2015-01-01

Combining Semiparametric Regression and Kriging for Prediction of PM2.5 Pollutant Levels at Unmonitored Locations with Meterological and Traffic Data

Justin Jonathan Strate
University of Texas at El Paso, jjstrate@miners.utep.edu

COMBINING SEMIPARAMETRIC REGRESSION AND KRIGING FOR PREDICTION OF PM2.5 POLLUTANT LEVELS AT UNMONITORED LOCATIONS WITH METEROLOGICAL AND TRAFFIC DATA

JUSTIN JONATHAN STRATE

Department of Mathematical Sciences

APPROVED:

Joan Staniswalis, Chair, Ph.D.

Wen-Whai Li, Ph.D.

Ori Rosen, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

to my

FAMILY

with love

COMBINING SEMIPARAMETRIC REGRESSION AND KRIGING FOR PREDICTION OF PM2.5 POLLUTANT LEVELS AT UNMONITORED LOCATIONS WITH METEROLOGICAL AND TRAFFIC DATA

by

JUSTIN JONATHAN STRATE

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2015

Acknowledgements

I would like to thank the faculty of the Mathematical Sciences Department at the Unviserity of Texas at El Paso for all their hard work, encouragement and persistence in guiding me throughout my experience here.
I specifically would like to thank Dr. Joan Staniswalis for her hard work, patience, advice, encouragement and dedication to my success. Dr. Staniswalis took the time to not only improve my statistics and mathematics, but to improve my programming, communication and writing. I would like to thank my other committee members for their feedback and guidance. Thank you to Mario Garcia and Dr. Wen-Whai Li for providing data that was used throughout this study. I would also like to thank Dr. Art Duval, Dr. Helmut Knaust, Dr. Panagis Moschopoulos, Dr. Ori Rosen, Dr. Naijun Sha, Dr. Xiaogang Su, and Dr. Amy Wagler for their instruction and the academic challenges they provided me.

Thank you to my friends and classmates, Behzad, Luis, Michael, Pei, Tun Lee, and Yaa. Thank you for your friendship, support, encouragement and guidance throughout this time here. I wish you the best in your future endeavors.

I would also like to thank my family who has persistently guided my throughout my life. To my parents for first sparking my interest in statistics. My brother, Nathan for your remarkable, selfless generosity. My brother Ben and sister Glori for your help in even the smallest tasks. And to the rest of my family for their love and support throughout, thank you. You have all played an integral role in my life for which I am forever grateful. Most of all, I would like to thank my beautiful wife, Abigail. You have been there for me in good times and bad. You have shown me love, support, patience, kindness and understanding. I look forward to our years ahead together.

Abstract

Particulate matter (PM) is defined by the Texas Commission on Environmental Quality (TCEQ) as "a mixture of solid particles and liquid droplets found in the air". These particles vary widely in size. Those particles that are less than 2.5 μm in aerodynamic diameter are known as Particulate Matter 2.5 or PM2.5. These particles are inhaled, and their health effects are still largely being studied. Past studies have assessed PM2.5 exposure of a population, yet individual exposure is more difficult to assess and may vary widely in a population. Recent studies have combined semiparametric models with kriging (Li et. al [2012]) to assess nitrogen dioxide exposure in California. These methods may prove valuable at predicting PM2.5 at unmonitored locations in El Paso and subsequently in assessing personal exposure to PM2.5 within our population.

Garcia (2010) provides us with a unique opportunity to estimate the spatial covariance of PM2.5 in the El Paso region. Past studies have established that PM2.5 varies spatially within a region based on local traffic variables (Smith et. al [2006]). Other studies have found meterological, variables such as wind speed play an important role in PM levels (Staniswalis et al.[2005]). First, we use meteorological variables to build a semiparametric model to estimate the mean PM2.5 at two monitored locations. Then in conjunction with traffic data and the spatial covariance structure of PM2.5, we use kriging of the residuals of the semiparametric models to predict PM2.5 at unmonitored locations.

Contents

Page

Acknowledgements v

Abstract vi

Table of Contents vii

Chapter

1 Introduction and Data 1

1.1 Particulate Matter in Ambient Air 1

1.2 Monitors for PM2.5 in El Paso 2

1.3 Meterological Data 6

1.4 Other Predictors for the Mean of log(PM2.5) 14

1.5 Data at Unmonitored Sites 15

1.6 Traffic Data 20

2 Parametric Regression 22

3 Nonparametric Regression Using Kernel Smoothing 28

3.1 Nonparametric Regression Using Kernel Smoothing 28

3.2 Kernels 29

3.3 Mean and Variance of $\hat{f}(x)$ 31

3.3.1 Mean of $\hat{f}(x)$ 31

3.3.2 Variance of $\hat{f}(x)$ 34

3.4 Asymptotic Mean Squared Error 35

3.5 Boundary Correction 36

3.5.1 Jackknifing 39

3.5.2 Local Linear Fitting 41

4 Nonparametric Regression using Splines 45

4.1 Smoothing Splines 45

4.2 B Splines . . . 49
4.3 P-Splines . . . 54
5 Additive Models . . . 56
5.1 Semiparametric Additive Models . . . 56
5.2 Fitting the Semiparametric Model by Backfitting . . . 57
5.3 Model Comparison . . . 60
5.3.1 Mean Square Prediction Error . . . 62
6 Semiparametric Modeling . . . 63
6.1 Semiparametric Model Selection . . . 63
6.1.1 Estimation at UTEP and Chamizal . . . 69
6.1.2 Estimation at D, J and K . . . 70
6.2 Comparison of UTEP and Chamizal Models . . . 72
7 Geostatistical Methods . . . 76
7.1 Inverse Distance Weighting . . . 76
7.2 The Multivariate Normal Distribution . . . 77
7.3 Estimation of Spatial Covariance . . . 78
7.3.1 Stationarity Assumptions . . . 78
7.3.2 The Exponential Model . . . 79
7.4 Kriging . . . 80
8 Application of Geostatistical Methods to Prediction of PM2.5 at Unmonitored Locations . . . 82
8.1 Inverse Distance Weighting . . . 82
8.2 Spatial Covariance of PM2.5 . . . 85
8.2.1 Fitting the Exponential Model of Correlation . . . 85
8.2.2 Residual Correlations and Covariances . . . 88
8.3 Kriging . . . 92
8.3.1 Traffic Data . . . 92
8.3.2 Kriging of the Residuals . . . 93

8.4 Summary 97
References 98
Curriculum Vitae 101

Chapter 1

Introduction and Data

1.1 Particulate Matter in Ambient Air

Particulate matter (PM) is defined by the Texas Commission on Environmental Quality (TCEQ) as "a mixture of solid particles and liquid droplets found in the air". These particles vary widely in size. Some are visible to the naked eye while others are visible only via an electron microscope. PM2.5 is defined as those particles that are less than 2.5 μm in aerodynamic diameter. These particles are airborne and inhaled by the population daily. Sources of these particles include vehicle exhaust, dust, residential and industrial activities.

Because PM2.5 is inhaled by the population, the impact on cardiovascular and respiratory health is of great concern. Researchers are trying to assess associations between PM2.5 exposure and adverse health consequences, such as asthma or cardiovascular hospital admissions, and mortality. However, a ubiquitous difficulty for these studies is in assessing PM2.5 exposure. Technology has yet to allow us to assess personal exposure for large scale studies. Past studies have assessed exposure often by averaging monitors in a specific region or zip code (if available). Better estimates of personal exposure could lead to a better understanding of the impact of PM2.5 on respiratory and cardiovascular health.

There are several air pollution monitors in El Paso, Texas. We are seeking to use meteorological variables to build a semiparametric model to estimate the temporal mean PM2.5 at a monitored location. Then in conjunction with traffic data and spatial correlation, use kriging of the residuals of this model to predict PM2.5 at an unmonitored location.

Past studies have used this approach with other pollutants such as nitrogen dioxide. Li et al. (2012) used cokriging of the residuals of land use regression models to predict nitrogen oxides in southern California. Models that relate to spatial data are often referred to as land-use regression models. Semiparametric land use regression models have been widely used in pollution studies. Using data from 2006, Gonzales et al. (2012) evaluated a semiparametric model that was developed in 1999 to describe the spatial-temporal gradient of nitrogen dioxide across El Paso County (Smith et al. [2006]). They found that despite the elapsed time period, the predicted nitrogen dioxide concentration gradients were similar.

1.2 Monitors for PM2.5 in El Paso

There are two sources of PM2.5 data: (i) Texas Commission on Environmental Quality (TCEQ) available through www.tceq.texas.gov, and (ii) data collected by Mario Garcia (2010) under the supervision of Dr. Wen-Whai Li, Department of Civil Engineering at UTEP. TCEQ data is publicly available data; see References for the full link. TCEQ operates air pollution monitors in El Paso County. TCEQ sites record various measurements of both meteorological and air pollution variables. Although a site may be run by TCEQ, it does not necessarily collect PM2.5 data, but it always collects meterological data. We use all TCEQ sites with PM2.5 data, namely, sites F (UTEP) and H (Chamizal). The data reported on by Garcia (2010) covered the time period August 2, 2006 to February 3, 2009. Garcia (2010) recorded at 8 sites: A, D, F, H, I J, K and L. Mario Garcia and Dr. Wen Whai-Li (Garcia 2010) obtained permission to place air pollution monitors at these locations. Some of the monitors were placed in El Paso Water Utilities (EPWU) facilities, because they provided secure locations well-suited to monitoring air pollution, and other monitors were placed at existing TCEQ sites. EPWU does not monitor air pollution as part of their operations. The sites pertinent to this study and the site operator are listed in Table 1.1 and individually described below.

Site A is located on the west side of El Paso. It is in a EPWU facility mostly surrounded by residential areas.

Site D is a TCEQ site, but does not collect PM2.5. This site is located in Northeast El Paso near Highway 54.

Site F corresponds to the UTEP site, which is also a TCEQ site providing PM2.5 data. This is located on the UTEP campus, which is close to Interstate 10.

Site H is the Chamizal site. It is also a TCEQ site providing PM2.5 data. The Chamizal site's region has the greatest traffic density. It is the closest site to a U.S. Mexico border crossing, and also the closest to the junction of Interstate 10 and Highway 54, referred to by resident's as the "Spaghetti Bowl".

Site I is located on the east side of El Paso in an EPWU facility. This site is located next to Album Park, one of the largest public parks in El Paso in the middle of a residential area.

Site J is a TCEQ site, but does not collect PM2.5. It is located in the Southeast of El Paso near Socorro, Texas. This site is also near the U.S. Mexico border.

Site K is also a TCEQ site, but does not collect PM2.5. It is located in El Paso's Lower Valley also near the U.S. Mexican border, close to Border Highway.

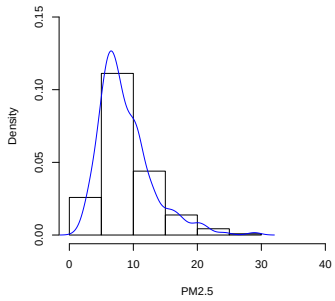
Site L is located in the Montana Vista community south of Highway 180 in an EPWU location.

Table 1.1: Sites of interest to this study, their single-letter code, and the source of available PM2.5 data.

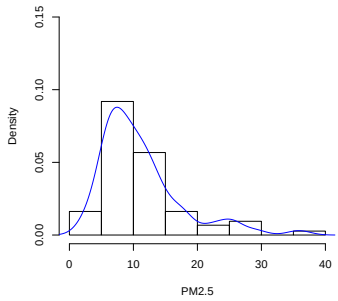
Site	Name	TCEQ	Garcia (2010)
A	Lindbergh		X
D	Skyline		X
F	UTEP	X	X
H	Chamizal	X	X
I	Album		X
J	Southeast		X
K	Lower Valley		X
L	Montana Vista		X

First, we describe the PM2.5 data obtained at TCEQ sites from August 8, 2006 until February 3, 2009, which corresponds to the time period studied in Garcia (2010). The TCEQ data were collected using an FRM sampler. At the UTEP TCEQ site, PM2.5 was recorded as a daily measurement every three days from August 8, 2006 to January 1, 2008. In other words, this is a daily measurement that occurs with two day gaps. From January 1, 2008 until February 3, 2009, PM2.5 was measured daily but with five day gaps instead of two. PM2.5 was recorded at the Chamizal site during this period as a daily measurement every six days. Both the Chamizal and UTEP PM2.5 levels showed right-skewness as depicted in Figure 1.1. The log transformation was effective in reducing this skewness as depicted in Figure 1.2.

There were a total of four missing observations at the UTEP site and five missing observations at the Chamizal site. The Chamizal site was not operational from June 25, 2007 to July 1, 2007. These were imputed using splines as depicted in Figure 1.3. See Chapter 4 for more information regarding splines.

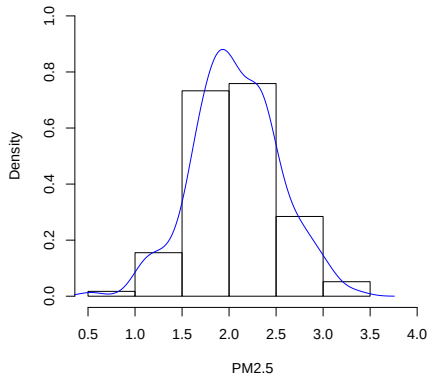


(a) UTEP PM2.5

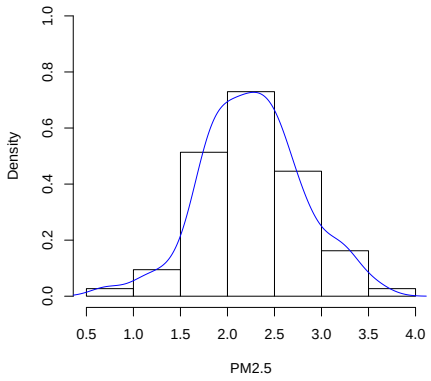


(b) Chamizal PM2.5

Figure 1.1: Density histograms of PM2.5 recorded at TCEQ sites: (a) UTEP and (b) Chamizal. Right-skewness is seen in both of these histograms.

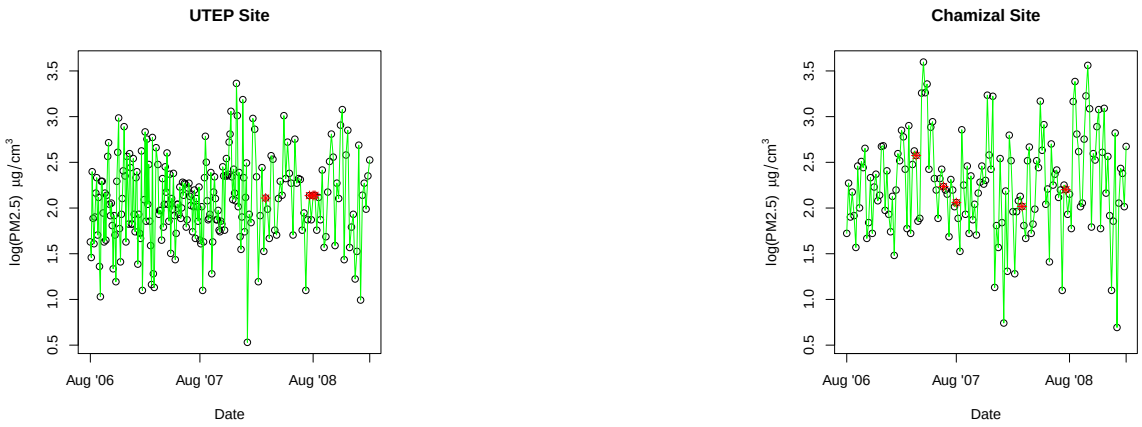


(a) UTEP log(PM2.5).



(b) Chamizal log(PM2.5).

Figure 1.2: Density histograms of log(PM2.5) recorded at TCEQ sites: (a) UTEP and (b) Chamizal.



(a) UTEP log(PM2.5).

(b) Chamizal log(PM2.5).

Figure 1.3: Time series plots of log(PM2.5) recorded at TCEQ sites: (a) UTEP and (b) Chamizal. The imputed values are shown in red.

1.3 Meterological Data

The above discussion focused on the PM2.5 data, and the location of the air pollution monitor sites. Since one specific aim of this thesis is to fit and validate semiparametric models for prediction of the temporal mean of PM2.5 by meterological variables, we now describe the meterological data available. Meterological data is available only from TCEQ. The available meterological data is summarized in Table 1.2. We obtained meterological data from TCEQ sites D, F, H, J, and K, but not A, I and L which are EPWU locations. We considered data in the time period from August 8, 2006 to February 3, 2009 to match the time period during which Garcia (2010) recorded his data.

Table 1.2: Meterological data available at each site: * indicates the site is a TCEQ site; X indicates the data is available. Meterological data is not available for EPWU sites A, I and L.

Site	Temperature	Wind Speed	Wind Direction	Humidity	Dew Point
D*	X	X	X		
F *	X	X	X	X	X
H *	X	X	X	X	X
J*	X	X	X		
K*	X	X	X		
Units	°F	mph	Degrees		°F

Temperature is measured hourly and recorded in Fahrenheit (°F) by TCEQ. Thus, there are 24 measurements of temperature every day, which we averaged to obtain average daily temperature:

$$X_1 = \frac{1}{24}\sum_{i=1}^{24} t_i.$$

Temperature follows a seasonal trend in El Paso as depicted in Figure 1.4.

(a) UTEP

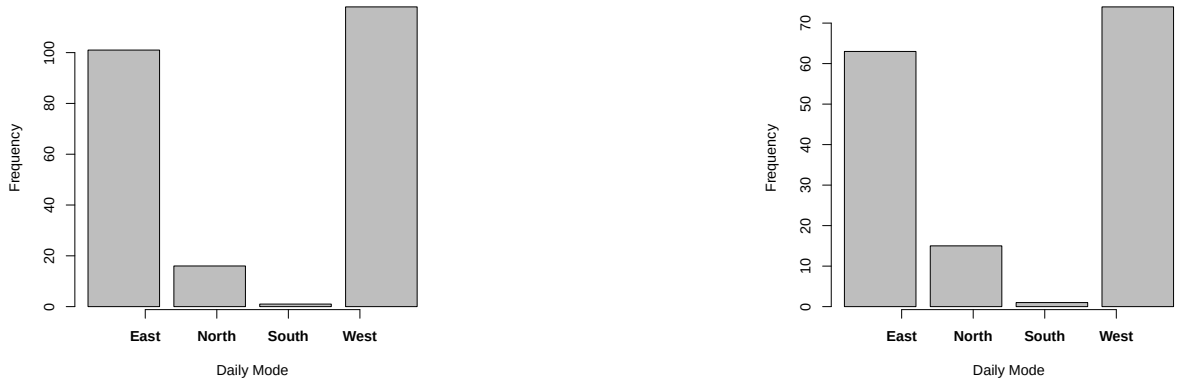
(b) Chamizal

Figure 1.4: Daily average temperature on days for which there is a PM2.5 measurement at the UTEP and Chamizal TCEQ sites.

During this time period, TCEQ measured wind direction hourly in degrees at all of their sites. We categorized the wind direction either as north, south, east or west depending on the angle measured. Letting θ denote the angle measured, the respective categories are defined as

$$\text{Direction Category} = \begin{cases} \text{North} & \text{if } \theta \in [315, 360] \cup [0, 45) \\ \text{East} & \text{if } \theta \in [45, 135) \\ \text{South} & \text{if } \theta \in [135, 225) \\ \text{West} & \text{if } \theta \in [225, 315). \end{cases}$$

Using this classification, we define the daily mode of wind direction as the most frequently occurring direction on a given day. Barcharts of the daily mode of wind direction are shown in Figure 1.5.



(a) UTEP

(b) Chamizal

Figure 1.5: Barcharts of daily modes of wind direction for the UTEP and Chamizal sites.

There was only one day on which the daily mode was south for both the UTEP and Chamizal sites. Considering this, three indicator variables were defined

$$X_2 = \begin{cases} 1 & \text{if the daily mode of wind direction is north or south} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if the daily mode of wind direction is east} \\ 0 & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if the daily mode of wind direction is west} \\ 0 & \text{otherwise.} \end{cases}$$

We also considered how much the wind direction is changing throughout the given day by looking at the daily entropy of the wind direction. We define the daily proportions

$$\begin{aligned} p_1 &= \frac{\text{number of north observations in the day}}{24} \\ p_2 &= \frac{\text{number of south observations in the day}}{24} \\ p_3 &= \frac{\text{number of east observations in the day}}{24} \\ p_4 &= \frac{\text{number of west observations in the day}}{24}. \end{aligned}$$

The daily entropy for the observed day is defined as

$$X_5 = -\sum_{j=1}^{4} p_j \log(p_j).$$

Entropy measures variability in wind direction. If the wind is constantly changing throughout the day, entropy will be higher than when the wind is blowing in the same direction all day. In fact, entropy is maximized when $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$. Box plots of entropy by the daily mode are shown in Figure 1.6. As Figure 1.6 depicts, east and west are the most common daily modes of wind direction, and the distribution of entropy is relatively similar. On the 14 days the wind direction mode was north, the mean entropy was considerably higher.

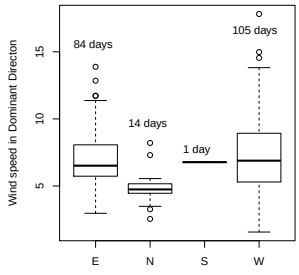
(a) UTEP

(b) Chamizal

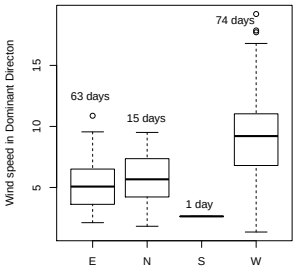
Figure 1.6: Boxplots of entropy of wind direction for the UTEP and Chamizal sites.

TCEQ also measured wind speed hourly in miles per hour during this time period. We defined the variable X_6 to be the average of the hourly measurements in the direction of the daily mode. For example, if the wind was blowing north for 23 hours of the day and west for only one hour, then the wind speed for this day is the average of the 23 measurements when the wind was blowing north. Boxplots of wind speed in the dominant direction are shown in Figure 1.7.

Relative humidity is measured as a percentage of moisture in the air, whereas dew point values represent the temperature at which the air will no longer be able to hold moisture. Both of these variables are measured daily. These values are highly correlated. TCEQ measures dew point in degrees Fahrenheit. The UTEP and Chamizal sites are relatively close to each other, and thus the meterological data is very correlated. This is depicted in Figure 1.8.



(a) UTEP



(b) Chamizal

Figure 1.7: Boxplots of wind speed in the dominant direction for the UTEP and Chamizal sites.

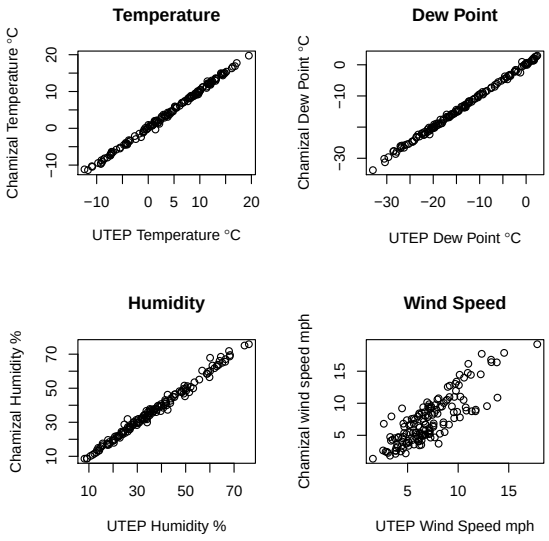


Figure 1.8: Comparison of meterological measurements taken at the UTEP and Chamizal sites.

Denoting the hourly measurements of humidity for a given day as $h_1, h_2, \ldots, h_{24}$ and the hourly dew point measurements for that respective day as $d_1, d_2, \ldots, d_{24}$, the variables X_7 and X_8 are defined as

$$X_7 = \frac{1}{24}\sum_{i=1}^{24} h_i$$
$$X_8 = \frac{1}{24}\sum_{i=1}^{24} d_i.$$

Because dew point and humidity are only available at the UTEP and Chamizal sites the average of the UTEP and Chamizal sites are used to fill in the dew point and humidity at other sites. Dew point and temperature were converted to degrees Celsius. Figure 1.8 suggests that dew point and humidity may correlate well throughout El Paso as they do correlate well between the UTEP and Chamizal sites. Summary statistics for PM2.5 at the UTEP and Chamizal sites and the meterological covariates are provided in Tables 1.3 through 1.6. Chamizal has higher levels of PM2.5 and greater variability than UTEP.

Table 1.3: UTEP TCEQ

	N	Minimum	5th percentile	25th percentile	Median	Mean
log(PM2.5)	236.00	0.53	1.28	1.79	2.07	2.07
X_1 Temperature	236.00	1.04	5.33	13.43	19.15	18.65
X_5 Entropy	236.00	0.00	0.00	0.44	0.82	0.73
X_6 Wind Speed	236.00	1.57	3.39	5.17	6.53	6.99
X_8 Dew Point	236.00	-18.78	-12.19	-5.60	0.90	1.57
X_7 Humidity	236.00	8.17	14.00	26.10	35.38	37.30

Table 1.4: UTEP TCEQ Continued.

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
log(PM2.5)	2.35	2.84	3.36	0.56	0.45
X_1 Temperature	24.87	29.04	33.73	11.44	7.71
X_5 Entropy	1.03	1.27	1.33	0.60	0.40
X_6 Wind Speed	8.54	11.82	17.82	3.38	2.62
X_8 Dew Point	9.75	14.97	18.35	15.35	8.84
X_7 Humidity	47.81	64.65	76.12	21.71	15.26

Table 1.5: Chamizal TCEQ

	N	Minimum	5th percentile	25th percentile	Median	Mean
log(PM2.5)	153.00	0.69	1.45	1.92	2.24	2.26
X_1 Temperature	153.00	2.85	4.97	13.27	18.80	18.47
X_5 Entropy	153.00	0.00	0.00	0.51	0.90	0.78
X_6 Wind Speed	153.00	1.35	2.54	4.67	6.70	7.38
X_8 Dew Point	153.00	-19.58	-12.89	-5.99	-0.34	0.98
X_7 Humidity	153.00	8.60	13.33	24.48	34.59	36.62

Table 1.6: Chamizal TCEQ Continued.

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
log(PM2.5)	2.58	3.22	3.60	0.66	0.53
Temperature	24.58	29.19	33.97	11.31	7.70
Entropy	1.12	1.29	1.38	0.61	0.40
Wind Speed	9.36	14.60	19.21	4.69	3.69
Dew Point	9.62	15.58	17.15	15.61	9.29
Humidity	46.83	66.48	75.84	22.35	16.10

1.4 Other Predictors for the Mean of log(PM2.5)

To account for the seasonal trend of PM2.5 and the population growth during the time period, we considered time in two fashions, one linear predictor corresponding to the day of the time period, ranging from 1 to 921 and a periodic time period that corresponds to the day of the year, ranging from 1 to 365. That is, the covariates X_8 and X_9 are defined as

$$X_9 = \text{day of time period } (1 - 921),$$
$$X_{10} = \text{day of year } (1 - 365).$$

The training data were divided into two categories of dates: August 1, 2006 to July 31, 2007 and August 1, 2007 to July 31, 2008. We also wanted to consider whether the measured day was a weekend day. We defined three additional indicator variables as

$$X_{11} = \begin{cases} 1 & \text{if the measurement date is between August 1, 2006-July 31 2007} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_{12} = \begin{cases} 1 & \text{if the measurement date is between August 1, 2007-July 31 2008} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$X_{13} = \begin{cases} 1 & \text{if the measurement date is a Saturday or Sunday} \\ 0 & \text{otherwise.} \end{cases}$$

A list of all the variables with a brief description is provided in Table 1.7.

Table 1.7: A list of all the covariates with a brief description of each.

Variable	Description
X_1	Average Daily Temperature
X_2	North or South wind direction
X_3	East wind direction
X_4	West wind direction
X_5	Entropy
X_6	Wind Speed in the direction of the daily mode
X_7	Average daily humidity
X_8	Average daily dew point
X_9	Day of time period (1-921)
X_{10}	Day of year (1-365)
X_{11}	August 1, 2006-July 31, 2007
X_{12}	August 1, 2007-July 31, 2008
X_{13}	Weekend day

1.5 Data at Unmonitored Sites

As described in Table 1.1, PM2.5 was recorded at sites D, J and K, thanks to the efforts of Garcia (2010). They used dichotomous samplers to record PM2.5, while TCEQ records PM2.5 using an FRM sampler. TCEQ records the covariates temperature, wind speed, and wind direction. The number of measurements of PM2.5 differs from that of the covariates because there is much more missing data in PM2.5. In Tables 1.8 through 1.13, N refers to the number of days the variable is recorded.

Table 1.8: Site D Garcia (2010)

	N	Minimum	5th percentile	25th percentile	Median	Mean
log(PM2.5)	102.00	0.46	1.07	1.45	1.75	1.85
Temperature	128.00	-1.58	4.89	11.53	18.48	17.89
Entropy	128.00	0.00	0.35	0.71	1.06	0.95
Wind Speed	128.00	2.20	2.49	4.01	5.74	6.91

Table 1.9: Site D Garcia (2010) Continued

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
log(PM2.5)	2.23	2.77	3.27	0.78	0.56
Temperature	24.10	30.77	33.35	12.57	8.08
Entropy	1.19	1.32	1.36	0.48	0.32
Wind Speed	8.87	16.19	20.82	4.86	4.10

Table 1.10: Site J Garcia (2010)

	N	Minimum	5th percentile	25th percentile	Median	Mean
log(PM2.5)	72.00	0.02	0.99	1.84	2.08	2.11
Temperature	128.00	-1.21	5.12	10.77	17.86	17.39
Entropy	128.00	0.00	0.40	0.84	1.05	0.99
Wind Speed	128.00	1.30	1.67	3.36	4.87	5.38

Table 1.11: Site J Garcia (2010) Continued

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
log(PM2.5)	2.38	3.09	4.19	0.54	0.60
Temperature	23.83	29.60	32.80	13.06	8.07
Entropy	1.22	1.33	1.36	0.39	0.30
Wind Speed	6.64	11.58	15.40	3.27	2.89

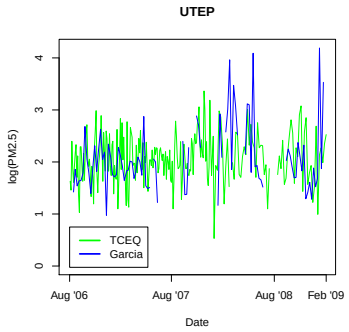
Table 1.12: Site K Garcia (2010)

	N	Minimum	5th percentile	25th percentile	Median	Mean
log(PM2.5)	99.00	-1.51	0.48	1.54	1.92	1.92
Temperature	124.00	-1.05	4.73	10.52	17.69	17.26
Entropy	124.00	0.00	0.29	0.76	1.04	0.94
Wind Speed	124.00	1.37	1.73	3.60	5.65	6.38

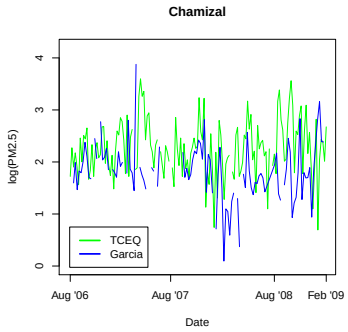
Table 1.13: Site K Garcia (2010)Continued

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
log(PM2.5)	2.44	3.14	4.17	0.90	0.88
Temperature	23.88	29.67	33.92	13.36	8.12
Entropy	1.18	1.33	1.37	0.42	0.33
Wind Speed	8.09	13.92	19.99	4.49	3.59

Garcia (2010) also recorded PM2.5 at the UTEP and Chamizal sites. Garcia (2010) recorded weekly averages of PM2.5 using dichotomous samplers. The calendar date assigned to the weekly averages was the beginning of the 7 day monitoring window. Thus, these recordings rarely aligned with TCEQ measurements. The UTEP site overlapped with 9 TCEQ measurements, and the Chamizal site only overlapped with 4 TCEQ measurements. Figure 1.9 shows time series plots from August 2006 to February 2009 of log(PM2.5) at the UTEP and Chamizal sites. These figures show both the data as recorded by Garcia (2010) and by TCEQ. As Figure 1.9 depicts, the seasonal trend for these measurements are similar but the measurement days rarely overlap. There is also much more missing data in Garcia (2010) and occasionally large deviations from the TCEQ data.



(a) UTEP



(b) Chamizal

Figure 1.9: log(PM2.5) as measured by TCEQ in green and Garcia (2010) in blue at (a) UTEP and (b) Chamizal sites.

Tables 1.14 through 1.17 show summary statistics for both Garcia (2010) and TCEQ of these measurements. Figure 1.10 plots all the log(PM2.5) data as recorded by Garcia. PM.25 seems to follow a seasonal trend with some sites particularly having higher or lower log(PM2.5) levels and variability than other sites.

Table 1.14: UTEP log(PM2.5) as recorded by TCEQ and Garcia (2010).

	N	Minimum	5th percentile	25th percentile	Median	Mean
Garcia	92.00	0.83	1.29	1.68	1.92	2.06
TCEQ	236.00	0.53	1.28	1.79	2.07	2.07

Table 1.15: UTEP log(PM2.5) as recorded by TCEQ and Garcia (2010).

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
Garcia	2.28	3.28	4.19	0.59	0.64
TCEQ	2.35	2.84	3.36	0.56	0.45

Table 1.16: Chamizal log(PM2.5) as recorded by TCEQ and Garcia (2010).

	N	Minimum	5th percentile	25th percentile	Median	Mean
Garcia	103.00	0.10	0.95	1.52	1.80	1.83
TCEQ	153.00	0.69	1.45	1.92	2.24	2.26

Table 1.17: Chamizal log(PM2.5) as recorded by TCEQ and Garcia (2010).

	75th percentile	95th percentile	Maximum	IQR	Standard Deviation
Garcia	2.12	2.80	3.88	0.60	0.55
TCEQ	2.58	3.22	3.60	0.66	0.53

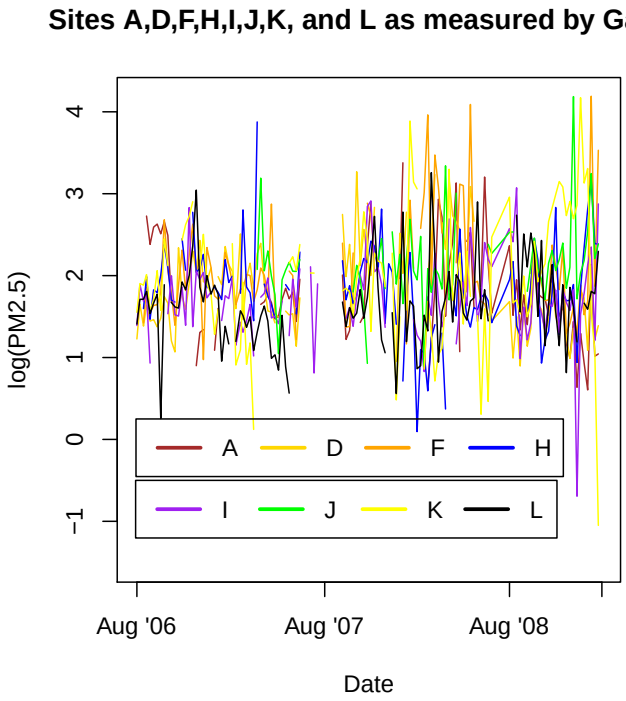


Figure 1.10: log(PM2.5) from August 2006 - February 3, 2009 as recorded by Garcia (2010).

1.6 Traffic Data

The Metropolitan Planning Office (MPO) measures the vehicle miles traveled or VMT's through the use of small devices called traffic counters. VMT's are generally recorded using a time period and a buffer radius. Olvera et al. (2012) found a buffer radius of 1 km "as best suitable for capturing the PM2.5 spatial variability". Since PM2.5 is measured at sites A, D, F, H, I, J, K and L, the VMT's and summary statistics of these measurements are shown in Tables 1.18 and 1.19. Site L does not have VMT measurement at this buffer radius.

Table 1.18: The VMT at a 1 km buffer radius for 2007 by site.

Site	2007 VMT
A	2530.85
D	2257.25
F	1550.72
H	5064.86
I	352.07
J	51.26
K	2459.50
L	Not Available

Table 1.19: Summary statistics regarding the 2007 VMT's.

Statistic	
N	7.00
Minimum	51.26
25th percentile	951.40
Median	2257.25
75th percentile	2495.17
Maximum	5064.86

Chapter 2

Parametric Regression

Suppose we are interested in a random variable Y that is related in the following form to a predictor variable x:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1 are constants and $\epsilon \sim N(0, \sigma^2)$. In this situation the relationship between x and $\mathbb{E}(Y|x)$ is linear in nature. Suppose we have n independent observations of data from this population (x_i, Y_i) for $i = 1, 2 \ldots n$. A scatterplot of this data may resemble Figure 2.1. Figure 2.1 shows there may be several lines that can serve to estimate the relationship between x and Y. However, some lines appear to better capture the linear nature of this relationship.

A common approach to this scenario is to define the estimator of the regression line by finding the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$ is minimized with respect to β_0, β_1. This approach is referred to as simple linear regression using least squares. Simple refers to the fact that there is only one predictor variable, while linear refers to the fact that $\mathbb{E}(Y|x)$ is linear in the parameters β_0 and β_1. For example, $Y = \beta_0 + \beta_1 x^2 + \epsilon$ is a simple linear model because $\mathbb{E}(Y|x) = \beta_0 + \beta_1 x^2$, which is linear in the parameters β_0 and β_1. We define $\hat{\boldsymbol{\beta}}^T = (\beta_0, \beta_1)$ to be

$$\hat{\boldsymbol{\beta}} = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.1}$$

As we will see, equation (2.1) has an explicit solution. However, let us first turn our attention to the case where we have more than one independent variable known as multiple linear regression. Now suppose we have a response variable that is linearly related to several predictor variables $x_1, \ldots, x_{p-1}$ as

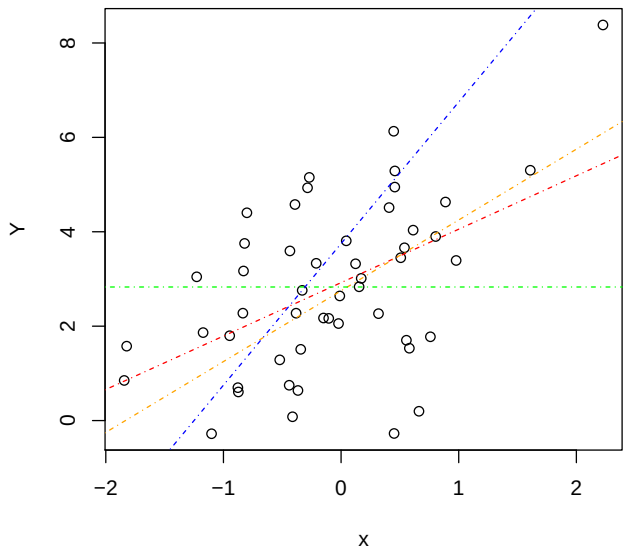


Figure 2.1: A linear relationship between x and Y. The red line is the line yielded by the least squares approach while the blue and orange lines were arbitrarily chosen. The green line is simply $\bar{Y}$ with a slope of zero.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and $\beta_0, \beta_1, \ldots \beta_{p-1}$ are fixed parameters. Notice that p is the number of parameters. To extend (2.1) to a situation when we have $p-1$ predictor variables, let us first introduce the vectors $\boldsymbol{Y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ as follows:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Also, let us introduce the matrices X and I_n as

$$X = \begin{pmatrix} 1 & x_{1,1} & \dots x_{1,p-1} \\ 1 & x_{2,1} & \dots x_{2,p-1} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots x_{n,p-1} \end{pmatrix}_{n \times p,} \qquad I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}_{n \times n.}$$

Then both $\boldsymbol{\epsilon}, \boldsymbol{Y} \in \mathbb{R}^n$ where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$ and $\boldsymbol{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$. If we are considering the scenario in which there are only two independent variables the relationship between the response and the predictors is represented by a plane. If there are more than two independent variables then the relationship is represented by a hyperplane. Naturally, we can extend the definition of $\hat{\boldsymbol{\beta}}$ in (2.1) as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} ||\boldsymbol{Y} - X\boldsymbol{\beta}||^2, \text{ where } ||\boldsymbol{Y} - X\boldsymbol{\beta}||^2 = (\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}). \tag{2.2}$$

We derive the explicit solution to (2.2). Consider

$$\begin{aligned} ||\boldsymbol{Y} - X\boldsymbol{\beta}||^2 &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\beta}^T X^T\boldsymbol{Y} - \boldsymbol{Y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T X^T\boldsymbol{Y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}, \end{aligned}$$

because $\boldsymbol{\beta}^T X^T\boldsymbol{Y}$ is a scalar $\boldsymbol{\beta}^T X^T\boldsymbol{Y} = \boldsymbol{Y}^T X\boldsymbol{\beta}$. If we let $Q(\boldsymbol{\beta}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T X^T\boldsymbol{Y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$, then minimizing Q by differentiating with respect to $\boldsymbol{\beta}$ and solving for $\hat{\boldsymbol{\beta}}$ in

$$\frac{\partial Q}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = -2X^T\boldsymbol{Y} + 2X^T X\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

From here conclude that

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T\boldsymbol{Y},$$

and the vector of fitted values $\hat{\boldsymbol{Y}}$ is

$$\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T\boldsymbol{Y} = H\boldsymbol{Y},$$

where $H = X(X^T X)^{-1} X^T$. The matrices H and $I_n - H$ are idempotent and symmetric.

Mean and Variance of $\hat{\boldsymbol{\beta}}$

The mean of $\hat{\boldsymbol{\beta}}$ can be found by

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[(X^T X)^{-1} X^T \boldsymbol{Y}] = (X^T X)^{-1} X^T \mathbb{E}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta}.$$

Thus, $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. Similarily we can find the variance of $\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}((X^T X)^{-1} X^T \boldsymbol{Y}) = (X^T X)^{-1} X^T \mathrm{Var}(\boldsymbol{Y})((X^T X)^{-1} X^T)^T \\ &= (X^T X)^{-1} X^T \big(\sigma^2 I_n\big) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.\end{aligned}$$

Model Selection

The variation in the response variable Y can be broken up about its sample mean as follows:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2. \tag{2.3}$$

The term $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is the squared distance between each observed value and its predicted value. We define the terms SSE, SSR and SST as

$$\mathrm{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad \mathrm{SSE} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \quad \mathrm{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2,$$

so we can rewrite (2.3) as

$$SST = \mathrm{SSE} + \mathrm{SSR}.$$

The coefficient of determination, R^2 is defined as

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}.$$

This is also equivalently defined as $R^2 = \frac{\text{SSR}}{\text{SST}}$. Notice that $R^2 \in [0, 1]$, and larger values of R^2 are preferred. When the predictor variables are correlated the model will suffer from large variance, although R^2 will increase as more variables are included in the model. A tradeoff is needed between increasing R^2 and limiting the number of variables in the model. An alternative to R^2 is the Akaike Information Criterion (AIC).

Akaike Information Criterion (AIC)

To introduce the Akaike Information Criterion (AIC), we first introduce the predictive sum of squares and generalized cross validation. The predictive sum of squares or $PRESS$ statistic is defined as

$$PRESS = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2$$

where $\hat{Y}_{i(i)}$ denotes the predicted $E(Y_i|x_i)$ given that the model has been fitted without the i^{th} observation. It can be shown that this is equivalent to

$$PRESS = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \right)^2,$$

where h_{ii} is the ith diagonal element of H (Kutner, Nachtsheim and Nester 2004). Generalized cross validation (GCV) replaces the h_{ii} elements with the average of the diagonal elements of H:

$$\bar{h} = \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{\text{tr}(H)}{n}.$$

Because H is idempotent, $\text{tr}(H) = \text{rank}(H)$. Thus, $\bar{h} = \frac{\text{tr}(H)}{n} = \frac{\text{rank}(H)}{n} = \frac{p}{n}$. The GCV of a model is defined as

$$GCV = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - \bar{h}} \right)^2 = \frac{SSE}{(1 - \frac{tr(H)}{n})^2} = \frac{SSE}{(1 - \frac{p}{n})^2}.$$

When $x \approx 0$, $\log(1 - x) \approx -x$, so if $n >> p$ we can establish that

$$\log(GCV) = \log \text{SSE} - 2\log(1 - \frac{p}{n})$$
$$\approx \log \text{SSE} + 2\frac{p}{n}.$$

Finally, AIC is defined as

$$AIC = n\ \log(SSE) + 2p.$$

Notice, it can be seen in two parts: a goodness-of-fit part in the $n\ \log(SSE)$ and a penalty for excessively large models.

The Akaike Information Criterion is very similar to the Bayesian Information Criterion (BIC) defined as

$$BIC = n\log(SSE) + p\log(n).$$

Both the AIC and the BIC have certain advantages and disadvantages. They are both widely used in stepwise variable selection. The important thing to realize is that lower $PRESS$ is preferable. Lower $PRESS$ generally implies lower GCV, which implies lower AIC and BIC. Thus, lower AIC and BIC are desirable.

Chapter 3

Nonparametric Regression Using Kernel Smoothing

3.1 Nonparametric Regression Using Kernel Smoothing

Suppose we have a random variable Y that is related to x in the following form:

$$Y = f(x) + \epsilon, \tag{3.1}$$

where f is a smooth function of x and ϵ is a noise random variable with constant variance σ^2 and mean 0. In simple linear regression f is assumed to be a known function of x that is linear in the unknown parameters. In nonparametric regression, we will make no such assumptions. Instead we seek to use the data, in this case $(x_i, Y_i) \quad i = 1, 2, \ldots n$ to estimate the function f without imposing a functional form.

If a parametric approach is applied with a misspecified model, results can be grossly misleading. However, as we will see, if a valid parametric model is applied, it is superior to our nonparametric estimate of f. For example, suppose $x \in [-1, 1]$ and $Y = f(x) + \epsilon$ and f is the function $f(x) = x^2$. Applying a simple linear model to this situation would be misleading; see Figure 3.1.

Nonparametric regression allows for any smooth functional form of f, usually quantified by the number of continuous derivatives that exist for the function $f(x)$. This is especially useful when variables do not appear to be related by a simple functional form. There

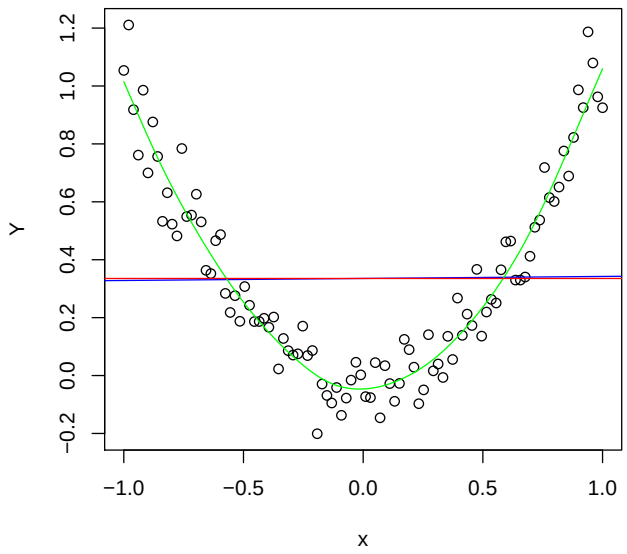


Figure 3.1: This figure illustrates applying a linear model $Y = \alpha + \beta x + \epsilon$ to the variables x and Y. The blue line is the ordinary least squares line, while the green line is a nonparametric fit. Clearly, the green line captures the quadratic relationship between x and Y.

are several methods of nonparametric regression. Here, we will focus on kernel regression. First, we will introduce some basic definitions.

3.2 Kernels

Generally speaking kernel functions have compact support on $[-1, 1]$, are symmetric functions and integrate to 1. The objective of a kernel function is to weight each observation. These properties are very useful for this purpose. For example, the Epanechnikov kernel is defined as $w(u) = \frac{3}{4}(1 - u^2)\mathbf{I}_{[-1,1]}(u)$. This is symmetric about 0 and integrates to one. Furthermore, notice that the largest weighted point is $(0, w(0))$. This implies that if we

consider $w(\frac{x-x_i}{b})Y_i$, where $b > 0$ is called the bandwidth, data points closer to x receive more weight than points further from x. The bandwidth serves to create a window about a targeted point x, such that when $x - x_i > b$, then $w(\frac{x-x_i}{b}) = 0$. Thus, points outside the window receive no weight. This discussion follows from Wand and Jones (1995).

The j^{th} moment of a kernel w_j is defined as

$$w_j = \int_{-\infty}^{\infty} u^j \, w(u) \, \mathrm{d}u \qquad\qquad j = 1, 2, \ldots.$$

The order of a kernel corresponds to the first nonzero moment. For example, the Epanechnikov kernel is a 2nd-order kernel because

$$\int_{-\infty}^{\infty} u \, w(u) \mathrm{d}u = \frac{3}{4} \int_{-1}^{1} u \, (1 - u^2) = 0,$$

but $\int_{-\infty}^{\infty} u^2 \, w(u) \mathrm{d}u = \frac{3}{4} \int_{-1}^{1} u^2 \, (1 - u^2) \mathrm{d}u > 0.$

Kernel functions of order $\nu > 2$ are called higher order kernels. Higher order kernels are sometimes referred to as "bias-reducing" kernels. We will see why this is the case later.

Returning to (3.1), suppose we wish to estimate the function f, but do not want to constrain f to a specific functional form, such as a linear function. We can use the estimator $\hat{f}(x)$ defined as

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} w(\frac{x-x_i}{b}) Y_i}{\sum_{i=1}^{n} w(\frac{x-x_i}{b})}. \tag{3.2}$$

Suppose we want to estimate f at the points $t_1, t_2 \ldots t_m$. Consider the vector $\hat{\boldsymbol{f}}(\boldsymbol{t})$ and the matrix S as follows.

$$\hat{\boldsymbol{f}}(\boldsymbol{t}) = \begin{pmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \vdots \\ \hat{f}(t_m) \end{pmatrix} \qquad S = \begin{pmatrix} \frac{w(\frac{t_1-x_1}{b})}{\sum_{i=1}^{n} w(\frac{t_1-x_i}{b})} & \frac{w(\frac{t_1-x_2}{b})}{\sum_{i=1}^{n} w(\frac{t_1-x_i}{b})} & \cdots & \frac{w(\frac{t_1-x_n}{b})}{\sum_{i=1}^{n} w(\frac{t_1-x_i}{b})} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{w(\frac{t_m-x_1}{b})}{\sum_{i=1}^{n} w(\frac{t_m-x_i}{b})} & \frac{w(\frac{t_m-x_2}{b})}{\sum_{i=1}^{n} w(\frac{t_m-x_i}{b})} & \cdots & \frac{w(\frac{t_m-x_n}{b})}{\sum_{i=1}^{n} w(\frac{t_m-x_i}{b})} \end{pmatrix}_{m \times n}$$

From (3.2),

$$\hat{\boldsymbol{f}}(\boldsymbol{t}) = S\boldsymbol{Y}. \tag{3.3}$$

Notice the similarity between (3.3) and the least squares fitted values $\hat{\boldsymbol{Y}} = H\boldsymbol{Y}$. However, the matrix S does not share the same properties as H. In fact, S is not idempotent, and is only symmetric when $m = n$ and $\boldsymbol{t} = \boldsymbol{x}$. (Hastie and Loader [1993])

3.3 Mean and Variance of $\hat{f}(x)$

3.3.1 Mean of $\hat{f}(x)$

To find the expected value of $\hat{f}(x)$, assume that f has at least k continuous derivatives, and w is a kernel of order k. Assume that all the $x_i \in (0, 1)$ are on a regular grid. The expected value of $\hat{f(x)}$ can be found as follows.

$$\mathbb{E}(\hat{f}(x)) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} w(\frac{x-x_i}{b})Y_i}{\sum_{i=1}^{n} w(\frac{x-x_i}{b})}\right)$$

$$= \frac{\frac{1}{nb}\sum_{i=1}^{n} w(\frac{x-x_i}{b})f(x_i)}{\frac{1}{nb}\sum_{i=1}^{n} w(\frac{x-x_i}{b})}$$

For large n and small b the denominator can be approximated as

$$\frac{1}{b}\left[\frac{1}{n}\sum_{i=1}^{n} w(\frac{x-x_i}{b})\right] \approx \frac{1}{b}\int_0^1 w\Big(\frac{x-v}{b}\Big)\,\mathrm{d}v.$$

This follows from the Riemann sum approximation by an integral. Substituting $u = \frac{x-v}{b}$

$$\frac{1}{b}\int_{\frac{x}{b}}^{\frac{x-1}{b}} w(u)(-b)\,\mathrm{d}u = \int_{\frac{x-1}{b}}^{\frac{x}{b}} w(u)\,du.$$

Notice that for $x \in (b, 1-b)$, $\frac{x}{b} > 1$ and $\frac{x-1}{b} < -1$. Thus, we can break up this integral as follows:

$$\int_{\frac{x-1}{b}}^{\frac{x}{b}} w(u)\, du = \int_{\frac{x-1}{b}}^{-1} w(u)\, du + \int_{-1}^{1} w(u)\, du + \int_{1}^{\frac{x}{b}} w(u)\, du$$

Assuming the kernel has compact support on $[-1, 1]$, the first and last integrals in this expression are 0. So we can write

$$\frac{1}{nb} \sum_{i=1}^{n} w(\frac{x - x_i}{b}) \approx \int_{-1}^{1} w(u)\, du = 1.$$

Thus, we can approximate the denominator of $\mathbb{E}(\hat{f}(x))$ with 1.

Similarly, for the numerator

$$\frac{1}{nb}\left[\sum_{i=1}^{n} w\Big(\frac{x - x_i}{b}\Big) f(x_i)\right] \approx \frac{1}{b} \int_{0}^{1} w\Big(\frac{x - v}{b}\Big) f(v) dv.$$

Substituting $u = \frac{x-v}{b}$,

$$\frac{1}{b} \int_{0}^{1} w\Big(\frac{x - v}{b}\Big) f(v) dv = \int_{\frac{x-1}{b}}^{\frac{x}{b}} w(u) f(x - ub)\, du.$$

For $x \in (b, 1-b)$, breaking this up as we did for the denominator,

$$\int_{\frac{x-1}{b}}^{\frac{x}{b}} w(u) f(x - ub) du = \int_{\frac{x-1}{b}}^{-1} w(u) f(x - ub) du + \int_{-1}^{1} w(u) f(x - ub) du + \int_{1}^{\frac{x}{b}} w(u) f(x - ub) du.$$

Again, the first and last integrals are 0 so this becomes

$$\int_{-1}^{1} w(u) f(x - ub) du.$$

Assuming $f(x)$ has k continuous derivatives, we can expand $f(x-ub)$ in a Taylor series about the point x to obtain

$$\int_{-1}^{1} w(u)f(x-ub)du$$

$$= \int_{-1}^{1} w(u)\left[f(x) + f'(x)(-ub) + f''(x)\left(\frac{(-ub)^2}{2}\right) + \ldots + f^{(k)}(x)\left(\frac{(-ub)^k}{k!}\right)\right]du + R(n,b)$$

where $R(n,b)$ is some small remainder which is a function of n,b and x. Further simplification shows that $\int_{-1}^{1} w(u)f(x-ub)\,du$ is about equal to

$$f(x)\int_{-1}^{1} w(u)du + f'(x)(-b)\int_{-1}^{1} u\,w(u)du + f''(x)\frac{(-b)^2}{2}\int_{-1}^{1} u^2\,w(u)du + \ldots$$

$$\ldots + f^{(k)}(x)\left(\frac{(-b)^k}{k!}\right)\int_{-1}^{1} u^k\,w(u)du + R(n,b).$$

Since w integrates to 1, the first term in this expression $f(x)\int_{-1}^{1} w(u)du = f(x)$. Because we have assumed that w is of order k we can conclude that $\int_{-1}^{1} u\,w(u)du$, $\int_{-1}^{1} u^2\,w(u)du, \ldots, \int_{-1}^{1} u^{k-1}w(u)du$ are all 0. Thus, we are left with

$$\int_{-1}^{1} w(u)f(x-ub)\,du = f(x) + f^{(k)}(x)\left(\frac{(-b)^k}{k!}\right)w_k + R(n,b).$$

We have shown that the expected value of $\hat{f}(x)$ is

$$\mathbb{E}(\hat{f}(x)) \approx f(x) + f^{(k)}(x)\left(\frac{(-b)^k}{k!}\right)w_k.$$

Notice that $\hat{f}(x)$ is a biased estimator for $f(x)$. The asymptotic bias is

$$\text{ABias}[\hat{f}(x)] = f^{(k)}(x)\left(\frac{(-b)^k}{k!}\right)w_k.$$

We'll denote $\frac{f^{(k)}(x)}{k!}w_k$ by C_1 and therefore the asymptotic bias can be written $(-b)^k C_1$ ignoring smaller order terms. Now, we can write the asymptotic bias

$$\text{ABias}[\hat{f}(x)] = (-b)^k C_1.$$

Since this is an increasing function of b, one might suggest reducing the bandwidth to minimize bias. Although this indeed reduces bias, we will see that shrinking the bandwidth will enlarge the variance of $\hat{f}(x)$.

3.3.2 Variance of $\hat{f}(x)$

Now, consider the variance of $\hat{f}(x)$,

$$\text{Var}(\hat{f(x)}) = \text{Var}\left(\frac{\frac{1}{nb}\sum_{i=1}^{n} w(\frac{x-x_i}{b})Y_i}{\frac{1}{nb}\sum_{i=1}^{n} w(\frac{x-x_i}{b})}\right)$$

$$= \left(\frac{1}{\frac{1}{nb}\sum_{i=1}^{n} w(\frac{x-x_i}{b})}\right)^2 \frac{1}{n^2b^2}\sum_{i=1}^{n} w^2\Big(\frac{x-x_i}{b}\Big)\text{Var}(Y_i)$$

The term $\left(\frac{1}{\frac{1}{nb}\sum_{i=1}^{n} w(\frac{x-x_i}{b})}\right)^2$ is approximately 1 as discussed when we found the mean of $\hat{f}(x)$. Recall that $\text{Var}(Y_i) = \sigma^2$, and thus this expression is approximately

$$\approx \frac{\sigma^2}{nb}\Big[\frac{1}{nb}\sum_{i=1}^{n} w^2\Big(\frac{x-x_i}{b}\Big)\Big].$$

Making similar integral approximations as we made when finding the mean of $\hat{f}$ for $x \in [b, 1-b]$ and assuming large n this is approximately

$$\approx \frac{\sigma^2}{nb}\Big[\int_{-1}^{1} w^2\Big(u)du\Big].$$

We will write the constant $\sigma^2 \int_{-1}^{1} w^2(u)du$ as C_2. Thus, the asymptotic variance of $\hat{f}(x)$ is

$$\text{AVar}(\hat{f}(x)) = \frac{C_2}{nb}$$

ignoring smaller order terms. Notice that as $b \to 0$, $\text{Var}(\hat{f}) \to \infty$. Thus, choosing b very small will reduce bias but increase variance, while choosing b very large will reduce variance but increase bias. Naturally, this leads us to consider the mean squared error of $\hat{f}(x)$.

3.4 Asymptotic Mean Squared Error

The asymptotic mean squared error of $\hat{f}(x)$ is $\text{AMSE}(\hat{f}(x)) = \text{AVar}(\hat{f}(x)) + \text{ABias}^2(\hat{f}(x))$. Thus, the AMSE of $\hat{f}(x)$ is given by

$$\text{AMSE}(\hat{f}(x)) = b^{2k}C_1^2 + \frac{C_2}{nb}.$$

Now, we can minimize AMSE by differentiating with respect to b as follows

$$\frac{d}{db}[\text{AMSE}]\bigg|_{b=b_0} = 0$$
$$2kb_0^{2k-1}C_1^2 - \frac{C_2}{nb_0^2} = 0.$$

Solving for b_0

$$b_0 = n^{\frac{-1}{2k+1}}\left(\frac{C_2}{2kC_1^2}\right)^{\frac{1}{2k+1}}.$$

Notice that b_0 depends inversely on n. Sometimes this is stressed by writing $b_0(n)$. Plugging in C_1 and C_2 we can write this as

$$b_0 = n^{\frac{-1}{2k+1}} \ \frac{(k!)^2\sigma^2\int_{-1}^{1} w^2(u)du}{2k(f^{(k)}(x)w_k))^2}.$$

This is the optimal bandwidth by minimal mean squared error criterion. If we substitute the expression b_0 into AMSE

$$\begin{aligned}
\text{AMSE}(b_0) &= \left[\left(\frac{C_2}{n2kC_1^2}\right)^{\frac{1}{2k+1}}\right]^{2k} C_1^2 + \frac{C_2}{n}\left(\frac{C_2}{n\ 2kC_1^2}\right)^{\frac{-1}{2k+1}} \\
&= \left(\frac{C_2}{n\ 2k}\right)^{\frac{2k}{2k+1}} C_1^{\frac{2}{2k+1}} \ + \ n^{\frac{-2k}{2k+1}}C_2^{\frac{2k}{2k+1}}\left(2kC_1^2\right)^{\frac{1}{2k+1}} \\
&= \left(\frac{C_2}{n}\right)^{\frac{2k}{2k+1}} \ C_1^{\frac{2}{2k+1}}\left[(2k)^{\frac{-2k}{2k+1}} + (2k)^{\frac{1}{2k+1}}\right]
\end{aligned}$$

Since $(2k)^{\frac{1}{2k+1}} = (2k)^{1-\frac{2k}{2k+1}}$ we can further simple this to

$$\text{AMSE}(b_0) = n^{\frac{-2k}{2k+1}} \; C_2^{\frac{2k}{2k+1}} \; C_1^{\frac{2}{2k+1}} \; (2k)^{\frac{-2k}{2k+1}} \Big(1 + 2k\Big).$$

Substituting in C_1 and C_2 into this expression

$$\begin{aligned}
\text{AMSE}(b_0) &= n^{\frac{-2k}{2k+1}} \left(\sigma^2 \int_{-1}^{1} w^2(u) du\right)^{\frac{2k}{2k+1}} \left(\frac{f^{(k)}(x)}{k!} w_k\right)^{\frac{2}{2k+1}} (2k)^{\frac{-2k}{2k+1}} \; \Big[1 + 2k\Big] \\
&= C_k \; n^{\frac{-2k}{2k+1}} \Big(\sigma^{2k} f^{(k)}(x)\Big)^{\frac{2}{2k+1}} \Big[\Big(\int_{-1}^{1} w^2(u) du\Big)^k \; w_k\Big]^{\frac{2}{2k+1}}
\end{aligned}$$

where

$$C_k = \frac{(2k)^{\frac{-2k}{2k+1}} \; (1 + 2k)}{(k!)^{\frac{2}{2k+1}}}.$$

The AMSE of the kernel estimator can be further reduced by a judicious choice of the kernel that minimizes the expression $T(w) = \left(\int_{-1}^{1} w^2(u) \, du\right)^k w_k$ for a given value of k. Such "optimal" kernels are provided by Gasser and Muller (1979). They also provide kernels that minimize $\int_{-1}^{1} w^2(u) \, du$ among all kernels of order k; these are called minimum variance kernels. The Epanechnikov kernel is an optimal kernel of order 2. (Wand and Jones [1993]).

3.5 Boundary Correction

As we have seen, kernel estimators suffer from bias. Particularly, kernel estimators suffer from greater bias in the boundary. Notice the kernel plotted in red with a bandwidth of 0.15 in Figure 3.2. The points outside the window, $[x_0 - b, x_0 + b]$ are weighted zero.

At the point x_0, we seek to plot the estimate $\hat{f}(x_0)$ that reflects the regression function $\mathbb{E}(Y|x = x_0) = f(x_0)$. The kernel weights the points relative to their distance from x_0. The points outside the window (shown in blue) get no weight, while the points inside the window get weight based on each point's individual distance from x_0. Notice that the

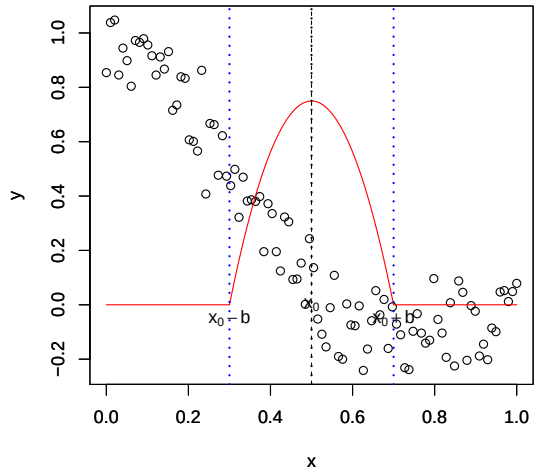


Figure 3.2: This figure illustrates that points close to x_0 are weighted more by the kernel than points further from $x_0 = 0.5$ and $b = 0.15$.

kernel is symmetric. Thus, it weights points to the left and right of x_0 equally. However, consider the left boundary where $x < b$ as depicted in Figure 3.3.

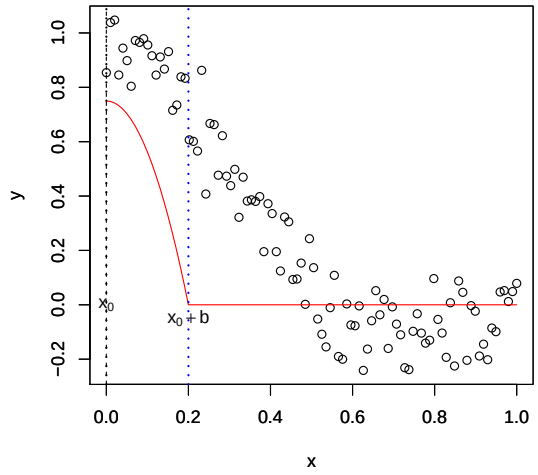


Figure 3.3: This figure shows the Epanechnikov kernel plotted with bandwidth, $b = 0.2$ and $x_0 = 0$.

This region where $x < b$ is referred to as the left boundary. Notice there are no points on the left side of the window and this has the effect of exacerbating the bias.

In the left boundary, $\frac{x}{b} \in (0, 1)$ thus, we cannot break up the integral $\int_{\frac{x-1}{b}}^{\frac{x}{b}} w(u)\, du$ as we did previously. Instead, we break it up as

$$\int_{\frac{x-1}{b}}^{\frac{x}{b}} w(u)\, du = \int_{\frac{x-1}{b}}^{-1} w(u)du + \int_{-1}^{\rho} w(u)du \tag{3.4}$$

where $\rho = \frac{x}{b}$. Again, assuming the kernel has compact support, the first integral is zero. The second integral cannot be broken up further since $\frac{x}{b} < 1$. If we compute $\mathbb{E}[\hat{f}(x)]$ for x in the left boundary, we arrive at

$$\mathbb{E}[\hat{f}(x)] \approx \frac{1}{W_0(\rho)} \int_{-1}^{\rho} f(x - ub)w(u)du$$

where $W_0(\rho) = \int_{-1}^{\rho} w(u)\,du$. Again using a Taylor series expansion about the point x, $\mathbb{E}[\hat{f}(x)]$ is approximately equal to

$$\frac{1}{W_0(\rho)}\int_{-1}^{\rho} w(u)\left[f(x) + (-ub)f'(x) + \frac{(-ub)^2 f''(x)}{2} + \ldots + \frac{(-ub)^k f^{(k)}(x)}{k!}\right]du + R(x,n,b)$$

$$= \left[f(x)\int_{-1}^{\rho} w(u)du + (-b)f'(x)\int_{-1}^{\rho} uw(u)du + (-b)^2\frac{f''(x)}{2}\int_{-1}^{\rho} u^2 w(u)du\right.$$

$$\left. \ldots + \frac{(-b)^k f^{(k)}(x)}{k!}\int_{-1}^{\rho} u^k w(u)du\right]/W_0(\rho) + R(x,n,b).$$

Let $W_j(\rho) = \int_{-1}^{\rho} u^j w(u)du$, for $j = 1, 2, \ldots k$. Because these integrals are no longer from -1 to 1 these terms do no correspond to the moments of the kernel. We can further simplify this as

$$f(x) + (-b)f'(x)\frac{W_1(\rho)}{W_0(\rho)} + \frac{(-b)^2}{2}f''(x)\frac{W_2(\rho)}{W_0(\rho)} + \ldots + \frac{(-b)^k}{k!}f^k(x)\frac{W_k(\rho)}{W_0(\rho)} + R(x,n,b)$$

$$= f(x) + (-b)f'(x)R_1(\rho) + \frac{(-b)^2}{2}f''(x)R_2(\rho) + \ldots + \frac{(-b)^k}{k!}f^k(x)R_k(\rho) + R(x,n,b)$$

where $R_j(\rho) = \frac{W_j(\rho)}{W_0(\rho)}$ for $j = 1, \ldots k$.

So the bias of the kernel estimator is of order b instead of b^k as in the interior. There are several fixes to this problem. We will discuss two approaches: 1.)Jackknifing and 2.) Local Linear Fitting.

3.5.1 Jackknifing

Consider when $x \in (0, b)$, $k = 2$ and the kernel estimator is constructed using two different bandwidths, b_1 and b_2. We can write the expected value of $\hat{f}(x; b_h)$ when x is in the interior as

$$\mathbb{E}(\hat{f}(x;b_h)) \approx f(x) + (-b_h)f'(x)R_1(\rho_h) + (-b_h)^2 f''(x)R_2(\rho_h) \tag{3.5}$$

where $\rho_h = \frac{x}{b_h}, \quad h = 1, 2$. We define the Jackknife estimator, $\tilde{f}(x; b_1, b_2)$ to be a convex linear combination of these estimators

$$\tilde{f}(x; b_1, b_2) = c\hat{f}(x; b_1) + (1 - c)\hat{f}(x; b_2) \tag{3.6}$$

where $c \in (0, 1)$ is a constant. The appropriate choice of c eliminates the middle term in (3.5), thus reducing the bias inside the boundary to an order of b^2. Consider $\mathbb{E}(\tilde{f}(x))$, for $x \in (0, b)$, this is given by

$$c\mathbb{E}\hat{f}(x, b_1) + (1 - c)\mathbb{E}\hat{f}(x, b_2)$$

$$\begin{aligned}
&\approx c\big[f(x) + (-b_1)R_1(\rho_1)f'(x) + \frac{(-b_1)^2}{2}R_2(\rho_1)f''(x)\big] \\
&\quad + f(x) + (-b_2)R_1(\rho_2)f'(x) + \frac{(-b_2)^2}{2}R_2(\rho_2)f''(x) \\
&\quad - c\big[f(x) + (-b_2)R_1(\rho_2)f'(x) + \frac{(-b_2)^2}{2}R_2(\rho_2)f''(x)\big].
\end{aligned}$$

$$= f(x) + f'(x)\Big[c\Big(b_2R_1(\rho_2) - b_1R_1(\rho_1)\Big) - b_2R_1(\rho_2)\Big] + \frac{f''(x)}{2}\Big[c\Big(-b_2^2R_2(\rho_2) + b_1^2R_2(\rho_1)\Big) + b_2^2\frac{R_2(\rho_2)}{2}\Big] \tag{3.7}$$

From here it follows that if c is given by

$$c = \frac{b_2R_1(\rho_2)}{b_2R_1(\rho_2) - b_1R_1(\rho_1)}.$$

Generally, for a kernel of order k, a convex linear combination of k kernel estimators with bandwidths $b_1, \ldots b_k$ is needed to reduce the bias in the boundary back to order b^k.

3.5.2 Local Linear Fitting

In local linear fitting, as the name implies, the objective is to fit a line within a window $[x_0 - b, x_0 + b]$. That is, we want to find $\hat{\alpha}$ and $\hat{\beta}$ such that

$$\min_{\alpha,\beta} \sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)\Big[Y_i - \Big(\alpha + \beta(x_i - x_0)\Big)\Big]^2. \tag{3.8}$$

Consider WSS defined as

$$WSS = \sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)\Big[Y_i - \Big(\alpha + \beta(x_i - x_0)\Big)\Big]^2.$$

Minimizing by differentiating with respect to α and β, we solve for $\hat{\alpha}$ and $\hat{\beta}$ in the following system of two linear equations:

$$\frac{\partial WSS}{\partial \alpha}\Big|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = -2\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)(Y_i - \alpha - \beta(x_i - x_0))\Big|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = 0$$

$$\frac{\partial WSS}{\partial \beta}\Big|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = -2\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)(x_i - x_0)(Y_i - \alpha - \beta(x_i - x_0))\Big|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = 0.$$

Therefore,

$$\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)Y_i - \hat{\alpha}\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big) - \hat{\beta}\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)(x_i - x_0) = 0 \tag{3.9}$$

$$\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)(x_i - x_0)Y_i - \hat{\alpha}\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)(x_i - x_0) - \hat{\beta}\sum_{i=1}^{n} w\big(\frac{x_i - x_0}{b}\big)(x_i - x_0)^2 = 0.$$

To solve for $\hat{\alpha}$ and $\hat{\beta}$ in (3.9) and to simplify these equations the following constants are

defined,

$$c_0 = \sum_{i=1}^{n} w(\frac{x_i - x_0}{b})Y_i \qquad c_1 = \sum_{i=1}^{n} w(\frac{x_i - x_0}{b})$$
$$c_2 = \sum_{i=1}^{n} w(\frac{x_i - x_0}{b})(x_i - x_0) \quad c_3 = \sum_{i=1}^{n} w(\frac{x_i - x_0}{b})(x_i - x_0)Y_i$$
$$c_4 = \sum_{i=1}^{n} w(\frac{x_i - x_0}{b})(x_i - x_0)^2.$$

It's important to realize that these are all functions of x_0. Now, we can write (3.9) as follows

$$c_0 - \hat{\alpha}c_1 - \hat{\beta}c_2 = 0$$
$$c_3 - \hat{\alpha}c_2 - \hat{\beta}c_4 = 0.$$

In matrix notation,

$$\begin{pmatrix} c_1 & c_2 \\ c_2 & c_4 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} c_0 \\ c_3 \end{pmatrix},$$

with solution

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} c_1 & c_2 \\ c_2 & c_4 \end{pmatrix}^{-1} \begin{pmatrix} c_0 \\ c_3 \end{pmatrix}$$

$$= \frac{1}{c_1c_4 - c_2^2} \begin{pmatrix} c_4 & -c_2 \\ -c_2 & c_1 \end{pmatrix} \begin{pmatrix} c_0 \\ c_3 \end{pmatrix} = \frac{1}{c_1c_4 - c_2^2} \begin{pmatrix} c_0c_4 - c_2c_3 \\ c_1c_3 - c_0c_2 \end{pmatrix}.$$

The solutions for $\hat{\alpha}$ and $\hat{\beta}$ are

$$\hat{\alpha} = \frac{c_0c_4 - c_2c_3}{c_1c_4 - c_2^2},$$
$$\hat{\beta} = \frac{c_1c_3 - c_0c_2}{c_1c_4 - c_2^2}.$$

Finally, the estimate of $\mathbb{E}(Y|x = x_0)$ is given by $\hat{\alpha} + \hat{\beta}(x - x_0)|_{x=x_0}$. So at each point x_0 where we desire to estimate $f(x)$, we use the intercept $\hat{\alpha}$ of a locally fitted line. The slope

of the locally fitted line at this particular point can be estimated using $\hat{\beta}$.

Turning our attention to the expression for $\hat{\alpha}$, notice that only the constants c_0 and c_3 involve Y. Now $\hat{\alpha}$ can be written as

$$\hat{\alpha} = \frac{\sum_{j=1}^{n} \left[w\left(\frac{x_j - x_o}{b}\right) c_4 - w\left(\frac{x_j - x_0}{b}\right)(x_j - x_0)c_2 \right] Y_j}{c_1 c_4 - c_2^2}.$$

The term $w\left(\frac{x_j - x_o}{b}\right)\Big[c_4 - (x_j - x_0)c_2\Big]$ is the new weight for the point (x_j, Y_j) and $c_1c_4 - c_2^2$ is the normalization constant for the weights to sum to 1. This is called the equivalent kernel see Figure 3.4. Again, notice that points far from x_0 receive relatively smaller weight than points close to x_0. Thus, the equivalent kernel for local linear fitting is

$$K(x_j, x_0, b) = w\left(\frac{x_j - x_o}{b}\right)\Big[c_4 - (x_j - x_0)c_2\Big]$$

and $\hat{f}(x_0)$ is given by

$$\hat{f}(x_0) = \frac{\sum_{j=1}^{n} K(x_j, x_0, b) Y_j}{\sum_{j=1}^{n} K(x_j, x_0, b)}.$$

The equivalent kernel for $x < b$ has negative lobes that make the first moment zero (kernel of order 2). In general, to obtain a boundrary corrected kernel of order-k, a polynomial of degree k is fit locally to the data. The equivalent kernel resulting from this is a kernel of order k in the boundary.

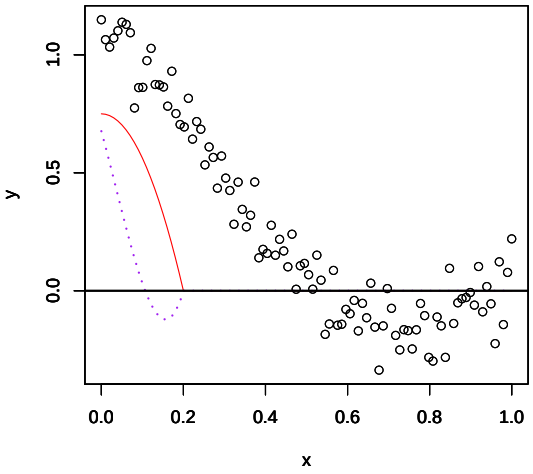


Figure 3.4: This figure shows the Epanechnikov kernel plotted in red (solid), and the equivalent kernel plotted in the purple (dashed), both use the bandwidth $b = 0.2$ for $x_0 = 0$.

Chapter 4

Nonparametric Regression using Splines

4.1 Smoothing Splines

An alternative approach to nonparametric regression is smoothing splines. The discussion here follows Eubank (1999). Smoothing splines fit data by a compromise between closeness to the data and smoothness of an estimated function. Suppose we have data (t_j, Y_j) $j = 1, 2, \ldots n$ where $0 < t_1 < t_2 < \ldots < t_n < T$. Among all functions f which have a continuous m^{th} derivative that is square integrable, we define the penalized residual sum of squares as

$$\text{PRSS}(f, \lambda) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - f(t_i)]^2 + \lambda \int_0^T [f^{(m)}(t)]^2 dt \tag{4.1}$$

where $\lambda \geq 0$. Lower values of PRSS are preferable. The first term measures closeness to the data, while the second term measures smoothness of the fitted function. The Sobolev space of order m is defined as the set

$$W_2^m[0,T] = \{f \mid f^{(m)}\text{is absolutely continuous and} \int_0^T [f^{(m)}(t)]^2 dt < \infty\}.$$

Our objective is to minimize PRSS over all such functions, thereby creating a function that is close to the data yet smooth. Since, the second term measures smoothness, λ is referred to as the smoothing parameter. When $\lambda = 0$, the solution to this problem is any function in the Sobolev space that interpolates the data. When $\lambda \to \infty$ and $m = 2$ the solution is the least squares line. The solution to (4.1) is a natural spline of degree $2m - 1$

which we will define later.

To define a spline, first define the function $(x-a)_+$ as follows

$$(x-a)_+ = \max(0, x-a).$$

This is simply $x-a$ when $x > a$ and 0 otherwise. This function is plotted in Figure 4.1 for $a = 0.5$ below.

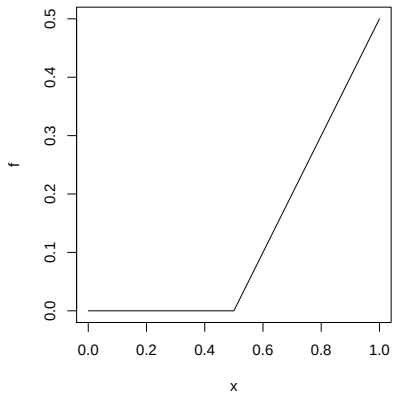


Figure 4.1: The function $(x-0.5)_+$

A spline of order r with knots at $t_1, \ldots t_n$ is any function of the form:

$$S(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{i=1}^{n} \delta_i (t-t_i)_+^{r-1} \qquad t \in [0, T] \tag{4.2}$$

for some real valued coefficients $\theta_1, \theta_2, \ldots \theta_{r-1}$ and $\delta_1, \delta_2, \ldots \delta_n$. For example, suppose we are given the two knots t_1 and t_2, where $t_1 < t_2$. An example of a spline of order 3 is a function of the form

$$S(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \delta_1 (t-t_1)_+^2 \; + \; \delta_2 (t-t_2)_+^2.$$

A spline of order r has three important properties:

1. It is a polynomial of degree $r-1$ on the interval $[t_i, t_{i+1}]$.

2. It has $r-1$ continuous derivatives.

3. The r^{th} derivative has jumps at the knots.

This is also established in Hastie, Tibshirani and Friedman (2009). A natural spline of order $r = 2m$ is a spline that is a piecewise polynomial of order m on the intervals $[0, t_1)$ and $(t_n, T]$. For example, a natural cubic spline is linear on the intervals $[0, t_1)$ and $(t_n, T]$. We further define the space $S^r(t_1, t_2, \ldots t_n)$ to be the collection of all splines of order r with knots at $t_1, \ldots, t_n$. The set $NS^r(t_1, t_2, \ldots t_n)$ is defined as the set of all natural splines or order r with knots at $t_1, \ldots, t_n$.

Now, consider finding the unique minimizing natural spline $S \in NS^r(t_1, t_2, \ldots t_n)$ of

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - S(t_i)]^2 + \lambda \int_0^T [S^{(m)}(t)]^2 \, dt, \tag{4.3}$$

for $r = 2m$. Let $\psi_1(t), \psi_2(t), \ldots, \psi_n(t)$ be a basis for $NS^r(t_1, t_2, \ldots t_n)$, then there exists coefficients $\theta_{0,j}, \ldots \theta_{m-1,j}; \delta_{1,j}, \delta_{2,j}, \ldots \delta_{n,j}$ such that

$$\psi_j(t) = \sum_{i=0}^{m-1} \theta_{i,j} t^i + \sum_{i=1}^{n} \delta_{i,j}(t - t_i)_+^{2m-1}.$$

This follows from the fact that $NS^r(t_1, \ldots, t_n)$ is a subspace of $S^r(t_1, \ldots, t_n)$. Lyche and Shumaker (1978 [4]) showed that for $f \in W_2^m[0, T]$ and $S(t) = \sum_{j=1}^{n} \beta_j \psi_j(t)$,

$$\int_0^T f^{(m)}(t) S^{(m)}(t) \, dt = (-1)^m (2m-1)! \sum_{i=1}^{n}\sum_{j=1}^{n} \beta_j \delta_{i,j}. \tag{4.4}$$

Using this result Lyche and Shumaker further showed that

$$\int_0^T [S^{(m)}(t)]^2 \mathrm{d}t = (-1)^m (2m-1)! \sum_{i=1}^{n} S(t_i)(\sum_{j=1}^{n} \beta_j \delta_{i,j}).$$

Let $\boldsymbol{S}(\boldsymbol{t}) = \boldsymbol{\Psi}\boldsymbol{\beta}$ according to

$$\begin{pmatrix} S(t_1) \\ S(t_2) \\ \vdots \\ S(t_n) \end{pmatrix} = \begin{pmatrix} \psi_1(t_1) & \psi_2(t_1) \dots \psi_n(t_1) \\ \psi_1(t_2) & \psi_2(t_2) \dots \psi_n(t_2) \\ \vdots & \vdots \dots \vdots \\ \psi_1(t_n) & \psi_2(t_n) \dots \psi_n(t_n) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

and define the matrix $\boldsymbol{G}$ as

$$(\boldsymbol{G})_{i,j} = \delta_{i,j}(-1)^m(2m-1)!.$$

Using the previous results we can write (4.3) as

$$\frac{1}{n}||\boldsymbol{Y} - \boldsymbol{\Psi\beta}||^2 + \lambda\boldsymbol{\beta^T\Psi^T G\beta}. \tag{4.5}$$

Notice that this depends on a given λ, thus we will denote the unique minimizer to this by $\boldsymbol{\beta}_\lambda$. This can be written as

$$\boldsymbol{\beta}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \{\frac{1}{n}(\boldsymbol{Y} - \boldsymbol{\Psi\beta})^T(\boldsymbol{Y} - \boldsymbol{\Psi\beta}) \ + \ \lambda\boldsymbol{\beta}^T\boldsymbol{\Psi}^T\boldsymbol{G\beta}\}$$

The solution to equation (4.5) is analogous to the solution to ridge regression with minimizer with respect to $\boldsymbol{\beta} \in \mathbb{R}^n$

$$\boldsymbol{\beta}_\lambda = (\boldsymbol{\Psi} + n\lambda\boldsymbol{G})^{-1}\boldsymbol{Y}.$$

This shows that $S_\lambda(t) = \sum_{i=1}^n \beta_{\lambda,i}\psi_i(t)$ is in fact the minimizer among all $S \in NS^r(t_1, \dots, t_n)$. It can be shown that this is in fact the minimizer of (4.1) over the entire Sobolev space. We can rewrite (4.1) as

$$\frac{1}{n}\sum_{i=1}^n \left[Y_i - S_\lambda(t_i) + S_\lambda(t_i) - f(t_i)\right]^2 + \lambda\int_0^T \left[S_\lambda^{(m)}(t) - \left(S_\lambda^m(t) - f^{(m)(t)}\right)\ dt\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i - S_\lambda(t_i)\Big]^2 + \lambda\int_0^T \Big[S_\lambda^{(m)}(t)\Big]^2\ dt \tag{4.6}$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\Big[S_\lambda(t_i) - f(t_i)\Big]^2 + \lambda\int_0^T \Big[S_\lambda^{(m)}(t) - f^{(m)(t)}\Big]^2\ dt$$
$$+ 2\Big\{\frac{1}{n}\sum_{i=1}^{n}\big[Y_i - S_\lambda(t_i)\big]\big[S_\lambda(t_i) - f(t_i)\big] - \lambda\int_0^T S_\lambda^{(m)}(t)\big[S_\lambda^{(m)}(t) - f^{(m)(t)}\big]\ dt\Big\}.$$

It can be shown that the cross-product does not depend on f, thus $S_\lambda(t)$ is the unique minimizer for all $f \in W_2^m[0,T]$. From what has been established so far, this cross-product depends on f through

$$\frac{1}{n}\sum_{i=1}^{n}\Big[Y_i - S_\lambda(t_i)\Big]f(t_i) - \lambda\int_0^T S_\lambda^{(m)}(t)\ f^{(m)}(t)\ dt$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i - S_\lambda(t_i)\Big]f(t_i) - \lambda\sum_{i=1}^{n}\big[\sum_{j=1}^{n}\beta_{\lambda,j}\delta_{i,j}\big]f(t_i)(-1)^m(2m-1)!.$$

Using matrix notation,

$$= \frac{1}{n}\big((Y - \boldsymbol{\Psi}\boldsymbol{\beta}_\lambda)^T\boldsymbol{f} - \lambda(\boldsymbol{G}\boldsymbol{\beta}_\lambda)^T\boldsymbol{f}$$

$$= \frac{1}{n}\big[(Y - \boldsymbol{\Psi}\boldsymbol{\beta}_\lambda) - \lambda\boldsymbol{G}\boldsymbol{\beta}_\lambda\big]^T\boldsymbol{f}.$$

By definition of $\boldsymbol{\beta}_\lambda$, the term $(Y - \boldsymbol{\Psi}\boldsymbol{\beta}_\lambda) - \lambda\boldsymbol{G}\boldsymbol{\beta}_\lambda = \mathbf{0}$.

Since the cross product is 0, and the first term in 4.6 does not involve f it follows that $f = S_\lambda$ is in fact the minimizer of (4.1) over the entire Sobolev space.

4.2 B Splines

Previously we said that the functions 1, t,..., t^{r-1}, $(t-t_1)_+^{r-1}$, ..., $(t-t_n)_+^{r-1}$ form a basis referred to as the power basis for splines of order r with knots $t_1, \ldots, t_n$. The B-spline basis

is an alternative to the power basis. In fact the solution to

$$\min_{f \in W_2^2} \quad \frac{1}{n}\sum_{j=1}^{n}[Y_j - f(t_i)]^2 + \lambda \int_0^T [f^{(2)}(t)]^2 \mathrm{d}t.$$

can be written using the B-spline basis. The B-spline basis functions of order k are defined recursively as follows.

$$N_{i,1}(t) = \begin{cases} 1 & \text{if} \quad t \in [t_i, t_{i+1}) \\ 0 & \text{otherwise} \end{cases}$$

$$N_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} N_{i,k-1}(t) \; + \; \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} N_{i+1,k-1}(t) \tag{4.7}$$

for $i = -(k-1), \ldots 0, \ldots, n$. To develop the B-spline basis for $S^k(t_1, \ldots, t_n)$ we have to define $2k$ additional knots as follows

$$0 = t_{-(k-1)}, \ldots t_{-1}, t_0; t_{n+1}, \ldots, t_{n+k} = T.$$

Suppose we have knots at $0.1, \ldots, 0.9$, these knots are referred to as the interior knots. Take $k = 1$, we have to define 2 additional knots t_0 and t_{n+1}. Practically for implementation, these are defined as $t_0 = t_1 - \epsilon$ and $t_{n+1} = t_n + \epsilon$ where ϵ is a small non-negative real number. In this case we will choose ϵ to be 0, so we will be "stacking" knots at the endpoints. A spline of order k with $k+1$ coincident knots is defined to be the zero function. The B-spline basis of order 1 includes $N_{0,1}(t) = \mathbf{I}_{[0,0.1)}(t)$, $N_{1,1}(t) = \mathbf{I}_{[0.1,0.2)}(t)$, $\ldots, N_{9,1}(t) = \mathbf{I}_{[0.9,1)}(t)$. These are simply indicator functions depicted in Figure 4.2.

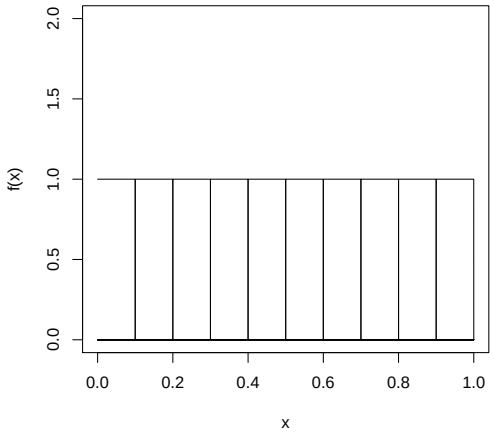


Figure 4.2: This figure illustrates B-splines of order 1, with knots at $0.1, \ldots, 0.9, \in [0, 1]$.

Now, to develop the B-spline basis of order $k = 2$ we have to define four additional knots, define $t_{-1} = t_0 = 0$ and $t_{n+1} = t_{n+2} = 1$. Consider $N_{-1,2}$. Using (4.7) $N_{-1,2}$ is computed using $N_{-1,1}$ and $N_{0,1}$. However, $N_{-1,1}(t)$ has two coincident knots so it is the zero function, so

$$N_{-1,2}(t) = \frac{t_1 - t}{t_1 - t_0} N_{0,1}(t) = \frac{0.1 - t}{0.1} N_{0,1}(t).$$

This is simply a line with slope of $\frac{-1}{0.1} = -10$ and y-intercept at 1. Next, consider $N_{0,2}$. Using (4.7)

$$N_{0,2}(t) = \frac{t - t_0}{t_1 - t_0} N_{0,1}(t) + \frac{t_2 - t}{t_2 - t_1} N_{1,1}(t) = \frac{t}{0.1} N_{0,1}(t) + \frac{0.2 - t}{0.1} N_{1,1}(t).$$

Notice that the support of $N_{0,2}$ covers 2 intervals of the knots. B-splines of order k are of degree $k - 1$ and cover k intervals if they have full support on those intervals.

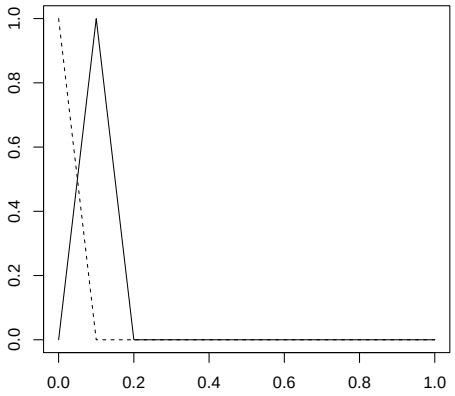


Figure 4.3: The functions $N_{-1,2}(t)$ (dashed line) and $N_{0,2}$ (solid line).

Similarly, for $N_{1,2}(t) = \frac{t-0.1}{0.1}N_{0,1}(t) + \frac{0.3-t}{0.1}N_{1,1}(t)$. Proceeding in this fashion we arrive at $N_{9,2}(t)$ which is given by

$$N_{9,2}(t) = \frac{t-0.9}{0.1}N_{9,1}(t)$$

since $N_{10,1}(t) = 0$. The function $N_{9,2}$, like $N_{-1,2}$ does not have full support but it is an upward sloping line with a slope of 10 and y-intercept of -9. The B-spline basis of order 2 is shown in Figure 4.3.

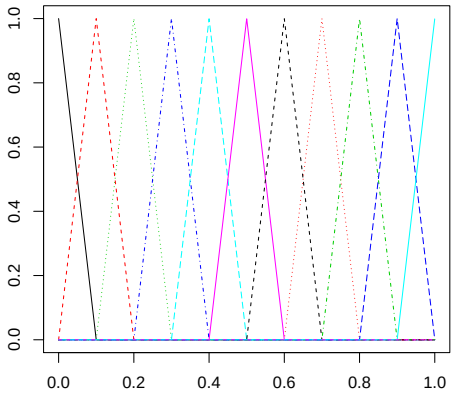


Figure 4.4: This figure illustrates the $n + k = 11$ B-splines of order 2 with knots at $0.1, \ldots, 0.9$.

The B-sline basis of order 4 is called the cubic B-spline basis because it is of degree 3. This basis is most often used because it has 2 continuous derivatives. These are smooth functions shown in Figure 4.5. There is seldom a need to go beyond cubic B-splines.

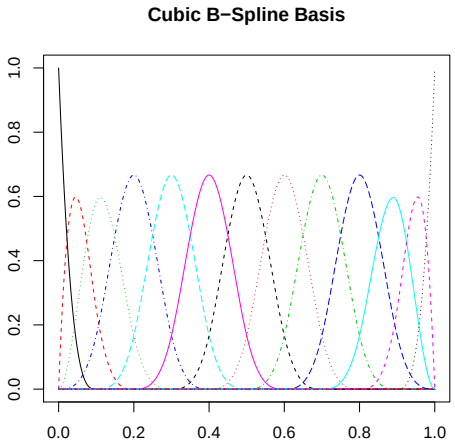


Figure 4.5: This figure illustrates the $n + k = 13$ cubic B-spline basis with interior knots at $0.1, \ldots, 0.9$.

4.3 P-Splines

Instead of penalizing the second derivative of the spline, P-splines (Penalized Splines) penalize differences between adjacent coefficients (Eilers and Marx [1996]). The following notation is useful in defining P-splines.

$$\begin{aligned}\triangle\beta_j &= \beta_j - \beta_{j-1} \\ \triangle^r\beta_j^r &= \triangle^{r-1}\beta_j^{r-1} - \beta_{j-1}^{r-1}\end{aligned}$$

Suppose we have a set of k B-splines, B_j for $j = 1, 2, \ldots, k$. Let p denote the order of the penalty. Given data (x_i, Y_i), for $i = 1, 2, \ldots, n$, the approach of P-splines is to minimize

$$\text{PSS} = \sum_{i=1}^{n}\Big(Y_i - \sum_{j=1}^{k}\beta_j B_j(x_i)\Big)^2 + \lambda\sum_{j=p+1}^{k}(\triangle^p\beta_j)^2. \tag{4.8}$$

For example, for cubic P-splines, $p = 2$, we seek to minimize

$$\text{PSS} = \sum_{i=1}^{n}\Big(Y_i - \sum_{j=1}^{k}\beta_j B_j(x_i)\Big)^2 + \lambda\sum_{j=3}^{k}(\triangle^2\beta_j)^2. \tag{4.9}$$

For cubic splines the penalty becomes

$$\begin{aligned}\triangle^2\beta_j &= \triangle\beta_j - \triangle\beta_{j-1} \\ &= \beta_j - 2\beta_{j-1} + \beta_{j-2}.\end{aligned}$$

Define the vector $\boldsymbol{\beta}$ and the matrices B, D and X as

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{k\times 1} \qquad D = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix}_{k-2 \times k}$$

$$B = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \dots & B_k(x_n) \\ B_1(x_2) & B_2(x_2) & \dots & B_k(x_2) \\ \vdots & \vdots & \dots & \vdots \\ B_1(x_n) & B_2(x_n) & \dots & B_k(x_n) \end{pmatrix}_{n\times k}.$$

So we can write (4.9) as

$$\text{PSS} = ||\boldsymbol{Y} - B\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T D^T D \boldsymbol{\beta}.$$

Defining $\hat{\beta}$ as

$$\hat{\beta} = \underset{\boldsymbol{\beta}}{\text{argmin}}\, ||\boldsymbol{Y} - B\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T D^T D \boldsymbol{\beta},$$

we find that the solution is given by

$$\hat{\beta} = (B^T B + \lambda D^T D)^{-1} B^T \boldsymbol{Y}.$$

Chapter 5

Additive Models

5.1 Semiparametric Additive Models

Suppose we have two independent variables, one of which we wish to fit parametrically and the other nonparametrically to predict a univariate response. For example, in Figure 5.1 we see a clear linear relationship between Y and x while a nonlinear relationship between Y and t.

(a) x and Y

(b) t and Y

Figure 5.1: This figure depicts two predictor variables x and t, each plotted with the response Y. The relationship between x and Y appears to be a linear relationship, while the relationship between t and Y is nonlinear.

Modeling $\mathbb{E}(Y|x,t)$ as

$$Y = \beta x + f(t) + \epsilon \tag{5.1}$$

where $\epsilon \sim N(0, \sigma^2)$, this model is referred to as a semiparametric model because x is fitted parametrically while t is fitted nonparametrically. It is an additive model because the effect of x and t on Y is additive. We can extend this model easily. Suppose we have q variables, $x_1, \ldots, x_p$ of which we wish to fit parametrically and $t_1, \ldots, t_{q-p}$ of which we wish to fit nonparametrically. Then we could write the model as follows

$$Y = \sum_{i=1}^{p} \beta_i x_i + \sum_{j=1}^{q-p} f_j(t_j) + \epsilon. \tag{5.2}$$

For example, if we desire to fit 2 variables parametrically and 2 variables nonparametrically then this becomes $Y = \beta_1 x_1 + \beta_2 x_2 + f_1(t_1) + f_2(t_2) + \epsilon$.

Imposing the parametric form reduces the number of variables that will be fitted nonparametrically. Parametric models are unbiased, and they are superior to nonparametric models if the parametric model is correct. However, parametric models can be too inflexible compared to nonparametric models.

5.2 Fitting the Semiparametric Model by Backfitting

Suppose we have n observations following the modeling assumption of (5.1); that is we have observations (x_i, t_i, Y_i) for $i = 1, 2, \ldots n$. We can fit this model using backfitting. Define the additional vectors

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \boldsymbol{f} = \begin{pmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_n) \end{pmatrix}.$$

First, we suppose $\boldsymbol{f}$ is known, then $\tilde{\boldsymbol{Y}}$ is computed as

$$\tilde{Y} = \boldsymbol{Y} - \boldsymbol{f},$$
$$\hat{\beta} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\tilde{\boldsymbol{Y}}, \text{ and}$$

$$\boldsymbol{f_1} = H(\boldsymbol{Y} - \boldsymbol{f}) = H\tilde{\boldsymbol{Y}}, \tag{5.3}$$

where $H = \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T$. Notice the similarity between this and the ordinary least squares approach. Here $\boldsymbol{f_1}$ is an estimate of $\boldsymbol{x}\beta$.

Next, we suppose β is known. If β was known consider that

$$\boldsymbol{Y} - \beta\boldsymbol{x} = \boldsymbol{f(t)} - \boldsymbol{\epsilon}.$$

Define the estimator of $\boldsymbol{f}$ to be

$$\boldsymbol{f_2} = S(\boldsymbol{Y} - \beta\boldsymbol{x}), \tag{5.4}$$

where S is the smoothing matrix defined in Chapter 3. This discussion motivates the backfitting algorithim that proceeds by iterating between (5.3) and (5.4) to convergence.

Backfitting

1.) Initialize $\boldsymbol{f_1}^{(0)} = \beta_0\boldsymbol{x}$ and compute $\boldsymbol{f_2}^{(0)} = S(\boldsymbol{Y} - \boldsymbol{f_1}^{(0)})$.

2.) Update step:

$$\boldsymbol{f_1}^{(i+1)} = H(\boldsymbol{Y} - \boldsymbol{f_2}^{(i)})$$
$$\boldsymbol{f_2}^{(i+1)} = S(\boldsymbol{Y} - \boldsymbol{f_1}^{(i+1)})$$

3.) Repeat the update step until convergence.

Convergence can be defined as $||\boldsymbol{f}_1^{i+1} - \boldsymbol{f}_1^i|| < \epsilon$ and $||\boldsymbol{f}_2^{i+1} - \boldsymbol{f}_2^i|| < \epsilon$ where ϵ is a small positive real number. Here, $\boldsymbol{f_1}$ is the estimate of $\boldsymbol{x}\beta$ and $\boldsymbol{f_2}$ is the estimate of $\boldsymbol{f}$. In this case, $\hat{\beta}$ and $\boldsymbol{f}$ can be solved explicitly in

$$\hat{\beta} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}H(\boldsymbol{Y} - \boldsymbol{f})$$
$$\boldsymbol{f} = S(\boldsymbol{Y} - \boldsymbol{x}\hat{\beta}).$$

The explicit solution for $\hat{\beta}$ and $\hat{\boldsymbol{f}}$ is given by (Hastie and Tibshirani: 1990 pg. 118)

$$\hat{\beta} = (\boldsymbol{x}^T(I_n - S)\boldsymbol{x})^{-1}\boldsymbol{x}^T(I_n - S)\boldsymbol{Y}$$
$$\hat{\boldsymbol{f}} = S(\boldsymbol{Y} - \boldsymbol{x}\hat{\beta}).$$

We can extend the backfitting algorithm to the situation which there q variables, p of which we wish to fit parametrically and variables, $q - p$ of which we wish to fit non-parametrically, $t_1, \ldots, t_{q-p}$. This is done by writing a system of equations of the the form $\boldsymbol{f}_j = S_j(\boldsymbol{Y} - \sum_{i \neq j}^{q} \boldsymbol{f}_j)$, where S_j is a smoother matrix and may be H if the j^{th} variable is desired to be fitted parametrically. That is, we establish the following system of equations

$$\begin{aligned}
\boldsymbol{f}_1^{(i+1)} &= S_1(\boldsymbol{Y} - \boldsymbol{f}_2^{(i)} - \ldots - \boldsymbol{f}_q^{(i)}) \\
\boldsymbol{f}_2^{(i+1)} &= S_2(\boldsymbol{Y} - \boldsymbol{f}_1^{(i+1)} - \boldsymbol{f}_3^{(i)} \ldots - \boldsymbol{f}_q^{(i)}) \\
\vdots \quad &\quad \vdots \quad \quad \vdots \\
\boldsymbol{f}_q^{(i+1)} &= S_q(\boldsymbol{Y} - \boldsymbol{f}_1^{(i+1)} - \boldsymbol{f}_2^{(i+1)} - \ldots - \boldsymbol{f}_{q-1}^{(i+1)}).
\end{aligned}$$

These functions are updated recursively after initialization.

Parametric models are impeded by multicollinearity. In the presence of multicollinearity regression coefficients are estimated with a great deal of variance. Multicollinearity inflates the variance of the estimators and thus the overall model. Additionally, multicollinearity

can present computational difficulties as well. In the case of perfect multicollinearity, when there is an exact correspondence with one or more of the predictor variables, the matrix $X^T X$ will not be invertible. With semiparametric models concurvity is the analogous concept to multicollinearity in parametric regression. Concurvity inflates the variance of the estimates, and imposes computational difficulties.

The R software package implements the backfitting algorithm when the data are Gaussian; otherwise it uses the local-scoring algorithm, a closely related algorithm. The backfitting algorithm sets the default values for convergence at $\epsilon = 10^{-7}$, and the maximum number of iterations allowed for is 30. The original version in S used 10^{-3}, and the maximum number of iterations was 10. However, due to the work of Dominici et al. [2002], the convergence criteria was adjusted to be more stringent.

5.3 Model Comparison

Suppose we have two linear models:

$$\begin{aligned} &\text{(F) Model 1:} \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ &\text{(R) Model 2:} \quad Y = \beta_0 + \beta_1 x_1 + \epsilon. \end{aligned}$$

Because Model 1 contains Model 2, Model 2 is said to be nested in Model 1. Model 2 is also sometimes referred to as the reduced model. It can be shown under the Gaussian modeling assumptions that

$$\frac{SSE(R) - SSE(F)/(df_R - df_F)}{SSE(F)/(n - df_F)} \sim F_{df_R - df_F, n - df_F}, \tag{5.5}$$

where $SSE(R)$ and $SSE(F)$ refer to the SSE, and df_R and df_F refer to the degrees of freedom for the reduced and full models, respectively (Kutner, Nachtscheim and Neter [2004]). So we can use (5.5) to test whether the contribution of x_2 is significant.

Suppose we have two smooth estimates of $\boldsymbol{f}$ denoted $\hat{\boldsymbol{f}}_1 = S_1\boldsymbol{Y}$ and $\hat{\boldsymbol{f}}_2 = S_2\boldsymbol{Y}$. Define

$$\mathrm{RSS}_k = \sum_{i=1}^{n} \left(Y_i - \hat{f}_k(x_i) \right)^2$$
$$\gamma_k = \mathrm{tr}(2S_k - S_k S_k^T)$$

for $k = 1, 2$. We can compare the two estimates by computing a test statistic with an approximate $F_{\gamma_2-\gamma_1, n-\gamma_2}$ distribution as

$$\frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(\gamma_2 - \gamma_1)}{\mathrm{RSS}_2/(n - \gamma_2)}, \tag{5.6}$$

where $\mathrm{RSS}_1 > \mathrm{RSS}_2$. Notice the similarity between (5.6) and (5.5). In (5.6), γ_1 and γ_2 play the role of df_F and df_R, and in doing so are sometimes referred to as the degrees of freedom. However, it is important to note that only equation (5.5) can be established analytically to follow an F distribution. Nonetheless, (5.6) serves as a useful comparison of models. For example, this approach can be taken in comparing two semiparametric models, or a nonparametric estimate to a parametric estimate or two nonparametric fits with differing smoothness (Chambers and Hastie [1992]).

The GAM package in R provides a summary table for additive models fitted in R. This summary table contains an ANOVA table for the nonparametric and parametric variables with F values. For the variables fitted nonparametrically, Statistical Models in S (Chambers and Hastie [1992]) describes how this F-value is computed;

> "For each nonparametric term in the model the nonlinear component is set to zero and the parametric part of the model is refit by weighted least squares, holding the other nonlinear components fixed."

5.3.1 Mean Square Prediction Error

Another approach to compare models is to partition the data into a training and testing data sets. Suppose there are n observations, then we can use the first k "training" observations to build a model and the remaining $n-k$ observations to validate the model. Once, a model has been developed we can validate the model by computing the mean squared prediction error as

$$\text{MSPR} = \frac{1}{n-k} \sum_{i=k+1}^{n} (Y_i - \hat{Y}_i)^2. \tag{5.7}$$

Models with lower mean squared prediction errors are thought to be better models.

Chapter 6

Semiparametric Modeling

6.1 Semiparametric Model Selection

Semiparametric models were fitted using the GAM package in R version 3.1.1 for the UTEP and Chamizal TCEQ sites. Both of these models were fitted using the TCEQ training data corresponding to August 3, 2006 to July 31, 2008. The remaining TCEQ time period August 1, 2008 - February 2, 2009 was set aside for validation. All the meteorological covariates- temperature (X_1), entropy (X_5), wind speed (X_6), humidity (X_7), dew point (X_8) and day of year ($X_{10} : \quad 1 - 365$) were fitted nonparametrically. The remaining covariates, day of time period, ($X_9 : \quad 1 - 921$), wind direction, (X_2, X_3, and X_4), time period (X_{11}, X_{12}) and weekend day (X_{13}) were fitted parametrically. Table 6.1 lists all the variables, a brief description and how they were initially fitted into the model.

Table 6.1: A list of all the covariates with a brief description of each and how they were initially fitted in the model. * is marked next to the variables forced into the regression model.

Variable	Description and Initially Fitted
X_1	Average Daily Temperature-Nonparametrically
X_2	North or South wind direction- Parametrically
X_3	East wind direction-Parametrically
X_4	West wind direction-Parametrically
X_5	Entropy-Nonparametrically
X_6	Wind Speed in the direction of the daily mode- Nonparametrically
X_7	Average daily humidity-Nonparametrically
X_8	Average daily dew point-Nonparametrically
X_9 *	Day of time period (1-921)-Parametrically
X_{10} *	Day of year (1-365)-Nonparametrically
X_{11}	August 1, 2006-July 31, 2007-Parametrically
X_{12}	August 1, 2007-July 31, 2008- Parametrically
X_{13}	Weekend day- Parametrically

Model selection was carried out via Backward Variable Selection with a stay criterion of 0.1. Gonzales et al. (2007) also used this stay criterion for model selection. The model selection proceeded in 6 steps for both the UTEP and Chamizal models. These steps and the actions taken are outlined in Tables 6.2 and 6.3. The time variables X_9 and X_{10} were forced into the model to explain changes in log(PM2.5) levels over time that may be attributable to seasonal, population, and environmental changes in the El Paso community over time.

Table 6.2: Backwards Variable Selection for UTEP Model

Step	Action Taken
Step 0	Full Model Fitted
Step 1	Dropped $X_{11} - X_{12}$
Step 2	Changed X_7 from nonparametric to linear
Step 3	Dropped X_{13}
Step 4	Moved X_5 to be linear
Step 5	Dropped $X_2 - X_4$
Step 6	Dropped X_5
Alternative Model	Dropped X_7
Final Model	Dropped X_5 and X_7

Table 6.3: Backwards Variable Selection for Chamizal Model

Step	Action Taken
Step 0	Full Model Fitted
Step 1	Changed X_5 from nonparametric to linear
Step 2	Changed X_7 from nonparametric to linear
Step 3	Dropped $X_{11} - X_{12}$
Step 4	Dropped X_{13}
Step 5	Dropped $X_2 - X_4$
Step 6	Dropped X_5
Alternative Model	Dropped X_7
Final Model	Dropped X_5 and X_7

The final model selected for both sites, with X_9 and X_{10} forced into the model, was of the form

$$\log(\text{PM2.5}) = \beta X_9 + f_1(X_1) + f_3(X_6) + f_5(X_8) + f_6(X_{10}) + \epsilon.$$

Recall from Table 6.1, X_1 is average daily temperature, X_6 is wind speed in the direction of the daily mode of wind direction, X_8 is average daily dew point, X_9 $(1 - 921)$ is day of time period and X_{10} $(1 - 365)$ was day of year.

Mean square prediction errors were computed using the validation data set from August 1, 2008 to February 2, 2009 using each model developed by Backwards Variable Selection. Figures 6.1 and 6.2 describe the mean squared prediction errors computed using the validation data set throughout the Backwards Variable Selection process. Figures 6.1 and 6.2 show that all mean squared predictions errors were between 0.94 and 1.06 using the validation time period, suggesting that variables dropped from the full model were not important predictors.

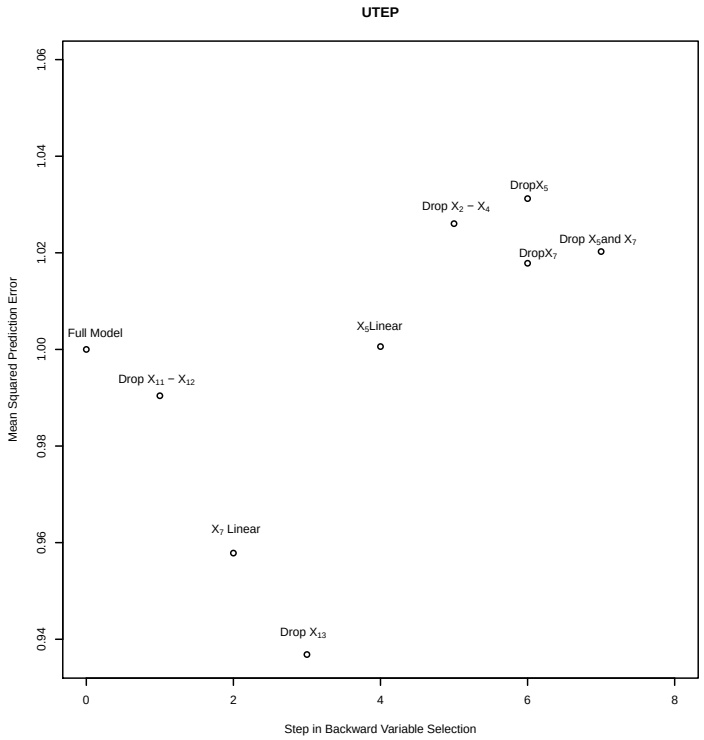


Figure 6.1: Mean square prediction errors for UTEP models developed via Backwards Variable Selection. Each description above the mean square prediction error describes the change made from the previous model.

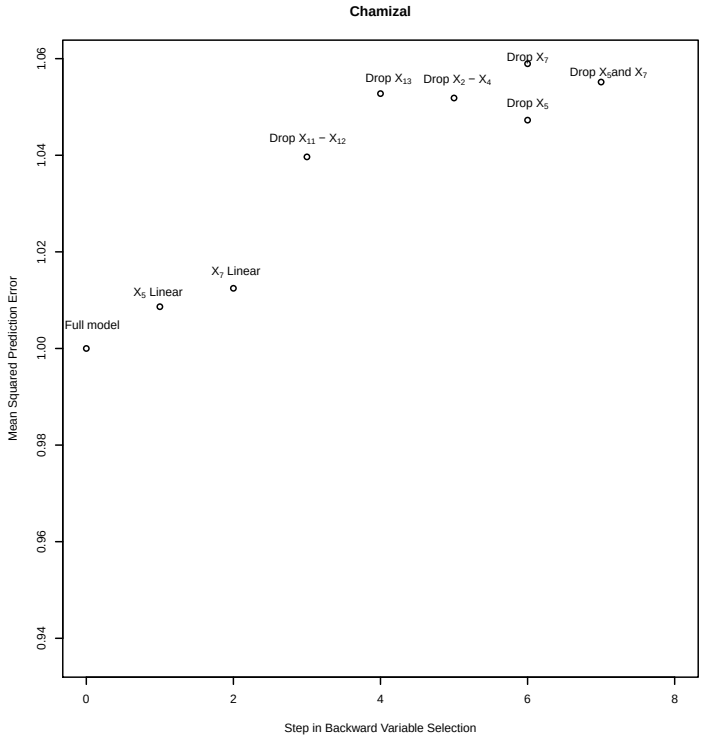


Figure 6.2: Mean square prediction errors for Chamizal models developed via backward variable selection. Each description above the mean square prediction error describes the change made from the previous model.

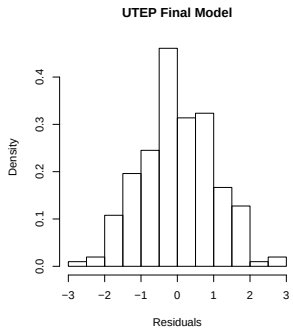
Histograms of the standardized residuals are shown in Figure 6.3. We examined the residuals $e(.)$ for both the UTEP and the Chamizal models for temporal autocorrelation. The temporal autocorrelation refers to the correlation between the residuals $e(.)$ at times t and $t + h$. The temporal autocorrelation was examined using the sample semivariogram. The semivariogram is defined by

$$\gamma(t, h) = \frac{1}{2}\text{Var}[e(t) - e(t + h)]. \tag{6.1}$$

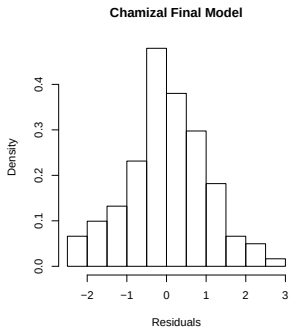
If the residual process is second-order stationary, then the semivariogram $\gamma(t, h)$ only depends on h, in which case the semivariogram is related to the covariance function $C(.)$ of the residuals according to

$$\gamma(h) = C(0) - C(h). \tag{6.2}$$

Therefore, a semivariogram $\gamma(h)$ of a second-order stationary process that is constant across values of h indicates there is no temporal autocorrelation. The sample semivariogram of the residuals indicates there does not appear to be significant autocorrelation over time; see Figure 6.4.

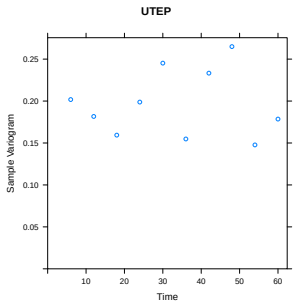


(a) UTEP

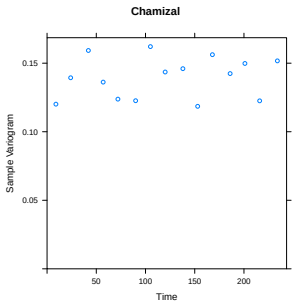


(b) Chamizal

Figure 6.3: Density histograms of the standardized residuals for the final UTEP and Chamizal models. Both histograms resemble each other.



(a) UTEP



(b) Chamizal

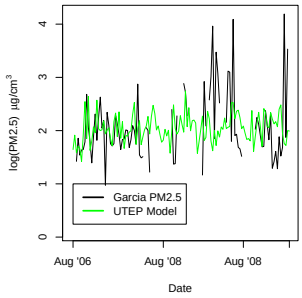
Figure 6.4: The sample variograms for the residuals of the UTEP and Chamizal models. There does not appear to be any correlation between the residuals over time.

6.1.1 Estimation at UTEP and Chamizal

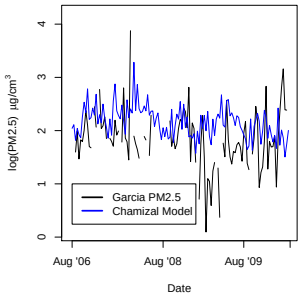
The UTEP and Chamizal semiparametric models were also used to make predictions on all days when Garcia's data were obtained. Modeling performance was evaluated using the relative root mean square prediction error. Relative root mean square prediction error is defined as the square root of the mean square prediction errors divided by the mean of these two sites. For example, the relative root mean squared prediction error for the UTEP site using the UTEP model is the square root mean squared prediction error divided by $\frac{\bar{x}_u+\bar{x_c}}{2}$, where $\bar{x}_u$ and $\bar{x}_c$ are the mean log(PM2.5) of the UTEP and Chamizal sites, respectively. This quantity is used in Smith et. al. (2006) to evaluate the predictions obtained by cross-validation. However, there is no need to cross-validate here, since the semiparametric models were developed with the TCEQ data and we are computing the mean squared error of prediction for log(PM2.5) using an independent data set (Garcia [2010]). The relative root mean square prediction errors for the UTEP and Chamizal sites are shown in Table 6.4. UTEP has smaller relative root mean square prediction error, but the standard deviation of log(PM2.5) was also lower at the UTEP site (sd=0.45) as compared to the Chamizal site (sd=0.53); see Tables 1.4 and 1.6. The observed and predicted values are displayed in Figure 6.5. The predictions follow the trend of log(PM2.5) as recorded by Garcia (2010), except for the time period at the Chamizal site when the Garcia (2010) monitors had a negative drift.

Table 6.4: Relative root mean square error of predictions for log(PM2.5) obtained by Garcia using the UTEP and Chamizal models.

	Relative Root MSEP
UTEP Garcia (2010)	0.19
Chamizal Garcia (2010)	0.33



(a) UTEP



(b) Chamizal

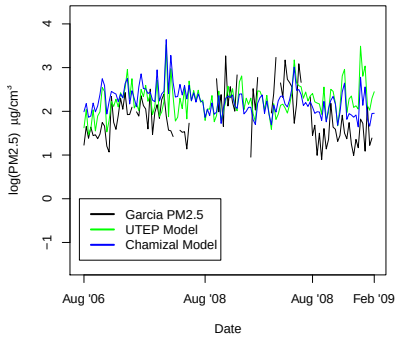
Figure 6.5: Time series plot of log(PM2.5) as recorded by Garcia (2010) and the predicted log(PM2.5) using the UTEP and Chamizal GAM models.

6.1.2 Estimation at D, J and K

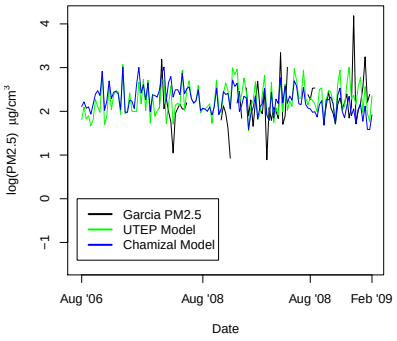
The UTEP and Chamizal semiparametric models were used to predict log(PM2.5) at sites D, J and K with the weather variables recorded at these sites by TCEQ. The relative root mean square error of prediction are shown in Table 6.5 for sites D, J and K using the UTEP and Chamizal models. Here, the relative root mean square error of prediction is computed relative to the means of sites D, J and K. The predicted and observed values are shown in Figure 6.6. As Figure 6.7 shows, the predictions from the UTEP and Chamizal models are similar with correlations of 0.7299 (D), 0.7081 (J), and 0.7523 (K). In Chapter 7, we describe Inverse Distance Weighting to combine the UTEP and Chamizal model to obtain predictions at sites D, J and K.

Table 6.5: Relative root mean square error of prediction for log(PM2.5) for sites D, J and K using the UTEP and Chamizal models.

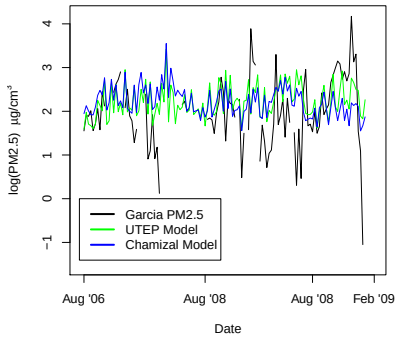
	UTEP Model	Chamizal Model
D	0.38	0.37
J	0.37	0.35
K	0.48	0.49



(a) Site D

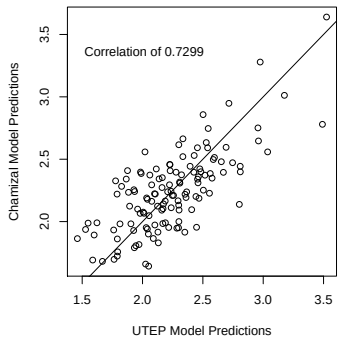


(b) Site J

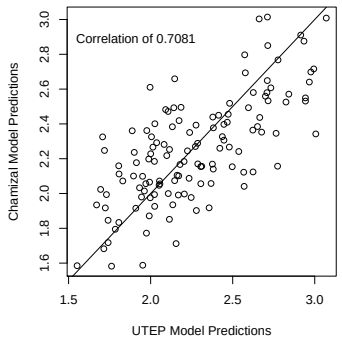


(c) Site K

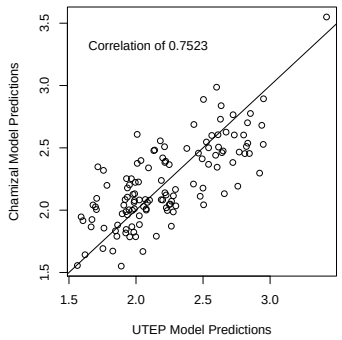
Figure 6.6: Time series plot of log(PM2.5) for the sites D, J and K as recorded by Garcia (2010) and predicted log(PM2.5) using the UTEP and Chamizal models.



(a) Site D



(b) Site J



(c) Site K

Figure 6.7: Comparison of the predictions for log(PM2.5) by the UTEP and Chamizal models at sites D, J and K.

6.2 Comparison of UTEP and Chamizal Models

The only parametric variable in the final UTEP model and Chamizal model is X_9 (time $1 - 921$), which was forced into the model. The estimate of its coefficient and the standard error are shown in Table 6.6. The estimates have opposite signs, positive for the UTEP model and negative for the Chamizal model. However, the standard errors are rather large, so these coefficients are not significantly different from 0. Nonetheless, this term is believed

to explain changes in log(PM2.5) levels over time that may be attributable to population and environmental changes in the El Paso community over time.

Table 6.6: Estimates of the coefficient of X_9 for both the UTEP and Chamizal Models and their standard errors.

Model	Estimate	Standard Error
UTEP	1.3×10^{-4}	1.4×10^{-4}
Chamizal	-3.7×10^{-4}	2.0×10^{-4}

Partial residual plots were obtained using the "visreg" package available in R version 3.1.1. Partial residual plots follow from the work done by Landwehr et al. (1984) and references therein. The fits for the covariates fitted nonparametrically are shown for the UTEP model in Figure 6.8 and the Chamizal model in Figure 6.9. Examination of these figures indicates that the covariates dew point and wind speed that were fitted nonparametrically have similar effects for both the UTEP and Chamizal models. Specifically, log(PM2.5) has a negative quadratic relationship with dew point, and a positive quadratic relationship with wind speed. Other studies in the El Paso region have also found that inversions in the atmosphere associated with low wind speeds trap urban pollution, while high wind speeds result in the entrainment of dust and sand, i.e. high levels of PM2.5. Staniswalis et al. (2005) describes "effects of particles in the ambient aerosol during El Paso sandstorms is believed different from that of particles present during still-air conditions resulting from atmospheric temperature inversions".

Time X_{10} (1-365) was also not statistically significant, but forced into the model with a p-value of 0.29 for the UTEP model and a p-value of 0.7 for the Chamizal model. The partial residual plots for this variable are not similar. The partial residual plots for temperature both show a peak during cooler temperatures. Temperature was statistically significant for the UTEP model (p-value of 0.014), but not statistically significant for the Chamizal model (p-value of 0.064). Nonetheless, temperature stayed in the model because the stay criterion was 0.1.

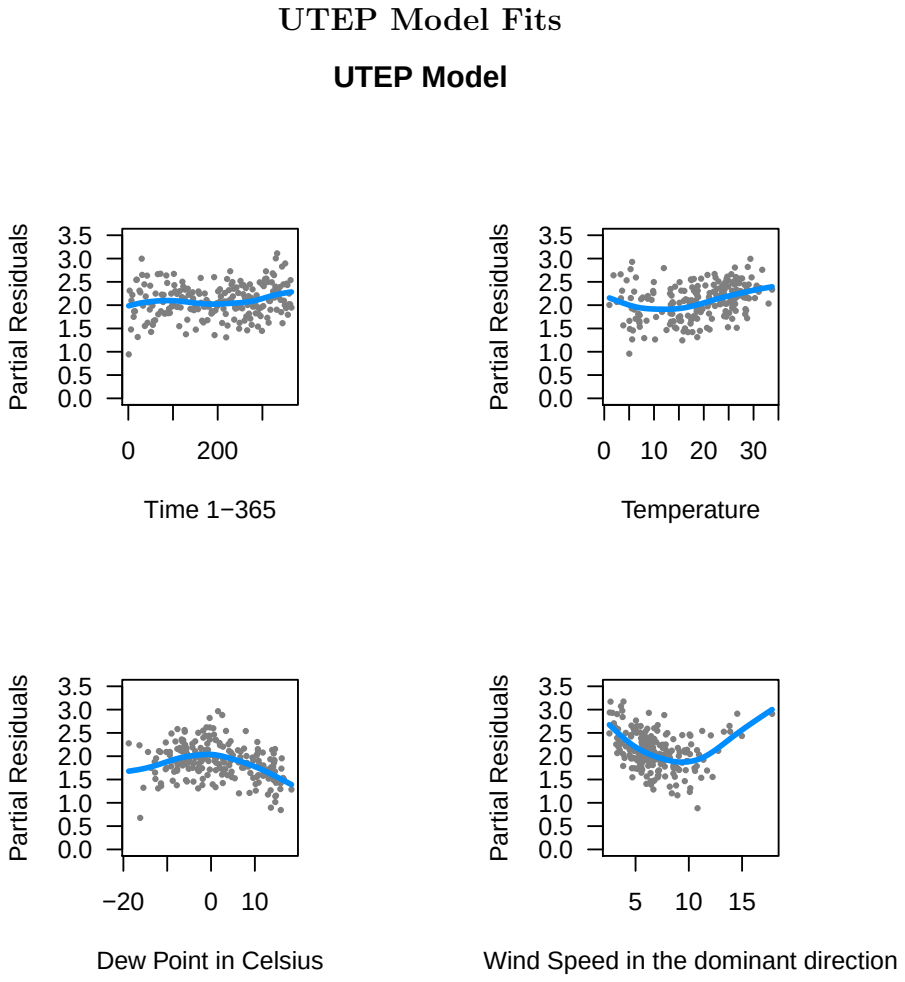


Figure 6.8: Partial residuals versus predictors for the UTEP Model

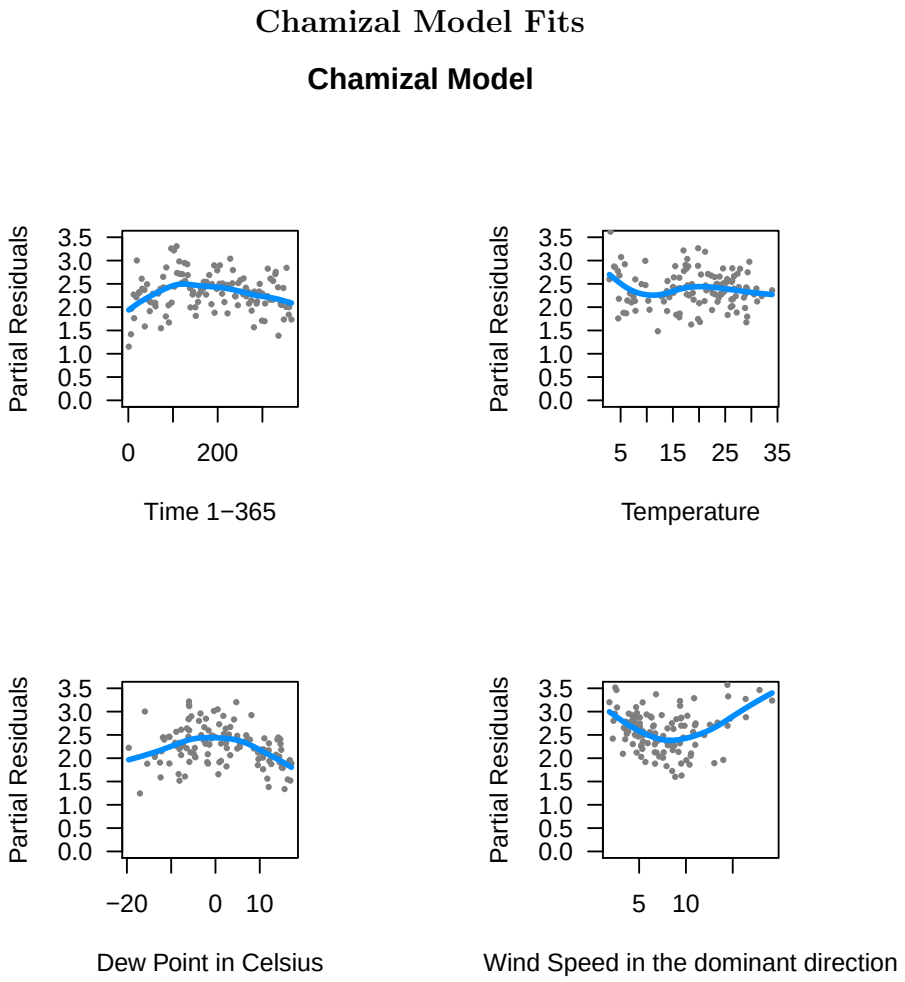


Figure 6.9: Partial residuals versus predictors for the Chamizal Model

Chapter 7

Geostatistical Methods

7.1 Inverse Distance Weighting

In geostatistics, observations are often made at sites $s_1, s_2, \ldots, s_m$. The notation and the content follows from Bivand, Pebesma, and Gomez-Rubio (2008). The values of interest at these sites are denoted $Z(s_1), Z(s_2), \ldots, Z(s_m)$. We denote these as vectors with

$$\boldsymbol{s}^T = \begin{pmatrix} s_1 & s_2 & \ldots & s_m \end{pmatrix} \quad \text{and } \boldsymbol{Z(s)}^T = \begin{pmatrix} Z(s_1) & Z(s_1) & \ldots & Z(s_m) \end{pmatrix}.$$

Often in geostatistics, we are interested in predicting $Z(s_0)$ at an unmonitored location s_0. We will reserve the notation s_o for an unmonitored location. If there is more than one unmonitored location then we use the vector $\boldsymbol{s_0}$ to denote the vector of values corresponding to unmonitored locations.

An intuitive approach to predicting at an unmonitored location is to take a weighted average of the observed levels based on distance to the unmonitored site. This approach is referred to as inverse distance weighting (IDW). The IDW estimator for a site s_0 is given by

$$\hat{Z}^{IDW}(s_0) = \frac{\sum_{i=1}^{m} w(s_i) Z(s_i)}{\sum_{i=1}^{m} w(s_i)} \tag{7.1}$$
$$\text{where} \quad w(s_i) = ||s_i - s_0||^{-l}.$$

Typically l is taken to be 2. Notice the similarity between equation (7.1) and (3.2). Both are weighted averages that are weighted inversely according to the distance to the desired

point to be estimated. The weights depend solely on distance rather than any spatial covariance structure between the sites.

7.2 The Multivariate Normal Distribution

Meteorological, geological, epidemiological and even sociological data often pertain to the space in which the data were recorded. Observations may correlate according to the geographical distance between them. In our case air pollution levels at different locations may correlate. The multivariate normal distribution is a useful distribution because it relates the components of a random vector to the other components. Specifically, the conditional distribution of unknown components of a vector given that the other components are known has been explicitly established.

Let $\boldsymbol{X}^T = \begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix}$ and suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_p \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \dots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \dots & \sigma_{p,p} \end{pmatrix}.$$

Here $\sigma_{i,i}$ denotes the variance of X_i, while $\sigma_{i,j}$ is the covariance between X_i and X_j; $i, j = 1, 2 \dots, p$. The probability density function of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big((\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big).$$

Often, we may be given a portion of the vector $\boldsymbol{X}$ and be interested in the distribution of the remaining portion. For example, partitioning $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1} \\ \dots \\ \boldsymbol{X_2} \end{pmatrix},$$

where $\boldsymbol{X_1}$ is $q \times 1$ and $\boldsymbol{X_2}$ is $(p-q) \times 1$, induces a partition on the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \dots \\ \boldsymbol{\mu_2} \end{pmatrix} \quad \Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right).$$

Here $\boldsymbol{\mu_1}$ is the $q \times 1$ mean vector of $\boldsymbol{X_1}$, $\boldsymbol{\mu_2}$ is the $(p-q) \times 1$ mean vector of $\boldsymbol{X_2}$, Σ_{11} is the $q \times q$ covariance matrix of $\boldsymbol{X_1}$, Σ_{12} is the $q \times (p-q)$ covariance matrix between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, and Σ_{22} is the $(p-q) \times (p-q)$ covariance matrix of $\boldsymbol{X_2}$. Suppose we are given $\boldsymbol{X_2} = \boldsymbol{x_2}$, then the distribution of $\boldsymbol{X_1}$ conditional on $\boldsymbol{X_2} = \boldsymbol{x_2}$ is given by

$$\boldsymbol{X_1}|\boldsymbol{X_2} = \boldsymbol{x_2} \sim N_q\Big(\boldsymbol{\mu_1} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x_2} - \boldsymbol{\mu_2}),\ \ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big). \tag{7.2}$$

A proof of this result is given in Johnson and Wichern (2007, Result 4.6). We can look upon the conditional mean of $\boldsymbol{X_1}|\boldsymbol{X_2} = \boldsymbol{x_2}$ as the mean of $\boldsymbol{X_1}$ adjusted by the covariance between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ or equivalently adjusted by the correlation between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. If $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are independent then Σ_{12} is a 0 matrix, and the conditional mean of $\boldsymbol{X_1}|\boldsymbol{X_2} = \boldsymbol{x_2}$ is simply the mean of $\boldsymbol{X_1}$.

7.3 Estimation of Spatial Covariance

7.3.1 Stationarity Assumptions

To estimate spatial correlation we make assumptions regarding the spatial covariance of the process being observed, in this case the log(PM2.5) levels being recorded by the monitors. Suppose we have log(PM2.5) measured at two sites that are a distance h apart, $Z(s)$ and $Z(s+h)$. If we assume the process is a second-order stationary process, then

1. The mean function is constant over the space.
2. $\text{Cov}\big(Z(s), Z(s+h)\big) = C(h)$, where $C(h)$ is some function of h.

Furthermore, istropy assumes the covariance is the same in any direction. This discussion has followed from Gaetan and Guyon (2010).

7.3.2 The Exponential Model

Suppose we have n observations at each of the sites $s_1, s_2, \ldots, s_m$. Let z_{ij} denote log(PM2.5) measured at time i at site j for $i = 1, 2, \ldots n$ and $j = 1, 2, \ldots, m$. One approach to compute the covariance between the sites to is to compute the matrix S as

$$S = \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1m} \\ s_{21} & s_{22} & \ldots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \ldots & s_{mm} \end{pmatrix},$$

$$\text{where} \quad s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (z_{ji} - \bar{z_i})(z_{jk} - \bar{z_k}), \quad \text{for } i, k = 1, 2, \ldots, m.$$

From this matrix we can obtain the matrix of correlations R,

$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \ldots & 1 \end{pmatrix},$$

$$\text{where} \quad r_{jk} = \frac{\sum_{i=1}^{n} (z_{ji} - \bar{z_i})(z_{jk} - \bar{z_k})}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \quad \text{for } i, k = 1, 2, \ldots, m.$$

We can obtain S from R, or R from S by

$$S = D^{\frac{1}{2}} R D^{\frac{1}{2}} \tag{7.3}$$
$$R = (D^{\frac{1}{2}})^{-1} S (D^{\frac{1}{2}})^{-1},$$

where

$$D^{1/2} = \begin{pmatrix} \sqrt{s_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{s_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{s_{mm}} \end{pmatrix}.$$

It is of interest to model correlations between sites. One approach is to model the correlation between two sites as

$$\log r_{jk} = -a\, d(s_j, s_k), \tag{7.4}$$

where $d(s_j, s_k)$ is a function of distance between the sites s_j and s_k. This is called the exponential model. Schabenberger and Gotway (2005) "strongly feel that the exponential model has earned its place among the isotropic covariance models for modeling spatial data". As the distance between s_j and s_k increases the correlation decreases exponentially. If we have observations at sites $s_1, s_2, \dots, s_m$, and we want to predict the correlation between site $Z(s_0)$ and, for example, $Z(s_1)$ we can use the model in (7.4) as long as we know the distance between the two sites. Estimation of a can be done via least squares. This is carried out using the correlations between UTEP and sites A, D, H (Chamizal), I, J, K, and L as well as between Chamizal and sites A, D, F (UTEP), H, I, J, K and L in the next chapter.

7.4 Kriging

Suppose the response variable at a particular site s is linearly related to a set of predictors as

$$\boldsymbol{Z(s)} = X\boldsymbol{\beta} + e(\boldsymbol{s}) \tag{7.5}$$

where $\mathbb{E}e(\boldsymbol{s}) = 0$, X is the $n \times p$ design matrix and $\boldsymbol{\beta}$ is the vector of unknown parameters. Notice the similarity between equation (7.5) and the linear model. Unlike the linear model the covariance of the residuals is not assumed to be a multiple of the identity matrix, but rather the spatial covariance structure of the $\boldsymbol{Z(s)}$ is denoted by V. Let $\boldsymbol{x(s_0)}$ denote the vector of the predictors at the unmonitored site s_0 and let $\boldsymbol{v}$ denote the covariance between $Z(s_0)$ and $\boldsymbol{Z(s)}$. The kriging estimator of an unmonitored site $Z(s_0)$ when V is known is given by

$$\hat{Z}(s_0) = \boldsymbol{x(s_0)}^T\hat{\boldsymbol{\beta}} + \boldsymbol{v}^T V^{-1}\Big(Z(\boldsymbol{s}) - X\hat{\boldsymbol{\beta}}\Big), \tag{7.6}$$
$$\text{where} \quad \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T V^{-1} Z(\boldsymbol{s}).$$

When there are no predictors ($p = 1$) and the mean is unknown, the estimation is referred to as ordinary kriging. Simple kriging refers to kriging when $\boldsymbol{\beta}$ is assumed to be known and not estimated as $\hat{\boldsymbol{\beta}}$. Notice the similarity between equation (7.6) and the mean in equation (7.2). Under the model assumptions of (7.5), $\hat{Z}(s_0)$ is the mean of $Z(s_0)$ adjusted by the spatial correlation between $Z(s_0)$ and the $\boldsymbol{Z(s)}$. When $Z(s_0)$ and $\boldsymbol{Z(s)}$ are independent, $\boldsymbol{v} = \boldsymbol{0}$ and the conditional mean of $Z(s_0)$ given $Z(\boldsymbol{s})$ is simply the mean of $Z(s_0)$. Accordingly, it is imperative to estimate V well. Approaches to estimation of V involve modeling correlations between sites, for example, in terms of the distance between the sites.

Chapter 8

Application of Geostatistical Methods to Prediction of PM2.5 at Unmonitored Locations

8.1 Inverse Distance Weighting

We predict the temporal mean of log(PM2.5) at an unmonitored location as a function of meterological and weather covariates. Recall that two models, one for the UTEP site and one for the Chamizal site, were built to estimate log(PM2.5) based on weather and time covariates using TCEQ data from August 1, 2006- July 31, 2008. The UTEP and Chamizal models predicted log(PM2.5) at sites D, J and K using the weather covariates obtained at each respective site with the exception of dew point and humidity which were taken to be the average of the UTEP and Chamizal measurements. We can estimate the mean log(PM2.5) at an unmonitored site by weighting the predictions of the UTEP and Chamizal models based on the distance to the unmonitored site using Inverse Distance Weighting, see equation (7.1). This allows the mean log(PM2.5) at an unmonitored site to be a function of the meteorological covariates and time. However, this requires the distance between the sites. The coordinates of the sites are listed in Table 8.1. The distance between two sites was computed using the Haversine distance. The R package "sp" provides the function "spDists" which was used to compute the distance in kilometers (km) between two sites using each site's latitude and and longitude coordinates.

Table 8.1: Lattiudes and longitudes for sites A, D, F, H, I, J, K and L.

Site	Latitude	Longitude
A	31.84808	-106.5833
D	31.89389	-106.4256
F	31.76806	-106.5011
H	31.76556	-106.455
I	31.78345	-106.3612
J	31.66194	-106.3031
K	31.70361	-106.3558
L	31.80492	-106.1664

The Inverse Distance Weighted predictions for sites D, J and K shown in Figure 8.1 were computed by Inverse Distance Weighting of the predictions based on the UTEP and Chamizal semiparametric regression models. The relative root mean square prediction errors for sites D, J and K using Garcia's data are provided in Table 8.2. For sites D and K, the IDW predictions have lower relative root mean square prediction errors than the predictions from the UTEP model alone and the Chamizal model alone. The UTEP model predictions for site J had a slightly lower relative root mean square prediction error than the IDW predictions by 0.02. Predictions were not made for sites A, I and L, because meterological data were not available.

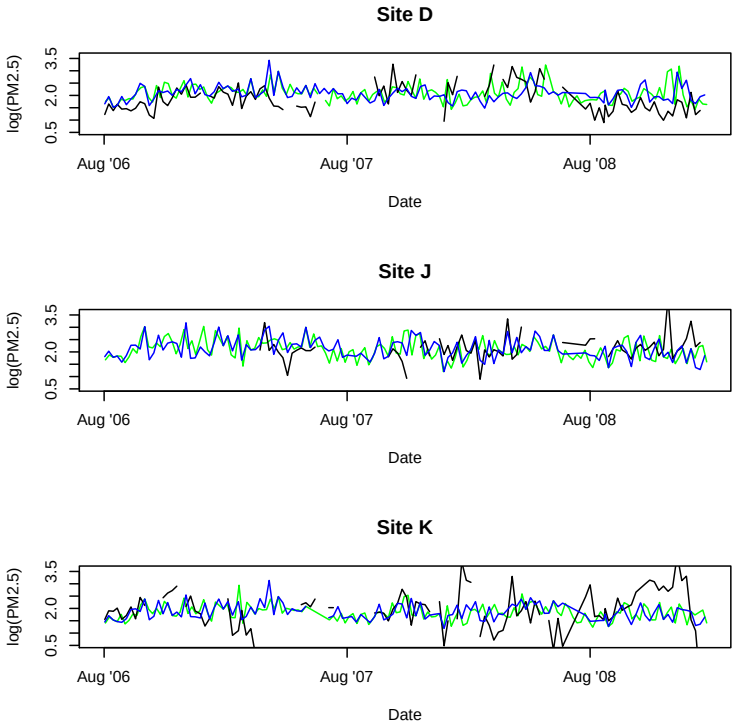


Figure 8.1: IDW predictions for sites D, J and K. The green lines are the IDW predictions on days where there is a TCEQ measurement at UTEP and Chamizal, while the blue lines are the IDW predictions on days for which Garcia took measurements. Garcia's log(PM2.5) is shown in black.

Table 8.2: Relative root mean square prediction errors for sites D, J and K.

	D	J	K
IDW	0.33	0.33	0.46
UTEP Model	0.38	0.31	0.48
Chamizal Model	0.37	0.33	0.49

8.2 Spatial Covariance of PM2.5

8.2.1 Fitting the Exponential Model of Correlation

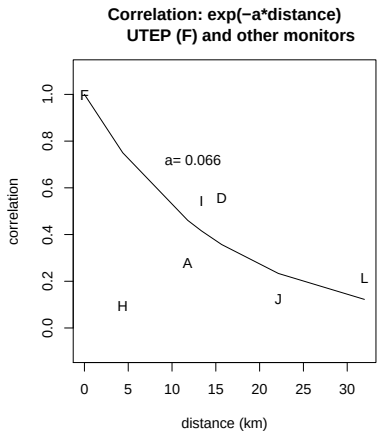
We need to take the spatial covariance of log(PM2.5) into account in the estimation of log(PM2.5). We obtained correlations between UTEP and all other sites for which PM2.5 is available using Garcia's data. These are sites A, D, F, H, I, J, K and L. Site K had a negative correlation with UTEP of -0.35. This is an unusual correlation. Raysoni et. al (2013) showed positive correlations between the monitors in El Paso, Texas. In a personal conversation with Dr. Wen-Whai Li, who was a coauthor in the Raysoni et al (2013) publication, he reiterated that this finding was unusual. Because this was unusual we omitted it in estimation of a in (7.4). The correlations with the UTEP log(PM2.5) and distances to the UTEP site are shown in Table 8.3. The correlations and distances relative to Chamizal were also obtained. Site D correlated negatively with the Chamizal site with a correlation of -0.16. Accordingly, this was omitted as well in estimation of a in (7.4). The remaining correlations with the Chamizal log(PM2.5) and distances to the Chamizal site are shown in Table 8.4. We fitted two exponential models by applying ordinary least squares without an intercept term to estimate a in (7.4), one for the UTEP site and one for the Chamizal site. Using the correlations with the UTEP site we estimated a to be 0.066, while using Chamizal data we estimated a to be 0.124. Figure 8.2 depicts $r_{jk} = \exp\big(-a\, d(s_j, s_k)\big)$ for both these models.

Table 8.3: Correlations and distances in kilometers relative to the UTEP site. The negative correlation for site K is marked with * to indicate that it was omitted in fitting the exponential model for the correlation between monitors.

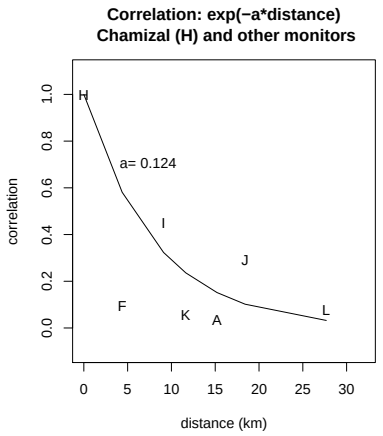
Site	Correlation with UTEP	Distance to UTEP (km)
A	0.28	11.79
D	0.56	15.66
F	1.00	0.00
H	0.09	4.38
I	0.54	13.36
J	0.13	22.14
K*	-0.36	15.51
L	0.22	31.96

Table 8.4: Correlations and distances in kilometers relative to the Chamizal site. The negative correlation for site D is marked with * to indicate that it was omitted in fitting the exponential model for the correlation between monitors.

Site	Correlation with Chamizal	Distance to Chamizal (km)
A	0.04	15.20
D*	-0.16	14.48
F	0.09	4.38
H	1.00	0.00
I	0.45	9.10
J	0.29	18.41
K	0.06	11.64
L	0.08	27.68



(a) UTEP



(b) Chamizal

Figure 8.2: The exponential models as described by Schabenberger and Gotway (2005) fitted using distances relative to UTEP and Chamizal sites.

In the interest of a single exponential model we combined the distances and correlations in Tables 8.3 and 8.4 and estimated a via ordinary least squares without an intercept term to be $a = 0.091$. Notice this is between our previous two estimates for a, namely 0.066 (UTEP) and 0.124 (Chamizal). This is depicted in Figure 8.3. This exponential model is useful because it allows us to predict the correlation of log(PM2.5) between any unmonitored location and TCEQ sites within El Paso, Texas based on the distances to the monitored locations.

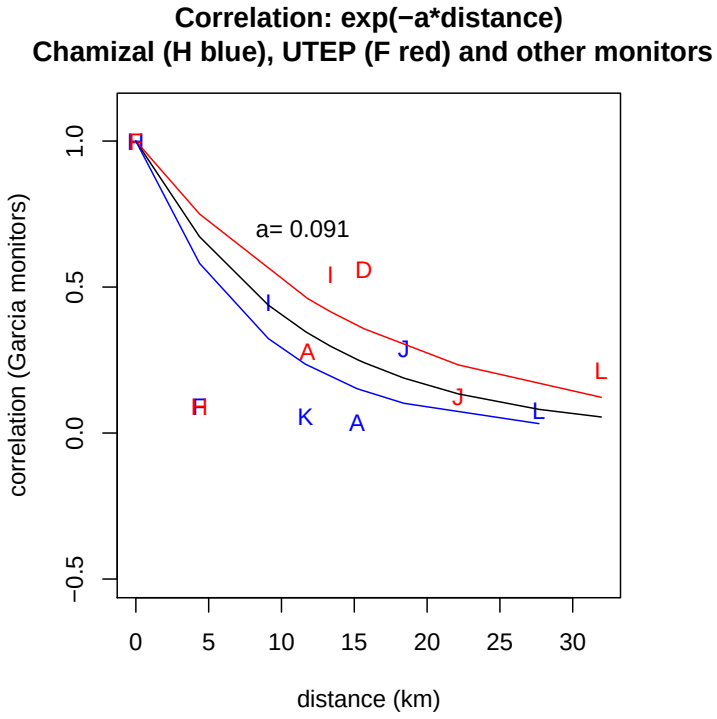


Figure 8.3: Correlations between sites. The red letters reflect the distance and correlation between sites relative to the UTEP site, while the blue letters reflect correlations and distances relative to the Chamizal site. The black line is the exponential model fit to the correlations for both sites.

8.2.2 Residual Correlations and Covariances

Residuals, e_i, of predictions at site k on day i are computed as

$$e_i(k) = Z_i(k) - \hat{Z}_i^{IDW}(k)$$

where $Z_i(k)$ is the observed log(PM2.5) on day i at site k and $\hat{Z}_i^{IDW}(k)$ is the predicted log(PM2.5) from Inverse Distance Weighting. We computed all residuals using Garcia's data Z_i. For the UTEP and Chamizal sites the residuals were computed using the prediction from the semiparametric model built using TCEQ data for that respective site.

Histograms of these standardized residuals are shown in Figure 8.4. We computed the correlation of residuals between UTEP and all other sites, as well as between the Chamizal and all other sites. The correlations between the residuals relative to the UTEP and the Chamizal sites are shown in Tables 8.5 and 8.6 respectively. The variances of the residuals are given in Table 8.7.

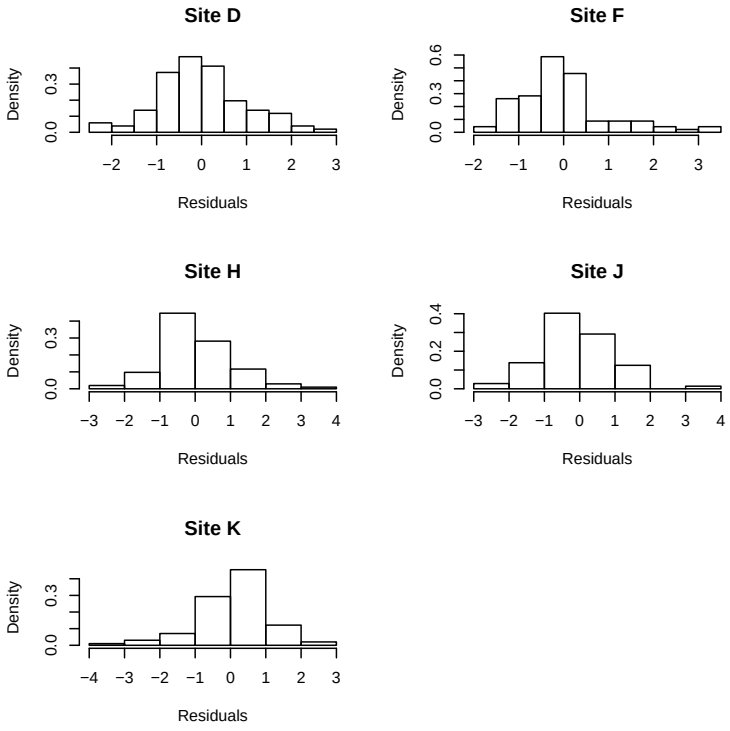


Figure 8.4: Standardized residuals for sites D, F, H, J, and K. Residuals for sites F (UTEP) and H (Chamizal) are computed using Garcia (2010) data and the predictions from the sites' respective model fitted with TCEQ data. Sites D, H and J predictions are computed using the IDW.

Table 8.5: Distances to UTEP site and correlations with UTEP residuals.

	Distance To UTEP	Correlation with UTEP residuals
D	15.66	0.32
F	0.00	1.00
H	4.38	0.55
J	22.14	0.23
K	15.51	-0.08

Table 8.6: Distances to Chamizal site and correlations with Chamizal residuals.

	Distance To Chamizal	Correlation with Chamizal residuals
D	15.66	0.00
F	0.00	0.55
H	4.38	1.00
J	22.14	0.33
K	15.51	0.19

Table 8.7: Variance of the residuals from prediction of log(PM2.5) using Inverse Distance Weighting. The residuals are calculated on days for which Garcia (2010)recorded log(PM2.5).

Site	Variance of Residuals from modeling temporal variation
D	0.3689
F (UTEP)	0.1370
H (Chamizal)	0.2164
J	0.4366
K	0.8269

At this point, let's pause to compare the correlations between sites calculated from log(PM2.5); see Tables 8.3 and 8.4, with the correlations between residuals reported above in Tables 8.5 and 8.6 . Figure 8.5 shows the correlations between the residuals relative to the UTEP site versus distance (red letters) and those relative to Chamizal site versus distance (blue letters). The dashed line is the exponential model for correlation using log(PM2.5) with $a = 0.91$, whereas the solid black line is the exponential model fitted with the correlations between residuals, $a = 0.22$.

Once the temporal variation in log(PM2.5) is subtracted from the data, the spatial correlation decreases more rapidly. If we desire to estimate the spatial correlation between two sites, $a = 0.091$ would be used if the response is log(PM2.5). If instead, the response is log(PM2.5) after removal of the temporal variation as estimated by our semiparametric model, then $a = 0.22$ in the exponential model for correlation.

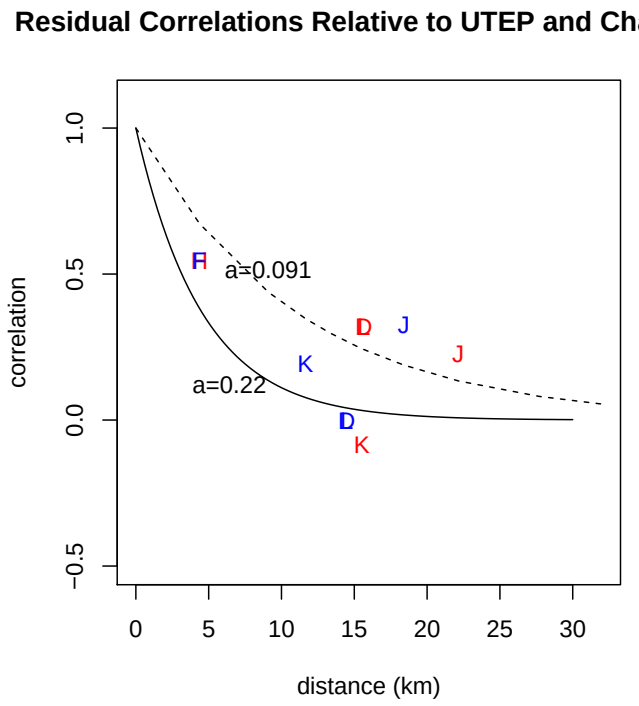


Figure 8.5: log(PM2.5) residual correlations relative to UTEP (red) and Chamizal (blue). The dashed line is the exponential model for the correlations of log(PM2.5). The solid line is the exponential model for residual correlations.

8.3 Kriging

8.3.1 Traffic Data

Traffic is believed to be a major contributor to PM2.5 levels. Figure 8.6 shows the mean of log(PM2.5) for 2007 plotted against the VMT's of a 1 km buffer radius for sites A, D, F, H, I, J, and K for 2007. A least squares line is also shown in the solid line, along with the grand mean of all the means of the sites. As you can see the relationship here is not very strong. This may be due to the fact that there are other local contributors to PM2.5 such as power plants, an oil refinery and the adjacent Ciudad Juarez, Mexico. Since VMT did

not explain much variation in log(PM2.5), we did not pursue using it as an explanatory variable for the spatial variation of log(PM2.5).

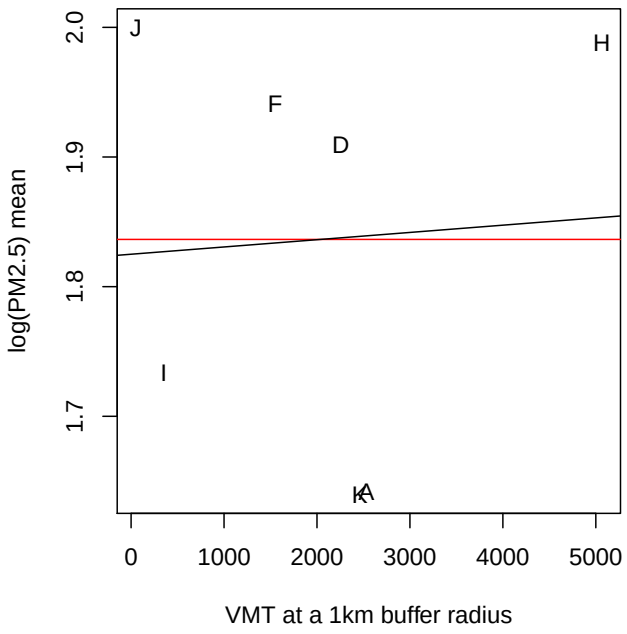


Figure 8.6: VMT in 2007 at a 1 km buffer radius for sites A, D, F, H, I, J, and K shown on the x-axis and the mean of log(PM2.5) for each site on the y-axis. The black line is the least squares line, while the red line is the mean of the site means.

8.3.2 Kriging of the Residuals

We have shown that we can model the temporal mean log(PM2.5) at an unmonitored site as a function of meterological and time covariates. Next, kriging the residuals from IDW for prediction of log(PM2.5) will be used to take into account the spatial covariance between an unmonitored site (site D, J or K) and a monitored site (UTEP [F] and Chamizal [H]).

Simple kriging is used because it is assumed that the residuals from IDW have a known mean of 0. We define the matrix and vector of covariances

$$S_{FH} = \begin{pmatrix} s_{FF} & s_{FH} \\ s_{FH} & s_{HH} \end{pmatrix}$$
$$\boldsymbol{s}_{i,FH}^T = \begin{pmatrix} s_{iF} & s_{iH} \end{pmatrix} \text{ for sites } i = \text{D, J and K.}$$

We also define the vector

$$E^T = \begin{pmatrix} \boldsymbol{e}_F & \boldsymbol{e}_H \end{pmatrix},$$

where $\boldsymbol{e}_F$ and $\boldsymbol{e}_H$ are the vector of residuals for the UTEP and Chamizal sites, respectively.

The kriged estimates for the unmonitored locations are given by

$$\hat{Z}(i) = \hat{Z}^{IDW}(i) + \boldsymbol{s}_{i,FH}^T S_{FH}^{-1} E^T. \quad \text{for sites } i = \text{D, J and K.} \tag{8.1}$$

We can obtain kriged estimates for sites D, J and K at days when measurements are available for the UTEP and Chamizal monitoring sites. Figure 8.7 shows the kriged estimates for sites D, J and K on days for which Garcia took measurements (in blue) and on days which TCEQ has measurements (in green). Equation (8.1) was used to obtain these predictions that take into account both spatial and temporal variation of log(PM2.5). The covariances used in equation (8.1) were those obtained from the residual correlations reported in Tables 8.5 and 8.6, and the residual variances reported in Table 8.7.

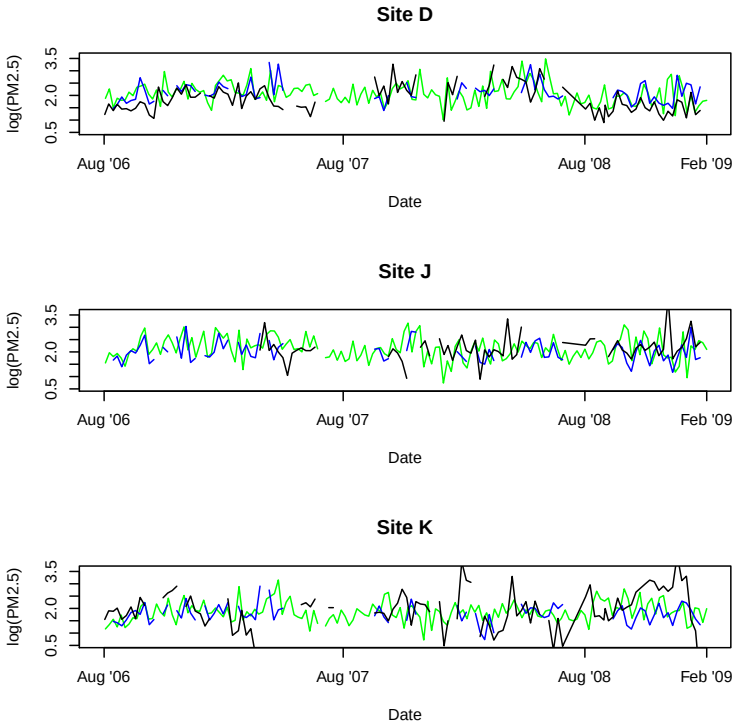


Figure 8.7: The observed log(PM2.5) as recorded by Garcia (2010) is shown in black. Displayed are log(PM2.5) predictions obtained from equation (8.1) for TCEQ days (green) and Garcia (2010) days (blue).

The relative root mean square prediction errors for days when Garcia has measurements are shown in Table 8.8. As shown in Table 8.8 the kriged estimates have the lowest relative root mean squared prediction errors in all sites D, J and K. This improvement is most likely attributed to the fact that kriging takes into account the spatial correlation between the monitored sites (UTEP and Chamizal) and the unmonitored sites (D, J and K).

Table 8.8: Relative root mean squared prediction errors for sites D, J and K.

Method	D	J	K
UTEP Model	0.38	0.31	0.48
Chamizal Model	0.37	0.33	0.49
IDW	0.33	0.33	0.46
Simple Kriging	0.18	0.22	0.41

The Simple Kriging estimates obtained in Table 8.8 employed the data-based correlations and variances of residuals. However, we could estimate the correlations between residuals using the exponential model. To do this we also have to estimate the variance of the residuals for sites D, J and K. A natural choice for this variance is the average of the UTEP and Chamizal residuals. Using the estimates for correlations between UTEP and D, J and K residuals and between Chamizal and D, J and K residuals, and using the average of the UTEP and Chamizal residuals variance, we applied (8.1). The results are listed in Table 8.9. As you can see the relative root mean square prediction errors are about the same as the IDW predictions. This result is not suprising because they are both weighted estimates based solely on distance, whereas the simple kriging estimates are based on the data.

Table 8.9: Relative root mean squared prediction errors for sites D, J and K.

Method	D	J	K
UTEP Model	0.38	0.31	0.48
Chamizal Model	0.37	0.33	0.49
IDW	0.33	0.33	0.46
Simple Kriging	0.18	0.22	0.41
Simple Kriging via the Exponential Model	0.37	0.38	0.47

8.4 Summary

The approach for prediction of log(PM2.5) at unmonitored locations considered in this thesis is unique in that it used a combination of semiparametric models to explain temporal variation combined with simple kriging to account for spatial correlation. This allows us to obtain a better estimate of personal PM2.5 exposure. For example, this procedure allows us to predict PM2.5 exposure at a particular school within El Paso, and thereby study the impact of PM2.5 exposure on children's health.

A challenge for this research was using data collected using different monitoring technologies (FRM and dichotomous samplers) that were measured on different dates. Also, the dichotomous samplers of Garcia (2010) collected weekly averages of PM2.5, whereas FRM measurements recorded by TCEQ are daily average PM2.5 measurements. The meteorological covariates pertained to daily measurements to match the TCEQ data. Garcia (2010) was the best available data for estimation of correlations used in kriging. Other traffic variables may also serve to improve estimation of log(PM2.5) by kriging, for example Johnson et al. (2010) considered distance to a petroleum facility in a land use regression model.

References

[1] R. Bivand, E. Pebesma, and V. Gomez-Rubio, "Applied Spatial Data Analysis with R". Springer. 2008. pp. 191-235.

[2] R. Chambers and T. Hastie, "Statistical Models in S". Chapman and Hall/CRC. Pacific Grove, California, 1992, pp. 249–306.

[3] F. Dominici, A. McDermott, S. Zegar and J. Samet, "On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health", *American Journal of Epidemiology*, 2002, Vol 156, pp. 1–11.

[4] P. Eilers and B. Marx. "Flexible Smoothing with Penalties", *Statistical Science*, 1996. Vol. 11, No. 2, 89-121.

[5] R. Eubank. "Nonparametric Regression and Spline Smoothing". Marcel Dekker. New York, NY, 1999. pp. 227-284.

[6] C. Gaetan and X. Guyon, "Spatial Statistics and Modeling". Springer. New York, NY, 2010, pp. 1–52.

[7] M. Garcia. (2010) "Assessing Annual and Seasonal Spatial Variability of Ambient PM10 using Linear Regression Analysis in a United States-Mexico Urban Sparwl". (Master's Thesis). Retrieved from "http://digitalcommons.utep.edu".

[8] T. Gasser and M. Mueller "Kernel Estimation of Regression Functions". *Smoothing Techniques for Curve Estimation*, 1979 Vol. 757 pp. 23–69

[9] M. Gonzales, O. Myers, L. Smith, H. Olvera, S. Mukerjee, WW. Li, N. Pingitore, M. Amaya, S. Burchiel, and M. Berwick. "Evaluation of Land Use Regression Models for NO2 in El Paso, Texas, USA", *Science of The Total Environment*, 2012, No. 432, pp. 135–142.

[10] T. Hastie and C. Loader, "Local Regression: Automatic Kernel Carpentry", *Statist. Sci. Statistical Science*, 1993, 8.2, pp. 139–143.

[11] T. Hastie, R. Tibshirani, and J. Friedman "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". New York: Springer, 2009. Print. pp. 151.

[12] T. Hastie and R. Tibshirani " Generalized Additive Models", Chapman and Hall/CRC, Boca Raton, Florida. 1990.

[13] M. Johnson, V. Isakov, J. Touma, S. Mukerjee and H. Ozkaynak, "Evaluation of land-use regression models to predict air-quality concentrations in an urban-area", *Atmospheric Environment*, 2010, Vol 40, pp. 3660–3668.

[14] R. Johnson and D. Wichern, "Applied Multivariate Statistical Analysis", Pearson. Upper Saddle River, New Jersey, 2007, pp. 149–168.

[15] Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. "Graphical Methods for Assessing Logistic Regression Models", *Journal of the American Statistical Association*, Vol. 79 (385); 1984, pp. 61-71

[16] M. Kutner, C. J. Nachtscheim, and J. Neter *Applied Linear Regression Models*, McGraw-Hill Irwin, 2004.

[17] L. Li, J. Wu, M. Wilhelm, and B. Ritz, "Use of Generalized Additive Models and Cokriging of Spatial Residuals to Improve Land-use Regression Estimates of Nitrogen Oxides in Southern California". *Atmospheric Environment*, 2012, Vol 55, pp. 220–228.

[18] T. Lyche and L. Shumaker, "Computation of Smoothing and Interpolating Natural Splines via Local Bases", *Journal of Applied Numerical Analysis*, 1973, Vol 10, pp. 1027–1038.

[19] H. Olvera, M. Garcia, WW. Li, H. Yang, MA. Amaya, O. Myers, S. Burchiel, M. Berwick, and NE. Pingitore. "Principal component analysis optimization of a PM2.5

land use regression model with small monitoring network", *Science of the Total Environment*. 2012. 425:27-34

[20] A. Raysoni, T. Stock, J. Sarnat, T. Sosa, S. Sernat, F. Holguin, R. Grenwald. B. Johnson and WW. Li "Characterization of traffic-related air pollutant metrics at four schools in El Paso, Texas, USA: Implications for exposure assessment and siting schools in urban areas." *Atmospheric Environment*, 2013, 80, pp. 140–151.

[21] O. Schabenberger and C. Gotway. *Statistical Methods for Spatial Data Analysis*, Chapman and Hall/CRC. Boca Raton, Florida, 2005.

[22] L. Smith, S. Mukerjee, M. Gonzales, C. Stallings, L. Neas, G. Norris, and H. Ozkaynak "Use of GIS and Ancillary Variables to Predict Volatile Organic Compound and Nitrogen Dioxide Levels at Unmonitored Locations." *Atmospheric Environment*, 2006, 40 (20), pp. 3773–87.

[23] J.G. Staniswalis, N.J. Parks, J.O. Bader, Munoz and Y. Maldonado. "Temporal analysis of airborne particulate matter: Reveals a dose-rate effect on mortality in El Paso; Indications of differential toxicity for different particle mixtures". *Journal of Air & Waste Management Association*, 2005, 55, pp. 893-902.

[24] Texas Commision on Environmental Quality. "Air Pollution from Particulate Matter." *Retrieved from www.tceq.state.tx.us/airquality/sip/criteria-pollutants/sip-pm*, Web. 19 May 2015

[25] M. P. Wand and M. C. Jones, *Kernel Smoothing*, 1st ed. Bury St. Edmunds, Suffolk: St. Edmundsburry, 1995.

Curriculum Vitae

Justin Jonathan Strate was born in El Paso, Texas. He studied Biological Sciences at the University of Texas at El Paso (UTEP). As an undergraduate, he worked as a Chemistry Peer-leader and for his family owned business.

Upon graduation from the University of Texas at El Paso, Justin was inundated with facts and figures about the world around him. It was then he discovered his passion for statistics and mathematics. Justin enrolled in Summer of 2013 to pursue his Masters in Statistics.

Upon graduation Justin plans to pursue a career in Biostatistics to take part in advancing medicine and health care using his passion for statistics and data.

Contact Information:

justin.strate@gmail.com