THESIS

ANALYSIS AND MODELING OF ACID NEUTRALIZING CAPACITY IN THE MID-ATLANTIC HIGHLANDS AREA

Submitted by

Brett R. Kellum

Department of Statistics

In partial fulfillment of the requirements

for the degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2003

COLORADO STATE UNIVERSITY

ABSTRACT

Acid Neutralizing Capacity (ANC) is a measure of a solution's ability to buffer itself against acidification and is used to monitor the effect of acid rain on watersheds. From 1993 to 1996, the U.S. Environmental Protection Agency collected ANC values and other data for 579 stream sites in the Mid-Atlantic Highlands Area of the Eastern United States. Combining these data with information from Geographic Information Systems (GIS) and Thematic Mapper satellite imagery, a model for predicting ANC using remotely sensed predictors was developed. The goal was to be able to predict ANC at unobserved stream locations. Several issues and concerns regarding the available data were examined, along with exploratory data analyses to investigate the relationships between ANC and the possible predictors. Several types of models were considered; in particular, a multiple regression model was selected. Our analyses determined the presence of anisotropic correlation between the residuals from the multiple regression analysis. The effects of changing direction and span of correlation in the autocorrelation function were investigated and a final variogram model for explaining the observed spatial correlation was selected.

In conclusion, nine remotely sensed predictors (elevation, felsic and carbonate bedrock, and percent pasture, percent quarry, percent probable row crops, percent woody wetlands, percent emergent wetlands, and percent high density urban area in the watershed above a site) were determined to be useful predictors of ANC in the MAHA and an exponential autocorrelation function was selected with a correlation range of approximately 100 miles, sill and nugget variance of 0.21 and 0.104, respectively, and direction of maximal correlation at 45° with a span of 35°. Here distance is measured in terms of the Euclidean distance between sites. In addition to the results of our analyses, several areas of possible future work will be presented.

DISCLAIMER

This research has been supported by Environmental Protection Agency cooperative agreement R-82909501. The research presented herein has not been submitted to or approved by the Environmental Protection Agency and are the opinions and conclusions of the author alone.

ACKNOWLEDGEMENTS

Thanks to:

Dr. Alan Herlihy of Oregon State University, for his unique insight into and understanding of acid neutralizing capacity in the MAHA region.

Dr. Dave Theobald of Colorado State University, for introducing me to the capabilities of Geographic Information Systems and coordinate projection systems.

Mary Kneeland of Colorado State University, for her assistance with acquiring necessary data from Geographic Information Systems.

Dr. Jennifer Mueller of Colorado State University, for her willingness to be a part of my graduate committee.

And Special Thanks to:

Dr. N. Scott Urquhart of Colorado State University, for his guidance, expertise, objectivity, and encouragement throughout the course of this research.

Dr. Jennifer Hoeting of Colorado State University, for her constant encouragement, guidance, and instruction throughout the course of this research. I couldn't have done this without her patience, understanding, assistance, and expertise.

DEDICATION

I wish to dedicate this work to my wonderful wife, Morgan, whose love and support keep me going each and every day. She has given so much of herself to see that I complete my degree, and it is time that I be able to give something back to her. I love you, Morgan!

CONTENTS

I. Introduction 1

II. ANC in the Mid-Atlantic Highlands Area 3
A. Overview 3
B. Acid Neutralizing Capacity 4
C. Landscape and Watershed Characteristics 6
D. Availability of Data for Sampled Sites 10

III. Overview of ANC and Predictors of ANC in the MAHA Region 13
A. Acid Neutralizing Capacity 13
B. Relationships Among Predictors of ANC 18
C. An Initial Model to Predict ANC 20

IV. Selecting Predictors of ANC for a Non-Spatial Model 24
A. Final Set of Predictors 24
B. Model Selection Procedures 24
C. Multicollinearity 27
D. Some Concerns About the Inferences From the Final – ANC Regression Model 31

V. Modeling the Spatial Correlation Between ANC Values 34
A. Alber's Equal Area Projection Coordinates 35
B. Basics of Spatial Correlation Modeling 36
C. Isotropic Spatial Correlation 38
D. Anisotropic Spatial Correlation 39
E. Anisotropic Spatial Model Selection 42
F. Inclusion of Quadratic Trend of Location in *Final – ANC* 45
G. Summary 46

VI. Conclusions and Future Work 47
A. Conclusions 47
B. Future Work 48
1. Measuring Predictive Ability of *Final – Spatial* model 48
2. Weighted Least Squares 49
3. Bayesian Model Averaging 50
4. Increased Number of Model Predictors 50
5. Stratification / Small Area Estimation 52

References 54

Appendices 56
Appendix A 57
Appendix B 58
Appendix C 59
Appendix D 75

I. Introduction

As the population of the United States and the reach of civilization continue to expand, more and more of the environment is being impacted by man. As this impact increases, mankind's responsibility to maintain a healthy environment increases as well. In 1990, Congress passed the Clean Air Act Amendments, toughening the environmental standards set forth by the Clean Air Act of 1970 and amended in 1977. These amendments mandated the monitoring of several environmental indicators, including the acidification of lakes and streams. The acidification of lakes and streams has a significantly harmful effect on the surrounding ecosystems (Stoddard, et al., 2003). The monitoring of water acidification can be done by analyzing acid neutralizing capacity (ANC) of bodies of water, for acid neutralizing capacity measures the ability of a solution to buffer itself against acidification (Stoddard, et al., 2003).

In order to monitor and explain the impact of mankind on lakes and streams, watersheds must be monitored. It is always best to have precise or exact information for all stream sites of interest, but acquiring precise information is not always practical or economical. It is not always physically possible to visit a given site in order to ascertain that exact information. Nor is it fiscally possible to frequently visit every site. Therefore, establishing a method to accurately predict the characteristics of a given stream site from a great distance would be very beneficial. Is it possible to predict the characteristics of a given stream site based solely upon remotely sensed predictors, i.e. influential site characteristics that can be gathered without actually visiting the stream site itself? This is the central motivating question of the research presented here.

This paper will present the data used and the problems encountered when gathering and combining data from different sources. A model for predicting acid neutralizing capacity using a set of remotely sensed predictors for the Mid-Atlantic Highlands (MAHA) region of the U.S. will be presented, followed by a discussion of several possible areas of future work to consider.

In Section II, we will examine the availability and distribution of data in the MAHA region. The data were combined from several different sources, which created several problems that needed to be addressed. The available data and procedure used for creating the final data set will be discussed.

Section III will cover the initial analysis of this final data set, exploring the characteristics of acid neutralizing capacity and several predictor variables. Initial relationships between ANC and the predictors will be pursued, as well as the inter-relationships between the predictor variables themselves.

Section IV will examine some preliminary modeling of ANC using linear regression methods. We will describe our methods of model selection, look at the issue of multicollinearity in the predictors, and discuss some initial results from the linear regression modeling.

In Section V, we will describe the need to include spatial considerations in any modeling of ANC in this region. We will discuss what we mean by isotropic and anisotropic spatial correlation and create a model for predicting ANC using these spatial considerations.

Finally, Section VI will examine several of the possible areas for further research and questions that arose over the course of this study that require further examination.

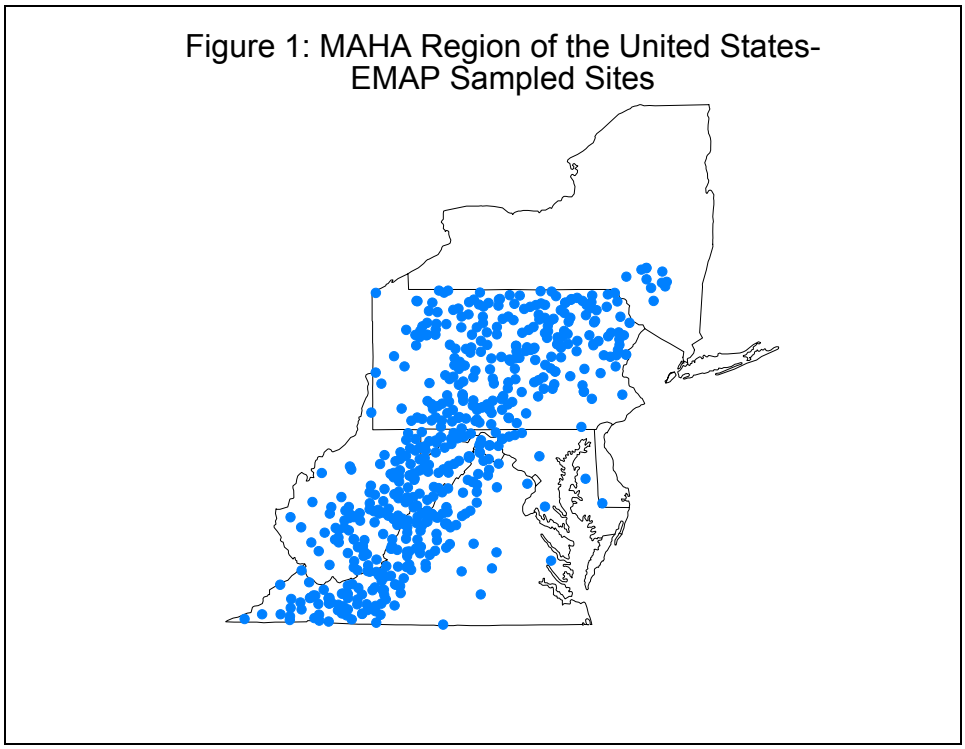
II. ANC in the Mid-Atlantic Highlands Area

A. Overview

In order to create a model for predicting acid neutralizing capacity using remotely sensed predictors, data were used from the Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) study of the Mid-Atlantic Highlands region. EMAP is a "research program to develop the tools necessary to monitor and assess the status and trends of national ecological resources" (Environmental Monitoring and Assessment Home Page, 2002). The data were collected by teams of researchers at probability selected stream sites in the Mid-Atlantic Highlands Area (MAHA) of the eastern United States. This region consists of most of the states of Pennsylvania, Virginia, and West Virginia. Parts of Delaware, Maryland, and New York were also included in the study area (Figure1).

Within this region, teams collected data from various stream sites over the course of a four-year period from 1993 through 1996. The time of year in which the samples were collected was restricted to the months of May through July, which are typically "low flow" months for this region, to minimize the seasonal variability in stream characteristics (Jones, Riitters, et.al, 1997).

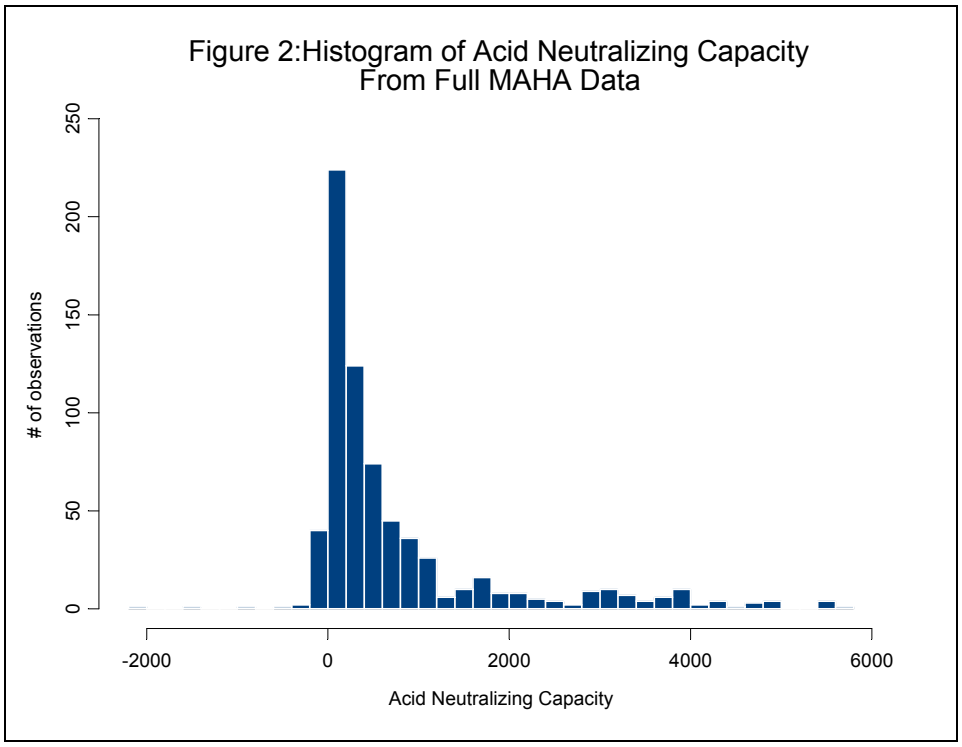
The data collected by EMAP in the MAHA region resulted in a very large amount of information on hundreds of different variables, including landscape characteristics at and surrounding each site, counts and characteristics of biologic life at the site, and the chemistry composition of the stream water. Our focus was on one of the water chemistry measurements, ANC. In addition to the data from EMAP, data were gathered from a distance via satellite imagery or Geographic Information Systems (GIS) modeling.



The goal of this study was to create a model for predicting ANC using explanatory variables that could be acquired from a distance. Further, it was intended that the method used in creating this model could be reproduced and updated when more information became available, for one concern we encountered was the lack of complete information at many of the sampled sites. This lack of information limited the scope and accuracy of the final model. Therefore, the methods and procedures used in this study were performed in such a way that as more data is gathered, the study within this region could be reproduced with this new information.

B. Acid Neutralizing Capacity

Acid neutralizing capacity was evaluated at 579 unique sites (Figure 1). Some sites were visited twice in the same year, once every year, once every other year, or just one time total during this four-year period. The number of repeated visits depended on



the role of the site in the EMAP monitoring design. ANC ranged from –2195 to + 5620 μeq/L. Figure 2 indicates that the distribution of acid neutralizing capacity was highly skewed to the right. A majority of ANC values lie between 0 and 1000, with only 2 or 3 extremely large, negative ANC values observed. Further, many positive values of ANC lie between 1000 and 6000. Although negative values of ANC were unusual, such values are legitimate because they indicate samples that were acidic when they were collected.

Some studies have limited their research to streams and/or lakes with ANC values between 0 and 200 (Brewer, Sullivan, Cosby, and Munson, 2002), for these bodies of water are typically considered sensitive to acid deposition. While there are many sites where ANC does fall into that range, approximately 61% sites had values of ANC greater than 200.

Table 1: Sites and ANC values of five identified outliers

Site ID	PA839	WV797	WVT02	PA029	WV529
ANC	-2120	-1460	-803	-436	-326

By limiting the scope of the research to only those sites that would be considered sensitive to acid deposition, the analysis would fail to take into account what may contribute to such large values of ANC. Therefore, it was deemed worthwhile to pursue a model to predict ANC over the whole range of values, and not limit the scope of the model to a certain range of ANC.

It was shown that the distribution of ANC was highly skewed to the right (Figure 2). Even though there were many extreme positive values of ANC in the data, few sites had extreme negative values of ANC. Five sites in particular were identified as unexplained outliers, and eliminated from further analysis, for the final model described below could not accurately predict ANC for these sites. The identified sites and levels of ANC are shown in Table 1.

C. Landscape and Watershed Characteristics

The driving motivation for this study was to create a model to predict acid neutralizing capacity using only remotely sensed model predictors. The first identified remotely sensed predictors consisted of 15 separate classifications of landscape characteristics as identified by Thematic Mapper (TM) satellite imagery. These variables are considered watershed scale variables. This means that using the TM satellite image, the watershed above a certain stream point is broken down into 30-meter resolution

Table 2: Thematic Mapper landscape classifications

Barren Areas	Developed Areas	Cultivated Areas	Vegetated Areas	Water Areas
Beach	Urban (Low Intensity)	Pasture	Deciduous Forest	Water
Mines	Urban (High Intensity)	Probable Row Crops	Evergreen Forest	
Quarry		Row Crops	Mixed Forest	
Transitional			Emergent Wetlands	
			Woody Wetlands	

pixels, with each pixel then classified into one of 15 different landscape classifications (Table 1). (For further information on how each class was defined, please see Appendix A). For example, a quarry value of 5% at a sampled site means that 5% of the pixels above that given site were classified as belonging to a quarry. Further, if, at a given site, deciduous forest has a value of 65, this indicates that 65% of the pixels (i.e. 65% of the watershed) above that given site is classified as deciduous forest.

Thematic Mapper satellite imagery information was available for 515 individual sites in five states. At the time of this analysis, Thematic Mapper imagery data were not available for sites in New York. For the 515 sites where TM data were available, data were reported in one or both of two different forms:

- The number of pixels per watershed, separated by landscape classification
- The percentage of each watershed classified as one of the 15 specific landscape types.

All of the sites did have the percentage allocation, but not all of them gave the pixel information. Therefore, it appeared that using the percentages per watershed would be most useful due to the greater amount of information available. Two classes, barren beach areas and barren mine areas, were eliminated from the analysis. Only five sites

possessed mine drainage in the watershed above the site, one of which had already been eliminated from the analysis (Site ID #PA029). Also, barren beach areas was not included in the analysis because none of the sampled sites had any beach area located in the watershed above the site.

Bedrock geology is another factor that is related to acid neutralizing capacity. The importance of bedrock geology in predicting acid neutralizing capacity was shown by Sullivan, et al. (2002). In analyzing ANC, Sullivan, et al. (2002) demonstrated that bedrock geology was the most significant indicator of acid neutralizing capacity in the Southern Appalachian Mountain Region of the United States. Virginia and West Virginia were included in this analysis. This result indicated that bedrock geology should be considered in the MAHA region as well; especially due to the overlapping states found in the Southern Appalachian and Mid-Appalachian regions. While the EMAP data did not include the bedrock geology at each site, this was available from Geographic Information Systems (GIS) models as determined by state geologic surveys. From GIS models, both the geological class of bedrock and specific type of rock were supplied. Five classes of bedrock were included in this analysis: Carbonate, Argillace, Siliceous, Mafic, and Felsic rock. Along with these five classes, a sixth class of bedrock geology was included, called Unclassified. This "Unclassified" class consisted of geology that could not be classified as any of the five above. For example, sand and gravel were not included in any of the five classes above, so they were lumped together in the "Unclassified" class.

The bedrock geology information initially provided for this analysis consisted of geology at the sampled site. Data on the geology above each sampled site was not summarized or available at the time of our initial analysis, but would have been more

useful. The characteristics of any stream site depend upon the characteristics of the watershed above that site. Although the bedrock geology at the sampled site was informative, knowing the bedrock geology in the watershed above the site would most likely have given a clearer picture of the impact on ANC of bedrock geology (See Section IV for further discussion on this subject).

Also available from GIS modeling were elevation and Strahler stream order. From previous analyses (e.g. Sullivan, et al, 2002), it was anticipated that site elevation would be a significant predictor of acid neutralizing capacity. It should be noted that site elevation was also included in the original MAHA data, but information was not complete for all sites.

Strahler stream order (Strahler, 1964) is a method of designating the location of a stream segment in a stream network. A Strahler order of 1 indicates a headwater reach of a given stream. When two streams of order 1 merge, a stream is subsequently classified as a second order stream. Where two streams of order 2 merge, a stream is subsequently classified as a third order stream. And so on. In general, when two streams of the same order merge together, a stream section with a higher Strahler order code is created. When two streams of different order merge together, the resulting stream has Strahler order equal to the maximum of the two previous orders. Strahler order is a very approximate measure of stream size.

Finally, GIS modeling generated locations for the sampled sites using Alber's Equal-Area Projection Coordinates (Alber's Equal Area Conic, 2002). Latitude and longitude were also available for each site, but we elected to use the Alber's coordinate system to map locations. Both coordinate systems do tend to distort Euclidean distance,

but Euclidean distance using Alber's Equal-Area Projection coordinates tends to be less distorted than Euclidean distance computed using Latitude and Longitude (Theobald, 2002). For example, in the MAHA region, one degree of longitude is approximately 75 miles in length. One degree of latitude is approximately 60 miles in length. If a given site is selected and the number of sites with a radius of one degree are counted, the maximum distance between sites could be anywhere between 60 and 75 miles away. Therefore, the interpretation of Euclidian distance from one point to another depends not only upon the distance in degrees but the angle between the sites. Alber's Projection Coordinates have the same interpretation of distance in both the N/S and E/W directions

Thematic Mapper satellite imagery was available for 515 of the possible sites, and 488 of those 515 sites included data for both satellite imagery and ANC.

The following predictors formed the basis of the prediction model analysis:

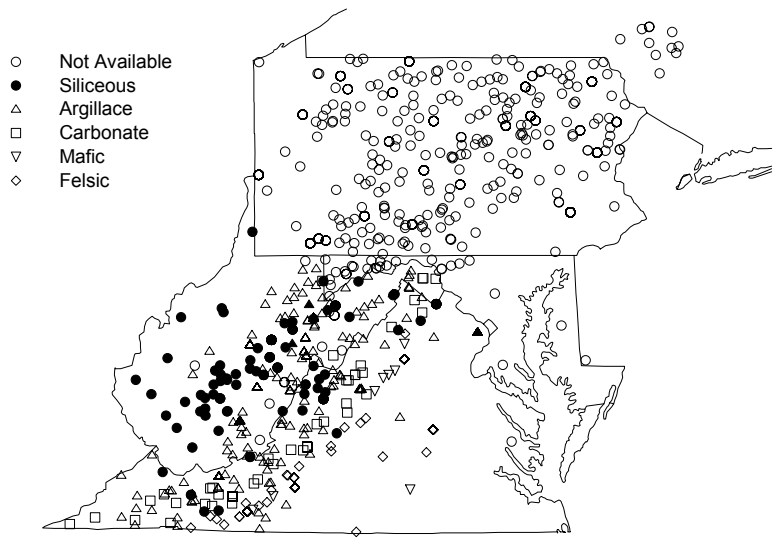
- Elevation
- Strahler Stream Order
- Bedrock Geologic Class
- Alber's Projection Coordinates
- 15 Thematic Mapper Satellite Imagery classifications (Table 1)

D. Availability of Data for Sampled Sites

Incomplete data was a concern in this analysis. It was previously stated that EMAP had sampled 579 sites resulting in 699 individual observations of the response, ANC. After the elimination of the five identified outliers, the data consisted of 574 sites

resulting in 694 observations. Elevation, Strahler stream order, and Alber's projection coordinates were available for all 574 sites. But bedrock geology and TM satellite

Figure 3: ANC Sites Identified by Bedrock Geologic Class



imagery data were not available for all 574 sites. In fact, a vast majority of sites were missing one of those two groups of variables.

Knowing that geology was most likely a significant predictor of ANC, we needed the bedrock geology for each of the sites included in the analysis. But at the time of the original data analysis, bedrock geologic data for sampled sites in Pennsylvania were not available. Thus, all sites for which bedrock geology was unavailable were eliminated from the analysis (Figure 3). This resulted in the elimination of 293 sampled sites, or about half of the possible observations, leaving only 345 of the original 699 observations.

The number of sites available for analysis continued to decrease when Thematic Mapper satellite imagery information was added to the data. Including the available TM information at sampled sites left us with 292 observations at 242 sites.

The issue of repeated observations at sample sites also needed to be addressed. Out of the 242 sites where complete predictor and response information were available, only 21 of these sites were visited more than once. The concern was that the multiple observations at these 21 sites would have more influence on the eventual model by weighting these sites more than the 221 others. To ensure that each site would contribute equal weight to the eventual model estimate, all observations except the first visit to each site were eliminated.

Out of the 242 identified sites, four belonged to the "unclassified" geologic class. Due to the small number of sites within this class and the lack of a cohesive interpretation for this geologic class, these three sites were not included in the analysis.

Therefore, after starting out with 699 observed values of ANC in the MAHA region, the final data set consisted of 238 observed values at 238 individual sites, with complete response and predictor information at each site. This data set will be referred to as DARM (Data Assisting Remote Modeling).

III. Overview of ANC and Predictors of ANC in the MAHA Region

A. Acid Neutralizing Capacity

Even after several observations were eliminated from the EMAP study, acid neutralizing capacity for the DARM data set still had a significantly skewed distribution (Figure 4). The summary statistics for the distribution of ANC are listed in Table 3. The range of ANC in DARM was very large, from –194 to 5620 μeq/L. Further, the median of 336 indicated that a majority of the sites were indeed not sensitive to acidification. Only 30.7% of the observed values of ANC were between 0 and 200, the range that would be considered sensitive. Therefore, it appeared that the original decision to look at the full range of ANC levels was most likely the correct one.

All of the 238 sites included in the DARM data set described above were located in Virginia and West Virginia (Figure 5). In Figure 5, circles are centered on the sampled site with diameter size representing the relative magnitude of positive levels of ANC. Triangles represent the location of a sampled site with the size of the triangle representing the relative magnitude of negative levels of ANC. The large, positive values of ANC tended to lie at the base of the eastern slope of the Appalachian Mountains. There were also some large values of ANC at the base of the western slope of the Appalachian Mountains as well, but they were not nearly as numerous. Small ANC levels were scattered throughout the region. Sites with negative ANC values, of which there are only nine, all appear near the peak of the Appalachian Mountains.

Figure 4: Distribution of ANC from DARM

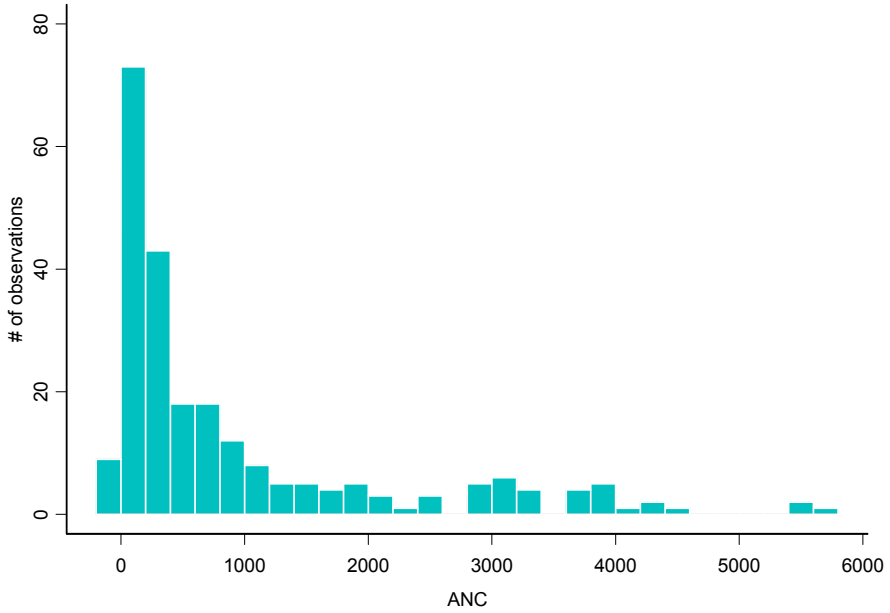


Figure 5: ANC Sites Designated by Magnitude
(Circle = Positive, Triangle = Negative)

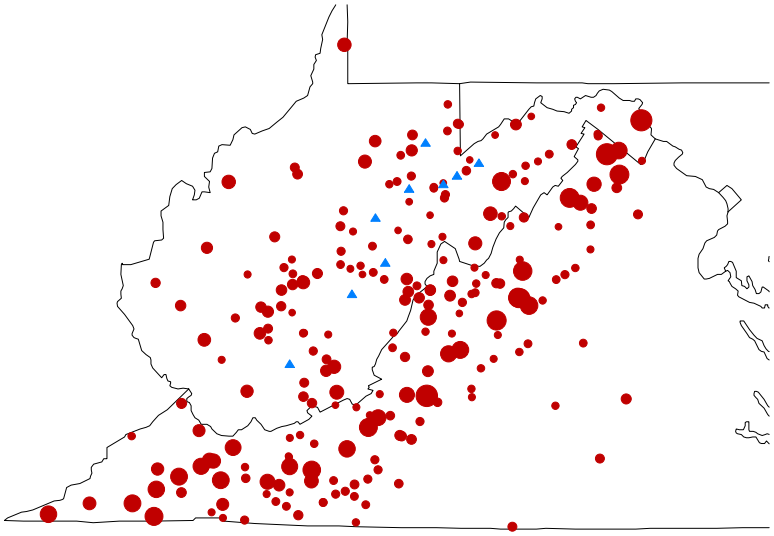


Table 3: Summary Statistics for ANC (μeq/L)

Mean	924
Median	336
Minimum	-194
1st Quartile	130
3rd Quartile	1138
Maximum	5620

The distribution of ANC in each of the Thematic Mapper classes was skewed as well. By comparing the mean and median of each of the thirteen Thematic Mapper classes, each distribution appeared highly skewed to the right, except for Deciduous Forest (Table 4). Four out of the thirteen TM classes did not appear in over half of the watersheds sampled, demonstrated by having a median of zero. This absence of certain landscape characteristics in the watershed above a sampled site was very prevalent in each of the TM variables, except for Deciduous Forest, which was identified somewhere above each of the sites in DARM. Deciduous Forest was the dominant land cover in the MAHA region, for the average percentage of each watershed above a sampled site covered by deciduous forest is 63.8%.

Finally, the distribution of elevation ranged from a minimum of 81 meters above sea level to a maximum of 1172 meters (Table 4), and was approximately normally distributed.

Plots of each of the predictors versus acid neutralizing capacity were created in an attempt to detect any initial relationships between the two. Few of these plots revealed any bivariate relationship between the predictors and ANC. The only predictor that revealed any sort of significant relationship was bedrock geology (Figure 6). Sites that

Table 4: Summary Statistics of Percentage of Watershed TM Classifications

TM Variable	Mean	Median	Minimum	1st Quartile	3rd Quartile	Maximum
Transitional	0.36	0.01	0.00	0.00	0.16	9.32
Quarry	0.43	0.00	0.00	0.00	0.00	26.15
Emergent	0.06	0.00	0.00	0.00	0.03	2.12
Woody	0.12	0.00	0.00	0.00	0.02	5.66
Deciduous	63.80	67.73	6.57	48.70	81.81	98.86
Mixed	12.08	10.57	0.00	6.96	15.97	46.90
Evergreen	5.84	2.45	0.00	0.66	7.54	73.51
Probable	7.99	3.20	0.00	0.21	12.51	72.22
Row Crops	2.33	0.72	0.00	0.11	2.42	21.29
Pasture	6.27	1.74	0.00	0.03	8.79	56.00
Water	0.14	0.03	0.00	0.00	0.13	3.46
Urban – Low	0.51	0.01	0.00	0.00	0.12	29.97
Urban – High	0.07	0.00	0.00	0.00	0.00	7.74

Table 5: Summary Statistics for Elevation (m)

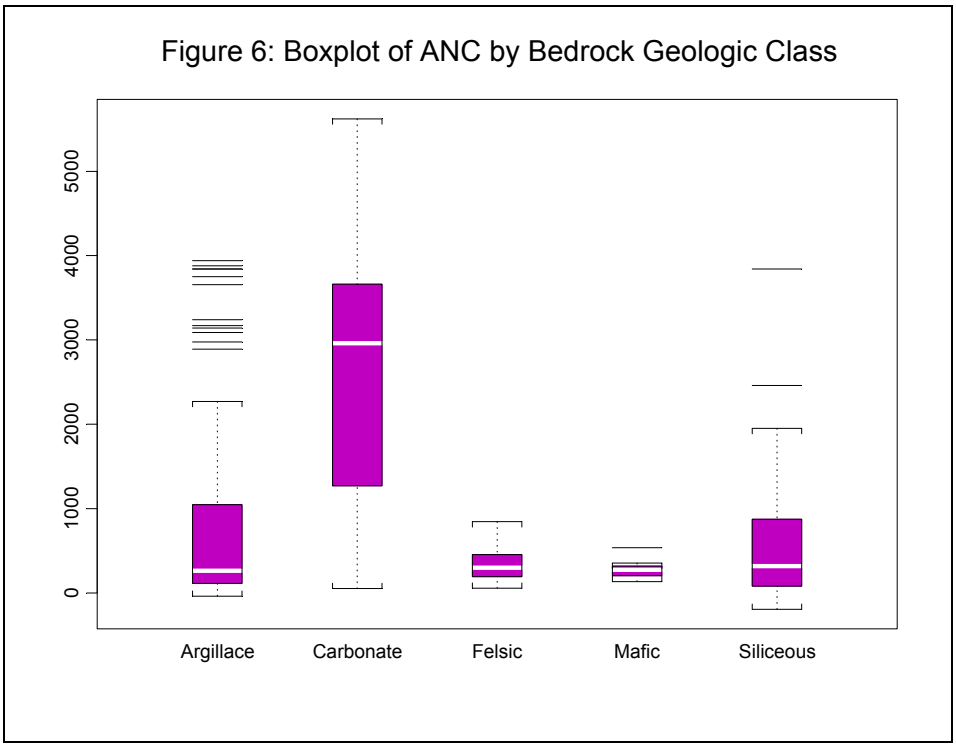
Mean	Median	Minimum	1st Quartile	3rd Quartile	Maximum
548.5	534	81	369.8	707.8	1172

Table 6: Bedrock Geologic Class – Approximate Percentage of 238 Sites

Geologic Class	Argillace	Carbonate	Felsic	Mafic	Siliceous
Percentage of DARM Sites	46%	13%	11%	3%	27%

Table 7: Strahler Stream Order – Approximate Percentage of 238 Sites

Strahler Stream Order	Order 1	Order 2	Order 3
Percentage of DARM Sites	41%	26%	33%



lay on carbonate bedrock appeared to have significantly higher values of ANC associated with them than each of the four other geologic classes.

A nonparametric Kruskal-Wallis test indicated an association between bedrock geologic class and acid neutralizing capacity. A test of the equality of variances within geologic classes indicated that the variances should be regarded as unequal. Therefore, the nonparametric Kruskal – Wallis rank test was used instead of a typical analysis of variance. The p-value for this test was less than 0.001. So, it could be concluded that there was a significant difference between the geologic classes. Multiple pairwise comparisons between geologic classes were constructed using the average class rank in order to see which of the geologic classes were significantly different (Neter, Kutner, Nachtsheim, and Wasserman (NKNW), 1996, p.777-780). Table 8 gives the results of these comparisons at a family error rate of 0.05. Each of the pairwise comparisons involving carbonate did not include the value of zero, indicating a

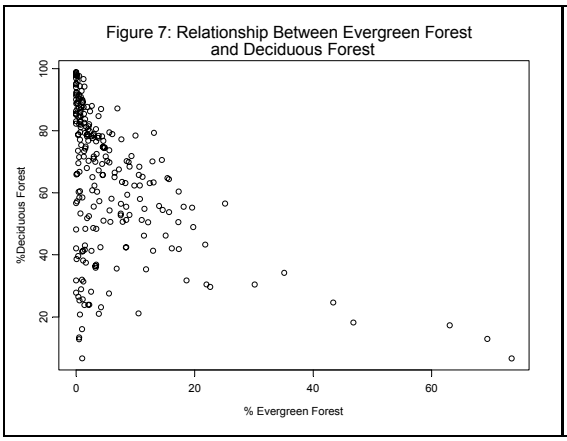
Table 8: Pairwise Rank Comparisons of Bedrock Geologic Classes

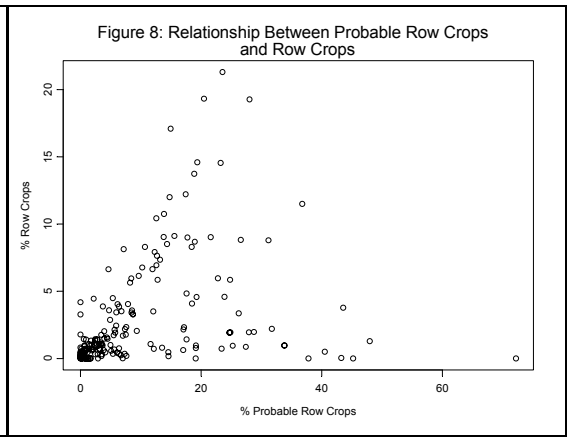
Pairwise Comparison	Lower Bound	Upper Bound	Significant Difference?
Argillace – Carbonate	-119.3	-40.7	Y
Argillace – Felsic	-37.6	48.0	N
Argillace – Mafic	-61.4	80.2	N
Argillace – Siliceous	-28.5	34.3	N
Carbonate – Felsic	33.2	137.2	Y
Carbonate – Mafic	12.8	166.0	Y
Carbonate – Siliceous	41.6	126.2	Y
Felsic – Mafic	-74.3	82.7	N
Felsic – Siliceous	-46.9	44.2	N
Mafic - Siliceous	-78.0	67.0	N

significant difference between ANC at sampled sites that lie on carbonate bedrock and those that do not. No other geologic classes had ANC values that were significantly different.

B. Relationships Among Predictors of ANC

Although none of the predictors had a significant linear relationship with ANC, several of the predictors did appear to have slight relationships with one another. There was an apparent negative relationship between percent deciduous forest and percent evergreen forest (Figure 7). This relationship can be easily explained. There exists only a





finite area in each watershed. If part of a watershed is covered exclusively by deciduous forest, then the available area to be covered by evergreen forest is diminished. The strength of the negative relationship between percent deciduous forest and percent evergreen forest may in part due to deciduous forest making up a majority of the examined watersheds. With such a large percentage of the total area covered by deciduous forest, the smaller percentage must be split between twelve (fourteen) different landscape types.

This negative relationship was not as common between each of the other TM classifications. In fact, many of the TM variables shared positive, rather than negative, relationships.

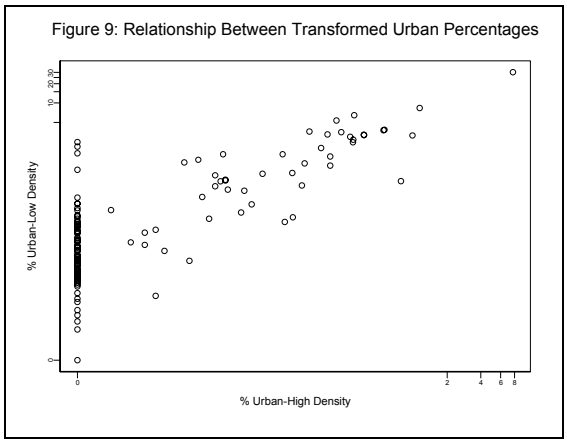
One of the most interesting positive relationships was between percent row crops and the portion of each watershed designated as percent probable row crops (Figure 8). It appeared from the figure that not one but two relationships were present. For some watersheds, as percent row crops increases, percent probable row crops increases as well, but at a slower rate. In other watersheds, the relationship is reversed such that as percent probable row crops increases, percent row crops increases at a slower rate. The reason for this positive relationship most likely lies in the difficulty interpreting and classifying land

cover based upon satellite imagery. Probable row crops "will sometimes be confused with other areas, such as grasslands that were not green during times of spring data acquisitions" (Vogelmann, 2002). The line between classifying a 30m resolution pixel as row crops or as probable row crops may be very slight. In retrospect, it may have proper to combine these two TM predictors into one class, but that was not done in this analysis.

In discussions with Dr. Alan Herlihy of Oregon State University, it was pointed out that none of the predictors included in the analysis are completely independent. In fact, each of the Thematic Mapper classifications are related to each other, as well as to geology and elevation. For example, certain bedrock types may be found predominantly at certain elevations and may be related to certain types of soils. Each of the predictors blend together and influence one another in some way. Therefore, at all stages of analysis, care was taken to examine the relationships between predictors to see if any significant relationships could be determined.

C. An Initial Model to Predict ANC

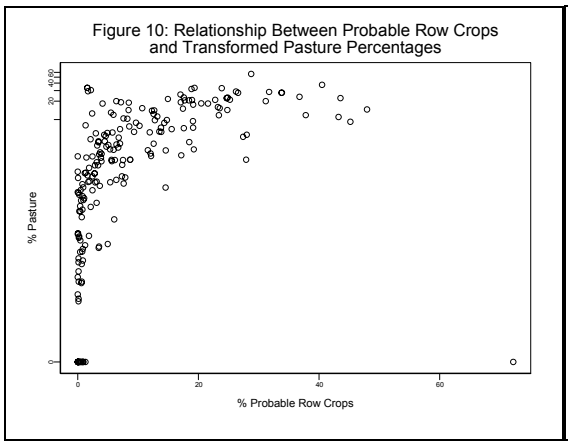
To further explore the relationship between ANC and the explanatory variables, an initial linear regression model was fit using all identified predictor variables. After fitting this model and examining the corresponding residual plots, it appeared that several of the variables needed to be transformed in order to meet the basic assumptions of a multiple regression model. In fact, the only four of the continuous variables that were not transformed were elevation, percent probable row crops, percent row crops, and percent deciduous forest. The other ten continuous variables were all transformed using a natural

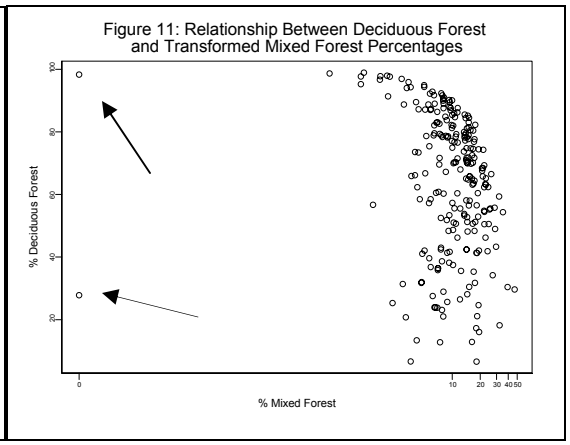


log to maintain consistency of transformation type. All but one of the Thematic Mapper classification variables did have zeros in them, representing 0% of a watershed classified as, say, emergent wetlands. Therefore, in order to perform a natural log transform, an arbitrary constant of 0.001 was added to each value of those percentages. Also, the response, ANC, was transformed using a natural log. An arbitrary constant of 500 was added to each value of ANC in order to make all values of the response positive. Once these transformations were applied, our normal regression assumptions were approximately satisfied. Throughout the rest of the analysis, these transformations were used to perform said analysis.

After transforming the necessary predictors, some of the relationships between predictors became stronger and more obvious. The question of whether or not this made a difference in the analysis is something that will be considered in Section III. Here we consider basic relationships between some of the transformed predictors.

Percent urban –high density and percent urban – low density display a positive, linear relationship (Figure 9). Approximately 79 percent of the sites have zero percent urban – high density in the watershed above. Those



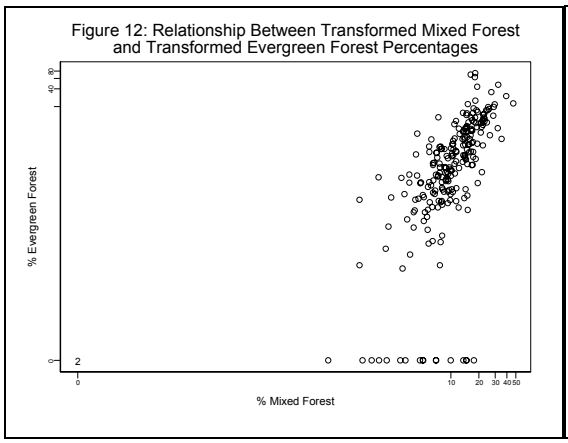


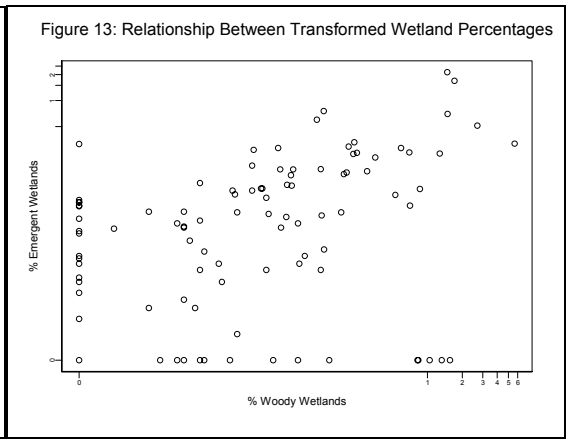
sites with zero percent urban – high density have an average of 1% of the watershed classified as urban – low density.

The relationship between percent row crops and percent pasture appeared to be curvilinear (Figure 10), if we ignored the point in the lower right hand corner that has a large percentage of row crops and no land designated as pasture in the watershed above it. This curvilinear relationship was likely caused by the fact that percent pasture was natural log transformed but residual plots for percent probable row crops did not indicate such a transformation was necessary.

The relationship between Deciduous Forest and Mixed Forest (Figure 11) may be explained in a manner similar to the relationship between the percent agriculture variables (Figure 10). There did appear to be a curvilinear relationship between Deciduous Forest and Mixed Forest, but the curve was not as pronounced as the above relationship between percent probable row crops and natural log transform of percent pasture. Two sites had large percentages of deciduous forest, but no mixed forest above the site; arrows point to these point in Figure 11.

Further, there appeared to be a positive linear relationship between percent mixed forest and percent evergreen forest (Figure 12). After disregarding the two apparent





influential points in the lower left hand corner, the linear correlation between these two predictors was 0.63.

Finally, a positive, linear relationship between the two percent wetland predictor variables was apparent (Figure 13). When examining only those sites where both wetlands classifications did appear in the watershed, the linear correlation between these two variables is 0.71. This relationship will be further pursued in Section IV.

It was stated earlier that none of these predictors are independent of one another, a fact demonstrated by Figures 9 though 13. One would anticipate that watersheds that are suitable for growing crops would see large percentages of all three classifications of agricultural land cover: row crops, probable row crops, and pasture. Further, land where woody wetlands are present would likely be near land where emergent wetlands are found. Are some predictors related in such a way that they are providing the same or overlapping information about ANC levels? This question will be addressed in Section IV.

IV. Selecting Predictors of ANC for a Non-Spatial Model

Can the observed relationships between ANC and the model predictors examined in Sections II and III be combined in such a way that an accurate prediction of ANC in the MAHA region can be made? In this section, we will examine an initial attempt at using the identified predictors to create a model for predicting ANC. We will also perform some preliminary analysis on the results and accuracy of this model.

A. Final Set of Predictors

The final DARM data set included 13 Thematic Mapper classification variables, five classes of bedrock geology, three classes of Strahler stream order, and elevation. Thus, there are 22 possible predictors. Because both bedrock geology and Strahler stream order are categorical variables, dummy indicator variables were created for four of the five bedrock geology classes (Argillace, Carbonate, Felsic, and Siliceous) and two of the three Strahler stream order classes (Stream order 2 and 3). The appropriate transformations of the predictors and response identified during the exploratory analysis were used in selecting a model.

B. Model Selection Procedures

Although initial analysis examining the relationships between ANC and the model predictors gave us an idea of which predictors should be include, it was unknown which combination of predictors, in conjunction, would yield the most accuracy. In order to find the best set of model predictors for predicting ANC, linear regression was initially used to select a model. Two different model selection procedures returned similar results.

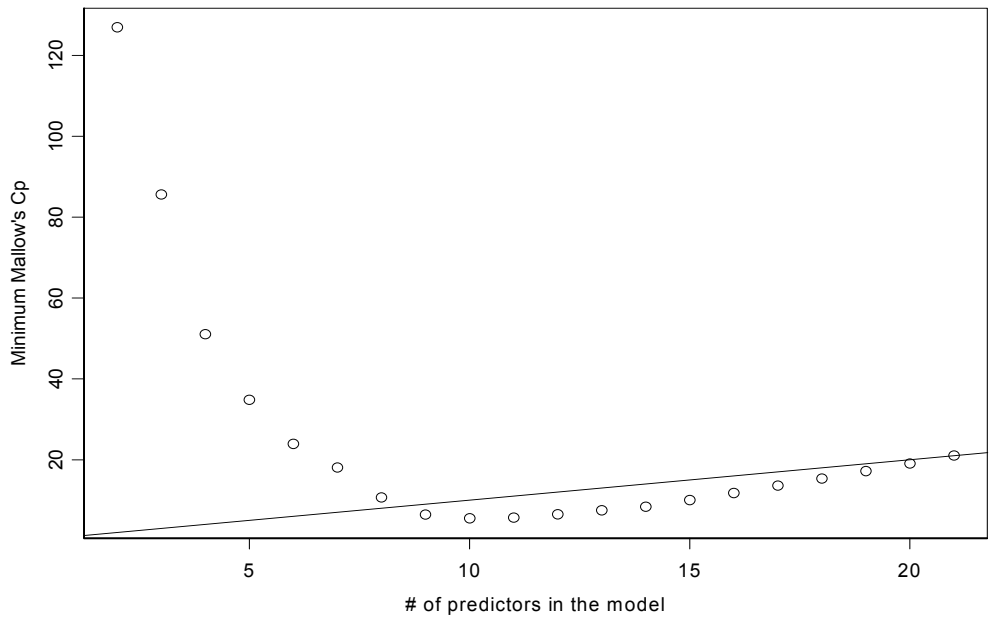
Table 5: List of Final Model Predictors (*Final – ANC)*

% Pasture
Elevation
% Quarry
% Probable Row Crops
% Woody Wetlands
% Emergent Wetlands
% Urban – High Density
Carbonate Bedrock
Felsic Bedrock

Efroymson's forward stepwise model selection procedure (Seber, 1977, p.376-377) was performed with all 20 predictors, using the Splus function 'stepwise'. The 'forward' process begins with a model consisting of an intercept only. Each forward step consists of adding to the current model that predictor variable giving the largest reduction of the residual sum of squares provided a certain level of significance or F-to-enter value. The stepwise process of the model selection procedure refers to the possible addition or deletion of predictors in the current model. When an independent variable is added to the model, partial correlations between the variables in the model and the response are considered to determine whether any of the predictors can be eliminated (based upon a specified F-to-exit value). Therefore, the number of predictors, p, increases and/or decreases at each step based upon the specified selection parameters, typically resulting in a "final" predictive model for the response. Using the default settings in Splus of an F-to-enter and F-to-exit value of 2, the final model produced by this selection procedure produced the subset of predictors of the natural log transform of ANC given in Table 5. This set of predictors will be referred to as *Final – ANC*.

The Leaps and Bounds model selection procedure (Furnival, 1974) was used in addition to Efroymson's forward stepwise model selection. The Leaps and Bounds

Figure 14: Results from Leaps and Bounds Model Selection Procedure



selection procedure examined all possible subsets of predictors, returning the best models of each subset size based upon a designated criteria. The criterion considered was minimum Mallow's Cp. Mallow's Cp estimates the total model mean square error divided by the true error variance (NKNW, 1996, p.341-345).

The leaps and bounds procedure using Mallow's Cp returned a minimum value of 5.48. This model corresponded to the model selected by Efroymson's stepwise procedure. Using the Cp criterion for model selection, Cp values for a model greater than the number of predictors, *p*, indicates model bias, or an inflation of the error variance. When Cp is less than the number of model predictors, the model is "interpreted as showing no bias." This result is attributed to sampling error. Figure 14 plots the minimum Cp values for the best model of each size, with the estimated intercept included as one of the predictors. The line, p = Cp, is included for perspective. Figure 14 displays that the set *Final – ANC* produces the overall minimum Cp in the Leaps and Bounds model selection procedure.

Table 6: Liner Regression Parameter Estimates for *Final – ANC* Model

Predictor	Coefficient	Standard Error	P-value
(Intercept)	7.506	0.157	<0.001
Probable	0.012	0.004	<0.001
Pasture	0.051	0.011	<0.001
Urban High Density	0.063	0.019	<0.001
Emergent Wetlands	0.046	0.021	0.033
Woody Wetlands	-0.073	0.019	<0.001
Quarry	0.020	0.012	0.084
Carbon	0.542	0.101	<0.001
Felsic	-0.273	0.102	0.008
Elevation	-0.001	<0.001	<0.001
$R^2 = 0.5781$		Overall p-value < 0.0001	

The form of *Final – ANC* is:

$$\mathbf{Y} = X \cdot \boldsymbol{\beta} + \mathbf{e} \tag{1}$$

where:

- $\mathbf{Y}$ = a vector of responses, ln(ANC + 500)
- X = a (238 x 10) design matrix of predictors
- $\boldsymbol{\beta}$ = a vector of model parameters
- $\mathbf{e} \sim N(0, \sigma^2 \cdot I)$

Using the designated model predictors, the estimated model parameters, standard errors and p-values are given in Table 6.

C. Multicollinearity

As discussed above, a concern about this model is the possible interrelationships between the predictors. One manifestation of this problem is multicollinearity. The term

multicollinearity describes a situation where predictor variables are correlated among themselves. When multicollinearity is present among predictors in a given model, the estimated regression coefficients tend to have large sampling variability (NKNW, 1996, p.290). This means that the estimation of the true regression coefficients will vary drastically from sample to sample, indicating little information regarding the true regression parameters.

The possibility of multicollinearity within the predictors of *Final – ANC* was examined by calculating the Variance Inflation Factor (VIF) of each predictor in *Final ANC.* The variance inflation factor for predictor, *k*, is equal to

$$\mathrm{VIF}_k = \frac{1}{\left(1 - R^2{}_k\right)}$$

where $R^2{}_k$ is the coefficient of multiple determination when a predictor X_k is regressed on the *p*-2 other predictors in the model (NKNW, 1996, p.385-388). Typically, a variance inflation factor of 10 or more indicates the presence of multicollinearity between two or more of the model predictors. The variance inflation factors of the final nine predictors, plus one, were examined. It was previously noted that the two % urban TM variables appeared to be related. Therefore, % urban – low density was included in order to see if this previously identified association significantly altered the model *Final – ANC.* The results of the VIF analysis are listed in Table 7. None of the calculated variance inflation factors were greater than ten. Therefore, multicollinearity did not appear to be present within these ten predictors, at least as measured by the VIF.

Table 7: Variance Inflation Factors of *Final – ANC* plus % Urban – Low Density

Predictor	VIF
% Probable Row Crops	1.73
% Pasture	1.90
% Emergent Wetlands	2.37
% Woody Wetlands	2.12
% Urban – Low Density	3.10
% Urban – High Density	2.20
% Quarry	1.27
Elevation	1.17
Carbonate Bedrock	1.34
Felsic Bedrock	1.13

Although none of the VIF values were greater than 10, it was noted that the highest VIF values belonged to the two percent urban area predictors, as well as the two percent wetlands predictors. The relationships between these variables and the effect of including and/or excluding certain combinations of these variables were examined further. When all of the above ten predictors were included in a regression model, each predictor was significant at the $\alpha = 0.10$ level of significance except for percent urban – low density and percent quarry. When percent urban – low density (which had a p-value of 0.41) was eliminated the model, the resulting predictors were the same ones identified in the *Final – ANC* model.

When both of the percent urban predictors were included in the model, only percent urban – high density was significant. If one of the two percent urban predictors were included in a model, then the included percent urban predictor was significant at the $\alpha = 0.01$ level of significance. The implications of this will be further described below.

The relation between percent emergent wetlands and percent woody wetlands was different from the relationship between the percent urban predictors. When both percent

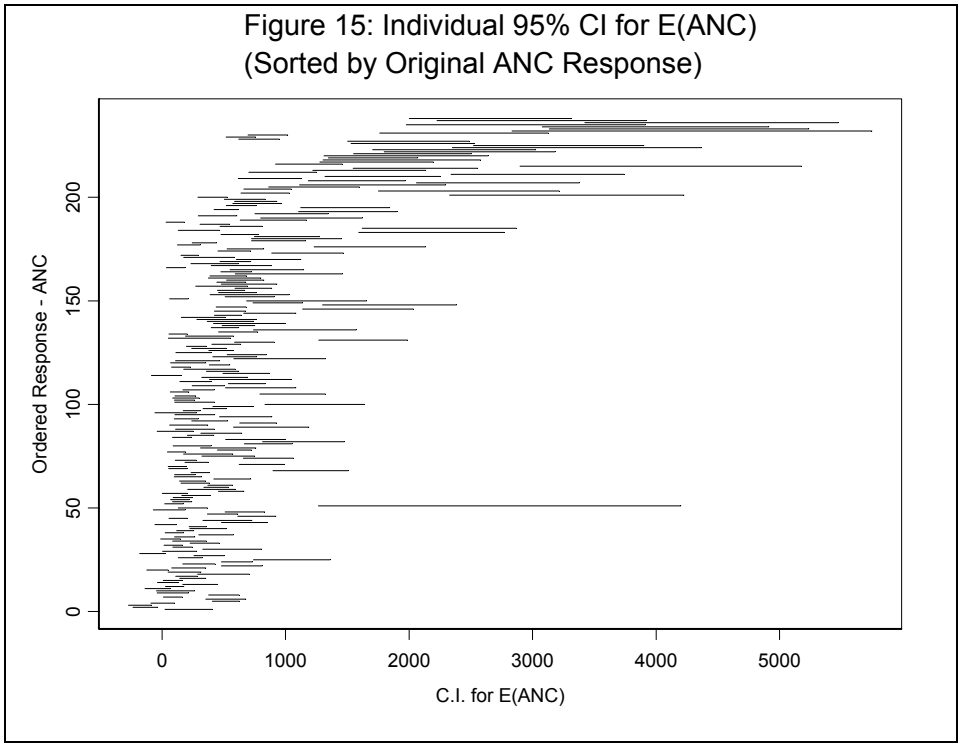
wetland predictors were included in a model, both predictors were significant, but their regression coefficients indicated opposite effects on acid neutralizing capacity. Percent woody wetlands had a negative coefficient and percent emergent wetlands had a positive coefficient.

Percent woody wetlands was always significant at the 0.01 level of significance in the model without percent emergent wetlands. But when excluding percent woody wetlands from the possible regression models, percent emergent wetlands quickly became insignificant with p-values greater than 0.50. Further, the coefficient was always negative for each percent wetlands predictor when only one of the two was included in a model.

Therefore, the following conclusions regarding the relationships between the percent wetlands and percent urban predictor variables can be made:

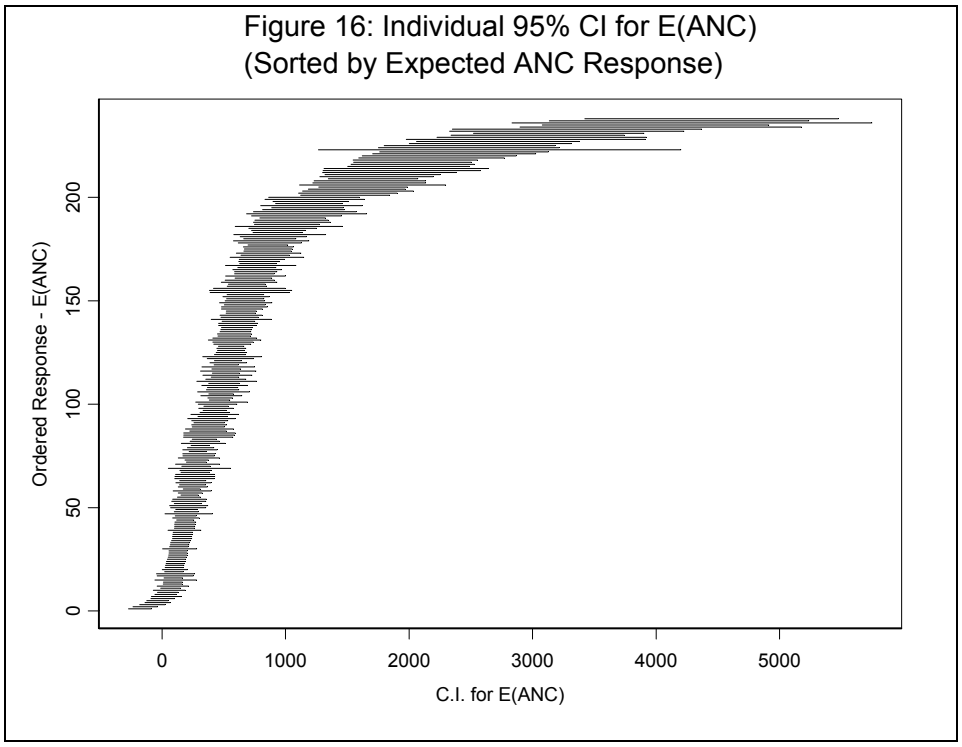
- The two percent urban variables provide nearly equivalent information about ANC; only one of the two needs to be included in the final model.
- Percent emergent wetlands was not a significant predictor of ANC, but its negative relationship with percent woody wetlands did significantly affect the predictive ability of a regression model. Due to this relationship, either both or neither of them should be included in the final model. The scientific reasoning for this is unknown.

Therefore, given the available information regarding the location and landscape surrounding each of the sampled stream sites, the simplest and most informative least squares regression model for predicting acid neutralizing capacity included the predictors identified as *Final – ANC*.



D. Some Concerns About The Inferences From The *Final – ANC* Regression Model

How well did the *Final – ANC* model acid neutralizing capacity? Figure 15 displays individual 95% confidence intervals for the expected value of ln(ANC+500), sorted by actual levels of ANC. The first confidence interval corresponds to the lowest observed value of ANC. As the observed level of ANC increases, the expected value of ANC tends to increase as well. This is not a strong relationship. Several low values of observed ANC are associated with very high predicted levels of ANC at that site. Likewise, some high levels of observed ANC predict low levels of ANC based upon the model, *Final – ANC.* This indicates that *Final – ANC* may not do a satisfactory job of predicting acid neutralizing capacity. Note the confidence intervals presented here are individual confidence intervals



and thus do not account for the probability of a type I error when 238 intervals are computed.

A potential problem is presented in Figure 16. Figure 16 displays the same individual 95% confidence intervals for the expected value of ANC sorted by the expected value of ANC. When the expected value of ANC is low, the confidence intervals are narrower than when the expected value of ANC is high. There may be two possibilities for explaining this observation. First, the distribution of ANC is highly skewed, and a large majority of the data lies between 0 and 1000 μeq/L. Therefore, the uncertainty surrounding the estimates of ANC at lower levels is smaller due to the increased amount of available information. Second, the variance of the response, ANC, is not constant. Previous studies have indicated that ANC is heterscedastic, i.e. as ANC

increases from zero, the variance increases as well (Stodderd, Urquhart, Newell, and Kugler, 1996). This issue will be considered further in Section VI.

The concerns described here led us to consider extensions of the multiple regression model described above. In Section V, we consider a model to account for spatial correlation between observed ANC values.

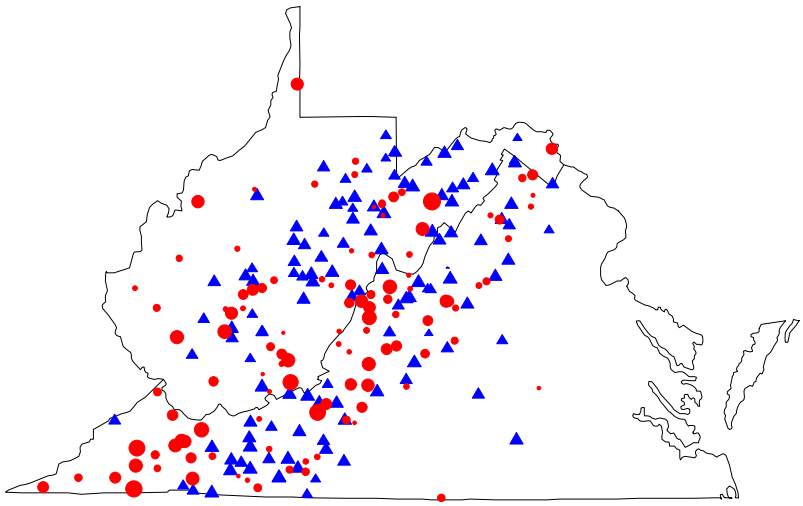
V. Modeling the Spatial Correlation Between ANC Values

The multiple regression model with the *Final – ANC* predictors did not appear to do a satisfactory job of explaining acid neutralizing capacity in this portion of the MAHA region. The ecological nature of the research led to a question of possible spatial correlation in the hope of better explaining the observed variability of ANC. Given the model *Final – ANC*, can it be concluded that there is spatial correlation in the residuals of this model, i.e. is there spatial correlation present in acid neutralizing capacity that cannot be explained by a linear regression model with independent errors (Figure 17)?

The environmental nature of the problem indicated the possibility of spatial correlation between ANC at streams that are in close proximity. The closer stream sites are to one another, the more likely they will exhibit similar features. Analysis of spatial correlation examines the relationship of the response between sites based upon the distance, and possibly direction, between those given sites. The model *Final – ANC* assumed that the responses at each sampled site were independent, but sampled sites that were close to one another likely shared characteristics such as elevation, topography, bedrock geology, etc. This would mean that the errors of our linear regression model would likely be dependent. If this were true, knowing ANC at one site may tell us something of the value of ANC at another site. By choosing to investigate the spatial relationship between sampled sites, we hoped to explain more of the observed variation in ANC than was accounted for by the *Final – ANC* model.

In order to examine whether spatial correlation existed within the residuals of the model *Final – ANC*, we elected to examine spatial correlation as measured by Euclidean distance rather than limit our analysis to spatial correlation between sites within a single

Figure 17: Residuals of Regression Model by Magnitude
(Circle = Positive, Triangle = Negative)



stream or watershed. Sites located on the same stream are likely related. Stream sites located close to one another but on separate streams may also be related, but this relationship may be different from the relationship between ANC values for different streams. Unfortunately, the DARM data did not have sufficient data for individual streams or watershed to model these relationships. Therefore we considered spatial correlation based on Euclidean Distance.

A. Alber's Equal Area Projection Coordinates

The location of each sampled site was identified using Alber's Equal Area Projection coordinates and Euclidean distances between sampled sites were calculated using these coordinates as well. Both projection coordinates were centered and divided by 100,000. This was done to maintain the equal area relationship between the X (East and West direction) and Y (North and South direction) coordinates of the projection

system as well as make using the coordinates more manageable numerically. On this new scale, a Euclidean distance of one represents approximately 65 miles, a distance of two represents 130 miles, etc. The maximum distance between any two sampled sites was approximately 350 miles, or roughly 5.2 on our coordinate scale. Therefore, we considered a maximum Euclidean distance of 2.6 for our analyses of spatial correlation, or half the observed maximum distance (Webster and Oliver, 2001).

B. Basics of Spatial Correlation

Spatial correlation is typically estimated using variograms or semi-variograms. A variogram estimates the covariance between two sites separated by a given distance, *h*. The semi-variogram is:

$$\gamma(\mathbf{h}) = (1/2)\mathrm{var}[(Y(\mathbf{s}) - Y(\mathbf{s} + \mathbf{h})] \qquad (2)$$

where

$$Y(\mathbf{s}) = \mu + \varepsilon(\mathbf{s})$$

is a stationary random process with mean μ and covariance function

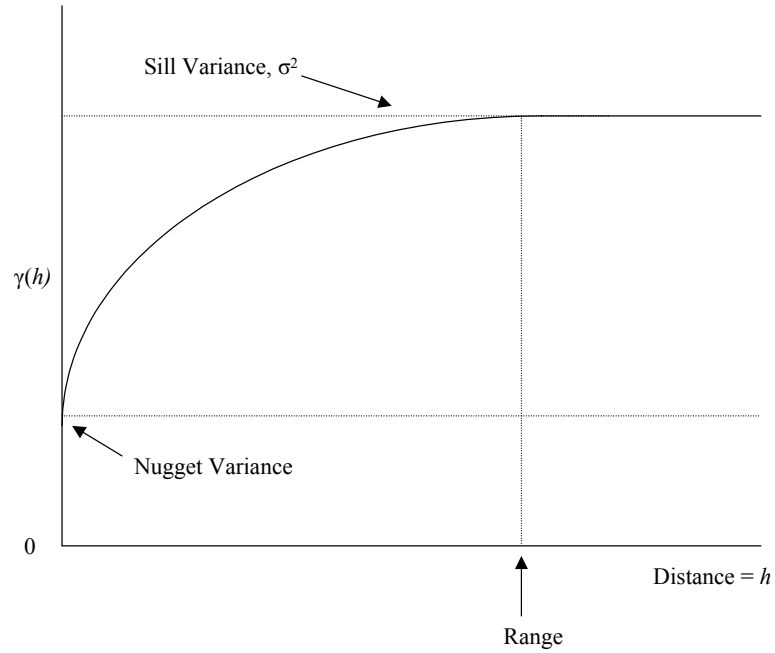
$$C(\mathbf{h}) = E[\varepsilon(\mathbf{s})\ \varepsilon(\mathbf{s}+\mathbf{h})].$$

The semi-variogram is one half the variogram, but the two terms are used interchangeably (Webster and Oliver, 2001, p.54). If the spatial process is second-order stationary, then the variogram and covariance are equivalent with the following relationship,

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

Second-order stationarity means that the mean of the spatial process is location invariant and the autocorrelation between two values of the spatial process depends only on the

Figure 18: Sill Variance, Nugget Variance, and Correlation Range



distance between the locations (Thompson, 2001, p.13). For our model, Y(s) is the ANC value at location s and μ is the regression function estimated from (1) assuming independent errors.

There are three characteristics that are typically included in any spatial covariance function: the nugget variance, sill variance, and correlation range (Figure 18). The nugget variance, or nugget effect, occurs when a spatial process is discontinuous as the distance, *h*, approaches zero, i.e.

$$\lim_{h \to 0} \gamma(0) \neq 0$$

This nugget effect could possibly be attributed to measurement errors in the data values or to very small-scale irregularities near the sampled site (Ripley, 1981, p.50). The sill variance is typically the *a priori* variance of the process, σ^2 (Webster and Oliver, 2001, p.111).

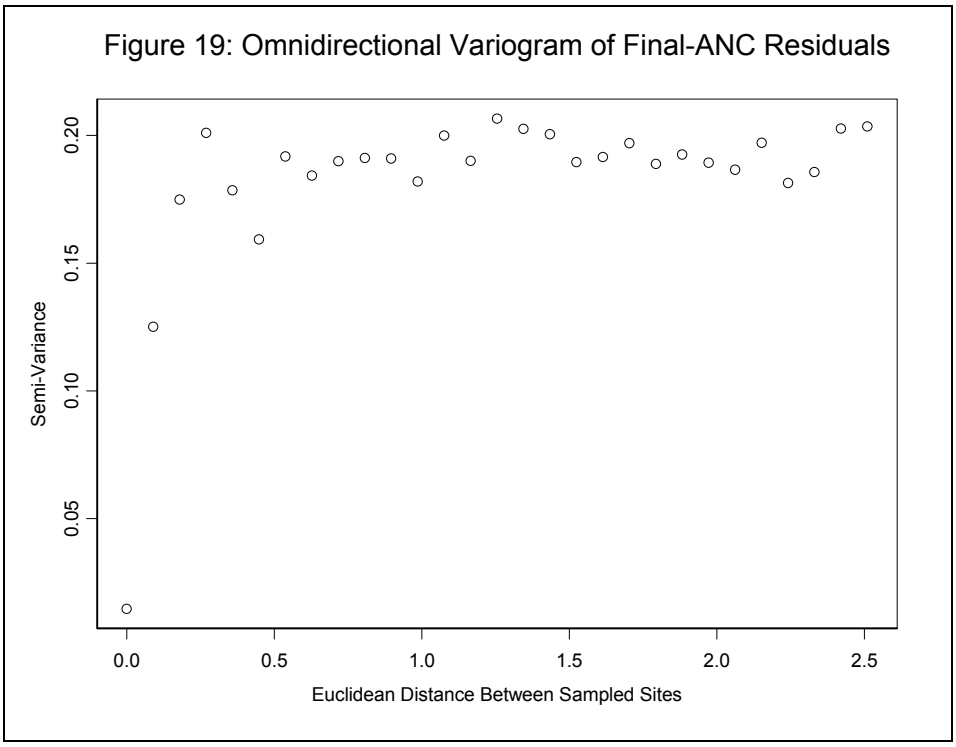
The correlation range is the range of spatial dependence. Sites separated by a distance greater than the correlation range are assumed to be spatially independent.

If spatial correlation is isotropic, then the correlation between two sites is based solely on the distance between those sites regardless of the directional relationship between them. The relationship between all sites and a given sampled site a set distance from that given site is considered the same, no matter what the direction. This is also termed omni-directional spatial correlation. Anisotropic spatial correlation examines correlation based on distance and a given direction from one site to another. If anisotropy is present, the correlation between a given site and all sites a set distance from that site changes as the angle of the relationship changes. This is also termed directional spatial correlation.

The analysis of spatial correlation was performed using several functions from the Spatial Library for Splus created by Dr. Robin Reich and Dr. Richard Davis of Colorado State University. All of the variograms and variogram modeling was performed using their Splus code.

C. Isotropic Spatial Correlation

At a maximum distance of 2.6, the omni-directional semi-variogram (Figure 19) indicated that there was little or no spatial correlation present in the residuals of *Final – ANC*. Therefore, it appeared that *Final – ANC* did account for the assumed spatial correlation in the response. However, further analysis indicated that an omni-directional analysis was deemed inappropriate for this region.



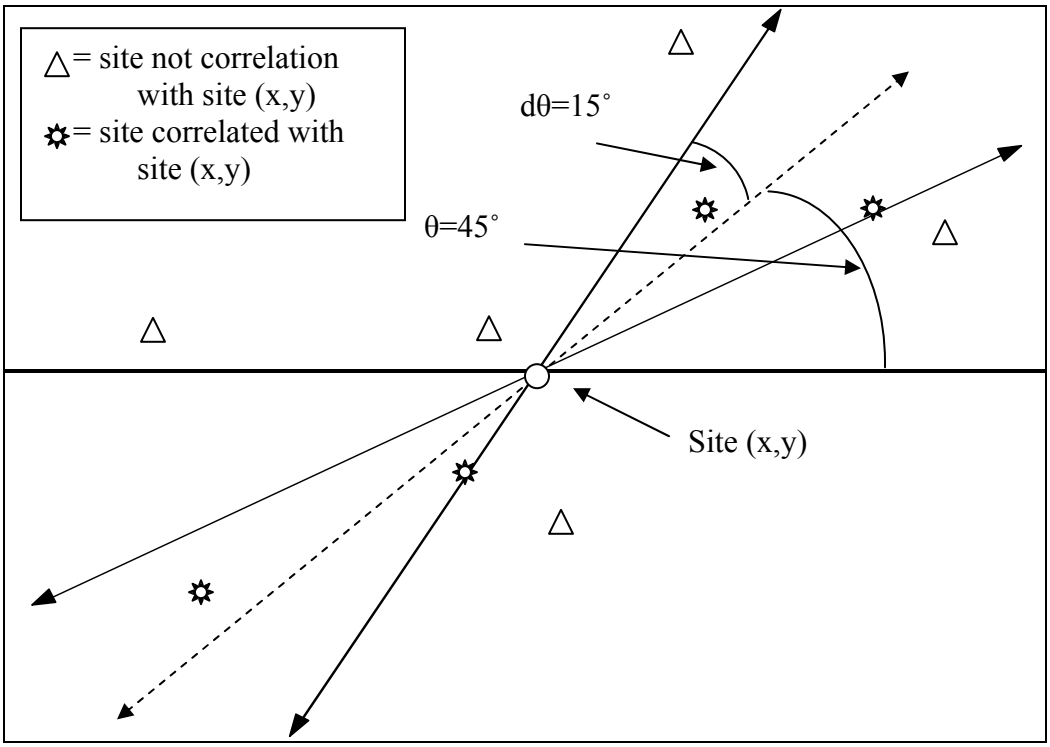
D. Anisotropic Spatial Correlation

The general direction of sampled sites in the MAHA region of Virginia and West Virginia closely follows the Appalachian Mountains, which run in a SW-NE direction. Any spatial correlation in this region would likely run in that specific SW-NE direction, rather than in all directions.

Anisotropic spatial correlation looks at the spatial correlation in a given direction from any given point. Given a certain direction and span from a sampled site, only points that fall within that span are used to calculate the correlation between sites. The angle of the span, $d\theta$, is the size of the angle above and below the angle of correlation, θ. If a site does not lie within the specified angle and span, the correlation between the two sites is assumed to be zero. Figure 20 displays the presented concept. In Figure 20, the angle of

correlation is θ = 45˚ and the span is dθ = 15˚. Any sites within the angles 30˚ and 60˚ of a given site are examined for correlation. Further, the correlation is examined

Figure 20: Anisotropic Spatial Correlation for Site at Longitude and Latitude (x,y).



in both directions. When the angle of correlation is 45˚, 45˚ + 180˚ = 225˚ is also included in the correlation/covariance calculation. Therefore, from a given site, possible spatial correlation exists only with sites that lie between 30˚ and 60˚ or 210˚ and 240˚ from the given site(x,y). If the angle between two sites does not fall within the stated ranges, it is assumed no correlation exists between those two sites.

In order to examine the possible anisotropic (directional) spatial correlation found in the residuals of the *Final – ANC* regression model, semi-variograms were estimated at different angles (θ = 0º to 165º) and different spans (dθ = 5º to 60º).

At different angles (θ), the spatial correlation within the residuals shifted. As expected, the semi-variograms for θ = 0º to 165º showed vast differences. In particular,

there appeared to be significant spatial correlation between the angles (θ) of 40° and 60°. At all other angles the spatial correlation was minimal.

One specific result based upon the angle of correlation was given special notice. At 90° angles to the aforementioned range of maximum spatial correlation (40° to 60°), no correlation in the residuals was present until a Euclidean distance between points of approximately 1.5. At this point, the residuals appeared to become correlated again. While the direct cause of the association was not known, it was conjectured that the apparent correlation between points at this distance whose relationship was between 130° and 150° was due in some respects to the Appalachian Mountains themselves. Stream sites that share similar characteristics yet are located on opposite sides of the Appalachian Mountains may be the cause of this observed correlation. While the angle of maximal correlation runs parallel to the mountains, the angle at which this possible “hole effect" was observed runs perpendicular to the mountains.

Two different trends were observed when examining the effect of increasing the span (dθ) at a given angle of correlation (θ). As the span increased, the variability within the semi-variogram decreased for dθ between 5° and 30°. This was most likely due to the increased number of pairs of points included in the calculations for the semi-variogram, increasing precision. At very small spans (i.e. 5° to 15°), there may not have been enough information available to get a clear picture of the spatial correlation between points at a given distance of separation. But as the size of the span increased, the observed uncertainty in the spatial correlation at a given distance decreased, creating a “smoother” semi-variogram.

Secondly, it appeared that given a specific angle of correlation with a span between 15° and 60°, the effect of the span on the possible spatial correlation was minimal. At a given angle of correlation, the semi-variograms differed only slightly as the span increased within the stated range (Appendix C).

E. Anisotropic Spatial Model Selection

Semi-variograms (2) were calculated at several different combinations of angle and span in order to the find the direction and span of maximal spatial correlation. Many of these combinations produced semi-variograms that were similar in appearance. Therefore, in order to select the "best" model based upon the available information, two classes of covariance functions were fit to each semi-variogram: spherical and exponential.

The spherical variogram function is (Webster and Oliver, 2001, p.114):

$$\gamma(h) = \begin{cases} c_0 + (c - c_0)\left\{\left(\frac{3h}{2d}\right) - \frac{1}{2}\left(\frac{h}{d}\right)^3\right\} & \text{if } 0 \le h \le d \\ \text{c} & \text{if } h > d \end{cases}$$

where:

c = sill variance

c_0 = nugget variance

d = correlation range

The exponential variogram function is (Webster and Oliver, 2001, p.121):

$$\gamma(h) = \left\{ c_0 + (c - c_0)\left(1 - e^{-\frac{h}{d}}\right) \qquad if\ h \geq 0 \right.$$

where:

c = sill variance

c_0 = nugget variance

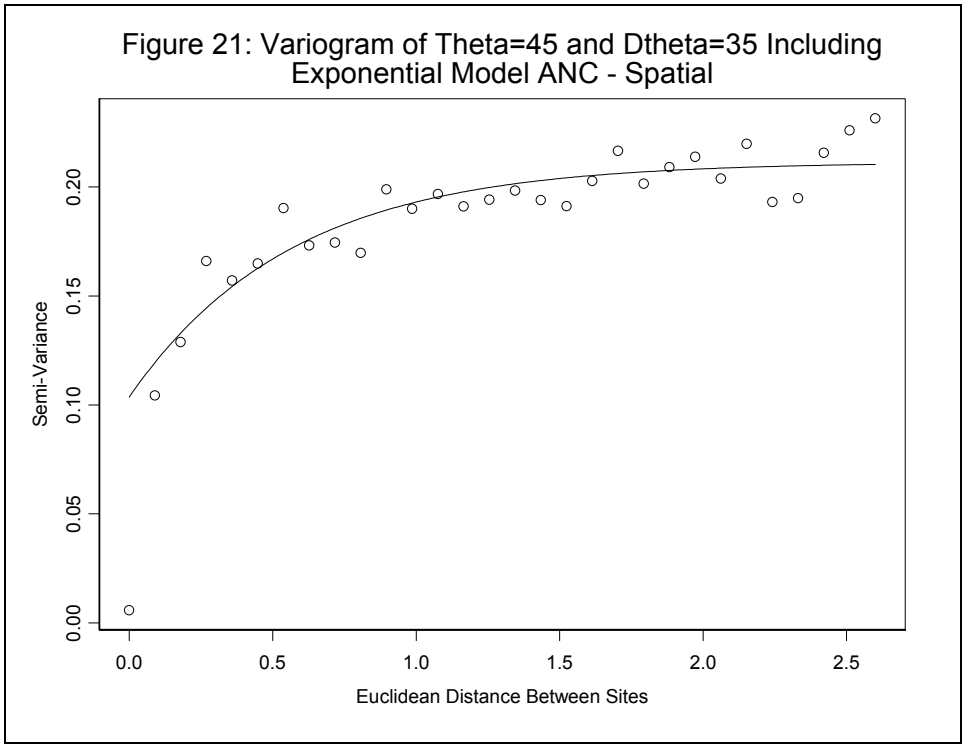
d = range parameter

$3 \cdot d$ = correlation range

For selection of the optimal combination of direction and span, the first point of each semi-variogram was eliminated from the covariance function estimation procedure. This point, (0,0), was produced by default by the Splus function 'variogrm', indicating that at a distance of $h = 0$, the estimated variance is zero. This was determined to be inappropriate because acid neutralizing capacity at a given site is not constant. There is variability in ANC at a given site. Therefore, this point was disregarded and only the estimated correlations at positive distances were used to calculate the covariance functions.

The function 'fitvar1' was used to fit each of the exponential and spherical covariance models in Splus (Code attached in Appendix D. 'Fitvar1' represents the described modification from the Splus function 'fitvar'). If the covariance class was designated as exponential, the 'fitvar' function estimated the parameters of the

exponential function to a given semi-variogram model. This Splus function used the non-linear minimization algorithm 'nlminb' to find the best combination of nugget variance,



sill variance, and correlation range for a given model type that minimizes the function error (Webster and Oliver, 1996, p.56-58). The function also estimated Akaike's Information Criterion (AIC) (Akaike, 1973). AIC is an order selection criterion used for comparing different models. Models with a small AIC relative to other models balance accurate estimation of the response and the tendency to overfit the model.

For each variogram, the 'fitvar1' function calculated the best estimated spherical and exponential variogram models, and Akaike's Information Criterion was examined for each of the estimated models. The exponential anisotropic variogram model with an angle of correlation of 45º and a span of 35º had the lowest AIC measure. This model also had the smallest mean square error (Figure 21).

Table 8: Model statistics of Quadratic Trend of Location (*Final – ANCquad)*

Predictor	Coefficient	Standard Error	P-value
X	-0.078	0.051	0.129
Y	-0.082	0.041	0.048
X^2	0.019	0.051	0.706
Y^2	-0.070	0.041	0.089
X*Y	0.085	0.059	0.151

F. Inclusion of Quadratic Trend of Location of *Final – ANC*

In addition to examining anisotropic spatial correlation within the residuals of *Final – ANC,* a quadratic trend of location was added to the model *Final – ANC*. Did the inclusion of this quadratic trend using Alber's projection coordinates eliminate the observed anisotropic spatial correlation (model *Final – ANCquad)*? The coefficients, standard errors, and p-values of the quadratic trend are presented in Table 8. Including the quadratic trend of location had little effect on the observed spatial correlation present in the residuals. The isotropic and anisotropic semi-variograms of residuals from *Final – ANCquad* indicated a decrease in the omnidirectional and directional spatial correlation from the observed correlation in *Final – ANC,* but the spatial correlation was not eliminated. Therefore, the previously identified anisotropic spatial analysis was deemed more appropriate for modeling acid-neutralizing capacity in this portion of the MAHA region.

G. Summary

Therefore, the final model anisotropic spatial model, *ANC - Spatial,* took the following form:

$$\mathbf{Y} = \mathrm{X} \cdot \boldsymbol{\beta} + \mathbf{e}$$

where:

- $\mathbf{Y}$ = a vector of responses, ln(ANC + 500)
- X = a (238 x 10) design matrix of predictors
- $\boldsymbol{\beta}$ = a vector of model parameters, estimated by the model coefficients in Table 6.
- $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathrm{C}(\mathbf{h}))$

If the angle between site_i and site_j is between 10° and 80° or 190° and 260°, then

$$\gamma(h) = \begin{cases} 0.104 + (0.211 - 0.104)(1 - e^{-\frac{h}{0.565}}) & \text{if } h \geq 0 \end{cases}$$

If the angle between site_i and site_j does not lie within 10° and 80° or 190° and 260°, then $\gamma(h) = 0$.

The *a priori* variance of this process is 0.211, the sill variance. This is the asymptote of the covariance function which best explains the spatial relationship. The estimated nugget variance was 0.104, or approximately 49% of the sill variance. Finally, the estimated correlation range was approximately 1.69 with a range parameter for the exponential covariance function of 0.565. Any sites separated by more than a Euclidean distance of 1.69 were not considered correlated (Figure 21).

IV. Conclusions and Future Work

A. Conclusions

The monitoring of acid neutralizing capacity in the Mid-Atlantic Highlands region of the eastern United States was mandated by the 1990 Amendments to the Clean Air Act. This monitoring takes time, money, and resources to accurately perform. These costs could be significantly decreased if an accurate model using remotely sensed information were available. In an attempt to meet this goal, I have examined several possible remotely sensed model predictors of acid neutralizing capacity, as well as the interrelationships between those predictors. Using multiple regression methods, an initial model to predict ANC was selected. Although this model did a fair job of explaining the variability in ANC, further analysis was required.

A spatial analysis of the residuals of this multiple regression model indicated the presence of anisotropic spatial correlation. The correlation was examined at different angles and spans, resulting in a direction of maximal correlation of 45° with a span of 35°. We then estimated a possible regression model with a directional spatial correlation structure from data gathered by the Environmental Protection Agency's Environmental Monitoring and Assessment Program, Geographic Information Systems, and Thematic Mapper Satellite Imagery.

This study concluded that directional spatial correlation between sampled sites needed to be accounted for. By not considering spatial relationships between sites, much of the variability in acid neutralizing capacity would remain unexplained. Although this result was significant, more research needs to be done in this area. Several different areas

of future and continuing work were identified, and it is hoped that the results and conclusions of this study will be beneficial in future analysis within this region.

B. Future Work

During research into spatial modeling of acid neutralizing capacity in the MAHA region of the eastern United States, several areas of further research and investigation were identified.

1. Measuring Predictive Ability of *Final – Spatial*

Future researchers may want to consider one or more methods for measuring the ability of *Final – Spatial* to predict acid neutralizing capacity in the MAHA region. First, how well does the model predict those observations used to create the model itself? Some form of cross-validation procedure would be helpful in assessing the accuracy of the model. A cross-validation procedure would take out one observation, refit the model using the procedure used to create *Final – Spatial*, and then see how accurately this model predicts the one withdrawn observation. Ideally, the procedure used to create *Final – Spatial* would prove robust, indicating that the accuracy of *Final – Spatial* was due to the modeling method and not a byproduct of the data alone.

The second possible method for measuring the predictive ability of *Final – Spatial* is to apply the model to similar areas within the MAHA region, such as Pennsylvania, that were not included in the analysis due to lack of information. Many sampled sites in Pennsylvania only lacked bedrock geologic data, and were eliminated from analysis on this basis alone. During the course of this research, geologic information

became available for the state of Pennsylvania. Therefore, a large majority of sampled sites in Pennsylvania now have complete predictor and response information. These data could be used to test the predictive ability of *Final – Spatial* in those areas not used in creating the model.

The third possible method of measuring the predictive ability of *Final – Spatial* involves predicting acid neutralizing capacity at unsampled sites in Virginia and West Virginia. Using the resources available through GIS and satellite imagery in conjunction with *Final - Spatial*, predictions of acid neutralizing capacity could be identified at specified stream sites. Once these sites have been identified, research teams could take samples at those sites, comparing the ANC results from the sample with the prediction made by *Final – Spatial*.

2. Weighted Least Squares

Stoddard, et al. (1990) showed that the variability of acid neutralizing capacity increases as the expected value of ANC tends away from zero. The variance of ANC increases as the absolute value of ANC increases. This was shown by looking at repeated measurements at several bodies of water, then measuring and comparing the mean and variance at each body of water. The solution of Stoddard, et al. (1990) to minimize or eliminate the effect of heteroscedasticity of the response, ANC, was to perform a Weighted Least Squares analysis (e.g., NKNW, 1996, p.400-409), weighting each stream by its inverse variance. In order to perform this analysis for the data considered here, further research must be done in order to get an accurate estimate of the variability at

each level of ANC. In the current analysis, this variability could not be accurately estimated due to so few repeated observations at sampled sites.

3. Bayesian Model Averaging

Using the given predictors identified previously, preliminary results indicated significant model uncertainty in the data. While a spatial linear regression model was identified, initial analysis using Bayesian Model Averaging (Hoeting, 1999) indicated that no one subset of predictors could sufficiently predict acid neutralizing capacity in this region. We used Adrian Raftery's 'bic.reg' Splus function (Statlib) to estimate the posterior model probability for each subset of predictors. While a few predictors were clearly identified as significant, this preliminary analysis indicated that the "best" models included a diverse selection of possible model predictors. No one combination of model predictors produced a large model posterior probability. Therefore, some form of Bayesian Model Averaging approach using available model predictors may produce more accurate estimates of acid neutralizing capacity in the MAHA region as compared to estimates from a single model.

4. Increased Number of Model Predictors

Due to the exploratory nature of this research and its role the initial stages of the STARMAP program at Colorado State University (STARMAP, 2002), lack of knowledge regarding available information played a key role in hindering research and model construction. The greatest unknown quantity was the vast information available through Geographic Information Systems. Misguided assumptions as well as

miscommunication further played a role in decreasing the effectiveness of this analysis. GIS modeling provided elevation, Strahler Stream Order, Alber's Projection Coordinates, and bedrock geology at each stream site. Most of these predictors displayed relationships with acid neutralizing capacity, but there may exist more and better predictors that are available and will likely increase the accuracy of a model explaining ANC.

It is correct to say that bedrock geology is an important predictor of ANC, as has been shown. But the bedrock geology variable used in this study did not take into account the geologic impact on streams of different types of bedrock over which the stream flows before it gets to the designated site. For example, the geologic bedrock at a given site may be argillace, but the majority of geologic bedrock in the watershed above that site may be carbonate, mafic, siliceous, etc. The nutrients picked up by the water as it flows over carbonate above a site resting on argillace bedrock has not been taken into account in our analysis. Like the fifteen Thematic Mapper landscape variables, percentage breakdowns of bedrock geology within the watershed above a stream site would provide better information on the effect of bedrock geology on acid neutralizing capacity.

In discussions with Dr. Alan Herlihy, he also indicated that stream slope also has a significant impact on acid neutralizing capacity. Several different classifications of slope are available through GIS models, such as slope of the stream at the sampled site, watershed slope from highest point in the watershed to the sample site, and maximum stream slope.

The effect of these variables on acid neutralizing capacity within the MAHA region is unknown. But with a larger pool of information to work with, the probability of

creating a more accurate model for ANC increases. Therefore, future analysis using these expanded and more precise predictors would likely be beneficial.

5. Stratification / Small Area Estimation

Future work should also be done in researching the relationship between geographic features or spacing and ANC. There are several issues that could be addressed in this area. First, the concept of distance needs to be more closely examined. One concern is whether or not Euclidean distance is the best measure for calculating distance for these stream-based data? First, when dealing with coordinate systems, such as Alber's Projection or Latitude and Longitude, distortion will occur when calculating Euclidean distance. Second, when modeling stream data, some form of hydrologic distance between sites may hold more meaning. Further, what is the maximum distance of separation at which we would expect sites to still be correlated, i.e. would it be realistic to expect to see a spatial relationship between two sites that are 60 miles apart? Therefore, this issue and concept of distance between stream sites needs to be more thoroughly examined.

The area over which the analysis is done could also be reexamined. Should a model cover the entire MAHA region, or should the MAHA region be broken down into smaller, more homogeneous units? One possibility is modeling sites separated by Hydrologic Unit Code (HUC) as specified by the United States Geological Survey (USGS). ANC could also be modeled after stratifying the MAHA region by identified ecoregion. However, two problems will arise with either of these stratification procedures. First for the EMAP data used in this study, the number of sampled sites per HUC or ecoregion is extremely small, indicating the necessary implementation of small

area estimation statistical techniques (Ghosh and Rao, 1994). Second, the scale of the stratification for HUCs must be determined. Hydrologic Unit Codes are typically eight digit numbers that operate much like postal zip codes, with each number representing a given area. The first six digits of the HUC describe a given region, and the last two describe a subregion. Ecoregion can also be broken down from general regions to more specific subregions. How large stratification is necessary to create homogeneous strata suitable for analyzing ANC? The strata must be large enough to encompass enough information from sampled sites to create a workable model, but not so large that the strata are no longer homogeneous.

References

Acid Deposition Standard Feasibility Study Executive Summary (2002). 29 October 2002. U.S. Environmental Protection Agency. 19 Nov. 2002 <http://www.epa.gov/airmarkets/articles/depfeas/>

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood Principle", *2nd International symposium on Information Theory*, B.N. Petrov and F. Csaki (eds.), Akademiai Kiado, Budapest, 267-281.

Alber's Equal Area Conic (2002). Universiy of Texas at Austin Maps Library. 19 November 2002. <http://www.lib.utexas.edu/maps/albers_equal_area.jpg>

Brewer, P.F, Sullivan, T.J, Cosby, J., and Munson, R. (2002). "Acid Deposition Effects To Forests and Streams in the Southern Appalachian Mountains". *Southern Appalachian Mountains Initiative.* 19 November 2002 <http://www.saminet.org/reports/AWMAacid_326.htm>

Environmental Monitoring and Assessment (EMAP) Home Page (2002). 7 November 2002. U.S. Environmental Protection Agency (EPA). 19 November 2002 <http://www.epa.gov/docs/emap/>

Furnival, G. and R. Wilson (1974). "Regression by leaps and bounds." *Technometrics*, 16 499-511.

Ghosh, M. and Rao, J.N.K. (1994). "Small Area Estimation: An Appraisal," *Statistical Science*, 9:1, 55--93.

Herlihy, Alan (2002). Personal Interview. 23 September 2002.

Hoeting, J. A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). "Bayesian Model Averaging: A Tutorial (with discussion)," *Statistical Science*, 14 :4, 382--417. Corrected version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Jones, K. Bruce, Riitters, K.H., et.al. (1997). *An Ecological Assessment of the United States Mid-Atlantic Region: A Landscape Atlas*. U.S. Environmental Protection Agency Publication EPA/600/R-97/130.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. 4th Ed. Boston: McGraw-Hill.

Reich, R. and Davis, R. (2000). *Quantitative Spatial Analysis*. Coursepack for ST523. Fort Collins, CO: Colorado State University.

Ripley, B. (1981). *Spatial Statistics*. New York: Wiley.

Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: Wiley.

STARMAP: Space-Time Aquatic Resources Modeling and Analysis Program Home Page (2002). Colorado State University 25 November 2002. <http://www.stat.colostate.edu/~nsu/starmap/>

Statlib: Data, Software, and News from the Statistics Community (2002). The Department of Statistics at Carnegie Mellon Univeristy. <http://lib.stat.cmu.edu/>

Stoddard, J. L., Urquhart, N.S., Newell, A. D., and Kugler, D. (1996). "The Temporally Integrated Monitoring of Ecosystems (TIME) project design – 2. Detection of regional acidification trends". *Water Resources Research,* 32 (8): 2529-2538.

Stoddard, J.L., Kahl, J.S., Deviney, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, J.R. Webb, J.R., and Webster, K.E. (2003). *Response of Surface Water Chemistry to the Clean Air Act Amendments of 1990.* EPA/620/R-02/004. U.S. Environmental Protection Agency, Washington, DC. <http://www.epa.gov/ord/htm/CAAA-2002-report-2col-rev-4.pdf>

Strahler, A.N. (1964). Quantitative geomorphology of drainage basins and channel networks, section 4-II, *Handbook of Applied Hydrology*, V.T. Chow (ed.), McGraw-Hill, New York, 4-39.

Sullivan, T. J., et al. (2002). "Spatial Distribution of Acid-Sensitive and Acid-Impacted Streams in Relation to Watershed Features in the Southern Appalachian Mountains". *Water Resources Research*, to appear.

Theobald, Dave (2002). Personal Interview. October 2002.

Thompson, Sandra E. (2001). *Bayesian Model Averaging and Spatial Prediction*. PhD thesis, Colorado State University.

Vogelmann, J. (2002). Land cover data layer for EPA Region III. 19 November 2002 <http://www.webmapping.org/data/ahr_utm_27.htm>

Webster, R. and M.A. Oliver. (2001). *Geostatistics for Environmental Scientists*. West Sussex, England: Wiley.

.

Appendices

A.	Thematic Mapper Classifications	57
B.	Procedure Used in Acquiring Final Data Set – DARM	58
C.	Estimated Variograms Using Different Directions and Spans	59
D.	Selected Splus Code Used in Analysis	75

Appendix A

Thematic Mapper Classifications

Water: all area of open water, generally with less than 30% cover of vegetation/land cover.

Urban – low density: approximately 50-80% constructed material; approximately 20-50% vegetation cover; high percentage of residential development typifies this class.

Urban – high density: 20% or less vegetation, high percentage (80-100%) building materials; typically low percentage of residential development in this class.

Pasture: areas characterized by high percentages of grasses and other herbaceous vegetation that is regularly mowed for hay and/or grazed by livestock; predominantly hay fields and pastures, but also currently includes golf courses and city parks.

Row Crops: areas regularly tilled and planted, often on an annual or biennial basis; corn, cotton, sorghum, vegetable crops.

Probable Row Crops: sometimes confused with other areas, such as grasslands that were not green during times of spring data acquisitions.

Evergreen Forest: of trees present, 70% or higher conifers.

Deciduous Forest: of trees present, 70% or higher deciduous tree species.

Mixed Forest: both conifers and deciduous tree species present, with neither particularly dominant.

Woody Wetlands: wetlands with substantial amount of woody vegetation present, either trees or shrubs.

Emergent Wetlands: wetlands without a substantial amount of woody vegetation present, usually with substantial amounts of herbaceous vegetation.

Quarry: all quarry areas, including sand and gravel operations, except some spectrally dark coal areas in northern Pennsylvania.

Transitional: areas likely to change to other land cover categories, such as clear cuts.

Appendix B

Procedure Used in Acquiring Final Data Set - DARM

Data Set #1 (896 observations)

- This was the data set sent to me by Dave Theobold with data for all sites within the MAHA region
- Data set contained the following:
    - ANC values from EMAP study
    - Elevation, geology (in form of Class and Type), Strahler Stream Order, and Albers projection coordinates (X and Y) from GIS modeling
- Changed ANC to double from character
- Eliminated all sites w/ no ANC values (i.e. no response)

Data Set #2 (699 observations)

- There were 699 observations from which ANC was found
- Eliminated five identified outliers

Data Set #3 (694 observations)

- Eliminated all observations where geologic class was not available

Data Set #4 (345 observations)

- Merged Data Set #3 and Thematic Mapper Data Set (TM.IMAGE)

Data Set #5 (292 observations)

- Eliminated all but the first visit to each site

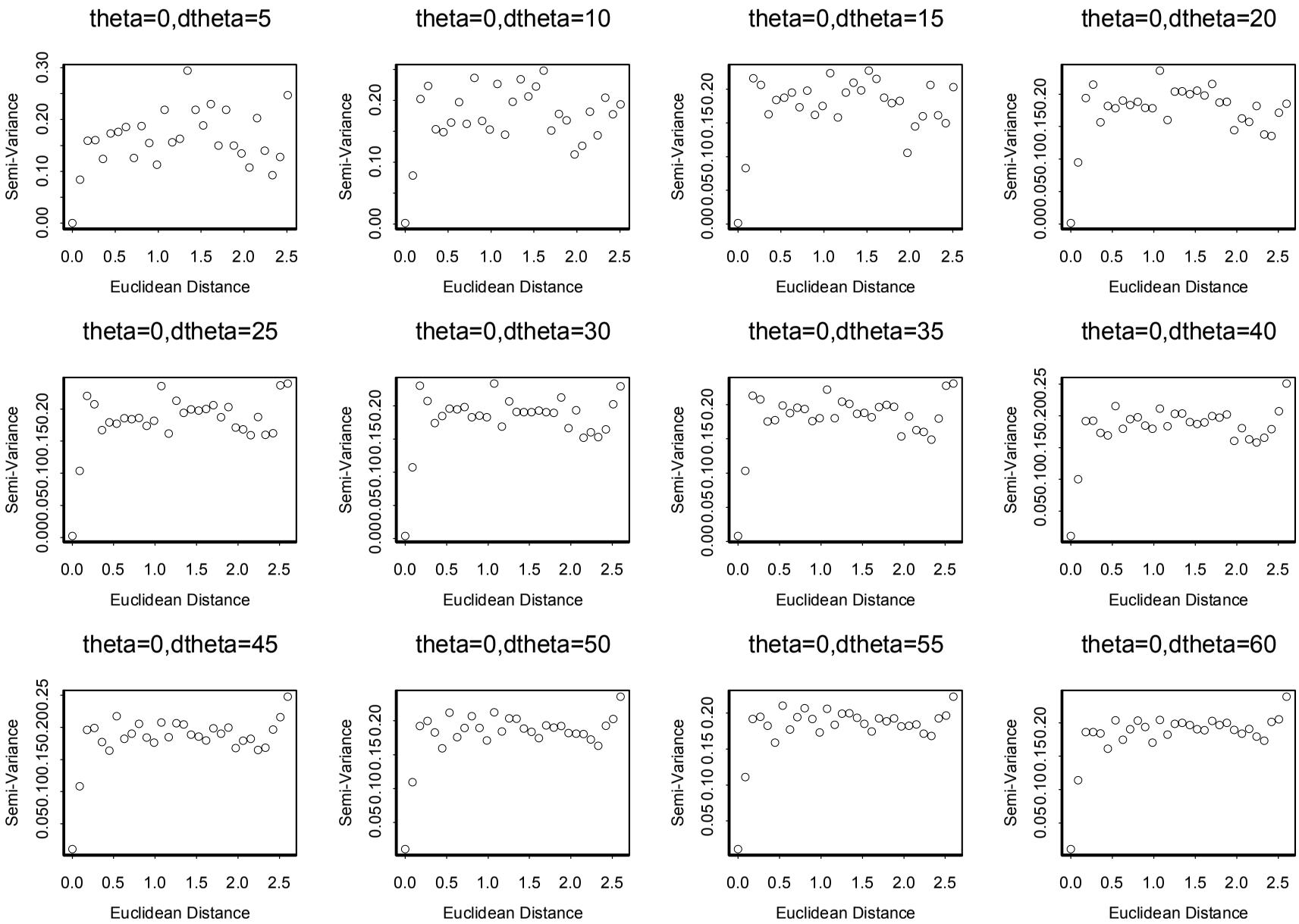
Data Set #6 (242 observations)

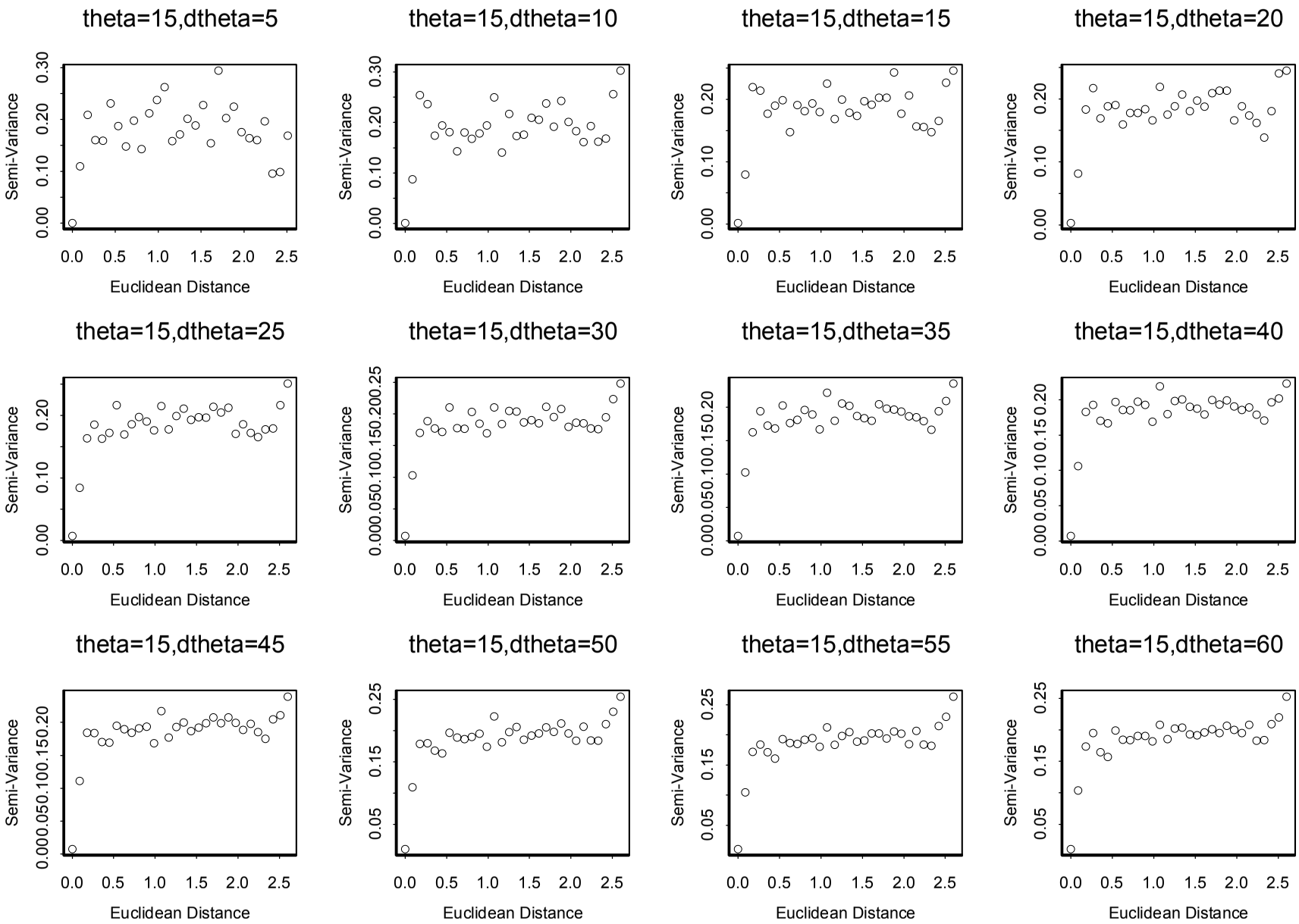
- Eliminated all sites belonging to "Unclassified" geologic class

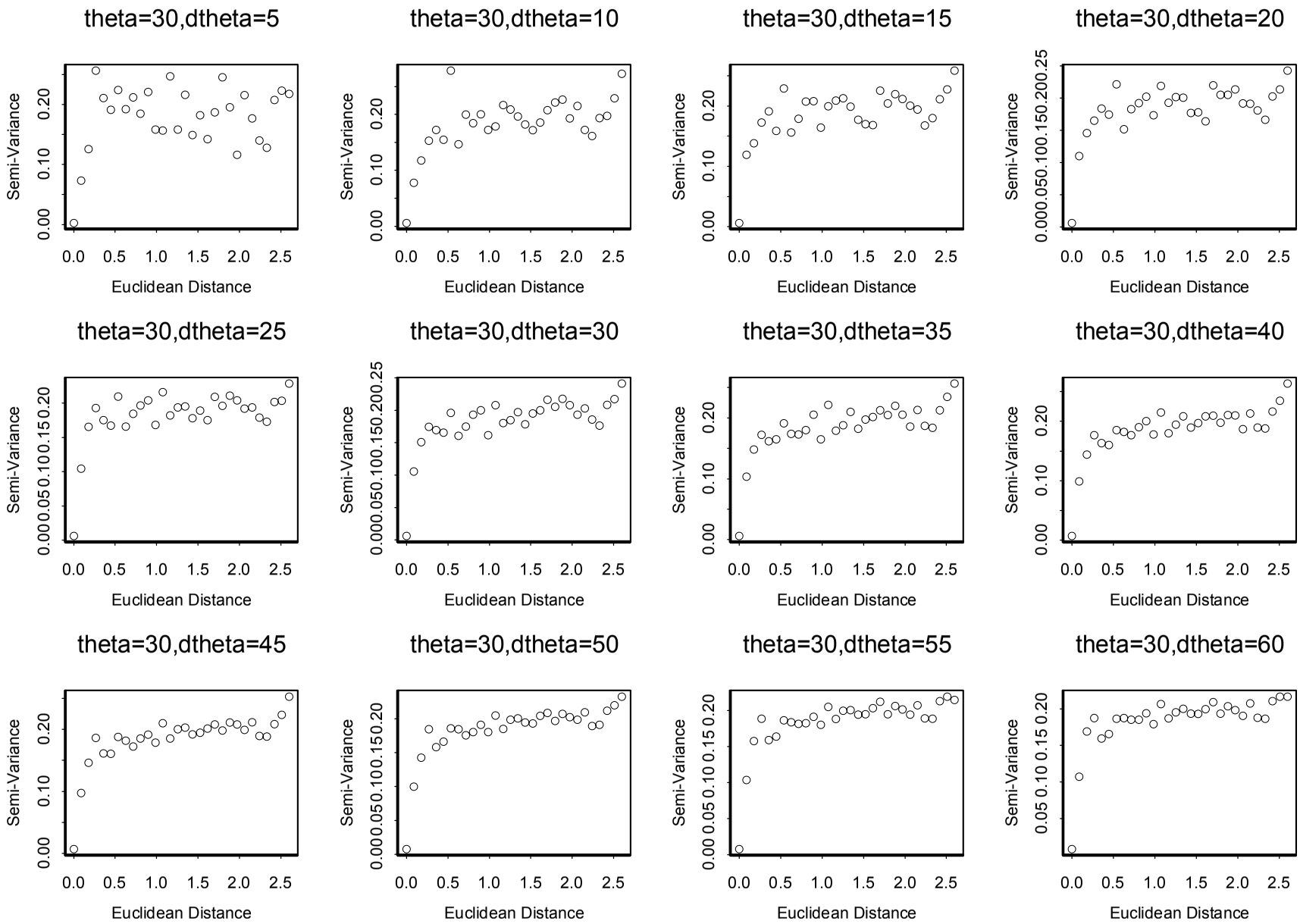
DARM Set (238 observations)

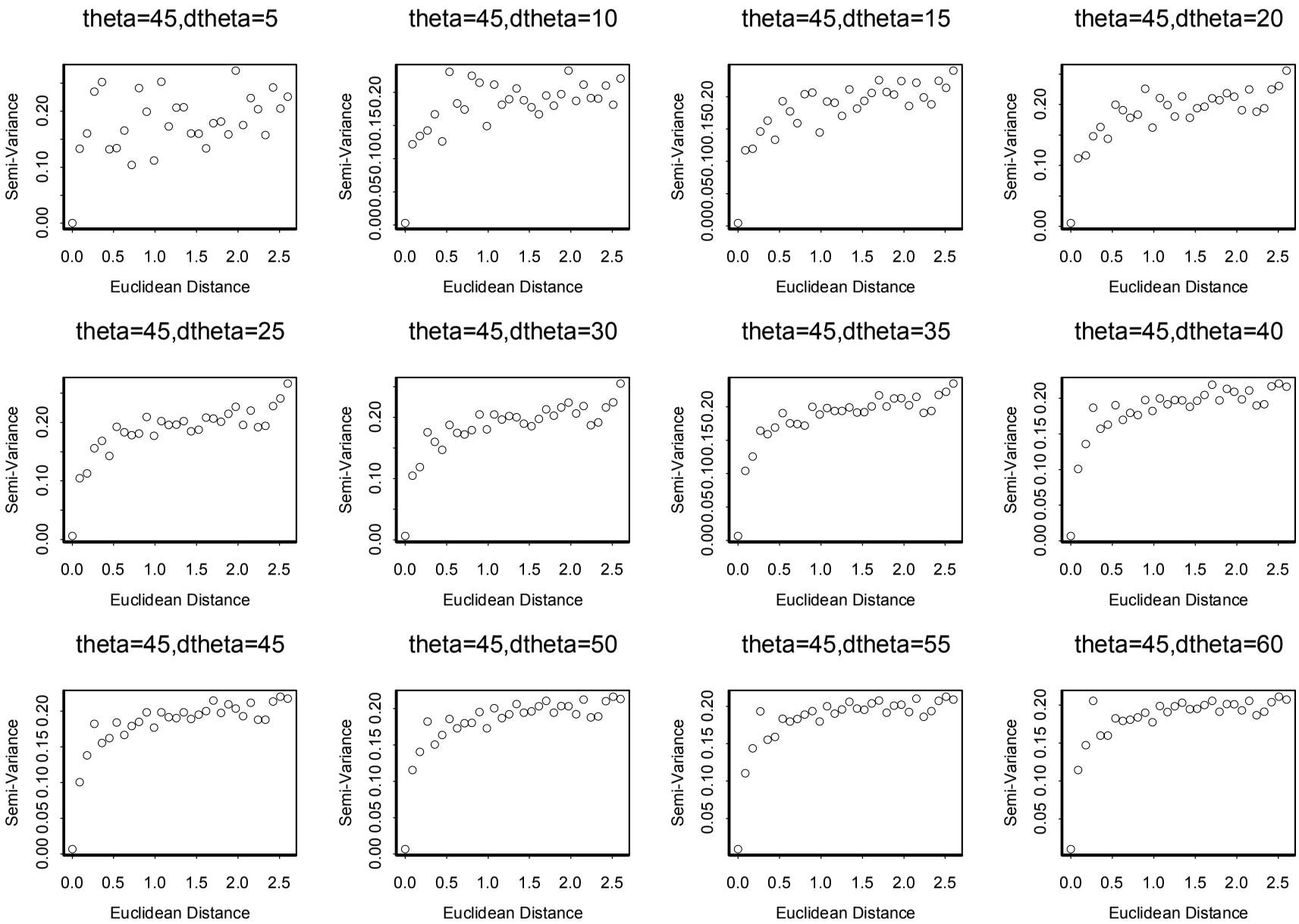
Appendix C

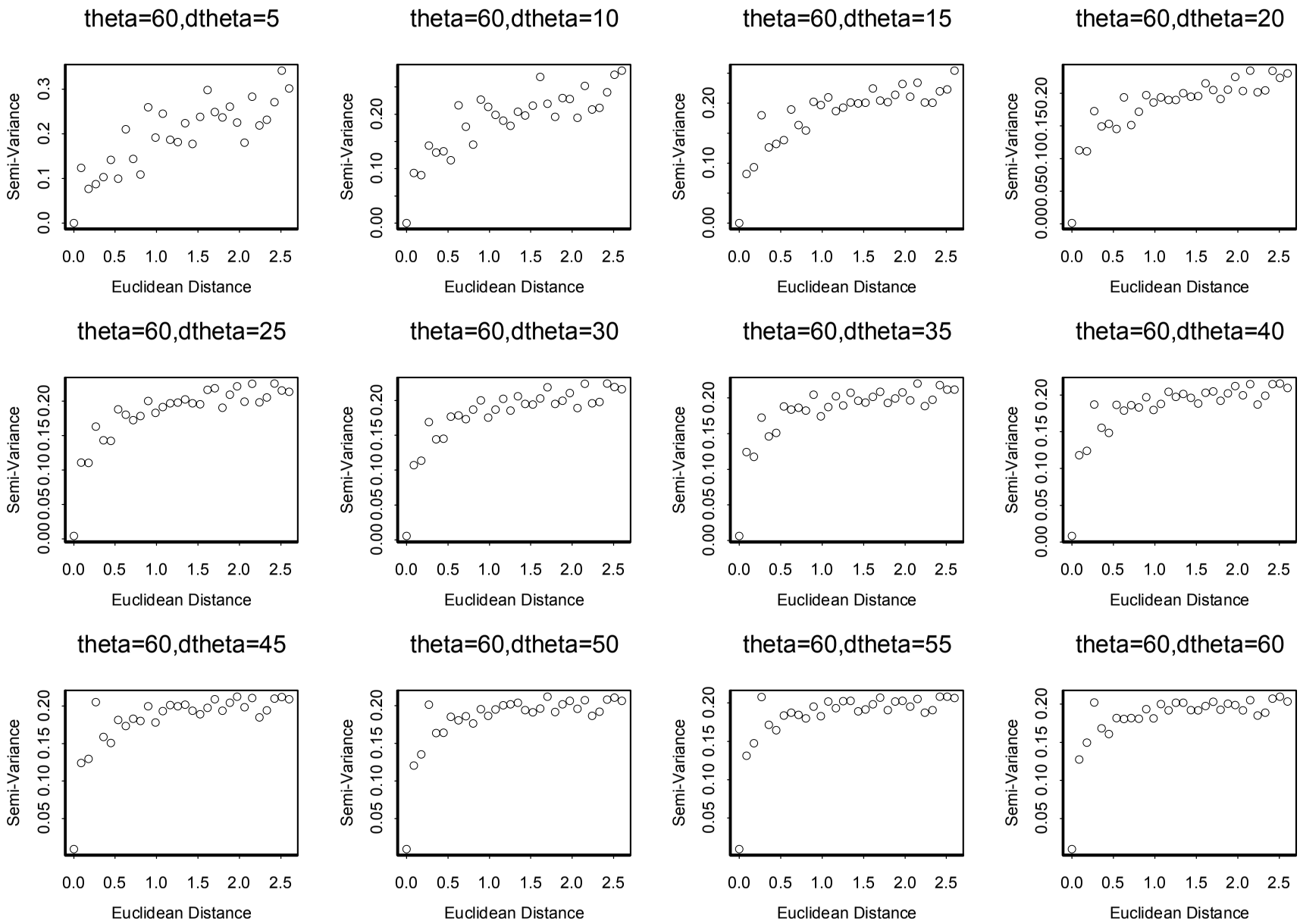
Estimated Variograms Using
Different Directions (theta) and Spans (dtheta)

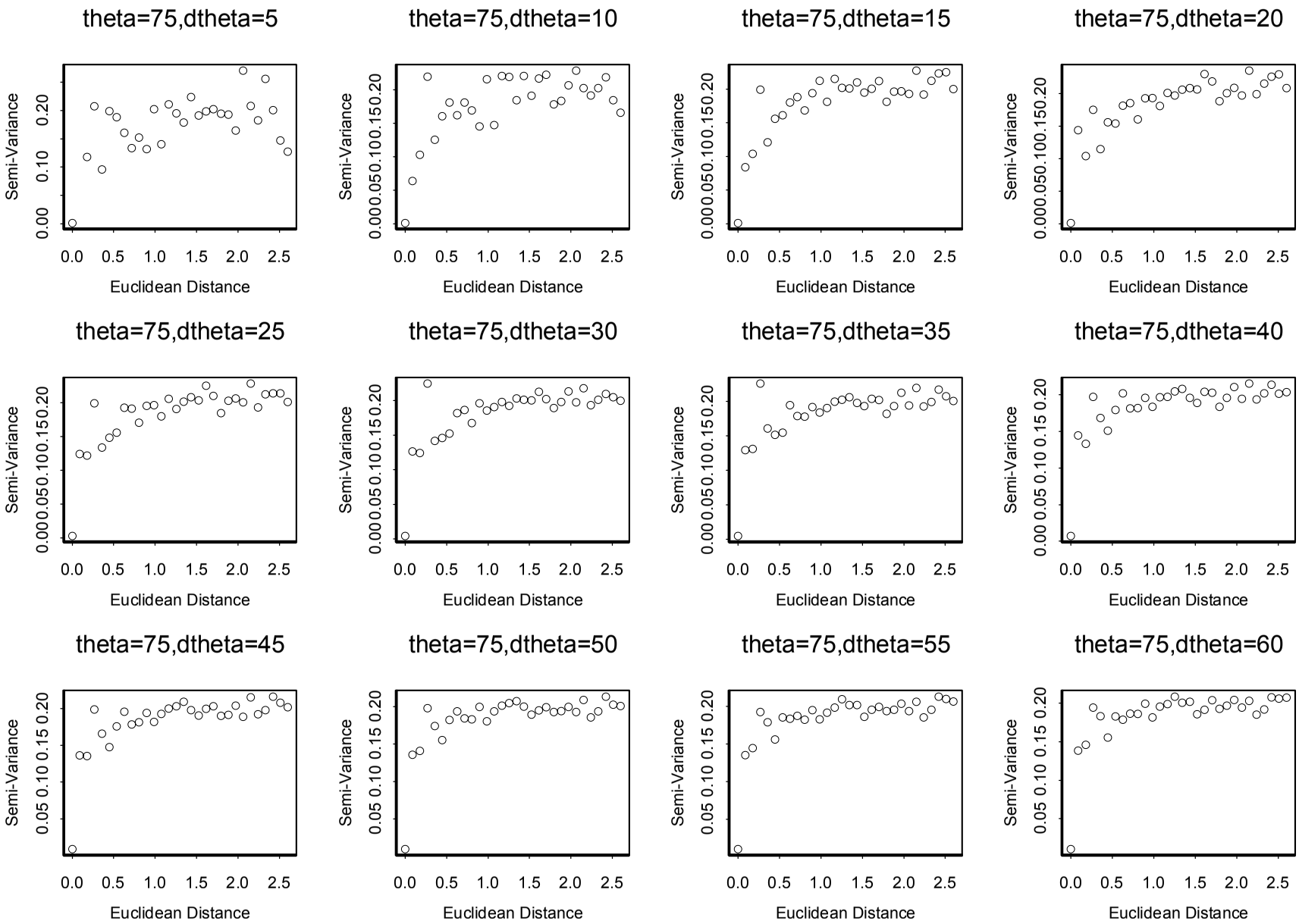

theta=0,dtheta=5
theta=0,dtheta=10
theta=0,dtheta=15
theta=0,dtheta=20
theta=0,dtheta=25
theta=0,dtheta=30
theta=0,dtheta=35
theta=0,dtheta=40
theta=0,dtheta=45
theta=0,dtheta=50
theta=0,dtheta=55
theta=0,dtheta=60
Semi-Variance
Euclidean Distance

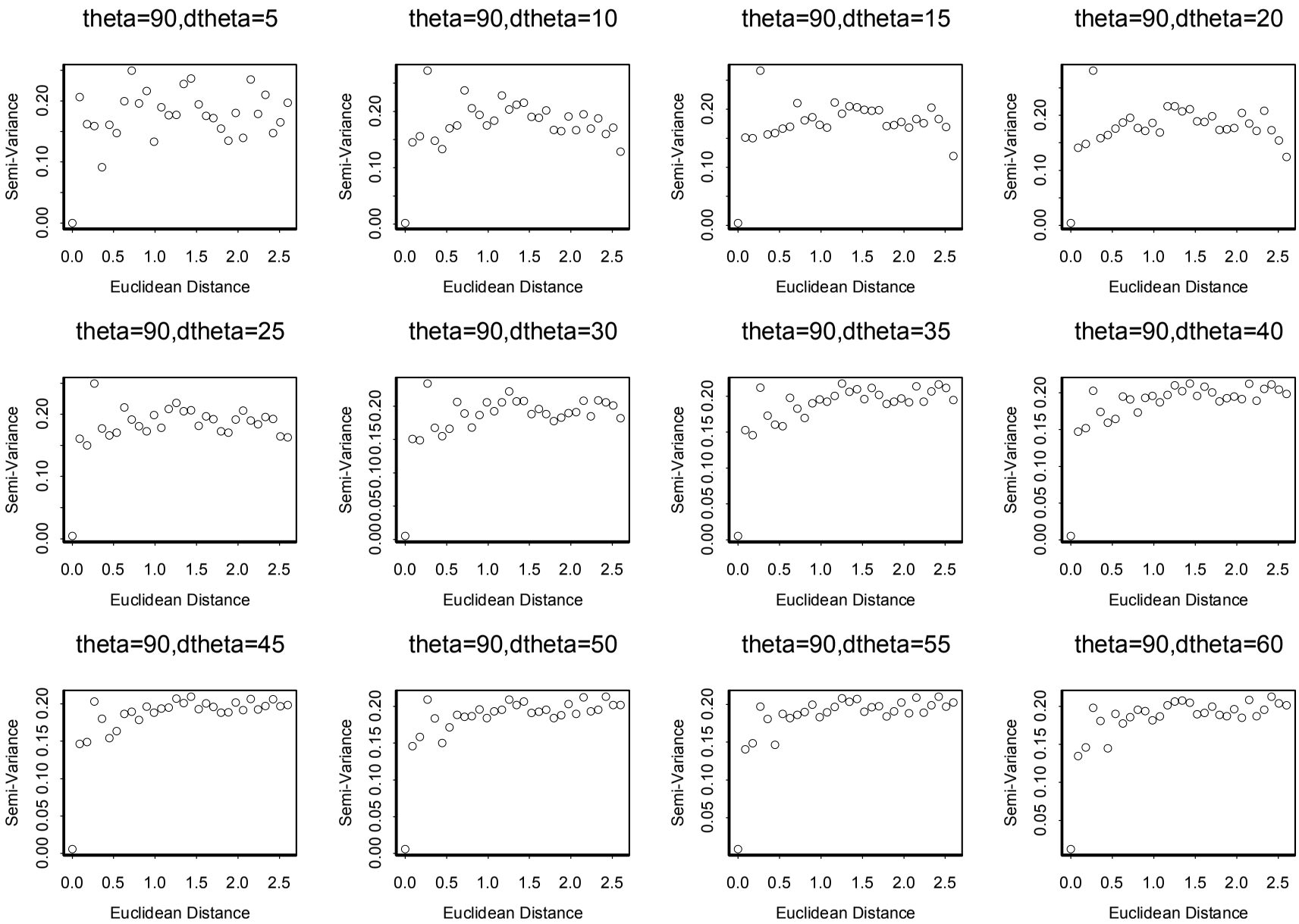

theta=15,dtheta=5
theta=15,dtheta=10
theta=15,dtheta=15
theta=15,dtheta=20
theta=15,dtheta=25
theta=15,dtheta=30
theta=15,dtheta=35
theta=15,dtheta=40
theta=15,dtheta=45
theta=15,dtheta=50
theta=15,dtheta=55
theta=15,dtheta=60
Semi-Variance
Euclidean Distance

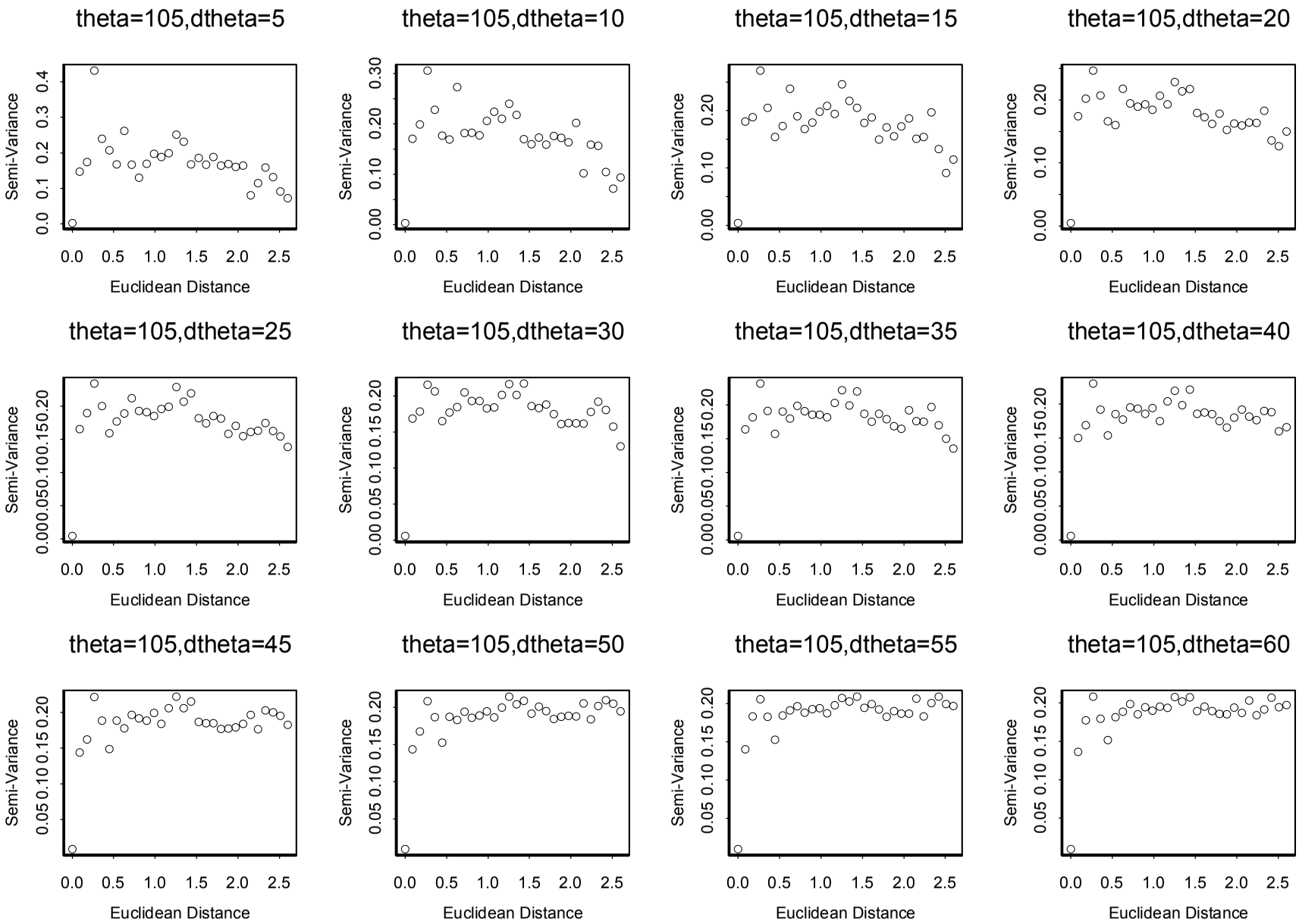

theta=30,dtheta=5
theta=30,dtheta=10
theta=30,dtheta=15
theta=30,dtheta=20
theta=30,dtheta=25
theta=30,dtheta=30
theta=30,dtheta=35
theta=30,dtheta=40
theta=30,dtheta=45
theta=30,dtheta=50
theta=30,dtheta=55
theta=30,dtheta=60
Semi-Variance
Euclidean Distance

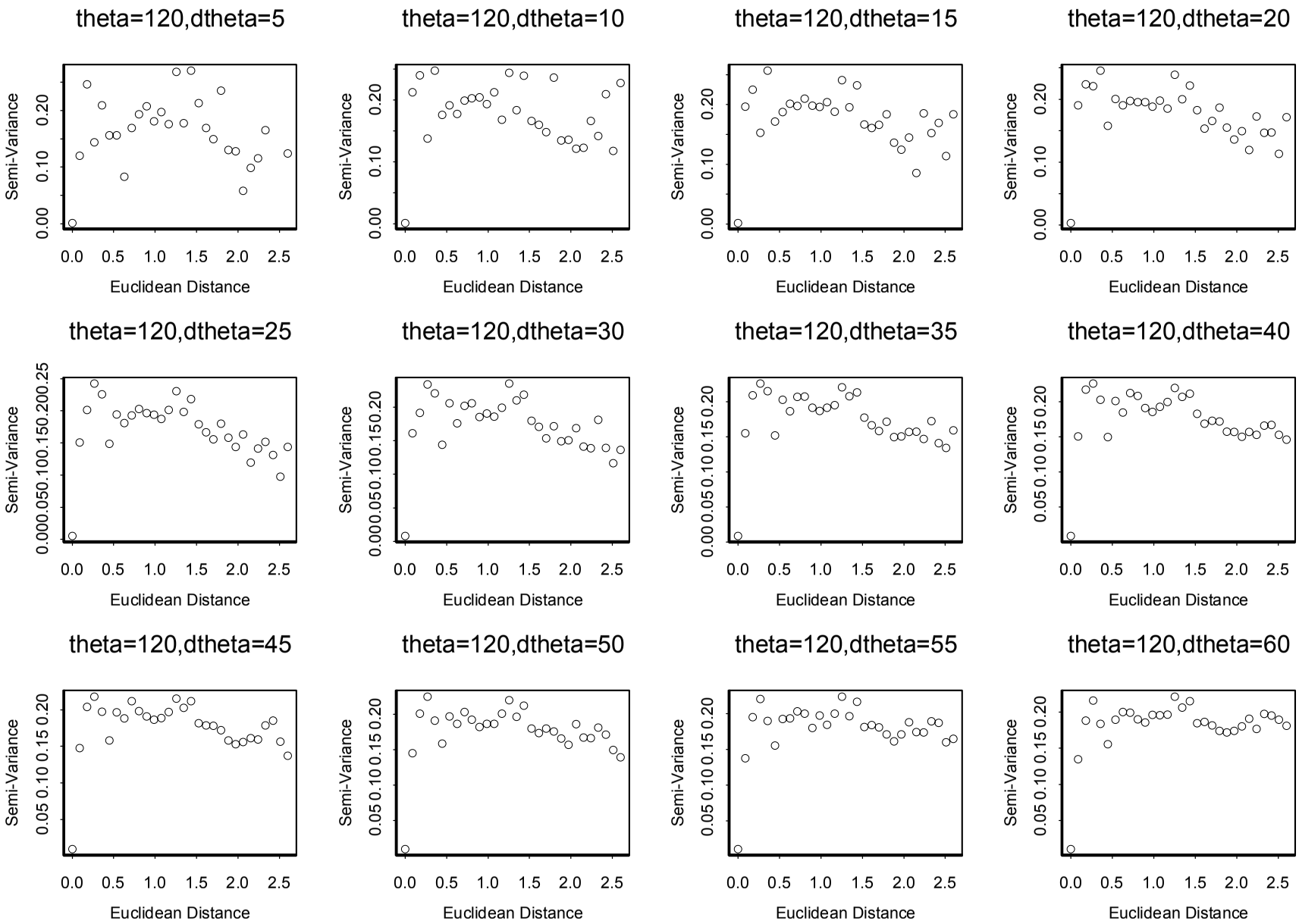

theta=45,dtheta=5
theta=45,dtheta=10
theta=45,dtheta=15
theta=45,dtheta=20
theta=45,dtheta=25
theta=45,dtheta=30
theta=45,dtheta=35
theta=45,dtheta=40
theta=45,dtheta=45
theta=45,dtheta=50
theta=45,dtheta=55
theta=45,dtheta=60
Semi-Variance
Euclidean Distance
0.0 0.5 1.0 1.5 2.0 2.5

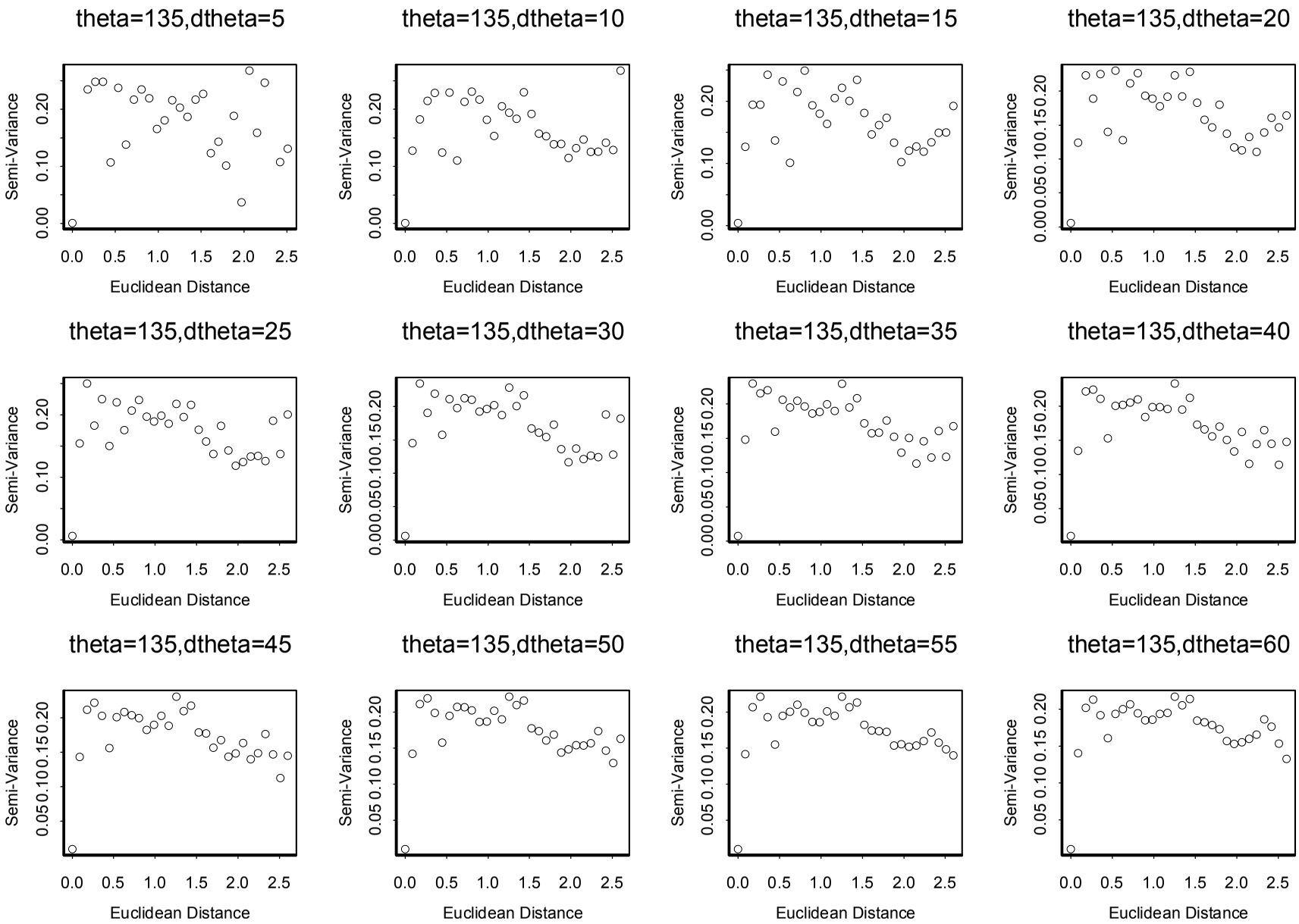

theta=60,dtheta=5
theta=60,dtheta=10
theta=60,dtheta=15
theta=60,dtheta=20
theta=60,dtheta=25
theta=60,dtheta=30
theta=60,dtheta=35
theta=60,dtheta=40
theta=60,dtheta=45
theta=60,dtheta=50
theta=60,dtheta=55
theta=60,dtheta=60
Semi-Variance
Euclidean Distance

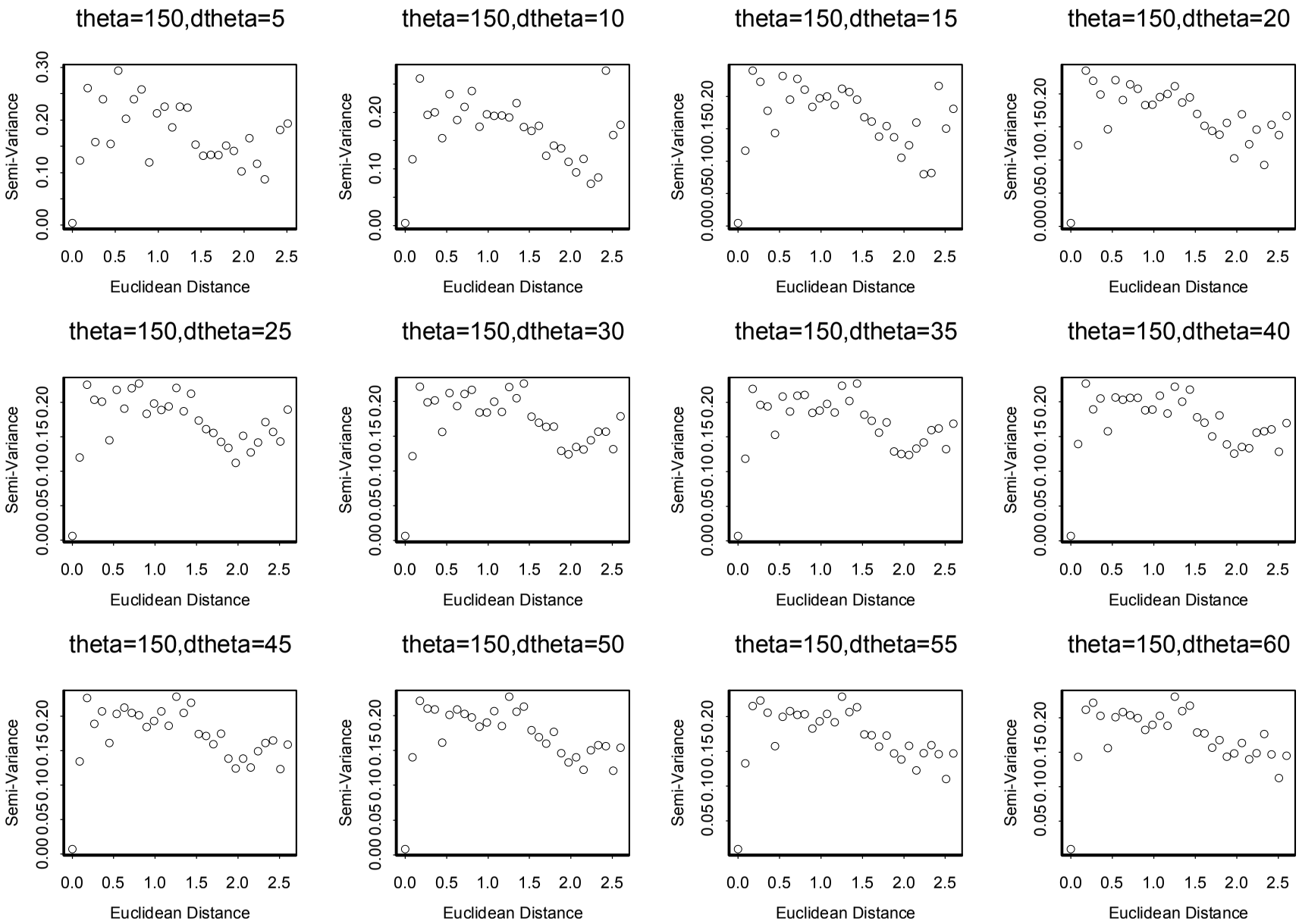

theta=75,dtheta=5
theta=75,dtheta=10
theta=75,dtheta=15
theta=75,dtheta=20
theta=75,dtheta=25
theta=75,dtheta=30
theta=75,dtheta=35
theta=75,dtheta=40
theta=75,dtheta=45
theta=75,dtheta=50
theta=75,dtheta=55
theta=75,dtheta=60
Semi-Variance
Euclidean Distance
0.0 0.5 1.0 1.5 2.0 2.5

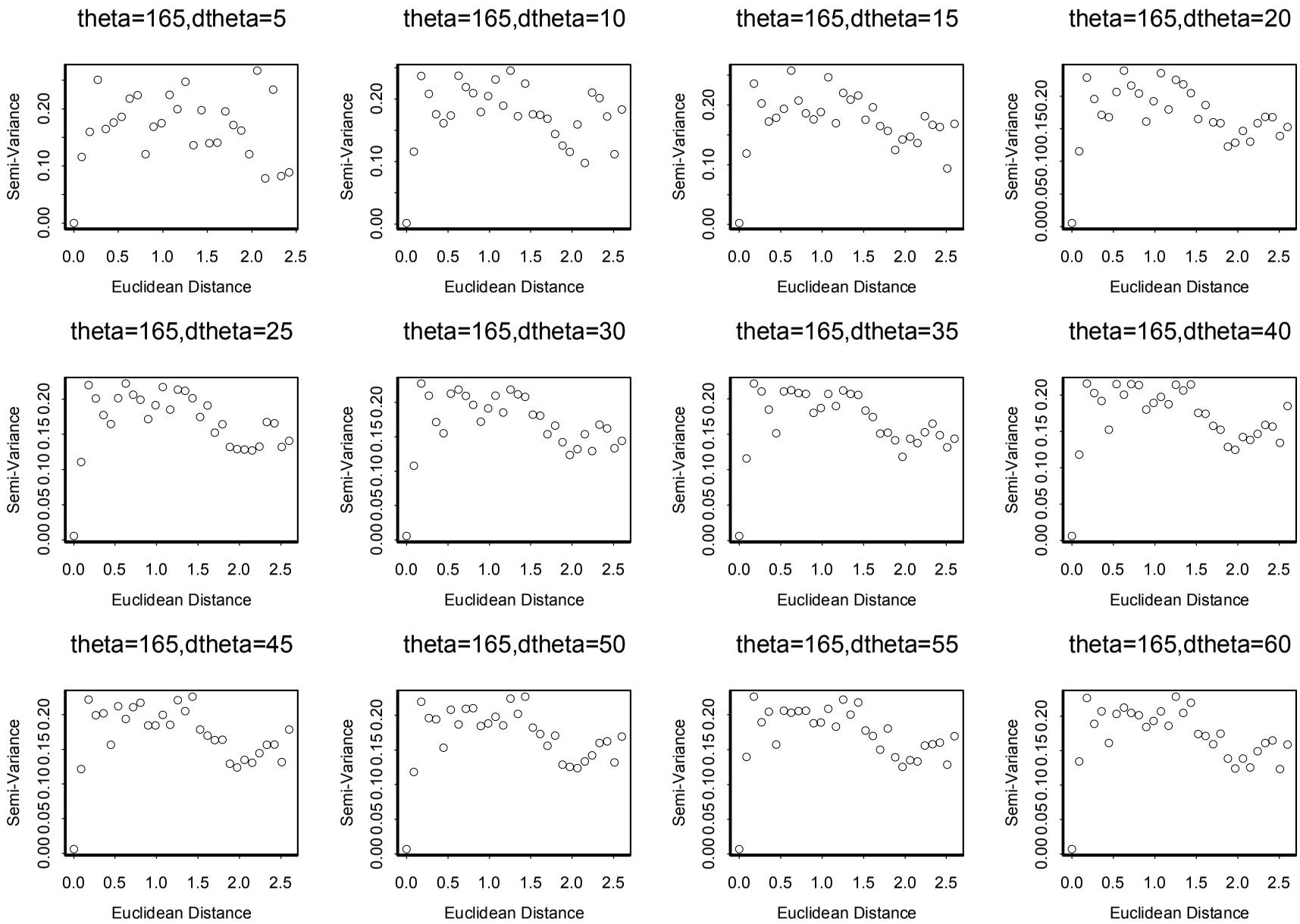

theta=90,dtheta=5
theta=90,dtheta=10
theta=90,dtheta=15
theta=90,dtheta=20
theta=90,dtheta=25
theta=90,dtheta=30
theta=90,dtheta=35
theta=90,dtheta=40
theta=90,dtheta=45
theta=90,dtheta=50
theta=90,dtheta=55
theta=90,dtheta=60
Semi-Variance
Euclidean Distance
0.0 0.5 1.0 1.5 2.0 2.5

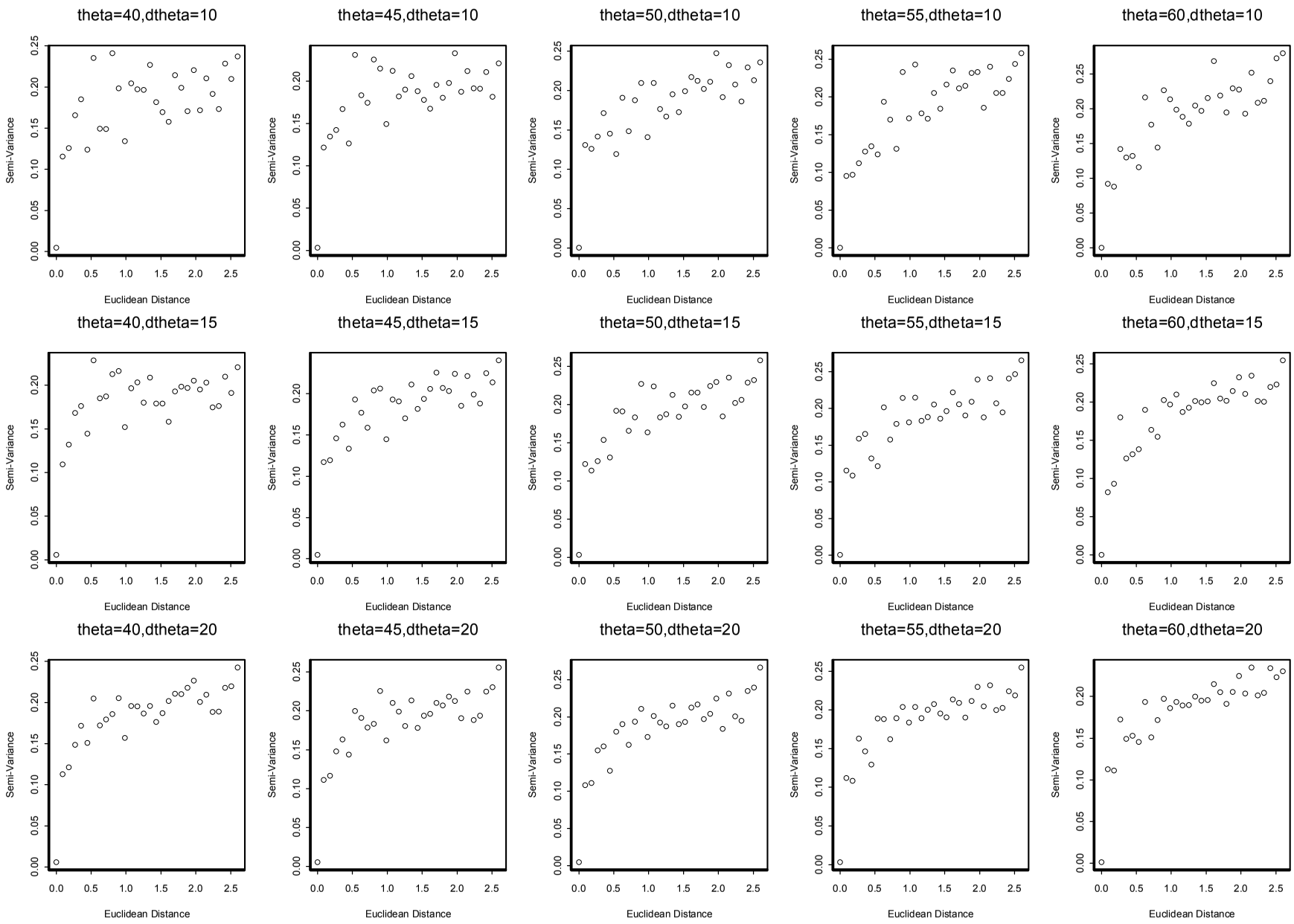

theta=105,dtheta=5
theta=105,dtheta=10
theta=105,dtheta=15
theta=105,dtheta=20
theta=105,dtheta=25
theta=105,dtheta=30
theta=105,dtheta=35
theta=105,dtheta=40
theta=105,dtheta=45
theta=105,dtheta=50
theta=105,dtheta=55
theta=105,dtheta=60
Semi-Variance
Euclidean Distance

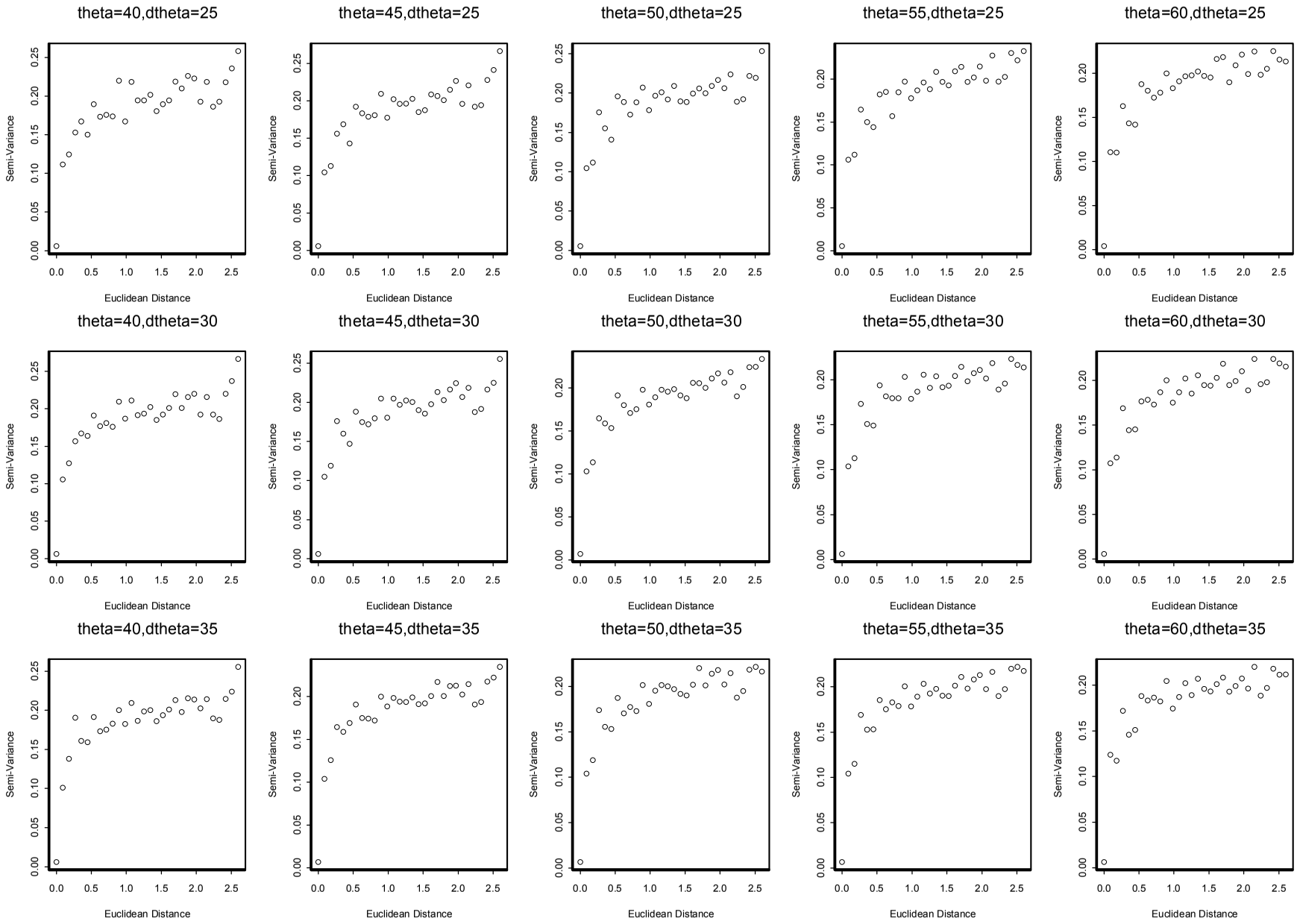

theta=120,dtheta=5
theta=120,dtheta=10
theta=120,dtheta=15
theta=120,dtheta=20
theta=120,dtheta=25
theta=120,dtheta=30
theta=120,dtheta=35
theta=120,dtheta=40
theta=120,dtheta=45
theta=120,dtheta=50
theta=120,dtheta=55
theta=120,dtheta=60
Semi-Variance
Euclidean Distance

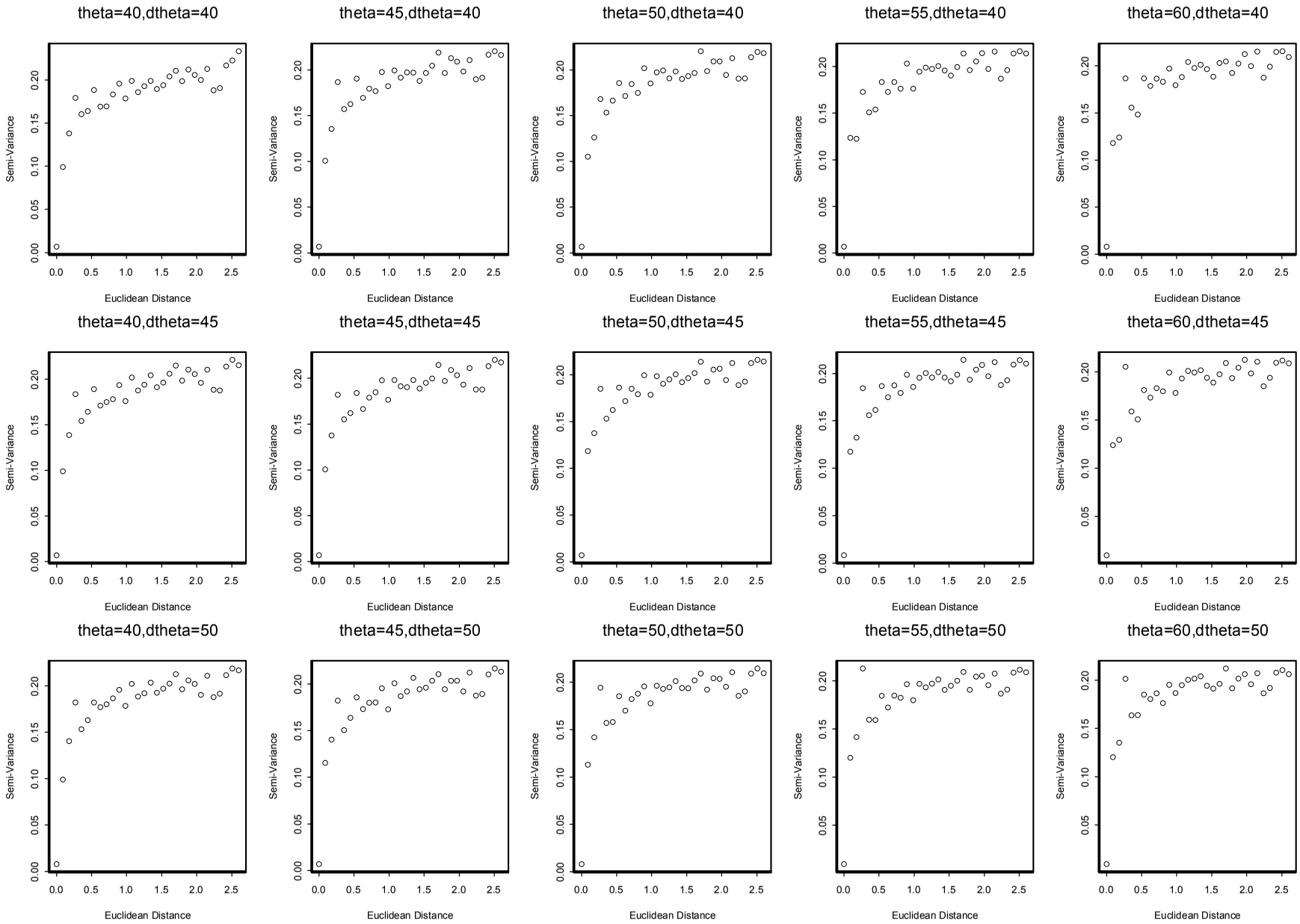

theta=135,dtheta=5
theta=135,dtheta=10
theta=135,dtheta=15
theta=135,dtheta=20
theta=135,dtheta=25
theta=135,dtheta=30
theta=135,dtheta=35
theta=135,dtheta=40
theta=135,dtheta=45
theta=135,dtheta=50
theta=135,dtheta=55
theta=135,dtheta=60
Semi-Variance
Euclidean Distance


theta=150,dtheta=5
theta=150,dtheta=10
theta=150,dtheta=15
theta=150,dtheta=20
theta=150,dtheta=25
theta=150,dtheta=30
theta=150,dtheta=35
theta=150,dtheta=40
theta=150,dtheta=45
theta=150,dtheta=50
theta=150,dtheta=55
theta=150,dtheta=60
Semi-Variance
Euclidean Distance


theta=165,dtheta=5
theta=165,dtheta=10
theta=165,dtheta=15
theta=165,dtheta=20
theta=165,dtheta=25
theta=165,dtheta=30
theta=165,dtheta=35
theta=165,dtheta=40
theta=165,dtheta=45
theta=165,dtheta=50
theta=165,dtheta=55
theta=165,dtheta=60
Semi-Variance
Euclidean Distance


theta=40,dtheta=10
theta=45,dtheta=10
theta=50,dtheta=10
theta=55,dtheta=10
theta=60,dtheta=10
theta=40,dtheta=15
theta=45,dtheta=15
theta=50,dtheta=15
theta=55,dtheta=15
theta=60,dtheta=15
theta=40,dtheta=20
theta=45,dtheta=20
theta=50,dtheta=20
theta=55,dtheta=20
theta=60,dtheta=20
Semi-Variance
Euclidean Distance
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.5
1.0
1.5
2.0
2.5


theta=40,dtheta=25
theta=45,dtheta=25
theta=50,dtheta=25
theta=55,dtheta=25
theta=60,dtheta=25
theta=40,dtheta=30
theta=45,dtheta=30
theta=50,dtheta=30
theta=55,dtheta=30
theta=60,dtheta=30
theta=40,dtheta=35
theta=45,dtheta=35
theta=50,dtheta=35
theta=55,dtheta=35
theta=60,dtheta=35
Semi-Variance
Euclidean Distance


theta=40,dtheta=40
theta=45,dtheta=40
theta=50,dtheta=40
theta=55,dtheta=40
theta=60,dtheta=40
theta=40,dtheta=45
theta=45,dtheta=45
theta=50,dtheta=45
theta=55,dtheta=45
theta=60,dtheta=45
theta=40,dtheta=50
theta=45,dtheta=50
theta=50,dtheta=50
theta=55,dtheta=50
theta=60,dtheta=50
Semi-Variance
Euclidean Distance

Appendix D

Selected Splus Code Used in Analysis

Listing of Coded Predictors and Response

```
pasture_log(darm$pasture+.001)
probable_darm$probable
row.crops_darm$row.crops
decid_darm$decid
mixed_log(darm$mixed+.001)
evergreen_log(darm$evergreen+.001)
urb.hi_log(darm$urb.hi+.001)
urb.low_log(darm$urb.low+.001)
emergent_log(darm$emergent+.001)
woody_log(darm$woody+.001)
quarry_log(darm$quarry+.001)
transition_log(darm$transition+.001)
water_log(darm$water+.001)
elev_darm$elev
carbon_darm$carbon
felsic_darm$felsic
mafic_darm$mafic
arg_darm$arg
silic_darm$silic
ord2_darm$ord2
ord3_darm$ord3
X_scale(darm$X,scale=F)/100000
Y_scale(darm$Y,scale=F)/100000
ANC_log(darm$ANC+500)
lat_darm$LAT.DD
lon_darm$LON.DD1
```

Final – ANC Model

```
model_lm(ANC~probable+pasture+urb.hi+emergent+woody+quarry+carbon+felsic+elev)
```

Code Used in Model Selection

```
pred_cbind(pasture,probable,row.crops,decid,mixed,evergreen,urb.hi,urb.low,
   emergent,woody, transition, quarry,water,elev,carbon,felsic,arg,silic,ord2,
   ord3)
pred1_cbind(pasture,probable,urb.hi,emergent,woody,quarry,elev,carbon,
   felsic,X,Y,X^2,Y^2,X*Y)
efroym_stepwise(pred,ANC)
efroym
par(mfrow=c(2,2))
r2_leaps(pred,ANC,method="r2",nbest=4)
r2
plot(r2$size,r2$r2/100)
title("Leaps: R-squared")
mallow_leaps(pred,ANC,method="Cp",nbest=4)
mallow
plot(mallow$size,mallow$Cp)
title("Leaps: Mallow's Cp")
abline(0,1)
```

Code for VIF (from Dr. Jennifer Hoeting) and Multicollinearity Investigation

```
VIF<-function(X) {
#Computes Variance Inflation Factors for a Predictor Matrix
#   See page 386 of NKNW for more on computations
#INPUTS:
#X is a matrix (or data frame) of the predictors (no column of ones).

   cat("REMINDER: Your input matrix should not include the response\n")
   a<-1/sqrt(dim(X)[1]-1)*scale(X)
   b<-cbind(diag(solve(t(a)%*%a)))
   dimnames(b)<-list(dimnames(X)[[2]],"VIF")
   return(b)
}

pred1_cbind(probable,pasture,urb.hi,urb.low,elev,quarry,emergent,woody,
       carbon,felsic)
VIF(pred1)
```

Code for Variogram estimation (Original code found in Spatial Library created by Dr. Robin Reich and Dr. Richard Davis)

```
variogrm_function(x, y, z, nint, iso = T, theta = 0, dtheta = 0, dmax = 0)
{
       n <- length(x)
       rx <- range(x)
       ry <- range(y)
       .Fortran("frset",
              as.single(rx[1]),
              as.single(rx[2]),
              as.single(ry[1]),
              as.single(ry[2]))
       z <- .Fortran("variogram",
              xp = single(nint),
              yp = single(nint),
              nint = as.integer(nint),
              as.double(x),
              as.double(y),
#               if(krig$np == 0) as.double(krig$z) else as.double(krig$w
z),
              as.double(z),
              as.integer(length(x)),
              iso = as.logical(iso),
              theta = as.single(theta),
              dtheta = as.single(dtheta),
              nsv = integer(nint),
              dist = single((n * (n + 1))/2),
              indi = logical((n * (n + 1))/2),
              dmax = as.single(dmax))
       ni <- z$nsv[1:z$nint]
       xp <- z$xp[1:z$nint]
       yp <- z$yp[1:z$nint]
       plot(xp, yp, type = "p")
       invisible(list(x = xp, y = yp, ni = ni, type = "var"))
}
```

Adjusted Code for Variogram Model Fitting (Original code found in Spatial Library created by Dr. Robin Reich and Dr. Richard Davis)

```
fitvar1_function(dt, a, b, cc, model, wt = F)
{
   x <- dt$x[-1]
   y <- dt$y[-1]
   ni <- dt$ni[-1]
   if(wt == F)
      ni[] <- 1
   k <- 3
   prm <- cbind(a, b, cc)
   if(dt$type == "var") {
      lb <- cbind(0, 0, 0)
   }
   else {
      lb <- cbind( - Inf,  - Inf, 0)
   }
   if(model == "sph") {
      mdl <- sph
   }
   else if(model == "exp") {
      mdl <- fexp
   }
   else if(model == "gau") {
      mdl <- gau
   }
   else {
      mdl <- lin
      prm <- cbind(a, b)
      lb <- cbind( - Inf,  - Inf)
      k <- 2
   }
   res <- nlminb(start = prm, obj = mdl, x = x, y = y, ni = ni, lower = lb)
   parm <- res$parameters
   alpha <- parm[1]/(parm[1] + parm[2])
   n <- length(x)
   ndf <- n - k
   ssmle <- res$objective/n
   r2 <- 1 - res$objective/(var(y) * (n - 1))
   like <-  - n/2 * log(ssmle * 2 * pi) - 0.5 * n
   cat(" Least Squares Estimate \n")
   cat("\n Nugget = ", round(parm[1], 4))
   if(model == "lin") {
      cat("\n Slope  = ", round(parm[2], 6))
   }
   else {
      cat("\n Sill   = ", round((parm[1] + parm[2]), 6))
      cat("\n Range  = ", round(parm[3], 6))
      cat("\n alpha  = ", round(alpha, 6))
      cat("\n s.e.   = ", round(sqrt(parm[1] + parm[2]), 6),
         "\n")
   }
   cat("\n Log(like)   = ", round(like, 4))
   aic <- -2 * like + 2 * (n - ndf)
   aicc <- -2 * like + (2 * (n - ndf) * n)/(n - (n - ndf) - 1)
   sc <- -2 * like + (n - ndf) * log(n)
   cat("\n AIC         = ", round(aic, 4))
   cat("\n AICC        = ", round(aicc, 4))
   cat("\n Schwartz    = ", round(sc, 4), "\n")
```

```
   invisible(list(nugget = parm[1], sill = (parm[1] + parm[2]),
      range = parm[3], alpha = alpha, se = sqrt(parm[1] +
      parm[2]), slope = parm[2], model = model))
}
```

Code For Creating Several of the Presented Figures

(Several of the figures required adjustment using the Splus PC version graphical user interface)

July4 = DARM

July5 consists of DARM sites with positive ANC

July5neg consists of DARM sites with negative ANC

```
hist(site.geo.elev$ANC,nclass=35, ylim=c(0,250),xlim=c(-2200,6000),
   xlab="Acid Neutralizing Capacity",ylab="# of observations")
title("Figure 2:Histogram of Acid Neutralizing Capacity
From Full MAHA Data", cex=1)

arch_split(july4$ANC,july4$class)
what_boxplot(arch$Argillace,arch$Carbonate,arch$Felsic,arch$Mafic,
   arch$Siliceous,
   names=c("Argillace","Carbonate","Felsic","Mafic","Siliceous"))
   title("Figure 2: Boxplot of ANC by Bedrock Geologic Class")

plot(probable,pasture, xlab="% Probable Row Crops",ylab="% Pasture",axes=F)
   pas_pretty(range(exp(pasture)-.001))
   axis(side=2,at=log(pas+.001), lab=pas, srt=90,cex=0.6)
   pro_pretty(range(probable))
   axis(side=1,at=pro, lab=pro, srt=0,cex=0.6)
   title("Figure 10: Relationship Between Probable Row Crops
and Transformed Pasture Percentages",cex=1)

plot(mixed, decid, xlab="% Mixed Forest",ylab="% Deciduous Forest",axes=F)
   dec_pretty(range(decid))
   axis(side=2,at=dec, lab=dec, srt=90,cex=0.6)
   mix_pretty(range(exp(mixed)-.001))
   axis(side=1,at=log(mix+.001), lab=mix, srt=0,cex=0.6)
   title("Figure 11: Relationship Between Deciduous Forest
and Transformed Mixed Forest Percentages",cex=1)

plot(mixed[-c(35,202)], evergreen[-c(35,202)],xlab="% Mixed Forest",ylab="%
   Evergreen Forest",axes=F,
   xlim=range(mixed),ylim=range(evergreen))
   points(mixed[35],evergreen[35],pch=35)
   ever_pretty(range(exp(evergreen)-.001))
   axis(side=2,at=log(ever+.001), lab=ever, srt=90,cex=0.6)
   mix_pretty(range(exp(mixed)-.001))
   axis(side=1,at=log(mix+.001), lab=mix, srt=0,cex=0.6)
   title("Figure 12: Relationship Between Transformed Mixed Forest
and Transformed Evergreen Forest Percentages",cex=1)

plot(woody, emergent,xlab="% Woody Wetlands",ylab="% Emergent Wetlands",axes=F)
   wood_pretty(range(exp(woody)-.001))
   axis(side=1,at=log(wood+.001), lab=wood, srt=0,cex=0.6)
   emerge_pretty(range(exp(emergent)-.001))
   axis(side=2,at=log(emerge+.001), lab=emerge, srt=90,cex=0.6)
   title("Figure 13: Relationship Between Transformed Wetland
   Percentages",cex=1)

plot(urb.hi, urb.low,xlab="% Urban-High Density",ylab="% Urban-Low
   Density",axes=F)
```

```
   high_pretty(range(exp(urb.hi)-.001))
   axis(side=1,at=log(high+.001), lab=high, srt=0,cex=0.6)
   low_pretty(range(exp(urb.low)-.001))
   axis(side=2,at=log(low+.001), lab=low, srt=90,cex=0.6)
   title("Figure 9: Relationship Between Transformed Urban Percentages", cex=1)


hist(july4$ANC,nclass=30,ylim=c(0,80),xlab="ANC",ylab="# of observations")
title("Figure 5: Distribution of ANC from DARM")

plot(july4$evergreen,july4$decid,xlab="% Evergreen Forest",ylab="%Deciduous
   Forest")
title("Figure 7: Relationship Between Evergreen Forest
and Deciduous Forest", cex=1)
tree.lm_lm(july4$decid~july4$evergreen)
plot(july4$probable,july4$row.crops, xlab="% Probable Row Crops",ylab="% Row
   Crops")
title("Figure 8: Relationship Between Probable Row Crops
and Row Crops",cex=1)
crop.lm_lm(july4$row.crops~july4$probable)

library(maps)
map(region=c('West Virginia','Virginia','Maryland'),xlim=c(-84,-
   77),ylim=c(36,40.25))
points(july5$LON.DD1,july5$LAT.DD,col=8,pch=16)
points(july5neg$LON.DD1,july5neg$LAT.DD,col=6,pch=17)
title("ANC Sites Designated by Magnitude
(Circle = Positive, Triangle = Negative)")


map(region=c('Pennsylvania','West Virginia','Virginia','Maryland','New
   York','Delaware'))
points(site.geo.elev1$LON.DD1,site.geo.elev1$LAT.DD,col=6,pch=16)
title("Figure 1: MAHA Region of the United States-
EMAP Sampled Sites",cex=1.1)
title("ANC Sites Identified by Bedrock Geologic Class",cex=1.1)
```

Anisotropic Model Fitting Using Spherical and Exponential Covariance Functions

```
date()
h_seq(0,2.6,by = .01)
par(mfrow=c(3,5))
mse_matrix(0,30,1)
msespher_matrix(0,30,1)
source("g:/st523/st523.q") #This is the spatial library of Dr. Reich and Dr.
   Davis

var1_variogrm(X,Y,resid,30,dmax=max(h),theta=40,dtheta=10)
title("theta=40,dtheta=10")
abline(var(resid),0)
fit1_fitvar1(var1,.05,.22,0.8,model="exp")
lines(expvar(h,fit1))
expfit1_expvar(var1$x,fit1)
mse[1]_sum((expfit1$y-var1$y)^2)
fit1s_fitvar1(var1,.05,.22,2,model="sph")
lines(sphervar(h,fit1s),lty=2)
spherfit1_sphervar(var1$x,fit1s)
msespher[1]_sum((spherfit1$y-var1$y)^2)
```

```
var2_variogrm(X,Y,resid,30,dmax=max(h),theta=45,dtheta=10)
title("theta=45,dtheta=10")
abline(var(resid),0)
fit2_fitvar1(var2,.05,.22,0.8,model="exp")
lines(expvar(h,fit2))
expfit2_expvar(var2$x,fit2)
mse[2]_sum((expfit2$y-var2$y)^2)
fit2s_fitvar1(var2,.05,.22,2,model="sph")
lines(sphervar(h,fit2s),lty=2)
spherfit2_sphervar(var2$x,fit2s)
msespher[2]_sum((spherfit2$y-var2$y)^2)

var3_variogrm(X,Y,resid,30,dmax=max(h),theta=50,dtheta=10)
title("theta=50,dtheta=10")
abline(var(resid),0)
fit3_fitvar1(var3,.05,.22,0.8,model="exp")
lines(expvar(h,fit3))
expfit3_expvar(var3$x,fit3)
mse[3]_sum((expfit3$y-var3$y)^2)
fit3s_fitvar1(var3,.05,.22,2,model="sph")
lines(sphervar(h,fit3s),lty=2)
spherfit3_sphervar(var3$x,fit3s)
msespher[3]_sum((spherfit3$y-var3$y)^2)

var4_variogrm(X,Y,resid,30,dmax=max(h),theta=55,dtheta=10)
title("theta=55,dtheta=10")
abline(var(resid),0)
fit4_fitvar1(var4,.05,.22,0.8,model="exp")
lines(expvar(h,fit4))
expfit4_expvar(var4$x,fit4)
mse[4]_sum((expfit4$y-var4$y)^2)
fit4s_fitvar1(var4,.05,.22,2,model="sph")
lines(sphervar(h,fit4s),lty=2)
spherfit4_sphervar(var4$x,fit4s)
msespher[4]_sum((spherfit4$y-var4$y)^2)

var5_variogrm(X,Y,resid,30,dmax=max(h),theta=60,dtheta=10)
title("theta=60,dtheta=10")
abline(var(resid),0)
fit5_fitvar1(var5,.05,.22,0.8,model="exp")
lines(expvar(h,fit5))
expfit5_expvar(var5$x,fit5)
mse[5]_sum((expfit5$y-var5$y)^2)
fit5s_fitvar1(var5,.05,.22,2,model="sph")
lines(sphervar(h,fit5s),lty=2)
spherfit5_sphervar(var5$x,fit5s)
msespher[5]_sum((spherfit5$y-var5$y)^2)

var6_variogrm(X,Y,resid,30,dmax=max(h),theta=40,dtheta=15)
title("theta=40,dtheta=15")
abline(var(resid),0)
fit6_fitvar1(var6,.05,.22,0.8,model="exp")
lines(expvar(h,fit6))
expfit6_expvar(var6$x,fit6)
mse[6]_sum((expfit6$y-var6$y)^2)
fit6s_fitvar1(var6,.05,.22,2,model="sph")
lines(sphervar(h,fit6s),lty=2)
spherfit6_sphervar(var6$x,fit6s)
msespher[6]_sum((spherfit6$y-var6$y)^2)

var7_variogrm(X,Y,resid,30,dmax=max(h),theta=45,dtheta=15)
title("theta=45,dtheta=15")
```

```
abline(var(resid),0)
fit7_fitvar1(var7,.05,.22,0.8,model="exp")
lines(expvar(h,fit7))
expfit7_expvar(var7$x,fit7)
mse[7]_sum((expfit7$y-var7$y)^2)
fit7s_fitvar1(var7,.05,.22,2,model="sph")
lines(sphervar(h,fit7),lty=2)
spherfit7_sphervar(var7$x,fit7s)
msespher[7]_sum((spherfit7$y-var7$y)^2)


var8_variogrm(X,Y,resid,30,dmax=max(h),theta=50,dtheta=15)
title("theta=50,dtheta=15")
abline(var(resid),0)
fit8_fitvar1(var8,.05,.22,0.8,model="exp")
lines(expvar(h,fit8))
expfit8_expvar(var8$x,fit8)
mse[8]_sum((expfit8$y-var8$y)^2)
fit8s_fitvar1(var8,.05,.22,2,model="sph")
lines(sphervar(h,fit8s),lty=2)
spherfit8_sphervar(var8$x,fit8s)
msespher[8]_sum((spherfit8$y-var8$y)^2)

var9_variogrm(X,Y,resid,30,dmax=max(h),theta=55,dtheta=15)
title("theta=55,dtheta=15")
abline(var(resid),0)
fit9_fitvar1(var9,.05,.22,0.8,model="exp")
lines(expvar(h,fit9))
expfit9_expvar(var9$x,fit9)
mse[9]_sum((expfit9$y-var9$y)^2)
fit9s_fitvar1(var9,.05,.22,2,model="sph")
lines(sphervar(h,fit9s),lty=2)
spherfit9_sphervar(var9$x,fit9s)
msespher[9]_sum((spherfit9$y-var9$y)^2)

var10_variogrm(X,Y,resid,30,dmax=max(h),theta=60,dtheta=15)
title("theta=60,dtheta=15")
abline(var(resid),0)
fit10_fitvar1(var10,.05,.22,0.8,model="exp")
lines(expvar(h,fit10))
expfit10_expvar(var10$x,fit10)
mse[10]_sum((expfit10$y-var10$y)^2)
fit10s_fitvar1(var10,.05,.22,2,model="sph")
lines(sphervar(h,fit10s),lty=2)
spherfit10_sphervar(var10$x,fit10s)
msespher[10]_sum((spherfit10$y-var10$y)^2)

var11_variogrm(X,Y,resid,30,dmax=max(h),theta=40,dtheta=20)
title("theta=40,dtheta=20")
abline(var(resid),0)
fit11_fitvar1(var11,.05,.22,0.8,model="exp")
lines(expvar(h,fit11))
expfit11_expvar(var11$x,fit11)
mse[11]_sum((expfit11$y-var11$y)^2)
fit11s_fitvar1(var11,.05,.22,2,model="sph")
lines(sphervar(h,fit11s),lty=2)
spherfit11_sphervar(var11$x,fit11s)
msespher[11]_sum((spherfit11$y-var11$y)^2)

var12_variogrm(X,Y,resid,30,dmax=max(h),theta=45,dtheta=20)
title("theta=45,dtheta=20")
abline(var(resid),0)
fit12_fitvar1(var12,.05,.22,0.8,model="exp")
```

```
lines(expvar(h,fit12))
expfit12_expvar(var12$x,fit12)
mse[12]_sum((expfit12$y-var12$y)^2)
fit12s_fitvar1(var12,.05,.22,2,model="sph")
lines(sphervar(h,fit12s),lty=2)
spherfit12_sphervar(var12$x,fit12s)
msespher[12]_sum((spherfit12$y-var12$y)^2)


var13_variogrm(X,Y,resid,30,dmax=max(h),theta=50,dtheta=20)
title("theta=50,dtheta=20")
abline(var(resid),0)
fit13_fitvar1(var13,.05,.22,0.8,model="exp")
lines(expvar(h,fit13))
expfit13_expvar(var13$x,fit13)
mse[13]_sum((expfit13$y-var13$y)^2)
fit13s_fitvar1(var13,.05,.22,2,model="sph")
lines(sphervar(h,fit13s),lty=2)
spherfit13_sphervar(var13$x,fit13s)
msespher[13]_sum((spherfit13$y-var13$y)^2)

var14_variogrm(X,Y,resid,30,dmax=max(h),theta=55,dtheta=20)
title("theta=55,dtheta=20")
abline(var(resid),0)
fit14_fitvar1(var14,.05,.22,0.8,model="exp")
lines(expvar(h,fit14))
expfit14_expvar(var14$x,fit14)
mse[14]_sum((expfit14$y-var14$y)^2)
fit14s_fitvar1(var14,.05,.22,2,model="sph")
lines(sphervar(h,fit14s),lty=2)
spherfit14_sphervar(var14$x,fit14s)
msespher[14]_sum((spherfit14$y-var14$y)^2)


var15_variogrm(X,Y,resid,30,dmax=max(h),theta=60,dtheta=20)
title("theta=60,dtheta=20")
abline(var(resid),0)
fit15_fitvar1(var15,.05,.22,0.8,model="exp")
lines(expvar(h,fit15))
expfit15_expvar(var15$x,fit15)
mse[15]_sum((expfit15$y-var15$y)^2)
fit15s_fitvar1(var15,.05,.22,2,model="sph")
lines(sphervar(h,fit15s),lty=2)
spherfit15_sphervar(var15$x,fit15s)
msespher[15]_sum((spherfit15$y-var15$y)^2)

stamp()

var16_variogrm(X,Y,resid,30,dmax=max(h),theta=40,dtheta=25)
title("theta=40,dtheta=25")
abline(var(resid),0)
fit16_fitvar1(var16,.05,.22,0.8,model="exp")
lines(expvar(h,fit16))
expfit16_expvar(var16$x,fit16)
mse[16]_sum((expfit16$y-var16$y)^2)
fit16s_fitvar1(var16,.05,.22,2,model="sph")
lines(sphervar(h,fit16s),lty=2)
spherfit16_sphervar(var16$x,fit16s)
msespher[16]_sum((spherfit16$y-var16$y)^2)

var17_variogrm(X,Y,resid,30,dmax=max(h),theta=45,dtheta=25)
title("theta=45,dtheta=25")
abline(var(resid),0)
```

```
fit17_fitvar1(var17,.05,.22,0.8,model="exp")
lines(expvar(h,fit17))
expfit17_expvar(var17$x,fit17)
mse[17]_sum((expfit17$y-var17$y)^2)
fit17s_fitvar1(var17,.05,.22,2,model="sph")
lines(sphervar(h,fit17s),lty=2)
spherfit17_sphervar(var17$x,fit17s)
msespher[17]_sum((spherfit17$y-var17$y)^2)

var18_variogrm(X,Y,resid,30,dmax=max(h),theta=50,dtheta=25)
title("theta=50,dtheta=25")
abline(var(resid),0)
fit18_fitvar1(var18,.05,.22,0.8,model="exp")
lines(expvar(h,fit18))
expfit18_expvar(var18$x,fit18)
mse[18]_sum((expfit18$y-var18$y)^2)
fit18s_fitvar1(var18,.05,.22,2,model="sph")
lines(sphervar(h,fit18s),lty=2)
spherfit18_sphervar(var18$x,fit18s)
msespher[18]_sum((spherfit18$y-var18$y)^2)

var19_variogrm(X,Y,resid,30,dmax=max(h),theta=55,dtheta=25)
title("theta=55,dtheta=25")
abline(var(resid),0)
fit19_fitvar1(var19,.05,.22,0.8,model="exp")
lines(expvar(h,fit19))
expfit19_expvar(var19$x,fit19)
mse[19]_sum((expfit19$y-var19$y)^2)
fit19s_fitvar1(var19,.05,.22,2,model="sph")
lines(sphervar(h,fit19s),lty=2)
spherfit19_sphervar(var19$x,fit19s)
msespher[19]_sum((spherfit19$y-var19$y)^2)

var20_variogrm(X,Y,resid,30,dmax=max(h),theta=60,dtheta=25)
title("theta=60,dtheta=25")
abline(var(resid),0)
fit20_fitvar1(var20,.05,.22,0.8,model="exp")
lines(expvar(h,fit20))
expfit20_expvar(var20$x,fit20)
mse[20]_sum((expfit20$y-var20$y)^2)
fit20s_fitvar1(var20,.05,.22,2,model="sph")
lines(sphervar(h,fit20s),lty=2)
spherfit20_sphervar(var20$x,fit20s)
msespher[20]_sum((spherfit20$y-var20$y)^2)

var21_variogrm(X,Y,resid,30,dmax=max(h),theta=40,dtheta=30)
title("theta=40,dtheta=30")
abline(var(resid),0)
fit21_fitvar1(var21,.05,.22,0.8,model="exp")
lines(expvar(h,fit21))
expfit21_expvar(var21$x,fit21)
mse[21]_sum((expfit21$y-var21$y)^2)
fit21s_fitvar1(var21,.05,.22,2,model="sph")
lines(sphervar(h,fit21s),lty=2)
spherfit21_sphervar(var21$x,fit21s)
msespher[21]_sum((spherfit21$y-var21$y)^2)

var22_variogrm(X,Y,resid,30,dmax=max(h),theta=45,dtheta=30)
title("theta=45,dtheta=30")
abline(var(resid),0)
fit22_fitvar1(var22,.05,.22,0.8,model="exp")
lines(expvar(h,fit22))
expfit22_expvar(var22$x,fit22)
```

```
mse[22]_sum((expfit22$y-var22$y)^2)
fit22s_fitvar1(var22,.05,.22,2,model="sph")
lines(sphervar(h,fit22s),lty=2)
spherfit22_sphervar(var22$x,fit22s)
msespher[22]_sum((spherfit22$y-var22$y)^2)

var23_variogrm(X,Y,resid,30,dmax=max(h),theta=50,dtheta=30)
title("theta=50,dtheta=30")
abline(var(resid),0)
fit23_fitvar1(var23,.05,.22,0.8,model="exp")
lines(expvar(h,fit23))
expfit23_expvar(var23$x,fit23)
mse[23]_sum((expfit23$y-var23$y)^2)
fit23s_fitvar1(var23,.05,.22,2,model="sph")
lines(sphervar(h,fit23s),lty=2)
spherfit23_sphervar(var23$x,fit23s)
msespher[23]_sum((spherfit23$y-var23$y)^2)

var24_variogrm(X,Y,resid,30,dmax=max(h),theta=55,dtheta=30)
title("theta=55,dtheta=30")
abline(var(resid),0)
fit24_fitvar1(var24,.05,.22,0.8,model="exp")
lines(expvar(h,fit24))
expfit24_expvar(var24$x,fit24)
mse[24]_sum((expfit24$y-var24$y)^2)
fit24s_fitvar1(var24,.05,.22,2,model="sph")
lines(sphervar(h,fit24s),lty=2)
spherfit24_sphervar(var24$x,fit24s)
msespher[24]_sum((spherfit24$y-var24$y)^2)

var25_variogrm(X,Y,resid,30,dmax=max(h),theta=60,dtheta=30)
title("theta=60,dtheta=30")
abline(var(resid),0)
fit25_fitvar1(var25,.05,.22,0.8,model="exp")
lines(expvar(h,fit25))
expfit25_expvar(var25$x,fit25)
mse[25]_sum((expfit25$y-var25$y)^2)
fit25s_fitvar1(var25,.05,.22,2,model="sph")
lines(sphervar(h,fit25s),lty=2)
spherfit25_sphervar(var25$x,fit25s)
msespher[25]_sum((spherfit25$y-var25$y)^2)

var26_variogrm(X,Y,resid,30,dmax=max(h),theta=40,dtheta=35)
title("theta=40,dtheta=35")
abline(var(resid),0)
fit26_fitvar1(var26,.05,.22,0.8,model="exp")
lines(expvar(h,fit26))
expfit26_expvar(var26$x,fit26)
mse[26]_sum((expfit26$y-var26$y)^2)
fit26s_fitvar1(var26,.05,.22,2,model="sph")
lines(sphervar(h,fit26s),lty=2)
spherfit26_sphervar(var26$x,fit26s)
msespher[26]_sum((spherfit26$y-var26$y)^2)

var27_variogrm(X,Y,resid,30,dmax=max(h),theta=45,dtheta=35)
title("theta=45,dtheta=35")
abline(var(resid),0)
fit27_fitvar1(var27,.05,.22,0.8,model="exp")
lines(expvar(h,fit27))
expfit27_expvar(var27$x,fit27)
mse[27]_sum((expfit27$y-var27$y)^2)
fit27s_fitvar1(var27,.05,.22,2,model="sph")
lines(sphervar(h,fit27s),lty=2)
```

```
spherfit27_sphervar(var27$x,fit27s)
msespher[27]_sum((spherfit27$y-var27$y)^2)

var28_variogrm(X,Y,resid,30,dmax=max(h),theta=50,dtheta=35)
title("theta=50,dtheta=35")
abline(var(resid),0)
fit28_fitvar1(var28,.05,.22,0.8,model="exp")
lines(expvar(h,fit28))
expfit28_expvar(var28$x,fit28)
mse[28]_sum((expfit28$y-var28$y)^2)
fit28s_fitvar1(var28,.05,.22,2,model="sph")
lines(sphervar(h,fit28s),lty=2)
spherfit28_sphervar(var28$x,fit28s)
msespher[28]_sum((spherfit28$y-var28$y)^2)


var29_variogrm(X,Y,resid,30,dmax=max(h),theta=55,dtheta=35)
title("theta=55,dtheta=35")
abline(var(resid),0)
fit29_fitvar1(var29,.05,.22,0.8,model="exp")
lines(expvar(h,fit29))
expfit29_expvar(var29$x,fit29)
mse[29]_sum((expfit29$y-var29$y)^2)
fit29s_fitvar1(var29,.05,.22,2,model="sph")
lines(sphervar(h,fit29s),lty=2)
spherfit29_sphervar(var29$x,fit29s)
msespher[29]_sum((spherfit29$y-var29$y)^2)

var30_variogrm(X,Y,resid,30,dmax=max(h),theta=60,dtheta=35)
title("theta=60,dtheta=35")
abline(var(resid),0)
fit30_fitvar1(var30,.05,.22,0.8,model="exp")
lines(expvar(h,fit30))
expfit30_expvar(var30$x,fit30)
mse[30]_sum((expfit30$y-var30$y)^2)
fit30s_fitvar1(var30,.05,.22,2,model="sph")
lines(sphervar(h,fit30s),lty=2)
spherfit30_sphervar(var30$x,fit30s)
msespher[30]_sum((spherfit30$y-var30$y)^2)
stamp()

date()


fitrange_matrix(0,30,1)
fitrange[1]_fit1$range
fitrange[2]_fit2$range
fitrange[3]_fit3$range
fitrange[4]_fit4$range
fitrange[5]_fit5$range
fitrange[6]_fit6$range
fitrange[7]_fit7$range
fitrange[8]_fit8$range
fitrange[9]_fit9$range
fitrange[10]_fit10$range
fitrange[11]_fit11$range
fitrange[12]_fit12$range
fitrange[13]_fit13$range
fitrange[14]_fit14$range
fitrange[15]_fit15$range
fitrange[16]_fit16$range
fitrange[17]_fit17$range
fitrange[18]_fit18$range
```

```
fitrange[19]_fit19$range
fitrange[20]_fit20$range
fitrange[21]_fit21$range
fitrange[22]_fit22$range
fitrange[23]_fit23$range
fitrange[24]_fit24$range
fitrange[25]_fit25$range
fitrange[26]_fit26$range
fitrange[27]_fit27$range
fitrange[28]_fit28$range
fitrange[29]_fit29$range
fitrange[30]_fit30$range

fitrange1_matrix(0,30,1)
fitrange1[1]_fit1s$range
fitrange1[2]_fit2s$range
fitrange1[3]_fit3s$range
fitrange1[4]_fit4s$range
fitrange1[5]_fit5s$range
fitrange1[6]_fit6s$range
fitrange1[7]_fit7s$range
fitrange1[8]_fit8s$range
fitrange1[9]_fit9s$range
fitrange1[10]_fit10s$range
fitrange1[11]_fit11s$range
fitrange1[12]_fit12s$range
fitrange1[13]_fit13s$range
fitrange1[14]_fit14s$range
fitrange1[15]_fit15s$range
fitrange1[16]_fit16s$range
fitrange1[17]_fit17s$range
fitrange1[18]_fit18s$range
fitrange1[19]_fit19s$range
fitrange1[20]_fit20s$range
fitrange1[21]_fit21s$range
fitrange1[22]_fit22s$range
fitrange1[23]_fit23s$range
fitrange1[24]_fit24s$range
fitrange1[25]_fit25s$range
fitrange1[26]_fit26s$range
fitrange1[27]_fit27s$range
fitrange1[28]_fit28s$range
fitrange1[29]_fit29s$range
fitrange1[30]_fit30s$range

exponent_cbind(mse,fitrange)
sphere_cbind(msespher,fitrange1)
fitting_cbind(exponent,sphere)
```