



Article

Spatiotemporal Interpolation Methods for the Application of Estimating Population Exposure to Fine Particulate Matter in the Contiguous U.S. and a Real-Time Web Application

Lixin Li ^{1,*}, Xiaolu Zhou ², Marc Kalo ¹ and Reinhard Piltner ³

¹ Department of Computer Sciences, Georgia Southern University, Statesboro, GA 30460, USA; marc.kalo@gmail.com

² Department of Geology and Geography, Georgia Southern University, Statesboro, GA 30460, USA; xzhou@georgiasouthern.edu

³ Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460, USA; rpiltner@georgiasouthern.edu

* Correspondence: lli@georgiasouthern.edu; Tel.: +1-912-478-7646

Academic Editor: Kim Natasha Dirks

Received: 28 March 2016; Accepted: 4 July 2016; Published: 25 July 2016

Abstract: Appropriate spatiotemporal interpolation is critical to the assessment of relationships between environmental exposures and health outcomes. A powerful assessment of human exposure to environmental agents would incorporate spatial and temporal dimensions simultaneously. This paper compares shape function (SF)-based and inverse distance weighting (IDW)-based spatiotemporal interpolation methods on a data set of PM_{2.5} data in the contiguous U.S. Particle pollution, also known as particulate matter (PM), is composed of microscopic solids or liquid droplets that are so small that they can get deep into the lungs and cause serious health problems. PM_{2.5} refers to particles with a mean aerodynamic diameter less than or equal to 2.5 micrometers. Based on the error statistics results of k-fold cross validation, the SF-based method performed better overall than the IDW-based method. The interpolation results generated by the SF-based method are combined with population data to estimate the population exposure to PM_{2.5} in the contiguous U.S. We investigated the seasonal variations, identified areas where annual and daily PM_{2.5} were above the standards, and calculated the population size in these areas. Finally, a web application is developed to interpolate and visualize in real time the spatiotemporal variation of ambient air pollution across the contiguous U.S. using air pollution data from the U.S. Environmental Protection Agency (EPA)'s AirNow program.

Keywords: fine particulate matter (PM_{2.5}); spatiotemporal interpolation; shape function; Inverse Distance Weighting (IDW); cross validation; population exposure; web application; visualization; real-time air pollution

1. Introduction

Particulate matter (PM) is the generic term for a broad class of chemically and physically diverse substances that exist as discrete particles (liquid droplets or solids) over a wide range of sizes [1]. Some particulates occur naturally, originating from volcanoes, dust storms, forest and grassland fires, living vegetation, and sea spray. Some other particulates are made from human activities, such as the burning of fossil fuels in vehicles, power plants, and various industrial processes [2]. The United States Environmental Protection Agency (EPA) first established national ambient air quality standards for PM in 1971. The published evidence supports an association between PM and an increased

risk of mortality. It has been shown that those with cardiovascular or respiratory conditions and the youth and elderly are the most susceptible to the adverse effects of PM. The pollutant class studied in this paper is specifically fine particulate matter, or PM_{2.5}, which refers to particles with a mean aerodynamic diameter less than or equal to 2.5 micrometers. PM_{2.5} is considered one of the most unhealthy particulate air pollutants because it is more likely to be toxic and can be breathed more deeply into the lungs. PM_{2.5} has been associated with visibility reduction [3,4], acute stroke mortality [5], and daily mortality in many U.S. cities [6].

In order to find the association between air pollutants such as PM_{2.5} and health effects, researchers need to estimate pollutant concentrations in the continuous space-time domain. Since concentration values are typically measured only at discrete monitoring sites and at certain time instances, estimation of pollutant concentrations at unmeasured locations and times is needed. Implementing an appropriate interpolation method is critical to the assessment of relationships between air pollution exposure and health outcomes.

Spatial interpolation has been well developed and widely used in Geographic Information Systems (GIS). It is used to estimate values at unknown locations based upon values that are spatially sampled. Traditional spatial interpolation models have been extensively investigated over the years. Popular spatial interpolation methods are Inverse Distance Weighting (IDW) [7,8], shape functions [9,10], radial basis functions [11], spline [12], natural neighbor [13], trend surfaces [14], Kriging [15], model-data fusion (sometimes called *analysis*) [16,17], and optimal interpolation [18]. IDW, shape functions, radial basis functions, splines, natural neighbor, and trend surfaces are deterministic methods. They provide no indication of the extent of possible errors. Their output is fully determined by the parameter values and the inputs. There are no strict assumptions about the variability or randomness of a feature. These methods are relatively simple to implement. On the other hand, Kriging, model-data fusion, and optimal interpolation are stochastic methods that possess some inherent randomness. The same set of parameter values and inputs will lead to an ensemble of different outputs. Stochastic methods provide probabilistic estimates. One of the advantages of stochastic methods is that they treat clusters more like single points and assign individual points within a cluster less weight than isolated data points, which helps to compensate for the effect of data clustering. In the field of atmospheric data analysis, model-data fusion and optimal interpolation methods are developed to include physics and chemistry of an air quality model in the interpolation mechanism and thus achieve better prediction and representation of air quality.

Nowadays, modern sensors are able to monitor different variables (such as particulate matter, sulfur dioxide, and ozone) at an increasing temporal resolution, resulting in rich spatiotemporal data sets. This calls for appropriate theories and methods to deal with these data sets to gain a better understanding of the observed spatiotemporal processes. Traditionally, many GIS researchers treat space and time separately [19]. They simply reduce spatiotemporal interpolation problems to spatial interpolation problems by assuming that time can be incorporated by conducting a sequence of snapshots of spatial interpolations. Since the spatiotemporal interpolation considers the additional time attribute, it can provide more accurate predictions than pure spatial interpolation. However, adding the temporal domain implies that variability in space and time must be modeled, which is more complicated than modeling purely spatial or purely temporal variability. A review of some air pollution exposure assessment methods utilized in epidemiological studies and the use of GIS for resolving problems with spatiotemporal attributes can be found in [20]. Other work on spatiotemporal interpolation are presented in the literature [9,21–28].

The main challenge presented by the spatiotemporal interpolation relates to the spatiotemporal dependence structure, i.e., the relative importance of time with reference to space. A powerful assessment of human exposure to air pollution would incorporate spatial and temporal dimensions. The temporal dimension of environmental exposure analysis is often ignored, underemphasized, or isolated from the spatial domain mainly due to the few efficient and effective tools to interpolate complex spatiotemporal datasets. The popular ArcGIS software (version 10.3, ESRI, Redlands,

CA, USA) cannot handle spatiotemporal interpolation and is computationally inefficient with large datasets.

This paper has three goals. First, it investigates and compares two different spatiotemporal interpolation methods for an actual set of PM_{2.5} data measured by U.S. EPA monitoring sites in the contiguous United States: shape function (SF)-based vs. Inverse Distance Weighting (IDW)-based methods using the so-called *extension approach*. The extension approach has been proposed in [9] to integrate space and time simultaneously by extending spatiotemporal interpolation problems into higher dimensional spatial interpolation problems. SF and IDW are originally deterministic spatial interpolation methods. Since they can be extended to higher dimensions, they are both suitable for the extension approach. Furthermore, IDW is one of the most commonly used interpolation methods [7,23,29–31] for GIS applications. Although SF was initially from engineering, it has shown great interpolation performance in various GIS application data such as real estate data [9] and air pollution data [23,24,32]. Second, after obtaining the comparison results of the SF-based and IDW-based spatiotemporal interpolation methods, we apply the better method to estimate population exposure to PM_{2.5} in the contiguous United States using interpolated daily PM_{2.5} concentration values at the centroids of census block groups. Third, we aim to develop a web application to interpolate and visualize in real time the spatiotemporal variation of ambient air pollution (including but not limited to PM_{2.5}) across the contiguous U.S. using air pollution data from the U.S. EPA's AirNow program.

2. Methods

2.1. Shape Function-Based Spatiotemporal Interpolation Using the Extension Approach

Shape functions (SF) have been popular and utilized in engineering applications such as finite element algorithms [10,33]. Just like other traditional spatial interpolation methods used in GIS such as IDW [8] and Kriging [15], SF-based methods assume a stronger correlation among points that are closer than those farther apart. Therefore, SF-based methods can be spatial interpolation methods for GIS applications [9,25,34–38]. In addition, because the computational complexity of SF-based methods is linear, they can be efficient interpolation methods for large data sets.

2.1.1. General Formula of the SF-Based 3D Spatial Interpolation Method

In order to apply SF-based interpolation methods, a mesh that divides the total domain into a finite number of simple sub-domains or elements should be generated. For a 3D spatial problem, a mesh composed of tetrahedral elements should be generated if one wants to use shape functions for tetrahedra to interpolate unknown values in the 3D (x, y, z) coordinate system. Considering the tetrahedral element in Figure 1, the SF-based interpolation result w at an unknown point (x, y, z) located inside the tetrahedron can be obtained by using the measurement values w_1, w_2, w_3 , and w_4 at the four known locations, which serve as the corner vertices of the tetrahedron as in [9]:

$$w(x, y, z) = N_1(x, y, z)w_1 + N_2(x, y, z)w_2 + N_3(x, y, z)w_3 + N_4(x, y, z)w_4 \quad (1)$$

where N_1, N_2, N_3 and N_4 are the following linear shape functions:

$$N_1(x, y, z) = \frac{\mathcal{V}_1}{\mathcal{V}}, N_2(x, y, z) = \frac{\mathcal{V}_2}{\mathcal{V}}, N_3(x, y, z) = \frac{\mathcal{V}_3}{\mathcal{V}}, N_4(x, y, z) = \frac{\mathcal{V}_4}{\mathcal{V}} \quad (2)$$

$\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3$ and \mathcal{V}_4 are the volumes of the four sub-tetrahedra $w_2w_3w_4, w_1w_3w_4, w_1w_2w_4$, and $w_1w_2w_3$, respectively; and \mathcal{V} is the volume of the bounding tetrahedron $w_1w_2w_3w_4$ as shown in Figure 1.

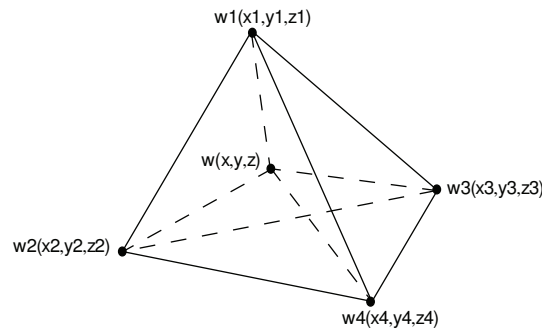


Figure 1. A tetrahedral element. Computing 3D shape functions by tetrahedral volume divisions. w_1 , w_2 , w_3 and w_4 are measured values, while the value w at the location (x, y, z) is unknown and needs to be interpolated.

It can be seen from Figure 1 that \mathcal{V}_1 is the volume of the sub-tetrahedron with four corner vertices as the unknown point (x, y, z) and three known points 2–4. Suppose the unknown point moves closer to the known point 1. \mathcal{V}_1 is increasing, while \mathcal{V}_2 , \mathcal{V}_3 and \mathcal{V}_4 are decreasing, which lead to the increment of N_1 and decrement of N_2 , N_3 and N_4 . In an extreme case, when the unknown point moves to the exact location 1, the weight of N_1 becomes 1 and the other three weights N_2 , N_3 and N_4 become 0. Similar observations can be made that any of the other three known points 2–4 will contribute a heavier weight in interpolating the value at the unknown point when the unknown point gets closer to this particular known point.

In finite element methods, shape functions of different orders (linear, quadratic, cubic, etc.) are used. In engineering, finite elements are used to approximate processes governed by differential equations such as deformations and stresses in a car. Whereas in engineering the nodal values at the corners of finite elements are all unknown and have to be computed from a large system of equations, in GIS applications, the nodal values at the element corner points come from measured data collections. The common point between finite element and GIS applications is that with nodal values the interpolation function can be evaluated for the complete domain. The size of the finite elements depends on the gradients and the changes in the function. For high gradients and oscillating functions, more elements of smaller size are needed. For the data interpolation, the situation is similar: if we expect high gradients and a lot of changes in a relatively small area, then we would ideally need a sufficiently high number of discrete data values to result automatically in a larger number of smaller elements.

2.1.2. Extension Approach of the SF-Based Spatiotemporal Interpolation Method

Although spatial interpolation methods are well developed and widely adopted in various GIS applications [39–42], the traditional spatial interpolation methods face many challenges when handling spatiotemporal data because of the addition of the time attribute of the data set. One of the major challenges is that traditional GIS researchers tend to treat space and time separately when interpolation needs to be conducted in the continuous space-time domain. The primary strategy identified from the literature is to reduce spatiotemporal interpolation problems to a sequence of snapshots of spatial interpolations [19]. However, integrating space and time simultaneously is anticipated to yield better interpolation results than treating them separately for certain typical GIS applications [43].

In order to integrate space and time simultaneously for a spatiotemporal interpolation, the *extension approach* has been proposed in [9] and reviewed in [37,38]. This approach treats time as another dimension in space, thereby extending the spatiotemporal interpolation problem into a higher-dimensional spatial interpolation problem. Applications using the extension approach can be found in [9,32,44,45]. To develop the extension approach for SF-based interpolation methods,

we substitute the z variable in Equations (1) and (2) by ct , where t is the time variable and c is a factor of [spatial distance unit/time unit]. Equations (3) and (4) define our SF-based spatiotemporal interpolation method for 2D space and 1D time problems:

$$w(x, y, ct) = N_1(x, y, ct)w_1 + N_2(x, y, ct)w_2 + N_3(x, y, ct)w_3 + N_4(x, y, ct)w_4 \quad (3)$$

where N_1, N_2, N_3 and N_4 are the following linear shape functions:

$$N_1(x, y, ct) = \frac{V_1}{V}, N_2(x, y, ct) = \frac{V_2}{V}, N_3(x, y, ct) = \frac{V_3}{V}, N_4(x, y, ct) = \frac{V_4}{V} \quad (4)$$

Please note that there are some assumptions and resulting limitations for this approach. We assume that there are sufficient data measurements in space and time so that simple functions can be used to describe what is happening between two measurements. If there would be a relatively large time interval or data are scarcely sampled in space, and the type of data under consideration have the potential of strong oscillations between the points in space and time, we would not be able to use a simple linear function to interpolate from one space-time point to the next point. Therefore, before using this simple spatiotemporal approach, we have to make sure that the process that we analyze cannot show strong oscillations, and that we have sufficient measurements in space and time. We are only using the method to evaluate for events that already happened. We are not trying to predict the future with this method.

2.2. IDW-Based Spatiotemporal Interpolation Using the Extension Approach

Inverse Distance Weighting (IDW) is also known as Shepard's method [7,8]. Similar to SF-based interpolation methods, IDW is based on Tobler's First Law of Geography [46], which states: "Everything is related to everything else, but near things are more related than distant things", page 236. IDW is generally considered a spatial interpolation method, but this paper applies IDW to spatiotemporal interpolation by using the *extension approach* and treating time as a third dimension [9,37].

2.2.1. General Formula of the IDW-Based Spatial Interpolation Method

According to [47], the general formula of the IDW-based interpolation method in 2D space is:

$$w(x, y) = \sum_{i=1}^N \lambda_i w_i, \quad \lambda_i = \frac{\left(\frac{1}{d_i}\right)^p}{\sum_{k=1}^N \left(\frac{1}{d_k}\right)^p} \quad (5)$$

where $w(x, y)$ is the interpolated value at the unknown (or unsampled) location (x, y) , N is the number of nearest known points surrounding (x, y) , w_i are the measurement values at the nearest known points of (x_i, y_i) (with $1 \leq i \leq N$), λ_i are the weights assigned to w_i , d_i are the Euclidean distances between each (x_i, y_i) and (x, y) , and p is the exponent that influences the weighting of w_i on w .

2.2.2. Extension Approach of the IDW-Based Spatiotemporal Interpolation Method

The formula of the extension approach of IDW used in this paper is

$$w(x, y, ct) = \sum_{i=1}^N \lambda_i w_i, \quad \lambda_i = \frac{\left(\frac{1}{d_i}\right)^p}{\sum_{k=1}^N \left(\frac{1}{d_k}\right)^p} \quad (6)$$

where

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2 + c^2(t_i - t)^2} \quad (7)$$

and c is a factor defined as [spatial distance unit/time unit]. Compared with Equation (5), Equation (6) replaces $w(x, y)$ with $w(x, y, ct)$ and calculates the distance d_i using the 3D Euclidean distance between (x_i, y_i, ct_i) and (x, y, ct) .

2.3. Cross Validation

The first goal of this paper is to compare whether SF-based or IDW-based spatiotemporal interpolation using the extension approach is more accurate in interpolating an actual set of data of daily fine particulate matter $PM_{2.5}$ in the contiguous United States. K-fold cross validation [48] is used in this paper for this purpose.

2.3.1. K-Fold Cross Validation

Classic validation divides the full data set into two data sets: a training data set and a validation data set. The validation data set is used for estimating the performance of the interpolation method based on the training data set. The interpolation method with the smallest error is selected as the best method. However, a potential flaw is that we may miss some characteristics in the full data set and make an inaccurate estimate of our model's interpolation ability. Thus, k-fold cross validation is used to avoid this limitation. In this framework, the full data set is randomly split into k equal-sized data sets, with one group as the validation set and the remaining k-1 groups together forming the training set. This is repeated k times. In practice, 10-fold ($k = 10$) cross-validation is accepted as providing a highly accurate estimate of a model's prediction errors. For large data sets, this approach may be computationally expensive.

Using 10-fold cross validation, the $PM_{2.5}$ data set in our experiment is partitioned to ten nearly equally sized folds randomly. Ten iterations of training and validation are performed such that, within each iteration, a different fold of the data is held-out for validation while the remaining nine folds are used for learning. More specifically, within each iteration, the following two actions are taken:

1. The points in one fold (test data) of the $PM_{2.5}$ data set are interpolated using the remaining nine folds (training data). Therefore, each point in the test data will have both the original $PM_{2.5}$ concentration measurement and an interpolated $PM_{2.5}$ concentration value.
2. Error statistics are calculated to compare the original and interpolated $PM_{2.5}$ values in the test data.

2.3.2. Error Statistics

The error statistics used in this paper are: MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MARE (Mean Absolute Relative Error). They are defined as follows:

$$\begin{aligned} MAE &= \frac{\sum_{i=1}^N |I_i - O_i|}{N} & MSE &= \frac{\sum_{i=1}^N (I_i - O_i)^2}{N} \\ RMSE &= \sqrt{\frac{\sum_{i=1}^N (I_i - O_i)^2}{N}} & MARE &= \frac{\sum_{i=1}^N \frac{|I_i - O_i|}{O_i}}{N} \end{aligned} \quad (8)$$

where N is the number of observations, I_i is the interpolated value, and O_i is the original measurement value. For each iteration of 10-fold cross validation, we have the assumption that a different set of training data has true measurements. From the mathematical point of view, it is reasonable to calculate averages of 10 sets of error statistics. We use \overline{MAE} , \overline{MSE} , \overline{RMSE} and \overline{MARE} to denote the average error statistics results in this paper.

In addition, we use an R^2 error statistic, which is also known as *the coefficient of determination*. The regular R^2 error statistic measures how close the data are to the fitted regression line, whereas the R^2

in [49] measures how close the data are to the 1–1 line. In this paper, we use the R^2 error statistic defined in [49]:

$$R_{CV}^2 = \max\left(0, 1 - \frac{RMSE^2}{MSE_{obs}}\right) \quad MSE_{obs} = \frac{\sum_{i=1}^N (O_i - \bar{O})^2}{N} \quad (9)$$

where \bar{O} is the mean of the original values. As for the other error statistics in Equation (8), the average of the ten R_{CV}^2 results needs to be calculated. We use $\overline{R_{CV}^2}$ to denote the average R_{CV}^2 result in this paper.

2.4. Linking $PM_{2.5}$ to Census Population

We collected population data at the census block group level. To map $PM_{2.5}$ and population spatial distribution, we created choropleth maps based on the interpolated values as well as population data. Because $PM_{2.5}$ may exhibit different spatial patterns in different seasons, we also investigated the seasonal variations. In order to associate the $PM_{2.5}$ values with the national standard, the revised U.S. EPA National Ambient Air Quality Standards for $PM_{2.5}$ in 2006 were adopted in this paper. We conducted spatial queries to identify areas where annual and daily $PM_{2.5}$ are above the standard and calculated the population size in these areas.

3. Experimental Data

The data used in this study are daily $PM_{2.5}$ concentrations measured in 2009 by U.S. EPA monitoring sites in the contiguous United States.

3.1. $PM_{2.5}$ Data Set with Measurements

The data coverage contains locations of the monitoring sites, the daily concentration measurements of $PM_{2.5}$, and the days of the measurements. We obtained a number of data sets from the U.S. EPA website [50] and reorganized them into a data set with the schema (id , x , y , [time], w), where x and y are the longitude and latitude coordinates of the monitoring sites, [time] is (year, month, day) when a $PM_{2.5}$ measurement is taken, and w is the measured $PM_{2.5}$ value. The reorganized data set has some entries with zero $PM_{2.5}$ values, which means no measurements were available at a particular site and on a particular day. After all the zero entries are deleted, there are 146,125 daily measurements at 955 monitoring sites. The monitoring sites are illustrated as stars (*) in Figure 2.



Figure 2. U.S. Environmental Protection Agency (EPA) monitoring sites. These monitoring sites have $PM_{2.5}$ (fine particulate matter) measurements across the contiguous United States in 2009.

3.2. Census Block Group Data Set to Interpolate

In our experiment, we want to interpolate daily PM_{2.5} concentration values in 2009 at the centroids of all the 207,630 census block groups in the contiguous United States. Census block groups are statistical divisions of census tracts and are generally defined to contain between 600 and 3000 people. They are the smallest geographical unit for which the United States Census Bureau publishes sample data. Our experimental data set with locations to compute interpolation has the format of (id, x, y) with id as the identification number of a census block group and (x, y) as the longitude and latitude coordinates of the centroid of a census block group. Since PM_{2.5} concentration values at the centroid of each census block group and on each day in 2009 are not measured, there are $207,630 \times 365 = 75,784,950$ PM_{2.5} values to be interpolated.

The motivation of interpolating at the small geographic level of the census block group is that we aim to link the interpolation results with the census block group population data in the same year for the second goal of this paper. As discussed in the Results section of the paper, we analyze population exposure to PM_{2.5} and estimate the U.S. population with unhealthy PM_{2.5} exposure. In future work, such estimates are important and we plan to link them to a variety of health outcomes to evaluate PM_{2.5}'s adverse impact on human health.

4. Results

4.1. Cross Validation Results of the SF-Based Method

4.1.1. Choice of Time Scale

In order to decide on an appropriate time scale for the SF-based method using the extension approach, we tested four time scales as shown in Table 1. The factor c in the table is from Equations (3) and (4).

Table 1. Four times scales tested for the PM_{2.5} (fine particulate matter) data set.

Time	Scale A ($c = 1$)	Scale B ($c = 1/10$)	Scale C ($c = 1/5$)	Scale D ($c = 1/15$)
01/01/2009	1	0.1	0.2	0.067
01/02/2009	2	0.2	0.4	0.133
01/03/2009	3	0.3	0.6	0.2
01/04/2009	4	0.4	0.8	0.267
...
12/31/2009	365	36.5	73	24.333

A challenge of using the extension approach for spatiotemporal interpolation is the correlation between space and time, and which choice of the factor c is optimal for a particular data set. This is an open question and a research topic in GIS that has been rarely studied. In this paper, authors tested only four possible time scales in Table 1. More research is needed to address this challenge in the future.

4.1.2. Cross Validation and Error Statistics

Ten-fold cross validation was implemented to test the four time scales in Table 1. Since there are ten iterations in 10-fold cross validation and a different fold of the data is held-out for validation during each iteration, the average of ten error statistics has been calculated for each error statistic in Equation (8).

Table 2 shows the results for the average error statistics (\overline{MAE} , \overline{MSE} , \overline{RMSE} , \overline{MARE} , and $\overline{R_{CV}^2}$) using the SF-based extension method for the PM_{2.5} data set. All of these five measures of error

statistics are based on interpolated and original values, I_i and O_i in Equations (8) and (9), but they have different sensitivity to error patterns. The ideal situation is that \overline{MAE} , \overline{MSE} , \overline{RMSE} , and \overline{MARE} are lowest, while R_{CV}^2 is the highest for the same time scale choice. If not, we need to make a choice according to the characteristics of the five error measures. MSE, RMSE, and R_{CV}^2 are sensitive to individual outliers. MAE is less sensitive to outliers but could not reflect the relative prediction errors. MARE is less sensitive to outliers and also incorporates the predictive mean to measure the error from a model prediction. The same size of an error is not acceptable for a small predicted mean but could be acceptable for a large predicted mean. MARE is a better choice to evaluate overall model performance. However, if outliers are major concerns, RMSE or R_{CV}^2 would be better choices.

Table 2. Error statistics for the PM_{2.5} data set using the shape function-based extension method and 10-fold cross validation before removing outliers.

Error Statistics	Scale A ($c = 1$)	Scale B ($c = 1/10$)	Scale C ($c = 1/5$)	Scale D ($c = 1/15$)
\overline{MAE}	3.1512	3.5576	3.2463	3.7307
\overline{MSE}	85.8621	78.5322	78.4890	77.1072
\overline{RMSE}	8.8832	8.6045	8.6067	8.5023
\overline{MARE}	3.2162	0.4158	0.3745	0.4365
R_{CV}^2	0.3079	0.3226	0.3138	0.3382

We produced a scattered plot to compare observed daily PM_{2.5} values with interpolated daily PM_{2.5} values across monitoring sites. Please see Figure 3. Descriptive statistics show that the original PM_{2.5} values contain 16 outliers with PM_{2.5} values above $250 \mu\text{g}/\text{m}^3$, which were much higher than the normal range. According to the National Ambient Air Quality Standards (NAAQS) established by the U.S. EPA under authority of the Clean Air Act, the 24 h standard for PM_{2.5} is met if the three-year average of the annual 98th percentile of values at designated monitoring sites in an area is less than or equal to $35 \mu\text{g}/\text{m}^3$ [51]. The PM_{2.5} values above $250 \mu\text{g}/\text{m}^3$ might be wrongly recorded or some short and extreme conditions happened. These conditions are not usual, so we removed these 16 outliers with PM_{2.5} values greater than $250 \mu\text{g}/\text{m}^3$ from the original 146,124 values. The new error statistics result after removing the outliers are recorded in Table 3.

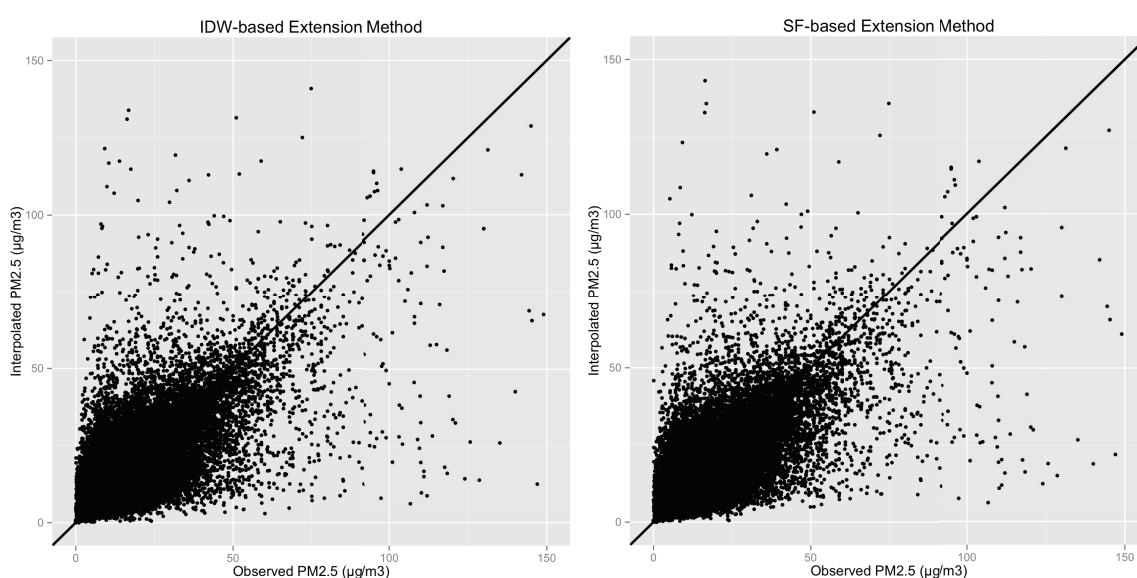


Figure 3. Scattered plots. Comparing observed daily PM_{2.5} values with interpolated daily PM_{2.5} values across monitoring sites across the contiguous United States in 2009.

Table 3. Error statistics for the PM_{2.5} data set using the shape function-based extension method and 10-fold cross validation after removing outliers.

Error Statistics	Scale A (c = 1)	Scale B (c = 1/10)	Scale C (c = 1/5)	Scale D (c = 1/15)
\overline{MAE}	3.0941	3.4976	3.1812	3.6751
\overline{MSE}	42.2910	37.7745	35.6601	39.2077
\overline{RMSE}	6.5032	6.1461	5.9716	6.2616
\overline{MARE}	3.2135	0.4128	0.3708	0.4349
$\overline{R^2_{CV}}$	0.4817	0.5371	0.5630	0.5195

Compared with Table 2, Table 3 shows better error statistics for all measures. Scale C outperformed the other three scales on all error statistics, except for Scale A on \overline{MAE} . However, Scale A performed significantly poorly on \overline{MARE} , \overline{RMSE} , $\overline{R^2_{CV}}$, and \overline{MSE} . Thus, Scale C is selected as the best time scale for daily PM_{2.5} interpolation using the SF-based extension method.

4.2. Cross Validation Results of the IDW-Based Method

4.2.1. Choice of Time Scale, Number of Neighbors, and Exponents

In order to choose an appropriate time scale for the IDW-based method using the extension approach and compare it with the SF-based method, the same four time scales in Table 1 were tested for the IDW-based method.

We evaluated 45 IDW methods with five choices for the number of nearest neighbors N (3, 4, 5, 6 and 7) and nine choices for the exponent p (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5 and 5.0).

4.2.2. Cross Validation and Error Statistics

Similar to evaluating the SF-based method, 10-fold cross validation was implemented to test the time scales, as well as the choices for the number of nearest neighbors N and the exponent p . The optimal average error statistics among the forty-five combinations of N and p are summarized in Table 4 for each chosen time scale, along with the values of N and p when the optimal averages were obtained. Based on Table 4, we choose Scale B as the best of the four time scales for the IDW-based method since it provides the lowest \overline{MARE} , \overline{MSE} , and \overline{RMSE} , as well as the second highest $\overline{R^2_{CV}}$.

Table 4. Error statistics for the PM_{2.5} data set using the IDW-based extension method and 10-fold cross validation. \overline{MAE} (Mean Absolute Error), \overline{MSE} (Mean Squared Error), \overline{RMSE} (Root Mean Squared Error), \overline{MARE} (Mean Absolute Relative Error) and $\overline{R^2_{CV}}$ are the optimal averages of the error statistics.

Error Statistics	Scale A (c = 1)	Scale B (c = 1/10)	Scale C (c = 1/5)	Scale D (c = 1/15)
\overline{MAE}	3.1586 (N = 4, p = 1.0)	3.2856 (N = 3, p = 2.0)	3.1070 (N = 3, p = 2.0)	3.4207 (N = 5, p = 2.5)
\overline{MSE}	75.3792 (N = 7, p = 1)	67.8379 (N = 7, p = 1.5)	68.0293 (N = 6, p = 1.0)	68.2309 (N = 7, p = 1.5)
\overline{RMSE}	8.3258 (N = 7, p = 1.0)	7.8888 (N = 7, p = 1.5)	7.8967 (N = 7, p = 1.0)	7.9143 (N = 7, p = 1.5)
\overline{MARE}	2.7005 (N = 7, p = 1.0)	0.3803 (N = 3, p = 5.0)	0.9717 (N = 3, p = 5.0)	0.3963 (N = 3, p = 2.5)
$\overline{R^2_{CV}}$	0.3789 (N = 7, p = 1.0)	0.4413 (N = 7, p = 1.0)	0.4416 (N = 7, p = 1.0)	0.4374 (N = 7, p = 1.0)

The decision of what values of N and p to use in order to achieve the best IDW interpolations possibly depends on the error statistic deemed most important to optimize. It should be noted

that [31] also discussed the character of the exponent and suggested that the exponent should be deduced from the form of pollution encountered. For air pollution, [31] suggests that elementary reasoning shows that the exponent should be 2 or 3, but more sophisticated considerations could show that the exponent may vary between 1 and 3. For our study, the best exponent could depend on the specific outcome or measure we wanted to model. Hence, we experiment with different exponents p (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5 and 5.0) in order to select the one with the best performance via error analysis. If it were only possible to run an interpolation for one choice of the number of nearest neighbors N and the exponent p (because of time constraints, lack of computational resources, etc.), then the configurations of ($N = 7, p = 1.0$) and ($N = 3, p = 5.0$) seem better than the other configurations that were tested. The configuration of ($N = 7, p = 1.0$) yields the second highest \overline{R}_{CV}^2 among all time scales with a very close result to the highest \overline{R}_{CV}^2 , whereas the configuration of ($N = 3, p = 5.0$) yields the least \overline{MARE} among all time scales. In order to further investigate the difference between configurations of ($N = 7, p = 1.0$) and ($N = 3, p = 5.0$) under *Scale B*, we conducted a further experiment to compare just these two configurations.

The comparison results are shown in the first two columns of Table 5. We consider the configuration of ($N = 3, p = 5.0$) better than ($N = 7, p = 1.0$) under *Scale B* because ($N = 3, p = 5.0$) yields a smaller \overline{MARE} . Similar to the SF-based interpolation method, we removed the same outliers, recomputed the error statistics for the configuration of ($N = 3, p = 5.0$) under *Scale B*, and recorded them in the third column of Table 5. All of the error statistics improved after removing outliers.

Table 5. Error statistics comparison of two configurations under *Scale B* using the IDW-based extension method and 10-fold cross validation.

Error Statistics	$N = 7, p = 1.0$	$N = 3, p = 5.0$	$N = 3, p = 5.0$
	Scale B ($c = 1/10$) before Removing Outliers	Scale B ($c = 1/10$) before Removing Outliers	Scale B ($c = 1/10$) after Removing Outliers
\overline{MAE}	3.4519	3.3378	3.2765
\overline{MSE}	68.0348	79.5497	37.5608
\overline{RMSE}	7.8909	8.6320	6.1287
\overline{MARE}	1.2594	0.3803	0.3773
\overline{R}_{CV}^2	0.4413	0.3359	0.5399

4.3. Comparison of SF-Based and IDW-Based Extension Methods

The first goal of this paper is to compare the performance of the SF-based and IDW-based spatiotemporal interpolation methods in order to find the most suitable method for the $PM_{2.5}$ data. It is evident from the error statistics, as shown in Tables 2 and 3 for the SF-based method and Tables 4 and 5 for the IDW-based method, that *Scale C* using the SF-based method is the best interpolation method among all the methods that we have tested for the $PM_{2.5}$ data set. Both the SF-based and IDW-based methods see improvements in the accuracy of all error statistics, except \overline{MAE} , when choosing a different time scale than *Scale A*, with significant improvement of \overline{MARE} . The SF-based method outperforms the IDW-based method even in the IDW-based method's best scenarios, i.e., the combinations of the number of nearest neighbors and exponents that minimize the relevant error statistics. Therefore, we choose the SF-based extension method using *Scale C* to interpolate the $PM_{2.5}$ data set for population exposure analysis.

In addition to accuracy comparison based on cross validation, the SF-based spatiotemporal interpolation method using the extension approach is computationally efficient because the algorithm is linear according to Equations (3) and (4). On the other hand, the IDW-based method is non-linear according to Equation (6). Therefore, the IDW-based spatiotemporal interpolation method is not as computationally efficient as the SF-based method.

4.4. Population Exposure Analysis

The second goal of this paper is to evaluate the population exposure to fine particulate matter $PM_{2.5}$ in the contiguous United States. Annually updated population data are only available from the five-year American Community Survey at the census block group level. Therefore, we used census block groups in our analysis. The SF-based spatiotemporal interpolation using *Scale C* and the extension approach was implemented to compute a total of 75,784,950 ($207,630 \times 365$) $PM_{2.5}$ values at the centroids of 207,630 census block groups in the contiguous U.S. on each day in 2009. The interpolated census block group-level $PM_{2.5}$ was then linked to 2009 census block group population data.

To analyze the spatial relationship between the $PM_{2.5}$ concentration and the population distribution at the census block group level, we first plot the population distribution in Figure 4a. Second, we plot the annual $PM_{2.5}$ average values in Figure 4b. Several hotspots of high $PM_{2.5}$ values, such as central south California, the Idaho–Montana border, and some regions in Pennsylvania, are distinctively shown in Figure 4b.

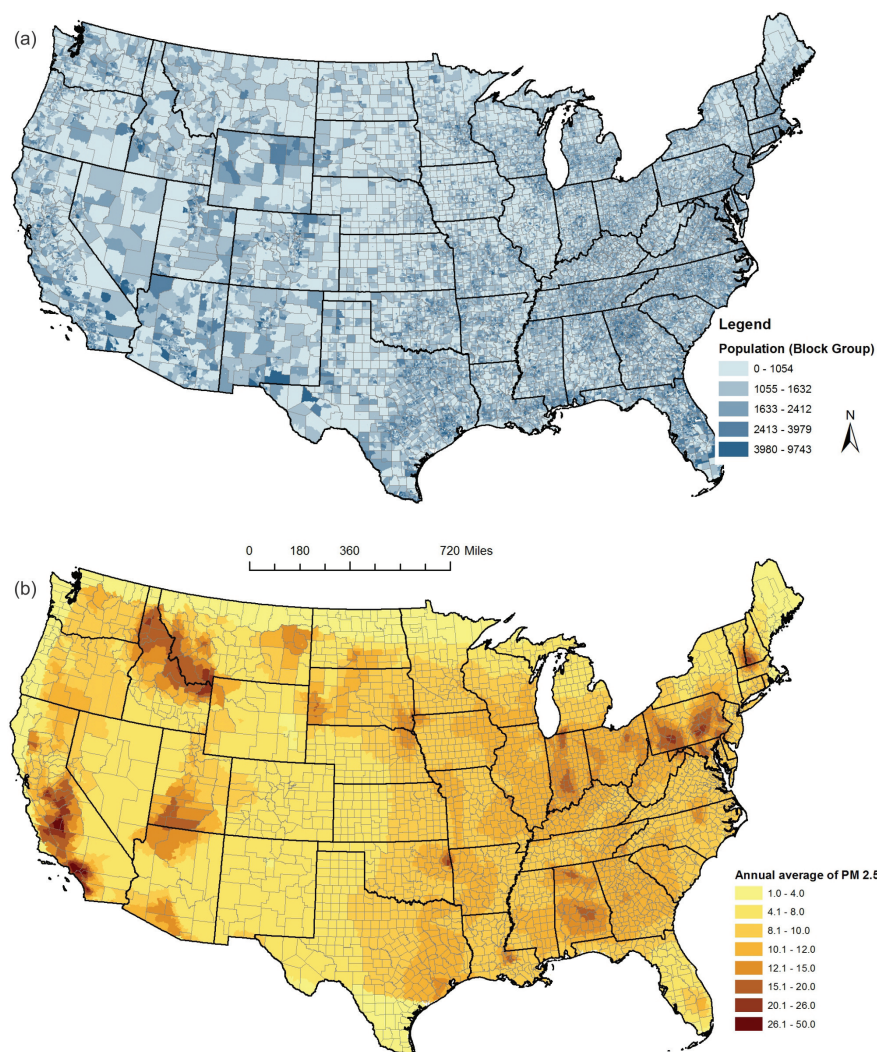


Figure 4. Spatial relationship between the $PM_{2.5}$ concentration and the population distribution across the contiguous United States in 2009. (a) population distribution; (b) annual average $PM_{2.5}$. We used natural breaks to define the color ramps. A lighter color represents a smaller value, while a darker color represents a higher value.

To investigate whether this pattern varies in different seasons, we break the annual average into seasonal averages. Because we used 2009 data, January, February, and December are combined as the winter season. Spring starts from March and ends in May. Summer starts from June and ends in August. The rest of the time is the fall season. Figure 5 shows the seasonal differences. We find that, in spring, the average PM_{2.5} values were high in the west and mountainous areas. The values substantially decreased in summer. In fall, the values increased in the southeast region. Some areas such as central south California had high PM_{2.5} values almost all year around.

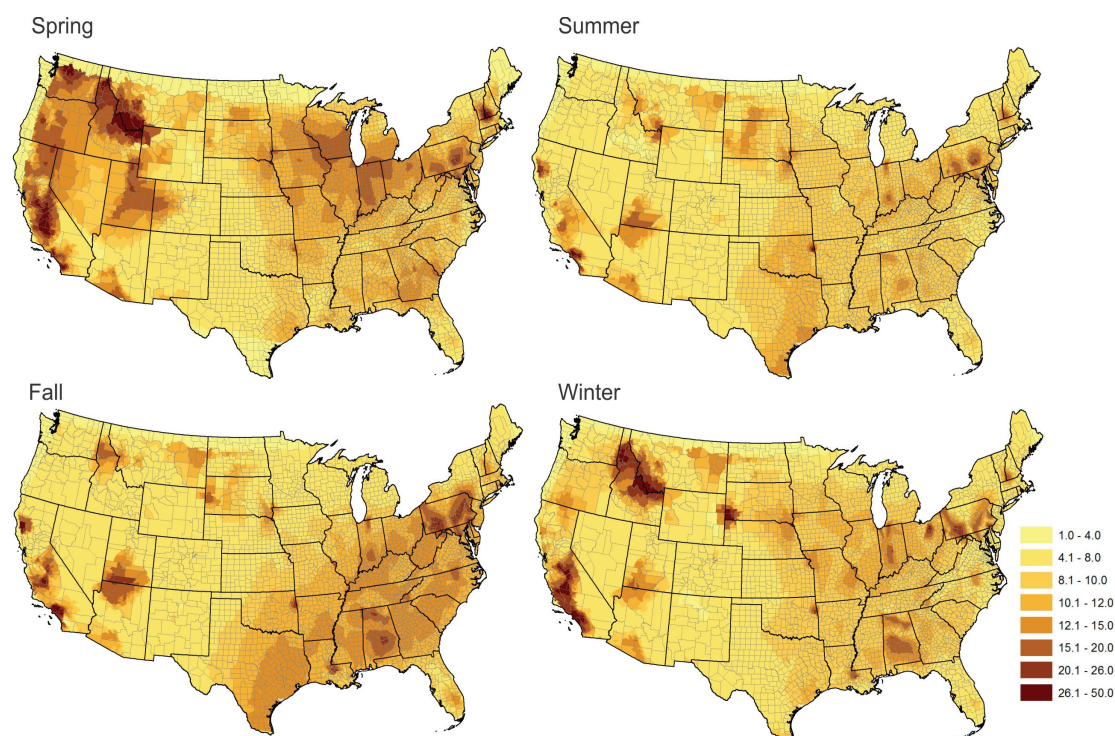


Figure 5. Seasonal differences. The average PM_{2.5} in different seasons across the contiguous United States in 2009. In order to make the color scheme consistent in four seasons, we manually defined the classification scheme. The legend shows the class ranges.

In addition, we observe from Figure 5 that a region near the Idaho–Montana border shows higher PM_{2.5} values during spring and winter than during summer and fall of the year 2009. To verify this pattern, we used the PM_{2.5} Federal Reference Method (FRM)/Federal Equivalent Method (FEM) Mass (88101) daily data (arithmetic mean value) from AirNow to plot the PM_{2.5} values in 2009 at two monitoring stations in Idaho and Montana, as shown in Figure 6. The trends at these two stations are consistent with what we observed in Figure 5. The reason for this pattern remains unclear, despite efforts elucidate its cause. More investigation on the cause for high PM_{2.5} values in this region in 2009 is needed in the future.

Finally, in order to associate the PM_{2.5} values with the national standard, the revised U.S. EPA National Ambient Air Quality Standards for PM_{2.5} in 2006 [51] were adopted in this paper:

- 35 micrograms per cubic meter ($35 \mu\text{g}/\text{m}^3$) for 24 h:
We identify block groups that have PM_{2.5} values greater than $35 \mu\text{g}/\text{m}^3$ for at least one day.
- 15 micrograms per cubic meter ($15 \mu\text{g}/\text{m}^3$) for the annual mean:
We identify block groups that have annual PM_{2.5} values greater than $15 \mu\text{g}/\text{m}^3$.

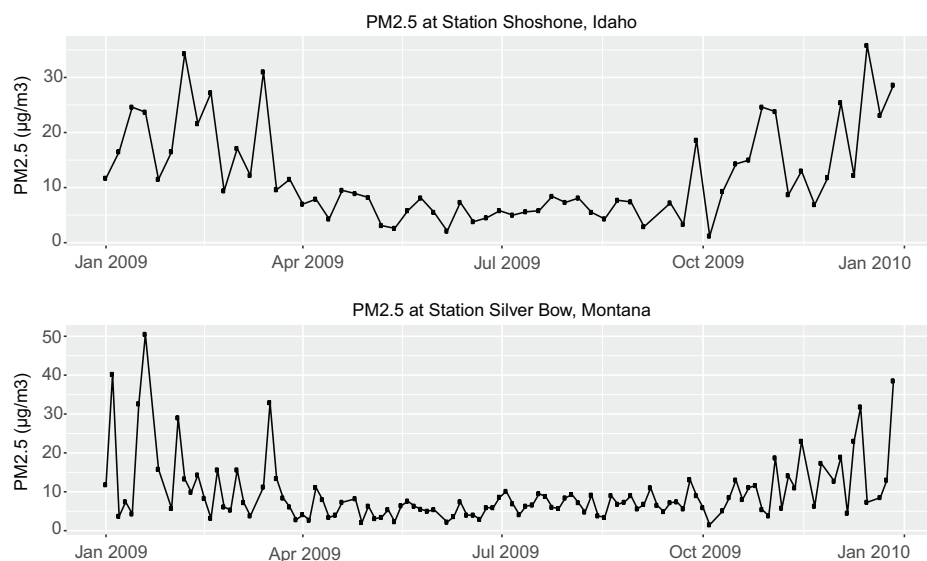


Figure 6. Verification of a spatial pattern in a region near the Idaho–Montana border in 2009. Plots of $PM_{2.5}$ measurements at two monitoring stations in Idaho and Montana using $PM_{2.5}$ daily data from AirNow in 2009.

Figure 7 shows the geographic distribution of such census block groups with the annual and/or 24 h exceeding the U.S. EPA National Ambient Air Quality $PM_{2.5}$ standards. The results suggest:

- there is a population of 27,809,017 (27.8 million) residing in 18,017 census block groups in the contiguous United States with an annual $PM_{2.5}$ exceeding the national standard of $15 \mu g/m^3$;
- more than one-third of the U.S. population (115,310,354) residing in 80,399 census block groups where $PM_{2.5}$ exceeded $35 \mu g/m^3$ for at least one day in 2009.

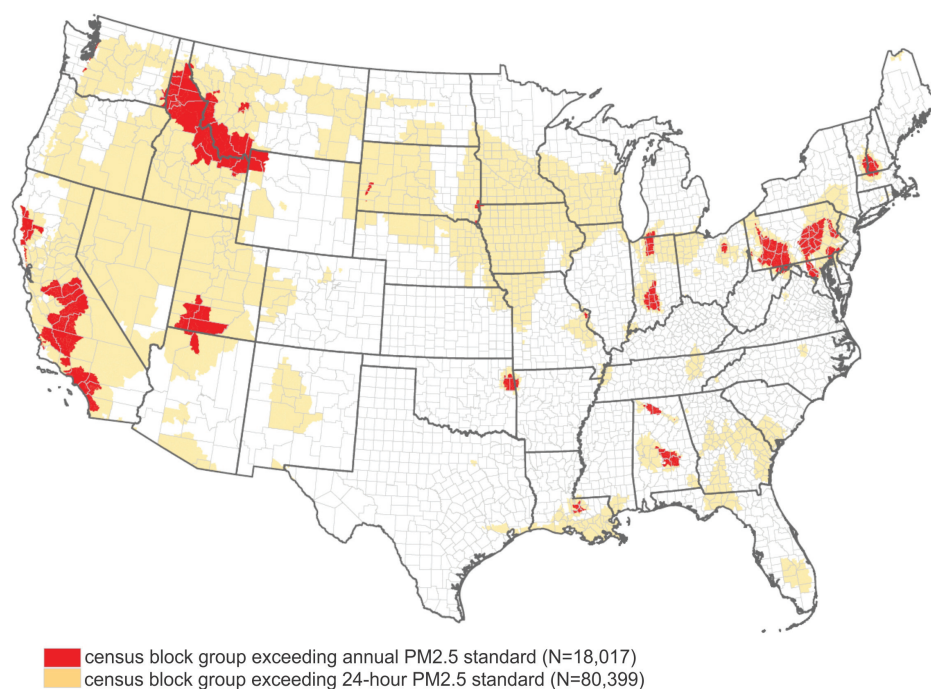


Figure 7. Geographic distribution of census block groups in the contiguous United States that exceeded the $PM_{2.5}$ air quality standards in 2009.

4.5. Web Application

The third goal of this paper is to develop a web application to interpolate and visualize in real time the spatiotemporal variation of ambient air pollution (including but not limited to $PM_{2.5}$) across the contiguous U.S. The web application is based on the MEAN framework [52]. This framework relies on the MongoDB database [53] to store the application's data, Express framework [54] to facilitate HTTP routing, AngularJS [55] to construct an MVC (Model View Controller) architecture to simplify building of responsive web pages, and NodeJS [56] to support the application. The use of MongoDB, Node.js, Express, and AngularJS provides a unified development approach. Each of the technologies is based on JavaScript which allows for more code reuse and less context switching for developers as they move between server side and client side application development.

In addition, a REST (REpresentational State Transfer) [57] Application Program Interface (API) is utilized to handle requests from clients, including user sign up and authentication, requests for interpolated pollution data, and requests for triangulations of measurement sites. A REST call is used to initiate the downloading of pollution data from the *AirNow* [58] File Transfer Protocol (FTP) server and initiate the triangulation and interpolation of the data using the SF-based method. *AirNow* is a U.S. EPA program that provides real-time observed air quality information across the U.S., Canada, and Mexico. It receives real-time air quality observations from over 2,000 monitoring stations and collects forecasts for more than 300 cities. The *AirNow* program includes a web services API for accessing current and historical pollution data [59]. However, queries to this service are generally rate limited to 500 per hour. Therefore, the web application presented in this paper uses an alternative FTP server method to access the *AirNow* data.

This web application uses an SF-based interpolation to compute and update any hour/parameter combination when data has not been updated. Using this method, the system can always include the data for the latest downloaded hour and may include data for previous hours if a time-based interpolation has been calculated. Triangulations are stored in the MongoDB database in a "triangles" collection. When a query is received, the web application can use a geospatial query supported by MongoDB to locate the containing triangle in the triangulation and interpolate the $PM_{2.5}$ concentration.

In order to use the web application, the user needs to sign up by filling out a simple form or log in if they already have an account at the website [60]. After successful log in, the user will see the screen in Figure 8. The screen includes an options menu on the left and an embedded Google Maps application on the right. The Google Maps application is the main panel used for visualization of pollution data, developed using the Google Maps API. When the user changes visualization options in the options menu, such as selecting the pollution parameter type, date, time, or visualization rendering parameters, the data in the Google Maps application will be updated automatically and responsively rendered.

Visualization of the pollution data is rendered on the client side by embedding a Google Maps application within the AngularJS application. Figure 9 shows the interpolated $PM_{2.5}$ concentrations across the contiguous U.S. on 22 March, 2016 at 18:00 GMT. This web application allows a user to visualize six air pollution parameters: O_3 (ppb), $PM_{2.5}$ ($\mu g/m^3$), PM_{10} ($\mu g/m^3$), CO (ppm), SO_2 (ppb), and NO_2 (ppb).

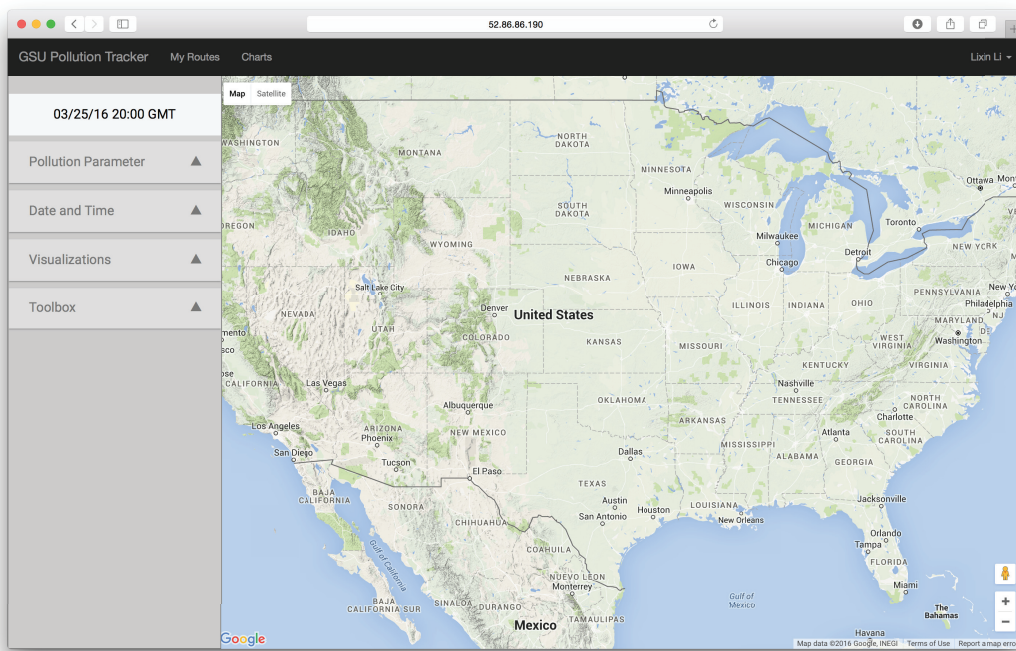


Figure 8. Web application. Map overview screen after logging in.

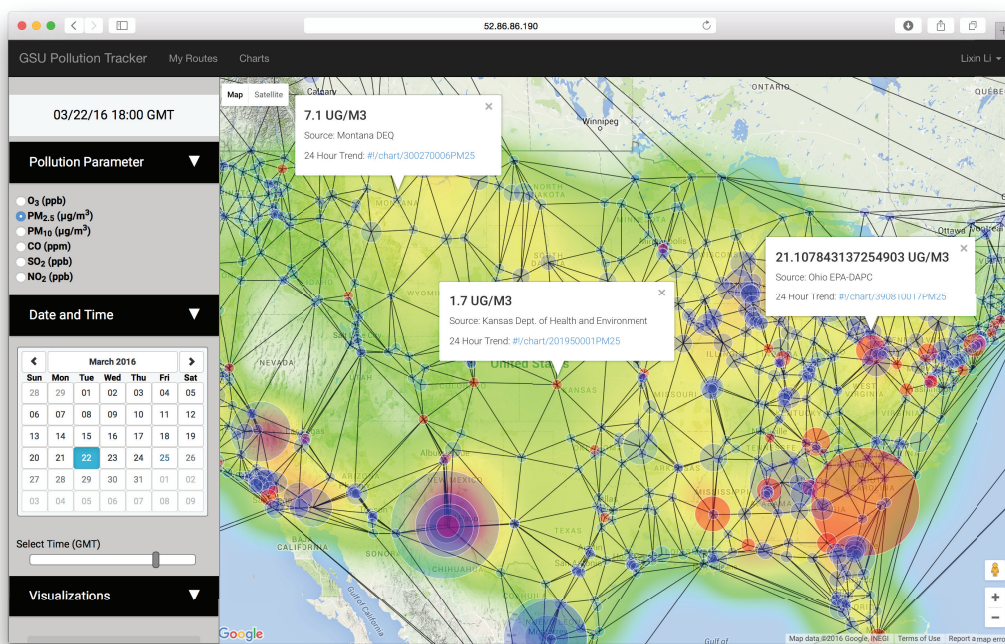


Figure 9. Web application. Rendering of $PM_{2.5}$ concentrations across the contiguous U.S. on 22 March 2016 at 18:00 GMT, including intensities at known measurement sites and the resultant triangulations used in the shape function (SF)-based interpolation method.

5. Discussion

Due to the technological advances and the societal need for analysis of physical phenomena that continuously change in space and time, such as weather and air quality variables, etc., the collection

and processing of spatiotemporal data becomes more and more important. There are significant spatial and temporal dependencies among these data, which are usually ignored or underemphasized by a purely spatial interpolation approach. Investigating the additional temporal information has the potential to improve the interpolation result. Therefore, developing appropriate spatiotemporal interpolations is critical to estimate missing values at points from neighboring observations by looking deep into the spatial and temporal correlations.

This study compares performance of the SF and IDW based spatiotemporal interpolation methods in order to find an interpolation method suitable for an actual set of daily $PM_{2.5}$ values in the contiguous U.S. This paper also explored population exposure to $PM_{2.5}$ in the contiguous U.S. by linking interpolated $PM_{2.5}$ at the centroids of census block groups to census population. Finally, we implemented a web application to interpolate and visualize in real time the spatiotemporal variation of ambient air pollution (including but not limited to $PM_{2.5}$) across the contiguous U.S. using air pollution data from the U.S. EPA's AirNow program. There are some limitations and future work with our study:

- This study is limited to investigating only four choices for time scales, five choices for the number of nearest neighbors, and nine choices for the exponents. In future work, we plan to apply machine learning methods to efficiently learn the best possible configurations in the model, using a lightning-fast cluster computing framework *Apache Spark* [61].
- The SF-based and IDW-based methods are deterministic methods. In this paper, we did not compare our methods with geostatistical interpolation methods such as Kriging, neural networks, and land use regression. In future work, we plan to develop multidimensional and stochastic spatiotemporal interpolation methods suitable for ambient air pollution data (NO_2 , O_3 , $PM_{2.5}$, and PM_{10}) by incorporating factors associated with the environmental exposure of interest, and then make comparisons with other commonly-applied geostatistical interpolation methods.
- Finally, there is a limitation in the currently implemented SF-based algorithm with respect to missing data close to some boundaries of the contiguous United States. For example, along the west coast in Oregon and Washington, there are monitoring stations relatively far away from the coastal border. Because of missing data, an unrealistic stripe next to the coast is visible in our map presentations of the interpolated results. In order to avoid this type of problem, we will need additional measurements along the coast, or use meshless interpolation methods such as IDW with a limited number of neighboring measurements in future work.

Additionally, in future work, we plan to link interpolated air pollution concentration values to a variety of health outcomes to evaluate air pollution's adverse impact on human health, as well as link the interpolated pollution values with individual GPS trajectory to better estimate personal-based air pollution exposure.

6. Conclusions

In conclusion, this study has made three contributions to the ambient air pollution and spatiotemporal interpolation research community.

First, using an actual set of daily $PM_{2.5}$ values measured by U.S. EPA monitoring sites in the contiguous United States, the performance of the SF and IDW based spatiotemporal interpolation methods is compared in order to find an interpolation method suitable for the $PM_{2.5}$ data. The SF-based interpolation method performed better overall than the IDW-based method for the daily $PM_{2.5}$ data.

Second, more than 75 million $PM_{2.5}$ spatiotemporal interpolation results are calculated using the SF-based spatiotemporal method in the contiguous U.S. at the fine geographic level of census block groups. The interpolation results are linked to 2009 census block group population data so that the population with unhealthy $PM_{2.5}$ exposure in the contiguous U.S is estimated. To map $PM_{2.5}$ and

population spatial distribution, we generated choropleth maps based on the interpolated values as well as the population data. Because PM_{2.5} may exhibit different spatial patterns in different seasons, we also investigated the seasonal variations. We conducted spatial queries to identify areas where annual and daily PM_{2.5} are above the standard and calculated the population size in these areas.

Third, this study implemented a web application to interpolate and visualize in real time the spatiotemporal variation of ambient air pollution (including but not limited to PM_{2.5}) across the contiguous U.S. using air pollution data from the U.S. EPA's AirNow program.

Acknowledgments: The authors of this paper would like to thank Xingyou Zhang for providing valuable input on population exposure analysis for this study. The authors would also like to thank Jason Franklin for his generous help on programming, generating IDW-related error statistics, and proofreading. Finally, the authors would like to thank Weitian Tong for giving his insight into error statistics and cross validation.

Author Contributions: Lixin Li and Xiaolu Zhou contributed to the design and coordination of the study. Reinhard Piltner contributed to the shape function-based interpolation methods. Marc Kalo implemented the web application. All authors collaboratively drafted and edited the manuscript, and approved the final version of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. EPA. Particulate Matter (PM). Available online: <https://www.epa.gov/pm-pollution> (accessed on 26 March 2016).
2. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2016.
3. Sloane, C.S.; Watson, J.; Chow, J.; Pritchett, L.; Willard Richards, L. Size-segregated fine particle measurements by chemical species and their impact on visibility impairment in Denver. *Atmos. Environ.* **1991**, *25*, 1013–1024.
4. Ghim, Y.S.; Moon, K.C.; Lee, S.; Kim, Y.P. Visibility Trends in Korea during the Past Two Decades. *J. Air Waste Manag. Assoc.* **2005**, *55*, 73–82.
5. Hong, Y.C.; Lee, J.T.; Kim, H.; Ha, E.H.; Schwartz, J.; Christiani, D.C. Effects of air pollutants on acute stroke mortality. *Environ. Health Perspect.* **2002**, *110*, 187–191.
6. Laden, F.; Neas, L.M.; Dockery, D.W.; Schwartz, J. Association of fine particulate matter from different sources with daily mortality in six U.S. cities. *Environ. Health Perspect.* **2000**, *108*, 941–947.
7. Robichaud, A.; Ménard, R. Multi-year objective analyses of warm season ground-level ozone and PM_{2.5} over North America using real-time observations and Canadian operational air quality models. *Atmos. Chem. Phys.* **2014**, *14*, 1769–1800.
8. Shepard, D. A two-dimensional interpolation function for irregularly spaced data. In *Proceedings of the 23rd National Conference ACM*; New York, NY, USA, 1968; pp. 517–524.
9. Li, L.; Revesz, P. Interpolation methods for spatio-temporal geographic data. *J. Comput. Environ. Urban Syst.* **2004**, *28*, 201–227.
10. Zienkiewicz, O.C.; Taylor, R.L. *Finite Element Method*; Butterworth Heinemann: London, UK, 2000; Volume 1.
11. Franke, C.; Schaback, R. Solving Partial Differential Equations by Collocation using Radial Basis Functions. *Appl. Math. Comput.* **1998**, *93*, 73–82.
12. De Boor, C. *A Practical Guide to Splines*; Springer: New York, NY, USA, 2001; Volume 27.
13. Sibson, R. A brief description of natural neighbor interpolation. In *Interpreting Multivariate Data*; Barnett, V., Ed.; John Wiley & Sons, Inc.: New York, NY, USA, 1981; Chapter 2, pp. 21–36.
14. Zurflueh, E.G. Applications of Two-dimensional Linear Wavelength Filtering. *Geophysics* **1967**, *32*, 1015–1035.
15. Krige, D. Two dimensional weighted moving average trend surfaces for ore evaluation. *J. Soc. Afr. Inst. Min. Metall.* **1966**, *66*, 13–38.
16. Blond, N.; Vautard, R. Three-dimensional ozone analyses and their use for short-term ozone forecasts. *J. Geophys. Res. Atmos.* **2004**, doi:10.1029/2004JD004515.

17. Pagowski, M.; Grell, G.A.; McKeen, S.A.; Peckham, S.E.; Devenyi, D. Three-dimensional variational data assimilation of ozone and fine particulate matter observations: some results using the Weather Research and Forecasting—Chemistry model and Grid-point Statistical Interpolation. *Q. J. R. Meteorol. Soc.* **2010**, *136*, 2013–2024.
18. Robichaud, A.; Ménard, R.; Zaitseva, Y.; Anselmo, D. Multi-pollutant surface objective analyses and mapping of air quality health index over North America. *Air Qual. Atmos. Health* **2016**, doi: 10.1007/s11869-015-0385-9.
19. Liao, D.; Pequet, D.J.; Duan, Y.; Whitsel, E.A.; Dou, J.; Smith, R.L.; Lin, H.M.; Chen, J.C.; Heiss, G. GIS Approaches for the Estimation of Residential-Level Ambient PM Concentrations. *Environ. Health Perspect.* **2006**, *114*, 1374–1380.
20. Zou, B.; Wilson, J.G.; Zhan, F.B.; Zeng, Y. Air pollution exposure assessment methods utilized in epidemiological studies. *J. Environ. Monit.* **2011**, *11*, 475–490.
21. Cressie, N.; Wikle, C.K. *Statistics for Spatio-temporal Data*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2011.
22. Pebesma, E. Spacetime: Spatio-temporal data in R. *J. Stat. Softw.* **2012**, *51*, 1–30.
23. Li, L.; Losser, T.; Yorke, C.; Piltner, R. Fast Inverse Distance Weighting-based Spatiotemporal Interpolation: A Web-based Application of Interpolating Daily Fine Particulate Matter PM_{2.5} in the Contiguous U.S. using Parallel Programming and k-d Tree. *Int. J. Environ. Res. Public Health* **2014**, *11*, 9101–9141.
24. Losser, T.; Li, L.; Piltner, R. A Spatiotemporal Interpolation Method Using Radial Basis Functions for Geospatiotemporal Big Data. In Proceedings of the 5th International Conference on Computing for Geospatial Research and Application, Washington, DC, USA, 4–6 August 2014; IEEE: Washington, DC, USA; pp. 17–24.
25. Revesz, P.; Wu, S. Spatiotemporal reasoning about epidemiological data. *Artif. Intell. Med.* **2006**, *38*, 157–170.
26. Anderson, S.; Revesz, P.Z. Efficient MaxCount and threshold operators of moving objects. *Geoinformatica* **2009**, *13*, 355–396.
27. Hussain, I.; Spöck, G.; Pilz, J.; Yu, H.L. Spatio-temporal interpolation of precipitation during monsoon periods in Pakistan. *Adv. Water Resour.* **2010**, *33*, 880–886.
28. Yu, H.L.; Wang, C.H. Quantile-Based Bayesian Maximum Entropy Approach for Spatiotemporal Modeling of Ambient Air Quality Levels. *Environ. Sci. Technol.* **2013**, *47*, 1416–1424.
29. Lu, G.Y.; Wong, D.W. An adaptive inverse-distance weighting spatial interpolation technique. *Comput. Geosci.* **2008**, *34*, 1044–1055.
30. Li, J.; Heap, A.D. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecol. Inform.* **2011**, *6*, 228–241.
31. De Mesnard, L. Pollution Models and Inverse Distance Weighting: Some Critical Remarks. *Comput. Geosci.* **2013**, *52*, 459–469.
32. Li, L.; Zhang, X.; Piltner, R. An Application of the Shape Function Based Spatiotemporal Interpolation Method on Ozone and Population Exposure in the Contiguous U.S. *J. Environ. Inform.* **2008**, *12*, 120–128.
33. Buchanan, G.R. *Finite Element Analysis*; McGraw-Hill: New York, NY, USA, 1995.
34. Revesz, P.; Li, L. Representation and Querying of Interpolation Data in Constraint Databases. In Proceedings of the Third National Conference on Digital Government Research, 19–22 May 2002; pp. 225–228.
35. Li, J.; Narayanan, R. A shape-based approach to change detection of lakes using time series remote sensing images. *IEEE Trans. Geosci. Remot. Sen.* **2003**, *41*, 2466–2477.
36. Gao, J.; Revesz, P. Voting prediction using new spatiotemporal interpolation methods. In Proceedings of the Seventh Annual International Conference on Digital Government Research, San Diego, CA, USA, 21–24 May 2006; pp. 293–300.
37. Li, L. *Spatiotemporal Interpolation Methods in GIS - Exploring Data for Decision Making*; VDM Verlag: Saarbrücken, Germany, 2009.
38. Revesz, P. *Introduction to Databases: From Biological to Spatio-Temporal*; Springer: New York, NY, USA, 2010.
39. Murphy, R.; Curriero, F.; Ball, W. Comparison of Spatial Interpolation Methods for Water Quality Evaluation in the Chesapeake Bay. *J. Environ. Eng.* **2010**, *136*, 160–171.

40. Rahman, H.; Alireza, K.; Reza, G. Application of Artificial Neural Network, Kriging, and Inverse Distance Weighting Models for Estimation of Scour Depth around Bridge Pier with Bed Sill. *J. Softw. Eng. Appl.* **2010**, *3*, 944–964.
41. Eldrandaly, K.A.; Abu-Zaid, M.S. Comparison of Six GIS-Based Spatial Interpolation Methods for Estimating Air Temperature in Western Saudi Arabia. *J. Environ. Inform.* **2011**, *18*, 38–45.
42. Xie, Y.; Chen, T.; Lei, M.; Yang, J.; Guo, Q.; Song, B.; Zhou, X. Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis. *Chemosphere* **2011**, *82*, 468–476.
43. Li, L.; Zhang, X.; Holt, J.B.; Tian, J.; Piltner, R. Estimating Population Exposure to Fine Particulate Matter in the Conterminous U.S. using Shape Function-based Spatiotemporal Interpolation Method: A County Level Analysis. *GSTF Int. J. Comput.* **2012**, *1*, 24–30.
44. Li, L.; Zhang, X.; Piltner, R. A Spatiotemporal Database for Ozone in the Conterminous U.S. In Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning, Budapest, Hungary, 15–17 June 2006; pp. 168–176.
45. Li, L. Constraint Databases and Data Interpolation. In *Encyclopedia of Geographic Information System*; Shekhar, S., Xiong, H., Eds.; Springer: New York, NY, USA, 2008; pp. 144–153.
46. Tobler, W.R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, *46*, 234–240.
47. Johnston, K.; Hoef, J.M.V.; Krivoruchko, K.; Lucas, N. *Using ArcGIS Geostatistical Analyst*; ESRI Press: Redlands, CA, USA, 2001.
48. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross Validation. In *Encyclopedia of Database Systems*; Özsu, M.T., Liu, L., Eds.; Springer: New York, NY, USA, 2009; pp. 532–538.
49. Keller, J.P.; Olives, C.; Kim, S.Y.; Sheppard, L.; Sampson, P.D.; Szpiro, A.A.; Oron, A.P.; Lindström, J.; Vedal, S.; Kaufman, J.D. A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environ. Health Perspect.* **2015**, *123*, 301–309.
50. EPA. Air Quality System (AQS). Available online: <http://www3.epa.gov/pm> (accessed on 26 March 2016).
51. EPA. Particulate Matter (PM) Standards – Table of Historical PM NAAQS. Available online: http://www3.epa.gov/ttn/naaqs/standards/pm/s_pm_history.html (accessed on 26 March 2016).
52. MEAN.IO. MEAN. Available online: <http://mean.io> (accessed on 26 March 2016).
53. MongoDB. Available online: <https://www.mongodb.org> (accessed on 26 March 2016).
54. Express. Available online: <http://expressjs.com> (accessed on 26 March 2016).
55. AngularJS. Available online: <https://angularjs.org> (accessed on 26 March 2016).
56. NodeJS. Available online: <https://nodejs.org> (accessed on 26 March 2016).
57. Zur Muehlen, M.; Nickerson, J.V.; Swenson, K.D. Developing web services choreography standards—the case of REST vs. SOAP. *Decis. Support Syst.* **2005**, *40*, 9–29.
58. EPA. AirNow. Available online: <http://www.airnow.gov> (accessed on 26 March 2016).
59. AirNow Web Services Documentation. Available online: <http://airnowapi.org/webservices> (accessed on 26 March 2016).
60. A Real-Time Web Application to Interpolate and Visualize Spatiotemporal Variation of Ambient Air Pollution Across the Contiguous U.S. Available online: <http://52.86.86.190:3000> (accessed on 26 March 2016).
61. Apache Spark. Available online: <https://databricks.com/spark> (accessed on 26 March 2016).

