DEVELOPING CRITERIA FOR EXTRACTING PRINCIPAL COMPONENTS AND ASSESSING MULTIPLE SIGNIFICANCE TESTS IN KNOWLEDGE DISCOVERY APPLICATIONS

Kellie Bliss Keeling, B.S., M.B.A.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 1999

APPROVED:

Robert J. Pavur, Major Professor and Chair
Michael G. Monticino, Minor Professor
Alan H. Kvanli, Committee Member
Maliyakal D. Jayakumar, Committee Member
Victor R. Prybutok, Ph.D. Coordinator of the Department of Business Computer Information Systems
John C. Windsor, Chair of the Department of Business Computer Information Systems
Richard E. White, Associate Dean and Doctoral Director of the College of Business
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Keeling, Kellie Bliss, Developing criteria for extracting principal components and assessing multiple significance tests in knowledge discovery applications. Doctor of Philosophy (Management Science), August, 1999, 229 pp., 62 tables, 52 illustrations, references, 100 titles.

With advances in computer technology, organizations are able to store large amounts of data in data warehouses. There are two fundamental issues researchers must address: the dimensionality of data and the interpretation of multiple statistical tests.

The first issue addressed by this research is the determination of the number of components to retain in principal components analysis. This research establishes regression, asymptotic theory, and neural network approaches for estimating mean and 95^{th} percentile eigenvalues for implementing Horn's parallel analysis procedure for retaining components. Certain methods perform better for specific combinations of sample size and numbers of variables. The adjusted normal order statistic estimator (ANOSE), an asymptotic procedure, performs the best overall. Future research is warranted on combining methods to increase accuracy.

The second issue involves interpreting multiple statistical tests. This study uses simulation to show that Parker and Rothenberg's technique using a density function with a mixture of betas to model *p*-values is viable for *p*-values from central and non-central t distributions. The simulation study shows that final estimates obtained in the proposed mixture approach reliably estimate the true proportion of the distributions associated with the null and nonnull hypotheses. Modeling the density of *p*-values allows for better control of the true experimentwise error rate and is used to provide insight into grouping

hypothesis tests for clustering purposes. Future research will expand the simulation to include *p*-values generated from additional distributions.

The techniques presented are applied to data from Lake Texoma where the size of the database and the number of hypotheses of interest call for nontraditional data mining techniques. The issue is to determine if information technology can be used to monitor the chlorophyll levels in the lake as chloride is removed upstream. A relationship established between chlorophyll and the energy reflectance, which can be measured by satellites, enables more comprehensive and frequent monitoring. The results have both economic and political ramifications.

ACKNOWLEDGMENTS

I thank Dr. Robert Pavur, my committee chair, for his academic example, guidance, and friendship throughout my doctoral studies. Particularly I acknowledge his advice, encouragement, and time during the creation of this work. Additionally, I appreciate the support and recommendations from my committee members Drs. Michael Monticino, Alan Kvanli, and Maliyakal Jayakumar.

Thanks to Dr. Samuel Atkinson, Professor of Environmental Sciences, for the use of his data from Lake Texoma. I also acknowledge the faculty and resources of the Business Computer Information Systems Department.

Special thanks to my husband, Kevin, whose sacrifice, unending encouragement, and partnership kept me heading toward completion of this work. Finally, thanks to my parents for instilling in me a passion for learning and always assisting me in achieving my goals.

TABLE OF CONTENTS

Page

TABLES vi

ILLUSTRATIONS ix

Chapter

1. INTRODUCTION 1
    Statement and Purpose of the Problem 4
        Purpose of Investigating Methodology for Estimating Mean Eigenvalues 4
        Purpose of Investigating Methodology for Estimating the Density of *P*-values 5
    Significance of the Study 6
    Framework of the Study 7

2. LITERATURE REVIEW 9
    Knowledge Discovery in Databases 10
    Principal Components 11
    Example of Number of Factors Problem 14
    Subset Selection Techniques 15
    Parallel Analysis 18
    Regression Procedures 22
    Neural Networks 28
    P-values and Multiple Significance Tests 29
    Mixture Models 34
    Real World Data Set 35
    Theoretical Development 36
    Research Questions 39
        Research Question 1 40
        Motivation for Research Question 1 40
        Research Question 2 41
        Motivation for Research Question 2 41
        Research Question 3 42
        Motivation for Research Question 3 42
        Research Question 4 43
        Motivation for Research Question 4 44

Research Question 5 44
Motivation for Research Question 5 45
Research Question 6 46
Motivation for Research Question 6 46
Summary of Literature Review 47

3. RESEARCH METHODOLOGY 48
Previous Regression Techniques 48
Back Propagation Neural Networks 51
P-value Plots and Mixture Models 55
Research Design for Estimating Eigenvalues 58
Developing a New Regression Model for Estimating Mean and 95th Percentile Eigenvalues 58
Developing a Neural Network Model for Estimating Mean and 95th Percentile Eigenvalues 58
Developing Asymptotic Estimates of Mean Eigenvalues 59
Assessment and Validation of New Eigenvalue Estimation Approaches 59
Simulation Design for Estimating the Density of P-values 60
Developing a Mixture Model to Estimate the Density of *P*-values 60
Assessment of the Simulations 61
Summary of Research Methodology 61

4. DATA ANALYSIS 62
Models Used in the Eigenvalue Analysis 62
Regression Models for Predicting Mean Eigenvalues 63
Neural Network Models for Predicting Mean Eigenvalues 65
Normal Scores and Asymptotic Approaches for Predicting Mean Eigenvalues 67
Understanding the Distribution of Eigenvalues 67
Average Values for Sample Eigenvalues from a Random Correlation Matrix 70
Regression Model to Predict 95th Percentile Eigenvalues 74
Neural Network Model for Predicting 95th Percentile Eigenvalues 75
Performance of the Previous and New Mean Eigenvalue Predictors 77
Performance of Models on Reduced Data Set 77
Performance of Models on Reduced Data Set Top p/3 86
Performance of Models on Full Data Set 88
Performance of Models on Full Data Set Top p/3 97
Performance of Models on Full Data Set Bottom 10% 99
Performance of the Previous and New 95th Percentile Eigenvalue Predictors . . 103
Summary of Performance of Eigenvalue Prediction Methods 109
Estimating the Density of P-values 109

Approaches for Analyzing *P*-values 111
Estimates of the Parameters 112
Simulation Study 114
Results of the Simulations 115
Summary of P-value Density Estimation 122
Real World Application 122
Summary of Real World Application 131
Summary of Data Analysis 132

ADDITIONAL FIGURES FOR CHAPTER 4 134

5. DISCUSSION OF RESULTS 144
Conclusions 144
Research Question 1 145
Research Question 2 146
Research Question 3 147
Research Question 4 148
Research Question 5 148
Research Question 6 149
Other Conclusions 150
Limitations of this Study 152
Future Research 153

APPENDIX 154

REFERENCES 222

TABLES

Table Page

1. Comparison of extraction rules 15

2. List of extraction rules 19

3. Bread volumes 32

4. NN and NNALL input parameters 66

5. NN95 input parameters 76

6. Tukey test on MSE - Reduced data set 81

7. Tukey test on MAX DEV - Reduced data set 82

8. Tukey test on MAPE - Reduced data set 84

9. Frequency of absolute errors using reduced data set 86

10. Tukey test on MSE - Reduced data set top p/3 87

11. Tukey test on MAX DEV - Reduced data set top p/3 89

12. Tukey test on MAPE - Reduced data set top p/3 90

13. Tukey test on MSE - Full data set 94

15. Tukey test on MAX DEV - Full data set 95

16. Tukey test on MAPE - Full data set 96

. Frequency of absolute errors using full data set 97

17. Tukey test on MSE - Full data set top p/3 98

20. Tukey test on MSE - Full data set bottom 10% 101

21. Tukey test on MAX DEV - Full data set bottom 10% 102

22. Tukey Test on MAPE - Full data set bottom 10% 104

23. Tukey test on MSE - 95th percentile 107

24. Tukey test on MAX DEV - 95th percentile 108

25. Tukey test on MAPE - 95th percentile 110

26. Means and standard deviations - B(2,6) 116

27. MAD and MSE - B(2,6) 117

28. Paired t-tests between initial and final estimates 119

29. Means and standard deviations - Central and noncentral t distributions 119

30. MAD and MSE - Central and noncentral t distribution 121

31. Paired t-tests between initial and final estimates 121

32. Number and proportion in groups 126

33. Combinations in null and smallest *p*-value groups 127

34. Recommendations for mean eigenvalue estimation 151

A1. R-Squares for reduced data set 155

A2. MADs for reduced data set 156

A3. MAPEs for reduced data set 157

A4. Randomized block for MSE - Reduced data set 158

A5. Randomized block for MAX DEV - Reduced data set 160

A6. Randomized block for MAPE - Reduced data set 162

A7. Randomized block for MSE - Reduced data set top p/3 164

A8. Randomized block for MAX DEV - Reduced data set top p/3 166

A9. Randomized block for MAPE - Reduced data set top p/3 168

A10. R-Squares on full data set 170

A11. MAD on full data set 172

A12. MAPE on full data set 174

A13. Randomized block for MSE - Full data set 176

A14. Randomized block for MAX DEV - Full data set 180

A15. Randomized block for MAPE - Full data set 184

A16. Randomized block for MSE - Full data set top p/3 188

A17. Randomized block for MAX DEV - Full data set top p/3 192

A18. Randomized block for MAPE - Full data set top p/3 196

A19. Randomized block for MSE bottom 10% 200

A20. Randomized block for MAX DEV bottom 10% 204

A21. Randomized block for MAPE bottom 10% 208

A22. R-Squares for 95^{th} percentile 212

A23. MADs for 95^{th} percentile 213

A24. MAPEs for 95^{th} percentile 214

A25. Randomized block for MSE - 95^{th} percentile 215

A26. Randomized block for MAX DEV - 95^{th} percentile 217

A27. Randomized block for MAPE - 95^{th} percentile 219

A28. Deleted observations from simulation study 221

ILLUSTRATIONS

Figure Page

1. Scree plot for career orientation survey 16

3. Demonstration of parallel analysis procedure 21

6. Errors when N=50, p=15 25

7. Errors when N=100, p=25 25

8. Errors when N=300, p=35 25

7. Errors when N=500, p=50 27

8. Errors when N=20, p=5 27

9. Errors when N=2000, p=10 27

10. Plot of *p*-values for bread making data 32

11. Functions of data mining 37

12. Techniques used in modeling 37

13. View of principal components analysis 38

14. Multiple hypothesis testing 39

15. Back propagation neural network model 52

16. Plot of *p*-values 56

17. Graph of sample eigenvalues for NOSE and simulation approach 71

18. Difference between NOSE and simulated eigenvalues N=50, p=20 71

19. R-Squares for reduced data set 79

20. MADs for reduced data set 79

21. MAPEs for reduced data set 79

22. R-Squares for full data set 91

23. MADs for full data set 91

24. MAPEs for full data set 91

25. R-Squares for 95th percentile 106

26. MADs for 95th percentile 106

27. MAPEs for 95th percentile 106

28. Sampling locations within five major zones of Lake Texoma 132

29. MSE for N = 50 - Reduced data set 135

30. MSE for N = 75 - Reduced data set 135

31. MSE for N = 100 - Reduced data set 135

32. MSE for N = 150 - Reduced data set 136

33. MSE for N = 200 - Reduced data set 136

34. MSE for N = 300 - Reduced data set 136

35. MSE for N = 400 - Reduced data set 137

36. MSE for N = 500 - Reduced data set 137

37. MAX DEV for N = 50 - Reduced data set 138

38. MAX DEV for N = 75 - Reduced data set 138

39. MAX DEV for N = 100 - Reduced data set 138

40. MAX DEV for N = 150 - Reduced data set 139

41. MAX DEV for N = 200 - Reduced data set 139

42. MAX DEV for N = 300 - Reduced data set 139

43. MAX DEV for N = 400 - Reduced data set 140

44. MAX DEV for N = 500 - Reduced data set 140

45. MAPE for N = 50 - Reduced data set 141

46. MAPE for N = 75 - Reduced data set 141

47. MAPE for N = 100 - Reduced data set 141

48. MAPE for N = 150 - Reduced data set 142

49. MAPE for N = 200 - Reduced data set 142

50. MAPE for N = 300 - Reduced data set 142

51. MAPE for N = 400 - Reduced data set 143

52. MAPE for N = 500 - Reduced data set 143

CHAPTER 1

INTRODUCTION

With advances in computer technology, organizations are able to store large amounts of data in data warehouses. The data are cleansed and integrated for use with techniques such as data mining to determine patterns in the information. There are many statistical techniques as well as artificial intelligence approaches such as neural networks that can be used to uncover complex patterns. These tools are capable of creating accurate statistical estimates, insightful descriptive measures, and models that can indicate an organization's current and future prospects. There are several fundamental issues researchers must address in examining information from large databases. One is the dimensionality of data and another is the interpretation of multiple statistical tests. A statistical approach used in the interpretation of dimensions, data reduction, and visualization in a data set is principal components analysis.

The first issue discussed in this research is the problem of assessing the dimensionality of a data set. Principal components analysis is used to assess the dimensionality of the data. There are many methods currently documented and used by researchers to determine the number of principal components to retain (Guttman 1954; Horn 1965; Cattell 1966). This research focuses on the implementation of Horn (1965)'s parallel analysis method. Parallel analysis requires the use of mean (or some other percentile) eigenvalues generated from random data. No closed form solution exists for

the percentiles of eigenvalues. In this research, several previously documented methods are compared with several new methods for predicting average and 95th percentile eigenvalues generated from random data. The new methods include regression, asymptotic theory, and neural networks.

A literature review reveals that business research has not yet embraced the use of the parallel analysis procedure though there have been several works that have supported its use (Hakstian, Rogers, and Cattell 1982; Zwick and Velicer 1986; Glorfeld 1995). Therefore, an example from the psychology literature is used to illustrate an application in which parallel analysis is useful. Müller and Wetzel (1998) present a study to determine if the Bech-Rafaelsen Melancholia Scale (BRMES) could be shown to be unidimensional for acutely schizophrenic patients. The scale has previously been confirmed to have one dimension for patients with major depression.

Their sample includes 132 acutely schizophrenic patients who completed the 11 item BRMES. Each item is scored on a 5-point scale. Principal components analysis results in eigenvalues of 4.0, 1.5, 1.2, and 0.8 for the first four principal components are 4.0, 1.5, 1.2, and 0.8. The authors use simulation to randomly draw ten samples of size 132 from normally distributed values with a mean and standard deviation equal to that of the sample. These simulated eigenvalues are compared to the sample data and lead to the retention of 3 components. The results conclude that the BRMES has three dimensions and is therefore not unidimensional for acutely schizophrenic patients.

The second issue facing researchers using large databases involves performing multiple statistical tests; for example, an experiment with multiple statistical tests to

determine the number of uncorrelated variables in a data set with a large number of variables. A consequence of multiple testing is that the experimentwise error rate can easily become inflated. Because of the inflation of the experimentwise error rate, several multiple comparison procedures have been proposed including Tukey's HSD procedure, Dunnet's C test, and the Fisher-Hayter test (Tukey 1953; Dunnet 1980; Hayter 1986). One popular solution in controlling for inflation of the type I error is to use an overall test, such as an overall F test in the analysis of variance procedure or a Bonferroni adjustment. However, though many researchers have used the Bonferroni adjustment, often researchers understand that this approach does not take into account the true number of null hypotheses.

This research investigates a viable approach to statistically estimating the distribution of the *p*-values and evaluates its performance in a simulation study. The knowledge of this distribution allows for better control of the true experimentwise error rate by estimating the true number of null hypotheses. This distribution is also used to provide insight into grouping hypothesis tests for clustering purposes in knowledge discovery applications with large databases.

Multiple testing and the problems associated with these are represented in the business literature in areas such as accounting, management, and economics (Nakamura, Nakamura, and Duleep 1990; Castañeda, Levin, and Dunham 1993; Lindsay 1997). A hypothetical illustration of the use of multiple testing in a one-factor design with the use of the Bonferroni adjustment is presented in Castañeda, Levin, and Dunham (1993).

Statement and Purpose of the Problem

This research addresses the problem of providing accurate and readily accessible estimates of mean and 95th percentile eigenvalues across a large range of numbers of variables and sample sizes. In addition, this research addresses the difficulties of examining results from multiple hypothesis tests occurring in applications using large databases. An evaluation of an estimation procedure of the density of *p*-values is conducted to allow for insight into the results from multiple tests.

Purpose of Investigating Methodology for Estimating Mean Eigenvalues

Researchers have regarded interpretation of eigenvalues and eigenvectors as being essential to decision making processes involving multivariate statistical analyses. This research investigates the interpretation of eigenvalues from correlation matrices. Correlation matrices are used frequently in the analysis of survey research data across a variety of disciples as well as in numerous applications such principal component regression analysis. A comparison of observed sample eigenvalues to expected eigenvalues from random data can provide insight into this interpretation. Several methods are available for obtaining a meaningful number of eigenvalues. These methods have often been criticized for retaining too many eigenvalues. In this case, some of the eigenvalues do not explain a sufficient amount of variation in the data.

Dimensionality is a term used to describe the minimum number of constructs or latent factors that underlie the data set. For example, if a set of variables has a

dimensionality of two, then two new variables which may be linear combinations of the existing variables would explain most of the variability of the data. Knowing the number of eigenvalues to retain in a study is important for understanding the dimensionality of data. Often visual representation is needed to obtain insight into relationships in a large database. Since graphs using pairs of principal components can be used for descriptive or clustering purposes, only essential principal components should be used. Horn (1965)'s parallel analysis method, which uses the mean (or another percentile) value of eigenvalues generated from random data, has been recommended. However, these random data eigenvalues may not be readily available unless a researcher has access to tables of previously simulated mean eigenvalues or to tables with coefficients to be used in a regression equation to approximate the eigenvalues. The methodology introduced in this paper allows for eigenvalue estimates to be obtained from standard tables of order statistics for the standard normal distribution as well as from two new regression equations that do not require extensive tables. In addition, eigenvalues estimated with neural networks are also investigated.

Purpose of Investigating Methodology for Estimating the Density of *P*-values

The problem of multiple tests or multiple steps in various statistical procedures has been a call for alarm in assessing the merits of procedures such as step-wise regression, path analysis, multiple comparisons, and data mining procedures. The problem here is that when many statistical tests are conducted; for example, several hundred or even

thousands, many tests will show significance merely because of the random nature of the experiment. Separating significant tests that occur because the alternative hypothesis is true from the those tests that are significant because of experimental error is a problem that has long plagued practitioners who conduct multiple statistical tests. While an attempt is usually made to control the experimentwise error rate in many studies, an approach in which the distribution of the *p*-values is estimated has received little attention. This distribution of *p*-values is used to provide insight into grouping hypothesis tests for clustering purposes in knowledge discovery with large databases. This approach is also beneficial to researchers wishing to obtain an experimentwise error rate which is neither too liberal nor too conservative.

Significance of the Study

1. Principal component scores can be used as input variables for further analysis of the data using other multivariate techniques such as cluster analysis, regression, and discriminant analysis. Researchers typically can interpret statistical analyses easier with uncorrelated variables than with correlated ones. Principal component scores are newly created variables that possess the desired characteristic of being uncorrelated (Sharma 1996). Therefore, the problem of multicollinearity, which occurs in many statistical procedures, can be circumvented by using principal components as independent variables. Principal component scores can also be used in variable subset selection procedures.

2. Parallel analysis provides a method of determining the number of components to retain. This research provides quick methods of computing mean and 95th percentile eigenvalues without extensive tables of coefficients or requiring a simulation study. These methods give researchers a rapid method of determining a starting point for the number of components problem.
3. Researchers need to be concerned with the experimentwise error rate when performing multiple hypothesis tests, but they may be unnecessarily reducing the value of the error rate. If an estimation of the density of *p*-values can be used based on a revised estimate of the true number of null hypotheses, then an experimentwise error rate can be properly determined. In addition, the distribution of *p*-values provides a procedure to group hypothesis tests into clusters for knowledge discovery with large databases.

Framework of the Study

A literature review in the beginning of chapter 2 provides the basis for the theoretical development of the research questions presented at the end of that chapter. Chapter 3 discusses the methodology used in creating several proposed techniques as well as existing methods and in testing the research questions that address the feasibility of these techniques. In chapter 4, these techniques are developed and illustrated in an application to a real world set of data that are currently under examination to address the future of a major recreational center. Chapter 5 answers the research questions and the

appendix provides additional tables to support the viability and limitations of techniques used in this investigation.

CHAPTER 2

LITERATURE REVIEW

This chapter includes a literature review of the research surrounding the problems associated with determining the dimensionality of data and with interpreting the results of a large number of statistical tests. First, a discussion of knowledge discovery in databases is presented. Knowledge discovery in databases (KDD) involves discovering useful knowledge from data (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Principal components analysis and the methods currently documented and used by researchers (Guttman 1954; Horn 1965; Cattell 1966) to retain components and select a subset of variables are introduced next. This research focuses on the implementation of Horn (1965)'s parallel analysis method. Several regression procedures have been proposed to predict eigenvalues for the parallel analysis procedure. These regression procedures are also documented. In addition, a neural network procedure is presented to predict eigenvalues and therefore some previous literature on neural networks is included. A description of *p*-values, multiple significance tests, and mixture models concludes the report of previous literature. This literature review is followed by discussion of a real world data set that is used to assess the applicability of the procedures. The theoretical development surrounding this study is presented next. The definition of the research questions along with the motivation for their study conclude this chapter.

Knowledge Discovery in Databases

The technology has been developed to collect and store large volumes of business data, but readily available procedures to analyze and understand the data are lacking. Traditionally, manual analysis and interpretation have been required to turn data into knowledge. With the increasing use of data warehouses to store large volumes of data, new technologies are being developed to address this shortcoming. This collection of tools is important to the emerging field of KDD. The KDD process involves discovering useful knowledge from data (Fayyad, Piatetsky-Shapiro, and Smyth 1996).

The steps of the KDD process can be summarized as follows:

1. Selecting the application domain and the data set: This step involves focusing prior knowledge, identifying the goals of the application, and focusing on a subset of variables.
2. Data cleaning and preprocessing: Data cleaning and preprocessing involves removing outliers or noise if applicable, collecting information to model noise, creating a strategy to handle missing data, and determining database management system (DBMS) issues such as data types and a schema.
3. Data Transformation: The data transformation step entails reducing and projecting data, finding useful features to represent the data, and using dimensionality reduction or transformation methods to reduce the number of variables or find invariant representations.

4. Data Mining: Data Mining involves first understanding the data well enough to choose from among numerous procedures such as statistical techniques, artificial intelligence approaches, or optimization algorithms. This choice is made based on obtaining a representative model that illustrates hidden insights into the data, new models to examine variables of interest, and patterns that can be studied in exploratory research. .
5. Interpretation/Evaluation: The interpretation/evaluation step involves interpreting the discovered patterns, visualizing the extracted patterns, translating the patterns into a form useful to users, taking actions based on knowledge, and resolving potential conflicts with previously believed (or extracted) knowledge.

As can be seen, transformation and data mining are important steps in the KDD process. These steps involve reducing or transforming the data and then determining patterns or fitting models to data. Steps one through three are performed in data warehousing.

Principal Components

Principal components analysis (PCA) is a technique to form new variables which are linear composites of the original variables. Essentially, a new set of orthogonal axes are identified. For example, if there were p variables, one may want to represent that data in a lower m-dimensional space where m is much less than p. Dimensional reduction is based on a mathematical transformation (eigenvalue decomposition) of a covariance or correlation matrix. The loss of information resulting from dimensional reduction is

typically calculated as the sum of the variances of the new variables not used to represent that data (Sharma 1996).

Exploratory factor analysis is one of the more commonly applied multivariate procedures in the behavioral and social sciences (Glorfeld 1995). The purpose of exploratory common factor analysis (ECFA) is to discover and represent underlying unobservable latent variables which are manifest by a larger set of observable variables. An assumption is that the number of factors is known a priori. While specialized EFCA procedures exist, PCA is often used to accomplish the primary objective of ECFA. The most important decision to be made in the application of ECFA is choosing the method determine the correct number of factors to retain (Fava and Velicer 1992; Glorfeld 1995).

A variety of rationales upon which to base the number of factors decision rule have been discussed in the literature (Hakstian and Muller 1973). For example, the algebraic criteria stems from the work on algebraic bounds on the rank of the reduced correlation matrix (Ledermann 1937; Albert 1944; Guttman 1954, 1958). The psychometric criteria is from Kaiser (1960, 1965) who specified that for standardized data the amount of variance extracted by each component should, at a minimum, be equal to the variance of at least one variable. Statistical criteria arise from inferential statistical procedures that specify the number of correct factors being the number that yields a reproduced dispersion or correlation matrix that has entries that are simultaneously within normal sampling error of the observed coefficients. Finally, another criteria involves psychological importance or interpretability.

Various researchers have proposed modifications and new approaches based on these rationales for deciding how many factors to retain. Barlett (1950) develops a procedure for methodically removing each successive eigenvalue until the null hypothesis on the remaining eigenvalues fails to be rejected. Cattell (1958) creates rules for determining how much variance should be accounted for before factoring is terminated. Horn (1965) uses random data generation to determine the number of non-error latent roots (eigenvalues). Catell (1966) outlines the scree test which is performed by plotting the observed eigenvalues on a graph with lines connecting the points. The point in the graph where the line forms an "elbow" is considered to be the point where the eigenvalue position equals the number of components to retain. Velicer (1976) creates an approach based on the matrix of partial correlations. The average squared partial correlation is calculated after each of the components has been partialed out. No further components are extracted when the average squared partial correlation matrix reaches a minimum which occurs when the residual matrix most closely resembles an identity matrix. More recent work includes Krzanowski and Kline (1995) who propose a leave-one-out cross-validation method for selecting the significant principal components. Finally, Zoski and Jurs (1996) propose a linear regression approach to match the visual solution of the scree plot. According to Sharma (1996) the most common rules include the eigenvalue greater than one rule (Kaiser 1960), the scree plot (Cattell 1966), and to a lesser extent, Horn's parallel analysis (Horn 1965) procedure. An example is presented next to illustrate the conflicting results generated by some common extraction techniques.

Example of Number of Factors Problem

Igbaria and Baroudi (1993) develop a short form of a career orientation survey originated by Schein (1985). The original survey consists of 41 items while the proposed short form scale contains 25 items. These include 15 items on the importance of an employee's career measured on a 5-point scale ranging from 1 (of no importance) to 5 (centrally important). Examples include: "To be in charge of a whole organization is …" and "Remaining in my area of expertise throughout my career is …" The survey also includes 10 career preference items measured using a 5-point scale ranging from 1 (not true at all) to 5 (completely true). Examples include: "I have always wanted to start and build up a business of my own." and "A career is worthwhile only if it enables me to lead my life in my own way."

Igbaria and Baroudi (1993) use factor analysis to determine if the new survey contains the same number of factors as the original survey. They have a sample of 396 members of Mid-Atlantic chapters of the Data Processing Management Association. The first 9 sample eigenvalues are 3.91, 2.53, 2.03, 1.51, 1.46, 1.18, 1.04, 0.97, and 0.92. Igbaria and Baroudi (1993) use Bartlett's (1950) test to conclude that there are 9 factors which matched the number of factors in the original survey. As a comparison, table 1 shows the results from using several of the number of factors techniques on their data.

The "eigenvalue greater than one" rule results in 7 factors. Figure 1 shows the results from a scree plot of the sample eigenvalues. The "elbow" of this data is a bit subjective. The result could be either 4 or 6 factors. To implement the parallel analysis procedure, the Longman, Cota, Holden, and Fekken (1989) regression equation is chosen

because of the odd number of observations and variables (N = 396, p = 25). The tables of simulated values do not include this combination of N observations and p variables. Using the regression equation to estimate the random eigenvalues resulted in the eigenvalues shown in figure 2. Therefore, the parallel analysis procedure indicates 5 factors. Table 1 summarizes these four techniques. Further discussion on the accuracy of these methods is presented in the discussion of the parallel analysis methodology.

Table 1. Comparison of extraction rules

Extraction Method	# Factors
Horn's Parallel Analysis	5
Catell's Scree Plot	4 or 6
Kaiser Eigenvalue > 1	7
Barlett's Test	9

Subset Selection Techniques

Choosing a subset of principal components or variables is an important part of many statistical procedures. Subset selection techniques are used in determining the number of components in principal components analysis, in determining the number of factors in factor analysis, and in determining of the number of variables in regression analysis. The decision for determining the number of components is based on accounting for most of the variation in the p variables with m components where m is much less than p. But usually all the values of the p variables are still needed since each component may be a function of all p variables. Therefore, it might be preferable instead of using m principal components to use m (or slightly more) variables (Jolliffe 1986).

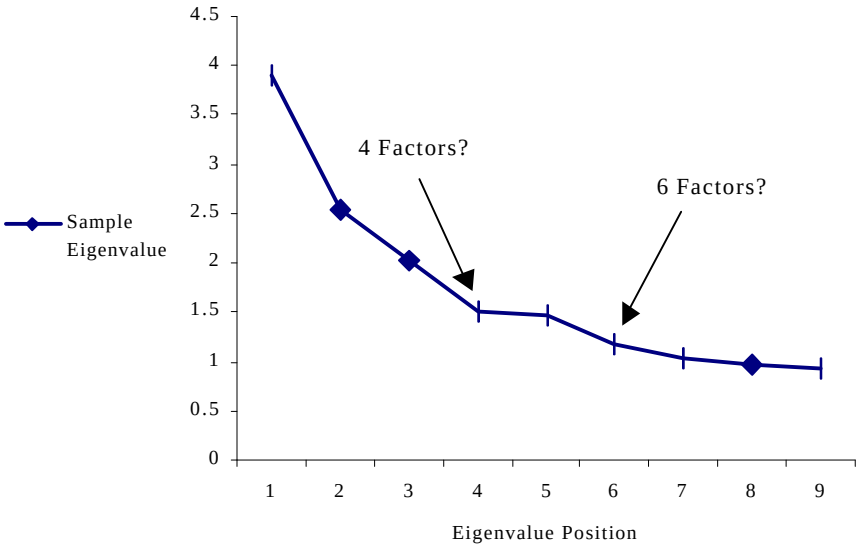

Figure 1. Scree plot for career orientation survey

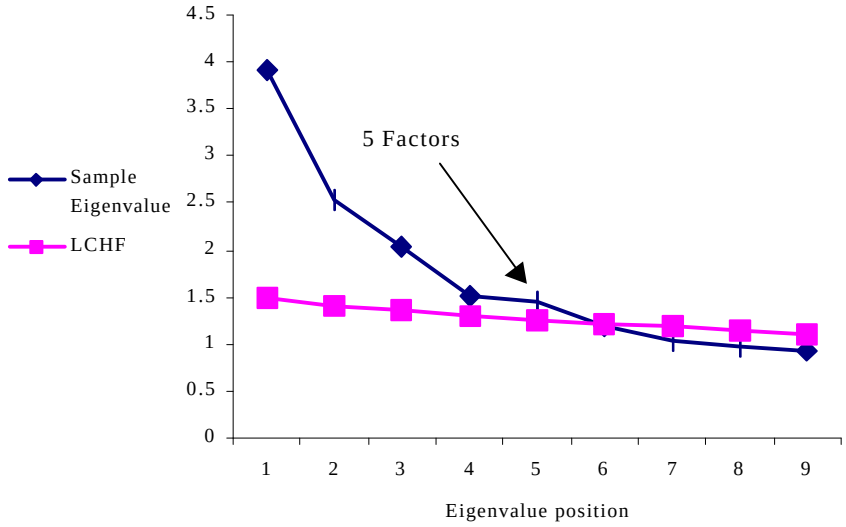

Figure 2. Parallel analysis using LCHF regression equation

There are several methods for selecting a subset of variables that are based on principal components. These are described by Jolliffe (1972, 1973, 1986). The first involves associating one variable with each of the last $m_1^* = p - m_1$ principal components and deleting those m_1^* variables. This could be done once or iteratively. The iterative procedure would involve performing a second principal components analysis on the m_1 remaining variables and deleting a further set of m_2^* variables. In this analysis, the number of variables to be deleted is chosen by a criterion based on the size of eigenvalues λ_k. Possible criterion for selecting m_1^*, m_2^*,... would be the average eigenvalues generated from random matrices or λ_k^* where

$$l_k^* = \frac{1}{p}\sum_{j=k}^{p}\frac{1}{j}$$ [1]

The reasoning supporting this method is that small eigenvalues correspond to near-perfect relationships between a subset of variables. Therefore, if one of the variables in this relationship is deleted, then little information will be lost. The choice of the variable to be deleted can be the variable with the highest coefficient in absolute value in the certain principal component.

A second method involves deleting a set of m* variables that are associated with the last m* principal components. The choice of m* is based on the size of the λ_k's. Jolliffe (1972) determined that this method was unsatisfactory in selecting an appropriate subset for some simple correlation structures. The final method associates one variable

[1]Tables are available for various values of p and k in Legendre and Legendre (1983, 406)

with each of the first m principal components successively. These m variables are retained and the remaining m* = p - m variables are deleted. This approach is an obvious complementary approach to the first method. In addition, in cases where there are groups of highly correlated variables, a single variable from each group should preserve most of the information given by that group.

Two other works have adopted a somewhat different approach to the variable selection problem. McCabe (1984) proposes a method of creating principal variables that would contain as much information as possible. Tanaka and Mori (1997) propose a modified principal components analysis that derives principal components which are linear combinations of a subset of variables based on the work of Rao (1964)'s principal components analysis of instrumental variables and Robert and Escoufier (1976)'s approach using the RV-coefficient. This recent work shows that principal components analysis continues to be refined in the literature.

Parallel Analysis

Estimating the expected value of the eigenvalue of a sample correlation matrix is be a mathematically intractable problem. Yet researchers need to make important decisions in multivariate statistical analyses that require interpretations based on the values of the sample eigenvalues. For example, the retention of factors in an exploratory factor analysis, the importance of only a few principal components, and the percentage of variation explained by several factors are issues typically resolved through an investigation of the eigenvalues of the sample correlation matrix (Johnson and Wichern 1992). As

discussed earlier, there have been several procedures that have been commonly used to determine the number of components to retain in principal components analysis and these are listed in table 2.

Table 2. List of extraction rules

Technique	Reference
Bartlett's test	Bartlett 1950
Eigenvalue greater than 1	Kaiser 1960
Parallel analysis	Horn 1965
Scree plot	Catell 1966
Minimum average partial	Velicer 1976

Zwick and Velicer (1986) present a comprehensive comparison of a number of procedures used to retain principal components. Velicer's Minimum Average Partial method is correct for 84% of the cases considered and when it is incorrect it tends to underestimate the number of factors. Fava and Velicer (1996) note that underextraction of factors is a more severe problem that overextraction. Zwick and Velicer (1986) conclude that although the scree plot is considered subjective it is still correct for 57% of the cases considered. When it is incorrect, it tends to overestimate the number of factors. The remainder of the procedures also tend to overestimate when incorrect. Bartlett's test is correct for 30% of the cases considered. The eigenvalue greater than one rule is highly inaccurate, only being correct for 22% of the cases considered. And finally, the parallel analysis method is correct for 92% of the cases considered. Therefore, the parallel analysis procedure performs comparatively well.

Another study that supports the use of parallel analysis is Hubbard and Allen (1987). Their study suggests that parallel analysis will typically extract fewer principal components than the Kaiser (1960) criterion and that components retained by this method are generally interpretable. These authors consider parallel analysis to be a more objective alternative than Cattell's (1966) scree test.

While there is a large amount of empirical support for using Horn's (1965) parallel analysis method (Hakstian, Rogers, and Cattell 1982; Zwick and Velicer 1986; Buja and Eyuboglu 1992; Glorfeld 1995), this technique is not widely known and is usually not included as an option in standard statistical packages such as SAS and SPSS. This paired with the lack of understanding of the problems with the most common techniques has caused the procedure to only slowly gain popularity. The computational complexity of performing Monte Carlo simulations required in the previous implementations of parallel analysis or having to refer to extensive tables of simulated values has also been a burden to researchers wishing to employ the method.

In addition, there is concern about whether to use the mean or a percentile, such as the 95^{th} percentile, as the reference mark to declare an eigenvalue as being important enough to warrant the retention of the corresponding principal component (Glorfeld 1995). Gorsuch (1983) noted that Barlett's (1950) chi-square test of the significance of a correlation matrix with unities as diagonal elements should be performed before the parallel analysis procedure is implemented. The rationale is that if the correlation matrix itself is truly random, then the odds are 50/50 that the first root will be higher than the root given by the parallel analysis procedure.

Briefly, the parallel analysis procedure can be described as follows. Plot the sample eigenvalues, with lines connecting the values, in order from largest to smallest on a scree plot. Similarly, include a plot of eigenvalues considered to have come from a random matrix (these could be the expected sample eigenvalues or, say, the 95th percentile of the distribution of the sample eigenvalues). The cutoff for determining the number of factors or principal components would be determined by the intersection of the two lines. Figure 3 illustrates the parallel analysis procedure.

Note that the two sample eigenvalues above the mean eigenvalues in the figure are the eigenvalues that a researcher would consider retaining. One drawback to the parallel analysis method is that it requires knowledge of either the mean eigenvalue or some selected percentile of the distribution of the sample eigenvalues. Therefore, several regression equations and tables of simulated eigenvalues have been produced by researchers to help determine estimates of the mean eigenvalue or selected percentile.

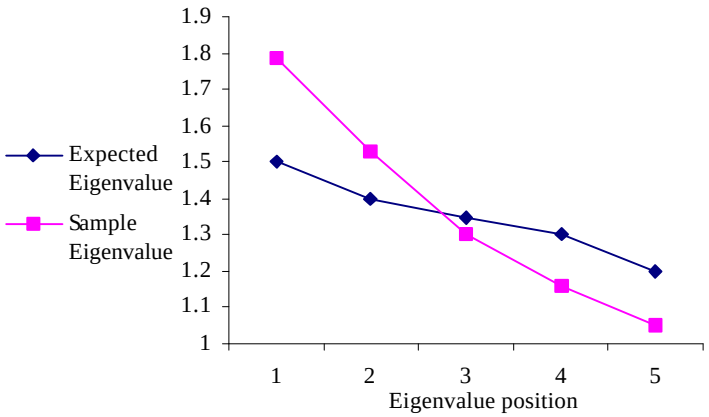

Figure 3. Demonstration of parallel analysis procedure

Regression Procedures

The development of regression equations has helped researchers to implement the parallel analysis procedure (Montanelli and Humphreys 1976; Allen and Hubbard 1986; Lautenschlager, Lance, and Flaherty 1989; Longman et al. 1989). Longman et al. (1989) compared their regression approaches to Allen and Hubbard (1986)'s regression model. Longman et al. (1989)'s study showed that their proposed model (LCHF) was more accurate than Allen and Hubbard (1986)'s model (AH). Lautenschlager, Lance, and Flaherty (1989) also published a regression model (LLF), but their model has not been extensively compared to the LCHF.

The AH equation is able to predict all but the last two eigenvalues for any value of k (eigenvalue position or eigenroot) less than or equal to 48. The AH equation has been created with sample sizes ranging from N = 30 to 1000. Cota, Longman, Holden, and Fekken (1991) note that several anomalies occur when using this iterative equation. The LLF model improves upon the performance of the AH equation by adding an additional term, p/N. Sample sizes of N = 50 to 1000 are used. This equation has been created to predict up to the 48^{th} eigenvalue. Longman et al. (1989) present the LCHF equation with two tables of coefficients; one table to allow for the prediction of the mean and the other table for the prediction of the 95^{th} percentile eigenvalue. This equation provides a noniterative approach to estimating the mean eigenvalue as well as the 95^{th} percentile of the eigenvalue. Their study uses sample sizes of N from 50 to 500.

Another approach to making parallel analysis more accessible is to use published tables of simulated eigenvalues, such as those presented in Buja and Eyuboglu (1992) and

Lautenschlager (1989). However, the combinations of sample size and number of variables is limited for these tables. For values of the sample size and the number of variables not presented in the published tables, a researcher can use interpolation. The accuracy of this procedure needs to be compared with that of regression equations to determine its merits (Cota, Longman, Holden, and Fekken 1993).

While regression models have not been provided for estimating the median value of an eigenvalue, the regression equations for the mean could be used to estimate the median. A comparison of the median eigenvalues and mean eigenvalues as presented in Buja and Eyuboglu (1992) and Lautenschlager (1989) reveals that the median and mean eigenvalues are very close in value for values of the sample size and number of variables that are found in these papers. The availability of the median and mean eigenvalues obtained through extensive simulation studies conducted by Buja and Eyuboglu (1992) and Lautenschlager (1989) facilitates determining the accuracy of the regression procedures as shown in the next four figures.

Figures 4, 5, 6, and 7 show the performance of the three regression equations within the bounds of N and p that are used in the creation of these equations. When N =50 and p = 15, figure 4 shows that the LCHF equation performs the best. This selection of N and p approaches the lower bound of the combinations of N and p. All three equations perform better at the beginning eigenvalue positions. The LLF equation, which was created as an improvement to the AH equation, performs better than the AH equation for k =1, 2, and 3. When k = 11, the LCHF equation begins to perform worse than the AH equation. The AH and LCHF equations perform within ± 0.07 for all values of k.

When N = 100 and p = 25, figure 5 demonstrates the errors occurring for each regression equation. The AH equation performs the best until k = 4. The LCHF equation errors are all within ± 0.06. The LLF equation begins to perform better than the AH equation when $k \geq 13$.

As N and p are increased, figures 6 and 7 show the performance of the three regression equations. Figure 6 shows a comparison of the errors when N = 300 and p = 35. The LCHF equation's errors are all within ± 0.015. The AH equation's errors are all within ± 0.10. The LLF equation performs better than the AH equation until k = 10 where the equation results in underestimation of the mean eigenvalue.

Figure 7 shows the errors when N = 500 and p = 50. The LCHF equation has errors that stay within ± 0.12. The LLF equation performs consistently better than the AH equation at this combination of N and p. The AH equation performs very poorly in this figure and at k = 2 the predicted value is already farther than 0.10 from the simulated value. The AH equation continues down to -12.5 when k = 33.

Figures 8 and 9 show the performance of the three regression equations outside of the bounds of N and p that were used to create the equations. Figure 8 shows that when N is small, that the LCHF equation still performs the best. The AH equation is closer than the LLF equation when k = 1, but when k = 2 and 3, their performance is similar.

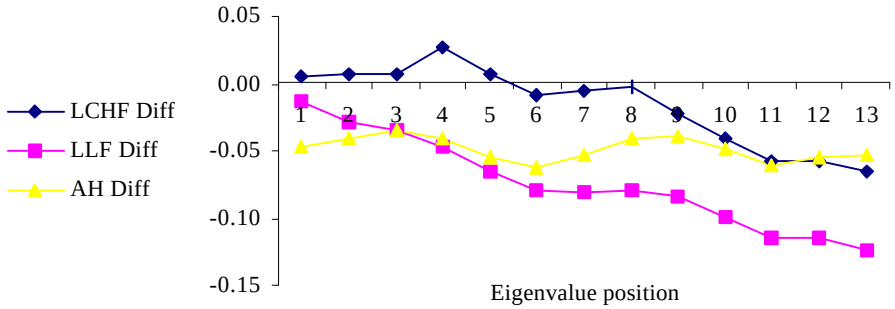

Figure 6. Errors when N=50, p=15

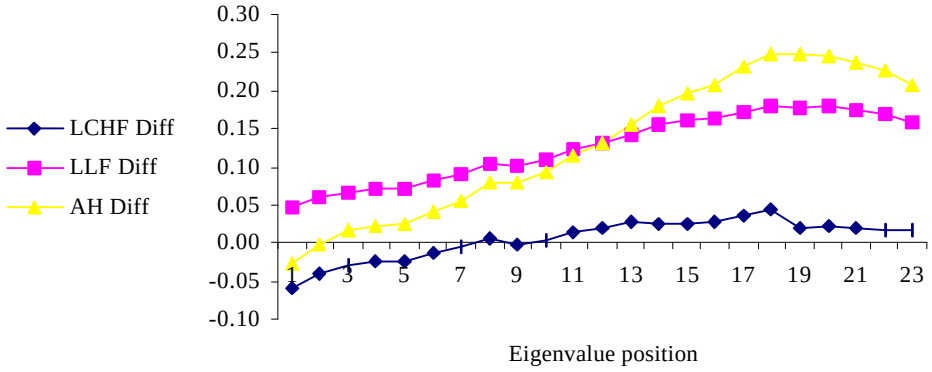

Figure 7. Errors when N=100, p=25

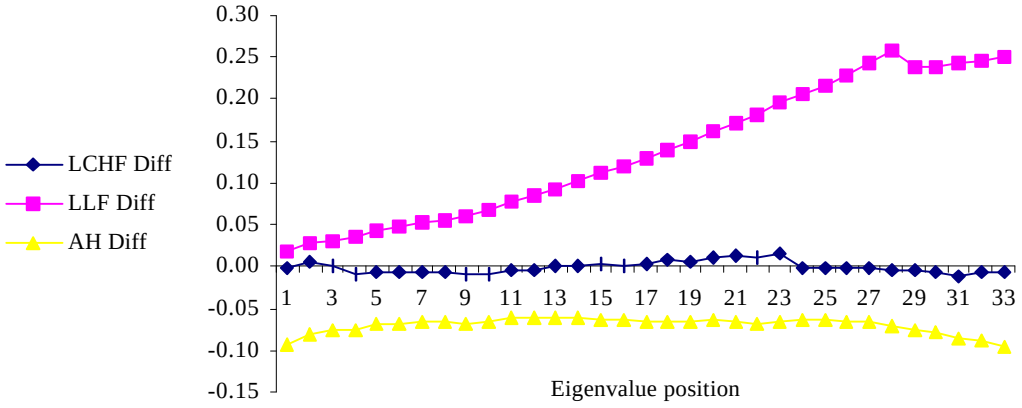

Figure 8. Errors when N=300, p=35

Figure 9 shows that when N is large, the LLF equation performs the best. All errors for the LLF are within ±0.07. This can be attributed to the fact that the LLF and AH equation have a larger upper bound for N. Both equations used values of N up to 1000. The LCHF equation only goes to values of N = 500. At this level of N, the correction that the LLF equation provides to the AH equation is evident.

The AH and LLF equations also contain anomalies within the bounds of N and p that are used to create the equations. While the LCHF equation does not have any anomalies within the bounds of N and p used to create its equation, as N and p increases past the specified bounds, anomalies do occur. Another problem with the previously proposed regression equations is the reliance on tables of coefficients in order to predict eigenvalues.

The LCHF equation performs the best within the bounds of N and p used to create the equation. This paper extends the LCHF equation by model building with second order terms. In addition, the new models do not need tables of coefficients for different values of k. The new models are also evaluated to determine if estimates of mean eigenvalues are accurate beyond the bound of the LCHF equation of N = 500.

For researchers concerned that a percentile value other than the median or mean should be used in parallel analysis to determine the dimensionality of the data, Glorfeld (1995) has suggested using a 95^{th} percentile. While several regression equations have appeared for estimating the mean value of an eigenvalue, only one regression equation, to date, has appeared for estimating a percentile other than the median. Longman et al.

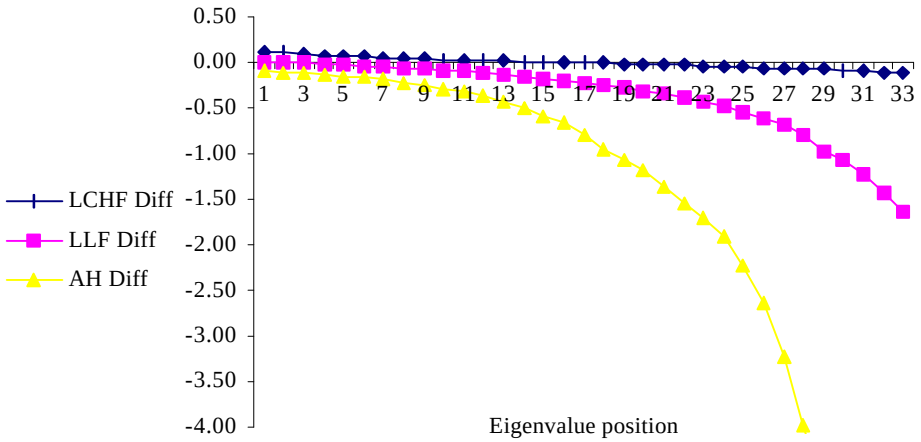

Figure 7. Errors when N=500, p=50

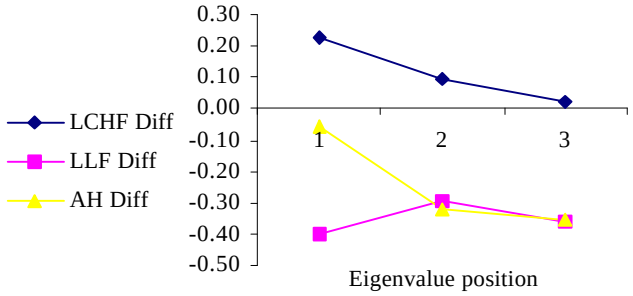

Figure 8. Errors when N=20, p=5

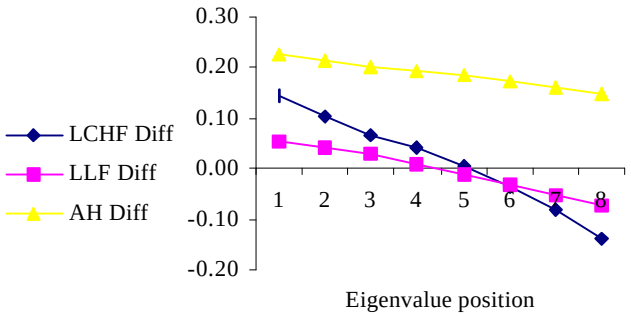

Figure 9. Errors when N=2000, p=10

(1989) have proposed a regression model for the 95th percentile. This study includes an examination of a new equation to predict 95th percentiles.

Neural Networks

Neural networks have not been previously used to predict eigenvalues. But, it is believed that they would be a good tool to use in the prediction of eigenvalues. Neural networks have been used in many business applications including airline security, futures, and banking (Brody 1990; Ruggiero 1994; Norton 1994). This technique has performed fairly well in comparisons with other statistical techniques in applications such as forecasting (Yi, Mitchell, and Prybutok 1996). Neural networks have many of the same underlying ideas as statistical regression analysis and can be thought of as a nonparametric regression method (Warner and Misra 1996). Since neural networks are data dependent, their performance often improves with increases in sample size (Warner and Misra 1996). Since the published regression procedures have performed relatively well, it is reasonable to explore the development of a neural network model to estimate mean eigenvalues.

Back propagation is one algorithm used in neural networks. The basic back propagation algorithm minimizes the mean squared error using the gradient descent method. Variations of this algorithm have been used in the literature. Selection of the number of hidden neurons and the selection of several parameters are important in implementing this algorithm. Back propagation is a powerful tool with applications in pattern recognition, dynamic modeling, sensitivity analysis, and the control of systems over time (Werbos 1990). Back propagation neural networks (BPNN) have been

successfully applied to modeling and regression situations such as household trip forecasts (Mitchell, Yi, and Govind 1996), control systems (Nguyen and Widrow 1990), and audio signal processing (Hoyt and Wechsler 1990). In this study, the BPNN is used to predict mean eigenvalues and the 95^{th} percentile of eigenvalues.

P-values and Multiple Significance Tests

The *p*-value is the value of alpha (type I error) at which the hypothesis test procedure changes conclusions based on a given set of data. The type I error rate is the probability that under the hypothetical null distribution a test statistic would produce an observed value that is considered extreme. Researchers would like to use statistical procedures that control the type I error and provide maximum power. The power of the test is the probability of making a correct decision when the null hypothesis is false. When many of these statistical procedures are used at a fixed significance level, the chance of detecting differences that are not real increases (Kirk 1995). For multiple comparison procedures, the inflation of the type I error has received much attention. Research in this area has spawned a number of multiple comparison procedures, including Tukey's HSD procedure, Dunnet's C test, and the Fisher-Hayter test (Tukey 1953; Dunnet 1980; Hayter 1986). Controlling for the experimentwise error rate has thus been a major concern for researchers.

Researchers have been warned of the problem of controlling for the experimentwise error rate in various statistical procedures including step-wise regression, path analysis, and data mining procedures. One popular solution in controlling for inflation

of the type I error is to use an overall test, such as an overall F test in analysis of variance, or a Bonferroni adjustment. However, studies such as Kromrey and Dickinson (1995) illustrate through a limited simulation study that these procedures are not satisfactory in controlling for experimentwise error.

Bobko (1986) suggests that, rather than routine application of a standard analysis, theory should guide the choice of analytic technique. All comparisons and statistical tests, if planned by the researcher's knowledge of the theoretical aspects of psychology, managerial decision making, and published findings of theoretical constructs, will have increased power and will reduce the experimentwise error rate. Many proponents of structural equation modeling (SEM) also have proposed that only theory-driven hypotheses be tested. Bollen (1989) states that SEM models need to be based on a theoretical relationships and thus be used to confirm relationships rather than to do exploratory analysis. Therefore the process should be limited by methods and theory, but controlling the type I error is also a useful device to limit searches. Users who do not control type I error across multiple tests of individual parameters may include parameters in the model that are statistically significant only because of random sample fluctuation (Green and Babyak 1997).

Often exploratory analysis needs to be performed without undue control of the experimentwise error rate. Even in a planned experiment with only one a priori hypothesis there may be hundreds or thousands of statistical tests. A researcher may think that the Bonferroni adjustment is a reasonable approach. But this procedure does not take into account an estimate of the number of hypothesis that may in fact be true. There may be

situations where the number of true null hypotheses is less than 20% of the entire number of statistical tests performed. Yet the Bonferroni adjustment will be used assuming that all of the statistical tests are coming from true null hypotheses.

Schweder and Spjotvoll (1982) proposed an approach using a *p*-value plot of cumulative *p*-values to determine the true number of null hypotheses. Schweder and Spjotvoll (1982) state that their technique is intended for informal inference and that it is difficult to make exact probability statements. An example of their technique is explained next.

To illustrate this procedure, the results of a bread-baking experiment are shown in table 3. These results are presented in Scheffé (1959, 143) and discussed in Duncan (1965) and Schweder and Spjotvoll (1982). The table lists the volumes in milliliters of loaves of bread made under controlled conditions from 100-gram batches of dough. The dough is made with 17 different varieties of wheat flour and contains x milligrams of potassium bromate, for x = 0, 1, 2, 3, and 4. The residual mean square is 1713.17 with 64 degrees of freedom. The overall F test is significant ($F = 28.34 > 1.80$) and there are 136 pairwise multiple comparisons that can be made between the 17 varieties of wheat flour. Using the Fisher Least Significant Difference rule with a significance level of 0.05, the FSD = 52.30. All of the differences which are less than the FSD, of which there were 34, are not significant. The 102 remaining do show a significant difference.

A plot of the *p*-values of the 136 pairwise tests comparing means is shown in figure 10. The plot begins on the left with the *p*-values falling close to a straight line. There is a line that is fitted visually in the figure. The slope of this line corresponds to

Table 3. Bread volumes

	Loaf Volume for x=					
Variety	0	1	2	3	4	Mean
1	950	1075	1055	975	880	987
2	890	980	955	865	825	903
3	830	850	820	770	735	801
4	770	815	765	725	700	755
5	860	1040	1065	975	945	977
6	835	960	985	915	845	908
7	795	900	905	880	785	853
8	800	860	870	850	850	846
9	750	940	1000	960	960	922
10	885	1000	1015	960	895	951
11	895	935	965	950	920	933
12	685	835	870	875	880	829
13	615	665	650	680	660	654
14	885	910	890	835	785	861
15	985	1075	1070	1015	1005	1030
16	710	750	740	725	720	729
17	785	845	865	825	820	828

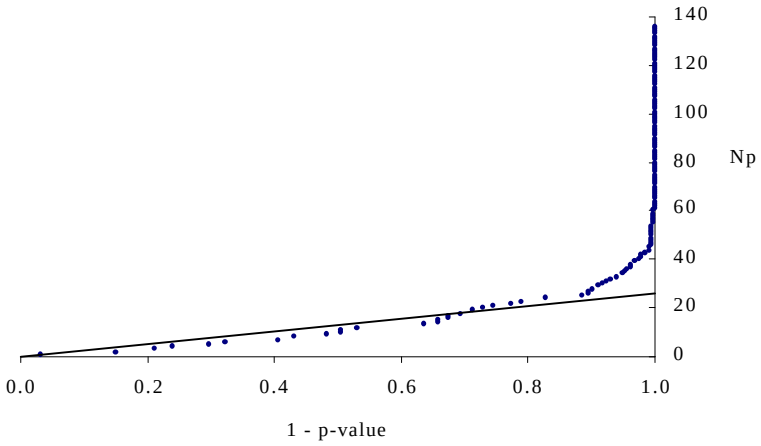

Figure 10. Plot of *p*-values for bread making data

where the line crosses at (1 - *p*-value) = 1 and estimates the true number of null hypotheses to be 25. The Fisher procedure concludes that there are 34 hypotheses that are not rejected. A conclusion made by Schweder and Spjotvoll (1982) is that the standard multiple comparison methods do not detect differences due to low power. Most likely, 9 of these are due to false null hypotheses. One would expect at most 1 or 2 false rejections at a significance level of 0.05. In order to determine a more exact method to identify null hypotheses that should clearly be rejected, Schweder and Spjotvoll (1982) propose an adjustment to the Bonferroni approach. Since there are approximately 25 null hypotheses, when aiming for an overall significance level α, a level of $\alpha/25$ should be used for the individual tests. This is an improvement over the traditional level of $\alpha/136$ for the individual tests which might be considered too stringent.

To make Schweder and Spjotvoll (1982)'s approach less subjective, Parker and Rothenberg (1988) have proposed a technique to provide more accurate information about the distribution of *p*-values coming from true and false null hypothesis. These authors propose using a density function with a mixture of betas to model *p*-values. The EM algorithm is used to obtain maximum likelihood estimates of the parameters. The next section presents an explanation of mixture models and the difficulties in using these models.

While the mixture approach to obtaining information about the distribution of *p*-values for the null and nonnull hypotheses appears to offer a fresh look at how to control for type I error, very little attention has been given to it in the literature, particularly in business problems. Parker and Rothenberg (1988) did not examine the accuracy of the

parameter estimates via a simulation study where estimates of the parameters can be manipulated. This research assesses the performance of their approach and implements their approach to create a distribution of *p*-values generated from central t and noncentral distributions as well as *p*-values from uniform and a beta distribution. The knowledge of this distribution of *p*-values allows for an estimate of the true experimentwise error rate.

Mixture Models

Mixtures of distributions have been used to model the distribution of data for a wide variety of important practical situations. Any situation where data can be viewed as arising from two or more distributions can be modeled using mixtures of distributions. However, obtaining the correct decomposition of a finite mixture model can be a very difficult problem. The basic difficulty is that a researcher typically never knows the form of distribution in the mixture nor does the researcher know the number of distributions in the mixture model. Another difficulty is that the probability density function in the mixture model may not have the property of identifiability. That is, the density functions may not uniquely determine the distribution being modeled. In addition, using a nonlinear optimization approach to obtain the estimates of the parameters in a mixture model may have convergence problems. Pearson (1894) discusses a method of moments approach in the case of a mixture of two univariate distributions with unequal variances. This method requires the solution of a ninth degree polynomial equation. Fowlkes (1979) speculated that the intractability of moment estimators and the absence of modern computer

technology caused researchers to focus their attention on graphical techniques for mixtures (Harding 1948; Cassie 1954; Bhattacharya 1967; Wilk and Gnanadesikan 1968).

As high speed computers become less expensive and more available, researchers are increasingly resorting to efficient optimization algorithms to obtain likelihood estimates of the parameters in a mixture distribution. Rao (1948) pioneered work on the detection of distributions in a mixture model and used likelihood estimation to implement the procedure. He examined a mixture of two univariate distributions with equal variances using Fisher's method of scoring. However, Butler (1986) noted that Newcomb (1886) suggested an iterative reweighting procedure, that is similar to the EM algorithm (Dempster, Laird, and Rubin 1977). Hasselblad (1966, 1969) continued Rao's work as he examined a mixture of g univariate normal distributions with equal variances and then mixtures of distributions from the exponential family. This study examines the distribution of *p*-values as a mixture of beta distributions.

Real World Data Set

A real world application using water data from a lake in the southern United States, Lake Texoma, is used to demonstrate the use of the eigenvalue and *p*-value techniques. The water in the lake is being considered for use as a water source for an area town. The level of chloride (salt) in the water must be reduced in order for the water to be used for drinking. But this reduction may reduce the chlorophyll level in the water. The chlorophyll-a concentration in the water is the best estimator for determining algal biomass, and thus the productivity, of a body of water (Atkinson, Acevedo, Dickson, and

Rolbecki 1998). There is concern that reducing the level of chlorophyll-a in the water will have a detrimental effect on plant and animal life in the region, particularly the striped bass fisheries. But chlorophyll-a is a relatively expensive parameter to measure manually.

The use of information technology to estimate the chlorophyll-a levels is planned in order to help determine the economic impact of controlling these variables. It is hypothesized that there is a relationship between reflectance of the electromagnetic energy at different wavelengths and the level of chlorophyll-a. If there is a relationship between the chlorophyll-a and the energy reflectance, then in the future satellites can be used to measure the energy output and the energy reflectance will be used to estimate the chlorophyll-a levels. The use of satellites for sampling could result in more frequent monitoring of the water as well as an expanded sampling area.

Theoretical Development

The theory surrounding the research in determining the dimensionality of data as discussed in this study begins with the functions of data mining. Data mining functions can be broken down into five main areas: associations, modeling, classification, time series, and sequential patterns (figure 11).

Modeling is the area of research that is discussed in this research. Modeling involves the relating of inputs to outputs based on previous examples. Modeling allows a researcher to make accurate predictions from complex examples with many variables. Several techniques have been used to accomplish the modeling functions. These are shown in figure 12 and include modeling nonlinear relationships with neural networks and

nonlinear regression. Multidimensional relationships are modeled with principal components analysis and multivariate statistics. In addition, linear and logistic regression can be used to model relationships. Hypothesis testing can be used in logistic regression, linear regression, and multivariate statistics.

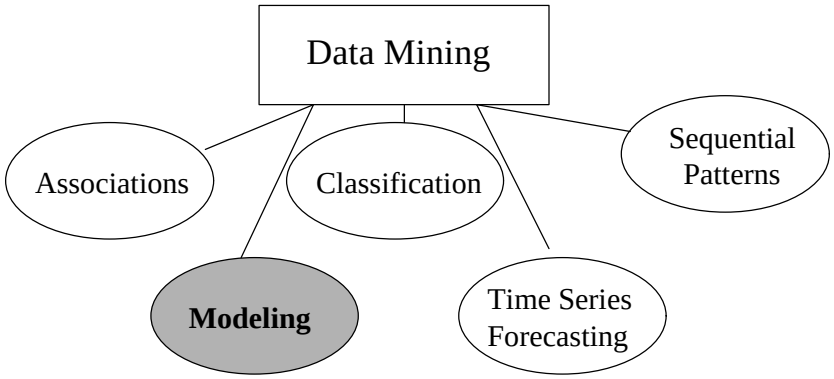

Figure 11. Functions of data mining

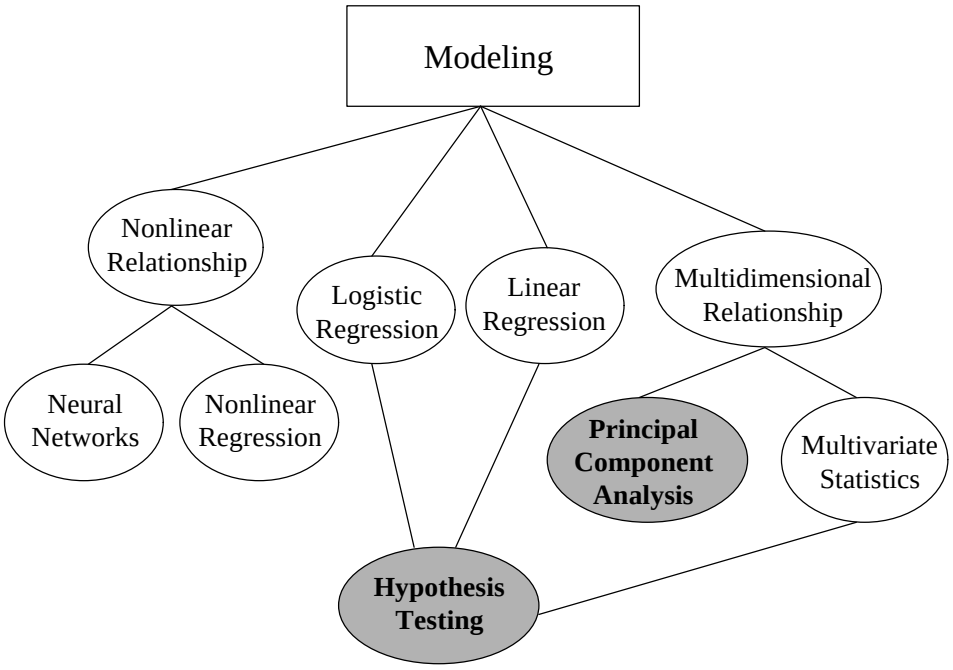

Figure 12. Techniques used in modeling

Figure 13 illustrates a grouping for principal component techniques. Horn's parallel analysis and Cattell's scree test methods are classified as a psychological importance or interpretability measures. The psychological importance approach considers the number of eigenvalues to retain by using a technique that does not require statistical significance testing. The Kaiser eigenvalue greater than one rule is classified as a psychometric measure which states that the latent root must exceed unity in order for the component to have positive alpha internal consistency. The Barlett's test for the equality of eigenvalues is a statistical measure. Statistical measures are created from inferential statistical procedures based on analyses of the reproduced covariance or correlation matrix.

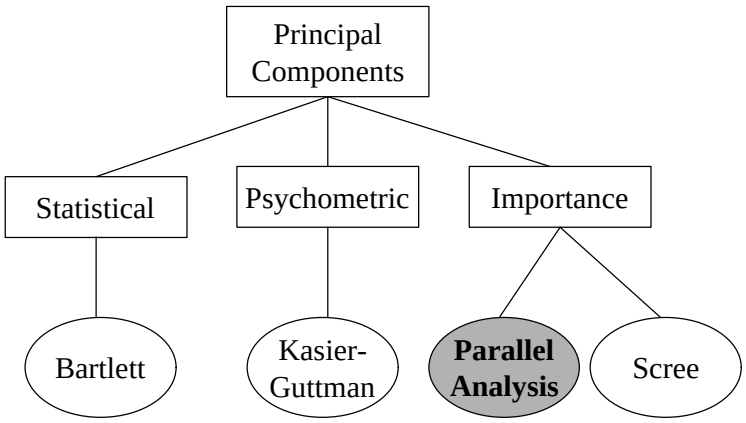

Figure 13. View of principal components analysis

Multiple hypothesis testing may be necessary when researchers are faced with databases with large numbers of variables. Figure 14 documents several adjustments to the experimentwise error rate. The estimation of the experimentwise error rate is an area of concern. Adjustments of the experimentwise error rate can occur with an overall F test

that can be followed with multiple comparison procedures such as Dunnet's C, Tukey's HSD, or the Fisher-Hayter test (Tukey 1953; Dunnet 1980; Hayter 1986). In addition, the Bonferroni adjustment can be implemented. The *p*-value approach provides an adjustment for the experimentwise error rate that is neither too liberal nor too conservative. It also provides a method for interpreting multiple statistical tests by the use of a clustering procedure.

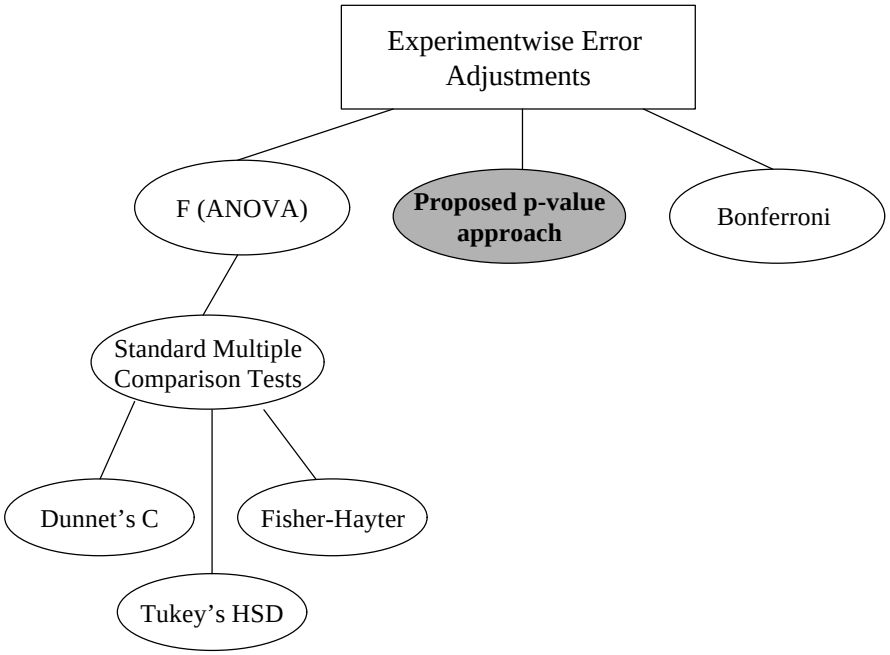

Figure 14. Multiple hypothesis testing

Research Questions

This research study involves the investigation of six research questions that deal with the issues of determining the dimensionality of data and exploring relationships in databases. The investigation of these questions help to create new methods of accurately

identifying the number of principal components to retain and addresses issues concerning the interpretation of multiple significance tests that may be associated with applications in KDD.

Research Question 1

Can an improved regression equation be found for predicting the mean eigenvalue? Would this improved regression equation require a table with an extensive list of coefficients? How does this compare to the two previously recommended regression equations?

Motivation for Research Question 1

The development of regression equations has helped researchers to implement the parallel analysis procedure (Montanelli and Humphreys 1976; Allen and Hubbard 1986; Lautenschlager, Lance, and Flaherty 1989; Longman et al. 1989). Longman et al. (1989) compared their regression approach to Allen and Hubbard (1986)'s regression models. Their study showed that the LCHF model is more accurate than Allen and Hubbard's model. Lautenschlager, Lance, and Flaherty (1989) also published a regression model (LLF), but their model has not been extensively compared to Longman et al. (1989).

The previously proposed regression models have been questioned with regard to their accuracy. In addition, large tables of coefficients are needed with the Longman et al. (1989) procedures. Furthermore, the LLF equation is an iterative process and therefore errors in the prediction could be compounded.

The Allen and Hubbard equation is examined in regards to anomalies by Cota et al. (1991). An anomaly is defined at the onset of degeneracy in the prediction of the eigenvalues. Degeneracy occurs when successively larger eigenvalues are predicted when in fact, the actual eigenvalues are decreasing. Further examination showed that, within the range of N and p used to construct their respective equations, the LCHF equation does not incur degeneracy while the LLF equation does show anomalies.

The previously proposed regression equations only cover data sets with 1000 observations for LLF and 500 observations for LCHF. These also only use 5 to 50 variables. If an organization has a data set that contains more than 50 variables, then a researcher would have to look to other procedures for determining the number of components to retain. The new regression equation is designed to extend the values of N and p used in the previous regression equations.

Research Question 2

Can an asymptotic prediction method be used to accurately predict mean eigenvalues? How does this method compare with the previously recommended regression equations and with the new regression equation?

Motivation for Research Question 2

Many papers in the literature have analytically investigated the distribution of eigenvalues of a random matrix. Anderson (1963) determined that the sample eigenvalues are asymptotically normal and asymptotically independent. This holds for eigenvalues from

a sample covariance matrix under the assumption that the eigenvalues of the population matrix are positive, distinct, and nonzero.

Glorfeld (1995) concluded from 5000 replications that the empirical distribution of the eigenvalues is approximately normal for the sample sizes he considered. The approach used in this paper is first to assume that the sample eigenvalues are approximately distributed normal with mean equal to one and variance determined by incorporating the change in variance for different values of N and p. Once the variance is estimated, the eigenvalues are estimated by using tables of normal order statistics and the distribution theory of percentiles. Since the tables of normal order statistics are fairly accessible to most researchers, this method has the advantage of being accessible and easy for researchers to calculate.

Research Question 3

Can a neural network prediction model be used to accurately predict mean eigenvalues? How does this method compare with the regression and asymptotic methods?

Motivation for Research Question 3

Neural networks – including BPNN – have been used in many business applications: airline security control (Brody 1990), stock portfolio management (Ruggiero 1994) and credit card fraud (Norton 1994). The BPNN model has been compared with other statistical techniques in applications such as forecasting and has performed well (Yi, Mitchell, and Prybutok 1996). Schalkoff (1997) stated that problems dealing with flawed,

missing, contradicting, fuzzy, or probabilistic data are suitable for artificial neural networks. These problems are characterized by some or all of the following: a high-dimensional problem space; complex, unknown, or mathematically intractable interactions between problem variables; and a solution space that may contain a unique solution.

Neural networks have many of the same underlying ideas as statistical regression analysis and can be thought of as a nonparametric regression method (Warner and Misra 1996). Neural networks can be useful when the functional relationship between the independent and dependent variables are not know. Since neural networks are data dependent, their performance often increases with sample size (Warner and Misra 1996). Regression analysis may perform well with smaller sample sizes and when theory or experience indicates an underlying relationship (Warner and Misra 1996). Since the published regression procedures have performed relatively well, it is reasonable to think that a neural network model has the potential to perform at least as well. Neural networks may outperform regression techniques due to their ability to fit nonlinear relationships.

Research Question 4

Can an improved regression equation to predict the 95^{th} percentile eigenvalue be formulated? Can a neural network topology be found which will be a viable approach to predicting the 95^{th} percentile eigenvalue?

Motivation for Research Question 4

The regression equation proposed by Longman et al. (1989) requires the use of extensive tables of coefficients. This study investigates the creation of a regression equation similar to the equation developed in Research Question 1. In addition, a neural network model is also investigated. The implications for using a higher percentage such as the 90^{th} or 95^{th} percentile is to produce a slightly more conservative number of components to be retained than using the mean eigenvalue approach. More conservative numbers would help to prevent the overextraction of factors. Overextraction has been reported to cause a negative effect by an overall degradation of the true component scores (Fava and Velicer 1992). They determined that the effects are accentuated if saturation is low or for small sample sizes. More study is needed to determine the effect of using a higher percentile because Wood, Tataryn, and Gorsuch (1996) have determined that overextraction is preferable to underextraction, assuming that no factor splitting occurs and that the false factors are eventually eliminated.

Research Question 5

In considering large numbers of independent variables, can a method be implemented to determine the number of true null hypotheses? Will a new method using a mixture of beta distributions be useful in determining the distribution of the *p*-values in studies which result in multiple hypotheses that may in fact be too numerous for traditional experimentwise error controlling procedures to perform satisfactorily?

Motivation for Research Question 5

When performing exploratory analysis researchers may need less stringent control over their experimentwise error rate. While a researcher may think that the Bonferroni adjustment is a reasonable approach, this procedure does not take into account an estimate of the true number of null hypotheses. The Bonferroni adjustment assumes that all of the statistical tests are coming from true null hypotheses. In the case of using large numbers of variables, even if a planned experiment only has one a priori hypothesis there may be hundreds or thousands of statistical tests.

To address the problem of controlling for experimentwise error, Schweder and Spjotvoll (1982) propose an approach using a *p*-value plot of cumulative *p*-values. *P*-values are the observed significance probabilities. Their idea was simple: the cumulative number of statistical tests plotted with the *p*-values that correspond to true null hypotheses forms a straight line and the point at which *p*-values mostly corresponde to false alternative hypotheses is where the *p*-values deviate from the line. Their approach of visually selecting a deviation point on the graph provides an estimate of the number of the true null hypotheses. Schweder and Spjotvoll (1982) state that their technique is intended for informal inference and that it is difficult to make exact probability statements.

To make Schweder and Spjotvoll (1982)'s approach less subjective, Parker and Rothenberg (1988) have proposed a density function with a mixture of betas to model the *p*-values. The EM algorithm is used to obtain maximum likelihood estimates of the parameters. Several examples of real-world data taken from the medical area were used in their paper to illustrate this approach.

While the mixture approach to obtaining information about the distribution of *p*-values for the null and alternative hypotheses appears to offer a fresh look at how to control for type I error, very little attention has been given to this procedure in the literature, particularly in business problems. Parker and Rothenberg (1988) did not examine the accuracy of the parameter estimates via a simulation study where estimates of the parameters can be manipulated. The purpose of this study is to assess the performance of their approach by using a simulation study with data sets consisting of *p*-values generated by beta distributions and with data sets consisting of *p*-values generated by t and noncentral t distributions. The data sets from the central and noncentral t distributions would correspond to *p*-values from popular statistical tests. The distribution of these *p*-values may only have an approximate beta distribution.

Research Question 6

Can the number of variables be reduced in the lake data set? What is the interpretation of multiple tests conducted on this real world data set?

Motivation for Research Question 6

The techniques in this study are designed to be useful to researchers. The demonstration of the applicability of this data shows that the techniques are viable in a real world situation.

Summary of Literature Review

This chapter presents a literature review surrounding the techniques in this study and outlines the research questions. Parallel analysis has been recommended by researchers as an extraction technique, but its lack of inclusion in statistical packages and the previous difficulties of performing simulations for the random eigenvalues has caused it to be ignored in the business literature. This research extends previously proposed regression techniques used in parallel analysis as well as investigates asymptotic and neural network approaches. Furthermore, the new techniques enhance the accessibility of parallel analysis to practitioners. The extensions to the parallel analysis techniques allow for an improved approach in addressing the number of factors to retain.

Currently, the business literature has given little attention to controlling the experimentwise error rate for multiple hypothesis tests. With the advent of large databases and growing interest in data mining/knowledge discovery techniques, an approach of controlling the experimentwise error rate is beneficial. In addition, the *p*-value approach may be useful; for example, for a corporation that is searching for new relationships within their customer base and must interpret a large number of statistical tests. The next chapter outlines the research methodology used to answer the research questions.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter discusses the research methodology used in this study. The chapter begins with a discussion of the previous techniques that are examined. These include regression techniques for predicting mean eigenvalues, back propagation neural networks, and *p*-value plots and mixture models. This is followed by a discussion of the research design. The procedures used to assess and validate the new eigenvalue estimation approaches are presented next. Finally, the simulation design implemented to estimate the distribution of the *p*-values is constructed and its validation procedures are detailed.

Previous Regression Techniques

The use of regression equations has made parallel analysis more accessible. There have been three major studies that have developed regression equations to estimate the mean eigenvalues from a sample correlation matrix. Notation used in these equations are 1) N = sample size, 2) p = the number of variables, and 3) k = the ordered position of the eigenvalue (ordered from largest to smallest). Allen and Hubbard (1986) published a regression equation based on previous work by Humphreys and Montanelli (1975) and Montanelli and Humphreys (1976) to estimate the mean value of an eigenvalue. This equation is able to predict all but the last two eigenvalues for any value of p less than or equal to 48. Sample sizes of N = 30, 60, 90, 120, 240, 500, and 1000 are used in

combination with p values from 5 to 50 in steps of 5 to develop the regression model. Since p represents the number of variables in the data, it also represents the maximum number of eigenvalues. In addition, only eigenvalues such that N > 3p/2 is satisfied are included in constructing the regression model. This equation provided an R^2 value of .931 for the first eigenvalue and .998 or above for the successive eigenvalues. It is noted that several anomalies exist in this iterative regression equation (Cota, Longman, Holden, and Fekken 1991). The equation[1] below is referred to as the AH equation.

(1) $$\log(l_k) = a_k + b_k \log(N - 1) + c_k \log[(p - k - 1)(p - k + 2)/2] + d_k \log(l_{k-1})$$

Equation (1) has been improved by Lautenschlager, Lance, and Flaherty (1989) to provide a more exact estimate of the first eigenvalue. Since factor analysis and psychometric texts had indicated the importance of an "adequate" p/N ratio, Lautenschlager et al. (1989) investigated the AH regression model with the additional predictor variable p/N. Sample sizes of N = 50, 75, 100, 150, 200, 300, 400, 500, 750, and 1000 are used in combination with p values from 5 to 50 in steps of 5. Combinations of N and p that did not satisfy $N \geq 3p/2$ are not examined. This equation can be used to predict up to the 48th eigenvalue. Therefore, the parameters (N and p) are similar to that in the AH study, with the main difference being the values of N that were selected. It is noted that use of this derived equation and equation (1) without Monte Carlo analyses

[1]The log indicates natural logarithm.

restricts the accuracy of the estimates of subsequent eigenvalues based on the accuracy of the first eigenvalue (Lautenschlager et al. 1989). The equation below is referred to as the LLF equation.

(2) $$\log(l_k) = a_k + b_k \log(N - 1) + c_k \log[(p - k - 1)(p - k + 2)/2] + d_k \log(l_{k-1}) + e_k p / N$$

Longman et al. (1989) present a regression equation and two tables of coefficients; one table to allow for the prediction of the mean and the other table for the prediction of the 95th percentile eigenvalue. This equation provides a noniterative alternative to the previous equations as well as providing the 95th percentile criterion eigenvalue. The use of the 95th percentile has been suggested by Longman et al. (1989) based on previous research (Skinner 1979). However, researchers need to decide the importance of using the mean and possibly overextracting the number of eigenvalues versus using a high percentile, such as the 95th and possibly underextracting eigenvalues. Their study uses sample sizes N = 50, 75, 100, 125, 150, 175, 200, 300, 400, and 500. The number of variables included were p = 5, 10, 15, 20, 25, 35, and 50. These parameter values are similar to that in Lautenschlager et al. (1989) with the main difference being that N stopped at 500. The eigenvalues for which k > 33 are not included in this study. This restriction is in place because Longman et al. (1989) determined that there is no variability in the value of the eigenvalue after the 33rd ordinal value. The resulting equation is suggested to be used for $k \leq 33$ and it is also not applicable for estimating the last two

eigenvalues. Longman et al. (1989) did not use the last two eigenvalues to obtain their regression model. The equation below is referred to as the LCHF equation.

(3) $$\log(l_k) = a_k \log(N) + b_k \log(p) + c_k[\log(N)\log(p)] + d_k$$

The equation using the 95th percentile table of coefficients will be referred to as the LCHF95 equation. Equations (2) and (3) have been shown to be accurate over the range of N and p for which they were created (Lautenschlager et al. 1989; Longman et al. 1989). A limitation to both of these equations is the dependency upon the tabled values of the coefficients. In the next chapter of this paper, regression equations that are noniterative and do not require the use of coefficient tables are presented.

Back Propagation Neural Networks

Neural networks are defined by their learning paradigm, topology, and algorithm (Bigus 1996). The back propagation network uses a supervised learning paradigm. In this paradigm, the network is fed a database that contains the problem and the solution. The network will then make a prediction and assess the error. This error is used by the algorithm to adjust the weights. A feedforward topology is used in the back propagation procedure as shown in figure 15.

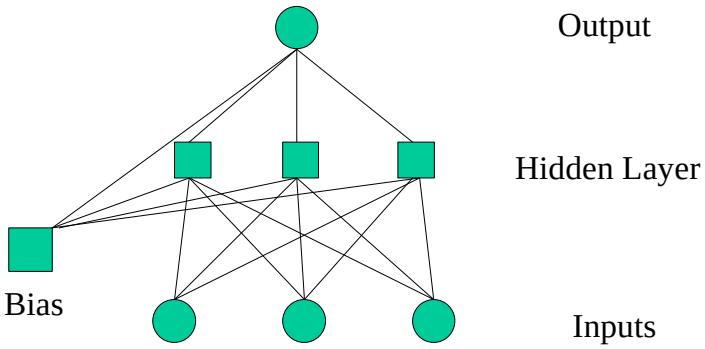

Figure 15. Back propagation neural network model

The data flows through the network in a single direction and the output is based on the current set of inputs. The back propagation neural network uses the generalized delta rule (back propagation) neural network. The generalized delta rule is a general-purpose algorithm that can be used to perform linear regression if no hidden layers are included. The addition of a single hidden layer will turn the linear neural network into a nonlinear model that is capable of performing multivariate logistic regression with multiple outputs (Bigus 1996).

The back propagation algorithm as described by Schalkoff (1997) consists of six steps. The notation used in this discussion follows:

$\underline{i}_i^p$: input pattern (vector) - the ith element of $\underline{i}^p$

$\underline{o}_j^p$: corresponding output pattern or response (vector) - the jth element of $\underline{o}^p$ where

$\underline{o}^p$ is the actual network output resulting from input $\underline{i}^p$ and the current set of weights $\underline{w}$

$\underline{w}_{ji}$: network weights (vector) - strength of interconnection from input unit i to output unit j

$\underline{t}_j^p$: desired (or target) system output (vector) - jth element of $\underline{t}^p$

$\tilde{o}_i^p$: generalized notation for ith input to neuron j to allow for hidden layers where

= o_i^p if input is output of another neuron

= i_i if input is direct input to the network

δ_j^p: sensitivity of the pattern error on the net activation of the jth unit

ε: learning rate

i: designates the ith input unit

j: designates the jth output unit

k: designates the kth hidden unit

n: number of hidden layers

p: designates the pth input/output pair

1. Initialize all weights, $\underline{w}$, in the network with random values.
2. Introduce an input (stimulus) vector to the network.
3. Propagate (feed forward) the input vector to compute the unit responses.

   $\underline{o}_j^p = f_j(w_{ij}, i^p)$

   where $f(net_j)$ is some transfer function such as the sigmoid function, i.e.,

$$f(net_j) = \frac{1}{1+e^{-net_j}} \text{ and } net_j = \sum_i w_{ji} i_i + bias_j$$

4. Compare the unit responses in the output layer with the target response.

5. Compute and propagate an error sensitivity measure backward from the output layer to the input layer.

Error measure $E^p = \frac{1}{2}\sum_j (t_j^p - o_j^p)^2$

If the E^p is less than maximum acceptable error, or if the maximum number of iterations were reached, then stop. If not, continue by adjusting the weights and returning to step 2.

Weight correction $\Delta^p w_{ji} = e d_j^p \tilde{o}_i^p$

Output units $d_j^p = (t_j^p - o_j^p) f_j'(net_j^p)$

Internal units $d_j^p = f_k'(net_k^p)\sum_n d_n^p w_{nk}$

where d_n^p is obtained from the output layer

Back propagation neural networks can be used for classification, modeling, and time-series forecasting. There are two major learning parameters that can be used to control the training process of a back propagation network. The learn rate determines whether the adjustments after each learning trial will be major or minor. The momentum is used to control possible oscillations in the weights. These might occur due to alternately signed errors. These two parameters usually produce the most impact on the training and performance of a neural network (Bigus 1996). The values chosen depend on the type of

problem that is presented. The sigmoid is the most widely used activation function for the back propagation algorithm (Cheng 1995).

P-value Plots and Mixture Models

To obtain an estimate of the number of statistical tests corresponding to true null hypothesis, Schweder and Spjotvoll (1982) recommended a graphical approach. These authors suggested visually drawing a line through a plot with the horizontal axis representing (1 - *p*-value)'s and the vertical axis representing the cumulative number of statistical tests up that point which had (1 - *p*-value)'s that were less than or equal to the (1 - *p*-value) on the horizontal axis. The line would ignore the points that started deviating from the line since these points would correspond to false null hypotheses. Their approach is based on the result that the distribution of the cumulative density function of a random variable is uniformly distributed. Their technique is performed visually and therefore is somewhat subjective.

If all of the statistical tests corresponded to true null hypotheses then the plot of (1 - *p*-value) and N_p, where N_p is the cumulative number of statistical tests, would approximate a straight line. The value of the straight line at (1 - *p*-value) = 1 would be an estimate of the number of true null hypotheses. Consider figure 16. Note that a turning point (a bend in the line) occurs around the *p*-value of 0.7. An estimate of the number of null hypotheses can be obtained by extending the straight line going through the (1 - *p*-values) of 0 to 0.7 to the (1 - *p*-value) of 1.0. The value of the straight line at this value is an estimate of the number of true null hypotheses. If a researcher wanted to use the

Bonferroni argument in controlling for the overall type I error for 500 tests in the situation illustrated by figure 16, then all tests would be given a significance level of $\alpha/250$ rather than a significance level of $\alpha/500$, representing an improvement by providing less stringent experimentwise error control.

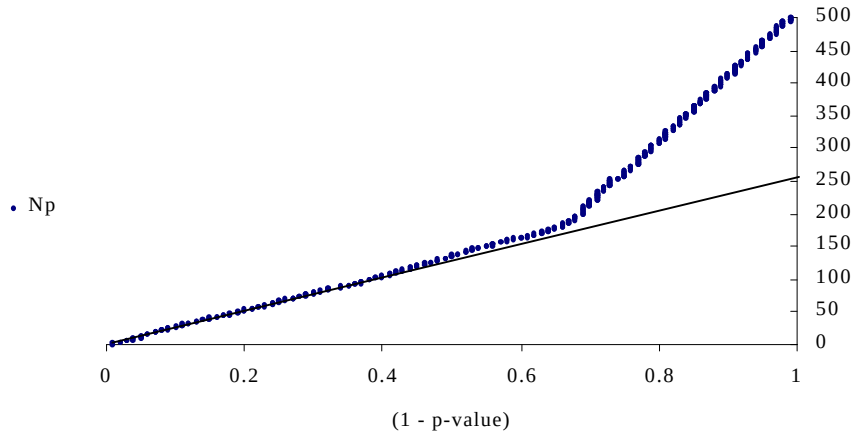

Figure 16. Plot of *p*-values

A later study by Parker and Rothenberg (1988) also illustrates the advantages of using a plotting procedure for estimating the number of hypothesis with true nulls. However, this study uses a mixture of distributions to estimate the density of the *p*-values. Parker and Rothenberg (1988) fit real data from the medical world to illustrate the potential of the procedure. No simulation study is presented in their paper and hence the true merits of approximating the density of the *p*-values with a mixture of beta distributions has not been assessed.

Hung, O'Neill, Bauer and Köhne (1997) show that the distribution of the *p*-values can be obtained for any test statistic. If the alternative hypothesis has a distribution of values which make it true, then the density function for the *p*-values may be somewhat

complicated and difficult to use in obtaining maximum likelihood estimates for the mixture density. The results in their paper are considered in using the beta distribution as an approximation for these more complicated density functions.

The mixture of density functions to model the distribution of *p*-values can be written as follows:

$$g_k(x_i) = p_0 B(1,1)(x_i) + \sum_{j=1}^{k} p_j B(r_j, s_j)(x_j)$$

where x_i represents a *p*-value, p_j represents a proportion of the *p*-values having a particular density function, and $B(r_j, s_j)(x_j)$ represents the value of the beta density function with parameters r and s. Note that a B(1,1)(x) is simply a uniform distribution. The beta distribution can be described as follows with the symbol Γ( ) representing the gamma function:

$$f(x) = (\Gamma(r+s) / (\Gamma(r)\Gamma(s)) x^{r-1}(1-x)^{s-1}$$

with shape parameters r > 0, s > 0. An obvious attraction to using the beta distribution is that the beta has a distribution where x ranges from 0 to 1. In addition, the beta distribution is

1. U shaped if r < 1 and s < 1
2. J shaped if (r - 1)(s - 1) < 0
3. otherwise is unimodal.

Since a polynomial distribution of sufficient order can be used to approximate any distribution, the sum of weighted betas is used to approximate a polynomial distribution.

Research Design for Estimating Eigenvalues

This research study investigates three new methods for estimating the mean eigenvalue: two regression approaches, two neural network approaches, and three asymptotic approaches. In addition, a regression procedure and a neural network is investigated to predict the 95th percentile of the eigenvalue. The basis for each of these new approaches is described next.

Developing a New Regression Model for Estimating Mean and 95th Percentile Eigenvalues

For the regression approach to estimating eigenvalues, model building approaches are used to determine a response surface to predict the mean and 95th percentile eigenvalues by specifying values of the sample size (N), number of variables (p), and eigenvalue position (k). These regression models include terms used in previous regression procedures as well as second order terms and their respective interactions. The new models are evaluated to determine if the extensive tables of coefficients used in prior models are necessary.

Developing a Neural Network Model for Estimating Mean and 95th Percentile Eigenvalues

For the neural network approach to estimating the mean and 95th percentile eigenvalues, a neural network architecture used in previously promising approaches to predicting real-world data is used. The most parsimonious model is retained. Since the use of neural networks has become more widespread, a neural network configuration could be

readily used by a practitioner to estimate eigenvalues by using sample size, number of variables, and position of eigenvalues as inputs. An advantage to this procedure is that tables of coefficients are not necessary as is required in the previously published regression models.

Developing Asymptotic Estimates of Mean Eigenvalues

For the asymptotic approach to estimating the mean eigenvalue, published tables of order statistics for the normal distribution are used. These order statistics for the normal distribution are available in handbooks of statistical tables. As the sample size increases, the distribution becomes normally distributed and estimates of the eigenvalues become mutually independent. Correcting for biases in the estimation, especially near the beginning and end of the ordering of the eigenvalues is addressed. The distribution theory of percentiles is also investigated to determine the role that this theory can play in obtaining better asymptotic estimates of the mean eigenvalues.

Assessment and Validation of New Eigenvalue Estimation Approaches

In this study, the three new mean eigenvalue approaches are assessed against two previously recommended regression procedures for estimating mean eigenvalues (LLF and LCHF). The AH equation is not used in the comparison since the LLF is an improved modified version of the AH equation. These comparisons occur over a variety of values for N, p, and k, with N ranging from 50 to 2000, p ranging from 5 to 80, and k ranging

from 1 to 78. The mean squared error, maximum absolute deviation, R-squares, and maximum absolute percentage error are used to determine the error in each prediction method. A frequency table of the errors is also used to classify the errors. The performance of the eigenvalue procedures is illustrated for specified values of N. In addition, a randomized block design is used for fixed values of N to perform multiple comparisons between the different techniques. The regression equation and neural network to predict the 95^{th} percentile eigenvalues are analyzed in a similar fashion. Since only the LCHF95 procedure includes an approach to predict 95^{th} percentile eigenvalues, the comparisons are made against this single equation.

Simulation Design for Estimating the Density of P-values

For the *p*-value approach, a model is investigated to determine a mixture of densities that can best describe the density of the *p*-values for the case of t-statistics and noncentral t-statistics. Finding initial estimates of the parameters of this mixture of densities is one of the contributions of this study. Good initial estimates may be important for convergence and computational efficiency. Nonlinear estimation of the parameters is used in finding maximum likelihood estimates.

Developing a Mixture Model to Estimate the Density of *P*-values

The approach in this study is to use a simulation study with data sets consisting of *p*-values generated by beta distributions and with data sets consisting of *p*-values generated by t and noncentral t distributions. The data sets from the central and noncentral

t distributions correspond to *p*-values from statistical tests. The distribution of these *p*-values may only have an approximate beta distribution. Furthermore, this study creates initial estimates for the case of a mixture of 2 components for the density function of the *p*-values.

Assessment of the Simulations

For the simulations, different combinations of sample size for the null and nonnull hypotheses are investigated. In order to determine the accuracy of the simulations, MADs and MSEs of the parameter estimates with respect to the true values of these estimates are calculated. In addition, paired t tests are used to determine whether the initial and final estimates are significantly different.

Summary of Research Methodology

This chapter presented the previous methods used to predict mean and 95th percentile eigenvalues. These include the LCHF, LLF, and LCHF95 regression equations. The motivation and methodology used to create the seven new procedures is discussed. The measures used to assess the validity of the new procedures are presented. In addition, the background of the investigation into the distribution of the *p*-values is examined. The need for the simulation is documented and the plan for performing and evaluating the simulation is discussed. The next chapter reports the results from these analyses.

CHAPTER 4

DATA ANALYSIS

This chapter discusses the data analysis procedures used in this study. First, the analysis of the eigenvalues is presented. The analysis begins with a presentation of the new regression, asymptotic theory, and neural network methods to predict mean eigenvalues. These models were created through experimentation and have been found to be reasonable. A comparison is made between the previous methods and the new models. New regression and neural network models are also presented to predict the 95^{th} percentile eigenvalue. A comparison of a single previous regression equation (Longman et al. 1989) to predict the 95^{th} percentile eigenvalue is compared with the new regression and neural network models.

The second part of the study includes a mixture approach for analyzing *p*-values. In addition, a procedure to estimate the proportion of true null and false null hypotheses is presented. A simulation study is conducted to determine the feasibility of this approach. The results from this analysis are presented. This chapter concludes with an example of real-world data showing the application of the procedures outlined in this study.

Models Used in the Eigenvalue Analysis

There have been several regression models used to predict mean eigenvalues (Allen and Hubbard 1986; Lautenschlager et al. 1989; Longman et al. 1989). The LLF

and LCHF regression equations are compared to the new methods in this study. This chapter begins with a discussion of the new regression, asymptotic, and neural network methods for predicting mean eigenvalues.

Regression Models for Predicting Mean Eigenvalues

Two models are used to create regression models for predicting mean eigenvalues. The first model is created with a reduced selection of values similar to the restrictions placed on the LCHF and LLF regression equations. The second model is created across all values of N in the Lautenschlager (1989) tables of simulated mean eigenvalues.

The first discussion is about the new model with the reduced selection of N and p values. For the new model, values of N are 50, 75, 100, 150, 200, 300, 400, and 500 and values of p are 5 to 50 in steps of 5. Performance of the model has been found to improve with the following restrictions: 1) The number of eigenvalues estimated was no greater than 34 when the number of variables are 40, 45, and 50. 2) The last two (three) eigenvalues were not estimated for 5 variables (greater than 5 variables). These restrictions are in contrast to the LCHF model which used the following restrictions: 1) The number of eigenvalues estimated was less than 33. 2) The sample size must be greater than or equal to 3p/2.

(4)

$$\log(l_k) = -0.130827 - 0.444853k - 0.008497k^2 + 0.639462\log(p) + 0.059901k\log(N) - 0.078631[\log(N)\log(p)] + 0.001488k^2\log(N) + 0.095875k\log(p) + 0.001576k^2\log(p) - 0.013331k[\log(N)\log(p)] - 0.000278k^2[\log(N)\log(p)]$$

Equation (4) is referred to as the REGEXT equation. This equation has 11 regression terms. This compares to 4 regression terms for the AH equation, 5 regression terms for the LLF equation, and 4 regression terms for the LCHF equation. While at first this may seem like a more cumbersome equation, one should note that the REGEXT equation is a function of k (the ordered eigenvalue number) and k is not incorporated into LCHF equation. An advantage to the REGEXT equation is that a table of coefficients is not necessary. For example the LCHF equation, while only having 4 regression terms, requires a table of 132 regression coefficients for eigenvalues numbered 1 through 33. The LLF equation requires a table of 240 regression coefficients for eigenvalues numbered 1 through 48.

The motivation for the REGEXT equation is that it is actually an extension of the LCHF equation. If all of the terms with k are eliminated from the REGEXT equation, then the resulting terms are from the LCHF equation. Now, if the first- and second-order terms of k and the interactions of these terms with each of the terms in the LCHF equation are included with the terms in the LCHF equation then the new model is the REGEXT equation. The use of first- and second-order terms of k in the LCHF model became a reasonable choice after experimentation with the columns of table 1 in Longman et al. (1989). A regression model has been analyzed with the dependent variable being one of the coefficients in Longman et al.'s table 1 (columns headed a, b, c, and d) and the independent variables being k, k^2, dummy variables representing either 70, 60, 50, 40, 30, or 20 (these are the number of data points used) and the interaction of these dummy

variables with k and $k^{2.}$. Since the regression model with second-order terms provides a good fit, the REGEXT model has been investigated with second-order terms of k.

The second regression model includes a more complete set of the simulated Lautenschlager (1989) mean eigenvalues. The motivation for this equation is similar to the REGEXT equation with the difference occurring in the increased number of combinations of N and p. For this model, values of N are 50, 75, 100, 150, 200, 300, 400, 500, 750, 1000, 1500, and 2000. The values of p are 5 to 50 in steps of 5 and also values of 60, 70, and 80. The restrictions on the REGEXT model are: 1) The last two eigenvalues are not estimated for the case in which there are 5 variables. 2) The last three eigenvalues are not estimated for the case in which there are more than 5 variables. This regression equation is shown in equation (5) also has 11 regression coefficients and it is referred to as REGEXTALL.

(5) $$\log(l_k) = -0.115055 - 0.335793k - 0.0078122k^2 + 0.540997\log(p) + 0.040136k\log(N) - 0.060734[\log(N)\log(p)] + 0.0012092k^2\log(N) + 0.069679k\log(p) + 0.0015843k^2\log(p) - 0.0084807k[\log(N)\log(p)] - 0.00024824k^2[\log(N)\log(p)]$$

Both of these regression equations performs well for certain combinations of N and p. This is shown later in this chapter and in the conclusions in chapter 5. The next section introduces the neural network models.

Neural Network Models for Predicting Mean Eigenvalues

With the relatively promising performance of the previous and new regression equations, the use of neural networks to predict mean eigenvalues is also explored. Two

models are created to mirror the conditions placed on the REGEXT and REGEXTALL regression equations. The same sets of N and p with their simulated values from Lautenschlager (1989) are used to create an NN and an NNALL neural network model. The neural network models are created with parameters common in the literature. The choice of parameters is defined by refinement towards a smaller root mean square. The NN model used N, p, k, and p/N as inputs with a single hidden layer with 2 neurons. The NNALL model used N, p, k and p/N as inputs with a single hidden layer with 3 neurons. Table 4 shows the input parameters and root mean square for the NN and NNALL models.

Table 4. NN and NNALL input parameters

	NN	NNALL
Learning Coefficient	1.2	1.8
Momentum	0.7	0.7
Output Coefficient	2.4	1.8
Root Mean Square	.006796	.008509

Systematic procedures are used to vary these input parameters until no improvement in the root mean square is gained. Both models use the same backpropagation options. First, bipolar inputs and the minimum/maximum table are used to store the highest and lowest values for each data field and to map the input data into values between -1 to 1. The learn rule is the delta-rule. This rule uses the differences between the actual and the desired outputs to make adjustments to the connection

weights. The transfer function transforms the weighted sum of the effective inputs to a potential output value. The sigmoid transfer is chosen because it is a commonly used, continuous, monotonic mapping of the inputs into values between 0 and 1. A random number seed of 257 is also used for both models. The performance of the neural network models is compared in a later section after the discussion of the asymptotic and order scores approaches to estimating the mean eigenvalues.

Normal Scores and Asymptotic Approaches for Predicting Mean Eigenvalues

Based on information in the literature about the distribution of the mean eigenvalues being approximately normal, several approaches are created to predict the mean eigenvalues. The introduction to these approaches begins with a discussion of the distribution of eigenvalues. After this discussion, the three approaches are described in detail.

Understanding the Distribution of Eigenvalues

Eigenvalues are also referred to in the literature as characteristic roots or latent roots. Many papers in the literature analytically investigate the distribution of eigenvalues of a random matrix. For example, Silverstein and Bai (1995) study the convergence of the empirical distribution function of eigenvalues from a random matrix and Romanazzi (1991) examine properties of the jackknife statistic for the eigenvalues of the covariance matrix. Anderson (1963) establishes the following large-sample distribution result for eigenvalues from a sample covariance matrix under the assumption that the eigenvalues of

the population matrix are positive, distinct, and nonzero. His result is as follows for sample size N, number of variables equal to p, and the ith ordered eigenvalue represented by l_k :

$$\sqrt{N}(\hat{l_i} - l_i) \rightarrow N(0,\ 2 * l_i^2)$$

$$CORR(\hat{l_i}, \hat{l_j}) \rightarrow 0, \text{ for } i \neq j$$

Thus the sample eigenvalues are asymptotically normal and asymptotically independent. The notation N(μ, σ^2), used above, represents a standard normal distribution with the first and second parameter being the mean and variance, respectively. While this result holds for large sample sizes, its accuracy needs to be investigated for samples that are not very large.

According to Anderson (1984), for random variables from a standard multivariate normal distribution (mean equal to zero and population covariance matrix $\sum$ = I), the joint density function of the ordered eigenvalues from largest to smallest, that is $f(l_1, l_2, \ldots, l_p)$ for eigenvalues from the usual sample covariance estimator multiplied by N-1, denoted as (N-1) $\hat{\sum}$, is as follows:

$$\frac{p^{.5p^2} \prod_{k=1}^{p} l_k^{.5(N-p-1)} \exp(-.5\sum_{k=1}^{p} l_k) \prod_{k<l} (l_k - l_l)}{2^{.5pN} \Gamma_p(.5N) \Gamma_p(.5p)}$$

for $l_1 \geq l_2 \geq \ldots \geq l_p \geq 0$ over the range where the density is not equal to 0. As Glorfeld (1995) emphasizes, finding the expected value of a particular eigenvalue using an analytical approach may be impractical for many researchers desiring a straightforward solution.

Glorfeld (1995) considers a nine-variable set of data with 90 observations and a 36-variable set of data with 180 observations. He generates both normal and nonnormal observations for these data sets and concluded from 5000 replications that the empirical distribution of the eigenvalues is approximately normal. The approach that is used in this discussion is first to assume that the sample eigenvalues are approximately distributed normal with mean equal to one and variance equal to some constant times 2/N. The reason for a mean of one is simply that the population correlation matrix is assumed to be the identity matrix. The reason for 2/N for a parameter in the variance formula follows from the asymptotic results of Anderson (1963). Furthermore the sample eigenvalues are assumed to be approximately independent. Next, the order statistics for the standard normal distribution as presented in the CRC Handbook of Tables for the Use of Order Statistics in Estimation (Harter and Balakrishnan 1996) are used to estimate the average value of the ith sample eigenvalue from a random correlation matrix. This order statistic must be multiplied by the standard deviation of the eigenvalue to estimate the average of the eigenvalue. The order statistics for standard normal random variables can be calculated from the following formula using numerical techniques. The notation p represents the number of variables and $X_{k,p}$ represents the kth order statistic.

$$E(X_{k,p}) = (p!/((p-k)!(k-1)!))\int_{-\infty}^{\infty} x[.5-\Phi(x)]^{k-1}[.5+\Phi(x)]^{p-k} f(x)dx$$

where $f(x) = \sqrt{1/(2p)}\exp(-x^2/2)$ and $\Phi(x) = \int_0^x f(x)dx$.

The CRC Handbook of Tables presents values of the calculations of the above formula for values of p from 1 to 100 and then for many selected values from 100 to 400. The results in this study illustrate that these tabled values of order statistics are useful in forming tables of expected eigenvalues similar to that in Lautenschlager (1989).

Average Values for Sample Eigenvalues from a Random Correlation Matrix

The average value of sample eigenvalues are presented in Lautenschlager (1989) starting with N=50. The study in this paper first compares the expected values of order statistics from a standard normal population to the average values of sample eigenvalues presented in the Lautenschlager (1989) paper. The averages presented in Lautenschlager (1989) need to be standardized in this comparison. So these averages are centered by subtracting 1 from each average and then dividing by $\sqrt{2/N}$, where $\sqrt{2/N}$ is an estimate of the asymptotic standard deviation as suggested by Anderson (1963). However, the variance of the sample eigenvalues is also a function of p as suggested by the joint density function of the eigenvalues. Intuitively and as suggested by CRC Handbook of Tables, the variance of the sample eigenvalues increases as p increases. Furthermore the variance of the sample eigenvalues also varies by the eigenvalues' order among the eigenvalues.

An appropriate constant needs to be found to adjust the value of $\sqrt{2/\mathrm{N}}$ used as an estimate of the standard deviation. A regression analysis, with no intercept term, has been performed to predict the tabled order statistics of a standard normal population with the standardized values of Lautenschlager. The constant $\sqrt{p/2}$ appears to be generally close to the regression beta coefficient. Therefore a reasonable estimate of the average of the kth sample eigenvalue is as follows:

a. Find the expected value of the kth order statistic of a random variable from a standard normal population (CRC Handbook of Tables).

b. Multiply the expected value found in part a by the constants $\sqrt{p/2}$ and $\sqrt{2/\mathrm{N}}$.

c. Add 1 to the value found in part b. If less than 0, then set the estimator to 0.

The estimator constructed from this procedure is referred to as the normal order statistic estimator (NOSE).

After viewing the differences between the values of the average eigenvalues in Lautenschlager (1989) and the estimates from the new procedure, a further adjustment is made. As illustrated in figures 17 and 18, the estimates of the new procedure were smaller in magnitude for the first few eigenvalues and smaller in magnitude for the last few eigenvalues. Figure 17 shows that for p = 20 and N=50, that the normal order statistic estimators for k between 1 and 5 and for k equal to 19 and 20 are less than the average eigenvalues obtained through simulation. The points near the ith position labeled ‘change

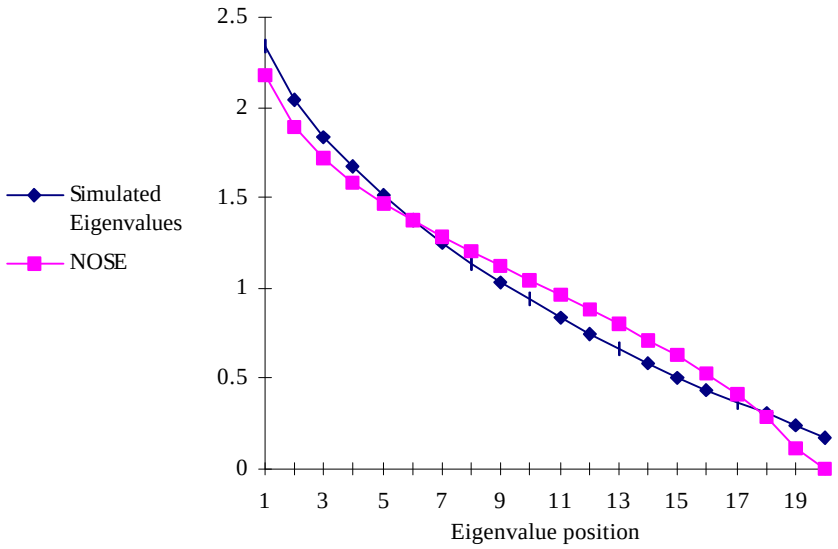

Figure 17. Graph of sample eigenvalues for NOSE and simulation approach

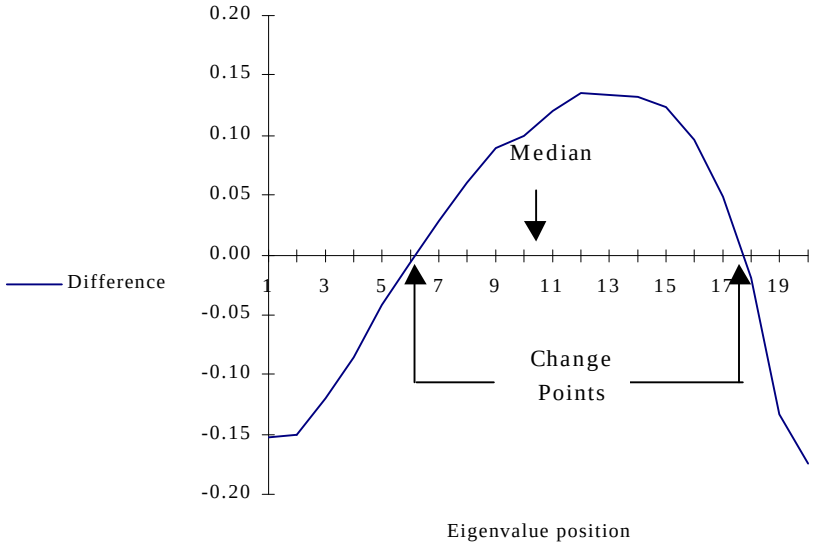

Figure 18. Difference between NOSE and simulated eigenvalues N=50, p=20

points' illustrate where the average values from simulation agree with the NOSE. Also for the eigenvalues in the middle (that is, when k is around [p+1]/2), the new procedure appeared to sometimes produce $\hat{l}_k$ s that were much larger than the average eigenvalues tabled in Lautenschlager (1989).

Therefore, the following additional adjustment is suggested after the NOSEs , represented by $\hat{l}_k$ s below, are calculated:

a. If $\hat{l}_k \geq 1.6$ then replace $\hat{l}_k$ by
$1 + (\hat{l}_k - 1) * (1 + (1/3)(p/N))$.

b. If $k \leq (p+1)/2$ and if either of the following two conditions hold:
1. $\hat{l}_k \leq 1.2$ and $p > 10$
2. $\hat{l}_k \leq 1.10$ and $p \leq 10$ then replace $\hat{l}_k$ by
$1 + [(\hat{l}_k - 1) * (1 + (1/3)(p/N))] - [(1/3)*(p/N)]$.

c. If $\hat{l}_k \geq .40$ and $k > (p+1)/2$ then replace $\hat{l}_k$ by
$(1 - (1/3)(p/N)) * \hat{l}_k$

d. If $\hat{l}_k \leq .15$ and $\hat{l}_k > 0$ then replace $\hat{l}_k$ by
$(N/p) * \hat{l}_k$.

e. If $\hat{l}_k < 0$ then replace $\hat{l}_k$ by 0.

This estimator is called the adjusted normal order statistic estimator (ANOSE). The rules used in the procedure for calculating ANOSE have been established by trial and error procedures and by viewing figures such as those presented in figures 17 and 18. It is recognized that the order statistics near the smallest and largest statistics have larger

variances than those order statistics near the middle. In addition, the term p/N was used in the adjustments since this same term appears in step b of the procedure to calculate the NOSE. Also, it is noted that p and N are the only two parameters in the joint density function of the sample eigenvalues.

To make further use of normal approximation theory, an additional procedure that does not rely on the use of tabled values of order statistics is outlined. This procedure assumes that the kth order statistic is a reasonable estimator of the (p - k + 1)/(p + 1)th percentile. This procedure to find the average of the kth sample eigenvalue is presented below.

a. Find the (p - k + 1)/(p+1)th percentile of a standard normal random variable.

b. Multiply the expected value found in part a by the constants $\sqrt{p/2}$ and $\sqrt{2/\mathrm{N}}$.

c. Add 1 to the value found in part b. If less than 0, then set the estimator to 0.

The estimator constructed from this procedure is termed the normal approximation estimator (NAE). The next section outlines the development of the models to predict 95th percentile eigenvalues.

Regression Model to Predict 95th Percentile Eigenvalues

The motivation for the 95th percentile regression equation follows the reasoning for the mean regression equation. This model is created using the values of N and p used to construct the LCHF regression equations. The simulated 95th percentile eigenvalues are taken from Cota, Longman, Holden, Fekken, and Xinaris (1993). These authors used a

Monte Carlo simulation approach to predict the eigenvalues. For the model, values of N are 50, 75, 100, 125, 150, 175, 200, 300, 400, and 500 and values of p are 5, 10, 15, 20, 25, 35, and 50. In keeping with the restrictions placed on the LCHF model, the regression equation has been developed over all combinations of N and p selected for the study such that k ≤ 33. Also, the last two eigenvalues for each value of p are not included in developing this regression equation. Equation (6) is referred to as the REGEXT95 equation.

(6)
$$\log(l_k) = 0.060911 - 0.426233\,k - 0.007825\,k^2 - 0.01718\log(\mathrm{N}) + 0.626761\,\log(p) - 0.07933\,[\log(N)\log(p)] + 0.055645\,k\log(N) + 0.001447\,k^2\log(N) + 0.088855\,k\log(p) + 0.001507\,k^2\log(p) - 0.011903\,k[\log(N)\log(p)] - 0.000283\,k^2[\log(N)\log(p)]$$

The REGEXT95 equation has 12 regression coefficients. This compares to 4 regression coefficients for the LCHF equation. As with the other regression models, while this may seem like a more cumbersome equation, the cause is the incorporation of the value of k. An advantage to the REGEXT95 equation is that a table of coefficients is not necessary. For example the LCHF equation, while only having 4 regression coefficients, requires a table of 132 regression coefficients for eigenvalues numbered 1 through 33.

Neural Network Model for Predicting 95th Percentile Eigenvalues

The results of the regression equation used to predict the 95th percentile lead to the proposition that a neural network model might also predict eigenvalues with a similar level of accuracy. For this neural network model, the combinations of N and p of the simulated

values from Cota, Longman, Holden, Fekken, and Xinaris (1993) that are used in the regression approach are used to create an NN95 neural network model. The development of this model is similar to the development of the NN and NNALL models using parameters common in the literature. The choice of parameters is defined by refinement toward a smaller root mean square. The NN95 model used N, p, k and p/N as inputs with a single hidden layer with 2 neurons. Table 5 shows the input parameters and root mean square for the NN95 model.

Table 5. NN95 input parameters

	NN95
Learning Coefficient	1.9
Momentum	.6
Output Coefficient	2.4
Root Mean Square	.009099

The input parameters have been varied until no improvement in the root mean square appeared. The NN95 model uses bipolar inputs and the minimum/maximum table, the delta-rule as the learn rule, the sigmoid transfer function, and a random number seed of 257. The performance of this model with the REGEXT95 and LCHF95 models is compared later in this chapter.

Performance of the Previous and New Mean Eigenvalue Predictors

The performance of these previous and new models to predict mean eigenvalues are assessed by several measures. First the methods are compared showing the R-squares between the predicted and SIME values across each eigenvalue position. The mean absolute prediction error (MAPE) and mean absolute deviation (MAD) are also calculated across each eigenvalue position. Tukey multiple comparisons are also performed to show the performance of the estimators within each sample size. The mean squared error (MSE), maximum absolute deviation (MAX DEV), and MAPE are used in the Tukey tests. In some instances, line graphs are also presented to further enhance the comparisons.

Performance of Models on Reduced Data Set

A reduced data set is used to compare the performance of the previous and new eigenvalue prediction methods. This is done because of the limits placed on N and p for the LCHF and LLF equations during their creation. Another reason for the reduced data set is that these are the values of N and p that are more commonly be used in traditional factor and principal components analysis. Similar limits are also placed on the creation of the REGEXT and the NN models in this study. Therefore this reduced data set used for comparison uses sample sizes of N = 50, 75, 100, 150, 200, 300, 400, and 500. The number of variables include p = 5, 10, 15, 20, 25, 35, and 50. In addition, combinations that did not satisfy $N \geq 3p/2$ are excluded from the comparisons. The restrictions on the eigenvalue position are that the last two eigenvalues are not included and $k \leq 33$. The

LCHF, LLF, NOSE, ANOSE, NAE, REGEXT, NN, REGEXTALL, and NNALL models are compared. Performance is evaluated with R-squares, MAPEs, MADs, and MAX DEVs. The analysis begins with graphs of their performance using the R-squares, MAPEs, and MADs. Tables A1-A3 include the complete set of values used to create the graphs and are included in the appendix.

Figure 19 shows that most of the procedures stay above an R-square of 0.90 for the eigenvalue position (k) lower than 24. The ANOSE performs well for all of the values of k except 33. The adjustment to the NOSE in the form of the ANOSE does increase its accuracy in most of the analyses. The LLF method performs the worst and when k > 23 the performance begins to drop off to 0.44 at k = 33. The performance of the REGEXT and REGEXTALL is similar to the LCHF regression equation. The NN and NNALL neural networks perform similarly to the LCHF and REGEXT, and perform even better when k > 17.

Figure 20 shows a comparison of the MADs using a line graph. This figure shows that the LLF equation again performs increasingly worse up to 0.43 at k = 33. The remainder of the procedures stay below 0.10 after the first eigenvalue position. The best performers are the LCHF, NN, REGEXT, and ANOSE. The NNALL performs well after k > 3. The NAE and NOSE show similar performances with the NOSE performing better on the first few eigenvalue positions.

Figure 21 shows the relative prediction error by using the MAPE. Again, the LLF begins to perform worse as k increases beyond 12 up to 75.5 when k = 33. Most of the procedures stay below the 10% mark. Once k >23 several of the procedures begin to

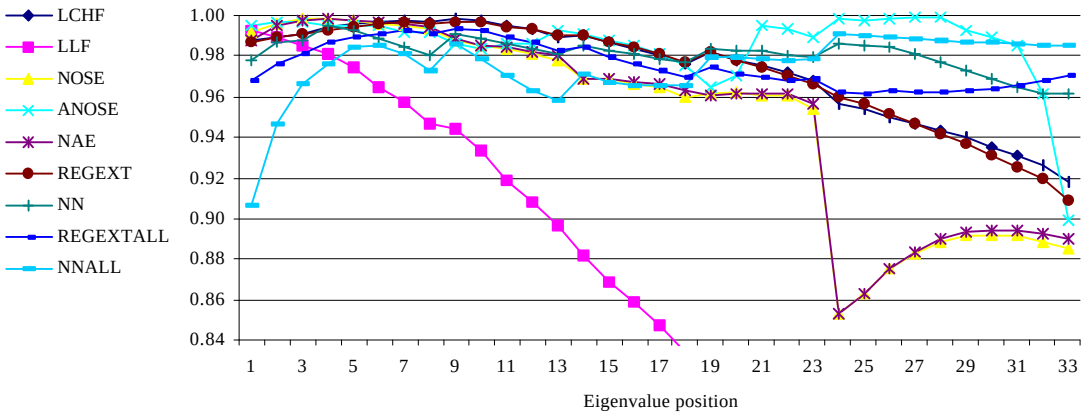

Figure 19. R-Squares for reduced data set

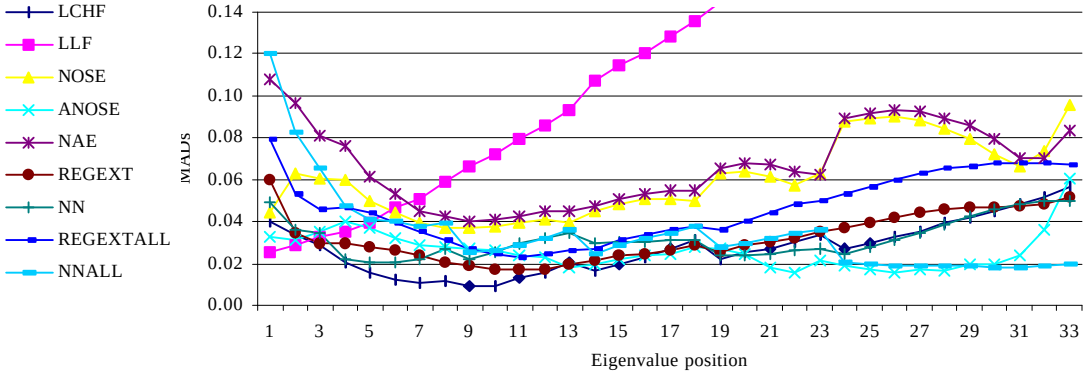

Figure 20. MADs for reduced data set

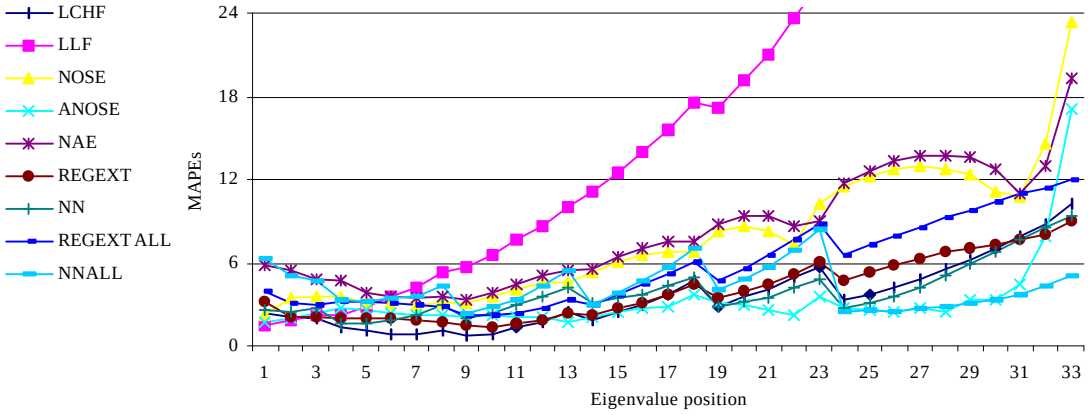

Figure 21. MAPEs for reduced data set

show a larger error. The ANOSE again is a good performer but begins to increase in error when k > 31. The LCHF, REGEXT, NN, REGEXTALL, and NNALL all perform similarly. The NNALL performs better for the last eigenvalue positions.

A second part of the analysis includes Tukey multiple comparison tests on the MSE, MAX DEV, and MAPE. The Tukey multiple comparison procedure is used because it is the most widely used procedure and has more power than the popular Scheffé test which is only recommended if some nonpairwise comparisons are of interest (Kirk 1995). These Tukey analyses are done with a randomized block design. For each value of N, the MSE (and also MAX DEV and MAPE) is calculated for a particular number of variables p across all eigenvalues. The factor in the analyses is the method of prediction (LCHF, REGEXT, etc.) and the blocks are the number of variables p. The complete tables of the MSEs, MAX DEVs, and MAPEs are included in the appendix in tables A4-A6. The output from the Tukey multiple comparisons tests is shown in tables 6-8.

The Tukey tests for the MSE in table 6 shows that for N = 75, 400, and 500 that the different techniques all perform similarly. For N = 200 and 300, the LLF technique has a higher MSE than the other techniques. For the smaller value of N = 50, the NNALL equation performs the worst. The ANOSE equation is never different from the LCHF equation in any of these tests on the MSE.

The MAX DEV Tukey tests are shown in table 7. For N = 75, 400 and 500 there is no difference between the techniques. For N = 300, the LLF equation performs worse than the other techniques. For N = 50, the NNALL performs the worst and the LCHF, ANOSE, NOSE, LLF and REGEXT perform the best. For N = 100, there are increased

Table 6. Tukey test on MSE - Reduced data set

N	Critical Distance	Alpha = 0.05			Means joined by a double line are not significantly different				
50	0.0114	ANOSE	LCHF	REGEXT	NOSE	REGEXTALL	LLF	NN	NAE	NNALL
		0.0011	0.0015	0.0029	0.0077	0.0079	0.0083	0.0109	0.0130	0.0271
75	0.0793	LCHF	ANOSE	REGEXT	REGEXTALL	NN	NOSE	NNALL	NAE	LLF
		0.0007	0.0018	0.0018	0.0040	0.0040	0.0101	0.0103	0.0141	0.0555
100	0.0090	LCHF	ANOSE	REGEXT	REGEXTALL	NN	NNALL	NOSE	NAE	LLF
		0.0012	0.0014	0.0020	0.0022	0.0027	0.0051	0.0058	0.0084	0.0133
150	0.0105	ANOSE	LCHF	REGEXT	NN	REGEXTALL	NNALL	NOSE	NAE	LLF
		0.0010	0.0010	0.0015	0.0016	0.0019	0.0021	0.0028	0.0044	0.0133
200	0.0102	LCHF	ANOSE	REGEXT	NN	NNALL	NOSE	REGEXTALL	NAE	LLF
		0.0007	0.0009	0.0011	0.0012	0.0013	0.0018	0.0022	0.0029	0.0133
300	0.0054	LCHF	REGEXT	ANOSE	NN	NOSE	NNALL	NAE	REGEXTALL	LLF
		0.0002	0.0005	0.0007	0.0008	0.0010	0.0014	0.0016	0.0022	0.0094
400	0.0206	ANOSE	NOSE	REGEXT	LCHF	NN	NAE	NNALL	REGEXTALL	LLF
		0.0005	0.0007	0.0007	0.0007	0.0008	0.0011	0.0016	0.0017	0.0166
500	0.0742	ANOSE	NOSE	NAE	NN	REGEXTALL	NNALL	REGEXT	LCHF	LLF
		0.0004	0.0005	0.0008	0.0010	0.0012	0.0017	0.0017	0.0021	0.0498

Table 7. Tukey test on MAX DEV - Reduced data set

N	Critical Distance	Alpha = 0.05				Means joined by a double line are not significantly different				
50	0.1167	LCHF	ANOSE	NOSE	LLF	REGEXT	NN	REGEXTALL	NAE	NNALL
		0.0647	0.0742	0.1205	0.1233	0.1307	0.2144	0.2199	0.2234	0.3640
75	0.2902	LCHF	ANOSE	REGEXT	NN	NOSE	REGEXTALL	NAE	NNALL	LLF
		0.0517	0.0990	0.1109	0.1236	0.1323	0.1907	0.2071	0.2606	0.2761
100	0.0900	LCHF	NN	REGEXT	ANOSE	NOSE	REGEXTALL	NAE	LLF	NNALL
		0.0618	0.0735	0.0857	0.0906	0.1024	0.1155	0.1604	0.1687	0.1858
150	0.0698	NN	LCHF	ANOSE	REGEXT	NOSE	REGEXTALL	NNALL	NAE	LLF
		0.0533	0.0573	0.0624	0.0639	0.0689	0.0725	0.0989	0.1115	0.1279
200	0.0731	NNALL	LCHF	NN	REGEXT	ANOSE	NOSE	REGEXTALL	NAE	LLF
		0.0483	0.0510	0.0515	0.0536	0.0543	0.0562	0.0704	0.0874	0.1357
300	0.0620	LCHF	REGEXT	NOSE	ANOSE	NN	NNALL	NAE	REGEXTALL	LLF
		0.0272	0.0400	0.0429	0.0439	0.0465	0.0582	0.0660	0.0686	0.1532
400	0.1649	NOSE	ANOSE	NN	LCHF	NAE	REGEXT	REGEXTALL	NNALL	LLF
		0.0352	0.0384	0.0487	0.0520	0.0543	0.0558	0.0601	0.0682	0.1950
500	0.3419	NOSE	ANOSE	NAE	NN	REGEXTALL	NNALL	REGEXT	LCHF	LLF
		0.0297	0.0341	0.0449	0.0501	0.0507	0.0727	0.0859	0.0875	0.2915

differences between the techniques. The NAE, LLF, and NNALL perform the worst. For N = 200, the LLF was significantly worse than the others and the REGEXTALL and NAE were not different from the LLF.

The MAPE Tukey tests are also performed and are shown in table 8. For N = 75 and N = 500, there is no significant difference between the techniques. For N = 300, the LLF performs significantly worse than the others. For N = 50, there was a significant difference between several of the techniques. The ANOSE performs better than the NAE and NNALL and the NNALL is significantly worse than the other methods.

Graphs of the MSE values for each method have been created for the sample sizes in the comparison. Figures 29-36 are located at the end of chapter 4. These graphs show that the LLF equation has a much larger MSE for most cases when the sample size gets larger and the eigenvalue position is higher. The larger errors of the LLF cause are truncated so that the other methods can be compared. In most cases, the remainder of the procedures perform similarly. For N = 50, all of the values have an MSE below 0.025. Once N increases, the value of the MSE for most methods stays below 0.01. For N = 300 and 400, all but the LLF stay below 0.005. The differences between the methods are most pronounced for the first and last few eigenvalues.

The MAX DEVs are also used to create graphs comparing each method. These are shown in figures 37-44 at the end of the chapter. For N = 50, 75, and 100, the values of the MAX DEV stayed below 0.5. The NNALL and NAE procedures do not perform as well for the smaller sample sizes. The LCHF performs consistently with lower values than the other procedures until the sample size reaches 400 and 500. Once N is 150 or

Table 8. Tukey test on MAPE - Reduced data set

N	Critical Distance	Alpha = 0.05				Means joined by a double line are not significantly different				
50	5.68	ANOSE	LCHF	REGEXT	REGEXTALL	NOSE	NN	LLF	NAE	NNALL
		2.88	4.42	5.10	8.31	8.40	9.80	10.07	10.13	15.91
75	13.73	LCHF	REGEXT	REGEXTALL	ANOSE	NN	NNALL	NOSE	NAE	LLF
		1.77	2.69	3.15	3.41	4.91	8.08	8.85	10.02	14.75
100	5.97	ANOSE	LCHF	REGEXTALL	REGEXT	NN	NNALL	NOSE	NAE	LLF
		2.98	3.04	3.47	3.72	4.05	5.15	6.23	7.18	11.32
150	6.15	ANOSE	LCHF	NN	REGEXT	NNALL	REGEXTALL	NOSE	NAE	LLF
		2.29	2.82	3.18	3.40	3.45	3.76	4.30	5.32	10.13
200	5.59	ANOSE	LCHF	NNALL	REGEXT	NN	NOSE	REGEXTALL	NAE	LLF
		2.06	2.33	2.72	2.91	3.09	3.31	4.07	4.23	9.65
300	3.67	LCHF	REGEXT	ANOSE	NOSE	NN	NNALL	NAE	REGEXTALL	LLF
		1.08	1.56	1.76	2.46	2.47	2.94	3.24	4.01	7.89
400	4.20	ANOSE	REGEXT	NOSE	LCHF	NN	NAE	NNALL	REGEXTALL	LLF
		1.66	1.87	2.05	2.05	2.33	2.73	3.19	3.51	7.33
500	8.41	ANOSE	NOSE	NAE	NN	REGEXTALL	REGEXT	NNALL	LCHF	LLF
		1.47	1.75	2.35	2.60	2.95	3.14	3.29	3.51	8.79

larger, all procedures except the LLF perform below 0.20 for sample sizes greater than 200. The ANOSE and NOSE perform the best for samples sizes 400 and 500. Overall there is a tendency for higher MADs as the number of variables increases within a certain sample size.

The final set of graphs for the reduced data set are using the MAPE values to compare the performance of the nine different methods. These are at the end of the chapter in figures 45-52. When N gets larger than 150, all models have MAPEs less than 8% except for the LLF equation. As indicated by the MAX DEV also, the larger numbers of variables have slightly higher errors in most cases. The NAE and NOSE have larger error values for the smaller N values, but when N is greater than 300, their performance improves. For N = 400 and 500, the ANOSE has MAPE values less than 3%.

A table of frequencies of the absolute errors is also presented. This table breaks down the number of absolute errors that fall into specific size ranges. The results are shown in table 9. For this reduced data set, the ANOSE, LCHF, REGEXT, and NN procedures all perform similarly with 99.0%, 98.6%, 98.4%, and 97.4%, respectively, of their errors below 0.1. The ANOSE and LCHF both have close to 88% of their errors that are smaller than 0.05. All of the procedures except the LLF have more than 80% of their errors below 0.1. In addition to evaluating these models with the reduced data set, a smaller selection of values from the reduced data set was examined. These comparisons are presented in the next section.

Table 9. Frequency of absolute errors using reduced data set

	LCHF	Rel. Freq.	LLF	Rel. Freq.	NOSE	Rel. Freq.	ANOSE	Rel. Freq.	NAE	Rel. Freq.
> = .3	0	0.0%	55	5.6%	0	0.0%	0	0.0%	5	0.5%
.2 to < .3	0	0.0%	81	8.2%	19	1.9%	1	0.1%	29	3.0%
.1 to < .2	14	1.4%	186	18.9%	123	12.5%	9	0.9%	155	15.8%
.05 to < .1	104	10.6%	281	28.6%	237	24.1%	112	11.4%	282	28.7%
< .05	864	88.0%	379	38.6%	603	61.4%	860	87.6%	511	52.0%
Total	982		982		982		982		982	

	REGEXT	Rel. Freq.	NN	Rel. Freq.	REGEXTALL	Rel. Freq.	NNALL	Rel. Freq.
> = .3	1	0.1%	0	0.0%	4	0.4%	8	0.8%
.2 to < .3	2	0.2%	4	0.4%	4	0.4%	14	1.4%
.1 to < .2	13	1.3%	22	2.2%	35	3.6%	64	6.5%
.05 to < .1	168	17.1%	141	14.4%	292	29.7%	147	15.0%
< .05	798	81.3%	815	83.0%	647	65.9%	749	76.3%
Total	982		982		982		982	

Performance of Models on Reduced Data Set Top p/3

In addition to the comparisons made with the reduced data set, additional comparisons are made that include the top p/3 eigenvalues from the reduced data set. The rationale for this is to determine which methods performed best on the first one-third eigenvalues since often when determining the number of principal components, only the first few eigenvalues are examined. In order to evaluate the performance, first Tukey multiple comparisons were performed similar to those performed on the reduced data set.

Table 10 shows the Tukey test on the MSE values. Table A7 in the appendix includes all of the actual values used in the Tukey test. For N = 500, there is no significant difference in the models. The LCHF and ANOSE perform the best for N = 50 and 75. For these same N values, the NAE and NNALL perform the worst using the MSE criteria. The NNALL method is significantly different from the others when N =

Table 10. Tukey test on MSE - Reduced data set top p/3

N	Critical Distance	Alpha = 0.05				Means joined by a double line are not significantly different				
50	0.0146	LCHF	ANOSE	REGEXT	LLF	NOSE	NN	REGEXTALL	NAE	NNALL
		0.0011	0.0016	0.0040	0.0050	0.0075	0.0090	0.0116	0.0185	0.0366
75	0.0143	LCHF	ANOSE	NN	REGEXT	REGEXTALL	NOSE	LLF	NNALL	NAE
		0.0011	0.0022	0.0031	0.0032	0.0075	0.0090	0.0091	0.0144	0.0173
100	0.0059	LCHF	ANOSE	NN	REGEXT	LLF	REGEXTALL	NOSE	NNALL	NAE
		0.0011	0.0019	0.0020	0.0023	0.0028	0.0030	0.0053	0.0066	0.0110
150	0.0034	LCHF	NNALL	NN	ANOSE	REGEXT	REGEXTALL	NOSE	LLF	NAE
		0.0011	0.0017	0.0017	0.0017	0.0018	0.0019	0.0025	0.0038	0.0057
200	0.0027	LCHF	NNALL	REGEXT	NN	ANOSE	NOSE	REGEXTALL	LLF	NAE
		0.0009	0.0010	0.0013	0.0013	0.0016	0.0017	0.0021	0.0032	0.0038
300	0.0013	NN	LCHF	REGEXT	LLF	NOSE	ANOSE	NNALL	REGEXTALL	NAE
		0.0003	0.0004	0.0008	0.0010	0.0010	0.0012	0.0018	0.0022	0.0022
400	0.0011	NN	LCHF	NOSE	REGEXT	LLF	ANOSE	NAE	REGEXTALL	NNALL
		0.0002	0.0004	0.0007	0.0007	0.0007	0.0010	0.0016	0.0018	0.0026
500	0.0026	NN	NOSE	ANOSE	LCHF	REGEXT	NAE	REGEXTALL	LLF	NNALL
		0.0004	0.0005	0.0008	0.0009	0.0011	0.0012	0.0014	0.0021	0.0029

400. The NN model performs significantly better than the NNALL and REGEXTALL for N = 300, 400, and 500.

The MAX DEV is also used to compare the methods. Table 11 shows the output from the Tukey test and the appendix contains a complete listing of the MAX DEV values in table A8. The Tukey test shows that for N = 500, there is no significant difference between the models. The NNALL has significantly higher deviations than the LCHF, ANOSE, LLF, REGEXT, and NOSE for N < 150. For values of N = 300 and 400, the NNALL, REGEXTALL, and NAE all have a significantly higher deviation than the other methods.

The final Tukey comparison on the reduced data set using the top p/3 values is made using the MAPE. The results are shown in table 12. For N = 200, there is no significant difference between the methods. The LCHF, REGEXT, ANOSE, NOSE, NN, and LLF are not significantly different. The NNALL performs worse for the larger values of N. The ANOSE performs better than the NOSE until the sample size is greater than 150.

Performance of Models on Full Data Set

In order to fully evaluate the performance of the new estimators of the mean eigenvalue, a set of comparisons in this study focuses on the full set of values available to compare to the predicted eigenvalues. Because of the restrictions placed on the LCHF and LLF equations as well as the REGEXT and NN models, they are not included in this

Table 11. Tukey test on MAX DEV - Reduced data set top p/3

N	Critical Distance	Alpha = 0.05				Means joined by a double line are not significantly different				
50	0.1202	LCHF	ANOSE	LLF	NOSE	REGEXT	NN	REGEXTALL	NAE	NNALL
		0.0426	0.0684	0.0860	0.1205	0.1307	0.1946	0.2199	0.2234	0.3588
75	0.1098	LCHF	ANOSE	LLF	NN	REGEXT	NOSE	REGEXTALL	NAE	NNALL
		0.0500	0.0853	0.0905	0.1084	0.1103	0.1323	0.1899	0.2071	0.2511
100	0.0713	LCHF	NN	LLF	REGEXT	ANOSE	NOSE	REGEXTALL	NAE	NNALL
		0.0524	0.0647	0.0671	0.0801	0.0801	0.1021	0.1137	0.1604	0.1758
150	0.0443	LCHF	NN	ANOSE	REGEXT	NOSE	REGEXTALL	LLF	NNALL	NAE
		0.0487	0.0519	0.0620	0.0626	0.0687	0.0690	0.0735	0.0901	0.1115
200	0.0396	NNALL	LCHF	NN	REGEXT	ANOSE	NOSE	REGEXTALL	LLF	NAE
		0.0417	0.0467	0.0496	0.0530	0.0543	0.0557	0.0611	0.0676	0.0874
300	0.0238	LCHF	NN	REGEXT	NOSE	LLF	ANOSE	NNALL	REGEXTALL	NAE
		0.0265	0.0267	0.0400	0.0426	0.0430	0.0439	0.0567	0.0587	0.0660
400	0.0249	NN	LCHF	NOSE	ANOSE	LLF	REGEXT	REGEXTALL	NAE	NNALL
		0.0203	0.0286	0.0345	0.0384	0.0391	0.0464	0.0533	0.0543	0.0682
500	0.0438	NOSE	NN	ANOSE	NAE	LCHF	REGEXTALL	LLF	REGEXT	NNALL
		0.0291	0.0313	0.0341	0.0449	0.0461	0.0465	0.0503	0.0612	0.0727

Table 12. Tukey test on MAPE - Reduced data set top p/3

N	Critical Distance	Alpha = 0.05				Means joined by a double line are not significantly different				
50	3.1222	LCHF	ANOSE	REGEXT	NOSE	NN	REGEXTALL	LLF	NAE	NNALL
		1.7981	2.0465	2.8768	3.6832	4.0443	4.0671	4.5157	5.9880	7.7079
75	3.1241	LCHF	ANOSE	REGEXT	NN	REGEXTALL	NOSE	LLF	NNALL	NAE
		1.5517	2.2824	2.6734	2.7255	2.8227	3.7891	3.7898	4.9346	5.6317
100	2.0331	LCHF	ANOSE	NN	REGEXTALL	REGEXT	LLF	NOSE	NNALL	NAE
		1.8263	2.3695	2.5254	2.5636	2.7315	3.1068	3.3059	3.5346	5.0458
150	2.0686	LCHF	NNALL	ANOSE	NN	NOSE	REGEXT	REGEXTALL	LLF	NAE
		1.9808	1.9835	2.4782	2.5598	2.5646	2.6880	2.7425	3.5006	4.1385
200	2.0563	LCHF	NNALL	NOSE	REGEXT	NN	ANOSE	REGEXTALL	LLF	NAE
		1.8609	2.0382	2.1572	2.4328	2.4340	2.4618	3.0862	3.1926	3.5877
300	1.4085	NN	LCHF	NOSE	REGEXT	LLF	ANOSE	NAE	NNALL	REGEXTALL
		1.1474	1.3306	1.8456	1.8537	1.9162	2.2842	3.0501	3.1590	3.3467
400	1.4125	NN	LCHF	REGEXT	LLF	NOSE	ANOSE	NAE	REGEXTALL	NNALL
		0.8243	1.0575	1.4999	1.5615	1.7090	2.1804	2.7647	3.0769	3.7349
500	1.9140	NN	NOSE	LCHF	REGEXT	ANOSE	LLF	NAE	REGEXTALL	NNALL
		1.3504	1.5302	1.7202	1.7702	1.9927	2.1263	2.4695	2.7752	3.9558

analysis. Therefore, the NOSE, ANOSE, NAE, REGEXTALL, and NNALL models are included in these comparisons.

This full data set includes sample size values of N = 50, 75, 100, 150, 200, 300, 400, 500, 750, 1000, 1500, and 2000. The values of p range from 5 to 50 in steps of 5 and also p = 60, 70, and 80. The one exclusion is that the last two eigenvalues are not included. In order to compare these models, R-squares, MAPEs, MADs, and MAX DEVs are created. The first look at the full data set includes graphs of R-squares, MAPEs, and MADs presented by eigenvalue position. Tables A10-A12 with the complete set of values used to create the graphs are included in the appendix.

Figure 22 shows the R-square comparisons. The NNALL performs the best and the ANOSE performs second best. All of the R-squares are above 0.81. The performance of the NAE and NOSE estimators are similar. Their values are above 0.90 for $k < 48$. The MADs are presented in figure 23. The MAD values are the smallest for the NNALL procedure. The ANOSE also performs well but for the larger eigenvalues its performance begins to decline. The NAE and REGEXTALL do not perform comparatively well for the first eigenvalue positions. The NAE, NOSE, and ANOSE begin to perform worse for the ending eigenvalue positions with the NAE reaching a MAD of 0.20 for k = 78. Figure 24 includes the MAPEs for the 5 different models. The NAE model begins to increase in error for the last few eigenvalue positions up to a very large 275% for k = 78. The remainder of the eigenvalues perform under 25% for most of the eigenvalue positions. All of the models perform similarly with prediction errors under 10% error for $k < 28$.

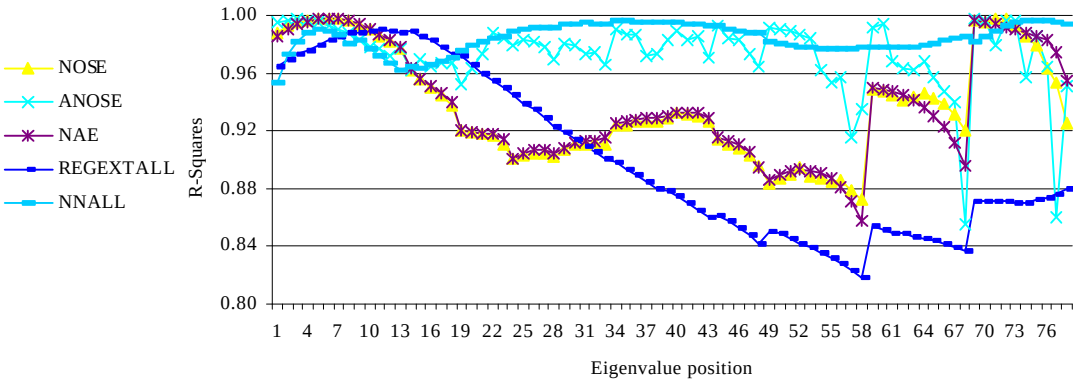

Figure 22. R-Squares for full data set

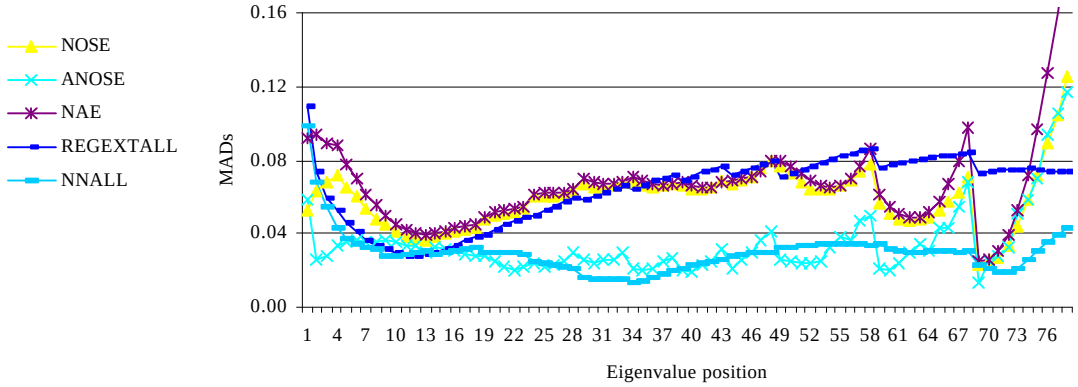

Figure 23. MADs for full data set

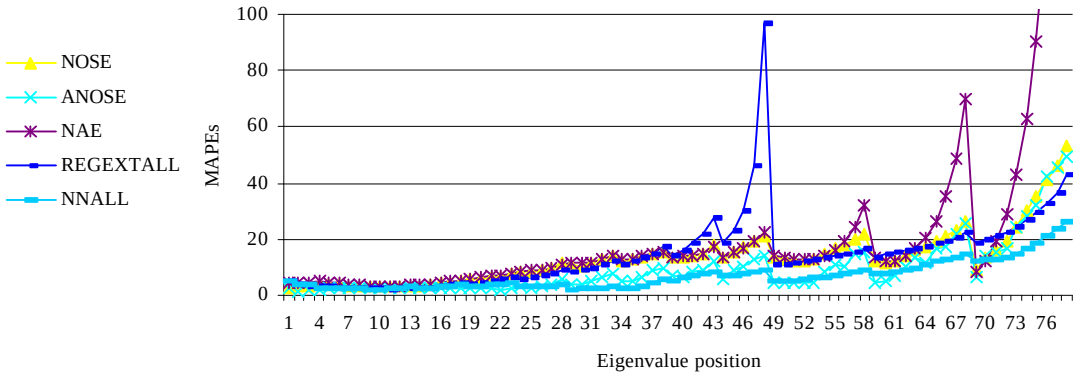

Figure 24. MAPEs for full data set

In addition to the examination of the previous graphs, Tukey multiple comparison tests are performed in a similar fashion as those used on the reduced data set. A randomized block design has been created for each value of N. The complete tables of the MSEs, MAX DEVs, and MAPEs are included in the appendix in tables A13-A15. The Tukey multiple comparison output is shown in tables 13,15 and 16.

The Tukey tests for the MSE in table 13 show that the REGEXTALLN is significantly higher than the other methods for N =400, 1500, and 2000. The NNALL is significantly higher for N=750. The ANOSE always performs in the lowest group of MSEs. For N > 50 and N < 500, the ANOSE and NNALL are not significantly different.

The next Tukey test uses the MAX DEV values and the results are in table 15. For N = 300 and 400, there are no significant differences between the methods. For N = 750, the NNALL is significantly worse than the others. For N = 1500 and 2000, the REGEXTALL is significantly worse than the others. The NNALL performs the best when N = 200. The ANOSE and NOSE are significantly better than the others when N = 50 and N = 75.

The MAPE values are used to create a final Tukey test shown in table 16. In this analysis, for N = 50, there is no significant difference between the methods. The ANOSE is in the smallest group for all values of N. The REGEXTALL MAPE values were significantly different from the other 4 methods when the sample size was 400, 500, 1500 and 2000. For N = 200, the NNALL and ANOSE have significantly smaller MAPE values.

Table 13. Tukey test on MSE - Full data set

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	0.0153	ANOSE	REGEXTALL	NNALL	NOSE	NAE
		0.005971	0.017564	0.021676	0.025437	0.032611
75	0.0082	ANOSE	REGEXTALL	NNALL	NOSE	NAE
		0.002121	0.004034	0.008631	0.012182	0.016167
100	0.0083	ANOSE	NNALL	REGEXTALL	NOSE	NAE
		0.002967	0.004115	0.004780	0.012796	0.017845
150	0.0040	NNALL	ANOSE	REGEXTALL	NOSE	NAE
		0.001617	0.001752	0.003250	0.006627	0.008737
200	0.0023	NNALL	ANOSE	REGEXTALL	NOSE	NAE
		0.001217	0.001508	0.003739	0.004251	0.005261
300	0.0014	ANOSE	NNALL	NOSE	NAE	REGEXTALL
		0.001207	0.001383	0.002293	0.002867	0.003937
400	0.0010	ANOSE	NOSE	NNALL	NAE	REGEXTALL
		0.000980	0.001493	0.001582	0.001895	0.003186
500	0.0007	ANOSE	NOSE	NAE	NNALL	REGEXTALL
		0.000793	0.001086	0.001376	0.001681	0.002274
750	0.0005	ANOSE	REGEXTALL	NOSE	NAE	NNALL
		0.000535	0.000579	0.000627	0.000803	0.001344
1000	0.0004	ANOSE	NOSE	REGEXTALL	NAE	NNALL
		0.000407	0.000443	0.000486	0.000562	0.000845
1500	0.0017	NNALL	ANOSE	NOSE	NAE	REGEXTALL
		0.000274	0.000280	0.000283	0.000350	0.004831
2000	0.0052	NOSE	ANOSE	NNALL	NAE	REGEXTALL
		0.000206	0.000211	0.000248	0.000256	0.013805

Table 15. Tukey test on MAX DEV - Full data set

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	0.1051	ANOSE	NOSE	NAE	NNALL	REGEXTALL
		0.143412	0.242255	0.361820	0.427366	0.437289
75	0.0685	ANOSE	NOSE	REGEXTALL	NAE	NNALL
		0.107729	0.159913	0.235827	0.236095	0.275254
100	0.0871	ANOSE	NOSE	NNALL	REGEXTALL	NAE
		0.148733	0.163539	0.184894	0.226723	0.251568
150	0.0541	NNALL	ANOSE	NOSE	REGEXTALL	NAE
		0.095824	0.112891	0.119179	0.128578	0.171748
200	0.0352	NNALL	ANOSE	NOSE	REGEXTALL	NAE
		0.052324	0.103990	0.105646	0.106292	0.121133
300	0.0346	NNALL	ANOSE	NOSE	REGEXTALL	NAE
		0.063788	0.083383	0.085883	0.092826	0.093844
400	0.0265	NOSE	ANOSE	NAE	NNALL	REGEXTALL
		0.068046	0.072156	0.074350	0.074579	0.080968
500	0.0225	NOSE	NAE	ANOSE	REGEXTALL	NNALL
		0.056134	0.060904	0.062934	0.068471	0.079211
750	0.0179	REGEXTALL	NOSE	NAE	ANOSE	NNALL
		0.039496	0.041828	0.046080	0.048630	0.070538
1000	0.0152	NOSE	NAE	ANOSE	NNALL	REGEXTALL
		0.036160	0.037412	0.041533	0.050878	0.055023
1500	0.0347	NNALL	NAE	NOSE	ANOSE	REGEXTALL
		0.025670	0.028072	0.031453	0.033123	0.147184
2000	0.0614	NAE	NOSE	ANOSE	NNALL	REGEXTALL
		0.024048	0.030033	0.030791	0.031123	0.245996

Table 16. Tukey test on MAPE - Full data set

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	24.018	ANOSE	NNALL	NOSE	NAE	REGEXTALL
		12.30	16.26	22.49	23.98	33.18
75	6.137	REGEXTALL	ANOSE	NNALL	NOSE	NAE
		4.116623	6.289144	7.894862	13.277493	14.136032
100	11.698	NNALL	REGEXTALL	ANOSE	NOSE	NAE
		6.993568	7.112562	7.692902	14.272840	22.997519
150	4.077	NNALL	ANOSE	REGEXTALL	NOSE	NAE
		3.498180	4.660089	5.709528	9.095906	10.466473
200	2.648	NNALL	ANOSE	REGEXTALL	NOSE	NAE
		2.817899	3.655272	6.484456	6.618448	7.187608
300	1.681	ANOSE	NNALL	NOSE	NAE	REGEXTALL
		2.868290	3.091306	4.445838	4.949471	6.464438
400	1.294	ANOSE	NNALL	NOSE	NAE	REGEXTALL
		2.508792	3.300153	3.467655	3.927384	5.603239
500	1.046	ANOSE	NOSE	NAE	NNALL	REGEXTALL
		2.189717	2.872128	3.285748	3.382375	4.624535
750	0.760	ANOSE	REGEXTALL	NOSE	NAE	NNALL
		1.777104	2.084359	2.127740	2.479825	3.002709
1000	0.761	ANOSE	REGEXTALL	NOSE	NAE	NNALL
		1.537325	1.746879	1.756441	2.065960	2.395941
1500	1.187	ANOSE	NNALL	NOSE	NAE	REGEXTALL
		1.261718	1.340439	1.370936	1.617331	5.483181
2000	2.029	NNALL	ANOSE	NOSE	NAE	REGEXTALL
		1.090134	1.105651	1.172944	1.386085	9.108421

A final comparison is made by looking at a frequency table of the size of the absolute errors for each of the models using the full data set. The NNALL and ANOSE procedures perform similarly with 96.8% and 96.1%, respectively, of their errors being below 0.1 and 81.9% and 84.6% being below 0.05. The remainder of the procedures have more than 80% of their errors being less than 0.1.

Table 17. Frequency of absolute errors using full data set

	NOSE	Rel. Freq.	ANOSE	Rel. Freq.	NAE	Rel. Freq.	REGEXTALL	Rel. Freq.	NNALL	Rel. Freq.
> = .3	38	0.7%	2	0.0%	58	1.1%	49	1.0%	15	0.3%
.2 to < .3	176	3.5%	28	0.5%	198	3.9%	82	1.6%	33	0.6%
.1 to < .2	657	12.9%	171	3.4%	713	14.0%	519	10.2%	135	2.6%
.05 to < .1	1011	19.8%	584	11.5%	1093	21.4%	1509	29.6%	741	14.5%
< .05	3218	63.1%	4315	84.6%	3038	59.6%	2941	57.7%	4176	81.9%
Total	5100		5100		5100		5100		5100	

Performance of Models on Full Data Set Top p/3

The analysis of the full data set is further refined to only include the top p/3 eigenvalue positions. This procedure is similar to the one for the reduced data set. The same five models compared using the full data set are used in this analysis: NOSE, ANOSE, NAE, REGEXTALL, NNALL. These procedures are analyzed with Tukey multiple comparison procedures. The full set of data used for the analysis of MSE, MAX DEV, and MAPE can be found in tables A16-A18 in the appendix.

Tukey tests are performed on the full data set top p/3 values using the MSE, MAX DEV, and MAPE values. For the MSE values, the results are shown in table 20. When N = 50, the ANOSE performs significantly better than the other methods. For the sample

Table 17. Tukey test on MSE - Full data set top p/3

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	0.017	ANOSE	NOSE	REGEXTALL	NNALL	NAE
		0.0084	0.0262	0.0374	0.0381	0.0433
75	0.009	ANOSE	REGEXTALL	NOSE	NNALL	NAE
		0.0022	0.0097	0.0118	0.0138	0.0213
100	0.007394	ANOSE	NNALL	REGEXTALL	NOSE	NAE
		0.003685898	0.00647164	0.009528995	0.012293304	0.019642217
150	0.003449	NNALL	ANOSE	REGEXTALL	NOSE	NAE
		0.0017	0.0022	0.0041	0.0058	0.0098
200	0.001901	NNALL	ANOSE	REGEXTALL	NOSE	NAE
		0.0013	0.0021	0.0034	0.0036	0.0061
300	0.001	ANOSE	NOSE	NNALL	REGEXTALL	NAE
		0.0019	0.0020	0.0020	0.0030	0.0034
400	0.001	NOSE	ANOSE	REGEXTALL	NAE	NNALL
		0.0014	0.0016	0.0023	0.0023	0.0027
500	0.001	NOSE	ANOSE	REGEXTALL	NAE	NNALL
		0.0011	0.0014	0.0017	0.0017	0.0029
750	0.001	REGEXTALL	NOSE	ANOSE	NAE	NNALL
		0.0007	0.0007	0.0010	0.0010	0.0023
1000	0.001	REGEXTALL	NOSE	NAE	ANOSE	NNALL
		0.0004	0.0005	0.0007	0.0008	0.0012
1500	0.000	NNALL	NOSE	NAE	ANOSE	REGEXTALL
		0.0003	0.0004	0.0005	0.0005	0.0013
2000	0.001	NOSE	NAE	ANOSE	NNALL	REGEXTALL
		0.0003	0.0003	0.0004	0.0005	0.0032

sizes N = 500 and 750, the NNALL method performs significantly higher than the others. For N = 1500 and 2000, the REGEXTALL equation performs significantly worse. The ANOSE method is always in the smallest group of MSE values.

When examining the MAX DEV values, the Tukey test shows several differences. The results are in table 18. The ANOSE and NOSE are significantly different from the others for sample sizes N = 50 and 75. When N = 1000, there are no significant differences. For N = 500 and 750, the NNALL is higher than the others. For N = 1500 and 2000, the REGEXTALL is significantly higher. The NNALL performs significantly better than the others when N = 200.

The final Tukey test includes the MAPE values. For this selection of top p/3 eigenvalue positions, the output is in table 19. When N = 50, the NAE and NNALL are significantly higher than the others. For N = 75, 100, and 150, the NAE is higher than the rest of the methods. When N = 500 and 750, the NNALL performs worse than the other four techniques. For sample size N = 2000, the REGEXTALL performs the worst. The ANOSE technique is in the smallest group of MAPE values for all sample sizes except N = 1000.

Performance of Models on Full Data Set Bottom 10%

A final analysis is made using the full data set of values. This analysis examines the bottom 10% of the eigenvalue positions. The rationale is to determine which methods could produce estimates for the lower eigenvalues which could in turn be used for various variable subset selection applications. The NOSE, ANOSE, NAE, REGEXTALL and

NNALL techniques are compared. The Tukey multiple comparison procedures are used. The MSE, MAX DEV, and MAPE values are tested. The full listing of values used for the Tukey tests can be found in the appendix in tables A19-A21.

The first examination is of the MSE values for the bottom 10% of eigenvalue positions in the full data set. The output is shown in table 20. For N = 1000, there are no significant differences. For N = 1500 and 2000, the REGEXTALL performs significantly worse than the other MSE values. The REGEXTALL performs significantly better when the sample size is 150. The REGEXTALL is in the lower group for all sample sizes except N = 1500 and 2000.

Tukey tests are also performed on the MAX DEV values. The output is shown in table 21. For N = 50 and 75, there is no difference in the techniques. For N = 200, the NNALL and REGEXTALL performs better than the others. The REGEXTALL performs the worst for the sample sizes of N =1500 and 2000. The NNALL is always included in the smallest group. The REGEXTALL is also in this group except for N = 1500 and 2000.

The MAPE values are also examined. Since some of the SIME values are close to zero for the bottom 10% of the eigenvalues, this skews the results of the MAPE calculations. Therefore, when the SIME value is less than 0.01, these values are eliminated from the analysis. This includes combinations where N = 50 and p = 40, 45 or 50 and N = 100 and p = 70 or 80. The results of this analysis are shown in table 22. For N = 50 and 75, there is no difference between the techniques. For N = 150, the REGEXTALL and NNALL performs better than the others. For N = 400, the NNALL performs the best.

Table 20. Tukey test on MSE - Full data set bottom 10%

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	0.024	ANOSE	NAE	NOSE	REGEXTALL	NNALL
		0.0046	0.0051	0.0070	0.0155	0.0350
75	0.016	REGEXTALL	NAE	ANOSE	NOSE	NNALL
		0.0045	0.0086	0.0101	0.0123	0.0219
100	0.012	REGEXTALL	NAE	NNALL	ANOSE	NOSE
		0.0022	0.0107	0.0111	0.0134	0.0148
150	0.009	REGEXTALL	NNALL	NAE	ANOSE	NOSE
		0.0015	0.0062	0.0116	0.0152	0.0170
200	0.008	NNALL	REGEXTALL	NAE	ANOSE	NOSE
		0.0031	0.0031	0.0111	0.0143	0.0157
300	0.009	NNALL	REGEXTALL	NAE	NOSE	ANOSE
		0.0012	0.0049	0.0082	0.0124	0.0151
400	0.005	NNALL	REGEXTALL	NAE	ANOSE	NOSE
		0.0011	0.0045	0.0058	0.0092	0.0094
500	0.004	NNALL	REGEXTALL	NAE	NOSE	ANOSE
		0.0012	0.0032	0.0040	0.0071	0.0076
750	0.003	REGEXTALL	NNALL	NAE	NOSE	ANOSE
		0.0004	0.0015	0.0021	0.0038	0.0044
1000	0.002	NNALL	NAE	NOSE	REGEXTALL	ANOSE
		0.0013	0.0014	0.0026	0.0027	0.0030
1500	0.010	NNALL	NAE	NOSE	ANOSE	REGEXTALL
		0.0005	0.0008	0.0015	0.0017	0.0268
2000	0.029	NNALL	NAE	NOSE	ANOSE	REGEXTALL
		0.0001	0.0005	0.0010	0.0012	0.0750

Table 21. Tukey test on MAX DEV - Full data set bottom 10%

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	0.082	ANOSE	NAE	NOSE	REGEXTALL	NNALL
		0.0961	0.0978	0.1072	0.1383	0.1640
75	0.084	REGEXTALL	NAE	ANOSE	NNALL	NOSE
		0.0853	0.1264	0.1387	0.1395	0.1501
100	0.076	REGEXTALL	NNALL	NAE	ANOSE	NOSE
		0.0559	0.1028	0.1455	0.1602	0.1640
150	0.075	REGEXTALL	NNALL	NAE	ANOSE	NOSE
		0.0512	0.0810	0.1529	0.1851	0.1859
200	0.072	NNALL	REGEXTALL	NAE	ANOSE	NOSE
		0.0556	0.0631	0.1536	0.1810	0.1850
300	0.084	NNALL	REGEXTALL	NAE	NOSE	ANOSE
		0.0328	0.0760	0.1373	0.1744	0.1914
400	0.065	NNALL	REGEXTALL	NAE	ANOSE	NOSE
		0.0365	0.0720	0.1226	0.1596	0.1602
500	0.060	NNALL	REGEXTALL	NAE	ANOSE	NOSE
		0.0393	0.0613	0.1033	0.1421	0.1426
750	0.045	REGEXTALL	NNALL	NAE	NOSE	ANOSE
		0.0257	0.0415	0.0760	0.1072	0.1119
1000	0.039	NNALL	NAE	REGEXTALL	NOSE	ANOSE
		0.0395	0.0618	0.0668	0.0882	0.0927
1500	0.042	NNALL	NAE	NOSE	ANOSE	REGEXTALL
		0.0242	0.0462	0.0673	0.0718	0.1770
2000	0.068	NNALL	NAE	NOSE	ANOSE	REGEXTALL
		0.0106	0.0384	0.0566	0.0603	0.2854

For N = 1500 and 2000, the REGEXTALL performs the worst. The NNALL is in the smallest group for all sample sizes except N = 750. The next section is about the performance of the predictors of the 95th percentile eigenvalues.

The MAPE values are also examined. Since some of the SIME values are close to zero for the bottom 10% of the eigenvalues, this skews the results of the MAPE calculations. Therefore, when the SIME value is less than 0.01, these values are eliminated from the analysis. This includes combinations where N = 50 and p = 40, 45 or 50 and N = 100 and p = 70 or 80. The results of this analysis are shown in table 22. For N = 50 and 75, there is no difference between the techniques. For N = 150, the REGEXTALL and NNALL performs better than the others. For N = 400, the NNALL performs the best. For N = 1500 and 2000, the REGEXTALL performs the worst. The NNALL is in the smallest group for all sample sizes except N = 750. The next section is about the performance of the predictors of the 95th percentile eigenvalues.

Performance of the Previous and New 95th Percentile Eigenvalue Predictors

The performance of the previous LCHF95 regression equation, the REGEXT95 regression equation, and NN95 neural network are assessed by several measures including R-squares, MAPEs, MADs, and MSEs, and MAX DEVs. Tukey multiple comparisons are also made to further enhance the conclusions. The analysis begins with graphs of the R-Squares, MADs, and MAPEs for the three different techniques. The full listing of values used to create the graphs are in the appendix in tables A22-A24.

Table 22. Tukey Test on MAPE - Full data set bottom 10%

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different			
50	37.974	ANOSE	NAE	NOSE	NNALL	REGEXTALL
		41.23	43.89	49.40	70.25	72.71
75	23.191	REGEXTALL	NNALL	NAE	ANOSE	NOSE
		31.75	38.94	45.61	47.82	51.52
100	21.051	REGEXTALL	NNALL	NAE	ANOSE	NOSE
		11.34	24.14	40.84	43.45	46.26
150	16.609	REGEXTALL	NNALL	NAE	ANOSE	NOSE
		10.21	16.78	34.61	36.72	39.77
200	14.259	NNALL	REGEXTALL	NAE	ANOSE	NOSE
		7.66	11.71	25.04	27.45	29.47
300	9.176	NNALL	REGEXTALL	NAE	NOSE	ANOSE
		3.97	12.99	15.10	18.62	19.76
400	5.934	NNALL	NAE	REGEXTALL	NOSE	ANOSE
		4.49	10.42	11.14	13.19	13.70
500	4.304	NNALL	NAE	REGEXTALL	NOSE	ANOSE
		5.05	7.75	8.62	10.07	11.06
750	2.717	REGEXTALL	NAE	NNALL	NOSE	ANOSE
		2.38	4.86	5.32	6.29	7.13
1000	2.526	NAE	NOSE	NNALL	ANOSE	REGEXTALL
		3.67	4.66	4.76	5.33	6.43
1500	4.940	NAE	NNALL	NOSE	ANOSE	REGEXTALL
		2.53	2.59	3.18	3.55	19.51
2000	8.554	NNALL	NAE	NOSE	ANOSE	REGEXTALL
		0.92	2.04	2.53	2.78	31.11

Figure 25 shows the R-squares for the techniques used to predict the 95th percentile eigenvalues. All of the techniques have R-squares above 0.90. The REGEXT95 and LCHF95 regression approaches perform similarly across all eigenvalue positions. The NN95 neural network procedure performs better than the regression procedures in predicting the final eigenvalue positions. The MADs are compared in figure 26. For most of the eigenvalue positions, the three prediction methods have similar MADs. The REGEXT95 did not perform as well for the first eigenvalue. The NN95 performs slightly worse for the eigenvalue positions 8 to 16. The MAPE values are also considered and the comparison graph is shown in figure 27. The REGEXT95 again performs worse for the first eigenvalue, but after the first it has values similar to the LCHF95. Once k > 19, the NN95 procedure performed better than both regression equations.

Tukey multiple comparison procedures are used to evaluate the MSE, MAX DEV, and MAPE as is done in the comparisons for the mean eigenvalues. The complete set of values used to calculate the tests are included in the appendix in tables A25-A26. Table 23 shows the Tukey test for the MSE values. For N = 100, 125, 150, 175, 200, 400, and 500, there is no significant difference between the three methods. The NN95 procedure does perform worse than the LCHF95 for N = 50, 75, and 300. In table 24 the analysis of the MAX DEV values is presented. For N = 100, 125, 150, 175, 200, 300, and 400, there is no significant difference between the methods. For N = 50 and 75, the LCHF95 procedure performs better than the other methods. For N = 500, the NN95 procedure performs significantly better than the REGEXT95 method.

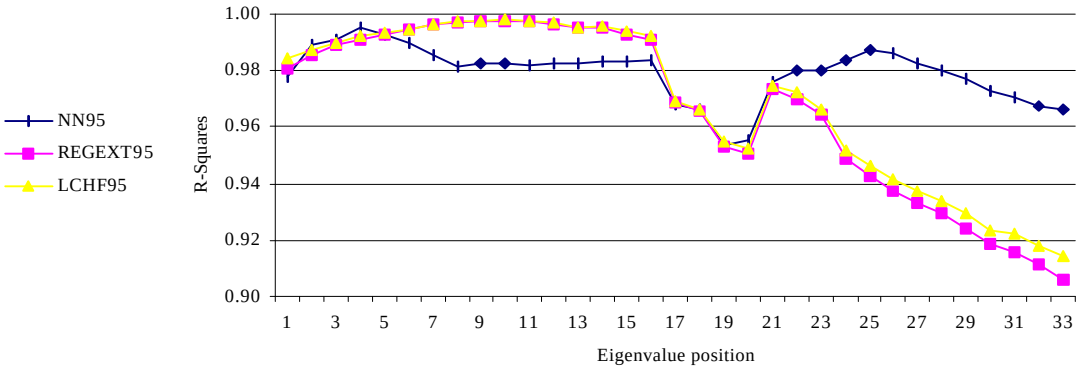

Figure 25. R-Squares for 95th percentile

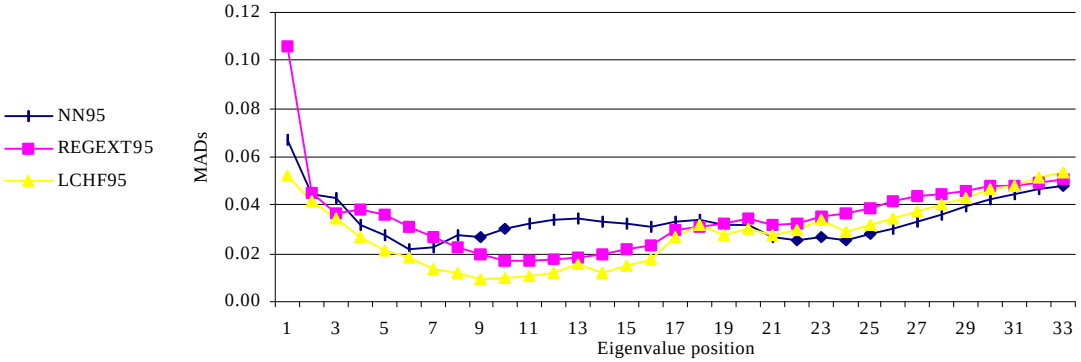

Figure 26. MADs for 95th percentile

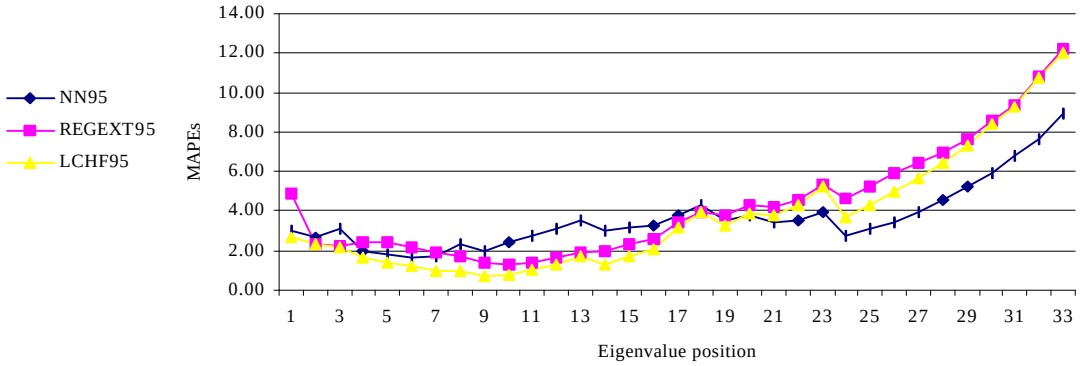

Figure 27. MAPEs for 95th percentile

Table 23. Tukey test on MSE - 95th percentile

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different	
50	0.006	LCHF95	REGEXT95	NN95
		0.0042	0.0100	0.0118
75	0.003	LCHF95	REGEXT95	NN95
		0.0007	0.0032	0.0046
100	0.003	LCHF95	REGEXT95	NN95
		0.0007	0.0022	0.0031
125	0.002	LCHF95	REGEXT95	NN95
		0.0009	0.0020	0.0024
150	0.002	LCHF95	REGEXT95	NN95
		0.0018	0.0023	0.0024
175	0.001	LCHF95	REGEXT95	NN95
		0.0008	0.0014	0.0017
200	0.001	LCHF95	REGEXT95	NN95
		0.0007	0.0011	0.0014
300	0.001	LCHF95	REGEXT95	NN95
		0.0002	0.0006	0.0010
400	0.001	LCHF95	REGEXT95	NN95
		0.0006	0.0008	0.0009
500	0.001	NN95	LCHF95	REGEXT95
		0.0012	0.0018	0.0019

Table 24. Tukey test on MAX DEV - 95th percentile

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different	
50	0.110	LCHF95	NN95	REGEXT95
		0.1329	0.3136	0.3339
75	0.088	LCHF95	NN95	REGEXT95
		0.0550	0.1585	0.1761
100	0.069	LCHF95	NN95	REGEXT95
		0.0535	0.0991	0.1202
125	0.056	LCHF95	NN95	REGEXT95
		0.0571	0.0665	0.0966
150	0.033	NN95	LCHF95	REGEXT95
		0.0761	0.0864	0.0884
175	0.038	LCHF95	NN95	REGEXT95
		0.0552	0.0590	0.0752
200	0.031	LCHF95	NN95	REGEXT95
		0.0498	0.0542	0.0661
300	0.027	LCHF95	NN95	REGEXT95
		0.0313	0.0571	0.0582
400	0.018	LCHF95	NN95	REGEXT95
		0.0484	0.0501	0.0661
500	0.037	NN95	LCHF95	REGEXT95
		0.0536	0.0782	0.0918

The final comparison is made with the MAPE values. The output is shown in table 25. For N = 50, 100, 125, 150, 175, 400, and 500, there is no significant difference between the methods. For N = 75 and 200, the LCHF95 procedure performs significantly better than the NN95 procedure. For N = 300, the NN95 procedure performs worse than the other two procedures. For many of the sample sizes, the three different measures of error are not significantly different across the three prediction techniques.

Summary of Performance of Eigenvalue Prediction Methods

The first part of this chapter presents 7 new methods for predicting mean eigenvalues and 2 new methods for predicting 95th percentile eigenvalues. These methods have been motivated by previous methods and by the literature on the distribution of the mean eigenvalues. The methods have been compared based on their R-squares, MADs, MAPEs, MSEs, MAX DEVs, and frequencies of errors. The second part of this chapter presents the analysis and results from the simulation of the density of *p*-values.

Estimating the Density of P-values

The problem of the type I error in multiple tests or multiple steps in various statistical procedures has caused the examination of ways to control this error. The investigation of the distribution of *p*-values can be useful in controlling the type I error. Approaches by Schweder and Spjotvoll (1982) and Parker and Rothenberg (1988) have been explained in chapters 2 and 3. In the next section, these approaches are introduced

Table 25. Tukey test on MAPE - 95th percentile

N	Critical Distance	Alpha = 0.05	Means joined by a double line are not significantly different	
50	4.608	LCHF95	REGEXT95	NN95
		5.36	5.79	6.52
75	1.838	LCHF95	REGEXT95	NN95
		1.50	2.68	4.25
100	2.205	LCHF95	REGEXT95	NN95
		2.10	3.10	3.61
125	2.018	LCHF95	REGEXT95	NN95
		2.54	3.32	3.48
150	1.646	LCHF95	REGEXT95	NN95
		2.93	3.46	3.61
175	1.272	LCHF95	REGEXT95	NN95
		2.34	2.93	3.21
200	0.929	LCHF95	REGEXT95	NN95
		2.11	2.63	3.06
300	0.970	LCHF95	REGEXT95	NN95
		1.00	1.49	2.53
400	1.164	LCHF95	REGEXT95	NN95
		1.80	1.89	2.18
500	1.501	NN95	REGEXT95	LCHF95
		2.35	3.06	3.18

again and the simulation study to test the Parker and Rothenberg (1988) procedure is explained. In addition, initial estimates for the proportions in the case of a mixture of 2 densities are presented. In the following two sections, the simulation study and the results are presented.

Approaches for Analyzing *P*-values

The Schweder and Spjotvoll (1982) graphical approach to obtain an estimate of the number of statistical tests corresponding to true null hypothesis is an informal approach. Parker and Rothenberg (1988) expanded on the work by Schweder and Spjotvoll (1982) to make it less subjective. They propose using a mixture of betas to model the density of *p*-values. A simulation was not done to test their approach. Therefore, this study includes a simulation as well as a procedure for finding initial estimates for the case of a mixture of 2 densities for the density function of the *p*-values.

The simulation study consists of data sets with *p*-values generated by beta distributions and with data sets consisting of *p*-values generated by t and noncentral t distributions. The data sets from the central and noncentral t distributions would correspond to *p*-values from statistical tests. The distribution of these *p*-values may only have an approximate beta distribution. Furthermore, this study proposes initial estimates for the case of a mixture of 2 densities for the density function of the *p*-values. An advantage to having these initial estimates is that they can be used as starting values for the parameters in a numerical optimization approach to obtaining maximum likelihood estimates of the density function.

The mixture of density functions to model the distribution of *p*-values was presented in chapter 3. For problems associated with estimating the parameters of a mixture model, see McLachlan and Basford (1988). These authors state that convergence of the estimates of a mixture density function may be very slow with algorithms such as the EM, and that poor choices of the initial estimates may exacerbate this situation. They also state that the problem of testing for the number of components in the mixture has been a very difficult problem and has not been completely resolved. The next section outlines the procedure for estimating the parameters.

Estimates of the Parameters

A procedure is presented next to obtain initial estimates of the proportions and beta parameters. To first obtain an estimate of the proportion of true null hypotheses, a sequence of estimates of Np are used. These estimates are obtained using a fraction with the numerator being the cumulative number of tests corresponding to (1 - *p*-value)s that are less than or equal to the specified value and the denominator being the (1 - *p*-value). This sequence of estimates of Np is continually obtained until two of the estimates are higher than three standard deviations of the previous (1 - *p*-value)'s. Then p_0 is estimated by Np/N where N is the total number of hypotheses. Since this paper only considers the two density case, an estimate of p_1 is simply 1 - $p_{0.}$ Next the estimates of the beta parameters r and s are obtained by using matching moments. That is, the following formulas are used.

$$\hat{r} = \bar{x} \left[\bar{x} (1 - \bar{x}) / s^2 \right] - 1$$
$$\hat{s} = (1 - \bar{x}) \left[\bar{x} (1 - \bar{x}) / s^2 \right] - 1$$

where the sample mean is the estimate of the average of the *p*-values corresponding to false null hypotheses and the s^2 is the sample variance of these *p*-values. To get these sample means and sample variances, first estimates are obtained of the sample mean and variance of the *p*-values up to the point where Np is estimated. Then the mean and variance of all of the *p*-values are obtained. From these estimates, the estimates of the sample mean and variance of the *p*-values corresponding to false null hypotheses can be obtained from formulas for the mean and variance of a mixture model with two densities.

In mixture theory, there are issues that present difficulties to practitioners wishing to correctly interpret these models in the context of an application. First, there is the problem of knowing how many distributions should be in a mixture model. Second, there is the possibility of obtaining distributions within the mixture model that may in fact be approximately represented by a mixture of distributions already used in a mixture model. The initial estimates may be useful in protecting against this problem of identifiability. Since it is possible for a beta distribution to in fact be mixture of betas, there is a potential problem with identifying and interpreting the correct distributions that can explain the distribution of the data (Rao 1992, Titterington, Smith, and Makov 1985). To understand how a beta distribution can be represented by other betas, the following example is presented. A beta distribution with r and s assumed to be integers can be defined as follows:

$$B(r, s)(x) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1-x)^{s-1}$$ where $\Gamma(r)=(r-1)!$ if r is an integer

Now a mixture of betas $(r/(r+s))B(r+1, s) + (s/(r+s))B(r, s+1)$ can be written as

$$(r/(r+s))\frac{(r+s)!x^{r}(1-x)^{s-1}}{r!(s-1)!} + (s/(r+s))\frac{(r+s)!x^{r-1}(1-x)^{s}}{(r-1)!s!}$$

with the coefficients r/(r+s) and s/(r+s) representing the probability of the associated distribution occurring. This mixture simplifies to

$$\frac{(r+s)!x^{r-1}(1-x)^{s-1}}{(r-1)!(s-1)!} = B(r,s)(x)$$

If the estimate obtained from the optimization program for determining estimates of the beta parameters is much different from the initial estimates proposed in this study, then the obtained beta distributions should be further examined to assess the uniqueness of the beta distributions. A discussion of the simulation study is presented next.

Simulation Study

In this study, (1 - *p*-values) are generated by a beta distribution with an r = 2 and s= 6 to represent *p*-values corresponding to false null hypotheses. In addition, *p*-values are generated by a central t distribution used to test the zero-mean null hypothesis with 25 degrees of freedom and from the noncentral t distribution with the same number of degrees of freedom and noncentrality parameter equal to 1 to represent *p*-values from a false null hypothesis. A uniform distribution is always used to generate *p*-values

corresponding to true null hypotheses. Six different combinations of sample size are used for the number of true and false null hypotheses and these are as follows:

Sample Size Notation	True Null Hypothesis Sample Size	False Null Hypothesis Sample Size
200_200	200	200
300_100	300	100
500_500	500	500
750_250	750	250
1000_1000	1000	1000
1500_500	1500	500

All estimates of the parameters of the mixture are obtained from 50 replications of the experimental situation. If parameter estimates are not reasonable, possibly due to convergence difficulty, these estimates are deleted and are listed in the appendix in table A28. The nonlinear optimization procedure in SAS is used to find the maximum likelihood estimates of the parameters.

Results of the Simulations

To understand how accurate the estimates of the mixture model are for the distribution of the *p*-values, the mean and standard deviation of the estimates, the MSE and MAD estimates (for estimating the deviation from the true parameter values), and paired t tests on initial and final parameter estimates are presented. The notation “init” , “final”, and “real” are used with the parameters in the tables in this section to denote the initial estimates of the parameters using the technique of estimation outlined, the estimates

obtained after a nonlinear optimizing program is used in SAS (proc nlp), and the estimates that the parameters have in the true model.

Table 26 shows the estimates of the relative proportions of true and false null hypotheses and the estimates of the parameters of the $B(r_1, s_1)$ density function for the case where the (1 - *p*-values) are generated from a beta distribution with r = 2 and s = 6. Both the initial estimates and the final estimates are presented along with their standard deviations. Several observations have been deleted because the data produced estimates for one or more parameters that were clearly out of line with reasonable estimates for the parameters. A researcher would probably not use these values and thus they are eliminated from the means. A list of these deleted values for each of the combinations of sample sizes is presented in the appendix in table A28.

Table 26. Means and standard deviations - B(2,6)

Sample Size		p0init	p0final	p1init	p1final	r1init	r1final	s1init	s1final
200_200	Mean	0.50	0.50	0.50	0.50	2.35	2.10	7.35	6.41
	StDev	0.06	0.05	0.06	0.05	0.88	0.36	3.37	1.44
300_100	Mean	0.72	0.74	0.28	0.26	1.94	2.44	5.21	7.70
	StDev	0.10	0.08	0.10	0.08	1.47	0.88	4.88	3.96
500_500	Mean	0.51	0.50	0.49	0.50	2.37	2.06	7.51	6.22
	StDev	0.05	0.03	0.05	0.03	0.86	0.21	3.58	0.81
750_250	Mean	0.73	0.75	0.27	0.25	1.99	2.06	5.92	6.28
	StDev	0.07	0.04	0.07	0.04	1.25	0.44	5.10	1.80
1000_1000	Mean	0.50	0.50	0.50	0.50	2.14	2.06	6.52	6.21
	StDev	0.04	0.02	0.04	0.02	0.54	0.18	2.13	0.63
1500_500	Mean	0.74	0.75	0.26	0.25	2.40	2.16	7.46	6.54
	StDev	0.05	0.03	0.05	0.03	1.23	0.36	4.95	1.36

The estimates for p_0 and p_1 should be either .5 or .75 depending on whether the sample combination has equal numbers of true and false null hypotheses or unequal numbers. For the combination of 300_100, the estimates are off the most. This same combination also produced the estimates of r and s of the beta density function that are off the most from the values of r=2 and s=6. For all estimates, the standard deviation of the final estimates are smaller than that of the initial estimates.

To gain further insight into the results in table 26, table 27 illustrates the MAD and MSE of the parameter estimates with respect to the true values of these estimates. As expected, the large sample size combination of 1000_1000 showed the smallest values and thus illustrates that for this large of a sample size the estimates (particularly the final estimates) are very accurate.

Table 27. MAD and MSE - B(2,6)

Sample Size		p0init - realp0	p0final - realp0	r1init - realr1	r1final - realr1	s1init - reals1	s1final - reals1
200_200	MAD	0.05	0.04	0.75	0.28	2.86	1.01
	MSE	0.00	0.00	0.88	0.14	12.98	2.20
300_100	MAD	0.08	0.06	1.13	0.74	4.11	3.07
	MSE	0.01	0.01	2.11	0.94	23.97	18.25
500_500	MAD	0.04	0.03	0.62	0.16	2.50	0.54
	MSE	0.00	0.00	0.86	0.04	14.83	0.69
750_250	MAD	0.06	0.03	0.94	0.34	3.57	1.34
	MSE	0.01	0.00	1.54	0.19	25.45	3.26
1000_1000	MAD	0.03	0.02	0.43	0.13	1.60	0.46
	MSE	0.00	0.00	0.31	0.03	4.72	0.43
1500_500	MAD	0.04	0.02	0.95	0.30	3.84	1.12
	MSE	0.00	0.00	1.63	0.15	26.14	2.09

For all estimates, the standard deviation of the final estimates are smaller than that of the initial estimates. The MSE and MAD for the initial estimates of the s parameter of the beta density function are much larger than that of the final estimates. The MSEs for the estimates of the relative proportions of the true and null hypotheses are rather small, thus indicating that either the initial or final estimates are fairly accurate.

Table 28 illustrates the results of the paired t test for testing whether the initial and final estimates are significantly different. Significant differences appear the most in the small sample combinations. For the sample size combination of 300_100, 200_200, and 500_500 all of the initial estimates for r and s for the beta density function are significantly different from that of the final estimates. At the 0.05 significance level, only for the sample size combination of 750_250 is there a significant difference between the initial and final estimates of the proportion of true and null hypotheses. In general, the results in this table illustrate that in many cases the initial estimates may be almost as good as using the final estimates, particularly for the estimates of the proportions.

Tables 29, 30, and 31 give insight into the accuracy of the estimates of the mixture density function with data (possibly from a real-world situation) where multiple t tests are performed independently. These three tables are similar to tables 26, 27, and 28 except that the *p*-values are generated from a central and noncentral t distribution. Using just the data from the noncentral t distribution, maximum likelihood estimates for the r and s parameters of the beta distribution are found to be 0.66 and 2.45, respectively. These parameters are taken to be the realr1 and reals1 values used in table 30.

Table 28. Paired t-tests between initial and final estimates

Sample Size		p0final vs. p0init	r1final vs r1init	s1final vs. s1init
200	Correlation	0.60	0.44	0.32
	t Stat	-0.18	-2.23	-2.06
	two-tail *p*-value	0.86	0.03	0.04
300_100	Correlation	0.64	0.08	-0.08
	t Stat	1.82	2.09	2.65
	two-tail *p*-value	0.08	0.04	0.01
500	Correlation	0.73	0.47	0.55
	t Stat	-0.69	-2.80	-2.78
	two-tail *p*-value	0.49	0.01	0.01
750_250	Correlation	0.56	0.27	0.27
	t Stat	2.03	0.44	0.50
	two-tail *p*-value	0.05	0.66	0.62
1000	Correlation	0.53	0.46	0.38
	t Stat	1.02	-1.12	-1.09
	two-tail *p*-value	0.31	0.27	0.28
1500_500	Correlation	0.52	-0.01	-0.06
	t Stat	1.52	-1.29	-1.23
	two-tail *p*-value	0.14	0.20	0.22

Table 29. Means and standard deviations - Central and noncentral t distributions

Sample Size		p0init	p0final	p1init	p1final	r1init	r1final	s1init	s1final
200_200	Mean	0.59	0.58	0.41	0.42	0.73	0.72	3.32	3.44
	Stdev	0.07	0.12	0.07	0.12	0.19	0.07	1.68	1.32
300_100	Mean	0.76	0.74	0.24	0.26	0.79	0.84	2.95	5.70
	Stdev	0.09	0.18	0.09	0.18	0.37	0.26	2.46	6.33
500_500	Mean	0.57	0.58	0.43	0.42	0.66	0.69	2.68	3.21
	Stdev	0.06	0.09	0.06	0.09	0.14	0.06	1.07	1.28
750_250	Mean	0.77	0.69	0.23	0.31	0.77	0.72	3.00	3.32
	Stdev	0.07	0.21	0.07	0.21	0.29	0.07	2.04	2.39
1000_1000	Mean	0.55	0.57	0.45	0.43	0.64	0.68	2.44	2.84
	Stdev	0.05	0.05	0.05	0.05	0.08	0.03	0.67	0.55
1500_500	Mean	0.77	0.71	0.23	0.29	0.76	0.71	3.39	3.11
	Stdev	0.06	0.22	0.06	0.22	0.33	0.06	3.02	1.71

In contrast to table 26, table 29 shows that the initial estimates of the proportions of true and false null hypotheses generally have smaller standard deviations than the final estimates. Also, the standard deviations for the estimates of the r and s parameters appear to be somewhat closer than they were in table 26. For the sample size combination 300_100 the estimates of r and s for the final estimates are the farthest off, similar to the case with *p*-values generated from the beta distribution. For the sample size combination 1000_1000, the estimates of the parameters are the closest to the real values, similar to that in table 26.

From table 30, the MAD and MSE for the initial estimates of the proportions appear to be somewhat more accurate than that of the final estimates of the proportions. In addition, the initial estimates of the r and s parameters appear to be competitive with the final estimates of these parameters. In table 31, at the 0.05 significance level, the sample size combinations of 1000_1000, 1500_500, and 750_250 revealed a significant difference between the initial and final estimates of the proportions of true and false null hypotheses. This is important because the initial estimates appear to be giving somewhat more accurate estimates than the final estimates. Significant differences in the initial and final estimates for either the r or s estimates are found in three of the sample size combinations: 1000_1000, 500_500, and 300_100.

Table 30. MAD and MSE - Central and noncentral t distribution

Sample Size		p0init - realp0	p0final - realp0	r1init - realr1	r1final - realr1	s1init - reals1	s1final - reals1
200_200	MAD	0.09	0.12	1.27	1.28	2.93	2.58
	MSE	0.01	0.02	1.66	1.64	9.94	8.28
300_100	MAD	0.08	0.14	1.21	1.16	3.68	4.67
	MSE	0.01	0.03	1.59	1.42	15.27	39.25
500_500	MAD	0.08	0.10	1.34	1.31	3.38	2.92
	MSE	0.01	0.01	1.82	1.71	12.12	9.39
750_250	MAD	0.06	0.15	1.23	1.28	3.45	3.25
	MSE	0.01	0.05	1.60	1.65	13.04	12.76
1000_1000	MAD	0.06	0.08	1.36	1.32	3.56	3.16
	MSE	0.00	0.01	1.86	1.75	13.13	10.31
1500_500	MAD	0.05	0.13	1.24	1.29	3.69	3.12
	MSE	0.00	0.05	1.64	1.67	15.75	11.18

Table 31. Paired t-tests between initial and final estimates

Sample Size		p0final vs. p0init	r1final vs r1init	s1final vs. s1init
200_200	Correlation	0.52	0.12	0.17
	t Stat	-0.40	-0.21	0.39
	P(T<=t) two-tail	0.69	0.84	0.70
300_100	Correlation	0.64	0.08	-0.08
	t Stat	1.82	2.09	2.65
	P(T<=t) two-tail	0.18	0.51	0.00
500_500	Correlation	0.82	-0.06	0.05
	t Stat	1.73	1.71	2.22
	P(T<=t) two-tail	0.09	0.09	0.03
750_250	Correlation	0.65	-0.20	-0.05
	t Stat	-3.08	-1.07	0.67
	P(T<=t) two-tail	0.00	0.29	0.51
1000_1000	Correlation	0.56	0.11	0.10
	t Stat	3.39	16.21	3.39
	P(T<=t) two-tail	0.00	0.00	0.00
1500_500	Correlation	0.74	-0.07	0.33
	t Stat	-2.44	-1.08	-0.64
	P(T<=t) two-tail	0.02	0.29	0.52

Summary of P-value Density Estimation

The results of the simulation study are presented to assess the applicability of the suggested mixture of densities. Overall, the results from these tables show that for *p*-values coming from central and noncentral t distributions, the initial estimates may in many cases be viable alternatives to estimating the parameters of the mixture model.

Real World Application

A set of data collected from Lake Texoma has been used to apply these procedures to real world data. The data set contains measures of the chlorophyll-a level and hyperspectral data on the measurement of energy coming out of the water. These samples have been collected manually from a boat. It is hypothesized that there is a relationship between the reflectance of electromagnetic radiation and the level of chlorophyll-a. If this can be proven true then, as an alternative to manual sampling, information technology could be used to measure the electromagnetic radiation. Eventually, the plan is to pay for commercial satellites to take these measurements and to use these measurements to predict the chlorophyll-a levels. With satellites, monitoring of the water could be done more frequently and at more locations throughout the lake. Advancements in satellite imaging by Kodak are allowing multispectral imagery to expand beyond the traditional three bands (blue, green, and red) to include a fourth band (near-infrared) on a single integrated array (Jurgens 1999). The four bands are used in this analysis.

Lake Texoma's water consists of numerous particles, wildlife, and algae. Each of these substances reflect energy from wavelengths (light visible to the eye as well as infrared light) to a certain degree. Because of the turbidity of the water, the amount of energy reflected by wavelengths from algae is difficult to measure. A well known formula for the energy of wavelengths is the following:

$$Etotal = E_a + E_t + E_r$$

where Etotal is the total amount of energy of the light or wavelengths, E_a is the energy absorbed by the algae, E_t is the amount of energy transmitted by the algae, and E_r is the amount reflected.

According to Raven and Johnson (1996), the amount of energy transmitted by algae is concentrated in the blue and red bandwidths. This energy transmitted peaks at approximately 415 nm and the a range of 400-440 for blue light. Two other definitions that will be discussed in the analysis include the red edge which is the side of the red wavelength bandwidth closest to the infrared wavelengths and the yellow edge which is the side of the red wavelength bandwidth closest the green spectrum. The green part of the spectrum may not be as useful as the other colors because plants do not transmit green.

There are political as well as economic ramifications to the study of this relationship. Therefore, the study of this data will affect the future of the lake. The current state of the lake is recreational use. But there is a federal plan in place to remove salt from the Red River which feeds into Lake Texoma in order to provide water to users further downstream. The reduction of chloride will affect the chlorophyll-a levels and

could affect life in the water. Opponents to the plan point out that potential damage could occur to Lake Texoma's multi-million-dollar striped bass fishery (Hodge 1996). Therefore, careful monitoring of the level of chlorophyll-a is warranted. There are two issues that are addressed using this data. First, can a relationship be established between the energy reflectance and chlorophyll-a? Second, are all of the stations and date/time locations useful in the analysis of the chlorophyll-a data?

In this study there are 2 large data sets investigated, one for the 0.5 and one for the 1 meter data, to try to find any relationship between the chlorophyll-a and the reflectance measurements. The data have been collected at 5 different sampling stations and on different days across a 10 month time period. The data were also collected at 0.5 and 1 meter depths and the chlorophyll-a data were taken in 10 replicates. There were 752 wavelengths ranging from 400.04 to 799.97.

The second derivative of the reflectance data is taken to transform the following the work of Atkinson, Acevedo, Dickson, and Rolbecki (1998) who found a significant correlation between chlorophyll-a and the second derivative of the reflectance data. The wavelengths can be broken into the four colors: blue (400-499 nanometers (nm)), green (500 - 599 nm), red (600 - 699 nm), and infrared (700 - 799 nm). The approach in this study is to create all possible tuples from values in each of the four colors. Within the 752 wavelength values collected, successive values such as 400.04 and 400.62 result in similar values. In order to reduce the number of tuples, every 18^{th} wavelength was selected. This results in 42 wavelength values. In addition, there are 32 station/time periods for the 0.5 meter data and 38 for the 1 meter data. The difference in these is accounted for by the

removal of data for particular dates, depths, and/or wavelengths for which there are inaccurate measurements due to cloud cover or other anomalies. The chlorophyll-a data has been matched to each date/station/wavelength combination and an average of the 10 chlorophyll-a replicates is used in the analysis.

These tuples are each used to regress chlorophyll-a on the second derivative of the reflectance data. There were 10,890 regressions for the 0.5 meter depth and 12,100 regressions for the 1 meter depth. The highest R-squares reached 0.50 for the 0.5 meter data and 0.41 for the 1 meter data. The *p*-values from these tests have been fitted with the mixture approach presented in this study. In this application, 3 betas plus the uniform are used in the mixture. There were 5,011 tests for the 0.5 meter data and 4,802 tests for the 1 meter data in the smallest *p*-value group which could correspond to those tests that show the strongest relationship between the energy reflectance and chlorophyll-a. This smallest *p*-value group has *p*-values ranging from 0.0006 to 0.282 for the 0.5 meter data and 0.001 to 0.059 for the 1 meter data.

The group considered to come from the null hypothesis contained 2,916 of the tuples for the 0.5 meter data and 525 for the 1 meter data. This null group had *p*-values ranging from 0.282 to 0.374 and 0.707 to 0.999 for the 0.5 meter data and 0.511 to 0.986 for the 1 meter data. The skip in the range of p-values in the 0.5 meter data occurs because beta distributions essentially cancel each other out and resemble the uniform distribution for 0.707 to 0.999. A summary of the number and proportion in each group is shown in table 32. In order to ascertain information from these results, these tuples have

been examined until several combinations are identified as belonging to each group. These groups are labeled the"smallest *p*-value" group and the "null" group.

Table 32. Number and proportion in groups

	0.5 Meter		1 Meter	
	Number of Observations	Proportion	Number of Observations	Proportion
Null Group	2916	27%	525	4%
Smallest *P*-Value Group	5011	46%	4802	40%

Table 33 shows the main combinations that have been identified in each group for the 0.5 meter and 1 meter data. For the 0.5 meter data, there are 5 combinations identified in the smallest *p*-value and 3 in the null group. The 1 meter data have 6 combinations in the smallest *p*-value group and 2 in the null group. These combinations are created by selecting a wavelength from each of the colors in a row. For example, combination 1 for the 0.5 meter smallest *p*-value group indicates that all tuples containing the infrared wavelength 734.90 are in this group. Combination 2 in the smallest *p*-value group for the 1 meter data indicates that all combinations that contain green wavelengths of either 585.21 or 594.87 are in this group.

In addition, frequencies of the occurrences of each wavelength in the analysis have been examined for the null and smallest *p*-value groups in each data set. Based on Table 33 and the frequency analysis, some conclusions are drawn. The 0.5 meter data will be examined first. This data could be more accurate because the extra 0.5 meter depth for the 1 meter data may be shifting the graph of the amount of reflected light due to reflectance from other substances in the water at the 1 meter depth. In the smallest *p*-

Table 33. Combinations in null and smallest *p*-value groups

		0.5 Meter				1 Meter			
Combination		Blue	Green	Red	Infrared	Blue	Green	Red	Infrared
Smallest *P*-Value Group	1	All	All	All	735	All	All	643	All
	2	435 486	All	643 652 661	All	All	585 595	All	All
	3	All	All	605	762 780	All	All	652 661	726 771
	4	All	All	698	735	404	All	All	717
	5	All	546 595	All	735 761 780	496	All	All	735
	6					All	All	671	726
Null Group	1	404 415 425 435 466 496	556 566 576 585	624 633 680 689 698	717 753 771	415 425 435 445 455 465 476	506 516 526 536 546 556	680	744 753 77` 788
	2	446 456 486	566 576 585	689 698	726	415 425 435 445 456 466 476	536 546 556	605	744 753 762 788
	3	446 456 486	566 576 585	624	726				

Note: Combinations are created by selecting one value from each color

value group, specific values of red occur most often, then blue, and then green. The red range from 642 - 661 is most predominant. The red wavelengths of 604 and 698 also appear. These two values fall into the yellow edge and red edge, respectively, mentioned earlier in this section. Red also appears The blue ranges from 442 - 456 and 470 - 491 in combination with the red range of 642 - 661 is unique to this group. The green range of 542 - 548 and 594 also occur frequently. This value of green at 594 occurs where a dip in the graph of reflectance may occur. The infrared wavelengths that occur most often are 735 and 761.

In examining the null group for the 0.5 meter data, the red ranges from 621 - 635 and 670 - 691 appear to be most frequent in being associated with regression equations that are not good predictors of chloropyll-a. In addition, infrared 752 is not a good predictor across several combinations of blue, green, and red.

The 1 meter data are also examined. The smallest *p*-value group has a high frequency of the red range from 642 - 663. This range has a dip in the graph of the energy reflectance for these wavelengths. In addition, the blue ranges from 400 - 407 and 491 - 498 also appear most frequently. The green range of 585 - 594 and the infrared wavelength of 725 also occur most often. The null group for the 1 meter data is a small group with only 525 combinations out of 12,100. Therefore, the patterns are harder to detect.

This technique can be compared to the Bonferroni approach. If the significance level is set at 0.05 and the 0.5 meter data are used, then each of the 10,890 tests would be evaluated using $\alpha = .05/10,890 = .0000005$. The smallest *p*-value for the 0.5 meter data is

0.00062. In this case, none of the tests would be considered significant and the result would be that none of the tuples are able to predict the chlorophyll-a levels.

The approach in this study using the mixture of betas allows a researcher to relax the restriction of the experimentwise error rate and using a clustering procedure based on the mixture proportions in order to discover patterns in the data. Further testing can be concentrated on the combinations identified to build models that may more accurately predict chlorophyll-a. Although the patterns discovered do not completely describe these groups, they are insights into the characteristics that are exhibited by types of blue, green, red, and infrared wavelengths that appear to have either a strong or very weak relationship with chlorophyll-a.

The lake data are also used to determine which locations and dates could be eliminated from the analysis. The 1 meter data are examined across 1029 wavelengths from 260.07 to 853.42. There are 38 station/date valid combinations. Principal components analysis is used to determine if the data could be represented with a fewer number of dimensions. A varimax rotation is used to fit the groups to the factors. Since the sample size N = 1029 and there are p = 38 variables, the ANOSE procedure is used to predict eigenvalues to use for Horn (1965)'s parallel analysis procedure. From the results of the simulation study, the ANOSE procedure performs well for similar combinations of N and p. The first four random ANOSE eigenvalues are 1.415, 1.337, 1.292, and 1.259. The first sample eigenvalues are 24.114, 6.165, 1.680, and 1.128. Therefore, 3 dimensions should be retained.

The first dimension consists of all observations in November and January. In addition, for station 3, September, the beginning of May and the end of June are also included. For station 9, there is an additional observation in the beginning of June. For station 17, the observations are from March, April, and the beginning of May. For station 22, the months include September, March and the beginning of June. Finally, for station 24, March, April and the beginning of June are also included. The second dimension consists of March, end of May, and beginning of June for station 3. For station 9, it includes September, March, and both samples for May. For station 17, only the beginning of June. For station 24, September and both samples for May are included.

The final dimension is station 3 in April. This station is in a more sediment-rich section of the river entering into the lake. This may point to a difference in the relationship between the reflectance and chlorophyll-a due to the turbidity in the water during April. There are other circumstances that may account for the creation of these dimensions. The flow that goes past the points at different times of the year may affect the reflectance. In addition, some of the areas such as station 3 are more constricted and this may cause the reflectance to react more as river conditions rather than lake conditions. Therefore, even though these dimensions do not appear to mean anything, it may be that additional analysis of the flow or turbidity in the water at certain times/locations would provide additional explanation.

The next situation uses a subset selection procedure on this data set. The goal is to eliminate variables so that 95% of the variability would still be explained under the assumption that the variables are independent. The last three eigenvalues predicted by the

ANOSE procedure are 0.726, 0.686, and 0.625. These explain approximately 5% of the data. Note that we have 0.0536 =(0.726 + 0.6863 + 0.6254)/ 38 . In this case, 3 variables can be deleted leaving at least 95% of the variation of the data to be explained by the other variables.

Examination of the smallest eigenvalues results in the possible deletion of the following 3 variables: Station 3 measurements in September and March and station 24 measurements in April. Figure 28 shows that both station 3 and station 24 are located farthest from the main lake body which could account for differing conditions at these locations. In addition, most of the growth has taken place before September and will occur after March, so the data could be stagnant at these station/date combinations. Those knowledgeable about this data and the effects of the locations of the sampling stations and the dates of the collection periods can determine if these stations/dates should be dropped from the analysis in order to provide a better prediction model for chlorophyll-a. This analysis illustrates that if samples had to be reduced that these station/date locations would be primary candidates for deletion.

Summary of Real World Application

The two situations in this real world application demonstrate how the concepts discussed in this paper can be applied to data mining applications. The further analysis of this data is beyond the scope of this paper, but this initial analysis shows how these techniques can support a researcher in examining relationships within a data set.

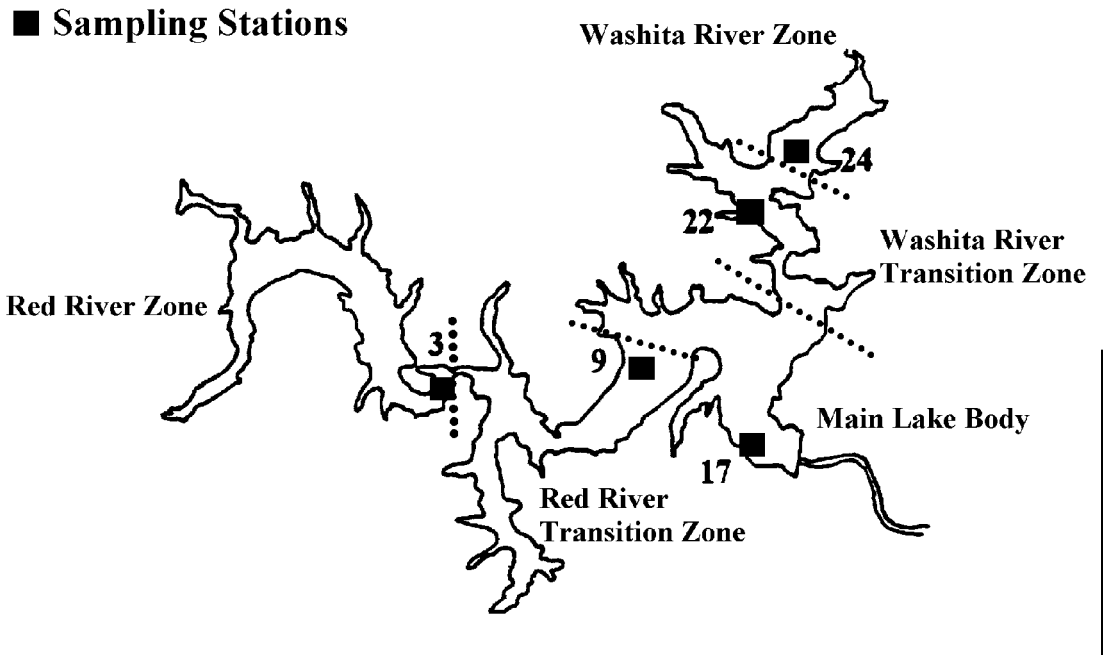

Figure 28. Sampling locations within five major zones of Lake Texoma (adapted from Atkinson, Acevedo, Dickson, and Rolbecki 1998)

Summary of Data Analysis

This chapter begins with a discussion of the models that were used to predict mean eigenvalues. The 2 regression models, the 2 neural network models, and the 3 asymptotic theory models are explained and then compared with the previous approaches. These comparisons are made using a full data set of the values from Lautenschlager (1989) and then a reduced data set that would include values of sample size and numbers of variables used in many applications. The next section explains the new regression and neural network models used to create 95^{th} percentile eigenvalues. These are compared to the previous regression equation estimates.

A third section outlines the simulation study on the density of *p*-values and the estimation procedure for the proportions in this mixture of betas. The final section introduces a real world data set on Lake Texoma and shows how the techniques presented can be used in a real application. The last chapter in this study reports the conclusions of the research questions presented in chapter 2.

ADDITIONAL FIGURES FOR CHAPTER 4

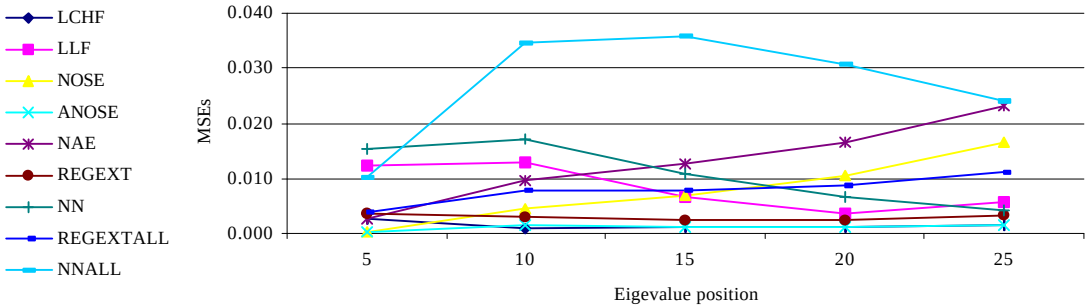

Figure 29. MSE for N = 50 - Reduced data set

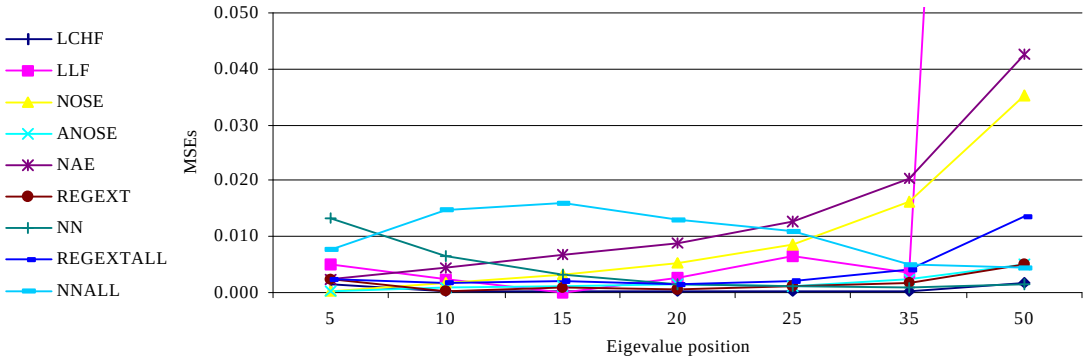

Figure 30. MSE for N = 75 - Reduced data set

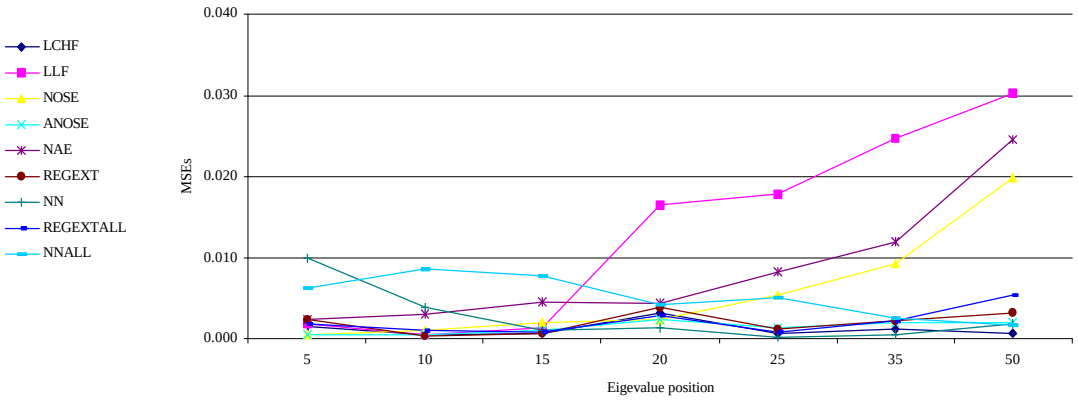

Figure 31. MSE for N = 100 - Reduced data set

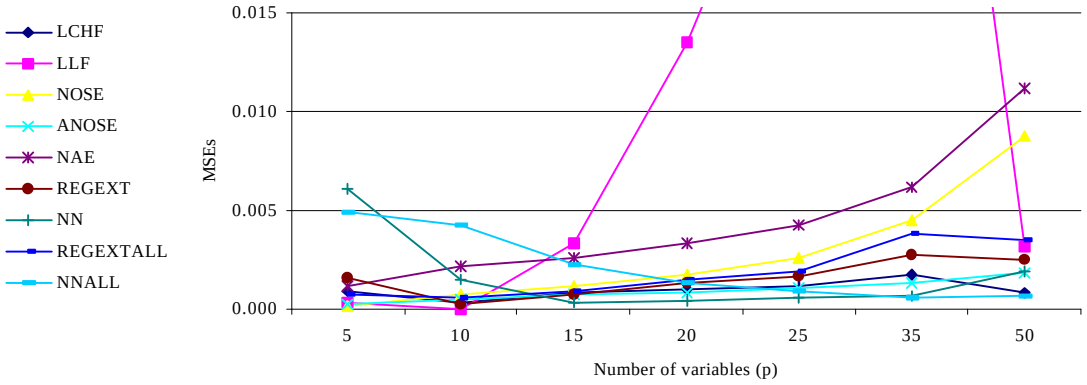

Figure 32. MSE for N = 150 - Reduced data set

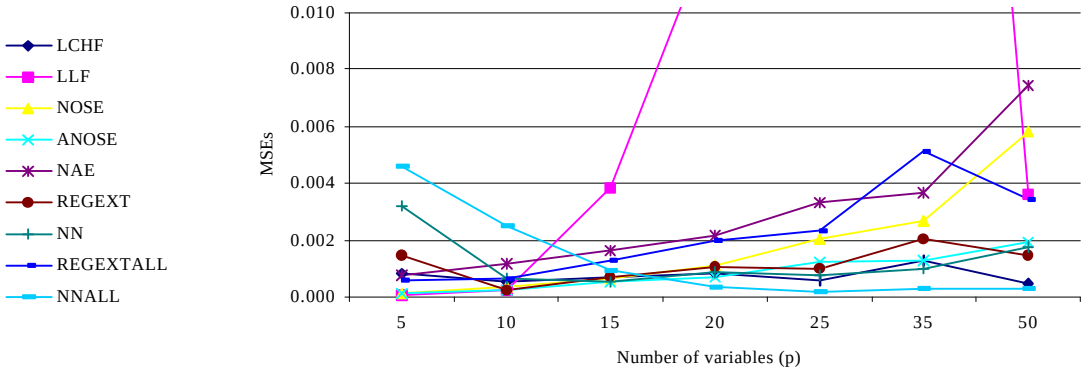

Figure 33. MSE for N = 200 - Reduced data set

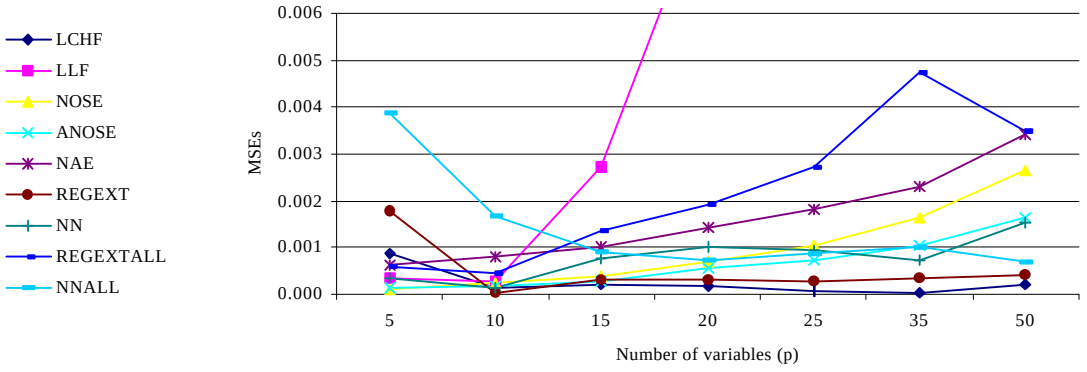

Figure 34. MSE for N = 300 - Reduced data set

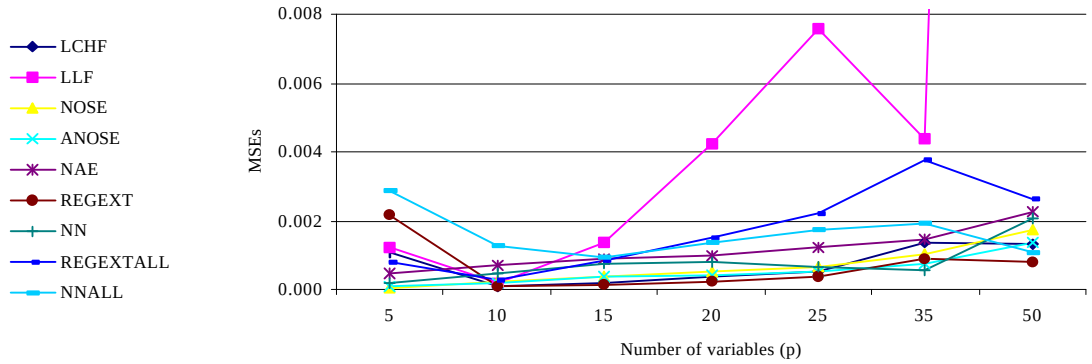

Figure 35. MSE for N = 400 - Reduced data set

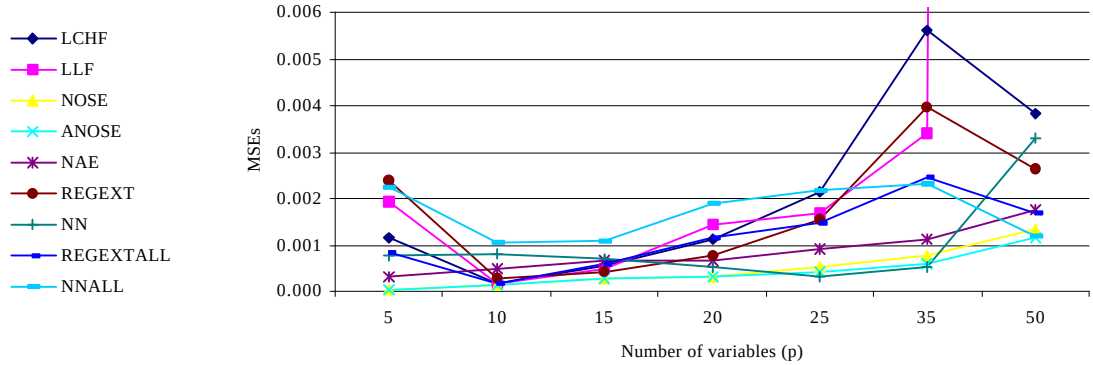

Figure 36. MSE for N = 500 - Reduced data set

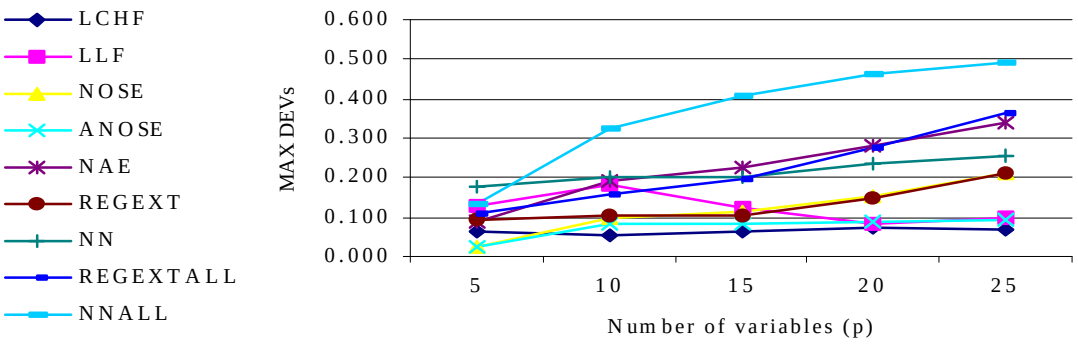

Figure 37. MAX DEV for N = 50 - Reduced data set

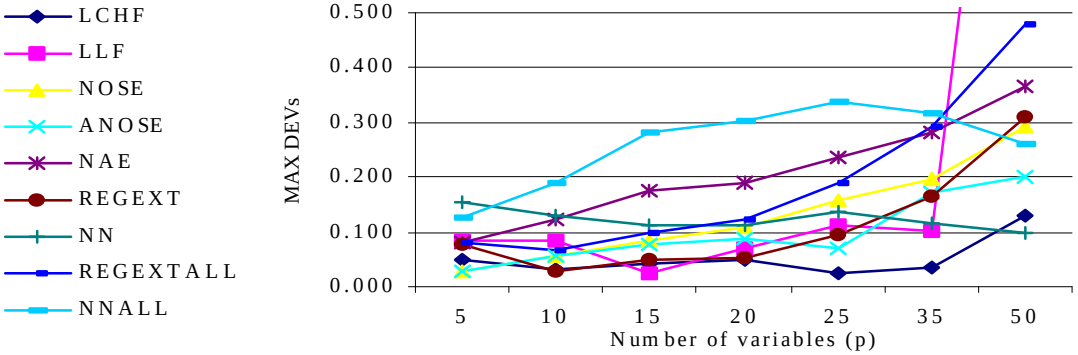

Figure 38. MAX DEV for N = 75 - Reduced data set

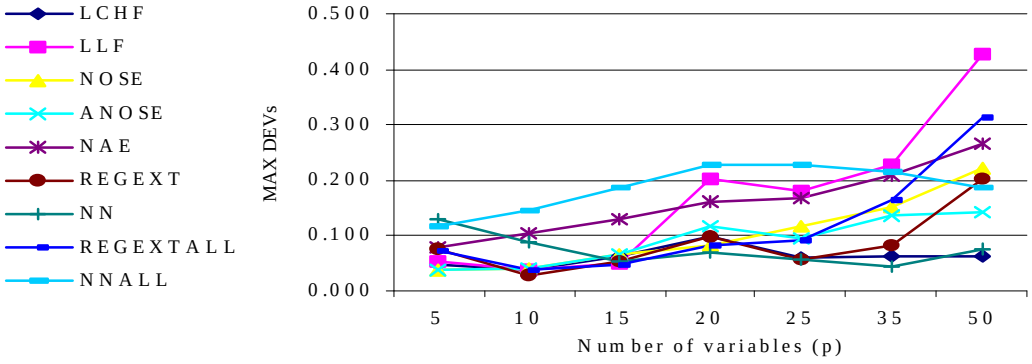

Figure 39. MAX DEV for N = 100 - Reduced data set

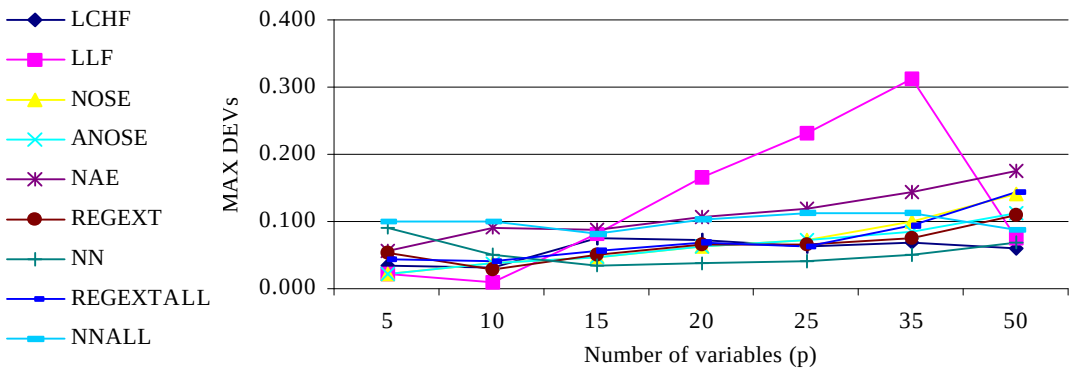

Figure 40. MAX DEV for N = 150 - Reduced data set

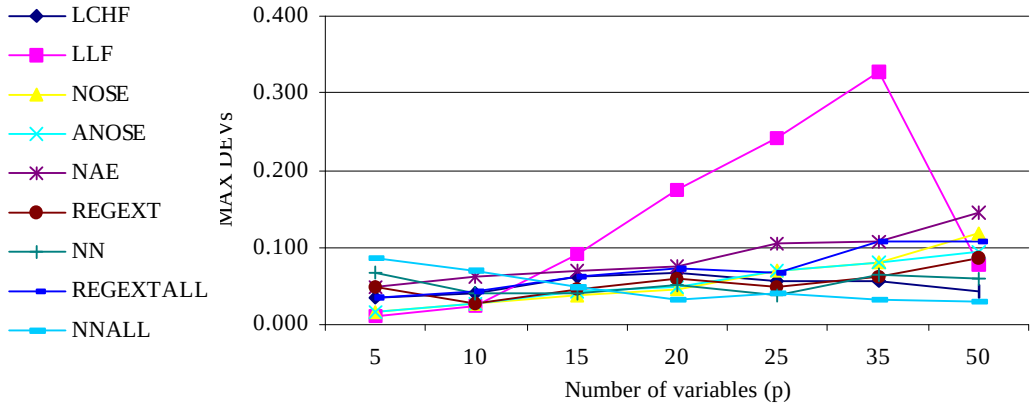

Figure 41. MAX DEV for N = 200 - Reduced data set

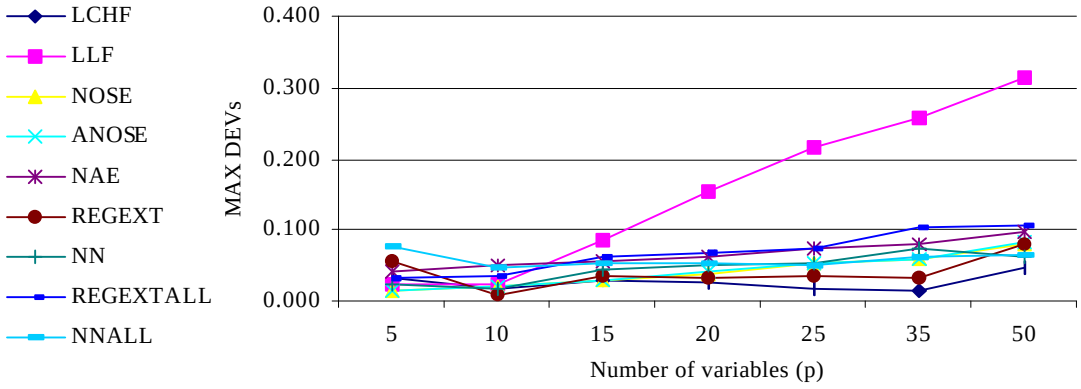

Figure 42. MAX DEV for N = 300 - Reduced data set

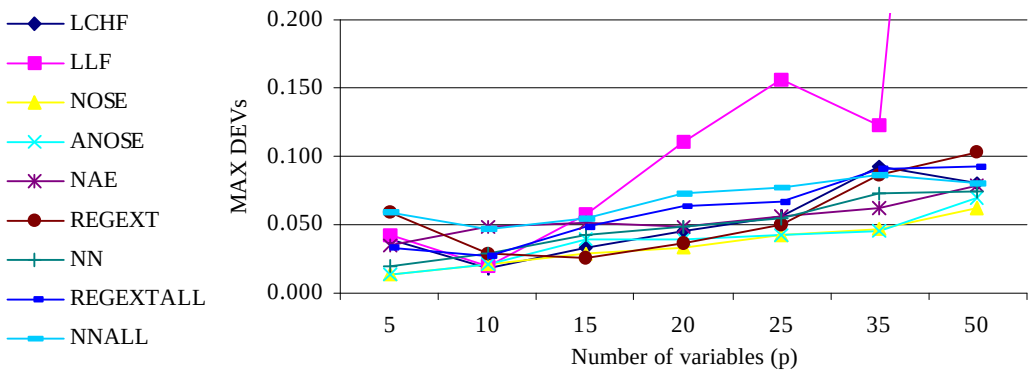

Figure 43. MAX DEV for N = 400 - Reduced data set

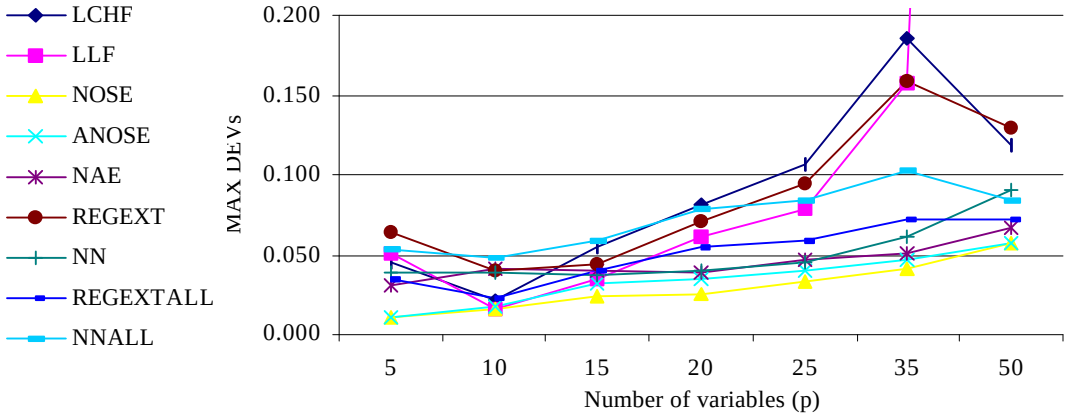

Figure 44. MAX DEV for N = 500 - Reduced data set

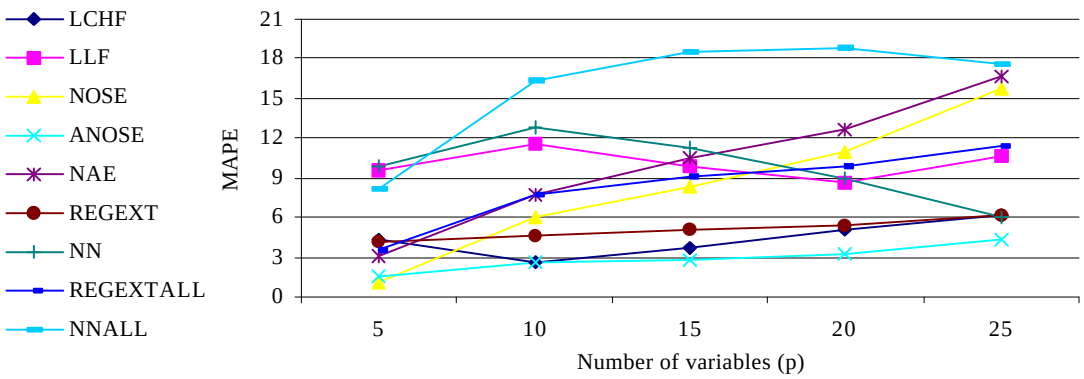

Figure 45. MAPE for N = 50 - Reduced data set

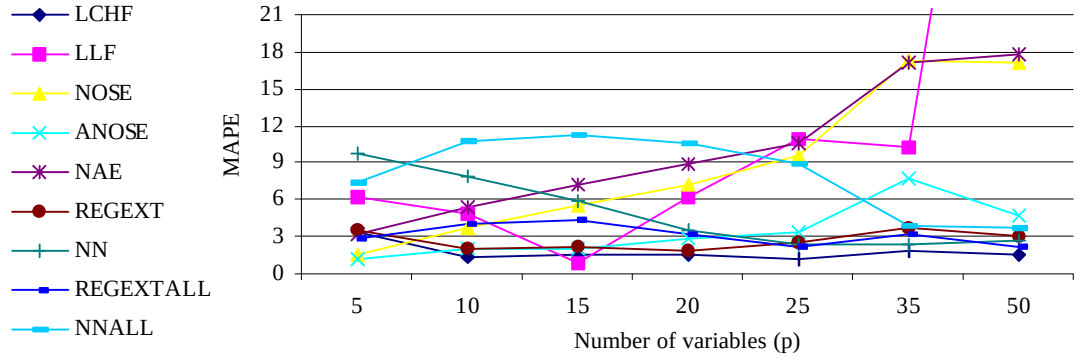

Figure 46. MAPE for N = 75 - Reduced data set

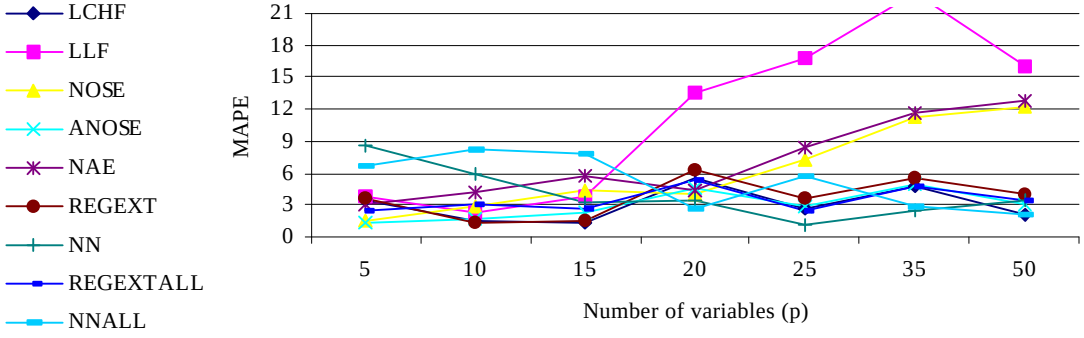

Figure 47. MAPE for N = 100 - Reduced data set

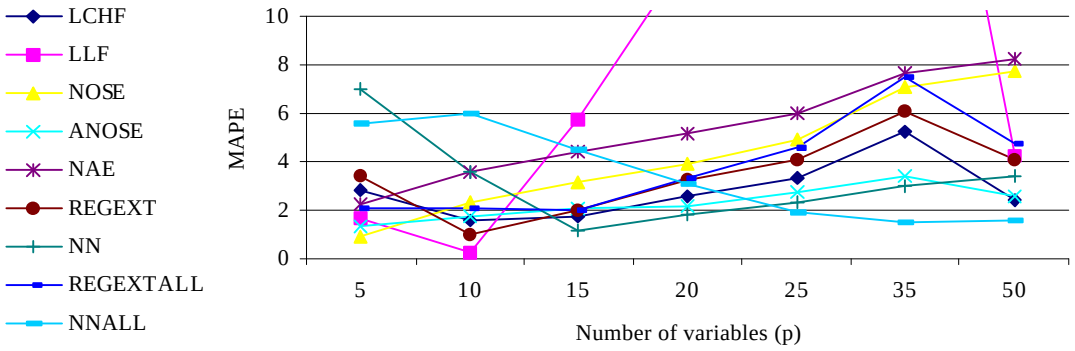

Figure 48. MAPE for N = 150 - Reduced data set

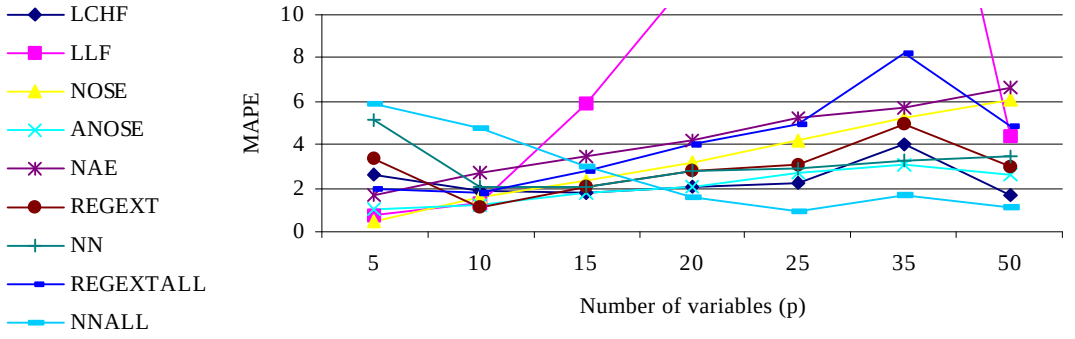

Figure 49. MAPE for N = 200 - Reduced data set

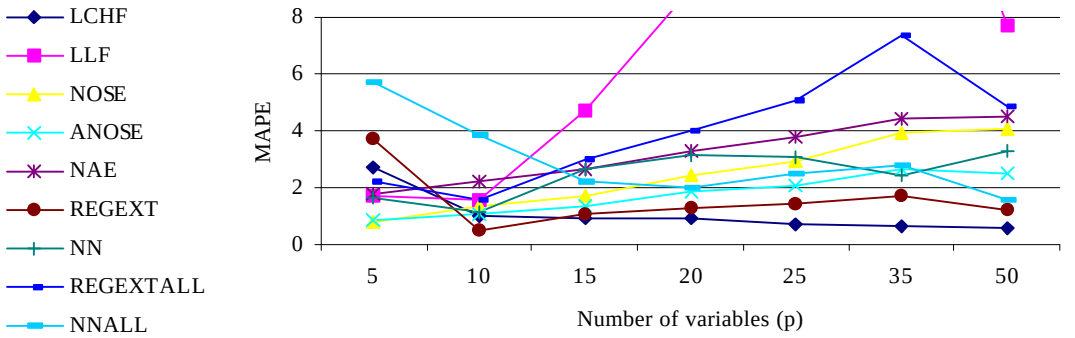

Figure 50. MAPE for N = 300 - Reduced data set

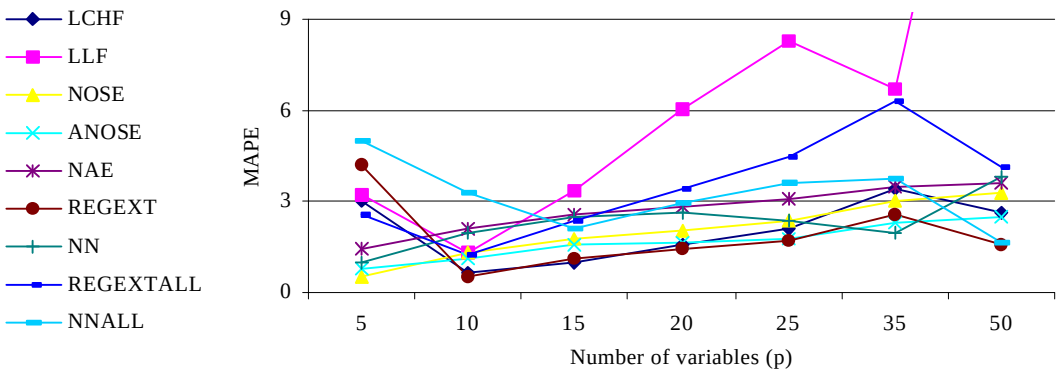

Figure 51. MAPE for N = 400 - Reduced data set

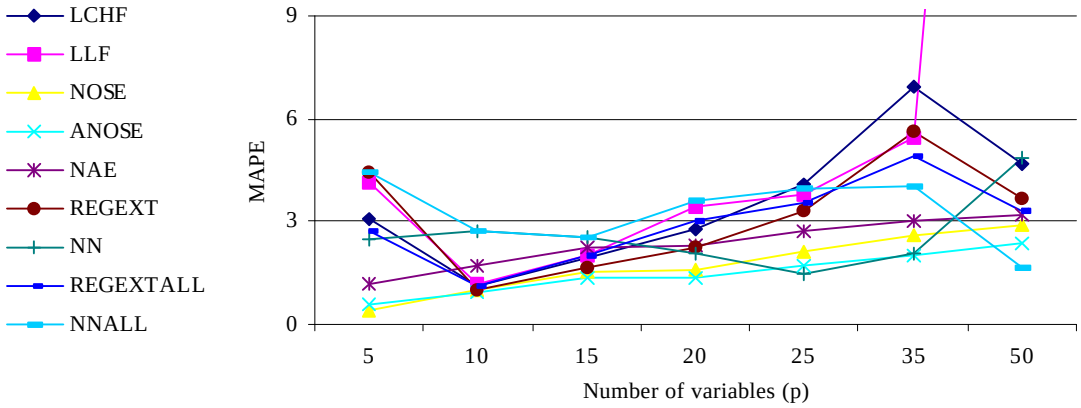

Figure 52. MAPE for N = 500 - Reduced data set

CHAPTER 5

DISCUSSION OF RESULTS

This chapter discusses the results of this dissertation research. The estimators to predict the mean and 95th percentile eigenvalues have been compared to the previous regression approaches and the published tables of simulated values. There are many cases where the new approaches perform as well or better than the previous techniques. A simulation was performed to analyze the density of *p*-values as a mixture of 2 beta distributions. The results show that this procedure is viable for estimating *p*-values from central and non-central t distributions. In addition, an application to real world data is used to demonstrate these concepts. The conclusions of these analyses are presented next.

Conclusions

The conclusions presented in this section are based on the results in chapter 4. For the estimation of the eigenvalues, the previous and new methods have been compared based on their prediction errors using the tabled simulated values from Lautenschlager (1989) as the actual values. These comparisons are made based on several subsets of the data. These subsets of N and p values are used to make recommendations for using a particular method based on the particular data in the study.

For the simulation study testing the mixture model for the *p*-values, the results are analyzed using MADs and MSEs of the parameter estimates with respect to the true

values of these estimates. The analysis of real world data using the methodology investigated in this study has also been presented in chapter 4. In this section, each research question from chapter 2 is presented separately and conclusions are made.

Research Question 1

Can an improved regression equation be found for predicting the mean eigenvalue? Would this improved regression equation require a table with an extensive list of coefficients? How does this compare to the two previously recommended regression equations?

In this study, two regression equations are presented, the REGEXT and the REGEXTALL equations. An advantage to these equations is that they do not require tables with extensive lists of coefficients. When using the reduced data set for the comparisons, the REGEXT equation performs similarly to the LCHF equation and outperforms the LLF equation on almost all of the comparisons. Therefore it is determined that an improved equation can be found.

The REGEXTALL equation has been created to use the full data set from Lautenschlager (1989). It does not perform better than the REGEXT equation on the reduced data set. But, when examining the full data set, the REGEXTALL equation does perform comparably well with the ANOSE method. The ANOSE is discussed under research question 3 and is determined to perform the best overall. Therefore, comparison to the ANOSE is part of the conclusions. In most cases, for sample sizes of 150, 200, 750, and 1000, the REGEXTALL is not significantly different from the ANOSE

procedure. When examining the top p/3 eigenvalues for the full data set, the REGEXTALL is not significantly different from the ANOSE for sample sizes of 100, 150, 200, 400, 500, 750, and 1000. And finally, for the comparisons made using the bottom 10 % of the full data set, the REGEXTALL performs well for all sample sizes except 1500 and 2000. Therefore it is determined that the REGEXTALL equation does not perform well when the sample size is above 1000.

Research Question 2

Can an asymptotic prediction method be used to accurately predict mean eigenvalues? How does this method compare with the previously recommended regression equations and with the new regression equation?

To answer this research question, two approaches to estimating eigenvalues have been created based on theory that eigenvalues are asymptotically normal. The first approach uses order statistics to estimate the eigenvalues. This procedure is called the NOSE. An adjustment to the procedure is also implemented and this new method is called the ANOSE. A second approach uses the distribution theory of percentiles and is called the NAE.

The NAE and NOSE procedures are similar when comparing across the eigenvalue positions. The ANOSE procedure performs better overall and is often better than the LCHF and REGEXT equations, especially for the higher eigenvalue positions. The NAE and NOSE procedures perform worse when the sample size is small. When looking at the reduced data set, all three procedures perform well if the sample size is greater than 50.

The ANOSE performs very well in almost all cases when examining the full data set. The NOSE performs similarly to the ANOSE if the sample size is greater than 150 and the NAE performs similarly if the sample size is greater than 300. When examining the bottom 10 % of the eigenvalue positions, the NAE performs well if the sample size is greater than 400 and the ANOSE and NOSE perform well if the sample size is greater than 750. If the sample size is 750 or below, the REGEXTALL procedure performs the best on the bottom 10% data set.

Research Question 3

Can a neural network prediction model be used to accurately predict mean eigenvalues?

This study creates two neural network procedures for predicting mean eigenvalues. The NN uses a reduced set of values where the sample size is 500 or less and the number of variables is no greater than 50. The NNALL uses the complete set of simulated values from the Lautenschlager (1989) study. Both procedures perform similar to the LCHF, ANOSE, and REGEXT models on the reduced data once N was greater than 50. For the reduced data set with the top p/3 eigenvalues, the NN still performs similarly to the REGEXT and ANOSE for sample sizes greater than 50. The NNALL performs comperable when N = 150 and 200. When N > 200, the NNALL performs worse. This is similar to the performance of the REGEXTALL equation.

When examining the eigenvalue position graphs using the full data set, the NNALL and ANOSE perform the best. The NNALL performs the best on the full data set when

the sample size is 1000, 1500, and 2000. Its performance for these sample sizes is similar to the ANOSE. And finally, when looking at the bottom 10 % of eigenvalue positions, the NNALL performs the best for almost all the cases. The one exception is when the sample size is 750 when the REGEXTALL and NAE have the smallest prediction errors.

Research Question 4

Can an improved regression equation to predict the 95^{th} percentile eigenvalue be formulated? Can a neural network topology be found which will be a viable approach to predicting the 95^{th} percentile eigenvalue?

The REGEXT95 and NN95 methods are compared to the LCHF95 in chapter 4. The results show that when comparing based on the eigenvalue position, the REGEXT95 equation performs similarly to the LCHF95 equation. The REGEXT95 equation does have the advantage of not requiring a table of coefficients in order to predict the eigenvalues. The Tukey tests show that in many cases there is no difference between the three techniques. The NN95 and REGEXT95 equations do not perform as well as the LCHF95 when the sample size is small. In most of the other cases, there is no significant difference between the three techniques.

Research Question 5

In considering large numbers of independent variables, can a method be implemented to determine the number of true null hypotheses? Will a new method using a mixture of beta distributions be useful in determining the distribution of the *p*-values in

studies which result in multiple hypotheses that may in fact be too numerous for traditional experimentwise error controlling procedures to perform satisfactorily?

This study uses simulation to validate the use of a mixture of beta distributions to model *p*-values from central and noncentral t distributions as well as to model *p*-values from the uniform distribution and a beta distribution. In addition, a procedure is presented that creates estimates of the proportions for the mixture. The results show that this procedure does have merit and would allow a researcher to estimate the number of true null hypotheses.

Research Question 6

Can the number of variables be reduced in the lake data set? What is the interpretation of multiple tests conducted on this real world data set?

The Lake Texoma data have been used to demonstrate the procedures outlined in this dissertation. The first goal is to reduce the number of variables in the data set. In this case, the variables represent location/date combinations. If it is determined that 95% of the variation in the data should be explained by the data, then the variables associated with the last 3 eigenvalues can be removed. Since the ANOSE procedure has proven to be the best in the case of this particular sample size and number of variables, it is used to estimate the last 3 eigenvalues. In the case of using exploratory data mining techniques such as a large number of regressions, the reduction of the number of variables from 38 to 35 can be a cost savings measure for the researcher. The second goal in analyzing that data is to determine if a relationship exists between the energy reflectance and the chlorophyll-a

levels. A large number of regressions to predict chlorophyll-a with the energy reflectance data have been performed and the resulting *p*-values have been analyzed with the mixture approach. The results identified a set of reflectance that could possibly be removed from the analysis in order refine the data for further study of relationship of chlorophyll-a and the energy reflectance.

Other Conclusions

Some final conclusions about the eigenvalue prediction methods are noted in this section. If the number of observations in a study is between 50 and 500 and the number of variables is less than 50, then the ANOSE, NN, REGEXT, or LCHF procedures could be used. If a researcher wants to estimate the bottom 10% of the eigenvalues, then the REGEXTALL or NNALL should be used up to sample sizes of 1000 at which point the ANOSE, NOSE, or NAE procedure should be used. But if the researcher does not have access to tables of order statistics, then the NAE procedure should be used. If the researcher has a large sample size of 1000 or more, then the ANOSE or NNALL should be used.

Table 34 is presented to allow a researcher quick access to recommended eigenvalue prediction methods based on the characteristics of their study. Each method compared in this study is rated using the scale: Recommended, acceptable, discouraged, or highly discouraged. The entries in this table are based on the multiple comparison procedures and graphs created in chapter 4 and are somewhat subjective. The first column gives boundaries for sample size and numbers of variables. The first six rows

Table 34. Recommendations for mean eigenvalue estimation

	LCHF[a]	LLF[a]	NOSE[b]	ANOSE[b]	NAE	REGEXT	NN[c]	REGEXTALL	NNALL[c]
N=50 to 150 p <= 50	♦	◊	◊	♦	◊	♦	×	×	×
N=50 to 150 50 < p <= 80	⊗	⊗	◊	♦	◊	⊗	⊗	×	×
150 < N <= 500 p <= 50	♦	⊗	♦	♦	♦	♦	♦	◊	◊
150 < N <= 500 50 < p <= 80	♦	⊗	◊	♦	◊	♦	♦	◊	◊
500 < N <= 750 p <= 80	⊗	⊗	♦	♦	♦	⊗	⊗	×	×
1000 < N <= 2000 p <= 80	⊗	⊗	♦	♦	♦	⊗	⊗	⊗	♦
Top p / 3 N=50 to 150 p <= 50	♦	♦	♦	♦	×	♦	♦	♦	◊
Top p / 3 N=50 to 150 50 < p <= 80	⊗	⊗	◊	♦	×	⊗	⊗	◊	◊
Top p / 3 150 < N <= 500 p <= 50	♦	♦	♦	♦	×	◊	♦	◊	×
Top p / 3 150 < N <= 500 50 < p <= 80	⊗	⊗	◊	♦	×	⊗	⊗	◊	×
Top p / 3 500 < N <= 2000 p <= 80	⊗	⊗	♦	♦	♦	⊗	⊗	×	◊
Bottom 10% N=50 to 1000 p <= 80	⊗	⊗	×	×	×	⊗	⊗	♦	♦
Bottom 10% 1000 < N <= 2000 p <= 80	⊗	⊗	♦	♦	♦	⊗	⊗	×	♦

[a]Requires table of coefficients

[b]Requires table of order statistics

[c]Neural network software makes implementation easier

♦ Recommended

◊ Acceptable

× Discouraged

⊗ Highly discouraged

recommend procedures for general use. The next five rows recommend procedures when the first p/3 eigenvalue positions are of interest. The final two rows detail recommendations when the bottom 10% of eigenvalue positions are to be examined. The final sections of this study contain the limitations of this research and the plans for future research.

Limitations of this Study

This study does include several limitations. Only the mean eigenvalue and 95th percentile of the eigenvalues are considered. Further research may be warranted in considering a methodology for estimating other percentiles. The question of comparing the number of components extracted versus the number of known components is not addressed. The neural network models are designed to be as parsimonious as possible. All three models presented used only 4 inputs: N, p, k, and p/N to create their models. In addition, only a single hidden layer is used with either 2 or 3 hidden neurons. If this restriction had not been used and the model included more inputs such as the logNlogp and other interaction terms used in the regression equations, for example, the models may have produced a better fit.

The inability of assessing the true distribution of *p*-values in any real world data set restricts this study to assessing the ability of the mixture approach to fit only the *p*-values generated from a t and a noncentral t distribution as well as *p*-values for the uniform and a beta distribution. The use of the mixture to form a distribution has inherent problems. In practice, the value of k (number of densities) is never known in the mixture model for

estimating the density of the *p*-values. Also, the question of identifiability can cause the results of a mixture to be misleading. Another point to be made is that the *p*-value technique should be used more for exploratory analysis because the clusters generated by the distribution in the mixture may not be the true representation of the underlying population. Finally, the beta distribution may not be a good fit for all the possible distributions from which *p*-values could be generated.

Future Research

This study showed that there are other techniques that can be used to predict eigenvalues of a correlation matrix from random data. The regression, asymptotic, and neural network approaches each perform well for certain combinations of N and p. Future research should examine combining these models to see if more accurate results can be achieved across a greater range of sample size and numbers of variables.

The ANOSE procedure performs the best overall. Further refinement of those adjustments made to the NOSE to create the ANOSE should be examined. In addition, the neural network models have been created using very parsimonious models and a basic selection of back propagation parameters. It could be that better fitting networks could be found by using additional input options. Finally, continuation with the simulation study should be performed to expand it to include *p*-values generated from additional distributions.

APPENDIX

Table A1. R-Squares for reduced data set

	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
1	0.9877	0.9930	0.9917	0.9948	0.9880	0.9868	0.9783	0.9678	0.9063
2	0.9892	0.9895	0.9962	0.9966	0.9950	0.9892	0.9872	0.9765	0.9463
3	0.9909	0.9852	0.9981	0.9969	0.9978	0.9911	0.9876	0.9810	0.9665
4	0.9943	0.9811	0.9980	0.9953	0.9981	0.9926	0.9948	0.9871	0.9759
5	0.9957	0.9744	0.9975	0.9949	0.9979	0.9944	0.9924	0.9891	0.9843
6	0.9969	0.9651	0.9961	0.9949	0.9966	0.9959	0.9882	0.9906	0.9850
7	0.9976	0.9571	0.9950	0.9919	0.9955	0.9966	0.9847	0.9922	0.9813
8	0.9971	0.9467	0.9930	0.9918	0.9939	0.9962	0.9804	0.9910	0.9726
9	0.9980	0.9440	0.9877	0.9857	0.9881	0.9971	0.9910	0.9933	0.9859
10	0.9976	0.9332	0.9852	0.9832	0.9855	0.9967	0.9887	0.9923	0.9788
11	0.9953	0.9188	0.9839	0.9862	0.9842	0.9946	0.9859	0.9897	0.9701
12	0.9935	0.9079	0.9815	0.9860	0.9823	0.9931	0.9837	0.9868	0.9634
13	0.9896	0.8963	0.9780	0.9925	0.9800	0.9901	0.9809	0.9830	0.9580
14	0.9904	0.8818	0.9692	0.9912	0.9691	0.9900	0.9854	0.9842	0.9714
15	0.9869	0.8684	0.9687	0.9878	0.9687	0.9868	0.9825	0.9793	0.9676
16	0.9839	0.8586	0.9665	0.9853	0.9671	0.9841	0.9811	0.9758	0.9659
17	0.9802	0.8473	0.9649	0.9815	0.9663	0.9809	0.9790	0.9727	0.9658
18	0.9758	0.8342	0.9598	0.9755	0.9629	0.9771	0.9774	0.9699	0.9656
19	0.9818	0.8061	0.9611	0.9650	0.9607	0.9821	0.9832	0.9745	0.9792
20	0.9780	0.7901	0.9620	0.9701	0.9618	0.9778	0.9826	0.9714	0.9791
21	0.9755	0.7765	0.9609	0.9950	0.9612	0.9748	0.9824	0.9698	0.9791
22	0.9718	0.7585	0.9602	0.9935	0.9612	0.9706	0.9802	0.9683	0.9780
23	0.9678	0.7423	0.9544	0.9889	0.9564	0.9662	0.9799	0.9684	0.9787
24	0.9567	0.6325	0.8533	0.9981	0.8528	0.9598	0.9857	0.9620	0.9910
25	0.9540	0.6115	0.8631	0.9978	0.8629	0.9565	0.9851	0.9611	0.9901
26	0.9502	0.5927	0.8754	0.9984	0.8756	0.9519	0.9841	0.9627	0.9896
27	0.9470	0.5646	0.8830	0.9995	0.8836	0.9469	0.9809	0.9624	0.9884
28	0.9430	0.5339	0.8888	0.9988	0.8898	0.9417	0.9768	0.9624	0.9873
29	0.9402	0.5176	0.8918	0.9922	0.8934	0.9371	0.9727	0.9634	0.9865
30	0.9352	0.5009	0.8916	0.9896	0.8938	0.9312	0.9691	0.9640	0.9869
31	0.9313	0.4771	0.8913	0.9853	0.8943	0.9252	0.9650	0.9655	0.9862
32	0.9265	0.4588	0.8886	0.9612	0.8926	0.9194	0.9615	0.9680	0.9851
33	0.9181	0.4368	0.8849	0.8991	0.8903	0.9089	0.9613	0.9704	0.9854

Table A2. MADs for reduced data set

	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
1	0.0405	0.0256	0.0445	0.0330	0.1077	0.0601	0.0488	0.0791	0.1207
2	0.0334	0.0289	0.0629	0.0314	0.0963	0.0340	0.0363	0.0531	0.0824
3	0.0286	0.0329	0.0607	0.0353	0.0811	0.0298	0.0344	0.0458	0.0656
4	0.0201	0.0348	0.0601	0.0402	0.0765	0.0294	0.0221	0.0468	0.0472
5	0.0159	0.0397	0.0501	0.0365	0.0614	0.0281	0.0202	0.0439	0.0412
6	0.0122	0.0465	0.0444	0.0319	0.0528	0.0259	0.0201	0.0396	0.0400
7	0.0102	0.0511	0.0390	0.0284	0.0452	0.0236	0.0224	0.0350	0.0378
8	0.0117	0.0589	0.0370	0.0280	0.0425	0.0205	0.0270	0.0313	0.0393
9	0.0088	0.0663	0.0366	0.0268	0.0399	0.0186	0.0217	0.0270	0.0253
10	0.0094	0.0718	0.0380	0.0258	0.0408	0.0170	0.0253	0.0245	0.0266
11	0.0127	0.0797	0.0396	0.0235	0.0428	0.0169	0.0295	0.0230	0.0284
12	0.0157	0.0860	0.0408	0.0227	0.0449	0.0175	0.0320	0.0244	0.0323
13	0.0204	0.0933	0.0392	0.0176	0.0449	0.0197	0.0341	0.0264	0.0361
14	0.0160	0.1074	0.0453	0.0198	0.0473	0.0211	0.0294	0.0267	0.0246
15	0.0196	0.1147	0.0480	0.0225	0.0508	0.0234	0.0303	0.0311	0.0285
16	0.0227	0.1207	0.0509	0.0237	0.0536	0.0248	0.0299	0.0338	0.0318
17	0.0269	0.1282	0.0510	0.0248	0.0550	0.0265	0.0310	0.0362	0.0346
18	0.0314	0.1363	0.0503	0.0281	0.0545	0.0286	0.0312	0.0380	0.0377
19	0.0218	0.1447	0.0626	0.0274	0.0652	0.0258	0.0238	0.0358	0.0282
20	0.0251	0.1533	0.0643	0.0249	0.0681	0.0283	0.0240	0.0399	0.0296
21	0.0271	0.1603	0.0617	0.0180	0.0671	0.0301	0.0243	0.0444	0.0323
22	0.0305	0.1707	0.0572	0.0157	0.0638	0.0323	0.0264	0.0480	0.0340
23	0.0339	0.1791	0.0631	0.0211	0.0622	0.0349	0.0274	0.0502	0.0360
24	0.0271	0.2062	0.0875	0.0187	0.0894	0.0366	0.0244	0.0529	0.0201
25	0.0292	0.2201	0.0893	0.0169	0.0919	0.0393	0.0278	0.0566	0.0198
26	0.0325	0.2361	0.0898	0.0155	0.0932	0.0420	0.0310	0.0598	0.0187
27	0.0353	0.2560	0.0881	0.0169	0.0922	0.0442	0.0345	0.0628	0.0188
28	0.0392	0.2825	0.0840	0.0163	0.0891	0.0458	0.0388	0.0656	0.0191
29	0.0417	0.3106	0.0795	0.0196	0.0857	0.0463	0.0425	0.0665	0.0187
30	0.0451	0.3288	0.0716	0.0200	0.0793	0.0466	0.0457	0.0677	0.0182
31	0.0486	0.3560	0.0662	0.0238	0.0702	0.0476	0.0481	0.0683	0.0181
32	0.0516	0.3937	0.0738	0.0362	0.0705	0.0486	0.0499	0.0676	0.0190
33	0.0566	0.4322	0.0960	0.0604	0.0834	0.0512	0.0503	0.0671	0.0198

Table A3. MAPEs for reduced data set

	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
1	2.41	1.45	2.26	1.75	5.76	3.26	2.61	3.95	6.31
2	2.16	1.80	3.45	2.01	5.45	2.13	2.45	3.09	5.09
3	1.97	2.26	3.54	2.37	4.80	2.08	2.70	3.02	4.75
4	1.39	2.26	3.60	2.74	4.65	2.01	1.60	3.24	3.37
5	1.14	2.79	3.14	2.61	3.88	2.00	1.64	3.22	3.20
6	0.92	3.55	3.03	2.34	3.61	1.94	1.82	3.08	3.44
7	0.83	4.22	2.96	2.21	3.43	1.82	2.21	2.92	3.64
8	1.10	5.28	3.12	2.28	3.63	1.74	2.93	2.86	4.33
9	0.75	5.68	3.07	2.10	3.29	1.43	1.93	2.24	2.38
10	0.90	6.50	3.58	2.18	3.81	1.40	2.41	2.21	2.79
11	1.33	7.64	4.11	2.07	4.45	1.58	3.01	2.30	3.36
12	1.75	8.71	4.59	2.10	5.11	1.83	3.57	2.74	4.29
13	2.43	10.05	4.63	1.74	5.48	2.35	4.23	3.35	5.42
14	1.87	11.14	5.33	2.05	5.59	2.27	3.08	2.98	3.03
15	2.43	12.53	6.00	2.52	6.42	2.73	3.47	3.74	3.86
16	2.97	13.92	6.60	2.75	7.09	3.15	3.77	4.41	4.72
17	3.75	15.59	6.79	2.90	7.55	3.69	4.30	5.18	5.72
18	4.67	17.51	6.75	3.73	7.53	4.45	4.89	6.07	7.05
19	2.88	17.17	8.25	3.11	8.75	3.41	2.89	4.72	4.12
20	3.54	19.15	8.67	2.94	9.44	4.00	3.17	5.57	4.80
21	4.05	21.03	8.27	2.64	9.42	4.47	3.48	6.53	5.74
22	4.94	23.59	7.29	2.21	8.70	5.16	4.22	7.70	6.98
23	5.75	26.07	10.22	3.64	9.08	6.03	4.76	8.74	8.41
24	3.31	25.18	11.48	2.66	11.78	4.74	2.69	6.55	2.45
25	3.65	28.00	12.21	2.60	12.63	5.28	3.15	7.25	2.54
26	4.23	31.34	12.75	2.52	13.31	5.86	3.64	7.93	2.53
27	4.78	35.40	13.00	2.72	13.74	6.35	4.26	8.59	2.69
28	5.59	40.70	12.79	2.50	13.74	6.81	5.09	9.25	2.90
29	6.22	46.02	12.35	3.40	13.59	7.07	5.92	9.75	3.11
30	7.04	50.77	11.07	3.36	12.71	7.26	6.81	10.36	3.29
31	7.88	57.39	10.74	4.41	11.04	7.61	7.65	11.00	3.67
32	8.83	65.91	14.64	7.91	12.95	8.08	8.57	11.41	4.30
33	10.25	75.52	23.32	17.03	19.26	9.01	9.44	12.02	5.08

Table A4. Randomized block for MSE - Reduced data set

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					50				
5	0.00275	0.01221	0.00024	0.00038	0.00271	0.00360	0.01547	0.00405	0.01010
10	0.00085	0.01303	0.00437	0.00153	0.00974	0.00289	0.01700	0.00784	0.03471
15	0.00110	0.00662	0.00686	0.00130	0.01261	0.00234	0.01086	0.00789	0.03588
20	0.00129	0.00374	0.01038	0.00114	0.01664	0.00241	0.00664	0.00863	0.03078
25	0.00142	0.00568	0.01660	0.00138	0.02331	0.00317	0.00428	0.01112	0.02402
					75				
5	0.00162	0.00495	0.00039	0.00032	0.00240	0.00241	0.01326	0.00232	0.00769
10	0.00029	0.00245	0.00171	0.00092	0.00458	0.00042	0.00661	0.00190	0.01492
15	0.00039	0.00012	0.00326	0.00111	0.00681	0.00091	0.00335	0.00217	0.01588
20	0.00042	0.00261	0.00530	0.00134	0.00902	0.00070	0.00150	0.00149	0.01316
25	0.00018	0.00642	0.00849	0.00109	0.01263	0.00109	0.00119	0.00221	0.01097
35	0.00028	0.00344	0.01617	0.00234	0.02027	0.00164	0.00082	0.00403	0.00506
50	0.00173	0.36872	0.03530	0.00517	0.04275	0.00517	0.00153	0.01361	0.00453
					100				
5	0.00156	0.00184	0.00052	0.00049	0.00240	0.00239	0.00995	0.00189	0.00621
10	0.00043	0.00054	0.00105	0.00059	0.00309	0.00031	0.00378	0.00098	0.00859
15	0.00069	0.00133	0.00207	0.00108	0.00449	0.00070	0.00101	0.00090	0.00772
20	0.00318	0.01640	0.00236	0.00237	0.00434	0.00393	0.00142	0.00280	0.00424
25	0.00069	0.01780	0.00534	0.00131	0.00820	0.00122	0.00025	0.00092	0.00511
35	0.00110	0.02462	0.00922	0.00200	0.01195	0.00221	0.00051	0.00223	0.00249
50	0.00074	0.03024	0.01975	0.00199	0.02448	0.00324	0.00191	0.00540	0.00165
					150				
5	0.00095	0.00036	0.00017	0.00026	0.00120	0.00157	0.00606	0.00078	0.00490
10	0.00037	0.00002	0.00074	0.00054	0.00218	0.00022	0.00148	0.00060	0.00427
15	0.00086	0.00335	0.00117	0.00077	0.00260	0.00074	0.00031	0.00093	0.00226
20	0.00103	0.01348	0.00172	0.00085	0.00329	0.00134	0.00043	0.00147	0.00132
25	0.00118	0.02711	0.00259	0.00112	0.00428	0.00167	0.00055	0.00195	0.00089
35	0.00173	0.04593	0.00453	0.00134	0.00616	0.00274	0.00066	0.00386	0.00059
50	0.00082	0.00318	0.00878	0.00180	0.01120	0.00252	0.00191	0.00346	0.00065
					200				
5	0.00081	0.00007	0.00009	0.00014	0.00078	0.00147	0.00321	0.00057	0.00459
10	0.00050	0.00022	0.00032	0.00023	0.00115	0.00021	0.00066	0.00063	0.00252
15	0.00070	0.00383	0.00067	0.00054	0.00165	0.00068	0.00053	0.00130	0.00092
20	0.00083	0.01327	0.00110	0.00070	0.00214	0.00104	0.00089	0.00198	0.00033
25	0.00061	0.02607	0.00206	0.00123	0.00333	0.00099	0.00073	0.00235	0.00016
35	0.00126	0.04639	0.00269	0.00125	0.00369	0.00203	0.00096	0.00510	0.00031
50	0.00044	0.00360	0.00581	0.00190	0.00742	0.00145	0.00174	0.00341	0.00028

Table A4-*Continued*

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					300				
5	0.00087	0.00036	0.00010	0.00013	0.00064	0.00177	0.00036	0.00061	0.00388
10	0.00014	0.00029	0.00023	0.00018	0.00081	0.00004	0.00015	0.00044	0.00166
15	0.00022	0.00271	0.00038	0.00029	0.00100	0.00030	0.00076	0.00137	0.00091
20	0.00017	0.00870	0.00071	0.00054	0.00143	0.00031	0.00101	0.00193	0.00075
25	0.00009	0.01688	0.00106	0.00072	0.00182	0.00030	0.00094	0.00273	0.00087
35	0.00005	0.02500	0.00164	0.00105	0.00231	0.00034	0.00075	0.00476	0.00100
50	0.00020	0.01195	0.00265	0.00165	0.00343	0.00042	0.00152	0.00349	0.00071
					400				
5	0.00107	0.00121	0.00006	0.00009	0.00045	0.00218	0.00017	0.00078	0.00285
10	0.00007	0.00021	0.00022	0.00017	0.00070	0.00012	0.00046	0.00026	0.00128
15	0.00017	0.00137	0.00038	0.00038	0.00091	0.00016	0.00075	0.00083	0.00092
20	0.00036	0.00422	0.00050	0.00040	0.00100	0.00025	0.00080	0.00152	0.00138
25	0.00053	0.00758	0.00068	0.00052	0.00120	0.00038	0.00068	0.00221	0.00174
35	0.00139	0.00437	0.00103	0.00075	0.00147	0.00092	0.00056	0.00378	0.00191
50	0.00133	0.09722	0.00173	0.00137	0.00225	0.00079	0.00207	0.00263	0.00109
					500				
5	0.00116	0.00193	0.00004	0.00005	0.00033	0.00240	0.00078	0.00084	0.00224
10	0.00019	0.00016	0.00013	0.00013	0.00049	0.00027	0.00082	0.00019	0.00105
15	0.00055	0.00051	0.00028	0.00029	0.00067	0.00042	0.00069	0.00059	0.00110
20	0.00114	0.00143	0.00031	0.00030	0.00067	0.00079	0.00054	0.00118	0.00188
25	0.00215	0.00169	0.00053	0.00044	0.00091	0.00154	0.00032	0.00147	0.00218
35	0.00561	0.00340	0.00078	0.00059	0.00111	0.00396	0.00054	0.00247	0.00230
50	0.00384	0.33925	0.00134	0.00116	0.00174	0.00264	0.00331	0.00167	0.00121

Table A5. Randomized block for MAX DEV - Reduced data set

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					50				
5	0.06306	0.12778	0.02424	0.02424	0.08607	0.09347	0.17514	0.10652	0.13181
10	0.05352	0.18193	0.09885	0.08118	0.18989	0.10176	0.20223	0.15932	0.32518
15	0.06610	0.12315	0.11548	0.08593	0.22473	0.10217	0.20328	0.19633	0.40745
20	0.07334	0.08329	0.15290	0.08652	0.27882	0.14690	0.23699	0.27300	0.46347
25	0.06754	0.10059	0.21116	0.09303	0.33725	0.20944	0.25454	0.36430	0.49188
					75				
5	0.04825	0.08512	0.02973	0.02973	0.08021	0.07651	0.15582	0.08073	0.12782
10	0.03003	0.08347	0.05536	0.05536	0.12346	0.02980	0.12937	0.06533	0.19118
15	0.04058	0.02314	0.08468	0.07790	0.17492	0.04764	0.11146	0.10029	0.28170
20	0.05102	0.07112	0.11012	0.08898	0.18845	0.05205	0.11427	0.12374	0.30433
25	0.02538	0.11276	0.15694	0.06949	0.23577	0.09507	0.13847	0.19169	0.33957
35	0.03549	0.10266	0.19835	0.17124	0.28014	0.16513	0.11794	0.29382	0.31863
50	0.13134	1.45448	0.29123	0.20044	0.36708	0.31005	0.09809	0.47944	0.26096
					100				
5	0.04771	0.05334	0.03695	0.03695	0.08068	0.07451	0.13049	0.07279	0.11618
10	0.03855	0.03793	0.04140	0.04140	0.10578	0.03002	0.08732	0.03715	0.14411
15	0.06428	0.04940	0.06767	0.06767	0.12984	0.05382	0.05288	0.04683	0.18519
20	0.09676	0.20264	0.08278	0.11572	0.16287	0.09958	0.06813	0.08319	0.22684
25	0.05983	0.18093	0.11585	0.09360	0.16859	0.05658	0.05780	0.09319	0.22708
35	0.06275	0.22810	0.15205	0.13522	0.20944	0.08198	0.04312	0.16359	0.21433
50	0.06262	0.42875	0.22038	0.14339	0.26557	0.20372	0.07465	0.31191	0.18671
					150				
5	0.03350	0.02087	0.02167	0.02167	0.05737	0.05392	0.08999	0.04427	0.09895
10	0.03228	0.00996	0.03845	0.03845	0.09026	0.02797	0.04886	0.03944	0.09961
15	0.07348	0.08132	0.04794	0.04737	0.08787	0.04954	0.03528	0.05577	0.08056
20	0.07075	0.16538	0.06302	0.06302	0.10679	0.06432	0.03804	0.06866	0.10326
25	0.06372	0.23028	0.07143	0.07048	0.11788	0.06587	0.04175	0.06163	0.11153
35	0.06725	0.31290	0.09973	0.08427	0.14440	0.07650	0.04944	0.09290	0.11167
50	0.05993	0.07483	0.13991	0.11178	0.17560	0.10951	0.06994	0.14497	0.08645
					200				
5	0.03432	0.01054	0.01612	0.01612	0.04704	0.04911	0.06678	0.03414	0.08636
10	0.04068	0.02453	0.02609	0.02609	0.06144	0.02632	0.04018	0.04218	0.07037
15	0.06175	0.09207	0.03777	0.04333	0.07086	0.04546	0.04017	0.06052	0.04768
20	0.06654	0.17446	0.04629	0.04944	0.07641	0.05804	0.05189	0.07382	0.03315
25	0.05504	0.24218	0.07055	0.07055	0.10381	0.04751	0.03881	0.06778	0.04076
35	0.05547	0.32792	0.07942	0.07931	0.10645	0.06276	0.06329	0.10739	0.03118
50	0.04305	0.07825	0.11713	0.09526	0.14564	0.08587	0.05965	0.10706	0.02847

Table A5-*Continued*

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					300				
5	0.03328	0.02311	0.01586	0.01586	0.04111	0.05503	0.02325	0.03297	0.07654
10	0.01792	0.02236	0.02222	0.02218	0.05023	0.00976	0.01878	0.03520	0.04730
15	0.03062	0.08578	0.03095	0.03095	0.05596	0.03468	0.04308	0.06226	0.05307
20	0.02748	0.15357	0.03959	0.04249	0.06297	0.03292	0.04937	0.06802	0.05196
25	0.01905	0.21638	0.05248	0.05248	0.07433	0.03433	0.05421	0.07294	0.05112
35	0.01495	0.25732	0.05944	0.05944	0.08062	0.03365	0.07468	0.10282	0.06341
50	0.04741	0.31402	0.07958	0.08419	0.09704	0.07991	0.06192	0.10569	0.06394
					400				
5	0.03897	0.04289	0.01298	0.01298	0.03484	0.05951	0.02010	0.03270	0.05971
10	0.01769	0.01939	0.02067	0.02067	0.04889	0.02890	0.02944	0.02739	0.04730
15	0.03342	0.05684	0.02848	0.03869	0.05192	0.02627	0.04208	0.04840	0.05469
20	0.04490	0.11105	0.03300	0.03922	0.04894	0.03628	0.04830	0.06349	0.07242
25	0.05646	0.15618	0.04230	0.04284	0.05548	0.05057	0.05389	0.06662	0.07656
35	0.09187	0.12322	0.04726	0.04521	0.06190	0.08591	0.07207	0.09028	0.08646
50	0.08069	0.85578	0.06149	0.06930	0.07847	0.10290	0.07488	0.09188	0.08050
					500				
5	0.04472	0.05082	0.01070	0.01070	0.03026	0.06436	0.03819	0.03402	0.05331
10	0.02194	0.01620	0.01639	0.01727	0.04118	0.03969	0.03906	0.02280	0.04864
15	0.05434	0.03419	0.02386	0.03221	0.04028	0.04462	0.03685	0.04034	0.05865
20	0.08170	0.06176	0.02592	0.03468	0.03817	0.07019	0.03998	0.05482	0.07819
25	0.10643	0.07933	0.03315	0.03997	0.04612	0.09511	0.04495	0.05895	0.08389
35	0.18501	0.15675	0.04113	0.04694	0.05105	0.15815	0.06126	0.07260	0.10223
50	0.11860	1.64146	0.05669	0.05693	0.06718	0.12895	0.09022	0.07169	0.08375

Table A6. Randomized block for MAPE - Reduced data set

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					50				
5	4.3353	9.6307	1.0452	1.4942	3.1097	4.1427	9.9532	3.5022	8.1535
10	2.6723	11.5718	5.9495	2.6065	7.7380	4.6314	12.7956	7.7356	16.3555
15	3.7650	9.9023	8.3069	2.7370	10.4237	5.1190	11.2254	9.0449	18.5604
20	5.0718	8.5777	10.9136	3.2684	12.7215	5.3938	8.9314	9.8861	18.9020
25	6.2463	10.6577	15.7767	4.2746	16.6708	6.2186	6.0930	11.3588	17.5823
					75				
5	3.4006	6.2478	1.4688	1.1376	3.2222	3.4839	9.7691	2.8104	7.4701
10	1.2694	4.8114	3.7269	2.0565	5.3827	1.9380	7.8882	4.0329	10.6910
15	1.4796	0.9175	5.4893	2.0948	7.1593	2.2539	5.7999	4.3667	11.2940
20	1.5782	6.1658	7.1592	2.8884	8.9246	1.8929	3.5223	3.2336	10.5808
25	1.2239	10.9884	9.6431	3.3153	10.6477	2.6006	2.3003	2.2547	8.9858
35	1.9308	10.1968	17.2545	7.6585	17.0625	3.7378	2.3788	3.1772	3.8935
50	1.4943	63.9410	17.1787	4.6945	17.7574	2.9460	2.6927	2.2060	3.6448
					100				
5	3.4061	3.8133	1.5976	1.4178	3.1492	3.6857	8.6060	2.5179	6.7094
10	1.5312	2.2160	2.8213	1.6650	4.2798	1.3561	5.9554	3.0247	8.2082
15	1.3226	3.8013	4.3454	2.2945	5.7757	1.5869	3.2424	2.5944	7.8376
20	5.4883	13.6339	4.0850	4.5578	4.3840	6.3582	3.5254	5.4047	2.5948
25	2.5931	16.7929	7.3266	2.8586	8.3572	3.5341	1.1710	2.5219	5.7187
35	4.7571	23.0020	11.2881	5.0523	11.5784	5.5646	2.4606	4.7345	2.8858
50	2.1688	15.9518	12.1446	3.0289	12.7070	3.9559	3.4176	3.4852	2.0682
					150				
5	2.8114	1.6996	0.8903	1.3018	2.2117	3.3758	7.0251	2.0652	5.6055
10	1.6168	0.2739	2.3425	1.7108	3.5604	0.9786	3.5820	2.0563	6.0027
15	1.7914	5.7853	3.1901	2.0615	4.4334	2.0372	1.1416	2.0336	4.5002
20	2.5712	12.4957	3.9021	2.2020	5.1293	3.2296	1.8274	3.3314	3.0418
25	3.3048	18.9539	4.9301	2.7334	6.0219	4.0806	2.3491	4.6160	1.9005
35	5.2105	27.4716	7.1114	3.3984	7.6282	6.0672	2.9713	7.5020	1.5086
50	2.4451	4.2213	7.7215	2.6220	8.2657	4.0437	3.3887	4.7397	1.5779
					200				
5	2.6162	0.7114	0.5012	1.0160	1.6738	3.3771	5.1747	1.9618	5.9312
10	1.8488	1.2686	1.5896	1.2364	2.6686	1.0975	2.0481	1.7876	4.7377
15	1.7846	5.9330	2.3606	1.7632	3.4444	2.0677	2.0194	2.7804	3.0147
20	2.0767	11.8046	3.1642	2.0661	4.2327	2.7582	2.8442	3.9765	1.6155
25	2.2681	17.7300	4.2246	2.6860	5.2366	3.0981	2.8701	4.9637	0.9006
35	4.0609	25.6719	5.1942	3.0818	5.7056	4.9527	3.2899	8.1829	1.6482
50	1.6672	4.4105	6.1177	2.5972	6.6180	3.0224	3.4113	4.8488	1.1602

Table A6-*Continued*

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					300				
5	2.7135	1.7326	0.8108	0.8239	1.7910	3.7450	1.6658	2.1860	5.6949
10	1.0307	1.5607	1.3256	1.1027	2.2268	0.4892	1.1131	1.5682	3.8788
15	0.8935	4.7181	1.7480	1.3649	2.6438	1.0942	2.6218	2.9893	2.1986
20	0.9633	9.0141	2.4129	1.8290	3.2910	1.2840	3.1209	4.0183	1.9994
25	0.7182	13.1881	2.9286	2.0473	3.7963	1.4178	3.0767	5.0998	2.5007
35	0.6290	17.3324	3.8946	2.6666	4.4310	1.6991	2.4091	7.3693	2.7937
50	0.6069	7.7173	4.1020	2.5162	4.4711	1.2072	3.2869	4.8285	1.5366
					400				
5	3.0157	3.2351	0.5581	0.8207	1.4217	4.2368	1.0155	2.5792	4.9839
10	0.6637	1.3343	1.3100	1.1010	2.1006	0.5188	1.9898	1.2772	3.2650
15	0.9733	3.3731	1.7431	1.5805	2.5340	1.0858	2.4714	2.3439	2.0989
20	1.5960	6.0583	2.0590	1.6120	2.8346	1.4190	2.6384	3.4317	2.9492
25	2.0794	8.2473	2.3629	1.7554	3.1114	1.6963	2.3791	4.4505	3.5838
35	3.4343	6.6858	3.0144	2.3118	3.4873	2.5682	1.9986	6.3088	3.7555
50	2.6043	22.3938	3.2909	2.4638	3.6066	1.5517	3.7929	4.1654	1.6638
					500				
5	3.0669	4.1480	0.4291	0.5740	1.2112	4.4365	2.4846	2.7017	4.4573
10	1.1203	1.1560	1.0249	0.9375	1.7406	1.0330	2.7107	1.1528	2.6998
15	1.9325	2.0250	1.5245	1.3326	2.2307	1.6406	2.5376	2.0344	2.5631
20	2.7791	3.4350	1.6139	1.3580	2.2907	2.2777	2.0948	3.0111	3.6320
25	4.0834	3.7653	2.1024	1.7159	2.7448	3.2909	1.4571	3.5513	3.9946
35	6.9388	5.4555	2.6271	2.0040	3.0443	5.6179	2.0920	4.9250	4.0388
50	4.6568	41.5592	2.9133	2.3959	3.2094	3.6635	4.8283	3.2980	1.6290

Table A7. Randomized block for MSE - Reduced data set top p/3

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					50				
5	0.00334	0.01016	0.00030	0.00030	0.00401	0.00512	0.00787	0.00607	0.00646
10	0.00068	0.00364	0.00445	0.00270	0.01367	0.00293	0.00823	0.00744	0.03379
15	0.00019	0.00173	0.00718	0.00243	0.01962	0.00247	0.01017	0.00940	0.05041
20	0.00022	0.00262	0.01007	0.00167	0.02378	0.00366	0.00976	0.01362	0.04878
25	0.00097	0.00674	0.01556	0.00112	0.03121	0.00601	0.00883	0.02127	0.04338
					75				
5	0.00200	0.00380	0.00046	0.00046	0.00348	0.00344	0.00775	0.00346	0.00337
10	0.00047	0.00067	0.00159	0.00159	0.00644	0.00038	0.00218	0.00121	0.01129
15	0.00078	0.00012	0.00360	0.00239	0.01117	0.00179	0.00298	0.00285	0.02193
20	0.00081	0.00106	0.00498	0.00249	0.01284	0.00159	0.00209	0.00275	0.02144
25	0.00021	0.00219	0.00810	0.00157	0.01739	0.00229	0.00260	0.00528	0.02159
35	0.00012	0.00029	0.01475	0.00143	0.02610	0.00322	0.00162	0.01057	0.01330
50	0.00323	0.05538	0.02925	0.00562	0.04350	0.00969	0.00236	0.02639	0.00805
					100				
5	0.00187	0.00134	0.00073	0.00073	0.00354	0.00330	0.00641	0.00283	0.00256
10	0.00072	0.00027	0.00097	0.00097	0.00444	0.00053	0.00205	0.00091	0.00519
15	0.00168	0.00041	0.00219	0.00229	0.00724	0.00160	0.00083	0.00147	0.00921
20	0.00143	0.00281	0.00336	0.00192	0.00909	0.00209	0.00076	0.00191	0.01048
25	0.00083	0.00627	0.00532	0.00258	0.01147	0.00155	0.00045	0.00166	0.00988
35	0.00030	0.00695	0.00842	0.00253	0.01562	0.00208	0.00067	0.00318	0.00634
50	0.00073	0.00155	0.01629	0.00225	0.02530	0.00463	0.00265	0.00927	0.00265
					150				
5	0.00087	0.00042	0.00025	0.00036	0.00180	0.00186	0.00504	0.00107	0.00246
10	0.00057	0.00003	0.00082	0.00097	0.00330	0.00041	0.00144	0.00086	0.00190
15	0.00200	0.00068	0.00121	0.00157	0.00409	0.00161	0.00074	0.00204	0.00157
20	0.00196	0.00280	0.00160	0.00170	0.00477	0.00230	0.00063	0.00262	0.00188
25	0.00147	0.00698	0.00241	0.00211	0.00582	0.00196	0.00053	0.00205	0.00163
35	0.00072	0.01128	0.00409	0.00218	0.00816	0.00202	0.00068	0.00205	0.00131
50	0.00031	0.00450	0.00744	0.00286	0.01203	0.00237	0.00258	0.00273	0.00085
					200				
5	0.00063	0.00008	0.00013	0.00017	0.00118	0.00155	0.00309	0.00064	0.00316
10	0.00077	0.00019	0.00029	0.00039	0.00168	0.00038	0.00113	0.00115	0.00127
15	0.00160	0.00041	0.00070	0.00114	0.00265	0.00144	0.00087	0.00277	0.00073
20	0.00171	0.00220	0.00096	0.00143	0.00293	0.00185	0.00094	0.00325	0.00053
25	0.00074	0.00578	0.00220	0.00259	0.00472	0.00106	0.00044	0.00211	0.00032
35	0.00069	0.00834	0.00247	0.00224	0.00481	0.00149	0.00091	0.00265	0.00037
50	0.00021	0.00512	0.00538	0.00318	0.00842	0.00132	0.00185	0.00213	0.00035

Table A7-*Continued*

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					300				
5	0.00100	0.00027	0.00014	0.00019	0.00095	0.00214	0.00028	0.00071	0.00288
10	0.00019	0.00014	0.00024	0.00029	0.00120	0.00007	0.00008	0.00082	0.00170
15	0.00056	0.00013	0.00036	0.00055	0.00157	0.00070	0.00029	0.00279	0.00192
20	0.00038	0.00078	0.00072	0.00110	0.00207	0.00068	0.00031	0.00295	0.00170
25	0.00018	0.00190	0.00112	0.00150	0.00257	0.00056	0.00029	0.00273	0.00161
35	0.00004	0.00288	0.00168	0.00208	0.00313	0.00045	0.00033	0.00271	0.00181
50	0.00037	0.00067	0.00263	0.00278	0.00407	0.00067	0.00073	0.00255	0.00119
					400				
5	0.00122	0.00089	0.00010	0.00012	0.00068	0.00253	0.00021	0.00075	0.00260
10	0.00010	0.00011	0.00024	0.00029	0.00107	0.00022	0.00025	0.00045	0.00174
15	0.00002	0.00009	0.00043	0.00075	0.00147	0.00018	0.00007	0.00166	0.00214
20	0.00003	0.00027	0.00049	0.00078	0.00139	0.00023	0.00008	0.00241	0.00283
25	0.00005	0.00040	0.00070	0.00108	0.00167	0.00028	0.00007	0.00241	0.00301
35	0.00022	0.00020	0.00104	0.00145	0.00192	0.00033	0.00006	0.00258	0.00359
50	0.00144	0.00298	0.00186	0.00235	0.00282	0.00113	0.00047	0.00214	0.00201
					500				
5	0.00147	0.00160	0.00006	0.00007	0.00050	0.00292	0.00094	0.00083	0.00217
10	0.00024	0.00009	0.00015	0.00022	0.00075	0.00048	0.00073	0.00030	0.00177
15	0.00014	0.00007	0.00033	0.00058	0.00108	0.00027	0.00033	0.00115	0.00240
20	0.00012	0.00013	0.00032	0.00061	0.00094	0.00024	0.00007	0.00193	0.00347
25	0.00032	0.00011	0.00054	0.00086	0.00120	0.00034	0.00004	0.00184	0.00368
35	0.00094	0.00037	0.00084	0.00121	0.00147	0.00073	0.00008	0.00192	0.00427
50	0.00328	0.01216	0.00147	0.00197	0.00221	0.00251	0.00095	0.00155	0.00229

Table A8. Randomized block for MAX DEV - Reduced data set top p/3

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					50				
5	0.06306	0.11058	0.02424	0.02424	0.08607	0.09347	0.12370	0.10652	0.10602
10	0.03494	0.09351	0.09885	0.08118	0.18989	0.10176	0.15451	0.15932	0.32518
15	0.02762	0.06539	0.11548	0.08593	0.22473	0.10217	0.20328	0.19633	0.40745
20	0.02410	0.05988	0.15290	0.08652	0.27882	0.14690	0.23699	0.27300	0.46347
25	0.06314	0.10059	0.21116	0.06392	0.33725	0.20944	0.25454	0.36430	0.49188
					75				
5	0.04825	0.06872	0.02973	0.02973	0.08021	0.07651	0.11742	0.08073	0.06168
10	0.03003	0.02945	0.05536	0.05536	0.12346	0.02585	0.06108	0.05973	0.19118
15	0.04058	0.02314	0.08468	0.07790	0.17492	0.04764	0.11146	0.10029	0.28170
20	0.05102	0.04622	0.11012	0.08898	0.18845	0.05205	0.11427	0.12374	0.30433
25	0.02538	0.07123	0.15694	0.06949	0.23577	0.09507	0.13847	0.19169	0.33957
35	0.02310	0.03126	0.19835	0.07495	0.28014	0.16513	0.11794	0.29382	0.31863
50	0.13134	0.36368	0.29123	0.20044	0.36708	0.31005	0.09809	0.47944	0.26096
					100				
5	0.04771	0.04308	0.03695	0.03695	0.08068	0.07451	0.10517	0.07279	0.05929
10	0.03855	0.01909	0.04140	0.04140	0.10578	0.03002	0.05998	0.03494	0.13124
15	0.06428	0.03145	0.06767	0.06767	0.12984	0.05382	0.04394	0.04683	0.18519
20	0.06144	0.06970	0.08050	0.07822	0.16287	0.05979	0.06813	0.07296	0.22684
25	0.05983	0.10412	0.11585	0.09360	0.16859	0.05658	0.05780	0.09319	0.22708
35	0.03216	0.11797	0.15205	0.09963	0.20944	0.08198	0.04312	0.16359	0.21433
50	0.06262	0.08452	0.22038	0.14339	0.26557	0.20372	0.07465	0.31191	0.18671
					150				
5	0.02976	0.02087	0.02167	0.02167	0.05737	0.05392	0.08610	0.04427	0.06989
10	0.03228	0.00996	0.03845	0.03845	0.09026	0.02797	0.04886	0.03944	0.06702
15	0.07348	0.04252	0.04737	0.04737	0.08787	0.04954	0.03528	0.05577	0.08056
20	0.07075	0.08413	0.06302	0.06302	0.10679	0.06432	0.03804	0.06866	0.10326
25	0.06372	0.12454	0.07048	0.07048	0.11788	0.06587	0.04175	0.06163	0.11153
35	0.04348	0.15779	0.09973	0.08156	0.14440	0.06732	0.04308	0.06856	0.11167
50	0.02710	0.07483	0.13991	0.11178	0.17560	0.10951	0.06994	0.14497	0.08645
					200				
5	0.02558	0.01054	0.01612	0.01612	0.04704	0.04911	0.06678	0.03414	0.07488
10	0.04068	0.02453	0.02609	0.02609	0.06144	0.02632	0.04018	0.04218	0.04801
15	0.06175	0.03116	0.03646	0.04333	0.07086	0.04546	0.04017	0.06052	0.03554
20	0.06654	0.07707	0.04436	0.04944	0.07641	0.05804	0.05189	0.07382	0.03315
25	0.05504	0.11239	0.07055	0.07055	0.10381	0.04751	0.03243	0.06377	0.04076
35	0.04790	0.13914	0.07942	0.07931	0.10645	0.05846	0.05628	0.07534	0.03118
50	0.02916	0.07825	0.11713	0.09526	0.14564	0.08587	0.05965	0.07816	0.02847

Table A8-*Continued*

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					300				
5	0.03328	0.01972	0.01586	0.01586	0.04111	0.05503	0.02261	0.03297	0.06601
10	0.01792	0.01694	0.02218	0.02218	0.05023	0.00976	0.01411	0.03520	0.04730
15	0.03062	0.01937	0.03095	0.03095	0.05596	0.03468	0.02307	0.06226	0.05307
20	0.02748	0.05012	0.03799	0.04249	0.06297	0.03292	0.02758	0.06802	0.05196
25	0.01905	0.07039	0.05248	0.05248	0.07433	0.03433	0.02825	0.06717	0.05112
35	0.00998	0.08419	0.05944	0.05944	0.08062	0.03365	0.03360	0.07199	0.06341
50	0.04741	0.04026	0.07958	0.08419	0.09704	0.07991	0.03802	0.07327	0.06394
					400				
5	0.03897	0.03340	0.01298	0.01298	0.03484	0.05951	0.02010	0.03270	0.05971
10	0.01769	0.01612	0.02067	0.02067	0.04889	0.02890	0.02422	0.02739	0.04730
15	0.00605	0.01317	0.02848	0.03869	0.05192	0.02240	0.01558	0.04840	0.05469
20	0.00904	0.03082	0.03206	0.03922	0.04894	0.02318	0.01669	0.06349	0.07242
25	0.01449	0.03797	0.04230	0.04284	0.05548	0.03647	0.01735	0.06662	0.07656
35	0.03310	0.02619	0.04374	0.04521	0.06190	0.05124	0.01415	0.07066	0.08646
50	0.08069	0.11592	0.06149	0.06930	0.07847	0.10290	0.03424	0.06396	0.08050
					500				
5	0.04472	0.04277	0.01070	0.01070	0.03026	0.06436	0.03819	0.03402	0.05331
10	0.02194	0.01449	0.01639	0.01727	0.04118	0.03969	0.03906	0.02280	0.04864
15	0.01769	0.01468	0.02386	0.03221	0.04028	0.03402	0.02820	0.04034	0.05865
20	0.02254	0.01882	0.02481	0.03468	0.03817	0.03769	0.01978	0.05482	0.07819
25	0.03496	0.01609	0.03163	0.03997	0.04612	0.04909	0.01858	0.05895	0.08389
35	0.06237	0.02497	0.03990	0.04694	0.05105	0.07479	0.02320	0.05934	0.10223
50	0.11860	0.22016	0.05669	0.05693	0.06718	0.12895	0.05240	0.05532	0.08375

Table A9. Randomized block for MAPE - Reduced data set top p/3

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					50				
5	4.506	7.993	1.063	1.063	4.159	5.022	6.085	5.035	5.573
10	1.689	4.053	3.593	2.951	5.849	2.511	5.128	3.933	8.874
15	0.716	2.543	4.129	2.449	6.700	2.109	3.901	3.681	9.631
20	0.768	3.129	4.440	1.959	6.431	2.291	2.982	3.700	8.055
25	1.312	4.861	5.191	1.811	6.801	2.451	2.126	3.985	6.406
					75				
5	3.621	5.046	1.390	1.390	4.020	4.295	6.736	3.916	4.710
10	1.461	1.939	2.467	2.467	4.803	1.369	3.295	2.034	5.433
15	1.519	0.415	3.154	2.598	5.485	2.861	2.514	2.761	6.559
20	1.231	2.110	3.690	2.591	5.494	2.441	1.271	2.286	6.095
25	0.785	2.935	4.074	2.029	5.585	2.412	1.445	2.395	5.265
35	0.470	0.959	5.219	1.839	6.494	2.067	1.526	2.527	3.046
50	1.775	13.125	6.529	3.063	7.541	3.269	2.293	3.839	3.434
					100				
5	3.544	3.023	1.840	1.840	4.168	4.323	6.312	3.664	4.191
10	1.910	1.253	1.947	1.947	4.063	1.728	3.383	2.318	3.999
15	2.179	1.392	2.905	3.173	5.041	2.582	1.821	2.664	4.321
20	1.888	3.603	3.217	2.450	4.962	2.959	1.063	2.710	4.245
25	1.316	5.294	3.774	2.745	5.303	2.405	0.793	2.130	3.961
35	0.883	5.192	4.156	2.488	5.464	2.345	1.445	1.900	2.633
50	1.064	1.991	5.301	1.942	6.320	2.779	2.862	2.559	1.394
					150				
5	2.535	1.755	1.135	1.596	3.117	3.479	6.020	2.414	3.441
10	1.792	0.350	1.902	2.358	3.718	1.541	2.853	2.222	2.963
15	2.583	1.753	2.383	2.856	4.255	2.886	1.909	3.252	2.070
20	2.581	3.706	2.428	2.629	4.067	3.196	1.446	3.411	1.834
25	1.956	5.786	2.814	2.826	4.267	2.694	1.195	2.956	1.244
35	1.454	7.026	3.257	2.489	4.498	2.575	1.490	2.559	1.017
50	0.964	4.128	4.033	2.594	5.047	2.444	3.005	2.383	1.315
					200				
5	2.208	0.757	0.752	1.107	2.511	3.262	4.823	1.906	4.579
10	2.091	0.836	1.238	1.588	2.877	1.493	2.734	2.749	2.783
15	2.597	1.408	1.831	2.535	3.526	2.863	2.188	4.014	2.003
20	2.435	3.124	2.094	2.769	3.588	2.983	1.816	4.055	1.650
25	1.433	5.419	2.830	3.442	4.167	2.111	1.260	3.094	1.035
35	1.456	6.216	2.709	2.869	3.856	2.353	1.632	3.151	1.098
50	0.805	4.588	3.646	2.921	4.588	1.964	2.584	2.634	1.120

Table A9-*Continued*

p	LCHF	LLF	NOSE	ANOSE	NAE	REGEXT	NN	REGEXTALL	NNALL
					300				
5	2.823	1.438	0.965	1.208	2.435	4.020	1.330	2.270	4.696
10	1.153	0.891	1.195	1.436	2.587	0.730	0.657	2.404	3.454
15	1.748	0.761	1.328	1.830	2.787	1.963	1.264	4.210	3.462
20	1.403	1.927	1.813	2.505	3.105	1.943	1.173	4.139	3.140
25	0.952	3.135	2.102	2.780	3.272	1.735	1.048	3.808	2.896
35	0.442	3.810	2.544	3.142	3.495	1.385	0.904	3.468	2.629
50	0.793	1.452	2.972	3.089	3.670	1.200	1.655	3.128	1.835
					400				
5	3.144	2.710	0.787	0.973	2.083	4.442	1.091	2.412	4.588
10	0.678	0.802	1.267	1.450	2.499	0.735	1.260	1.761	3.585
15	0.283	0.735	1.596	2.153	2.896	0.981	0.525	3.307	3.778
20	0.380	1.178	1.604	2.143	2.766	1.167	0.577	3.806	4.073
25	0.432	1.317	1.783	2.466	2.840	1.100	0.527	3.678	3.997
35	0.697	0.941	2.213	2.836	2.963	0.891	0.422	3.542	3.861
50	1.789	3.247	2.713	3.241	3.306	1.184	1.369	3.032	2.262
					500				
5	3.443	3.679	0.594	0.744	1.767	4.818	2.660	2.574	4.246
10	1.168	0.734	0.998	1.284	2.122	1.440	2.261	1.447	3.677
15	0.900	0.566	1.436	1.895	2.627	0.823	1.437	2.800	4.016
20	0.712	0.734	1.320	1.894	2.338	0.801	0.420	3.514	4.587
25	1.049	0.696	1.730	2.293	2.590	0.797	0.292	3.268	4.394
35	1.777	1.490	2.093	2.661	2.759	1.261	0.437	3.125	4.261
50	2.993	6.986	2.540	3.179	3.084	2.452	1.947	2.699	2.509

Table A10. R-Squares on full data set

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
1	0.988183	0.995148	0.984908	0.964939	0.953570
2	0.992158	0.996822	0.990605	0.969729	0.972648
3	0.994575	0.997226	0.993808	0.973278	0.981863
4	0.996135	0.996716	0.995656	0.975273	0.987473
5	0.997467	0.995308	0.997352	0.979349	0.990369
6	0.997704	0.992713	0.997801	0.982351	0.989315
7	0.997369	0.990072	0.997623	0.985090	0.986402
8	0.996248	0.987348	0.996696	0.987138	0.980327
9	0.993012	0.982849	0.993498	0.987560	0.982256
10	0.989591	0.978364	0.990079	0.989110	0.977082
11	0.985800	0.976383	0.986265	0.989799	0.971175
12	0.981695	0.971747	0.982255	0.989219	0.966357
13	0.976440	0.976818	0.977492	0.987676	0.962494
14	0.962163	0.962036	0.962629	0.988389	0.963921
15	0.955377	0.968937	0.955790	0.985609	0.963402
16	0.949970	0.963950	0.950541	0.982257	0.965571
17	0.944514	0.967334	0.945622	0.977973	0.968060
18	0.937966	0.966487	0.940317	0.973019	0.971165
19	0.920005	0.952427	0.920115	0.972046	0.975411
20	0.918828	0.962846	0.919144	0.966115	0.978816
21	0.917275	0.972440	0.918115	0.960091	0.981948
22	0.916219	0.987331	0.918022	0.954872	0.983979
23	0.910520	0.984344	0.914142	0.949198	0.985299
24	0.900134	0.978630	0.900495	0.944427	0.989246
25	0.903269	0.982674	0.904198	0.939125	0.990273
26	0.904840	0.981071	0.906618	0.934691	0.990921
27	0.904084	0.978238	0.907102	0.928548	0.991177
28	0.901656	0.969897	0.903930	0.923038	0.991217
29	0.907037	0.980060	0.908345	0.919025	0.994123
30	0.910602	0.979024	0.912187	0.914530	0.994431
31	0.910117	0.973221	0.912595	0.909795	0.994615
32	0.911368	0.974232	0.912832	0.905410	0.994446
33	0.910797	0.965467	0.914896	0.900269	0.994278
34	0.923604	0.990547	0.924848	0.898115	0.996135
35	0.924293	0.986852	0.926111	0.893344	0.995953
36	0.926721	0.986416	0.927631	0.889182	0.995548
37	0.926340	0.972296	0.928993	0.884232	0.995176
38	0.925802	0.972432	0.928826	0.879185	0.994687
39	0.929292	0.983290	0.930606	0.878589	0.995624
40	0.932026	0.988416	0.932778	0.874751	0.994742
41	0.931074	0.982690	0.933121	0.869990	0.993798

Table A10-*Continued*

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
42	0.930397	0.984868	0.932548	0.865154	0.993513
43	0.926192	0.970557	0.928547	0.859605	0.992744
44	0.914171	0.992039	0.915616	0.861330	0.992059
45	0.910448	0.983959	0.913212	0.857173	0.990522
46	0.908013	0.983273	0.910889	0.852165	0.989494
47	0.902718	0.972677	0.905734	0.847625	0.988287
48	0.895350	0.964342	0.894963	0.841878	0.987527
49	0.883222	0.991742	0.885440	0.850330	0.981266
50	0.887123	0.989682	0.889625	0.848577	0.980203
51	0.889252	0.989003	0.892089	0.845047	0.979234
52	0.894732	0.986051	0.893732	0.842082	0.978378
53	0.888305	0.984496	0.892070	0.839161	0.977335
54	0.886794	0.962033	0.891180	0.836027	0.976866
55	0.885252	0.952897	0.887247	0.831783	0.976963
56	0.885904	0.956753	0.881018	0.827976	0.976926
57	0.878064	0.914834	0.870633	0.823248	0.976995
58	0.872124	0.935437	0.857637	0.818003	0.978047
59	0.948371	0.991893	0.950153	0.853532	0.977602
60	0.946984	0.993731	0.948987	0.851786	0.977711
61	0.944892	0.967647	0.947156	0.849677	0.978162
62	0.941676	0.962864	0.944284	0.848970	0.978118
63	0.943226	0.962430	0.940860	0.847125	0.978386
64	0.945598	0.968602	0.936635	0.844970	0.979093
65	0.942152	0.956835	0.930072	0.844687	0.980506
66	0.938211	0.947349	0.922335	0.842287	0.982546
67	0.931817	0.939412	0.911268	0.839500	0.984115
68	0.920468	0.854904	0.895099	0.836761	0.985192
69	0.995831	0.997761	0.995831	0.870882	0.982098
70	0.994573	0.996195	0.994573	0.870831	0.985005
71	0.997808	0.978817	0.993322	0.870682	0.989088
72	0.997407	0.996151	0.991553	0.871528	0.992102
73	0.992654	0.996835	0.990158	0.870322	0.994751
74	0.985094	0.957381	0.988289	0.870360	0.996290
75	0.979135	0.984940	0.985519	0.872054	0.996687
76	0.962907	0.964231	0.982654	0.873868	0.996386
77	0.953871	0.860002	0.974194	0.875948	0.995492
78	0.925383	0.950551	0.954940	0.879158	0.994029

Table A11. MAD on full data set

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
1	0.052939	0.058069	0.091568	0.108807	0.098264
2	0.062905	0.025703	0.093468	0.073856	0.067945
3	0.068117	0.027674	0.089253	0.058972	0.05463
4	0.071528	0.033763	0.088213	0.052535	0.042814
5	0.065425	0.035623	0.078031	0.045562	0.037155
6	0.059918	0.036539	0.069985	0.040767	0.034355
7	0.053564	0.035599	0.061749	0.0365	0.032236
8	0.048141	0.035615	0.055248	0.033369	0.031825
9	0.044637	0.036177	0.050244	0.031388	0.027486
10	0.040834	0.035348	0.045287	0.02966	0.027519
11	0.03838	0.033836	0.042267	0.028208	0.028274
12	0.036984	0.033385	0.040282	0.028224	0.029633
13	0.036243	0.029751	0.039472	0.029114	0.031112
14	0.038575	0.032471	0.040344	0.029539	0.028616
15	0.039865	0.031185	0.041665	0.031732	0.029955
16	0.041305	0.030936	0.042961	0.03364	0.030906
17	0.042307	0.028689	0.044227	0.03611	0.031307
18	0.042878	0.027644	0.044909	0.038775	0.032308
19	0.04788	0.027554	0.049249	0.039375	0.029363
20	0.049831	0.024514	0.051354	0.042185	0.029518
21	0.051063	0.021949	0.052997	0.04498	0.029804
22	0.051602	0.019706	0.053985	0.047218	0.029384
23	0.053847	0.021829	0.05492	0.049198	0.029096
24	0.059944	0.023512	0.061566	0.049937	0.024739
25	0.060579	0.022292	0.06271	0.052461	0.023729
26	0.059996	0.023	0.062681	0.054589	0.022561
27	0.060876	0.025202	0.062534	0.057149	0.021675
28	0.063678	0.029506	0.064258	0.05925	0.0207
29	0.067443	0.025407	0.069639	0.058135	0.016537
30	0.065361	0.024261	0.068416	0.060236	0.015703
31	0.065391	0.025539	0.067478	0.062251	0.014951
32	0.065754	0.025964	0.06736	0.064387	0.015097
33	0.068005	0.029648	0.06788	0.066487	0.015455
34	0.067929	0.021228	0.07135	0.064292	0.013648
35	0.066183	0.020126	0.068819	0.066552	0.014518
36	0.065114	0.020793	0.06734	0.06857	0.016414
37	0.065713	0.025232	0.066529	0.070405	0.018405
38	0.067129	0.027285	0.066825	0.072284	0.019998
39	0.06576	0.020329	0.068448	0.069181	0.021159
40	0.063896	0.019134	0.066284	0.071145	0.02291
41	0.063984	0.023019	0.065115	0.07331	0.024368

Table A11-*Continued*

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
42	0.065182	0.025184	0.064991	0.075125	0.025047
43	0.068768	0.032068	0.067884	0.077059	0.025758
44	0.067463	0.020812	0.069353	0.071545	0.028006
45	0.068936	0.02586	0.069683	0.073454	0.028967
46	0.071035	0.029882	0.070697	0.07556	0.029467
47	0.075224	0.036847	0.073745	0.077433	0.030017
48	0.079535	0.040794	0.079622	0.079148	0.029967
49	0.076712	0.025761	0.079961	0.0705	0.032757
50	0.073109	0.025206	0.076723	0.07235	0.032957
51	0.068465	0.024247	0.072493	0.074532	0.033134
52	0.064098	0.023814	0.068515	0.076501	0.033669
53	0.063941	0.024826	0.066008	0.078174	0.034049
54	0.063942	0.032973	0.06496	0.080049	0.034254
55	0.06599	0.038427	0.066305	0.082058	0.034444
56	0.068534	0.036659	0.069759	0.083449	0.034593
57	0.073443	0.047102	0.076601	0.084964	0.034784
58	0.077499	0.049374	0.085765	0.086206	0.033743
59	0.056326	0.02137	0.061193	0.075658	0.034115
60	0.050954	0.019837	0.054906	0.077262	0.032026
61	0.047714	0.024201	0.050303	0.078722	0.030597
62	0.047324	0.029741	0.049102	0.07969	0.029845
63	0.047538	0.034213	0.049087	0.080717	0.030166
64	0.048825	0.030849	0.051998	0.081753	0.030676
65	0.053063	0.042896	0.0577	0.082	0.030921
66	0.057639	0.043127	0.066879	0.082609	0.030436
67	0.0624	0.054883	0.079194	0.08315	0.03013
68	0.071277	0.067801	0.097327	0.083965	0.030191
69	0.022888	0.013704	0.025058	0.073118	0.023405
70	0.025811	0.025267	0.02606	0.073715	0.020601
71	0.02715	0.027988	0.030486	0.074271	0.01884
72	0.03334	0.032504	0.038976	0.074386	0.019324
73	0.044123	0.050812	0.052988	0.075057	0.021126
74	0.058363	0.058299	0.072256	0.075284	0.025981
75	0.071829	0.069709	0.097155	0.074665	0.030603
76	0.089061	0.094038	0.127819	0.074	0.035176
77	0.104393	0.105428	0.161288	0.073787	0.039583
78	0.125835	0.116446	0.201092	0.073504	0.043009

Table A12. MAPE on full data set

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
1	2.538382	2.829654	4.245703	5.080094	4.874025
2	2.979551	1.467762	4.645465	3.844278	3.989269
3	3.475105	1.759174	4.687533	3.394699	3.61535
4	3.833908	2.215322	4.824558	3.204709	2.87126
5	3.689329	2.392035	4.452395	2.97875	2.589603
6	3.571418	2.482266	4.202934	2.816725	2.51495
7	3.386019	2.492265	3.924583	2.643911	2.490029
8	3.240543	2.545876	3.755454	2.5431	2.666719
9	3.049413	2.563009	3.421641	2.312872	2.014317
10	3.004476	2.597098	3.31642	2.252377	2.145657
11	3.046076	2.564372	3.347878	2.232118	2.369992
12	3.150383	2.60224	3.45221	2.359096	2.712656
13	3.23111	2.372883	3.600864	2.605594	3.132968
14	3.499163	2.63706	3.654968	2.486959	2.468242
15	3.869306	2.653758	4.054325	2.835098	2.786373
16	4.252277	2.718725	4.452905	3.189269	3.09829
17	4.577869	2.597117	4.850782	3.647049	3.418125
18	4.827191	2.759254	5.120043	4.188963	3.872778
19	5.538919	2.663869	5.726729	3.936962	3.033729
20	6.044125	2.469423	6.296609	4.450406	3.276522
21	6.408973	2.449107	6.768451	5.009454	3.565277
22	6.622928	2.238312	7.081785	5.589785	3.862606
23	7.620413	2.695282	7.591981	6.179238	4.22701
24	8.279537	2.96204	8.598947	5.912192	2.985406
25	8.626572	2.846633	9.073954	6.549322	3.073347
26	8.691493	2.950347	9.271076	7.19988	3.168212
27	9.618554	3.556749	9.696441	8.006591	3.385313
28	11.23625	5.658116	11.21086	8.887057	3.654569
29	10.69993	3.897634	11.24851	8.093972	2.240801
30	10.50475	3.649543	11.25271	8.878353	2.306878
31	11.52575	4.888959	11.66983	9.744561	2.374873
32	12.85283	6.095538	13.04813	10.82962	2.648429
33	14.24173	7.720169	14.09172	12.12614	3.0108
34	11.85974	4.880151	12.79996	10.72542	2.284577
35	12.76772	5.374499	12.99063	11.97053	2.695399
36	13.676	6.427546	14.03379	13.28379	3.358973
37	14.7013	8.762928	14.77308	14.94447	4.259382
38	15.50886	9.397493	15.48113	17.17784	5.48997
39	13.29908	6.977262	13.36618	14.16656	5.277041
40	13.4644	6.609338	13.87244	15.82581	6.340718
41	14.19153	8.363717	14.32231	18.35505	7.167859

Table A12-*Continued*

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
42	14.92529	8.721534	14.49022	21.7631	7.610085
43	17.81351	12.42343	17.5562	27.42337	8.083957
44	13.60376	6.035202	13.56036	18.84364	6.818017
45	15.69071	8.734036	15.19912	23.26685	7.273487
46	17.36776	10.17657	16.94172	30.44676	7.763693
47	19.466	12.98805	19.05925	45.87301	8.367931
48	21.40141	14.33832	22.41408	96.96384	9.013923
49	13.38491	4.597255	14.02715	10.57724	4.94939
50	12.96211	4.507667	13.71132	11.05016	5.16035
51	12.25039	4.305636	13.13007	11.64246	5.413034
52	12.02324	4.582798	12.67456	12.2477	5.777433
53	13.0556	4.72342	13.1097	12.8204	6.174817
54	14.49841	8.306659	14.23084	13.51362	6.587719
55	16.44547	11.10381	16.3193	14.22187	7.109289
56	17.92084	10.47179	19.45759	14.85636	7.736992
57	20.17237	14.61132	24.33372	15.59066	8.509554
58	22.03075	15.42647	31.73909	16.42539	9.162635
59	12.00123	4.579843	13.24703	13.76885	7.451766
60	11.55058	4.815516	12.14605	14.50749	7.682721
61	12.26338	7.029	12.30883	15.24783	8.11941
62	14.21336	10.54443	13.91657	15.93468	8.767641
63	15.8668	13.17558	16.40143	16.63478	9.849672
64	17.16282	11.83588	20.49073	17.57798	11.19975
65	19.21228	16.58248	26.49325	18.37026	12.09986
66	21.19209	17.22869	35.39342	19.42903	12.83016
67	22.7725	21.68946	48.54281	20.6594	13.78072
68	26.55001	25.53624	69.80378	22.4764	15.06226
69	8.722019	6.445164	8.050508	18.48751	12.32259
70	13.76339	14.04239	12.19093	19.70798	12.72695
71	15.85439	15.66288	18.92015	21.1437	12.82129
72	19.29365	16.62322	28.65578	22.51272	13.50526
73	24.31913	24.25838	42.93835	24.57687	14.58083
74	30.25894	28.42054	62.86085	26.91381	16.54277
75	35.03002	32.21437	90.49694	29.36431	18.6072
76	41.30139	42.04737	130.9067	32.8894	20.95745
77	46.23708	45.81892	184.887	36.47421	23.51481
78	53.14535	49.37806	275.0229	42.96141	26.26657

Table A13. Randomized block for MSE - Full data set

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.00024	0.00038	0.00271	0.00405	0.01010
10	0.00437	0.00153	0.00974	0.00784	0.03471
15	0.00686	0.00130	0.01261	0.00789	0.03588
20	0.01038	0.00114	0.01664	0.00863	0.03078
25	0.01660	0.00138	0.02331	0.01112	0.02402
30	0.02451	0.00365	0.03209	0.01505	0.01755
35	0.03291	0.00480	0.04133	0.02019	0.01305
40	0.04278	0.00827	0.05192	0.02657	0.01225
45	0.05106	0.01486	0.06067	0.03241	0.01541
50	0.06466	0.02240	0.07509	0.04188	0.02302
			75		
5	0.00039	0.00032	0.00240	0.00232	0.00769
10	0.00171	0.00092	0.00458	0.00190	0.01492
15	0.00326	0.00111	0.00681	0.00217	0.01588
20	0.00530	0.00134	0.00902	0.00149	0.01316
25	0.00849	0.00109	0.01263	0.00221	0.01097
30	0.01198	0.00150	0.01624	0.00304	0.00798
35	0.01617	0.00234	0.02027	0.00403	0.00506
40	0.02038	0.00316	0.02504	0.00562	0.00346
45	0.02478	0.00399	0.02991	0.00779	0.00328
50	0.02935	0.00545	0.03476	0.00978	0.00390
			100		
5	0.00052	0.00049	0.00240	0.00189	0.00621
10	0.00105	0.00059	0.00309	0.00098	0.00859
15	0.00207	0.00108	0.00449	0.00090	0.00772
20	0.00236	0.00237	0.00434	0.00280	0.00424
25	0.00534	0.00131	0.00820	0.00092	0.00511
30	0.00633	0.00140	0.00890	0.00144	0.00307
35	0.00922	0.00200	0.01195	0.00223	0.00249
40	0.01236	0.00200	0.01514	0.00299	0.00195
45	0.01477	0.00261	0.01797	0.00347	0.00138
50	0.01809	0.00306	0.02238	0.00404	0.00126
60	0.02358	0.00435	0.03123	0.00571	0.00198
70	0.03100	0.00666	0.04351	0.01085	0.00371
80	0.03964	0.01066	0.05839	0.02390	0.00579

Table A13-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	0.00017	0.00026	0.00120	0.00078	0.00490
10	0.00074	0.00054	0.00218	0.00060	0.00427
15	0.00117	0.00077	0.00260	0.00093	0.00226
20	0.00172	0.00085	0.00329	0.00147	0.00132
25	0.00259	0.00112	0.00428	0.00195	0.00089
30	0.00363	0.00127	0.00528	0.00282	0.00063
35	0.00453	0.00134	0.00616	0.00386	0.00059
40	0.00605	0.00181	0.00768	0.00424	0.00053
45	0.00744	0.00227	0.00895	0.00451	0.00046
50	0.00914	0.00208	0.01056	0.00457	0.00045
60	0.01285	0.00313	0.01484	0.00405	0.00063
70	0.01643	0.00349	0.02036	0.00435	0.00134
80	0.01969	0.00388	0.02620	0.00813	0.00274
			200		
5	0.00009	0.00014	0.00078	0.00057	0.00459
10	0.00032	0.00023	0.00115	0.00063	0.00252
15	0.00067	0.00054	0.00165	0.00130	0.00092
20	0.00110	0.00070	0.00214	0.00198	0.00033
25	0.00206	0.00123	0.00333	0.00235	0.00016
30	0.00212	0.00105	0.00327	0.00399	0.00030
35	0.00269	0.00125	0.00369	0.00510	0.00031
40	0.00398	0.00160	0.00511	0.00524	0.00022
45	0.00458	0.00175	0.00560	0.00601	0.00025
50	0.00596	0.00219	0.00697	0.00586	0.00020
60	0.00805	0.00242	0.00891	0.00557	0.00056
70	0.01063	0.00295	0.01164	0.00469	0.00171
80	0.01301	0.00354	0.01414	0.00531	0.00376
			300		
5	0.00010	0.00013	0.00064	0.00061	0.00388
10	0.00023	0.00018	0.00081	0.00044	0.00166
15	0.00038	0.00029	0.00100	0.00137	0.00091
20	0.00071	0.00054	0.00143	0.00193	0.00075
25	0.00106	0.00072	0.00182	0.00273	0.00087
30	0.00126	0.00082	0.00196	0.00385	0.00108
35	0.00164	0.00105	0.00231	0.00476	0.00100
40	0.00192	0.00122	0.00248	0.00581	0.00099
45	0.00258	0.00146	0.00316	0.00608	0.00062
50	0.00290	0.00168	0.00342	0.00654	0.00050
60	0.00413	0.00209	0.00460	0.00645	0.00056
70	0.00562	0.00302	0.00605	0.00554	0.00153
80	0.00726	0.00249	0.00760	0.00509	0.00362

Table A13-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.00006	0.00009	0.00045	0.00078	0.00285
10	0.00022	0.00017	0.00070	0.00026	0.00128
15	0.00038	0.00038	0.00091	0.00083	0.00092
20	0.00050	0.00040	0.00100	0.00152	0.00138
25	0.00068	0.00052	0.00120	0.00221	0.00174
30	0.00085	0.00063	0.00132	0.00297	0.00198
35	0.00103	0.00075	0.00147	0.00378	0.00191
40	0.00137	0.00100	0.00180	0.00433	0.00151
45	0.00170	0.00123	0.00206	0.00477	0.00120
50	0.00191	0.00134	0.00229	0.00515	0.00083
60	0.00264	0.00164	0.00293	0.00531	0.00066
70	0.00348	0.00203	0.00373	0.00505	0.00125
80	0.00458	0.00255	0.00478	0.00444	0.00305
			500		
5	0.00004	0.00005	0.00033	0.00084	0.00224
10	0.00013	0.00013	0.00049	0.00019	0.00105
15	0.00028	0.00029	0.00067	0.00059	0.00110
20	0.00031	0.00030	0.00067	0.00118	0.00188
25	0.00053	0.00044	0.00091	0.00147	0.00218
30	0.00059	0.00048	0.00092	0.00211	0.00255
35	0.00078	0.00059	0.00111	0.00247	0.00230
40	0.00102	0.00079	0.00131	0.00287	0.00195
45	0.00121	0.00093	0.00150	0.00319	0.00146
50	0.00147	0.00109	0.00178	0.00335	0.00098
60	0.00189	0.00143	0.00207	0.00382	0.00075
70	0.00260	0.00171	0.00279	0.00368	0.00109
80	0.00323	0.00209	0.00334	0.00379	0.00234
			750		
5	0.00004	0.00006	0.00025	0.00134	0.00074
10	0.00007	0.00007	0.00029	0.00005	0.00055
15	0.00020	0.00021	0.00047	0.00014	0.00094
20	0.00023	0.00023	0.00047	0.00029	0.00172
25	0.00032	0.00032	0.00054	0.00035	0.00217
30	0.00037	0.00034	0.00056	0.00042	0.00238
35	0.00047	0.00043	0.00066	0.00042	0.00215
40	0.00062	0.00053	0.00081	0.00040	0.00171
45	0.00070	0.00058	0.00089	0.00045	0.00129
50	0.00081	0.00067	0.00095	0.00050	0.00101
60	0.00110	0.00093	0.00121	0.00065	0.00058
70	0.00146	0.00121	0.00153	0.00092	0.00075
80	0.00177	0.00138	0.00181	0.00159	0.00148

Table A13-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.00002	0.00003	0.00017	0.00172	0.00017
10	0.00005	0.00006	0.00020	0.00008	0.00025
15	0.00011	0.00012	0.00028	0.00004	0.00072
20	0.00016	0.00018	0.00033	0.00013	0.00121
25	0.00025	0.00026	0.00041	0.00024	0.00142
30	0.00030	0.00029	0.00045	0.00034	0.00149
35	0.00042	0.00039	0.00056	0.00048	0.00125
40	0.00041	0.00039	0.00053	0.00056	0.00111
45	0.00051	0.00046	0.00062	0.00065	0.00081
50	0.00056	0.00051	0.00065	0.00067	0.00060
60	0.00079	0.00070	0.00085	0.00056	0.00035
70	0.00095	0.00083	0.00099	0.00035	0.00049
80	0.00123	0.00107	0.00125	0.00049	0.00112
			1500		
5	0.00000	0.00001	0.00006	0.00243	0.00003
10	0.00003	0.00003	0.00012	0.00043	0.00010
15	0.00007	0.00007	0.00016	0.00045	0.00024
20	0.00012	0.00013	0.00023	0.00121	0.00033
25	0.00015	0.00016	0.00025	0.00231	0.00038
30	0.00019	0.00020	0.00028	0.00380	0.00034
35	0.00024	0.00025	0.00032	0.00560	0.00025
40	0.00029	0.00030	0.00037	0.00737	0.00015
45	0.00033	0.00032	0.00039	0.00885	0.00009
50	0.00039	0.00039	0.00045	0.00995	0.00006
60	0.00049	0.00048	0.00053	0.01001	0.00015
70	0.00062	0.00059	0.00064	0.00714	0.00047
80	0.00075	0.00070	0.00075	0.00325	0.00096
			2000		
5	0.00001	0.00002	0.00009	0.00339	0.00026
10	0.00004	0.00004	0.00012	0.00108	0.00017
15	0.00005	0.00006	0.00014	0.00147	0.00011
20	0.00008	0.00008	0.00015	0.00322	0.00008
25	0.00011	0.00012	0.00018	0.00631	0.00005
30	0.00013	0.00014	0.00019	0.01029	0.00003
35	0.00018	0.00019	0.00024	0.01528	0.00003
40	0.00021	0.00022	0.00027	0.02019	0.00004
45	0.00024	0.00025	0.00028	0.02470	0.00008
50	0.00027	0.00028	0.00029	0.02801	0.00013
60	0.00035	0.00036	0.00038	0.02999	0.00037
70	0.00044	0.00045	0.00045	0.02326	0.00069
80	0.00055	0.00054	0.00055	0.01229	0.00116

Table A14. Randomized block for MAX DEV - Full data set

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.02424	0.02424	0.08607	0.10652	0.13181
10	0.09885	0.08118	0.18989	0.15932	0.32518
15	0.11548	0.08593	0.22473	0.19633	0.40745
20	0.15290	0.08652	0.27882	0.27300	0.46347
25	0.21116	0.09303	0.33725	0.36430	0.49188
30	0.26556	0.13400	0.38109	0.44687	0.48519
35	0.30685	0.15549	0.44821	0.55635	0.49194
40	0.35996	0.20122	0.50053	0.65334	0.48371
45	0.40642	0.24876	0.55853	0.75755	0.48892
50	0.48113	0.32376	0.61308	0.85932	0.50412
			75		
5	0.02973	0.02973	0.08021	0.08073	0.12782
10	0.05536	0.05536	0.12346	0.06533	0.19118
15	0.08468	0.07790	0.17492	0.10029	0.28170
20	0.11012	0.08898	0.18845	0.12374	0.30433
25	0.15694	0.06949	0.23577	0.19169	0.33957
30	0.17182	0.09405	0.26485	0.24766	0.34128
35	0.19835	0.17124	0.28014	0.29382	0.31863
40	0.23313	0.14600	0.30205	0.34910	0.29605
45	0.26779	0.14411	0.34402	0.42647	0.29103
50	0.29123	0.20044	0.36708	0.47944	0.26096
			100		
5	0.03695	0.03695	0.08068	0.07279	0.11618
10	0.04140	0.04140	0.10578	0.03715	0.14411
15	0.06767	0.06767	0.12984	0.04683	0.18519
20	0.08278	0.11572	0.16287	0.08319	0.22684
25	0.11585	0.09360	0.16859	0.09319	0.22708
30	0.12146	0.08399	0.16848	0.11388	0.21166
35	0.15205	0.13522	0.20944	0.16359	0.21433
40	0.18645	0.12562	0.23115	0.21750	0.21377
45	0.19140	0.18500	0.24177	0.25672	0.19367
50	0.22038	0.17400	0.26557	0.31191	0.18671
60	0.24967	0.21681	0.37829	0.40750	0.15588
70	0.30412	0.29415	0.50033	0.51080	0.13839
80	0.35581	0.36339	0.62759	0.63237	0.18981

Table A14-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.02424	0.02424	0.08607	0.10652	0.13181
10	0.09885	0.08118	0.18989	0.15932	0.32518
15	0.11548	0.08593	0.22473	0.19633	0.40745
20	0.15290	0.08652	0.27882	0.27300	0.46347
25	0.21116	0.09303	0.33725	0.36430	0.49188
30	0.26556	0.13400	0.38109	0.44687	0.48519
35	0.30685	0.15549	0.44821	0.55635	0.49194
40	0.35996	0.20122	0.50053	0.65334	0.48371
45	0.40642	0.24876	0.55853	0.75755	0.48892
50	0.48113	0.32376	0.61308	0.85932	0.50412
			75		
5	0.02973	0.02973	0.08021	0.08073	0.12782
10	0.05536	0.05536	0.12346	0.06533	0.19118
15	0.08468	0.07790	0.17492	0.10029	0.28170
20	0.11012	0.08898	0.18845	0.12374	0.30433
25	0.15694	0.06949	0.23577	0.19169	0.33957
30	0.17182	0.09405	0.26485	0.24766	0.34128
35	0.19835	0.17124	0.28014	0.29382	0.31863
40	0.23313	0.14600	0.30205	0.34910	0.29605
45	0.26779	0.14411	0.34402	0.42647	0.29103
50	0.29123	0.20044	0.36708	0.47944	0.26096
			100		
5	0.03695	0.03695	0.08068	0.07279	0.11618
10	0.04140	0.04140	0.10578	0.03715	0.14411
15	0.06767	0.06767	0.12984	0.04683	0.18519
20	0.08278	0.11572	0.16287	0.08319	0.22684
25	0.11585	0.09360	0.16859	0.09319	0.22708
30	0.12146	0.08399	0.16848	0.11388	0.21166
35	0.15205	0.13522	0.20944	0.16359	0.21433
40	0.18645	0.12562	0.23115	0.21750	0.21377
45	0.19140	0.18500	0.24177	0.25672	0.19367
50	0.22038	0.17400	0.26557	0.31191	0.18671
60	0.24967	0.21681	0.37829	0.40750	0.15588
70	0.30412	0.29415	0.50033	0.51080	0.13839
80	0.35581	0.36339	0.62759	0.63237	0.18981

Table A14-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.01298	0.01298	0.03484	0.03270	0.05971
10	0.02067	0.02067	0.04889	0.02739	0.04730
15	0.02848	0.03869	0.05192	0.04840	0.05469
20	0.03300	0.03922	0.04894	0.06349	0.07242
25	0.04230	0.04284	0.05548	0.06662	0.07656
30	0.04338	0.04425	0.05527	0.07684	0.08539
35	0.04726	0.04521	0.06190	0.09028	0.08646
40	0.05709	0.06007	0.07230	0.10010	0.08860
45	0.06146	0.07847	0.07493	0.10702	0.09209
50	0.07581	0.09348	0.07847	0.11008	0.08050
60	0.11567	0.11567	0.09115	0.11248	0.07554
70	0.15390	0.15390	0.12767	0.10841	0.05827
80	0.19259	0.19259	0.16479	0.10878	0.09199
			500		
5	0.01070	0.01070	0.03026	0.03402	0.05331
10	0.01639	0.01727	0.04118	0.02280	0.04864
15	0.02386	0.03221	0.04028	0.04034	0.05865
20	0.02592	0.03468	0.03817	0.05482	0.07819
25	0.03315	0.03997	0.04612	0.05895	0.08389
30	0.03666	0.04195	0.04753	0.06342	0.09682
35	0.04113	0.04694	0.05105	0.07260	0.10223
40	0.04966	0.05446	0.06041	0.08179	0.09713
45	0.04949	0.06478	0.06108	0.08736	0.09558
50	0.06402	0.08018	0.06718	0.08977	0.08375
60	0.09450	0.11072	0.07256	0.09676	0.08956
70	0.12594	0.12594	0.10248	0.09403	0.06520
80	0.15833	0.15833	0.13346	0.09347	0.07680
			750		
5	0.00904	0.01106	0.02501	0.04180	0.03005
10	0.01337	0.01337	0.03183	0.01077	0.03749
15	0.02007	0.02575	0.03604	0.02153	0.05183
20	0.02160	0.02916	0.03321	0.03099	0.06748
25	0.02450	0.03318	0.03521	0.03473	0.08048
30	0.02712	0.03807	0.03580	0.03643	0.09314
35	0.03148	0.03900	0.03931	0.03654	0.09182
40	0.03611	0.04248	0.04489	0.03363	0.09082
45	0.04050	0.04463	0.04986	0.03255	0.08730
50	0.04771	0.05539	0.04856	0.03269	0.08621
60	0.06740	0.08113	0.05394	0.04774	0.07443
70	0.09019	0.10431	0.07103	0.06186	0.06663
80	0.11466	0.11466	0.09436	0.09221	0.05932

Table A14-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.00677	0.00830	0.02059	0.04811	0.01551
10	0.01139	0.01151	0.02448	0.01704	0.02522
15	0.01545	0.01935	0.02611	0.01259	0.03780
20	0.01806	0.02366	0.02702	0.02377	0.05213
25	0.02230	0.02825	0.03052	0.03690	0.05945
30	0.02515	0.03250	0.03381	0.04683	0.06617
35	0.02967	0.03780	0.03864	0.05980	0.06791
40	0.03516	0.03928	0.03594	0.06640	0.06808
45	0.03833	0.04123	0.03834	0.07152	0.06655
50	0.04591	0.04591	0.03962	0.07325	0.06191
60	0.06011	0.06396	0.04366	0.06807	0.05199
70	0.07002	0.08367	0.05342	0.08367	0.03690
80	0.09178	0.10451	0.07420	0.10733	0.05180
			1500		
5	0.00242	0.00350	0.01215	0.05642	0.00650
10	0.00843	0.01028	0.01898	0.04061	0.01666
15	0.01151	0.01425	0.01797	0.04374	0.02549
20	0.01584	0.01863	0.02173	0.08076	0.02819
25	0.01728	0.02207	0.02331	0.11377	0.02990
30	0.01989	0.02488	0.02606	0.14610	0.02892
35	0.02479	0.02826	0.02622	0.17836	0.02715
40	0.03285	0.03285	0.02971	0.20640	0.02236
45	0.03939	0.03939	0.03135	0.22831	0.02071
50	0.04262	0.04262	0.03249	0.24042	0.02012
60	0.05386	0.05386	0.03480	0.24111	0.02299
70	0.06357	0.06357	0.03867	0.20405	0.03289
80	0.07644	0.07644	0.05150	0.13335	0.05183
			2000		
5	0.00585	0.00664	0.01563	0.07272	0.02399
10	0.00919	0.01074	0.01759	0.06109	0.02350
15	0.01209	0.01216	0.01838	0.07094	0.02120
20	0.01346	0.01593	0.01824	0.12414	0.01472
25	0.01479	0.01840	0.01995	0.18190	0.01576
30	0.02119	0.02119	0.02048	0.23800	0.01522
35	0.02268	0.02405	0.02452	0.29030	0.02217
40	0.03058	0.03058	0.02519	0.33572	0.02386
45	0.03616	0.03616	0.02493	0.37429	0.02869
50	0.04261	0.04261	0.02622	0.40125	0.03320
60	0.04871	0.04871	0.02987	0.41731	0.04984
70	0.06276	0.06276	0.03138	0.36917	0.05875
80	0.07035	0.07035	0.04023	0.26113	0.07371

Table A15. Randomized block for MAPE - Full data set

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	1.04523	1.49417	3.10968	3.50215	8.15352
10	5.94953	2.60655	7.73801	7.73563	16.35554
15	8.30694	2.73698	10.42368	9.04492	18.56040
20	10.91358	3.26839	12.72153	9.88605	18.90203
25	15.77666	4.27459	16.67081	11.35884	17.58234
30	22.44148	10.42540	23.50617	14.80065	15.11763
35	29.12485	14.84587	30.33690	20.81710	10.17762
40	35.86950	21.16664	37.10057	32.43553	9.32915
45	43.52954	27.66922	44.57977	55.08664	19.85181
50	51.96585	34.46421	53.56667	167.17373	28.53124
			75		
5	1.46884	1.13761	3.22216	2.81039	7.47008
10	3.72695	2.05655	5.38269	4.03286	10.69100
15	5.48931	2.09485	7.15930	4.36675	11.29402
20	7.15918	2.88841	8.92457	3.23364	10.58082
25	9.64305	3.31533	10.64766	2.25467	8.98580
30	12.96572	4.85338	13.38185	2.58333	7.03283
35	17.25449	7.65853	17.06249	3.17717	3.89355
40	21.17445	9.94008	21.30580	4.36654	1.39759
45	24.99920	12.71224	25.13146	5.66228	5.73340
50	28.89372	16.23446	29.14233	8.67861	11.86953
			100		
5	1.59757	1.41781	3.14915	2.51785	6.70942
10	2.82130	1.66500	4.27981	3.02468	8.20819
15	4.34538	2.29446	5.77571	2.59438	7.83763
20	4.08504	4.55778	4.38404	5.40466	2.59475
25	7.32659	2.85864	8.35725	2.52193	5.71874
30	8.82255	3.66988	9.47018	3.79581	3.96152
35	11.28809	5.05226	11.57837	4.73448	2.88582
40	14.37228	6.12823	14.36987	5.15382	2.05056
45	17.07793	7.78354	17.47436	5.17217	2.01986
50	20.23818	9.94522	22.04443	4.95079	3.76657
60	25.53189	13.41380	34.39567	5.01875	9.52017
70	31.08891	18.24229	57.92223	13.18102	15.80084
80	36.95123	22.97881	105.76665	34.39298	19.84230

Table A15-*Continued*

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	0.89031	1.30182	2.21167	2.06524	5.60546
10	2.34247	1.71084	3.56044	2.05630	6.00269
15	3.19010	2.06147	4.43335	2.03357	4.50018
20	3.90206	2.20200	5.12928	3.33136	3.04178
25	4.93014	2.73343	6.02194	4.61601	1.90047
30	6.01690	3.09833	6.80229	6.10911	1.44134
35	7.11145	3.39845	7.62816	7.50201	1.50857
40	8.70592	4.21800	8.97857	8.31388	1.59309
45	10.28207	5.17915	10.40090	8.84855	1.52333
50	12.04722	5.30456	12.02209	9.15422	1.19218
60	16.04085	7.48537	16.38583	8.35542	2.58125
70	19.74399	9.83124	22.32584	5.78827	5.22083
80	23.04331	12.05650	30.16377	6.04994	9.36518
			200		
5	0.50118	1.01603	1.67375	1.96182	5.93123
10	1.58962	1.23636	2.66863	1.78757	4.73773
15	2.36063	1.76322	3.44438	2.78042	3.01467
20	3.16423	2.06607	4.23271	3.97652	1.61552
25	4.22460	2.68603	5.23662	4.96368	0.90060
30	4.47651	2.75497	5.20226	6.85095	1.41820
35	5.19422	3.08182	5.70562	8.18290	1.64816
40	6.57958	3.63656	7.02879	8.72567	1.26042
45	7.36956	4.01222	7.67402	9.72218	1.17428
50	8.71424	4.59047	8.87020	9.93803	0.94293
60	11.08123	5.37075	10.98692	10.07248	2.12575
70	14.04327	6.77978	14.00392	8.81696	4.50216
80	16.74094	8.52426	16.71111	6.51873	7.36102
			300		
5	0.81080	0.82391	1.79104	2.18597	5.69486
10	1.32561	1.10269	2.22677	1.56818	3.87884
15	1.74801	1.36494	2.64385	2.98926	2.19861
20	2.41288	1.82896	3.29101	4.01834	1.99945
25	2.92861	2.04730	3.79625	5.09985	2.50072
30	3.32712	2.34836	4.01102	6.34454	2.86428
35	3.89463	2.66662	4.43099	7.36926	2.79374
40	4.31371	2.95794	4.70252	8.38648	2.64371
45	5.18118	3.32939	5.44513	8.85795	1.94199
50	5.62576	3.63738	5.84663	9.47897	1.46366
60	7.08421	4.33419	7.16497	9.89438	2.09114
70	8.66587	5.41378	8.64528	9.39057	3.94139
80	10.47753	5.43231	10.34769	8.45396	6.17458

Table A15-*Continued*

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.55814	0.82066	1.42174	2.57922	4.98388
10	1.31000	1.10101	2.10058	1.27716	3.26503
15	1.74311	1.58049	2.53398	2.34395	2.09889
20	2.05898	1.61196	2.83456	3.43169	2.94921
25	2.36288	1.75541	3.11144	4.45051	3.58380
30	2.72641	2.08455	3.32560	5.39479	3.81038
35	3.01437	2.31177	3.48735	6.30883	3.75552
40	3.57800	2.64566	3.91302	6.90451	3.22227
45	4.03884	2.94813	4.30787	7.44207	2.58643
50	4.36388	3.15512	4.59213	7.94023	1.91244
60	5.34166	3.58398	5.45061	8.40079	1.90165
70	6.37621	4.15976	6.40165	8.43927	3.37926
80	7.60705	4.85578	7.57546	7.92909	5.45324
			500		
5	0.42911	0.57399	1.21117	2.70172	4.45728
10	1.02494	0.93746	1.74062	1.15282	2.69978
15	1.52453	1.33258	2.23071	2.03444	2.56311
20	1.61386	1.35800	2.29074	3.01107	3.63204
25	2.10238	1.71592	2.74479	3.55129	3.99459
30	2.24493	1.76967	2.75202	4.42785	4.34684
35	2.62710	2.00404	3.04426	4.92502	4.03885
40	3.01256	2.32292	3.29524	5.47432	3.60483
45	3.35477	2.55900	3.61359	5.86878	2.91570
50	3.75971	2.84521	3.99347	6.15850	2.08543
60	4.33659	3.25246	4.44155	6.80998	1.86688
70	5.26417	3.67273	5.32253	6.90208	3.09505
80	6.04301	4.12233	6.03404	7.10109	4.67048
			750		
5	0.48433	0.62043	1.13368	3.46234	2.59408
10	0.75544	0.71060	1.35064	0.64306	1.83069
15	1.27255	1.19355	1.85697	0.88631	2.59351
20	1.35752	1.21137	1.89791	1.34007	3.62652
25	1.63066	1.45836	2.11487	1.48228	4.05833
30	1.73317	1.49049	2.15001	1.72422	4.22789
35	1.99466	1.71460	2.32765	1.77927	3.95007
40	2.27840	1.90519	2.57644	1.80291	3.38880
45	2.45324	2.00318	2.70709	1.96788	2.78159
50	2.66330	2.16490	2.84268	2.15359	2.16067
60	3.16606	2.52548	3.29204	2.53606	1.66382
70	3.70599	2.93396	3.78329	3.12206	2.49871
80	4.16530	3.17025	4.20446	4.19660	3.66054

Table A15-*Continued*

	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.36870	0.45600	0.93778	3.94309	1.23102
10	0.65133	0.65478	1.19428	0.73271	1.34242
15	0.91381	0.92232	1.43266	0.52235	2.40886
20	1.16129	1.08026	1.62295	1.03231	3.09455
25	1.43980	1.31405	1.85325	1.38865	3.35570
30	1.54294	1.40011	1.90484	1.66519	3.40691
35	1.83503	1.62764	2.13442	1.95700	3.02923
40	1.83231	1.59199	2.06981	2.12418	2.76531
45	2.04499	1.75952	2.24975	2.28330	2.17375
50	2.16453	1.87550	2.33284	2.32265	1.63516
60	2.60409	2.18368	2.72323	2.05473	1.44844
70	2.90652	2.38504	2.98342	1.17128	2.10027
80	3.36840	2.73433	3.41826	1.51197	3.15561
			1500		
5	0.13827	0.24252	0.61064	4.72931	0.48648
10	0.45666	0.47937	0.90565	1.60551	0.81372
15	0.73481	0.70106	1.12297	1.71114	1.36417
20	1.03304	0.99986	1.38497	2.77338	1.64041
25	1.11352	1.05439	1.42071	3.87168	1.78190
30	1.24140	1.18079	1.52952	5.04124	1.66278
35	1.38721	1.29050	1.63082	6.18789	1.37686
40	1.51088	1.39657	1.70998	7.22948	1.01327
45	1.61001	1.47343	1.77166	8.00158	0.69286
50	1.79217	1.60028	1.92116	8.63508	0.67457
60	2.01689	1.79917	2.11919	8.86772	1.15115
70	2.27561	2.01759	2.34019	7.64570	1.99788
80	2.51171	2.16682	2.55785	4.98165	2.76965
			2000		
5	0.32313	0.34626	0.73376	5.50701	1.37819
10	0.52972	0.54373	0.92175	2.60400	1.17466
15	0.64583	0.63709	1.01616	3.17951	0.97633
20	0.81266	0.78305	1.08879	4.67114	0.84039
25	0.95401	0.92283	1.21273	6.47807	0.65732
30	1.00309	0.97427	1.22838	8.27955	0.50371
35	1.18099	1.11974	1.40013	10.19715	0.34377
40	1.30478	1.23567	1.47228	11.80999	0.44258
45	1.38517	1.30284	1.52395	13.15676	0.65565
50	1.44246	1.34671	1.54320	14.13206	0.90936
60	1.67924	1.55255	1.77327	14.93737	1.52432
70	1.87865	1.71567	1.93882	13.46280	2.09225
80	2.10854	1.89306	2.16587	9.99407	2.67321

Table A16. Randomized block for MSE - Full data set top p/3

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.00030	0.00030	0.00401	0.00607	0.00646
10	0.00445	0.00270	0.01367	0.00744	0.03379
15	0.00718	0.00243	0.01962	0.00940	0.05041
20	0.01007	0.00167	0.02378	0.01362	0.04878
25	0.01556	0.00112	0.03121	0.02127	0.04338
30	0.02552	0.00255	0.04490	0.03432	0.03915
35	0.03334	0.00440	0.05450	0.04655	0.03146
40	0.04477	0.01069	0.06760	0.06105	0.03016
45	0.05352	0.02155	0.07910	0.07780	0.03960
50	0.06763	0.03616	0.09503	0.09659	0.05750
			75		
5	0.00046	0.00046	0.00348	0.00346	0.00337
10	0.00159	0.00159	0.00644	0.00121	0.01129
15	0.00360	0.00239	0.01117	0.00285	0.02193
20	0.00498	0.00249	0.01284	0.00275	0.02144
25	0.00810	0.00157	0.01739	0.00528	0.02159
30	0.01188	0.00179	0.02296	0.00810	0.01928
35	0.01475	0.00143	0.02610	0.01057	0.01330
40	0.01870	0.00202	0.03075	0.01461	0.00932
45	0.02456	0.00281	0.03843	0.02158	0.00839
50	0.02925	0.00562	0.04350	0.02639	0.00805
			100		
5	0.00073	0.00073	0.00354	0.00283	0.00256
10	0.00097	0.00097	0.00444	0.00091	0.00519
15	0.00219	0.00229	0.00724	0.00147	0.00921
20	0.00336	0.00192	0.00909	0.00191	0.01048
25	0.00532	0.00258	0.01147	0.00166	0.00988
30	0.00573	0.00229	0.01211	0.00186	0.00721
35	0.00842	0.00253	0.01562	0.00318	0.00634
40	0.01089	0.00191	0.01869	0.00492	0.00495
45	0.01328	0.00213	0.02180	0.00666	0.00339
50	0.01629	0.00225	0.02530	0.00927	0.00265
60	0.02223	0.00400	0.03228	0.01576	0.00308
70	0.03006	0.00850	0.04104	0.02681	0.00658
80	0.04033	0.01581	0.05272	0.04663	0.01261

Table A16-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	0.00025	0.00036	0.00180	0.00107	0.00246
10	0.00082	0.00097	0.00330	0.00086	0.00190
15	0.00121	0.00157	0.00409	0.00204	0.00157
20	0.00160	0.00170	0.00477	0.00262	0.00188
25	0.00241	0.00211	0.00582	0.00205	0.00163
30	0.00356	0.00236	0.00747	0.00191	0.00148
35	0.00409	0.00218	0.00816	0.00205	0.00131
40	0.00547	0.00266	0.00972	0.00181	0.00105
45	0.00660	0.00292	0.01110	0.00212	0.00087
50	0.00744	0.00286	0.01203	0.00273	0.00085
60	0.01042	0.00281	0.01559	0.00508	0.00117
70	0.01416	0.00289	0.01988	0.01007	0.00207
80	0.01761	0.00351	0.02380	0.01938	0.00354
			200		
5	0.00013	0.00017	0.00118	0.00064	0.00316
10	0.00029	0.00039	0.00168	0.00115	0.00127
15	0.00070	0.00114	0.00265	0.00277	0.00073
20	0.00096	0.00143	0.00293	0.00325	0.00053
25	0.00220	0.00259	0.00472	0.00211	0.00032
30	0.00205	0.00196	0.00465	0.00298	0.00055
35	0.00247	0.00224	0.00481	0.00265	0.00037
40	0.00367	0.00280	0.00646	0.00192	0.00030
45	0.00405	0.00259	0.00695	0.00234	0.00035
50	0.00538	0.00318	0.00842	0.00213	0.00035
60	0.00657	0.00312	0.00985	0.00372	0.00097
70	0.00794	0.00285	0.01130	0.00652	0.00259
80	0.01034	0.00317	0.01394	0.01196	0.00551
			300		
5	0.00014	0.00019	0.00095	0.00071	0.00288
10	0.00024	0.00029	0.00120	0.00082	0.00170
15	0.00036	0.00055	0.00157	0.00279	0.00192
20	0.00072	0.00110	0.00207	0.00295	0.00170
25	0.00112	0.00150	0.00257	0.00273	0.00161
30	0.00131	0.00167	0.00279	0.00311	0.00204
35	0.00168	0.00208	0.00313	0.00271	0.00181
40	0.00180	0.00219	0.00306	0.00269	0.00189
45	0.00254	0.00263	0.00403	0.00235	0.00138
50	0.00263	0.00278	0.00407	0.00255	0.00119
60	0.00343	0.00295	0.00509	0.00316	0.00105
70	0.00453	0.00332	0.00631	0.00449	0.00209
80	0.00543	0.00323	0.00730	0.00754	0.00517

Table A16-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.00010	0.00012	0.00068	0.00075	0.00260
10	0.00024	0.00029	0.00107	0.00045	0.00174
15	0.00043	0.00075	0.00147	0.00166	0.00214
20	0.00049	0.00078	0.00139	0.00241	0.00283
25	0.00070	0.00108	0.00167	0.00241	0.00301
30	0.00088	0.00124	0.00178	0.00273	0.00375
35	0.00104	0.00145	0.00192	0.00258	0.00359
40	0.00139	0.00195	0.00231	0.00217	0.00292
45	0.00179	0.00233	0.00265	0.00212	0.00273
50	0.00186	0.00235	0.00282	0.00214	0.00201
60	0.00245	0.00274	0.00336	0.00234	0.00150
70	0.00287	0.00294	0.00388	0.00319	0.00183
80	0.00367	0.00337	0.00472	0.00466	0.00424
			500		
5	0.00006	0.00007	0.00050	0.00083	0.00217
10	0.00015	0.00022	0.00075	0.00030	0.00177
15	0.00033	0.00058	0.00108	0.00115	0.00240
20	0.00032	0.00061	0.00094	0.00193	0.00347
25	0.00054	0.00086	0.00120	0.00184	0.00368
30	0.00067	0.00104	0.00128	0.00214	0.00456
35	0.00084	0.00121	0.00147	0.00192	0.00427
40	0.00110	0.00154	0.00167	0.00169	0.00374
45	0.00129	0.00180	0.00196	0.00165	0.00319
50	0.00147	0.00197	0.00221	0.00155	0.00229
60	0.00183	0.00235	0.00239	0.00174	0.00177
70	0.00236	0.00269	0.00304	0.00205	0.00158
80	0.00280	0.00308	0.00342	0.00306	0.00321
			750		
5	0.00006	0.00009	0.00037	0.00144	0.00075
10	0.00008	0.00012	0.00046	0.00004	0.00102
15	0.00022	0.00039	0.00074	0.00031	0.00171
20	0.00025	0.00047	0.00066	0.00062	0.00261
25	0.00034	0.00063	0.00069	0.00071	0.00323
30	0.00044	0.00072	0.00076	0.00082	0.00397
35	0.00053	0.00087	0.00085	0.00074	0.00372
40	0.00071	0.00105	0.00104	0.00058	0.00315
45	0.00082	0.00120	0.00121	0.00059	0.00265
50	0.00092	0.00127	0.00120	0.00057	0.00222
60	0.00120	0.00170	0.00149	0.00058	0.00126
70	0.00151	0.00207	0.00174	0.00067	0.00110
80	0.00170	0.00221	0.00194	0.00104	0.00197

Table A16-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.00003	0.00005	0.00025	0.00190	0.00018
10	0.00005	0.00008	0.00030	0.00009	0.00039
15	0.00012	0.00023	0.00043	0.00006	0.00096
20	0.00019	0.00036	0.00048	0.00013	0.00146
25	0.00027	0.00050	0.00054	0.00015	0.00181
30	0.00036	0.00059	0.00062	0.00018	0.00220
35	0.00051	0.00080	0.00075	0.00014	0.00200
40	0.00051	0.00082	0.00070	0.00018	0.00187
45	0.00062	0.00093	0.00082	0.00021	0.00155
50	0.00064	0.00099	0.00081	0.00026	0.00119
60	0.00093	0.00133	0.00106	0.00035	0.00061
70	0.00102	0.00146	0.00115	0.00063	0.00053
80	0.00127	0.00177	0.00139	0.00099	0.00139
			1500		
5	0.00000	0.00001	0.00009	0.00271	0.00003
10	0.00003	0.00005	0.00019	0.00078	0.00001
15	0.00007	0.00012	0.00024	0.00044	0.00007
20	0.00014	0.00024	0.00031	0.00039	0.00016
25	0.00018	0.00032	0.00033	0.00039	0.00026
30	0.00025	0.00042	0.00038	0.00049	0.00032
35	0.00030	0.00051	0.00042	0.00059	0.00029
40	0.00038	0.00062	0.00048	0.00078	0.00017
45	0.00044	0.00068	0.00052	0.00098	0.00011
50	0.00052	0.00081	0.00060	0.00123	0.00004
60	0.00062	0.00095	0.00068	0.00188	0.00012
70	0.00073	0.00110	0.00076	0.00257	0.00061
80	0.00090	0.00128	0.00089	0.00360	0.00157
			2000		
5	0.00002	0.00003	0.00014	0.00416	0.00036
10	0.00004	0.00006	0.00018	0.00198	0.00026
15	0.00006	0.00009	0.00021	0.00167	0.00014
20	0.00010	0.00015	0.00020	0.00143	0.00005
25	0.00013	0.00021	0.00022	0.00147	0.00003
30	0.00018	0.00029	0.00025	0.00175	0.00003
35	0.00024	0.00039	0.00033	0.00217	0.00007
40	0.00028	0.00046	0.00034	0.00238	0.00010
45	0.00033	0.00052	0.00037	0.00287	0.00017
50	0.00037	0.00059	0.00039	0.00318	0.00027
60	0.00047	0.00074	0.00049	0.00456	0.00078
70	0.00057	0.00086	0.00056	0.00582	0.00144
80	0.00070	0.00102	0.00069	0.00808	0.00257

Table A17. Randomized block for MAX DEV - Full data set top p/3

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.02424	0.02424	0.08607	0.10652	0.10602
10	0.09885	0.08118	0.18989	0.15932	0.32518
15	0.11548	0.08593	0.22473	0.19633	0.40745
20	0.15290	0.08652	0.27882	0.27300	0.46347
25	0.21116	0.06392	0.33725	0.36430	0.49188
30	0.26556	0.08578	0.38109	0.44687	0.48519
35	0.30685	0.12377	0.44821	0.55635	0.49194
40	0.35996	0.18504	0.50053	0.65334	0.48371
45	0.40642	0.24876	0.55853	0.75755	0.48892
50	0.48113	0.32376	0.61308	0.85932	0.50412
			75		
5	0.02973	0.02973	0.08021	0.08073	0.06168
10	0.05536	0.05536	0.12346	0.05973	0.19118
15	0.08468	0.07790	0.17492	0.10029	0.28170
20	0.11012	0.08898	0.18845	0.12374	0.30433
25	0.15694	0.06949	0.23577	0.19169	0.33957
30	0.17182	0.08234	0.26485	0.24766	0.34128
35	0.19835	0.07495	0.28014	0.29382	0.31863
40	0.23313	0.11755	0.30205	0.34910	0.29605
45	0.26779	0.14411	0.34402	0.42647	0.29103
50	0.29123	0.20044	0.36708	0.47944	0.26096
			100		
5	0.03695	0.03695	0.08068	0.07279	0.05929
10	0.04140	0.04140	0.10578	0.03494	0.13124
15	0.06767	0.06767	0.12984	0.04683	0.18519
20	0.08050	0.07822	0.16287	0.07296	0.22684
25	0.11585	0.09360	0.16859	0.09319	0.22708
30	0.12146	0.08399	0.16848	0.11388	0.21166
35	0.15205	0.09963	0.20944	0.16359	0.21433
40	0.17255	0.08281	0.23115	0.21750	0.21377
45	0.19140	0.11913	0.24177	0.25672	0.19367
50	0.22038	0.14339	0.26557	0.31191	0.18671
60	0.24967	0.21681	0.29871	0.40750	0.15588
70	0.30412	0.29415	0.37131	0.51080	0.13839
80	0.35581	0.36339	0.41141	0.63237	0.18981

Table A17-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	0.02167	0.02167	0.05737	0.04427	0.06989
10	0.03845	0.03845	0.09026	0.03944	0.06702
15	0.04737	0.04737	0.08787	0.05577	0.08056
20	0.06302	0.06302	0.10679	0.06866	0.10326
25	0.07048	0.07048	0.11788	0.06163	0.11153
30	0.08964	0.08494	0.13316	0.05963	0.10541
35	0.09973	0.08156	0.14440	0.06856	0.11167
40	0.11961	0.09858	0.15289	0.09612	0.10790
45	0.13789	0.09393	0.16424	0.11298	0.09064
50	0.13991	0.11178	0.17560	0.14497	0.08645
60	0.16904	0.15442	0.19964	0.21409	0.07866
70	0.20595	0.19268	0.23837	0.29473	0.07826
80	0.22333	0.24833	0.25658	0.36393	0.08292
			200		
5	0.01612	0.01612	0.04704	0.03414	0.07488
10	0.02609	0.02609	0.06144	0.04218	0.04801
15	0.03646	0.04333	0.07086	0.06052	0.03554
20	0.04436	0.04944	0.07641	0.07382	0.03315
25	0.07055	0.07055	0.10381	0.06377	0.04076
30	0.06541	0.06541	0.09811	0.07384	0.04490
35	0.07942	0.07931	0.10645	0.07534	0.03118
40	0.09209	0.09187	0.12306	0.06386	0.03609
45	0.09720	0.08877	0.12461	0.06537	0.03553
50	0.11713	0.09526	0.14564	0.07816	0.02847
60	0.12900	0.12335	0.15670	0.13501	0.04524
70	0.15021	0.16455	0.17555	0.18494	0.07629
80	0.16976	0.20647	0.19375	0.23832	0.10419
			300		
5	0.01586	0.01586	0.04111	0.03297	0.06601
10	0.02218	0.02218	0.05023	0.03520	0.04730
15	0.03095	0.03095	0.05596	0.06226	0.05307
20	0.03799	0.04249	0.06297	0.06802	0.05196
25	0.05248	0.05248	0.07433	0.06717	0.05112
30	0.05141	0.05141	0.07999	0.07535	0.06297
35	0.05944	0.05944	0.08062	0.07199	0.06341
40	0.06462	0.06462	0.07850	0.07512	0.07439
45	0.07375	0.07375	0.09238	0.06934	0.06534
50	0.07958	0.08419	0.09704	0.07327	0.06394
60	0.09045	0.10036	0.10467	0.07649	0.05299
70	0.10927	0.12069	0.12564	0.11311	0.06376
80	0.11686	0.14957	0.13323	0.15246	0.10247

Table A17-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.01298	0.01298	0.03484	0.03270	0.05971
10	0.02067	0.02067	0.04889	0.02739	0.04730
15	0.02848	0.03869	0.05192	0.04840	0.05469
20	0.03206	0.03922	0.04894	0.06349	0.07242
25	0.04230	0.04284	0.05548	0.06662	0.07656
30	0.04224	0.04425	0.05527	0.06857	0.08539
35	0.04374	0.04521	0.06190	0.07066	0.08646
40	0.05709	0.05709	0.07230	0.06406	0.08860
45	0.06146	0.06926	0.07493	0.06436	0.09209
50	0.06149	0.06930	0.07847	0.06396	0.08050
60	0.07071	0.09817	0.08390	0.06560	0.07554
70	0.08214	0.11153	0.09594	0.07819	0.05827
80	0.09209	0.13664	0.10271	0.10878	0.09199
			500		
5	0.01070	0.01070	0.03026	0.03402	0.05331
10	0.01639	0.01727	0.04118	0.02280	0.04864
15	0.02386	0.03221	0.04028	0.04034	0.05865
20	0.02481	0.03468	0.03817	0.05482	0.07819
25	0.03163	0.03997	0.04612	0.05895	0.08389
30	0.03666	0.04195	0.04753	0.05996	0.09682
35	0.03990	0.04694	0.05105	0.05934	0.10223
40	0.04966	0.05446	0.06041	0.05447	0.09713
45	0.04949	0.05419	0.06108	0.05474	0.09558
50	0.05669	0.05693	0.06718	0.05532	0.08375
60	0.06137	0.09256	0.07080	0.05574	0.08956
70	0.07196	0.10204	0.07998	0.06875	0.06520
80	0.08371	0.13548	0.08683	0.08379	0.07680
			750		
5	0.00904	0.01106	0.02501	0.04180	0.03005
10	0.01337	0.01337	0.03183	0.00864	0.03749
15	0.01998	0.02575	0.03604	0.02153	0.05183
20	0.02160	0.02916	0.03321	0.03099	0.06748
25	0.02423	0.03318	0.03521	0.03473	0.08048
30	0.02712	0.03807	0.03580	0.03643	0.09314
35	0.03042	0.03900	0.03931	0.03654	0.09182
40	0.03611	0.04248	0.04489	0.03363	0.09082
45	0.04050	0.04463	0.04986	0.03255	0.08730
50	0.04771	0.04771	0.04856	0.03123	0.08621
60	0.05699	0.07448	0.05394	0.04774	0.07443
70	0.07630	0.09889	0.06002	0.06186	0.06663
80	0.07958	0.10776	0.06314	0.09221	0.05932

Table A17-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.00677	0.00830	0.02059	0.04811	0.01551
10	0.00986	0.01151	0.02448	0.01704	0.02522
15	0.01493	0.01935	0.02611	0.00937	0.03780
20	0.01806	0.02366	0.02702	0.01384	0.05213
25	0.02133	0.02825	0.03052	0.01616	0.05945
30	0.02428	0.03250	0.03381	0.01622	0.06617
35	0.02967	0.03780	0.03864	0.01815	0.06791
40	0.03516	0.03928	0.03594	0.02612	0.06808
45	0.03833	0.04123	0.03834	0.03757	0.06655
50	0.04591	0.04591	0.03962	0.04501	0.06191
60	0.06011	0.06011	0.04366	0.06150	0.05199
70	0.06899	0.08367	0.04822	0.08367	0.03636
80	0.07639	0.09470	0.05175	0.10733	0.05180
			1500		
5	0.00242	0.00350	0.01215	0.05642	0.00650
10	0.00824	0.01028	0.01898	0.04061	0.00433
15	0.01123	0.01425	0.01797	0.03566	0.01234
20	0.01465	0.01863	0.02173	0.03912	0.01826
25	0.01728	0.02207	0.02331	0.04260	0.02283
30	0.01989	0.02488	0.02606	0.04964	0.02737
35	0.02479	0.02826	0.02622	0.05789	0.02715
40	0.03285	0.03285	0.02971	0.06351	0.02236
45	0.03939	0.03939	0.03135	0.07099	0.01988
50	0.04262	0.04262	0.03249	0.08194	0.01444
60	0.05386	0.05386	0.03480	0.09926	0.01632
70	0.06357	0.06357	0.03741	0.11802	0.03289
80	0.07644	0.07644	0.04111	0.13335	0.05183
			2000		
5	0.00585	0.00664	0.01563	0.07272	0.02399
10	0.00919	0.01074	0.01759	0.06109	0.02350
15	0.00993	0.01216	0.01838	0.06255	0.02120
20	0.01290	0.01593	0.01824	0.06286	0.01472
25	0.01473	0.01840	0.01995	0.07071	0.01576
30	0.02119	0.02119	0.02048	0.07613	0.01522
35	0.02268	0.02405	0.02452	0.08806	0.02217
40	0.03058	0.03058	0.02519	0.09384	0.02386
45	0.03616	0.03616	0.02493	0.10206	0.02869
50	0.04261	0.04261	0.02622	0.10942	0.03320
60	0.04871	0.04871	0.02987	0.13078	0.04984
70	0.06276	0.06276	0.03138	0.14380	0.05875
80	0.07035	0.07035	0.03513	0.16278	0.07371

Table A18. Randomized block for MAPE - Full data set top p/3

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	1.06280	1.06280	4.15947	5.03496	5.57346
10	3.59331	2.95150	5.84927	3.93328	8.87381
15	4.12937	2.44861	6.69961	3.68144	9.63058
20	4.43993	1.95879	6.43063	3.70043	8.05524
25	5.19074	1.81084	6.80082	3.98527	6.40646
30	6.14793	2.36253	7.70009	5.12536	5.66797
35	6.72004	2.86663	8.03521	5.60169	5.29404
40	7.91958	4.94300	9.05772	6.20962	6.50946
45	8.10304	7.01144	9.21756	6.62046	8.47050
50	8.88174	9.12059	9.86952	7.03667	10.87149
			75		
5	1.39025	1.39025	4.02023	3.91596	4.71041
10	2.46713	2.46713	4.80336	2.03419	5.43266
15	3.15415	2.59794	5.48515	2.76136	6.55899
20	3.68962	2.59097	5.49367	2.28646	6.09525
25	4.07428	2.02855	5.58458	2.39544	5.26484
30	4.76515	2.09666	6.35230	2.30829	4.62770
35	5.21884	1.83884	6.49429	2.52673	3.04644
40	5.56270	1.87054	6.74320	2.81468	2.04585
45	6.05282	2.12745	7.17504	3.41548	2.74416
50	6.52922	3.06329	7.54087	3.83889	3.43388
			100		
5	1.84024	1.84024	4.16761	3.66400	4.19072
10	1.94731	1.94731	4.06280	2.31804	3.99894
15	2.90475	3.17344	5.04057	2.66363	4.32071
20	3.21720	2.44995	4.96240	2.70979	4.24480
25	3.77433	2.74530	5.30270	2.13022	3.96061
30	3.83285	2.56282	5.32827	1.82659	3.30390
35	4.15649	2.48842	5.46450	1.90039	2.63285
40	4.61680	2.19014	5.73950	2.04280	1.92950
45	4.92730	1.98978	6.03762	2.16971	1.43351
50	5.30093	1.94167	6.32004	2.55884	1.39376
60	5.93971	2.26785	6.81695	3.78110	2.03023
70	6.70593	3.82045	7.44894	5.67294	3.05381
80	7.45268	5.54076	8.13232	8.50860	4.07324

Table A18-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	1.13467	1.59575	3.11670	2.41402	3.44107
10	1.90236	2.35759	3.71830	2.22228	2.96324
15	2.38314	2.85557	4.25498	3.25220	2.07014
20	2.42755	2.62907	4.06669	3.41097	1.83436
25	2.81397	2.82646	4.26720	2.95578	1.24446
30	3.24893	2.73348	4.68272	2.71146	1.20705
35	3.25745	2.48930	4.49795	2.55937	1.01664
40	3.67955	2.71591	4.77430	2.21537	1.03518
45	3.91049	2.72169	5.02047	2.20562	1.22512
50	4.03276	2.59351	5.04744	2.38263	1.31461
60	4.58203	2.35878	5.45220	3.21471	1.81833
70	5.07294	2.11721	5.84043	4.65421	2.48794
80	5.42073	1.97287	6.11089	6.81804	3.25227
			200		
5	0.75178	1.10738	2.51063	1.90643	4.57900
10	1.23827	1.58841	2.87706	2.74885	2.78341
15	1.83146	2.53471	3.52622	4.01360	2.00291
20	2.09403	2.76944	3.58848	4.05549	1.64952
25	2.82984	3.44232	4.16702	3.09439	1.03482
30	2.68255	2.82927	4.00004	3.51784	1.36427
35	2.70926	2.86928	3.85633	3.15065	1.09831
40	3.24280	3.11603	4.31120	2.52676	0.95224
45	3.29918	2.71200	4.34746	2.82318	1.00721
50	3.64605	2.92095	4.58785	2.63402	1.11951
60	3.83711	2.71998	4.70845	3.39863	1.85742
70	4.06659	2.44007	4.81993	4.46278	2.95583
80	4.44612	2.29148	5.13529	6.07509	4.26520
			300		
5	0.96494	1.20795	2.43530	2.26971	4.69621
10	1.19508	1.43597	2.58662	2.40368	3.45408
15	1.32791	1.82969	2.78679	4.20993	3.46173
20	1.81343	2.50521	3.10516	4.13923	3.14034
25	2.10229	2.77989	3.27157	3.80800	2.89604
30	2.33299	2.93686	3.45401	3.85986	3.02943
35	2.54368	3.14204	3.49541	3.46831	2.62895
40	2.58937	3.06300	3.39542	3.26759	2.32855
45	3.06169	3.19417	3.83684	3.03133	1.89533
50	2.97217	3.08858	3.67010	3.12804	1.83546
60	3.22943	2.88518	3.89509	3.50636	1.95447
70	3.45098	2.92893	4.03454	4.15377	2.89077
80	3.71401	2.80728	4.25543	5.31726	4.36569

Table A18-*Continued*

p	NOSE	ANOSE	NAE	REGEXTAL	NNALL
			400		
5	0.78726	0.97292	2.08266	2.41207	4.58811
10	1.26653	1.45018	2.49868	1.76147	3.58475
15	1.59630	2.15346	2.89622	3.30696	3.77795
20	1.60377	2.14338	2.76642	3.80571	4.07268
25	1.78310	2.46632	2.83979	3.67844	3.99742
30	2.11400	2.65369	2.99159	3.77405	4.20351
35	2.21314	2.83595	2.96280	3.54217	3.86141
40	2.42774	3.20071	3.12048	3.12485	3.17382
45	2.75525	3.29911	3.32066	3.06422	2.75211
50	2.71283	3.24082	3.30633	3.03173	2.26189
60	2.97565	3.23144	3.46576	3.15785	2.15256
70	3.04134	3.16087	3.49897	3.66034	2.76664
80	3.34387	3.13308	3.75983	4.38843	4.18121
			500		
5	0.59362	0.74420	1.76671	2.57376	4.24614
10	0.99793	1.28391	2.12173	1.44700	3.67694
15	1.43571	1.89487	2.62731	2.80026	4.01606
20	1.32032	1.89397	2.33805	3.51406	4.58719
25	1.73032	2.29283	2.58968	3.26792	4.39425
30	1.92569	2.50568	2.63649	3.46122	4.74502
35	2.09319	2.66073	2.75914	3.12489	4.26086
40	2.27067	2.90419	2.76270	2.91135	3.73847
45	2.45217	3.09763	2.99797	2.80182	3.24163
50	2.54034	3.17864	3.08379	2.69857	2.50940
60	2.66989	3.22860	3.05173	2.85909	2.14389
70	2.88962	3.25389	3.26459	3.03862	2.60504
80	3.05693	3.23888	3.39794	3.67038	3.74583
			750		
5	0.67645	0.86947	1.65047	3.49876	2.53782
10	0.77165	0.96781	1.71616	0.54189	2.81817
15	1.21295	1.62649	2.22073	1.42457	3.45879
20	1.22415	1.71241	2.04689	2.06167	4.07094
25	1.47021	2.03266	2.07762	2.13966	4.23564
30	1.61457	2.11516	2.14047	2.24988	4.60096
35	1.75341	2.33014	2.21266	2.05045	4.24555
40	1.96222	2.53083	2.35593	1.81966	3.65367
45	2.07477	2.65398	2.51183	1.82143	3.23511
50	2.10844	2.67986	2.42139	1.74967	2.64170
60	2.33855	2.96225	2.64699	1.69508	1.80422
70	2.53264	3.19544	2.79993	1.73793	2.16625
80	2.62462	3.17143	2.87382	2.00700	3.09162

Table A18-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.50300	0.65063	1.35662	4.06727	1.23807
10	0.63662	0.78601	1.46989	0.71612	1.71392
15	0.87138	1.19263	1.74681	0.65851	2.63031
20	1.11310	1.55504	1.80591	0.98484	3.06386
25	1.36447	1.86248	1.89042	1.02665	3.22610
30	1.53180	1.99020	1.96634	1.07524	3.54409
35	1.75973	2.27430	2.09831	0.93329	3.19520
40	1.69282	2.26432	1.98530	1.03003	2.95555
45	1.84743	2.36759	2.13759	0.97412	2.55692
50	1.83218	2.41015	2.10679	1.03193	2.03274
60	2.15606	2.73431	2.38563	0.93682	1.38845
70	2.17029	2.77734	2.38298	1.15573	1.64282
80	2.39583	3.00369	2.58385	1.23586	2.65876
			1500		
5	0.15746	0.25833	0.86601	4.93653	0.45455
10	0.47390	0.62238	1.16874	2.41944	0.26954
15	0.74437	0.92026	1.37382	1.65446	0.64607
20	1.03426	1.34273	1.53525	1.37653	0.95266
25	1.14513	1.53461	1.51115	1.24228	1.23084
30	1.29670	1.70878	1.61772	1.38871	1.37762
35	1.38397	1.83284	1.65702	1.49795	1.21584
40	1.51245	1.99874	1.73258	1.72548	0.91430
45	1.62117	2.07885	1.81068	1.91057	0.64835
50	1.74796	2.24293	1.92072	2.14950	0.42523
60	1.84585	2.38881	2.01566	2.64838	0.83911
70	1.95784	2.53298	2.08617	3.20459	1.92289
80	2.09338	2.67097	2.20444	3.99712	3.02636
			2000		
5	0.38449	0.46094	1.00044	6.09528	1.70817
10	0.57252	0.68540	1.18084	3.96593	1.40614
15	0.61886	0.81542	1.26822	3.42310	0.89767
20	0.86794	1.07320	1.23375	2.99027	0.50098
25	0.96522	1.26682	1.27007	2.84633	0.34252
30	1.09718	1.45073	1.31584	3.08504	0.28984
35	1.26759	1.64883	1.49171	3.28717	0.54624
40	1.33109	1.74088	1.51438	3.38936	0.69950
45	1.42426	1.84816	1.58467	3.72994	0.97040
50	1.47166	1.92351	1.59140	3.89071	1.23967
60	1.64640	2.13656	1.76332	4.71981	2.11885
70	1.73775	2.25186	1.83556	5.43339	2.87354
80	1.90659	2.42393	2.00892	6.62837	3.81770

Table A19. Randomized block for MSE bottom 10%

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.00035	0.00042	0.00251	0.00807	0.08123
10	0.00435	0.00380	0.00280	0.02891	0.11255
15	0.01649	0.00404	0.00529	0.02213	0.07628
20	0.01634	0.01064	0.01208	0.01925	0.04856
25	0.01284	0.01020	0.01113	0.01448	0.02189
30	0.00990	0.00831	0.00870	0.01401	0.00735
35	0.00587	0.00519	0.00508	0.01254	0.00055
40	0.00262	0.00262	0.00262	0.01253	0.00070
45	0.00112	0.00112	0.00112	0.01196	0.00112
50	0.00015	0.00015	0.00015	0.01141	0.00015
			75		
5	0.00018	0.00011	0.00195	0.00304	0.05394
10	0.00194	0.00276	0.00219	0.01205	0.06574
15	0.00807	0.00822	0.00283	0.00745	0.04300
20	0.02302	0.01048	0.00781	0.00502	0.02800
25	0.01839	0.01429	0.01448	0.00315	0.01475
30	0.02084	0.01996	0.01487	0.00295	0.00716
35	0.01712	0.01504	0.01306	0.00210	0.00159
40	0.01469	0.01259	0.01203	0.00272	0.00004
45	0.01092	0.00954	0.00919	0.00284	0.00100
50	0.00823	0.00812	0.00725	0.00379	0.00345
			100		
5	0.00031	0.00003	0.00217	0.00172	0.04244
10	0.00136	0.00197	0.00167	0.00515	0.04136
15	0.00407	0.00413	0.00152	0.00300	0.02720
20	0.03868	0.04086	0.01879	0.00431	0.00157
25	0.02233	0.01583	0.00928	0.00052	0.00998
30	0.02403	0.01845	0.01889	0.00032	0.00507
35	0.02297	0.02316	0.01592	0.00045	0.00207
40	0.02245	0.01865	0.01828	0.00033	0.00050
45	0.01794	0.01543	0.01564	0.00035	0.00001
50	0.01728	0.01466	0.01565	0.00052	0.00059
60	0.01214	0.01124	0.01152	0.00127	0.00489
70	0.00651	0.00651	0.00651	0.00337	0.00577
80	0.00271	0.00271	0.00271	0.00716	0.00271

Table A19-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	0.00015	0.00003	0.00122	0.00021	0.02521
10	0.00055	0.00073	0.00148	0.00161	0.02148
15	0.00254	0.00331	0.00121	0.00030	0.01082
20	0.00706	0.00777	0.00192	0.00038	0.00640
25	0.01065	0.01087	0.00386	0.00123	0.00386
30	0.03377	0.03407	0.00890	0.00218	0.00209
35	0.02520	0.02540	0.01368	0.00323	0.00100
40	0.02617	0.01814	0.02059	0.00295	0.00059
45	0.02533	0.02372	0.01872	0.00295	0.00024
50	0.02641	0.02074	0.02239	0.00224	0.00005
60	0.02538	0.02096	0.02197	0.00114	0.00050
70	0.02090	0.01754	0.01926	0.00024	0.00224
80	0.01690	0.01476	0.01589	0.00099	0.00579
			200		
5	0.00009	0.00001	0.00093	0.00013	0.01723
10	0.00042	0.00062	0.00091	0.00030	0.01070
15	0.00145	0.00190	0.00089	0.00024	0.00481
20	0.00484	0.00640	0.00131	0.00123	0.00191
25	0.00617	0.00678	0.00205	0.00251	0.00140
30	0.01280	0.01344	0.00505	0.00423	0.00061
35	0.01660	0.02320	0.00764	0.00601	0.00035
40	0.02518	0.01277	0.01290	0.00599	0.00024
45	0.02544	0.03310	0.01717	0.00686	0.00010
50	0.02779	0.01853	0.02198	0.00591	0.00004
60	0.02882	0.02479	0.02485	0.00403	0.00005
70	0.02835	0.02235	0.02510	0.00206	0.00054
80	0.02632	0.02192	0.02356	0.00038	0.00177
			300		
5	0.00008	0.00002	0.00067	0.00030	0.00938
10	0.00020	0.00024	0.00081	0.00008	0.00377
15	0.00105	0.00139	0.00053	0.00066	0.00043
20	0.00221	0.00300	0.00059	0.00165	0.00017
25	0.00358	0.00458	0.00119	0.00359	0.00024
30	0.00674	0.00754	0.00242	0.00541	0.00027
35	0.00890	0.00993	0.00385	0.00726	0.00027
40	0.01299	0.01395	0.00602	0.00842	0.00026
45	0.01537	0.05896	0.00785	0.00958	0.00022
50	0.02129	0.01039	0.01173	0.00922	0.00011
60	0.02833	0.03557	0.01979	0.00875	0.00013
70	0.02980	0.02678	0.02393	0.00599	0.00020
80	0.03079	0.02418	0.02680	0.00341	0.00053

Table A19-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.00008	0.00003	0.00062	0.00049	0.00566
10	0.00016	0.00014	0.00075	0.00006	0.00133
15	0.00048	0.00050	0.00072	0.00030	0.00014
20	0.00145	0.00194	0.00044	0.00131	0.00045
25	0.00245	0.00313	0.00081	0.00293	0.00103
30	0.00452	0.00572	0.00157	0.00428	0.00110
35	0.00553	0.00618	0.00221	0.00579	0.00105
40	0.00831	0.00947	0.00359	0.00702	0.00093
45	0.01025	0.01103	0.00504	0.00814	0.00072
50	0.01351	0.01464	0.00704	0.00807	0.00045
60	0.02024	0.01964	0.01196	0.00825	0.00032
70	0.02708	0.02571	0.01725	0.00708	0.00030
80	0.02816	0.02097	0.02337	0.00519	0.00050
			500		
5	0.00003	0.00001	0.00035	0.00073	0.00300
10	0.00009	0.00010	0.00052	0.00008	0.00032
15	0.00046	0.00050	0.00057	0.00019	0.00022
20	0.00100	0.00130	0.00039	0.00063	0.00097
25	0.00179	0.00227	0.00061	0.00169	0.00185
30	0.00311	0.00400	0.00098	0.00255	0.00198
35	0.00409	0.00507	0.00161	0.00343	0.00175
40	0.00609	0.00693	0.00255	0.00438	0.00162
45	0.00701	0.00799	0.00324	0.00509	0.00122
50	0.00974	0.01107	0.00490	0.00530	0.00091
60	0.01473	0.01610	0.00846	0.00619	0.00071
70	0.01917	0.02646	0.01184	0.00576	0.00050
80	0.02483	0.01649	0.01633	0.00542	0.00067
			750		
5	0.00003	0.00001	0.00029	0.00079	0.00083
10	0.00009	0.00009	0.00037	0.00003	0.00005
15	0.00025	0.00026	0.00039	0.00010	0.00092
20	0.00061	0.00078	0.00023	0.00011	0.00208
25	0.00093	0.00114	0.00037	0.00015	0.00249
30	0.00175	0.00222	0.00054	0.00014	0.00267
35	0.00229	0.00284	0.00086	0.00016	0.00245
40	0.00316	0.00394	0.00122	0.00018	0.00189
45	0.00358	0.00444	0.00151	0.00025	0.00153
50	0.00554	0.00676	0.00264	0.00019	0.00146
60	0.00782	0.00892	0.00419	0.00037	0.00107
70	0.01051	0.01152	0.00616	0.00079	0.00091
80	0.01338	0.01441	0.00837	0.00207	0.00107

Table A19-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.00002	0.00001	0.00020	0.00083	0.00016
10	0.00004	0.00003	0.00028	0.00017	0.00022
15	0.00015	0.00016	0.00027	0.00069	0.00110
20	0.00036	0.00046	0.00017	0.00154	0.00184
25	0.00068	0.00081	0.00031	0.00228	0.00209
30	0.00111	0.00139	0.00033	0.00340	0.00213
35	0.00152	0.00186	0.00057	0.00405	0.00192
40	0.00224	0.00278	0.00084	0.00502	0.00169
45	0.00264	0.00323	0.00112	0.00528	0.00139
50	0.00366	0.00449	0.00166	0.00546	0.00127
60	0.00512	0.00621	0.00263	0.00442	0.00095
70	0.00686	0.00783	0.00386	0.00194	0.00097
80	0.00901	0.01029	0.00548	0.00019	0.00127
			1500		
5	0.00000	0.00000	0.00009	0.00082	0.00001
10	0.00003	0.00002	0.00022	0.00094	0.00014
15	0.00011	0.00011	0.00017	0.00375	0.00050
20	0.00023	0.00026	0.00019	0.01010	0.00052
25	0.00037	0.00044	0.00017	0.01609	0.00068
30	0.00063	0.00078	0.00019	0.02569	0.00062
35	0.00084	0.00101	0.00031	0.03384	0.00051
40	0.00128	0.00156	0.00046	0.04404	0.00043
45	0.00145	0.00176	0.00058	0.05013	0.00034
50	0.00200	0.00243	0.00085	0.05631	0.00032
60	0.00296	0.00357	0.00145	0.05424	0.00036
70	0.00396	0.00476	0.00214	0.03750	0.00055
80	0.00506	0.00604	0.00294	0.01459	0.00093
			2000		
5	0.00001	0.00001	0.00011	0.00060	0.00003
10	0.00003	0.00002	0.00019	0.00225	0.00003
15	0.00008	0.00008	0.00013	0.00878	0.00005
20	0.00015	0.00018	0.00009	0.02282	0.00006
25	0.00030	0.00034	0.00015	0.03905	0.00004
30	0.00047	0.00057	0.00016	0.06429	0.00003
35	0.00057	0.00068	0.00021	0.08731	0.00002
40	0.00089	0.00107	0.00031	0.11611	0.00001
45	0.00097	0.00116	0.00037	0.13576	0.00002
50	0.00150	0.00180	0.00063	0.15495	0.00002
60	0.00201	0.00239	0.00096	0.16147	0.00004
70	0.00274	0.00326	0.00144	0.12193	0.00022
80	0.00340	0.00404	0.00192	0.05911	0.00061

Table A20. Randomized block for MAX DEV bottom 10%

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	2.163	2.415	7.024	12.207	39.899
10	14.117	11.271	7.686	38.009	74.917
15	34.005	17.574	19.252	44.657	82.494
20	53.786	39.598	45.450	61.853	97.500
25	63.165	51.415	59.155	72.213	87.466
30	87.266	78.777	81.879	115.121	81.459
35	91.299	87.570	86.770	164.928	28.037
			75		
5	1.177	1.072	5.761	6.697	30.268
10	7.991	7.636	7.203	20.262	47.439
15	17.157	16.903	11.167	19.452	46.911
20	37.206	28.659	22.873	20.674	49.133
25	44.021	37.796	37.804	19.869	43.418
30	64.408	62.681	52.928	25.513	39.875
35	69.437	64.748	60.086	25.928	22.568
40	85.786	78.585	78.986	40.739	4.830
45	88.673	81.121	83.195	52.583	29.920
50	99.352	99.028	96.091	85.734	75.013
			100		
5	2.038	0.526	5.866	4.744	26.011
10	6.004	5.804	5.838	11.827	33.747
15	10.533	10.412	6.902	10.263	32.533
20	35.773	37.677	24.854	12.240	7.399
25	34.106	29.774	22.082	5.625	27.925
30	47.278	40.127	39.412	5.295	23.917
35	54.116	54.384	44.536	7.181	17.719
40	68.087	62.728	60.453	7.459	10.434
45	70.731	65.867	64.990	9.113	1.725
50	84.442	77.688	80.426	14.437	15.967
60	95.770	92.951	93.930	36.521	68.147

Table A20-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	1.229	0.625	4.166	1.500	19.012
10	3.499	3.422	5.483	5.756	21.882
15	6.661	7.297	5.332	2.593	17.378
20	13.051	14.781	6.932	3.178	15.492
25	17.928	18.659	11.694	5.954	13.291
30	35.308	36.964	19.128	10.084	11.035
35	35.429	36.683	25.818	13.556	8.165
40	45.210	34.998	37.629	14.714	7.184
45	49.322	48.142	41.277	15.515	4.829
50	59.972	52.375	53.913	14.917	1.904
60	73.470	63.649	68.184	12.411	10.532
70	83.667	74.419	80.418	8.129	28.740
80	92.217	85.406	90.002	24.422	58.675
			200		
5	1.022	0.265	3.489	1.212	15.285
10	2.829	2.782	4.002	2.040	14.376
15	4.740	5.332	3.721	2.284	10.554
20	9.648	11.395	5.282	5.653	7.386
25	11.663	12.819	7.433	8.533	6.507
30	18.796	20.899	11.827	12.995	4.610
35	23.256	27.837	15.761	16.460	3.440
40	32.749	26.051	23.283	18.392	3.221
45	35.848	41.515	28.766	20.725	1.875
50	44.337	32.595	37.628	21.156	1.338
60	55.741	50.254	50.448	19.980	2.094
70	66.527	57.526	62.045	15.104	9.147
80	75.910	67.544	71.845	7.641	19.804
			300		
5	0.947	0.386	2.891	1.771	10.958
10	1.830	1.810	3.659	1.178	7.953
15	3.492	3.903	2.827	3.350	2.577
20	5.733	6.610	3.573	5.959	1.678
25	7.386	8.456	4.828	9.473	2.275
30	11.326	12.890	7.120	12.768	2.582
35	13.547	15.126	9.331	15.618	2.664
40	18.046	20.531	12.237	17.953	2.734
45	20.823	35.877	14.774	19.922	2.906
50	26.952	21.738	19.860	21.047	2.211
60	36.505	41.674	29.665	22.557	2.676
70	43.815	42.267	38.237	20.665	3.665
80	51.709	45.573	47.338	16.613	6.668

Table A20-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.989	0.568	2.643	2.249	8.351
10	1.491	1.479	3.399	0.914	4.521
15	2.759	2.320	3.262	1.842	1.239
20	4.390	4.872	2.858	4.992	2.656
25	5.685	6.370	3.694	7.956	4.542
30	8.230	9.702	5.270	10.420	5.163
35	9.336	10.559	6.301	12.567	5.168
40	12.685	14.887	8.263	14.630	5.103
45	14.295	16.125	10.237	16.412	4.835
50	17.922	20.449	12.870	17.290	4.060
60	24.539	26.009	18.727	19.030	3.765
70	31.886	33.032	25.198	18.993	3.777
80	37.213	31.721	32.791	17.489	5.161
			500		
5	0.555	0.219	2.006	2.907	6.002
10	1.154	1.146	2.788	0.923	2.018
15	2.636	2.276	2.678	1.490	1.685
20	3.390	3.789	2.576	3.220	4.155
25	4.654	5.159	2.996	5.680	6.037
30	6.320	7.614	3.954	7.431	6.610
35	7.467	8.494	4.981	9.024	6.504
40	9.949	11.571	6.481	10.680	6.526
45	10.859	12.586	7.555	11.856	5.860
50	13.722	15.970	9.754	12.788	5.379
60	18.376	21.001	13.836	14.857	5.141
70	23.037	27.529	18.009	15.363	4.476
80	28.761	26.447	23.166	15.861	5.295
			750		
5	0.626	0.402	1.788	2.961	3.109
10	1.114	1.109	2.219	0.634	0.765
15	1.855	1.620	2.219	0.977	3.659
20	2.490	2.765	1.807	1.085	5.852
25	3.212	3.314	2.356	1.371	6.622
30	4.336	4.945	2.748	1.419	7.181
35	5.072	5.656	3.394	1.605	7.074
40	6.185	7.182	4.171	1.704	6.474
45	6.681	7.782	4.594	2.105	5.954
50	8.682	10.356	6.104	1.928	6.079
60	11.124	13.057	8.180	2.821	5.473
70	13.693	15.707	10.466	4.354	5.188
80	16.663	18.828	13.169	7.975	5.773

Table A20-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.466	0.299	1.459	3.002	1.339
10	0.669	0.667	1.931	1.144	1.620
15	1.346	1.179	1.833	2.879	3.893
20	1.841	2.051	1.384	4.713	5.298
25	2.581	2.580	2.124	5.958	5.846
30	3.244	3.711	2.023	7.605	6.100
35	3.875	4.234	2.749	8.518	5.943
40	4.791	5.555	3.191	9.843	5.773
45	5.277	6.006	3.714	10.286	5.350
50	6.482	7.697	4.494	10.800	5.278
60	8.084	9.638	5.899	10.162	4.772
70	9.965	11.653	7.511	6.840	4.942
80	12.017	14.027	9.404	1.890	5.751
			1500		
5	0.171	0.079	0.968	2.903	0.343
10	0.491	0.490	1.669	3.239	1.186
15	1.081	0.966	1.402	6.894	2.499
20	1.557	1.422	1.347	11.967	2.677
25	1.862	1.851	1.512	15.388	3.119
30	2.259	2.577	1.485	20.113	3.060
35	2.742	2.878	1.923	23.512	2.835
40	3.373	3.745	2.263	27.582	2.688
45	3.588	3.974	2.495	29.888	2.439
50	4.397	4.958	2.995	32.475	2.419
60	5.522	6.337	3.971	33.026	2.653
70	6.600	7.730	4.910	28.381	3.314
80	7.730	9.117	5.949	18.214	4.431
			2000		
5	0.372	0.289	1.059	2.298	0.572
10	0.435	0.435	1.508	5.079	0.567
15	0.913	0.827	1.221	10.418	0.736
20	1.061	1.150	0.899	17.506	0.790
25	1.647	1.640	1.336	23.335	0.690
30	1.977	2.047	1.335	30.818	0.574
35	2.191	2.295	1.617	36.470	0.525
40	2.725	3.010	1.811	43.024	0.440
45	2.864	3.126	1.977	47.140	0.455
50	3.564	4.026	2.406	51.293	0.511
60	4.206	4.735	3.067	54.165	0.779
70	5.102	5.804	3.789	48.357	1.975
80	5.817	6.712	4.483	34.591	3.389

Table A21. Randomized block for MAPE bottom 10%

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			50		
5	0.02546	0.02883	0.05407	0.11633	0.33278
10	0.08715	0.08715	0.07473	0.18917	0.36281
15	0.21480	0.09999	0.10427	0.17990	0.31217
20	0.17500	0.17500	0.17500	0.16032	0.23931
25	0.16500	0.16500	0.16500	0.13988	0.15817
30	0.13400	0.13400	0.13400	0.13018	0.08615
35	0.10800	0.10800	0.10800	0.11830	0.02949
40	0.08000	0.08000	0.08000	0.11264	0.03614
45	0.05900	0.05900	0.05900	0.11488	0.05900
50	0.02400	0.02400	0.02400	0.12109	0.02400
			75		
5	0.01919	0.01482	0.05121	0.07436	0.27073
10	0.05487	0.07434	0.06328	0.12701	0.28185
15	0.14932	0.14932	0.06626	0.11112	0.23976
20	0.24836	0.15036	0.14555	0.09452	0.19302
25	0.21600	0.21600	0.21600	0.08272	0.14592
30	0.20900	0.20900	0.16902	0.07787	0.10184
35	0.18391	0.17124	0.16100	0.06805	0.05028
40	0.15529	0.14600	0.14600	0.07248	0.01162
45	0.13805	0.13200	0.13200	0.07136	0.03301
50	0.12687	0.12430	0.11400	0.07364	0.06731
			100		
5	0.02231	0.00749	0.05468	0.05668	0.23835
10	0.04560	0.06271	0.05472	0.08463	0.22388
15	0.10532	0.10532	0.05541	0.08184	0.20216
20	0.29416	0.29416	0.20513	0.07891	0.05642
25	0.27666	0.22462	0.17841	0.03940	0.12233
30	0.24200	0.24200	0.24200	0.03071	0.09021
35	0.23795	0.23787	0.19300	0.03325	0.06591
40	0.19700	0.19700	0.19700	0.03270	0.03552
45	0.18500	0.18500	0.18500	0.03754	0.00802
50	0.17400	0.17400	0.17400	0.04289	0.02654
60	0.14700	0.14700	0.14697	0.05392	0.08000
70	0.12100	0.12100	0.12100	0.06874	0.10266
80	0.08400	0.08400	0.08400	0.08532	0.08400

Table A21-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			150		
5	0.01662	0.00652	0.04037	0.02038	0.17994
10	0.02430	0.03770	0.04644	0.05147	0.16496
15	0.08294	0.09798	0.04646	0.02748	0.12710
20	0.14091	0.14091	0.06821	0.02933	0.10047
25	0.19133	0.19133	0.11112	0.05005	0.08940
30	0.34500	0.34500	0.17172	0.06292	0.06645
35	0.30200	0.30200	0.22679	0.07655	0.05207
40	0.26300	0.26300	0.26300	0.07339	0.04014
45	0.25392	0.23038	0.22700	0.07695	0.02956
50	0.23200	0.23200	0.23200	0.06626	0.01593
60	0.20896	0.20300	0.20300	0.05194	0.02811
70	0.18700	0.18700	0.18700	0.02734	0.05907
80	0.16900	0.16900	0.16449	0.05090	0.09933
			200		
5	0.01173	0.00405	0.03704	0.01512	0.14879
10	0.02407	0.03501	0.03686	0.02433	0.11715
15	0.06040	0.07351	0.04305	0.02013	0.09159
20	0.11655	0.13020	0.05359	0.04419	0.05932
25	0.14784	0.14784	0.07837	0.06594	0.06121
30	0.20716	0.20716	0.13196	0.08128	0.04065
35	0.25226	0.30752	0.17189	0.09660	0.03706
40	0.30233	0.16765	0.21724	0.09486	0.02930
45	0.30400	0.30400	0.26174	0.10249	0.02230
50	0.27500	0.27500	0.27500	0.09567	0.01472
60	0.25400	0.25400	0.25400	0.08457	0.01006
70	0.23100	0.23100	0.23100	0.06431	0.03176
80	0.21924	0.21661	0.20500	0.03127	0.05949
			300		
5	0.01109	0.00589	0.03111	0.02241	0.10738
10	0.01518	0.02093	0.03214	0.01045	0.07173
15	0.05316	0.06336	0.03163	0.03278	0.03164
20	0.07918	0.09069	0.02935	0.05087	0.02080
25	0.11334	0.12535	0.05661	0.07241	0.02329
30	0.15098	0.15098	0.08958	0.08526	0.02362
35	0.18854	0.18854	0.12293	0.09536	0.02197
40	0.22201	0.22201	0.15253	0.10883	0.02341
45	0.25405	0.56736	0.18099	0.11472	0.01907
50	0.29218	0.16625	0.21577	0.11291	0.01328
60	0.32700	0.32700	0.28166	0.11131	0.01544
70	0.28700	0.28700	0.28700	0.09474	0.02070
80	0.27300	0.27300	0.27300	0.07641	0.03380

Table A21-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			400		
5	0.00998	0.00635	0.03184	0.02890	0.08519
10	0.01567	0.01460	0.03036	0.00890	0.04379
15	0.02816	0.03646	0.03520	0.02338	0.01929
20	0.06358	0.07329	0.02828	0.04401	0.02820
25	0.09333	0.10392	0.04421	0.06410	0.04113
30	0.12543	0.13645	0.07226	0.07551	0.04082
35	0.15014	0.15014	0.09332	0.08810	0.03951
40	0.17830	0.17830	0.11813	0.09777	0.03691
45	0.21149	0.21149	0.14822	0.10434	0.03004
50	0.23717	0.23717	0.17100	0.10533	0.02424
60	0.29625	0.28031	0.22476	0.10576	0.02494
70	0.35152	0.32569	0.27520	0.10212	0.02475
80	0.32100	0.32100	0.32100	0.08739	0.03546
			500		
5	0.00650	0.00333	0.02326	0.03160	0.06018
10	0.01139	0.01283	0.02452	0.01197	0.02359
15	0.02967	0.03666	0.03034	0.02028	0.02077
20	0.05250	0.06085	0.02749	0.03308	0.03844
25	0.07946	0.08880	0.03552	0.04929	0.04923
30	0.10437	0.11436	0.05681	0.06250	0.05329
35	0.13036	0.14068	0.07953	0.06952	0.04720
40	0.15316	0.15316	0.09934	0.07794	0.04518
45	0.17632	0.17632	0.11973	0.08389	0.03937
50	0.20222	0.20222	0.14303	0.08502	0.03434
60	0.25742	0.25742	0.19347	0.09264	0.03133
70	0.30153	0.37710	0.23326	0.09042	0.03008
80	0.34871	0.22386	0.27648	0.08816	0.03799
			750		
5	0.00658	0.00445	0.02101	0.03274	0.03221
10	0.01137	0.01133	0.02210	0.00637	0.00950
15	0.02049	0.02552	0.02554	0.01721	0.03385
20	0.04096	0.04714	0.02167	0.01730	0.05040
25	0.05681	0.06394	0.02576	0.02154	0.05440
30	0.07855	0.08644	0.03972	0.01982	0.05578
35	0.09808	0.10656	0.05658	0.02004	0.05257
40	0.11301	0.12192	0.06907	0.02429	0.04713
45	0.12878	0.13796	0.08257	0.02684	0.04197
50	0.15671	0.16602	0.10838	0.02107	0.04019
60	0.19099	0.19099	0.13878	0.03003	0.03822
70	0.22930	0.22930	0.17356	0.04003	0.03901
80	0.26258	0.26258	0.20360	0.05738	0.04423

Table A21-*Continued*

p	NOSE	ANOSE	NAE	REGEXTALL	NNALL
			1000		
5	0.00500	0.00339	0.01759	0.03413	0.01543
10	0.00786	0.00670	0.01715	0.01824	0.01885
15	0.01560	0.01954	0.02135	0.03949	0.03727
20	0.03110	0.03601	0.01902	0.05394	0.04779
25	0.04874	0.05449	0.02371	0.06507	0.05002
30	0.06282	0.06928	0.02919	0.07855	0.05066
35	0.08011	0.08718	0.04417	0.08705	0.04686
40	0.09516	0.10273	0.05710	0.09545	0.04308
45	0.11133	0.11930	0.07131	0.09981	0.04035
50	0.12791	0.13619	0.08606	0.10038	0.03864
60	0.15711	0.16574	0.11189	0.09399	0.03620
70	0.18599	0.18599	0.13772	0.07168	0.04039
80	0.21839	0.21839	0.16732	0.03046	0.04779
			1500		
5	0.00242	0.00134	0.01215	0.03515	0.00487
10	0.00724	0.00520	0.01498	0.03923	0.01550
15	0.01359	0.01635	0.01729	0.07882	0.02519
20	0.02364	0.02712	0.01873	0.11870	0.02708
25	0.03572	0.03987	0.01831	0.15553	0.02990
30	0.04789	0.05263	0.02043	0.19025	0.02892
35	0.05979	0.06506	0.03045	0.22179	0.02539
40	0.07285	0.07861	0.04178	0.24719	0.02213
45	0.08339	0.08956	0.05072	0.26865	0.02071
50	0.09562	0.10217	0.06145	0.28063	0.02012
60	0.12086	0.12800	0.08394	0.27636	0.02379
70	0.14457	0.15214	0.10516	0.23388	0.02957
80	0.16744	0.17525	0.12574	0.15497	0.04087
			2000		
5	0.00385	0.00307	0.01363	0.03245	0.00610
10	0.00719	0.00564	0.01376	0.05811	0.00571
15	0.01133	0.01346	0.01517	0.11613	0.00936
20	0.01975	0.02246	0.01324	0.17690	0.01081
25	0.03173	0.03498	0.01695	0.23437	0.00848
30	0.04119	0.04494	0.01741	0.29274	0.00728
35	0.04868	0.05289	0.02327	0.34904	0.00765
40	0.06058	0.06521	0.03367	0.39427	0.00564
45	0.06716	0.07217	0.03886	0.43548	0.00792
50	0.08261	0.08798	0.05302	0.45594	0.00606
60	0.10071	0.10669	0.06874	0.46590	0.00992
70	0.12176	0.12824	0.08763	0.40754	0.01978
80	0.13935	0.14622	0.10324	0.29111	0.03251

Table A22. R-Squares for 95th percentile

	NN95	REGEXT95	LCHF95
1	0.97762	0.98078	0.98414
2	0.98917	0.98580	0.98727
3	0.99091	0.98909	0.98978
4	0.99509	0.99081	0.99195
5	0.99293	0.99284	0.99350
6	0.98979	0.99460	0.99472
7	0.98558	0.99626	0.99632
8	0.98123	0.99726	0.99744
9	0.98257	0.99786	0.99789
10	0.98227	0.99785	0.99789
11	0.98214	0.99759	0.99779
12	0.98229	0.99653	0.99692
13	0.98274	0.99489	0.99511
14	0.98316	0.99527	0.99590
15	0.98343	0.99292	0.99381
16	0.98349	0.99111	0.99191
17	0.96778	0.96847	0.96921
18	0.96600	0.96583	0.96601
19	0.95378	0.95301	0.95459
20	0.95525	0.95086	0.95239
21	0.97618	0.97342	0.97485
22	0.98001	0.97012	0.97216
23	0.98006	0.96443	0.96633
24	0.98390	0.94890	0.95184
25	0.98709	0.94297	0.94624
26	0.98644	0.93736	0.94181
27	0.98275	0.93322	0.93739
28	0.98021	0.92942	0.93379
29	0.97740	0.92408	0.92957
30	0.97293	0.91858	0.92334
31	0.97078	0.91594	0.92258
32	0.96726	0.91120	0.91786
33	0.96654	0.90601	0.91449

Table A23. MADs for 95th percentile

	NN95	REGEXT95	LCHF95
1	0.06700	0.10587	0.05256
2	0.04445	0.04503	0.04193
3	0.04327	0.03705	0.03487
4	0.03170	0.03788	0.02678
5	0.02736	0.03625	0.02141
6	0.02196	0.03080	0.01809
7	0.02250	0.02657	0.01372
8	0.02727	0.02242	0.01229
9	0.02714	0.01967	0.00936
10	0.03049	0.01709	0.00977
11	0.03254	0.01706	0.01067
12	0.03381	0.01798	0.01205
13	0.03453	0.01833	0.01586
14	0.03308	0.01989	0.01206
15	0.03253	0.02216	0.01504
16	0.03079	0.02306	0.01800
17	0.03340	0.02974	0.02678
18	0.03399	0.03092	0.03141
19	0.03205	0.03233	0.02735
20	0.03193	0.03461	0.03061
21	0.02697	0.03153	0.02775
22	0.02567	0.03227	0.02974
23	0.02680	0.03504	0.03365
24	0.02559	0.03657	0.02875
25	0.02817	0.03868	0.03178
26	0.03009	0.04187	0.03459
27	0.03292	0.04347	0.03761
28	0.03619	0.04442	0.04013
29	0.03930	0.04599	0.04296
30	0.04213	0.04776	0.04674
31	0.04482	0.04826	0.04822
32	0.04687	0.04959	0.05150
33	0.04821	0.05049	0.05397

Table A24. MAPEs for 95th percentile

	NN95	REGEXT95	LCHF95
1	3.012	4.882	2.646
2	2.664	2.321	2.351
3	3.056	2.223	2.121
4	1.989	2.411	1.628
5	1.836	2.445	1.371
6	1.590	2.127	1.190
7	1.751	1.897	0.945
8	2.304	1.695	0.965
9	1.979	1.412	0.696
10	2.396	1.294	0.809
11	2.770	1.413	1.000
12	3.111	1.655	1.247
13	3.503	1.889	1.719
14	3.020	1.946	1.284
15	3.198	2.342	1.692
16	3.256	2.604	2.104
17	3.819	3.466	3.151
18	4.271	3.933	3.941
19	3.526	3.814	3.241
20	3.788	4.327	3.847
21	3.453	4.229	3.777
22	3.528	4.559	4.312
23	3.973	5.358	5.205
24	2.739	4.668	3.710
25	3.098	5.210	4.327
26	3.437	5.885	4.961
27	3.920	6.409	5.657
28	4.553	6.991	6.454
29	5.201	7.670	7.311
30	5.889	8.598	8.444
31	6.797	9.389	9.307
32	7.665	10.807	10.734
33	8.959	12.202	12.043

Table A25. Randomized block for MSE - 95th percentile

p	NN95	REGEXT95	LCHF95
		50	
5	0.01728	0.01351	0.00577
10	0.01700	0.00494	0.00058
15	0.01270	0.00428	0.00093
20	0.00884	0.00442	0.00133
25	0.00610	0.00527	0.00166
35	0.00499	0.01031	0.00420
50	0.01587	0.02761	0.01467
		75	
5	0.01293	0.00598	0.00204
10	0.00645	0.00126	0.00023
15	0.00371	0.00163	0.00029
20	0.00201	0.00133	0.00034
25	0.00193	0.00179	0.00015
35	0.00145	0.00267	0.00020
50	0.00354	0.00757	0.00172
		100	
5	0.01120	0.00320	0.00089
10	0.00380	0.00105	0.00029
15	0.00111	0.00113	0.00060
20	0.00057	0.00138	0.00088
25	0.00099	0.00189	0.00063
35	0.00098	0.00239	0.00091
50	0.00289	0.00440	0.00075
		125	
5	0.00860	0.00226	0.00074
10	0.00237	0.00084	0.00041
15	0.00063	0.00125	0.00095
20	0.00054	0.00144	0.00085
25	0.00081	0.00209	0.00131
35	0.00085	0.00259	0.00148
50	0.00288	0.00332	0.00074
		150	
5	0.00660	0.00188	0.00069
10	0.00154	0.00059	0.00058
15	0.00045	0.00090	0.00121
20	0.00064	0.00136	0.00121
25	0.00096	0.00199	0.00145
35	0.00081	0.00225	0.00157
50	0.00601	0.00730	0.00559

Table A25-*Continued*

p	NN95	REGEXT95	LCHF95
		175	
5	0.00522	0.00187	0.00065
10	0.00107	0.00052	0.00052
15	0.00047	0.00083	0.00073
20	0.00082	0.00125	0.00091
25	0.00094	0.00152	0.00105
35	0.00078	0.00195	0.00134
50	0.00236	0.00208	0.00058
		200	
5	0.00353	0.00159	0.00069
10	0.00072	0.00041	0.00048
15	0.00064	0.00080	0.00074
20	0.00091	0.00102	0.00077
25	0.00105	0.00128	0.00078
35	0.00076	0.00145	0.00098
50	0.00236	0.00149	0.00038
		300	
5	0.00076	0.00139	0.00070
10	0.00025	0.00024	0.00015
15	0.00070	0.00038	0.00019
20	0.00091	0.00037	0.00022
25	0.00112	0.00046	0.00021
35	0.00065	0.00039	0.00004
50	0.00236	0.00072	0.00022
		400	
5	0.00039	0.00173	0.00107
10	0.00034	0.00016	0.00008
15	0.00049	0.00017	0.00010
20	0.00048	0.00031	0.00024
25	0.00039	0.00048	0.00039
35	0.00057	0.00149	0.00115
50	0.00347	0.00141	0.00127
		500	
5	0.00076	0.00187	0.00133
10	0.00058	0.00034	0.00024
15	0.00041	0.00041	0.00037
20	0.00023	0.00086	0.00088
25	0.00021	0.00172	0.00166
35	0.00100	0.00503	0.00467
50	0.00516	0.00325	0.00331

Table A26. Randomized block for MAX DEV - 95th percentile

p	NN95	REGEXT95	LCHF95
		50	
5	0.179946	0.195301	0.110518
10	0.243867	0.179805	0.04522
15	0.306782	0.20237	0.057803
20	0.329138	0.245702	0.06988
25	0.327761	0.298779	0.066586
35	0.357679	0.474345	0.188695
50	0.450163	0.74102	0.391821
		75	
5	0.160368	0.130424	0.060176
10	0.112666	0.08194	0.026527
15	0.173582	0.104465	0.033019
20	0.170457	0.109562	0.05214
25	0.184583	0.157658	0.025191
35	0.153208	0.232738	0.036377
50	0.154455	0.415921	0.151743
		100	
5	0.146682	0.095206	0.03456
10	0.07598	0.067276	0.028558
15	0.094299	0.061035	0.060966
20	0.086735	0.073784	0.08176
25	0.119169	0.110586	0.042733
35	0.076037	0.149377	0.055785
50	0.094704	0.284045	0.07034
		125	
5	0.123961	0.078737	0.030337
10	0.058905	0.049411	0.032957
15	0.050741	0.063481	0.068553
20	0.051213	0.070018	0.077335
25	0.046429	0.07391	0.073565
35	0.044564	0.117488	0.061676
50	0.090023	0.223322	0.054976
		150	
5	0.105928	0.070945	0.028857
10	0.058367	0.042581	0.044032
15	0.044946	0.059703	0.081085
20	0.039746	0.064746	0.092254
25	0.049701	0.091687	0.085068
35	0.053986	0.07515	0.06141
50	0.180082	0.214093	0.212211

Table A26-*Continued*

p	NN95	REGEXT95	LCHF95
		175	
5	0.094014	0.071876	0.028335
10	0.05345	0.040195	0.042051
15	0.040764	0.054765	0.066112
20	0.045767	0.064264	0.070138
25	0.04547	0.061489	0.070612
35	0.056296	0.074935	0.060172
50	0.077484	0.159074	0.04913
		200	
5	0.075905	0.06373	0.027871
10	0.043181	0.032956	0.040474
15	0.039508	0.04917	0.065364
20	0.046293	0.060083	0.069639
25	0.04732	0.058621	0.051882
35	0.060554	0.062412	0.050246
50	0.066478	0.135645	0.042995
		300	
5	0.037165	0.058399	0.029515
10	0.022614	0.024977	0.015388
15	0.044869	0.035984	0.028181
20	0.049976	0.0373	0.035883
25	0.09099	0.048132	0.040514
35	0.074551	0.06973	0.015067
50	0.079237	0.133013	0.054358
		400	
5	0.033278	0.064403	0.042666
10	0.032666	0.033952	0.017451
15	0.037849	0.027562	0.023223
20	0.040307	0.040819	0.040011
25	0.046417	0.054734	0.048533
35	0.063229	0.103309	0.083874
50	0.097161	0.138231	0.08327
		500	
5	0.045247	0.063573	0.047076
10	0.050488	0.048197	0.026149
15	0.044385	0.05019	0.042823
20	0.037953	0.064319	0.067819
25	0.035822	0.097125	0.091598
35	0.052259	0.170186	0.161944
50	0.109333	0.149222	0.110074

Table A27. Randomized block for MAPE - 95th percentile

p	NN95	REGEXT95	LCHF95
		50	
5	10.146	5.539	5.007
10	10.201	3.766	1.687
15	8.446	4.066	3.054
20	6.367	4.118	4.344
25	4.568	4.872	5.608
35	2.608	11.113	10.390
50	3.311	7.051	7.442
		75	
5	8.496	3.887	3.351
10	6.773	2.259	0.943
15	4.332	2.089	1.107
20	2.099	1.988	1.224
25	2.350	2.466	1.158
35	2.217	3.668	1.406
50	3.504	2.382	1.279
		100	
5	8.021	2.902	2.375
10	5.436	2.001	1.234
15	2.208	2.130	1.126
20	1.325	2.821	1.645
25	1.899	3.705	2.360
35	2.661	4.627	3.978
50	3.685	3.480	1.968
		125	
5	7.369	2.782	2.268
10	4.294	1.834	1.534
15	1.418	2.416	1.717
20	1.852	3.143	2.061
25	2.637	4.208	3.189
35	2.933	5.266	4.749
50	3.864	3.601	2.236
		150	
5	6.602	2.722	2.226
10	3.324	1.618	1.741
15	1.466	2.071	1.968
20	2.354	3.050	2.264
25	3.001	4.037	3.189
35	2.933	5.017	4.616
50	5.567	5.725	4.518

Table A27-*Continued*

p	NN95	REGEXT95	LCHF95
		175	
5	5.845	2.639	2.168
10	2.525	1.672	1.687
15	1.760	2.136	1.597
20	2.661	2.916	2.059
25	3.069	3.578	2.778
35	2.902	4.519	4.108
50	3.728	3.081	1.958
		200	
5	4.889	2.744	2.283
10	1.829	1.466	1.643
15	2.120	2.057	1.624
20	2.799	2.628	1.857
25	3.205	3.245	2.446
35	2.708	3.746	3.394
50	3.848	2.555	1.528
		300	
5	2.190	2.720	2.313
10	1.311	1.117	1.068
15	2.348	1.324	0.697
20	2.777	1.324	0.816
25	2.948	1.631	0.927
35	2.155	1.454	0.527
50	3.975	0.826	0.623
		400	
5	1.302	3.193	2.826
10	1.472	0.635	0.630
15	1.872	1.046	0.748
20	1.958	1.399	1.235
25	1.660	1.739	1.643
35	2.204	3.179	3.105
50	4.820	2.040	2.433
		500	
5	1.890	3.501	3.163
10	1.906	0.910	1.212
15	1.674	1.401	1.565
20	1.073	2.233	2.450
25	1.161	3.282	3.467
35	2.831	6.173	6.167
50	5.930	3.944	4.229

Table A28. Deleted observations from simulation study

Sample Size	B(2, 6) p0init	p1init	r1init	s1init	p0final	p1final	r1final	s1final
200_200	NONE							
300_100	0.82	0.18	7.74	28.19	0.77	0.23	1.75	4.75
	0.82	0.18	25.52	77.54	0.77	0.23	2.54	6.49
500_500	0.60	0.40	8.05	32.99	0.53	0.47	2.04	6.55
	0.60	0.40	8.71	31.20	0.52	0.48	2.13	6.22
750_250	0.79	0.21	9.05	35.50	0.78	0.22	2.88	9.28
	0.83	0.17	12.74	36.68	0.81	0.19	3.24	8.76
	0.76	0.24	18.86	66.33	0.72	0.28	2.29	6.61
	0.82	0.18	17.74	79.17	0.78	0.22	1.97	6.49
1000_1000	0.57	0.43	5.11	18.14	0.51	0.49	2.15	6.51
1500_500	0.79	0.21	12.87	50.13	0.76	0.24	2.19	6.78
	0.78	0.22	20.33	79.88	0.73	0.27	1.66	4.87
	Noncentral and Central t							
1000_1000	0.46	0.54	0.53	1.43	0.00	1.00	0.66	1.09
1500_500	0.72	0.28	0.46	1.18	0.00	1.00	0.78	1.00
	0.69	0.31	0.73	1.94	0.00	1.00	0.80	1.06
	0.87	0.13	9.47	66.03	0.84	0.16	0.77	3.20
200_200	0.51	0.49	0.43	1.18	0.00	1.00	0.66	1.04
	0.50	0.50	0.52	1.51	0.00	1.00	0.66	1.08
	0.56	0.44	0.58	1.53	0.00	1.00	0.73	1.10
	0.52	0.48	0.57	1.59	0.00	1.00	0.69	1.10
	0.46	0.54	0.66	1.83	0.00	1.00	0.68	1.17
	0.51	0.49	0.67	1.91	0.00	1.00	0.73	1.19
300_100	0.88	0.12	2.94	20.92	0.94	0.06	1.48	31.71
	0.62	0.38	0.65	1.31	0.00	1.00	0.80	1.05
	0.78	0.22	0.47	1.41	0.00	1.00	0.80	1.01
	0.63	0.37	0.60	1.45	0.00	1.00	0.81	1.11
500_500	0.70	0.30	22.45	100.00	0.66	0.34	0.69	4.05
	0.44	0.56	0.52	1.34	0.00	1.00	0.68	1.11
	0.49	0.51	0.71	2.10	0.00	1.00	0.71	1.18
750_250	0.87	0.13	6.89	73.96	0.85	0.15	0.77	6.00
	0.70	0.30	0.60	1.27	0.00	1.00	0.83	1.03
	0.73	0.27	0.53	1.29	0.00	1.00	0.79	1.00
	0.60	0.40	0.63	1.32	0.00	1.00	0.81	1.07
	0.74	0.26	0.50	1.47	0.00	1.00	0.79	1.02

REFERENCES

Albert, A. A. 1944. The matrices of factor analysis. *Proceedings of the National Academy of Sciences* 30: 90-95.

Allen, S. J. and R. Hubbard. 1986. Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research* 2: 393-398.

Anderson, T. W. 1963. Asymptotic theory for principal components analysis. *Annals of Mathematical Statistics* 34: 122-148.

_____. 1984. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Atkinson, S. F., M. F. Acevedo, K. L. Dickson, and D. A. Rolbecki. 1998. Distinguishing between chlorophyll-a and suspended solids in lake water using hyperspectral data. *Proceedings of SPIE Europto Series* 3499: 97-104.

Bartlett, M. S. 1950. Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section* 3: 77-85.

Bhattacharya, C.G. 1967. A simple method for resolution of a distribution into its Gaussian components. *Biometrics* 23: 115-135.

Bigus, J. P. 1996. *Data Mining with Neural Networks*. New York: McGraw-Hill.

Bobko, P. 1986. A solution to some dilemmas when testing hypotheses about ordinal interactions. *Journal of Applied Psychology* 71, no. 2: 323-326.

Bollen, K. 1989. *Structural Equations With Latent Variables*. New York: Wiley.

Brody, H. 1990. The neural computer. *Technology Review* August/September: 43-49.

Buja, A. and N. Eyuboglu. 1992. Remarks on parallel analysis. *Multivariate Behavioral Research* 27, no. 4: 509-540.

Butler, R. W. 1986. Predictive likelihood inference with applications (with discussion). *J.R. Statistical Society* B, 48: 1-38.

Cassie, R. M. 1954. Some uses of probability paper for the graphical analysis of polymodal frequency distributions. *Australian Journal of Marine and Freshwater Res* 5: 513-522.

Castañeda, M. B., J. R. Levin, and R. B. Dunham. 1993. Using planned comparisons in management research: A case for the Bonferroni procedure. *Journal of Management* 19, no. 3 : 707-724.

Cattell, R. B. 1958. Extracting the correct number of factors in factor analysis. *Educational and Psychological Measurement* 18: 791-838.

_____. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* 1: 245-76.

Cheng, C. 1995. A multi-layer neural network model for detecting changes in the process mean. *Computers in Industrial Engineering* 28, no. 1: 51-61.

Cota, A. A., R. S. Longman, R. R. Holden, and G. C. Fekken. 1991. Anomalies in the Allen and Hubbard parallel analysis procedure. *Applied Psychological Measurement* 15: 95-97.

_____. 1993. Comparing different methods for implementing parallel analysis: A practical index of accuracy. *Educational and Psychological Measurement* 53: 865-876.

Cota, A. A., R. S. Longman, R. R. Holden, G. C. Fekken, and S. Xinaris. 1993. Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement* 53: 585-596.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statistical Society* B, 39: 1-38.

Duncan, D. B. 1965. A bayesian approach to multiple comparisons. *Technometrics* 7: 171-222.

Dunnet, C. W. 1980. Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. *Journal of the American Statistical Association* 75: 789-795.

Fava, J. L. and W. F. Velicer. 1992. The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research* 27: 387-415.

_____. 1996. The effects of underextraction in factor and component analysis. *Educational and Psychological Measurement* 56: 907-929.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39, no. 11: 27-34.

Fowlkes, E. B. 1979. Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association* 74: 561-575.

Glorfeld, L. W. 1995. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement* 55, no. 3: 377-93.

Gorsuch, R. L. 1983. *Factor Analysis*. Hillsdale, N.J. : L. Erlbaum Associates.

Green, S. B. and M. A. Babyak. 1997. Control of type I errors with multiple tests of constraints in structural equation modeling. *Multivariate Behavioral Research* 32: 39-51.

Guttman, L. 1954. Some necessary conditions for common factor analysis. *Psychometrika* 19: 149-61.

_____. 1958. To what extent can communalities reduce rank? *Psychometrika* 23: 297-308.

Hakstian, A. R., and V. Muller. 1973. Some notes on the number of factors problem. *Multivariate Behavioral Research* October: 461-475.

Hakstian, A. R., W. T. Rogers, and R. B. Cattell. 1982. The behavior of number of factors rules with simulated data. *Multivariate Behavioral Research* 17: 193-210.

Harding, J. P. 1948. The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of Marine Biol. Association* U.K., 28: 141-153.

Harter, H. and N. Balakrishnan. 1996. *CRC Handbook of Tables for the Use of Order Statistics in Estimation*. New York: CRC Press.

Hasselblad, V. 1966. Estimation of parameters for a mixture of normal distributions. *Technometrics* 8: 431-444.

_____. 1969. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association* 64: 1459-1471.

Hayter, A. J. 1986. The maximum familywise error rate of fisher's least significant difference test. *Journal of the American Statistical Association* 81: 1000-1004.

Hodge, L. D. 1996. Briny brouhaha. *Texas Parks and Wildlife* May: 19-27.

Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30: 179-85.

Hoyt, J. D. and H. Wechsler. 1990. An examination of the application of multi-layer neural networks to audio signal processing. *Proceedings of the International Joint Conference on Neural Networks* 305-310.

Hubbard, R. and S. J. Allen. 1987. An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research* 15, no. 2: 173-190.

Humphreys, L. G., and R. G. Montanelli. 1975. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research* 10: 193-205.

Hung, H. M. J., R. T. O'Neill, P. Bauer and K. Köhne. 1997. The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics* 53: 11-22.

Igbaria, M., and Baroudi, J. J. 1993. A short-form measure of career orientations: A psychometric evaluation. *Journal of Management Information Systems* 10: 131-154.

Johnson, R. A., and D. W. Wichern. 1992. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

Jolliffe, I. T. 1972. Discarding variables in principal components analysis. I: Artificial data. *Applied Statistics* 21: 160-173.

_____. 1973. Discarding variables in a principal components analysis. II: Real data. *Applied Statistics* 22: 21-31.

_____. 1986. *Principal Components Analysis*. New York: Springer-Verlag.

Jurgens, F. 1999. Picture Perfect. *Imaging Notes* 14: 18-21.

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20: 141-151.

_____. 1963. Image analysis. In *Problems in measuring change*, ed. C. W. Harris. Madison: University of Wisconsin Press.

_____. 1965. Psychometric approaches to factor analysis. *Proceedings of the 1964 Invitational Conference on Testing Problems*. Princeton: Educational Testing Service.

Kirk, R. 1995. *Experimental design: Procedures for the behavioral sciences*. New York: Brooks/Cole Publishers.

Krzanowski, W. J. and P. Kline. 1995. Cross-validation for choosing the number of important components in principal components analysis. *Multivariate Behavioral Research* 30: 149-165.

Kromrey, J. D. and W. B. Dickinson. 1995. The use of an overall F test to control type I error rates in factorial analyses of variance: Limitations and better strategies. *Journal of Applied Behavioral Science* 31, no. 1: 51-64.

Lautenschlager, G. J. 1989. A comparison of alternatives to conducting monte carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research* 24, no. 3: 365-395.

Lautenschlager, G. J., C. E. Lance, and V. L. Flaherty. 1989. Parallel analysis criteria: Revised regression equations for estimating the latent roots of random data correlation matrices. *Educational and Psychological Measurement* 49: 339-345.

Ledermann, W. 1937. On the rank of the reduced correlation matrix in multiple-factor analysis. *Psychometrika* 2: 85-93.

Legendre, P. and L. Legendre. 1983. *Numerical Ecology*. Amsterdam: Elsevier.

Lindsay, R. M. 1997. Lies, damned lies and more statistics: The neglected issue of multiplicity in accounting research. *Accounting and Business Research* 27: 243-258.

Longman, R., A. A. Cota, R. R. Holden, and G. C. Fekken. 1989. A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research* 24, no. 1: 59-69.

McCabe, G. P. 1984. Principal Variables. *Technometrics* 26: 137-144.

McLachlan, G. J. and K. E. Basford. 1988. *Mixture Models: Inference and Applications to Clustering.* New York, NY: Marcel Decker, Inc.

Mitchell, D. L., J. Yi, and S. Govind. 1996. Household trip forecasting using neural networks: An exploratory investigation. In *1996 Annual Meeting Proceedings: Decision Sciences Institute*: 1117-1119.

Montanelli, R. G., and L. G. Humphreys. 1976. Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A monte carlo study. *Psychometrika* 41: 341-348.

Müller, M. J. and H. Wetzel. 1998. Dimensionality of depression in acute schizophrenia: a methodological study using the Bech-Rafaelsen Melancholia Scale (BRMES). *Journal of Psychiatric Research* 32: 369-378.

Nakamura, A. O., M. Nakamura, and H. O. Duleep. 1990. Alternative approaches to model choice. *Journal of Economic Behavior and Organization* 14: 97-125.

Newcomb, S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal and Mathematics* 8: 343-366.

Nguyen, D. H. and B. Widrow. 1990. Neural networks for self-learning control systems. *IEEE Control Systems Magazine* April: 18-23.

Norton, M. 1994. Close to the nerve (credit card fraud). *Banking Technology* UK, 11, no. 10: 28-31.

Parker, R. A., and R. B. Rothenberg. 1988. Identifying important results from multiple statistical tests. *Statistics in Medicine* 7: 1031-1043.

Pearson, K. 1894. Contribution to the mathematical theory of evolution. *Phil. Trans.* A, 185: 71-110.

Rao, B. L. S. P. 1992. *Identifiability in Stochastic Models*. Academic Press.

Rao, C. R., 1948. The utilization of multiple measurements in problems of biological classification. *J. R. Statistical Society* B, 10: 159-203.

_____. 1964. The use and interpretation of principal components analysis in applied research. *Sankhya* A: 329-358.

Raven, P. H. and G. B. Johnson. 1996. *Biology*. 4th ed. Chicago: Wm. C. Brown Publishers.

Robert, P. and Y. Escoufier. 1976. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied Statistics* 25: 257-265.

Romanazzi, M. 1991. Jackknife estimation of the eigenvalues of the covariance matrix. *Computational Statistics and Data Analysis* 15: 179-198.

Ruggiero, M. A. 1994. Interpreting feedback to build a better system. *Futures: The Magazine of Commodities and Options* 23, no. 8: 46-48.

Schalkoff, R. J. 1997. *Artificial Neural Networks*. New York: McGraw-Hill.

Scheffé, H. 1959. *The Analysis of Variance*. New York: Wiley.

Schein, E. H. 1985. *Career Anchors: Discovering Your Real Values*. San Diego: University Associates.

Schweder, T. and E. Spjotvoll. 1982. Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* 69: 493-502.

Sharma, S. 1996. *Applied Multivariate Techniques*. New York: John Wiley & Sons, Inc.

Silverstein, J. W. and Z. D. Bai. 1995. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis* 54: 175-192.

Skinner, H.A. 1979. Dimensions and clusters: A hybrid approach to classification. *Applied Psychological Measurement* 3: 327-341.

Tanaka, Y. and U. Mori. 1997. Principal components analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematical and Management Sciences* 17: 61-89.

Titterington, D. M., A. F. M. Smith, and V. E. Makov. 1985. *Statistical Analysis of Finite Mixture Distributions*. Wiley.

Tukey, J. W. 1953. The problem of multiple comparisons. Unpublished manuscript, Princeton University, Princeton, N. J.

Velicer, W. F. 1976. Determining the number of components from the matrix of partial correlations. *Psychometrika* 41: 321-327.

Warner, B. and M. Misra. 1996. Understanding neural networks as statistical tools. *American Statistical Association* 50, no. 4: 284-293.

Werbos, P. 1990. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE* 78, no. 10: 1550-1560.

Wilk, M. B. and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55: 1-17.

Wood, J. M., D. J. Tataryn, and R. L. Gorsuch. 1996. Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Measurement* 1, no. 4: 354-365.

Yi, J., D. Mitchell, and V. R. Prybutok. 1996. Comparison of neural network model with conventional statistical models for prediction of Houston's maximum ozone levels. In *Decision Sciences Institution Southwest Region: Theory and Applications Proceedings, 27th Annual Conference* 57-59.

Zoski, K. W., and S. Jurs. 1996. An objective counterpart to the visual scree test for factor analysis: The standard error scree. *Educational and Psychological Measurement* 56, no. 3: 443-451.

Zwick, W. R., and W. F. Velicer. 1986. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin* 99: 432-442.