described in Section 3.2.1). The researchers implement the bootstrap on each of the locations independent of the other locations. While this is fine for setting individual SILs, making use of spatial dependence within the bootstrap would be a more appropriate way to define a national SIL. Note that some measures have been taken to account for temporal dependence (i.e., insuring that observations sampled in each iteration of the bootstrap are observations from the same quarter), but nothing for spatial dependence.

ii. The discussion of the lack of evidence for regional SIL's is weak. Figures 11 and 14 show strong spatial dependence. Additionally, I would expect that different types of monitors (i.e., those with different sampling frequencies) will exhibit different relative uncertainties. I'd expect that monitors with less frequent measurements are more variable (and this is supported in Table 2) and therefore regional SILs could be considered for the different types of

monitors. The discussion for the desirability a national SIL is solid, but the spatial plots do not give enough evidence that regional SILs would be unreasonable to define.

iii. *The discussion in the final paragraph of page 28 (Section 3.2.1) is poor. They are comparing two very different types of variation: variation between locations and variation within a location. This discussion should be revised or removed.*

b. *Consider a "Significant Impact Threshold." While the 50% CI for the SIL is well motivated as a value for insuring no difference (and the need for this type of a value rather than a threshold), the SIL will be used instead as a threshold limit, when in actuality, it's extremely plausible that values beyond the SIL will not "cause or contribute to an air quality violation." Providing a second level – or a threshold – of "will likely cause or contribute to an air quality violation" (e.g., a level corresponding to 99% or 99.9% CI) would be very valuable for decision makers in managing the individual cases (e.g., rather than the vague 1.2 vs 1.3 descriptions given in the current draft guidance document).*

A few typos:

- *Page 13, final paragraph: "normal distribution" and "Normal Distribution" are both used.*
- *Fourth line of the paragraph under Section 2.2.2.3: "…then **the** mean and the value…"*
- *Page 19: there's an out of place bolded "Error! Bookmark not defined."*
- *Parenthetical statement at the top of page 22: If q=50%, then the percentiles listed are correct. However, they are not correct for any value of q. The statement should read "the lower bound is the (50-q/2)% percentile and the upper bound is the (50+q/2)% percentile."*

Comments from peer reviewer 3

Response to charge questions:

1. **Are the relevant technical aspect of the statistical procedure clearly described?**
   a) **Are input data (EPA's AQS) and their characteristics sufficiently described?**

   In my opinion, the document presents the air quality data in a clear way: the description of the network design is very informative, as well as the description of the different types of spatial scale monitors employed for the two pollutants. Also the discussion of the issue of spatial and temporal variability were well presented and discussed.
   Potentially, a more extended explanation as for why the middle scale is not considered an appropriate spatial scale for PM2.5 could be useful.

   b) **Is it clear what is being estimated?**

   In general the description of the estimation procedure is rather clear, although there are parts of the documents on the estimation procedure that would benefit from a more thorough explanation.
   In more details: the document defines clearly the DV for the two pollutants and determines explicitly what the DVs are in relations to the different NAAQS. The document also clearly explains how the DVs are calculated in the resampled datasets: in particular the extended explanation on page 21 is really helpful. The explanation on how confidence intervals corresponding to different confidence levels are determined in the bootstrap framework is also rather clear. Less clear are the description of the statistics computed and presented in the Results section. Specifically, the document often refers to "difference between the bootstrapping CI value and the actual design value for a single monitoring site". This is quite confusing since a CI is an interval and thus defined by two bounds, while the actual design value at a monitoring site is a number, hence the term difference is rather ambiguous: is the difference between the design value and the upper bound of the bootstrapping CI or the difference between the design value and the lower bound of bootstrapping CI? The label on the horizontal axis of Figure 4 seems to indicate that both differences were calculated (similarly for the axis of Figure 6), however both the text in page 23 and 25 as well as the caption to Figure 4 and 6 is ambiguous. Similarly, the middle panel of Figure 4 and the bottom two panels in Figure 5 are rather confusing and do not present information on the quantities being estimated in an unambiguous way.

   c) **Is the bootstrap procedure described in sufficient detail to allow reproduction?**

   I believe that the explanation of the calculation of a bootstrap CI provided in page 21 clarified greatly the description of the bootstrap procedure given in page 20 and provided enough detail for reproduction.

2. **Are the descriptions of statistical concepts clear and accurate?**
   a) **Are the description of statistical significance and significance testing clearly and sufficiently described to assist the layperson in understanding the analysis?**

In general I think the document does a very good job at presenting statistical concepts to the layperson. The idea of a sample being a representative of the population, the concept of hypothesis testing, the interpretation of the results of an hypothesis test, and the concept statistical significance were all well described. To my opinion, in certain parts the document is not completely precise from a statistical point of view, and I think that a revision of the document to address and correct these slight imprecisions would be ideal.
For example, on page 13 when the document discusses the derivation of confidence intervals, the way the text is written seems to imply that all confidence intervals are derived based on sampling distributions and Central Limit Theorem. While all confidence intervals are derived based on the asymptotic behavior of the sampling distributions, the Central Limit Theorem is a theorem that states the asymptotic behavior of the sampling distribution only of the mean of independent random variables. Thus it could only be used to derive confidence intervals of parameters that can be thought as the mean of a sequence of independent random variables. Calculation of the confidence intervals for other parameters, such as for example the variance, is not based on the Central Limit Theorem, although it is based on the asymptotic behavior of the sampling distribution of the variance.
A second small imprecision is on page 18 when the document discusses assessing the air quality variability: in section 2.2.2.3 it uses the incorrect language "the CI of the sample mean": confidence intervals are not intervals for the sample estimators, but they are intervals for the population parameters. Hence, there "the CI of the sample mean" should be replaced with "CI of the mean".
Besides these small imprecision, the description of statistical concepts is quite clear.

   b) **Do the examples provided in the TBD illustrate the concepts of statistics sufficiently for the layperson to understand the analysis?**

I think that the examples in the document are instrumental for the layperson to completely grasp and understand the statistical concepts presented in the document. I also think that they are well explained and presented.

3. **Are the assumptions and choices in the analysis clearly described?**
a) Are the assumption and choices in the analysis sufficiently documented?
I don't think that the assumptions underlying the analyses are always sufficiently discussed. For example, an underlying assumption of bootstrapping, at least in the implementation of bootstrapping used in the analysis reported in the document, is that the data is considered to be observations of independent random variables. The

document does not explicitly state this underlying assumption, which will translate into assuming that ozone and PM2.5 daily monitoring values at a given sites are independent. This is a strong assumption underlying bootstrapping that the document does not mention openly.
On the other hand, other choices, such as bootstrapping the data within each year independently, resampling data from each 3-month period have been clearly explained and documented.

b) Does the document sufficiently describe the sensitivity of results to the choices and assumptions in the analysis? For example, are the technical considerations that support the policy decision to aggregate the variability to a single national value clearly articulated?

I have found this part of the document (e.g. Section 4) very unclear and not well explained, especially in comparison with the rest of the document. To my opinion sensitivity of the results to the choices and assumptions of the analyses are not discussed at all, and I think that these two points should be addressed in a revised version of the document.

4. Are the procedure appropriate for the analytical goals?

a) Is bootstrapping an appropriate technique to quantify the variability in the air quality design value statistics? Is the bootstrapping analysis a reasonable approach to inform a policy determination of Significant Impact Level (e.g. threshold levels)?

I think that in a nutshell, as general procedure, bootstrap is an appropriate technique to quantify the variability in the air quality design value statistics, especially given that the design value statistics are based on percentiles of the distributions. Thus, given that the sampling distributions of the DV might not be available, bootstrapping can be a mean to quantify the variability and thus derive CI. I also believe that bootstrapping analysis is a reasonable approach to determine Significant Impact Level.
My point of contention with the analysis is that I am not sure that I completely agree with the way bootstrap has been implemented. In particular, I believe that ozone and PM2.5 concentration values at a site are fairly correlated from day to day, and thus the air quality data for a given site might display a significant auto-correlation at lag 1 (meaning that concentrations of ozone measured at a site a day apart are very likely significantly correlated), and might have a significant auto-correlation at longer lags depending on the season. Bootstrapping, in the way it has been implemented in the document, according to the document description, is based on the assumption that the observations are independent, which might not be the case for ozone concentrations. The sampling frequency of PM2.5 concentrations at the monitoring sites might render the PM2.5 data independent, however it is an assumption that should be verified.
Thus, while on a conceptual level, I think that bootstrapping could be used as a reasonable approach for deriving SILs, I think that in the actual

implementation of the bootstrapping method, it needs to be attested whether the observed ozone and PM2.5 concentration data within each 3-month period is independent. In case the assumption of independence is violated, bootstrapping method for temporally correlated data should be used in deriving the re-sampled datasets.

5. In your assessment is there need for further analysis or clarification? Do you have suggestions for improving the document?

As mentioned in the reply to Charge Question 4 above, I believe that there is need for further analysis. In particular I think that the issue of temporal autocorrelation in the data at each site has to be investigated and necessary correction to the bootstrap techniques should be implemented.

In terms of improvement to the documents, I think that the first two sections of the documents are well written and presented and, except for the few corrections suggested above, I do not see much need for improvements in those sections. I believe that the presentation of the results in Section 3 could be improved by clearly stating what are the statistics computed. Finally, as mentioned in the reply to question 3, I believe that Section 4 of the document is quite unclear and the document would improve greatly if a more exhaustive explanation of the considerations in Section 4 is provided.

United States Environmental Protection Agency	Office of Air Quality Planning and Standards Air Quality Analysis Division Research Triangle Park, NC	Publication No. EPA-454/S-18-001 March, 2018