A GIS Data Quality Case Study: 
Accuracy Comparison of Length Measurements Using GIS/NED and DMI

Technical Report

Department of Civil Engineering 
North Carolina State University
August, 2001

by
William Rasdorf
Hubo Cai
# TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

1. INTRODUCTION
   1.1 Background
   1.2 Geographic Information Systems (GISs)
   1.3 Data Quality
   1.4 Spatial Data Accuracy
   1.5 Work by Others
   1.6 Case Study
   1.7 Report Organization

2. MODELS AND TOOLS
   2.1 Elevation Models
      2.1.1 The Point Model
      2.1.2 The Contour Line Model
      2.1.3 DEM/DTM
      2.1.4 TIN (Triangulated Irregular Network)
      2.1.5 The Mathematical Model
   2.2 Linework Models
      2.2.1 CAD Files
      2.2.2 Thematic Map/Coverage
   2.3 Measurements and Distance Measurement Instrument
      2.3.1 Distance Measurement Instrument
   2.4 GIS Software
   2.5 GIS Surface Analysis
      2.5.1 Calculation of Elevations
      2.5.2 Calculation of Surface Lengths

3. DOT CASE STUDY
   3.1 General Background of the DOT Case Study
   3.2 Purpose of the DOT Case Study
   3.3 Scope and Guiding Questions
   3.4 Significance of the DOT Case Study

4. DATA SOURCES
   4.1 NED Data
   4.2 Photo-edited Road Layer
   4.3 DMI (Distance Measurement Instrument) Data
   4.4 Link-node Database
5. DATA ANALYSIS
   5.1 Surface Analysis
   5.2 Removing Suspect Links and Calculating Additional Attributes
      5.2.1 Suspect Links
      5.2.2 Removing Suspect Links and Calculating Additional Attributes
   5.3 Grouping and Statistical Analysis

6. ERRORS AND CONTROL
   6.1 Data Quality and Errors in GIS
   6.2 Errors from DMI
   6.3 Errors from NED
   6.4 Errors from the GIS Road Layer
   6.5 Errors from the surface length calculation

7. ANALYSIS RESULTS
   7.1 Frequency Analysis
   7.2 Descriptive Statistics
   7.3 RMSE and 95% Confidence Level

8. CONCLUSIONS

9. GLOBAL OBSERVATIONS AND IMPACT

10. RECOMMENDATIONS

11. REFERENCES

12. APPENDICES
    Appendix A. Sample Data (Attribute Table of Fully Attributed Road Layer)
    Appendix B. Distribution of Links
       B.1 All Interstate Highway Links
       B.2 Interstate Highway Links with Links Touching County Boundaries Removed
       B.3 Distribution of Links from the Suspect Link Group 2
       B.4 Distribution of Links from Group G1
       B.5 Distribution of Links from Group G2
       B.6 Distribution of Links from Group G3
       B.7 Distribution of Links from Group G4
       B.8 Distribution of Links from Group G5
       B.9 Distribution of Links from Group A
       B.10 Distribution of Links from Group B
       B.11 Distribution of Links from Group C
       B.12 Distribution of Links from Group D
       B.13 Distribution of Links from Group E
    Appendix C. Descriptive Statistics for the 5 Groups Based on Slope
       C.1 Descriptive Statistics for Group G1
       C.2 Descriptive Statistics for Group G2
C.3  Descriptive Statistics for Group G3
C.4  Descriptive Statistics for Group G4
C.5  Descriptive Statistics for Group G5

Appendix D. Descriptive Statistics for the 5 Groups Based on Length
  D.1  Descriptive Statistics for Group A
  D.2  Descriptive Statistics for Group B
  D.3  Descriptive Statistics for Group C
  D.4  Descriptive Statistics for Group D
  D.5  Descriptive Statistics for Group E

Appendix E. The Presentation Slides
Appendix F. The Program Code of Surface Length Calculations in AML
1. INTRODUCTION

This section introduces the national infrastructure information system and concepts of geographic information systems (GISs), both of which are useful tools in infrastructure management. Key issues in GIS applications, including data quality and spatial data accuracy, are described concisely. The context of a transportation case study is established considering related works from other researchers. Finally, the structure of this technical report is outlined.

1.1 Background

The nation’s civil engineering infrastructure includes, among many other things, roads, bridges, and other transportation components. These infrastructure components are integrally linked to our society and our environment. Two key data items in any infrastructure information management system are (1) the location of an object and (2) the spatial relationships between objects.

Increasingly, we are relying on computer tools and databases for information about infrastructure to enable us to better develop, use, and maintain it. Prominent among these tools are GISs which are optimized to input, store, and manipulate graphical spatial data. However, even with the recent advances in the abilities of GISs, one key problem is still data quality and accuracy issues (Rasdorf 2000).

1.2 Geographic Information Systems (GISs)

A GIS can be defined as a computer-based tool set for collecting, storing, retrieving, transforming and displaying spatial data from the real-world for a particular set of purposes (Burrough 1998). A GIS has at least five components: people, data, hardware, software, and procedures.

Spatial data is what GISs work with. Spatial (or geographical) data represents phenomena from the real-world in terms of (1) their positions with respect to a known coordinate system, (2) their attributes that are unrelated to position (such as thickness, cost, incidence of accidents, etc.), and, (3) their spatial interrelations with each other which describe how they are linked together (this is known as topology and describes space and spatial properties such as connectivity) (Burrough 1998). The availability and accuracy of data directly affect the analysis results of GIS applications.

From a functional point of view, a GIS must be able to realize at least four tasks:

1. Data input and verification,
2. Data storage and database management,
3. Data output and presentation, and
4. Data transformation.

GIS has benefited the field of transportation for many years due to the rich spatial data in transportation (Fletcher 1987, Nyerges and Dueker 1988, Dueker and Kjerne 1989, Ries 1993, Vonderohe et al. 1993, Dueker and Butler 1998). With the rapid developments in GIS, the transportation community has an unprecedented opportunity over the next few years to obtain,
use, and distribute spatial data by applying GISs (Fletcher 2000). Thus, GIS tools have become a critical tool in the transportation community.

1.3 Data Quality
Data quality is an important factor to any information system whose goal is to effectively and accurately convey information. The US Spatial Data Transfer Standard (SDTS 1997) defines “quality” as:

“Quality is an essential or distinguishing characteristic necessary for cartographic data to be fit for use.”

The quality of data is one of the most important factors for information systems that are intended to conduct analysis and support decision-making. It is clearly recognized that the validity of analysis is based on the validity of data. Unfortunately, any engineering data is an abstraction of the real-world phenomena it represents. It is not always the case that a phenomenon can be exactly visualized, understood, and modeled in an information system. In other words, errors and deviations are inherent in information management systems, which leads to the necessity for data quality control.

Three parameters - data lineage, completeness, and consistency - are often used as parameters for assessing spatial data quality and providing accuracy information for metadata that describes the data itself and is known as data about data. Data lineage describes the source of the derived data, derivation methods, and all transformations employed in producing the final data. As defined in the US Spatial Data Transfer Standard, completeness describes the objects represented and the abstract universe of all such objects (SDTS 1997). Consistency refers to the harmonious uniformity of agreement among parts of a database. It is the lack of apparent contradiction in the data.

In addition, data quality needs and levels can vary from person to person, organization to organization, or from application to application. Currently, it is most often the responsibility of the user to decide if a data set is sufficient to meet their or their organization’s quality requirements, and the standard may differ by application or use (Rasdorf 2000).

1.4 Spatial Data Accuracy
As mentioned before, GISs work with spatial data. Errors and uncertainties are intrinsic in spatial data and need to be addressed properly (Burrough 1998). Spatial data accuracy is one data quality measurement that is composed of positional accuracy (whether objects are in their correct positions), attribute accuracy (e.g., length of linear objects), and interrelationship accuracy that describes the relationships among objects.

Errors in spatial data can occur at various stages of data acquisition and transformation. Errors in perception can occur at the conceptual stage. Errors and approximations in determining a geographical location might depend on surveying skills, the provision of hardware and the choice of map projections and spheroids. Errors in the measurement of attributes depend on phenomenon variation, the accuracy of the measurement device, or observer bias and can occur during the recording of raw data. Errors can occur when data are stored in the computer
due to space limitations or storage differences. Errors can also occur when deriving new attributes from existing data by applying logical or mathematical rules that are flawed or that use low quality data.

The usual view of errors and uncertainties is that they are bad. This is not necessarily so, however, because it can be very useful to know how errors and uncertainties occur so that they can be controlled and possibly reduced. Also, a thorough knowledge and understanding of errors and error propagation can be used to improve our understanding of spatial patterns and processes (Burrough 1998).

In summary, a variety of different kinds of errors and uncertainties contribute to spatial data in accuracy and need to be carefully controlled in any GIS applications. These data errors and uncertainties will occur even with the recent advances in technologies to more accurately acquire raw data and in computer hardware and GIS software which process and analyze data. It is an important goal to provide a quantified approach to give users, analysts, and decision-makers information about the quality and accuracy of the data being provided, analyzed, and used. This is the topic of this study.

1.5 Work by Others
Spatial data accuracy and related issues have been long investigated by researchers. Some researchers tried to identify error sources and error ranges (Amrhein and Schut 1990, Goodchild 1991a, b, Goodchild 1994, Goodchild and Jeansoulin 1998, Burrough 1998, Beard 1999). Some of them delved into the propagation of errors in spatial modeling (Heuvelink 1999). In addition, the definition of formal spatial data standards was also proposed and discussed (Backe 1996). As a result, some researchers have proposed solutions for modeling spatial data accuracy (Goodchild et al. 1992, Hunter and Goodchild 1995, Goodchild 1996a, Hunter and Goodchild 1996, Chapman et al. 1997). A number of mathematical models have been used to model positional data errors. An example of these models is the method of least-square adjustments, which is well known in surveying, mapping, and GISs, and provides optimal estimation of the accuracy of positional data observations. Also, new positional data error models are being developed; for example, see Chapman et al. (1997) for a description of a new mathematical model they developed for positional data error.

However, most studies on errors are still at the research level (Fisher 1995, Goodchild and Gopal 1989, Heuvelink 1993, Lodwick et al. 1990) although systematic studies of spatial data quality are now being published (Guillill and Morrison 1995). In addition, most studies related to data quality and accuracy issues serve general GIS purposes from a theoretical point of view. There are no available systematic models and algorithms for transportation GIS users to determine data quality in transportation GIS applications. In this study, a transportation case study from state department of transportation is completely examined to provide an in-depth view of spatial data quality and accuracy in transportation GIS applications.

1.6 Case Study
The transportation community is seeing increasing numbers of GIS applications due to its rich spatial data processing, analysis, and presentation capabilities. In spite of an enormous existing base of GIS applications in transportation, there are presently no systematic and
quantified models and algorithms to provide transportation GIS users with information about spatial data accuracy and its effects in analysis. For example, a spatial query might tell us that the length of a road from one intersection to another is 5.26 miles. But is it? Presently we often accept that 5.26 miles on faith as being accurate. But how accurate?

In our study a GIS application was used as a case study to provide an in-depth view into one transportation data quality issue – length measurement. This case study developed a GIS program to calculate 3-dimensional road length based on elevation data and on a highway network, and then assessed the accuracy of the resulting distance measurement by comparing those GIS results with highly accurate DMI (distance measurement instrument) measurements. In addition to accuracy assessment, error sources were tentatively identified, control mechanics were discussed, and conclusions were drawn.

The context of the case study was interstate highways. Although this is a small proportion of the overall roadway network, the number of records is significant enough to serve quality assessment purpose. This is even truer when taking into consideration the confidence in DMI data quality for interstate highways. In order to make the amount of data in the case study manageable, we limited the test case size. However, the case study is broadly representative and the results provide useful insights into transportation distance measurement and storage accuracy.

### 1.7 Report Organization

The text of this report is divided into ten sections. The “INTRODUCTION” provides background information about the infrastructure management system, concepts of GIS, and data quality issues. Also, the context of the study was established within the research environment illustrated in our literature review. The “MODELS AND TOOLS” section illustrates GIS data models and software that are used to model surface. Information about the case study is provided in the “CASE STUDY” section which is followed by the section “DATA SOURCES” which described the major data sets used in the case study.

The case study analysis procedure was described in the section titled “DATA ANALYSIS” and flow charts are included. In the “ERRORS AND CONTROL” section error sources are tentatively identified and control mechanisms are discussed. The results of three statistical methods used to assess the accuracy are summarized in the “RESULTS” section. The “CONCLUSIONS” and “GLOBAL OBSERVATIONS AND IMPACT” sections identify what we have learned and in these sections important points are viewed in a broader view. Recommendations about what NCDOT should do now and a proposal of a future study are included in the section titled “RECOMMENDATIONS”. Figures and tables are embedded throughout the report to enhance understanding and clarify points.

The actual customized GIS program, the sample data, additional analysis results (tables and charts), and the study presentation slides, were included in the “APPENDICES”.

2. MODELS AND TOOLS

This section introduces the concepts behind topological elevation, surface modeling, and road and highway network modeling. It also establishes a relationship among them. Equipment that would be used to acquire accurate distance measurements (e.g. DMI) is described. Finally, a purely computational alternative (GIS approach) is proposed and discussed in detail in this section.

2.1 Elevation Models

Five different ways of modeling elevation data are presented herein: point model, contour lines model, TIN (triangulated irregular network), DEM/DTM (digital elevation model/digital terrain model), and mathematical model. The examples used in this section to illustrate the concepts are extracted directly from the paper *An Exercise in Digital Elevation/Terrain Models: From Point to Mathematics* (Massasati 2000).

2.1.1 The Point Model

In a point model, all elevations from sampling points are captured by storing the position and elevation of every sampling point. Each point has three accompanying numeric values, X, Y, and Z. The X and Y values identify the point position in a planar or geodetic coordinate system while the Z value represents the elevation for that point. Elevations for any other points are interpolated from the elevations of sampling points. Figure 1 provides an example of point model that represents the great pyramid of Khufu (Gizah, Egypt). The dimensions of the pyramid are simplified and reduced entirely to values 750 feet at the base and 450 feet in height. The orientation is assumed to be north-south and east-west. The coordinate of the southwest corner is zero east, zero north, and zero elevation. The entire point model for this structure consists of the X, Y, and Z coordinates of 5 points. Nothing more.

![Figure 1. The Point Model for the Pyramid of Khufu](image)

2.1.2 The Contour Line Model

A contour is an imaginary line of constant elevation on the ground surface. If the locations of several closely spaced ground points of equal elevation are plotted on a drawing, the line joining these points is called a contour line. Thus, a terrain surface may be represented by
series of contour lines with fixed interval, the vertical elevation difference between neighboring contours.

The contour line model is widely applied in surveying and mapping to represent surfaces on topographic maps like those produced by the USGS (USGS 2001d). Elevations for points along contour lines can be read directly from the elevation values represented by the contour line itself. The elevations of points between contour lines are obtained through interpolation. Figure 2 provides an example contour line model to represent the great pyramid of Khufu assuming a 100 foot vertical elevation interval, resulting in five contour lines. It must be noted that the very top of the pyramid is a single point. This model assumes that the pyramid is purely triangular in shape for the purposes of this discussion.

![Contour Line Model for the Pyramid of Khufu](image)

**Figure 2. The Contour Line Model for the Pyramid of Khufu**

### 2.1.3 DEM/DTM

DEM/DTM is the most commonly used format to represent terrain surfaces in a *raster* data structure. In a raster data format, space is divided into either regular or irregular units called cells. The most common raster data structure has either rectangular or square cells, each being the same size. Using a raster data structure model DEM/DTM stores a single elevation data value for every cell that is supposed to represent the elevation for the entire continuous cell surface. The elevations used for the surface are interpolated from the original sampling points. In addition, there are four different interpolation methods commonly used to derive the cell elevations: inverse distance weighted (IDW) interpolation, Spline interpolation, Kriging interpolation, and Trend interpolation.

The inverse distance weighted (IDW) interpolation method assumes that each sample point has a local influence that diminishes with distance. In estimating the value for a given cell, it gives greater weight to points closer to the cell than to those farther away.
The Spline interpolation method is a general-purpose interpolation method that fits a minimum-curvature surface to the sample points. The surface passes exactly through the sample points.

The Kriging interpolation method is a powerful statistical interpolation method used in diverse applications, such as health sciences, geochemistry, and pollution modeling. This technique was co-developed by Georges Matheron (a French geomathematician) and D. G. Krige (a South African mining engineer) to analyze geochemical data to find gold deposits (it worked). Kriging is based on the theory of regionalized variables, which states that naturally occurring attributes tend to reflect random quality, but at the same time, continuous and spatially dependent traits. The theory also states that objects closer together will tend to reflect dissimilar values to some range of the distance.

Finally a Trend surface interpolator creates surfaces that show the regional tendencies of the variable in question (the z value). Instead of producing a surface that fits reality well, the Trend surface interpolator was designed to display gradual long-range variations in the data. The Trend surface interpolator uses a polynomial regression to fit a least-squared surface to the input points. This finds the single best-fit equation to generate the entire surface. The Trend surface interpolation creates smooth surfaces.

DEM/DTM (a raster format data model to represent a terrain surface) possesses the advantages belonging to a raster data structure. First, it is easy to understand and visualize. Next it is also in a form that is readily processable by spatial analysis. However, even with a wide variety of available compacting methods such as chain codes, run length codes, block codes, and quadtrees and binary trees, one major continuing problem with all raster data structures is still their data redundancy. Figure 3 illustrates the DEM/DTM representing the surface of the great pyramid of Khufu using a 100×100 gridsize. The minimum elevation within a cell is used as the elevation for that cell for illustration purposes.

![Figure 3. The DEM/DTM for the Great Pyramid of Khufu](image)

2.1.4 TIN (Triangulated Irregular Network)

The Triangulated Irregular Network (or TIN) was designed by Peuker and co-workers (Perker et al. 1978) for digital elevation modeling to accurately model surfaces while avoiding the redundancies of raster format elevation models. A TIN is a vector format terrain model that
uses a sheet of continuous, connected triangular facets (cells) that are based on a Delaunay triangulation of irregularly spaced nodes or observation points. In TINs, a terrain surface is approximated with this series of irregular triangular facets. Each facet stores coordinates and elevation for its three nodes. Elevations for all other points in the facet are readily derivable by interpolation from these three nodes. Figure 4 provides the surface of the great pyramid represented by a TIN.

TINs are modeled with a vector data structure similar to those used for polygon networks. The main difference is that TIN data structure does not have to make provisions for islands or holes. A TIN overcomes some disadvantages inherent in raster data structures, particularly the problem of data redundancy. It also solves another problem inherent in a raster representation’s inability to represent areas with differing relief complexity without changing cell sizes. Thus both data redundancy and fixed cell sizes are the key problems with the raster model.

It is worth pointing out that TINs and DEM/DTMs are mutually convertible.

![Figure 4. The TIN Model for the Great Pyramid of Khufu](image)

2.1.5 *The Mathematical Model*

The mathematical model represents a terrain surface by depicting the relationship between an elevation and its X and Y coordinates analytically. For example, using the numerical values of Figures 1 and 4, each surface (of the pyramid) can be represented by the following mathematical model.

\[ Z = a \times X + b \times Y + c \]

Where:
- X, Y, and Z are surface point coordinates and
- a, b, and c are constants that can be determined for each surface by solving the equation using the three corner points (Figure 5).
Figure 5 illustrates the equations for representing the pyramid of Khufu as an example of how the mathematical model works. Consider the calculation procedure for the surface of Triangle 4 from Figure 4 as shown in the following equations.

\[
\begin{align*}
0 &= a*0 + b*0 + c & (1) \\
0 &= a*750 + b*0 + c & (2) \\
450 &= a*375 + b*375 + c & (3)
\end{align*}
\]

By solving the three equations one finds that \( b = 1.2 \), \( a = 0 \), and \( c = 0 \) and the resulting equation is:

\[ Z = 1.2*Y, \text{ where: } 0<Y<375 \text{ and } 0<Z<450 \]

Likewise, the equations for the other three surfaces are:

\[
\begin{align*}
Z &= 1.2*X \text{ where: } 0<X<375 \text{ and } 0<Z<450 \\
Z &= 1.2*(750-Y), \text{ where: } 375<Y<750 \text{ and } 0<Z<450 \\
Z &= 1.2*(750-X), \text{ where: } 375<X<750 \text{ and } 0<Z<450
\end{align*}
\]

The equations for the line that borders the triangles can be obtained by simply solving for any two surfaces.

Line AE:

\[ Z = 1.2*X = 1.2*Y \text{ or } X = Y, \text{ where } 0<=X<=375, \text{ and } 0<=Y<=375 \]

Line BE:

\[ Z = 1.2*X = 1.2*(750-Y) \text{ or } X + Y = 750, \text{ where } 0<=X<=375, \text{ and } 375<=Y<=750 \]

Line CE:

\[ Z = 1.2*(750-X) = 1.2*(750-Y) \text{ or } X + Y = 750, \text{ where } 0<=X<=375, \text{ and } 375<=Y<=750 \]

Line DE:

\[ Z = 1.2*Y = 1.2*(750-X) \text{ or } X + Y = 750, \text{ where } 375<=X<=750, \text{ and } 0<=Y<=375 \]

Figure 5 fully illustrates the mathematical model representing the surface for the great pyramid of Khufu as described above.

2.2 Linework Models

There are many ways to represent linework using computers. One is CAD (computer aided design) software, which was developed initially for drawing and cartographic purposes and has been optimized to do so. Another is GIS, in which linework and other features are captured as thematic maps (or coverages). The reason we must consider CAD files is that
much useful linear spatial information was originally stored using CAD files. In order to use the available spatial legacy data stored in CAD files instead of repeating the data input tasks already completed during the development of those CAD files, GIS must be capable of using existing CAD files without losing important information.

![Mathematical Model for the Pyramid of Khufu](image)

**Figure 5. The Mathematical Model for the Pyramid of Khufu**

### 2.2.1 CAD Files

Computer aided design (or CAD) has been to facilitate design work by providing fast, flexible, and easy to use computerized drawings and maps. Traditional paper drawings or maps are end products. They are not easy to change and update. With CAD, these time consuming and often interrelated tasks are easy to complete by working on parts of a drawing instead of redrawing it entirely. With the rapid advances in its 3D capabilities, CAD could even provide a virtual view of the actual structure during the design phase while still allowing rapid changes and updates.

Despite its strength, CAD has always had limited capabilities in supporting such analysis as quantity takeoff in construction. CAD files are still the end products of design yet they are mostly “lines on paper.” CAD does not encapsulate “knowledge” per se. The reason for incorporating CAD files into GIS applications is that CAD files have a lot of spatial information embedded in its lines that is useful for GIS. With the ability to extract information from CAD files into GIS software, we can avoid repeating the time-consuming data input task.

Many commercial CAD programs are available today. The two most commonly used in transportation are AutoCAD and MicroStation. In order to extract information from CAD files for GIS applications the GIS must be capable of converting those files into a format or formats (e.g., shapefile or coverage) that provide full support to the GIS spatial analysis capabilities.
2.2.2 Thematic Map/Coverage

In GIS, all real-world phenomena are stored in thematic maps (or coverages). A thematic map is a map that displays selected kinds of information relating to specific themes, such as city and county boundaries, land use, streets, population density, railroad, and so on. This simple but extremely powerful and versatile concept has proven invaluable for solving many real-world problems from tracking delivery vehicles, to recording details of planning applications (Burrough 1998).

Within the concept of a thematic map layer, real-world objects are internally represented as a cell or as a number of cells in raster data structures. Or, they are abstracted into point, line, or polygon features which are represented internally in vector data structures. Unlike CAD, GIS thematic layers are geo-referenced, which locates spatial data with respect to a common frame, to enable spatial analysis (Burrough 1998). In addition, each thematic layer is accompanied with one attribute table that stores descriptive characteristics for its features. In a vector data structure, each feature serves as one record in the attribute table. If necessary, additional tables can be linked with thematic layers.

Unlike CAD, thematic maps are not the end products in GIS. They just store information that is necessary for analysis and decision making. Maps, views, reports, displays, etc. can all be extracted from these thematic layers to meet user needs without changing the supporting thematic maps.

2.3 Measurements and Distance Measurement Instrument

As noted earlier, real-world objects are explicitly represented in GIS by three object types (points, lines, and polygons) using a vector data structure. In other words, any real-world objects are captured and abstracted as points, lines, or polygons, depending on the characteristics of these objects, measurement scales, and their roles in the analysis. Different measurements accompany different type of objects. Point objects are referred to using coordinates to specify their locations. In addition to locational coordinates, length is an essential measurement for line objects, while perimeter and area are considered to be important measurements for polygon objects.

Point features are spatial phenomena, each of which occurs at only one location in space (Demers 2000). Each point feature is said to be discrete in that it can occupy only a given point in space at any time. Point objects are assumed to have no spatial dimension (no length or width) although each can be referenced by its locational coordinates. Points then are said to have “0” dimensionality. However, in the real-world, point objects do have some spatial dimensions, otherwise, they would not be observed. We refer to point features using their positional coordinates.

Linear or line objects are conceptualized as occupying only a single dimension in coordinate space. These “one-dimensional” objects may be any objects that are fundamentally long and skinny, such as roads, canals, rivers, power lines, pipelines, etc. (DeMers 2000). One very important feature in addition to locational coordinates for line features is the length that allows us to measure their spatial extent. Other than length, which is the major concern of this case study, shapes and orientations of linear objects could also be measured by GIS.
Polygon features represent objects that occupy both length and width. These objects are two-dimensional areas. Two important attributes these objects possess are perimeter and area. As with lines, we can describe a polygon’s shape and orientation, but in addition, we can now describe the amount of territory (area) a polygon occupies.

In addition, if we add a third dimension (height) to polygon features, we could represent the existing terrain as surfaces. Similarly, another important attribute (volume) could be calculated if needed.

2.3.1 Distance Measurement Instrument

In this case study we were interested in one particular linear object -- road segments. More specifically, we were interested in one particular attribute of those linear objects -- length measurement. A distance measurement instrument (DMI) measures a linear path directly (length of lines). A global positioning system (GPS) is another technology that is used to obtain positional data from which a length can theoretically be obtained. GPS provides accurate locational coordinates for points (location of points) along linear objects while DMI measures the spatial extent (length) of these linear objects. The integration of DMI and GPS provides a highly accurate distance measurement system for a highway network that both measures as well as positions a road.

By driving cars equipped with DMI and GPS, surface length (or actual length between any two points) is directly read from DMI and positioned by GPS. This method is known to have high precision and accuracy (±1 foot per mile as specified in the manufacturer’s specifications of the DMI currently available to NCDOT). But these technologies are very time and labor consumers and thy are expensive. In many applications, this combined approach is either not possible or not practical. Examples include real-time emergency vehicle routing, automatic vehicle location (AVL), and monitoring of construction equipment.

2.4 GIS Software

As noted earlier, GIS can be defined as a computer-based tool set for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real world for a particular set of purposes (Burrough 1998). An essential component of any GIS is software. The conceptual functions defined in GIS are actually realized within GIS software.

Currently, there is a lot of commercially available GIS software. Examples include ArcView, ArcInfo, and ArcSDE from ESRI (Environmental Systems Research Institute, Inc.), MapInfo from MapInfo Corporation, and MGE (Modular GIS Environment) from Intergraph Corporation. There is also commercially available GIS software to meet the needs of specific fields, for example, TransCAD from Caliper Corporation is designed specifically for transportation GIS applications. In addition, software developers are still working hard to produce improvements or extensions to strengthen existing programs and create new GIS software.

Even with its advances in 3-D simulations, GIS still stores data in 2-dimensional formats. In other words, 3-dimensional length measurements for linear objects are not readily available.
from GIS. GIS must use functions provided by GIS software to calculate surface lengths. The accuracy and precision of the calculations depend not only on the accuracy of existing data sets, they also depend on the GIS software involved in the calculation procedure.

It must be noted that any GIS software has its own strengths and weaknesses. There is no single GIS program that is capable of providing all the spatial functionality any user might wish to have. When selecting GIS software to develop customized applications, two important factors must be taken into consideration, one is cost, and the other is needs. The user should select appropriate program based on his/her own situation.

2.5 GIS Surface Analysis

In this study, functions provided by ArcInfo and ArcView from ESRI are used to calculate surface lengths for road linework. These functions operate on two inputs, vector linework data and elevation data (either in raster or vector format).

2.5.1 Calculation of Elevations

ArcInfo or ArcView represents surfaces in two basic formats, a raster format (grid file) and a vector format (TIN). Grid elevation files can be converted into TINs, and vice versa. In a grid file, a grid is used to divide a space into regular rectangular units (or cells). All cells have the same size. By interpolating elevations from sampling points which may or may not be located within the cell, each cell is assigned an elevation value. It is worth pointing out here that USGS is providing raster elevation data (e.g. DEMs) with uniformly spaced sampling points. ArcView or ArcInfo converts DEMs directly into a grid file whose resolution (cell size) coincides with the sampling spacing. The interpolation is completed by USGS when producing those DEMs with uniformly spaced sampling points.

In a grid file, the elevation for an arbitrary point is actually interpolated from elevations of the surrounding grid cells (evaluated as a point at the center of the cell), using a bilinear interpolator. An example is provided in Figure 6. On the left we see elevations for a set of points that represent the center points of the grid’s cells. On the right we see an enlarged illustration of the interpolation process.

In order to interpolate the elevation value for the point E, the GIS takes the elevation values of four surrounding cells (evaluated at the cell center) to apply bilinear interpolation. First, the elevation for the two intermediate points (F and G) in the vertical direction are calculated by linear interpolation. The elevation for the left intermediate point F is calculated as:

\[ 1035 + \left( \frac{22.44}{30.00} \right) \times (1041 - 1035) = 1039.49 \]

The elevation for the right intermediate point G is similarly calculated as:

\[ 1048 + \left( \frac{22.44}{30.00} \right) \times (1060 - 1048) = 1056.98 \]

And finally the elevation for the point E is calculated using point F and G and, performing a linear interpolation in the horizontal direction as:

\[ 1039.49 + \left( \frac{22.25}{30.00} \right) \times (1056.98 - 1039.49) = 1052.46 \]
In a TIN model, as noted earlier, the terrain surface is represented using a network of irregular triangles. The three nodes of each triangle store elevations and coordinates. As demonstrated in both the earlier TIN and The Mathematical Model discussions, mathematical functions describing relationship between X and Y coordinates and elevation (Z value) can be derived from the three nodes for each triangular surface. In interpolating elevation values for a point, as was illustrated in Figure 6, the X and Y coordinates of that point are taken into consideration to decide which surface this point belongs to. The relationship between coordinates and elevation for that surface is then used to interpolate the elevation value for the new point.

In general, even though grid and TIN are two different data structures used by ArcInfo and ArcView to represent surfaces, both of them actually store elevations for some points throughout the surface. Elevations for points other than these points are interpolated when necessary rather than stored. This concept makes it possible to simulate continuous surfaces with a limited number of available elevations.

![Figure 6. A sample of Interpolating Elevation for Point](image)

2.5.2 Calculation of Surface Lengths

ESRI ArcInfo and ArcView software use the same function, SURFACELENGTH, to calculate surface lengths for linear features based on elevation data and vector polyline data. The elevation data for this function could be raster elevation data (grid files) or vector elevation data (TINs).

When using grid elevation data to calculate surface lengths for linear paths, a sample distance must be decided before performing calculations. This sample distance defines the distance along each linear object to obtain Z values. Suppose, for example, we wished to know the true three dimensional distance from point A to point B in Figure 7. The polyline feature is divided into segments based on the sample distance. Elevations for the start node and the end node (for example, nodes a and b for segment a-b) for each segment (a-b, b-c, c-d) are first interpolated from grid elevation data by applying the bilinear interpolator explained earlier.
Now, the surface length for each segment is calculated by visualizing the curved line of Figure 7 as straight line as in Figure 8 and using geometry to calculate the surface length of each segment. In other words, series of straight lines are used to approximate the actual roadway surface and the overall length is obtained by summing surface lengths of all segments for each polyline feature. Figure 8 shows the method when used to calculate surface lengths with a sample distance of 20 meters, which coincides with the sample distance used in the case study.

In the case of using TINs to provide elevations for the SURFACELength function, no sample distance is required. Input vertices for the polyline features and triangle edge intersection locations serve as the nodes to divide links into segments. Again, these segments are assumed to be straight lines to approximate the actual polyline features. Elevations for the start node and the end node of each segment are interpolated from the triangle nodes by
applying the relationship functions. Surface lengths for segments belonging to the same feature are added together to achieve the overall surface length for this feature. The differences between using grid and TIN models include the method used to divide polyline feature into segments, the method used to represent surface, and the method used to interpolate elevations for surface points. The underlying geometry used to calculate segment lengths and feature length is the same for both and is that which is illustrated in Figure 8.

One important point to emphasize again is that the accuracy of the calculated surface lengths is based on the correct positions of the linework and the quality of the elevation data. Without the lines being in the correct locations, the calculations make no sense and are of little value.

3. DOT CASE STUDY

This section describes an actual NCDOT case study on the accuracy of distances derived from GIS and DEM. In this section, background information about the case study is provided, the purpose of the case study is explained and established, the scope of the study is defined, and its significance is discussed.

3.1 General Background of the DOT Case Study

The North Carolina Department of Transportation is updating its road inventory. One key item that needs to be updated is the road length measurement. Keeping exact road lengths in the road inventory database is critical because it is the data foundation for the maintenance of existing roads, for the planning, design, and construction of new roads, and for network analysis.

Even with the many existing technologies that are available to measure road lengths, the problem of determining accurate length is not as simple as it appears to be. An efficient way to measure road lengths with reasonable cost and a high degree of accuracy is desired. Since DMI is expensive and time consuming, another method using NED and GIS is being considered.

As we noted earlier, with its rapidly growing capabilities, GIS has significant general potentials in transportation (Vonderohe 1993, 1994). The transportation community has an unprecedented opportunity over the next few years to obtain, use, and distribute spatial data by applying GIS (Fletcher 2000). In this study we explore the use of GIS for obtaining highly accurate distance calculations in support of transportation analysis.

3.2 Purpose of the DOT Case Study

Currently, road lengths are most often measured by driving a car with a highly accurate DMI. This method consumes time, money, and personnel. An alternative to DMI emerges with the developments in computer technology, measurement technology, and GIS. The actual road length can be determined by combining elevation data with road layer data and then by performing a surface analysis that leads to computer generated road lengths (DeMers 2000). This approach is fast and economical. However, the question arises as to whether this method provides a reasonable degree of accuracy.
The purpose of this project is to determine whether or not there is a practical way to calculate accurate road lengths using surface analysis based on USGS NED data and the GIS road layer data. In addition, if the method is technically and economically feasible, it is still necessary to find out exactly how accurate this method is when compared with DMI data. The overall purpose of this study then is to decide whether GIS surface analysis provides an efficient, feasible, and economical method to determine road lengths with an appropriate level of accuracy.

### 3.3 Scope and Guiding Questions

Due to the limitation of photo-edited GIS data availability for all roads, this study focused on the length measurement of only the interstate highways within North Carolina.

The guiding questions we will answer in this report are:

1. Given NED files and the road layer, how can we deduce the surface length using GIS?
2. How accurate are the results of this method when compared with DMI data?
3. What are the factors that significantly influence accuracy?
4. Is this method a feasible, efficient, and economical one?

By answering these questions we can determine whether we believe the proposed method is an appropriate way to obtain data for road lengths and for maintaining length measurements in transportation databases.

### 3.4 Significance of the DOT Case Study

North Carolina has approximately 78,000 miles of state maintained roads. For this mileage the use of DMI is time consuming and costly even though it provides accurate data. GIS, on the other hand, is fast and relatively cheap, but the quality of the distance measurements for its lines (which represent the roads) is unknown. By conducting this study, our goal was to determine which method is better—efficient and reasonably accurate. Then we can choose an appropriate method to determine distance measurements in the future.

The problem of road length measurement is not specific to NCDOT. Departments of Transportation from other states also face similar situations. The results from this study are immediately applicable. When we take a broader view of this case study and its results, it provides important information about data quality and accuracy issues that are also a major concern within the GIS community (Beard 1999, Amrhein 1990, Heuvelink 1999). Currently, there is no systematic description of how much accuracy different technologies can achieve. With this project as a case study, we have a useful and in-depth view of this problem.

### 4. DATA SOURCES

The DATA SOURCES section describes the four major data sets that are involved in this case study. They are NED data (DEMs provided by USGS for all states in the U.S.), a GIS photo-edited road layer, DMI data, and geometry and topology from a link-node database.
4.1 NED Data

The USGS National Elevation Dataset (NED) has been developed by merging the highest-resolution, best-quality elevation data available across the United States into a seamless raster format (USGS 2001a). NED is the result of the USGS effort to provide 1:24,000-scale Digital Elevation Model (DEM) data for the conterminous US and 1:63,360-scale DEM data for Alaska (USGS 2001a, b).

The National Elevation Dataset is designed to provide National elevation data in a seamless raster form with a consistent datum, elevation unit, and projection (USGS 2001b). In the NED assembly process, data corrections were made to perform edge matching and fill silver areas of missing data. The elevation values were converted to decimal meters as a consistent unit of measure. The North American Datum of 1983 (NAD83) was consistently used as the horizontal datum, and all the data were recast in a geographic projection (USGS 2001b). NED has a resolution of one arc-second (approximately 30 meters) for the conterminous United States, Hawaii, and Puerto Rico and a resolution of two arc-seconds for Alaska (USGS 2001b).

NED is essentially a DEM (digital elevation model) in which terrain elevations for ground positions are sampled at either a regular or irregular horizontal interval (Anderson 1998, USGS 2001). The terrain surface is represented by using an array with X, Y, and Z values. DEMs are readily convertible into a grid file with most commercial GIS software.

NED data involved in this study was purchased from USGS in 1998. It was developed based on 7.5-minute DEMs, which were derived from hypsographic data (contour lines) and/or photogrammetric methods using USGS 7.5-minute topographic quadrangle maps. DEM files from USGS with most parts having 30-meter by 30-meter grid spacing while some parts had 10-meter by 10-meter grid spacing. Figure 9 shows a sample NED data file in text format and illustrates how NED data is converted into a grid file for further calculations. It provides elevations in a raster data structure for the GIS analysis in this case study.

The first two rows shown in Figure 9 specify the number of columns and rows in the file (5 rows and 5 columns for this sample file). The xllcorner and yllcorner values specify the coordinates of the start point (the leftmost, lowest point). The cellsize indicates the 30 meter grid resolution. NODATA_value specifies –9999 as a given indicator that no elevation data exists for that cell. Following that is a set of numbers that specify the elevations for each cell in the sequence of row by row, top to bottom, and left to right within each row (in this example, in the sequence of (row 1, column 1), (row 1, column 2), …, (row 1, column 5), then (row 2, column 1), (row 2, column 2), …, to (row 5, column 5)).

4.2 Photo-edited Road Layer

The North Carolina DOT maintains the roadway network in sets of CAD (Computer Aided Design) files in a DGN (MicroStation Design File) file format. These CAD files are readily usable in GIS by geo-referencing. The results of CAD-GIS conversion can be stored and viewed as a thematic layer in GIS. This type of data is called road layer data in this study. The ability of GIS software to easily convert from existing CAD files saves the significant effort that otherwise would be necessary to repeat the intimidating tasks of linework and data input.
One problem for road layer data is that the CAD files had already been “smoothed” for producing maps. For example, a road might have been moved from its actual position for better visualization purposes. In other words, the road layer data extracted directly from these CAD files is not capable of truly representing the 3D topography of the roads. It contains errors and deviations that can propagate into further analysis (Heuvelink 1999, Donohoo 1990) if not corrected. In order to reliably use this CAD data source, it has to be revised or corrected. The revision process uses aerial photos to correct the road layer by digitizing these photos and matching them with the linework. The concept is illustrated in Figure 10, in which the GIS line representing the N-S running road is correctly placed directly over that road. However the diagonal line is clearly not over the center of the road and, therefore, is in the wrong location. Correcting the GIS linework in this manner is referred to as photo revising.
4.3 DMI (Distance Measurement Instrument) Data

This type of data was acquired by driving cars with highly precise and accurate DMI along roads to measure their lengths. This activity is relatively expensive and time consuming. However, DMI data is the most accurate length measurement (± 1 foot per mile as specified in the manufacturer’s specifications) presently available to NCDOT and thus it was used as the basis for comparison for this study. DMI data is available for all interstate highways in NC. It was obtained during a field measurement effort in the summer of 2000.

4.4 Link-node Database

Two essential components of any spatial specification are geometry and topology. This spatial information for the NC roadway network is available from a link-node database, which was used to obtain geometry and topology for all of the road segments involved in this study. Figure 11 illustrates the graphical definitions of links, nodes, and roads for this study.
Table 1 provides summarized information about data sources, data models, and metadata for the four data sets presented in section 4. Appendix A provides sample data in the format of an attribute table.

Table 1. Summarized Information of Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Data Model</th>
<th>Metadata</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>NED</td>
<td>Grid (Raster)</td>
<td>Based on 7.5-minute DEM, NAD83 datum, 1:24,000 scale, 30-by 30-meter resolution, purchased in 1998.</td>
<td>USGS</td>
</tr>
<tr>
<td>Road Layer</td>
<td>Coverage (Vector)</td>
<td>Revised from CAD drawings during project construction.</td>
<td>NCDOT, GIS</td>
</tr>
<tr>
<td>DMI</td>
<td>Distance (Attribute)</td>
<td>Collected Summer 2000, for all interstate highways, most accurate data</td>
<td>NCDOT, GIS</td>
</tr>
<tr>
<td>Topology &amp; Geometry data</td>
<td>Link-node Database</td>
<td>Existing database form PMU (under updating)</td>
<td>NCDOT, PMU</td>
</tr>
</tbody>
</table>

5. DATA ANALYSIS

This section provides a detailed description of the complete analysis procedure that was undertaken in this study, with flowcharts and step-by-step operations specified. The overall procedure is partitioned into three parts: surface analysis; removing suspect links and calculating additional attributes; and grouping and statistical analysis. The steps within each part are described in detail in the following subsections.

5.1 Surface Analysis

The goal of surface analysis is to calculate the actual length of the links comprising the highway network. To do so we use a horizontal distance obtained from the GIS linework, a vertical distance obtained from NED data, and a GIS function to calculate a 3-dimensional true slope distance. This calculated distance can then be compared to a slope distance actually measured using DMI. Figure 12 shows the flowchart for this sequence of activities.
(1) The original NED data file is converted into a 30-meter by 30-meter grid file (This concept was previously illustrated in Figure 9).

(2) CAD files are matched with aerial photos as discussed in Section 4.2. If roads on the CAD files do not match roads on the aerial photos, they are revised by digitizing the aerial photos. The product of this step is a cleaned road layer with all roads in their correct 2D spatial positions. The cleaned road layer is then converted from a CAD file into a GIS shapefile. By next building topology (a built-in GIS capability), the length of all links (2D planimetric length) is generated automatically by the GIS software.

(3) Geometry and topology information from the link-node database and actual DMI link lengths are attached to the road layer as attributes. After this step, the cleaned road layer has both 2-dimensional planimetric length (generated by the GIS software) and 3-dimensional length (DMI data) as attributes.

(4) The cleaned road layer is overlaid with the 30-meter by 30-meter elevation grid file. By using the NED elevations from the grid file and performing surface analysis, GIS calculates the actual 3-dimensional length for all links based on the 2-dimensional length of the cleaned road layer and the elevation information from the grid file. At this point we have a revised road layer with planimetric length, DMI length (actual length), and GIS generated surface length which can be used for further analysis.

GIS applications written in both ARC Macro Language and ArcView Avenue are capable of calculating surface length mentioned in step 4 above by combining the grid file (which has the elevation data) and a polyline coverage (which has the line locations). NCDOT typically uses a program written in ARC Macro Language that incorporates the $\text{SURFACELENGTH}$ function. This function approximates links with series of 20 meter straight line segments to calculate an overall surface length. Figure 8, which was previously discussed in Section 2.5.2, shows how a link that is longer than 20 meters is divided into multiple 20 meter or less

![Figure 12. Flowchart of Surface Analysis](image-url)
segments and how the surface lengths for the segments are calculated and added together to provide the overall link surface length. The elevations for segment nodes are actually interpolated from NED using the bilinear interpolation method that was previously discussed in Section 2.5.1 and illustrated with an example as shown in Figure 6.

Surface analysis is finished with the Revised Road Layer as the end product.

5.2 Removing Suspect Links and Calculating Additional Attributes

It is recognized that all data sources have inherent errors. Some are from the datasets themselves, while some are introduced during the analysis procedure (Donohoo 1990, Ngan 1995, Burrough 1998). What we are trying to do here is to remove dataset errors so that we are left only with errors that result from the elevation analysis itself. We can then evaluate those errors on their own merit. Thus, since we are working with a link node database and a link node GIS roadway network we wish to identify and remove links that have suspect length measurements.

5.2.1 Suspect Links

Three groups of suspect links were identified for removal from the test case data set in this study. The first group of links that needed to be removed were links that touch county boundaries. These are known to have uncertain length measurements due to the inability to exactly pinpoint the county boundary and subsequently, the start and end nodes for the links on each side of the boundary. They are not valid for accuracy assessment with any degree of certainty. They were removed from analysis by applying the 1st filter shown in Figure 13. The concept is illustrated in Figure 14.

The second group of suspect links were those which had a 2D planimetric length greater than their 3D surface length. This is clearly an impossible situation. It is caused by inherent system errors in integrating raster elevation data with vector roadway data to calculate surface length. These are inherent system errors in the GIS raster – vector conversion process. They are discussed further in the next subsection.

The third group of suspect links were identified by taking into consideration that all interstate links are relatively short flat links (64% of all links are less than 1 mile in length; 83% are less than 1.5 mile in length). For these links there should be no significant difference between planimetric length and surface length. If there is a significant difference there is a problem either with the line in the GIS or the value in the database. Thus, something is wrong with the line and no valid analysis can result from using it. For example, consider a link whose planimetric length in the database might be 0.50 mile but whose GIS surface length is 1.0 mile. This link is clearly suspect. There is a problem with the link itself. Thus, the link would be a poor choice to use to evaluate accuracy. Any links with that difference exceeding a tolerance level are not valid for accuracy assessment. These links were removed from accuracy assessment by applying the 3rd filter shown in Figure 13.

Figure 15 shows how many suspect links were removed for all of the suspect groups identified above. Note that the third filter removed 51 links. This was with an absolute difference between planimetric length and surface length of 0.1 mile. Previously we set that difference
tolerance at 0.01 mile and removed 383 links. However, the NCDOT advisory group felt this
tolerance difference was too constraining so that we relaxed it to 0.1 mile.

5.2.2 Removing Suspect Links and Calculating Additional Attributes

The goals for removing suspect links and calculating additional attributes were to remove
those three groups of suspect links identified above and to derive new attributes that are
necessary for analysis from existing attributes. Figure 13 illustrated the flowchart for this
sequence of activities. This section further enumerates and clarifies these activities.

(5) All the links touching county boundaries were individually identified and then
removed by overlaying the revised road layer with the county boundary layer. Figure 14 illustrates how this 1st filter works.

(6) All the necessary attributes for future analyses were added, the second filtering
was performed, links were grouped, and statistical analysis was completed. New
attributes include:

(a) A2DMILE, the result of converting LENGTH (computer generated
planimetric length) from units of meters into units of miles;

(b) D_PMU_A2D, the difference between A2DMILE and PMUMILE (the
actual length measurement from DMI);

(c) Slope and SLOPE%, the slope of links in both decimal and percent formats;

(d) DIFFERENCE, the absolute difference between PMUMILE and LENMILE
(GIS length measurement); and,

(e) DPERMILE, which indicates the proportional difference between
PMUMILE and LENMILE in the format of mile/mile.

(7) All the newly added attributes were calculated using the available data. The
equations used to calculate values for these attributes are the following (in
ArcView format):

\[
\begin{align*}
A2DMILE &= \text{LENGTH}/1609 \\
\text{DIFFERENCE} &= (\text{LENMILE} - \text{PMUMILE}).\text{ABS} \\
\text{DPERMILE} &= \text{DIFFERENCE}/\text{PMUMILE} \\
\text{D}_{\text{PMU}}\text{A2D} &= (\text{PMUMILE} - \text{A2DMILE}).\text{ABS} \\
\text{SLOPE} &= ((\text{SLENGTH} \times \text{SLENGTH}) - (\text{LENGTH} \times \text{LENGTH}))/\text{LENGTH} \\
\end{align*}
\]

(8) The third filter was applied to remove the third group of suspect links – links
having an absolute difference between 2D length and 3D length exceeding the
tolerance. These suspect links were removed from the fully attributed road layer,
which lead to a final remaining road layer.
One thing to observe after performing the step (7) calculations is that 66 links out of a total of 2336 links (2.8%) end up with a LENGTH (2-dimensional length) that is longer than SLENGTH (3-dimensional length). These links are specified as the special group 2 links mentioned in the previous section. The explanation for this result is that the links within this group are relatively short links (46 links out of 66, or 70%, are less than 0.05 mile; 60 out of 66, or 91%, are less than 0.5 mile). GIS uses two different data models (raster and vector, respectively) to calculate LENGTH and SLENGTH. For short links, there is inherent system error because of scale and storage difference. Table 2 provides descriptive statistics for length and error attributes for these 66 links. Further analyses did not include these 66 links.
Table 2. Descriptive Statistics for 66 Links

<table>
<thead>
<tr>
<th></th>
<th>LENMILE</th>
<th>PMUMILE</th>
<th>A2DMILE</th>
<th>DIFFERENCE</th>
<th>DPERMILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.16</td>
<td>0.17</td>
<td>0.16</td>
<td>0.02</td>
<td>0.28</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.37</td>
<td>0.38</td>
<td>0.37</td>
<td>0.04</td>
<td>0.63</td>
</tr>
<tr>
<td>Range</td>
<td>1.82</td>
<td>1.93</td>
<td>1.82</td>
<td>0.22</td>
<td>4.67</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.83</td>
<td>1.94</td>
<td>1.83</td>
<td>0.22</td>
<td>4.67</td>
</tr>
<tr>
<td>Count</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
</tr>
</tbody>
</table>

Figure 15 shows the numbers of links removed for the three suspect link groups to obtain a “cleaned” road layer as the end product. That cleaned road layer is composed of a set of links that are deemed to be valid and reasonable data set that can be used for further accuracy assessment.

5.3 Grouping and Statistical Analysis

Working with the cleaned road layer data set all links were grouped based on slope and length. The reason for doing so was to determine whether slope and length are significant factors influencing length measurement accuracy, and if so, what impact they might have. After grouping, statistical analyses including frequency analysis, descriptive statistics, and RMSE, were conducted for all groups. The overall procedure is illustrated in Figure 16. For the locational distribution of the total 10 groups (groups G1 to G5 and A to B), the reader is referred to Appendix B.

(1) These 185 links are links that touch county boundaries;
(2) These 66 links are links that have 2 dimensional length greater than 3 dimensional length;
(3) These 51 links are links that have the difference between 2 dimensional length and 3 dimension length exceeding tolerance level.

Figure 15. Removing Suspect Links from Further Analyses

(9) All the links in the cleaned road layer were sorted based on SLOPE, which resulted in 5 sub categorizations as groups G1 to G5. The slope range and the number of links in each group are as follows.

<table>
<thead>
<tr>
<th>Group</th>
<th>Slope Range</th>
<th>Number of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>0% =&lt; SLOPE &lt; 2%</td>
<td>511</td>
</tr>
<tr>
<td>G2</td>
<td>2% =&lt; SLOPE &lt; 4%</td>
<td>900</td>
</tr>
<tr>
<td>G3</td>
<td>4% =&lt; SLOPE &lt; 6%</td>
<td>415</td>
</tr>
<tr>
<td>G4</td>
<td>6% =&lt; SLOPE &lt; 8%</td>
<td>119</td>
</tr>
</tbody>
</table>
Group G5  \[8\% \leq \text{SLOPE} < 24\%\] 89 Links

Totally: 2034 Links

All links with a SLOPE greater than or equal to 8% were grouped together because the number of these links was so small that the analysis results would otherwise have been statistically insignificant. If we had divided Group G5 into 2% slope ranges, there would not have been enough links within the new groups to maintain the validity of statistical analysis.

(10) All links in the cleaned road layer were also grouped based on PMUMILE, the actual length (obtained from DMI), which resulted in another 5 groups, shown below as groups A to E. The length range and the number of links in each of these groups are as follows.

<table>
<thead>
<tr>
<th>Group</th>
<th>PMUMILE Range</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>0.00 mile (\leq) PMUMILE (\leq) 0.10 mile</td>
<td>166</td>
</tr>
<tr>
<td>Group B</td>
<td>0.10 mile (&lt;) PMUMILE (\leq) 0.50 mile</td>
<td>565</td>
</tr>
<tr>
<td>Group C</td>
<td>0.50 mile (&lt;) PMUMILE (\leq) 1.00 mile</td>
<td>561</td>
</tr>
<tr>
<td>Group D</td>
<td>1.00 mile (&lt;) PMUMILE (\leq) 2.00 mile</td>
<td>598</td>
</tr>
<tr>
<td>Group E</td>
<td>2.00 mile (&lt;) PMUMILE (\leq) 5.00 mile</td>
<td>144</td>
</tr>
</tbody>
</table>

Totally: 2034 Links

(11) Frequency analysis, descriptive statistics, and RMSE analyses were performed for each group. The analyses and assessments were then summarized together to presents the results.

---

**Figure 16. Grouping and Statistical Analysis**
6. ERRORS AND CONTROL

This section introduces the issues related to errors in this GIS data quality case study. Four major error sources encountered in this study are identified and control mechanisms for dealing with these errors are also discussed.

6.1 Data Quality and Errors in GIS

As mentioned earlier, data quality is always a critical issue in GIS applications. It is an important factor in the process of effectively and accurately conveying both spatial and nonspatial information (FGDC 1998). Errors are inherent in any GIS applications. Some are from datasets that are involved in analysis and some are from analysis methods themselves. In this study, major error sources were identified and controlled through careful data preparation and by applying screening filters during the analysis. It was recognized though that not all errors can be eliminated from the accuracy assessment in this case study. The following subsections discuss the nature of some of these errors.

6.2 Errors from DMI

DMI data was supposed to be the most accurate data (± 1 foot per mile as specified in a manufacture’s specifications) in this study and it was used as the base reference for accuracy assessment. The GIS length calculations were compared with that DMI data to determine how accurate the GIS data was. DMI is capable of providing accurate distance measurement between two locations by driving a car equipped with a highly precise and accurate DMI along the actual roads.

The primary errors that occur with DMI come from the operator’s failure to exactly pinpoint the county boundaries and subsequently, the start and end nodes for links touching county boundaries. This operator error occurs, in part, because county boundaries are just conceptual features instead of physical features like intersections. Even with intersections there is often the operator’s interpretation of where the center of the intersection is. With roads there is the question of precisely where they begin and end. Nearly all of these questions are related to the use of DMI rather than to its accuracy.

DMI operator errors were excluded from our accuracy assessment (for county boundaries) simply by applying the 1st filter, wherein all links touching county boundaries were removed. These links comprise the 1st group of suspect links that need to be removed before analysis. By removing this suspect group of links from the accuracy assessment, DMI provided a valid reference for comparisons based on our knowledge.
6.3 Errors from NED

NED data involved in this study came from USGS. This set of data is the result of USGS efforts to provide seamless raster elevation data for the United States. It is essentially a DEM. DEMs capture terrain surfaces with regular or irregular sampling points with X, Y, and Z values. Elevations for positions other than those sampling points are interpolated from elevations of those sampling points. Thus, there are many inherent errors in these elevation models.

First, the elevations for all sampling points have errors. The degree of accuracy for these points depends on the equipment, method, and personnel involved in measuring the elevations. Second, elevation values for points with missing values are obtained from interpolation. Interpolation is just an approximate method that assumes that the elevation for one unknown point is related to its neighboring points. In addition, by using points to represent the continuous terrain surface and interpolating known intermediate elevations from known sampling points, a second order of approximation is introduced.

The errors mentioned here not can be eliminated from the study, but their importance be evaluated. That was the purpose of the study and the last section presented those results.

6.4 Errors from the GIS Road Layer

The vector data representing interstate highways, the GIS road layer, was originally stored as CAD files in DGN format. Some roads in these files are not shown in their actual locations. This is obvious when matching aerial photos with the linework in the CAD files. It is obvious that without roads in their actual locations, surface length calculations would generate totally different (and incorrect) lengths and the resulting accuracy assessment would its validity.

This set of errors can be controlled very well by matching aerial photos with CAD files and correctly placing all the linework directly over the roadway centerline. Still it must be recognized that by digitizing roads to assure that they are in their correct locations, another set of errors – digitizing errors, are introduced into road layer.

6.5 Errors from the Surface Length Calculation

This set of errors is related to the method itself. As illustrated in Figure 8, in order to calculate the surface length for the links, each link was divided into shorter segments with sample distances of 20 meters. Each segment is assumed to be straight. Surface length for each segment was calculated based on trigonometry using the horizontal distance and vertical distance (difference in elevations). Then, segment lengths were added together to give the total length for each link.

First of all, approximating links using a series of straight line segments introduced errors in calculating the surface length. In addition, the elevations used in getting the vertical distance between two points came from interpolation as was illustrated in Figure 6 with an example. This set of errors is due to the limitations of the ESRI SURFACELENGTH function. The existence of these errors is one of the reasons for assessing accuracy before applying this method to update distance measurements.
7. ANALYSIS RESULTS

This section represents the results for the three statistical methods used to assess data accuracy in the case study: frequency analysis, descriptive statistics, and RMSE. Charts and tables are provided in this section to better understand the analysis results.

7.1 Frequency Analysis

Frequency provides an in-depth way to view error distribution using percentages. It represents the accuracy for every group by counting the number and calculating the percentage of links within certain error ranges for each group. Summary tables and charts showing error trends are provided to illustrate the relationships between error and slope and between error and length.

Table 3 summarizes the results of frequency analysis for groups based on slope. Figures 17 and 18 illustrate the comparisons of error ranges for groups based on slope from perspectives of both absolute error and proportional error, respectively. Observations on both of these grouping perspectives are summarized in this section.

Observations from frequency analysis for groups based on slope are:

1. Accuracy is good for all groups. The performance of DIFFERENCE (absolute error) is better than DPERMILE (proportional error). The explanation for this is that since most links are short links (about 35% of all cleaned links are less than 0.5 mile long and 63% are less than 1 mile long) errors are somewhat exaggerated in a proportional format by one is dividing the absolute by an ever decreasing length.

2. Accuracy is good for all groups.

   • From the aspect of absolute error (DIFFERENCE), Group G5 is the group that shows the worst accuracy. However, it still has 72% of its links within a 0.01 mile error range, 81% of its links within a 0.02 mile error range, and 88% of its links within a 0.03 mile error range. The remaining four groups have about 85% of their links within a 0.01 mile error range and about 95% of links within a 0.03 mile error range.

   • From the perspective of proportional error (DPERMILE), Groups G4 and G5 have similar accuracy and they are worse than the other groups. But together, these 2 groups still have about 57% of their links within a 0.01 mile/mile error range and 73% of their links within a 0.03 mile/mile error range.

3. Accuracy decreases as the slope increases as is shown from the data. Error occurs primarily within high slope groups.
Table 3. Frequency Analysis for Groups Based on Slope

<table>
<thead>
<tr>
<th>Groups</th>
<th>Group G1</th>
<th>Group G2</th>
<th>Group G3</th>
<th>Group G4</th>
<th>Group G5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of Links</td>
<td>511</td>
<td>900</td>
<td>415</td>
<td>119</td>
<td>89</td>
<td>2034</td>
</tr>
<tr>
<td>&lt;=0.01 mile</td>
<td>427</td>
<td>766</td>
<td>352</td>
<td>94</td>
<td>64</td>
<td>1703</td>
</tr>
<tr>
<td>%</td>
<td>84%</td>
<td>85%</td>
<td>85%</td>
<td>79%</td>
<td>72%</td>
<td>84%</td>
</tr>
<tr>
<td>&lt;=0.02 mile</td>
<td>475</td>
<td>828</td>
<td>382</td>
<td>108</td>
<td>72</td>
<td>1865</td>
</tr>
<tr>
<td>%</td>
<td>93%</td>
<td>92%</td>
<td>92%</td>
<td>91%</td>
<td>81%</td>
<td>92%</td>
</tr>
<tr>
<td>&lt;=0.03 mile</td>
<td>486</td>
<td>856</td>
<td>394</td>
<td>111</td>
<td>78</td>
<td>1925</td>
</tr>
<tr>
<td>%</td>
<td>95%</td>
<td>95%</td>
<td>95%</td>
<td>93%</td>
<td>88%</td>
<td>95%</td>
</tr>
<tr>
<td>&lt;=0.01 mile/mile</td>
<td>389</td>
<td>618</td>
<td>277</td>
<td>64</td>
<td>53</td>
<td>1401</td>
</tr>
<tr>
<td>%</td>
<td>76%</td>
<td>69%</td>
<td>67%</td>
<td>54%</td>
<td>60%</td>
<td>69%</td>
</tr>
<tr>
<td>&lt;=0.02 mile/mile</td>
<td>431</td>
<td>712</td>
<td>322</td>
<td>80</td>
<td>60</td>
<td>1605</td>
</tr>
<tr>
<td>%</td>
<td>84%</td>
<td>79%</td>
<td>78%</td>
<td>67%</td>
<td>67%</td>
<td>79%</td>
</tr>
<tr>
<td>&lt;=0.03 mile/mile</td>
<td>449</td>
<td>763</td>
<td>345</td>
<td>84</td>
<td>67</td>
<td>1708</td>
</tr>
<tr>
<td>%</td>
<td>88%</td>
<td>85%</td>
<td>83%</td>
<td>71%</td>
<td>75%</td>
<td>84%</td>
</tr>
</tbody>
</table>

Figure 17. Comparison of Absolute Error from Frequency Analysis for Groups Based on Slope
Table 4 summarizes the results of frequency analysis for groups based on length. Figures 19 and 20 illustrate the comparisons of error ranges for groups based on length from the perspective of both absolute error and proportional error, respectively.

**Table 4. Frequency Analysis for Groups Based on Length**

<table>
<thead>
<tr>
<th>Groups</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of Links</td>
<td>166</td>
<td>565</td>
<td>561</td>
<td>598</td>
<td>144</td>
</tr>
<tr>
<td>&lt;=0.01 mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of links</td>
<td>145</td>
<td>455</td>
<td>476</td>
<td>510</td>
<td>117</td>
</tr>
<tr>
<td>%</td>
<td>87%</td>
<td>81%</td>
<td>85%</td>
<td>85%</td>
<td>81%</td>
</tr>
<tr>
<td>&lt;=0.02 mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of links</td>
<td>150</td>
<td>501</td>
<td>525</td>
<td>555</td>
<td>134</td>
</tr>
<tr>
<td>%</td>
<td>90%</td>
<td>89%</td>
<td>94%</td>
<td>93%</td>
<td>93%</td>
</tr>
<tr>
<td>&lt;=0.03 mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of links</td>
<td>155</td>
<td>522</td>
<td>536</td>
<td>574</td>
<td>138</td>
</tr>
<tr>
<td>%</td>
<td>93%</td>
<td>92%</td>
<td>96%</td>
<td>96%</td>
<td>96%</td>
</tr>
<tr>
<td>&lt;=0.01 mile/mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of links</td>
<td>95</td>
<td>233</td>
<td>400</td>
<td>535</td>
<td>138</td>
</tr>
<tr>
<td>%</td>
<td>57%</td>
<td>41%</td>
<td>71%</td>
<td>89%</td>
<td>96%</td>
</tr>
<tr>
<td>&lt;=0.02 mile/mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of links</td>
<td>95</td>
<td>307</td>
<td>490</td>
<td>570</td>
<td>143</td>
</tr>
<tr>
<td>%</td>
<td>57%</td>
<td>54%</td>
<td>87%</td>
<td>95%</td>
<td>99%</td>
</tr>
<tr>
<td>&lt;=0.03 mile/mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of links</td>
<td>95</td>
<td>364</td>
<td>523</td>
<td>582</td>
<td>144</td>
</tr>
<tr>
<td>%</td>
<td>57%</td>
<td>64%</td>
<td>93%</td>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Observations from frequency analysis for groups based on length are:

(1) The performance of DIFFERENCE is better than DPERMILE, which confirms the first observation from groups based on slope.

(2) Accuracy is good for all groups.

- From the perspective of absolute error (DIFFERENCE), Group E is the group with the worst accuracy. However, it still has 81% of its links within a 0.01 mile error range, 93% within a 0.02 mile error range, and 96% links within a 0.03 mile error range. Other groups have about 85% of their links within a 0.01 mile error range, 92% within a 0.02 mile error range, and 95% within a 0.03 mile error range.

- From the perspective of proportional error (DPERMILE), Groups A and B have similar accuracy and they are worse than the other groups. But together, these 2 groups still have about 55% of their links within a 0.02 mile/mile error range and 60% within a 0.03 mile/mile error range.

(3) Accuracy increases as the length increases. Error occurs primarily with short length groups.

Figure 19. Comparison of Absolute Error from Frequency Analysis for Groups Based on Length
7.2 Descriptive Statistics

For each group, descriptive statistics are obtained for the attributes LENMILE (GIS length measurement), PMUMILE (DMI length measurement), A2DMILE (computer generated planimetric length), SLOPE, DIFFERENCE (error in the unit of mile), and DPERMILE (error in mile/mile). Statistical parameters are used to describe the general trend of different length measurements, slopes, and errors. Parameters include Mean, Variance, Standard Error, Median, Range, Confidence Level, and Count. For a complete list of descriptive statistics, please refer to Appendices C and D.

One important observation is that most links are relatively short. Average values of link length are less than 1 mile for 4 out of 5 groups based on slope. From groups based on length, those having a length less than 1 mile (A, B, and C) contain a total of 1076 links, which is about 63% of all links. This observation supports the explanation of why the performance of DIFFERENCE is better than DPERMILE. Another important observation is that descriptive statistics reinforce the observations from frequency analysis.

7.3 RMSE and 95% Confidence Level

RMSE is a kind of generalized standard deviation. From a statistical point of view, RMSE is useful in describing differences between two data sets. In the spatial world, this statistic is widely used to describe positional accuracy, including horizontal and vertical accuracy (FGDC 1998). However, with its ability to depict the difference between two datasets, RMSE was deemed to be a particularly useful tool for evaluating the accuracy of our results. Thus, RMSE has been calculated for all groups in this study to reveal the difference between DMI length
measurements and GIS length calculations. The following equation is used to calculate RMSE.

\[
RMSE = \left[ \frac{(e_1 + e_2 + e_3 + \ldots + e_n)}{n} \right]^{\frac{1}{2}}; \quad e_1, e_2, e_3, \ldots, e_n ---- errors
\]

In this study, RMSE was calculated for both absolute error (DIFFERENCE) and proportional error (DPERMILE). While assuming the normal distribution of errors, a 95% confidence level was determined by multiplying RMSE by 1.96, which corresponds to \( Z_{0.025} \) in a standard normal distribution. The 95% confidence level indicates the error range in which an occurrence has a 95% probability of falling. Table 5 provides RMSE and 95% confidence level values for all groups based on slope. Table 6 provides the same information for all groups based on length. Figures 21 and 22 provide the comparisons of RMSE with both the absolute error and the proportional error for groups based on slope and length, respectively.

### Table 5. RMSE and 95% Confidence Accuracy for Groups Based on Slope

<table>
<thead>
<tr>
<th>Groups</th>
<th>RMSE</th>
<th>95% Confidence Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIFFERENCE</td>
<td>DPERMILE</td>
</tr>
<tr>
<td>G1</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>G2</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>G3</td>
<td>0.02</td>
<td>0.12</td>
</tr>
<tr>
<td>G4</td>
<td>0.02</td>
<td>0.39</td>
</tr>
<tr>
<td>G5</td>
<td>0.03</td>
<td>0.28</td>
</tr>
</tbody>
</table>

**Figure 21. Comparison of 95% Confidence Accuracy for Groups Based on Slope**
Table 6. RMSE and 95% Confidence Accuracy for Groups Based on Length

<table>
<thead>
<tr>
<th>Groups</th>
<th>RMSE</th>
<th>95% Confidence Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIFFERENCE</td>
<td>DPERMILE</td>
</tr>
<tr>
<td>A</td>
<td>0.02</td>
<td>0.51</td>
</tr>
<tr>
<td>B</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>C</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>D</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>E</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Figure 22. Comparison of 95% Confidence Accuracy for Groups Based on Length

From these two tables and two figures, similar observations may be deduced as were previously deduced by frequency analysis.

(1) The performance of DIFFERENCE is better than DPERMILE.

(2) Accuracy is good.

- Considering absolute error (DIFFERENCE), links of groups G1, G2, G3, and G4 (based on slope) have a 95% probability of having less than a 0.03 mile error as indicated by the 95% confidence level. Links from group G5 have a 95% probability of having less than a 0.06 mile absolute error. Groups based on length indicate that a link from groups A, B, or E has a 95% probability of
having less than a 0.04 mile error while a link from group C or D has a 95% probability of having less than a 0.03 mile error.

- Considering proportional error (DPERMILE), links from groups G1, G2, and G3 (based on slope) have a 95% probability of having less than approximately a 0.23 mile/mile proportional error. Groups B, C, D, and E have the 95% confidence accuracy as 0.19 mile/mile, 0.05 mile/mile, 0.02 mile/mile, and 0.01 mile/mile, respectively. Groups A, G4, and G5 have the worst results, which correspond to the groups with the shortest links and the group with the highest slopes.

(3) Within groups errors occur primarily with high slopes and short lengths, especially groups G5 and A.

8. CONCLUSIONS

Data quality/accuracy were addressed in this study by comparing results from a GIS analysis with DMI data. In order to assure the validity of analysis, it was necessary to initially remove invalid data so that it was not involved in the accuracy assessment. To do so, three filters were applied to remove links with questionable lengths from the accuracy assessment study. Three methods (descriptive statistics, frequency analysis, and RMSE) were used to determine whether or not the GIS method provides good accuracy of distance measurement when compared with DMI data. It was also determined that the accuracy is very good and that when there are errors they occur primarily within groups with high slopes and short lengths.

This study leads to a number of conclusions. First, the method of using GIS plus NED is a technically feasible way to determine lengths using surface analysis functions provided by GIS tools. Programs may be written using either ArcView Avenue or ARC Macro Language that are capable of calculating surface length based on a shapefile (coverage) and NED. In addition, there are also readily available extensions from the ESRI (Environmental Systems Research Institute, Inc.) website such as Surface Tools for Points, Lines and Polygons (developed by Jeff Jenness from US Forest Service), which do the same thing, but do not require the user to write programs using Avenue or Macro Language (Jenness 2001). Second, the GIS method is accurate as demonstrated by the analysis results. Third, most errors occur with links that have short lengths and high slopes.

It was also recognized that the origin of errors might be either the data sets or the method itself. The grid file that contains elevation data is built using NED, which is acquired by sampling and interpolating. Thus, that original NED data contains both horizontal and vertical errors. These are the errors that we were trying to quantify and assess. Although the matching and digitizing process corrects it, the road layer also has inherent digitizing errors. Some of these errors can be excluded from the accuracy assessment for the proposed method by using filters. In short, it is best to remove bad data from the dataset prior to analysis so that the results are not prejudiced.

Finally, this study revealed that the proposed method of using GIS + NED to calculate surface lengths for road links provides accurate results. It can be readily used to calculate road lengths
for transportation databases with the currently available DEMs. This proposed method would not only improve the quality of transportation databases, it would also save money and time for state DOTs.

9. GLOBAL OBSERVATIONS AND IMPACT

Accurate mileages are critical for the effective operations of any state DOT. This study identified a path to take to acquire accurate mileages. These mileages measure the slope distance along the path of the roadway. Clearly, one way to obtain these mileages is through the use of DMI. However in those cases (and for states) where this is not practical, an alternative approach has emerged using DEMs.

A slope distance is calculated from an accurate elevation difference between two points and an accurate two dimensional planimetric distance between them. The accuracy of the slope distance is thus a function of the accuracy of these two linear and elevation distances.

The results of this study show that when you have accurate planimetric distances, elevations obtained from DEM data will give a degree of slope distance accuracy that is acceptable to most DOT applications. This tells us that we could immediately use current DEM data and straightforward algorithms to calculate accurate road mileages if we have accurate planimetric distances.

But there is even more good news. Geographers are continually conducting research to improve the accuracy of elevation models. Instrumentation is evolving that adds improved measurement accuracy to model accuracy. Thus, given that existing DEM data gives us good results, the future looks even better with respect to accuracy.

It must be noted, however, that the use of calculated distance, as opposed to driven distance, requires that all GIS linework be photo-revised. In doing so all lines will then be accurate to the same degree, and they will be accurate to the accuracy level of the 1:24,000 aerial photographs, thus introducing a significant amount of consistency into planimetric distance values. This will be very unlike the current situation where one distance value might be highly accurate and one might be highly inaccurate and it is difficult to have the degree of accuracy. Using DEM all the measurements will have the same degree of accuracy, thus improving user confidence in the distances.

This consistency is important. Still, it is somewhat fixed in time. No state is going to repeatedly photo-revise all of its roads as this is not practical. Thus, once roads are photo-revised the planimetric distances are unlikely to be changed very often. More accurate elevations, on the other hand, can be obtained whenever available, and new mileages can be calculated quickly using the algorithms mentioned herein. Accuracy of mileages can only get better as technology improves.

What does all of this tell us? Among other things that:

- We must get accurate, photo-revised and edited placement of lines in our GIS and CAD software;
• We can use DEM data to calculate roadway mileages rather than having to use DMI to obtain them;
• Even greater improvements in elevation accuracy are on the way and this will improve the performance of our future applications; and,
• That not to take these steps will inhibit our ability to automate in other areas and our ability to respond to customer data confidence needs will be diminished.

10. RECOMMENDATIONS

The analysis results of this case study for NCDOT lead to a number of recommendations. First, NCDOT can use the GIS/NED approach to obtain distance measurements for all road segments in North Carolina. This method provides enough accurate distance measurements for current applications and its cost, in terms of time, labor, and equipment, is less than using a DMI approach.

However, roads must be in their correct locations in the GIS roadway model before applying the GIS calculation algorithm. The accuracy of distance calculations from GIS depends directly and completely on the correct locations of roadway linework. The GIS Unit should consider expediting the matching and digitizing procedure, which is currently under way, to assure that all roads are correctly positioned prior to performing the distance calculations.

As mentioned earlier, the accuracy of GIS calculations depend on both the roadway network and on the elevation data. This case study did find that the current available elevation data (NED) is accurate enough for our purposes. However, there is no doubt that some errors from the elevation data did contribute to some inaccuracy in the results. A more accurate elevation data set (LIDAR -- Light Detection and Ranging) might provide more accurate distance calculations for road segments. Thus, a study using a sample portion of a LIDAR data set is proposed here to determine if the use of LIDAR might provide better results when compared with NED data. Then, it can also be determined whether LIDAR data should be used in place of NED data.

11. REFERENCES


