

Statistical Methods for the Evaluation and Monitoring of Traveler Information System Data
Quality

A Dissertation

Presented to
the faculty of the School of Engineering and Applied Science
University of Virginia

in partial fulfillment
of the requirements for the degree

Doctor of Philosophy

by

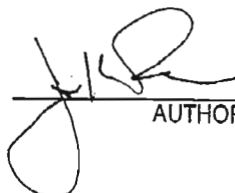
James K. Richardson

August

2012

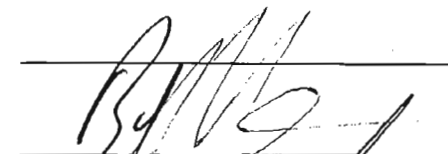
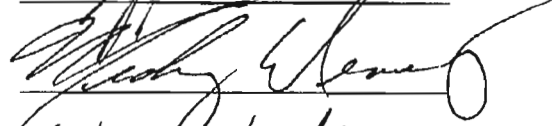
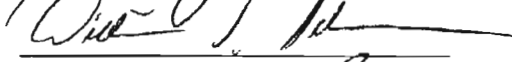
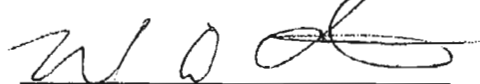
APPROVAL SHEET

The dissertation
is submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

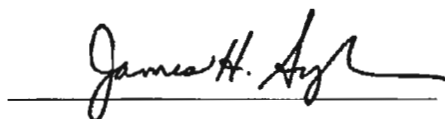

AUTHOR

The dissertation has been read and approved by the examining committee:


Advisor

Accepted for the School of Engineering and Applied Science:



Dean, School of Engineering and Applied Science

August
2012

Abstract

The size and complexity of modern transportation networks have created new challenges for the planning and operation of transportation infrastructure. Advancements in sensor technologies have greatly assisted civil engineers with this task. However, along with advances in sensor technologies come new challenges in understanding the application and scope of the data generated by these technologies. One of the most promising sensing technologies today can generate estimates of link travel time from probe-vehicle samples. These estimates of link travel time can be used for a number of important applications in transportation engineering such as performance measurement. However, these estimates of travel time rely on sample observations of traffic parameters from a variety of sources over large spatial and temporal extents. Therefore, data quality validation and monitoring of travel time estimates is an important task that civil engineers will require today and in the future. This dissertation presents an overview of the steps in a data quality evaluation and proposes new methods in three critical areas: link benchmark estimation, link selection, and data quality monitoring. The methods build on existing statistical techniques and apply these methods to the problem of Traveler Information System data quality evaluation and monitoring.

Acknowledgements

I would like to acknowledge key individuals who helped me with understanding the complexity of this problem and with acquiring important data for this research. First and foremost, I would like to acknowledge my advisor, Dr. Brian Smith for his patience and consistent mentorship. He has given me the opportunity to work on an interesting and challenging problem and helped me learn to think like a research scientist. I am sincerely grateful. Next, I would like to acknowledge Shawn Turner and Tony Voigt at the Texas Transportation Institute for opening doors and providing access to important data and research. I would also like to acknowledge my colleagues, Dr. Michael Fontaine, Simona Babiceanu, and Tim Ngov at the University of Virginia for assisting with data acquisition, mapping, and technical support. I would like to acknowledge the Virginia Department of Transportation and in particular the Virginia Center for Transportation Innovation and Research for its commitment to advanced research in Intelligent Transportation Systems. I am grateful for permission to use data from VDOT's Bluetooth sensors and their Inrix data archive. I would like to thank my dissertation committee for its guidance and input when I needed it. And finally, I would like to acknowledge my wife, Anna for her support and understanding while I completed this journey. Tak pige.

Table of Contents

Chapter 1 Introduction.....	2
1.1 Sensor Technology and Traffic Monitoring.....	4
1.2 Travel Time.....	5
1.3 Traveler Information Systems.....	6
1.4 Performance Measurement.....	9
1.5 Data Quality in Transportation Engineering.....	10
1.6 Research Motivation.....	11
1.7 Research Contributions.....	12
1.8 Outline of Dissertation.....	12
Chapter 2 Literature Review.....	14
2.1 Time Line of Relevant Research.....	14
2.2 Data Quality in Transportation Engineering.....	16
2.3 Traveler Information System Data Quality Evaluation Design.....	17
2.4 Developing Guidelines for Data Quality Evaluations.....	19
2.4.1 Benchmarking.....	20
2.4.2 Link Selection.....	21
2.4.3 Data Quality Estimation and Monitoring.....	21
2.5 Conclusions of Literature Review.....	22
Chapter 3 Travel Time Distribution and Variance.....	23
3.1 Distribution of Link Travel Time and Space-mean-speed.....	23
3.2 Mean of Travel Time.....	25
3.3 Space-mean-speed and Time-mean-speed.....	26

3.4 Variance of Travel Time and Space-mean-speed.....	27
3.5 Empirical Assessment of Factors Influencing Travel Time Distribution and Variance.....	28
3.6 Roadway Functional Classification.....	30
3.7 Minimum Sample Size.....	34
3.8 Roadway design factors affecting travel time variance	35
3.8.1 ADT per Lane.....	36
3.8.2 Access Point Density.....	36
3.8.3 Link Length.....	37
3.8.4 A Multivariate Regression Model of Roadway Factors on Travel Time Variance.....	37
3.9 Correlation between Travel Time Variance and Space-mean-speed.....	38
3.10 Spatial Correlation.....	40
3.11 Conclusions of Travel Time Distribution and Variance.....	43
Chapter 4 Travel Time Data Management and Processing.....	44
4.1 Basic Data Model for Managing Traveler Information Data.....	45
4.2 Travel Time Data Structure.....	45
4.2.1 Spatial Dimension of Data Structure.....	46
4.2.2 Temporal Dimension of Data Structure.....	46
4.2.3 Source Dimension of Data Structure.....	47
4.3 Spatial and Temporal Coordinate System.....	47
4.3.1 Spatial and Temporal Alignment.....	49
4.4 Conclusions of Travel Time Data Management and Processing.....	51
Chapter 5 Travel Time Benchmarking.....	53
5.1 Measuring Travel Time.....	53
5.1.1 Inferential Techniques for Measuring Travel Time.....	54

5.1.2 Direct Measurement of Travel Time.....	55
5.1.3 Probe Vehicle Monitoring.....	56
5.1.4 Automated Vehicle Identification.....	57
5.2 Estimation of Benchmark Travel Time.....	58
5.2.1 Local Averaging.....	60
5.2.2 Locally Weighted Regression (LOESS).....	63
5.2.3 Selecting the "best" value of Lambda in LOESS.....	66
5.3 Interval Estimation of Benchmark Travel Time.....	70
5.3.1 Interval Estimates from Local Averages.....	70
5.3.2 Interval Estimates from LOESS.....	71
5.4 Comparison of Local Averaging and LOESS.....	73
5.5 Residual Analysis.....	74
5.6 Potential Pitfalls with the LOESS Method.....	75
5.7 Conclusions of Travel Time Benchmarking Methods.....	77
Chapter 6 Link Selection by Maximum Entropy.....	78
6.1 Introduction to Maximum Entropy Sampling.....	78
6.2 Link Covariance.....	80
6.2.1 Computation of Covariance Matrix and Determinant.....	82
6.3 Evaluation of Maximum Entropy Sampling Method.....	84
6.3.1 Measuring the Performance of MES.....	85
6.3.2 Simulated Data Set Design.....	86
6.3.3 Experimental Design.....	88
6.3.4 Experimental Results.....	89
6.3.5 Discussion of Results from Simulated Data.....	91

6.4 Evaluation of MES using Empirical Data From Northern Virginia.....	92
6.4.1 Description of Study Area and Data.....	92
6.5 Results.....	94
6.6 Conclusions.....	96
Chapter 7 Proposed Benchmark Link Selection Method.....	98
7.1 Network Stratification.....	98
7.2 Selecting Candidate Segments.....	99
7.3 Maximizing Network Coverage.....	100
7.4 Selecting a Solution.....	100
7.5 Case Study in Northern Virginia.....	100
7.6 Conclusions of Proposed Benchmark Link Selection Method	104
Chapter 8 Measuring Traveler Information System Errors.....	105
8.1 Error Definitions.....	105
8.2 Selecting an Error Metric.....	105
8.2.1 A Simple Model of a Traveler Information System.....	106
8.2.2 Empirical Evaluation of Error Metrics.....	108
8.3 An Idealized Traveler Information System.....	111
8.4 Alternatives to Error Bias and Mean Absolute Error.....	112
8.5 Conclusions of Error Measurement.....	116
Chapter 9 Monitoring TIS Data Quality.....	117
9.1 Introduction to Data Quality Monitoring.....	117
9.2 A Data Quality Monitoring Information Architecture.....	119
9.2.1 Sensing Infrastructure.....	120
9.2.2 Road Network Meta-data.....	120

9.2.3 Traveler Information System.....	120
9.2.4 Analysis and Reporting Engine.....	121
9.3 Data Transmission and Processing.....	121
9.4 Statistical Methods of Quality Control.....	121
9.5 Process Monitoring.....	123
9.5.1 X-bar Charts.....	123
9.5.2 S-charts.....	125
9.5.3 Monitoring Errors by Location (Space) and by Time.....	127
9.6 Process Capability.....	130
9.7 Conclusions of Traveler Information System Data Quality Monitoring.....	132
Chapter 10 Conclusions and Contributions.....	134
10.1 Opportunities for Further Research.....	137

Index of Tables

Table 2.1: Previous TIS Data Quality Evaluations.....	19
Table 3.1: Results of Shapiro-Wilk Test of Normal Distribution on Houston Links.....	24
Table 3.2: Houston Arterial Segments.....	30
Table 3.3: Minimum Sample Size and CV.....	35
Table 3.4: Fitted Regression Model.....	38
Table 3.5: Description of Links Analyzed for Spatial Correlation.....	40
Table 5.1: Example of Binned Travel Time Observations.....	60
Table 5.2: RMSE (mph) Comparison.....	73
Table 5.3: Data Availability (% of intervals).....	74
Table 6.1: Description of Speed Profiles.....	86
Table 6.2: Description of Experiments.....	89
Table 6.3: Experimental Results.....	90
Table 6.4: Experiment E2: TIS Reported Speeds for Maximum and Minimum Entropy Solutions.....	91
Table 6.5: Bluetooth Link Descriptions.....	92
Table 6.6: Empirical Results.....	95
Table 7.1: Example of Solutions to Entropy and Coverage Decision Criteria.....	101
Table 8.1: Common Error Metrics.....	105
Table 9.1: Example of Process Capability Using Northern Virginia Data.....	132

Index of Figures

Figure 1: Illustration of Data Fusion Process.....	8
Figure 2: Distribution of CV on a Houston arterial roadway.....	31
Figure 3: Houston Freeway Links.....	32

Figure 4: Distribution of Travel Time CV in Houston.....	33
Figure 5: Relationship between Space-mean-speed and CV travel time for 3 links in Houston.....	39
Figure 6: Spatial Correlation Link 110-04177.....	41
Figure 7: Spatial Correlation Link 110+04152.....	42
Figure 8: Basic Data Structure of Travel Time Management System.....	45
Figure 9: Spatial and Temporal Coordinate System.....	48
Figure 10: Spatial Alignment.....	50
Figure 11: Spatial and Temporal Aggregation.....	51
Figure 12: Example of GPS points transmitted by trucks.....	57
Figure 13: Time Series observations of space-mean-speed.....	59
Figure 14: Sampled Data.....	61
Figure 15: Local Averaging at 5-minute intervals.....	63
Figure 16: LOESS functions fitted to sample data.....	66
Figure 17: 10 fold cross-validation MSE for a trial data set.....	68
Figure 18: 10-fold cross validation MSE repeated different data set.....	69
Figure 19: Local Averaging with Confidence Intervals.....	71
Figure 20: LOESS Confidence Intervals with 5-minute population mean.....	72
Figure 21: Quantile-quantile plot for a LOESS model.....	75
Figure 22: Example of LOESS with sparse data.....	76
Figure 23: Ground-truth Speed Profiles.....	87
Figure 24: Example of TIS and Ground Truth Speeds.....	88
Figure 25: Location of Bluetooth Links.....	93
Figure 26: Comparison of Empirical TIS and Bluetooth Speeds on Selected Links.....	94
Figure 27: Empirical Results: Subset $n = 5$	96

Figure 28: Link Selection -- Candidate Solution #1.....	102
Figure 29: Link Selection -- Candidate Solution #2.....	103
Figure 30: Link Selection -- Candidate Solution #3.....	104
Figure 31: Speed and Travel Time Errors.....	107
Figure 32: Example Plot of Empirical Observations.....	109
Figure 33: Empirical Comparison of Error Metrics.....	110
Figure 34: An idealized TIS model.....	112
Figure 35: Contour Plot of Relative Errors in Travel Time.....	113
Figure 36: Comparison of Models by Relative Errors in Travel Time.....	115
Figure 37: Traditional Quality Assurance.....	118
Figure 38: Data Quality Assurance.....	118
Figure 39: Example Architecture of a Data Quality Monitoring System.....	119
Figure 40: X-bar Chart for TIS Data Quality Monitoring.....	124
Figure 41: Example S-chart.....	126
Figure 42: Boxplots of Errors for 3 Days by Observed Segment.....	127
Figure 43: Observed Errors by Time for 3 Segments.....	128
Figure 44: Plot of TIS Estimates vs. LOESS fit.....	129
Figure 45: Process Capability versus Variability.....	131

List of Expressions

Expression	Description	Units (typical)
\bar{u}_s	space-mean-speed	mph
\bar{u}_T	time-mean-speed	mph
\bar{u}_s^*	TIS estimate of space-mean-speed	mph
\bar{T}	average travel time	seconds or minutes
\bar{T}^*	TIS estimate of average travel time	seconds or minutes
q	vehicle flow	vehicles per hour
k	density	vehicles per mile
μ	population mean	
σ^2, σ	population variance, standard deviation	
\bar{X}	sample mean	
s^2, s	sample variance, standard deviation	
CV	coefficient of variation	none
θ	a parameter or statistic	varies

Chapter 1 Introduction

Transportation infrastructure plays a critical role in modern society. The movement of people and goods is a fundamental need of all societies and one that transportation infrastructure facilitates. However, increasing demands for mobility and accessibility due to population growth, changes in land-use, structural changes in the economy, and other factors put pressure on existing infrastructure to satisfy mobility needs. In order to satisfy these needs and their many competing objectives it is imperative that transportation networks be planned, designed, and operated using a systemic, performance-based, data-driven approach.

One of the important roles of modern transportation engineers is to study the performance characteristics of the networks that they manage. Today, transportation engineers must collect, manage, and analyze data about the performance of the network and its key components. And, by understanding the performance of the transportation **system**, objective decisions can be made about how to best allocate resources to satisfy future objectives.

Until relatively recently, transportation infrastructure in the United States was designed and built as “stand-alone” infrastructure. While roads could facilitate vehicle movement and traffic signals could control movements at intersections, information about the performance of the infrastructure could only be acquired by labor-intensive field studies or surveys. This usually meant engineers were sent out to observe vehicles over some period of time and this data was used to help understand the performance of the infrastructure being studied. These field studies gave a very limited snapshot of performance at a few sites in the network.

Intelligent Transportation Systems (ITS) is the branch of transportation engineering concerned with the development and application of new sensing and information technologies used in

transportation engineering. By the mid-20th Century, new forms of electronic sensing and information technologies were being developed that would significantly change how transportation engineers collected data about network performance. ITS engineers helped introduce important sensor-based applications to improve traffic planning and operations.

Traveler Information Systems (TIS) are a more recent development within ITS that focus on the application of real-time and historical data to communicate the performance of links in the transportation network. Examples of Traveler Information System technologies include Dynamic Message Signs (DMS), 511 services, and color-coded traffic maps which can all be used to communicate to road users expected travel conditions. Other forms of TIS applications such as incident detection algorithms and Closed-caption TV (CCTV) can be used by transportation system operators to monitor system performance and respond to events.

Due to the size and complexity of modern transportation systems there is a need for technologies which can efficiently and automatically report important data about system performance. ITS is playing an increasingly larger role in the planning and operation of transportation systems. For example, ATRs are used today to collect traffic volume data on important links in the road network. Traffic Operations Centers (TOCs) use a variety of applications built on sensing technologies such as cameras and loop detectors to monitor and manage the transportation network. However, as with any sensing technology there are limitations to the quality of the data reported by the sensor.

There is a growing body of research in Civil Engineering examining the quality of data from sensors used in monitoring the transportation system. Much of this research has been focused on the quality of data reported by fixed point sensors such as inductive-loop detectors. However, new types of sensor data based on mobile wireless devices such as GPS, Cellular Phones, and Bluetooth devices are now emerging. It is critical that the quality of data from these systems be validated and monitored for it

to be used in transportation engineering applications. This dissertation is focused on the validation and monitoring of the quality of travel time estimates from Traveler Information Systems. The following sections present a background and overview of the problem.

1.1 Sensor Technology and Traffic Monitoring

Today, one of the most widely used sensing technologies in ITS is the inductive-loop detector. This is a type of magnetic coil that is installed beneath the pavement of a road. This type of sensor is capable of detecting the presence of a vehicle by registering the disturbance of an electro-magnetic field. A single-loop detector is capable of measuring flow (veh/time) and occupancy (% of time sensor is occupied). From these measurements, estimates of volume and density can be derived.

For decades, researchers have studied uninterrupted vehicular traffic flow. The fundamental equation of traffic flow relates link-speed to volume and density. The volume (q) is given in units of vehicles-per-time and density (k) is given in vehicles-per-distance. When these variables are known the average speed over the distance measured can be computed directly as:

$$\bar{u}_s = q/k \quad (1)$$

where \bar{u}_s (i.e. space-mean-speed) is specified in units of distance-per-time.

While inductive-loop technology can be used to estimate volume and density and therefore link-speeds, there are limitations to how widely loop-detectors can be deployed. There are also calibration and maintenance needs for each loop-detector. Thus, while the fixed-point sensor can measure important variables that traffic engineers need in order to understand network performance they are not easily deployed and maintained over very large networks.

1.2 Travel Time

One of the key performance indicators of many types of transportation systems is travel time. This is the time that it takes to complete a trip from one point in the network to any other point in the network. For example, the average time it takes for a vehicle to travel from one point in a road network to another point is an indication of how that part of the network is performing.

Travel time is related to other important traffic parameters. We can write the travel time of a single vehicle as t and the average travel time of a group of vehicles as \bar{T} . When the distance traveled, d , is known the average link-speed or space-mean-speed can be computed as $\bar{u}_s = d / \bar{T}$. And, as mentioned in the previous section, space-mean-speed can be related to other important parameters of network performance such as volume and density.

Travel time can also be used to measure reliability. Researchers have found that travel time reliability is a very important performance characteristic for users of a transportation network. Travel time reliability indicators summarize important distributional characteristics of travel time in a network. For example, the planning-time index and buffer-time index are two reliability measures used to characterize the reliability of a network path. Both of these indexes measure the reliability of a trip by relating an upper quantile (e.g. 95th percentile) of a travel time distribution to its median. Thus, these indexes require a large sample of travel times over a specified link to estimate the distribution.

In addition to reliability and traffic flow models, travel time is an important variable used in transportation planning models, signal optimization, signal coordination, and calibration of microscopic traffic models. In short, travel time has importance for road users, for transportation engineers, and for policy makers.

1.3 Traveler Information Systems

Traveler Information Systems (TIS) are systems that provide relevant travel-related information to users of the system. There are three key components of any TIS: sensing infrastructure, data processing systems, and communications infrastructure.

Sensing infrastructure may include fixed-point sensors such as inductive-loop detectors, side-fire radar, and video cameras. Sensing infrastructure also includes new types of mobile sensors that are becoming more widely available as inputs. For example, data from mobile GPS, cellular phones, and Bluetooth devices are now widely used in Traveler Information Systems. The data processing systems used in TIS include database management systems (DBMS) for managing and storing sensor data, data fusion engines, and specialized statistical software for processing sensor data. Finally, communications infrastructure includes the many ways in which traveler information is communicated to end users. For example, digital message signs (DMS) are used on freeways to broadcast estimated travel-time, color-coded maps are available on information websites to communicate the current state of monitored facilities, and raw data can be delivered to users via XML data feeds for further processing and analysis.

Early generations of TIS were mostly used to help inform the traveling public of road conditions. Services such as 511 combined input from sensor infrastructure with oversight from human operators to make available timely updates on traffic conditions in monitored regions. Other services with more specialized features such as trip-planning tools have been developed in metropolitan regions throughout the United States.

Today, Traveler Information Systems are being evaluated for use by transportation engineers as a potential new form of sensor data for use in planning and operations. As mentioned in the previous section, travel time is an important parameter used in many areas of transportation engineering.

Therefore, the availability of real-time and historical travel time on links throughout the transportation network is very desirable. Modern Traveler Information Systems now offer the ability to provide estimates of travel time on freeways and signalized arterials at very high spatial and temporal resolution. Essentially, transportation engineers can now obtain estimates of link travel time on virtually any road segment at any time.

The development of this new generation of TIS has largely been undertaken by private-sector technology firms that specialize in wireless communication and data processing. These proprietary systems can acquire sensor readings from large fleets of mobile sensors and fuse this data with sensor readings from fixed-point sensors in order to generate estimates of travel time.

The travel time estimation process is really just a different type of sensor. It is not a sensor in the traditional understanding but rather a *sensor-process* that may fuse together sensor observations from a number of different sensor types. Figure 1, below, illustrates the process of data fusion used to generate travel time estimates.

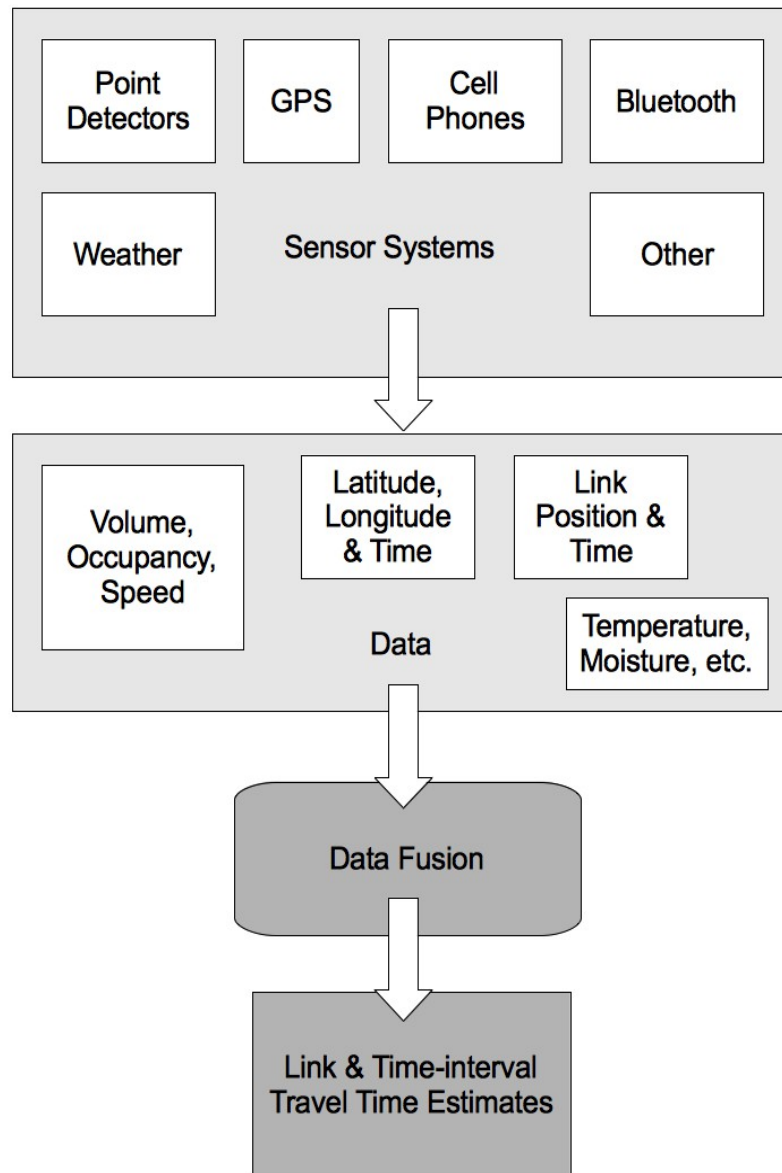


Figure 1: Illustration of Data Fusion Process

It can be seen that travel time estimates can be generated from a number of different sensor types. Each sensor type can generate different measurements of traffic conditions such as time-mean-speed, space-mean-speed, occupancy, volume, as well as environmental parameters and historical data.

All of these data types can be input into a data fusion process to generate estimates of travel time on a link during a time-interval.

Transportation agencies can therefore acquire massive amounts of travel-time data without the need to deploy or maintain any sensor infrastructure. However, there is still a need to evaluate and monitor the quality of the data.

1.4 Performance Measurement

Performance measurement has been identified as a critical need for the management of modern transportation infrastructure. A report on system operations performance measures for the Virginia Department of Transportation in 2007 identified travel time as a desired basis for future performance measures. However, the report notes that at the time of the writing travel time data was not yet available on a state-wide basis. The authors note that travel time based performance measures reflected information that was meaningful to both system operators and users.

A list of performance measures that use travel time as a variable

- Delay per Traveler
- Travel Time Index
- Buffer Index
- Planning Time Index
- Total Delay
- Percent of Congested Travel

The individual measures indicate performance of the system from the perspective of an

individual user. Area measures indicate aggregate performance over a geographical region. To illustrate how travel time is used in a performance measure, the expression for buffer index is given below in Equation 2.

$$\text{Buffer Index (\%)} = \frac{T_{95} - \bar{T}}{\bar{T}} * 100 \% \quad (2)$$

The Buffer Index (BI) gives a measurement of how reliable a trip is. It expresses the amount of extra buffer time needed to be on time for 95% of trips. For example, a BI of 10% indicates that a traveler should allow an extra 10% of time over the average trip time to arrive at the destination with 95% reliability.

1.5 Data Quality in Transportation Engineering

There has always been a need for data quality evaluation and monitoring in transportation engineering. All models make certain assumptions about the data used in the model. Thus, there is always a need to validate that the data used in a model is consistent with the assumptions of the model. Today, there is a need for validation and monitoring of travel time data quality because of the massive amounts of data now generated by these systems. As the amount of data grows, there is also a concern about how to monitor the quality of the data being collected.

The state of Virginia maintains approximately 57,000 miles of roadway of which approximately 9,200 miles are classified as interstate or primary roads. Today's Traveler Information Systems can provide estimates of travel time on nearly any link of this network at any time of the day. If travel time is estimated for every link in this system at frequent (e.g. every 5 - 10 minutes) intervals then there would be 100 or more estimates per link per day. This would result in a database of millions of estimates per year for the entire system.

Clearly, it would be impossible to validate every individual travel time estimate in such a system. Therefore, sampling strategies are needed that can provide robust estimates of data quality needed to validate and monitor the quality of the travel time estimates generated by the system.

1.6 Research Motivation

The motivation for this research comes out of the need to evaluate and monitor the quality of travel-time data from Traveler Information Systems. For example, a transportation agency may require that the average error of the travel time estimates be less than 5 mph. Thus, there is a need to evaluate the quality of the data. However, the way in which data-quality evaluations are designed and data-quality is measured is an open area of research. A number of different approaches to measuring travel-time data quality have been used in past evaluations. This research is therefore motivated by the need to provide a comprehensive framework for the evaluation of travel-time data quality.

A review of current practices in data-quality evaluations identified the following areas where further research was required:

1. Benchmark travel time estimation methods were based on sample averages over fixed time-windows. There is a need for further research into candidate methods for benchmark estimation.
2. The choice of where to collect benchmark data in the road-network is typically not driven by a quantitative or objective process. A more rigorous sampling design that satisfies an objective optimality criterion is required.
3. The assessment of data-quality did not follow best-practices established in other fields such as statistical process control. Typically, data quality evaluations used aggregations of observed errors in units of miles-per-hour to compute a point estimate of a data quality metric. However, the aggregation of observed errors is subject to uncertainty because of

sampling variation and process variation over time and space. Therefore, there is a need to develop a data-quality estimation process that can control for the various sources of uncertainty.

1.7 Research Contributions

The key contributions from this research are:

1. Empirical evaluation and insight into the distribution and variance of link travel time
2. Evaluation of existing benchmarking methods and development of a proposed smoothing method based on Locally Weighted Regression.
3. Development of a method to select links in a transportation network for benchmark data collection that uses a quantitative and objective selection criterion. The method is evaluated against other sampling designs and shown to provide optimal sampling properties.
4. A data-quality evaluation method that controls for both sampling and process variance. The method follows best practices developed in the field of statistical process control.

1.8 Outline of Dissertation

- Chapter 2 provides a literature review of relevant research for the dissertation.
- Chapter 3 covers the distribution and variance of travel time
- Chapter 4 develops a basic method for managing and processing travel time data necessary for data quality evaluations
- Chapter 5 investigates the measurement of travel time and the performance of benchmarking methods
- Chapter 6 introduces the concept of maximum entropy sampling for the problem of link

selection in a data quality evaluation

- Chapter 7 applies the maximum entropy sampling method developed in Chapter 6 to a network in Virginia
- Chapter 8 examines the role of error measurement in travel time data quality evaluations and examines the performance of different error measurements.
- Chapter 9 provides a background on the application of statistical quality control techniques to the problem of Traveler Information System data quality evaluations.
- Chapter 10 discusses the research contributions of the dissertation

Chapter 2 Literature Review

The previous chapter described the background and need for this research. In this chapter a review of the literature on traffic monitoring and TIS data quality assessments is presented. The purpose of the literature review is to provide a context and background for the research presented in this dissertation.

2.1 Time Line of Relevant Research

This timeline is provided to give the reader an historical perspective of the evolution of relevant research to the problem of data quality evaluations of Traveler Information Systems. The timeline provides a general overview of research in traffic monitoring, sensor data, and how the field has developed in the past twenty years.

1995 - 2000

During this time period and well into the 2000's, researchers were looking at questions dealing with loop detector accuracy, calibration, and deployment. There was also an emphasis on estimation and prediction of traffic conditions from point sensor readings. Algorithms to estimate travel time from point sensor readings were developed.

Management of sensor data is also an active area of research. Archived Data Management Systems (ADMS) are developed and adopted by state DOTs for managing real-time and historical data from loop detectors.

2000 - 2005

Early prototypes of Traveler Information Systems are developed. Rapid growth in cell phone usage leads to innovation in probe vehicle traffic monitoring. Travel time

estimation from wireless location technology (WLT) is evaluated. The results indicate that the early generation of WLT is not able to estimate travel time or traffic conditions with sufficient accuracy for deployment in traffic management applications.

2005 - 2010

Proliferation of wireless devices continues and the widespread adoption of GPS for freight logistics becomes a new potential source of probe data. During this period of time other wireless communication protocols such as RFID and Bluetooth are shown to be useful technologies for automated vehicle reidentification.

Private-sector technology companies begin to seriously compete for contracts with DOTs and the federal government to provide real-time and historical estimates of link travel time. Procurement of data from these companies requires validation of the data quality. Data quality evaluations are conducted at numerous test sites throughout the United States and globally.

Data management and processing infrastructure such as ADMS were developed to manage a specific type of data from a specific type of sensor. These systems must now be adapted to deal with non-traditional sensor data and the possibility of new types of data that had not yet been considered. The Regional Integrated Transportation Information System (RITIS) is developed at the University of Maryland in coordination with the regional transportation authorities of the metropolitan Washington, DC area.

2010 - Present

Private sector Traveler Information Systems such as Inrix Corporation are widely adopted by state DOTs for display on DMS signs on major freeways and for

transportation performance measurement. The Texas Transportation Institute (TTI) adopts Inrix as its source of link speed estimates for the Urban Mobility Report which focuses on congestion measurement in urban areas in the United States. The I-95 Corridor Coalition continues its ongoing validation of Inrix in collaboration with the University of Maryland.

2.2 Data Quality in Transportation Engineering

As the number of different data sources grows and the volume of data generated by these sources grows, methods for measuring data quality have become a real need for transportation engineers. Data quality has been defined by various authors[1], [2] to be a multi-dimensional measure that includes the following dimensions:

1. Accuracy
2. Confidence
3. Availability
4. Coverage

As more and more "information products" become available, data quality is also being evaluated using statistical sampling methods. Quality control researcher, Joesph Juran, defined data quality to be:

Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features. [2]

A formalized framework for statistical quality control was developed from the work of researchers such as Juran and Deming. This framework, commonly known as Total Quality

Management (TQM) emphasizes integrating quality control methods throughout the production and supply chain of manufacturing companies. Along with the Six Sigma program, these quality control programs have been widely adopted by businesses as a critical tool to remain competitive.

In 1992, the Total Data Quality Management (TDQM) program was formally launched at MIT. This program was created to provide a formal framework for research into data quality.[3] The framework extends TQM to the domain of data quality. It outlines a cyclical approach to continuous quality improvement as:

Define, Measure, Analyze, Improve[3]

Researchers have applied the methods of statistical quality control to a wide range of problems in data quality. For example, the application of data quality methods to sensor data has been investigated by a number of researchers. Methods for automated processing, and quality assurance of data from weather stations were developed by Williams, et. al.[4] Other researchers have proposed the application of machine learning methods such as Bayesian networks to automate data quality assessment of large historical databases of sensor data.[5]

An application of automated statistical quality control methods to a problem more specific to transportation engineering was investigated by Turochy and Smith. They looked at the problem of monitoring the quality of data reported by loop detectors which are a primary source of sensor data on many freeways in the United States. Because detectors report both volume and occupancy, they developed a multi-variate quality control model. The proposed model used a statistical test based on Hotelling's T^2 to identify points outside of a multi-dimensional ellipsoid.[6]

2.3 Traveler Information System Data Quality Evaluation Design

Between the mid-2000's and the present day there have been numerous TIS data quality

evaluations conducted in the United States and globally. While most evaluations shared the overarching goal of measuring the "quality" of the data provided by the TIS, the design of each evaluation differed in key areas. For example, the way in which benchmark observations were measured and processed and the techniques used to estimate the benchmark values differed among evaluations. To some extent these differences can be explained by the different sensing technologies used in collecting the benchmark observations. On the other hand, there were also differences in how the data was treated after measurement.

Evaluations often did not report a specific methodology or decision criterion for link selection. Most frequently, evaluators selected links that were known to be congested from local expert knowledge or of interest for some similar reason. Link selections were simply stated in the reports and not explicitly justified by any particular selection methodology or criteria.

Measurement of errors varied considerably between evaluations. In all cases, errors were measured as the distance between the benchmark value and the TIS estimate. However, aggregation of errors varied. The most commonly used aggregation of errors were the Mean Absolute Error (MAE) and the Error Bias. In some cases, Mean Square Error (MSE) was used and in other cases statistical tests such as a paired t-test were used. In all cases, the aggregated errors were reported as fixed statistics with no estimates of standard error for the reported error estimates.

An overview of some of the TIS data quality evaluations conducted in the United States in the last 10 years is provided below in Table 2.1.

Service Provider	Location	Evaluator	Benchmark Technique	Accuracy Criteria	Link selection criteria
Dayton, OH[7]	Dayton, OH	ODOT, et.al.	Spot-speed from radar, Bluetooth re-id, floating car	MAE,	Expert opinion
Inrix[8]	I-95 Corridor	University of Maryland	Bluetooth	MAE, Speed error bias	Expert opinion
AirSage[9]	Virginia	University of Virginia	Floating car, CCTV and loop detectors	MAE, Speed error bias	Expert opinion
CellInt	Atlanta, GA	URS, Georgia DOT	Floating car and point sensors	Paired t-test of means	Expert opinion

Table 2.1: Previous TIS Data Quality Evaluations

2.4 Developing Guidelines for Data Quality Evaluations

In 2009 a pooled-fund study[10] was initiated to study "best practices" for travel time data quality evaluations. The study examined how previous evaluations had approached the problem, examined existing technologies, and considered statistical techniques. The study was sponsored by and received input from a diverse group of stakeholders including a number of state DOTs, FHWA, and the North-American Traffic Working Group (NATWG) which included representatives from private-sector firms involved in developing relevant technologies. Out of this study the following outline of a data quality evaluation was designed.

1. Select Links and Time-intervals
 - a) Decide which links will be evaluated
 - b) Determine how frequently to collect travel time observations
2. Select a benchmark sampling technique

3. Collect benchmark data for each link and time-interval
4. Compare TIS estimate with benchmark and compute an error
5. Aggregate measured errors
6. Determine data quality from error observations

The final report of the pooled-fund study offered only a set of standardized guidelines that DOTs could use to evaluate Traveler Information Systems. While the report provided agencies with an overview of statistical tools that could be used in a data quality evaluation, the report did not fully investigate the wide range of statistical methods that can be applied to the problem. The three main areas of research left open by the pooled fund study include:

1. Benchmarking techniques
2. Link selection
3. Data quality evaluation and monitoring

2.4.1 Benchmarking

The literature on benchmarking in Traveler Information System data quality evaluations has been focused to a great extent on questions of sample size. Turner and Holdener investigated minimum sample sizes for TIS evaluations based on the 85th percentile of the distribution of travel time Coefficient of Variation (CV). They examined travel time samples from Houston's toll-tag data set to suggest a minimum sample size for probe-based information systems.[11]

Green, Smith, et.al. also examined sample size requirements for probe-based data sources. Specifically, they looked at sample size requirements for estimates of space-mean-speed on links with non-normal speed distributions. They found that good approximations of point estimates based on the

normal distribution could be justified for non-normal distributions by the Central Limit Theorem.[12]

However, few researchers have examined the problem of benchmark estimation from the perspective of time-series data. Rather, the approach most often used is to view each time-interval as a unique population and estimate the mean for that population based on the sample observations.

2.4.2 Link Selection

The selection of links for benchmarking in a data quality evaluation is a critical decision in the evaluation process. Yet, there exists little research to support this critical need. For probe-based monitoring systems, Tanikella investigated a sampling design using stratification.[13] The network was stratified by the expected Average Daily Traffic of a link for estimating average link speed. The method was evaluated using data from a simulated network in Virginia. The method showed improved accuracy and coverage of estimates based on stratified samples of probe vehicles. However, the method was designed for estimating link speeds from samples of probe vehicles and tested against a simulated network. Traveler Information Systems provide a rich historical database of link speeds that can be used to develop a sampling method better focused on data quality evaluation.

2.4.3 Data Quality Estimation and Monitoring

The I-95 Corridor Coalition's evaluation of Inrix is currently one of the largest evaluations of a Traveler Information System in both spatial and temporal scope. The evaluation has examined the accuracy of travel time estimates on links in several states along the I-95 corridor over a time span of several years. The evaluation reports average measures of error bias and absolute speed error binned by the observed average speed for each link that is monitored.[8] The terms of the contract state that a +/- 5 mph average error is acceptable within each of the speed bins. The bins are defined as below 30 mph, 30-45 mph, 45-60 mph, and above 60 mph. An error is measured as the distance from the bounds of a

95% confidence interval of the mean for the current time-interval. These errors are aggregated and reported as error bias and absolute speed error in the defined speed bins.

2.5 Conclusions of Literature Review

The literature review indicates that while data quality monitoring is an active area of research within civil engineering, there are still open questions around the design of specific methods such as benchmarking, link selection, and data quality evaluation. The background research in data quality methods indicates that statistical methods developed in quality control have the potential to be successfully applied to this problem domain. The literature review also indicates that the three specific subject areas of benchmarking, link selection, and quality evaluation require further methodological development.

Chapter 3 Travel Time Distribution and Variance

The design of an effective travel time data quality evaluation framework requires an understanding of the distribution and variance in link travel time. To investigate this, one year of travel time observations on over 100 links in the Houston metropolitan region were analyzed. The analysis was conducted to obtain an understanding of the distribution of travel time and its variance.

3.1 Distribution of Link Travel Time and Space-mean-speed

Link travel time is defined as the time to travel across a specified link. Traveler Information Systems estimate the average travel time of a vehicle population across a specified link during a specified time-interval. One should distinguish between the distribution of travel time of individual vehicles across a link and the distribution of average travel time or space-mean-speed. In data quality evaluations it is the latter distribution that is of more practical importance.

Average link travel time is typically modeled as a normally distributed random variable. Modeling average link travel time as a normal random variable can be justified by the Central Limit Theorem. The distribution of the sample mean will converge on a normal distribution given enough sample draws from any distribution. When sampling from a normal distribution, the distribution of the sample mean can be assumed to be normally distributed. When sampling from any other distribution, the distribution of the sample mean will approximate a normal distribution as the sample size increases. Thus, the sample mean can generally be considered to be normally distributed under a wide range of conditions.

Empirical data do give some support for this assumption. A random sample of travel time and space-mean-speed observations from 20 freeway links in Houston was analyzed at 5-minute and 15-

minute intervals over 3 days. A Shapiro-Wilk test of the normal distribution was performed on each 5-minute and 15-minute sample of travel time and space-mean-speed observations. The null hypothesis of the test is that the sample is taken from a normal distribution.[14] The null hypothesis was rejected if the p-value was less than .05 (i.e. 95%). There were 24,215 samples at 5-minute intervals and 8,405 samples at 15-minute intervals.

The results indicate that the distribution of travel time and speed is influenced by the length of the time-interval of the observation and the average speed during that time-interval. For example, at a 5-minute interval when the average speed was between 0 and 30 mph, the distribution of observed travel times was found to be normal in about 45% of the cases. When speeds were greater than 60 mph the distribution was found to be normal in about 75% of the cases. The full results of the tests are shown below in Table 3.1.

	Percentage of Samples Identified as Normally Distributed			
	Speeds		Travel Times	
Average Speed	5-minutes	15-minutes	5-minutes	15-minutes
0 - 30 mph	16%	2%	45%	18%
30 - 45 mph	38%	7%	40%	12%
45 - 60 mph	66%	50%	64%	48%
60+ mph	77%	65%	75%	62%

Table 3.1: Results of Shapiro-Wilk Test of Normal Distribution on Houston Links

These results indicate that at shorter time-intervals the population of vehicle speeds or travel times is more likely to be normally distributed than at longer time-intervals. The practical implication of this is that when collecting sample observations of travel time on a link the shorter the time-interval and the closer to free-flow speed the more likely the distribution will be close to a normal distribution.

However, the sample size also plays a role in the distribution of the sample mean. While the data indicate that in congested traffic flow, the distribution of vehicle travel time is less likely to be

normally distributed, the data do show that the sample size tends to increase as speeds approach the congested state. For example, the average sample size for a 5-minute sample when the average speed was between 0 - 30 mph was 36 observations whereas the average sample size when speeds were greater than 60 mph was 19 observations. This indicates that while congested traffic flow is less likely to be normally distributed, the sample sizes are likely to be larger than in free flow traffic. This can be understood within the context of traffic flow theory from the relationship between speed, volume, and density. In congested traffic flow, speeds decrease while density increases and to a limited extent volumes increase. These factors contribute to greater sampling rates.

To summarize these findings, the Central Limit Theorem states that the distribution of the sample mean will converge on a normal distribution when sampling from any population. When sampling from a normal distribution the sample mean will be normally distributed. When sampling from other distributions the sample mean will approximate a normal distribution as the sample size increases. Empirical data support this assumption and indicate that in free-flow traffic a normal distribution in speeds or travel times is likely. In congested traffic flow the normal distribution is less likely, however, as sample sizes increase the normal distribution becomes a reasonable approximation for the distribution of the sample mean.

3.2 Mean of Travel Time

The average travel time or average speed over a link (i.e. space-mean-speed) is a fundamental statistic used in all data quality evaluations. It is a measure of the central location of the distribution of travel time or speed on a link during a specified time-interval.

The arithmetic mean of link travel time is given as:

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n t_i \quad (3)$$

where \bar{T} is the mean travel time, t_i is the travel time of the i^{th} vehicle and n is the number of vehicles

The mean of link travel time is directly related to the distance traveled. In general, the longer the link the greater the average travel time. Therefore, the space-mean-speed, which is the average rate or speed of vehicles over a link is often used in place of the average travel time because it can be compared among links of differing lengths.

3.3 Space-mean-speed and Time-mean-speed

The space-mean-speed of a link is easily computed from the average travel time of a link if the length of the link is known. The space-mean-speed is also the harmonic mean of vehicle speeds. The harmonic mean is one of three types of means which include the arithmetic and geometric means. The time-mean-speed of a link is the arithmetic mean of speeds over a link. The harmonic mean, on the other hand is an average of rates and is used when averaging the observed speeds of vehicles on a link or when converting an arithmetic mean of travel time to speed. In general, the space-mean-speed is used to describe the "average speed" over a link.

The space-mean-speed of a link can be calculated in one of two ways. First, the inverse of the arithmetic mean of travel time can be multiplied by the distance traveled as:

$$\bar{u}_s = \frac{d}{\bar{T}} \quad (4)$$

where d is the distance traveled or link length, and \bar{T} is the arithmetic mean of travel time

Space-mean-speed can also be calculated as a harmonic mean of speeds. This is given as:

$$\bar{u}_s = n \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{s_i} \right)^{-1} \quad (5)$$

where s_i is the i^{th} average speed of the vehicles, and n is the number of vehicles

In practice, it is often simpler to compute the average travel time and convert this value to space-mean-speed. However, there are situations where the link length may not be known and the average speeds are the only observations available. In these cases, a harmonic mean would be the appropriate method to calculate the space-mean-speed for a link.

The relationship between space-mean-speed and time-mean-speed has been investigated in the literature. Using Wardrop's relationship between time-mean-speed and space-mean-speed[15] Rakha and Zhang showed that space-mean-speed will always be greater than time-mean-speed (arithmetic mean of speeds) by a factor related to the variance in speeds.[16] Specifically, they noted that:

$$\bar{u}_s \approx \bar{u}_T - \frac{\sigma_T^2}{\bar{u}_T} \quad (6)$$

Thus, when the variance in speeds is relatively small compared to the average speed the space-mean-speed and time-mean-speed will be close in magnitude. This is analogous to a situation such as free flow traffic where the average vehicle speed on a freeway link may be 65 mph. Yet, when there is a large variance in speeds relative to the average vehicle speed such as during the breakdown in traffic flow from a free-flow state to a congested state, the difference between space-mean-speed and time-mean-speed may be significant.

3.4 Variance of Travel Time and Space-mean-speed

The variance (σ^2) of link travel time and space-mean-speed is another key statistic used in a data quality evaluation. In general, the magnitude of the variance of travel time is of less interest because it is directly related to the distance traveled. For example, the variance in travel time of a 1-mile link will generally be much smaller than the variance in travel time of a 10-mile link. For this

reason, the co-efficient of variation (CV) is a more effective measure of link travel time variation. The CV of link travel time is the ratio of the standard deviation to the mean and is given as:

$$CV = \frac{\sigma}{\mu} \quad (7)$$

where σ is the standard deviation, and μ the mean .

Typically, the standard deviation and mean are sample statistics and the CV is therefore a random variable. However, point estimates of the CV from sample statistics are frequently used in the scientific literature and the sampling distribution of CV is not considered.

The variance of space-mean-speed was also explored by Rakha and Zhang. They found that the variance in space-mean-speed was typically larger than the variance in time-mean-speed by a factor related to the variance in time-mean-speed. They formulated the following relationship:

$$\sigma_s^2 = \sigma_T^2 + \left(\frac{\sigma_T^2}{\bar{u}_T} \right)^2 \quad (8)$$

Thus, the divergence in variance between space-mean-speed and time-mean-speed is related to the variance in time-mean-speed. Again, this can be related to the state of traffic flow. When there is a large variance in time-mean-speed relative to the average speed we can expect the variance in space-mean-speed to be greater than the variance in time-mean-speed.

These insights are important for data quality evaluations. The design of an effective data quality evaluation requires an understanding of the variance of travel time and space-mean-speed. Link selection, benchmarking, and data quality assessment depend on sampling and inference which require an understanding of the sampling and process variance.

3.5 Empirical Assessment of Factors Influencing Travel Time Distribution and Variance

An empirical assessment was conducted to provide greater understanding of the factors

influencing the distribution and variance of link travel time. The empirical assessment examined observations of travel time from Houston, TX. Observations of link travel time were aggregated at 5-minute time-intervals and the following statistics were computed for all links:

1. Average travel time and space-mean-speed
2. Variance in travel time and variance in space-mean-speed
3. Coefficient of Variation of travel time
4. Sample size
5. Confidence interval of the sample mean

The following insights were found after examining this data set.

1. The distribution of travel time on signalized arterials is significantly different from the distribution of travel time on limited access freeways.
2. The 90th percentile CV of travel time on freeways was approximately 0.10. This indicates that 90% of the observed time-intervals had a standard deviation that was less than 10% of the sample mean.
3. Median CV values of travel time on arterials was approximately 0.3
4. The minimum sample size for average freeway travel time estimation is approximately 7 observations when the CV in travel time is 0.10 or less.
5. Of the 10% of time-intervals from the Houston data set with a travel time CV greater than 0.10 there was an indication of a negative correlation between travel time variance and space-mean-speed
6. The links where higher travel time variance was most likely to occur were links that had high Average Daily Traffic (ADT) counts and links with a higher density of access points (on and off

ramps).

7. The length of the link was also found to be a factor correlated with travel time variance.

Generally, the longer the link length the smaller the variance in travel time.

These insights will be discussed in more detail in the following sections.

3.6 Roadway Functional Classification

The investigation examined roadway functional classification as a factor related to travel time variance. Two functional classes of roadway were evaluated:

1. Signalized arterial roadways
2. Limited access freeways

The signalized arterial travel times were measured on two arterial segments in Houston over 5 days in 2011. Table 3.2, below describes the arterial roads where travel times were observed.

Segment	Directions	Length	Signal Density	Land Use
Gessner btw. Westheimer & Briar Forest	NB, SB	0.7 miles	1 / mile	Commercial, Residential
Briar Forest bw. Dairy Ashford & Eldridge	EB, WB	1.1 mile	1 / mile	Commercial, residential

Table 3.2: Houston Arterial Segments

Travel time observations on these road segments were collected by Bluetooth re-identification. The segments were both approximately 1 mile in length with low signal density (approximately 1 signal per mile). The land use included significant residential and commercial development. The median CV travel time for each directional segment was approximately 0.3. The distribution of CV when observed at 15-minute intervals indicated a left-skewed distribution with a fairly high density of

time-intervals with high CV travel time. The distribution of CV at 15-minute intervals over the 5-day observation period for Gessner NB is shown below in Figure 2.

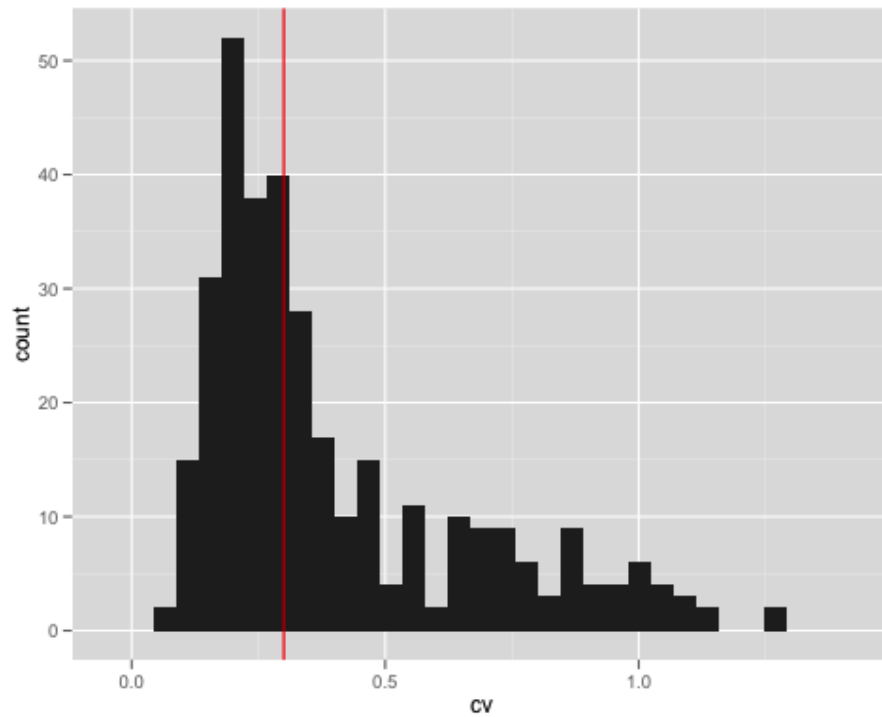


Figure 2: Distribution of CV on a Houston arterial roadway

The freeway data was acquired by toll-tag AVI observation on the freeway system throughout the Houston metropolitan region. The database contains approximately 274 million individual observations of travel time on various links throughout the Houston freeway system. An overview of the system is shown below in Figure 3.



Figure 3: Houston Freeway Links

The travel time CV was computed at 5-minute intervals for links where the sample size of the sample was greater than the following statistical threshold:

$$n \geq \left[\frac{t_{\alpha/2, n-1} \frac{s}{\bar{X}}}{e} \right]^2 \quad (9)$$

where $t_{\alpha/2, n-1}$ is the Student T statistic, and e is the error tolerance as a percentage

The error tolerance, e , was set at 10% or 0.10. Using this criterion, samples where the variance was too large to estimate the mean within +/- 10% error were rejected. In total there were approximately 14 million 5-minute samples across 117 links that were analyzed. The distribution of travel time CV for the links is shown below in Figure 4.

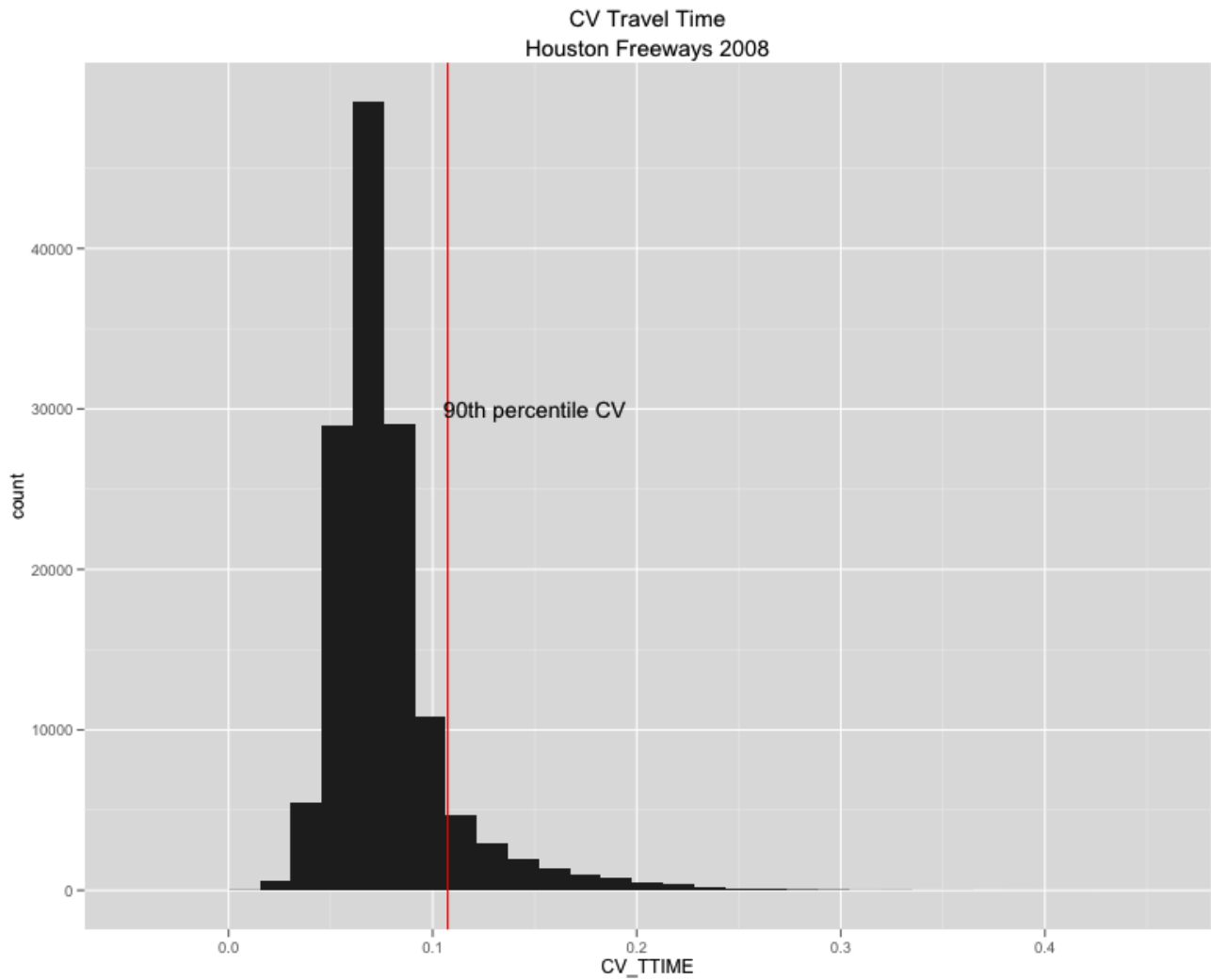


Figure 4: Distribution of Travel Time CV in Houston

The 90th percentile CV for this data set was approximately 0.11. This indicates that in approximately 90% of the observed 5-minute intervals the standard deviation of the travel time on that link during that time-interval was 10% or less of the average travel time.

The importance of this finding is that only a relatively small percentage of the observed time-intervals had a standard deviation in travel time greater than 10% of the mean. From a sampling perspective this is an important insight because it provides an estimate of the number of observations required for statistical inference of the population mean.

3.7 Minimum Sample Size

The number of sample observations required to estimate the population mean is a function of the sample mean, variance, and error tolerance. When the population is approximately normally distributed, the required sample size is given as in Equation 9 . Thus, if we assume that the CV of travel time for a given link at a 5-minute interval will be approximately 10% or less we can compute an upper bound on the required sample size for statistical inference. If the error tolerance is set to be 10% the approximate minimum sample size can be estimated through an iterative procedure.

The Student T statistic is parameterized by α and df which are the significance threshold of the test (e.g. 95%) and the degrees of freedom. In the univariate case this is simply the number of observations minus one. The minimum sample size must be computed iteratively for different levels of variation in the samples. The process is described for one example.

If the CV is 0.05 and the error tolerance is 10% we can compute the minimum sample size, n_{\min} , by plugging in different values of n into Equation 9 and comparing this value to the current value of n . When $n = 3$, $\alpha = .05$, the statistic is 4.3 and the computed value of n_{\min} is 4.6. Since this value is greater than the current value of n , we repeat with the next largest value of n . At $n = 4$ the computed value of n_{\min} is 2.53 which is less than the current value of n and therefore the minimum

sample size for a $CV = .05$ and an error tolerance of 10% is 4 observations.

Table 3.3 gives an approximate lower bound on the minimum sample size for different levels of CV.

CV	Student T statistic	n_min
0.05	3.18	4
0.1	2.45	7
0.15	2.01	12

Table 3.3: Minimum Sample Size and CV

From this table it can be seen that for any time-intervals where the CV in travel time is approximately 0.10 or less the required number of sample observations to estimate the population mean with 10% error tolerance is 7 observations or more.

A typical freeway segment in a metropolitan area may carry on average 15,000 - 20,000 vehicles per lane per day (veh/lane/day). Lane capacity of freeway travel lanes has been estimated to be approximately 2,000 vehicles per hour. Therefore, as lane volumes approach capacity there are approximately 167 vehicles per lane per 5-minute interval. If a sampling technology is capable of identifying only 5% of this volume then the number of identified vehicles would be approximately 8 vehicles per time-interval per lane. A freeway segment will typically have 3-4 lanes and therefore the sample size for a heavily traveled freeway segment may approach 20 or more observations. Based on the calculations of minimum sample size given in Table 3.3 this would be more than enough observations to estimate the population mean from the sampled data.

3.8 Roadway design factors affecting travel time variance

The following candidate factors were identified through a literature review and through collaboration with experts in the field.

1. Average Daily Traffic per Lane (ADT / Lane)
2. Access Point Density
3. Segment Length

Each of these factors will be defined and discussed in detail below.

3.8.1 ADT per Lane

ADT is the Average Daily Traffic for a specified road segment and is an important statistic used in a number of different areas of transportation engineering such as planning, operations, and safety. The effect of ADT on travel time variance was studied using the data from Houston. In order to account for different link designs the ADT value was normalized by the number of lanes.

The data from Houston, however, indicates that ADT per lane as a factor alone does not have a very strong correlation with variance in travel time. Random samples of travel time observations in 5-minute intervals were selected from the Houston data set. The correlation between ADT per lane and travel time CV was approximately 0.005.

3.8.2 Access Point Density

On a limited access freeway the entry and exit points to the freeway are the on and off ramps. These access points are a known contributor to variations in the flow of traffic on freeways. As vehicles enter and leave the traffic stream they must accelerate or decelerate to join or leave the flow. At a microscopic level this will affect the behavior of other vehicles in the stream. These perturbations contribute to variation in the flow at a macroscopic level.

An analysis of data from Houston showed that as the number of access points per unit distance

increases the relative variation in travel time also increases. For a random sample of data from the Houston data-set the correlation between travel time CV and access point density was approximately 0.15.

3.8.3 *Link Length*

Link length is the length over which travel time is measured. The definition of a link is somewhat arbitrary in that it can be defined to be the distance between two points of interest in the network. In this sense, link length is not a design feature, but it is a factor that affects measured variation in travel time.

Link length is also a factor that is controllable in the sampling design of a travel time data quality evaluation. For example, when selecting the links for benchmark data collection, the link length can be considered as a factor that will affect travel time variation. By understanding the contribution of link length to travel time variation, sampling designs can better control for this source of variation in the design.

Empirical data indicate that as the distance increases the relative variation in travel time decreases. The correlation between link length and CV travel time for a random sample of observations from Houston was -0.13.

3.8.4 *A Multivariate Regression Model of Roadway Factors on Travel Time Variance*

Since, by themselves, none of the explored roadway design factors are strong predictors of travel time variance, a multivariate linear model of CV travel time was evaluated. The model was specified as:

$$CV_{time} = \alpha + \beta_1 \text{ADT}/\ln + \beta_2 \text{access-dens} + \beta_3 \text{link-leng} + \epsilon$$

The model was fit to random sample of observations from the Houston data set ($n = 54,504$).

The resulting fitted model is shown below in Table 3.4.

	Estimate	Significance
Intercept	0.07821	Yes
ADT per lane	-0.00001	Yes
Length	-0.00254	Yes
Access points per mile	0.00494	Yes

Table 3.4: Fitted Regression Model

The adjusted R-square for the model was 0.036. The intercept of the model indicates that the expected CV in travel time is around 0.08. However, due to the extremely small magnitude of the coefficient estimates and the small R-square value for the model, the factors do not appear to explain a great deal of variance in travel time based on the observations from Houston.

3.9 Correlation between Travel Time Variance and Space-mean-speed

In the Houston data there was evidence of a negative correlation between travel time variance and space-mean-speed. The negative correlation can be understood in the context of traffic engineering as meaning that on freeway links as congestion forms the variance in travel time can be expected to increase. Higher variance in travel times leads to a larger minimum sample size required for estimating the population mean travel time. Therefore understanding the conditions which leads to higher variance in travel time is critical to the design of an effective data quality evaluation framework.

The 5-minute travel time data from Houston was analyzed to assess the correlation between space-mean-speed and travel time variance. To assess the relationship, travel time CV was used as a surrogate for travel time variance. As discussed earlier, the travel time CV is a non-dimensional

measure of relative variance and is used as a surrogate for link travel time variance because its magnitude is independent of link length.

The correlation between travel time CV and space-mean-speed for 1-month of data ($n=727,845$) from Houston freeways was found to be -0.25. This indicates that as average freeway speeds decrease from free-flow the variance in travel time is likely to increase. Figure 5, below, illustrates this relationship for three links in the Houston data set.

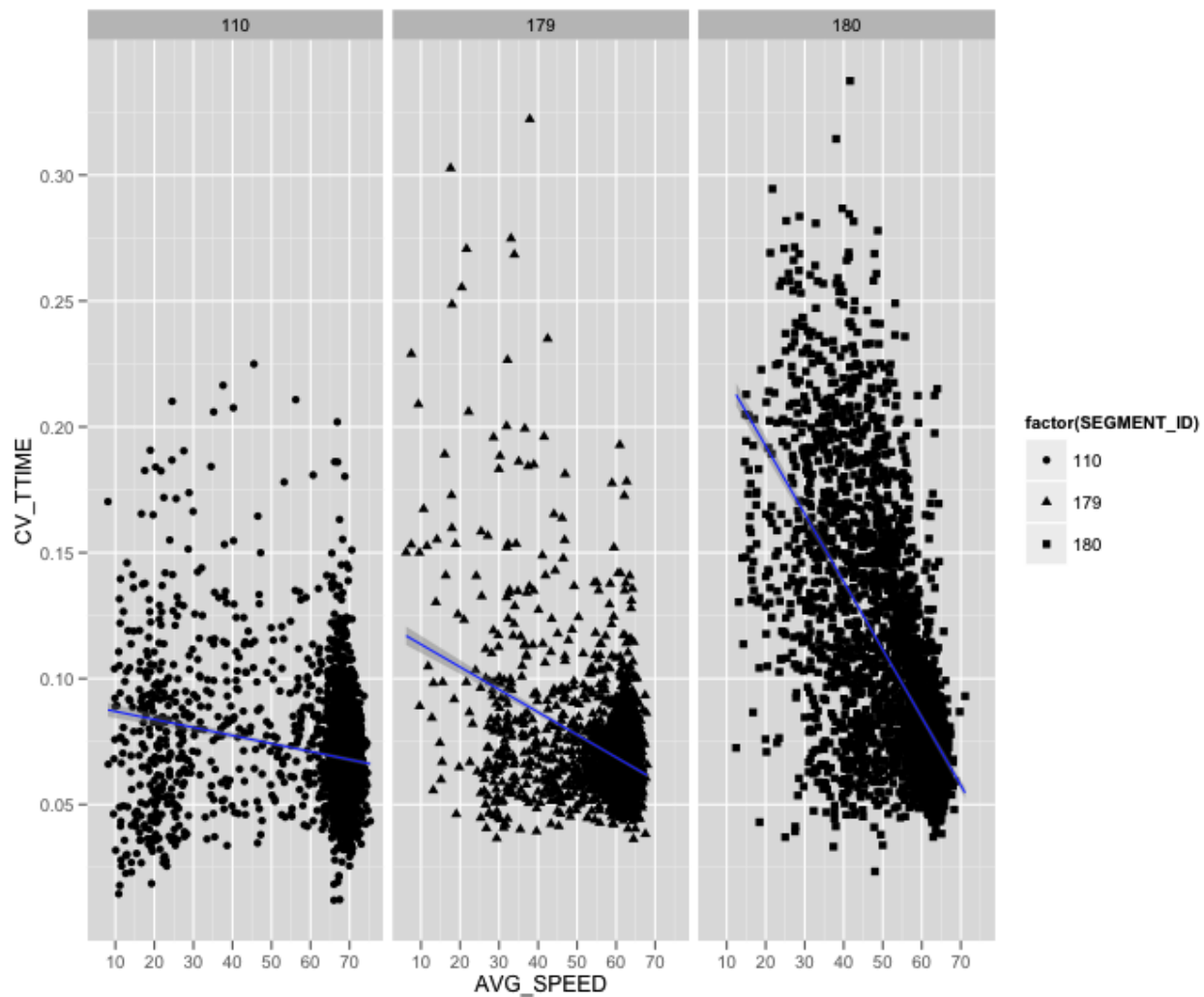


Figure 5: Relationship between Space-mean-speed and CV travel time for 3 links in Houston

This correlation is relatively weak and should not be interpreted as having predictive meaning. However, as a general conclusion it can be used to support the finding that congested traffic flow generally is connected to greater variance in travel times.

3.10 Spatial Correlation

While the variance in travel time or space-mean-speed on a single link in the network is an important statistic, it is also important to consider how link speeds correlate in space. For example, links that are connected or in some way close in distance may have more highly correlated speeds than links that are far apart. Thus, spatial correlation is another important statistic to understand when designing a data quality evaluation.

Spatial correlation can measure how strongly related any two links in a network are. A positive correlation indicates that as speeds on one link increase the expected change in speeds on the other link is also positive.

An analysis of link speed data from a TIS covering a network of links in northern Virginia indicated that speed correlations exist and that as distance increases, the correlations decrease. In other words, neighboring links tend to have highly correlated speeds and distant links tend to have less correlated speeds.

To illustrate this effect link speed correlations are plotted for two links: one on I-66 Eastbound and one on I-95 Northbound. These links were chosen randomly to investigate correlations. Descriptive statistics for the links are provided in Table 3.5.

Link_ID	Length (mi)	Road	Position	Links in Corridor
110-04177	1	I-66 EB	26	43
110-04152	1.26	I-95 NB	18	32

Table 3.5: Description of Links Analyzed for Spatial Correlation

The "position" of the link, shown in Table 3.5, indicates the ordinal position of the link in the corridor. The first link would be numbered 1, the second 2, and so on. The spatial correlation structure is plotted as a line graph below in Figure 6.

The same plot is shown for the second link that was examined. A similar correlation structure can be seen in Figure 7.

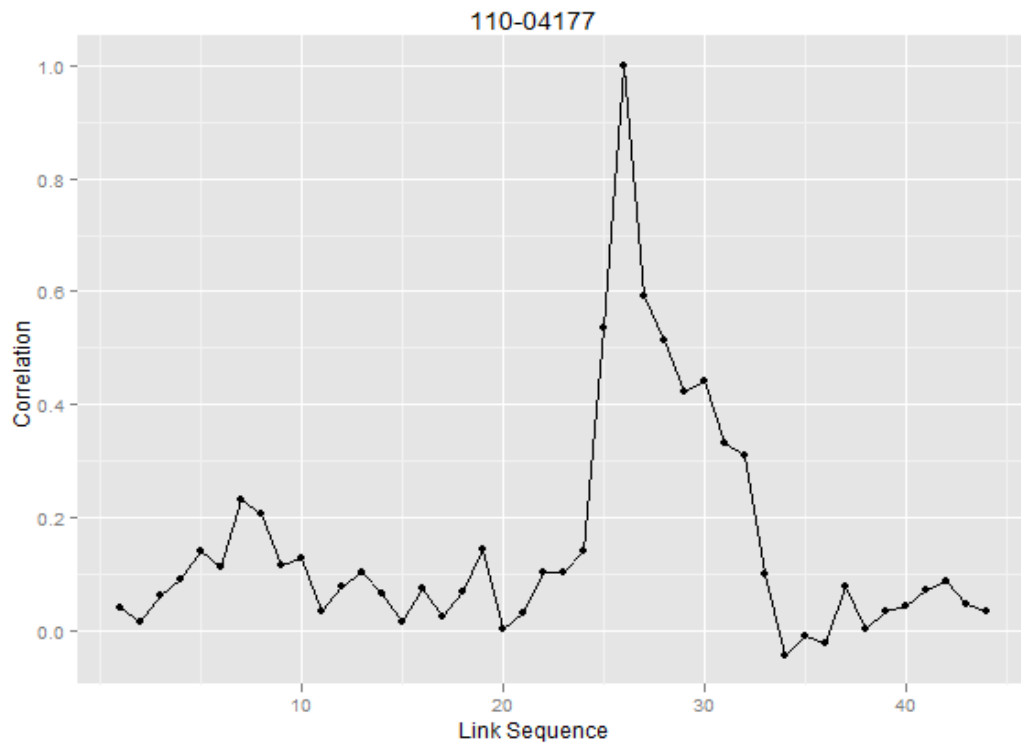


Figure 6: Spatial Correlation Link 110-04177

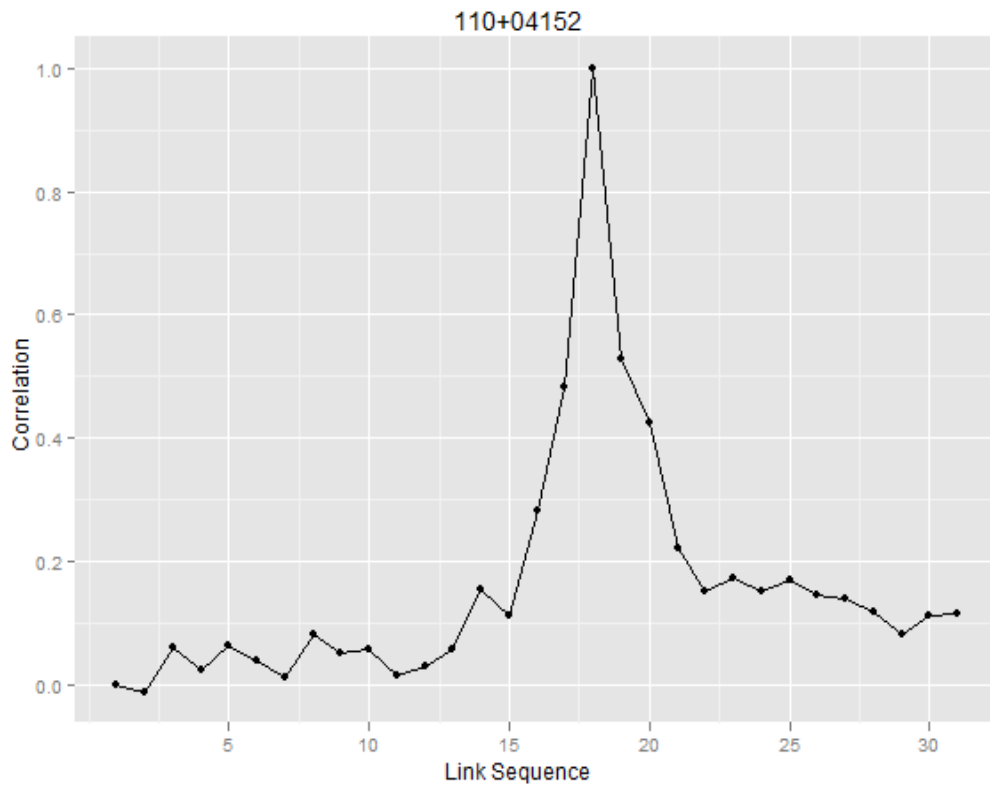


Figure 7: Spatial Correlation Link 110+04152

It can be seen from the plots of link correlations that links upstream and downstream tend to have a positive correlation with the reference link but that as the distance (i.e. # of links) increases from the reference link the correlations decrease.

This leads to the more general insight that while link correlation structure will vary depending on a number of different factors the distance between links does appear to have a consistent effect in reducing correlations. And, more importantly, this leads to the insight that benchmark link selection and methods need to consider the effect of link correlations on data quality estimates.

3.11 Conclusions of Travel Time Distribution and Variance

The statistical analysis of link travel time and space-mean-speed indicates that there are key factors in a transportation network that can be related to variance in travel time. The strongest factor appears to be the functional classification of the roadway. Arterial links are characterized by much greater variance than freeways due to the presence of factors such as delay due to signals, turning movements, and greater access density. Other factors such as space-mean-speed, ADT, freeway access point density, and link length were also investigated and found to have a relationship with travel time variance. However, only space-mean-speed appears to have a significant relationship with variance in travel time. The empirical correlation between space-mean-speed and CV travel time based on the sampled observations was -0.25. This indicates that as speeds decrease, variance in travel time increases. Finally, it was also shown that spatial correlation among links in transportation networks is a factor that should be considered when selecting links for evaluation.

Chapter 4 Travel Time Data Management and Processing

The evaluation of travel time data quality requires that travel time estimates from the TIS and from the benchmark data sources be managed in a way that facilitates the evaluation process. For offline evaluations data may be downloaded and managed in data structures such as flat-files. However, for real-time data quality monitoring and for evaluations of large transportation networks it is necessary that the data be managed in a way that facilitates automated processing of data and evaluation of data quality.

Relational database systems are an effective tool for managing real-time and archival data generated by Traveler Information Systems. The relational data model typically requires that data be "normalized" according to established database design practices such as 3rd Normal Form. The essence of this design is that "entities" be identified by a "key" and that relationships between entities be identified using these keys. These relationships can be thought of as the dimensions of a data structure.

Data warehousing is another branch of relational database systems that is focused on the specifics of managing large quantities of historical data. These systems take a slightly different approach to the modeling problem and relax many of the design criteria required by 3rd normal form. A typical data warehouse will structure data according to a "star schema" or "snowflake schema" design. [17] These designs place a central "fact" table at the center of the model and include "dimension" tables that support the fact table. The fact table is generally the table which contains all of the "measures" of interest. In the case of a travel time database system the main measure of interest is link travel time or space mean speed.

4.1 Basic Data Model for Managing Traveler Information Data

A travel time database system can be modeled with three dimensions: space, time, and source. The spatial dimension identifies the link in a network to which the travel time estimate refers, the temporal dimension identifies the time-interval for the estimate, and source dimension identifies the source of the estimate. The basic structure of the relational system is shown below in Figure 8.

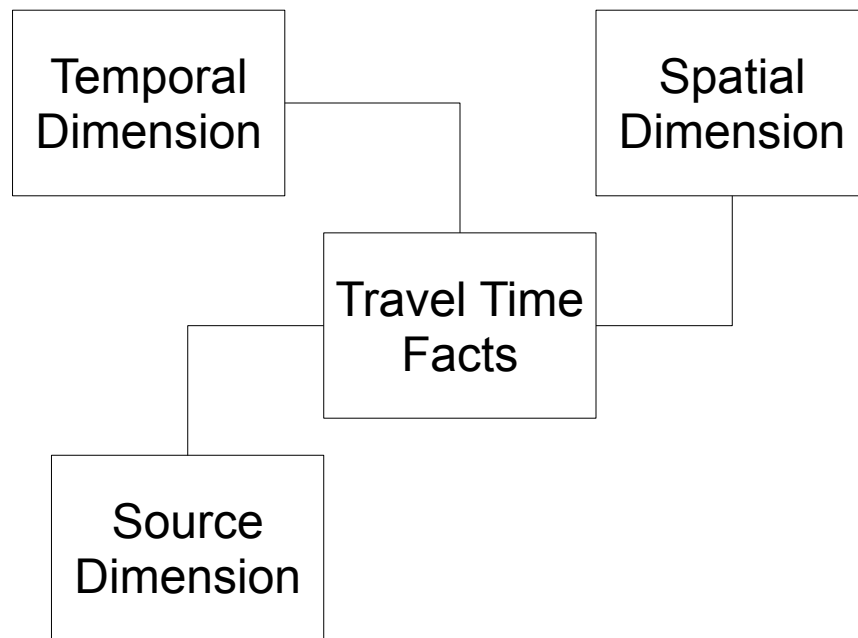


Figure 8: Basic Data Structure of Travel Time Management System

A travel time "fact" is therefore "keyed" by a link_id, time_id, and a source_id.

4.2 Travel Time Data Structure

This section describes the data structure used to manage and process travel time estimates. The data structure is designed to be flexible in that multiple sources of estimates can be managed while allowing aggregation of estimates along spatial and temporal dimensions at different resolutions. The

next section describes how this data structure can be utilized.

4.2.1 Spatial Dimension of Data Structure

The spatial dimension refers to the units of spatial measure over which travel time is estimated. Typically, these units are referred to as "links" in a network. A link may refer to the extent of roadway between two signals or two freeway access points in a road network. More generally, a link is the connection between two nodes in a network.

In the evaluation of travel time data quality, links are typically predefined by the source of the travel time data. A private-sector vendor may define links using a proprietary scheme. But currently most link definitions are following the "Traffic Message Channel" standard. This is a link definition standard in the United States that is managed by a consortium of data providers. The TMC standard divides a road network into base links that are identified uniquely by a character string and are defined by the latitude/longitude coordinates of the start and end points of the link. The links are directional so that a northbound link can be distinguished from a southbound link. More information about the TMC standard can be obtained from vendor specific documentation.

4.2.2 Temporal Dimension of Data Structure

The temporal dimension refers to the units of time over which travel time is estimated. An estimate of travel time on a specified link must refer to some time-interval as the relevant unit of time. For example, a travel time estimate may be relevant for the period of time between 9:00 AM and 9:05 AM on Wednesday March 18, 2012. Each time-interval should be "keyed" by a unique identifier.

The time-interval between 9:00 AM and 9:05 AM on Wednesday March 18, 2012 would then be identified by:

CALENDAR_DATE_KEY + FIVE_MINUTE_INTERVAL_KEY

4.2.3 *Source Dimension of Data Structure*

The source dimension defines the attributes of the source of the travel time estimates. This dimension is needed when a data archive is managing travel time estimates from multiple sources. For example, using the model described in this research, a travel time archive can manage estimates from multiple TIS sources as well as from a benchmark data source such as Bluetooth recorders.

4.3 Spatial and Temporal Coordinate System

The spatial and temporal coordinate system is the set of coordinates that identify a travel time or link-speed estimate in the spatial and temporal dimensions. Space and time are often measured on a continuous scale. However, a TIS will provide data at only discrete points in the space/time field. Therefore, the spatial and temporal dimensions are defined by a set of discrete intervals.

Spatial resolution describes the level of detail about spatial data obtainable from a data source. The spatial resolution of a TIS is the base set of elemental links in the network over which travel time estimates are provided. These are the smallest units of space over which travel time is estimated.

Similarly, temporal resolution describes the detail obtainable from a data source on the temporal dimension. The temporal resolution is the minimum time-interval over which travel times are estimated. For example, an agency may require that the TIS provide estimates of travel time every 5 minutes. This is then the most granular temporal resolution that the system provides.

The link and time-interval therefore defines the coordinates in space/time dimensions of a single estimate of travel time provided by the TIS. Figure 9, below, illustrates this concept.

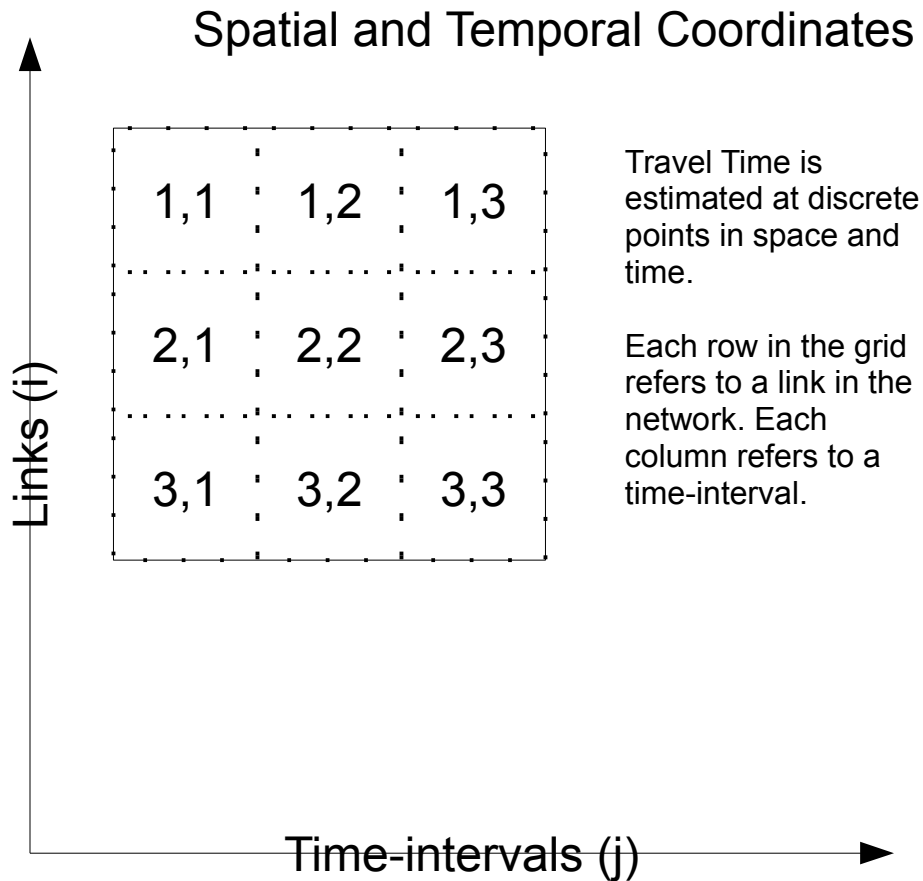


Figure 9: Spatial and Temporal Coordinate System

In Figure 9, each cell refers to a unique estimate of travel time or link-speed at a point in space and time. For example, cell (1,1) refers to the estimate on link #1 at time-interval #1. If we assume that the coordinates are ordered such that cell (1,1) precedes cell (1,2) in time and cell (1,1) precedes cell (2,1) in space then we can aggregate estimates in space, time or both. The process of aggregating estimates in space and time is often used to estimate average link speeds at a different spatial or temporal resolution.

Spatial aggregation is used to "combine" estimates from neighboring links in order to estimate travel time over a longer segment. For example, if we want to estimate the average travel time over link #1 in Figure 10, we would combine the estimates at cells (1,1), (1,2), (1,3). The method of combining

these estimates depends on the units of measurement. If we are combining travel times then we can simply sum up the estimated average travel time in each cell. If we are combining space-mean-speed then we would use a harmonic mean of the estimated average speeds. The equations for spatial aggregation in units of time or speed are given below.

Spatial Aggregation by sum of Travel Times (units of time):

$$\sum_{i=1}^m T_i \quad (10)$$

Spatial Aggregation by Harmonic Mean (units of speed):

$$HM(u) = n * \left(\sum_{i=1}^m 1/u_i \right)^{-1} \quad (11)$$

Temporal aggregation is the process of combining estimates of travel time or link-speed over a longer time-interval. For example, the estimates of link-speed can be aggregated a one-hour time-interval to estimate average hourly speeds for a given road segment. Temporal aggregation is simply the arithmetic mean of the observed travel times or link-speeds of a given link.

4.3.1 *Spatial and Temporal Alignment*

The benchmarking process may be estimating travel time at a different spatial and temporal resolution than the TIS. For example, benchmark data may be collected over an extent of roadway that covers the spatial extent of three links as defined by the TIS. Figure 10, below shows an example of this situation.



Figure 10: Spatial Alignment

In the example above, three estimates of travel time (TMC1 - TMC3) must be spatially aggregated for comparison with the benchmark estimate.

Similarly, the temporal resolution of the benchmark estimates must be aligned with the TIS. If the TIS is providing estimates of link travel time every 5 minutes, but the benchmark estimation process requires 15 minutes of observations then there will be 3 TIS estimates that must be aggregated.

Therefore, the process of temporal and spatial alignment involves aggregating TIS estimates over spatial and temporal dimensions. Along the spatial dimension, a harmonic mean of link-speeds is computed and along the temporal dimension an arithmetic mean is computed. The result of this process is then a scalar value that represents the TIS estimate of link speed at the same spatial and temporal resolution as the benchmark data. Figure 11, below illustrates this process in two dimensions.

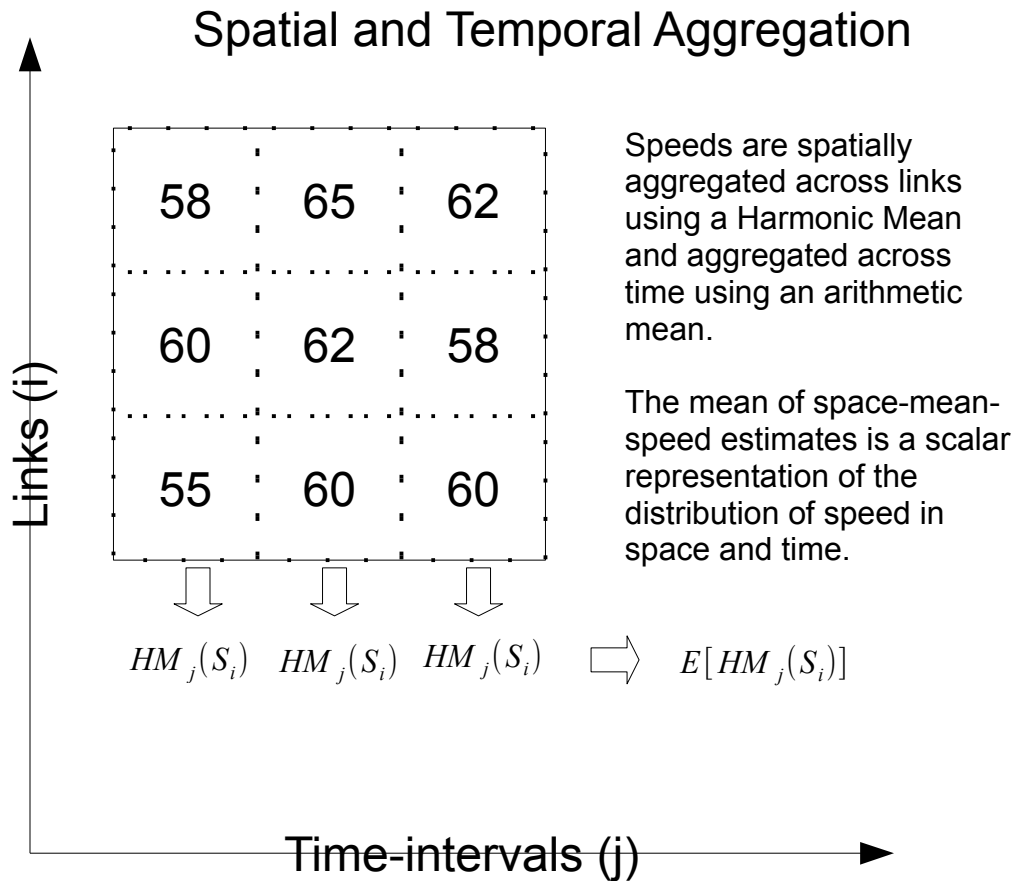


Figure 11: Spatial and Temporal Aggregation

Figure 11 shows an estimated speed provided by a TIS in each cell. The cells refer to coordinates in the space/time dimensions. First, spatial aggregation is performed resulting in a new estimate of link-speed at each time-interval. Next, temporal aggregation is performed to yield the final estimate of link-speed for the desired spatial and temporal resolution. This final value is a scalar value that estimates the average travel time or link-speed as estimated by a TIS.

4.4 Conclusions of Travel Time Data Management and Processing

In this chapter it was shown that travel time data can be managed using a relational database model with the travel time measurements modeled as "facts" and the supporting dimensions being the

spatial, temporal, and source dimensions. This approach allows for online and historical analysis of large databases of travel time estimates and comparison with a benchmark data source. It was also shown that Traveler Information Systems transmit data at discrete points in time and space. Therefore, there is a discrete coordinate system that can be used to identify a single travel time estimate from a TIS. This coordinate system can also be used for spatial and temporal alignment when comparing estimates at different spatial and temporal resolutions.

Chapter 5 Travel Time Benchmarking

The purpose of benchmarking is to establish a value, often referred to as “ground-truth” that reflects the actual conditions being estimated. Benchmarking is used in many areas of engineering and science to establish a reference value for some process or material. For example, computer engineers will benchmark the speed of a CPU by measuring the time to complete standardized processing tasks. A structural engineer will benchmark the strength of a material by subjecting it to different strain tests. Benchmarking is also used in transportation engineering to establish reference values for a data quality evaluation.

Transportation engineers often use the term “ground truth” to mean benchmark value. Although the term “ground truth” implies “truth”, it is in fact just an estimate of a statistic. In travel time data quality studies the ground truth or benchmark value most often used is the average travel time or space-mean-speed over a specified link during a specified time interval.

In order to estimate a benchmark value, first some empirical observations must be collected. Next, the collected observations are processed such that an estimate of the benchmark statistic can be calculated. This chapter will review some of the existing methods of benchmark data collection and the statistical techniques used to estimate the benchmark value.

5.1 Measuring Travel Time

The way in which benchmark travel time is measured has changed as sensing technologies have changed. Travel time can be directly measured by observing the time it takes for a vehicle to traverse a link or it can be inferred from measurements of related variables such as spot-speed, volume, and density.[18] The following sections briefly discuss the difference between inference and direct

measurement of travel time.

5.1.1 Inferential Techniques for Measuring Travel Time

The most widely used inferential technique involves estimating travel time from loop-detector measurements of speed, volume, and occupancy at points along a link. In the case of a single-loop detector, the speed is estimated by using a calibrated "g factor" to estimate density from detector occupancy. In the case of a double-loop detector trap the space-mean-speed is computed directly at the point of the detector. In both cases, these are estimates of space-mean-speed at a specific point along the link. Thus, they are point estimates of speed.

Given a series of detectors, a model can be specified to estimate travel time from the point estimates of speed along the link. One of the more popular models in the literature is the "zone of influence" model which estimates the travel time as a linear function of the estimates at each point along the link. For example, each detector has a specified zone of influence extending half-way to the next detector. The speed estimated at the detector is assumed to be constant in this zone. Given the length of the zone the travel time can be computed directly from the speed estimate.[19]

Inferential techniques were frequently used because the most common types of sensors used in the field were only capable of estimating point measurements of speed or volume. For example, single-loop detectors can estimate speeds from volume and occupancy readings averaged over time. This estimate of speed can be combined with other estimates of speed along the same link and an estimate of average speed over the entire link can be calculated.

One significant problem with this technique of estimating travel time is that each link requires a uniquely calibrated model. The parameters of such a model are not easily transferred to other links due to sensitivities in the model to detector spacing, and detector calibration. Finally, estimation of the

standard error of the estimates is also somewhat problematic and often not considered.

5.1.2 *Direct Measurement of Travel Time*

Direct measurement of travel time requires some method of observing actual vehicle trip times. Before the widespread adoption of wireless devices, travel time was typically measured by either a floating car run or by license plate matching.

The idea behind the floating car run was that a trained driver operating a monitored vehicle would drive along a link and record the travel time. The assumption with this approach was that the trained driver could approximate the middle of the travel time distribution by careful navigation within the traffic stream. For example, the drivers were instructed to pass one vehicle for every vehicle that passed the driver's vehicle.

This approach was used frequently in travel time based studies because it required relatively simple equipment (e.g. a stop-watch and a driver). It was also thought that by following this approach a reasonable estimate of the average travel time could be approximated from the floating car. However, this approach was limited in the number of samples that could be collected over time.

Another approach that was used in field studies of travel time was license plate matching. In this scenario, two people would be positioned at ends of the studied link and would record vehicle license plate numbers and descriptions. From this information, random samples of vehicle travel time could be collected by matching up the recorded license plate numbers and time. If a large enough sample was collected, a statistically significant estimate of the benchmark average travel time could be estimated by a null-hypothesis test.

Floating car runs and license plate matching were the early prototypes of what now are the dominant methods of travel time data collection: Automated Vehicle Identification (AVI) and probe

vehicle monitoring. Each of these newer methods can collect statistically significant samples of travel time observations over long periods of time without requiring people to be in the field. They work by anonymously tracking the position of vehicles via wireless communication. Thus, travel time can be directly measured between designated points of interest.

5.1.3 Probe Vehicle Monitoring

Probe vehicle monitoring refers to any type of technologies that can monitor the position of vehicles in a traffic stream over time. Such technologies can be loosely described as wireless location technologies and this includes GPS and cell-phone positioning. Many of these technologies can now wirelessly and in real-time transmit position and time to a central computer system for processing and storage.

Currently, one of the largest sources of probe vehicle data is the commercial trucking and delivery sector. Many companies have equipped their vehicle fleet with GPS devices capable of wirelessly transmitting the current position of a vehicle at frequent time intervals to a central computing system. The data can be used by the companies for logistics and real-time tracking of the vehicle fleet.

This has also been shown to be a valuable source of data for monitoring traffic conditions. As the position of each probe vehicle is updated the distance traveled between reports can be calculated and the travel time can be directly computed as the difference in time between reports. Since the distance is also available the space-mean-speed can also be computed from the probe data. With a large enough fleet of probe vehicles reporting their position over time it becomes possible to derive very accurate estimates of link travel time. Figure 12, below shows an example of GPS points transmitted by commercial trucks traveling on I-95 Northbound in Virginia.

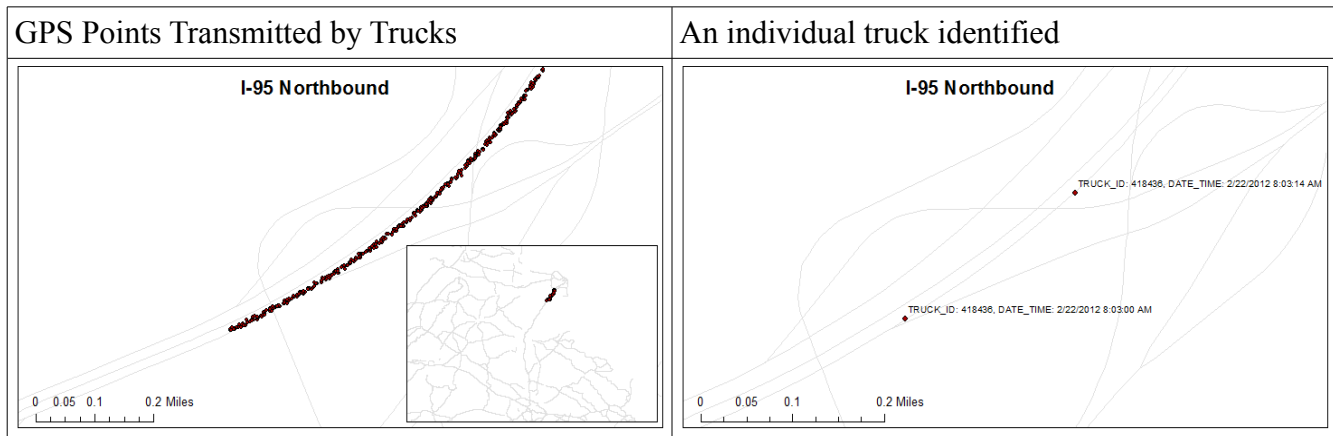


Figure 12: Example of GPS points transmitted by trucks

5.1.4 Automated Vehicle Identification

Automated Vehicle Identification (AVI) technologies include any technologies capable of remotely detecting a vehicle at ends of a specified link the road network. These technologies work by sensing a unique identifier associated with a vehicle at one end of the link and then sensing the same identifier again at the end of the link. The identifier value can be matched at each sensor position and the time difference is the travel time of that vehicle. Since the position of the sensors is fixed the distance traveled is known and the space-mean-speed can be calculated for each vehicle that is identified.

Currently, the two most popular types of AVI technologies are Bluetooth Reidentification (BTR) and Toll-tag Reidentification. BTR works by deploying a small computer with an antenna capable of detecting a bluetooth signal at ends of a link that is to be monitored. Each Bluetooth signal has a unique identifier embedded in the signal and the computer will register the identifier if a signal is detected. A growing percentage of the vehicle population currently carries some kind of bluetooth device on board. When these devices are powered on they transmit a signal. Studies have shown that approximately 5-

10% of freeway traffic can be identified using this approach. This can generate samples large enough for statistical estimation.

Toll-tag Reidentification functions similarly to BTR in that each vehicle is carrying a toll-tag transponder that responds to Radio Frequency Identification (RFID) signals. The RFID chip in the transponder can be used to uniquely identify the vehicle at points where toll-tag sensing equipment has been deployed.

5.2 Estimation of Benchmark Travel Time

Benchmark travel time is an estimate of the average travel time of vehicles on a link during a specified time-interval. As observations of travel time or space-mean-speed are recorded they form a series of sample observations in time. Figure 13, below illustrates a time-series plot of link-speed observations.

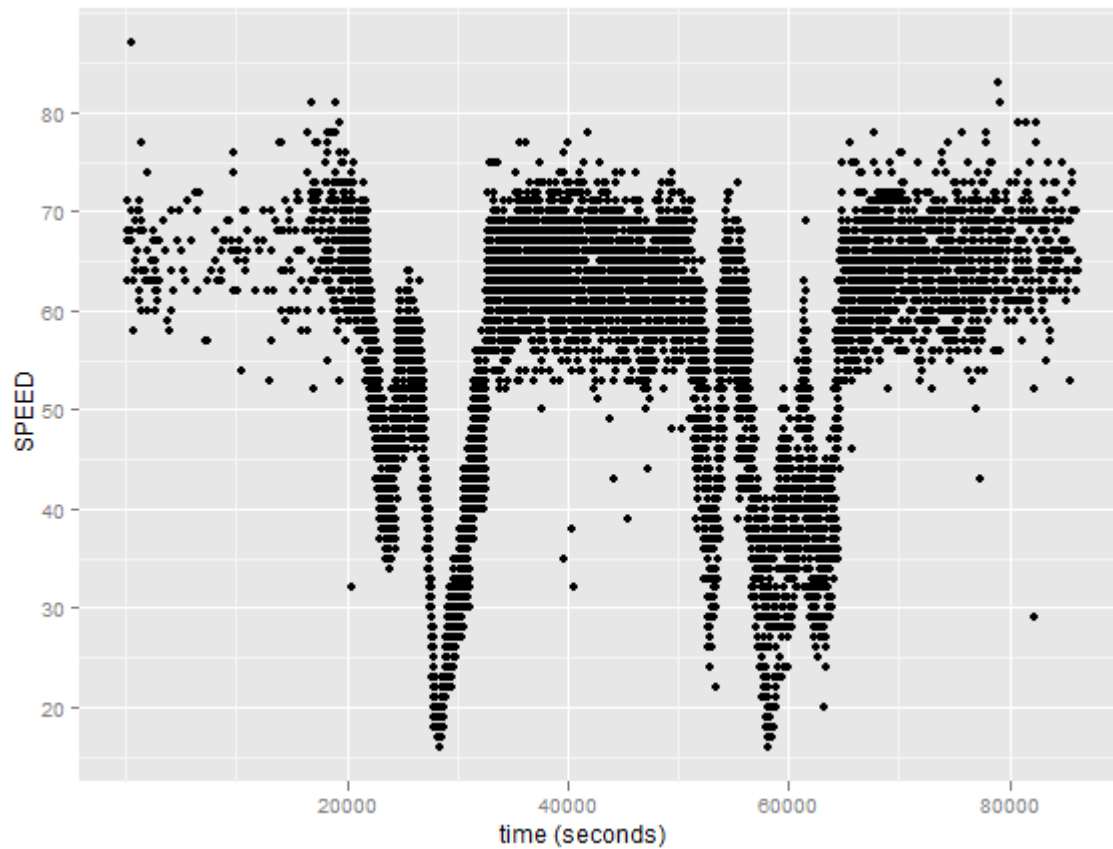


Figure 13: Time Series observations of space-mean-speed

It can be seen from the plot that fitting a global function to this series would be complex and unlikely to generalize well to other time periods or other links. A first-order linear regression of speed against time would explain very little of the variance in the data. High order polynomials, sinusoidal functions, and other global functions would also require more model complexity and lead to an over-fit model.

Another approach that can be used with time-series data is to fit a number of local parametric models to the observations. Each local model is fitted to a specific "neighborhood" of points. The estimate at any point in time is then a function of the relevant local model.

There are two types of local models that were investigated in this research:

1. Local Averaging
2. Locally weighted linear regression (LOESS)

5.2.1 Local Averaging

Local averaging involves specifying a desired time-interval bin-size and then organizing the observed travel time so that each observation falls into a particular bin. Once the data has been binned then an estimate of the benchmark can be computed from the binned data. For example, the average travel time in each time-interval bin and a confidence interval can be computed. Local averaging is the most commonly used benchmarking method in past data quality evaluations.

Table 5.1, below shows a series of observations of travel time. Each observation is assigned to a 5-minute bin and an estimate of the average travel time for that bin is computed as the average of the observed values in that bin.

Time of Observation	Bin #	Travel Time (s)
08:00:00 AM	480	65
08:02:00 AM	480	60
08:03:00 AM	480	66
08:06:00 AM	481	68

Table 5.1: Example of Binned Travel Time Observations

The estimate of benchmark travel time in bin # 480 would be the average of the observed travel times in that bin or $(65+60+66)/3 = 63.7$ seconds.

However, the choice of a bin size is somewhat problematic. For example, when using a 5-minute bin the number of observations in a particular 5-minute time-interval may be too small for confident statistical calculations. On the other hand increasing the bin size has the effect of smoothing

out fluctuations in the traffic state. At 15-minute bin-sizes, sample sizes may be sufficient for statistical analysis but a 15-minute time-interval may average out sudden fluctuations in the traffic state that may be valuable to detect.

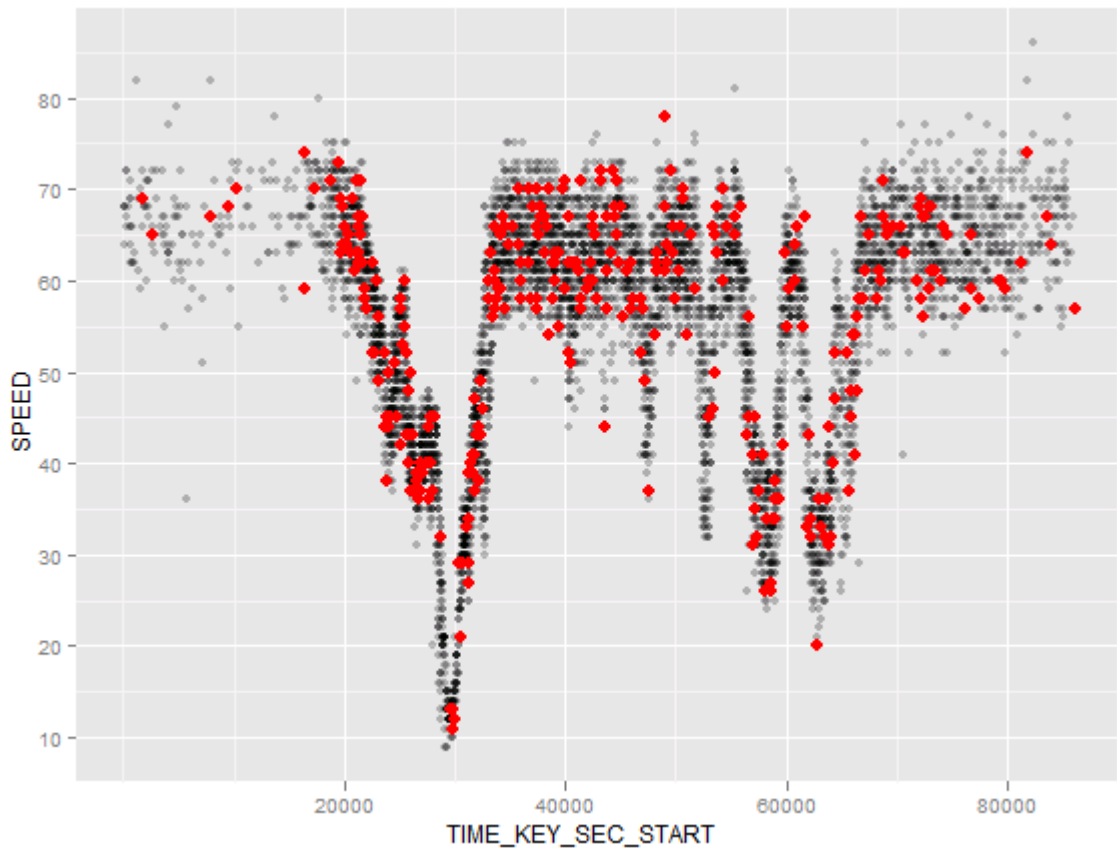


Figure 14: Sampled Data

To illustrate how local averaging works we can start by taking a random sample of the original data set shown in Figure 13. In this case, a sample of only 5% of the original data was drawn. The resulting sample is shown in red plotted on top of the original data in Figure 14.

Local averaging results in a discrete set of points on the time axis. Each point represents the estimate of average travel time for a particular time-interval. When the independent variable (e.g. time) is considered as a continuous variable then some assumptions must be made about how the response

values change with the independent variable. For example, the local average for the time between 08:00 AM and 08:05 AM was found to be 60 seconds. Now, we want to know what the average was at 08:03 AM.

We can view this problem in one of two ways. First, we can assume that the response is constant for all levels of the independent variable. Thus, the average travel time at all times between 08:00 AM and 08:05 AM is estimated to be 60 seconds. This seems to be somewhat problematic in the sense that we know that the response (e.g. average travel time) may fluctuate within the 5-minute time-interval.

We can also use linear interpolation between observed response levels to estimate response levels on a continuous scale. If we know the estimated average travel time for the preceding and following 5-minute intervals we can interpolate a line between each level and use that function as an estimate of the response for continuous levels of the independent variable. For example, given the following data we can build a local linear function (with t as the time in minutes) for the last two time-intervals:

Time-interval	Sample mean	Local linear function
07:55 - 08:00	59	
08:00 - 08:05	61	$y = 0.4*t + 59$
08:05 - 08:10	65	$y = 0.8*t + 61$

This results in a series of lines connecting the computed averages at each time-interval bin. Figure 15, below shows the result of local averaging at 5-minute intervals for the sampled data shown in Figure 14. The resulting response level estimates are the average of the sampled data at 5-minute intervals and the lines are the computed interpolation lines between each response level. The result is a very jagged series.

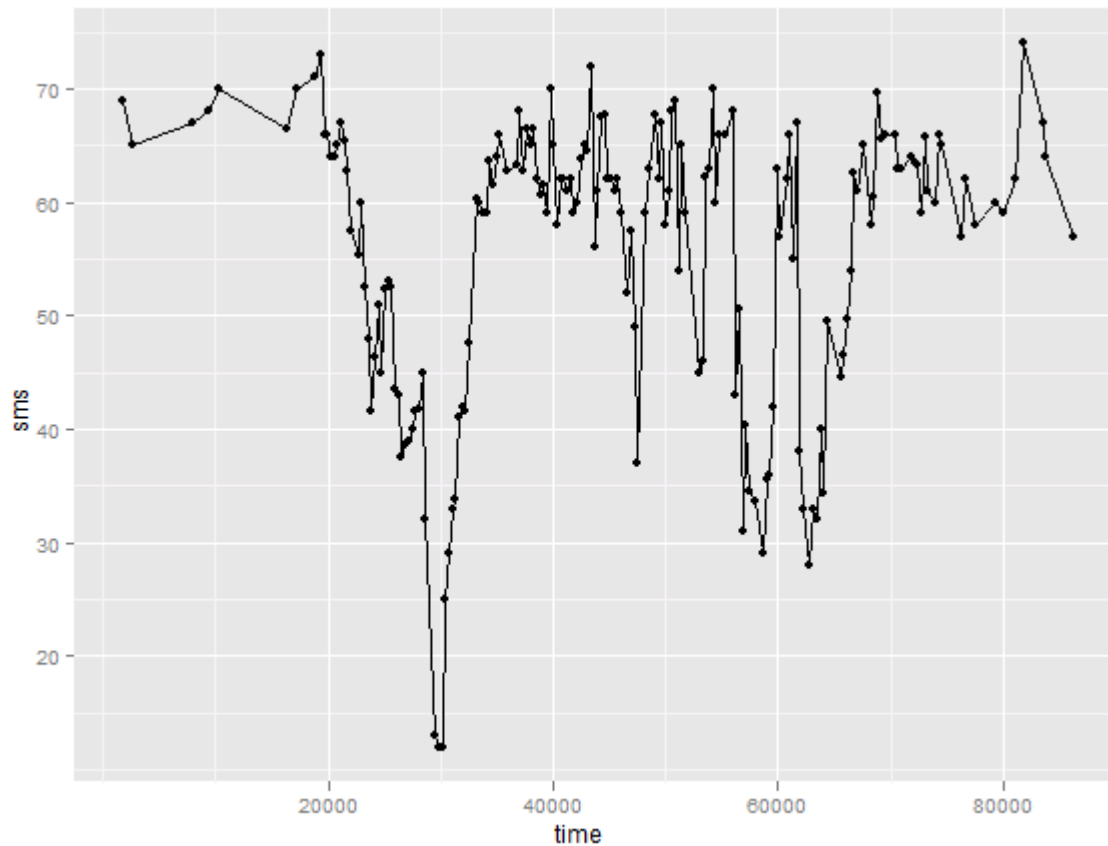


Figure 15: Local Averaging at 5-minute intervals

This is a type of local linear regression where the local neighborhood is fixed at a certain time-interval. This model seems to be an improvement over the constant response model for local bins. However, again, the model assumes that the response changes linearly between fixed points on the X-axis. It also makes an assumption about the appropriate bin-size for the local neighborhoods. Many of these problems are dealt with in Locally Weighted Regression which is a type of local averaging described in the next section.

5.2.2 Locally Weighted Regression (LOESS)

Locally weighted regression is a smoothing technique that was developed by Cleveland and Cleveland and Devlin.[20] It is a computationally intensive technique that until relatively recently was

impractical for most large data sets due to the required processing time. With the memory and CPU speed of modern desktop computers it is now practical for LOESS to be applied to most reasonably sized data sets.

LOESS has been applied in a wide range of time-series modeling problems. Eisele, Rilett, et. al. used LOESS to estimate mean travel time and confidence interval boundaries for time-series observations of travel time in Houston.[21] They applied this method to evaluate the effectiveness of real-time monitoring of a freeway system using travel time observations from an AVI system. However, the investigation did not evaluate the performance of the LOESS method relative to other estimation techniques. They concluded by recommending the LOESS due to its effectiveness and ease of implementation.

LOESS uses a nearest neighbors window to fit a linear regression model to a set of weighted observations. The nearest neighbors window is determined by a smoothing parameter, often denoted as λ (λ) which specifies the proportion of observations to be used in fitting the linear regression model. Typically λ is suggested to be between 0.50 and 0.75. However, an optimal value of λ can also be determined by minimizing square error using cross validation. This approach will be discussed later in this chapter.

The linear regression model is specified as a low-order polynomial. Most often a second order polynomial is used in the regression function. A higher order function may be selected but is discouraged in the statistical literature due to problems with over-fitting the data. Lower order models result in a linear model and a constant model which result in models similar to the local averaging approach described in the last section. The second order model is widely cited as the most appropriate model for most applications.

The weighting function of the LOESS algorithm is a tri-cube weighting function. The tri-cube is

specified as:

$$w(x) = \begin{cases} (1 - |x|^3)^3 & |x| < 1 \\ 0 & |x| \geq 1 \end{cases}$$

where x is specified as a scaled distance.

The algorithm proceeds by:

- 1) For each observation (x_i, y_i), determine the nearest neighbors (x, y) given the value of lambda
- 2) Fit a weighted second order regression function to the selected observations
- 3) Compute the estimated response at x_i from the fitted function
- 4) Repeat for every x_i

The result of this process is a vector of fitted points y at each level of the independent variable x . Estimates of response values between observed levels of x can be computed by the estimated response levels for the functions passing through the point of interest.

Figure 16, below illustrates a LOESS function fitted to the sampled data. It can be seen that the function is a smoothed fit to the sample data and does not have the same jagged appearance that the local averaging had. The degree of smoothing in LOESS is controlled by the lambda parameter and selecting the best value for lambda will be discussed in the next section.

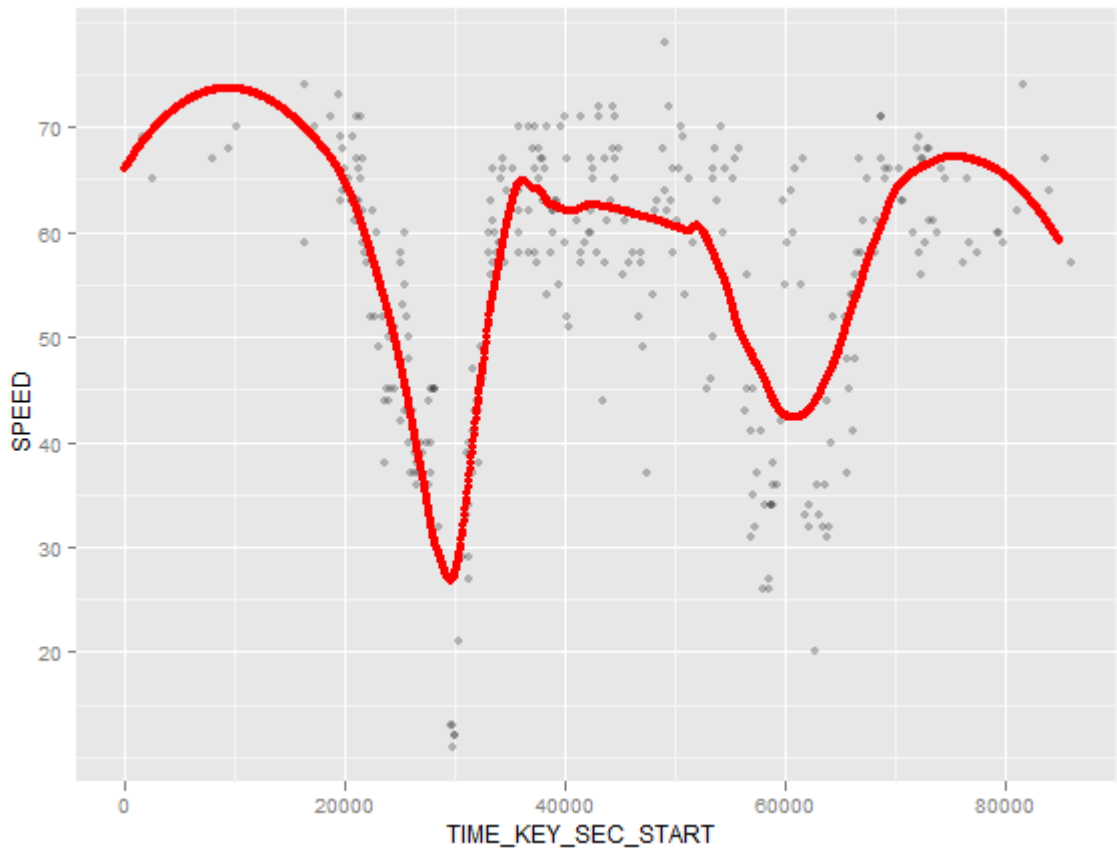


Figure 16: LOESS functions fitted to sample data

5.2.3 Selecting the "best" value of Lambda in LOESS

As described in the last section, the lambda parameter of the LOESS algorithm determines the proportion of observations to include in the nearest neighbor window. When lambda is very small (e.g. < 0.1) only points very close by will be used to fit the regression function. This results in a highly fitted function at each point and will tend to "chase" random fluctuations in the data. A lambda value that is very large (e.g. $> .75$) will include a large proportion of the data when fitting the local regression function at each level of x . This results in a very "smooth" function that may not capture important shifts in the signal being monitored.

The choice of lambda is therefore an important decision in fitting a LOESS model to a set of

observations. One method for choosing lambda suggested in the literature is to use 10-fold cross validation to estimate the mean square error (MSE) for a range of lambda values. The value of lambda that minimizes the MSE using 10-fold cross validation is the optimal value for a particular data set.[22]

N-fold cross validation is a resampling technique that can be used to calculate a robust estimate of MSE for a particular data set and a model. One of the challenges of fitting models is to avoid over-fitting the model. A highly specified model that closely follows the sampled data is unlikely to fit well to unseen observations. Cross validation helps to avoid the over-fitting problem by first fitting the model to a subset of the data and reserving the remainder for validation. For example, the data may be divided up randomly into 10 "folds" and the model will be fit to only 9 of those folds. The remaining fold is used to estimate the out-of-sample MSE.

To find an "optimal" value for lambda in fitting a LOESS model to time-series data 10-fold cross validation can be used to estimate the MSE for different levels of lambda. The procedure works as follows:

1. Choose a test value for lambda
2. Divide up data set randomly into 10 folds
3. Set aside one fold and use the other 9 for fitting at the current lambda value
4. Compute the total squared error of the model against the remaining fold
5. Repeat steps 3,4 for each of the folds
6. Average the observed squared errors for all ten trials
7. Repeat steps 3 - 6 for a new value of lambda
8. Select the value of lambda with the lowest MSE

Following this procedure we can determine an "optimal" lambda value for a particular set of observations. Figure 17, below shows the computed 10-fold cross validation MSE for different values of lambda. In this case the plot shows that the smaller values of lambda would be preferred.

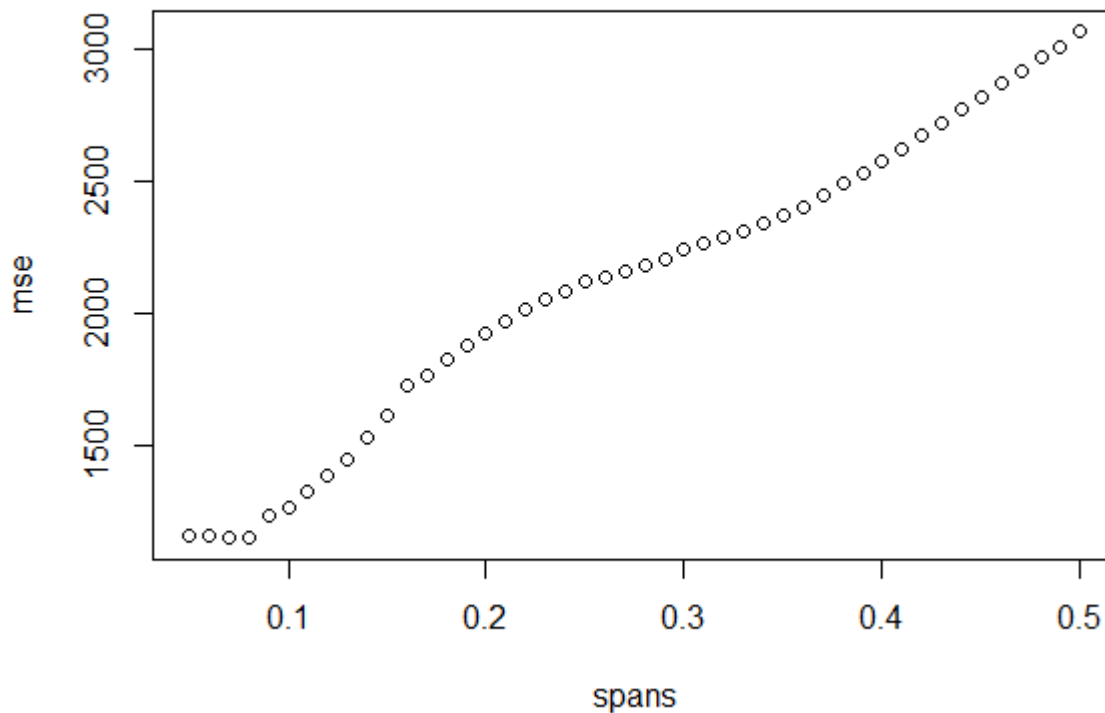


Figure 17: 10 fold cross-validation MSE for a trial data set

If we examine, another data set a different value of lambda may be optimal. For example, MSE was estimated by cross-validation for the same link on a different day and resulted in the following relationship between the smoothing parameter lambda and MSE shown below in Figure 18.

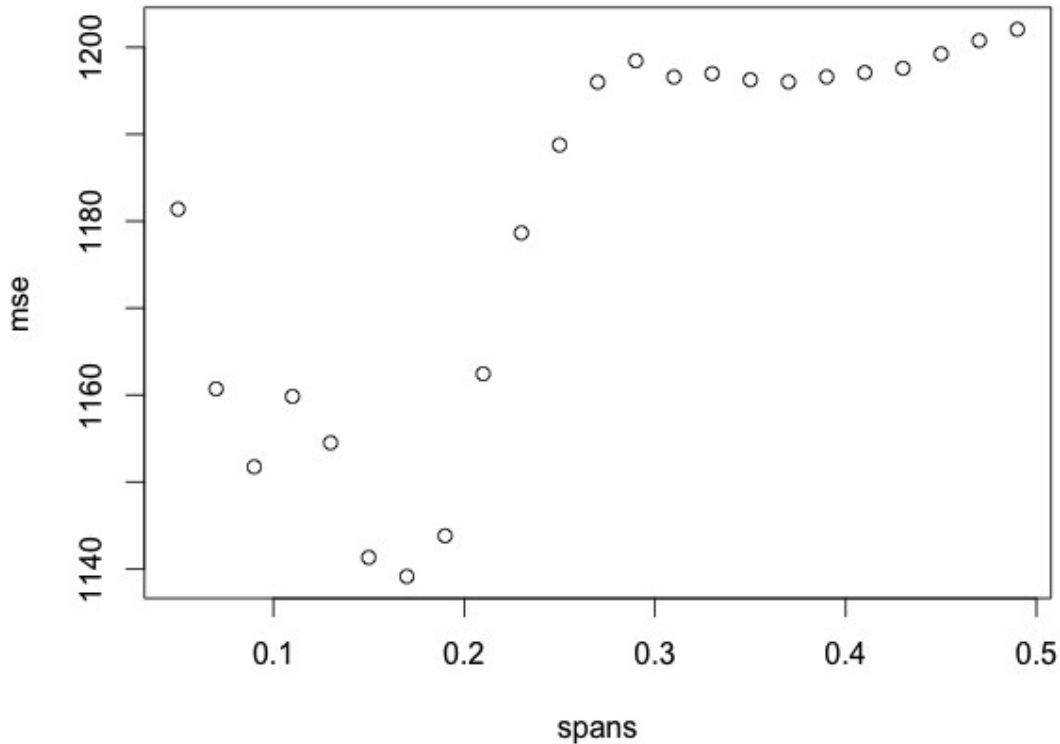


Figure 18: 10-fold cross validation MSE repeated different data set

In general, what we find is that different smoothing parameters work for different data sets. However, rather than find a unique value of lambda for each possible data set we can look for a "generally" optimal value. The smoothing parameter was allowed to vary between .05 and .5 and tested on 10 days of space-mean-speed observations on three different links from the Houston travel time data set. Each day of data was reduced to a sub-sample so that only 5, 10, or 15% of the data was used in the cross-validation tests.

The results indicate that generally lower values of lambda had lower MSE. However the lower the value of lambda the more "complex" the final model because smaller values of lambda mean that each regression function is fit to only a narrow band of observations. To avoid over-fitting it is

recommended in the literature to set lambda between 0.25 and 0.75 depending on the data. However, these tests showed that 0.25 was too high of a value and indicated that values as small 0.05 would often be optimal.

Based on the empirical tests and the recommendations in the statistical literature, the value of lambda recommended for fitting space-mean-speed observations on freeway links is between 0.10 - 0.20. This range of values offer the best compromise between low MSE in cross-validation tests and avoidance of unnecessary model complexity.

5.3 Interval Estimation of Benchmark Travel Time

The previous section introduced two methods for determining point estimates of benchmark travel time. The methods apply local linear models and estimate the benchmark from the local observations. An interval estimate may also be computed using either of these methods. An interval estimate provides "confidence bounds" on the likely range of values for the "true" benchmark.

5.3.1 *Interval Estimates from Local Averages*

The local average technique described in the last section uses a time windowing technique to determine which observations to average. Given a sample of observations from a specified time-interval, an interval estimate of the population mean can be computed using:

$$\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \quad (12)$$

Given the level of alpha, this interval provides a range of values that the true population mean should fall in with approximate probability equal to 1 - alpha.

However, due to problems with the sample size in time-intervals where only a few (e.g. $n < 4$)

observations are available, the confidence interval may be unreasonably wide or not computable.

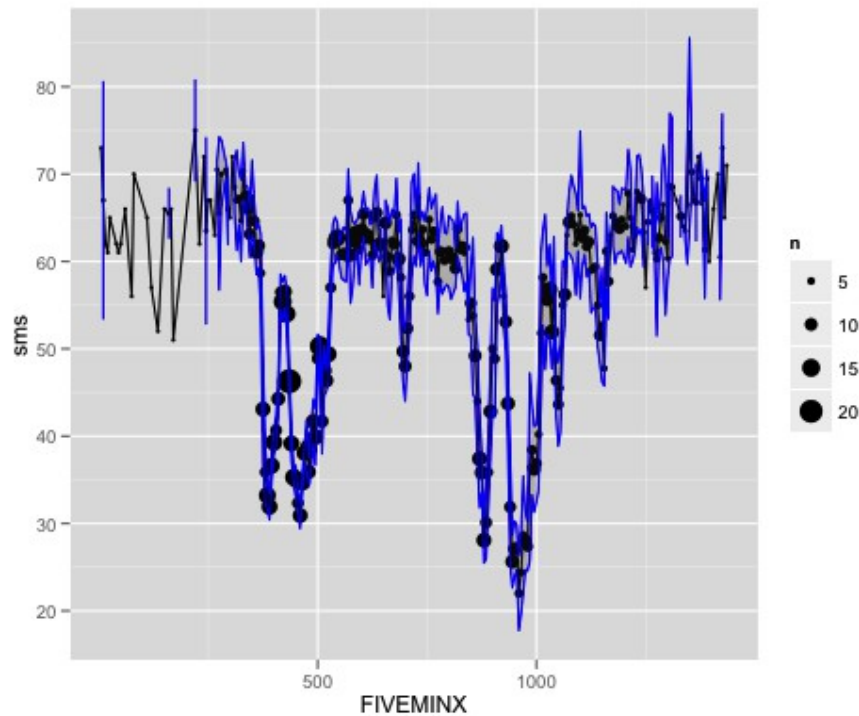


Figure 19: Local Averaging with Confidence Intervals

Figure 19, above, shows a plot of local averaging at 5-minute intervals with the confidence interval indicated in blue and the points indicating the 5-minute sample mean with sample size indicated by the point size. It can be seen from the plot that regions along the X-axis did not have sufficient samples for an interval estimate or had large intervals due to small samples and large variance. The confidence intervals appear to work best in congested flow where a larger number of observations tend to cluster.

5.3.2 Interval Estimates from LOESS

Interval estimates of fitted points from a LOESS function are slightly more complicated to compute. A bootstrap estimate of the standard error can be computed from the residuals of the points

included in the fitted function. The bootstrapped standard error can then be used in a Student T confidence interval. An alternative approximate confidence interval can also be computed using an estimate of the residual variance multiplied by the sum of the computed weights for the points in the fitted function.

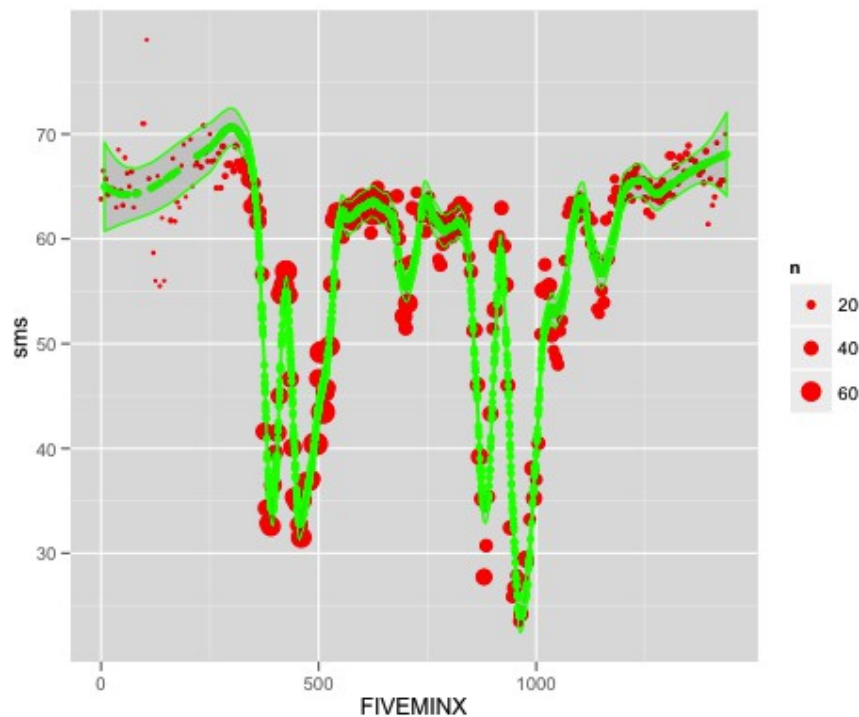


Figure 20: LOESS Confidence Intervals with 5-minute population mean

Figure 20, above shows a LOESS fit to sampled observations of space-mean-speed. The green shaded area indicates the confidence interval for the fitted curves. The red points indicate the population mean at 5-minute intervals with sample size indicated by the size of the points.

5.4 Comparison of Local Averaging and LOESS

The local averaging and LOESS methods of benchmark estimation have different strengths and weaknesses. Local averaging is technically a simple method and can be done using simple tools such as a spreadsheet program. LOESS requires more sophisticated tools like a statistical programming language such as R.

On the other hand, local averaging produces a fairly jagged estimation of benchmark values with the need for linear interpolation between points. LOESS solves this problem by applying second order polynomial regression functions between points and results in a smooth interpretation of the benchmark values.

The choice of methods depends on the tools available to the analyst but also the requirements of the evaluation. In order to better understand the relative performance of each method the RMSE of each method was estimated against a sample of 10 days of AVI data taken from the Houston data set. Each day's sample was first reduced by randomly selecting 5, 10, and 15% of the observations. This sample was treated as the data that would be "visible" to the algorithm. The full data set was used to compute the "true" benchmark values for comparison with each benchmark estimation procedure. The benchmark values were computed at 5 and 15 minute intervals. The RMSE results for the Local Average and LOESS methods are shown below in Table 5.2.

Sample Size	Local Average		LOESS	
	5-minutes	15-minutes	5-minutes	15-minutes
5%	3.9	3.1	3.7	3.6
10%	3.3	2.6	3.3	3.5
15%	3.1	2.1	3.2	3.5

Table 5.2: RMSE (mph) Comparison

In addition to the RMSE comparison, it was also useful to compare these methods by the

percentage of time-intervals where an estimate could be computed. The local averaging method may fail to report a value if no observations are available for that time window. LOESS, on the other hand will estimate the value based on the fitted polynomial regression function.

Results of the data availability evaluation are shown below in Table 5.3.

Sample Size	Local Average		LOESS	
	5-minutes	15-minutes	5-minutes	15-minutes
5%	60%	83%	96%	96%
10%	77%	92%	99%	98%
15%	84%	95%	99%	99%

Table 5.3: Data Availability (% of intervals)

5.5 Residual Analysis

A residual analysis was conducted for the LOESS models to verify that they met the standard assumptions of independent and normally distributed errors. The analysis indicates that for the data evaluated, the LOESS models do meet these assumptions. The normal quantile-quantile plot, shown below in Figure 21 indicates that the residuals follow an approximately normal distribution.

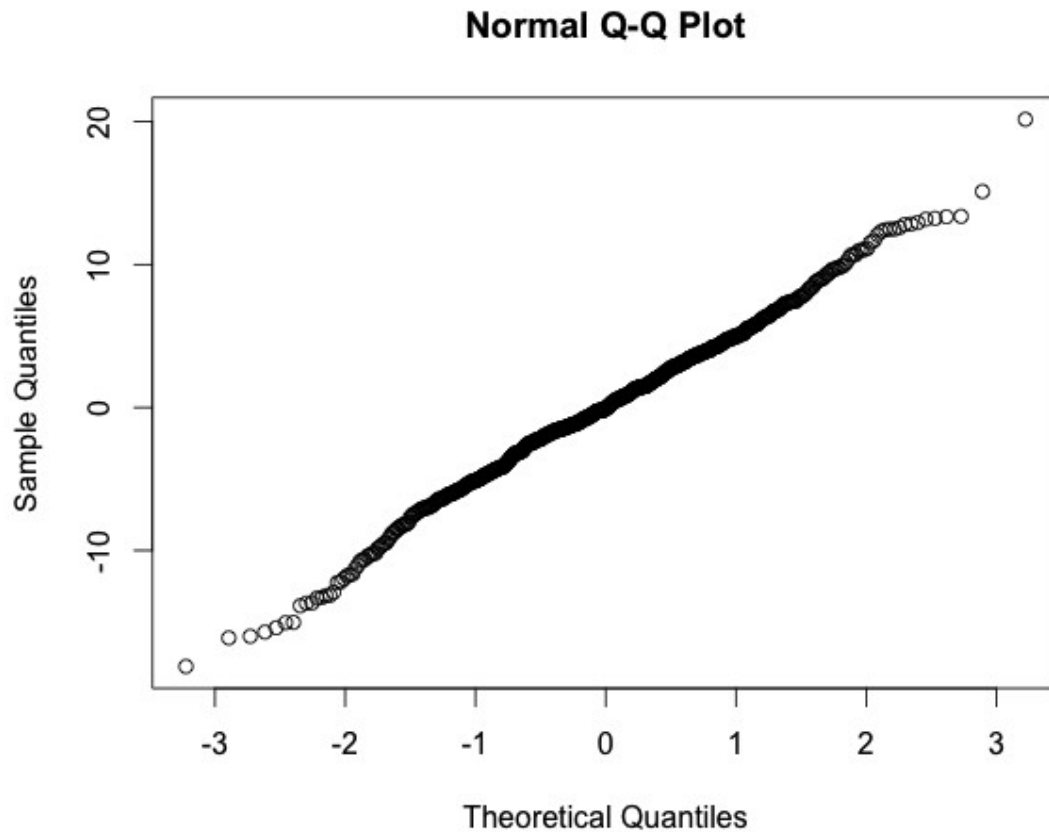


Figure 21: Quantile-quantile plot for a LOESS model

5.6 Potential Pitfalls with the LOESS Method

The LOESS method of fitting to time-series data works well with sparse data but can be sensitive to the value of lambda (smoothing parameter). In very sparse data sets a value of lambda that is too small will create a situation where the number of data points available to estimate a fit are fewer than the effective number of parameters. This can occur, for example, when there are fewer than 30 observations in a day and lambda is set at 0.10. In general, the highest value of lambda that does not "oversmooth" should be selected for the fit. In cases where, for example, a sensor failed and only a

small number of observations were reported in the course of a day, it may be necessary to adjust the lambda parameter upwards or to consider invalidating that day's data collection. An example of a sparse data set of 65 observations of space-mean-speed in one day (about 1 observation every 20 minutes on average) with the fitted LOESS curve with $\lambda = 0.15$ is shown below in Figure 22.

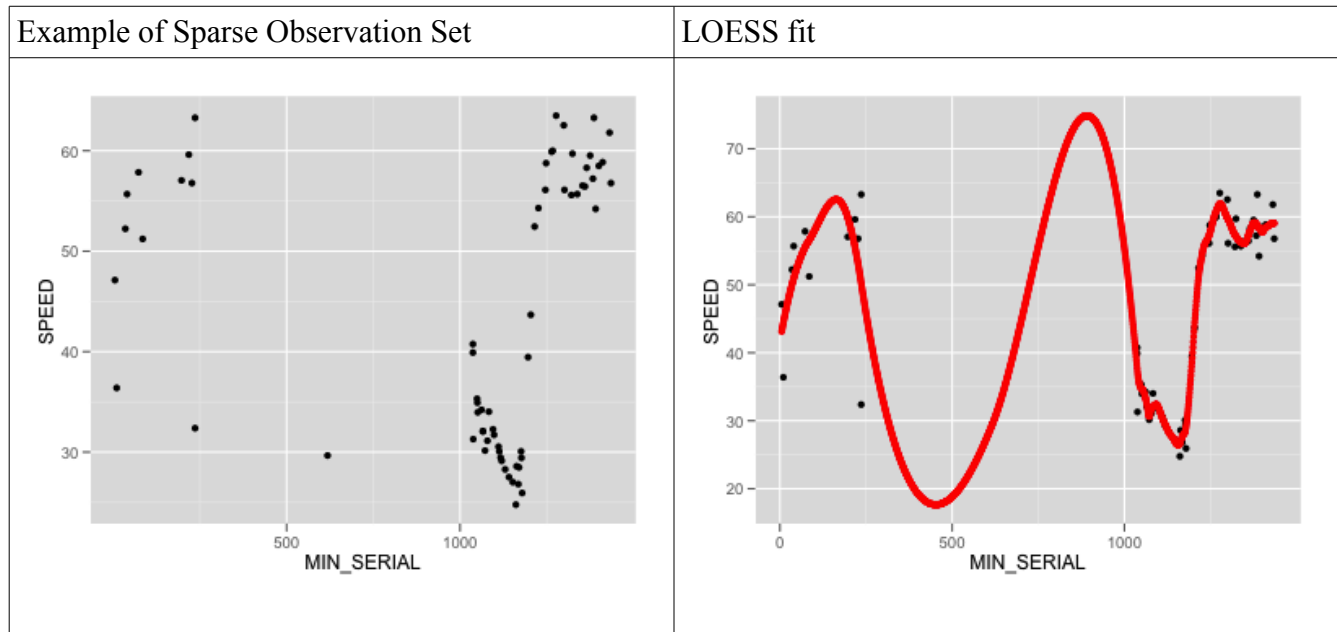


Figure 22: Example of LOESS with sparse data

It can be seen that for some unknown reason the sensors recorded very few observations in the morning and the middle of the day and then began to record more observations late in the day. In this situation, the LOESS function will diligently fit a curve to the existing observations. However, for the purposes of benchmarking, such a set of observations would not be appropriate. A simple test to avoid this situation would be to require that there be at least a few observations during each hour of monitoring for the data to be considered valid. Determining appropriate data densities for fitting LOESS curves to re-identification data would be a useful area of further research.

5.7 Conclusions of Travel Time Benchmarking Methods

The simplicity of calculation with local averaging comes at the expense of accuracy and coverage when data is sparse and benchmark estimates are required at short time-intervals. The local averaging method requires that sample observations be available in order to compute an average for the current time-interval. When data are sparse or frequent updates of the benchmark are required then the LOESS method is superior to local averaging. Local averaging does offer lower RMSE when data is abundant and benchmarks are not required at frequent intervals. However, even with greater density of data and larger time-interval bins, the coverage of local-averaging is inferior to LOESS. The trade-off between the methods appears to come down to a consideration of technical complexity and relative accuracy of the point estimates. The local averaging method offers simplicity and accuracy at the expense of poor coverage and poor performance in sparse data and frequent time-intervals. LOESS requires more computational complexity yet offers a robust method of benchmarking that performs well in sparse data and near real-time applications.

Chapter 6 Link Selection by Maximum Entropy

In many past data quality evaluations, links were selected by “expert opinion” such that the selected links represented a subset of interest to the evaluating agency. This type of sampling design can be categorized as a judgment sample which is a form of non-probability based sampling. Judgment sampling is typically used in cases where randomized designs are impractical or not feasible. However, one of the main disadvantages of a judgment sample is that it does not satisfy any optimality criteria for parameter estimation such as minimization of bias and variance. Another disadvantage of judgment sampling is that there is no quantifiable metric or objective basis for selection of a link beyond the expert's opinion.

An alternative approach that should be considered is to select a subset of links from the network that maximizes the amount of “information” in a sample drawn from those links. This approach is called Maximum Entropy Sampling (MES) and is based on Shannon Entropy which can be used to quantify the amount of information in a signal.

6.1 Introduction to Maximum Entropy Sampling

Maximum Entropy Sampling designs have been studied by a number of different researchers. Shewry & Wynn proposed MES for the optimal design of spatial sampling.[23] Ainslie, Reuten, et.al applied an entropy-based optimization technique to the design of a monitoring network for air pollutants.[24] Bueso & Angulo have also studied MES designs for optimal design of spatial monitoring networks.[25] Stohlgren, et. al. used maximum entropy sampling to develop methods for optimal selection of sites for monitoring environmental conditions over a large geographic region in the United States.[26] Husain and Khan applied Shannon entropy to the design of an optimum air quality

monitoring network in Saudi Arabia. [27]

Shewry & Wynn note that often the goal of experimental design is to acquire as much information as possible about a system parameter through a limited number of observations of a system response. With this perspective in mind, the optimality of an experimental design can be judged by the expected information content of the experiment.

According to information theory, the information contained in a random event is inversely related to the probability of the event. In other words, the more likely it is that an event occurs the less information it provides. Mathematically, the information associated with an observation of X is simply:

$$I(X=x) = -\log[P(X=x)] \quad (13)$$

where $P(X=x)$ is the probability that X takes the value x .

The fundamental quantity used in MES designs is Shannon Entropy which is the expected value of the information content of a random variable X . Shannon Entropy is typically written as:

$$H(X) = E_X[I(X)] = E_X[-\log f(X)] \quad (14)$$

for a random variable, X with probability density/mass function $f(X)$.

Now, consider the case of a system that is comprised of a set of jointly distributed random variables.

Let $S = \{X_1, X_2, \dots, X_n\}$ be the set of random variables that describe the system.

Shewry & Wynn show that if we consider the full joint distribution of the random variables in S , the entropy contained in the system is a finite and deterministic quantity. Therefore, if we partition the system into two sets, the sum of the entropy contained in each set will equal the entropy contained in the system.

Following this line of reasoning then, the entropy of the system can be expressed as:

$$Ent(X_S) = Ent(X_s) + E_{X_s}[Ent(X_{\bar{s}} | X_s)] \quad (15)$$

where X_S is the full joint distribution of X
and X_s is the subset of interest

This equation can be interpreted as meaning that the total entropy in the system is equivalent to the sum of the entropy in the selected subset, X_s , plus the expected value of the conditional entropy in the remaining subset given the selected subset. Thus, an optimal design should **minimize** the entropy in the remaining (i.e. unobserved) subset by **maximizing** the entropy in the observed subset.

For the Gaussian case, the entropy in the system can be expressed as:

$$Ent(X_s) = \text{constant} + \log \det(\Sigma_s) \quad (16)$$

where \det is the matrix determinant
and Σ_s is the covariance matrix of X_s

The covariance matrix, Σ , is an “N x N” square matrix which represents the variance of each random variable as well as the covariance of each joint distribution in the system. The covariance matrix also has some convenient properties that make it ideal for computation in optimization problems.

Using the fact that system entropy is a function of the determinant of the covariance matrix, Shewry & Wynn then showed that an optimal sampling design is a problem of maximizing the entropy contained in the selected subset of the system. Or in other words, the problem is equivalent to maximizing the determinant of the covariance matrix of the sampled variables. Therefore, the optimization problem becomes:

$$\text{Select } s \subset S \text{ such that } \arg \max_s \log \det(\Sigma_s) \quad (17)$$

6.2 Link Covariance

The application of MES designs to TIS data quality evaluations requires that the subset of links with the largest log-determinant of the system covariance matrix be selected for sampling. Of course, the availability of a suitable covariance matrix to represent the joint distribution of space-mean-speed across all links in the network is critical to the success of this method. Therefore, the key challenge in

applying this technique to TIS data quality evaluations is to find a good estimate of the covariance matrix.

There are three approaches to estimating the covariance matrix that can be considered:

- 1) Model-based estimation
- 2) Estimation from empirical data and extrapolation
- 3) Estimation from TIS historical archives

While analytical models of link travel time can be developed to estimate the covariance matrix they suffer from the problem that it is difficult to calibrate an analytical model representing just a single link in a network. The problem of calibrating a model to represent travel time variance over several hundred links in a regional network would be enormous and likely to be very error prone. For a smaller network, it should be possible to estimate the covariance matrix from a simulation if one is available. But developing a simulation model for a large geographic region would be a huge undertaking.

On the other hand it is possible to estimate the covariance matrix by collecting empirical data from a representative subset of the links in the network and extrapolating. This approach offers the advantage of estimating the covariance matrix from measured data but requires some way to extrapolate measurements to unmonitored links. Extrapolation from empirical data would likely result in a poor estimate unless a large enough subset of links were monitored.

However, if a sufficiently large data set of estimates from a Traveler Information System is available from a data archive, then the covariance matrix derived from the archive should be a close approximation of the “true” covariance matrix. This approach offers many of the advantages of a model-based approach in terms of link coverage with some of the advantages of empirical estimation if the TIS is considered to be reasonably accurate in its estimates.

6.2.1 Computation of Covariance Matrix and Determinant

Travel time data is usually stored and managed in a relational database system. One of the challenges in computing the covariance matrix is to transform the data from the relational structure (i.e. row-based) to a column-based structure. In effect this de-normalizes the normalized data structure commonly used for storage in relational databases.

In the relational structure, travel time or space-mean-speed estimates are keyed by the unique combination of LINK_ID and TIME_ID which represents the point in time and space that the estimate was made. This data structure must be transformed so that each link is represented by a column and the speed observations for each link are stored in the relevant column and ordered by the TIME_ID key.

Relational data structure

LINK_ID	TIME_ID	SPEED
X1	T1	60
X2	T2	45
X3	T3	50

Measurements are keyed by unique identifiers for space (LINK_ID) and time (TIME_ID).

De-normalized matrix structure

TIME_ID	X1	X2	X3
T1	60		
T2		45	
T3			50

Each column represents the measurements of a single link. Each row represents measurements at all links at a single point in time.

For this problem we used the R-language for data transformation and statistical calculations. A package called “reshape” can be used to transform a relational structure to a column-based structure.

Once the relational data has been transformed into a column-based structure, the covariance matrix can be easily calculated using the “cov” function in R or any statistical language. The covariance matrix of a multidimensional distribution is specified as:

$$\Sigma = E[(X - E[X])(X - E[X])^T] \quad (18)$$

where Σ is an $N \times N$ square matrix
 X is an $M \times N$ matrix of observations
 with M rows representing the time-intervals
 and N columns representing the links in the network

Recall that, in the covariance matrix the diagonal elements represent link variance and the off-diagonal elements represent the covariance of any two link pairs.

Once we have a covariance matrix for the network we then need to compute the determinant. Algorithms for computing the determinant of a square matrix have been implemented in a number of popular scientific programming languages such as R and Matlab. For this research we used the R programming language which uses the LU decomposition algorithm to compute the determinant.

The determinant of the covariance matrix is a scalar value which can be interpreted as the amount of “disorder” in the system.[29] If we choose an arbitrary subset of links from the network, we can represent the covariance matrix for the subset by selecting out the relevant cells from the full covariance matrix. Thus, the problem becomes an iterative problem where we exhaustively search through all possible subsets in the network to find the subset with the largest determinant of the covariance matrix.

For a relatively small network this not a computationally challenging problem. The number of possible subsets of size n in a network of size N can be computed as “N choose n” or $\frac{N!}{n!(N-n)!}$.

For problems where N is large (e.g. > 40), as n increases the sample space quickly becomes too large for an exhaustive search and more sophisticated optimization techniques are required. Although a

discussion of optimization techniques is beyond the scope of this research there are excellent papers by Ko, Lee and others investigating this topic.[30], [31]

An overview of the process using an exhaustive search is shown below.

1. Retrieve relational data for all links in the network over desired time period and transform from relational structure into column-based structure
2. For some arbitrary subset size (i.e. n) enumerate all feasible subsets of links in the network
3. Compute the determinant of the covariance matrix for all possible subsets and choose the subset with largest log-determinant

6.3 Evaluation of Maximum Entropy Sampling Method

The focus of this research was on the application of MES design to the problem of optimal link selection for ground-truth data collection in a TIS data quality evaluation. Therefore, two aspects of this sampling design were investigated. First, we investigated the performance of this sampling method in selecting an optimal subset of links using simulated data. The simulated data was designed to mimic a real-world transportation network and a noisy Traveler Information System. We defined an optimal subset of links as the subset which sampled the largest proportion of TIS error in the system.

Next, MES was evaluated using empirical data collected in the field. The empirical data was used to assess whether the MES method would show similar performance characteristics as with the simulated data. The goal of this evaluation was to determine whether a real-world TIS would generate a suitable covariance matrix for determining an optimal subset of links.

During each time-interval the TIS error was calculated as the squared error in units of $(sec/mi)^2$. This measures deviations in units of time (seconds) per unit distance (mile) which is arguably more important to users of a TIS than errors in speed (mph). For example, a 5 mph error when the ground-truth speed is 30 mph is quite larger in units of time than a 5 mph when ground-truth speed

is 60 mph. Thus, we prefer to measure errors in units of time per unit distance. The expression for TIS errors is given below.

$$E_{i,j} = (3600/u_{TIS} - 3600/u_{GT})^2 = (\text{sec/mi})^2 \quad (19)$$

where $E_{i,j}$ is the squared error of link i at time j
 and u is the space-mean-speed (mph) of each source
 and 3600 is in units seconds per hour

6.3.1 Measuring the Performance of MES

One way to measure the performance of this sampling design is to measure the percentage of total observable TIS error that is present in a sample. By this measure, subsets of links which sample a higher proportion of total TIS error have better performance.

We can compute the sum of the squared errors for all links in the network as:

$$SS_{\text{Network}} = \sum_{i=1}^N \sum_{j=1}^M E_{i,j} \quad (20)$$

where SS_{Network} is the sum of the squared errors

N is the number of links in the network

M is the number of time-intervals and $E_{i,j}$ is the squared error on link i at time j

The sum of squared errors for a selected subset can be computed similarly as:

$$SS_{\text{subset}} = \sum_{i=1}^n \sum_{j=1}^M E_{i,j} \quad (21)$$

where n is the number of links in the subset

M is the number of time-intervals and $E_{i,j}$ is the squared error on link i at time j

The sum of squares for the selected subset can be divided by the sum of squares for the network. This ratio gives a percentage of total system error that is sampled by the selected subset.

$$\% \text{ of } SS_{\text{Network}} = \frac{SS_{\text{subset}}}{SS_{\text{Network}}} \quad (22)$$

This quantity is similar to the coefficient of determination (R-square) often used to evaluate

regression models.

6.3.2 Simulated Data Set Design

We first evaluated the performance of MES against simulated data where ground truth speeds followed one of 5 typical “speed profiles” and TIS speeds were determined by a stochastic model. Using this approach it could be seen how much of the TIS error could be sampled from the subsets selected by the MES algorithm.

A network of $N = 40$ links was simulated where ground-truth speed for each link followed one of 5 speed profiles. Table 6.1 describes the speed profiles.

Speed Profile	Description
No congestion	Speeds between 55 – 65 mph
AM congestion	Speeds slow down to 25 mph during AM peak only
PM congestion	Speeds slow down to 25 mph during PM peak only
Random	Starting at 60 mph speeds can fluctuate +/- 5 mph during each time-interval
All congestion	Speeds between 25 – 35 mph

Table 6.1: Description of Speed Profiles

A graphical example of the speed profiles is shown below in Figure 23.

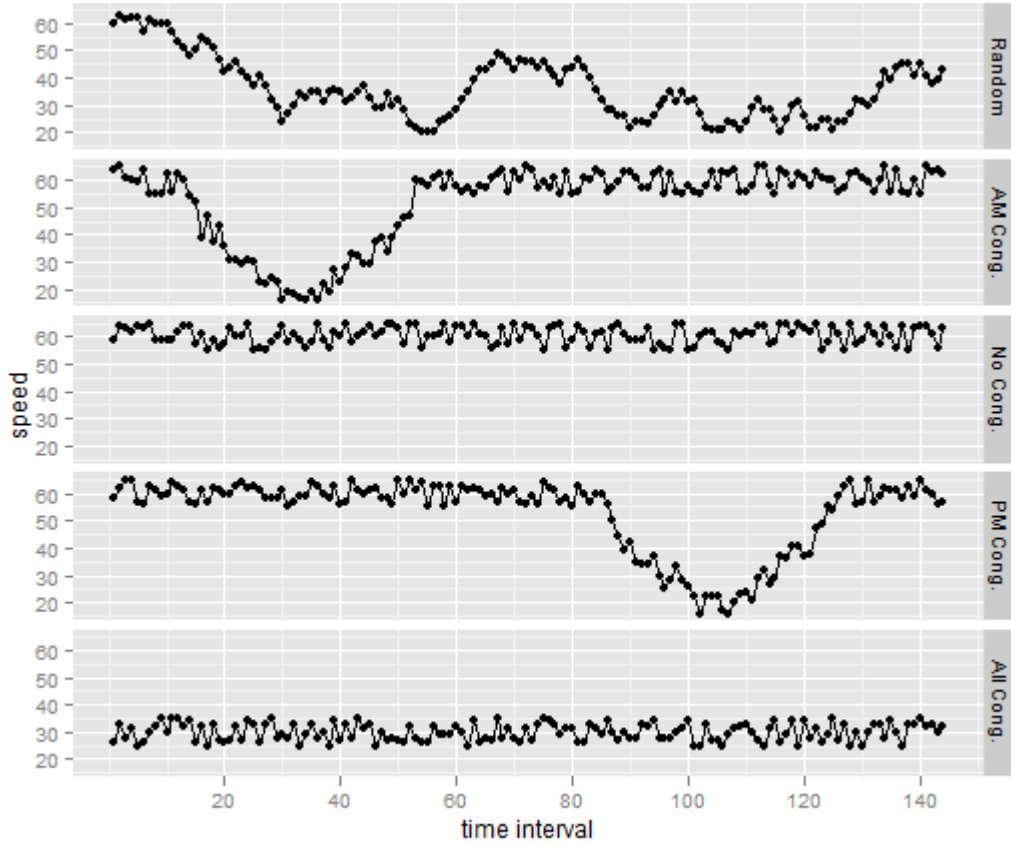


Figure 23: Ground-truth Speed Profiles

For each link in the network a series of $M = 144$ values of space-mean-speed was generated. The 144 values represented a typical day's worth of data where each value would be the estimated space-mean-speed over a 5-minute interval. This would therefore cover 12 hours of data.

The TIS was modeled as a noisy estimate of the ground truth speed. The model was specified as:

$$u_{TIS} = u_{GT} + \epsilon \quad (23)$$

where u is the space-mean-speed
 and $\epsilon = N(\mu, \sigma^2)$ is normally distributed noise
 with bias = μ and variance = σ^2

The TIS speed was bounded between 10 and 75 mph so that a large random error would not create an unrealistic estimated speed (e.g. -5 mph or 100 mph).

An example of the speeds from a TIS modeled as $N(0,4)$ is plotted against ground truth for 3 simulated links shown below in Figure 24.

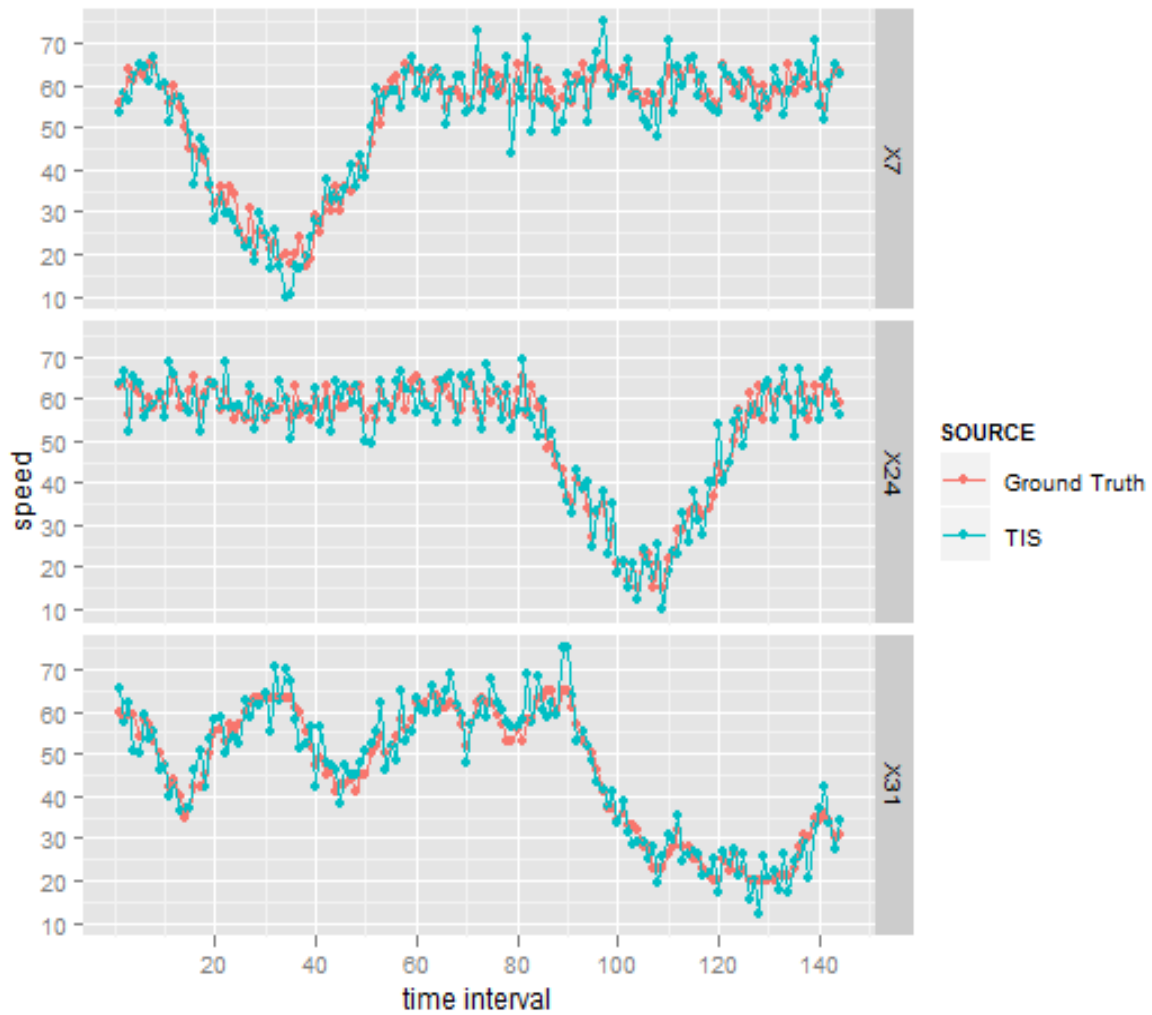


Figure 24: Example of TIS and Ground Truth Speeds

The figure shows that the simulated TIS gives a noisy estimate of the ground truth speed and reflects conditions similar to real-world conditions.

6.3.3 Experimental Design

The MES method was tested on a range of simulated data sets. We evaluated the sampling

method using subsets of size $n = 3 - 5$ links in a network of 40 links. For a subset of size $n=4$ in a network of $N=40$ links there are 91,390 possible subsets that can be selected. A subset of size $n=5$ would yield 658,008 possible subsets. Recall, that for each subset the covariance matrix and log-determinant for that subset must be calculated.

To illustrate the performance of this sampling method we show the results of three experiments using simulated data. The experiments are described in Table 6.2 below.

Table 6.2: Description of Experiments

Experiment	Description
E1	Only "no congestion" speed profile selected with TIS modeled as $N(0,4)$.
E2	Equal probability of all speed profiles with TIS modeled as $N(0,4)$.
E3	Equal Mix of AM, PM links with TIS modeled as $N(2,4)$.

6.3.4 Experimental Results

Table 6.3, below, shows the results of the three experiments for subsets of sizes $n = 3 - 5$. The table compares the percentage of total observable error in the MES sample with the “best” sample. Recall, that we wish to maximize the amount of TIS error in the sampled links.

Table 6.3: Experimental Results

Experiment	Subset size (n)	Max Log-determinant	% of Observable Error in Selected Subset	Max % of Observable Error among all Subsets	Ratio of Selected Subset to Max Subset	Percentile Rank of Selected Subset
E1	3	10.6	9.2	9.4	0.98	99
	4	14.11	12.1	12.5	0.96	99
	5	17.38	14.5	15.1	0.96	99
E2	3	22.61	12.3	17.4	0.7	88
	4	29.45	18.9	25.4	0.74	96
	5	36.61	21.5	26.9	0.8	97
E3	3	21.72	12.1	13.8	0.87	99
	4	28.18	16.1	18.5	0.87	99
	5	34.56	19.8	22.9	0.86	99

The results show that in all three experiments the MES design selected subsets of links that sampled more TIS error (Equation 10) than 90% of all other subsets (by Percentile Rank). The one exception was Experiment E2 at subset size $n=3$. In this experiment the MES design selected a subset that ranked at the 88th percentile among all possible subsets and sampled 70% of the maximum observable error. However, as the subset size increased in this experiment it can be seen that the Percentile Rank also increased. In general, it appears that the MES design can be expected to sample more observable error than approximately 90% of other feasible subsets for a range of experimental conditions and subset sizes.

On average, the ratio of TIS error observed in the MES sample to maximum observable error was 0.86. These results indicate that across a range of experimental conditions the MES design selected subsets of links that sampled approximately 86% of the maximum observable error that could be sampled at that subset size.

6.3.5 Discussion of Results from Simulated Data

The Maximum Entropy Sampling method was used to select a most informative subset of links from a simulated network. Using only the data from a TIS modeled as a stochastic function of the ground truth speed, the sampling method consistently selected a subset of links which captured a large percentage of TIS error among all possible subsets of the same size. To illustrate these results two plots of TIS speeds from Experiment E2 (subset-size $n=5$) are shown below in Table 6.4. In each plot, the speeds reported by the TIS are shown for the selected subset. The first plot shows the Maximum Entropy subset and the second plot shows the Minimum Entropy subset. The plots illustrate that the Maximum Entropy Sampling design selected a subset that sampled a wide variation of link speeds whereas the minimum entropy solution selected links with very little variation in speeds.

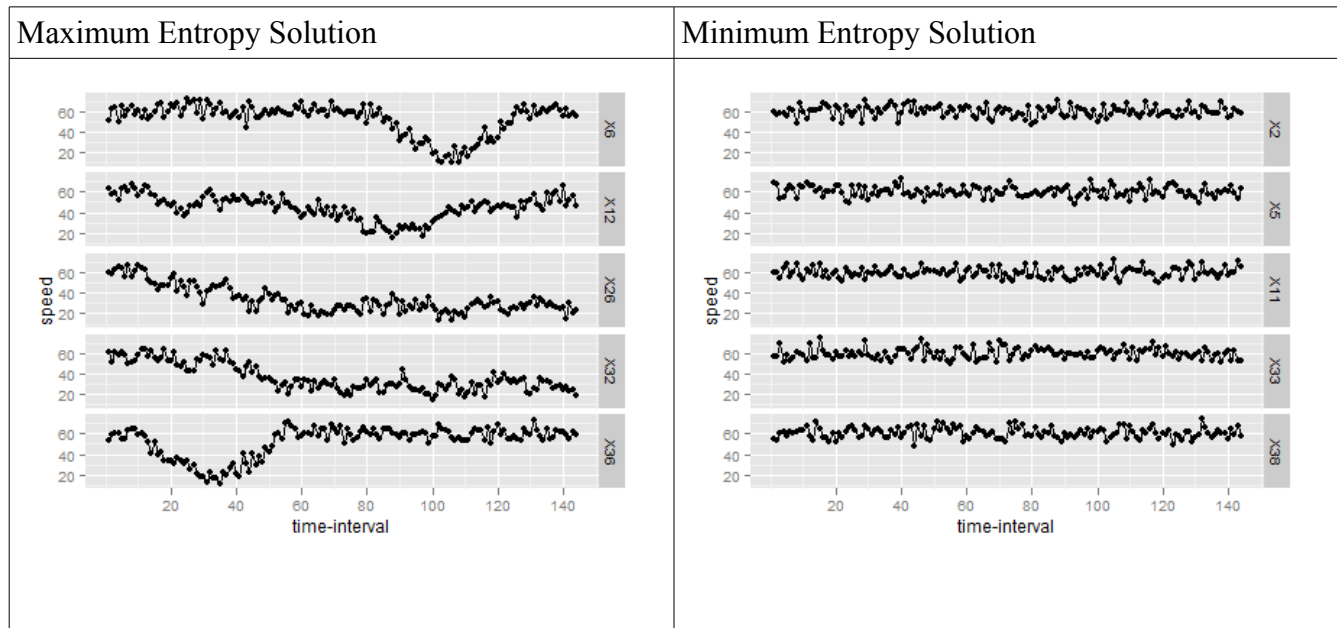


Table 6.4: Experiment E2: TIS Reported Speeds for Maximum and Minimum Entropy Solutions

6.4 Evaluation of MES using Empirical Data From Northern Virginia

6.4.1 *Description of Study Area and Data*

After evaluating MES using a simulated data set we evaluated the method using empirical data. We collected estimates of travel time from a private-sector TIS which provides data to the Virginia Department of Transportation (VDOT). For ground-truth estimates, a set of 12 links in Northern Virginia was selected. The links were selected based on the availability of Bluetooth sensors deployed on I-66 and I-95. The speed data from the TIS corresponds to a link identification system called “Traffic Message Channel” (TMC). The estimates from the TIS were combined by TMC to compute an overall estimate for the corresponding Bluetooth link. Table 6.5, below, provides an overview of the selected links including the starting and ending TMC identifiers.

BT_ID	Road	Dir	TMC Start	TMC End	LENGTH (mi)
2521	I-66	EB	110-04185	110N04181	10.5
2522	I-66	WB	110+04182	110+04186	10.5
2523	I-66	EB	110N04181	110N04178	4.8
2524	I-66	WB	110+04179	110+04182	4.8
2529	I-66	EB	110N04175	110N04163	10.3
2545	I-66	WB	110+04179	110+04186	15.86
2547	I-66	EB	110N04181	110N04175	10.12
3052	I-95	NB	110+04148	110+04158	18.4
3105	I-95	NB	110+04155	110+04158	14.5
3106	I-95	SB	110-04153	110N04151	3.3
3128	I-95	SB	110N04128	110-04120	8.09
3129	I-95	SB	110-04155	110N04148	15.06

Table 6.5: Bluetooth Link Descriptions

A map showing the road segments (I-66, I-95) and the position of the Bluetooth stations is shown below in Figure 25.

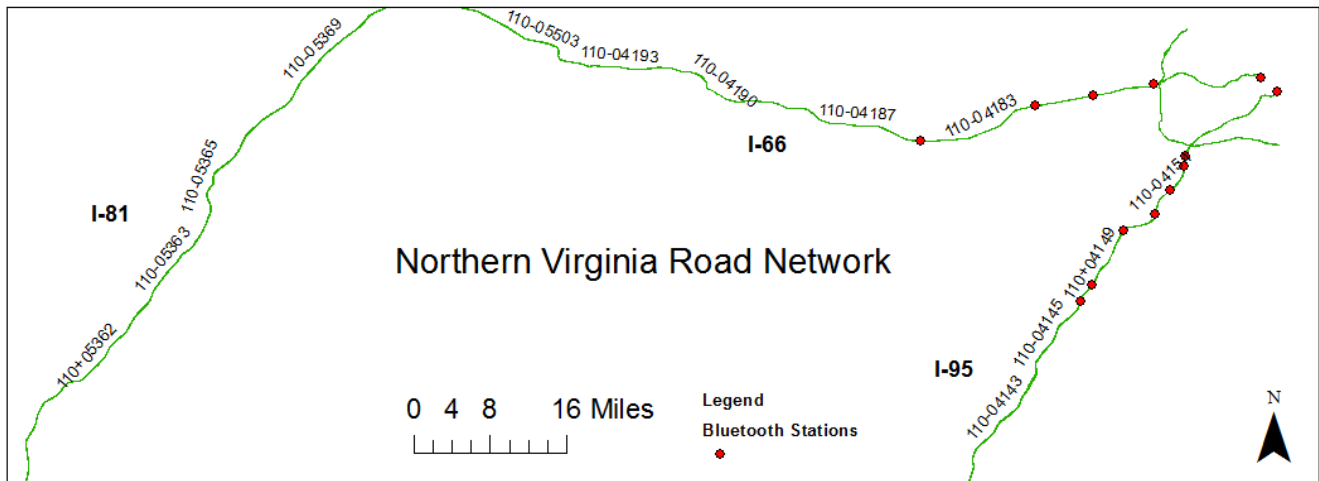


Figure 25: Location of Bluetooth Links

To estimate ground truth speed, the average travel time across each link was measured by Bluetooth reidentification samples. Each sample was aggregated at 5-minute intervals. The observations from the Traveler Information System were also aggregated at 5-minute intervals for comparison with the Bluetooth data. The starting and ending points of the Bluetooth links were determined by the position of the Bluetooth sensing equipment which was previously deployed in the field.

Data was collected on all links between December 5 - 9, 2011 and between 7 AM and 7 PM when there were generally sufficiently large sample sizes for computing the ground truth estimates. Plots of the TIS versus Bluetooth speeds for 3 of the links are shown in Figure 26, below.

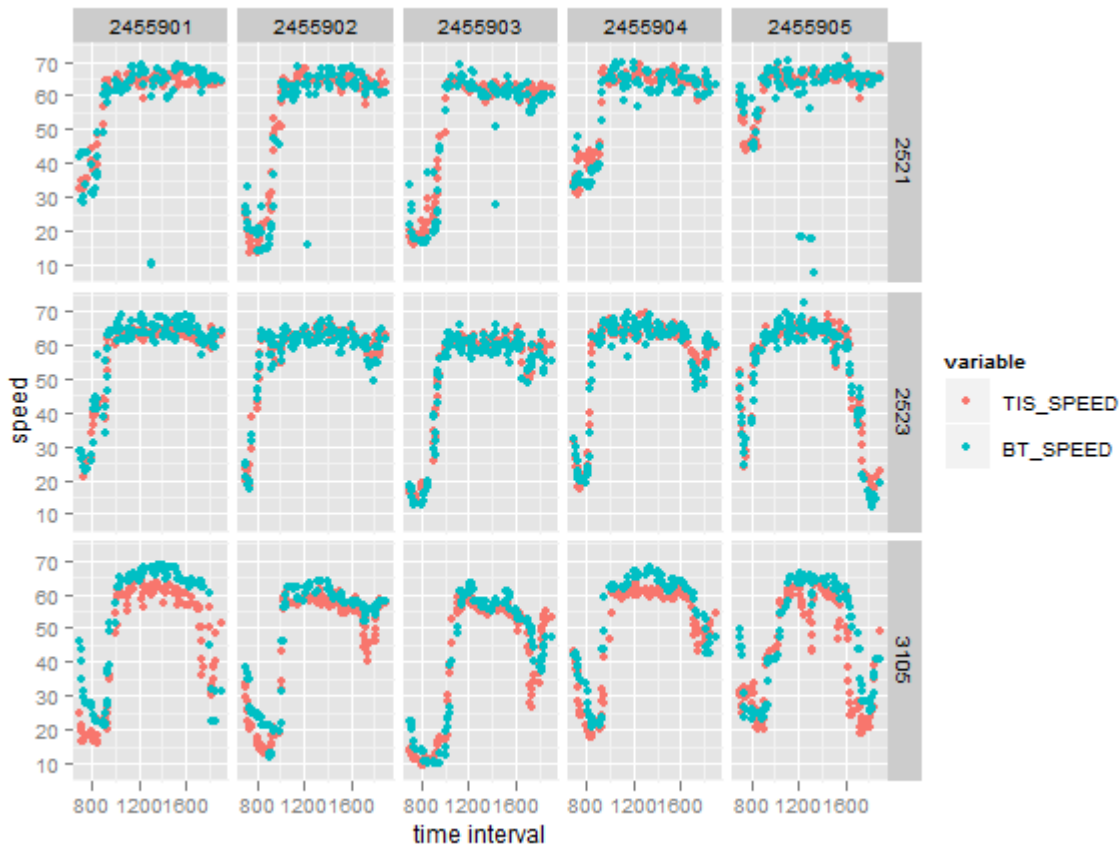


Figure 26: Comparison of Empirical TIS and Bluetooth Speeds on Selected Links

6.5 Results

In Table 6.6 we present the results of computing the optimal subset of links in the network of $N = 12$ for a range of subset sizes using the empirical data from the Bluetooth observations and the Traveler Information System. For a subset of size $n = 4$ there are 495 possible subsets and for $n = 5$ there are 792 possible subsets that could be selected.

Subset size	Max Log-determinant	% of Observable Error in Selected Subset	Max % of Observable Error among all Subsets	Ratio of Selected Subset to Max Subset	Percentile Rank of Selected Subset
3	24.74	47	50	0.94	98
4	31.62	48	62	0.77	90
5	38.36	64	75	0.85	97

Table 6.6: Empirical Results

The ratio of observed error to total observable error for the selected subset ranged between 0.77 and 0.94 using the empirical data. The MES design selected subsets which sampled more error than 90% of the other possible subsets in all three cases. This can be compared with the results from the simulated data where the ratio fell in the range of 0.7 – 0.98 and the percentile rank was greater than 90% in all but one experimental case.

These results show that the Maximum Entropy Sampling design using empirical data had similar performance characteristics to the simulated data set described earlier in the paper. The MES design consistently identified subsets of links that contained a relatively high percentage of observable TIS errors compared with a randomly selected subset of links. Figure 27, below illustrates this effect by showing a plot of Log-determinant versus % of Total Error for a subset of size $n = 5$. Each point in the plot represents one of the 792 possible subsets of $n = 5$ links in a network of $N = 12$ links.

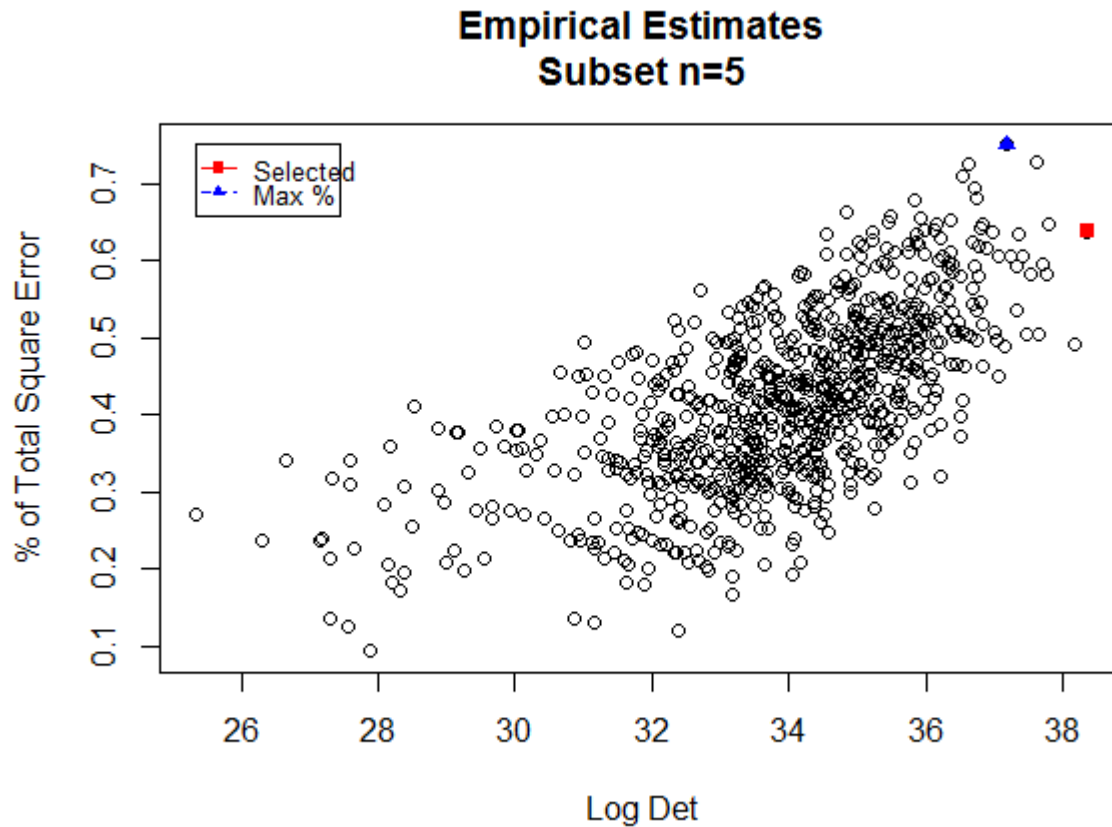


Figure 27: Empirical Results: Subset $n = 5$

6.6 Conclusions

The maximum entropy sampling design has been shown to be an effective method of selecting an optimal subset of links in a transportation network for ground-truth data collection in a data quality evaluation. The problem was shown to be an optimization problem of selecting the subset with the highest log determinant of the covariance matrix. This approach was applied to two transportation data sets to assess its effectiveness. First, MES was tested using a simulated data set where ground-truth speeds were determined by a specified speed profile and the TIS was modeled as a stochastic function of the ground-truth. Next, MES was tested against empirical data collected from a small transportation network in Northern Virginia.

The results show that MES designs can select subsets of links which sample a large percentage of TIS error. This method can therefore be used as an objective decision making criterion for choosing a set of links for ground truth data collection.

Chapter 7 Proposed Benchmark Link Selection Method

In the previous chapters, several factors contributing to variance in travel times were identified and a method of selecting links based on Maximum Entropy Sampling design was proposed. In this chapter a complete framework for selecting links in a network for data collection is proposed. The framework first uses network stratification to partition the network into different "strata" for benchmark data collection. Within each strata, the Maximum Entropy sampling design can be applied to select the most "informative" links.

7.1 Network Stratification

In Chapter 3, several roadway design factors related to variance in travel times were identified. Of these factors, roadway functional classification was shown to be the factor with the strongest relationship to travel time variance. In addition to this factor, the factors of ADT per lane, access point density, and link length were also identified as factors related to travel time variance. Among these four factors, roadway functional classification most clearly resulted in the largest shift in observed travel time variance. Among, the three other factors, no one particular factor stood out as contributing the most to travel time variance. However, a combination of the three factors appeared to have a weak relationship.

One strategy for network stratification would be to partition the network first by roadway classification and then by a combination of the three remaining factors. For example, the network could first be partitioned into freeways and arterials and secondarily into freeway segments based on the levels of the three remaining factors. The advantage of this approach is that it utilizes readily available roadway inventory data. However, there does not appear to be a very strong relationship between these

factors and travel time variance. Rather, travel time variance appears to be more strongly related to variance in average speeds.

The Maximum Entropy Sampling (MES) design uses estimates of average speeds from a Traveler Information System to select the most informative links from a set of candidate links. Therefore, this method can be applied after first stratifying the network by functional classification.

7.2 Selecting Candidate Segments

The MES method requires a set of candidate segments to evaluate. A TIS will provide estimates of travel time over a "base" set of links. Currently, the TMC standard defines the spatial extent of these links. In the future another standard may be possible. In many cases, the length of a TMC link will be less than one mile. This makes it too short for many benchmark estimation techniques such as BTR and toll-tag AVI. Therefore, TMC links will need to be appended together to form a longer segment for evaluation. This is done by spatial alignment, described in Chapter 4.

The choice of minimum and maximum segment length for benchmark data collection can be guided by the data collection technologies in use. For example, a reasonable lower and upper bound on segment length when using Bluetooth reidentification appears to be somewhere between 3 - 10 miles. Segments shorter than 3 miles may introduce too much measurement error. Links greater than 10 miles may begin to have problems with reduced sample sizes due to vehicles exiting the traffic stream or outliers due to vehicles stopping on the side of the road.

The "optimal" segment length will not be known until the entropy can be calculated. Therefore, a set of candidate segments should be created from the existing network. The candidate segments can be created by dividing the network into sets of segments of differing lengths. The network can be divided into sets of segments of length 5, 10, and 15 miles. These candidates are then collected together

and evaluated by calculating the entropy of each candidate.

7.3 Maximizing Network Coverage

A secondary goal of segment selection is network coverage. This is the percentage of the network that is covered by the monitored segments. If there are 10 links in the network, each 1-mile in length, then each link covers 10% of the network. All things being equal, a segment which covers more of the network should be preferred to a segment that covers less of the network. Thus, the percentage of network coverage for each candidate segment must also be computed.

7.4 Selecting a Solution

A set of segments that both maximizes entropy and coverage should be preferred to a set that does neither. In cases where a trade-off between entropy and coverage exists, a decision can be made based on a weighting function. For example, coverage could be valued more than entropy and the decision could simply be based on a linear combination of the entropy and coverage scores.

7.5 Case Study in Northern Virginia

To illustrate the application of this method a case study using link-speed estimates from Inrix for links in Northern Virginia was evaluated. In this case study, approximately 204 miles of freeway links were evaluated. The Inrix link segmentation follows the TMC standard and the average link length was 0.85 miles and there were a total of 238 links in the study area.

To apply this method, first, candidate benchmark segments must be identified. Routes were created by identifying continuous segments of TMC links that totaled approximately 4 or 10 miles in length. Each freeway in the study area was therefore divided into approximately 4 and 10 mile long segments. There were a total of 77 benchmark segments in the study area with 54 segments of 4 miles

in length and 23 segments of 10 miles in length.

For this case study, it was decided to identify an optimum set of 3 segments from the 77 candidate segments by applying two decision criteria. First, the entropy for each subset was calculated following the method described in Chapter 6. Second, the percentage of network coverage was calculated for each of the subsets. These two decision criteria were used to guide the choice of an optimal subset. By maximizing the entropy of the subset the most information would be from the benchmark sample set. And, by maximizing the network coverage, the largest physical extent of the network could be sampled.

The detailed results of the case study are presented below. The results indicate that an optimal subset which maximizes both entropy and network coverage can be found. To illustrate the results, three examples of subsets are presented below in Table 7.1.

Solution	Route IDs	Entropy	Network Coverage
1	40, 53, 68	16.14	16.25%
2	12, 53, 68	16.07	15.90%
3	1, 21, 26	7.79	5.16%

Table 7.1: Example of Solutions to Entropy and Coverage Decision Criteria

Solution #1, shown below in Figure 28 was the set of segments which maximized both entropy and network coverage. The selected segments included Northbound and Southbound sections of I-95 and a northbound section of I-495.

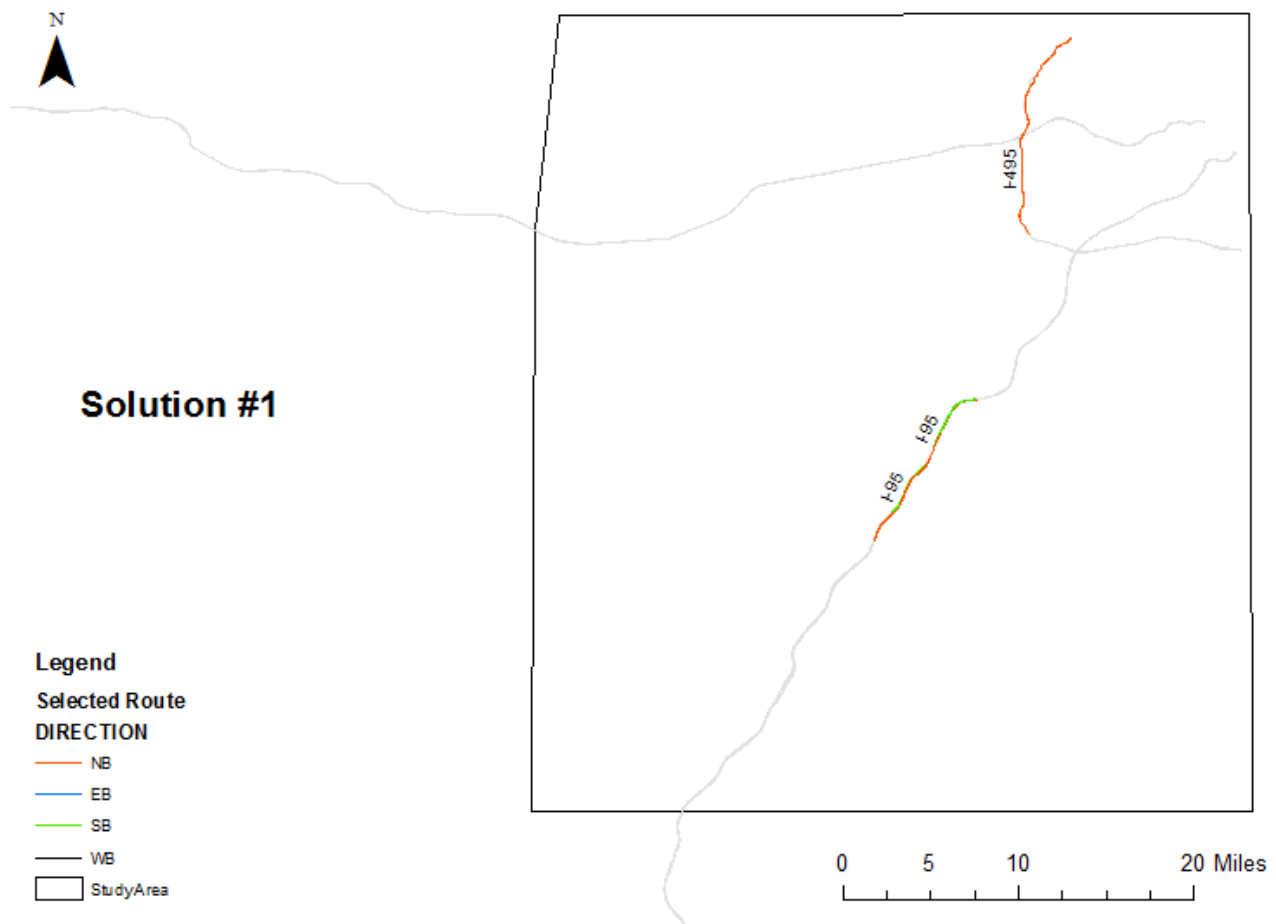


Figure 28: Link Selection -- Candidate Solution #1

Solution #2, shown below in Figure 29 had slightly less entropy and slightly less network coverage but included a segment from I-66. The selected segments included a Southbound section of I-95, an Eastbound section of I-66 and a Northbound section of I-495. This would be a candidate solution when, for example, it was required that I-66 be included in the solution.

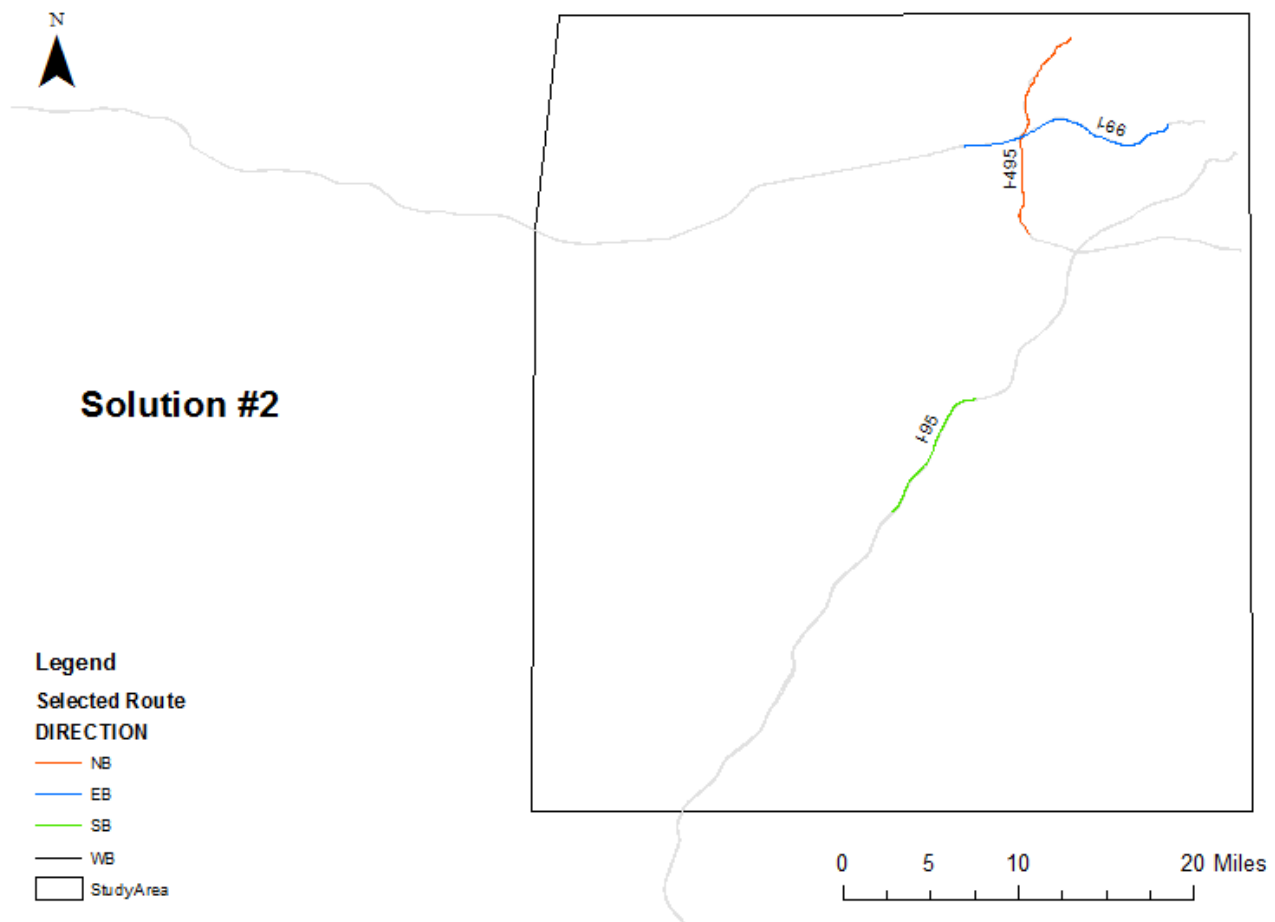


Figure 29: Link Selection -- Candidate Solution #2

Finally, to contrast solutions #1 and #2, we can look at a solution that did not have very much entropy and did not cover much of the network. Shown below in Figure 30 this solution included only Eastbound and Westbound sections of I-66 but the sections were relatively far west of the Beltway. This solution would be an example of a solution that did not provide much information gain from the sampled data.

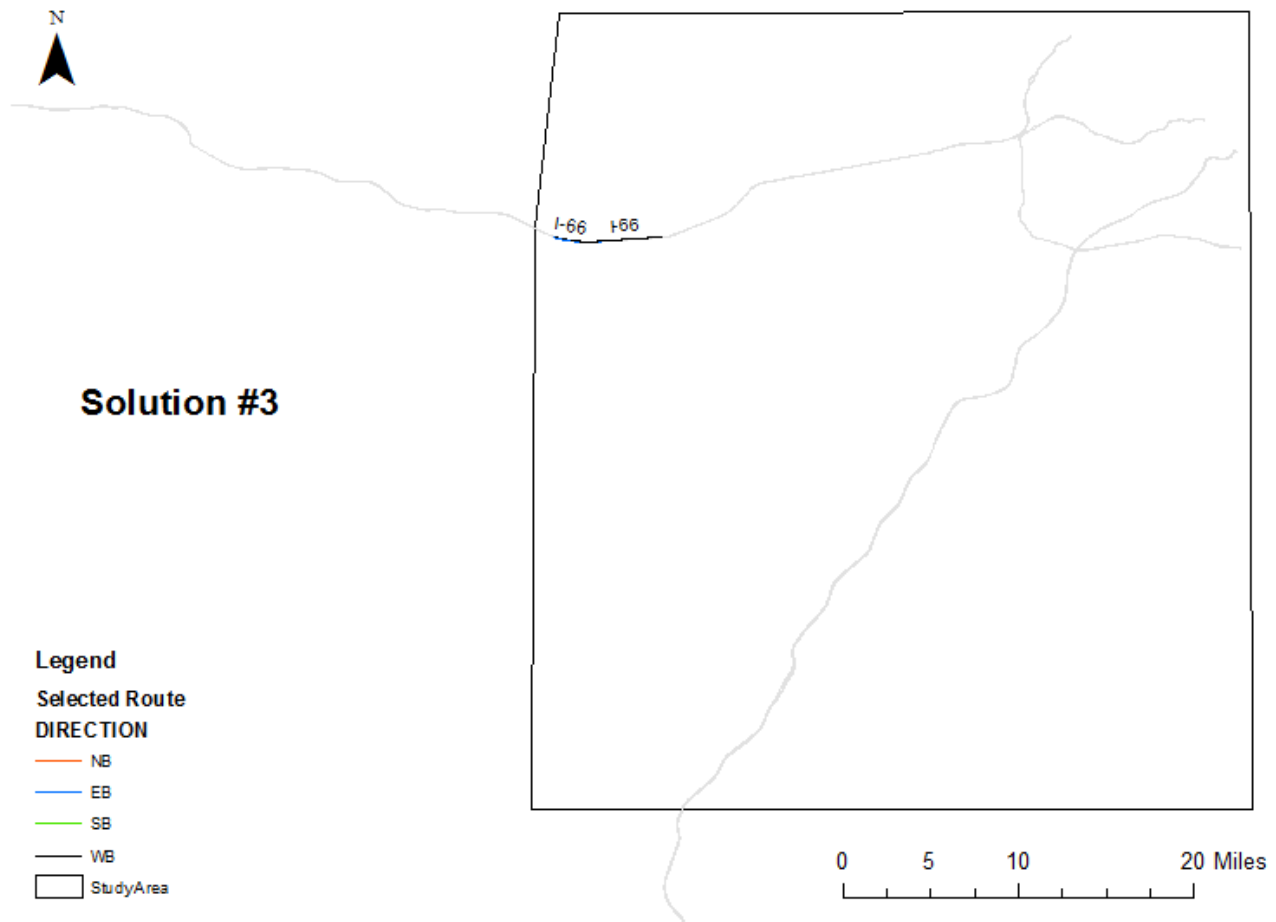


Figure 30: Link Selection -- Candidate Solution #3

7.6 Conclusions of Proposed Benchmark Link Selection Method

A complete benchmark link selection methodology was described in this chapter. By applying the Maximum Entropy sampling design described in the last chapter and combining it with a secondary decision criterion to maximize network coverage, a selection of links that provide maximal information and spatial coverage can be identified. This method can be used as an objective decision making method for selecting benchmark links for a data quality evaluation.

Chapter 8 Measuring Traveler Information System Errors

Errors are estimates of the distance between a TIS estimate and the benchmark value. There are a variety of ways in which errors can be measured and aggregated. This chapter will examine in detail the estimation of errors in a TIS data quality evaluation.

8.1 Error Definitions

Errors can be measured by a number of different functions. The most common error metrics include error bias, absolute error, squared error, relative error and absolute relative error. The definition of these error metrics is given below in Table 8.1.

Error Bias	$\hat{\theta} - \theta$
Absolute Error	$ \hat{\theta} - \theta $
Square Error	$(\hat{\theta} - \theta)^2$
Relative Error	$(\hat{\theta} - \theta)/\theta$
Absolute Relative Error	$ (\hat{\theta} - \theta)/\theta $

Table 8.1: Common Error Metrics

Each of these error metrics can be averaged such that the "average error" can be computed. For example, Mean Square Error (MSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) are all commonly used average error metrics for estimators.

8.2 Selecting an Error Metric

The choice of an error metric depends on the goals of the evaluation and the accepted data quality criteria. In Chapter 2, a review of past data quality assessments indicated that Average Error (Error Bias) and Mean Absolute Error (MAE) were the two most commonly used error metrics. In most

cases, errors were measured in units of miles-per-hour as these units are independent of link length and easily interpreted by stakeholders. However, most users of Traveler Information Systems are interested in the accuracy of estimates in travel time. There are several problems with assessing a system using Error Bias and Mean Absolute Error in units of mile-per-hour.

8.2.1 *A Simple Model of a Traveler Information System*

One of the simplest models we can assume about a TIS is that the TIS estimates space-mean-speed with some normally distributed error. This can be expressed as (**Model 1**):

$$\bar{u}_s^* = \bar{u}_s + \epsilon \quad (24)$$

where \bar{u}_s^* is the TIS estimate of space-mean-speed, \bar{u}_s is the ground truth
and $\epsilon \sim N(\mu, \sigma)$

Under this model the errors are normally distributed with a constant variance in units of speed (e.g. mph). However, it can be shown that under such a model, the travel time errors are not normally distributed and have an increasing variance as the space-mean-speed decreases. To illustrate this relationship, a simulation of 10,000 normally distributed errors $N(0,4)$ was created against a uniformly distributed sample of speeds between 20 and 80 mph. The error in travel time in seconds was calculated based on an assumed 1-mile link length. The results of the simulation are shown below in Figure 31.

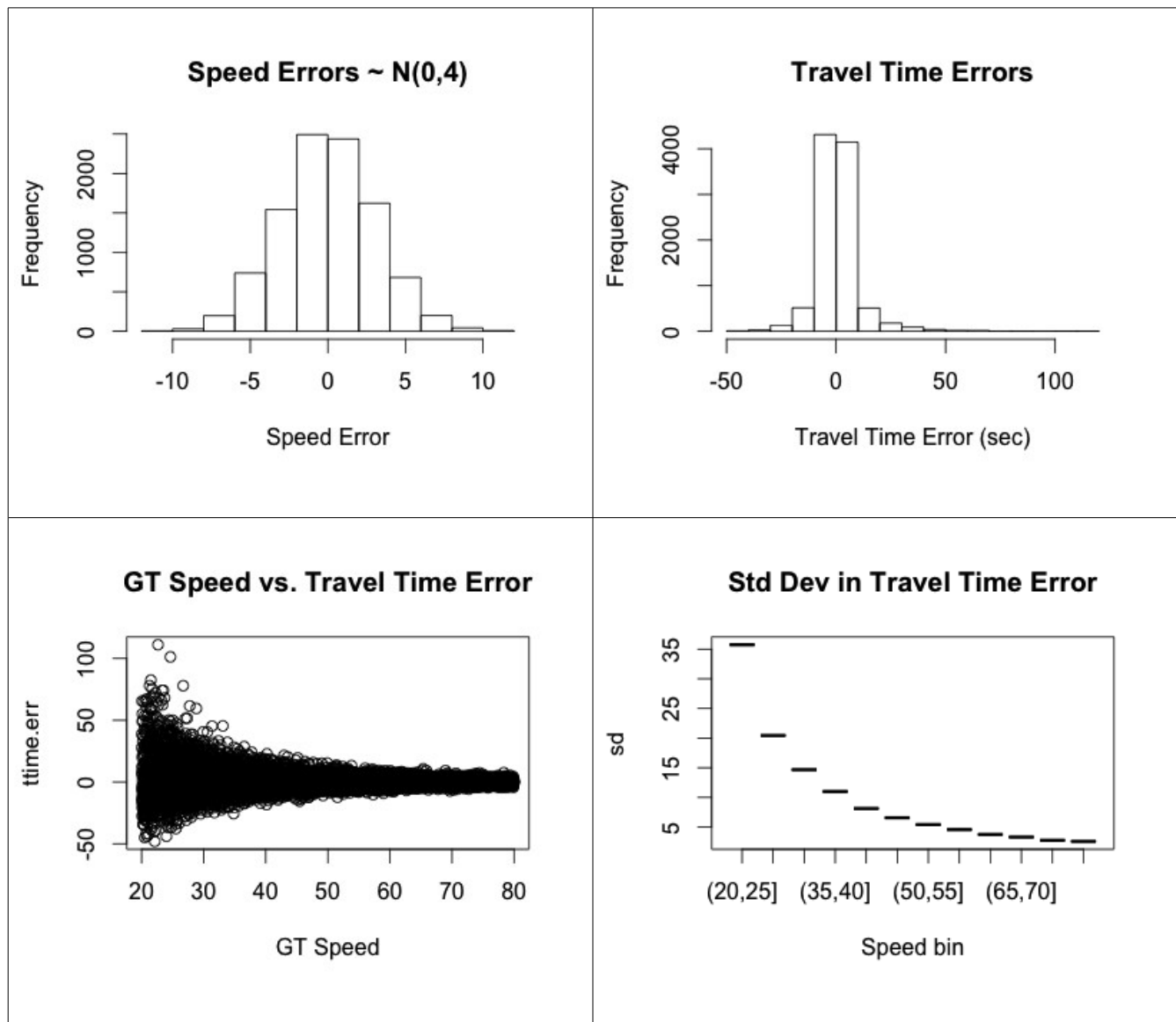


Figure 31: Speed and Travel Time Errors

The figures show that under the assumed model of a TIS that generates speed estimates with no bias and a constant variance, the variance in estimated travel times is not constant and increases as speeds decrease.

This leads to the important insight that estimates of speed error bias do not adequately describe the performance of a Traveler Information System. A system with no bias in speed but even a relatively small constant standard deviation (e.g. ≤ 4 mph) in speed will have a non-constant, increasing

variance in travel time errors as link-speeds decrease. This means that although on average the system will estimate travel time with little bias, the system will have very poor reliability for estimates of travel time in congested conditions.

8.2.2 *Empirical Evaluation of Error Metrics*

The last section described a hypothetical TIS that was modeled on normally distributed errors in speed with no bias and a constant variance. It was shown that estimating the system's speed bias is not a good estimator of performance in estimating link travel times. This is due to the inverse relationship between speeds and travel times. In order to verify that this is a reasonable assumption, data from a real Traveler Information System was compared with estimates of ground truth from a benchmark data source.

Estimates of space-mean-speed from Inrix were collected over 15 days on 11 links in Northern Virginia. These links were described in Chapter 7. The benchmark data source was Bluetooth reidentification. A LOESS model was fit to each link's observations for each day of observations. In total, 165 unique LOESS models were fit to the data. There were a total of 57,222 unique (i.e. link and time-interval) observations from Inrix. An example plot of one day of empirical observations for one link is shown below in Figure 32 with the red points indicating the benchmark estimate and the blue points indicating the estimate from Inrix.

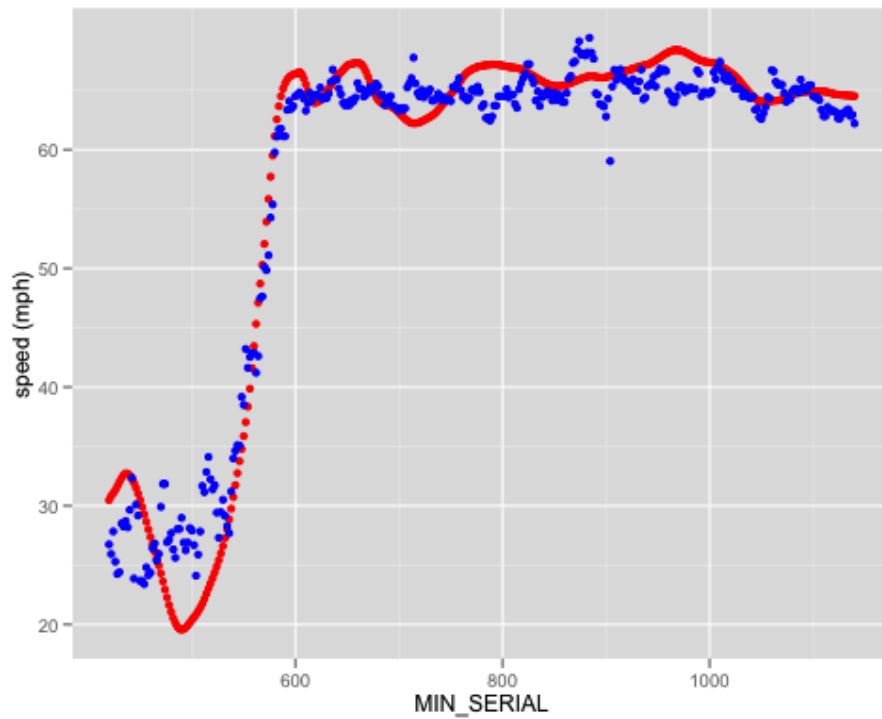


Figure 32: Example Plot of Empirical Observations

For each estimate that Inrix provided, the space-mean-speed was compared to the estimated space-mean-speed from the LOESS model for that link and that time-interval. The error bias was calculated in units of speed and in units of seconds-per-mile. The results, shown below in Figure 33 indicate an approximately normal distribution of speed error bias.

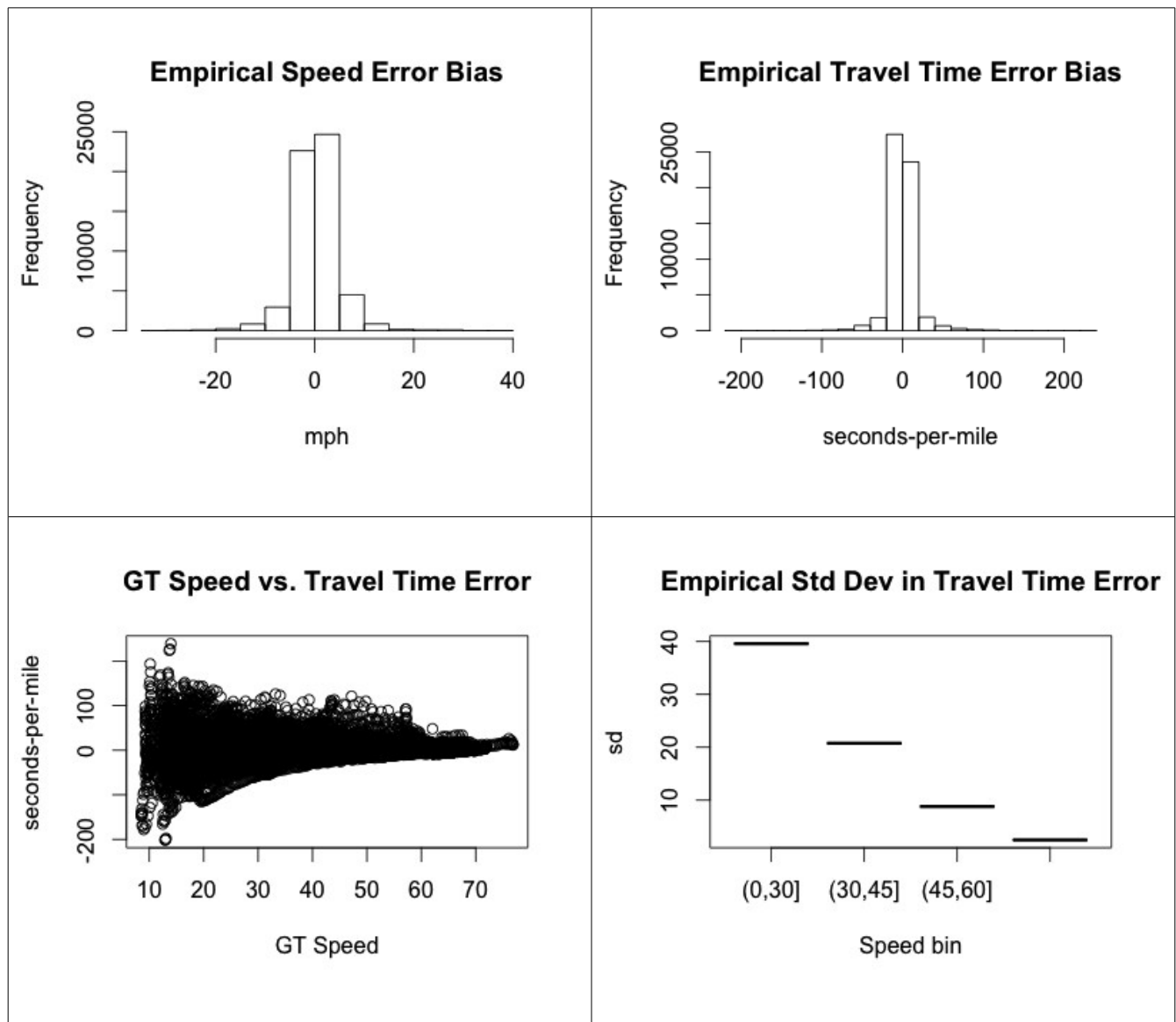


Figure 33: Empirical Comparison of Error Metrics

These results indicate that the assumption of normally distributed errors in speeds is reasonable. The results also indicate that assessing a TIS by estimating speed error bias may be misleading due to the inverse relationship between speed and travel time.

8.3 An Idealized Traveler Information System

Section 8.2 showed that under the assumption of a constant variance in speed errors, a Traveler Information System would have increasingly worse precision at slower ground truth speeds. This is due to the inverse relationship between travel time and speed. However, if we modify the model to allow the standard deviation to vary as a function of ground truth speed we get an idealized version of a traveler information system.

The idealized Traveler Information System would be increasingly precise in estimating space-mean-speed as the ground truth speeds decreased. For example, we could model the standard deviation as a simple linear function of ground truth speed with slope and intercept determined so that the line passes through the coordinates (20,1) and (80,5). Therefore, at 20 mph the standard deviation will be 1 mph and at 80 mph the standard deviation will be 5 mph. This relationship allows the model to be more precise at slower speeds. We can see what such a model would look like by looking at the relationship between ground truth speeds and errors in speed and travel time.

The model was specified as (**Model 2**):

$$\bar{u}_S^* = \bar{u}_S + \epsilon_A \quad (25)$$

where ϵ_A is a random normal variable with no bias and standard deviation of $\sigma = 0.07 \bar{u}_S - 0.33$

The results are shown below in Figure 34.

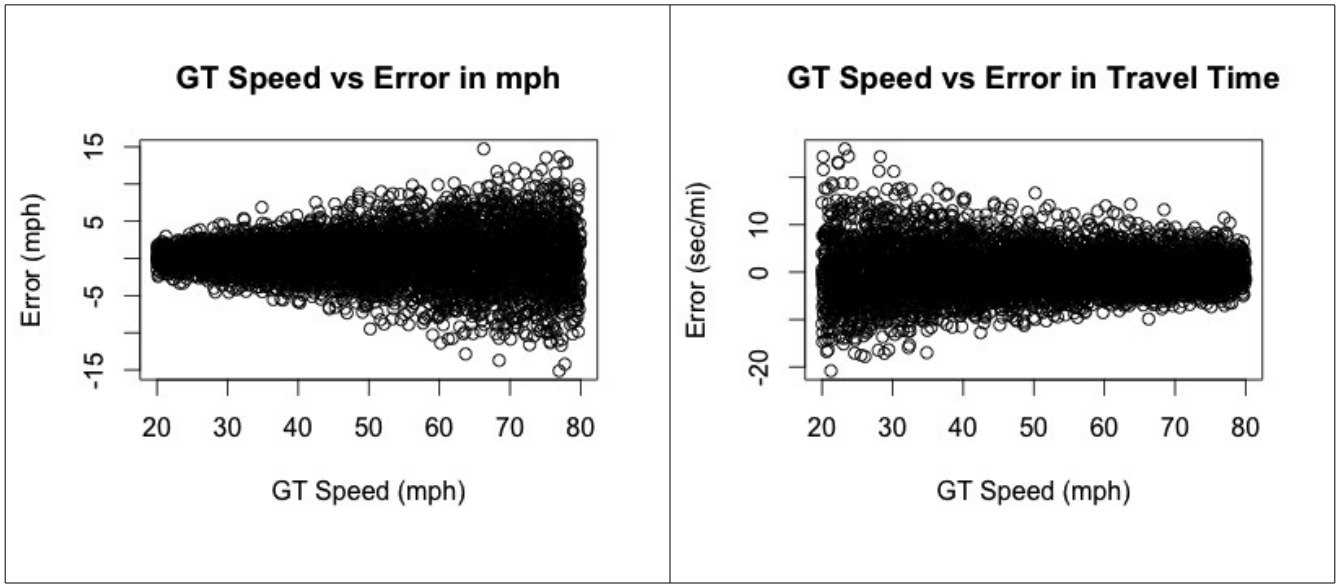


Figure 34: An idealized TIS model

The effect of the linear relationship between the precision of the estimate and the ground truth speed is that the precision of travel time is approaching a constant. The plot on the left shows the increasing precision in speeds as ground truth speeds decrease. The plot on the right shows the effect on the computed travel errors.

8.4 Alternatives to Error Bias and Mean Absolute Error

Mean Square Error is often selected as an error metric because it can be easily decomposed into a linear combination of variance and bias. For a given estimator, $\hat{\theta}$

$$MSE = E[(\hat{\theta} - \theta)^2] = Bias(\hat{\theta})^2 + Var(\hat{\theta}) \quad (26)$$

The minimization of variance and bias are commonly considered the optimal attributes of any estimator. For example, an estimator that minimizes bias but has high variance may be less desirable than an estimator that minimizes variance at the expense of large bias. An estimator with small bias but large variance is accurate but not precise. On the other hand, an estimator that has small variance but

large bias is precise but inaccurate. Such an estimator could be "corrected" or "adjusted" to account for its bias.

Relative error and absolute relative error are two other error metrics that deserve consideration. They provide a standardized measure of error that is independent of the magnitude. A 5 mph error when the ground truth speed is 25 mph is very different than a 5 mph error when the ground truth is 65 mph. Each of these errors can be expressed instead as a percentage relative to the ground truth.

Relative error in travel time and speed are closely related but not identical. When ground truth speeds are slow (e.g. < 25 mph), the relative error in travel time and speed begins to diverge. For example, an underestimate of 5 mph when ground truth is 25 mph is a relative error in speed of -20%. However, the same error in travel time is a 25% overestimate. Figure 35, below illustrates this relationship in a contour plot.

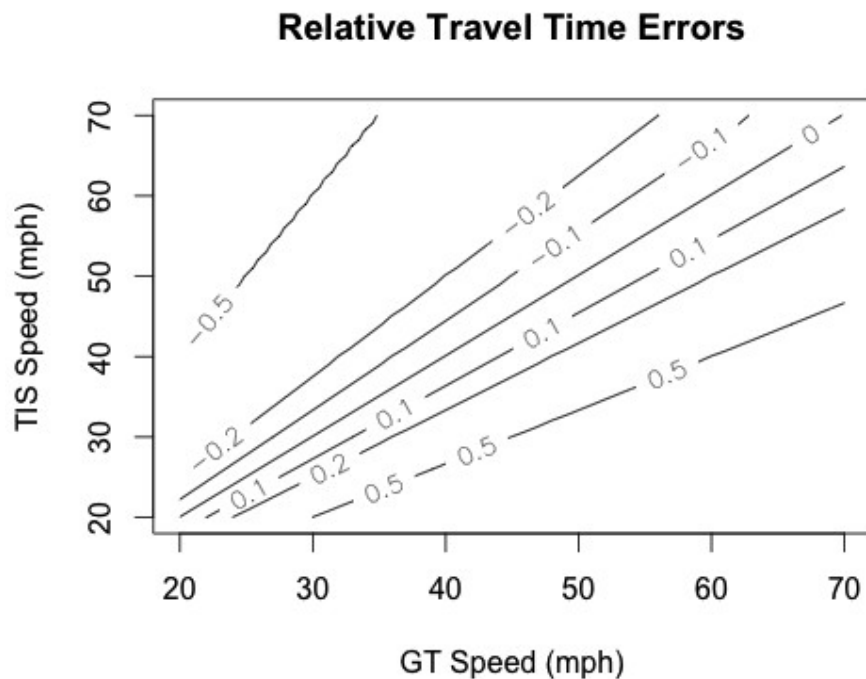


Figure 35: Contour Plot of Relative Errors in Travel Time

The contour lines indicate the relative error in travel time. For example, the bottom-most line is marked "0.5" which indicates that there is a 50% error in travel time when the ground truth speed is 30 mph and the TIS estimate is 20 mph. This would be only a 33% error in speed (i.e. $10/30$). The largest divergence occurs when the TIS significantly underestimates ground truth.

Relative travel time error is an appealing error metric because it can be directly compared between links of differing lengths and because it is easily understood. It is also an effective measure of performance for a Traveler Information System. It directly measures the errors in estimates of the quantity of interest (travel time) in a unit-less metric. A side-by-side comparison of the two models of a TIS (constant variance) and (linear variance) show how measuring errors in using relative errors in travel time can provide a better understanding of system performance.

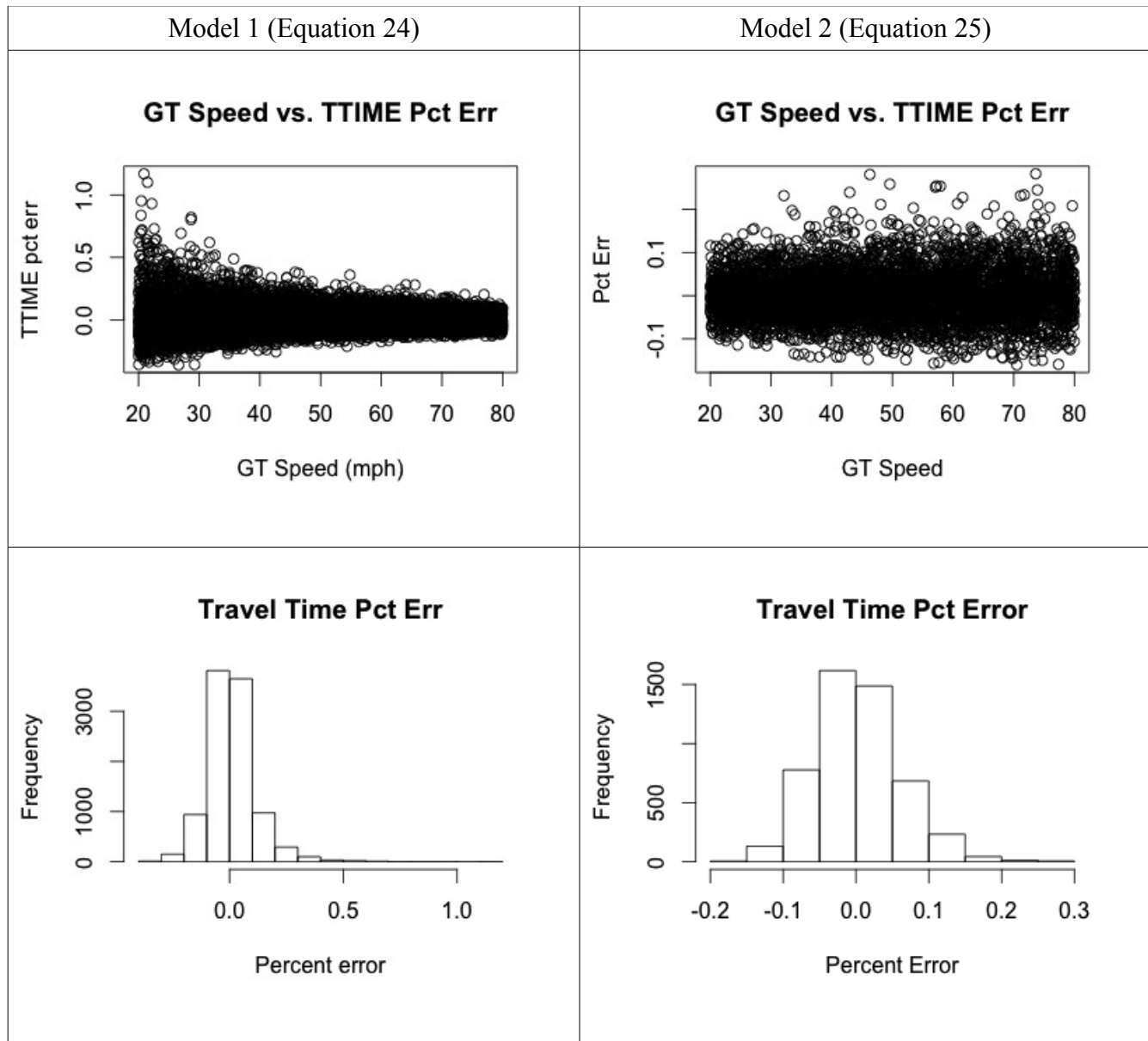


Figure 36: Comparison of Models by Relative Errors in Travel Time

Figure 36, above shows the distribution of relative error in travel time and the relationship between ground truth speed and relative error in travel time for the two models described in this chapter. It can be seen that the idealized model (Model 2) has an approximately normal distribution of relative errors in travel time with no bias.

8.5 Conclusions of Error Measurement

This chapter showed that an ideal Traveler Information System is one which minimizes the bias and variance of errors in travel time estimates. Evaluating a TIS by estimating the bias and variance of errors in speed is misleading because an unbiased system with constant variance in speed errors will have increasing variance in travel time errors.

A system which minimizes both bias and variance in travel time errors should be preferred to a system which minimizes only one of these criteria. However, the most likely scenario is that a system will have some amount of both bias and variance in its estimates. The question is how much of each and what is acceptable.

Further, it was also shown that assessing TIS performance by relative errors in travel time is preferred because they can be directly compared between links of differing lengths and they are easily interpreted by analysts and users of the system. The idealized model of a TIS is one where the speed estimates are increasingly precise as ground truth speeds decrease. This model has the property of an approximately normal distribution of relative travel time errors.

Chapter 9 Monitoring TIS Data Quality

The last chapter showed that error measurements play a key role in evaluating data quality. The observed errors are drawn from a sample in time and space. For example, in Chapter 6 and 7 a method for selecting links was described. Errors in estimated travel time observed on the selected links represent only a sample of the observable error in the entire network. Similarly, since benchmark data is collected over some finite period of time, the observed errors represent only a sample of observable errors over time.

Evaluating a TIS at a single point in time and space may not capture important variation in data quality. Therefore, it is important to establish a monitoring program for data quality that can detect shifts in quality levels in space or time. This chapter will outline the development of a set of automated tools that can be used to monitor the quality of data from a Traveler Information System in time and space.

9.1 Introduction to Data Quality Monitoring

Statistical quality assurance (SQA) is an area of statistical methods concerned with quality assurance. Traditionally, the application of SQA has been in the area of manufacturing, industrial processing, and other areas concerned with material output. However, there has also been research that shows that these methods can be adapted and applied to problems of information systems data quality. [32], [33]

A typical quality assurance schematic might look like Figure 37, below.

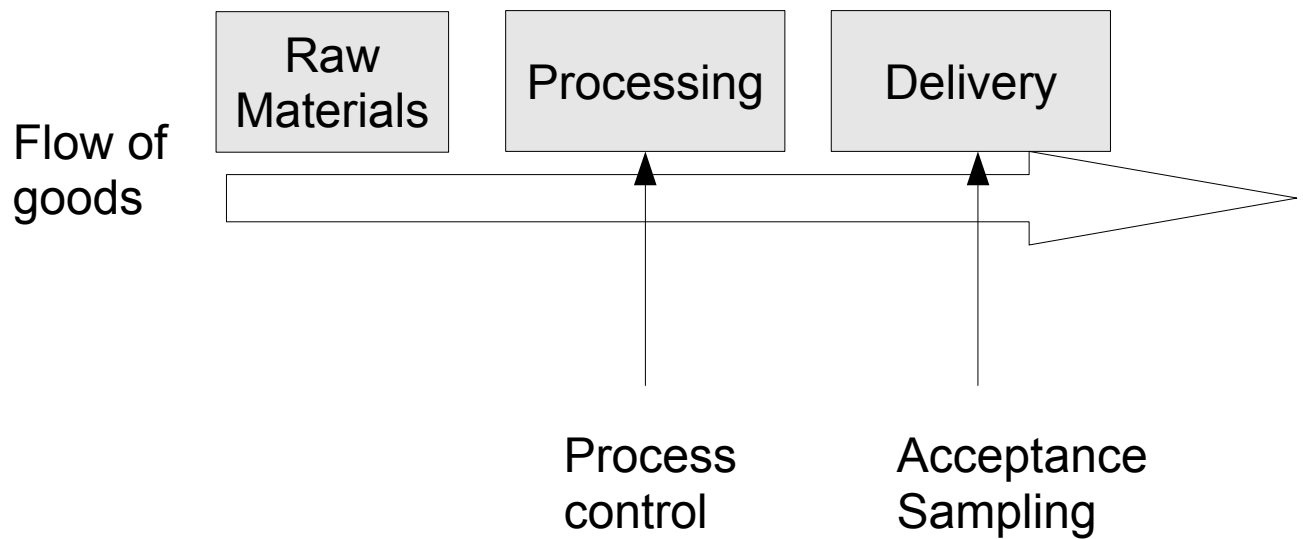


Figure 37: Traditional Quality Assurance

This model can be readily adapted for data quality assurance. Replacing the flow of goods with data and raw materials with sensors we can see that the processing and delivery stages are where data quality assurance can be applied. However, since the process generating the data is obscured from the evaluation, quality assurance must focus on the delivery end of the schematic.

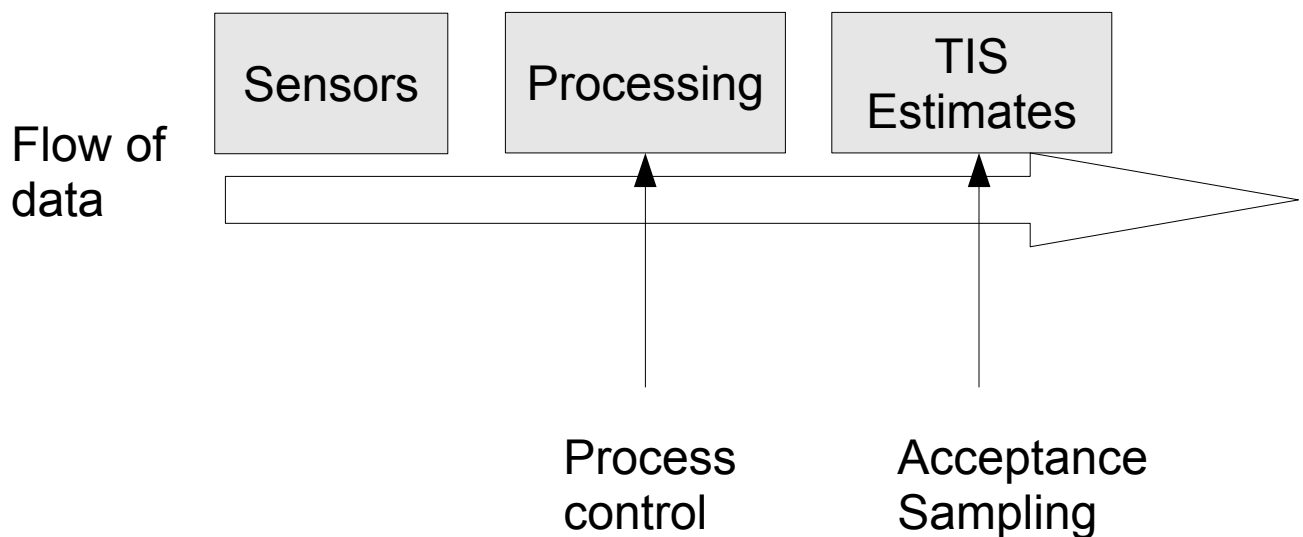


Figure 38: Data Quality Assurance

9.2 A Data Quality Monitoring Information Architecture

In order to "monitor" data quality, an information architecture must be in place to automate the processing of sensor data, and the ingestion of Traveler Information System estimates, road network meta-data, and generation of data quality reports. A schematic of a basic infrastructure is shown below in Figure 39.

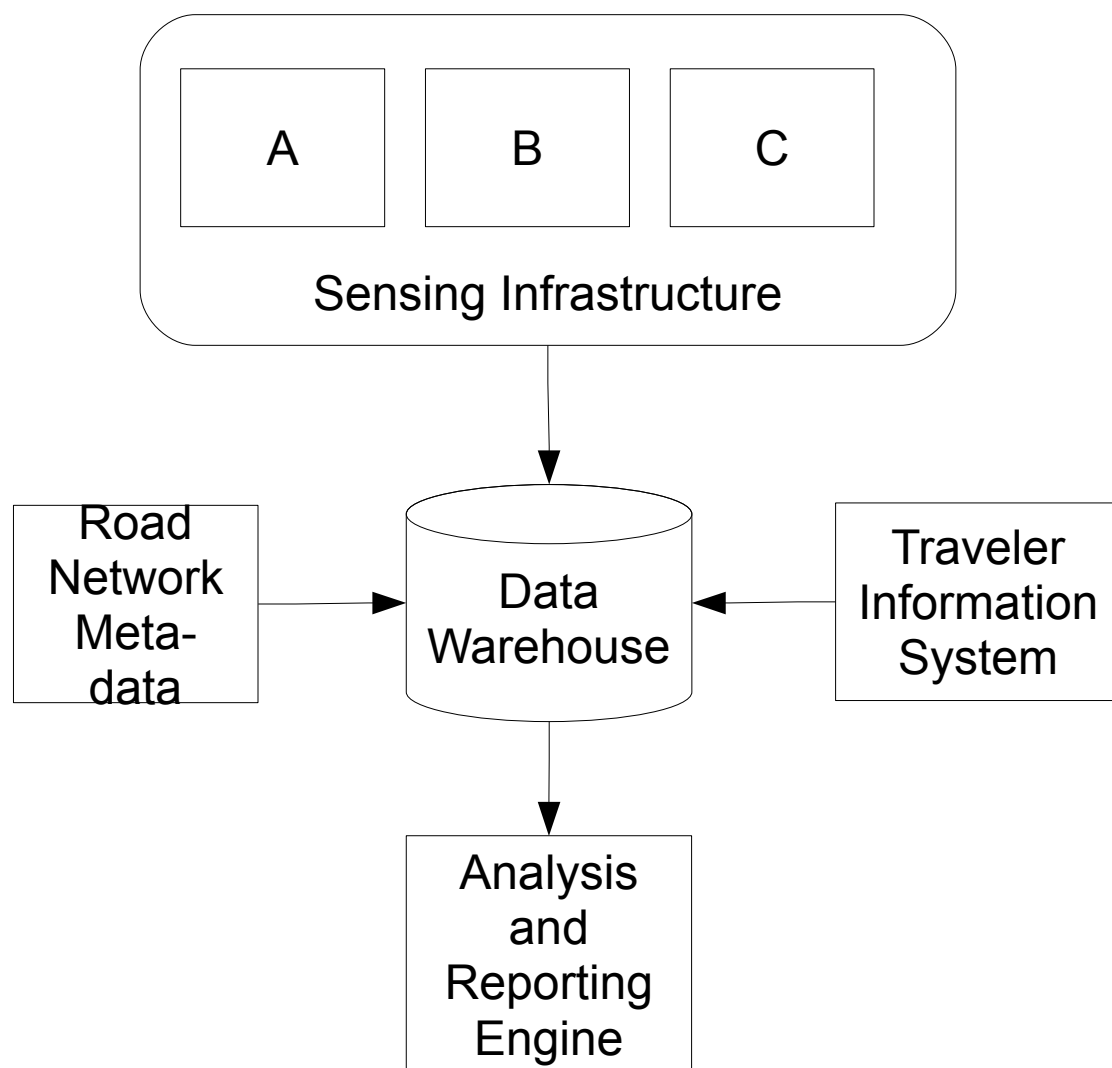


Figure 39: Example Architecture of a Data Quality Monitoring System

9.2.1 Sensing Infrastructure

The sensing infrastructure is used to collect the observations for estimates of benchmark travel time. The choice of where to collect observations and the methods and techniques for estimating benchmark travel time is described in Chapters 5 - 7 of this dissertation. Ideally, sensors should be "plugged in" to an existing communication network so that data can be transmitted back to a central processing computer. In the case that the sensors are not connected to a network the data can be retrieved manually and input in to the data warehouse. The data from each sensor is collected and organized by the data structure described in Chapter 4.

9.2.2 Road Network Meta-data

The road network meta-data is an important tertiary source of data that ideally covers many of the important attributes necessary for data quality monitoring. The meta-data should include attributes such as link length, latitude/longitude position of sensors, deployment date, and other attributes such as road design attributes, and link ADT. The meta-data can be used as a reporting feature for more advanced reports not directly discussed in this dissertation.

9.2.3 Traveler Information System

The Traveler Information System is the system that is being monitored. As the TIS generates estimates they are ingested into the data warehouse for comparison with a benchmark value. The basic data structure for managing TIS data is described in Chapter 4 of this dissertation.

9.2.4 *Analysis and Reporting Engine*

The analysis and reporting engine takes the data from the sensors and the TIS and runs the analysis to generate the appropriate measures of data quality. The reporting component takes the results of the analysis and presents them in graphical format for an analyst to review. The next section

The output of this system are reports that allow an analyst to monitor the quality of the Traveler Information System. The analysis and reporting engine is designed to be flexible so that new re-identification sensors can be easily added to the system and new links from the TIS can also be seamlessly added to the system.

9.3 Data Transmission and Processing

The layers which transmit data such as the sensing infrastructure and Traveler Information System can stream data to the data warehouse via XML (also called Extensible Markup Language). This is a basic coding of structured data that allows for exchange between different sub-systems and is easily transmitted over an internet connection. Processing of XML data is handled a loading process that can ingest XML and load it into the data warehouse. The coding of these components is fairly straightforward and need not be described in detail in this dissertation.

9.4 Statistical Methods of Quality Control

Statistical methods of quality assurance can be classified by two criteria[34]:

1. Timing of Inspection
2. Measurement of Quality Characteristic

Timing of inspection refers to the point in time during the production process when quality is monitored. In the case of a manufacturing process line, it may be possible to institute process control

procedures before delivery of goods. However, as mentioned earlier, because the process generating the data in a TIS is often not directly accessible, quality assurance must focus on the latter stages of the production process when the data has already been delivered.

The measurement of quality characteristics is the second criteria used in determining the type of quality control program. Measurement of quality characteristics can be further categorized into two types of quality measurements:

1. Variable inspection
2. Attribute inspection

Variable inspection refers to measurements of quality characteristics on a continuous scale. For example, a producer of material output may inspect the weight of products by comparison with a benchmark weight for the particular product. The difference in observed weights would be an example of a variable measurement.

Attribute inspection refers to measurements of quality characteristics on a categorical or discrete scale. Inspection by attributes classifies a product as conforming (C) or non-conforming (NC) based on a quality examination of the product.

The choice of variable or attribute inspection is driven by the type of data that the TIS generates. For example, a TIS that generates "state estimates" of traffic conditions can be monitored by attribute inspection. In such a scenario, the TIS would generate an estimate of the traffic state such as a "green-yellow-red" color to indicate the current conditions. Under attribute inspection, the traffic state would be classified as conforming or non-conforming by determining whether the TIS estimate of traffic conditions was in agreement with the benchmark estimate.

If the TIS is generating estimates of travel time on a continuous scale then a variable inspection

model would be appropriate. The remainder of this chapter focuses on an example of process monitoring by variable inspection. As a point of further research it would also be possible to develop similar techniques for attribute inspection.

9.5 Process Monitoring

9.5.1 *X-bar Charts*

The Shewart control charts are a widely used graphical technique of process monitoring. These charts have been applied in industrial processing and can be used as an effective monitoring technique for data quality. The X-bar chart is the most commonly used chart and an example of this chart is shown below in Figure 40.

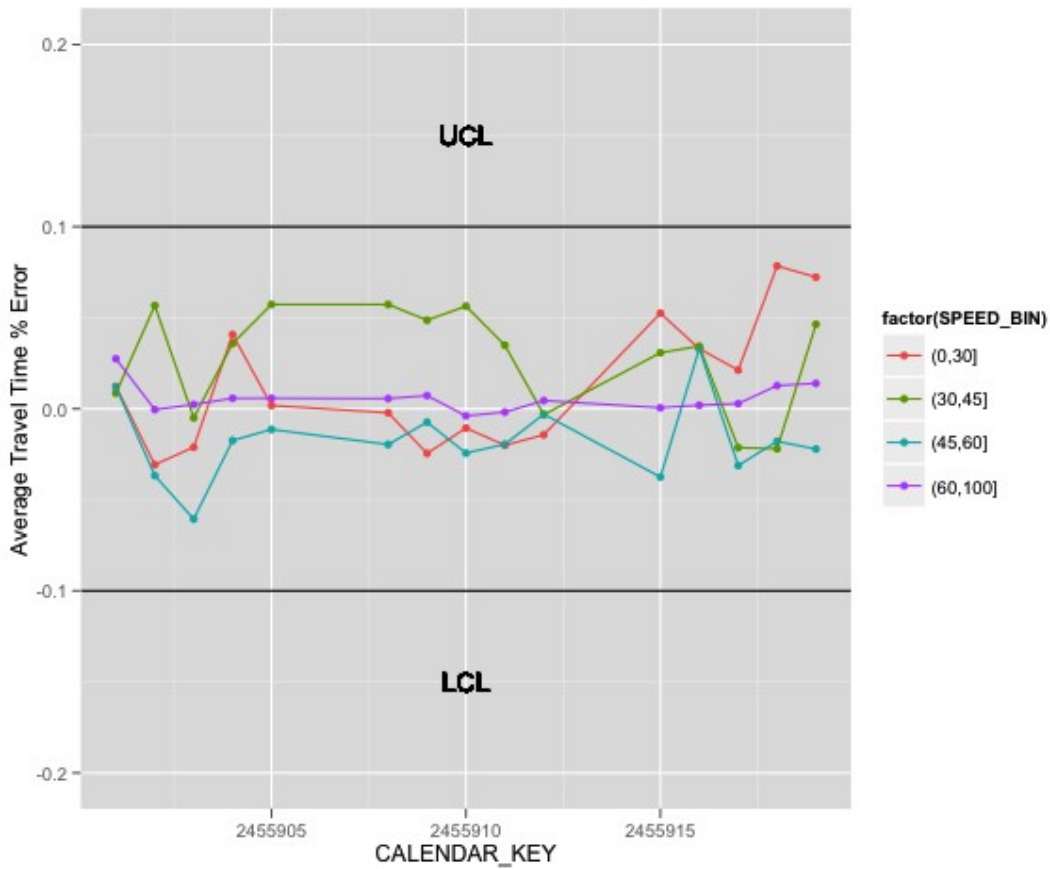


Figure 40: *X-bar Chart for TIS Data Quality Monitoring*

Control limits can be determined analytically or they can be specified. For a process which is in control with a normally distributed quality characteristic, the control limits are typically expressed as:

$$\bar{x} \pm 3\hat{\sigma} \quad (27)$$

where \bar{x} is the observed process mean and $\hat{\sigma}$ is the estimated standard deviation

The assumption with this specification of a control limit is that the process has a constant variance in the measured quality attribute. It was shown in the last chapter that the variance in travel time errors is most likely not constant but will increase as speeds decrease. Therefore, a specification limit can be specified based on engineering judgment. In this case, a specification limit of 10% is

suggested.

Since the variance in travel time errors is likely to be related to the speed of traffic, the observations can be grouped by speed bins. Figure 40 shows the average travel time percentage error for the 12 monitored segments in the Northern Virginia network. The data has been broken down into speed bins so that observations where the ground truth speed was between 0 and 30 mph are plotted in a different series than observations where the ground truth was greater than 60 mph. The solid black lines at the top and bottom of the chart are suggested control limits set to be $\pm 10\%$.

9.5.2 *S-charts*

A second type of control chart is the "S-chart" or process standard deviation chart. This is a chart which monitors the variance in the observations over time. This type of chart is an important analytical tool for monitoring the precision of the system. If a trend is noticed where process variance is increasing beyond a specified control limit then this is an indication that the precision of the system has degraded for some reason and requires investigation. An example of an S-chart is shown below in Figure 41.

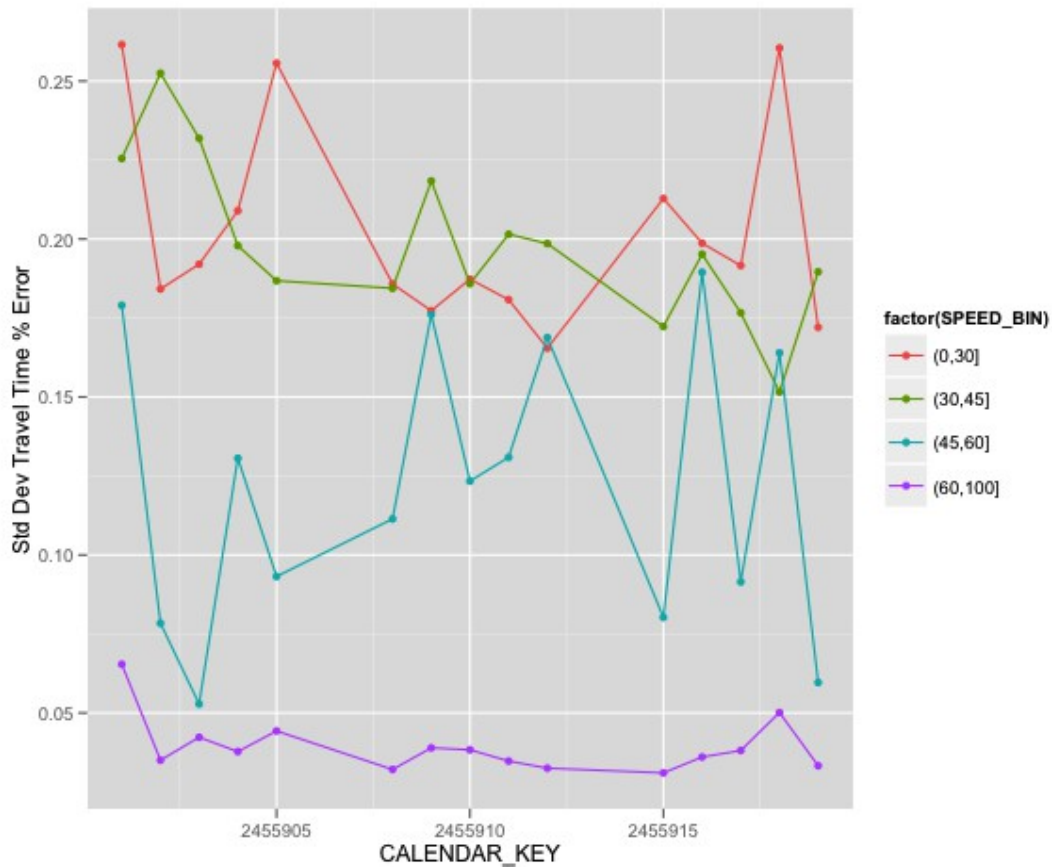


Figure 41: Example S-chart

Again, the chart shows plots of the standard deviation in the observed travel time percentage errors for the monitored road segments. The standard deviation is calculated each day. The data has also been divided into speed bins. It can be seen from the plot that TIS estimates in free flow traffic tend to be very precise (low variance) and the precision degrades as the ground truth speed decreases.

The X-bar and S-chart are the foundation of a process monitoring system. These two charts provide an overview of the bias and variance of the errors in relative travel time. Taken together they give a complete picture of system performance. The charts can indicate when trends such as worsening bias or variance are occurring and signal an analyst to investigate further.

9.5.3 Monitoring Errors by Location (Space) and by Time

The X-bar and S charts do average out variation that may occur between the monitored segments. For example, the plotted standard deviation on the S-chart for the first 3-4 days was higher than the rest of the days plotted. If we wanted to see why this happened we could also investigate the process characteristics on a segment by segment basis. This would give us a picture of variation across the spatial dimension.

To start, we can plot a boxplot of the 10 segments for the first three days in our data set. This results in the plot shown below in Figure 42.

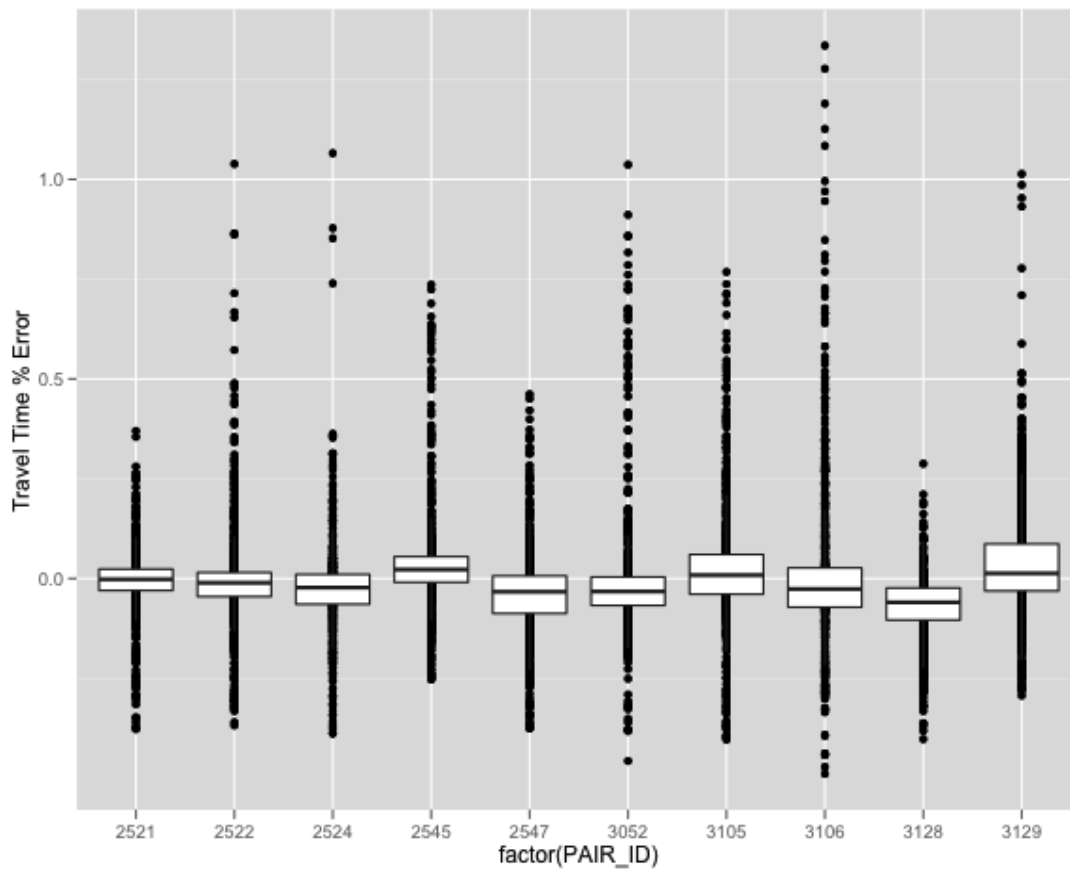


Figure 42: Boxplots of Errors for 3 Days by Observed Segment

Looking at the boxplot we might investigate a plot of errors versus time for three of the segments with the most variation. This appears to be segments 3052, 3106, and 3129. We can show the errors for these segments by time of day for the first three days of data.

An example of this is shown below in Figure 43.

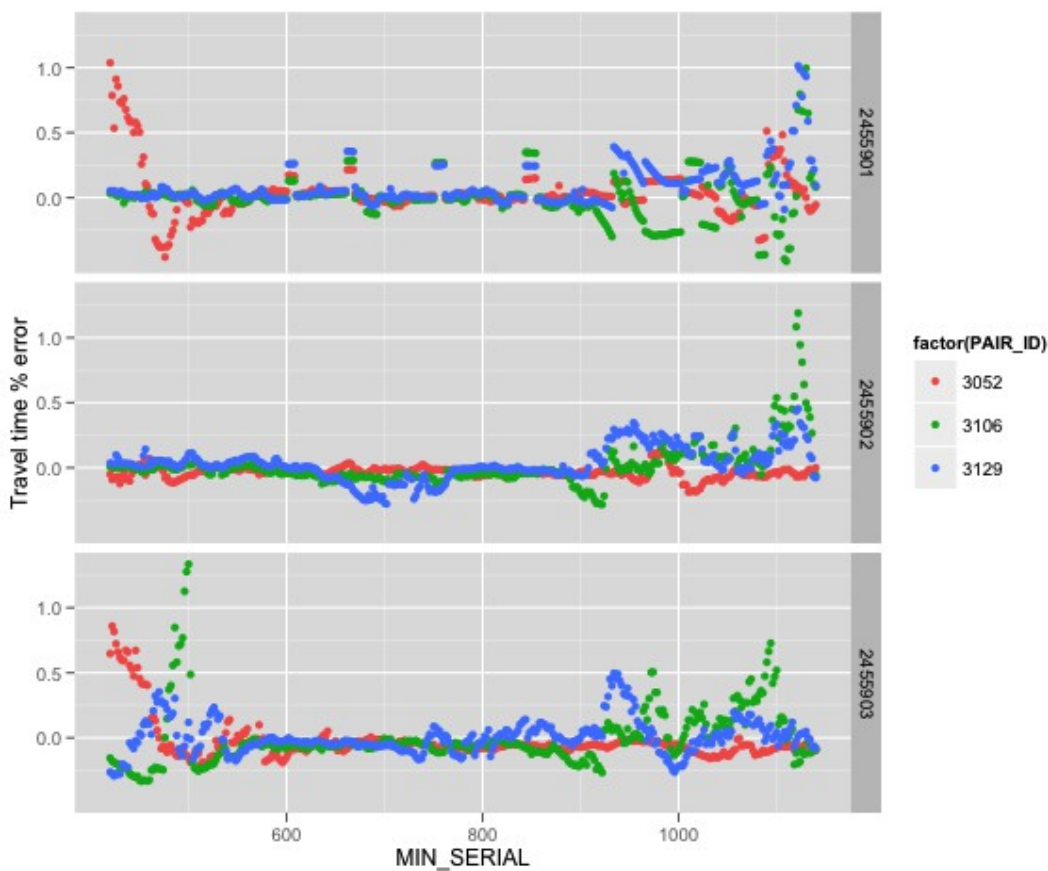


Figure 43: Observed Errors by Time for 3 Segments

Now we can see that segment 3052 had some large errors in the morning on the first day. So we can go ahead and plot the LOESS fit for that day against the Inrix estimates and include the individual Bluetooth re-identification data in the background to see how the LOESS fit looked. This is shown

below in Figure 44.

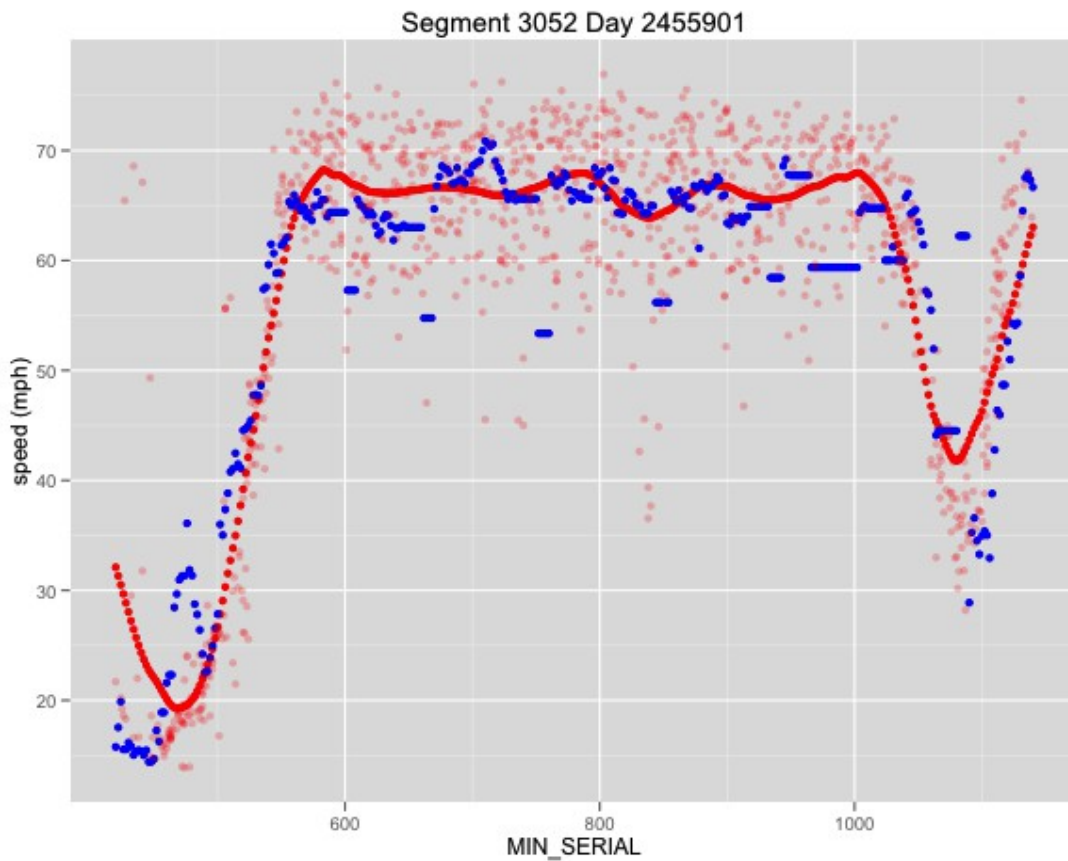


Figure 44: Plot of TIS Estimates vs. LOESS fit

What we can see from the plot is that the LOESS fit was influenced by some outlier observations in the early day. This may be a situation where a robust LOESS model would be appropriate. The errors, however, are not systemic and an analyst can easily verify that any minor deviations in the X-bar or S-bar chart are due to technical details such as this.

9.6 Process Capability

In addition to graphically monitoring the process quality characteristics, the capability of the process can be monitored by a Process Capability Index. Capability indexes have been an important component of quality control engineering since their introduction by Juran.[35] Process capability is a measure of the variability of the process in terms of the quality characteristic and the specification limits. This can be written as:

$$C_p = \frac{USL - LSL}{6\sigma} \quad (28)$$

This definition assumes that the quality characteristics are normally distributed with a constant standard deviation. If these assumptions are valid then the process capability index provides an upper bound on the proportion of output from the process that will fall within the range of the specification limits. If the index is equal to 1 we can expect that approximately 99.9% of the system output will fall within specification limits.

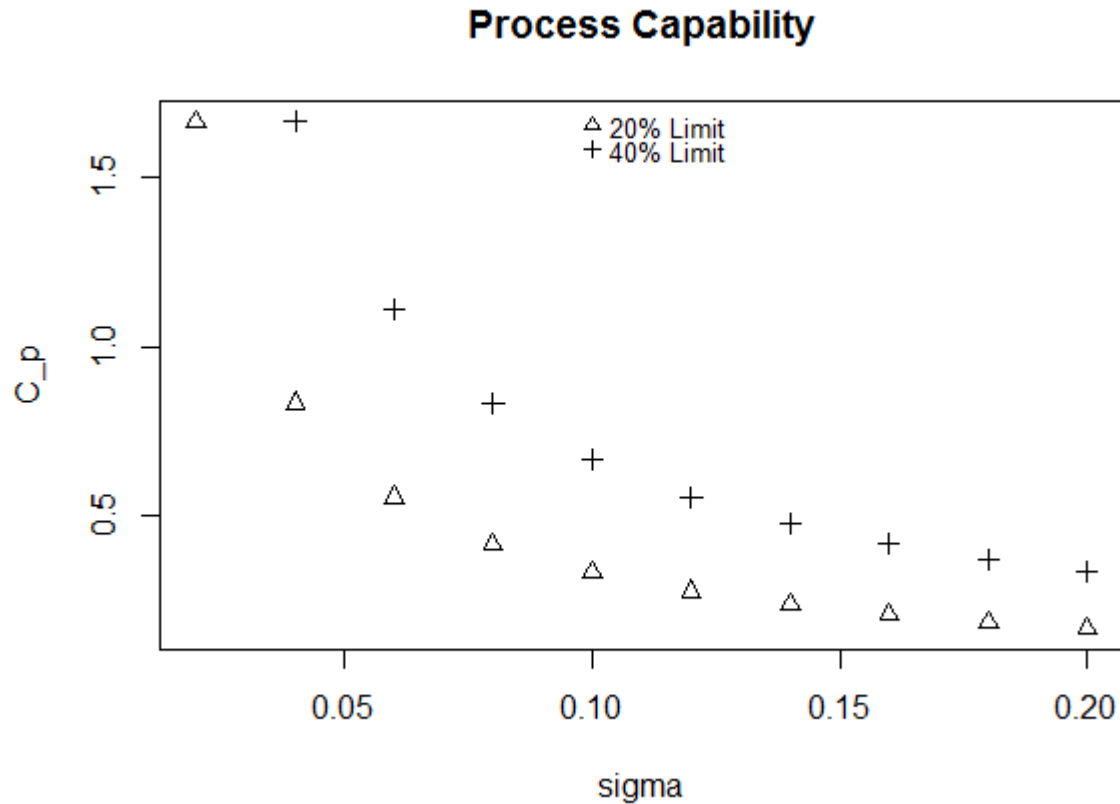


Figure 45: Process Capability versus Variability

It can be seen from Figure 45 that the process capability is dependent on the level of variability in the observed quality characteristic. Since travel time errors exhibit differing levels of variation by speeds, the process capability index should be evaluated by speed bins. The capability of the process in free flow conditions will be significantly different from the capability of the process in congestion.

If we use a process specification spread (USL - LSL) of 20% then the process capability for the northern Virginia data can be estimated as:

Speed Bin	C _p
60+ mph	1.11
45 - 60 mph	0.28
30 - 45 mph	0.19
0 - 30 mph	0.15

Table 9.1: Example of Process Capability Using Northern Virginia Data

It can be seen from Table 9.1 that the process capability decreases significantly when the speeds move from the free flow state to congestion. This reflects the fact that small errors in speed estimates generate large percentage errors in travel time estimates. And these large errors contribute to the large process variability.

The measurement of process capability provides an important method of monitoring the reliability of a Traveler Information System. If the process capability is worsening then the reliability of the system is worsening. And while, on average, the system may generate accurate estimates of travel time, the variability in those estimates will be quite large. Process capability can be monitored to see whether system reliability is improving over time.

9.7 Conclusions of Traveler Information System Data Quality Monitoring

This chapter presented an overview of how statistical quality control can be applied to the problem of monitoring Traveler Information System data quality. It was shown that the tools developed in quality control such as Shewhart charts of the process mean and variance have direct application to the problem of monitoring errors from a TIS. Furthermore, it was shown that these tools would provide transportation engineers the ability to monitor data quality trends over time and space. With the ability to "drill down" to errors at the segment level and to visually inspect the fit between a benchmark and the TIS estimates this chapter demonstrated a comprehensive set of tools for monitoring data quality. Finally, process capability indexes were suggested as a method of monitoring the reliability of a

traveler information system.

Chapter 10 Conclusions and Contributions

The goal of this research was to provide rigorous statistical methods for the evaluation of Traveler Information Systems. A literature review of past research indicated that there were three main areas of research that required investigation:

1. Benchmark Estimation
2. Link Selection
3. Data Quality Evaluation and Monitoring

In each area, this dissertation attempted to provide guidance or introduce new methods to the evaluation process. In the area of benchmark estimation, the existing method of local averaging was compared with a smoothing method based on Locally Weighted Regression (LOESS). The evaluation showed that while local averaging offers benefits in terms of simple computation and accuracy in dense samples, the method failed to provide estimates when samples were sparse and suffered accuracy problems with the presence of outlier observations. Although the LOESS method is more computationally complex, the method can be used to fit models to complex time-series data in sparse samples with relatively good accuracy compared with local averaging. LOESS also offers the benefit of predictions at time-intervals where no observations were available. The LOESS smoothing parameter, λ , can be selected by cross-validation, but an analysis of empirical data showed a parameter between 0.10 - 0.20 would be appropriate in most cases for link benchmark estimation. And, while the LOESS smoothing method has been applied in other ITS-related research, the performance of this method has not been fully investigated and compared with other more frequently employed methods such as local averaging. This dissertation evaluated the LOESS method against empirical data and provides practical insights into how the LOESS method performs for applications in benchmarking

Traveler Information Systems.

The process of link selection in TIS data quality evaluations has historically been driven by expert opinion which is a form of non-probability sampling. In this dissertation an objective and quantitative method for link selection was developed based on Maximum Entropy Sampling (MES) design. The method uses an existing historical database of space-mean-speed estimates from a Traveler Information System to estimate a covariance matrix and then finds subsets of links that maximize entropy based on the log-determinant of the subset of the covariance matrix. This has been shown to be an effective sampling strategy in a number of different applications with spatial covariance. The method was evaluated using simulated and empirical data and found to be effective in sampling observable TIS errors.

The MES method was then combined with a secondary criterion of network coverage in order to select from different candidate subsets of links. These criteria can be used to decide among different subsets so that both network coverage and entropy can be used to determine the optimal subset of links for monitoring. A case study of three candidate subsets of links in Northern Virginia was used as an example of how the method would work.

The literature review underscores the need for an objective method of link selection. The past data quality evaluations reviewed in this research all employed some form of expert opinion in selecting links for the evaluation. In many cases, experts may indeed select the same sets of links that an objective method such as MES would select. However, relying on experts to select links is not objective and can easily lead to questions of bias in an evaluation. An objective and quantitative method would not be as open to such critiques.

Furthermore, the problem of link selection for small networks is fairly trivial. One might expect that experts can identify the critical links for data collection in a small regional network. However, as

the scope of an evaluation grows it is unlikely that expert knowledge will be sufficient for consistently identifying critical links. The MES sampling method developed in this dissertation can be scaled and applied to problems much larger than local or regional networks. Thus, the method can be applied for sampling problems in large heterogeneous networks where local experts' knowledge may not be sufficient.

After investigating benchmark estimation and link selection, the problem of error measurement and data quality monitoring was investigated. The dissertation showed that measurement of errors from Traveler Information Systems is sensitive to the units of the errors due to the inverse relationship between travel time and space-mean-speed. It was shown that a system with normally distributed errors in units of speed with no bias and a constant variance will have an increasing variance in travel time errors as ground truth speeds decrease. This leads to the insight that evaluating a system by monitoring speed errors is not sufficient to monitor the quantity of interest which is travel time. Therefore, it was recommended that relative errors in travel time be monitored instead. Relative errors in travel time offer a unit-less measurement of error that can be compared among links of differing lengths and is easily understood by practitioners.

Statistical quality control tools such as X-bar and S charts can be used as an effective tool for monitoring the quality of data from a Traveler Information System. These tools give analysts the ability to track data quality across time. When quality issues do appear to occur, an analyst can also easily investigate quality problems by the spatial dimension by examining boxplots of errors by monitored segment. The system also offers the ability to "dig deeper" by investigating time-series plots of errors and examining individual plots of benchmark fits versus TIS estimates.

The methods investigated and developed in this dissertation support the basic framework of Traveler Information System data quality evaluation and monitoring. Because link travel time is such a

fundamental performance measure of transportation systems, how we measure, compare, and evaluate travel time estimates will have a significant impact on our ability to accurately assess the performance of transportation systems. The research presented in this dissertation provides insight into statistical methods that can be used to rigorously evaluate the quality of data generated by Traveler Information Systems.

10.1 Opportunities for Further Research

While one hopes that the research presented here is comprehensive there is clearly room for further research. Within the area of benchmarking there is a need for further refinement of the fitting algorithm so that outliers can be dealt with appropriately. The existing LOESS algorithm offers the ability to use robust regression as an alternative to least-squares regression and this would be an appropriate area for evaluation. There is also a need to better define an appropriate threshold for rejection of sensor data as being too sparse. Currently, this is not well defined.

Within the area of link selection, it would be valuable to investigate the ability to use advanced heuristic optimization algorithms to find optimal solutions to the maximum entropy problem. This has been investigated in the literature and would be valuable for monitoring large networks.

Finally, the tools within statistical quality control are numerous and there is plenty of opportunity to push for more research in this area. Monitoring data quality across large spatial and temporal dimensions is a problem where these tools can certainly make contributions. What was presented in this dissertation is hopefully a first step in that direction.

References

- [1] S. Turner, "Defining and Measuring Traffic Data Quality: White Paper on Recommended Approaches," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1870, no. -1, pp. 62–69, 2004.
- [2] T. C. Redman, *Data quality : the field guide*. Oxford: Digital Press, 2001.
- [3] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality*, vol. 1, no. 1, pp. 1–22, Jun. 2009.
- [4] M. Williams, D. Cornford, L. Bastin, R. Jones, and S. Parker, "Automatic processing, quality assurance and serving of real-time weather data," *Computers & Geosciences*, vol. 37, no. 3, pp. 353–362, Mar. 2011.
- [5] V. Sessions and M. Valtorta, "Towards a Method for Data Accuracy Assessment Utilizing a Bayesian Network Learning Algorithm," *Journal of Data and Information Quality*, vol. 1, no. 3, pp. 1–34, Dec. 2009.
- [6] R. E. Turochy and B. L. Smith, "Applying quality control to traffic condition monitoring," in *ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.00TH8493)*, Dearborn, MI, USA, pp. 15–20.
- [7] I. V. Schneider, H. William, S. M. Turner, J. Roth, and J. Wikander, "Statistical Validation of Speeds and Travel Times Provided by a Data Service Vendor," 2010.
- [8] I-95 Corridor Coalition, "Validation of Inrix Data: Two Year Summary Report June 2008 - June 2010." I-95 Corridor Coalition, Sep-2010.
- [9] M. Fontaine, B. Smith, A. Hendricks, and W. Scherer, "Wireless Location Technology-Based Traffic Monitoring: Preliminary Recommendations to Transportation Agencies Based on Synthesis of Experience and Simulation Results," *Transportation Research Record*, vol. 1993, no. 1, pp. 51–58, Jan. 2007.
- [10] "TPF > Browse TPF Studies > Detailed View," *Standard Test Procedure for Travel Time Data Quality Assessment*. [Online]. Available: <http://www.pooledfund.org/projectdetails.asp?id=426&status=4>. [Accessed: 01-Aug-2011].
- [11] S. Turner and D. Holdener, "Probe Vehicle Sample Sizes for Real-Time Information: The Houston Experience." 1995.
- [12] M. W. Green, M. D. Fontaine, and B. L. Smith, "Investigation of dynamic probe sample requirements for traffic condition monitoring," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1870, no. -1, pp. 55–61, 2004.
- [13] H. Tanikella and B. L. Smith, "An Investigation of the Application of Stratified Sampling in Probe-Based Traffic-Monitoring Systems," *Journal of Intelligent Transportation Systems*, vol. 14, no. 2, pp. 83–94, 2010.
- [14] "Anderson Darling and Shapiro Wilk tests," *Engineering Statistics Handbook*. [Online]. Available: <http://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm>. [Accessed: 01-May-2012].
- [15] A. May, *Traffic flow fundamentals*. Englewood Cliffs N.J.: Prentice Hall, 1990.
- [16] H. Rakha and W. Zhang, "Estimating traffic stream space mean speed and reliability from dual-

- and single-loop detectors,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1925, no. -1, pp. 38–47, 2005.
- [17] W. H. Inmon, *Building the data warehouse*. Indianapolis, Ind.: Wiley, 2005.
 - [18] W. Eisele and L. Rilett, “Travel-Time Estimates Obtained from Intelligent Transportation Systems and Instrumented Test Vehicles: Statistical Comparison,” *Transportation Research Record*, vol. 1804, no. 1, pp. 8–16, Jan. 2002.
 - [19] R. Li, G. Rose, and M. Sarvi, “Evaluation of Speed-Based Travel Time Estimation Models,” *Journal of Transportation Engineering*, vol. 132, no. 7, p. 540, 2006.
 - [20] W. S. Cleveland and S. J. Devlin, “Locally weighted regression: an approach to regression analysis by local fitting,” vol. 83, pp. 596–610, 1988.
 - [21] W. L. Eisele, L. R. Rilett, K. B. Mhoon, and C. Spiegelman, “Using Intelligent Transportation Systems Travel-Time Data for Multimodal Analyses and System Monitoring,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1768, no. -1, pp. 148–156, 2001.
 - [22] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning : data mining, inference, and prediction : Trevor Hastie, Robert Tibshirani, Jerome Friedman*. New York: Springer, 2009.
 - [23] M. C. Shewry and H. P. Wynn, “Maximum entropy sampling,” *Journal of Applied Statistics*, vol. 14, no. 2, pp. 165–170, 1987.
 - [24] B. Ainslie, C. Reuten, D. G. Steyn, N. D. Le, and J. V. Zidek, “Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants,” *Journal of Environmental Management*, vol. 90, no. 8, pp. 2715–2729, Jun. 2009.
 - [25] M. Bueso, J. Angulo, and F. Alonso, “A state-space model approach to optimum spatial sampling design based on entropy,” *Environmental and Ecological Statistics*, vol. 5, no. 1, pp. 29–44, 1998.
 - [26] T. J. Stohlgren, S. Kumar, D. T. Barnett, and P. H. Evangelista, “Using Maximum Entropy Modeling for Optimal Selection of Sampling Sites for Monitoring Networks,” *Diversity*, vol. 3, no. 2, pp. 252–261, 2011.
 - [27] T. Husain and H. U. Khan, “Shannon’s entropy concept in optimum air monitoring network design,” *Science of The Total Environment*, vol. 30, pp. 181 – 190, 1983.
 - [28] H. Wickham, “Reshaping Data with the reshape Package,” *Journal of Statistical Software*, vol. 21, no. 12, pp. 1–20, 2007.
 - [29] NIST, “6.5.3.2. Determinant and Eigenstructure,” *NIST/SEMATECH e-Handbook of Statistical Methods*. [Online]. Available: <http://www.itl.nist.gov/div898/handbook/pmc/section5/pmc532.htm>. [Accessed: 08-Feb-2012].
 - [30] C. W. Ko, J. Lee, and M. Queyranne, “An exact algorithm for maximum entropy sampling,” *Operations Research*, pp. 684–691, 1995.
 - [31] J. Lee, “CONSTRAINED MAXIMUM-ENTROPY SAMPLING.,” *Operations Research*, vol. 46, no. 5, pp. 655 – 664, 1998.
 - [32] R. Y. Wang, V. C. Storey, and C. P. Firth, “A framework for analysis of data quality research,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 7, no. 4, pp. 623–640, 1995.
 - [33] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, “Modeling Information Manufacturing Systems to Determine Information Product Quality,” *Management Science*, vol. 44, no. 4, pp. 462–484, Apr. 1998.
 - [34] H.-J. Mittag, *Statistical methods of quality assurance*, 1st English language ed. London ;;New York: Chapman & Hall, 1993.
 - [35] K. Palmer and K.-L. Tsui, “A review and interpretations of process capability indices,” *Annals of Operations Research*, vol. 87, pp. 31–47, 1999.

