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 many 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 pollutants, including $PM_{2.5}$. This project explores the relationship between the surface readings of $PM_{2.5}$ from coastal sites in California and Hawaii 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.

1 Introduction

$PM_{2.5}$ is particulate matter that is less than 2.5 micrometers in diameter and is often referred to as the greatest health risk of pollutants [2]. For this project, we used two different sources of data – Aerosol Optical Thickness (AOT) obtained from satellite measurements, and surface $PM_{2.5}$ measurements. AOT measurements refer to a quantitative measure of PM abundance in the atmospheric column and are also referred to as aerosol optical depth (AOD). The measurements are for the most part dominated by near surface emissions, [3].

Domestic sources of emissions are the primary cause of air pollution in the U.S.; however, there is potential for international flow of air pollution into the U.S. to be a contributing factor in some coastal cities having high recorded measurements of air pollution. The impact that international transport of air pollution has on our ability to attain air quality standards or other environmental objectives in the U.S. has yet to be fully characterized. In other words, cities in the western U.S. may unknowingly be receiving high pollutant

[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]EPA
[10]EPA
[11]CICS

values through no fault to themselves. The goal in this project is to establish a connection between AOT measurements from the AVHRR satellite to determine the impact of international transport of air pollution on $PM_{2.5}$ concentrations in coastal areas of the western U.S. We examined $PM_{2.5}$ sites that were on the coast. Most of our sites were situated in California, with a few in Washington and a handful in Hawaii.

Several previous studies have attempted to find correlations between $PM_{2.5}$ readings and AOD readings. Liu et al (2009) 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 find a much better agreement between $PM_{2.5}$ concentrations and EPA observations when they used a model with AOD information than with a non-AOD model. Li et al (2011) studied the approach to use AOD data to predict $PM_{2.5}$ concentrations,[4].

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 $\mu g/m^3$ and 80 $\mu g/m^3$ in Northern India and Eastern Asia, respectively. They found that the population-weighted concentrations in East Asia nearly doubled the global mean.

We aim to generate relationships between AOD measurements and surface $PM_{2.5}$ measurements at sites located on the west coast. In addition, we analyze times 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 help establish a visualization of high emission events in the state such as wildfires.

2 The Problem

2.1 Description of data

The first data set of Climate Data Record (CDR) of AOT was obtained from by the National Oceanic and Atmospheric Administration (NOAA) [1]. This data was collected using the Advanced Very High Resolution Radiometer (AVHRR) that provides an optical measure of aerosol column loading derived form the global ocean pixel-level PATMOS-x AVHRR clear-sky reflectance CDR at 0.63 μm channel [1]. This satellite provides global readings of oceanic measurements of AOT on a daily, as well as a monthly scale, for the years 1981-2009 [1].

The second data set consisting of the $PM_{2.5}$ measurements, was taken on land sites in California, Oregon, Washington, Alaska, and Hawaii. This was provided by the United States Environmental Protection Agency (EPA). $PM_{2.5}$ is so small that it can get lodged into lungs and make it difficult for people to breath. This creates an increase in respiratory problems. Common contributing factors to $PM_{2.5}$ include emissions for motor vehicles, power plants, wood burning, and dust from paved or unpaved roads, [2].

The AVHRR takes approximately 16 days to cover the entire earth. We, thus, have roughly two data values for each month of the year. The frequency of the $PM_{2.5}$ data is variable and ranges 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. Hence, we sometimes have months that have only two or less $PM_{2.5}$ data and/or no AOT data.

2.2 Challenges

Arvind: Here I would list the challenges associated with the data, lack of spatial and temporal coverage, etc. The differences between what we would like to have vs what we have instead.

3 The Approach

3.1 Data processing

Arvind: Describe what other preprocessing was done to the data.

No matter what PM2.5 sites we use, we need to find the closest AOD coordinates to the PM2.5 sites, and then to match the sites data for each day. This requires us to search all the observations in AOD dataset. The AOD data are very large and we cannot put all the data in hundreds of files into one matrix. Besides, AOD data have lots of missing data. So we need to clean the AOD data first.

Since the data we want to use are the data located on the west coast and Hawaii, first we used the longitude and latitude of the west coast and Hawaii to eliminate data from other locations that we won't use. Then, we dropped all the missing data and also changed the original format of the data into more readable one. Finally, our AOD data looks like Table 1.

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

Since the coverage of the AOT 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 location of all the $PM_{2.5}$ sites, whereas Figure 2 shows the geographical locations to 13 locations – four of these are found in California, whereas the rest can be found in Hawaii. In the rest of this report, we focus on only these 13 locations. Arvind: Can you also mention that we only chose overlapping temporal data.

3.3 Analzing $PM_{2.5}$ 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

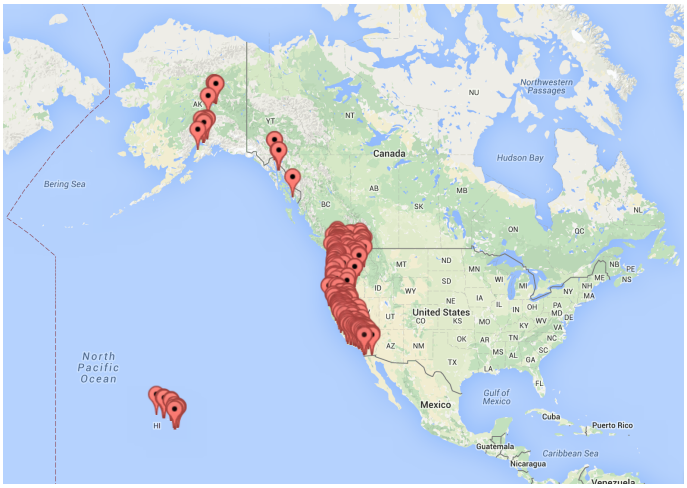


Figure 1: Arvind: Please fill in the caption

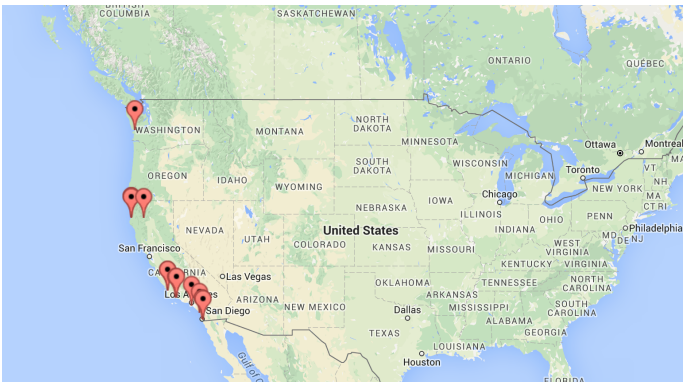
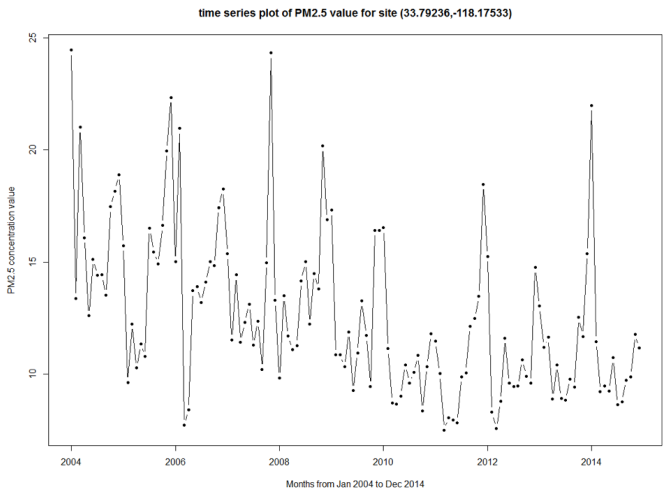


Figure 2: Arvind: The figure only shows the California locations, can the figure include the Hawaii sites as well? This map could also be zoomed in further. Also add a caption.

collected, with color varying depending on the intensity of the reading, with darker colors indicating a larger concentration of $PM_{2.5}$.

To figure out whether we can use previous PM2.5 value to predict future PM2.5 value, we decide to conduct time series analysis. We focus our attention on the PM2.5 site which located at latitude 33.79236 and longitude -118.175 (It is really close to Long Beach). We collect monthly average PM2.5 concentration value from Jan 2004 to Dec 2014 and plot the time series in the figure below.

There is a slightly downward trend over time about this time series and there seems to exist seasonality. We decide to use two different methods to fit the time series model. 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 one is ARIMA model. ARIMA is the short for autoregressive integrated



moving average. This model can be fitted to time series data either to better understand the data or to predict future points in the series. For the ARIMA method, we firstly adjust the time series by subtracting the estimated seasonal component and then apply the model to the adjusted series.

Arvind: The results of the experiments should be discussed in the next section. After analyzing this data we noticed (insert conclusion here... talk about it more...). The animation is available online (insert website-github?).

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

We noticed that there was a huge difference between the 13 sites we identified along the West Coast and the sites located on Hawaii. The sites on Hawaii are unique because Hawaii is surrounded on water. Thus we decided to build models separately for both the sites along the West Coast and the sites on Hawaii.

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 have the variability of sites. Therefore we used 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.

Date	lat_aot	long_aot	aot
2000-01-25	30	-126.6	-0.0836133733391762
...	...	...	...

Table 1: AOD data

3.4.2 AOD data and wind data

3.4.3 Match the AOD data with the PM2.5 data

As we are trying to find the relationships between AOD and PM2.5, we need to keep all the date the same, and locations closest. After matching these two datasets, we got the following as in Table table3.2.

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 PM2.5 match data

4 Computational Experiments

Arvind: I suggest organizing this section into a series of experiments. Here are a few suggestions. You can also follow the instructions at the end of the section.

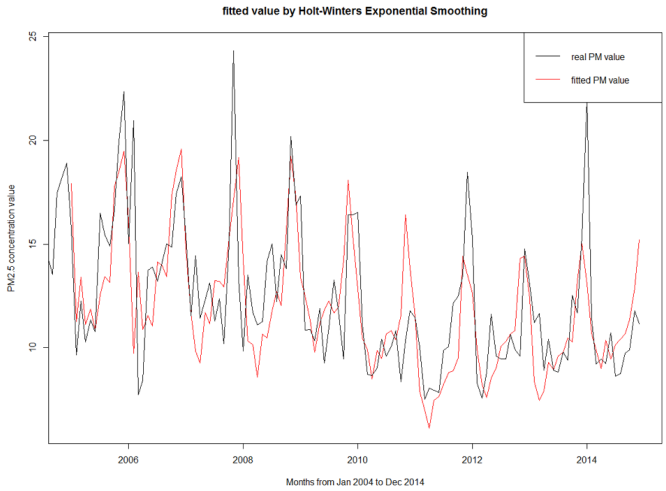
4.1 Experiment 1: Analysis of PM25 time series

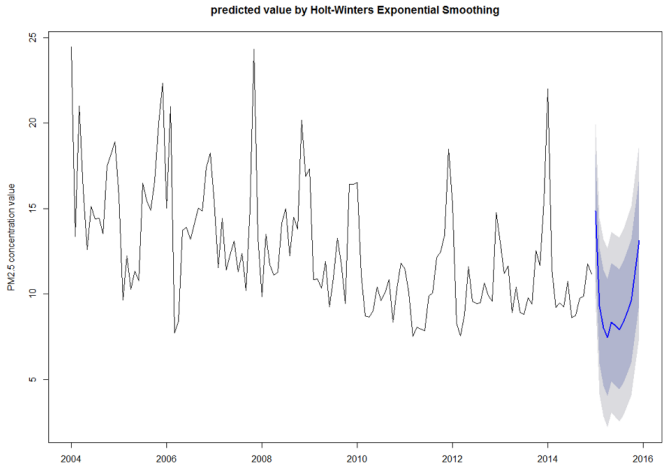
4.1.1 Holt-Winters Exponential Smoothing

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

Then we predict the PM2.5 concentration value for the whole year 2015. In the plot below, the blue line represents the predicted value for 2015. The shadow of deep color represents the 80% prediction interval for the predicted value and the shadow of light color represents the 95% prediction interval for the predicted value.

Now let?u 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 value for the year 2015, we find the mean square error between them is around 5.5. Becaues the mean of PM2.5 value during 2004-2014 is about 12.6, the mean square error is kind of large and this model is not very accurate.



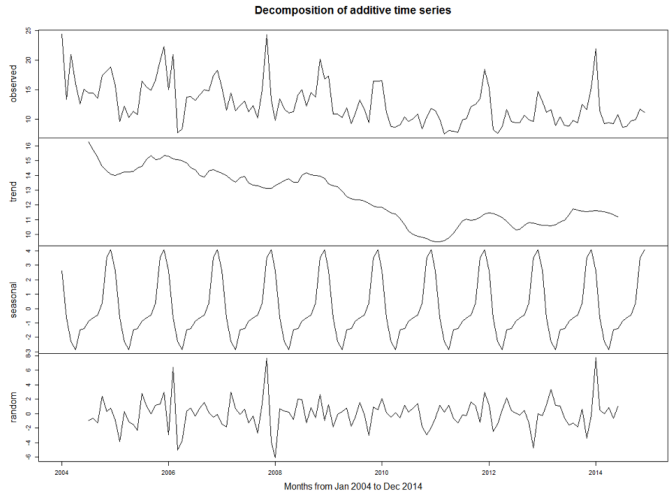


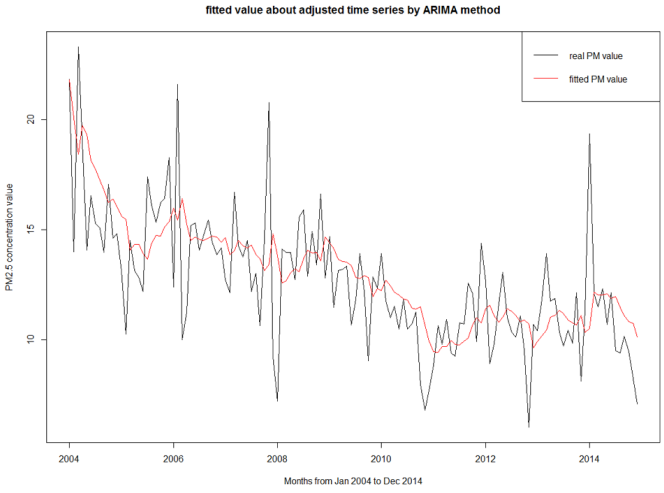
4.1.2 ARIMA model

Becaues there seems to exist seasonality in the PM2.5 value time series, we want to decompose the time series and apply ARIMA model to the adjusted time series. We first decompose the PM2.5 concentration value time series and plot it. In the plot below, the four subplots respectively represent observed value, overall trend component, seasonal component and random part. From the seasonal component, we can see that there does exist differences of PM2.5 value among different months. By subtracting the seasonal component from the original time series, we get adjusted PM2.5 concentration value time series.

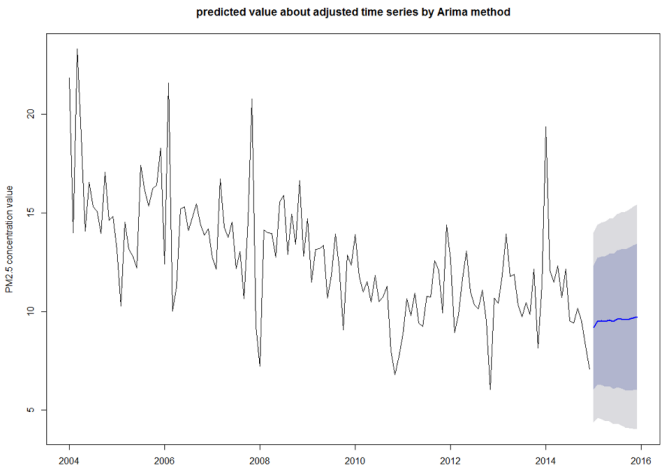
Then we draw fitted adjusted PM value based on ARIMA model. The black line represents the variation of real adjusted PM value during 2004-2014 and red line represents the variation of fitted adjusted PM value during the period. From this plot we can see that the fitted line is really rough.

Next, we predict the adjusted PM2.5 concentration value for the whole year 2015. Similar as before, , the blue line represents the predicted value for 2015. The shadow of deep color represents the 80% prediction interval for the predicted value and the shadow of light color represents the 95% prediction interval for the





predicted value.



Finally, we adding seasonal component to our predicted result and then compare it with the real PM value during 2004-2014. We find the mean square error is around 9.7. So 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 inspiration. For example, the PM 2.5 value will be different in different seasons. For next step

4.2 Experiment 2: PM 25 vs AOT data

4.2.1 Sites closest to the west coast

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

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

where $\hat{s}_i \sim N(0, 1.79^2)$ 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, which we agreed it is not a bad fit.

If we use both the AOD data and the wind data, we built a multivariate linear regression model. As in Figure XXX, different sites have different relationships between AOD and PM.

The following analysis is about the site (40.80178, -124.1621). For other sites, analysis should be similar.

$$\hat{PM} = -0.64 - 0.0176 \times AOD - 0.11 \times WindSpeed + 0.45 \times WindDirection + 0.017 \times Humidity - 0.16 \times AirTemp + Season + Year,$$

where Season and Year are factor variables. $R^2 = 0.758$.

We can see from the above plots that there are three outliers. After deleting these outliers, the model became:

$$\hat{PM} = 94.43 + 2.34 \times AOD - 1.11 \times WindSpeed + 0.033 \times WindDirection - 0.068 \times Humidity - 0.30 \times AirTemp + Season + Year,$$

where Season and Year are factor variables. $R^2 = 0.822$.

Then we did the stepwise to choose the best fit, we get the following model:

$$\hat{PM} = 3.42 - 1.07 \times WindSpeed + 0.036 \times WindDirection + Season,$$

where Season is a factor variable. $R^2 = 0.786$.

The results are better than the previous model. But we can see from those plots that the residuals still have some trend and some cluster. To figure out if this model is good, we did 5-folder cross validation. The results are as follows:

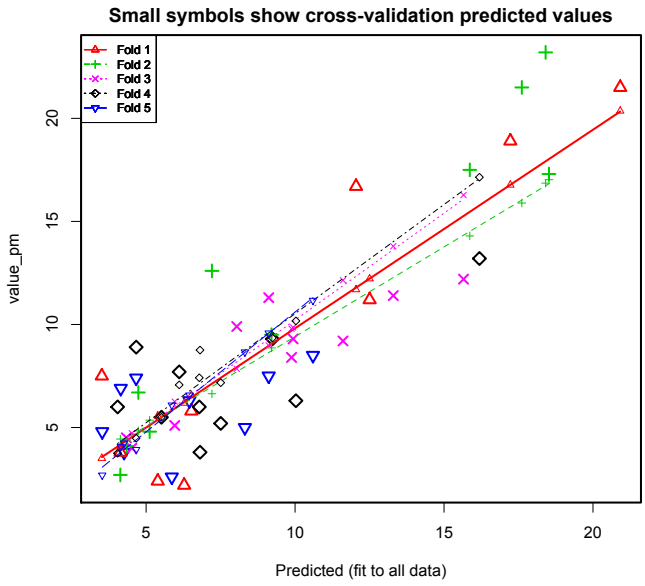


Figure 3:

The mean square error is 8.16, which is very high due to the average PM values is 8.75. So multivariate linear regression is not a good model.

4.2.2 Hawaii sites

Similarly, we can get the mixed effects model for Hawaii sites. Without the wind data,

$$\hat{PM}_{ij} = 6.205 + 7.780 \times AOD_{ij} + \hat{s}_i + \hat{d}_j,$$

where $\hat{s}_i \sim N(0, 4.05^2)$ and $\hat{d}_j \sim N(0, 2.83^2)$.

The correlation between the fitted $PM_{2.5}$ data and the true $PM_{2.5}$ data is 0.879, which is better than the west coast data.

4.2.3 relationships between AVHRR AOD and surface PM2.5 for sites closest to the west coast

For sites closest to the west coast, we finally found 4 sites that have common date information. Figure 4 is the plot regarding the AOD and PM2.5 by each site.

We can see from the above plot that the relationships between AOD and PM2.5 are different for different sites. So we consider building different models for each site to find the relationships.

4.2.4 relationships between AVHRR AOD and surface PM2.5 for Hawaii sites

For Hawaii sites, we finally found 9 sites that have common date information. Figure 5 is the plot regarding the AOD and PM2.5 by each site.

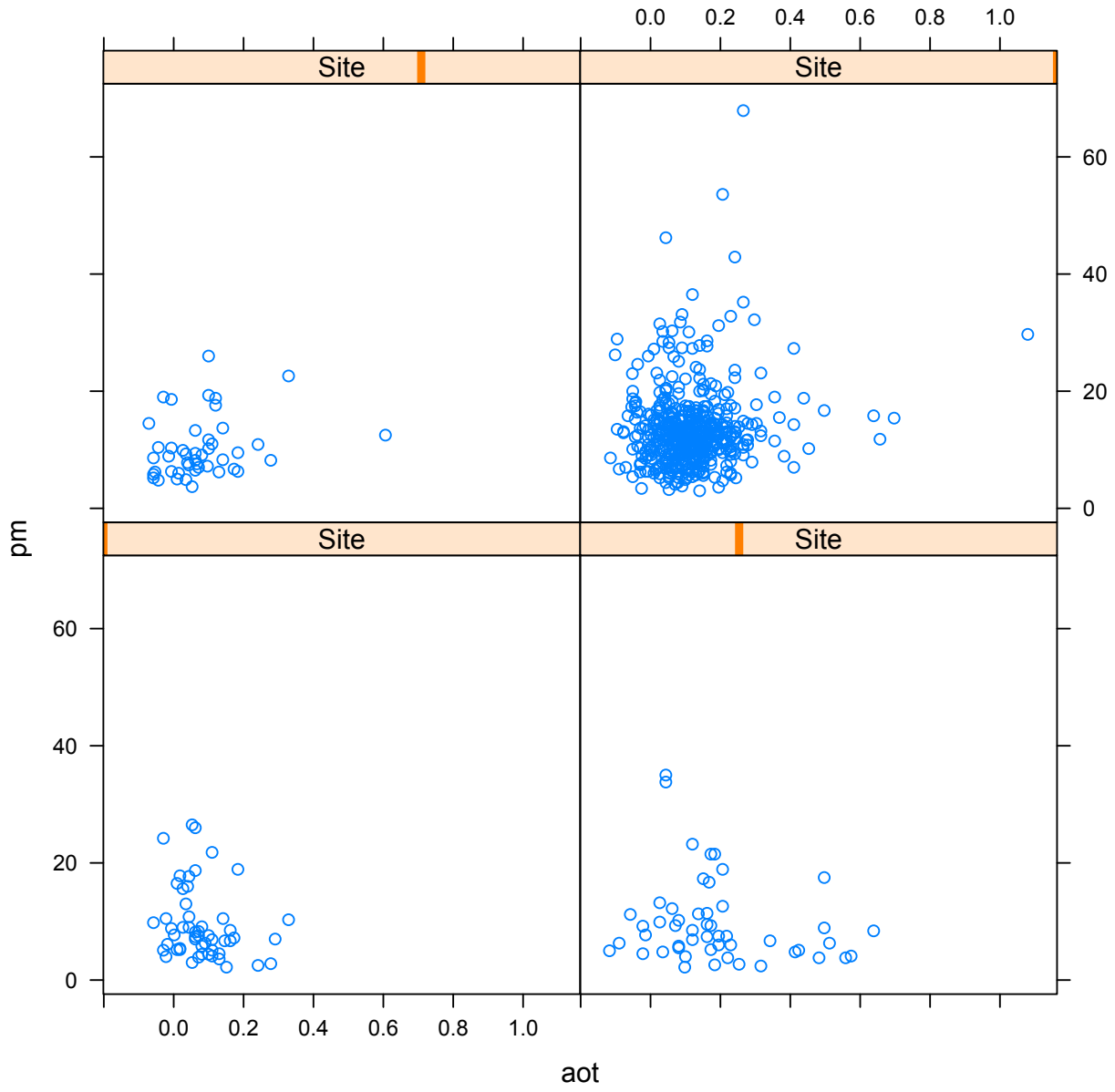


Figure 4: PM vs. AOD for each site close to the west coast

Similarly, we can see from the above plot that the relationships between AOD and PM2.5 are different for different sites. Like what we did for sites close to the west coast, we consider building different models for each site to find the relationships.

4.3 Spatial Statistics

We try to build spatial model for the AOT and PM data.In spatial statistics, we study the relations and variation in data with respect to its location. Spatial statistics on a geological data is carried out in two

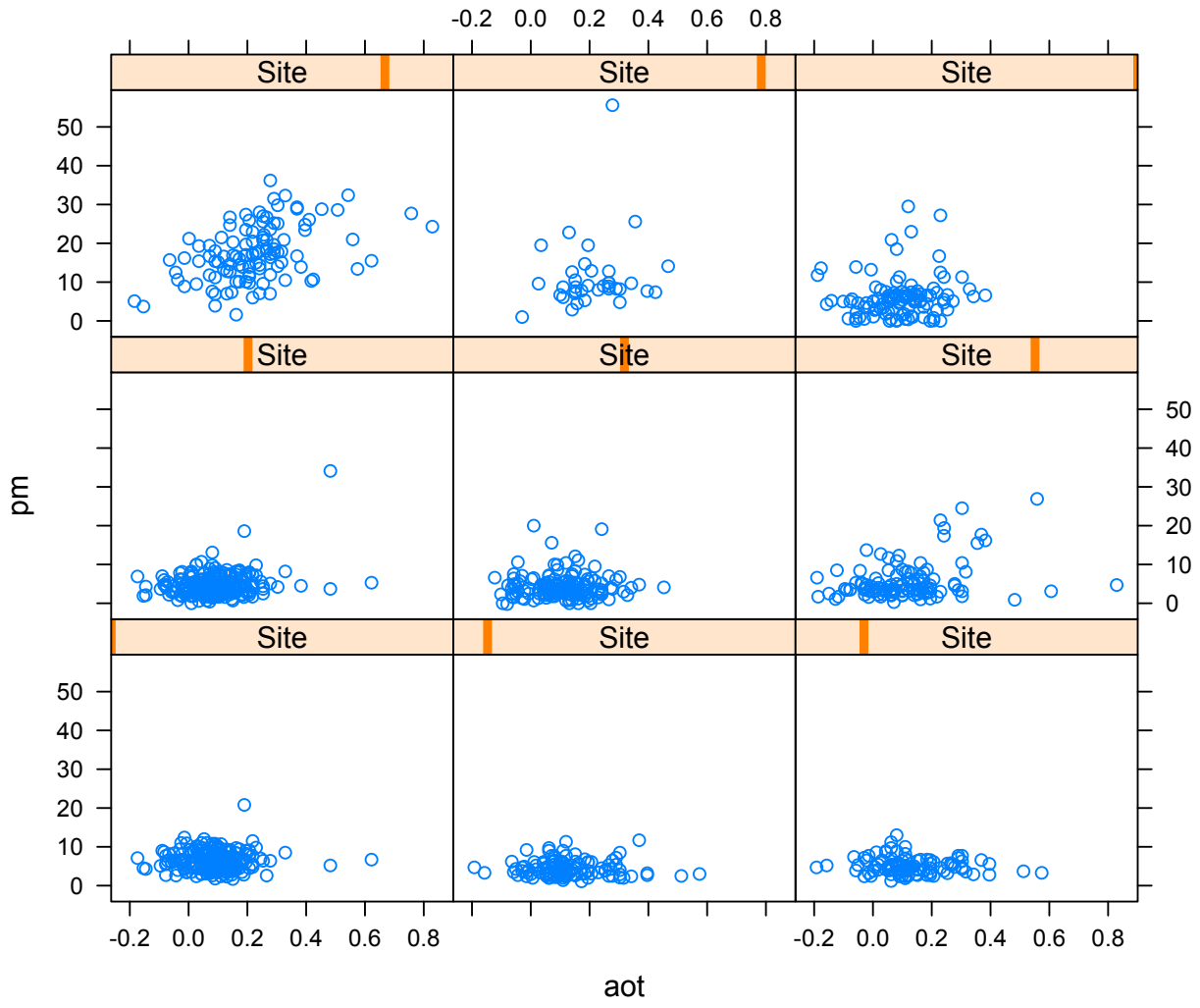


Figure 5: PM vs. AOD for each Hawaii site

stages:

- Analyze the dataset to build a relatedness (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 estimates values at unsampled locations. This process is called "Kriging". The interpolated values are obtained by Kriging are modeled using a Gaussian process directed by prior covariances. In contrast, the focus while using polynomial interpolation is to optimize the smoothness of the interpolated values.

We use ordinary kriging to interpolate AOD and PM2.5 values.

In ordinary kriging, the interpolated value is a weighted linear combination of sampled values. Ordinary kriging assumes constant mean and residual mean error to be 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 neighbouring sampled locations such that mean error for the model is exactly zero and modelled error variance is minimized. We have used 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 PM2.5 measurements. We use the daily $PM_{2.5}$ measurements values at these sites to make spatial interpolation plots. Similarly, spatial interpolation plots are constructed for AOD values.

Give enough details so that readers can duplicate your experiments.

- Describe the precise purpose of the experiments, and what they are supposed to show.
- Describe and justify your test data, and any assumptions you made to simplify the problem.
- Describe the software you used, and the parameter values you selected.
- For every figure, describe the meaning and units of the coordinate axes, and what is being plotted.
- Describe the conclusions you can draw from your experiments

5 Summary and Future Work

- Briefly summarize your contributions, and their possible impact on the field (but don't just repeat the abstract or introduction).
- Identify the limitations of your approach.
- Suggest improvements for future work.
- Outline open problems.

References

[1] National Centers for Environmental Information. National Oceanic and Atmospheric Administration. Department of Commerce, n.d. Web. 23 July 2016. *https://www.ncdc.noaa.gov/cdr/atmospheric/avhrr-aerosol-optical-thickness.*

[2] United States Environmental Protection Agency. AirData. EPA, 5 July 2016. Web. 23 July 2016. *https://www3.epa.gov/airdata/.*

[3] Liu, Yang, Christopher J. Paciorek, and Petros Koutrakis. *Estimating Regional Spatial and Temporal Variability of $PM_{2.5}$ Concentrations Using Satellite Data, Meteorology, and Land Use Information.* 6th ed. Vol. 117. N.p.: Environmental Health Perspectives, June 2009. Print.

[4] Lee, H J., Y Liu, B A. Coull, J Schwartz, and P Koutrakis. *A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations.* N.p.: Atmospheric Chemistry and Physics, 2011. Print.

[5] Donkelaar, Aaron von, Randall V. Martin, Michael Brauer, and Brian L. Boys. *Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter.* 2nd ed. Vol. 123. N.p.: Environmental Health Perspectives, 2015. Web. 26 July 2016. *http://dx.doi.org/10.1289/ehp.1408646*

[6] Li, Jing, Barbara E. Carlson, and Andrew A. Lacis. *How well do satellite AOD observations represent the spatial and temporal variability of $PM_{2.5}$ concentration for the United States?* N.p.: Atmospheric Environment, 2015. Print.