Fusing surface and satellite-derived PM observations to determine the impact of international transport on coastal $PM_{2.5}$ concentrations in the western U.S.

Neha Bora[1], Tuo Chen[2], Dana Cochran[3], Kelly Dougan[4], Gautam Sabnis[5] Chuanping Yu [6]

Faculty Mentors: Alen Alexanderian[7], Arvind Saibaba[8], Elizabeth Mannshardt[9], Brett Grant[10] Jessica Matthews[11]

Abstract

Long term exposure to $PM_{2.5}$ is associated with human health complications. Surface readings of $PM_{2.5}$ in the states on the West Coast of the United States have reported to be higher than allowed by the Clean Air Act. One possible reason for this is international transport of air pollution on $PM_{2.5}$. This project explores the relationship between the surface readings of $PM_{2.5}$ from coastal sites with the AOD measurements from the AVHRR in these regions. Once we found a correlation between these readings, we implemented a model to approximate $PM_{2.5}$ concentrations in the Pacific ocean, where surface readings are not possible. We model the variability in $PM_{2.5}$ concentrations using a Generalized Additive Model (GAM)

1 Introduction

Domestic sources of emissions are the primary cause of air pollution in the US. However, the international flow of air pollution into the US could potentially be a contributing factor in some coastal cities with high measurements of air pollution. The impact of international transport of air pollution on our ability to attain air quality standards or other environmental objectives in the US has yet to be fully understood. In other words, cities in the western US may unknowingly be receiving high pollutant values from external sources. To answer such questions, we need a comprehensive understanding of the transport in the atmosphere.

We focus on two different measures of pollution, namely $PM_{2.5}$ and Aerosol Optical Depth (AOD). $PM_{2.5}$ is particulate matter that is less than 2.5 micrometers in diameter and is often referred to as the greatest health risk of pollutants [?]. For this project, we used two different sources of data: AOD measurements obtained from the Advanced Very High Resolution Radiometer (AVHRR) satellite measurements, and surface $PM_{2.5}$

[1]Department of Mathematics, Iowa State University
[2]Department of Statistics, University of Florida
[3]Department of Mathematics, California State University Channel Islands
[4]Department of Mathematics, The State University of New York at Buffalo
[5]Department of Statistics, Florida State University
[6]Industrial and System Engineering, Georgia Institute of Technology
[7]North Carolina State University
[8]North Carolina State University
[9]Environmental Protection Agency
[10]Environmental Protection Agency
[11]National Oceanic and Atmospheric Administration

measurements. AOD measurements refer to a quantitative measure of the amount of light that is obstructed by particles in the atmospheric column and are also referred to as aerosol optical thickness (AOT). Because AOD measures any particle from the satellite to the earth's surface, the measurements of the amount of the specific pollutant $PM_{2.5}$ are, for the most part, measured by ground sites.

Previous work Several previous studies have attempted to find correlations between $PM_{2.5}$ readings and AOD readings. Liu et al looked into estimating $PM_{2.5}$ concentrations using the satellite AOD data, meteorology, and land use information in states surrounding Massachusetts [**?**]. In their study, they were able to predict $PM_{2.5}$ values better with an AOD model than with a non-AOD model. Lee et al developed a novel approach in which he used a mixed effects model to predict day-specific $PM_{2.5}$ concentrations based on AOD measurements [**?**]. Li et al [**?**] investigated if variability of AOD measurements can be used to infer space-time variability of PM readings. Their model resulted in good spatial agreement in the eastern region but not the central or western regions of the US. Overall, they concluded that the relationship between PM and AOD varies over different locations and times, and that a better prediction model would be one that focuses on a smaller region or time frame. Van Donkelaar et al (2015) looked for global trends of $PM_{2.5}$ concentrations from satellite data. Using a decadal mean over the years 2001-2010, in North America, they found a relatively higher concentration of PM in the east coast and in the San Joaquin valley of California. In Asia, they found extremely high concentration, over 60 μ g / m^3 and 80 μ g / m^3 in Northern India and Eastern Asia, respectively. They found that the population-weighted concentrations in East Asia nearly doubled the global mean. These studies suggest that AOD-based models can be used to predict PM concentrations on a daily basis, using a statistical model interpolating over a sufficiently small space and time scale.

Overview of main results The goal in this project is to establish a relationship between AOD measurements from the AVHRR satellite and $PM_{2.5}$ measurements. We focus our analysis on the coasts of California and Hawaii since at these sites there is a significant overlap between these two datasets. We aim to use AOD-$PM_{2.5}$ relationship to predict $PM_{2.5}$ concentrations over the Pacific Ocean. In addition, we analyze time series models of surface $PM_{2.5}$ measurements at each site. Spatial interpolation is also used on all $PM_{2.5}$ sites in California to help determine a trend in the readings. These images were used to visualize high emission events in the state such as wildfires, dust storms, etc. This analysis makes a significant progress in understanding the impact of international transport of air pollution on $PM_{2.5}$ concentrations in coastal areas of the western US.

2 The Problem

2.1 Description of data

The first dataset of Climate Data Record (CDR) of AOD was obtained from the National Oceanic and Atmospheric Administration (NOAA) [?]. Kelly: what does this next sentence mean??? This data was collected using AVHRR, which is an optical measure of aerosol column loading derived from the global ocean pixel-level PATMOS-x AVHRR clear-sky reflectance CDR at 0.63 μm channel [?]. The second data set, provided by EPA, contained the surface $PM_{2.5}$ measured in California, Oregon, Washington, Alaska, and Hawaii. [?].

The AVHRR takes approximately 16 days to make one revolution around the earth. We, thus, have roughly two data values for each month of the year at each pair of coordinates, for the years 1981-2009. The frequency of the $PM_{2.5}$ data collected from each site varies from once every day to once every six days. On certain occassions, the measurements from the satellite were found to be erroneous due to light reflection from cloud covers. Additionally, there are times when the $PM_{2.5}$ sensors malfunctioned resulting in no data. These points were appropriately removed from the datasets.

Additionally we have satellite data for wind speed, wind direction, air temperature, relative humidity, and height of the planetary boundary layer. This data was incorporated in our models to give more accurate results. The wind data was provided into a u wind and v wind format, where u represents wind blowing towards the east, and v represents wind blowing toward the north. The u wind needed to be scaled by a factor of 0.003052037, as indicated in the file information. The formulas to get the wind speed and the wind direction are below, where u_{wind} and v_{wind} represent the u and v values. [?]

$$\begin{aligned}\text{windspeed} &= \sqrt{u_{\text{wind}}^2 + v_{\text{wind}}^2},\\ \text{winddirection} &= \frac{180}{\pi}\arctan(-u_{\text{wind}}, -v_{\text{wind}}) + 180.\end{aligned}$$

2.2 Challenges

One of the challenges with the given data set was its size. For example, the file size for one year of AOD data was 9.47 GB, which exceeded the memory of our systems. Instead of using AOD data for one year in one file, we dowloaded daily AOD data. We restricted our AOD data set to the years 2006-2009, where we had access to the files in daily format.

The next challenge was with the format of the latitude and longitude of wind, air temperature, relative humidity, and height of the planetary boundary layer data. The data covered North America and was in the Lambert Conformal Conic map projection. However, the $PM_{2.5}$ and AOD data are in geographical coordinates. The geographical coordinates of the $PM_{2.5}$ sites were converted to Lambert Conformal Conic coordinates in order to extract the values of these meteorological variables.

The $PM_{2.5}$ sensor readings had some challenges as well. The main issue was that not all sensor sites pick up measurements on the same day. Thus, if we wanted to look at sensor readings on say Jan 1, 2008, we may only have 10 sites that produce measurements, but if we took a look at Jan 2, 2008, we may have 40 sites that produce measurements. Alen: Consider removing the next sentence. This proved to be a little challenging when it came time to analyze sensor readings for certain dates. Because of this, on many days, we had insufficient data to construct meaningful spatial interpolations.

3 The Approach

3.1 Data processing

No matter what $PM_{2.5}$ sites we consider using, we need to find the closest AOD coordinates to the $PM_{2.5}$ sites. In addition, we need to find dates when both types of data are available. This requires searching all the observations in the AOD dataset. The AOD yearly dataset is very large, and we cannot put all the data in hundreds of files into one matrix. Alen: Concerning this last sentence, I think this point has been made before. Additionally, AOD dataset has a lot of missing data, thus we need to clean the AOD data first. Since the data we want to use is located on the west coast, we used the longitude and latitude of the west coast to eliminate data from other locations that we will not use. We then dropped all the missing data and changed the original format of the data into more readable one. Table 1 visualizes our desired AOD data format.

Date	lat_aod	long_aod	aod
2000-01-25	30	-126.6	-0.0836133733391762
...	...	...	...

Table 1: AOD data. Alen: Do you need the second row with all the $\cdots$?

3.2 $PM_{2.5}$ sites adjacent to AVHRR grids

Since the coverage of the AOD data is over the oceans and the $PM_{2.5}$ data is collected over the land, there is no direct overlap between the two datasets. In order to compare the two data sets, we adopt the following strategy. First, of all the $PM_{2.5}$ measurement sites, we identify those that are close to the coast. Figure 1 shows the geographical locations to 13 locations, most of which lie in California. In the rest of this report, we focus on only 2 site locations. Future work will include examining the additionally 11 locations.

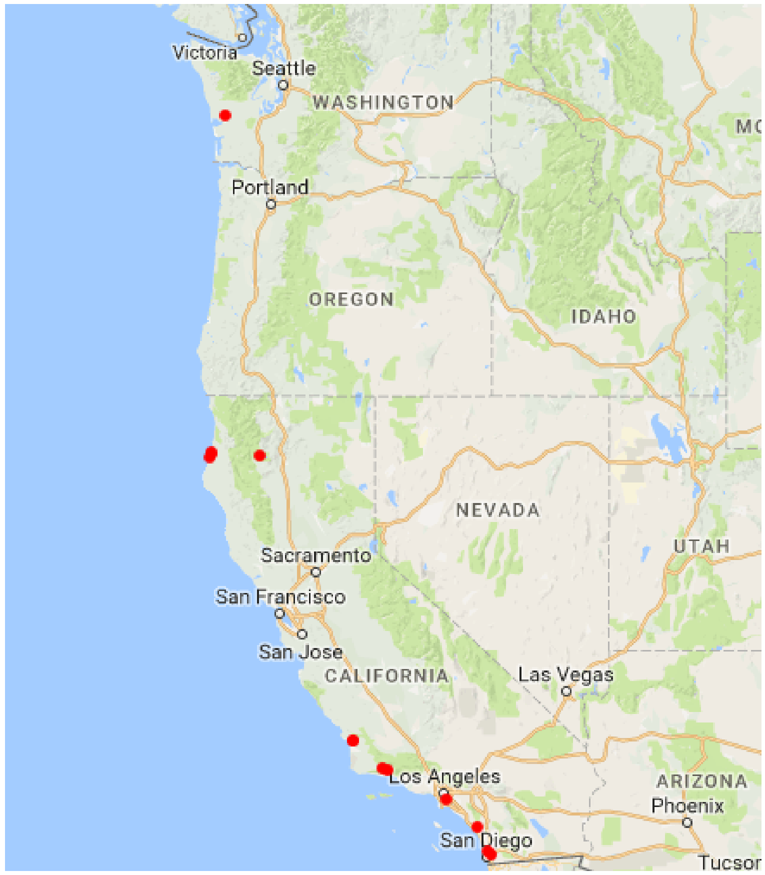


Figure 1: 13 Sensor cites Along West Coast.

3.3 Analyzing $PM_{2.5}$ trends

One method we used to analyze trends in the $PM_{2.5}$ data was to make an animation to show the change in the concentration of $PM_{2.5}$ over time. The animation plots points at each site where the $PM_{2.5}$ data was collected, with color varying depending on the intensity of the reading, with darker colors indicating a larger concentration of $PM_{2.5}$. Alen: Need to either show some snapshots from the animation, or at least point to a link where the animation can be downloaded.

To understand whether we can use measured $PM_{2.5}$ to predict future $PM_{2.5}$ values, we conducted a time series analysis. We focus our attention on the $PM_{2.5}$ site near Long Beach, California, which is located at latitude 33.79236 and longitude -118.175. We collect monthly averaged $PM_{2.5}$ values from January 2004 to December 2014, which are plotted in Figure 2.

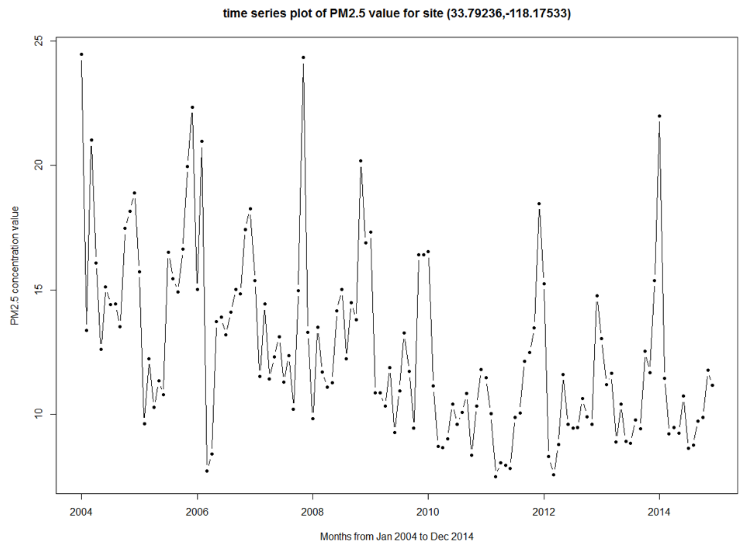


Figure 2: Time series plot of $PM_{2.5}$ value for site (33.79236, -118.17533).

Figure 2 suggests that there is a slightly downward trend over time and the time series seems to exhibit seasonality. We decided to use two different methods to fit the time series. One is Holt-Winters Exponential Smoothing. Exponential smoothing is a very popular scheme to produce a smoothed time series and it assigns exponentially decreasing weights as the observations get older [?]. Holt-Winters Exponential Smoothing can be used to make short-term forecasts on a time series that can be described using an additive model with increasing or decreasing trend and seasonality [?]. The second method used is the Autoregressive Integrated Moving Average, (ARIMA). This model can be fitted to time series data either to better understand the data or to predict future points in the series (wikipedia website) Alen: Fix the reference; also, if possible refer to something other than Wikipedia, you can probably use one of the references Wikipedia lists.. For the ARIMA method, we first adjust the time series by subtracting the estimated seasonal component and then apply the model to the adjusted time series. Alen: Mention the section in which the results of the analysis is provided.

3.4 Relationships between AVHRR AOD and surface $PM_{2.5}$

We noticed that there was a huge difference between the two sites that we identified along the West Coast, and we decided to build models for the sites along the West Coast. We also noticed that there were different relationships between AOD and PM in each site that can be illustrated in Figure 3. So we analyzed the relation for each site separately.

3.4.1 Only AOD data

Since there are many meteorological parameters varying from day to day, our statistical model must have the variability of the date. For each location, there are many different geographical properties, so our model must

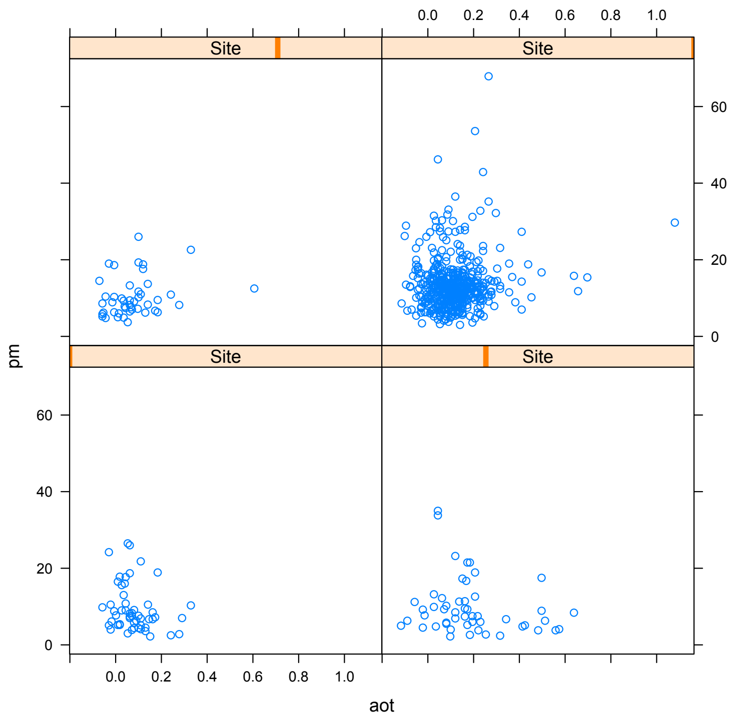


Figure 3: PM vs. AOD for west coast sites

have the variability of sites. Therefore, we used a mixed effects model to fit this relationship:

$$PM_{ij} = \alpha + \beta \times AOD_{ij} + s_i + d_j + \epsilon_{ij},$$

where PM_{ij} is the $PM_{2.5}$ concentration at a spatial site i on a specific day j, α is the fixed intercept, β is the fixed slope, AOD_{ij} is the AOD value at a spatial site i on a specific day j, $s_i \sim N(0, \sigma_s^2)$ is the random intercept of site i, $d_j \sim N(0, \sigma_d^2)$ is the random intercept of a specific day j, and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the error term at site i on a day j.

3.4.2 AOD data and wind data

As we have more information about the time-varying parameters such as air temperature, humidity, etc, it is not reasonable to simply treat them as random variables.

First, we tried the multivariate linear regression model to fit the relation. Since the result was not good enough Alen: not good enough is very vague, please be more specific., we tried another kind of model, the GAM model, which turned out to be a better model to fit the relationship.

3.4.3 Match the AOD data with the $PM_{2.5}$ data

As we are trying to find the relationships between AOD and $PM_{2.5}$, we need to keep all the date the same, and locations closest. After matching these two datasets, we got the following as in table 2. Alen: Do you need this subsection?

Date	lat_pm	long_pm	pm	lat_aot	long_aot	aot
2006-12-28	40.776944	-124.1775	17.8	40.9	-124.3	0.043815478682518
...	...	...	...	...	...	...

Table 2: AOD and $PM_{2.5}$ match data

4 Computational Experiments

4.1 Experiment 1: Analysis of PM25 time series

4.1.1 Holt-Winters Exponential Smoothing

We firstly draw fitted values plot based on Holt-Winters Exponential Smoothing. In Figure 4, black line represents the variation of real PM values during 2004-2014 and red line represents the variation of fitted PM values during the period. We can see that although the two lines have similar trends, the fitted line is not accurate enough at many points.

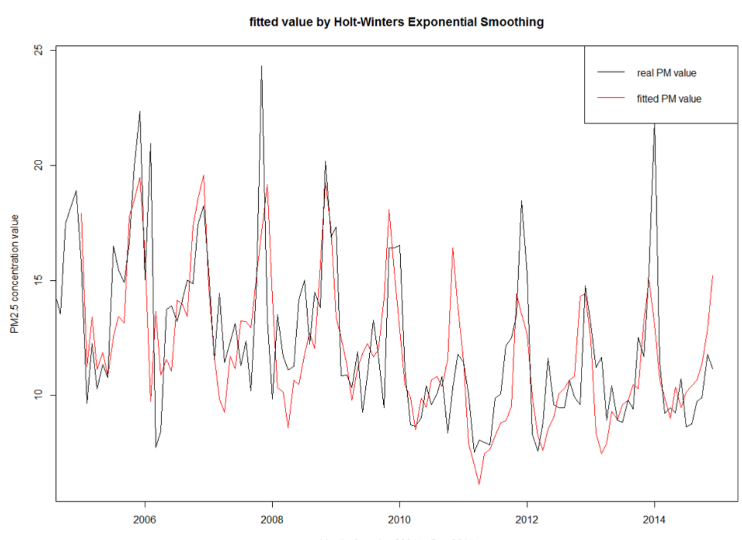


Figure 4: Fitted value by Holt-Winters Exponential Smoothing

Then we predict $PM_{2.5}$ values for the whole year 2015. In Figure 5, the blue line represents the predicted values for 2015. The shadow of deep color represents the 80% prediction interval for the predicted values and the shadow of light color represents the 95% prediction interval for the predicted values.

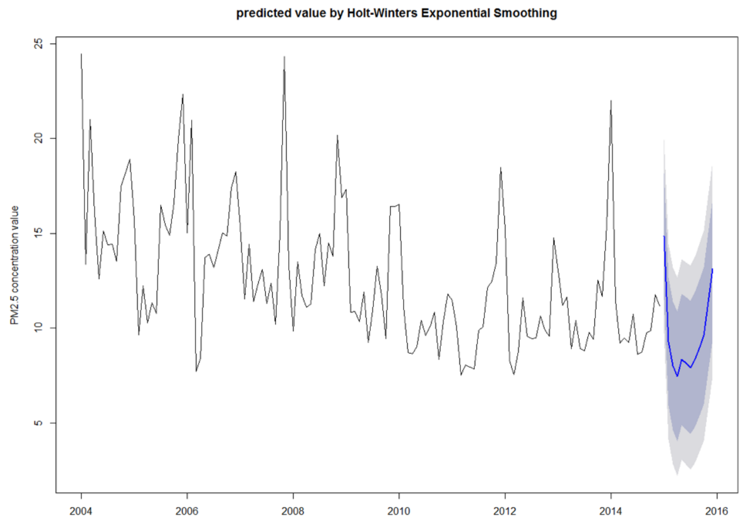


Figure 5: Predicted value by Holt-Winters Exponential Smoothin

Now let us exmaine the effect of our prediction. We decide to use mean square error as our criterion. Lower mean square error indicates we have a better prediction. By comparing the predicted results and the real PM values for the year 2015, we find the mean square error between them is around 5.5. Becaues the mean of $PM_{2.5}$ values during 2004-2014 is about 12.6, the mean square error is kind of large and this model is not very accurate.

4.1.2 ARIMA model

Becaues there seems to exist seasonality in the $PM_{2.5}$ time series, we want to decompose the time series and apply ARIMA model to the adjusted time series. We firstly decompose the $PM_{2.5}$ time series and plot it. In Figure 6, the four subplots respectively represent observed values, overall trend component, seasonal component and random part. From the seasonal component, we can see that there does exist differences among $PM_{2.5}$ values in different months. By subtracting the seasonal component from the original time series, we get adjusted $PM_{2.5}$ time series.

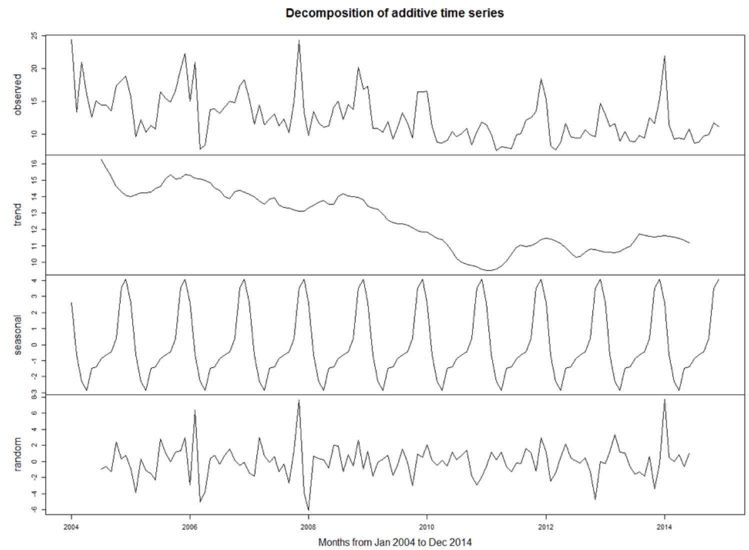


Figure 6: Decomposition of additive time series

Then we draw fitted adjusted PM values based on ARIMA model in Figure 7. The black line represents the variation of real adjusted PM values during 2004-2014 and red line represents the variation of fitted adjusted PM values during the period. From this plot we can see that the fitted line is really rough.

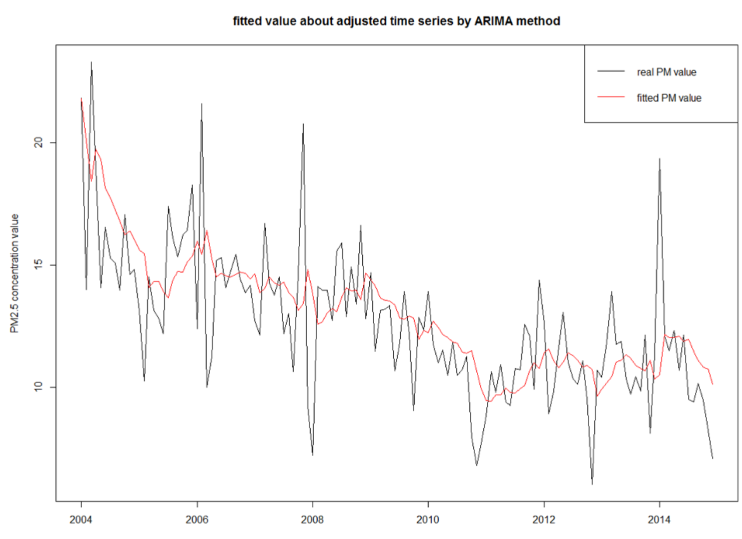


Figure 7: Fitted value about adjusted time series by ARIMA method

Next, we predict the adjusted $PM_{2.5}$ values for the whole year 2015. In Figure 8, similarly as before, the blue line represents the predicted values for 2015. The shadow of deep color represents the 80% prediction interval for the predicted values and the shadow of light color represents the 95% prediction interval for the predicted values.

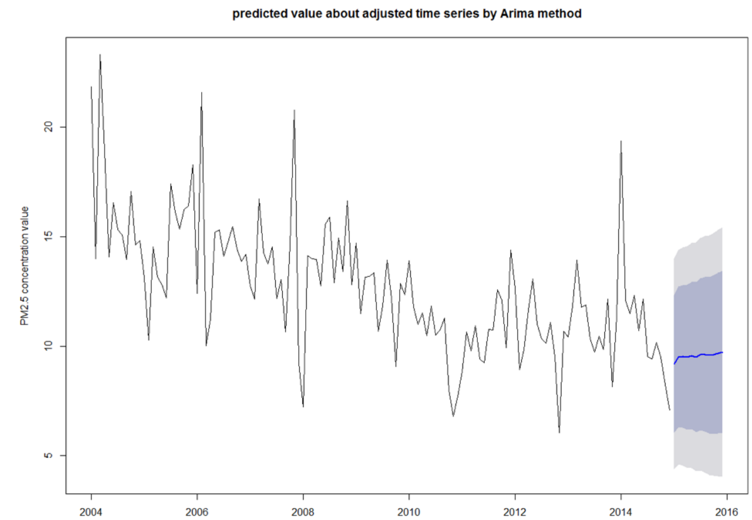


Figure 8: Predicted value about adjusted time series by Arima method

Finally, we add seasonal component to our predicted results and then compare them with the real PM values during 2015. We find the mean square error is around 9.7 so we believe the performance of ARIMA model is even not as good as Holt-Winters Exponential Smoothing.

In sum, the effect of prediction by time series is not very good. But it does give us some inspirations. For example, the PM 2.5 values are different in different seasons. For the next analytical step, we will combine $PM_{2.5}$ values and some other corvariates to construct a better model.

4.2 Experiment 2: PM 25 vs AOD data

Our main goal is to explore the relationship between PM measurements obtained from the ground sites and the AOD measurements obtained from the satellite near the coast. We want to ascertain whether satellite remote sensing can be used to assess PM air quality for areas where surface PM monitors are not available. The AOD measurements reflect the integrated amount of particles in the vertical column, and can be used as an input parameter in statistical models for predicting PM levels. Since time-varying parameters such as relative humidity, wind direction, wind speed and air temperature can influence the PM-AOD relationship, we want to formulate a statistical model that allows for day-to-day variability in this relationship.

We also noticed from Figure 9 that there was a huge difference between the 4 sites we identified along the West Coast, so we decided to either treat the *site* variable as a factor or build the model for each site along the West Coast.

Table 3 is notations for all the variables we used.

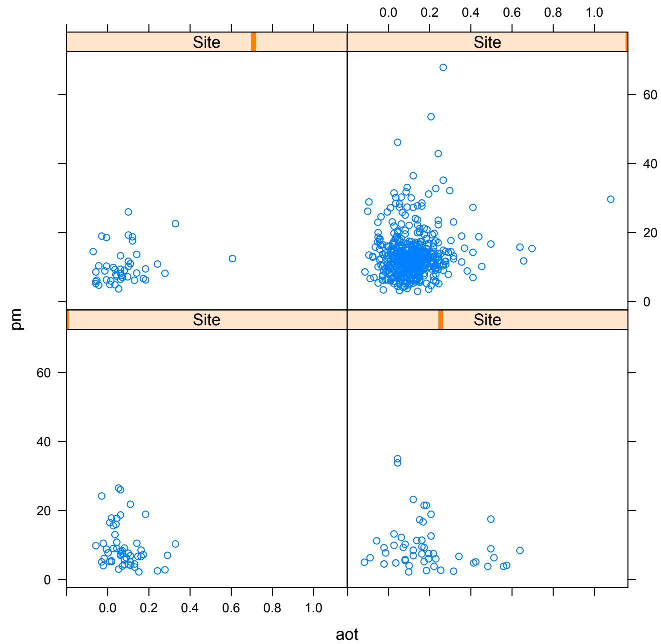


Figure 9: PM vs. AOD for the 4 west coast sites

Variable	Description
PM	Particulate Matter
AOD	Aerosol Optical Depth
U	Wind Speed
V	Wind Direction
H	Relative Humidity
$temp$	Air Temperature
se	Season (factor variable)
si	Site (factor variable)
d	Date (factor variable)
ye	Year (factor variable)

Table 3: Notations for variables

4.2.1 Only AOD data

Since there are many meteorological parameters varying from day to day, our statistical model must have the variability of the date. For each location, there are many different geographical properties, so our model must have the variability of sites. Therefore we used mixed effects model to fit this relationship:

$$PM_{ij} = \alpha + \beta \times AOD_{ij} + si_i + d_j + \epsilon_{ij},$$

where PM_{ij} is the $PM_{2.5}$ concentration at a spatial site i on a specific day j, α is the fixed intercept, β is the fixed slope, AOD_{ij} is the AOD value at a spatial site i on a specific day j, $si_i \sim N(0, \sigma_s^2)$ is the random intercept of site i, $d_j \sim N(0, \sigma_d^2)$ is the random intercept of a specific day j, and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the error term at site i on a day j.

4.2.2 AOD data and wind data

As we have more information about the time-varying parameters, like air temperature, humidity, etc, it is not reasonable to simply treat them as a random variable.

First, we tried the multivariate linear regression model to fit the relation.

$$PM = \beta_0 + \beta_{AOD} \times AOD + \beta_U \times U + \beta_V \times V + \beta_H \times H + \beta_{temp} \times temp + se + ye\epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

We also removed the outliers and did the stepwise regression to improve the model.

But as the result of this model was still not good enough, we tried another kind of model, the gam model, which turned out to be a better model to fit the relationship.

4.3 Experiment 2: PM 25 vs AOT data

4.3.1 Only AOD data

If we only use the AOD data, by using the approach in the Approach Section, we fitted the mixed effects model as follows.

$$\hat{PM}_{ij} = 10.51 + 3.60 \times AOD_{ij} + \hat{si}_i + \hat{d}_j,$$

where $\hat{si}_i \sim N(0, 1.79^2)$ and $\hat{d}_j \sim N(0, 3.04^2)$.

The correlation between the fitted $PM_{2.5}$ data and the true $PM_{2.5}$ data is 0.802, and $R^2 = 0.64$, which we agreed it is not a bad fit.

4.3.2 AOD data and wind data

Multivariate linear regression model If we use both the AOD data and the wind data, as we mentioned in the Approach Section, we first tried to build a multivariate linear regression model.

The following analysis is about the site (40.80178, -124.1621) since the relation is different for different sites as mentioned in Section 3. For other sites, analysis should be similar.

$$\hat{PM} = -0.64 - 0.0176 \times AOD - 0.11 \times U + 0.45 \times V + 0.017 \times H - 0.16 \times temp + se + ye,$$

$R^2 = 0.758.$

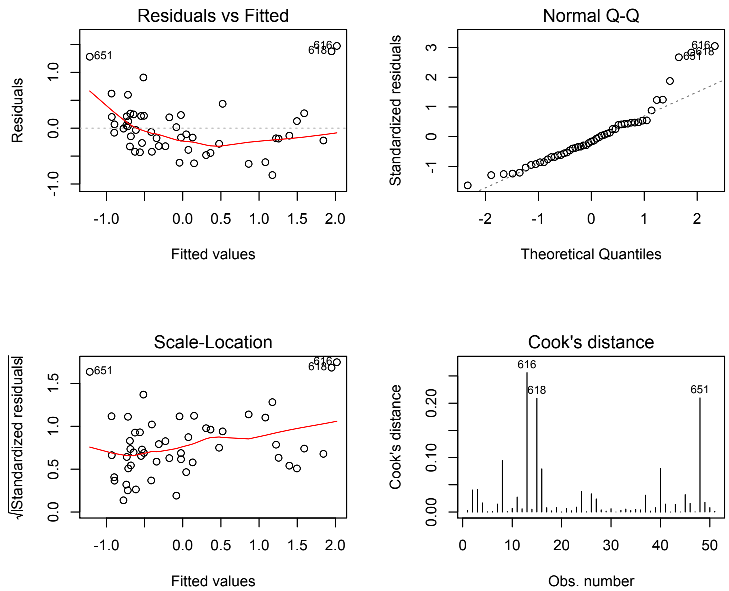


Figure 10: Diagnosis 2

We can see from figure 10 that there are three outliers. By deleting these outliers, the model became:

$$\hat{PM} = 94.43 + 2.34 \times AOD - 1.11 \times U + 0.033 \times V - 0.068 \times H - 0.30 \times temp + se + ye,$$

$R^2 = 0.822.$

Then we did the stepwise regression to choose the best subset of all the variables, and got the following model:

$$\hat{PM} = 3.42 - 1.07 \times U + 0.036 \times V + se,$$

$R^2 = 0.786.$

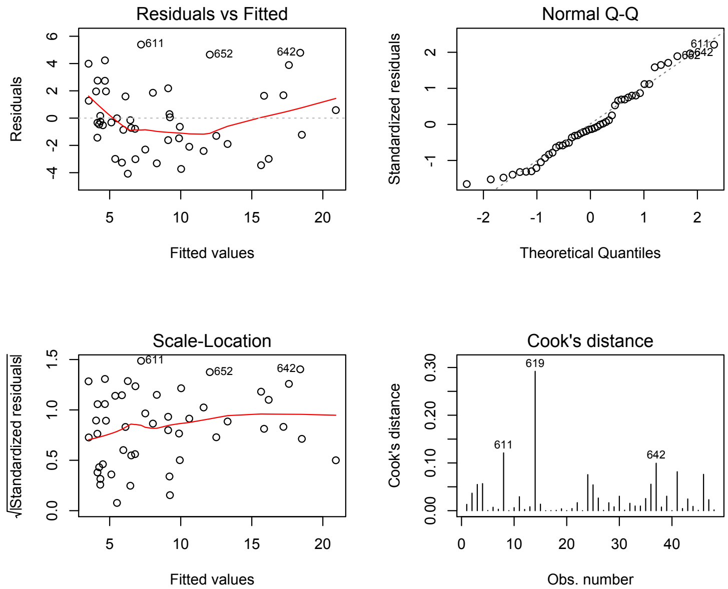


Figure 11: Diagnosis 2

Figure 11 are the diagnosis plots for the new model. Results obtained here were better than the previous model. But we can see from those plots that the residuals still have some trend and some cluster, which shows that this model violated some assumptions, like the equal variance assumption for the linear regression model. To figure out if this model is good, we did 5-fold cross validation. The results are described in figure 12.

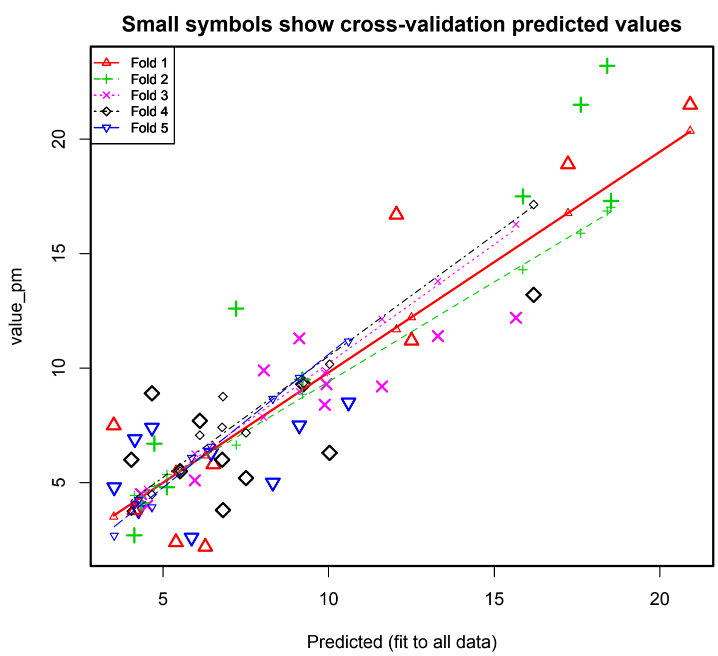


Figure 12: 5-fold cross validation for the multivariate linear regression model

The mean square error for cross validation is 8.16. As the average PM values is 8.75, 8.16 is pretty high as there would be $\frac{\sqrt{8.16}}{8.75} = 32.65\%$ error for PM value, which is high. Another reason that we agreed this is not a good model is that it actually did not include the AOD data as predictors. According to the meteorologic knowledge, AOD should play a key role in the model. Above all, we concluded that the multivariate linear regression model is not good.

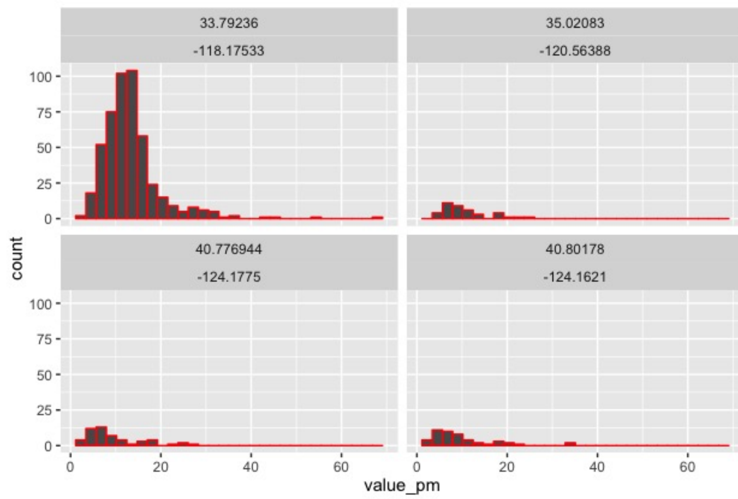


Figure 13: Histograms for $PM_{2.5}$ measurements at 4 sites in California. The titles on each subpanel represent the latitudes and longitudes for the sites respectively.

The fairly skewed nature of the $PM_{2.5}$ measurements, the response, and the fact that it is necessarily a positive quantity, suggest that some transformation maybe required if a Gaussian error model is to be used. Attempting to use a Gaussian model without transformation confirms this. The upper left normal QQ plot, in Figure, clearly shows a problem with the Gaussian assumption. Examining the plot on upper right of residuals versus fitted values reveals that the constant variance assumption is unreasonable. The lower left histogram of residuals confirms the pattern evident in the QQ plot: there are too many residuals in the lower tail which means that we tend to over estimate the $PM_{2.5}$ levels using the model. The lower right plot emphasizes the failure of the constant variance assumption. So we conclude that multiple linear regression is not the best approach to modeling the relationship between the response and predictor variables.

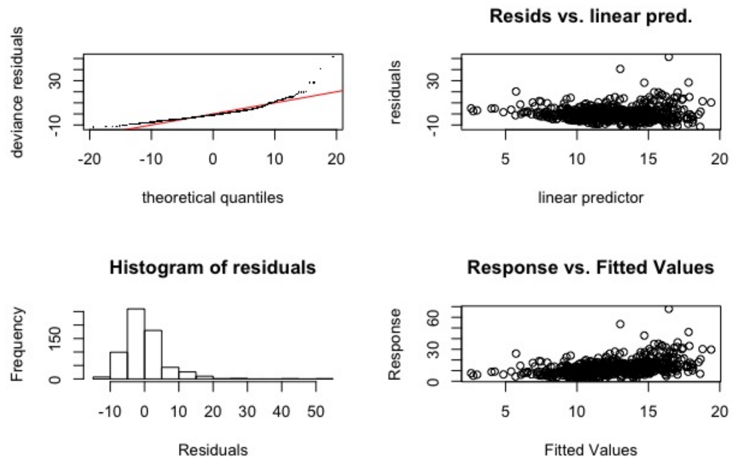


Figure 14: Some basic model checking plots for a model fitted to the $PM_{2.5}$ - AOD data

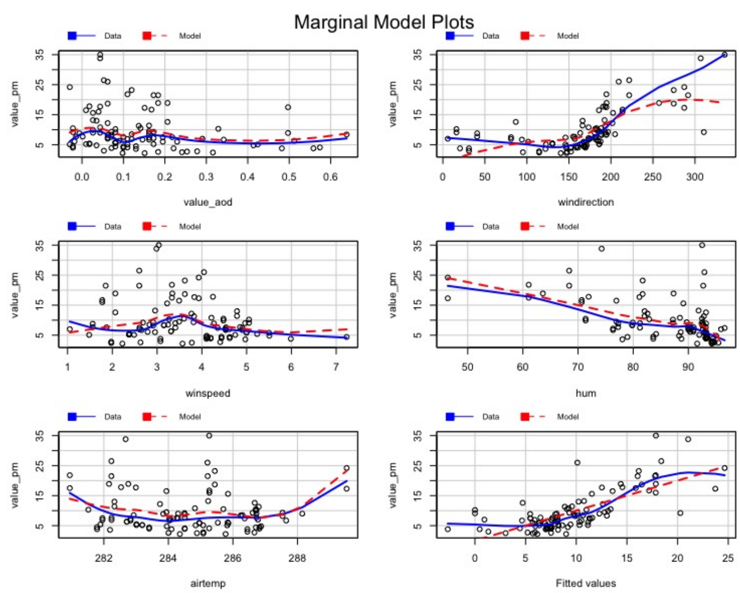


Figure 15: Marginal model plots for multiple linear regression

Generalized Additive Models (GAMs) employ a class of equations called "smoothers" or "scatterplot smoothers" that attempt to generalize data into smooth curves by local fitting to the subsections of the data. The basic idea behind GAMs can be described as follows. We calculate a smooth curve that goes through the data as well as possible while being parsimonious. Note that it is possible using a polynomial of high enough order to get a curve to go through every point. This makes the curve "wiggle" excessively, and not represent a parsimonious fit. The approach generally employed with GAMs is to divide the data into some number of intervals/sections, using "knots" as the end points of these intervals. Then a low order polynomial or spline function is to fit the data in an interval, with the added constraint that the second derivative of the function at the knots must be the same for both intervals sharing the knot. This eliminates sharp edges in the curve, and ensures that it is smooth and continuous at all points. As a practical matter, we can view GAMs as non-parametric curve fitters that attempt to achieve an optimal compromise between goodness-of-fit and parsimony of the final curve. One of the interesting aspects of GAMs is that they can only approximate the appropriate number of degrees of freedom, and that the number of degrees of freedom is often not an integer, but rather a real number with some fractional component. A second order polynomial (or quadratic equation) in a GLM uses two degrees of freedom (plus one for the intercept). A curve that is slightly less regular than a quadratic might require two and a half degrees of freedom (plus one for the intercept), but might fit the data better. The other aspects of GAMs that is different is that they don't handle interaction well. Rather than fit multiple variables simultaneously, the algorithm fits a smooth curve to each variable seperately and then combines the results additively, thus giving rise to the name "Generalized Additive Models.

We model the variability in $PM_{2.5}$ concentrations using generalized additive models (GAMSs) (Hastie and Tibshirani 1990). Our model can be described as follows,

$$\mathbb{E}(Y_{t,site} \mid AOD, U, V, H, temp, Season) = \beta_0 + \mathrm{AOD} + f_U(U) + f_V(V) + f_H(H) + f_{temp}(temp) + \mathrm{Season} \quad (1)$$

$Y_{t,\mathrm{site}}$ is the daily $PM_{2.5}$ concentration at a given site. All the covariates vary with time and site. μ is the random model intercept, $f_U(U)$, $f_V(V)$ are one dimensional smooth surfaces describing the impact of wind speed and direction on the AOD-$PM_{2.5}$ association. Season is modeled as a 4-level categorical variable because of its discrete values.

We fit the model with the *gam*() function in the *mgcv* package in R (Wood 2006). The package *mgcv* uses penalized regression splines for the f_j, so we can write

$$f_j(x) = \sum \beta_{jk}\phi_{jk}(x) = \boldsymbol{\beta}'\phi_j$$

for the basis functions ϕ_{jk} that determine the splines. Given the basis function representation of the f_j, (1) is now a parametric mean function with parameters $\beta = (\beta_0, \boldsymbol{\beta_1}', \ldots, \boldsymbol{\beta_n}')'$ and predictors that define the intercept and the splines that define the s_j.

The penalized least squares objective function for estimating $\boldsymbol{\beta}$ is

$$\|Y - X\boldsymbol{\beta}\|^2 + \sum_j \lambda_j \boldsymbol{\beta_j}'\boldsymbol{\beta_j}' \quad (2)$$

The values of λ_j are selected in an iterative algorithm to minimize a generalized cross validation criterion. This fit is done using the *gam* function in the *mgcv* package,

```
Family: gaussian
Link function: identity
```

```
Formula:
value_pm ~ (value_aod) + s(windirection) + s(hum) + s(airtemp) +
    season

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.085      1.065    3.84  0.00023 ***
value_aod     -2.283      2.330   -0.98  0.32993
season         2.485      0.384    6.47  5.2e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                 edf Ref.df     F p-value
s(windirection) 2.66   3.28 35.38 < 2e-16 ***
s(hum)          2.46   3.00  1.37 0.24676
s(airtemp)      3.07   3.81  5.73 0.00051 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.786   Deviance explained = 80.8%
GCV = 11.065  Scale est. = 9.8282    n = 100
```

The only visible parameter in this model is the intercept, estimated to be 4.085. The $\boldsymbol{\beta_j}$ are hidden inside the smoothers and are largely uninterpretable. For each of its predictors, a smoother was fit by the f functions. The default in the function f that does the smoothing uses *thin plate regression splines* that don't depend as much on the number of knots selected and also they generalize to smooths of more than one variable at a time. In the output, the "edf" is the equivalent degrees of freedom for each of the smooths. If we are using k basis functions for a smooth, then we would have k degrees of freedom (df) for the smooth. Penalization will generally reduce edf to a number smaller than k. The adjusted R^2 is the square of the correlation between the observed and fitted values, with an adjustment for degree of freedom, while the deviance explained appears to be the usual R^2.

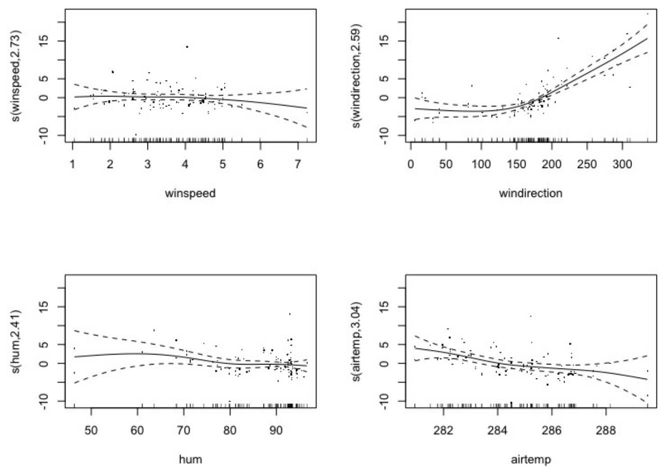


Figure 16: The estimates of the smooth functions shown without partial residuals. The dotted lines aound the solid lines show the Bayesian credible intervals.

In Figure 4, the solid line is the predicted value of the dependent variable as a function x axis. We plot two times the standard errors of the estimates (in dashed lines). The small lines along the x axis are the "rugs" showing the location of the sample plots. The result for all the predictors look relatively flat because the standard errors are not too large.

We also perform the likelihood ratio tests to compare different models and conclude that the model displayed on the previous page does the best job.

4.4 Experiment 3: Spatial Statistics

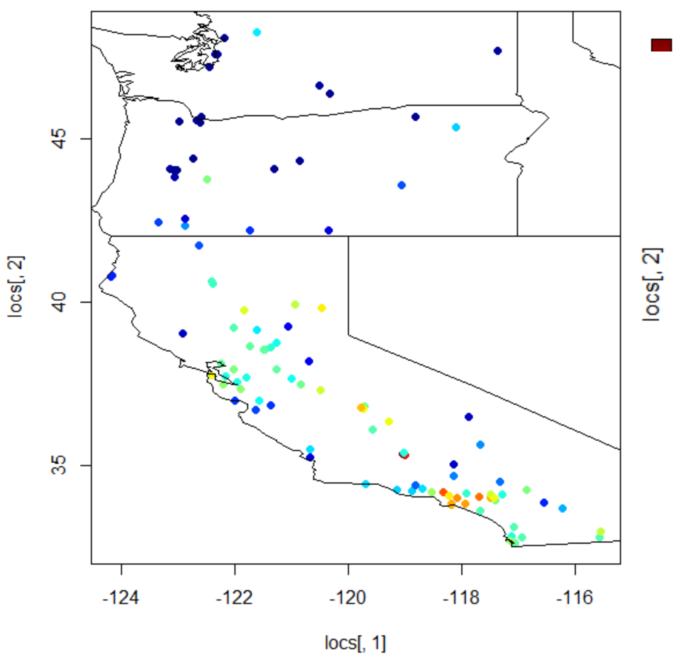
We try to build spatial model for the AOD and $PM_{2.5}$ data. In spatial statistics, we study the relations and variation in data with respect to its location. Spatial statistics on a geological set is carried out in two stages:

- Analyze the dataset to build a relation (covariance) amongst values based on their geographical location. In the language of statistics, this means building a variogram which models variance between values at two locations according to the distance and direction between them.

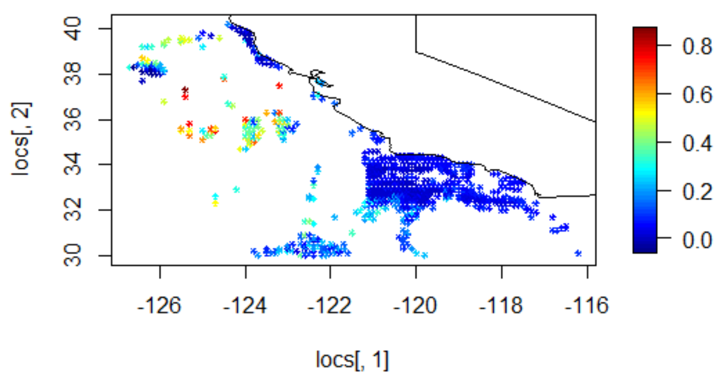
- The next step is to estimate values at unsampled locations. This process is called "Kriging". The interpolated values obtained by Kriging are modeled using a Gaussian process directed by prior covariances. In contrast to Kriging, using polynomial interpolation optimizes the smoothness of the interpolated values.

We use ordinary Kriging to interpolate AOD and $PM_{2.5}$ values. In ordinary Kriging, the interpolated value is a weighted linear combination of sampled values. Ordinary Kriging assumes a constant mean and residual mean error of 0. Along with this, ordinary Kriging aims to minimize the variance of error. The variogram gives the covariances between different values. Ordinary Kriging is obtained by using probability models that calculate the bias and error in variance, which can then be used to choose weights for neighboring sampled locations such that mean error for the model is exactly zero and modeled error variance is minimized. We have used a maximum likelihood probability model to estimate parameters. We obtain the spatial plots and standard error associated with the interpolations using this method of Kriging.

There are 284 sensor sites that recorded the $PM_{2.5}$ measurements. We use the daily $PM_{2.5}$ measurements at these sites to make spatial interpolation plots. Similarly, spatial interpolation plots are constructed for AOD values. Figure 17a below displays the 107 sensor sites where $PM_{2.5}$ data was collected on 7th January, 2009.



(a) January 7, 2009 PM sites



(b) January 6, 2009 AOD sites

Figure 17

Figure 18 displays the spatially interpolated plot obtained by using ordinary Kriging of the sample data and the standard error associated with the interpolated values. As expected, we see the error increases as distance from the sample data increases.