A SPATIAL EDITING AND VALIDATION PROCESS FOR SHORT COUNT TRAFFIC DATA

NCDOT Project Authorization Number HWY-2002-09
FHWA/NC/2006-59

FINAL REPORT

By:

Jacqueline M. Hughes-Oliver, Ph.D., Principal Investigator
Tae-Young Heo, Graduate Research Assistant

Of the

Department of Statistics
North Carolina State University
Raleigh

And

Shannon McDonald
Moorman, Kizer & Reitzel, Inc.

Created for the:

North Carolina Department of Transportation (NCDOT)
Raleigh

Submitted July 2006

Project Monitor: Kent L. Taylor, P.E., NCDOT, Traffic Survey Unit (TSU)

Technical Report Documentation Page

1. Report No. **FHWA/NC/2006-59**	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle **A Spatial Editing and Validation Process for Short Count Traffic Data**		5. Report Date **July 7, 2006** 6. Performing Organization Code
7. Author(s) **Jacqueline M. Hughes-Oliver, Tae-Young Heo, Shannon McDonald**		8. Performing Organization Report No.
9. Performing Organization Name and Address **Department of Statistics** **North Carolina State University** **Raleigh, NC, 27695** and **Moorman, Kizer & Reitzel, Inc.**		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address **North Carolina Department of Transportation** **Research and Analysis Group** **1 South Wilmington Street** **Raleigh, North Carolina 27601**		13. Type of Report and Period Covered **Final Report** **July 2001—December 2004** 14. Sponsoring Agency Code **2002-09**
Supplementary Notes:		

16. Abstract

The Traffic Survey Unit (TSU) manages 40,000 traffic monitoring stations, of which 25,000 are updated annually. These counts obtained by TSU play a crucial role in allocation of resources for the maintenance, upgrade, and expansion of traffic infrastructure. The need for reliable, edited, and validated traffic count data is well acknowledged by the Federal Highway Administration (FHWA) and the American Association of State Highway and Transportation Officials (AASHTO).

The research reported here addressed this need by developing a statistically defensible approach to achieving spatial continuity of traffic counts as part of the editing and validation process. The deliverables include GIS-formatted data that programmatically identify PTC stations that have anomalous counts. We also provide information for creating traffic continuity maps. Identification of problem areas is quick and reduces the burden on NCDOT staff.

As such, the project will significantly improve the process of validating traffic counts by increasing the accuracy of reported counts, by reducing the time delay between data collection and reporting, and by making it easy to provide customized reports of traffic counts to NCDOT departments and customers.

17. Key Words Annual average daily traffic (Bthvaa), Correlation analysis (Usd), Covariance (Usac), Data editing, Mathematical prediction (Ush), Portable traffic counters, Regression analysis (Uss), Spatial analysis, Statistical analysis (Us), Traffic continuity maps, Traffic counts (Btet), Traffic distribution (Bthj), Traffic volume (Bthv)	18. Distribution Statement

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 57	22. Price

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Disclaimer

The contents of this report reflect the views of the authors and not necessarily the views of either North Carolina State University or Moorman, Kizer & Reitzel, Inc. The authors are responsible for the statements made in this report. The contents do not necessarily reflect the official views or policies of the North Carolina Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

Acknowledgements

The authors are grateful to the following individuals who provided significant assistance and advice in support of this project:

Kent L. Taylor, PE, Chair
Susan Cosper
Larry Wikoff
Dr. Erik Brun
Bill Miller
Rodger Rochelle, PE
Dr. Moy Biswas, PE
Bill Schaller
Derry Schmidt
Bob Foyle
Richard Lakata, PE
Terry Byron

We are especially grateful to the Traffic Survey Unit at NCDOT for working hard to provide quality data for building models, for being generous with their offers of providing feedback and assistance, and for guiding development of the tool. In particular, we are indebted to Kent Taylor for constant support, Susan Cosper who first alerted the PI to this project need, and to Larry Wikoff for arranging data transmittal.

Finally, we acknowledge the support provided by our home institutions. North Carolina State University provided excellent computing resources and a stimulating research environment.

Executive Summary

The Traffic Survey Unit (TSU) manages 40,000 traffic monitoring stations, of which 25,000 are updated annually. These counts obtained by TSU play a crucial role in allocation of resources for the maintenance, upgrade, and expansion of traffic infrastructure. The need for reliable, edited, and validated traffic count data is well acknowledged by the Federal Highway Administration (FHWA) and the American Association of State Highway and Transportation Officials (AASHTO).

The research reported here addressed this need by developing a statistically defensible approach to achieving spatial continuity of traffic counts as part of the editing and validation process. The deliverables include GIS-formatted data that programmatically identify PTC stations that have anomalous counts. We also provide information for creating traffic continuity maps. Identification of problem areas is quick and reduces the burden on NCDOT staff.

As such, the project will significantly improve the process of validating traffic counts by increasing the accuracy of reported counts, by reducing the time delay between data collection and reporting, and by making it easy to provide customized reports of traffic counts to NCDOT departments and customers.

TABLE OF CONTENTS

Technical Report Documentation Page 2
Disclaimer 3
Acknowledgements 3
Executive Summary 4
TABLE OF CONTENTS 5
1. INTRODUCTION 7
1.1 Purpose 7
1.2 Background 7
1.3 Problem Definition 8
1.4 Literature Review 9
2. RESEARCH METHODOLOGY AND ITEMIZED TASKS 11
2.1 Research Objectives 11
2.2 Task 1: Further Review of Methods, Data, and Computing Options (Objectives 1, 2, 3) 11
2.3 Task 2: Spatial Modeling and Characterization, Using Current Seasonality Adjustments (Objectives 4, 5, 7, 8) 12
2.4 Task 3: Automation and GIS Implementation (Objectives 6, 7, 8) 13
2.5 Task 4: Implementation Guidelines (Objective 9) 13
2.6 Task 5: Spatial Modeling and Characterization to Explore Alternative Seasonality Adjustments (Objective 10) 14
3. ADDITIONAL DETAILS ON COMPLETION OF TASKS 15
3.1 Obtain and Explore Data, Task 1d 15
3.1.1 PTC Count Data 15
3.1.2 Auxiliary GIS Station Data and Land Use Data 16
3.1.3 Census Data 18
3.1.4 Distances Between PTC Stations 18
3.1.5 Scaling to the Entire State 20
3.2 Spatial Modeling and Characterization, Using Current Seasonality Adjustments, Task 2 20
3.2.1 Mean Model for Test Area 21
3.2.2 Mean Model for Entire State 23
3.2.3 Covariance Model for Test Area 25
3.2.4 Covariance Model for Entire State 31
3.2.5 Predictions and Prediction Intervals 36
3.3 Automation and GIS Implementation, Task 3 45
3.4 Implementation Guidelines, Task 4 48
4. FINDINGS AND CONCLUSIONS 50
5. RECOMMENDATIONS 53
6. IMPLEMENTATION AND TECHNOLOGY TRANSFER PLAN 54
7. CITED REFERENCES 55
8. BIBLIOGRAPHY 56
APPENDICES 58
1. Documentation of Data Edits, Queries, and Issues in Reconciliation 59
2. Procedure and Corrections of PTC Data 70

3. Data Description for Spatial Study 83
4. FY 2002 GIS Project Role and Assessment of Accomplishment and Budget Standing 93
5. Data Edits Determined From "List of 55" 99
6. New Urban and Municipal Field Value Assignments and Their Relation to Station Locations and the PTC Data Model Process 109
7. Recordings and Methodology To Calculating Capacity Variables For Unmatched Stations 114
8. Determining the Final Mean Model for the 5-County Test Area 125
9. Determining the Final Mean Model for the Statewide Area, Version 2 245
10. Preparations for November 19, 2004 Meeting with DOT Research Team 322
11. Primary Road Statewide Traffic Continuity Map 329
12. Automation With GIS Application Procedures 340

1. INTRODUCTION

1.1 Purpose

The volume of traffic on North Carolina's 79,043 miles of state roads is of premier importance when it comes to allocating resources for the maintenance, upgrade, and expansion of traffic infrastructure. An excellent illustration of this comes from considering pavement requirements for Interstate routes. Increased traffic on interstates leads to tougher requirements on thickness levels of pavement. More specifically, a **doubling of traffic volume requires an additional 1-2 inches of pavement, at a cost of about $30,000 per lane mile of pavement**. Pavement planning cannot proceed without knowledge of traffic volume. If traffic volume on an interstate is mistakenly recorded too high, then unnecessary adjustments may occur, causing extreme fiscal waste. On the other hand, if traffic volume is mistakenly recorded too low, then needed adjustments will not be initiated and the pavement will deteriorate more quickly than expected; repair costs exceed rehabilitation costs, so again there is fiscal waste.

The need for reliable, edited, and validated data on traffic volume is clear. As such, the Federal Highway Administration (FHWA) and the American Association of State Highway and Transportation Officials (AASHTO) have both issued guidelines related to the editing, validation, and analysis of data collected on traffic counts. This research project responds to the need by incorporating statistical techniques for the purpose of providing improved estimates and measures of uncertainty for annual average daily traffic counts.

1.2 Background

Traffic counts play a vital role in many aspects of daily life; they help determine the way people travel. Within North Carolina, **approximately 25,000 traffic counts were recorded annually on highways for the purpose of monitoring flow of traffic during 1998**. Counts are used to assess and define the current usage of roads in the state and can thus aid in the identification and development of transportation needs and plans. The Transportation Planning Branch of the North Carolina Department of Transportation (NCDOT) uses these counts in fulfilling its responsibilities of **identifying long-range transportation needs for the state and for providing assistance to Metropolitan Planning Organizations** (MPOs) as they respond to needs within their own areas of the state. The latter is accomplished through a comprehensive, coordinated, and cooperative transportation planning process between these agencies that is directly supported by the volume count data provided by the NCDOT. Additionally, **volume counts are published annually and are also available by calling a dedicated traffic count hotline**.

Traffic counting programs are housed within the Traffic Survey Unit (TSU) of NCDOT's Statewide Planning Branch. Two programs monitor traffic volume: the Coverage Count Program—PTC (also known as short-term volume counts) and the Continuous Volume Count Program—ATR.

> *Short-term volume counts* are produced by inexpensive portable traffic counters (PTCs) that are rotated through the more than 40,000 PTC stations across North Carolina. A PTC is installed at a station for a minimum of 48 hours and a maximum of 72 hours between Monday morning and Friday noon. The counter records the number of axle pairs (two axle hits = 1 axle pair) that cross a pneumatic tube sensor in daily totals for two to three days. This type of counter captures axle pair counts, not the volume count data needed for later analysis. Consequently, axle adjustment factors are used to convert daily axle pair counts to daily volume counts to account for vehicles in the traffic stream with 3 or more axles. Axle adjustment factors are the ratio of total traffic volume to the corresponding total number of axles divided by two. These are generated from vehicle

classification counts where the number of axles for each vehicle is known. These counts are more costly to collect and a factor generated at a single vehicle classification station is used at many PTC stations. The second adjustment performed transforms the daily volume counts to annual average daily traffic (AADT) volumes. This adjusts for seasonal variation typical for a location and provides a consistent measure of traffic regardless of the day of week or month the data was collected. This adjustment is performed using seasonal adjustment factors developed from continuous count monitoring stations operated by the NCDOT.

Continuous volume counts are produced by automatic traffic recorders (ATRs) that have sensors embedded in the pavement. ATRs provide continuous hourly volume counts for each lane of travel at a station. As such, ATRs serve a critical role for explaining the variability observed in traffic counts due to time of day, day of week, and month of year. Data is screened to identify typical travel patterns at each station and stations with common patterns are clustered to generate ATR Groups. Seasonal adjustment factors are generated for these groupings of ATR stations to provide a basis for factoring counts collected at PTC stations. Unfortunately, the cost of installing and maintaining ATRs limit this program to a sampling of stations (100 total) across North Carolina.

Traffic counts produced by the more than 40,000 PTC and the 100 ATR stations must be edited and validated to achieve consistency over time and across space. **Prior to this project, the process required manual and visual comparison of current counts to counts from previous years and neighboring stations. If a count was considered unusual, it was often modified to make it more similar to neighboring counts. This process was very slow, was prone to individual subjectivity and bias, and encouraged excessive manual adjustments to counts.** Improvements to this process can help those performing the editing and validation as well as those using the resulting data.

1.3 Problem Definition

Prior practice within NCDOT for editing and validating count data was to manually and visually calibrate traffic count data from each of the more than 40,000 counting stations with values from neighboring stations. Counts that were not consistent with their neighbors were often manually adjusted to achieve consistency. **Deficiencies of this process include**, but are not limited to, the following.

- There was subjectivity in deciding when a count needs to be manually adjusted.
- If a manual adjustment was needed, there was subjectivity in determining the amount of manual adjustment needed.
- The process encouraged excessive manual adjustments to counts that were not totally in line with their neighbors but were still within the level of variability of the data.
- It took a year to complete the process.
- Because the process was so slow, the window of opportunity for performing recounts of questionable data was often missed, so *ad hoc* manual adjustments were made.
- The process did not realize the recommendations from FHWA and AASHTO to incorporate spatial analysis.

The problem, and reason for this research project, was that NCDOT's process of editing and validating count data needed improvement to address the deficiencies listed above.

This report documents our system for objectively improving the editing and validation process for PTC counts in light of spatial patterns. The next subsection reviews related literature. Section 2 of this report is a general overview of research methodology for all tasks, while implementation details for all tasks are

provided in Section 3 with even more details (including computing code) provided in Appendices 1 to 12. Findings and conclusions are given in Section 4, and recommendations are made in Section 5. Technology transfer is discussed in Section 6 and references (cited and bibliography) are given in Sections 7 and 8.

1.4 Literature Review

The literature on traffic data editing and validation is very sparse, even though the need for such practices is universally recognized. Albright (1991) argues that different agencies have developed independent data collection and analysis procedures, where the emphasis has been on application of professional judgment. While this emphasis has been helpful when the professional is truly knowledgeable of specific roadway operational characteristics, it has also resulted in data inconsistency and lack of comparability. Albright (1991) subsequently proposed four principles for the development of national standards: base data integrity, measurement edits, consistent computation, and truth-in-data.

Truth-in-data is also a basic tenet in the AASHTO Guidelines for Traffic Data Programs, which will later be referred to as The Guidelines. Truth-in-data is described as "the disclosure of practice and estimate of data variability" that is "central to the Guidelines to ensure appropriate data quality and use" (AASHTO 1992, p. 8). The Traffic Monitoring Guide (FHWA 2001, p. 3-27) goes on to say

> Subjective editing procedures for identifying and imputing missing or invalid data are discouraged, since the effects of such data adjustments are unknown and frequently bias the resulting estimates. … Truth-in-data implies that agencies maintain a record of how data are manipulated, and that each manipulation has a strong basis in statistically rigorous analysis. Data should not be discarded or replaced simply because "they didn't look right." Instead, each State should establish systematic procedures that provide the checks and balances needed to identify invalid data, control how those invalid data are handled in the analysis process, and identify when those quality control steps have been performed.

The seminal source for *specific recommendations on avoiding the subjective input of professionals in the editing process* is The Guidelines. Their recommendations are given in two chapters: Chapter 4—Editing Traffic Data, and Chapter 8—Quality Control. Data edits result from the three broad validity concerns of machine malfunction, non-representative observations (data collected on holidays are atypical with respect to "usual" traffic patterns), and consistency with respect to other data collected over time and across space. The need for spatial consistency is the main topic of interest in this proposed research.

The Guidelines provide examples (using real data) of various data editing scenarios and make several recommendations:

- Traffic counts should be edited in context. The context is defined for some data by extensive histories, and for other data by other counts and land use information on the same roadway. (p. 38)
- Editing criteria should be as objective and nonjudgmental as possible. They may be based on expert judgment and experience, but should be capable of being reduced to algorithmic form so they can be programmed for computer editing of datasets. (p. 89)
- A specific standard should be established for the time between collection of field data and its submittal, tabulation and editing. This will ensure assessment in time for meaningful corrective action to take place, if necessary. (p. 89)
- If a traffic count is inconsistent with another traffic count during the same period on the same roadway, and no explanation of the count variance can be provided based on land use, the related

count should not be accepted for summarization until another count can be taken and compared. (p. 36)

- When there are inconsistencies, counts should also be initiated on the adjoining segments to confirm the validity of all counts. (p. 36)
- As states develop traffic data programs, computational methods for traffic statistics should be automated in a way that does not permit imputation. (p. 37)
- For states that currently employ imputation techniques:
  - The difference must be maintained between traffic measurements and imputed values. (p. 37)
  - Agencies should clearly document the method they used to impute missing values. (p. 37)
  - The extent of imputation in reported traffic summary statistics should be documented. (p. 37)

Several other articles offer insights on traffic data. Claramunt et al. (2000) propose a method for the real-time integration, manipulation, and visualization of urban traffic data. Their methods are based on observing the movements of several vehicles in space, or by observing changes in urban network properties. They do not address issues related to data editing or statistical modeling. Cunagin and Kent (1998) investigate the importance of traffic data variability on the reliability of traffic projections for pavement design. These activities are based on expressions for traffic data variability and the AASHTO reliability concept for pavement design.

The literature on spatial statistical modeling of traffic count data for the purpose of validating the counts is nonexistent. However, the level of activity in spatial statistical modeling is very high and applications are being forged in many areas. Xia et al (1999) come very close in that they developed a model for estimating AADT for non-state roads by using some of the same attributes that we used for this project. They do not, however, use their model to validate count data and they do not investigate the spatial correlation structure.

Finally, **the literature is replete with articles that investigate the development of seasonal factors**, either based on functional classifications or on cluster analysis methods. The findings very strongly support the need for seasonal adjustments of short count volume data. Sharma et al. (1996) find that estimation errors are very sensitive to the assignment of stations within clusters. Mohamad et al. (1998) use multiple regression to incorporate relevant demographic variables such as population, state highway mileage, per capita income, and the presence of interstate highways. Hu et al. (1998) use a simulation study to assess the bias, relative to continuously monitoring a site, of seasonally adjusting a short count.

2. RESEARCH METHODOLOGY AND ITEMIZED TASKS

2.1 Research Objectives

The **goal** of this research was to create a statistically defensible process for editing and validating traffic count data to achieve spatial continuity, and to automate this process in a GIS environment.

The following **research objectives** were established to ensure achievement of the aforementioned goal:
1. to learn from the experiences of other states where similar projects have already been established,
2. to gain familiarity with computing and database platforms at NCDOT and other local state agencies so that implementation and technology transfer will be as seem-less as possible,
3. to obtain and thoroughly explore NCDOT's count data, axle adjustment factors, seasonal adjustment factors, station locations, station classifications, and any other information that will aid in explaining the variability in the count data,
4. to develop a statistical model for characterizing the spatial continuity of traffic counts,
5. to obtain predictions and prediction intervals for traffic counts,
6. to display predicted traffic counts on primary routes for creating "traffic continuity" maps,
7. to identify and display anomalies in individual station counts so that various levels of "extremeness" can be recognized,
8. to identify and display anomalies in regional traffic counts for the purpose of highlighting regional changes in traffic flow, and
9. to support a continuous review process throughout the year, and
10. to investigate the feasibility of a statistical model that simultaneously characterizes spatial continuity of counts and performs seasonal adjustments.

Objectives 4 and 5 were primarily accomplished using advanced statistical modeling, while objectives 6, 7, and 8 were primarily dependent on GIS methodologies. However, success of this project depended on inter-weaving several methodologies from the two broad areas of statistics and GIS. Specifically, development of the statistical methods required full integration of GIS tools, and, even though the final products will be presented in a GIS environment, they are based on statistical methodologies. In other words, the statistics and GIS have been fundamentally linked.

2.2 Task 1: Further Review of Methods, Data, and Computing Options (Objectives 1, 2, 3)

Primary Components:
a. Review published literature.
b. Learn about practices and plans from other states.
c. Learn about computing options at NCDOT and other local state agencies.
d. Obtain and explore data.

Task Description:
a. & b. An exhaustive search of the literature and transportation research repositories for related articles guaranteed that this project did not amount to reinventing the wheel. This search was not limited to transportation related sources but also examined articles on spatial modeling and prediction of counts when covariates are available. This search investigated options for dealing with spatial data that does not clearly fall in any of the three typical categories of geostatistical, lattice, or point process.

More specifically, as of September 30, 2001, Tae-Young Heo and Jacqueline Hughes-Oliver completed reviews of 21 papers from sources such as Journal of Transportation Engineering, Transportation

Research Record, Journal of Transportation Planning and Technology, state DOT repositories, and other sources. Our literature search included practices from other states (notably Minnesota, Delaware, Florida, Iowa, Colorado, and Indiana) and from several other countries (including the UK, Norway, Hong Kong, and Canada). Most of these papers, selected because of their stated goal of improving estimation of AADT from short count data, focused on optimal ways of determining seasonal groupings or on alternative approaches, such as regression or neural networks, to the widely used clustering approach of deriving seasonal factors. Some of these papers also considered optimal selection of locations for ATR stations. The regression papers were the only ones that provided information of immediate value to this project. Based on the literature it appeared that the following variables may have very good power for explaining observed variation in traffic counts, at least for ATR data: number of lanes, city population, whether or not access to a road section is controlled, and route code. As a result, we included these variables in the exploratory and model development stages.

Between September and December of 2001, Tae-Young Heo and Jackie Hughes-Oliver continued reviewing the 1991 dissertation by Chih-Hsu Cheng entitled "Optimal Sampling for Traffic Volume Estimation" that was applied to Minnesota's DOT data. We obtained the data used in Cheng's study and recreated many of the tables and figures. A Bayesian analysis was also completed on this dataset for the purpose of learning the Bayesian BUGS software and for assessing the potential impact of a Bayesian analysis on this data. We found that the Bayesian and classical approaches gave comparable answers when vague priors were used.

The most relevant papers that were reviewed are listed in the bibliography.

c. Based on input from TSU personnel, an informal assessment was completed of computing power available to routine users of TSU's traffic reports. These users include NCDOT departments and city and local government agencies. The assessment provided information crucial for determining how we should standardize our primary product to be convenient to the greatest number of users.

The primary product was discussed at the Kick-Off meeting on July 25, 2001 with the original intention that it would be software to allow interactive visualization of both the count stations and their status of being labeled "extreme" or not as well as traffic continuity maps. At this time, however, NCDOT is not able to provide distributable mapping software for interactive visualization of GIS-formatted data, so we were required to modify our final product. The final product is now GIS-formatted data that could be used to allow interactive visualization of both the count stations and their status of being labeled "extreme" or not as well as traffic continuity maps, provided the user has appropriate visualization software. Because the GIS-formatted data will be downloadable from the internet and/or distributed via CD-ROM, the only computing equipment required is a PC with internet access or a CD drive. In the future, after NCDOT acquires the ability to provide distributable mapping software for interactive visualization, the final product of this project can and should be adjusted to the original intended format.

d. At the end of axle and seasonal adjustment factors to produce AADT data, the data was made available in two formats, plain text and in a GIS environment. This task required far more effort that originally anticipated but was very informative for the entire TSU process. **Details on this task are provided in Section 3.1.**

2.3 Task 2: Spatial Modeling and Characterization, Using Current Seasonality Adjustments (Objectives 4, 5, 7, 8)

Primary Components:
a. Develop and fit spatial model.
b. Perform model diagnostics.
c. Obtain predictions and prediction intervals.
d. Determination of station-specific anomalies and levels of extremeness.
e. Determination of regional anomalies and levels of extremeness.

Task Description:
a. & b. & c. With spatially correlated data, prediction at one site is based on a weighted average of observations at surrounding sites, where the method of averaging depends on the data type and on the correlation structure. Historical data was used to build and validate the model. This phase of the project provided various challenges.

First, spatial models are fairly common-place for continuous data, but not for count data of this sort. Appropriate distributions needed to be determined, and these decisions impacted how correlation between nearby stations was handled. Questions of whether to attempt to transform the data to achieve near-normality, how to estimate the model, and how to handle the very large dataset, all needed to be addressed. Additionally, the questions of whether to treat each station as an individual spatial location or as indexed by the larger road of which it is a part and how to determine distances between stations needed to be addressed.

This phase of the project relied heavily on the GIS structure and geo-referencing of the stations. Several iterations were required to obtain reasonable results. Programming was in the SAS Software, ArcView, and ArcGIS. **Detailed information is provided in Section 3.2.**

d. & e. Anomaly identification was accomplished using prediction intervals. **Detailed information is provided in Section 3.2.**

2.4 Task 3: Automation and GIS Implementation (Objectives 6, 7, 8)

Primary Components:
a. Program the methods developed in Task 2 for obtaining predictions and prediction intervals.
b. Program the methods developed in Task 2 for identifying station-specific anomalies.
c. Program the methods developed in Task 2 for identifying region-specific anomalies.
d. Program the methods for creating traffic continuity maps.

Task Description:
Details are provided in Section 3.3.

2.5 Task 4: Implementation Guidelines (Objective 9)

Primary Components:
a. Updating the spatial model in light of new data.
b. Details on making the software available to other departments within and outside NCDOT.

Task Description:
a. It is unnecessary and infeasible to update the spatial model every year. When enough new station data has been accumulated, say in five year intervals, it is recommended that this project be repeated to

update both the mean and covariance models, and hence the prediction intervals. This will ensure that prediction intervals do not become dated and they reflect recent changes to the transportation infrastructure. The detailed nature of this report will facilitate this updating process.

b. As explained for Task 1 above, we were required to modify our primary product from what was originally intended. The original intention was to create a customized GIS software package integrated with the traffic count data that would be distributed to other NCDOT departments, MPO's, and other users of the data. The software would have allowed visualization of the count stations in an interactive mapping environment with simplified GIS functions, and it would have displayed traffic continuity maps. Our current primary product falls just short of providing visualization software—we instead provide the GIS-formatted data that could be used as input to visualization software. The GIS-formatted data will be downloadable from the internet, and/or distributed via CD-ROM. This report contains elements of the type of GIS-formatted data provided to TSU and how this data may be used to obtain customizable reports. Details have been provided in Section 3.4.

2.6 Task 5: Spatial Modeling and Characterization to Explore Alternative Seasonality Adjustments (Objective 10)

Proposed Primary Components:

a. Develop and fit a spatial-temporal model.
b. Perform model diagnostics
c. Create "look up" table of new seasonality adjustments.
d. Compare new and old seasonality adjustments.

Proposed Task Description:

These activities will mirror the first two activities in Task 2, except with many additional complications. Later steps in this task are predicated by the success of earlier steps in the same task.

Comments:

This task was never completed, although some preliminary work was done towards it. Given the lengthy details in the project due to unanticipated difficulties in obtaining and editing the data, and the fact that seasonality adjustment factors were recently updated by another research project within the TSU, it was decided that this task would be omitted from the project. This decision was jointly made between NCDOT personnel and the project team.

No further references will be made to Task 5 for the remainder of this report.

3. ADDITIONAL DETAILS ON COMPLETION OF TASKS

3.1 Obtain and Explore Data, Task 1d

Several people worked diligently to get the data requested for this project. Two early meetings, one after the Kick-Off meeting on July 25, 2001 and the other on August 2, 2001, were devoted to discussing data needs, availability, and integrity. The four broad categories of data are:

1. *PTC count data:* actual counts, AADT or seasonally and axle adjusted counts, and non-attributed station descriptors such as name of road segment, route type, and county location.
2. *Auxiliary GIS station data and land use data:* some "duplicate" non-attributed station descriptor data from *PTC count data*, station locations with respect to the State Plane Coordinate System (which we will also refer to as quasi-latitude-longitude locations), data on land use in areas surrounding stations, roadway network and associated layers of the shapefile containing road segment attributes such as number of lanes, speed limit, or level of access control, and unique identifiers for linking station information to census data.
3. *Census data:* socio-economic and business indicators of regions surrounding PTC stations.
4. *Distances between PTC stations.*

PTC count data are collected and maintained almost entirely by TSU, but they needed extensive cleaning before being used in this project. The remaining data types were either collected by groups external to TSU or they were previously unavailable and needed to be calculated from rudimentary knowledge. Specifically, auxiliary GIS station data are collected and maintained by the GIS unit within NCDOT, while US census-derivative data was obtained from the GIS Unit in NCDOT and land use data was obtained from the GIS Unit in NCDOT as a UNC-derivative. Distances between stations had to be derived, and this turned out to be a very difficult task.

To ease the immediate burden of handling all the data, and to allow more intelligent decisions of exactly what kind of data and level of detail was needed for the entire state, a decision was reached to first consider a small five-county section of the state for detailed analysis. This test area consisted of Chatham, Orange, Durham, Wake, and Johnston counties. The next four subsections describe steps applied to obtain and edit the four data types, while the fifth subsection describes special handling required for obtaining data for the entire state, not just the five-county test area.

3.1.1 PTC Count Data

Larry Wikoff, with assistance from Susan Cosper, worked on the logic for extracting the PTC count data. The databases were rather large and complex, and many unexpected exceptions occurred. The road network (provided and maintained by NCDOT) had undergone a spatial and data integrity review, and several corrections and updates had to be made. In November 2001, Larry Wikoff transmitted the PTC count data for the test area as a space-delimited text file sent via electronic-mail.

Two extensive edits were required for this dataset. The first set of edits addressed issues described in the February 2002 report "Documentation of Data Edits, Queries, and Issues in Reconciliation" submitted by Jacqueline Hughes-Oliver and Tae-Young Heo, and included as Appendix 1. Discrepancies between non-attributed station descriptors in the PTC count data and auxiliary GIS station data, as well as missing entries or duplicate station entries within the PTC count data, were resolved. On May 6, 2002, two CDs were delivered in response to these data queries. The updated data, and the processes used to create it, was discussed in a meeting of the project team on May 16, 2002. Discussion during that meeting led to delivery of another CD and two reports, namely "Procedure and Corrections of PTC Data" by Shannon

McDonald (included as Appendix 2) and "Data Description for Spatial Study" by Larry Wikoff (included as Appendix 3). The data and reports were end-products of extensive cleaning by TSU and the GIS contractor, and they supersede all earlier versions. This was a major development in the project. Reaction to this development was provided by Shannon McDonald in the June 2002 report "FY2002 GIS Project Role and Assessment of Accomplishments and Budget Standing" included as Appendix 4.

The second set of edits was in response to an August 21, 2002 email from Jacqueline Hughes-Oliver and Tae-Young Heo. Using a preliminary model that predicted AADT counts as a function of station descriptors, GIS data, land use data, and census data, without acknowledging spatial correlations, 55 of the 3434 stations within the test area were identified as "outliers" because their AADT counts were not well predicted by the preliminary model. **The fact that only 1.6% of the stations were flagged indicated impressive early success of the project.** Nonetheless, it was important to verify the validity of the flags raised by these 55 stations. The research team proposed several alternatives for doing this, including:

- Augment the 2000 block-level census data with group-level attributes derived from 1990 block-group census data. This would provide benefits from a finer scale while retaining a large number of attributes per station, thus possibly leading to a more detailed and better predictive model. Unfortunately, the updated data was not yet available, so this path was not pursued.
- Investigate the 55 stations one-at-a-time to understand their unique perspectives. This was done, with the following findings:
    - Some stations suffered from time lags or even errors in their route classifications. For example, station 0501816 was listed as a local route in the LRS but not in CAD.
    - Some stations had incorrect IDs. For example, station attributes for stations 0910706 and 1080706 should be switched.
    - Some stations are unusual, but their unusual behavior could possibly be captured by an additional attribute that indicates existence within city limits. For example, urban routes within city limits might reasonably be expected to have higher volumes than urban routes outside city limits.
    - Some stations were incorrectly snapped. (This is a long-standing, already-identified issue that can be eradicated only from extensive, one-at-a-time checking of all stations, which is not feasible in the context of this project.)
    - Some stations are simply unusual and, short of customizing variables to capture their special nature, cannot be adequately modeled using a parsimonious structure. For example, some roads are low volume because alternate high-volume routes are available.

Investigation of the 55 stations revealed several long-reaching inconsistencies with how station data is obtained and recorded. Edits to account for some of these issues are thoroughly documented in the November 2002 report "Data Edits Determined From 'List of 55,'" prepared by Shannon McDonald and Susan Cosper and included as Appendix 5. The rationale for adding new station attributes as a consequence of these edits is provided in "New Urban and Municipal Field Value Assignments and Their Relation to Station Locations and the PTC Data Model Process," written November 2002 by Shannon McDonald and included as Appendix 6.

Following the second major update to the PTC count data, a new CD was delivered in October 2002 by Shannon McDonald. At this point, only 3,431 PTC stations were retained in the test area dataset because three stations were regarded as extremely difficult to predict their AADTs. This represented another major development in the project. In the interest of allowing additional progress on the project, the research team decided that no further edits would be conducted for the PTC count data of the test area.

3.1.2 Auxiliary GIS Station Data and Land Use Data

On November 27, 2001, Erik Brun, Susan Cosper, Jacqueline Hughes-Oliver, Shannon McDonald, Ann Strickland, Kent Taylor, and Larry Wikoff met to discuss issues, problems, and concerns that had arisen in relation to what kinds of GIS data are relevant and how distances should be calculated. Subsets of Erik Brun, Susan Cosper, Shannon McDonald, Kent Taylor, and Larry Wikoff had several subsequent face-to-face meetings, and the entire group had several detailed electronic and telephone discussions to further hammer out the issues. As a result, Erik Brun was able to transmit two related data files on December 19, 2001: one containing quasi-latitude-longitude locations of all stations; and the other containing count data, which consisted of data from the NCDOT Universe Database, land use variables, unique identifiers for linking stations to census data, and some road segment attributes such as number of lanes, speed limit, and level of access control.

At this point, we now had two sets of *PTC count data*, and difficulties were encountered in authenticating data from these and other sources, including the Universe Database, the LRS road layer, and the complete statewide road linework. Details are given in the February 2002 report "Data Edits, Queries, and Issues in Reconciliation" of Appendix 1. A few additional issues not included in the report are:

- In the five-county test area, three stations have been identified as having questionable values for their counties. Station 100 in sips/countyid 50 is currently listed as being in Johnston county, county90=163, tract90=9702, and group90=3. However, all other stations in Johnston county have county90=101 (not 163) and the census file does not contain a record for the combination of county90= 163, tract90=9702, and group90=3. Should this station really be listed as falling in Sampson County? This would change variables county, sips and countyid. Or should all of the variables county90, tract90, and group90 change for this station? Should station 82 in sips/countyid 18 have changes in variables county90, tract90, and group90? How about station 679 in sips/countyid 91? Several other stations outside the five-county test area also have questionable values for these variables.

- Can tract90 take the same value across different counties, that is, is this census variables defined within or across counties? For example, tract90=53404 is defined to be in Chatham, Durham, and Wake counties.

- The census file contains duplicate records for nine county90-tract90-group90 combinations. In all of these cases, one record contained mostly "zero" values for variables. The records containing zero values were deleted. For one county90-tract90-group90 combination, both records contained zero values and so both records were deleted.

As described in the previous subsection, two major edits were conducted. These edits are detailed in "Procedures and Corrections of PTC Data" (May 2002, Appendix 2) and "Data Edits Determined From 'List of 55'" (November 2002, Appendix 5). On April 16, 2002, Erik Brun submitted an updated version of the file containing count data from the NCDOT Universe Database, land use variables, unique identifiers for linking stations to census data, and road segment attributes such as number of lanes, speed limit, and level of access control. Additional updated data was received in May 2002, as described in the previous subsection.

During steps leading to creation of the "List of 55" outliers, it became apparent that more members of the project team needed the ability to individually query count stations for the purpose of identifying and understanding key features and deficits of the model. This knowledge would facilitate speedy and successful completion of the remainder of the project. As such, the GIS contractor Shannon McDonald delivered data and instruction for using NCDOT's GIS shapefile of PTC count stations. This half-day crash course was offered to the statisticians (Jacqueline Hughes-Oliver, Tae-Young Heo, Susan Cosper) during July 2002.

3.1.3 Census Data

The 1990 census data was received from Erik Brun as a comma-delimited text file on December 19, 2001. Census data was available at the group level, so all stations within a particular 1990 census group were assigned identical census values. Cross-referencing to this census informations was accommodated using the combination of four attributes: state90, county90, tract90, and group90.

We also obtained and experimented with the 2000 census data. At first, we replaced the 1990 census data with 2000 census data. This move was considered several months prior, but the research team decided that it may be too risky since the 2000 census data had not been completely screened for quality assurance. However, the fact that NCDOT's GIS Unit was working on completing the quality assurance for the 2000 census data, coupled with the fact that this 2000 census data was available on a much finer scale that would allow increased delineation in the model, encouraged us to replace 1990 census data with 2000 data for building the model. Unfortunately, the results offered no improvement beyond the model based on 1990 census data. The problem is that while the 2000 block-level census data is available on a much finer scale (a good thing), it has far fewer attributes than the 1990 block-group census data (a bad thing), thus resulting in a canceling effect that results in no improvement. In addition, some of the 2000 census blocks were so small that they only included road segments such as major highways. As a consequence, they were recorded as having null census values for things like population counts, and this affected the performance of predictive models. For these reasons, we discontinued investigations with the 2000 census data.

Edits to the 1990 census data were conducted along with edits as outlined in the previous two subsections for the count and GIS data. Relatively minor edits were needed, as expected, because the 1990 census data we chose to work with was quality assured.

3.1.4 Distances Between PTC Stations

Erik Brun and Shannon McDonald worked very hard to track down information for obtaining distances between stations. Distances between stations can be determined in a variety of ways. Here we consider two distances: point-to-point by straight line and point-to-point by the most likely traveled path. The point-to-point by straight line distance is very easy to calculate but is not expected to be particularly useful in this project. It is calculated as the simple Euclidean or "as the crow flies" distance using the quasi-latitude-longitude locations of the stations; it is the shortest distance between two points. The basic premise of a spatial analysis is that stations that are "close" to each other will have similar traffic patterns. But closeness is dependent on route connectivity, which is not usually the shortest distance between two points. On the other hand, how does one define the most likely traveled path?

On November 27, 2001, Erik Brun, Susan Cosper, Jackie Hughes-Oliver, Shannon McDonald, Ann Strickland, Kent Taylor, and Larry Wikoff met to discuss issues related to how distances should be calculated. In addition to not having a clear picture of how most-likely-traveled paths should be determined, another technical problem was encountered in determining the most efficient strategy for storing the required information on distances between stations. The team needed to find an effective strategy that would accommodate all 35,000 PTC stations, even though preliminary investigations were restricted to a test area of fewer than 3,500 stations. It was after several discussions that we agreed the source code approach was the most desirable. This approach was implemented for calculating the point-to-point by straight-line distances and the C+ source code was received from Erik Brun on December 19, 2001. These distances were useful for establishing baseline analyses.

The more interesting distances, however, are those along the most likely traveled path. Unfortunately, there was no universally agreed upon mechanism for determining such paths; we quickly realized that determining the distances could be a research project in itself. To complicate matters even more, some of the variables needed to determine most likely traveled paths were not even available. For example, much effort was expended in discussing route capacities and how these could be valuable in determining path preferences. But after further discussions we concluded that capacity would be too difficult to obtain and we instead decided to use variables that somehow provide information on, or are strongly related to, capacity. These variables include number of lanes, posted speed limit, and whether or not a segment has access control. At the time, these variables were included in the NCDOT Universe Database for about 80% of the stations, but they were absent for more than 10,000 stations. Because 20% missing-ness may invalidate the outcome of any statistical analysis, extended effort was applied to "filling in" these missing values. While detailed approaches were documented and implemented during 4th quarter 2001, they were not yet considered acceptable and work continued in this area. We searched for a much more complete Universe Database and/or more acceptable methodologies for filling in missing values. This would in turn lead to obtaining the point-to-point distances along the most likely traveled path. It is important to mention, however, that certain problems would persist even after the Universe Database was complete. Determining number of lanes was problematic because different directions of divided highways are entered as separate segments in the database. The consequence is that a segment on a divided highway may list the number of lanes as two when it should really be four. These interpretation difficulties cannot be avoided given the current setup of the database.

Several members of the research team met on February 21, 2002 to discuss determination of most likely traveled paths and their resultant distances. At this meeting, Kent Taylor provided several examples of most likely traveled paths and reviewed his reasoning for this designation. Shannon McDonald and Erik Brun used these examples as test cases for developing a routine to calculate distances along most likely traveled paths. During 1st quarter 2002, Shannon McDonald moved into office space in the same location as NCDOT's TSU and GIS units, and this allowed more efficient communication and faster progress towards completion of tasks. Additionally, the Universe Database had now been updated, with one consequence being that segment characteristics such as speed and number of lanes were more reliable. Details are given in the November 2002 report by Shannon McDonald "Recordings and Methodology To Calculating Capacity Variables For Unmatched Stations" in Appendix 7.

A consensus was reached during 2nd quarter 2002 that an optimal approach to determining most likely traveled paths was not attainable within the scope, budget and time requirements of this project. Instead, an intuitive (albeit ad hoc) approach was pursued. This approach had already been demonstrated to be far more effective than the naïve approach initially used to determine most likely traveled paths.

An October 28, 2002 meeting led to improvements in the AML code for determining these distances and for avoiding size limits during generation of the resulting very large files. The problems were not eliminated, however, and we anxiously awaited arrival of the new dedicated computer that would enhance the process of path generation. Shannon McDonald delivered distances for Chatham County in November 2002 and for Johnston County in December 2002. Distances for Orange County were delivered on January 27, 2003, for Durham County on March 14, 2003, and for Wake County on March 24, 2003. Files indicating which pairs of stations yielded "bad routes" were received on February 13, 2003, March 14, 2003, and March 24, 2003. These bad routes were mainly caused by misalignment in the GIS network. Inconsistencies found in the bad route files were reported, and these bad routes were subsequently replaced with corrected information by April 2003.

Most likely traveled paths were complete for the test area by April 2003. After analysis within the test area to determine maximum ranges, distances smaller than maximum ranges would be obtained for all stations in the entire state.

3.1.5 Scaling to the Entire State

All of the census data, land use data, station locations in State Plane Coordinates, shapefile, number of lanes, speed limit and level of access were provided for the entire state at once. PTC count data and distances between PTC stations were the only data types that were delivered in phases, first for the test area then for the entire state.

Statewide PTC count data was delivered in October 2002. It was very similar in format to the final version of the PTC count data for the test area, with one major exception. None of the modifications identified from the "List of 55" outliers were applied to this data. This caused some minor difficulties during the model fitting stage of the project because some attributes that were important in the mean model for the test area were either missing or incorrectly defined in the statewide PTC count data. The research team concluded that the slight degradation in the quality of the model was worth avoiding additional excessive delays in getting the data more complete. These details will be addressed when the models are updated in the future.

Distances between PTC stations for the entire state were delivered on September 30, 2003. A 100-mile buffer was used in that most likely traveled paths were determined only for station pairs that were less than 100 miles apart in Euclidean distance. Covariance model estimates from the test area indicated that effective spatial ranges were always much less than 100 miles, so allocating resources to determining distances beyond 100 miles would have been inefficient. Station pairs outside this buffer were assigned a distance of 100 miles. The following is a quote from Larry Wikoff on the generation, format and size of the resulting distance file,

> "- Using a program written by Bill Miller of GIS, the statewide text file has been merged with the Interstate-US text file, tossing out records with duplicate combinations of the FromStation and ToStation fields, and including only the fields for FromStation, ToStation, Distance, and Cost.
> - The merged text file has been sorted by FromStation and ToStation and written to Sorted.txt. It has 222,426,128 records, and is about 7.9G in size."

The resulting file was so large it had to be delivered on a computer. This computer was then used by Tae-Young Heo for the remainder of the project since he needed to continue accessing the file of distances during the modeling steps. "ToStation" is the UNIQ_ID for the originating point of the path, while "FromStation" is the endpoint of the path. Order was ignored in creating this file, so the record for station pair (1,2) also served as the record for station pair (2,1). Although this was a natural (and wise) thing to do, it was not without its problems. Stations lying on one-way roads might have a short distance when traveling from A to B but have a long distance when traveling from B to A. We ignored such problems.

Distances between PTC stations where both stations are on primary road segments were delivered February 6, 2004. Primary road segments can be either interstate, US, or NC routes. To obtain these distances, a primaries-only road network was created, and most likely traveled paths were limited to only traversing primary routes. This file was also subjected to a 100-mile buffer. The file was made accessible by ftp, with file size of 523 MB and 14,686,050 records.

3.2 Spatial Modeling and Characterization, Using Current Seasonality Adjustments, Task 2

Spatial modeling provides two component models, the mean model and the covariance model. The mean model explains variability in expected AADT as a function of systematic variations (e.g., route type, speed, number of lanes, land use in surrounding area, census information, etc.). For example, the mean model might explain how AADT increases or decreases according to whether the PTC station is located on an interstate or on a local road segment. On the other hand, the covariance model explains the amount of linear relationship that exists between AADT counts for two stations, where the strength of the relationship is usually a function of spatial proximity of the two stations. For example, two stations located 2 miles apart on an interstate might be expected to be more related than two stations located 20 miles apart on the same interstate.

The original intention of the principal investigator Jacqueline Hughes-Oliver was to fit a spatial model using a likelihood approach that would simultaneously estimate both the mean model and the covariance model. This was not possible for two major reasons:

1. There were significant delays in obtaining the data. A joint approach would require all the data, including distances between stations. While point-to-point by straight line distances were easy to obtain, they were not appropriate for this project. The more appropriate point-to-point by most likely traveled path distances were not complete for the test area until April 2003. We could not wait until then before beginning work on Task 2.
2. Size of the dataset (approximately 35,000 PTC stations) exceeded the capabilities of all software options for joint modeling. Even the test area data, which only contained 3,431 PTC stations, was too large. Not even the SAS system, which is well known for being able to handle large files, could accommodate the file for the kind of analyses we needed. The major difficulty came from needing to repeatedly calculate inverses and determinants of n x n matrices, where n is the number of stations. These matrices were functions of the point-to-point by most likely traveled path distances that needed to be stored since they could not be quickly calculated as needed in a just-in-time manner.

Although the joint modeling approach is generally considered to be optimal, the so-called estimated generalized least squares or variogram approach is also quite popular and effective. In this approach, one first estimates the mean model, applies it to obtain residuals (actual AADT counts minus mean-model-predicted AADT counts) for all stations, uses the residuals to estimate the covariance model, and then combines the estimated mean and covariance models to obtain predictions for each station. New residuals could then be obtained to iterate the process of estimating the covariance model, but this is not commonly done. Because of the limitations listed above, we use the estimated generalized least squares approach to separately estimate a mean model and a covariance model, as outlined in the following four subsections. The fifth subsection explains how we obtained predictions and prediction intervals to determine station-specific anomalies and levels of extremeness.

3.2.1 Mean Model for Test Area

The July 2003 report “Determining the Final Mean Model for the 5-County Test Area” by Jacqueline Hughes-Oliver and Tae-Young Heo (Appendix 8) details the fitting process. We provide a summary here.

We performed multiple linear regression using summarized census information and PTC station attributes. All programming was done using SAS/STAT® software, Version 8 of the SAS System for Windows. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Of the more than 300 possible attributes, the final mean model for the test area included only 60 attributes, five of which were summaries of the census data. Table 1 shows the attributes that contributed to the different census summaries, as obtained from principal components analysis using the varimax rotation of components from the covariance matrix. Factors were defined according to the largest absolute rotated coefficient.

Table 1: Summary of Rotated Principal Components Retained for Test Area

Principal Component	% variability explained	Attributes in Factor (- means negative impact on component, 0 means negligible impact on all components). Attribute names are explained in Appendix D of Appendix 8.
Factor2: Single	9.03	APARTMNT, ROOM_1_3, RNT_2550, ONEPERHH, RNT_5075, VEHICL_1, SEASONAL, PUBSEWER, INC_1525, ATTACHED, DUPLEX
Factor3: College	7.13	AGEIS_20, COLLEGE, AGE18_19, DORMITOR, SINGLE_, SAMESTAT, AGE21_24, ASIAN, ABLEENGL, AMIND, OTH_HEAT, O_NOINST, INSTREET, MENTAL
Factor4: Poor	5.05	CHILDPOV, NOVEHICL, RNT_LT25, PUBL_INC, INPOVRTY, BLACK, SINGWICH, INC_LT15, NODIPLOM, PUBTRANS, VAL_2550, INCPRCAP(-), INC_MEDN (-)
Factor5: Farm Life	3.10	LAND_KM, COALWOOD, BOTL_GAS, FARM_INC, VAL_LT25, F_MIGRNT, WATER_KM, RNT_MEDI (-), POPDEN (-), HOUSEDEN(-), HOUSHOLD (-), FAMILIES (-)
Factor6: Elderly	2.52	AGE65_74, AGE75_84, SOCS_INC, WIDOWED, RETI_INC, FUELKERO, BLTBFR70, NURSHOME, MEDYRBLT, O_INSTIT(0), SHELTERS(-)
Factor7: Wealthy	1.79	VAL_2C3C, VAL_GT3C, RNT_751K, RNT_GT1K, VAL_MEDI, CORRINST(-)
Factor1: Other	48.25	This component includes all other attributes, including middle income families with and without children and those who are employed.

Table 2 summarizes all 60 attributes appearing in the final model. "Baseline attributes" may be regarded as having coefficient zero. Positive coefficients imply that AADT increases as the attribute increases, while negative coefficients imply AADT decreases as the attribute increases. The best model required a power transformation of AADT, namely $(AADT)^{0.2}$. Diagnostics for this transformation indicated that the normality assumption was reasonably well supported. Therefore, the regression analysis actually modeled $(AADT)^{0.2}$ as a function of systematic variants. R-squared for this model was 0.7766, with Mallow's Cp of 31.1.

Table 2: Summary of Attribute Effects in Mean Model for Test Area

Category	*Baseline attribute***.** Other attributes ordered by coefficients, from largest to smallest, with "|" indicating a sign change
Census data	Factor1, Factor2, Factor7, Factor3 | Factor5
Route classification	*Local*. Interstate
County—urban #	*Wake—108*. Orange | Johnston, Durham—103
Lanes-by-route interaction	*8—US*. 8—interstate | 4—SR, 2—NC, 2—US, 2—SR, 2—local
Speed-by-route interaction	*70—interstate*. 55-US | 55—interstate, 25—SR
Lanes-by-speed-by-route interaction	*8—65—interstate*. 4—55—NC, 8—45—US, 2—25—SR, 2—45—US, 6—65—interstate, 6—45—US, 4—45—US | leftover, 2—55—SR, 4—35—NC, 3—35—

	US, 6—35—SR, 8—60—interstate, 3—35—SR, 3—30—SR
Counting cycle-by-road surface interact	*Variable—concrete*. Odd—soil, variable—asphalt, annual—asphalt, even—concrete, variable—concrete, even—soil, odd—concrete
Access control	*High, i.e., interstate*. | Medium
Year	*2001*. | 1999, 1998
Month	*Missing*. November | February, June, January, May
Day	*Missing*. | Thursday
Land use	*Transitional*. Other Ag Land, Indust & Comm Complxs, Trans Comm Util, Comm & Services, Mxd Urban or Built-up | Residential, Shrub & Brush Rangeland
City & urban location	*City—minor urban*. City—major | city—outside, outside—major, outside--outside

3.2.2 Mean Model for Entire State

The March 2005 report "Determining the Final Mean Model for the Statewide Area, Version 2" by Jacqueline Hughes-Oliver and Tae-Young Heo (Appendix 9) details the fitting process. We provide a summary here.

Again, we performed multiple linear regression using summarized census information and PTC station attributes. All programming was done using SAS/STAT software.

Of the approximately 550 possible attributes, the final mean model for the entire state included only 178 attributes, five of which were summaries of the census data. Table 3 shows the attributes that contributed to the different census summaries, as obtained from principal components analysis using the varimax rotation of components from the covariance matrix. Factors were defined according to the largest absolute rotated coefficient.

Table 3: Summary of Rotated Principal Components Retained for Statewide Area

Principal Component	% variability explained	Attributes in Factor (- means negative impact on component, 0 means negligible impact on all components). Attribute names are explained in Appendix D of Appendix 8.
Factor1: With Kids	44.68	BORNWEST, OTH_STAT, HISPANIC, ARMDFORC, AGE21_24, COM_LT15, ABLEENGL, BORNMIDW, MILIQUAR, BORNSOUT, WORKERS, MALE, WRK_HOME, WRKINCTY, AGEIS_20, AGELT_05, CARPOOL, SINGLE_, AGE25_34, AGE05_09, NATIVE, AGEGE_03, AGEGE_16, BORNNORT, ONLYENGL, ASIAN, ATTACHED, MARRWICH, AGE18_19, SOMECOLG, NATURAL_, INC_1525, MARRIED, WHITE, BLACK, ELEMSCND, PREPRIMA, PERINFAM, AGE10_14, PRUNIT34, SEPARATE, RNT_2550, DRVALONE, HIGHSCHL, COM_1529, VEHICL_1, ROOM_4_6, ELECTRIC, WAGE_SAL, VEHICL_2, AGE35_44, UNEMPLOY, PUBSEWER, COLLEGE, BLTBFR70, INC_2535, PUBWATER, CHILDPOV, AGE15_17, DUPLEX, PUBLWORK, APARTMNT, OTH_HEAT, PUBTRANS Note that this component includes middle income families with children.
Factor2: No Kids	13.41	PRIVWORK, EMPLOYED, OWNR_OCC, AGE45_54, TECHSALE, TRADE, MARRNOCH, PRUNIT12, INC_5075, INC_3550, EXECPROF, UTILITY, SELF_INC, VAL_501C, SERVICES, FIRE, SERVICE, DETACHED, BLTAFT84, AGE55_64, INTE_INC, BLT_8084, SAMEHOUS, BORN_INS, PROFSERV, ONEPERHH, BLT_7079, SELFWORK, COM_3044, MANUFACT, SAMECNTY, PRIMARY_, VAL_1C2C, COLGGRAD, RETI_INC, WRKEXCTY, ROOM_1_3 Note that this component includes middle income families with no children.
Factor3: Poor	6.98	PUBL_INC, NOVEHICL, INC_LT15, RNT_LT25, INPOVRTY, SINGWICH, VAL_LT25, NODIPLOM, AMIND, RNT_MEDI(-), VAL_MEDI(-), INCPRCAP(-), INC_MEDN(-)
Factor4: City Life	5.29	FAMILIES, HOUSHOLD, HOUSEDEN, POPDEN, PUBL_GAS SHELTERS, F_MIGRNT(-), BOTL_GAS(-), FARM_INC(-), COALWOOD(-), LAND_KM(-)
Factor5: Elderly	2.26	AGE75_84, AGE65_74, WIDOWED, SOCS_INC, FUELKERO, NURSHOME, MEDYRBLT, O_INSTIT
Factor6: Wealthy	1.88	RNT_751K, VAL_GT3C, VAL_2C3C, RNT_5075, RNT_GT1K, SEASONAL, WATER_KM, INSTREET(0), VAL_2550(-)
Factor7: Unstable	1.64	DORMITOR, SAMESTAT, MENTAL, CORRINST, O_NOINST

Table 4 summarizes all 178 attributes appearing in the final model. "Baseline attributes" may be regarded as having coefficient zero. Positive coefficients imply that AADT increases as the attribute increases, while negative coefficients imply AADT decreases as the attribute increases. The best model required a power transformation of AADT, namely $(AADT)^{0.15}$. Diagnostics for this transformation indicated that the normality assumption was reasonably well supported. Therefore, the regression analysis actually modeled $(AADT)^{0.15}$ as a function of systematic variants. R-squared for this model was 0.7287, with Mallow's Cp of 174.2.

Table 4: Summary of Attribute Effects in Mean Model for Statewide Area

Category	*Baseline attribute*. Other attributes ordered by coefficients, from largest to smallest, with "|" indicating a sign change
Census data	Factor4, Factor2, Factor1, Factor6 | Factor3
Route classification	*Local*. Interstate, US, SR
County—urban #	*Yancey*. Brunswick—109, Orange, Mitchell, Franklin, Guilford—107, Haywood, Brunswick, Hertford, Catawba, Wake, Northampton, Caldwell—114, Lenoir, Cabarrus—111, Richmond, Wake—108, Harnett—104, Lincoln, Onslow—112, Orange—101, Wilkes, Caldwell, Macon, Carteret, Pamlico, Forsyth, Rowan—111, Cabarrus, Burke—114, Burke, Davison, Cumberland—104, Durham—103, Vance, Pork, Swain, Jones, Hyde, Johnston, Catawba—114, Rockingham, Randolph—107, Gaston—105, Craven, Davison—107, Pitt, Yadkin, Buncombe, Henderson, Chatham | Scotland, Beaufort, Wayne, Robeson, Alamance—101, Rutherford, Pitt—115, Stokes, Tyrrell, Currituck, Avery, Bladen, Guilford, Chowan, Transylvania, Nash—116, Warren, Cherokee, Washington, Bertie, Caswell, Ashe, Camden, Lee, Harnett, Hoke, Madison, Edgecombe, Pasquotank, McDowell, Montgomery, Clay, Onslow, Randolph, Dare, Edgecombe—116, Union, Iredell, Graham, Granville, Moore, Martin, Anson, Alleghany, Guilford—106
Lanes-by-route interaction	*12—interstate*. 8—interstate, 6—NC, 5—SR, 4—NC, 4—SR | 3—US, 2—NC, 2—US, 2—SR, 2—local, 1—SR
Speed-by-route interaction	*70—interstate*. 45—SR, 35—SR | 50—interstate, 55—NC, 40—US, 35—NC, 60—US, 25—NC, 20—SR, 20—US, 25—local
Lanes-by-speed-by-route interaction	*12—55—interstate*. 2—20—US, 2—25—local, 6—50—US, 6—65—US, 5—45—SR, 4—55—SR, 4—60—NC, 2—50—SR, 3—55—NC, 3—45—US, 6—55—US, 4—55—NC | 4—45—NC, 2—60—US, 3—35—US, 2—25—US, 8—45—US, 8—60—interstate
Counting cycle-by-road surface interact	*Variable—soil*. Variable—Asphalt, Annual— Asphalt, Annual— Concrete, Variable— Concrete, Even— Concrete, Even— Soil, Odd— Concrete, Odd—Soil
Access control	*High, i.e., interstate*. | Medium
Year	*2001*. | 2000
Month	*Missing*. | October, January, June, November, July, May, September, August
Day	*Missing*. | Monday, Thursday
Land use	*Transitional*. Sandy Area (Non-Beach), Other Agricultural Land, Trans,Comm,Util, Indust & Commerc Cmplxs, Comm & Services, Other Urban Or Built-Up, Industrial | Forested Wetland, Evergreen Forest Land, Mixed Forest Land, Cropland And Pasture, Deciduous Forest Land, Confined Feeding Ops, Shrub & Brush Rangeland
Urban location?	*Urban*. | Rural

3.2.3 Covariance Model for Test Area

Using residuals from the mean model obtained for the test area, we estimated the covariance structure by first obtaining the empirical variogram. Let y_i represent $(\text{AADT})^{0.2}$ for the i^{th} station, $\hat{y}_i$ represent the predicted $(\text{AADT})^{0.2}$ as obtained from the mean model, $e_i = y_i - \hat{y}_i$ be the residual for the i^{th} station, and

i ranges from one to the sample size of $n = 3{,}431$ for the test area. Then the empirical variogram is a function of the distance between two stations and is obtained as

$$2\hat{\gamma}_{+}(h) = \frac{1}{|N_{+}(h)|}\sum_{N_{+}(h)}(e_i - e_j)^2 ,$$

where the sum is over all (i, j) pairs of stations that are h units apart and $|N_{+}(h)|$ is the number of such pairs. For irregularly spaced locations, as is the case for the PTC stations, the variogram estimator is usually smoothed to address the fact that stations typically are not separated by a small number of distinct distances. The smoothed empirical variogram estimator, which is our default estimator, is defined as a function of $h \in T(h(l))$, where the region $T(h(l))$ is some specified "tolerance" or buffer region around $h(l)$, for $l = 1, \ldots, k$. The equation is

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}(e_i - e_j)^2 ,$$

where $N(h)$ represents the set of (i, j) pairs falling in $T(h(l))$.

The empirical variogram estimates the true variogram

$$2\gamma(h) = \text{var}(e_i - e_j)$$

for stations *i* and *j* that have distance *h* units between them, where "var" represents variance. Having obtained empirical variograms, we then used both unweighted (ordinary) and weighted nonlinear least squares to fit several theoretical variograms, from which we selected the best fit. All programming was done using the SAS software. The effective range of a theoretical variogram model is defined as the distance at which the variogram increases 95% from its minimum nonzero value to its maximum possible or asymptotic value. This may be interpreted as the distance at which spatial correlation is ignorable or even zero. In other words, the effective range is the distance at which spatially recorded measurements may reasonably be regarded as nearly, if not completely, independent. As such, effective range is an important parameter to be estimated for any theoretical variogram model. Partial sill is the change in variogram between its maximum (or asymptotic) value and minimum non-zero value; a small partial sill indicates weak dependence or possibly independence. The nugget is the minimum non-zero value of the variogram, with variance being the sum of nugget and partial sill.

Under an assumption of weak stationarity, the covariance structure may be determined from the variogram using

$$2\gamma(h) = 2\,\text{var}(e_i) - 2\,\text{cov}(e_i, e_j),$$

where "cov" represents covariance. Weak stationarity assumes that any pair of sites separated by distance *h* will have the same covariance, irrespective of their actual locations. But is this a reasonable assumption? Two stations five miles apart on an interstate might well be more correlated than two stations five miles apart on a non-interstate road segment. If nonstationarity is an issue, can it be adequately addressed by choice of distance metric? Tae-Young and Jacqueline Hughes-Oliver investigated several options of distance metrics.

Euclidean Distances Are Not Appropriate. In lieu of the complete set of distances along most likely traveled paths (which were not yet available), Tae-Young Heo obtained sub-optimal road distances for subsets of stations and used these distances in modeling spatial structure. This work was conducted during 3rd quarter 2002. We classified our "road distances" as sub-optimal because they were based solely on route lengths, not likelihood of route selection. By taking this simplified approach, we were able to

conduct a preliminary investigation of our prior assumption that Euclidean distances would not be acceptable for building the covariance model.

Fitted theoretical variograms obtained using both Euclidean and road distances resulted in estimates as shown in Table 5. Using Euclidean distance, both the partial sill and effective range are estimated very small. As previously discussed, small partial sill and effective range indicate weak dependence or possibly independence. Both the partial sill and effective range are much larger when obtained from road distances. The summary is that Euclidean distances obscure correlation structure because of the manner in which they group pairs of stations.

Table 5. Parameter estimates for the exponential variogram fitted by nonlinear ordinary least squares to interstate stations in the test area based on both Euclidean and road distances.

Distance	*Partial Sill*	*Effective Range*	*Nugget*
Euclidean	0.0701	0.9243 miles	0.2281
Road	0.3821	3.7778 miles	0.249

To understand how this happens, consider Figure 1 where we display locations of all 82 interstate stations of the test area. More specifically, consider station pairs (1,2), (2,3) and (1,3) on the I-440 beltline. Existing evidence supports the belief that because station pairs (1,2) and (2,3) are about the same driving distance of *x* meters apart, their values of $(e_1 - e_2)^2$ and $(e_2 - e_3)^2$ will be similar. As the same time, the driving distance between stations 1 and 3 is about twice that for (1,2) and (2,3), so evidence suggests that $(e_1 - e_3)^2$ will be much larger than both $(e_1 - e_2)^2$ and $(e_2 - e_3)^2$. Recall that the computational formula for the empirical variogram requires grouping pairs of stations by their distances then averaging the values of $(e_i - e_j)^2$ within these groups. Based on road distances, station pairs (1,2) and (2,3) get grouped together while (1,3) is placed in a separate group. As seen in the hypothetical variogram cloud of Figure 2, these road distance groupings lead to a steadily increasing empirical variogram, which is suggestive of correlation. On the other hand, Euclidean distances group all three stations pairs (1,2), (2,3) and (1,3) together, so that the resulting averages lead to a variogram that increases at a slower rate. This, in turn, suggests there is reduced correlation than is indicated by road distances.

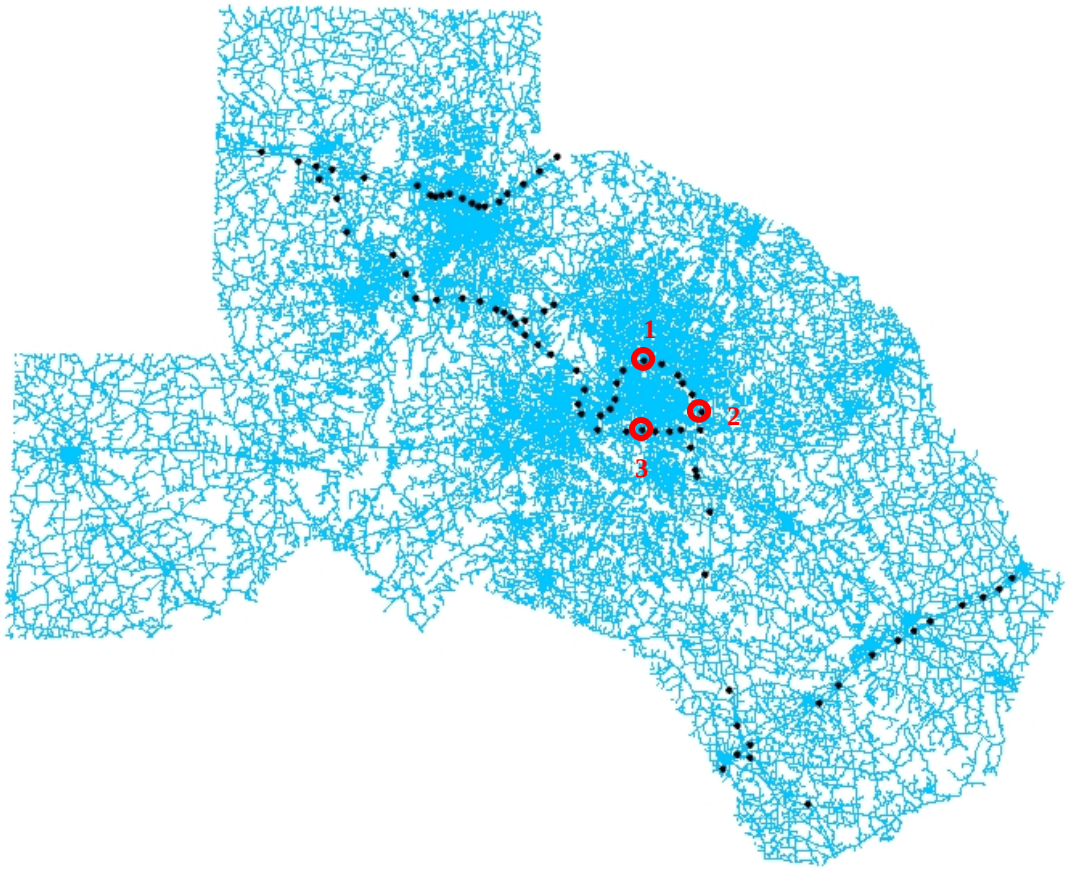


Figure 1. Interstate stations in the test area. Blue lines show road segments and bullets are PTC stations. East-west stations at the top of the figure are located on I-85. Northeast-southwest stations at the bottom right of the figure are on I-95. I-440 is the loop in the center of the figure, and the remaining stations are on I-40.

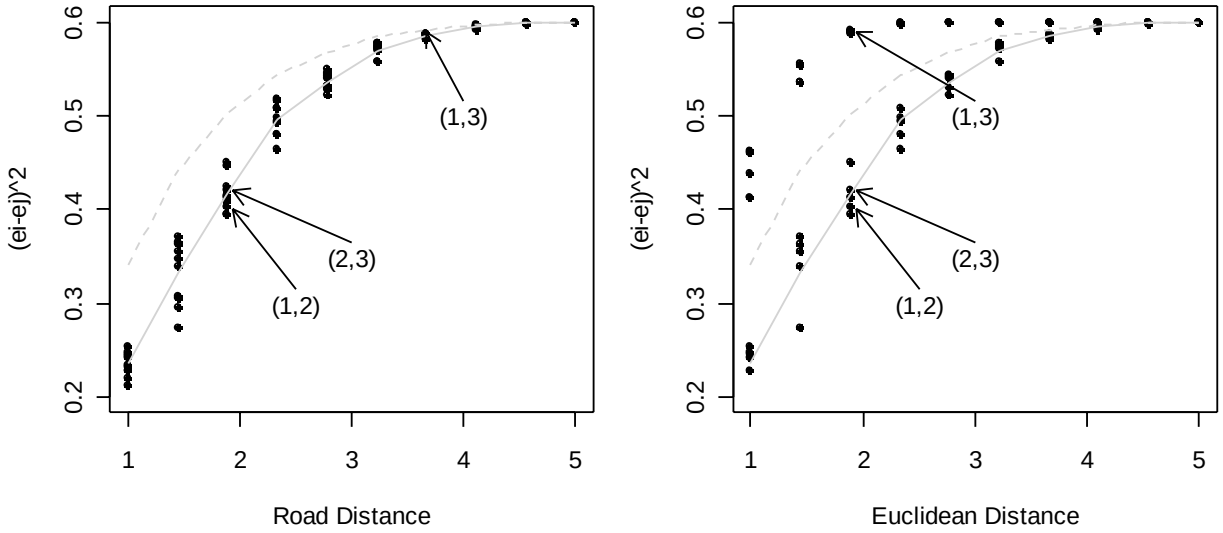


Figure 2. Hypothetical variogram clouds, where $(e_i - e_j)^2$ *is plotted as a function of distance between*

stations i and j, for road and Euclidean distances. The solid curve is the empirical variogram implied by road distances and the dashed curve is the empirical variogram implied by Euclidean distances.

Road Distances Are Reasonable. Among research team members, it was generally expected that effective range would be largest within the test area for interstate stations. US and NC stations were expected to have the next largest effective ranges. Because of significant heterogeneity among SR and local routes, no predictions were made concerning where their effective ranges would fall in an ordered list. To validate our expectations, and to investigate whether nonstationarity was an issue, we separately estimated the variogram for each of the different route types: interstate, US, NC, SR, and local. Because this work was also done prior to getting most likely traveled paths, we again used the sub-optimal road distances introduced above.

- We considered subsets of stations on: I-40, I-95, I-440, I-85, US-70, US-15/501, and NC-55.
- Using an exponential variogram structure, we obtained estimates of variances and effective ranges
- The results, in the order variance then effective range, are:
    - I-40: 0.72, 9.87 miles
    - I-95: 1.09, 10.72 miles
    - I-440: no spatial dependence
    - I-85: no spatial dependence
    - US-70: 1.19, 4.04 miles
    - US-15/501: no spatial dependence
    - NC-55: 0.82, 4.40 miles.

Given the high use of I-40 by commuters to and from the Research Triangle Park, the effective range of 9.9 miles for I-40 is not a surprise. Neither is the 10.7 mile effective range for I-95 given the long stretch of I-95 in Johnston County. The test area includes approximately 70 miles of I-40 and approximately 30 miles of I-95. Behavior for I-440 seems reasonable since this short route of approximately ten miles accommodates travelers who enter the highway and exit very soon thereafter; no spatial dependence is reasonable. Results for I-85 are also reasonable when one considers that we are still only considering stations in the test area of five counties, and the approximately 30-mile I-85 section within this area is heavily used for short-distance commuters in Durham. Similar arguments may be applied to the other subsets of count stations.

Distance or Time? Using most likely traveled paths for all stations in the test area as obtained by March 2003, we investigated dependence structures across route types using two “distance” metrics, namely, actual distance (meters along the path) and time-to-traverse (seconds to travel the path). A decision on the best distance metric was not trivial. Because a most likely traveled path can include road segments of varying classifications and speed limits, actual distance between stations A and B can be the same as actual distance between stations C and D, but time-to-traverse these paths can be quite different. Likewise, time-to-traverse paths A—B and C—D can be the same but actual distances can be different. The extent of disagreements between these metrics can be quite large, as is seen by the spread in Figure 3 and Figure 4 . Figure 3 and Figure 4 show scatter plots of actual distances versus time-to-traverse most likely traveled paths for Interstate and US routes. Graphs are similar for other route types. It is important to note that spread increases as you move away from the origin, thus suggesting that the two metrics are having diverging effects. Also important is the fact that the spread is larger for US routes, thus suggesting that choice of metric may have greater impact on US routes than on Interstate routes.

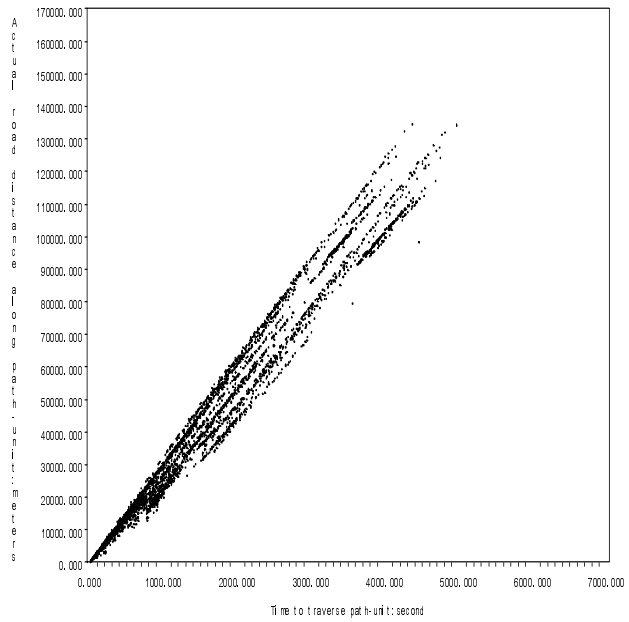


Figure 3. Actual distance on path (meters) vs. time-to-traverse path (seconds) for Interstate routes.

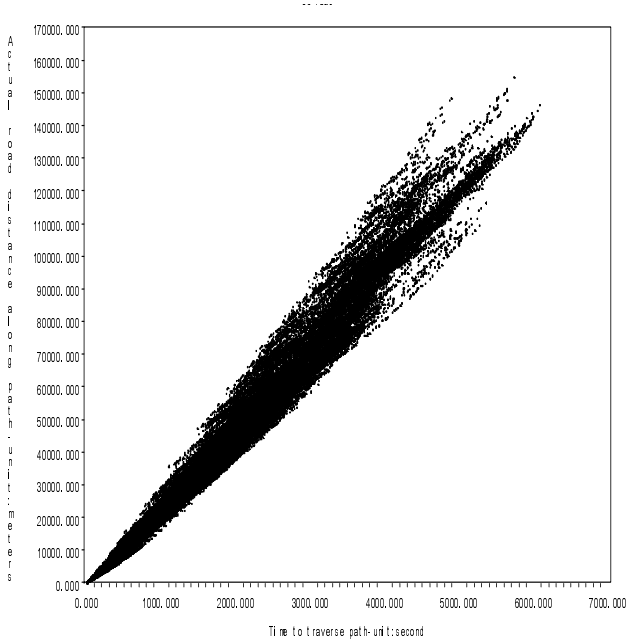


Figure 4. Actual distance on path (meters) vs. time-to-traverse path (seconds) for US routes.

Theoretical variogram models were fit to empirical variograms obtained from both actual distance and time-to-traverse along most likely traveled paths. During the last quarter of 2002 and the first three quarters of 2003, much effort was devoted to modeling the variogram using both actual distance and time-to-traverse along most likely traveled paths for the test area. Several theoretical variogram models were fit, including exponential, Gaussian, and spherical. These nonlinear models were fit using both ordinary and weighted least squares. Chatham and Johnston Counties were investigated separately and then together. Being the counties with fewest stations in the test area, we received their most likely traveled paths first. Their smaller sizes also allowed us to experiment with several options for computing variograms; timing studies were done to determine optimal strategies that would be needed for the larger counties within the test area and for the entire state.

Variograms were fit for individual route types and for all route types together. As previously mentioned, we were concerned about nonstationarity, which could be indicated by variograms that differ across route types, and about whether results are reasonable in that they are consistent with the expert opinion of TSU staff. Lag values for calculating empirical variograms were determined separately for each route type to ensure that each lag contained at least 30 points. Given that only 83 interstate stations existed in the test area, and so there are only 83x82/2=3,403 distances, empirical variograms for interstate stations were highly variable, especially when compared to SR routes, which consisted of 2,444 test area stations. This in turn implied varying degrees of uncertainty for theoretical variogram models built for the different route types.

Figure 5 shows effective ranges for the different route types using both distance and time as the metrics. In all cases, the Gaussian theoretical variogram model provided the best fit. All plotted ranges were obtained using nonlinear ordinary least squares, except distance-based range for US routes, which was obtained using nonlinear weighted least squares. From Figure 5, we see that distance-based ranges are more consistent with prior belief that, at least within the test area, interstate stations should have the largest ranges. For both distance-based and time-based approaches, effective ranges are smallest when all route types are combined. This is not surprising, since these variograms include station pairs that cross

route type categories. A station pair having an interstate and a local station is not expected to be as correlated as a station pair of two interstate stations.

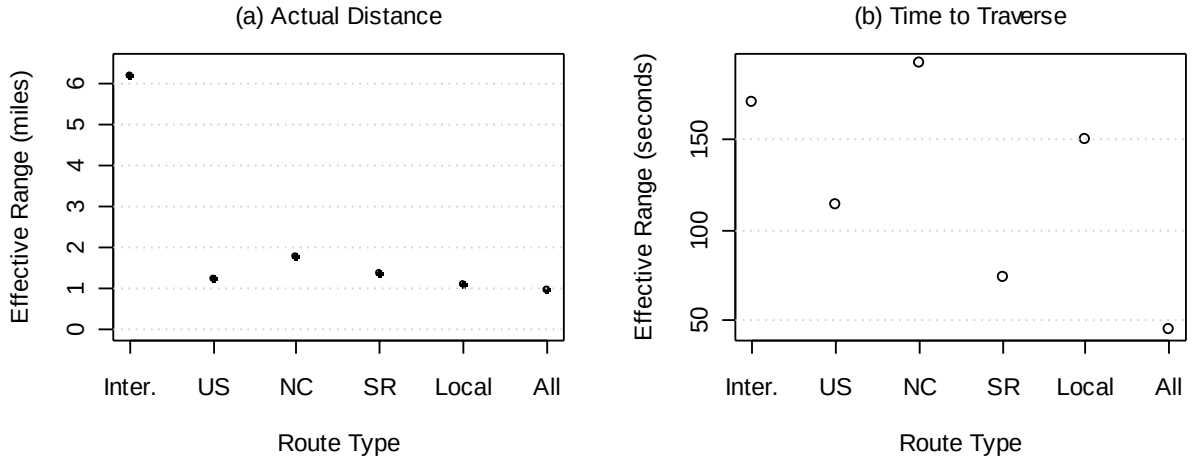


Figure 5. Estimated effective ranges of the Gaussian variogram model determined using both actual and time metrics along most likely traveled paths for each route type. All ranges, except for distance-based on US routes, were estimated using nonlinear ordinary least squares. Nonlinear weighted least squares was used for distance-based on US routes.

Our final decision was to model spatial covariance using actual distance along most likely traveled path with the Gaussian variogram

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \theta_0 + \theta_s \left\{ 1 - \exp\left[-\left(\frac{h}{\theta_r} \right)^2 \right] \right\} & h \neq 0 \end{cases},$$

having corresponding covariance function

$$C(h) = \begin{cases} \theta_0 + \theta_s & h = 0 \\ \theta_s \exp\left[-\left(\frac{h}{\theta_r} \right)^2 \right] & h \neq 0 \end{cases}$$

and correlation function

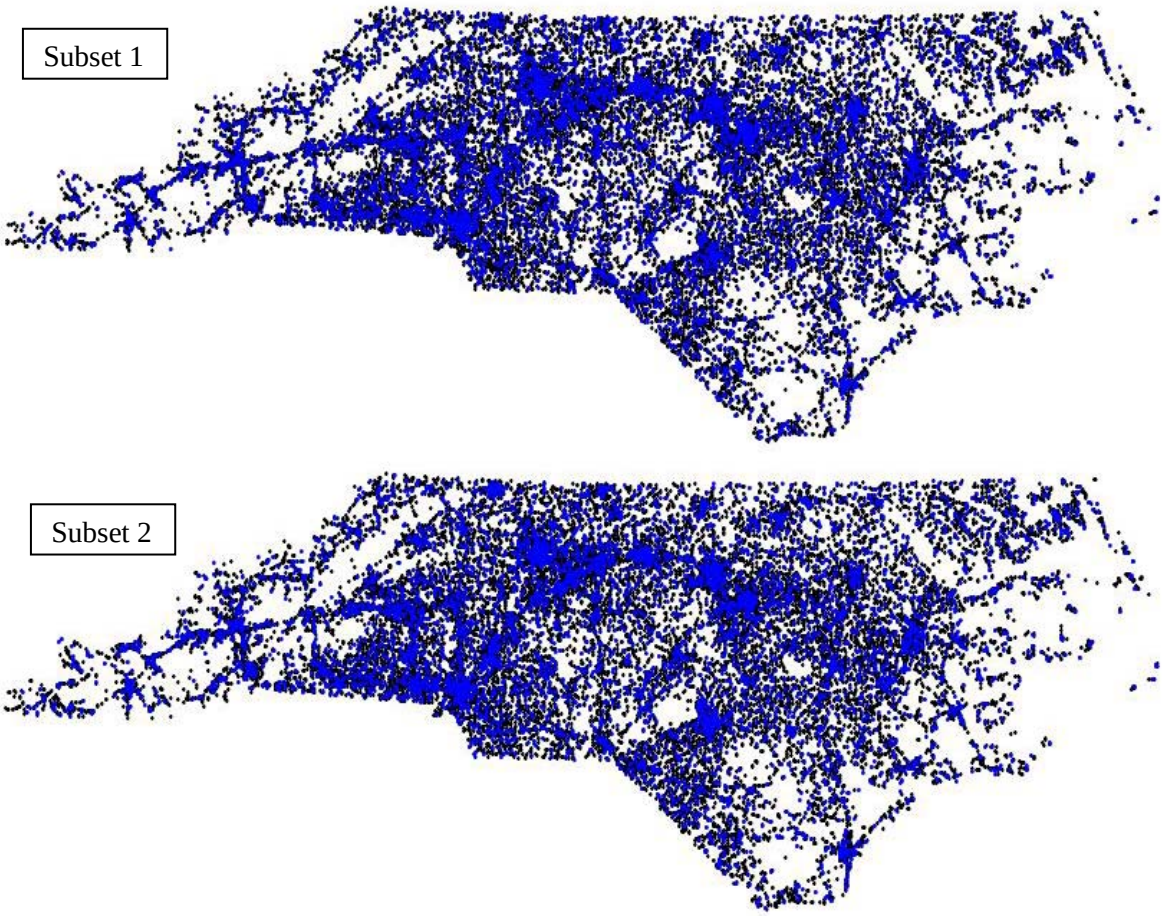
$$\rho(h) = \begin{cases} 1 & h = 0 \\ \frac{\theta_s}{\theta_0 + \theta_s} \exp\left[-\left(\frac{h}{\theta_r} \right)^2 \right] & h \neq 0 \end{cases}.$$

3.2.4 Covariance Model for Entire State

Primaries-and-Secondaries Network. There were 34,944 PTC stations to be modeled across the state. Starting with residuals from the statewide mean model, we attempted to estimate the all-route-types variogram as outlined in the previous subsection. Unfortunately, computations stalled. Even the NCDOT dedicated dual processor computer could not accommodate the required computations. Our only alternative was to use a subset of the 34,944 stations to build a covariance model. A 15% subseting

approach was undertaken because 5,248 stations was the largest subset for which we could perform the necessary computations.

Our subsetting was done using stratified random sampling. Strata were defined according to the combination of county name and route type. We considered it important to have representation in the subset from all route types within all counties. For each county and route type combination, 15% of their stations were randomly selected to be included in the subset. Three such subsets are displayed in Figure 6. Black bullets show all 34,944 PTC stations. Overlaid as blue bullets, we show the subsetted stations. It is clear that while representation is not perfect, it is also extremely good given the amount of savings in reducing to only 5,248 stations.



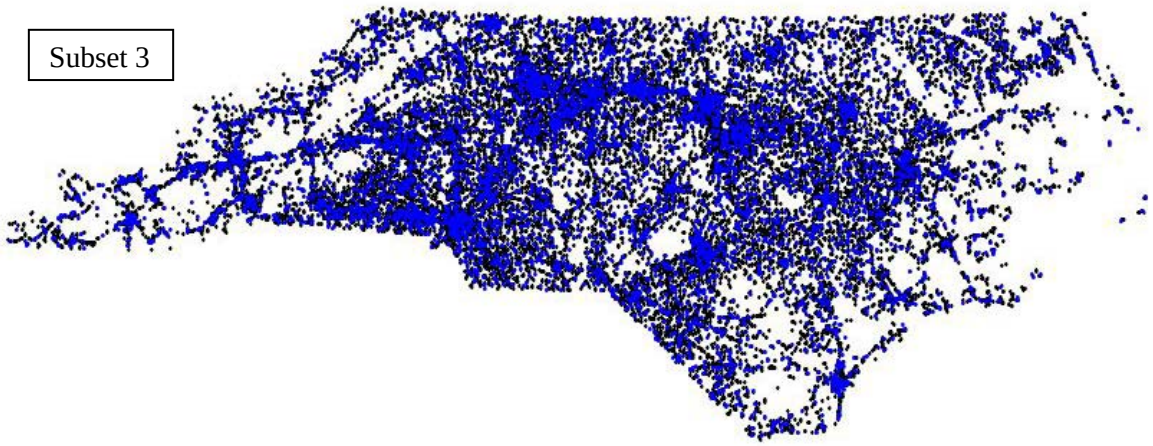


Figure 6. Three different 15% subsets of statewide PTC stations. Black: full set of 34,944 stations. Blue: selected for subset. These subsets were used to estimate the statewide covariance model.

Following findings from developing the covariance model for the test area, we used actual distances along most likely traveled paths to fit Gaussian variogram models to each of the subsets. Parameter estimates were obtained as

		Subset 1	Subset 2	**Subset 3**
Nugget	$\hat{\theta}_0$	0.102	0.100	**0.095**
Partial sill	$\hat{\theta}_s$	0.045	0.052	**0.050**
Variance	$\hat{\theta}_0 + \hat{\theta}_s$	0.147	0.152	**0.145**
(meters)	$\hat{\theta}_r$	1090.4	506.4	**676.1**
Effective range (meters)	$\sqrt{3}\hat{\theta}_r$	1888.6	877.1	**1171.0**
Effective range (miles)		1.17	0.55	**0.73**

We chose to use estimates from subset 3 since these lie between or close to estimates for the other two subsets. The resulting correlation function is plotted in Figure 7. It assigns correlations less than 0.30 even for stations as close as 0.15 miles apart. In other words, when viewed at the statewide level, PTC stations are not highly correlated after accounting for systematic variation.

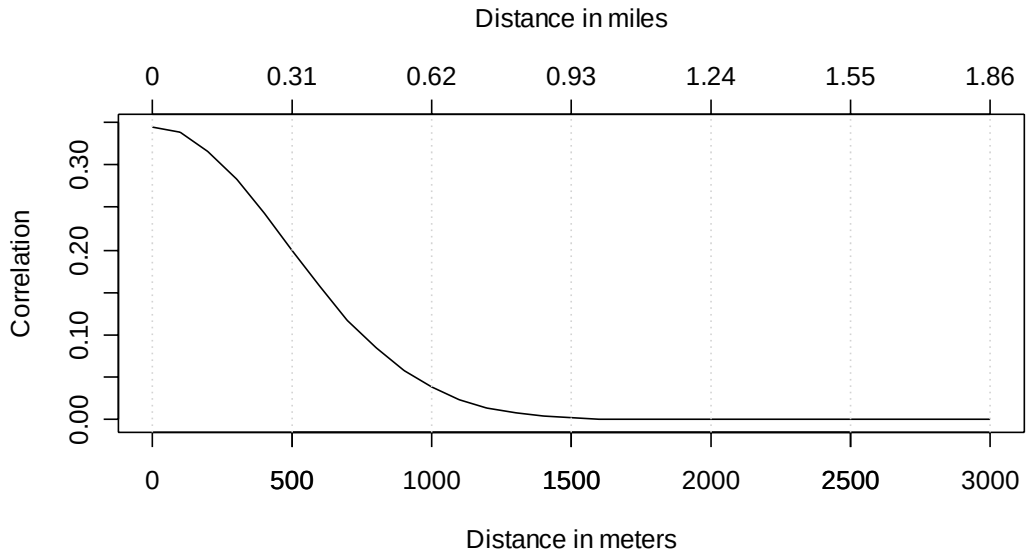


Figure 7. Fitted correlation function for the state, based on actual distance along most likely traveled paths for the primaries-and-secondaries network.

As an aside, we also used time-to-traverse most likely traveled paths in subset 3 to fit the Gaussian variogram. Variance was similar to those based on the distance metric:

$$\hat{\theta}_0 = 0.075, \qquad \hat{\theta}_s = 0.062, \qquad \text{variance} = \hat{\theta}_0 + \hat{\theta}_s = 0.137.$$

The estimate of effective range was $\sqrt{3}\hat{\theta}_r = \sqrt{3} \times 37.85 = 65.56$ seconds. Converting this effective range to miles is tricky. For stations on an interstate where the speed limit is 65 miles per hour, the effective range is $65.56 \times 65 / 3600 = 1.18$ miles. For stations on an SR route where the speed limit is only 35 miles per hour, the effective range is $65.56 \times 35 / 3600 = 0.64$ miles. The correlation function based on time-to-traverse is shown in Figure 8. Using time-to-traverse results in higher correlations than using the actual distance along most likely traveled paths.

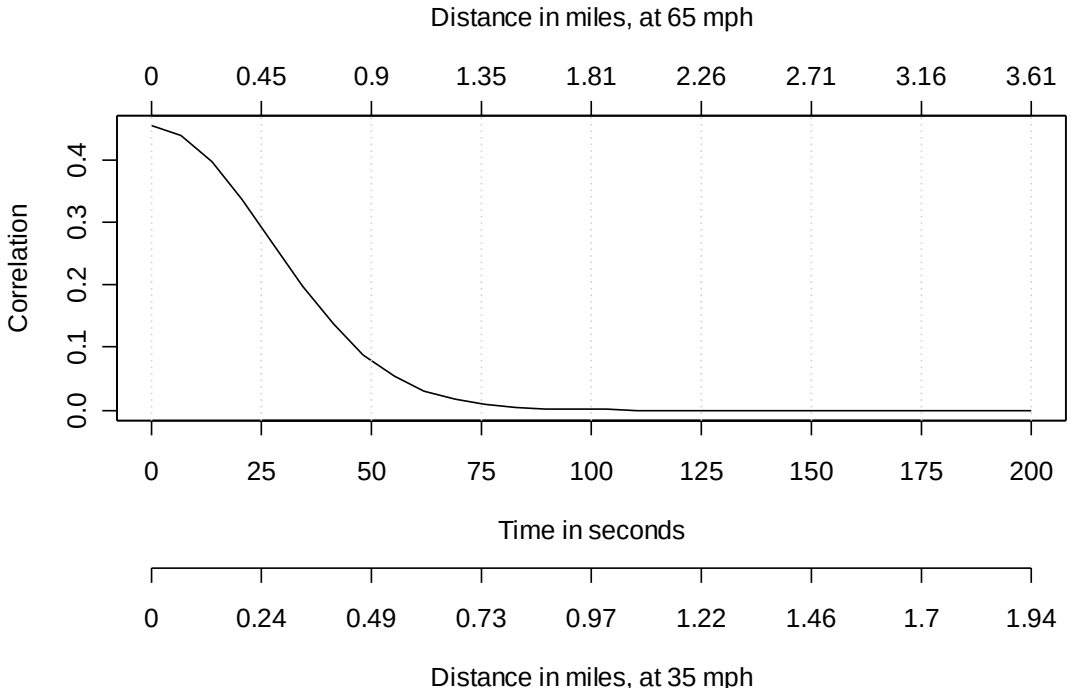


Figure 8. Fitted correlation function for the state, based on time-to-traverse along most likely traveled paths for the primaries-and-secondaries network.

Primaries-Only Network. In response to a request from Kent Taylor, we also considered the collection of stations that existed only on primary road segments. Primary road segments are either interstate, US, or NC routes. Most states other than North Carolina only monitor AADT on primary routes, so restricting investigation to a primaries-only network was not considered to be a major limitation.

We expected there would be a longer effective spatial range for stations in the primaries-only network, but wanted confirmation from the data. Only the time-to-traverse variogram was determined. The estimated Gaussian variogram model had

$$\hat{\theta}_0 = 0.090, \qquad \hat{\theta}_s = 0.024, \qquad \text{variance} = .114, \qquad \hat{\theta}_r = 370.9,$$

and effective ranges of 642.42 seconds, 11.6 miles at 65 miles per hour, and 6.2 miles at 35 miles per hour. The resulting correlation function, plotted in Figure 9, demonstrates that the maximum correlation among stations in the primaries-only network is significantly lower than among stations in the primaries-and-secondaries network, but the range is significantly higher for the primaries-only network.

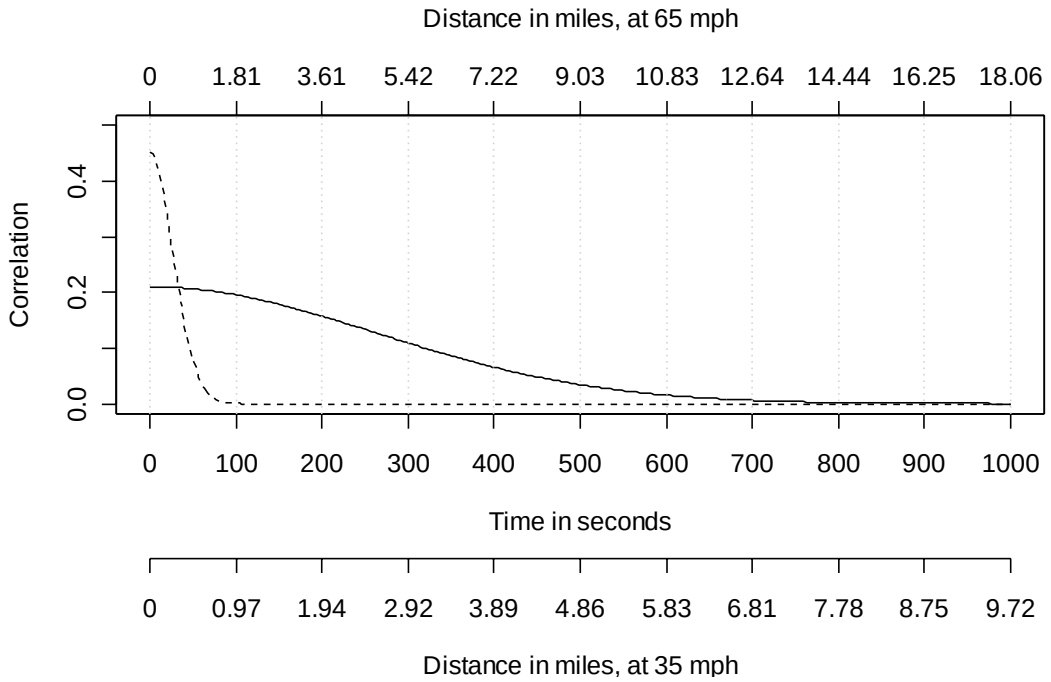


Figure 9. Fitted correlation function for the state, based on time-to-traverse along most likely traveled paths. The solid curve is from the primaries-only network and the dashed curve is from the primaries-and-secondaries network.

3.2.5 Predictions and Prediction Intervals

The major goal of this research project was to implement a strategy for identifying PTC stations that have anomalous counts. The proposed approach was to develop a spatial model that would separate out systematic and spatial variability in counts so that residual uncertainty may be identified. Having determined the "typical level of uncertainty" for a PTC station, if counts are observed that are in excess of this typical level of uncertainty, the PTC station will be flagged for further review by TSU staff. Jacqueline Hughes-Oliver proposed the use of two levels of prediction intervals to flag stations, 95% prediction intervals (95%PI) and 99% prediction intervals (99%PI). This results in a hierarchy of station classification:

- If observed AADT falls in the 95%PI, then FLAG=0. *No alert.*
- If observed AADT falls in the 99%PI but outside the 95%PI, then FLAG=1. *Level 1 alert.*
- If observed AADT falls outside the 99%PI, then FLAG=2. *Level 2 alert.*

This subsection details the calculation of prediction intervals. As a precursor to obtaining prediction intervals, this subsection also details the calculation of predictions.

Predictions and prediction intervals were required for three categories of PTC stations. These categories are:

1. PTC stations whose data was used to develop the models as described in this report and whose station information will be unchanged at the time that a prediction interval is needed.
2. PTC stations whose data was used to develop the models as described in this report but whose station information has been changed since this research project started. For example, the road segment might now have more lanes or a different speed limit.
3. PTC stations that were not included in this research project. For example, stations that were digitized after the start of the project.

For purposes of obtaining predictions and prediction intervals, categories 2 and 3 will both be classified as "new" stations—their predictions and prediction intervals will be determined in the same manner. Category 1 will be referred to as "old" stations. Predictions and prediction intervals for old stations were obtained and delivered by Tae-Young Heo in February 2004 for the network of all 34,944 PTC stations and in March 2004 for the network of stations on primary road segments. Procedures for obtaining predictions and prediction intervals for new stations were provided to TSU staff and Shannon McDonald on November 19, 2004.

Old Stations. Universal kriging is the most popular approach for obtaining predictions and prediction uncertainties from spatial models where the mean structure has to be estimated. Under an assumption of normality, universal kriging yields the best linear unbiased predictors under squared error loss. Unfortunately, because universal kriging is a perfect interpolator, using it on the old stations would result in predictions exactly equal to the observed AADT values used to build the model and prediction uncertainties would all equal zero. This would be entirely uninformative, so an alternative approach was necessary.

The formulas we used to obtain predictions yield best linear unbiased predictors under squared error loss *only if* the covariance structure is *known without error*. Because we had to estimate the covariance structure, there is clearly a disconnect with the assumptions, but this was the best possible approach for meeting project needs. The consequence is that we may underestimate the prediction uncertainty, which might lead to more flags than necessary. This conservative approach was deemed acceptable.

Let Y be the vector of observed $(\text{AADT})^{0.15}$ for all old stations and X be the matrix of attributes such that each column represents a variable in the mean model and each row corresponds to a station. The spatial model is

Mean model: $Y = X\beta + \varepsilon$, $\quad E(Y) = X\beta$

Covariance model: $\text{var}(Y) = \Sigma$, Σ is matrix from estimated statewide Gaussian covariance

Distributional assumption: ε is distributed as normal with $E(\varepsilon) = 0, \text{var}(\varepsilon) = \Sigma$.

Predictions were obtained as

$$\hat{Y}_{pred} = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

with matrix of prediction uncertainties determined by ignoring $\text{cov}(Y, \hat{Y}_{pred})$ to get

$$\text{var}(Y - \hat{Y}_{pred}) = \Sigma + X(X^T\Sigma^{-1}X)^{-1}X^T$$

and $(1-\alpha)100\%$ prediction intervals obtained as

$$\hat{Y}_{pred} \pm z_{\alpha/2}\sqrt{diag\left[\text{var}(Y - \hat{Y}_{pred})\right]},$$

where $z_{\alpha/2} = 1.96$ for 95% prediction intervals, $z_{\alpha/2} = 2.58$ for 99% prediction intervals, $diag[A]$ represents the vector formed by extracting the diagonal elements of matrix A, and $\sqrt{diag[A]}$ is the elementwise square root. Because these prediction intervals were for $(\text{AADT})^{0.15}$, we had to transform them back to the original scale. The final prediction intervals were

lower endpoints: $\left\{\hat{Y}_{pred} - z_{\alpha/2}\sqrt{diag\left[\text{var}(Y - \hat{Y}_{pred})\right]}\right\}^{1/0.15}$

upper endpoints: $\left\{\hat{Y}_{pred} + z_{\alpha/2}\sqrt{diag\left[\text{var}(Y - \hat{Y}_{pred})\right]}\right\}^{1/0.15}$.

Although the formulas given above are analytically attractive, they are computationally inconvenient. We actually converted the spatial model above to become free of a covariance model as follows:

- Decompose Σ as $\Sigma = \left(P^T\right)^{-1} P^{-1}$ such that $\Sigma^{-1} = PP^T$
- Then $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \Sigma)$ becomes $P^T Y = P^T X\beta + P^T \varepsilon$, $\varepsilon \sim N(0, \Sigma)$ or

$$Y^* = X^* \beta + \varepsilon^*, \ \varepsilon^* \sim N(0, I).$$

Having created a "new" response vector Y^* and "new" attribute matrix X^*, the REG procedure of SAS Software was then used to output prediction intervals without us needing to write code for the explicit formulas given above. SAS Software is advantageous in that it is well designed to handle large datasets and to invert large matrices. Unfortunately, not even the power of SAS Software could simultaneously obtain prediction intervals for all 34,944 stations in the statewide network. The difficulty came from sorting and merging distances from the 7.9 gigabyte matrix of most likely traveled paths between all $34944 \times 34943 / 2 = 6.1 \times 10^8$ station pairs in the state. In addition, we encountered issues with not having enough memory resources while attempting to obtain matrix P^T for decomposition of Σ. Once again, we were limited to working with only a segment of the data. Several experimental runs determined that a reasonable limit on segment size was less than 5,000 stations.

After consultation with TSU staff and adhering to not separating stations within a particular county, we arrived at a partition based on ten regions. These regions and their station counts are listed in Table 6 and displayed in Figure 10. For each of the ten regions, *Y* was redefined to be the vector of $(\text{AADT})^{0.15}$ in the region, not in the entire state; Σ and X were similarly redefined to match the region. Computing time was approximately 24 hours for each region. Prediction intervals and flag values as obtained for each of the ten regions were delivered on February 24, 2004. The file format is shown in Figure 11. Flags were distributed as shown in Table 7. Figure 12 is a sample flag map created by Tae-Young Heo using output for Region 7, which corresponds to the test area.

Table 6. Ten-region segmentation of PTC stations within counties for obtaining predictions across the statewide primaries-and-secondaries network.

Region 1	**# stations**
Cherokee	139
Graham	76
Clay	108
Swain	116
Macon	232
Jackson	173
Haywood	251
Transylvania	158
Madison	160
Buncombe	604
Henderson	294
Polk	160
Rutherford	448
Cleveland	552
Total	*3471*
Region 2	
Yancey	140
McDowell	243
Mitchell	114
Avery	121
Burke	375
Watauga	150
Caldwell	272
Ashe	196
Alleghany	146
Wilkes	373
Alexander	224
Catawba	576
Lincoln	360
Iredell	648
Total	*3938*
Region 3	
Gaston	814
Mecklenburg	1085
Rowan	561
Cabarrus	520
Union	535
Total	*3515*
Region 4	
Forsyth	1061
Davidson	680
Guilford	1011
Randolph	572
Alamance	772
Total	*4096*
Region 5	
Surry	410
Yadkin	229
Davie	216
Stokes	290
Rockingham	511
Caswell	197
Person	252
Granville	261

Vance	224
Warren	220
Franklin	307
Total	*3117*
Region 6	
Stanly	360
Anson	337
Montgomery	288
Richmond	301
Moore	458
Lee	229
Harnett	389
Scotland	288
Hoke	183
Cumberland	868
Total	*3701*
Region 7	
Orange	381
Durham	707
Chatham	320
Wake	1531
Johnston	489

Total	*3428*
Region 8	
Nash	521
Edgecombe	283
Wilson	362
Wayne	543
Greene	232
Lenoir	358
Total	*2299*
Region 9	
Robeson	674
Bladen	268
Sampson	508
Duplin	470
Craven	244
Jones	109
Carteret	133
Onslow	345
Pender	278
New Hanover	327
Brunswick	247
Columbus	442

Total	*4045*
Region 10	
Halifax	330
North Hampton	212
Hartford	208
Gates	158
Currituck	88
Camden	96
Pasquotank	165
Perquimans	121
Chowan	112
Bertie	210
Martin	247
Pitt	705
Washington	110
Tyrrell	60
Dare	95
Hyde	91
Beaufort	295
Pamlico	81
Total	*3384*
Total for 10 Regions	**34994**

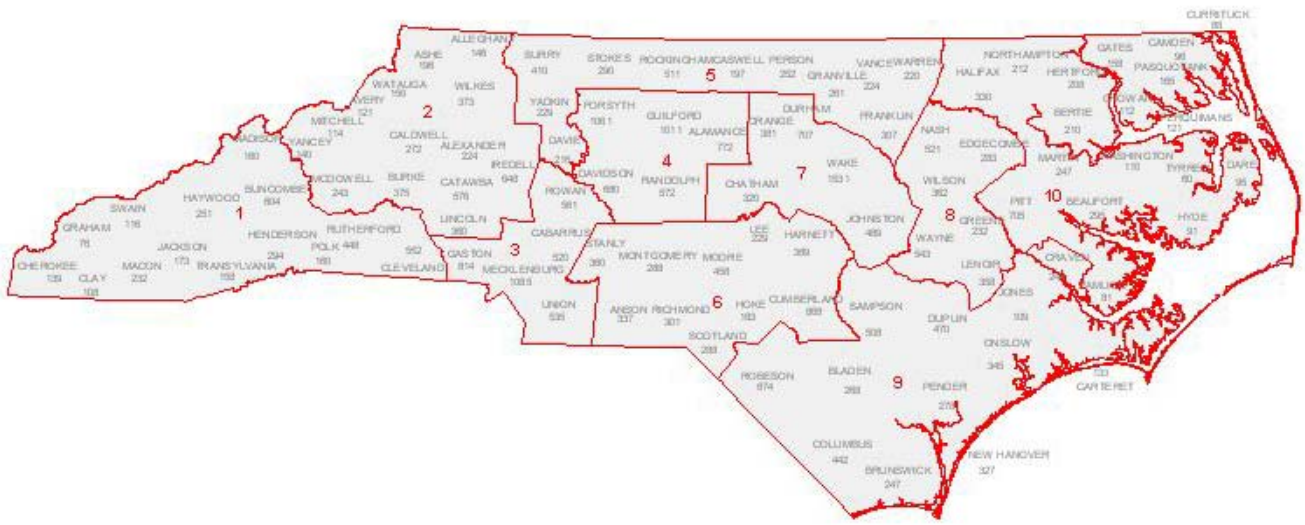


Figure 10. Ten-region segmentation for determining prediction intervals of the primaries-and-secondaries network.

LPRED95	*UPRED95*	*LPRED99*	*UPRED99*	*UNIQ_ID*	*aadt_actual*	*flag*
827.7471	15948.02	448.5096	22983.48	1	4801	0

840.2736	16080.73	456.1207	23159.83	2	5402	0
1153.088	21415.95	630.9604	30755.31	3	5803	0
949.6282	18039.15	516.5075	25961.96	4	6321	0
1583.019	25674.57	897.6915	36388.79	5	7684	0
696.7465	14276.32	371.3003	20695	6	1421	0
696.2873	14284.39	370.9322	20709.09	7	4449	0
1014.407	18293.94	559.4373	26197.46	8	5122	0
785.7608	15424.81	423.6226	22269.13	9	3245	0
842.8611	16111.81	457.6657	23202.05	10	2815	0
696.8376	14274.72	371.3734	20692.21	11	1791	0
693.4745	14334.02	368.6789	20795.81	12	2025	0
696.0162	14289.16	370.7148	20717.43	13	3156	0
603.9445	13067.23	317.0934	19039.61	14	2401	0
129.5896	5480.658	55.90043	8474.974	1505	80	1
84.47604	4003.241	35.15181	6249.271	1577	69	1
515.8869	11991.4	265.5439	17589.12	1100727	7199	0
2368.36	32915.57	1397.249	45940.18	1100728	34535	1
1061.002	19016.13	586.096	27215.37	1100729	476	2
7644.191	73343.03	4936.544	98497.47	1100730	14313	0

Figure 11. Sample records from the file containing reported prediction intervals. [LPRED95,UPRED95] is the 95% prediction interval, and [LPRED99,UPRED99] is the 99% prediction interval.

Table 7. Flag distribution for 10-region segmentation of the primaries-and-secondaries network.

Region	No Alert	Level 1 Alert	Level 2 Alert	Total
1	3322	119	30	3471
2	3740	154	44	3938
3	3336	116	63	3515
4	3917	110	69	4096
5	2965	112	40	3117
6	3534	130	37	3701
7	3270	111	47	3428
8	2184	91	24	2299
9	3855	141	49	4045
10	3227	120	37	3384
Total	*33350*	*1204*	*440*	*34994*

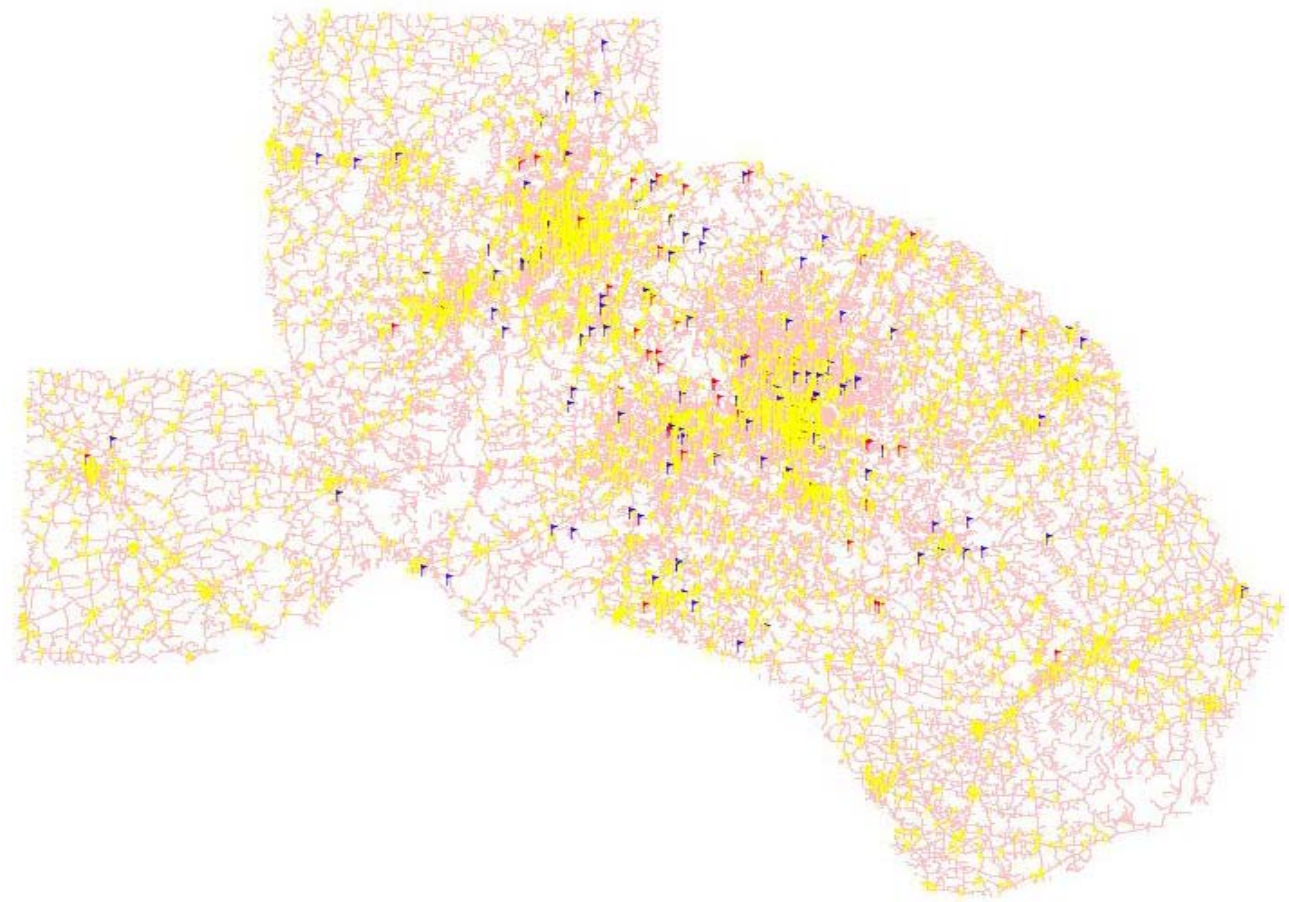


Figure 12. Prediction interval flag map for Region 7, the test area. Flag colors indicate the following: red—level 2 alert, blue—level 1 alert, yellow—no alert.

The segmentation approach, while computationally feasible, can create misleading results due to edge effects. More specifically, after segmentation stations near the border of a region do not benefit from information available in the adjoining region and this could impact either predictions or prediction uncertainties. To investigate the degree of impact, we focused on two counties, Johnston and Wake, placed them in a newly defined segment, and re-obtained their prediction intervals and flags. Of the 489 PTC stations in Johnston County, only two of them (0.4%) had changes in their flag values. Of the 1,531 PTC stations in Wake County, only 38 of them (2.5%) had changes in their flag values. Table 8 identifies those stations for which there were changes. The small percentages of affected stations and the fact that there were never any changes from a "no alert" flag to a "level 2 alert" flag suggested that edge effects may be very small.

Table 8. PTC stations in Johnston (2 of 489) and Wake (38 of 1,531) Counties whose flag vlaues changed as a consequence of being placed in a different region.

County	UNIQ_ID	Flag from Region 7	Flag from other region	Difference
Johnston	500144	1	0	1
Johnston	500543	1	0	1
Wake	910616	2	1	1
Wake	1080316	2	1	1
Wake	1080588	2	1	1

Wake	910648	1	0	1
Wake	910665	1	0	1
Wake	911512	1	0	1
Wake	911886	1	0	1
Wake	1080070	1	0	1
Wake	1080083	1	0	1
Wake	1080248	1	0	1
Wake	1080674	1	0	1
Wake	1080698	1	0	1
Wake	1080862	1	0	1
Wake	1080864	1	0	1
Wake	1080866	1	0	1
Wake	910666	1	2	-1
Wake	911537	1	2	-1
Wake	1080616	1	2	-1
Wake	1080715	1	2	-1
Wake	1080820	1	2	-1
Wake	1082218	1	2	-1
Wake	1082232	1	2	-1
Wake	910696	0	1	-1
Wake	910739	0	1	-1
Wake	912016	0	1	-1
Wake	1080267	0	1	-1
Wake	1080303	0	1	-1
Wake	1080329	0	1	-1
Wake	1080537	0	1	-1
Wake	1080592	0	1	-1
Wake	1080626	0	1	-1
Wake	1080751	0	1	-1
Wake	1080753	0	1	-1
Wake	1080844	0	1	-1
Wake	1080847	0	1	-1
Wake	1080870	0	1	-1
Wake	1080922	0	1	-1
Wake	1082223	0	1	-1

We also obtained predictions from larger regions. These three regions, listed in Table 9 and displayed in Figure 13, considered only stations on primary road segments. Primary road segments are either interstate, US, or NC routes. Most states other than North Carolina only monitor AADT on primary routes, so restricting investigation to a primaries-only network was not considered to be a major limitation. Prediction intervals from these three regions were delivered on March 14, 2004. Flags were distributed as shown in Table 10.

Table 9. Three-region segmentation of PTC stations within counties for obtaining predictions across the statewide primaries-only network.

Mountain (28 counties)

County	# Primary Stations
Alexander	30
Alleghany	46
Ashe	60
Avery	56
Buncombe	221
Burke	124
Caldwell	74
Catawba	115

Cherokee	44
Clay	28
Cleveland	143
Graham	25
Haywood	83
Henderson	80
Iredell	131
Jackson	67
Lincoln	102
Macon	57
Madison	60
Mcdowell	71
Mitchell	39
Polk	40
Rutherford	107
Swain	37
Transylvania	55
Watauga	58
Wilkes	79
Yancey	55
	2087

Piedmont (42 counties)

County	# Primary Stations
Alamance	213
Anson	70
Cabarrus	143
Caswell	59
Chatham	72
Cumberland	213
Davidson	141
Davie	73
Durham	206
Edgecombe	100
Forsyth	261
Franklin	91
Gaston	248
Granville	66
Greene	66
Guilford	275
Harnett	107
Hoke	30
Johnston	154
Lee	55
Lenoir	97
Mecklenburg	328
Montgomery	75
Moore	117
Nash	141
Orange	87
Person	52
Randolph	126
Richmond	64
Rockingham	187
Rowan	122
Scotland	70
Stanly	92
Stokes	80
Surry	110
Union	114
Vance	57
Wake	270
Warren	55
Wayne	145
Wilson	106
Yadkin	45
	5183

Coastal (30 counties)

County	# Primary Stations
Beaufort	68
Bertie	64
Bladen	112
Brunswick	80
Camden	25
Carteret	41
Chowan	28
Columbus	135
Craven	65
Currituck	25
Dare	28
Duplin	117
Gates	48
Halifax	137
Hertford	64
Hyde	25
Jones	32
Martin	76
New Hanover	129
Northampton	83
Onslow	88
Pamlico	20
Pasquotank	41
Pender	93
Perquimans	20
Pitt	206
Robeson	164
Sampson	113
Tyrrell	18
Washington	36
	2181

Total for 3 regions	**9451**

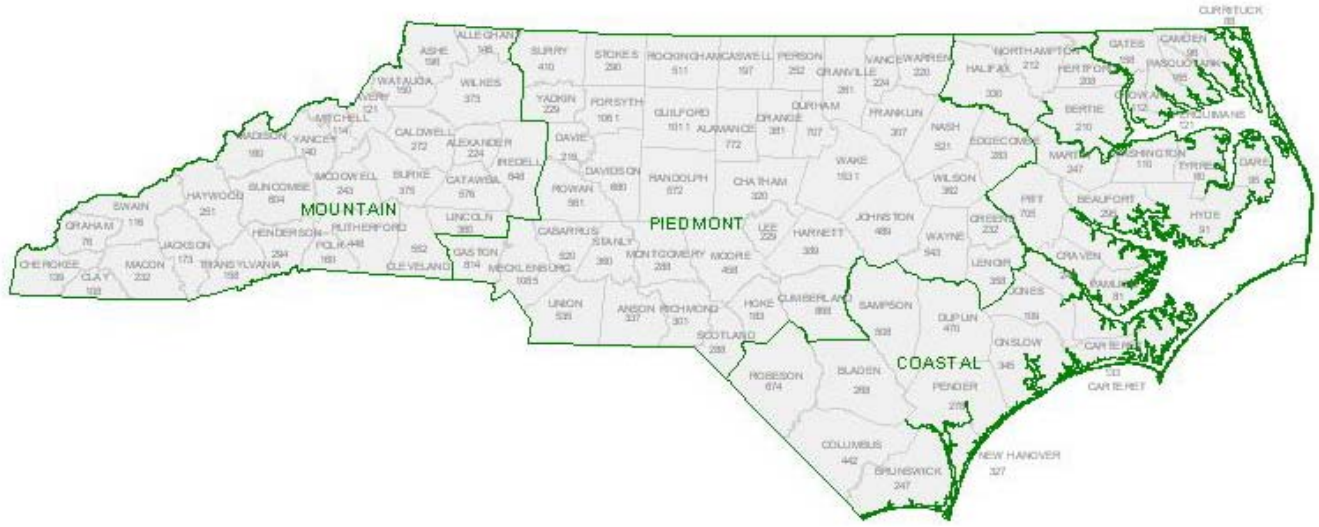


Figure 13. Three-region segmentation for determining prediction intervals for primaries-only network.

Table 10. Flag distribution for 3-region segmentation of the primaries-only network.

Region	No Alert	Level 1 Alert	Level 2 Alert	Total
Mountain	2024	45	18	2087
Piedmont	5057	92	34	5183
Coastal	2117	52	12	2181
Total	*9198*	*189*	*64*	*9451*

New Stations. The formulas used to obtain prediction uncertainties for old stations required availability of the covariance model through the matrix Σ. This matrix Σ was obtained as a function of the matrix of distances along most likely traveled paths. These distances would need to be determined from first principles for new stations that were not included in the data used for this research project. Stations whose information changed since the beginning of this project would also need to have their distances determined anew because identification of most likely traveled path depends on station attributes such as speed limit and number of lanes.

As reported earlier in this document, obtaining most likely traveled paths was an extremely difficult process that the research team was not willing to accept as a just-in-time feature of the final software deliverable. In a November 19, 2004 meeting of the research team, Jacqueline Hughes-Oliver suggested an alternative that does not require the covariance model. Essentially, it follows the same procedure as for old stations, but the Σ matrix is replaced by the diagonal matrix $\hat{\sigma}^2 I$.

A formulaic summary is given here, with additional details provided in Appendix 10, "Preparations for November 19, 2004 Meeting with DOT Research Team." Let X denote the matrix of attributes for the old stations (i.e., those stations that were used to build the mean model), X_0 denote the matrix of attributes for the new stations, Y denote the vector of $(\text{AADT})^{0.15}$ for the old stations, and Y_0 denote the vector of observed or unobserved $(\text{AADT})^{0.15}$ for which we want to get predictions $\hat{Y}_{0,pred}$ and associated prediction uncertainties. Then,

$$\hat{Y}_{0,pred} = X_0 (X^T X)^{-1} X^T Y \equiv X_0 \hat{\beta}$$

with matrix of prediction uncertainties determined by ignoring $\text{cov}(Y_0, \hat{Y}_{0,pred})$ to get

$$\text{var}(Y_0 - \hat{Y}_{0,pred}) = \hat{\sigma}^2 I + X_0 (X^T X)^{-1} X_0^T$$

and $(1-\alpha)100\%$ prediction interval for AADT as

lower endpoints: $\left\{\hat{Y}_{0,pred} - z_{\alpha/2}\sqrt{diag\left[\text{var}(Y_0 - \hat{Y}_{0,pred})\right]}\right\}^{1/0.15}$

upper endpoints: $\left\{\hat{Y}_{0,pred} + z_{\alpha/2}\sqrt{diag\left[\text{var}(Y_0 - \hat{Y}_{0,pred})\right]}\right\}^{1/0.15}$.

For a 95% prediction interval, uses $z_{\alpha/2} = 1.96$, and for a 99% prediction interval, use $z_{\alpha/2} = 2.58$. The vector $\hat{\beta}$ is simply the coefficient estimates from the final mean model for the statewide area and $\hat{\sigma}^2$ is 0.14489, the mean squared error (MSE) from that same mean model. Since there is no need to recompute $(X^T X)^{-1}$ for each set of new stations, Tae-Young Heo presented this 179×179 matrix to TSU and Shannon McDonald on November 23, 2004.

3.3 Automation and GIS Implementation, Task 3

Much effort was made during this project to further the ongoing effort by the NCDOT to standardize data within a relational database environment, coupled with GIS capabilities. The implementation of GIS proved essential in preparing, organizing and analyzing the information. The procedures used in preparing and organizing the model input data, along with the calculations of prediction intervals (using the mean model algorithm) have been automated within ArcGIS using a series of scripts written in Visual Basic for Applications (VBA). Details outlining these procedures are given in Appendix 12, "Automation With GIS Application Procedures."

The predictions and prediction intervals are stored in a comma-delimited text file. This file obtains the stations' unique identifiers and thus can be joined to the PTC station shapefile(s) for viewing in ArcGIS. Thus, with conventional ArcGIS tools, analysts can easily spot anomalies through logical selection and location. In turn, these stations can be compared with other stations within an area or along a route to ensure integrity.

The following screen shots illustrate a thematically mapped display that quickly alerts the analyst to station data anomalies.

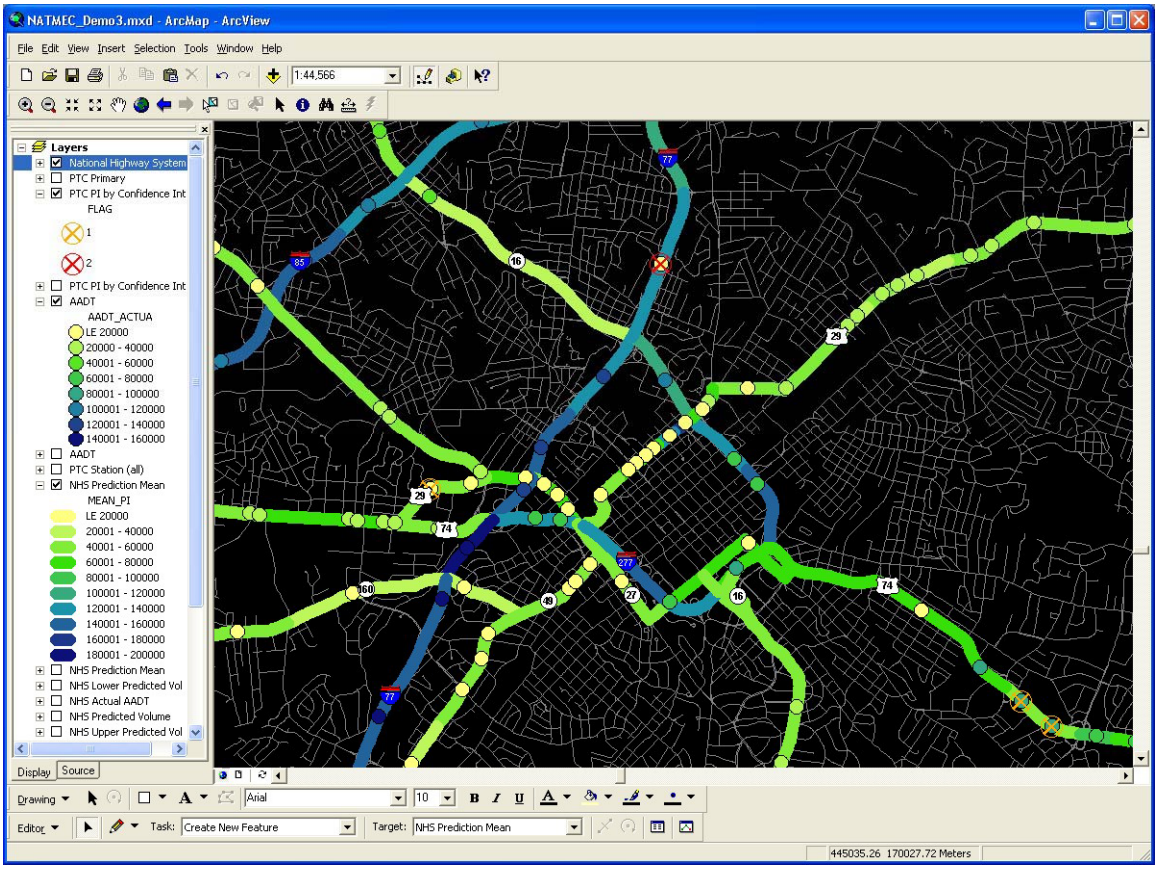


Figure 14: Thematic Map of Traffic Volume AADT vs. Predicted

In the above screenshot, the circles represent the PTC stations and are colored according to the AADT (actual) volumes recorded. In contrast, the road network is similarly colored by the prediction mean. The stations flagged (circles with 'X') indicate that they are outside the 95% (yellow) or 99% (red) prediction intervals. Observing the color difference between the AADT and mean prediction, the analyst can quickly determine that the station in red has an AADT volume less than the lower 99% prediction value. Those stations in yellow (bottom right) appear to have AADT volumes within the upper 95% and 99% prediction.

The next screenshot shows the outlier station AADT volume and that of its prediction intervals. In this map, the road layer has been mapped to show the prediction intervals at-a-glance.

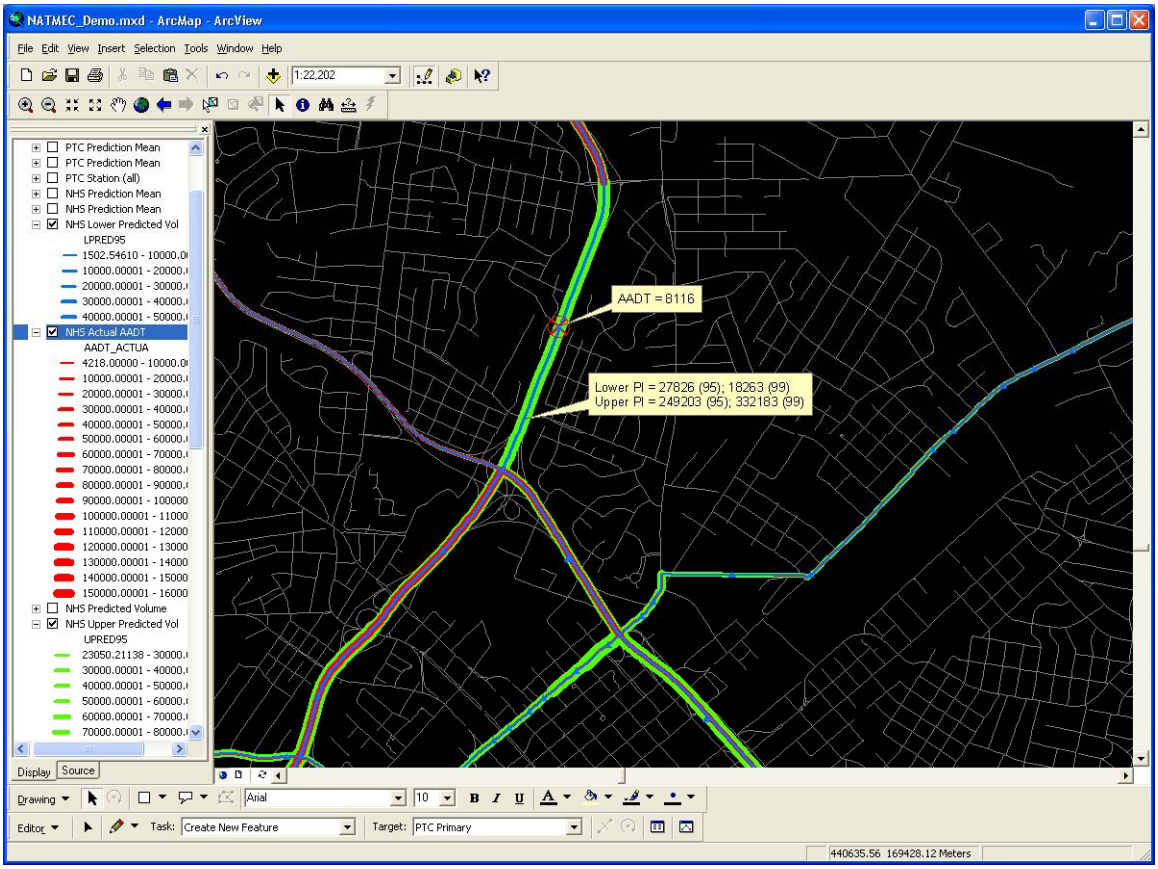


Figure 15: Outlier Station

Note the AADT of this station is 8116, while the lower 99% prediction interval value is 18,063. The road layer is drawn with relative thickness' that portray a viewable range of values. The blue and green represents the 95% prediction interval. The read represents the AADT value. If all three colors are visible, i.e. the red is between the blue and the green, then the station volume count is within the prediction interval. Notice the AADT representation (red) is not visible along the road segment(s) represented by the outlier station.

The following screenshot illustrates two stations that are outside the 95% prediction intervals.

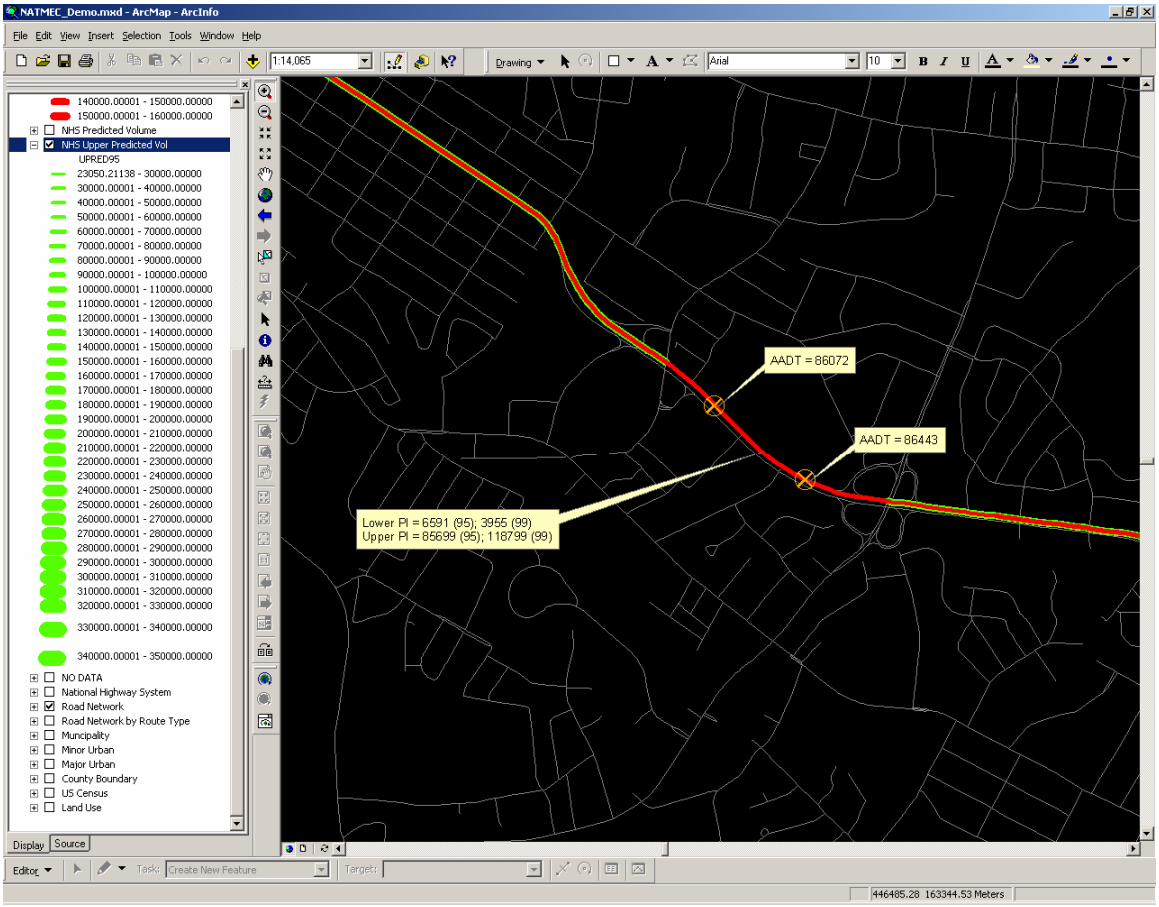


Figure 16: PTC Stations outside the 95% Prediction Intervals

In the above figure, the stations are located along a segment that appears red. This indicates that the AADT volume for each station is greater than, thus thicker than the upper 95% prediction interval represented by green. Note that the AADT values fall within the upper 95% and 99% prediction intervals for each station.

The seamless aspect of the GIS layers allows the analysts to view data across county lines as well. This aspect permits the analyst more flexibility when viewing regional anomalies.

Traffic continuity maps depicting traffic volumes along the primary routes is another example of the type of analysis typical of GIS. Assuming each station represents a particular segment of roadway, this data can be linked to the road layer. The road layer can thus be thematically mapped to show an intuitive display representing the predicted and/or actual travel of North Carolina's primary highways. The procedures for this exercise are outlined in Appendix 11, "Primary Road Statewide Traffic Continuity Map".

3.4 Implementation Guidelines, Task 4

All deliverable data derived from the procedures outlined in this report are available as stand-alone comma-delimited ASCII files and/or dBase (.dbf) files. The prediction interval data contains the station unique identifier and can be used in combination with any data that so contains the identifier.

ESRI ArcGIS versions 8.1, 8.2 and 8.3 were used in the creation of the application. It is not known at the time of this report whether newer versions of ArcGIS are compatable. Any ArcGIS module (ArcView, ArcEditor or ArcINFO) will be sufficient in creating the prediction intervals for new stations using the mean model. Modeling station data using the covariance model is not available at the time of this report due to the large amount of computations and limitations of the hardware systems.

Note that some procedures outlines in this report were created using AML. Recreation of such tasks will require the use of ArcINFO Workstation.

 The project file (.mxd) contains a button that launches the application for creating prediction intervals for existing stations within the PTC shapefile.

When executed, a new data frame is added to the table of contents (TOC). At the termination of execution, the themes will be removed from the TOC. At this point, the program has executed and the prediction intervals are available in comma-delimited format within the workspace.

The workspace is coded within the application. At the time of this report, the default workspace is 'C:\NCSU\GIS'. The subfolders within the workspace must be identical to that declared as variables within the application. To view or edit the pathnames, open the Visual Basic Editor in ArcGIS (Tools → Macros → Visual Basic Editor, or Alt+F11). All pathnames are declared as constants in the *Declarations* portion of the script.

The comma-delimited prediction interval output file may be added to the TOC within this or another project file without conversion. The data can be joined to the PTC shapefile or any other shapefile or data set that contains the PTC station unique identifier.

It is probable that the system may require a 64-bit architecture in order to run properly because of constraints with floating-point ranges. It is recommended that a dual-core 64-bit processor with adequate RAM (1 Gb or greater) be used to execute the algorithm. Viewing, editing, or otherwise using the data is possible with the minimum requirements needed to run ArcGIS.

Sample layers have been created for thematic mapping. These can be added to the TOC for intuitive analysis. It is predicted that these layers and the analytical interface will evolve to best suit the needs of the analysts. The required data is dynamic to allow this migration.

The application is written explicitly for the procedures and data as outlined in this report. As new data becomes available, it is likely the data will differ from that used in this exercise. In addition, new matrices will need to be calculated. These and other changes in the input data will invariably affect the code. Maintaining the code will be necessary whenever these changes occur.

4. FINDINGS AND CONCLUSIONS

The project led to several major findings, some of which suggest changes may be needed in areas other than the stated targeted areas of the project, which are spatially aligned edits of PTC data and traffic continuity maps. Major findings are summarized below, categorized by research tasks.

Task 1: Further Review of Methods, Data and Computing Options
As of late 2001, most research papers with stated goals of improving estimation of AADT from short count data are actually focused on determining seasonal adjustment factors. Unlike the current project, they do not seek to develop traffic continuity maps or actually edit short count data.

Current computing options within the NCDOT do not allow distribution of mapping software for interactive visualization. This caused a change to our primary product.

Data acquisition and editing was a much more significant ordeal than originally anticipated. There were three extensive components that each required major effort from within the TSU, within the research team, and between these two groups. Two of these components concerned reconciliation of the data with other sources and the third was creation of a new data type that did not previously exist within the NCDOT.

The first data reconciliation effort arose because of the many conflicts identified within and between the PTC count data (primarily maintained by the TSU) and the auxiliary GIS station data (primarily maintained by the GIS Unit). For example, several key attributes such as route type, number of lanes, and speed limit were missing or incorrect for some PTC stations, while other PTC stations had multiple conflicting records. Some of these issues were due to PTC stations being "snapped" to the wrong road segment in the GIS. While this snapping limitation was known to exist prior to the start of the project, the thorough data checking initiated by the project shed light on the severity, causes, and resolution of some problems.

The second data reconciliation effort occurred after a preliminary spatial characterization model was obtained during Task 2. Using a preliminary model that predicted AADT counts as a function of station descriptors, GIS data, land use data, and census data, without acknowledging spatial correlations, 55 of the 3434 stations within the test area were identified as "outliers" because their AADT counts were not well predicted by the preliminary model. Although these 55 stations represented only 1.6% of the stations studied at that time and this small fraction indicated that the model was already very successful in capturing AADT dependence on station attributes, we still wanted to know whether the 55 stations were special in ways that could be addressed in order to achieve an even better model. Further study of the 55 stations did indeed reveal additional ways that the model could be improved, for example, by creating additional attributes to indicate whether a PTC station is located within a municipality but outside the urban boundary. On the other hand, study of the 55 stations also revealed several long-reaching inconsistencies with how station data is obtained and recorded. The potential impact of these findings is far-reaching and will likely require much effort from NCDOT for complete resolution.

Spatial characterization of AADT from short count data is contingent on knowing distances between PTC stations along the road network. As such, the project required identification of shortest paths and distances, along the road network, between all pairs of stations in the state. Such data was not previously available in NCDOT and had to be obtained from first principles. Starting from the definition of shortest path, continuing to delivery of the distances, this sub-project resulted in several important findings. Computational improvements due to using AML code for determining distances were significant, but a dedicated dual-processor computer was needed to expedite computation to an acceptable level.

Transmittal of these distances then required moving the dedicated computer to North Carolina State University because the file size was almost 8GB with other 200 million records and was not easily accessible otherwise. This finding impacted later tasks in that Task 2 was limited to only being able to access portions of the distance matrix at a time, so modifications from earlier proposed approaches were necessary. In other words, the biggest finding was that computational adjustments would be needed in Task 2 to accommodate the massive data created by shortest path distances.

Task 2: Spatial Modeling and Characterization, Using Current Seasonality Adjustments
Some Task 2 findings have already been discussed above because of their impact on Task 1 activities. Additional findings will be presented below in three separate categories, one regarding the mean model, another regarding the covariance model, and the third regarding prediction intervals.

Data transformation was required to improve the fit of the mean model and to comply with modeling assumptions. As such, the statewide model was actually developed using $(AADT)^{0.15}$, but predictions were ultimately converted back to the original scale for the GIS implementation. Of the 550 variables available for building the model, only 178 were ultimately kept in the final model. The resulting coefficient of determination (R-squared) was an impressive 0.73, with correlation of 0.85 between observed and predicted $(AADT)^{0.15}$. The model captured impact on AADT of land use, census and demographic information for areas surrounding PTC stations, as well as station-specific attributes and day/time of data collection. The finding that day/time of data collection is important in the mean model, even after accounting for all the other variables that were included in building the model, suggests that enhancements are still needed for seasonality adjustments that convert PTC counts to an annualized count.

Residuals from the mean model were used to develop covariance models. Findings confirmed the need for distances along the road network, even though they were very difficult to obtain. In other words, covariance models built on Euclidean distances (shortest distances irrespective of network accessibility) were meaningless. The question of what to report as "distances" along the shortest path on the road network was also important and non-trivial. Our choices between actual distances on these paths versus time to travel these paths both had potential advantages. In the end, models fit to the data suggested that actual distances provide more reasonable interpretations. The variogram model selected was the Gaussian variogram. Not surprisingly, the effective range (or distance beyond which AADTs for pairs of PTC stations can reasonably be treated as independent) of six miles for interstate stations far exceeds the effective ranges for other station pairs, with the other route types having effective ranges between one and two miles. On a statewide level, the maximum correlation among stations in the primaries-only network (i.e., interstate, US or NC routes) is significantly lower than among stations in the primaries-and-secondaries network (i.e., all stations), but the effective range is significantly higher (almost by a factor of ten) for the primaries-only network. While a modest degree of covariance nonstationarity was indicated, the severity did not necessitate remedial actions, so we ultimately used covariance models developed simultaneously from all route types to obtain prediction intervals.

Predictions, 95% prediction intervals, and 99% prediction intervals were obtained for three categories of PTC stations: "old" stations, "new" but previously digitized stations, and "new" but newly digitized stations. Because of the computational challenges caused by the distance matrix, prediction intervals had to be determined in segments. Ten segments were needed for the primaries-and-secondaries network, while only three segments were needed for the primaries-only network. We studied the impacts of edge effects caused by segmentation and determined them to be minimal. In the primaries-and-secondaries network, 3.4% of the stations received a level 1 alert, which indicated that their observed AADT count was outside the 95% prediction interval but inside the 99% prediction interval, while 1.3% of the stations

received a level 2 alert, meaning that their observed AADT count was outside the 99% prediction interval. For the primaries-only network, there were 2.0% level 1 alerts and 0.7% level 2 alerts.

Task 3: Automation and GIS Implementation
This portion of the project focused more heavily on programming as opposed to model development. While several important programmatic findings led to improved code, there are no other findings comparable to those listed for Tasks 1 and 2.

Task 4: Implementation Guidelines
This portion of the project focused more heavily on programming as opposed to model development. While several important programmatic findings led to improved code, there are no other findings comparable to those listed for Tasks 1 and 2.

Conclusions
Succinctly put, the project significantly improves the process of editing and validating traffic count data. Specific improvements include:

- Reported counts are better in several ways:
  - determination of whether a count needs to be investigated is based on statistics and levels of uncertainty, not purely on the opinion of the data analyst,
  - adjustments, if necessary, are based on statistical predictions, not purely on the opinion of the data analyst,
  - over-control of the process through an excessive number of manual adjustments will no longer be a concern, and
  - recounts, if needed, will take place in a more timely fashion because the entire editing and validation process will be faster.
- The process has increased functionality:
  - the use of color to indicate unusual counts will enhance interpretability and presentation, and
  - expert knowledge will not be required for using the software.
- The process has increased flexibility: maps can be created at either the station level or at regional levels.
- The process is much faster, reducing turnaround time to 1-2 months instead of the current one year.
- The process and products allow and encourage easier sharing of data with other state and local government agencies by using computing tools and environments that are widely available.
- The process follows the recommendations from FHWA and AASHTO for incorporating spatial analysis.

5. RECOMMENDATIONS

PTC data collection is performed at monitoring stations annually or biennially and a natural question is how often should NCDOT repeat the exercises documented in this report. Our recommendation is that this update should be done every five to ten years in order to maintain relevance of the models. For areas of faster-than-average growth within the state, more level 1 and level 2 alerts might result due to the fact that the models created several years ago are no longer current for the time period of interest. While these extra alerts may be a nuisance that require unnecessary resources be allocated to investigate non-issues, they are actually the better of possible outcomes from out-of-date models. The more troublesome outcome is that a station does not get an alert when its new conditions actually demand such an alert. Urban fringe areas will likely be those regions that change most rapidly and hence be the regions where old models break down. Because transportation infrastructure is critical to continued growth, it will be very important to maintain up-to-date models in order to properly accommodate growth. The recommended frequency of five to ten years allows flexibility. Recall that this project generated two basic models, a mean model and a covariance model. The mean model was significantly less taxing to create, so it can be updated on a more frequent basis. The covariance model, on the other hand, was very difficult to create so it may need to be updated less frequently. This report contains sufficient details that allow re-creation of both the mean and covariance models.

On a more speculative note, we recommend several checks on model assumptions. There was indication during this project that covariance nonstationarity exists, but it was deemed minor enough to ignore. This will need to be checked for each determination of the covariance model. If covariance nonstationarity is significant, it may require that entirely separate models be built for each route type. We have already prototyped these models in this report—the primaries-only network is based on only interstate, US, and NC routes. Another technicality that requires attention is the impact of segmentation for creating the covariance model; the fewer the number of segments the better, but computational limitations must also be acknowledged. Other aspects that may need modification are the transformations needed for building the mean model. Here we applied a power of 0.15 to AADT for building the statewide model and we used a particular approach (principal components analysis) to create census and demographic summaries. Unfortunately, there is no guarantee that these same transformations will be appropriate for subsequent modeling.

In the future, after NCDOT acquires the ability to provide distributable mapping software for interactive visualization, the final product of this project can and should be adjusted to the original intended format. More specifically, we recommend enabling the GIS-formatted data (our current primary product) with the visualization software then building a user's manual for the software. Once this is done, additional guidance will then need to be provided on technical details for efficient transfer and distribution of the software and on creation of customized reports.

Finally, during the process of curating the data for this project, many findings were made regarding data acquisition, reconciliation of data stored in multiple locations, and data integrity. Some of these findings have already impacted long-term processing, but we recommend that other findings be reconsidered for future resolution.

6. IMPLEMENTATION AND TECHNOLOGY TRANSFER PLAN

What is the Primary Product?
The primary product is GIS-formatted data that will incorporate the spatial statistical model developed in research objective 4 for the purpose of creating the displays of predictions and anomalies as outlined in research objectives 6 to 8. More specifically, the GIS-formatted data programmatically identifies PTC stations that have anomalous counts and provides information for creating traffic continuity maps.

What are the Secondary Products?
Secondary products are the many written reports documenting all steps needed in the future for: (i) editing/reconciling the input data; (ii) generating distances along most likely traveled paths; (iii) updating the mean model; (iv) updating the covariance model; (v) obtaining predictions and prediction intervals; and (vi) automation and GIS implementation.

Who within NCDOT will use the Products, and why?
TSU, the direct clients for the proposed research, will use the products: to replace their current process of editing and validating traffic counts; to obtain traffic continuity maps; and to create volume count reports for their customers.

TSU will be the major users of the products, but they create so many reports for departments within NCDOT that indirect product usage will extend far and wide within NCDOT. For example, the Transportation Planning Branch uses reports created by TSU to develop models and long-range transportation plans, while the Project Development and Environmental Analysis Branch (PDEA) uses the reports for planning and developing alternatives for Transportation Improvement Program (TIP) projects.

What will it take for NCDOT customers to use the products?
With the new products, NCDOT will be able to create customized reports, using a GIS environment that encourages sharing and distribution, for both large and small customers. In order to effectively use the GIS-formatted data, users will need visualization software and general training for that software. Minimal additional training will be required to inform users of the potentially different and customizable reports that may be created using the GIS-formatted data that we provide.

7. CITED REFERENCES

Albright, D. 1991. Development of national highway traffic monitoring standards. *Transportation Research Record*, 1311, 85-87.

AASHTO. 1992. *AASHTO Guidelines for Traffic Data Programs*. Washington D.C.

Claramunt, C., Jiang, B., Bargiela, A. 2000. A new framework for the integration, analysis and visualisation of urban traffic data within geographic information systems. *Transportation Research Part C -- Emerging Technologies*, 8, 167--184.

Cunagin, W.D. and Kent, P.M. 1998. Reliability of traffic data. *Progress in Transportation Data 1998, Transportation Research Record*, 1625, 18--25.

FHWA. 2001. *Traffic Monitoring Guide*. DRAFT accessed March 2001 at http://www.fhwa.dot.gov////////ohim/tmguide/index.htm

Hu, P.S., Wright, T., and Esteve, T. 1998. Traffic count estimates for short-term traffic monitoring sites—Simulation study. *Progress in Transportation Data 1998, Transportation Research Record*, 1625, 26--34.

Mohamad, D., Sinha, K.C., Kuczek, T., and Scholer, C.F. 1998. Annual average daily traffic prediction model for county roads. *Land Use and Transportation Planning and Programming Applications, Transportation Research Record*, 1617, 69--77.

Sharma, S.C., Gulati, B.M., and Rizak, S.N. 1996. Statewide traffic volume studies and precision of AADT estimates. *Journal of Transportation Engineering—ASCE*, 122, 430--439.

Xia, Q., Zhao, F., Chen, Z., Shen, L.D., and Ospina, D. 1999. Estimation of annual average daily traffic for nonstate roads in a Florida County. *Transportation Research Record*, 1660, 32-40.

8. BIBLIOGRAPHY

Aldrin, M. (1995) A Statistical Approach to the Modeling of Daily Car Traffic, *Traffic Engineering and Control*, 36(9), 489-493.

Bolduc, D., Dagenais, M. and Gaudry, M. (1989), Spatially autocorrelated errors in origin-destination models: A new specification applied to aggregate mode choice, *Transportation Research Part B: Methodological*, 23(5), 361-372

Cheng, C., (1992) Optimum Sampling for Traffic Volume Estimation, Ph.D. Dissertation, University of Minnesota, Carlson School of Management, Minnesota

Deacon, J. A., Pigman, J. G. and Mohenzadeh, A., (1987) Traffic Volume Estimates and Growth Trends, UKTRP-87-32, Kentucky Transportation Research Program, University of Kentucky, Kentucky.

Goel, P. K., McCord, M. R., and Park, C. (2004). Exploiting Correlation to Improve AADT Estimation on Coverage Count Segments: Methodology and Numerical Results. *TRB* 2005.

Gutierrez Puebla, J. (1987), Spatial structures of network flows: A graph theoretical approach, *Transportation Research Part B: Methodological*, 21(6), 489-502

Lam, W. H. K., and Xu, J. (2000) Estimation of AADT from Short Period Counts in Hong Kong - A Comparison Between Neural Network Method and Regression Analysis, *Journal of Advanced Transportation*, 34, 249-268

Mohamad, D., Sinha, K. C., and Kuczek, T. (1998), An Annual Average Daily Traffic Prediction Model for County Roads, presented at Transportation Research Board, 77th Annual Meeting, January 11-15, 1998, Washington D.C.

Newell, G. F., (1993) Comments on spatial models of traffic, Transportation Research Part B: Methodological, Volume 27, Issue 3, Pages 185-188

Okutani, I. and Stephanedes, Y. J. (1984), Dynamic prediction of traffic volume through Kalman filtering theory, Transportation Research Part B: Methodological, 18(1),1-11

Saha, S.K. (1990) The Development of a Procedure to Forecast Traffic Volumes on Urban Segments of the State and Interstate Highway Systems, Ph.D. Dissertation, School of Civil Engineering, Purdue University.

Sharma, S.C., B.M. Gulati, and S. Rizak. (1996) Statewide Traffic Volume Studies and Precision of AADT Estimates. In Journal of Transportation Engineering, 122, 430?439.

Shen, L.D., Zhao, F., and Ospina, D. (1999) Estimation of Annual Average Daily Traffic for Off-System Roads in Florida, Research Report, Florida Department of Transportation, Tallahassee, FL

Xia, Q., Zhao, F., Chen, Z., Shen, L. D., and Ospina, D. (1999) Development of a Regression Model for Estimating AADT in a Florida County, Transportation Research Record, No.1660, Transportation Research Board, National Research Council, Washington, D.C., 32-40.

Zhao, F. and S. Chung (2001) Estimation of Annual Daily Traffic in a Florida County Using GIS and Regression, *TRB* 2001 Annual Meeting Paper, Washington.

Zhao, F., and Chung, C. (2001), Contributing Factors of Annual Average Daily Traffic in a Florida County: Exploration with Geographic Information System and Regression Models, *Transportation Research Record*, 1769, pp. 113-122.