Splicing of Multi-Scale Downscaler Air Quality Surfaces

Elizabeth Herman\textsuperscript{1}, Jeonghwa Lee\textsuperscript{2}, Kartik Lovekar\textsuperscript{3}, Fatemeh Norouzi\textsuperscript{4}, Dorcas Ofori-Boateng\textsuperscript{5}, Benazir Rowe\textsuperscript{6}, Jianhui Sun\textsuperscript{7}

Problem presenters: Elizabeth Mannshardt\textsuperscript{8}, Barron Henderson\textsuperscript{8} and Brett Gantt\textsuperscript{8}
Faculty Mentor: Brian Reich\textsuperscript{9}

Abstract

The United States Environmental Protection Agency (EPA) makes use of a suite of statistical data fusion techniques that combine ambient monitoring data with air quality model results to characterize pollutant concentrations for use in various policy assessments. Data fusion can overcome some of the spatial limitations of monitoring networks and benefit from the spatial and temporal coverage of air quality modeling. The current EPA air pollution prediction model uses a downscaler (DS) model to estimate pollutant concentrations on a national surface. Of interest are ways to improve the performance of the DS in certain areas of the continental US, particularly those with sparse monitor representation. The current methodology utilizes the same spatial range parameter across the continental United States. In order to capitalize on the strengths of spatial modeling capabilities, we consider predictions run on a regional scale. We do this by allowing for a flexible spatial range parameter. By allowing the spatial range parameter to be more localized, we may achieve more precise predictions. We use the output from the DS model run separately on the nine (9) National Oceanic and Atmospheric Administration (NOAA) climatic regions of the continental US with overlap. Findings show that regional and national DS air quality predictions differ significantly for regions where Air Quality Systems (AQS) density is low. We consider an assessment of the regional DS runs to the national runs, as well as a comparison to two (2) national monitoring station networks. Significant discrepancies are seen in the areas where the density of AQS stations is low, such as the Northwest, Northern Rockies, West and Southwest regions. A critical step is splicing the regions back together along the regional boundaries to create a smooth national air quality surface, that is applicable for regions with both low and high discrepancies. The method demonstrates visual smoothing of the two (2) overlapping surfaces and can be extended to handle higher dimensional overlap. Our smoothing parameter is based on the longitudinal distance to the edge of the DS regions. In our paper, we focus on providing smoothed estimates based on predictions from two regional DS runs and provide a comparison of three approaches to generate smoothed estimates for overlap regions.

1 Introduction

In 2016, the US EPA estimated that approximately 123 million people lived in counties where air quality concentrations were above the primary US EPA’s National Ambient Air Quality Standards (NAAQS) [7]. Additionally, in 2018, the World Health Organization also reported an estimate of about 7 million premature deaths caused by ambient air pollution for the same timeline [8]. One type of air pollution is called Particulate matter (PM). It is a complex mixture of extremely small particles and liquid droplets in the air. Research, over the years, has established linkage between exposure to Particulate Matter (PM) and health-related risks such as aggravated asthma, difficulty breathing, chronic bronchitis, decreased lung function, and premature death [3, 4, 6]. PM is also associated with environmental impacts such as visibility impairment and disruption to the natural nutrient and chemical balance of the soil. Specifically, PM\textsubscript{2.5} describes fine inhalable particles, with 2.5 micrometers and smaller diameters. We will focus on PM\textsubscript{2.5} for our assessment.

\textsuperscript{1}Department of Mathematics, North Carolina State University
\textsuperscript{2}Department of Statistics and Probability, Michigan State University
\textsuperscript{3}Department of Statistics, The Ohio State University
\textsuperscript{4}Department of Mathematics, Morgan State University
\textsuperscript{5}Department of Mathematical Sciences, The University of Texas at Dallas
\textsuperscript{6}Department of Mathematical Sciences, University of Nevada Las Vegas
\textsuperscript{7}Department of Statistics, University of Virginia
\textsuperscript{8}US Environmental Protection Agency
\textsuperscript{9}Department of Statistics, North Carolina State University
EPA’s AQS provides measurements of air pollutant concentrations from a monitoring network throughout the United States. The EPAs Community Multi-scale Air Quality (CMAQ) model includes simulations that combine current knowledge in atmospheric science and air quality modeling and multi-processor computing techniques to produce pollutant estimates on a grid surface. Each source of air quality information has different advantages and limitations including the spatial coverage of monitoring networks and the possible modeling bias and uncertainty associated with estimation of air quality. The US EPA currently utilizes a spatial prediction framework called “downscaling” [1, 2] to fuse these two data sources. The DS model fuses CMAQ output with AQS readings using a spatially weighted model that regresses monitoring data on a CMAQ derived regressor. The output of downscaling is estimates of air quality across the United States.

Figure 1: DS regional areas in the continental United States

DS is usually applied to a national scale, incorporating all data from the AQS sites located in the continental United States, yielding a smooth Air Quality surface map [1, 2]. One caveat is that for predictions in the Northwest region (see next section for detailed division of NOAA regions), the model also considers measurements from the AQS monitors in the Southeast region. In other words, the hypothesis is that while regions adjacent to each other are correlated, readings of PM$_{2.5}$ on the other side of the country should not be considered. By considering model runs on a regional scope, we could potentially see a reduction in computational time and increase in prediction accuracy. However, notice in Figure 1, the national surface is no longer smooth. Here we develop methodology to splice adjacent regions together, allowing for a continuous gradient between regions and smooth national surface. The approach considered here is to run the DS on a regional scale to allow for estimation of a regional scale parameter. Ultimately a smooth national surface is desired, so the regions are configured such that there are overlapping areas in-between the regions. That way, a specific grid point only includes information that is relevant to its corresponding region. The overlap allows for consideration of the transition between regions. An example of the overlap region is provided in Figure 2.

Our goal is to compare national and regional DS results and provide a methodology for smooth splicing for the points in overlap region. To splice regions, the methods we call horizontal mixed densities and horizontal mixed variables are introduced. By these methods, two different density functions or two different variables are combined according to the distance which will be defined in section 4.

In this paper we intend to build a methodological splicing framework/algorithm for regional estimates along regional boundaries to create smooth national surface. The rest of the paper is organized as follows: section
2 contains a description of the data, section 3 compares national versus regional DS results, section 4 outlines the methodology developed for splicing regional surfaces across a continuous gradient, and section 5 provides results and computational assessment of splicing. We summarize results and methodological discoveries in the conclusion.

2 Data

We are using regional and national DS predictions based on Community Multi-scale Air Quality (CMAQ) model and Air Quality System (AQS) measurements of air pollutant concentrations from a monitoring network throughout the United States for the first quarter in year 2014 as well as DS Model predictions generated using the two above mentioned data sources in the same time period. The DS model fuses pollution estimates from the CMAQ model with “ground-truth” data from AQS monitor stations to produce estimates of air quality across the United States. The two sources of information are valuable in different ways. The monitoring data are sparsely collected with some missing data, but provide direct measurement of the true pollutant value up to measurement error. AQS monitors are located on the ground and collect concentrations of pollutants (specifically PM$_{2.5}$), the pollutant we are interested in. While we do have actual readings of the concentration of PM$_{2.5}$, this data is usually collected in densely populated areas. CMAQ estimates grid averages with no missing values, but is subject to calibration error.

We use two types of DS output: National and Regional. The DS data contains location labels of the grid cells inside each region, the latitude and longitude of the grid cells center, as well as, mean and standard error of DS prediction for each cell. The regional DS data has the same structure as National DS data, where the main difference between the two is the amount of data being fed into the model as determined by the physical range of the NOAA region. NOAA Climate Zones [3], are a collection of states defined by NOAA in which the states have similar climates. Regional DS estimates were computed for the regions which roughly approximate two degrees outside of the corners of the rectangle encompassing NOAA climate region, more precisely:

<table>
<thead>
<tr>
<th>Region</th>
<th>Latitude Range</th>
<th>Longitude Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ohio Valley (OV)</td>
<td>(33,45)</td>
<td>(-97,-76)</td>
</tr>
<tr>
<td>North East (NE)</td>
<td>(36,50)</td>
<td>(-83,-65)</td>
</tr>
<tr>
<td>South East (SE)</td>
<td>(23,42)</td>
<td>(-88,-73)</td>
</tr>
<tr>
<td>Upper Midwest (UM)</td>
<td>(39,51)</td>
<td>(-99,-81)</td>
</tr>
<tr>
<td>South (S)</td>
<td>(24,42)</td>
<td>(-108,-86)</td>
</tr>
<tr>
<td>North Rockies (NR)</td>
<td>(38,51)</td>
<td>(-118,-93)</td>
</tr>
<tr>
<td>South West (SW)</td>
<td>(29,44)</td>
<td>(-116,-100)</td>
</tr>
<tr>
<td>North West</td>
<td>(40,51)</td>
<td>(-127,-109)</td>
</tr>
<tr>
<td>West</td>
<td>(31,44)</td>
<td>(-127,-112)</td>
</tr>
</tbody>
</table>
The next component of the data is obtained from DS model (DS) ran on regional surfaces. This model uses a grid of continental United States divided into 12km by 12km pieces. It is essentially a linear regression that incorporates data gathered from the AQS sites and other environmental parameters local to the grid cell. The output of DS is a mean estimate of the pollution and its standard error for each grid cell.

Source: NOAA National Climatic Data Center [5]

The DS model used here is a spatially varying weights model that regresses monitoring data on a derived regressor obtained by smoothing the entire CMAQ output with weights that vary both spatially and temporally. This adaptive smoothing of CMAQ was used to achieve stronger association with the monitoring data by taking advantage of useful spatial information in neighboring CMAQ cells surrounding the cell where the monitoring data occurs. Let \( Y(s, t) \) be the observed monitoring station data (AQS) at location \( s \) on day \( t \), and let \( X(B_k, t) \) be the observed numerical model output (CMAQ) for grid block \( B_k \) on day \( t \). The smoothed DS with spatially varying weights model assumes the prevalent model

\[
Y(s, t) = \beta_{0,t} + \beta_0(s, t) + \beta_{1,t}X(s, t) + \epsilon(s, t)
\] (1)
where $\beta_0(s, t)$ is a mean zero Gaussian process with covariance function specified in [2], and $\epsilon(s, t)$ is a white noise Gaussian process with variance $\tau^2_t$. Furthermore, $\tilde{X}(s, t)$ is monitor-specific weighted average of values from grid cell containing $s$ and some neighbors with the weights specified in [1, 2].

Intuitively, the SVW model assumes independence across time while the spatial dependence is governed by a Gaussian process with exponential covariance function. Markov chain Monte Carlo (MCMC) method is used to obtain a sample from the posterior distribution with vague or conjugate priors given to all model parameters. Draws are obtained using a Gibbs sampler where the Metropolis-Hastings accept-reject algorithm is used where necessary. For each model parameter, DS numerical model will output mean and standard error estimates, which denote the mean and standard error of a large number of MCMC samples after burn-in, respectively.

Another piece of data comes from Inter-agency Monitoring of Protected Visual Environments (IMPROVE) stations. Figure 4 shows the location of the IMPROVE monitors located throughout the United States. This data is an independent source, meaning we did not use IMPROVE to inform our DS runs. We will use this data to provide insight as we evaluate the DS on both the regional and national scales. As with our other data sources, we will focus our analysis on the first quarter of 2014. IMPROVE stations measure the concentrations of the pollutant PM$_{2.5}$ in the air.

Unlike the AQS sites, these sites are located in areas where one expects high air quality. For instance, the IMPROVE sites can be located in a National Parks or other rural areas. IMPROVE data-set contains variables such as quarterly averages of the PM$_{2.5}$ value, AQS$_{id}$, latitude and longitude.

3 Downscaler Comparative Assessment

Considering DS outputs such as mean and standard error, we noticed that the standard errors were high in the Northern Rockies region. Because mean estimates were constructed using MCMC, we decided to take a closer look at trace plots. The convergence is usually guaranteed by a sufficiently long run of the algorithm. We do some basic diagnostics to test whether the MCMC sample mean follows theoretical posterior distribution. For model parameter $\beta_{0,t}$ and $\beta_{1,t}$, vague conjugate prior distributions are assumed. Additionally, $\beta_{0,t}$ and $\beta_{1,t}$ are assumed to be independent in time and their posterior distribution should be Gaussian. We provide the following trace plots and QQ plots Figure 5 to justify the convergence of MCMC runs from the first quarter in Northern Rockies and Plains region.
Figure 5: Trace plots and QQ plots for model parameters $\beta_{0,t}$ and $\beta_{1,t}$ of the MCMC runs, from the first quarter data in Northern Rockies and Plains region.

Figure 6: Difference between Regional and National DS calculated using (2).

Although the air quality DS model shows relatively satisfactory agreement with AQS and CMAQ in general, the AQS monitoring stations which inform the DS model are concentrated in urban areas.

This can leave rural areas with less accurate estimates. Additionally, occasional pollution events may be correctly reported by a CMAQ model, but missed by the AQS monitor, leading to an under-estimation of pollution in an area in the DS model. The reverse is also possible - a pollution event may be modeled correctly by CMAQ, but missed by the closest AQS monitor because pollutants are only measured at that AQS site rather than areas between sites. This could result in over-estimation of pollution at an AQS site, but under-estimation in the affected areas. A third scenario is a pollution event that is missed by both the CMAQ model
and AQS sites, resulting in under-estimation of pollution in the downwind areas. EPA is interested in the identification of areas with discrepancies between the point-source monitoring network (AQS and IMPROVE) and DS model.

In order to judge whether DS adequately describes “ground-truth” monitoring data, we calculated their relative discrepancies. Relative discrepancy is determined by comparing the model estimates with AQS or IMPROVE observations. We define relative discrepancy through a Fractional Bias (FB) metric, specifically in form of 2.

\[
FB_{\text{site}_s, DS_k} = \frac{\text{site}_s - DS_k}{(\text{site}_s + DS_k)/2}, \tag{2}
\]

where site_s is the air pollutant readings from AQS or IMPROVE station at location s, and DS_k is the DS output from the k-th grid which includes the AQS or IMPROVE monitor s.

![Figure 7: Difference between the standard error for Regional and National DS calculated using (2).](image)

EPA is interested in the identification of areas with discrepancies between the point-source monitoring network (AQS and IMPROVE network) and DS model.

In Figure 6, we provide the comparison between mean predictions from National DS model and Regional DS model, respectively, while in Figure 7, the comparison between standard errors of prediction from National DS model and Regional DS model is shown, respectively. The distinctiveness of NW region, NR region and W region is in line with the conclusion from 8 and 9. The standard error intuitively characterizes the abundance of information we could get at certain location. Running DS model in National/Regional scale has different magnitude of model uncertainty, which is a noteworthy pattern awaits to be interpreted.

We present how DS model describes AQS measurements in Figure 8. DS outputs from the middle figure come from running the model in a national scale and other nine figures describe DS outputs by running the model in nine regions, respectively. From the figure, we could draw the conclusion that the application of the DS to regional domains (e.g., NOAA climate regions) does not in general improve the DS predictions, and even jeopardize DS performance in certain regions on West Coast. In NW region, National DS and Regional DS tend to have totally opposite conclusions. Specifically, most mis-predictions of Regional DS are mild over-estimations of AQS measurements, while on the other hand, National DS tend to under-estimates NW AQS stations.
In Figure 9, we show the discrepancy between National/Regional DS and IMPROVE station measurements. IMPROVE stations mainly focus on rural areas and are much more sparsely distributed compared to AQS monitoring system. Compared to 8, DS model performs worse in predicting rural areas (Note that the color scale in 9 is more extreme than 8). Most mis-predictions occur in regions on West Coast, and DS model seems to have a significant over-estimation of IMPROVE measurement.

In Figure 10, we provide the ratio of discrepancy between Regional DS and AQS over discrepancy between National DS and AQS. Note that we keep the sign for both discrepancy to discover possible sign changes. From the figure, we further strengthen the previous conclusion, which indicates consistency between DS and AQS overall, while there exist several AQS station with extreme values. The extreme cases do not exhibit distinctive spatial characteristics and are averagely scattered within each region. We provide similar ratio plot with IMPROVE sites in Figure 11, which indicates comparable discrepancy between IMPROVE measurements and Regional/National DS as well. No noticeable spatial pattern could be detected from this ratio plot either.

![Difference for Regional/National DS and AQS](image1)

Figure 8: Difference between the Regional/National DS values and AQS values calculated using (2).

![Difference for Regional/National DS and IMPROVE](image2)

Figure 9: Difference between the Regional/National DS values and IMPROVE calculated using (2).
4 Splicing Methodology

In this section, we consider the situation where we have the overlap between two regions, $R_1, R_2$. The generalization of the methods in this section for the case where two regions overlap primarily along latitudes is straightforward. For each monitoring site, we have DS predicted means, standard deviations from two regions. More specifically, for a monitoring site $s$ in the overlap, $\{\hat{\mu}_i(s), \hat{\sigma}_i(s)\}$ are predicted DS mean and predicted DS standard deviation at $s$ from DS region $i$. Now we consider the following methods.

4.1 Horizontal Mixed Density Method (HMD)

Assuming that for the site $s$ the distribution of DS from region $i$ is normal with mean $\hat{\mu}_i(s)$ and standard deviation $\hat{\sigma}_i(s)$, $i = 1, 2$. To merge the information from two different regions, we use model averaging where
the probability density function of PM$_{2.5}$ at $s$ is a weighted average of probability density functions (pdf) from regions 1 and 2.

$$f_s = w_1(s)f_{1,s} + w_2(s)f_{2,s}$$

where $f_i$ is a normal density function with $\mu = \hat{\mu}_i(s), \sigma = \hat{\sigma}_i(s)$. We set the weight as a function of a distance between $s$ and the outer boundary in the overlap, because as the site lies closer to the outer line, it is less affected by the information from the region. In other words, $w_i(s)$ is proportional to the distance of the site $s$ to the outer boundary of region $i$ in the overlap, because the larger the distance is, the further the point is toward the inner area of region $i$ and less influenced by adjacent region.

Figure 12: Distance from a site to the boundary

Figure 12 shows how the distance is calculated when two regions are overlapped side by side. If the intersection of two regions lies up and down in regions, then the distances are obtained by vertical distances of $s$ to the boundaries. The parameter $\phi$ is also introduced to adjust the effect of distance on weights. More specifically, we have the following weight for the density function from region $i$ at point $s$, $w_{i,s}$:

$$w_i(s) = \frac{e^{-\phi d(s,i)}}{e^{-\phi d(s,1)} + e^{-\phi d(s,2)}}$$

(3)

where $d(s,i)$ is the distance of point $s$ to outer boundary of region $i$ as explained above. This implies that the larger $\phi$ is, the bigger influence the distance has on the weight, making the distribution from each region have a stronger effect on near area. The smaller $\phi$ is, the more two distributions are blended in the overlap. In extreme case when $\phi$ is zero, the weights are simply evenly distributed among the densities, $w_i(s) = .5$ for $i = 1, 2$. If $\phi = \infty$ the probability density function (pdf) $f$ at $s$ simply becomes either $f_{1,s}$ or $f_{2,s}$ depending on which boundary is closer to $s$, and the weight has no effect on averaging. This is shown in Figure 13.
Now we have the following likelihood function for AQS $y_j$ at $s_j \in R1 \cap R2$, $j = 1, ..., n$.

$$L(y_1, ..., y_n; \phi) = \prod_{j=1}^{n} f(y_j) = \prod_{j=1}^{n} w_1(s)f_{1,s}(y_j) + w_2(s)f_{2,s}(y_j)$$  

$$= \prod_{j=1}^{n} e^{-\phi d(s_j,1)}f_{1,s_j}(y_j) + e^{-\phi d(s_j,2)}f_{2,s_j}(y_j)$$  

(4) (5)

Using method, $\hat{\phi}$ is obtained, and the new predicted mean at the site $s$ in the overlap becomes weighted means:

$$\hat{\mu}(s) = \int y f(y)dy = \int \hat{w}_1(s)f_{1,s}(y) + \hat{w}_2(s)f_{2,s}(y)$$  

$$= \int ye^{-\hat{\phi}d(s,1)}f_{1,s}(y) + ye^{-\hat{\phi}d(s,2)}f_{2,s}(y)dy$$  

$$= e^{-\hat{\phi}d(s,1)}\hat{\mu}_1(s) + e^{-\hat{\phi}d(s,2)}\hat{\mu}_2(s)$$  

$$= \hat{w}_1(s)\hat{\mu}_1(s) + \hat{w}_2(s)\hat{\mu}_2(s)$$  

(6) (7) (8) (9)

where $\hat{w}_i(s) = e^{-\hat{\phi}d(s,i)}/(e^{-\hat{\phi}d(s,1)} + e^{-\hat{\phi}d(s,2)})$, $i = 1, 2$.

### 4.2 Horizontal Mixed Variable Method (HMV)

Instead of averaging density functions, we now average two random variables. Assuming that we have two independent normal variables at site $s$ in the overlap, each from DS region $i$, $i = 1, 2$. In other words, for a site $s$ the information from DS region $i$ is assumed to be represented by $X_i(s) \sim N(\hat{\mu}_i(s), \hat{\sigma}_i(s))$, $i = 1, 2$. Our new combined information at site $s$ is then expressed as follows:

$$X_s = w_1(s)X_1(s) + w_2(s)X_2(s)$$

where the weight $w_1(s)$ is defined as in (3). With the assumption that $X_1(s)$ and $X_2(s)$ are independent normal variables, $X$ becomes normal variable with mean $\mu_\phi = w_1(s)\hat{\mu}_1(s) + w_2(s)\hat{\mu}_2(s)$ and variance $\sigma_\phi^2 = w_1(s)^2\hat{\sigma}_1(s)^2 + w_2(s)^2\hat{\sigma}_2(s)^2$. Therefore the following likelihood function is as follows:

$$L(y_1, ..., y_n; \phi) = \prod_{j=1}^{n} f(y_j; \phi)$$
where \( f(y; \phi) \) is Normal density function with mean \( \mu_\phi \) and variance \( \sigma_\phi^2 \).

The Maximum Likelihood Estimate (MLE) of \( \phi, \hat{\phi} \), is obtained and used to compute spliced estimates at the site \( s \) in the overlap as 
\[
\hat{\mu}(s) = \hat{w}_1(s)\hat{\mu}_1(s) + \hat{w}_2(s)\hat{\mu}_2(s),
\]
where 
\[
\hat{w}_i(s) = e^{-\hat{\phi}d(s,i)}/(e^{-\hat{\phi}d(s,1)} + e^{-\hat{\phi}d(s,2)}).
\]

### 4.3 Adaptive Horizontal Mixed Variable Method (AHMV)

The estimates given by HMV show a sharp divide at the edges of the intersection region. This is expected since the estimates do not take into account any values outside the intersection. To adjust for distance, we vary \( \phi \) linearly with distance from the centre of the intersection. Doing this, the degree of influence of distance on weight reduces as we approach edges of the intersection, thereby yielding smoother surfaces at intersection edges.

More specifically, we have the following function for \( \phi \):

\[
\phi(d(s,c)) = \beta_0 + \beta_1 d(s,c)
\]

where \( d(s,c) \) is the horizontal distance of \( s \) to the vertical center line, i.e., \( d(s,c) = |s_{lon} - c_{lon}| \) where \( s_{lon}, c_{lon} \) are longitude of point \( s \) and center line \( c \) respectively. This is explained in Figure 14.

![Figure 14: Distance from a site to the center](image)

Figure 15 is to compare the weight functions when \( \phi \) is constant and \( \phi \) is linear function as in (9). The black and blue graphs are weight functions with constant \( \phi \), 2, 4, respectively. The red graph is the weight function when \( \phi \) is the linear function as in (9) with \( \beta_0 = .3, \beta_1 = .5 \). It is seen that in red graph weights slowly vary in the center and then rapidly increase or decrease as it is moving away from the middle, whereas in the black and blue graphs the weights are either rapidly changing or slowly changing as it moves away from the center. Therefore by using this function for \( \phi \), we can make the surface in the middle of the overlap region blend more, while at the edge of the overlap blend less and give more weight to the mean that lies deeper inside the regional surface, compared to constant \( \phi \) as in HMV.
Figure 15: Weight functions for constant phi and function phi. (a) The black graph is the weight function when phi is constant, 2. (b) The blue graph is the weight function when phi is constant, .4. (c) The red graph is the weight function when phi is the linear function with $\beta_0 = .3, \beta_1 = .5$

5 Results

5.1 Smooth surfaces from three methods

Since the regional DS shows that the most discrepancy in the area is between NW and NR, we restrict ourselves to these regions, applying the three methods in the previous section on the overlap of NW and NR. The following results were obtained from each method:

Table 2 shows MLE of parameter in each method. For HMD and HMV, we have one parameter $\phi$ and for M3 we have $b_0, b_1$ in weight function. The MLE of $\phi$ is 1.450, 0.222 for HMD and HMV respectively implying that HMV smooths the data in the overlap more than HMD since it has larger $\hat{\phi}$. For M3, $\hat{b_0} = 0.157, \hat{b_1} = 0.029$. We apply these estimates to our models and obtain figures 16 and 18.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMD</td>
<td>$\phi$</td>
<td>1.450</td>
</tr>
<tr>
<td>HMV</td>
<td>$\phi$</td>
<td>0.222</td>
</tr>
<tr>
<td>AHMV</td>
<td>$(\beta_0, \beta_1)$</td>
<td>$(0.157, 0.029)$</td>
</tr>
</tbody>
</table>

By looking at Figure 16 and Figure 16 b, we can see that HMV shows smoother surface than HMD. The cause for the difference is explained by the respective functions in (Figure 17).

The black and red graph in Figure 17 are drawn with two normal densities, $N(0, 1), N(3, 1)$ with weight 0.3, 0.7 from HMD and HMV respectively. The blue graph is drawn with the same densities but with weights 0.1, 0.9 respectively with HMD. The means of the black and red density are the same, 0.21. Density function from HMV is normal density, symmetric, and has a peak at the mean. Density function from HMD is skewed to the left which has smaller weight, and has peak at 3. The peak in HMD is 3 when the second weight is larger than 0.5, highest when the weight is 0.5, and getting lowered making more skewed distribution as the weight is decreased. This can be seen by comparing the blue and black graph. It is that property that makes MLE
of $\phi$ in HMD larger than that of HMV, making the values in the overlap stay closer to the mean from closer area, thus producing less smooth surface.

With HMV and AHMV, we can see that the AHMV attains better smoothing near the outer boundary of NW region, since we saw the sharp edge on the outer boundary of NW on HMV. This sharpness becomes blurred in AHMV’s image, and the model smoothly blends two predictions together. This is expected as we make $\phi$ a linear function of distance from the center line. There is a sharp edge at the bottom of the overlap which can be eliminated using similar methods run by incorporating latitudinal distances, which is left for future work.

**Remark.** There are many other ways to splice two regions making smooth surface. If one uses different form of $\phi$, and/or different distance measures, different smooth surfaces might be obtained. In this study we only use either longitude of sites or latitude of sites for splicing two region to measure the distance but one can also use both latitude and longitude achieving smoothness in both direction horizontally and vertically. Also $\phi$ can be a quadratic form of distance yielding a perhaps smoother surface. We wish to mention that there could be a case where different smooth surfaces produced by different methods are all very smooth and at the same time they all fit the data well. In this case, splicing surfaces would be subjective. Nonetheless our method is intuitive and a useful tool when two different estimates on the same point needs to be combined.

We choose model 3 to run diagnostics since it takes more information than other two methods into account and looks smoother upon a visual inspection. To validate the method, we run it for the regions where regional DS predictions differ most from national estimates, the North West and Northern Rockies. The overlap of those 2 regions contains 5,978 grid cells and 45 AQS stations. We use the 45 AQS observations and corresponding DS predictions from regional DS estimates to fit the model using method 3. We use the R functions `mle` and `ggplot` for the same.

We compute Mean Square Error (MSE) for estimates obtained from the 3 models compared to the AQS and IMPROVE sites.

Compared to AQS data, the MSE is not very different between two models and is around 2.8. Compared to IMPROVE sites, the MSE is around 45. A possible implication is that DS is not accurate at rural sites, where
Figure 17: Density functions from different models: (a) The black graph is drawn with two normal densities, $N(0, 1), N(3, 1)$ with weight 0.3, 0.7 from method 1. (b) The red graph is drawn with the same densities and same weights but with method 2. (c) The blue graph is drawn with the same densities but with weights 0.1, 0.9 respectively with method 1.

Figure 18: AHMV applied on the intersection of NR and NW

IMPROVE should possibly be favoured. On the other hand, this is expected since we use AQS data to fit our model and this data has also been used by DS. Also, there are only 14 corresponding IMPROVE sites in the North West - Northern Rockies intersection, which may not be comprehensive.

Next, we perform a validation procedure by subsetting 80% of the AQS data points as the training set and the other 20% of the points as the test set. We compare the Mean Squared Error (MSE) of our smoothed predictions using full and subset of the available AQS data. The MSE for the estimates obtained using 80% of the data is 4.632, while the MSE for the estimates obtained using all 100% of the AQS stations in the overlap
of NR and NW is 4.64. The MSE of both estimates are very close, signaling that our estimates are likely not very sensitive to the little differences in data.

## 6 Summary and Future Work

Each of the three methods presented achieves different degrees of smoothness. Model 3 yields a surface based on the two regional DS model predictions and distance parameters estimated from the AQS data. For any two overlapping DS regions, the method produces a smooth surface while adjusting for both high and low discrepancies between them.

While our approach achieves a smooth surface, we only account for longitudinal distances. Taking latitude into account should improve the smoothing process and get rid of the sharp horizontal edges that are currently present in all 3 methods suggested. Since maps are inherently 2-dimensional, the first extension of the existing approach would be addition of the second dimension (latitudinal distance) to influence blending coefficient. Another extension would be considering more than double overlap regions and combining 3 or 4 predictions simultaneously. Since DS regions are rectangular, the corners could yield a group of rectangular areas with single, double, triple and quadruple overlaps. Next, inclusion of IMPROVE sites in the estimation procedure may provide additional precision of the resulting estimates. It is especially important for regions like Southwest, where AQS stations are mostly located on the edges of the region and the density of AQS stations is relatively low.

Our method could handle regions with both high and low density of AQS stations. It is motivated by providing a smooth surface of the map while optimizing for minimal discrepancy between the predictions and AQS readings. This implies that the method provides a way to optimally mask the discrepancies between regional and national DS predictions.

A high discrepancy in the Regional and National DS predictions signal that the DS model could be inherently local in nature and is highly dependent on the perspective, i.e. chosen size and location of the region as well as the density and topology of the AQS stations in that region. It is not clear yet to what extent all those variables affect the accuracy of the DS model and how they should be accounted, if deemed necessary, when choosing the shape and size of the regions on which the DS model is run.

## References


