Technical Report Documentation Page

1. Report No. FHWA/TX-09/0-5686-1	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Archiving, Sharing, and Quantifying Reliability of Traffic Data		5. Report Date October 2008 6. Performing Organization Code
7. Author(s) Dr. S. Travis Waller, Dr. Kara Kockelman, Dr. Dazhi Sun, Stephen Boyles, Dung-Ying Lin, ManWo Ng, Saamiya Seraj, Mohamad Tassabehji, Varunraj Valsaraj, Dr. Xiaokun Wang		8. Performing Organization Report No. 0-5686-1
9. Performing Organization Name and Address Center for Transportation Research The University of Texas at Austin 3208 Red River, Suite 200 Austin, TX 78705-2650		10. Work Unit No. (TRAIS) 11. Contract or Grant No. 0-5686
12. Sponsoring Agency Name and Address Texas Department of Transportation Research and Technology Implementation Office P.O. Box 5080 Austin, TX 78763-5080		13. Type of Report and Period Covered Technical Report September 2006–August 2008 14. Sponsoring Agency Code
15. Supplementary Notes Project performed in cooperation with the Texas Department of Transportation and the Federal Highway Administration.		
16. Abstract Vast quantities of transportation data are automatically recorded by intelligent transportations infrastructure, such as inductive loop detectors, video cameras, and side-fire radar devices. Such devices are typically deployed by traffic management centers (TMCs), and the data used for operational studies; however, such data are also highly valuable for transportation planning and other applications. This project considered how such data can best be stored and managed to accommodate multiple users, and multiple types of detector technologies. A modular system is developed, allowing data from multiple TMCs to be collected, translated into a common format, and placed in a central archive. Additionally, a novel method for quantifying data reliability is described, as error detection is critical when managing large quantities of data. Multiple techniques are also described for imputing missing data, or correcting erroneous data. Issues related to implementation are also discussed, along with innovative detector technologies that may be deployed in the near future, and thus must be considered when developing a flexible archival system.		
17. Key Words Traffic data archiving, intelligent transportation systems, data reliability, data imputation	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161; www.ntis.gov.	

19. Security Classif. (of report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of pages 162	22. Price

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

Archiving, Sharing, and Quantifying Reliability of Traffic Data

S. Travis Waller
Kara Kockelman
Dazhi Sun
Stephen Boyles
Dung-Ying Lin
ManWo Ng
Saamiya Seraj
Mohamad Tassabehji
Varunraj Valsaraj
Xiaokun Wang

CTR Technical Report:	0-5686-1
Report Date:	October 2008
Project:	0-5686
Project Title:	Utilizing the Data Collected at Traffic Management Centers for Planning Purposes through Non-Traditional Sources and Improved Equipment
Sponsoring Agency:	Texas Department of Transportation
Performing Agency:	Center for Transportation Research at The University of Texas at Austin

Project performed in cooperation with the Texas Department of Transportation and the Federal Highway Administration.

Center for Transportation Research
The University of Texas at Austin
3208 Red River
Austin, TX 78705

www.utexas.edu/research/ctr

Disclaimers

Author's Disclaimer: The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the Federal Highway Administration or the Texas Department of Transportation (TxDOT). This report does not constitute a standard, specification, or regulation.

Patent Disclaimer: There was no invention or discovery conceived or first actually reduced to practice in the course of or under this contract, including any art, method, process, machine manufacture, design or composition of matter, or any new useful improvement thereof, or any variety of plant, which is or may be patentable under the patent laws of the United States of America or any foreign country.

Notice: The United States Government and the State of Texas do not endorse products or manufacturers. If trade or manufacturers' names appear herein, it is solely because they are considered essential to the object of this report.

Engineering Disclaimer

NOT INTENDED FOR CONSTRUCTION, BIDDING, OR PERMIT PURPOSES.

Project Engineer: Kara Kockelman
Professional Engineer License State and Number: 93443
P. E. Designation: Research Supervisor

Acknowledgments

The authors would like to express their appreciation towards the Texas Department of Transportation for support of this research. In particular, thanks are due to Loretta Brown, Fabian Kalapach, Bill Knowles, Jim Neidigh, and Duncan Stewart for their assistance and leadership throughout this project.

Products

This report also contains products 0-5686-P1 (a possible implementation plan), 0-5686-P2 (a guidebook describing different detector technologies), and slides for 0-5686-P3 (a workshop for communicating the results of this research). Section 3.3 contains 0-5686-P1, and Appendices A and B contain 0-5686-P2 and 0-5686-P3, respectively.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Overview of ITS Data and Applications

Intelligent transportation systems (ITS) infrastructure automatically records vast amounts of traffic data, which is highly useful for a variety of applications if properly archived. Induction loops are still the most common detector used in urban areas, although newer technologies (such as video or infrared detection) continue to improve and have been successfully deployed. Although different technologies report different data, common implementations measure quantities such as traffic volumes, speeds, and occupancy, and may attempt to classify vehicles by weight or length. With automated devices, this data is typically collected continuously and at a relatively fine resolution, barring communication or technical failures.

It is not difficult to find applications for a large, well-maintained data set of this sort, especially in regions where spatial coverage is high. A common use is in operational studies, such as before-and-after evaluation of ramp meter deployment, or to determine an optimal schedule for reversible lanes. More recently, it has been suggested that transportation planners can use ITS data sets to assist in generating annual average daily traffic (AADT) counts for reporting to the Federal Highway Administration (FHWA). Other applications abound: volume counts are highly useful for calibrating planning models used by metropolitan planning organizations; for evaluating the effectiveness of work zone channelization in reducing driving speeds; and for measuring the impact of tolled or managed lanes at both the corridor-level and system-wide scales, to name just three. Such data can even be used to develop, test, and evaluate theoretical route choice and traffic flow models.

At present, ITS infrastructure is typically operated by a traffic management center (TMC), which maintains control and communication links with detectors. If the data is to be stored, the TMC then assumes responsibility for archiving the data and performing any quality control measures specified by agency policy. Although some TMCs then grant other users access to the data, internally or to the general public, in many cases it is difficult for others to obtain this data. This is usually not due to technical factors; rather, concerns about issues such as data reliability, responsibility for maintaining and providing support to users, and control over uses for the data pose larger obstacles to implementation of data sharing. In other cases, data sharing has simply not been identified as an agency priority.

This research described in this report addresses exactly these issues, providing guidance on how to organize and store data so it is useful to a broad spectrum of users, developing and testing data reliability and imputation algorithms to answer questions of quality and missing data, and describing some future trends in detector technologies to ensure that the system is useful well into the future. The remainder of this chapter describes data sharing applications in greater detail, particularly those relevant to planners employed by a state department of transportation (DOT); elucidates on implementation challenges, classifying them as "organizational," "methodological," or "technical;" provides brief discussion of a modular data archiving framework which can address these issues; and presents the organizational structure of this report.

1.2 Opportunities for Sharing Data

As briefly mentioned, many opportunities exist for using ITS data in new ways. In particular, a key motivation for this project was the possible use of ITS volume data to supplement automatic traffic recorders (ATRs) and tube counts collected by DOT planners to generate AADT estimates. Because all detector technologies in current use record traffic volumes more or less continuously throughout the entire year, the potential exists to obtain AADT counts without resorting to "factoring" or other estimation techniques, and at a greater number of locations, increasing both accuracy and spatial coverage.

Four prime advantages of using ITS data for this purpose are increased coverage, more accurate statistical inferences, diminished safety risks to agency personnel collecting data, and the elimination of inefficient "double counting" of traffic volumes by personnel in different agency departments.

The continuous recording of traffic data by ITS infrastructure offers much greater temporal coverage than short-term tube counts can provide. While some DOT planners also maintain permanent ATRs, these are typically fewer in number compared to the detectors operated by a local TMC. Thus, making use of both can lead to a great increase in spatial coverage as well. Improvements in both spatial and temporal coverage lead to greater redundancy and a larger source of data to draw from.

This in turns leads to more accurate statistical predictions. When only a short-term sample is available, historical scaling factors must be applied based on the day and month of each sample. This method is vulnerable to outliers in the observed data and other variations in observed traffic counts. Even when scaling factors are not needed, high spatial coverage allows interpolation in case of missing data, and even estimation of volumes in locations where no detector is present. Although current FHWA guidelines do not permit the use of interpolated data in lieu of actual measurements, several accurate statistical methods have been developed to accomplish this, suggesting that this policy may be revisited in the future.

Further, manually placing pneumatic tube counters can be dangerous, and expose agency personnel to unnecessary risk, as when placing tubes across a busy freeway ramp. The use of ITS detectors obviates this risk, as the traffic stream is only interrupted during installation and maintenance activities, which are typically accompanied by planned closures and changes in channelization.

Finally, it is not uncommon to see detectors or tube counts used for AADT counts in the vicinity of TMC sensors that collect similar data. The use of a common data repository eliminates the need for this "double counting," which is an inefficient use of agency resources and effort.

This data can also be applied by planners using traffic assignment models, which require calibrated volume-delay functions (VDFs). Rather than assuming a standard function to be used throughout the entire network, ITS data allows more accurate regional (and even corridor-level) specification of these functions, in principle allowing better calibration of planning models to observed counts and traveler behavior.

1.3 Organizational, Methodological, and Technical Challenges

Broadly speaking, challenges in implementing a central data archive can be classified as *organizational*, *methodological*, and *technical*. *Organizational* issues are related to how data should be stored, and how responsibilities should be assigned. These include determining the

workgroups or agencies that have primary responsibility for collecting, operating, and maintaining the archive; determining which users are authorized to access the archive; developing an interface allowing authorized users to retrieve data in a useful form; determining the level of aggregation (if any) performed on the data prior to storage; and documenting the protocols and formats used. A PostgreSQL database developed in this project provides a flexible basis for storing data, and generating different types of reports for users with differing needs. Several existing data archives were examined as case studies, and are discussed in Chapter 2 with an emphasis on organizational issues.

Methodological issues involve the use of statistics or other quantitative procedures to ensure the data is useful and generates useful models. The most significant issue is related to data quality—if incorrect data cannot be marked as such, the quality of the archive will suffer. Although it is not possible to correctly assess every single observation as either correct or incorrect, observation of general patterns and internal consistency can be used to flag data that are highly implausible or physically impossible. A second key issue involves estimation of missing or suspicious data, and developing statistical procedures that allow accurate imputation based on contemporaneous observations. Finally, new methods may be needed for other applications, such as calibration of VDFs. Novel algorithms are created and tested for these purposes, and also compared with existing methods.

At the same time, there are a number of *technical* challenges that must be addressed. In the short-term, communication protocols must be established to connect TMCs to the central archive. New communications infrastructure (such as fiber optic cable or wireless transmitters) may be required, depending on detector locations and existing communication links. In the long-term, advances in detector technology suggest that the archive should be able to accept data from multiple types of detectors. To this end, a common data format is developed in this project, allowing the archive to work with any detector whose data can be converted into this format.

These three types of challenges form the framework for this research project, as described in the following chapters.

1.4 Prototype Data Archive

This project included development of a prototype data archiving system, which was implemented on a small scale, receiving data from three detectors. Although larger-scale implementation will likely require structural changes, this prototype still demonstrates the key features of the proposed approach, shown in Figure 1.1.

Detector data is collected at participating TMCs, then transmitted at regular intervals to the central archive, followed by a preprocessing procedure: the data is converted into the common format, its reliability is assessed, and (optionally) a corrected estimate is made if the initial reading is missing or suspect. In all cases when an estimate is made, both the original and corrected readings are stored and marked as such. These assessments require knowledge of the network structure, which are coded when the archive is installed, and historical data values stored in the archive at an earlier time.

Following preprocessing, the data record is stored in the database. A variety of users can then access this data by generating reports, which are customized for individual applications. Supporting subroutines can be applied at this time, such as imputing data even at locations where no detector is present. These techniques are described more fully in Chapter 4.

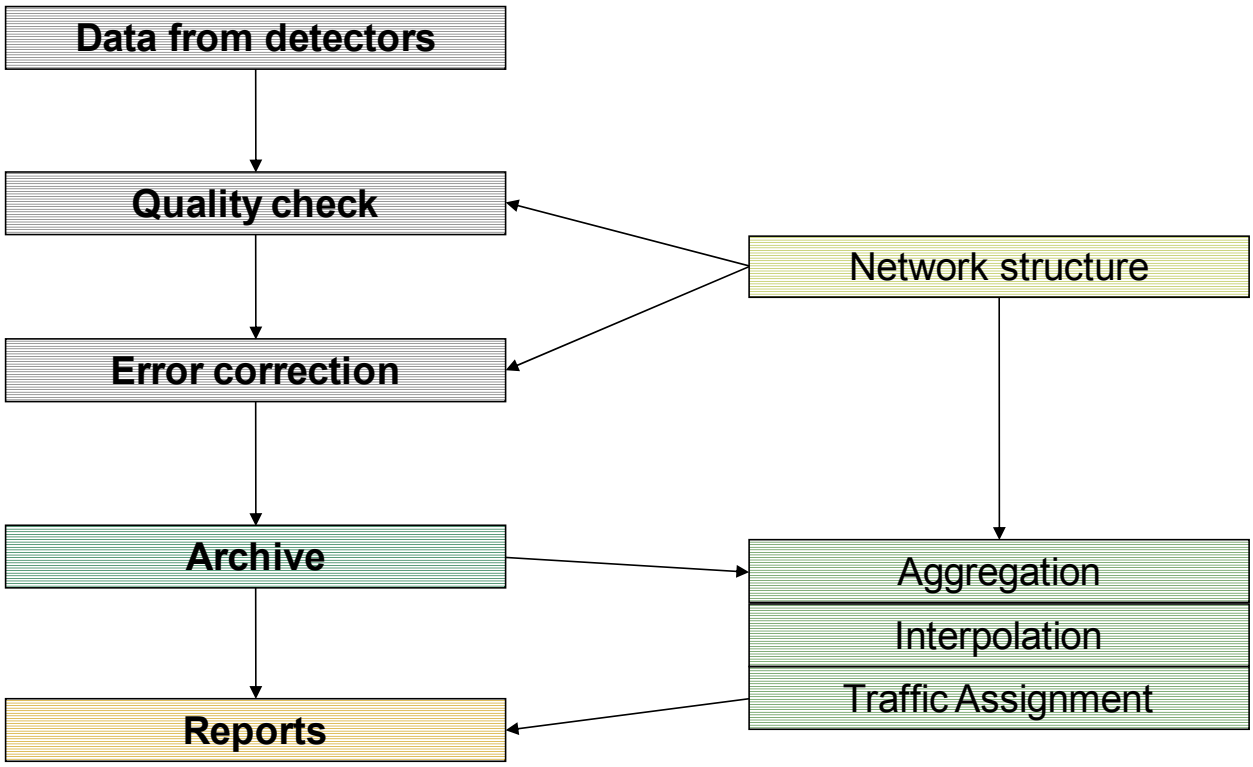


Figure 1.1: Prototype system schematic.

This basic structure can readily be adapted to larger-scale implementation involving multiple TMCs and detector types. In such cases, each TMC transmits its data directly to the central archive, maintaining a modular structure in which TMCs can be freely added or removed from the archive.

1.5 Outline

The remainder of this report is organized as follows: Chapter 2 describes past experience with data archives by other agencies, along with guidelines that have been developed for their implementation and a general overview of ITS data collection. Chapter 3 describes the prototype system in greater detail, along with a specific action plan. Chapter 4 focuses on the issues of data reliability and error correction, and presents a comparison of existing and newly-developed algorithms for these tasks. Chapter 5 describes field data tests conducted using the prototype system, and Chapter 6 summarizes the key findings.

Additional information can be found in six appendices: Appendix A provides information on current and emerging detector technologies, and Appendix B contains a survey form distributed to Texas TMCs regarding current practices in data sharing. Appendix C contains an analysis of variability in AADT count data collected in Texas. Appendix D demonstrates a potential application of this data, to calibrate a VDF used in traffic assignment. Appendix E describes the effect of reducing the amount of data stored, and using statistical techniques to estimate the omitted data. Although the results were promising, storage space does not appear to be a limiting factor in archive design, and thus this technique was not incorporated into the prototype system. Finally, Appendix F includes slides from a workshop that can be used to train

agency personnel in the methods developed in this report, and to communicate the most important research findings.

Chapter 2. ITS Data and Case Studies in Data Archiving

2.1 Introduction

This project's main goal is to determine how to use existing operational data for planning purposes. To this end, there are three major areas in which research needs to be directed: methodology, technology, and agency organization. But first, it is important to examine existing systems, to identify difficulties and key issues. For instance, using data in this way will require a central archive of sensor data, and there are many ways to implement such a system. Previous experience by other agencies can give crucial guidance in developing such a system for Texas.

Fundamentally, the issue of using ITS data for planning is one of data integration and sharing. Done effectively, this can greatly streamline the use of available resources. For instance, tube counts collected for planning purposes may duplicate loop detector data already being collected by TMCs, wasting resources and unnecessarily exposing technicians to danger when laying tubes on high-volume roads.

However, connecting data from different sources is often complicated. Hall (2003) highlighted several key components of successful data partnerships/integration, as seen in nine different states:

- Clarifying roles and responsibilities of partners
- Agreeing on data standards, and managing potentially conflicting data definitions and currencies
- Resolving equipment and connectivity issues and taking advantage of new technology
- Integrating data from different data sets
- Utilizing data with varying spatial accuracies and resolutions
- Archiving and managing large data sets
- Securing resources and funding, and sharing partnership costs
- Quantifying and qualifying the value, utility, and benefit of data partnering investments
- Addressing privacy and security concerns
- Obtaining management leadership and support
- Overcoming cultural and institutional barriers

These points outline the major issues involved in transportation data sharing, and should be kept in mind throughout the rest of this document. The remainder of this chapter proceeds as follows: first, the nature of ITS and planning data is discussed, with some discussion of the types of data collected and the differing needs associated with each use. Data quality issues are also addressed in this section. Next, a series of case studies is presented, each detailing a data archiving system, including its developers, users, and contents. Responses from a questionnaire distributed to TMCs in Texas are also included. Finally, the key barriers identified from the

implementation of these and other such systems are discussed, along with recommended strategies to overcome these obstacles.

2.2 ITS and Planning Data Collection

As ITS encompasses a broad range of technologies involved in transportation, there is a wide variety of data that can be collected through these means. Turner (2001) and Margiotta (2002) provide some description of these data and the following discussion summarizes these sources, as does Table 2.1.

Perhaps the most ubiquitous ITS data collection devices are loop detectors, which primarily measure volume and occupancy; certain loop configurations can measure speed directly as well. Other parameters of interest can be estimated from these measurements. Vehicle classification also can be attempted using such systems. These devices are located in the roadway itself, one per lane, and are commonly spaced ¼ mile – 1 mile apart in urban areas. This data is recorded continuously, and reported back to a central system regularly, typically at 20- to 60-second intervals.

Typical operational use of this data is made to automatically adjust ramp meter timing in real-time (Taylor and Meldrum, 2000), or to detect incidents or locations of heavy congestion. This may be made available to the public online, to the media, to transit agencies, to emergency dispatchers, or to other users who value up-to-date information on traffic conditions. Some agencies also archive this data to generate annual average daily traffic (AADT) counts, saturation flows, peak hour factors, and so on.

Similar data can be collected by video surveillance devices, although these are not as widespread as loop detectors. Electronic toll systems provide another means to measure dynamic traffic flows at particular points in the traffic network, and are becoming increasingly common.

Data collected by ITS devices are usually collected continuously, and at various points in the network; that is, they have broad spatial and temporal range. In this way, a large amount of data is collected. If this information is to be used for anything other than real-time use, it is vital to store it in an easily accessible form for later use. However, ITS measurements, with their wide spatiotemporal coverage, are of a different nature than typical planning volume measurements, which are collected at specific points in space and time. For instance, tube counts are usually performed only at select locations on particular dates.

Table 2.2 lists typical supply- and demand-side data used by planners. As this table indicates, the data needs of transportation planners go far beyond the volume and occupancy measurements that are routinely collected by loop detectors. Still, ITS data can be used to estimate some of this additional information as well. For instance, Ashok and Ben-Akiva (1993) describe a procedure to estimate dynamic origin-destination matrices. Electronic toll collection data has also been used towards this end.

Table 2.1: ITS Data Types (adapted from Turner, 2001, Table 3)

ITS data source	Primary data elements	Typical collection ITS-generated data equipment	Spatial coverage	Temporal coverage	Real-time uses
Freeway and Toll Collection					
Freeway traffic flow surveillance data	▮ volume ▮ speed ▮ occupancy	▮ loop detectors ▮ video imaging ▮ acoustic ▮ radar ▮ microwave	usually spaced at ≤1 mile; by lane	sensors report at 20- to 60-second intervals	▮ ramp meter timing ▮ incident detection ▮ congestion/queue identification
	▮ vehicle classification ▮ vehicle weight	▮ loop detectors ▮ weigh-in-motion ▮ video imaging ▮ acoustic	usually 50-100 per state; by lane	usually hourly	pre-screening for weight enforcement
Ramp meter and traffic signal preemptions	▮ time of preemption ▮ location	field controllers	at traffic control devices only	usually full-time	Priority to transit, HOV, and EMS vehicles
Ramp meter and traffic signal cycle lengths	▮ begin time ▮ end time ▮ location ▮ cycle length	field controllers	at traffic control devices only	usually full-time	Adapt traffic control response to actual traffic conditions
Visual and video surveillance data	▮ time ▮ location ▮ queue length ▮ vehicle trajectories ▮ vehicle classification ▮ vehicle occupancy	▮ cctv ▮ aerial videos ▮ image processing Technology	selected locations	usually full-time	▮ coordinate traffic control response ▮ congestion/queue identification ▮ incident verification
Vehicle counts from electronic toll collection	▮ time ▮ location ▮ vehicle counts	electronic toll collections equipment	at instrumented toll lanes	usually full-time	automatic toll collection
TMC-generated	▮ link congestion indices	TMC software	selected roadway	usually full-time	▮ incident detection
Traffic flow metrics	▮ stops/delay estimates		segments		▮ traveler information ▮ control strategies
Arterial Street					
Arterial traffic flow surveillance data	▮ volume ▮ speed ▮ occupancy	▮ loop detectors ▮ video imaging ▮ acoustic ▮ radar ▮ microwave	usually midblock at selected locations only ("system detectors")	Sensors report at 20- to 60-second Intervals	▮ progression setting ▮ congestion/queue identification
Traffic signal phasing and offsets	▮ begin time ▮ end time ▮ location ▮ up/downstream offsets	field controllers	at traffic control devices only	usually full-time	adapt traffic control response to actual traffic conditions

Table 2.2: Various Types of Planning Data (adapted from Jack Faucett Associates, 1997)

Supply	Demand
System Data ▮ Mileage and lanes ▮ Capacity ▮ Functional road class ▮ Nodes and segments ▮ Land use data for system expansion ▮ Intraurban truck routes *Service Data* ▮ Access ▮ Interurban access ▮ Intermodal access ▮ Data on service providers ▮ Fare or fee structure data ▮ Drayage services *Facilities Data* ▮ Inventory of facilities ▮ Delivery and pickup *Infrastructure Condition Data* ▮ Pavement data by highway route ▮ Any data pertinent to condition of routes, bridge, ramps, etc. that affect the efficiency of interurban truck access to the urban area or truck pick-up and delivery activities ▮ Age of various road classes	*Economic Activity Data* ▮ Employment data by SIC code and region ▮ Industrial operations ▮ Wholesalers and distributors ▮ Commodity data by SIC and geographic detail ▮ Export/import data by point of exit/entry *Demographic Data* ▮ Income data by household and region ▮ Vehicle ownership data by household and region ▮ Population and labor force data ▮ Household characteristics *Land Use Data* ▮ Acreage data ▮ Housing data ▮ Employment data ▮ Access data ▮ Zoning data *Travel Data* ▮ Trip generation data ▮ Trip distribution data ▮ Travel cost data ▮ Special generator data ▮ Traffic volume data ▮ VMT data *Travel Behavior Data* ▮ Mode choice data ▮ Route choice data ▮ User preference data ▮ Time-of-day for pickup and deliveries ▮ Carriers behavior data ▮ Intermodal agreements

Often, this data duplicates what is collected by operations personnel. One key reason for such duplication is the lack of an efficient means of sharing data. Accuracy may be another reason: if loop detectors malfunction, they may continue to report data and, in the absence of error-checking procedures, lead to skewed estimates. This may be more critical in the planning domain than in operations domain; for instance, incident detection or congestion monitoring requires only coarse estimates of vehicle speeds and occupancies. On the other hand, data requirements for operations such as real-time adaptive ramp metering may be more rigorous.

Turner (2001) suggests that data quality be defined as "the fitness of data for all purposes that require it," implying that "measuring data quality requires an understanding of all intended purposes for that data" (ibid.). In the context of operations data, the most common measure of data quality is completeness, or the number of samples available for aggregation. For instance, in Figure 2.1, the boldfaced 30's indicate that for each of the 15-minute aggregated samples, all thirty 30-second individual measurements are available.

```
Data for segment SEGK715001 for 07/15/2001
Number of Lanes: 4

# Time   Samp  Speed  Vol  Occ
00:01:51   30    47   575    6
00:16:51   30    48   503    5
00:31:51   30    48   503    5
00:46:51   30    49   421    4
01:01:52   30    48   274    5
01:16:52   30    42   275   14
...
```

Figure 2.1: The Advanced Regional Traffic Interactive Management & Information System (ARTIMIS) Reporting of Data Completeness (ARTIMIS archives; Turner, 2001)

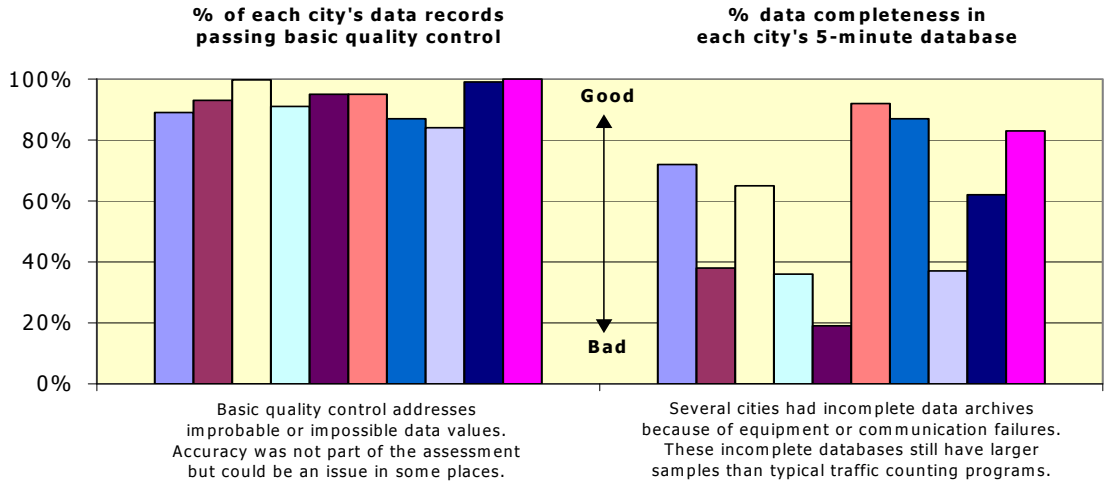


Figure 2.2: Quality and Completeness of Representative City Databases. (Turner, 2001)

Other systems, such as that used by the Washington State Department of Transportation, flag data as "good," "bad," or "suspect" (Ishimaru and Hallenbeck, 1999). These are identified through bounds checking (ensuring that observed occupancy, volume, and speed measurements meet basic physical feasibility requirements), noticing if measurements do not change (e.g., if a loop continually reports the same count, one may assume it is malfunctioning), and so on. Figure 2.2 displays quality and completeness statistics for several city databases.

From a planning perspective, data quality has received more attention, partially due to the different time scale involved: while many operations data needs are real-time and require very recent data, the time needed to perform quality checks is less burdensome for long-term planning applications. For instance, the Virginia Department of Transportation uses the following classification scheme:

Code 0 - Not Reviewed
Code 1 - Acceptable for Nothing
Code 2 - Acceptable for Qualified Raw Data Distribution
Code 3 - Acceptable for Raw Data Distribution
Code 4 - Acceptable for Use in AADT Calculation
Code 5 - Acceptable for All TMS Uses

Elsewhere, several European countries (the Netherlands, Switzerland, Germany, France, and the United Kingdom) perform automated data checking by comparing measured data to historical data for consistency (FHWA, 1997).

More sophisticated data checking measures might include consistency checking from period-to-period, from lane-to-lane, or verification against traffic flow theory (ibid.) These errors can arise from a number of sources, including environmental conditions, improper installation or calibration, communication failures, inadequate maintenance, and errors inherent in the chosen technology (Margiotta, 2002). All contribute to imperfect information, which must be addressed if this data is to be acceptable to planners.

2.3 Case Studies

The following case studies provide a representative look at several possible data archiving systems that have been implemented. Two separate systems are in place in Seattle, one operated by the state department of transportation and the other by a transit agency. In contrast to the other four studies reviewed here, Detroit's was designed for planning uses from the beginning. The archive used in the Minneapolis-St. Paul area had its genesis in a collaboration between the state and a university. The Maricopa County RADS system, in the Phoenix area, is the most recent and is still under development. Finally, California's PeMS system has a far broader scope, storing and integrating data collected throughout the entire state. Much of the information in this section comes from FHWA (2005).

A number of other regions also archive data; these include Atlanta, Chicago, New York City, Ft. Worth, Houston, Portland, San Antonio, Toronto, and the state of Virginia. These are not profiled here, in order to focus on five regions whose archival systems are particularly noteworthy. Information on the others can be found in FHWA (1999) and Bertini et al. (2005).

2.3.1 Seattle

ITS data from loop detectors and ramp meters in the Seattle metropolitan area is stored in an archive maintained by the Washington State Transportation Center (TRAC). When initiated in 1981, the goal of this archive was to provide ongoing data to evaluate and justify innovative traffic management measures such as HOV lanes and ramp metering. This data also is used to ensure that the reversible express lane schedules on I-5 and I-90 are optimal. Improvements to such technologies are tested using this data. One such example is the introduction of real-time fuzzy-logic ramp metering control (Taylor and Meldrum, 2000)

Seattle freeway loop detectors are polled for occupancy and volume readings at 20-second intervals, and this data is stored in an Oracle database. Five-minute aggregations of these data are stored in a flat-file database. Both of these databases store these values in binary form. A program called CDR has been developed to access this database, and allows one to retrieve data from selected loops during a given time period (Ishimaru and Hallenbeck, 1999).

Common uses for this data are WSDOT operational studies and agency publications regarding traffic counts (WSDOT, 2006). FHWA (2005) notes that this archive is also used for planning tasks, but is "planning-oriented in terms of how agencies plan for operations as opposed to the more traditional capital-improvements planning function." This archive is also used by regional planners, particularly the Puget Sound Regional Council (the local MPO), by consultants, and by researchers at institutions such as the University of Washington and The University of Texas at Austin.

Basic checks are performed to see if the data is consistent with fundamental physical requirements (such as jam density, or saturation flow). Failing data are flagged as "suspect" or "bad," although no suggestion is made for a more plausible value. Thus, it is the responsibility of those using the data to decide how to handle flawed data.

A second data archive is maintained by King County Metro, a transit agency also operating in the Seattle metropolitan area. This archive primarily consists of automated vehicle location (AVL) data reported by buses equipped with this technology (Casey et al., 1998; Wall and Dailey, 1999; Cathey and Dailey, 2003).

Some of this information is also revealed to the public using products such as BUSVIEW, which allows travelers to see real-time bus location (Figure 2.3). This is useful, for instance, to see if a bus is running late. However, others have realized the value of this data for estimating historical travel times and for air quality improvement strategies. In general, King County Metro is willing to share this data with anyone who requests it.

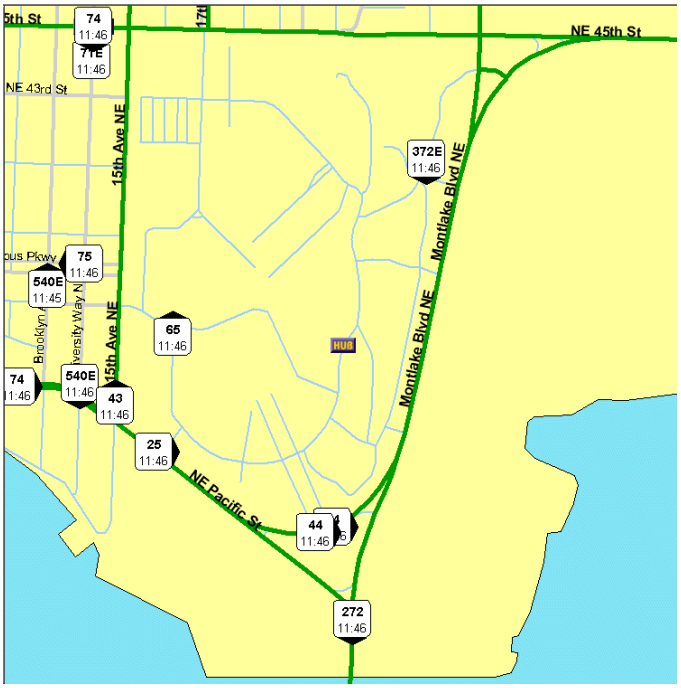


Figure 2.3: Sample BUSVIEW interface.

2.3.2 Detroit

The Michigan ITS (MITS) center stores data from loop detectors. An older system of loops reports data as 1-hour lane volumes; a newer double-loop system (in the Detroit region constituting the majority of system loops) reports volume, occupancy, and speed data at 2-minute

intervals. The original intent of this system was to simplify traffic counts, hence the 1-hour aggregation performed by the older loops. In contrast to some of the other systems described in this section, MITS was primarily constructed with planning aims in mind. Indeed, its primary users are Michigan Department of Transportation (MDOT) planners and the Southeast Michigan Council of Governments (SEMCOG).

Data are stored in a flat-file database, and quality control is performed automatically. If a loop is disabled (for instance, during maintenance), data is flagged accordingly. Data are also checked against historical values for consistency.

2.3.3 Minneapolis-St. Paul

The archived data management system (ADMS) operating in the Minneapolis-St. Paul region is a collaborative effort between the Minnesota Department of Transportation (MnDOT) and the University of Minnesota at Duluth (UM Duluth). Thus, its main users are MnDOT operations personnel and UM Duluth researchers. The system has been in operation since 1997 and has been used, for instance, to defend ramp metering programs to the state legislature and to provide data to university researchers.

Data are collected from loop detectors throughout the metropolitan area, compressed, and are loaded onto a UM Duluth FTP server daily. This archive is publicly accessible (ftp://tdrl.d.umn.edu/pub/tmcdata/) along with several utilities that can aggregate data and provide descriptive statistics; these can be downloaded from http://www.d.umn.edu/~tkwon/TDRLSoftware/Download.html (URLs current as of January 2007).

This data is stored in a flat file format, and is formatted in a manner similar to what is received by the TMC. Automatic quality control checking marks data as “good,” “suspect,” or “bad.”

2.3.4 Phoenix

Currently under development, the Maricopa County Arizona Regional Archive Data Server (Maricopa County RADS) will store traffic volumes, speeds, road closures, incident information, and other data. Main users of this system are expected to include Maricopa Association of Governments (MAG) planners, Arizona Department of Transportation (ADOT) ITS personnel, local traffic engineers, transit agencies, commercial vehicle operators, and private-sector information providers.

As the system is not yet operational, few details can be provided on specific database implementations or quality control procedures. However, the Internet is intended to be a key distribution point for this data. It also is anticipated that multiple database formats will be used, to facilitate use by multiple groups of users.

2.3.5 California

In contrast to the systems mentioned above, California’s freeway Performance Measurement System (PeMS) involves data obtained from freeway sensors, police dispatch systems, and weather information throughout the entire state, rather than just a single metropolitan area. PeMS was initiated jointly by the California Department of Transportation (Caltrans) and the University of California at Berkeley (UC Berkeley) (Varaiya, 2002). PeMS contains three large databases storing incident, weather, and freeway information. Processing and

interface layers allow access to this information in a variety of formats. The impetus for this system came from a 1997 white paper, and it was operational by 2002.

Because the system was initiated by operations personnel, most of the use made of this system by Caltrans is operational in nature, such as travel time prediction, congestion monitoring, and level of service analysis. This data also is used by university researchers, planning organizations (such as the San Diego Association of Governments), the public (via the Internet), and the media—for instance, the *Los Angeles Times* used this archive during a transit strike to report on its impacts.

Quality control is performed automatically, and inconsistent data are automatically replaced by estimates derived from other detectors in the vicinity.

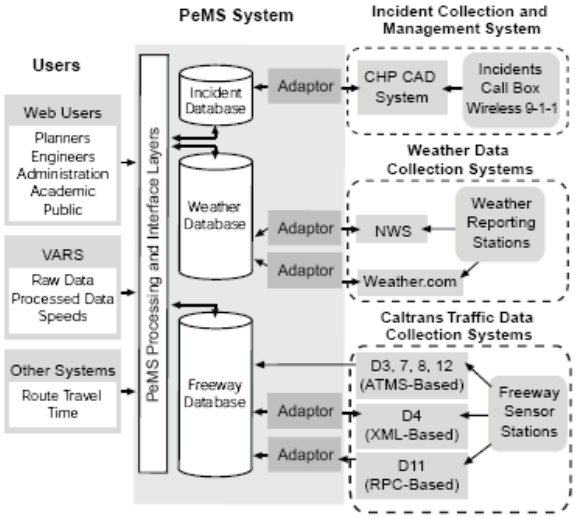


Figure 2.4: Schematic of Major PeMS Components (FHWA, 2005)

2.4 Institutional Barriers

As seen in the above case studies, although ITS solutions are frequently used for operations purposes, they are not typically considered in transportation planning. Institutional barriers tend to impede the linkage of operations and planning data. Issues include the endorsement of ITS data to peer agencies or the general public, devising a means of communication across geographic boundaries and between agencies, and coordination of data collection needs.

Turner (2001) suggests several reasons why data archiving is not as widespread as one might expect, given its potential benefits:

- Operations personnel tend to see their role as "crisis managers," overlooking the longer-term value of the data they use.
- Operations personnel may feel that others (e.g. planners) are the primary beneficiaries of archived data, and that responsibility for implementing such systems belongs to them.
- Planning personnel are unfamiliar with ITS data collection technologies, and as a result are uncomfortable using them.

- Data archiving was not considered when ITS systems were deployed in the past.
- Institutional issues relating to control, maintenance, and ownership of archived data.

Such obstacles are largely institutional in nature; the technology to archive and share data and is readily available. To help overcome these barriers and streamline the incorporation of ITS data into the planning process, the USDOT (2000) recommends the following specific strategies:

- Create an ITS committee involving regional stakeholders,
- Educate elected officials and transportation executives,
- Include ITS in MPO planning documents,
- Develop a program for regional ITS projects,
- Educate MPO staff,
- Educate other stakeholders,
- Educate the general public on specific ITS projects,
- Use ITS advocates in the region,
- Utilize the National ITS Architecture to develop a regional architecture,
- Use peer-to-peer networking,
- Involve academia in regional ITS planning
- Determine data collection needs for planning purposes, and
- Determine the most efficient and effective ways to distribute and apply ITS-generated data.

2.5 Data Archiving in Texas

To supplement the review of data archival systems implemented in other regions, a twelve-question survey was distributed to nine TMCs in Texas. (Appendix B contains a copy of the questions in the survey). Of these, five responses were received, from TMCs located in Austin, Dallas, El Paso, Fort Worth, and San Antonio. This section summarizes their replies, followed by some discussion of common elements in their responses.

2.5.1 Austin

The Austin TMC controls 75 closed-circuit television (CCTV) cameras and nearly 2500 inductive loop detectors. CCTV data is not archived, but loop detector data (including volume, occupancy, and speed measurements, as well as vehicle classification) is stored in an ASCII comma-separated file. These files are available online for up to two years, after which they remain archived on CD. Data are not stored if they are clearly in error, providing some basic data quality assurance.

Once stored, the data is retrieved as needed for particular projects. Typical uses of this data include congestion studies, volume forecasting performed by the Capital Area Metropolitan

Planning Organization, and detector maintenance. The Texas Transportation Institute (TTI) also accesses this data on a quarterly basis for its own studies, and has provided this TMC with recommendations for improving the usefulness and efficiency of the data format.

2.5.2 Dallas

The Dallas TMC uses approximately sixty microwave vehicle detectors and fifty video-based detectors to record speed, volume, and occupancy measurements, as well as vehicle classification. These data are stored in comma-delimited ASCII files, which are then compressed and archived online at a publicly-accessible website. Measurements obtained at a particular freeway location are only recorded if the detector in each lane reports valid data.

Usage statistics are not maintained for this data, but the North Central Texas Council of Governments and TTI both make regular use of this data for various research projects. Also, attempts to integrate this data with other regional sources are underway, including data imputation.

2.5.3 El Paso

The El Paso TMC operates over eighty CCTV cameras, and two types of automated vehicle detectors: 281 traditional loop detectors as well as 148 microwave detectors. The automated detectors record their data in comma-delimited text files, which are currently stored in separate locations. The loop detector data is archived on a dedicated server, which only maintains data for one week, although they indicate that this system is slowly being phased out in favor of microwave-based detectors. Microwave detector data, on the other hand, is stored on a separate computer, which is currently not integrated with the rest of the operation system. Currently there is no data checking performed, although research is underway to provide validation techniques for the microwave detectors. This data has been used for assorted traffic studies for three years; an example of a current use is a project to predict travel times.

2.5.4 Fort Worth

The Fort Worth TMC collects data from over 1500 loop detectors and 180 side-fire radar detectors; as the loop detectors age and stop functioning, they are being replaced with the radar detectors. All of these sensors report volume, occupancy, and vehicle classification; the radar detectors and selected pairs of closely-spaced loops also record speed information. These paired loops also perform error checking by ensuring that they report consistent results. The radar detectors employ specific noise reduction and anti-ghosting algorithms to counteract these sources of error. Currently, none of this data is stored in any permanent way, since the data archival component of the system proposal was not funded. Efforts are underway to add this to the system.

2.5.5 San Antonio

The San Antonio TMC makes use of a variety of detector types in their system: forty video image detection systems, five side-fire radar detectors (eighty more by the end of the year), and over 1600 loop detectors, which record speed, volume, and occupancy data. Statistical sampling is used to verify accuracy of the data. All of this data is initially stored on the servers collecting this data; after 24 hours, it is transferred permanently to disk arrays and also made available on a public FTP server for one year. Tape backups also exist to protect against any data

loss. Work is underway to develop databases to facilitate access to this data. This data is most often used for research and statistical purposes.

2.5.6 Opinions on Using Archived Data for Planning Purposes

The responders also were asked to give their opinion on the largest obstacles standing in the way of using ITS data for planning purposes; their responses are paraphrased below, in no particular order.

- **Managing a large amount of data.** In order to help, the vast quantities of ITS data need to be distilled to a useful summary.
- **Ensuring data accuracy**. Data standards for planners may be different, and there may be a lack of trust of ITS devices due to these issues.
- **Providing useful formats for all users**. Different clients need different information; for instance, planners want geocoded data by street and block, while operations personnel typically prefer locations identified by milepost or centerline station, while the devices may report their location in a latitude/longitude system.
- **Different data goals**. Planners are generally seeking system-wide information, such as trip origins and destinations; this is in contrast to operational demands focused on specific corridors or facilities, not the users themselves.

2.6 Conclusions

Developing a system that allows ITS data to be used for planning purposes carries tremendous potential, as the data already being collected by ITS devices can greatly expand the amount of information available to planners, while enabling the calibration and use of innovative transportation models with sufficient data requirements. The easiest way to accomplish this task is through the development of a centralized, automated data archive that stores this information, along with an easy-to-use program (or suite of programs) to enable ready access to this information.

The case studies profiled above provide some guidance as to the variety of such systems available, and possible applications. The diversity in these systems comes about from their different origins (whether initiated by operations personnel, planning personnel, or university researchers), different scopes of coverage (from volume data alone to databases containing weather and incident information as well) and a number of different quality control procedures (typically determined according to primary data use). Based on current practices in Texas, it seems that developing uniform data archiving formats and quality control measures can greatly facilitate this type of data sharing.

In the end, the barriers to implementing such a system are primarily institutional rather than technological. Therefore, it is crucial to clearly explain the benefits of such a system, and to design it with all of the involved parties in mind.

Chapter 3. Prototype System

3.1 Introduction

This chapter describes the prototype archive system in greater detail, first in terms of how data is collected, stored, and retrieved. This is followed by an example action plan presenting specific steps that could be taken to implement such a system.

This system interfaces with TMCs, which collect data directly from detectors and then transmit it to the archive. A modular design is proposed, in which TMCs only interface with the central archive, and in which all data processing algorithms are housed in the archive itself. This allows TMCs to be easily added or removed from the archive at any future point in time, for any reason. Furthermore, the proposed design is technologically flexible, and can accommodate a very broad range of current and future traffic detector designs.

3.2 Database Design and Data Formats

This section describes the purpose and components involved in the data archiving system. Illustrated schematically in Figure 3.1, the process begins as traffic detectors report data they collect. This data is then preprocessed, and undergoes a reliability testing (quality check) procedure to quantify confidence in the reading, based on fundamental, historical, and network-based considerations. (This procedure is described more fully in Chapter 4). Optionally, data that is considered unreliable can be replaced with interpolated data at this point—whether this is desirable or even permissible depends on the purposes for which the data will be used. Nevertheless, it is an option at this point.

Next, the data is stored in an archival database, to be accessed by operators generating "reports" (for instance, daily volumes for weekdays in March, for a given set of detectors). These reports consist of database queries, which return the desired data. Missing or suspect data can optionally be replaced by interpolated or estimated data at this stage as well. A web interface has been constructed to enable access from a variety of locations (Figure 3.2).

Recall that the system is designed with maximum future flexibility in mind, including the ability to handle data from multiple detector technologies with ease. To facilitate this, all incoming data are preprocessed into a common form, indicating the following information:

1. The detector ID number
2. The detector type
3. The information recorded by the detector (e.g., volume, speed, or occupancy)
4. The spatial location of the detector
5. The time span over which the data was collected

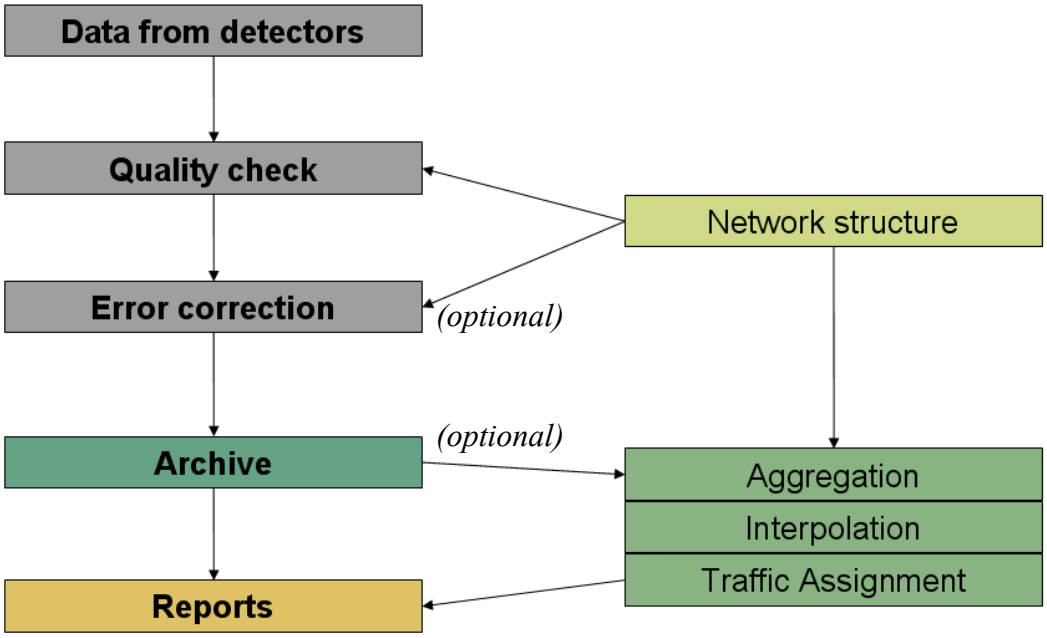
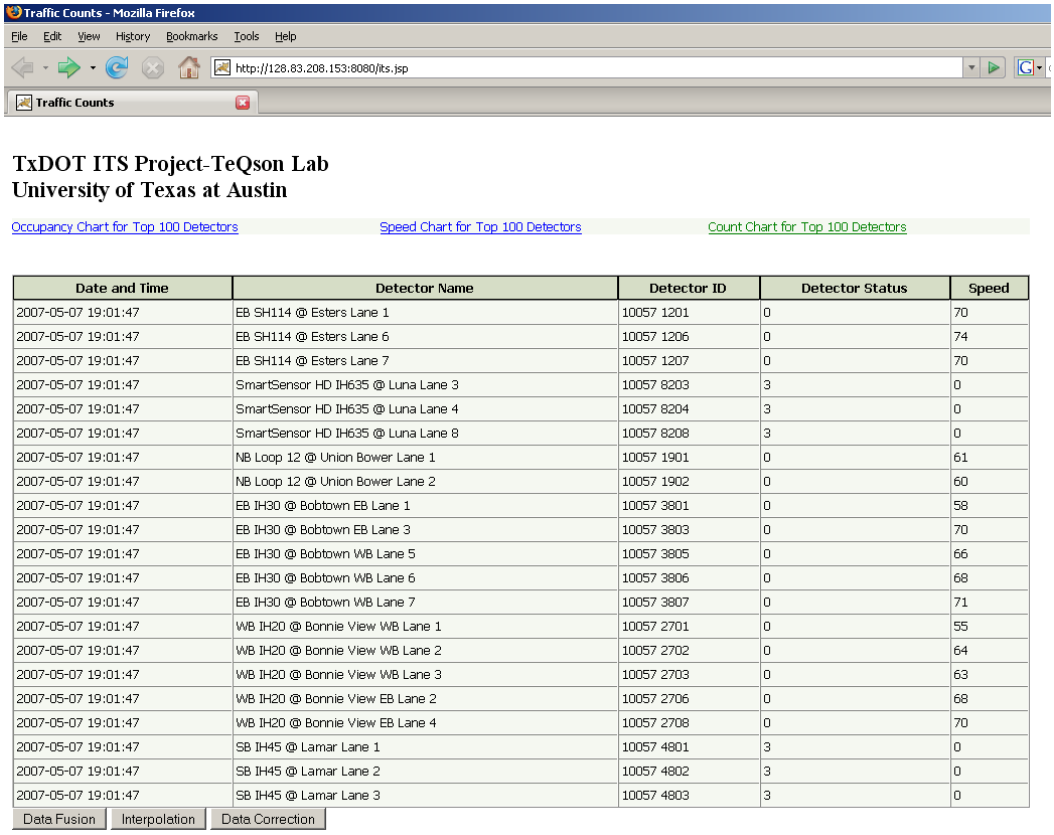


Figure 3.1: System design schematic



Date and Time	Detector Name	Detector ID	Detector Status	Speed
2007-05-07 19:01:47	EB SH114 @ Esters Lane 1	10057 1201	0	70
2007-05-07 19:01:47	EB SH114 @ Esters Lane 6	10057 1206	0	74
2007-05-07 19:01:47	EB SH114 @ Esters Lane 7	10057 1207	0	70
2007-05-07 19:01:47	SmartSensor HD IH635 @ Luna Lane 3	10057 8203	3	0
2007-05-07 19:01:47	SmartSensor HD IH635 @ Luna Lane 4	10057 8204	3	0
2007-05-07 19:01:47	SmartSensor HD IH635 @ Luna Lane 8	10057 8208	3	0
2007-05-07 19:01:47	NB Loop 12 @ Union Bower Lane 1	10057 1901	0	61
2007-05-07 19:01:47	NB Loop 12 @ Union Bower Lane 2	10057 1902	0	60
2007-05-07 19:01:47	EB IH30 @ Bobtown EB Lane 1	10057 3801	0	58
2007-05-07 19:01:47	EB IH30 @ Bobtown EB Lane 3	10057 3803	0	70
2007-05-07 19:01:47	EB IH30 @ Bobtown WB Lane 5	10057 3805	0	66
2007-05-07 19:01:47	EB IH30 @ Bobtown WB Lane 6	10057 3806	0	68
2007-05-07 19:01:47	EB IH30 @ Bobtown WB Lane 7	10057 3807	0	71
2007-05-07 19:01:47	WB IH20 @ Bonnie View WB Lane 1	10057 2701	0	55
2007-05-07 19:01:47	WB IH20 @ Bonnie View WB Lane 2	10057 2702	0	64
2007-05-07 19:01:47	WB IH20 @ Bonnie View WB Lane 3	10057 2703	0	63
2007-05-07 19:01:47	WB IH20 @ Bonnie View EB Lane 2	10057 2706	0	68
2007-05-07 19:01:47	WB IH20 @ Bonnie View EB Lane 4	10057 2708	0	70
2007-05-07 19:01:47	SB IH45 @ Lamar Lane 1	10057 4801	3	0
2007-05-07 19:01:47	SB IH45 @ Lamar Lane 2	10057 4802	3	0
2007-05-07 19:01:47	SB IH45 @ Lamar Lane 3	10057 4803	3	0

Data Fusion | Interpolation | Data Correction

Figure 3.2: Web interface to data archive

In particular, a common standard should be defined for recording spatial and temporal coordinates; latitude/longitude or facility/milepost are the most useful possibilities for encoding spatial information, while coordinated universal time (UTC) is a useful standard for recording times.

To design a flexible and efficient way of archiving traffic data, we utilize the relational open-source database PostgreSQL and employ the widely-used database normalization techniques (Codd, 1970 and 1971). The techniques are applied to organize relational database such that the duplication of information is minimized, redundancy is eliminated, and inconsistency can be discovered. It as well safeguards the database against different types of structural problems and abnormalities

The database outlined has four tables, as shown in Tables 3.1–3.4. The tables are at least in first, second, and third normal form (1NF, 2NF and 3NF), which means that the tables faithfully represent the relations of records and the appropriately address the dependency issues.

Table 3.1: Detector Details Table

	ID	Name	Detector Type
Type	INTEGER	TEXT	INTEGER
Modifiers	NOT NULL, UNIQUE	-	NOT NULL
Example	1	Mopac	1

Table 3.2: Detector Type Description Table

	Detector Type	Description
Type	INTEGER	TEXT
Modifiers	NOT NULL	NOT NULL
Example	1	Inductive Loop Detector

Table 3.3: Data Collected Table

	ID	Status	Date	Time	Volume	Speed	Occupancy
Type	INTEGER	INTEGER	DATE	TIME	INTEGER	DOUBLE PRECISION	DOUBLE PRESCISION
Modifiers	NOT NULL, UNIQUE	-	-	-	-	-	-
Example	1	1	01/30/08	14:00	0	60	0

Table 3.4: Status Description Table

	Status	Description
Type	INTEGER	TEXT
Modifiers	NOT NULL	NOT NULL
Example	1	Normal

3.3 Action Plan

3.3.1 Introduction

A centralized archive for traffic data can be successfully implemented in three phases, receiving data collected by multiple types of ITS detectors and allowing different users to generate custom reports for a variety of purposes.

The three phases can be summarized as follows:

Phase I. **Establish policies and standards for data storage and communications—**the desired functionality must be established, along with the hardware and communications infrastructure needed to support it. Leadership roles must also be assigned, and the archive's physical location must be identified.

Phase II. **Implement central data archive—**the chosen hardware must be identified, the database software initialized, and additional code must be written to implement error checking, error correcting algorithms, and provide an interface and reporting structure to allow access to the data.

Phase III. Integrate TMCs with central data archive—this phase must be performed once for each TMC that is connected to the archive, and again if an additional TMC is to be added. Programs must be written to convert data from the format used by the TMC's detectors into the standard format used by the archive, and the communication link between the TMC and archive must be established. Depending on the chosen error checking and error correcting routines, additional parameters may need to be specified at this point as well.

Further details of each phase are provided in this subsection, along with a list of specific tasks that must be accomplished in each phase.

3.3.2 Phase I. Establish policies and standards for data storage and communications

This first phase is concerned with preliminary matters, determining what data will be archived from what traffic management centers (TMCs) and how, the necessary communication infrastructure, the physical location(s) for the archive, and the management structure for operating and maintaining the archive. Although basic, substantial time should be invested at this stage to ensure that the archive is useful for both current and future needs. Future considerations to take into account include implementation of new traffic detector technologies, anticipated changes in data reporting requirements and standards, and new or proposed TMCs that may be built after implementing the archive. This phase is divided into four tasks, each of which is discussed in more detail below.

Task 1.1 Determine scope of data archive
Task 1.2 Determine communication and equipment needs
Task 1.3 Determine "chain of command"
Task 1.4 Identify physical location(s) for archive

Task 1.1 Determine scope of data archive

This task is further divided into three subtasks, each of which is concerned with identifying a key structural component of the archive:

Subtask 1.1.1—Specify desired functionality

At a minimum, one must decide (a) what basic data must be recorded (e.g., volume and speed), (b) the frequency at which data must be received (e.g., at least daily or hourly), (c) who may access the data (e.g., restricted to agency personnel or publicly available), and (d) how the data should be accessed (e.g., the structure of database queries, forms, and reports)

Subtask 1.1.2—Identify participating TMCs

Based on the functionalities specified in Subtask 1.1.1, as well as the desired scope of the archive and interest in participation, a set of TMCs will be identified for participation in the archiving system. Note that not all of these TMCs need to participate from the very beginning, as the overall implementation plan is modular and allows additional TMCs to be introduced to the system at any time.

Subtask 1.1.3—Specify data formats

According to the TMCs and data requirements selected in Subtasks 1.1.1 and 1.1.2, specific data formats will be identified, including encoding schemes for detector location(latitude/longitude vs. facility/milepost), data time (local time vs. UTC), and units of measurement for volume, speed, and density.

Task 1.2 Determine communication and equipment needs

In the previous task, the locations of participating TMCs were identified, along with the necessary data reporting requirements, including reporting frequency. Based on these, the appropriate mode(s) of communication (e.g., fiber optic, wireless, telephone, or radio) can be identified, along with the computer hardware needed for the central archive. In particular, enough storage space must be provided to store the data; a server, operating system, and software are needed to run the database program and communicate with users to generate reports; and backup and redundancy considerations, such as off-site storage or a redundant array of independent disks (RAID), to ensure continued access to the data in case of equipment failure.

Task 1.3 Determine "chain of command"

The departments and personnel responsible for implementing and maintaining this archive must be identified, within the context of the intended users, participants, and functionality.

Task 1.4 Identify physical location(s) for archive

The location of the hardware and software must be specified, as well as the location of any backup or redundancy options. Depending on the communication modes, it may be desirable to locate this in the proximity of one or more TMCs.

3.3.3 Phase II. Implement central data archive

The second phase is concerned with making the data archive operational, setting up the necessary hardware, software, and communications equipment. Note that integration with

individual TMCs is accomplished in a later phase. This division emphasizes the modular nature of the implementation plan, in that the central archive can operate independently of specific TMCs. This phase is divided into three tasks, each of which is described in further detail below:

Task 2.1 Install needed computational equipment and communications infrastructure
Task 2.2 Implement database and interface
Task 2.3 Enable remote access

Task 2.1 Install needed computational equipment and communications infrastructure

Installation of the equipment identified in Task 1.2 is accomplished during this task, physically establishing the database and preparing it for installation of software and communication with TMCs and end users.

Task 2.2 Implement database and interface

This task is divided into four steps, each corresponding to a software-related need that must be implemented.

Subtask 2.2.1—Initialize database

A suitable database platform must be identified, and the relevant fields and forms constructed, based on the specifications chosen in Phase I.

Subtask 2.2.2—Implement reliability assessment algorithms

Quality control algorithms for the data must be programmed and integrated with the database, such as the continuous set theoretic algorithm discussed in Chapter 4.

Subtask 2.2.3—Author routines to generate reports

Depending on the desired uses and the specific database platform, it may be necessary to write additional code to generate reports portraying data in the desired format, as well as the necessary forms to allow users to interface with the database.

Subtask 2.2.4—Implement interpolation/data correction scheme

One or more data correction and interpolation schemes should also be programmed and integrated within the database, with a clear option available to users as to whether interpolated data is appropriate for their application.

Task 2.3 Enable remote access

The final task in this phase is to activate and test communication links between the archive and other locations. In particular, TMCs must be able to access the archive to deposit data, and other users must be able to access the archive to generate reports and download traffic data.

3.3.4 Phase III. Integrate TMCs with central data archive

This phase is unique in that it needs to be performed several times, once for each TMC that will be connected to the archive. Once the data archive is operational, this step will need to be performed again if additional TMCs need to be connected. For each TMC, the following three steps need to be performed:

Task 3.1 Generate needed parameters for the central archive
Task 3.2 Develop routines for translating detector data to central archive format
Task 3.3 Establish communications link

Task 3.1 Generate needed data for the central archive (e.g. upstream/downstream; jam density & capacity; other detectors for interpolation)

The type and location of every detector operated by the TMC must be stored in the database before data archiving can begin. Furthermore, depending on the algorithms chosen for calculating data reliability and/or interpolating missing data, additional parameters must be specified, such as roadway capacity and jam density, or the IDs of upstream and downstream detectors.

Task 3.2 Develop routines for translating detector data to central archive format

Different detectors report data differently, and these need to be translated to a common format before transmission to the archive. Thus, it may be necessary to write a computer program to accomplish this conversion.

Task 3.3 Establish communications link

Finally, the communication link between the TMC and the archive must be created and tested. After this, the flow of data can commence.

Chapter 4. Data Reliability and Imputation

4.1 Introduction

No detector device is perfect, and thus the question of data quality is critical for any data recording and archival process. It is important to construct rigorous measures of reliability or confidence in traffic data; for instance, if archived data will be used to influence policy decisions through traffic studies, one should have high confidence in the validity of the measurements.

Section 4.2 of this chapter defines a general "reliability index," indicating the consistency of each data measurement with fundamental traffic relations, historical data, and upstream/downstream measurements. The reliability index is an integer ranging from zero (no confidence in the data; it is almost certainly wrong) to ten (very high confidence; it is very likely to be correct).

Following identification of suspicious data, it may be desirable to generate a more trustworthy estimate of the true value. Such procedures are also important for addressing problems of missing data. The research literature contains several examples of data imputation algorithms, and these are described and compared in Section 4.3, alongside three new algorithms developed in this project. Traffic data can also be estimated even in locations where no detector is present, using extrapolation techniques; these are described in Section 4.4.

4.2 Reliability Indices

It is desirable to apply the same metric to all data that are received. Thus, the reliability index is given a general definition that can be applied regardless of the type of detector. While this gives maximum flexibility in admitting innovative technologies, this requires that the reliability index not depend on the specific data type received (e.g., volume or occupancy). Further, there are multiple measures of consistency that are not easily compared: as an example, if the data is consistent with historical measurements, but not with upstream data, how should these two assessments be reconciled?

Continuous set theory (CST) provides a technical framework for making these assessments commensurable. Initially developed four decades ago, continuous set theory is based on two facts: it is often impossible to precisely classify measurements without arbitrariness; and decision-making must be made using such imprecise assessments. For instance, how should the thresholds for "historical consistency" be defined? One choice is to define a single interval of traffic volume, for which any measurement within that interval is deemed consistent, and any other measurement inconsistent. But with this definition of consistency, two volume measurements that are nearly identical can be classified differently, if one is just within the interval, and the other just outside it.

This is an issue because, fundamentally, "consistency" is an inherently imprecise concept that cannot properly be defined by discrete intervals. CST remedies this deficiency by allowing measurements to be both "consistent" and "inconsistent" to varying degrees, giving a fuller picture of the quality of the data.

As defined in this project, the reliability index is based on three separate consistency assessments. First, the data is checked for **fundamental** consistency: is it consistent with basic traffic laws? Are the volume and density measurements reasonable? Second, the **network** consistency is examined: how do the measured data compare to upstream and downstream

observations? Finally, the **historical** consistency is measured, according to previous records at the same location.

For each of these three checks, the data is classified among four categories: **probably correct** (PC), **maybe correct** (MC), **probably incorrect** (PI), and **absolutely incorrect** (AI). For instance, the data may be considered "probably correct" regarding network consistency, but "probably incorrect" regarding historical consistency. As mentioned above, CST allows for partial membership in multiple categories; for instance, the data may be two-thirds "probably correct," and one-third "maybe correct." A decision table and continuous set theoretical decision rule are then used to determine the overall reliability index, taking all of these measures into account.

The remainder of this section is organized as follows. First, there's a brief overview of continuous set theory; its concepts are introduced, along with a simple example. Next, the three consistency checks—fundamental, network, and historical—are described and defined in turn. Finally, the decision-making process is defined, and an example is given showing how this process can be applied to a hypothetical data measurement. For a fuller treatment of the mathematics of continuous set theory, see, for instance, von Altrock (1995).

4.2.1 Continuous Set Theory

Developed by Lofti Zadeh in 1965, continuous set theory directly addresses the notion that decisions must often be made based on inherently imprecise quantities. Although continuous set theory is a mathematically rigorous concept, CST-based classification does not accomplish anything that could not be done using previously-existing methods. Rather, its prime strength is its ability to model complicated decision problems using intuitive, natural language. This makes the process of calibrating and tuning models considerably easier, and facilitates comprehension of the model for all interested parties, regardless of specific expertise.

For instance, as explained below, a key element in CST decision making is the construction of a set of decision rules. In the context of traffic data archiving, one decision rule might be "if the data is probably correct (PC) according to fundamental rules, is probably incorrect (PI) when looking at nearby detectors, but may be correct (MC) historically, then, overall, the data may be correct (MC)." By phrasing the decision in natural language, the process of tuning is much easier: in this situation, if experience shows that this rule is expressing too much confidence in the data, it can be changed: "if the data is PC according to fundamental rules, PI when looking at nearby detectors, and MC historically, overall the data is probably incorrect (PI)." The mathematical details of exactly how this change affects the classification process are "under the hood," so to speak, and need not be fully understood by an operator calibrating the data. (Of course, these details are fully explained in this section). Note that such decision rules also provide an elegant solution to the problem of combining nominally incommensurable evaluations, by couching them in natural (although rigorously-defined) terminology.

CST is applied widely in automated decision-making contexts. For instance, many automobiles use CST to control automatic transmissions or braking, many thermostats use CST to control heating and air conditioning systems, and some dishwashers use CST to adjust cycle parameters. The common characteristic of all of these decision problems, and the reason why CST works well for these, is their need to account for multiple input parameters that may not fall neatly into clearly-defined categories. For instance, when controlling air conditioning, both outdoor temperature and current energy consumption levels are continuously-varying quantities

that are not well-suited to discrete categorization. For the remainder of this section, we use this example to illustrate how CST works. The following sections explain, in detail, how this procedure can be applied to transportation data archiving.

The first step in applying CST is to translate each of the input parameters into the corresponding linguistic parameters, a process called *fuzzification*. Continuing with the air conditioning example, we need to fuzzify the indoor and outdoor temperatures, as well as the current energy consumption. For instance, Figure 4.1 shows one way to fuzzify indoor air temperature among three categories ("cold," "warm," and "hot"). For each possible temperature and each category, we must decide to what degree that category describes the temperature, ranging from zero (not at all descriptive) to one (fully descriptive).

For instance, an indoor temperature of 50°F is almost universally considered cold, and not at all warm or hot; so we assign COLD = 1, WARM = 0, and HOT = 0. Likewise, an indoor temperature of 90°F is definitely hot, and not at all cold or warm; so we assign COLD = 0, WARM = 0, and HOT = 1. For intermediate temperatures, it is less clear; while a temperature of 65°F is certainly not hot, it is somewhat cold, and somewhat warm; so we assign COLD = 0.5, WARM = 0.5, and HOT = 0.

The figure shows the membership rules mapping each temperature to how COOL, WARM, and HOT that temperature is. This fuzzification rule is arbitrary, and it can be tuned as necessary to adapt to a particular situation.

Although not shown here, similar fuzzification rules must be developed for describing energy usage (for instance, into two categories, PEAK and OFF-PEAK; see Figure 4.2). Thus, at any moment in time, we observe the temperature and energy usage, and have a linguistic description of the current state as some combination of COOL, WARM, HOT, LOW energy consumption, and HIGH energy consumption.

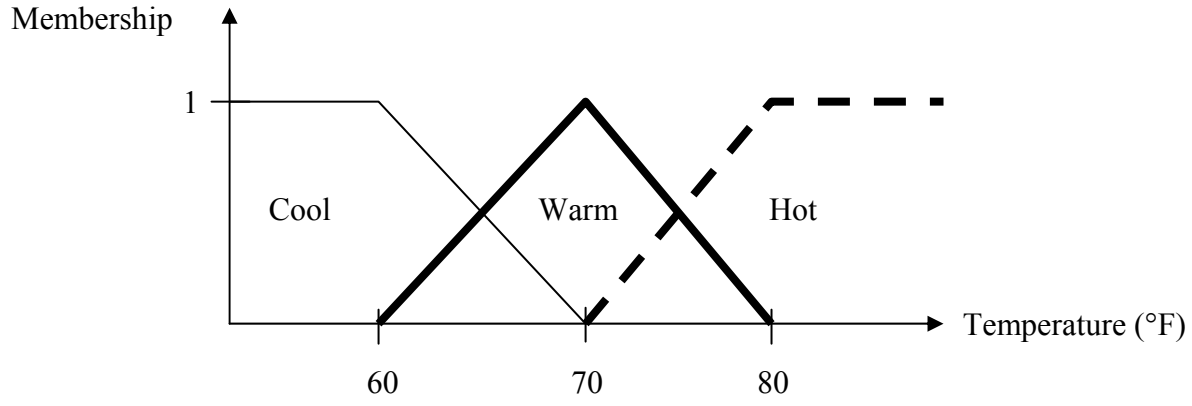


Figure 4.1: Fuzzification of indoor air temperature.

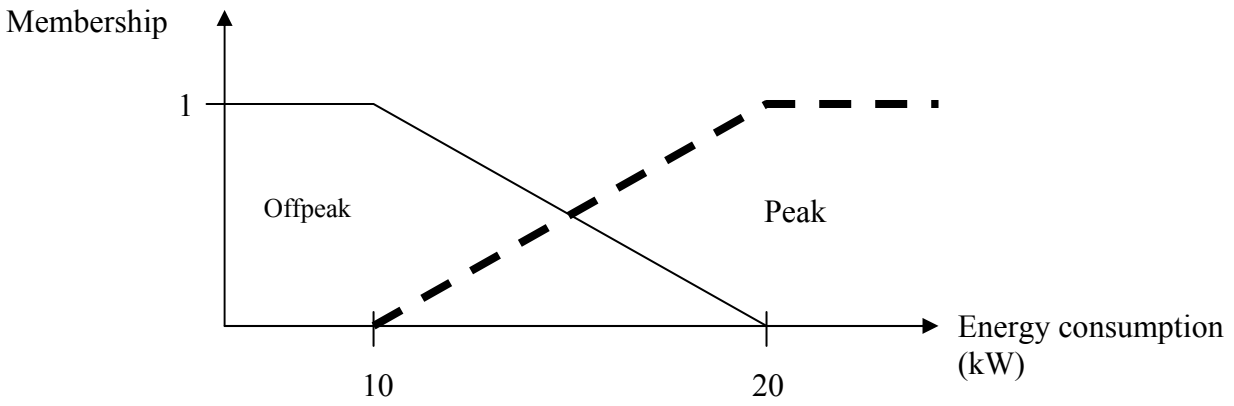


Figure 4.2: Fuzzification of energy consumption

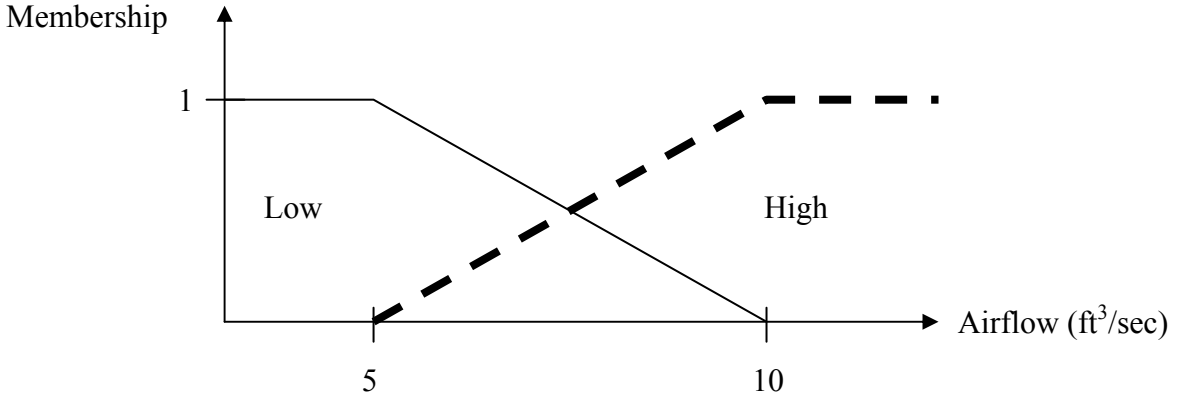


Figure 4.3: Defuzzification converting linguistic airflow to a numeric value

Table 4.1: Decision rules for air conditioning example

	OFF-PEAK energy use	PEAK energy use
COOL temperature	LOW airflow	LOW airflow
WARM temperature	HIGH airflow	LOW airflow
HOT temperature	HIGH airflow	HIGH airflow

The next step is to construct a *decision table*, showing what action should be taken for each combination of states. For this case, the desired output is the air flow through the air conditioning system (with a higher flow rate corresponding to faster cooling and higher energy expenditures); although this is a precise number, we fuzzify for the purpose of constructing a decision table (e.g., LOW airflow, and HIGH airflow; Figure 4.3), and one possible decision table is shown in Table 4.1. Note that this table frames the decision-making process in intuitive, natural-language terms.

At any given moment, the current temperature and energy use describe the "state" of the system. It is often the case that more than one "state" exists simultaneously due to fuzzification. For instance, if the indoor temperature is 75 degrees and the energy use is 25 kW, it is partially WARM and partially HOT, but energy use is entirely HIGH. That is, the states (WARM, PEAK)

and (HOT, PEAK) both exist partially, and the decisions for both of these states (LOW and HIGH airflow, respectively) should be taken into account. The exact method for accomplishing this is somewhat involved, and is described more fully below; but the following key properties are satisfied:

- All applicable states contribute to the final decision—both (HOT, PEAK) and (WARM, PEAK) influence the end result.
- If one state is more applicable than another, it should weigh more heavily in the decision—if it is mostly (HOT, PEAK) and only slightly (WARM, PEAK), the airflow should be more HIGH (the decision for HOT, PEAK) and less LOW (the decision for WARM, PEAK), and vice versa.
- The decision is *deterministic*; that is, it the same temperature and energy use always result in the same airflow rate.

Finally, a *defuzzification* process converts this combined decision into a crisp, numeric output (the rate of airflow). Note that this process is easy to tune; if, for instance, one implements this system and observes that the airflow is usually too high, one can either alter the decision table (to favor LOW airflow), or adjust the state definitions (perhaps the definitions of HIGH and LOW airflow are incorrect). The same advantages can be applied to transportation data as well, as described in the following sections.

Recall that the classification is made according to three inputs (fundamental consistency; network consistency; and historical consistency), so the decision table is three-dimensional, rather than two-dimensional as in the above example. The fuzzification procedures for fundamental, network, and historical consistency are described, followed by presentation of the decision table and defuzzification principle.

4.2.2 Fundamental Consistency

Fundamental consistency is a measure of data quality indicating the plausibility of observed data, based on basic physical constraints and laws of traffic flow theory. In particular, if q, u, and k denote the volume, space-mean speed, and density of traffic, the following relations must hold:

1. $q = uk$
2. $q \leq q_{max}$
3. $k \leq k_{jam}$

Dimensional analysis ensures that the first relation must hold; by definition, volume must equal the product of speed and density. The second relation requires that the flow be no greater than the capacity of q_{max} of the section, and the third requires that the density not exceed the jam density k_{jam}.

Although it might appear that these relations must hold in an absolute sense, continuous set theory is still applicable for several reasons. First, there is inherent instability and heterogeneity in a traffic stream: vehicles and driver behavior are not uniform, and traffic streams are continuously evolving; in fact, the entire notion of a continuous traffic stream is a modeling simplification. Thus, although these laws must hold on average at an aggregate level, the state at a particular detector location at any given moment may not satisfy these exactly.

Second, the notions of maximum capacity and jam density are not exact, and depend on driver and freeway characteristics that cannot be entirely observed, and which need not be stationary with time.

In fact, it is exactly for reasons such as these that continuous set theory is beneficial. It is not useful to either classify a data observation as either fully consistent (satisfying relations 1–3 exactly) or inconsistent (at least one of relations 1–3 is violated), but rather to have a continuously-varying measure of consistency, ranging from entirely consistent to entirely inconsistent.

As mentioned in the previous section, traffic data is fuzzified into four different states: probably correct (PC), maybe correct (MC), probably incorrect (PI), and absolutely incorrect (AI). We develop a "consistency score" for each of the relations 1–3 above, showing how plausible a given data observation is with respect to each of them. All consistency scores are nonnegative real numbers, with lower numbers indicating greater confidence. In particular, the respective consistency scores for the three relations are

1. $\left|\frac{q}{uk}-1\right|$
2. $\max\{q/q_{max}-0.9,\,0\}$
3. $\max\{k/k_{jam}-0.9,\,0\}$

The first relation penalizes any deviation from the requirement $q = uk$, while the second and third penalize excess volume and density, defined as a volume-capacity or density-jam density ratio exceeding 0.9. The *maximum* of these three consistency scores is taken to be the overall fundamental consistency score, that is,

$$FC(q,u,k)=\max\left\{\left|\frac{q}{uk}-1\right|,\frac{q}{q_{\max}},\frac{k}{k_{jam}}\right\}$$

which is then fuzzified according to the following relations (see Figure 4.4 for a graphical representation):

$$\begin{aligned} PC &= \max\{1-10FC,0\} \\ MC &= \max\{\min\{10FC,2-10FC\},0\} \\ PI &= \max\{\min\{10FC-1,3-10FC\},0\} \\ AI &= \max\{\min\{10FC-2,1\},0\} \end{aligned} \tag{4.1}$$

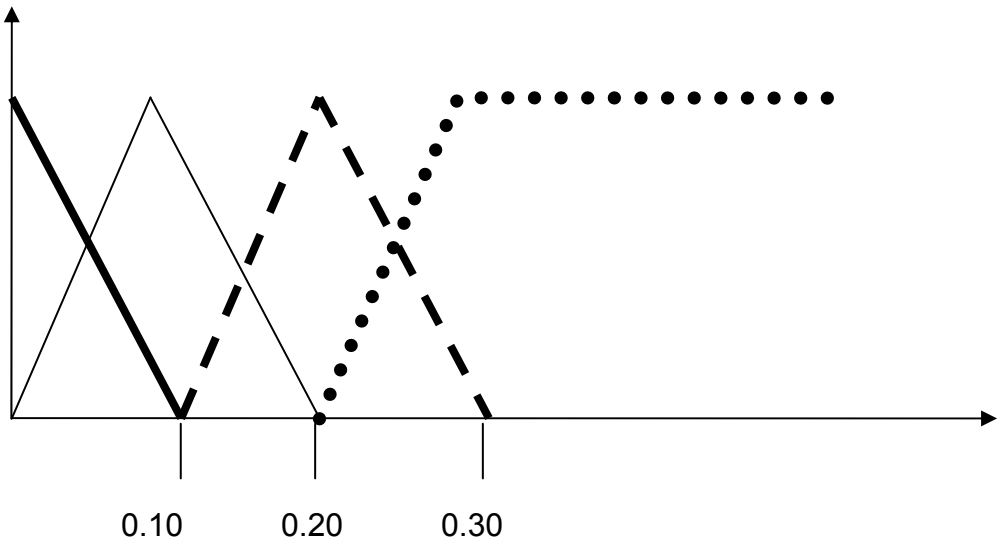


Figure 4.4: Fuzzification for fundamental consistency (regions from left to right are PC, MC, PI, and AI).

4.2.3 Network Consistency

Network consistency makes explicit use of the spatial relationships between road segments, especially the notions of "upstream" and "downstream" links. For access-controlled facilities with detectors on all on- and off-ramps, flow conservation implies that every vehicle detected by a sensor on the mainline must already have been detected by another detector upstream, and must be detected again downstream. For illustration, in Figure 4.5 any vehicle passing point C must have already passed point A or B, and must pass point D or E in the future.

Thus, when looking at traffic volume counts over a sufficiently long period of time (an hour or more), we expect the sum of the counts at A and B to equal both to the count at C, and to the sum of the counts at D and E, a property that is exploited in determining the network consistency score of a detector reading.

The network consistency score can either be calculated in real-time, or applied to an aggregated set of data offline. If performed in real time, as data is received, it is only possible to make a comparison with the upstream detectors (A and B in Figure 4.5) as the vehicles have not yet reached the next location downstream. In this case, we compare the values $V_A + V_B$ and V_C, where V_C is the current volume reading at detector C, and V_A and V_B are the volume readings at these detectors at a suitable point in the past. Given a current travel speed u, and distances d_A and d_B to detectors A and B, respectively, the appropriate time offsets are $t_A = d_A/u$ and $t_B = d_B/u$; it is appropriate to use interpolation if these time offsets do not exactly correspond to a past data measurement.

However, the previously-calculated reliability scores for the readings at A and B should also be taken into account: if the volume at either of these sites is highly unreliable, one should not expect the sum $A + B$ to be a reliable predictor of the volume at location C either.

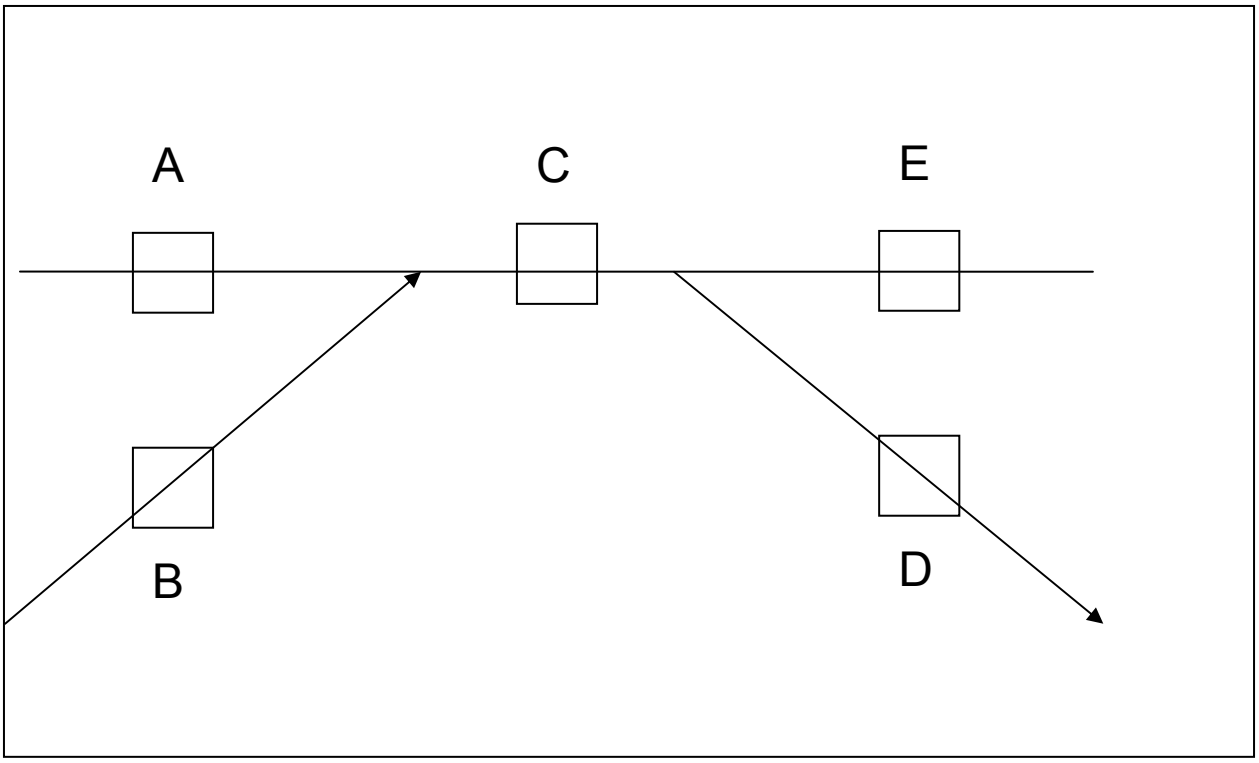


Figure 4.5: Demonstration of network consistency

Keeping this in mind, the network consistency score can be calculated as:

$$NC(q,u,k) = \left| \frac{RI_A + RI_B}{V_C} \times \frac{V_A + V_B - V_C}{20} \right|$$

where RI_A and RI_B are the reliability indices for A and B at the time offset determined by the present speed u. Effectively, this formula penalizes any deviation from true flow conservation ($[V_A + V_B]/V_C - 1$), scaled according to the maximum possible value of $RI_A + RI_B$. This value is then fuzzified according to the following relations (see Figure 4.6 for a graphical representation):

$$\begin{aligned} PC &= \max\{1 - 20NC, 0\} \\ MC &= \max\{\min\{20NC, 2 - 20NC\}, 0\} \\ PI &= \max\{\min\{20NC - 1, 2 - 10NC\}, 0\} \\ AI &= \max\{\min\{10NC - 1, 1\}, 0\} \end{aligned} \tag{4.2}$$

When there is no upstream detector, NC is set to be 0.075 to indicate that the data is 50% MC and 50% PI.

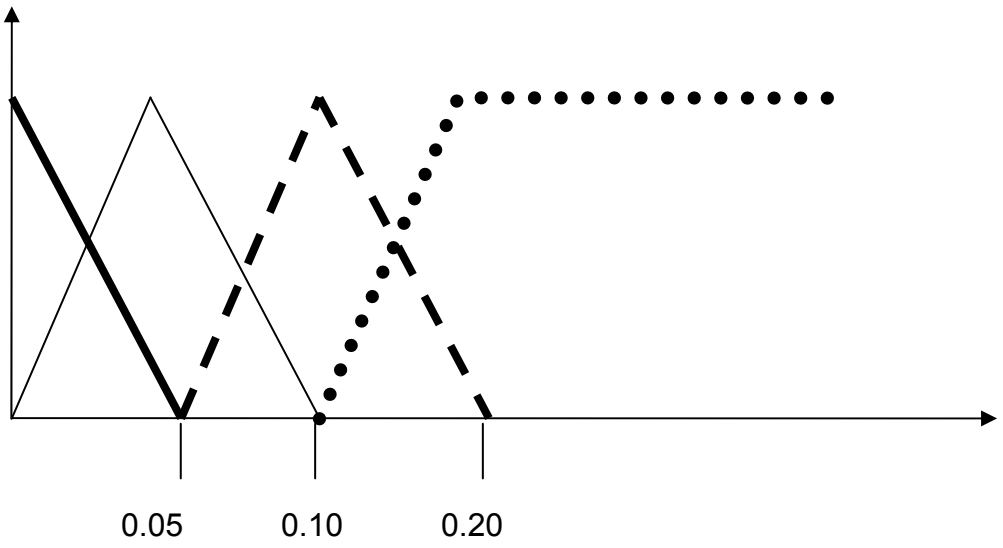


Figure 4.6: Fuzzification diagram for network consistency

4.2.4 Historical Consistency

The key notion of **historical consistency** is that past observations can be used to help identify suspect data in the present. In comparison to the more basic fundamental and network consistency checks, it is not difficult to imagine instances where a perfectly valid data reading might appear inconsistent given the historical record (due to a severe incident or extreme weather conditions, for instance). Such examples actually indicate the strength of the three-part consistency check: in such cases, the fundamental and network consistency scores can compensate for a low historical consistency score; while the historical consistency score can play a valuable role in detecting other anomalies not captured by the other two scores.

One must decide the data that is to be used for a historical consistency check; possible options are "all readings at this location for the same day of the week," "all readings at this location, at the same time of day (regardless of day of week)," "all readings at this location, at the same time of day and day of the year," etc. The appropriate choice depends on data availability, and the degree of aggregation for the counts being used.

These data form a set X of previous data (volume, speed, density, and volume difference with upstream detector), with which the current observation Y will be compared. In particular, for each element y_i of Y, the percentile p_i of this element will be calculated according to the set X, and each element is assigned a consistency score $c_i = 0.5 - |p_i - 0.5|$.

The overall historical **in**consistency score is taken to be the maximum of these:

$$HC(q,u,k,q_d) = \min_i \{0.5 - | p_i - 0.5 |\}$$

which is then fuzzified according to the following relations

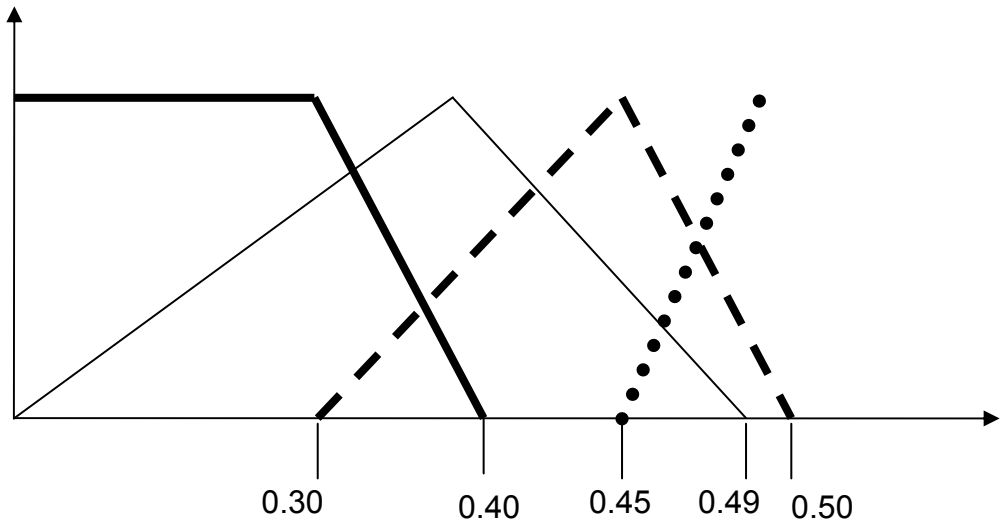


Figure 4.7: Fuzzification for historical consistency

$$
\begin{aligned}
PC &= \max\{\min\{1, 4 - 10HC\}, 0\} \\
MC &= \max\left\{\min\left\{2.5HC, \frac{49}{9} - \frac{100}{9}HC\right\}, 0\right\} \\
PI &= \max\left\{\min\left\{\frac{20}{3}HC - 2, 10 - 20HC\right\}, 0\right\} \\
AI &= \max\{20HC - 9, 0\}
\end{aligned}
\tag{4.3}
$$

4.2.5 Decision Table

In the previous three steps, three separate reliability assessments are made: fundamental, network, and historical. The "decision table" step combines these to produce an overall reliability assessment, using the concept of an "aggregate state." Aggregate states are simply the superposition of the input assessments, and the set of all aggregate states is the Cartesian product of the sets of component states. For instance, let the set $\{F_{PC}, F_{MC}, F_{PI}, F_{AI}\}$ represent the four possible states (probably correct, maybe correct, etc.) according to the fundamental consistency criterion, with $\{N_{PC}, N_{MC}, N_{PI}, N_{AI}\}$ and $\{H_{PC}, H_{MC}, H_{PI}, H_{AI}\}$ representing the possible states according to the network and historical consistency criteria.

An aggregate state represents all three of these components simultaneously. For instance, one possible aggregate state is (F_{PC}, N_{MC}, H_{PI}), corresponding to the case where the data is "probably correct" fundamentally, "maybe correct" from a network perspective, and "probably incorrect" from a historical perspective. Every possible combination is allowed; thus there are $4^3 = 64$ possible aggregate states, as shown in Table 4.2

Table 4.2: Enumeration of aggregate states.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	(F_{PC}, N_{PC}, H_{PC})	(F_{PC}, N_{PC}, H_{MC})	(F_{PC}, N_{PC}, H_{PI})	(F_{PC}, N_{PC}, H_{AI})
Maybe correct	(F_{PC}, N_{MC}, H_{PC})	(F_{PC}, N_{MC}, H_{MC})	(F_{PC}, N_{MC}, H_{PI})	(F_{PC}, N_{MC}, H_{AI})
Probably incorrect	(F_{PC}, N_{PI}, H_{PC})	(F_{PC}, N_{PI}, H_{MC})	(F_{PC}, N_{PI}, H_{PI})	(F_{PC}, N_{PI}, H_{AI})
Absolutely incorrect	(F_{PC}, N_{AI}, H_{PC})	(F_{PC}, N_{AI}, H_{MC})	(F_{PC}, N_{AI}, H_{PI})	(F_{PC}, N_{AI}, H_{AI})

(a) Table for data that are "probably correct" according to the "fundamental" criterion.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	(F_{MC}, N_{PC}, H_{PC})	(F_{MC}, N_{PC}, H_{MC})	(F_{MC}, N_{PC}, H_{PI})	(F_{MC}, N_{PC}, H_{AI})
Maybe correct	(F_{MC}, N_{MC}, H_{PC})	(F_{MC}, N_{MC}, H_{MC})	(F_{MC}, N_{MC}, H_{PI})	(F_{MC}, N_{MC}, H_{AI})
Probably incorrect	(F_{MC}, N_{PI}, H_{PC})	(F_{MC}, N_{PI}, H_{MC})	(F_{MC}, N_{PI}, H_{PI})	(F_{MC}, N_{PI}, H_{AI})
Absolutely incorrect	(F_{MC}, N_{AI}, H_{PC})	(F_{MC}, N_{AI}, H_{MC})	(F_{MC}, N_{AI}, H_{PI})	(F_{MC}, N_{AI}, H_{AI})

(b) Table for data that are "maybe correct" according to the "fundamental" criterion.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	(F_{PI}, N_{PC}, H_{PC})	(F_{PI}, N_{PC}, H_{MC})	(F_{PI}, N_{PC}, H_{PI})	(F_{PI}, N_{PC}, H_{AI})
Maybe correct	(F_{PI}, N_{MC}, H_{PC})	(F_{PI}, N_{MC}, H_{MC})	(F_{PI}, N_{MC}, H_{PI})	(F_{PI}, N_{MC}, H_{AI})
Probably incorrect	(F_{PI}, N_{PI}, H_{PC})	(F_{PI}, N_{PI}, H_{MC})	(F_{PI}, N_{PI}, H_{PI})	(F_{PI}, N_{PI}, H_{AI})
Absolutely incorrect	(F_{PI}, N_{AI}, H_{PC})	(F_{PI}, N_{AI}, H_{MC})	(F_{PI}, N_{AI}, H_{PI})	(F_{PI}, N_{AI}, H_{AI})

(c) Table for data that are "probably incorrect" according to the "fundamental" criterion.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	(F_{AI}, N_{PC}, H_{PC})	(F_{AI}, N_{PC}, H_{MC})	(F_{AI}, N_{PC}, H_{PI})	(F_{AI}, N_{PC}, H_{AI})
Maybe correct	(F_{AI}, N_{MC}, H_{PC})	(F_{AI}, N_{MC}, H_{MC})	(F_{AI}, N_{MC}, H_{PI})	(F_{AI}, N_{MC}, H_{AI})
Probably incorrect	(F_{AI}, N_{PI}, H_{PC})	(F_{AI}, N_{PI}, H_{MC})	(F_{AI}, N_{PI}, H_{PI})	(F_{AI}, N_{PI}, H_{AI})
Absolutely incorrect	(F_{AI}, N_{AI}, H_{PC})	(F_{AI}, N_{AI}, H_{MC})	(F_{AI}, N_{AI}, H_{PI})	(F_{AI}, N_{AI}, H_{AI})

(d) Table for data that are "absolutely incorrect" according to the "fundamental" criterion.

Since the reliability decision is based on an aggregate assessment, we need to determine the degree of membership of each of the aggregate states. Recall that continuous set theory allows partial set membership; that is, a particular piece of data might be 0.5 Probably Correct (Fundamental), 0.8 Maybe Correct (Network), and 0.3 Maybe Correct (Historical), what is the membership in the aggregate set (F_{PC}, N_{MC}, H_{MC})? By convention, this membership is defined to be the *lowest* of any of the components: in the case, 0.3. In equation form, the membership μ(F,N,H) of an aggregate state (F,N,H) is defined as

$$\mu(F,N,H) = \min\{\mu(F),\mu(N),\mu(H)\}$$

where μ(F) is the membership in the fundamental assessment, etc. Membership in all sixty-four aggregate states is calculated in this way. Note that the sum of all of these memberships need not equal one; this is not a concern.

Each aggregate state is associated with an overall assessment of data quality: if the data is probably correct (fundamental), maybe correct (network), and maybe correct (historical), what is the combined assessment? Table 4.3 shows one way to accomplish this, although this can certainly be tuned as necessary:

4.2.6 Defuzzification

The last step is to create the actual, numeric "reliability index" from the continuous aggregate state memberships determined by the decision table. To do this, a mapping from the aggregate assessment to the reliability index is needed, similar to the functions used to fuzzify the input data. Equations (4.4) are appropriate for this use:

Table 4.3: Decision rules for all aggregate states.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	Probably correct	Probably correct	Probably correct	Probably correct
Maybe correct	Probably correct	Probably correct	Maybe correct	Maybe correct
Probably incorrect	Maybe correct	Maybe correct	Maybe correct	Probably incorrect
Absolutely incorrect	Maybe correct	Maybe correct	Probably incorrect	Absolutely incorrect

(a) Table for data that are "probably correct" according to the "fundamental" criterion.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	Probably correct	Maybe correct	Maybe correct	Probably incorrect
Maybe correct	Probably correct	Maybe correct	Probably incorrect	Probably incorrect
Probably incorrect	Maybe correct	Maybe correct	Probably incorrect	Probably incorrect
Absolutely incorrect	Maybe correct	Probably incorrect	Probably incorrect	Absolutely incorrect

(b) Table for data that are "maybe correct" according to the "fundamental" criterion.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	Probably incorrect	Probably incorrect	Probably incorrect	Probably incorrect
Maybe correct	Probably incorrect	Probably incorrect	Probably incorrect	Absolutely incorrect
Probably incorrect	Probably incorrect	Probably incorrect	Probably incorrect	Absolutely incorrect
Absolutely incorrect	Probably incorrect	Absolutely incorrect	Absolutely incorrect	Absolutely incorrect

(c) Table for data that are "probably incorrect" according to the "fundamental" criterion.

Network ↓ / Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	Probably incorrect	Probably incorrect	Probably incorrect	Absolutely incorrect
Maybe correct	Probably incorrect	Probably incorrect	Absolutely incorrect	Absolutely incorrect
Probably incorrect	Probably incorrect	Absolutely incorrect	Absolutely incorrect	Absolutely incorrect
Absolutely incorrect	Absolutely incorrect	Absolutely incorrect	Absolutely incorrect	Absolutely incorrect

(d) Table for data that are "absolutely incorrect" according to the "fundamental" criterion.

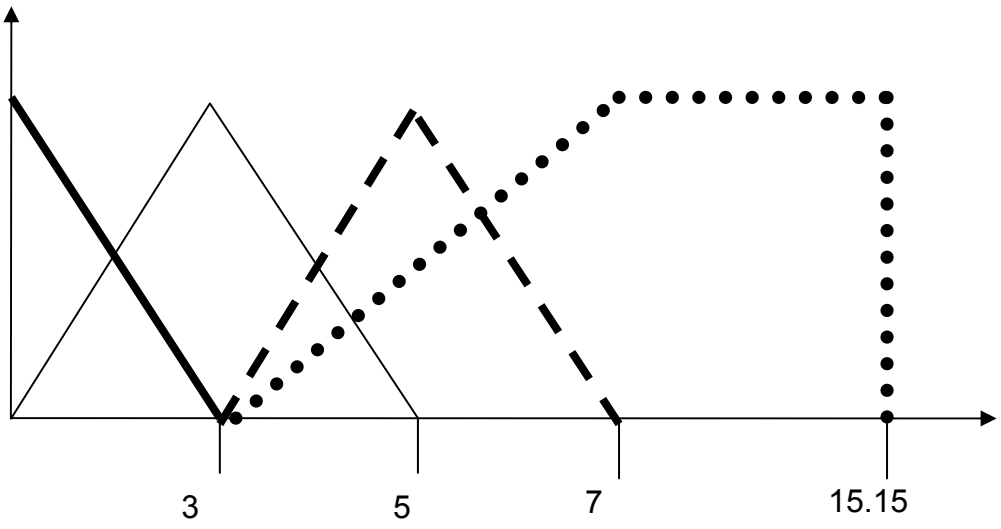


Figure 4.8: Fuzzification of reliability index

$$
\begin{aligned}
PC &= \max\left\{\min\left\{1, \frac{1}{2}RI - \frac{5}{2}\right\}, 0\right\} \\
MC &= \max\left\{\min\left\{\frac{1}{2}RI - \frac{3}{2}, \frac{7}{2} - \frac{1}{2}RI\right\}, 0\right\} \\
PI &= \max\left\{\min\left\{\frac{1}{2}RI, \frac{5}{2} - \frac{1}{2}RI\right\}, 0\right\} \\
AI &= \max\left\{\min\left\{1 - \frac{1}{2}RI, I\left(RI < 15.15\right)\right\}, 0\right\}
\end{aligned}
\tag{4.4}
$$

where *RI* denotes the overall reliability index, which we are trying to find. The most common method for doing this uses "centroid defuzzification," an algorithm based on the following intuition: for each of the sixty-four aggregate states, we know its degree of membership μ^*, as well as the aggregate decision (PC, MC, PI, or AI). In Figure 4.9, the centroid and area of the region underneath the appropriate curve and the horizontal line $\mu = \mu^*$. The area-weighted "center of mass" is then calculated, and its horizontal coordinate is taken to be the reliability index *RI*.

Note that the PC set is defined for reliability indices as high as 15.15, even though the reliability index is capped as 10. This is done to ensure that this maximum value of the reliability index can be attained, by allowing the *x*-coordinate for the centroid of the PC trapezoid to be as high as 10.

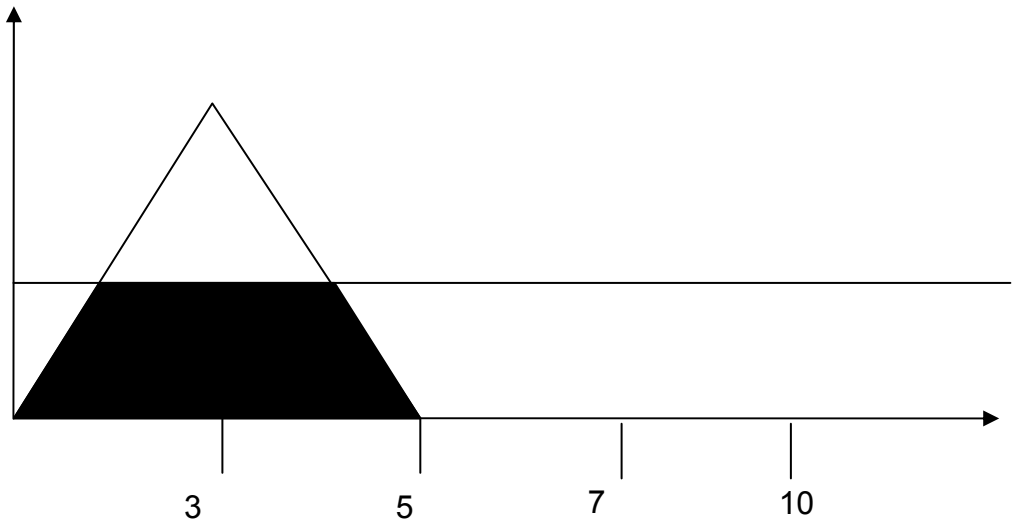


Figure 4.9: Area below the MC curve and the horizontal line $\mu = \mu*$

4.2.7 Example

This section demonstrates how the reliability index is calculated, working with five-minute aggregated data. Assume a detector records volume, speed, and traffic density on a two-lane freeway segment, with an estimated capacity of 4000 veh/hr, and an estimated jam density of 250 veh/mi.

Assume a detector on a freeway segment records the following data for a five-minute interval:

- Volume: 150 veh
- Speed: 55 mph
- Density: 30 veh/mi (many loop detectors report occupancy, which can be used to directly estimate density)

Another detector is located three quarters of a mile upstream of this detector, with one onramp in between. Forty seconds ago, this detector recorded a volume of 125 veh (with a reliability index of 7), and the detector on the onramp recorded a volume of 45 veh (with a reliability index of 2).

We begin by calculating the reliability scores for each of the three criteria:

Fundamental: First, scores are calculated for all three fundamental traffic requirements:

$$\left|\frac{q}{uk} - 1\right| = 0.09$$

$$\max\{q/q_{max} - 0.9,\ 0\} = 0$$

$$\max\{k/k_{jam} - 0.9,\ 0\} = 0$$

The maximum of these, or 0.09, is taken as the raw fundamental consistency score. Using the equations (1), this gives $\mu(F_{PC}) = 0.1$ and $\mu(F_{MC}) = 0.9$. (That is, according to the fundamental consistency criterion, the data is 0.1 "probably correct" and 0.9 "maybe correct.")

Network: The network consistency score is calculated as

$$\left|\frac{RI_A + RI_B}{V_C} \times \frac{V_A + V_B - V_C}{20}\right| = \left|\frac{9}{150} \times \frac{20}{20}\right| = 0.06$$

Using the equations (2), this gives $\mu(N_{MC}) = 0.8$ and $\mu(N_{PI}) = 0.2$. (That is, according to the fundamental consistency criterion, the data is 0.8 "maybe correct" and 0.2 "probably incorrect.")

Historical: To find historical consistency, one needs to know what percentile the current data form in the past data archive at this time and location. Assume that the volume, speed, and density data occur at the 43rd, 65th, and 37th percentiles, respectively; then the overall historical inconsistency score is calculated as

$$\min\{0.5 - |0.43 - 0.5|, 0.5 - |0.65 - 0.5|, 0.5 - |0.37 - 0.5|\} = 0.35$$

Using the equations (3), this gives $\mu(H_{PC}) = 0.5$, $\mu(H_{MC}) = 0.875$ and $\mu(N_{PI}) = 0.333$. (That is, according to the historical consistency criterion, the data is 0.5 "probably correct," 0.875 "maybe correct," and 0.3333 "probably incorrect.") The fact that these do not sum to one is not relevant.

Decision Table: Of the 64 aggregate states, there are twelve in which the observed data have a positive membership:

$\mu(F_{PC}, N_{MC}, H_{PC}) = \min\{0.1, 0.8, 0.5\} = 0.1$
$\mu(F_{PC}, N_{MC}, H_{MC}) = 0.1$
$\mu(F_{PC}, N_{MC}, H_{PI}) = 0.1$
$\mu(F_{PC}, N_{PI}, H_{PC}) = 0.1$
$\mu(F_{PC}, N_{PI}, H_{MC}) = 0.1$
$\mu(F_{PC}, N_{PI}, H_{PI}) = 0.1$
$\mu(F_{MC}, N_{MC}, H_{PC}) = 0.5$
$\mu(F_{MC}, N_{MC}, H_{MC}) = 0.8$
$\mu(F_{MC}, N_{MC}, H_{PI}) = 0.333$
$\mu(F_{MC}, N_{PI}, H_{PC}) = 0.2$
$\mu(F_{MC}, N_{PI}, H_{MC}) = 0.2$
$\mu(F_{MC}, N_{PI}, H_{PI}) = 0.2$

For each of these, we look up the appropriate rule in the decision table; and find the area and centroid of the region beneath that rule's graph and the horizontal line corresponding to the degree of membership, as shown in Table 4.4:

Thus, we calculate the x-coordinate of the total centroid to be 5.86, which we round up for a reliability index of 6.

Table 4.4: Degree of membership, areas, and centroids for aggregate states

State	μ	Decision	Area	Centroid (x-coordinate)
F_{PC}, N_{MC}, H_{PC}	0.1	PC	0.49	9.174
F_{PC}, N_{MC}, H_{MC}	0.1	PC	0.49	9.174
F_{PC}, N_{MC}, H_{PI}	0.1	MC	0.38	5
F_{PC}, N_{PI}, H_{PC}	0.1	MC	0.38	5
F_{PC}, N_{PI}, H_{MC}	0.1	MC	0.38	5
F_{PC}, N_{PI}, H_{PI}	0.1	MC	0.38	5
F_{MC}, N_{MC}, H_{PC}	0.5	PC	2.25	9.56
F_{MC}, N_{MC}, H_{MC}	0.8	MC	1.92	5
F_{MC}, N_{MC}, H_{PI}	0.333333	PI	1.111	3
F_{MC}, N_{PI}, H_{PC}	0.2	MC	1.92	5
F_{MC}, N_{PI}, H_{MC}	0.2	MC	1.92	5
F_{MC}, N_{PI}, H_{PI}	0.2	PI	0.72	3

4.2.8 Field Data Demonstration

In this section, the proposed reliability scheme is demonstrated using field data. The necessary code for the experimental testing is implemented in the Java language.

The reliability index calculated from the proposed CST is implemented on three randomly chosen detectors from Dallas. The locations of the detectors are depicted in Figure 4.10. These three detectors are installed on the eight lanes on westbound I-20, with an estimated capacity of 16,000 vehicles per hour and an estimated jam density of 1,000 vehicles per mile.

The upstream detector, middle detector and downstream detector are located near Dowdy Ferry Road, Bonnie View Road, and Houston School Road respectively. The travel time between upstream and middle detector is approximately eight minutes, while the travel time between middle and downstream detector is approximately two minutes. The experimental results are summarized in Table 4.5 and Figure 4.11:

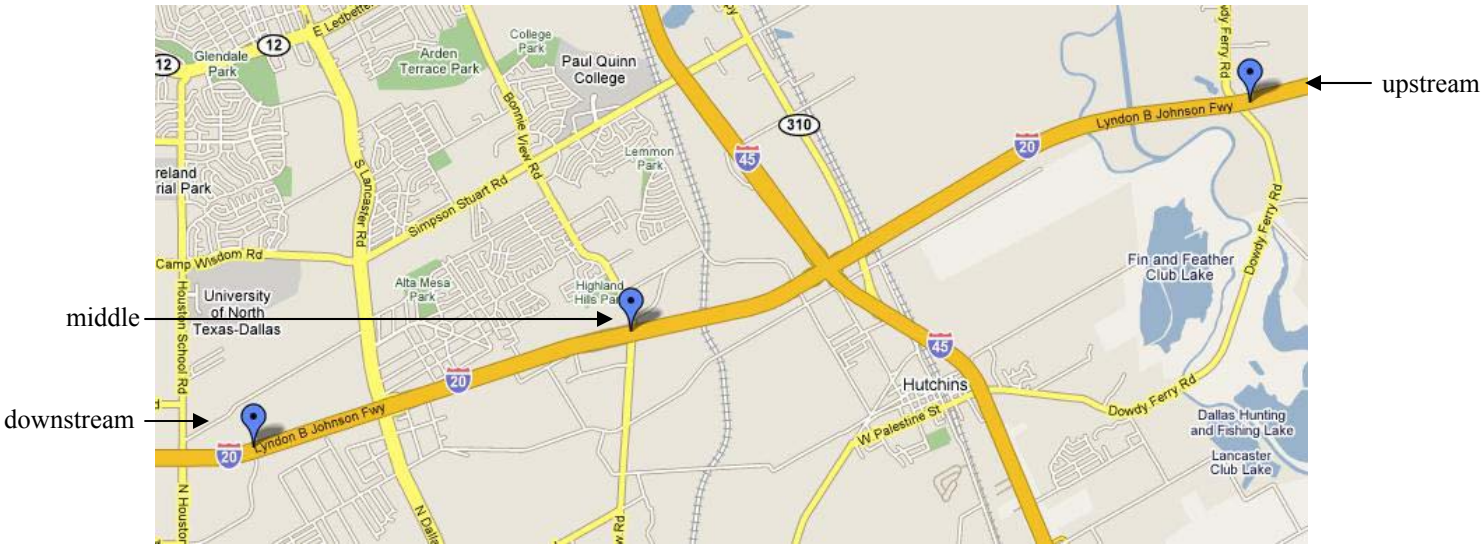


Figure 4.10: Detector Locations (from Google Maps)

Table 4.5: Observed distribution of reliability indices

RI	Upstream Detector		Middle Detector		Downstream Detector	
	Number	Percentage	Number	Percentage	Number	Percentage
0	42,567	31.45%	6,307	3.74%	6,293	3.84%
1	0	0.00%	3	0.00%	4	0.00%
2	831	0.61%	30	0.02%	47	0.03%
3	3,168	2.34%	351	0.21%	214	0.13%
4	3,230	2.39%	169	0.10%	190	0.12%
5	4,329	3.20%	12,898	7.64%	11,232	6.86%
6	31,645	23.38%	7,352	4.35%	5,108	3.12%
7	25,346	18.73%	8,906	5.28%	6,361	3.88%
8	19,136	14.14%	4,016	2.38%	4,565	2.79%
9	5,098	3.77%	11,934	7.07%	16,005	9.77%
10	0	0.00%	116,863	69.22%	113,815	69.47%

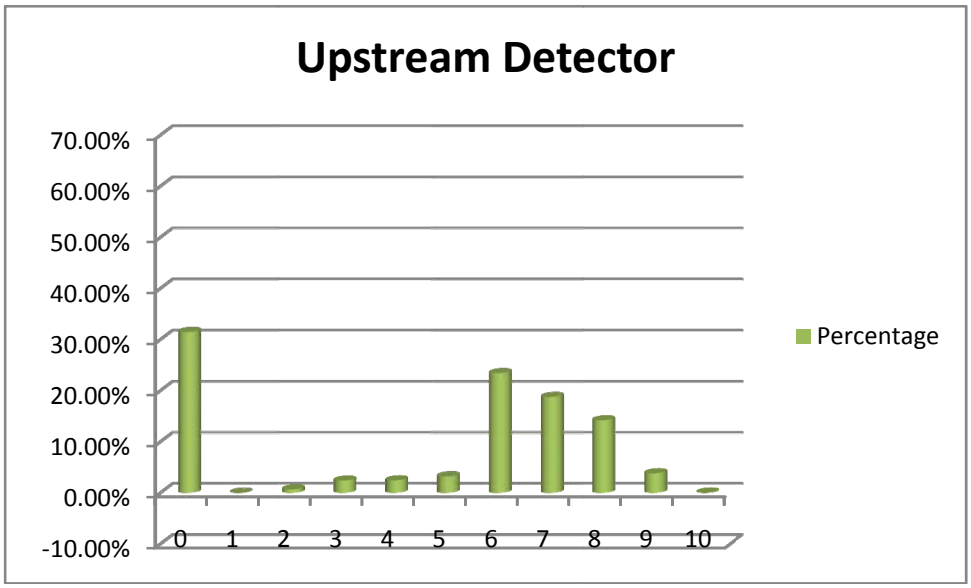


Figure 4.11: Upstream Detector Reliability Index Distribution

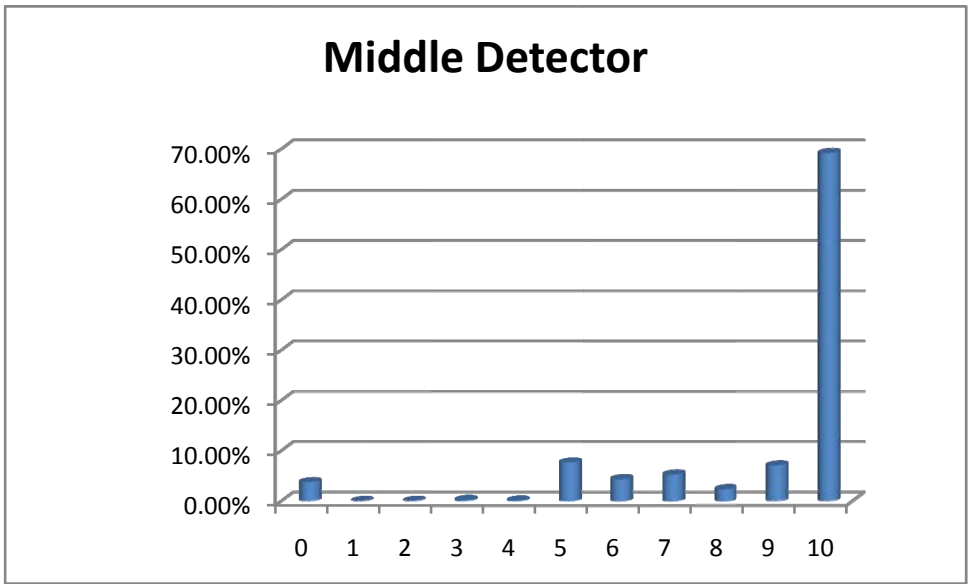


Figure 4.12: Middle Detector Reliability Index Distribution

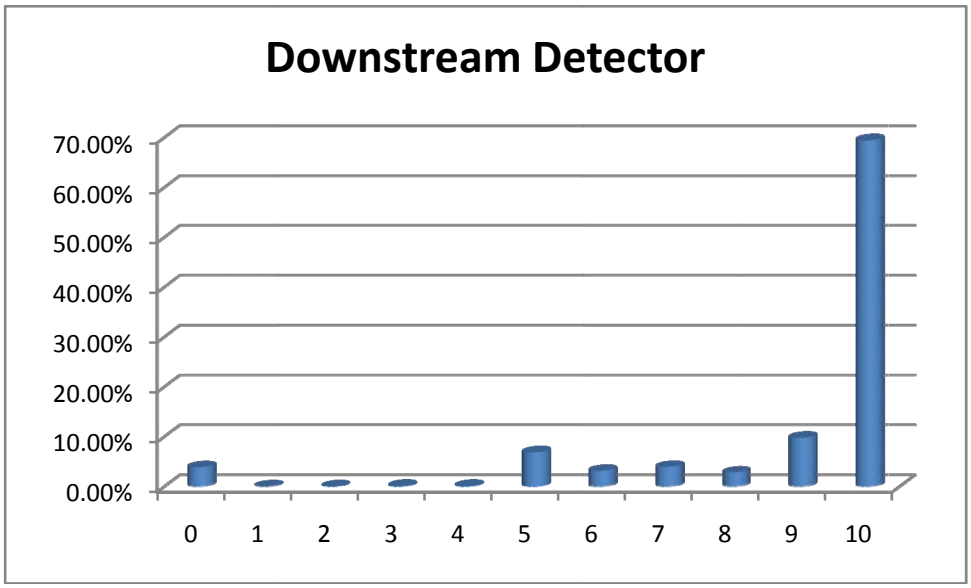


Figure 4.13: Downstream Detector Reliability Index Distribution

It should be noted that most of the low reliability index records ($RI \leq 1$) come from the records corresponding to the detector status of "NO data." Only nine records with $RI \leq 1$ come from records with status "Normal." We hence interpolate records with $RI \leq 1$ and with status "Normal" in later section.

As can be observed from the table, the upstream detector has the lowest percentage of high reliability index records since it does not have upstream data available for detailed calculation. For the middle and downstream detectors, 69.22% and 69.47% of the records are highly reliable based on the proposed CST.

4.3 Comparison of Imputation Methods

The amount of missing ITS data is often significant. Using archived data from San Antonio, Texas, Turner et al. (2000) identified almost a quarter of data records as either missing or "suspicious." This figure is not unusual for such systems: Nguyen and Scherer (2003) note that 25-30% of the detectors operated by the Virginia Department of Transportation are offline at any given time, and Kwon (2004) suggests that even functioning detectors often fail to report up to 5% of data. Clearly, missing data are problematic for any of the functions for which ITS data is to be used. Depending on the application, it may be desirable to replace missing or suspicious data with estimates of the true value, and several imputation techniques have been developed to accomplish this, as described in the next section.

This section provides a comparison of data imputation methods, using loop detector data from the Dallas, Texas metropolitan area to conduct an experiment regarding the accuracy of these procedures. Other factors affecting the suitability of these methods are also considered, such as the amount of data required, or the robustness to the specific types of events that commonly result in missing data.

4.3.1 Data Imputation Methods

The problem of missing data is well-known among transportation researchers and practitioners, and multiple approaches for estimating missing observations have been devised.

The simplest methods involve linear regression, using nearby (spatially and temporally) observed data to estimate the missing value, calibrated using a past corpus of data. Linear regression can be implemented in a number of ways. Chen et al. (2003) and Nguyen and Scherer (2003) recommend estimating a linear regression model based on neighboring detectors. Al-Deek and Chandra (2004) suggest estimating a set of linear regression models, each relating data at the missing detector to data at a nearby detector, and taking the median of all estimates generated in this way. The advantages of linear regression models are their simplicity, ease of computation, and ease of interpretation; however, they perform poorly when neighboring data is missing as well. Al-Deek and Chandra's method is more robust to missing data from individual neighboring detectors, but still cannot be used when all neighboring data is missing (for instance, when a power failure affects all detectors in a particular area).

Kwon (2004) suggests combining linear regression with non-normal Bayesian imputation, terming the procedure "nonnormal Bayesian linear regression" (NBLR). NBLR involves estimating a linear regression model as described above, along with the deviation between each past observation and the estimate the linear regression would have predicted. Missing data are then imputed by performing the linear regression and then applying a deviation sampled from the past set.

Other methods only use data from the missing detector to perform the imputation, thus avoiding the requirement that neighboring data also be available. Nguyen and Scherer (2003) mention that historical averages can be used to replace missing data, and Gold et al. (2000) describe a "factoring-up" approach, in which missing data are effectively set to the average of observed data at nearby time periods (for example, if two out of twelve data readings are missing from a one-hour block, the two missing data are set to the average of the ten that were observed). These approaches can be more reliable, in that they do not rely on neighboring data; however, but not considering other locations, they are less able to represent current conditions if they differ from historical norms.

Time-series approaches can also be applied, as described in Nguyen and Scherer (2003), in which spatiotemporal autocorrelations are calculated using observed data, and then used to impute the missing values. This approach has the advantage of building on a well-developed body of literature in this field (see, for instance, Hamilton, 1994).

Additionally, the CST-based reliability index developed in Section 4.2 can also be applied to impute missing data, by applying a search procedure to identify missing data values that maximize this reliability score. This approach has the advantage of considering multiple aspects of data reliability (consistency with historical measurements, with nearby observations, and with basic traffic flow theory).

While missing data are easy to detect, other researchers have considered the problem of identifying data that are present, but probably incorrect, and thus subject to imputation as well. This is commonly done by comparison with either historical data, or fundamental physical relationships. For instance, Payne et al. (1976) flag data with physically impossible values for volume, speed, and density, and Chen and May (1987) look for occupancy values far from the historical norms. Turner et al. (2000) and Chen et al. (2003) look at combinations of data that are impossible, such as zero volume and positive occupancy. More sophisticated methods have also been described, such as defining an acceptable set of volume/density values (Nihan et al., 1990),

data storage rates (Nihan et al., 2002), or statistical entropy (Al-Deek and Chandra, 2004). Observations from nearby detectors can also be used to mark suspicious data; such approaches can be found in Coifman (1999) and Vanajakshi and Rilett (2004).

4.3.2 Experimental Setup

This section describes the procedure used to compare the accuracy of imputation methods, starting with discussion of the data set used, and then presenting the methods used, along with specific details of implementation.

Data Set

To compare these methods, a sample of loop detector data was obtained from the Dallas Traffic Management System (DalTrans), operated by the Texas Department of Transportation, using a public website allowing downloading of archived data (http://ttidallas.tamu.edu/detectordataarchive/). A detector located on the I-20 freeway was chosen for a testbed location (Figure 4.14), recording data on the third travel lane in the westbound direction. The archive contains occupancy, speed, and volume (large vehicle and total) data, aggregated at the five-minute level. All available data from this detector between September 14, 2007 and December 31, 2007 were downloaded, a total of 24,577 observations.

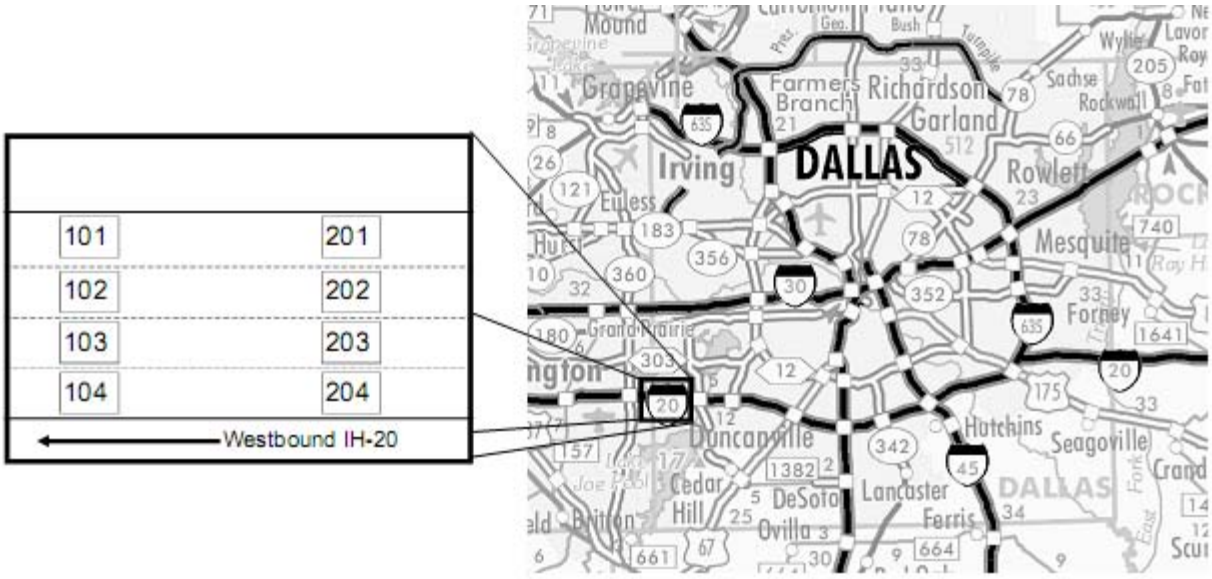


Figure 4.14: Detector locations (Map source: Texas Department of Transportation)

Ten percent of the sample (2457 observations) were randomly selected to serve as "missing" data, allowing each of the methods to be applied to impute a value, which could then be compared with the data actually observed. The remaining ninety percent of the sample was used to calibrate these models; that is, none of the "missing" observations were used to estimate any of the regression models or calculate any historical statistics. For the purposes of this comparison, only volume data were imputed and compared, although speed and occupancy data were used in some of the imputation models.

Methods Applied

Based on the past literature, eleven methods were applied to impute "missing" data.

Simple Linear Regression

Of the linear regression methods, the one described by Al-Deek and Chandra (2004) was adapted for use, because it is most robust to missing observations. Seven simple linear regression models were estimated, each relating the volume at detector 103 to the volume at one other nearby detector, as shown in Figure 4.14. That is, seven volume estimates $\hat{v}_1^{103}, \hat{v}_2^{103}, \ldots, \hat{v}_7^{103}$ are made using the equations

$$\hat{v}_1^{103} = \beta_0^{101} + \beta_1^{101} v^{101}$$
$$\hat{v}_2^{103} = \beta_0^{102} + \beta_1^{102} v^{102}$$
$$\hat{v}_3^{103} = \beta_0^{104} + \beta_1^{104} v^{104}$$
$$\hat{v}_4^{103} = \beta_0^{201} + \beta_1^{201} v^{201}$$
$$\hat{v}_5^{103} = \beta_0^{202} + \beta_1^{202} v^{202}$$
$$\hat{v}_6^{103} = \beta_0^{203} + \beta_1^{203} v^{203}$$
$$\hat{v}_7^{103} = \beta_0^{204} + \beta_1^{204} v^{204}$$

where v^i is the volume recorded at detector i, and β_0^i and β_1^i are estimated parameters for each detector.

We test two methods of generating an imputed value from these estimates: using either the average (SLR-AVG) or the median (SLR-MED). Note that if one or more of these estimates cannot be used because of missing data from another detector, the average or median is calculated using the remaining values.

Multiple Linear Regression

Al-Deek and Chandra (2004) also suggest making use of speed and occupancy data, as well as possible quadratic relationships among explanatory variables, when using linear regressions. As before, up to seven volume estimates are made, using the regression equations

$$\hat{v}_i^{103} = \beta_0^i + \beta_1^i v^i + \beta_2^i o^i + \beta_3^i s^i + \beta_4^i (v^i)^2 + \beta_5^i (o^i)^2 + \beta_6^i (s^i)^2 + \beta_7^i v^i o^i + \beta_8^i v^i s^i + \beta_9^i o^i s^i$$

for $i \in \{101,102,104,201,202,203,204\}$, where s^i and o^i respectively denote the speed and occupancy estimates. Again, either the average (MLR-AVG) or the median (MLR-MED) of these estimates can be used to produce the imputed value.

Local and Global Regression

Local and global linear regression models attempt to incorporate a greater geographic scope into the imputation process. Instead of using data from an adjacent location (or the same location), one can use volume readings from multiple locations in order to make the prediction.

Local linear regression uses locations in the vicinity of the detector missing data, such as ten locations immediately upstream. Global linear regression uses locations drawn from throughout the region.

Although local linear regression is expected to be more accurate, it is vulnerable to events that cause loss of data in an entire region, such as a localized power outage. Global regression models may help overcome this problem, by drawing on data from multiple sources. (It is important to note that global regression models are generally more subject to problems of missing data, because they rely on successful operation and communication between many different regions of detectors, rather than just one or two. However, they do help address the specific problem of power failures localized at the missing detector.)

Only one detector is chosen from each location because volume readings recorded from detectors at the same location are very highly correlated. Although correlation also exists between detectors at adjacent locations, the presence of on- or off-ramps between them mitigates this to some extent.

In this experiment, the detectors chosen for local and global regression can be seen in Figure 4.15. In each case, the detector in the third lane was chosen for making the estimates

$$\hat{v}^{103} = \beta_0 + \sum_{i=1}^{10} \beta_i v^i$$

where i indexes the detectors used (either local or global).

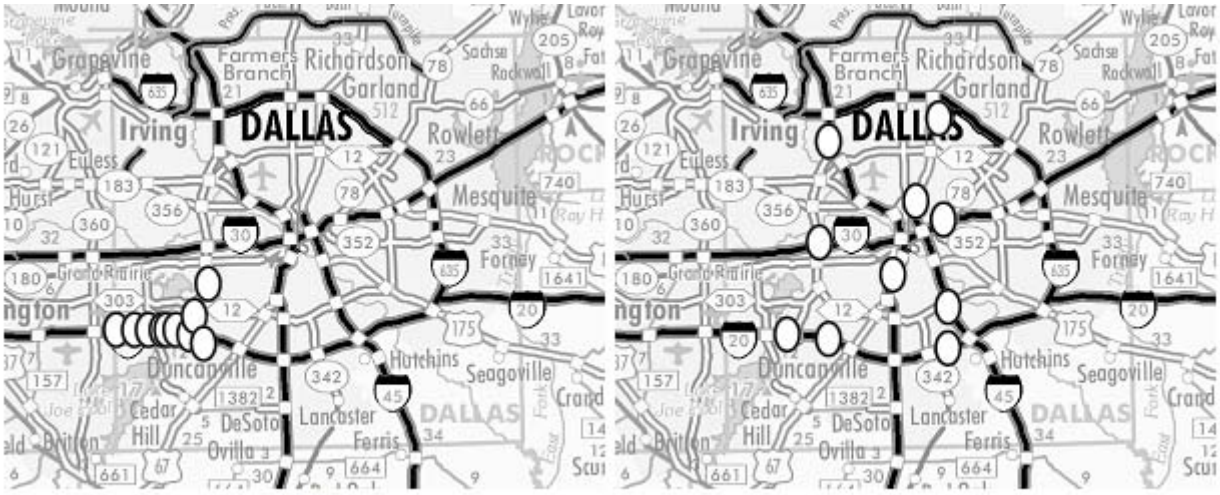


Figure 4.15: Detector locations for (a) local regression and (b) global regression

Nonnormal Bayesian Linear Regression

The NBLR technique suggested by Kwon (2004) can be implemented by modifying one of the above-mentioned regression techniques. After estimating a regression model using observations $1,2,\ldots,n$, one then calculates the deviations $d_i = v_i^{103} - \hat{v}_i^{103}$, where v_i^{103} and $\hat{v}_i^{103}$ respectively denote the actual volume for the *i*-th observation, and the value the regression model would have predicted.

Then, for the missing data, the regression model is applied, and the result is added to a deviation term randomly drawn from $d_1, \ldots, d_n$. The intent of this procedure is to more accurately reflect the variability existing in the data, and to remove any statistical bias from the regression equation.

Historical Imputation

All of the regression models described thus far rely on data from other locations in order to make a prediction. Historical imputation models, on the other hand, impute data using data measured only at the location where needed. While this has the advantage of robustness (this procedure can almost always be used), it carries the severe disadvantage of implicitly assuming typical operating conditions, and cannot use any spatial information to determine if conditions vary from past norms. However, ex post factoring approaches, such as that suggested by Gold et al. (2000), do allow limited temporal inferences to be made in this respect.

As implemented in this experiment, the readings from each detector are classified by time-of-day into 288 categories (the number of five-minute intervals in a day), generating a corpus of data that is used for historical imputation. Each missing observation is then imputed as either the average (HIST-AVG) or median (HIST-MED) of past observations from the same time-of-day; this follows the procedure given in Nguyen and Scherer (2003). Given additional data, further segmentation on day-of-week would be desirable.

A factoring approach similar to that in Gold et al. (2000) is also tested, in which each missing volume datum is set to the average of all present volume data from the same detector within the last hour; this approach is referred to as FACTOR in the following tests.

CST Imputation

Section 4.2 describes a CST-based system for quantifying confidence in ITS data on a continuous scale (0-10). This assessment is based on three criteria: fundamental consistency (do the data respect physical relationships, such as jam density or the requirement that volume be the product of speed and density), network consistency (are the data consistent with recent upstream measurements), and historical consistency (are the data reasonable given past observations at this location), which are then combined to generate an aggregate "reliability index."

Although developed to assess archived data and identify suspicious readings, this evaluation procedure can be used to assist with imputing missing data by treating it as a "black box" subroutine of a search procedure. Let $F(v^{103}, s^{103}, o^{103})$ denote the reliability index as a function of the volume, speed, and occupancy; the task is then to find values of v^{103}, s^{103}, and o^{103} which maximize F. F is not provably concave; however, any metaheuristic search (such as genetic algorithms, simulated annealing, or tabu search) may be used. For simplicity, we use the following local search algorithm to find a local maximum:

1. Generate initial values for $D = (v^{103}, s^{103}, o^{103})$, and calculate $F(D)$
2. Perturb these values by testing $D' = D + (\Delta v, 0, 0)$, $D - (\Delta v, 0, 0)$, $D + (0, \Delta s, 0)$, $D - (0, \Delta s, 0)$, and $D + (0, 0, \Delta o)$, $D - (0, 0, \Delta o)$ in turn.
3. If $F(D') > F(D)$ for any of these, set $D = D'$ and return to Step 2. Otherwise, terminate.

with pre-defined step sizes Δv, Δs, and Δo. In this implementation, the initial value is chosen by establishing s^{103} at its historical median, v^{103} so as to be consistent with flow conservation from upstream detectors, and from calculating density as the quotient of v^{103} and s^{103}, and then dividing by an assumed average vehicle length of 20 ft (6 m) to calculate estimated occupancy. One advantage of this procedure is that it estimates the volume, speed, and occupancy simultaneously, considering the relationship between the three; the methods described above must calculate each of these separately, with no explicit guarantee of consistency.

4.3.3 Results

Using the procedures described above, eleven methods for imputing data were applied: SLR-MED, SLR-AVG, MLR-MED, MLR-AVG, LOCAL, GLOBAL, NBLR, HIST-MED, HIST-AVG, FACTOR, and CST. This subsection presents results from these tests grouped by type; Section 4.3.6 presents a broader perspective on all eleven.

Linear Regression-Based Models

Seven of the models are based in linear regression; results are summarized in Table 4.6. Several results are apparent. First, for both the simple and multiple linear regression models, the average of the estimates outperformed the median, as measured both by the regression correlation coefficient (r^2) and root-mean square error (RMSE). This is interesting, as the median is often used in data imputation for its sensitivity to outliers (Al-Deek and Chandra, 2004).

Second, using multiple data from the same detector, and considering quadratic relations, as in MLR, provides a better estimate than the simple linear regression. This is not surprising, since more explanatory variables always improve a linear regression model; however, as seen by the adjusted correlation coefficients ($\bar{r}^2$), the improvement is more than would be expected by chance.

Third, both the LOCAL and GLOBAL models (as well as NBLR, which is built on LOCAL) appear to suffer from data availability issues that SLR and MLR do not—LOCAL and NBLR were only able to impute 2150 of the 2457 missing data (88%), while GLOBAL was only able to impute 755 of them (31%). By contrast, SLR and MLR were always able to impute a missing value. While this does suggest that the increased data requirements of LOCAL and GLOBAL may be a hindrance, the experimental setup nevertheless favored SLR and MLR in this manner. Since none of the imputed data replaced truly missing values (rather, missing values were simulated by randomly deleting observations), data was always available from other detectors at the same location, and thus SLR and MLR were always able to make imputations.

Fourth, the GLOBAL model appears to be significantly less accurate than the LOCAL model, as expected. Finally, the addition of the random deviation terms in NBLR seems to degrade the average performance of the prediction, but does succeed in reducing the bias (average signed difference between observed and predicted values) by two-thirds.

Figures 4.16 through 4.20 graphically illustrate the goodness-of-fit for these models, by plotting the estimated five-minute volume (vertical axis) against the value actually observed (horizontal axis) for these models. The median-based regressions are not shown, due both to their slight underperformance, and to their substantial similarity to the average-based regressions.

Table 4.6: Results from linear regressions

	SLR-MED	SLR-AVG	MLR-MED	MLR-AVG	LOCAL	NBLR	GLOBAL
n	2457	2457	2457	2457	2150	2150	755
r^2	0.9450	0.9508	0.9578	0.9625	0.9574	0.9155	0.8392
$\bar{r}^2$	0.9450	0.9508	0.9576	0.9624	0.9572	0.9151	0.8370
RMSE	8.80	8.33	7.73	7.29	7.77	10.9	15.0
Bias	-0.28	$+0.08$	$+0.17$	$+0.06$	$+0.06$	$+0.02$	$+0.58$

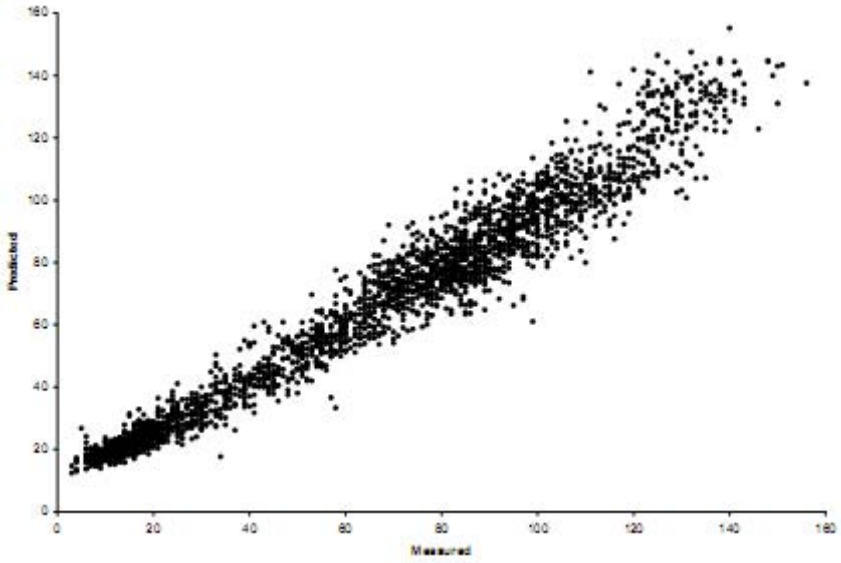


Figure 4.16: Plot of data estimated with SLR-AVG vs. observed data

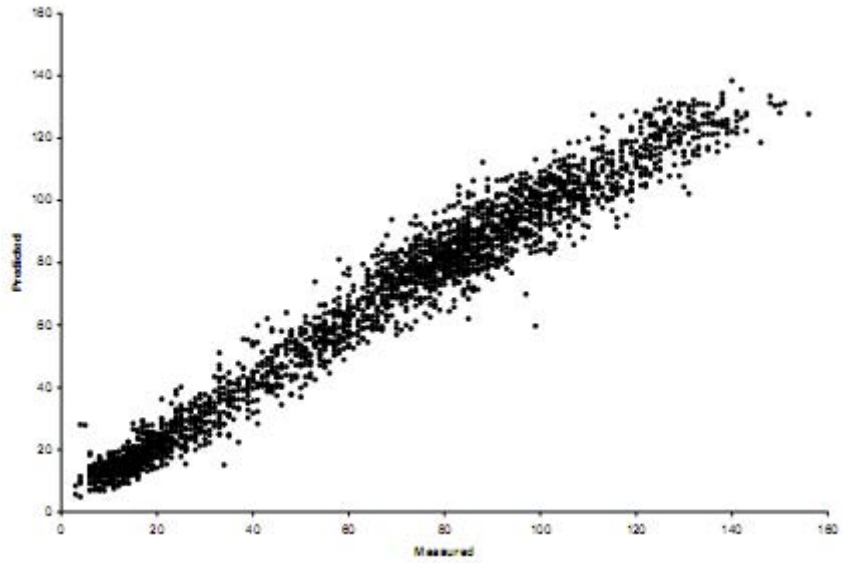


Figure 4.17: Plot of data estimated with MLR-AVG vs. observed data

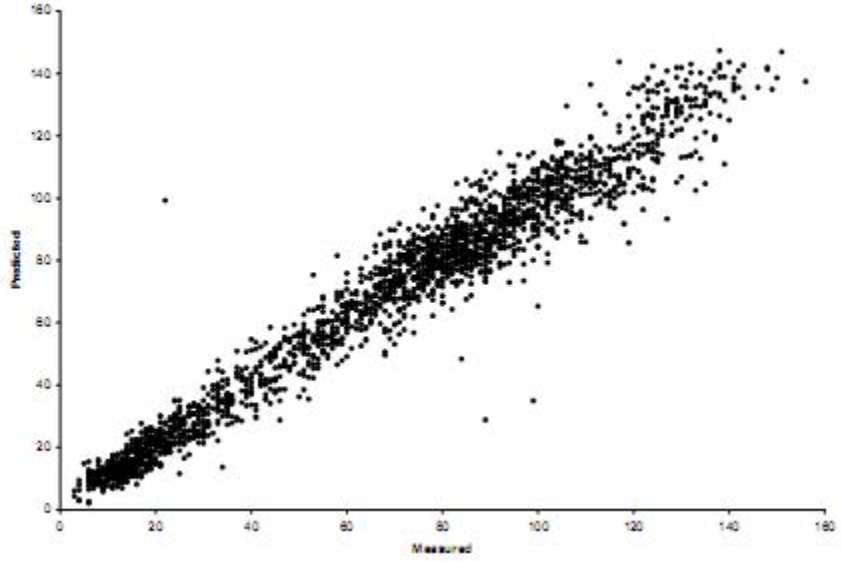


Figure 4.18: Plot of data estimated with LOCAL vs. observed data

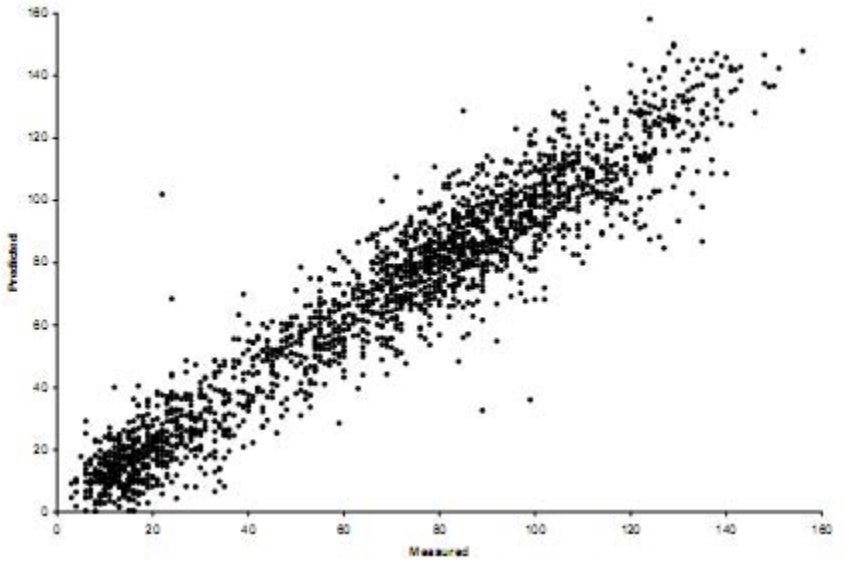


Figure 4.19: Plot of data estimated with NBLR vs. observed data

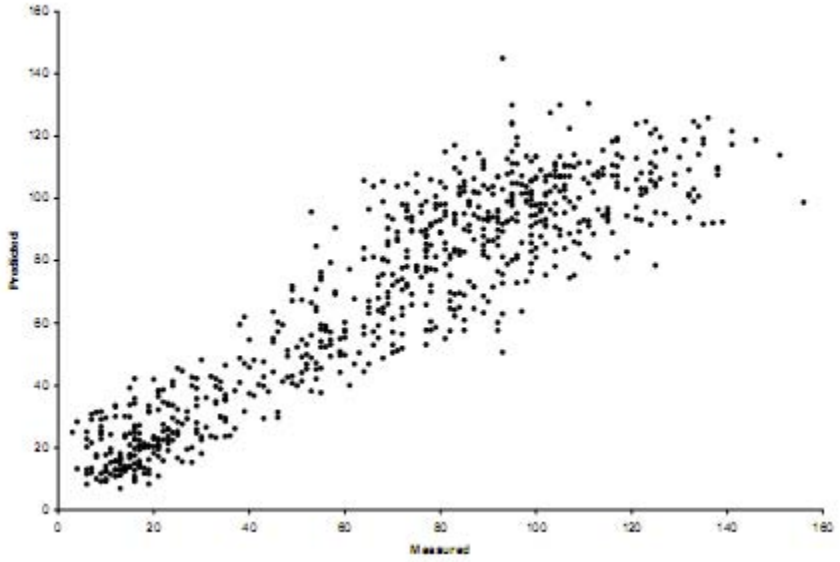


Figure 4.20: Plot of data estimated with GLOBAL vs. observed data

4.3.4 Historical Models

Table 4.7 and Figures 4.21 to 4.22 show the performance of the HIST-MED, HIST-AVG, and FACTOR models. Clearly, these models do not perform as well as the linear regression models, since they cannot use current data from other detectors to refine the prediction. Of the three, FACTOR performs the best since it can at least make use of data observed earlier at the current detector; however, the averaging process still leaves substantial error. Interestingly, HIST-MED and HIST-AVG both appear to have a substantial positive bias, overestimating the true volume by approximately twenty vehicles, while FACTOR does not suffer from this

problem. Also, using the average was more accurate than using the median when considering historical data, similar to the result found with the linear regressions.

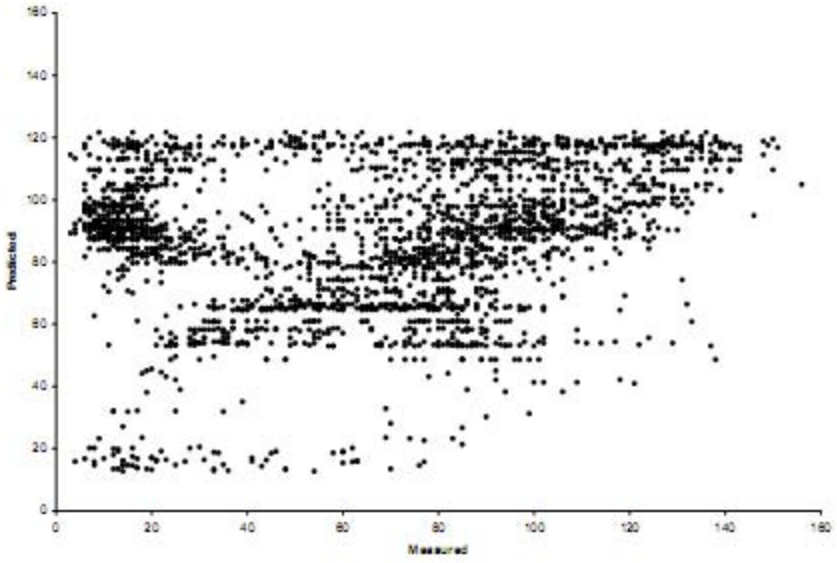


Figure 4.21: Plot of data estimated with HIST-AVG vs. observed data

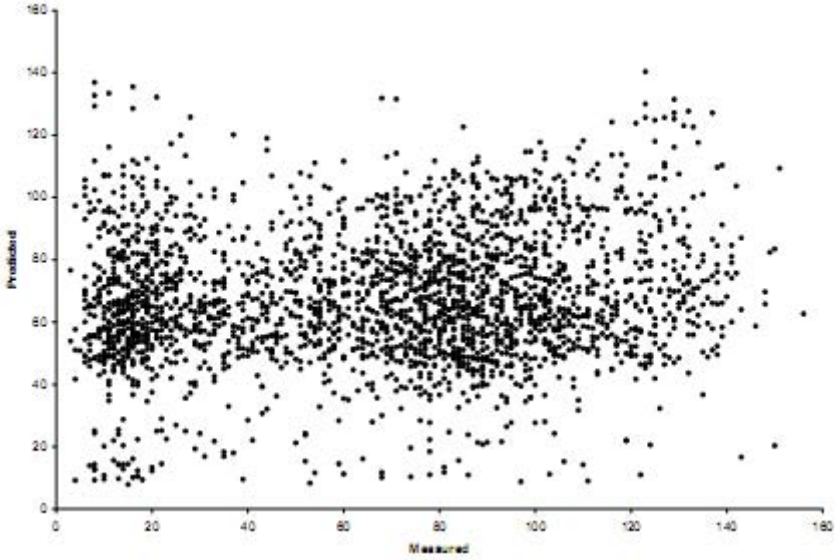


Figure 4.22: Plot of data estimated with FACTOR vs. observed data

Table 4.7: Results from other models

	HIST-MED	HIST-AVG	FACTOR	FUZZY
RMSE	45.15	44.21	40.37	35.01
Bias	+20.6	+20.3	+0.51	+2.02

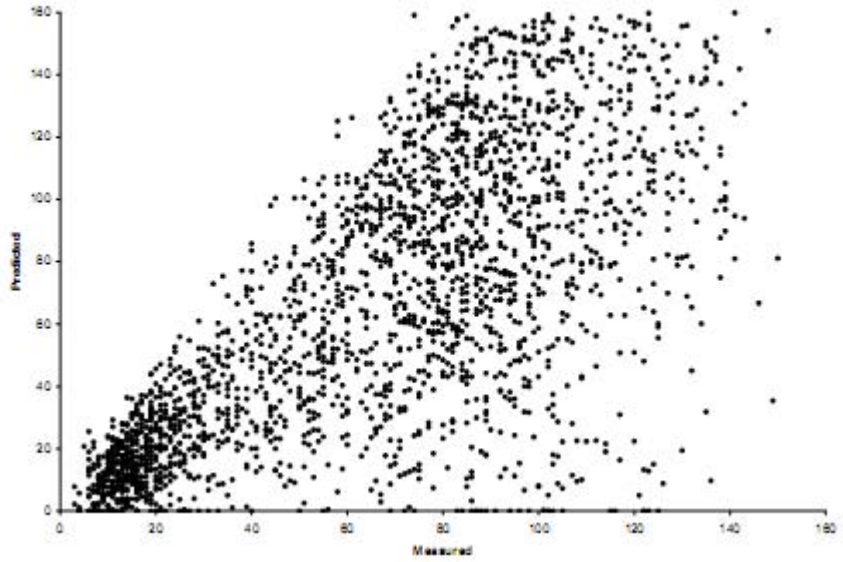


Figure 4.23: Plot of data estimated with CST vs. observed data

4.3.5 CST Model

Table 4.7 and Figure 4.23 show the performance of the CST imputation method. Although outperforming the historical models in terms of RMSE, it still falls significantly short of the performance of the linear regression models. Again, it should be noted that by only considering missing volume, the experimental setup did not favor one of CST's strengths, namely, its ability to simultaneously estimate consistent values of volume, speed, and occupancy.

4.3.6 Conclusion

To help improve the quality of traffic data, imputation algorithms have been developed to make estimates about missing data. This paper compared eleven such algorithms; eight of these existed in the past literature, and three (LOCAL, GLOBAL, and CST) were developed in the course of this research. Of these, the linear regression-based models were the most accurate in predicting the true values of randomly deleted observations. However, their data requirements are the most significant, and in practice it may be desirable to have a ``backup'' imputation model ready to use if all of the input data are not available. Conversely, the imputation methods based on historical observations have minimal data prerequisites, but are much less accurate. The performance of the CST imputation algorithm lie between these two, and more research is needed to fully determine its suitability for this application.

Although this paper gives guidance regarding such algorithms, there remains much future work in this area. First, more sophisticated regression models (such as nonparametric regression) or time series techniques can be compared alongside these eleven algorithms. Second, additional

data sets should be considered, to examine whether these findings can be generalized. Third, more realistic experiments replicating actual causes of missing data (such as power outages, or regular communication downtime for maintenance) can be conducted, providing a better view of how these algorithms might perform in practice.

4.4 Extrapolation by Kriging

While the imputation methods described in Section 4.3 are appropriate for estimating missing values of data at detector locations, they are less useful for estimating traffic data where *no* detectors are present. Kriging methods can be used to predict count values at unmeasured locations while also assessing the errors of these predictions. These methods rely on the notion of autocorrelation in error terms/unobserved factors over space, where the level of autocorrelation is a function of distance. Meanwhile, the values to be predicted may have their own predictive factors (e.g., number of lanes and facility type). These create a "trend" estimate, $\mu(s)$; so, in general, the spatial data can be expressed as follows:

$$Z(s)_i = \mu(s)_i + \varepsilon(s)_i, \tag{4.5}$$

where $Z(s)_i$ is the variable of interest (actual count here) and s gives location (x,y coordinates of site i. $Z(s)_i$ is composed of a deterministic trend $\mu(s)_i$ and a random error component $\varepsilon(s)$. These $\varepsilon(s)$ values are correlated over space. Features of "trend" (often called "drift" in other studies), or the expected value of $Z(s)$, divide Kriging methods into three categories: If $\mu(s)$ is constant across locations or unknown, one can rely on Ordinary Kriging. Trends that depend on explanatory variables and unknown regression coefficients must rely on Universal Kriging. If the trend is known, one has Simple Kriging. The "Geostatistical Analyst" tool and "Spatial Analyst" tool in ArcGIS can be used to fit and then apply these different Kriging methods.

Weak stationarity is assumed in all three of these methods, so that the correlation between $Z(s)$ and $Z(s + h)$ does not depend on actual locations, but only the distance h between the two sites. This is necessary to ensure replication. Furthermore, thanks to weak stationarity, the variance of $Z(s+h) - Z(s)$ equals $2\gamma(h)$ for any s and h, where $2\gamma(h)$ is used as the y-axis in a "variogram" and $\gamma(h)$ is used in a "semivariogram."

4.4.1 Universal Kriging

In Universal Kriging, $\mu(s)$ can be a deterministic function of any form. A simple assumption is to use a linear function where $\mu(s)=X\beta$, (where X contains explanatory variables like number of lanes and facility type). In contrast, ε(s) reflects unobserved variation (e.g., local land use patterns, presence of subway routes).

For purposes of prediction, Kriging is performed on the $Z(s)$ values. The sum of interpolated random component $\varepsilon(s)$ and the estimated $\mu(s)$ values together lead to the estimated $Z(s)$ values. The following section first introduces how the random component $\varepsilon(s)$ is estimated with Kriging.

4.4.2 Interpolating Random Components using Variograms

As noted above, weak stationarity ensures the following:

$$\gamma(h) = 1/2 \; var\,[Z(s+h) - Z(s)] \tag{4.6}$$

Where var $[Z(s+h) - Z(s)]$ is the variance (over all sites) between counts at sites s and $s + h$. The first step is to select an appropriate semivariogram model for a given dataset. There are several types of commonly used models such as exponential, spherical, and Gaussian models. The model specifications for these models are shown as equations (4.7)-(4.9) and plotted in Figure 4.24:

1. Exponential

$$\gamma(h) = \begin{cases} c_0 + c_1\left[1 - \exp\left(-\frac{h}{a}\right)\right] & \text{if } h > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

2. Spherical

$$\gamma(h) = \begin{cases} c_0 + c_1\left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right] & \text{if } 0 < h < a \\ c_0 + c_1 & \text{if } h > a \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

3. Gaussian

$$\gamma(h) = \begin{cases} c_0 + c_1\left[1 - \exp\left(-\frac{h}{a}\right)^2\right] & \text{if } h > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

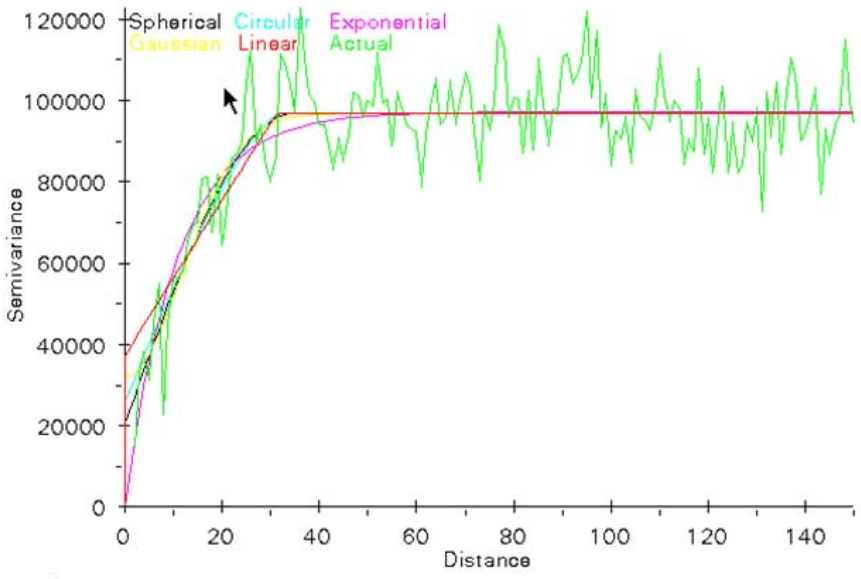


Figure 4.24: Several semivariance model specifications

These models all rely on three parameters that shape their functions and indicate spatial dependency. c_0 is called the "nugget effect." It reflects the discontinuity at the origin of the variogram caused by factors such as sampling error and short scale variability. (In theory the value of the variogram $\gamma(h)$ for $h = 0$ should be zero.) Here, a is called the "range." This scale

factor determines the threshold distance at which $\gamma(h)$ stabilizes flattens. c_0+c_1 is the maximum $\gamma(h)$ value, called the "sill," with c_1 referred to as the "partial sill" (Cressie, 1993). Figure 4.25 illustrates these parameters:

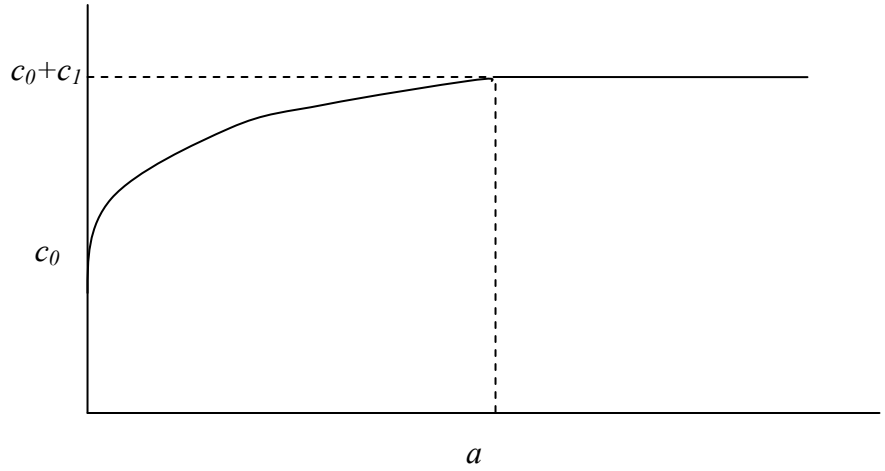


Figure 4.25: Illustration of Semivariogram

4.4.3 Estimation of Parameters

It is simpler to estimate the three shape parameters using Ordinary Kriging, when all $Z(s)$ values enjoy the same mean. In Universal Kriging, the vector of parameters β needs to be estimated (along with c_0, c_1 and a) because

$$\begin{aligned} &E[(Z(s_i)-Z(s_j))^2] \\ &= \text{var}\,[(Z(s_i)-Z(s_j)] + (\mu(s_i)-\mu(s_i))^2 \\ &= 2\gamma(s_i-s_j) + \sum k\beta_k(xk(s_i)-xk(s_j)))^2 \end{aligned} \qquad (4.10)$$

One approach is to use a serious of feasible general least square models (GLS) to estimate β and $\sum$ iteratively, where $\sum$ indicates the covariance matrix of error terms (ε) and is a function of a, c_0 and c_1.

Step 1. Obtain a starting value of β (e.g., $\beta=0$).

Step 2. Compute residuals $e=Z-X\beta$ (where X is the matrix of explanatory information across all sites).

Step 3. Estimate the variogram from residuals to get an estimate of $\sum$.

Step 4. Update estimate of β based on GLS: $\beta= (X'\sum^{-1} X)^{-1}X\sum^{-1}Z$.

Step 5. Repeat the above steps until parameter estimates converge.

Another approach to Universal Kriging is to use restricted maximum likelihood estimation (REML) by assuming the errors follow a normal distribution, so that dataset's log-likelihood enjoys the following proportionality:

$$LL \propto -|\sum|^{-1/2}[(Z-X\beta)'\ \sum^{-1}(Z-X\beta)] \qquad (4.11)$$

The likelihood can be maximized with respect to all unknown parameters using forms of Newton-Raphson or other optimization techniques.

In traditional kriging methods, the distance *h* refers to the Euclidean distance between the estimated and observed locations. Therefore, traditional kriging can only deal with continuous space with consistent characteristics. In many cases, however, non-Euclidean distances are more reasonable. For example, the transport of smog is blocked by hills and mountains; animals migrate around lakes, mountains, and settlements; and vehicles travel on road networks.

One approach to calculate such network distances based on vectors. Another, more convenient way is to use the "cost weighted distance," a geographic information system (GIS) raster function that calculates the cost of travel from one cell to another, subject to constraints.

However, no matter which approach is used, the calculation of network distance is not trivial. It is substantially more computationally intensive than simply using spatial coordinates to calculate the Euclidean distance. Thus, unless the region of interest involves a small-scale network with a few points/nodes (at most a few hundred), using Euclidean distance is preferable. If the network distance is to be used, it can be achieved by modifying the open-source code written by Lafleur (1998). after network distances between all points have been calculated. If the Euclidean distance is to be used, ArcGIS's toolbox can be applied.

Texas's SPTC or saturation count data have been used to forecast future year counts (temporal extrapolation) and between existing SPTC sites (spatial interpolation). Every year in Texas, close to 28,000 sites are monitored for 24 hours to obtain a day's traffic count. For the period 1999 through 2005, 27,363 of these sites have records count for all seven years. Table 4.8 provides some descriptive statistics of these traffic counts over different years.

Table 4.8: 24-HourTraffic Counts

Year	Number of Sites	Minimum	Maximum *	Mean	Standard Deviation
1998	27616	0	98040	5484.5	9483
1999	27663	0	99000	5762.5	10384
2000	27750	0	99000	5966.9	10656
2001	27905	0	99000	6181.5	11012
2002	27963	0	99000	6260.0	11036
2003	27921	0	99000	6389.6	11224
2004	27944	0	99000	6505.4	11404
2005	27910	0	99830	6676.1	11536
Over all years	27363	0	99000	6153.3	0

Note: The 99000 values shown here may not be real numbers, but they are consistent over years and cannot be distinguished from other numbers such as 99830, so they have been kept here, which may cause data quality problems.

The change in traffic counts over these seven years presents an approximately linear pattern. The magnitude or the slope of the change, however, varies significantly across different sites. In order to reasonably extrapolate future traffic counts, each site was analyzed using ordinary least square (OLS) regression based on all seven years' traffic counts. Assuming counts change linearly with time, two parameters can describe the equation for the traffic counts at each site. These are the slope and the average count (which together can also provide an intercept

term). Figure 4.26 is a histogram of all slope parameters for Texas' 27,363 SPTC sites. Due to the long right-side tail, the horizontal axis uses a logarithmic scale, and traffic counts at most sites increase by around 100 each year. For the 6799 sites experiencing volume reductions, the most common reduction in daily count per year is 20 vehicles.

Figure 4.27 shows the distribution of the traffic count averages (over 7 years) also with a logarithmically-scaled horizontal axis. As shown, most traffic sites average around 6,000 vehicles/day. To give a sense of relative change in counts, Figure 4.28 is a histogram of slope-to-mean count values.

After the mean count value and the slope are calculated, traffic counts at these saturation sites in future years can be extrapolated, assuming that the changes in traffic volumes follow a linear pattern. As an example, Figure 4.29 provides year 2006 for the predicted values for all 27,363 SPTC sites.

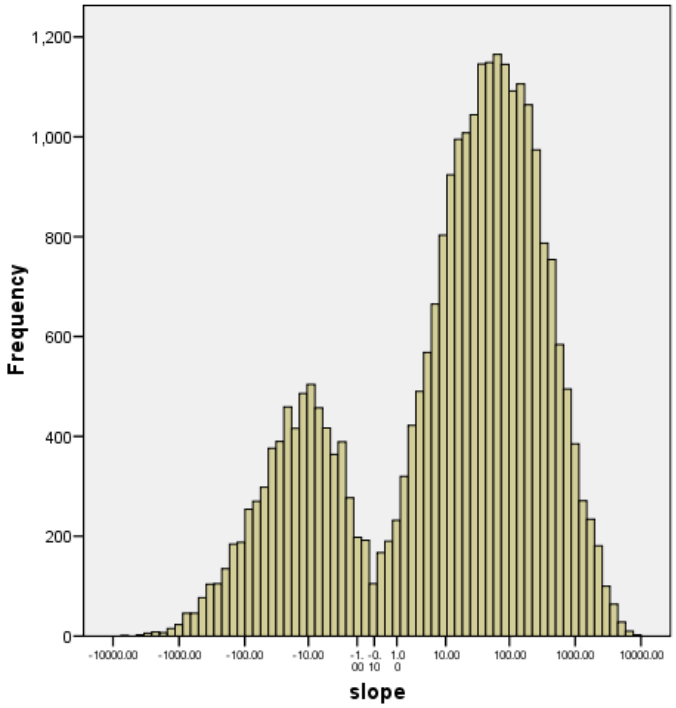


Figure 4.26: Distribution of Slope Parameters for 24 Hour Counts across SPTC Sites

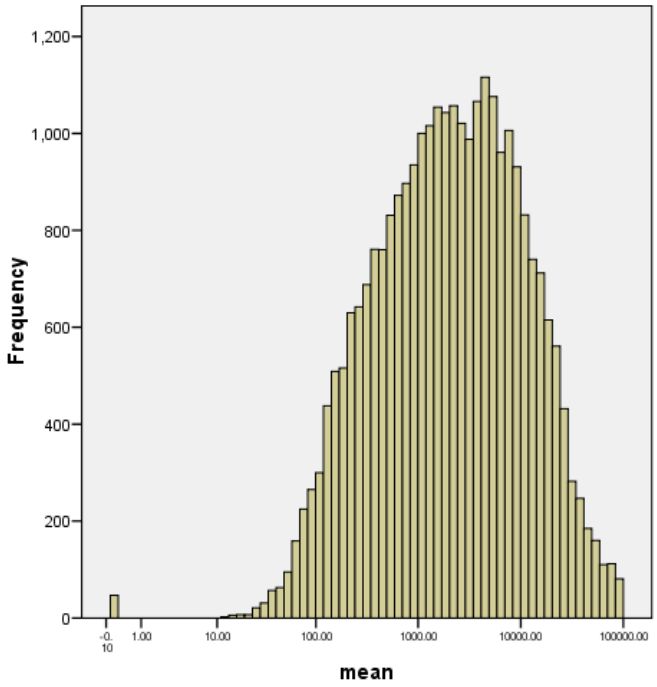


Figure 4.27: Distribution of Mean 24 Hour Traffic Counts across SPTC Sites

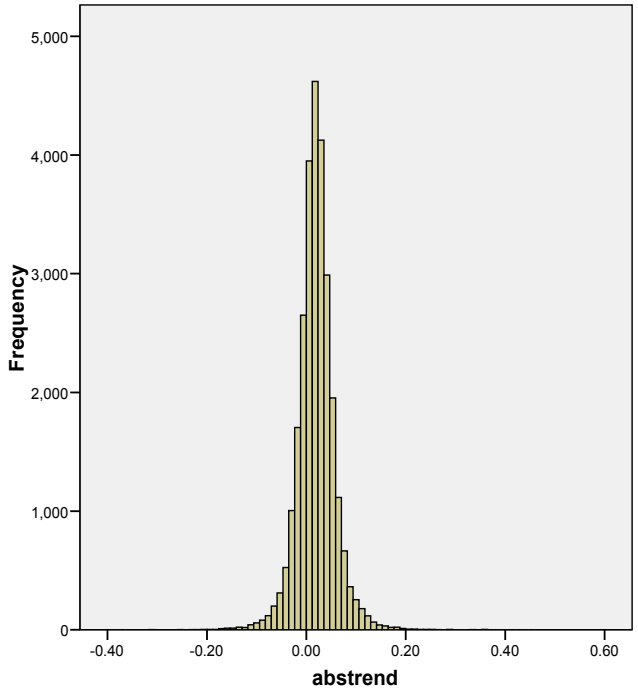


Figure 4.28: Distribution of Slope-to-Mean Count Values (Relative Change)

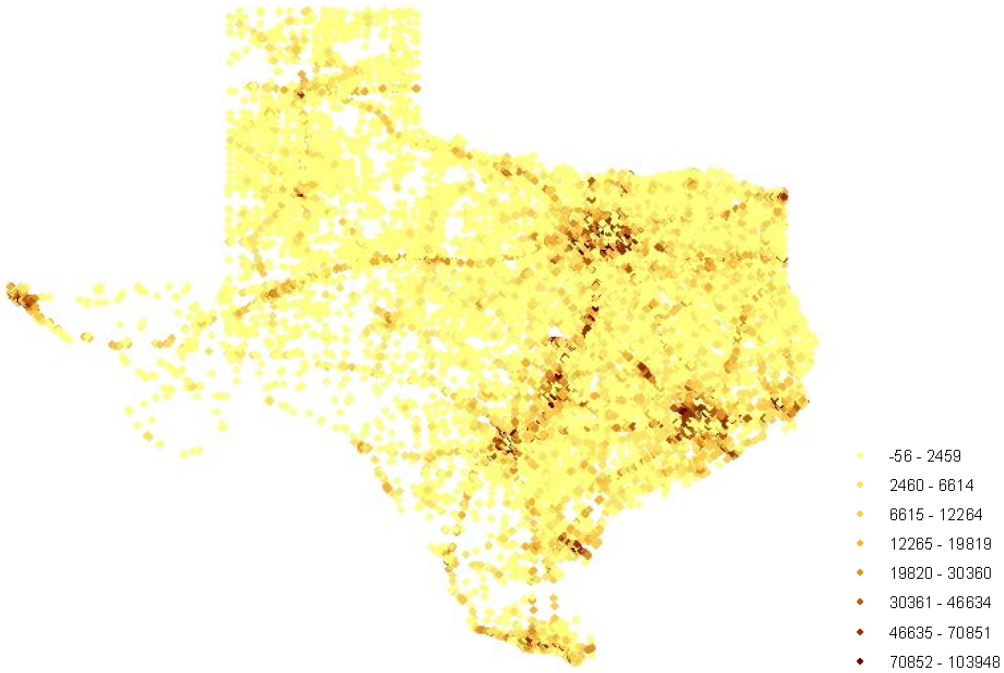


Figure 4.29: Predicted Counts for all SPTC Sites in 2006

4.4.4 Spatial Interpolation of Count Data via Kriging Analysis

Spatial Interpolation of AADT Data

Ideally, traffic counts at locations between SPTC and ATR sites should be estimated using universal Kriging based on considering the network distances between locations of interest. Due to data availability and computational time requirements, such an approach is not feasible with these Texas data. For example, it can be expected that the traffic counts at each site depend on variables like the number of lanes, speed limit, functional class, and area type. Unfortunately, such information was not availably linked to the traffic count dataset. The research team did receive a location file for all SPTC sites. Table 4.9 summarizes traffic count variables by districts changing patterns for sites in different districts.

Table 4.9: Patterns of Change for SPTC Values by Districts

District	Number of SPTC Sites	Change in 24 Hour Traffic Count/Year (Slope)	Average 24 Hour Traffic Count	Slope / Mean, (%)
Abilene	1075	34	2800	1.22%
Amarillo	1141	52	3754	1.38%
Atlanta	1209	83	4530	1.84%
Austin	1258	210	9636	2.18%
Beaumont	952	104	7159	1.45%
Brownwood	877	18	2307	0.77%
Bryan	1192	176	5568	3.17%
Childress	713	17	1278	1.36%
Corpus Christi	1181	197	5933	3.33%
Dallas	1433	321	10795	2.98%
El Paso	386	183	9301	1.97%
Fort Worth	1278	247	9764	2.53%
Houston	1246	354	12977	2.73%
Laredo	506	59	3522	1.69%
Lubbock	1661	39	2565	1.52%
Lufkin	1336	36	3724	0.96%
Odessa	728	55	3815	1.43%
Paris	1405	62	3466	1.78%
Pharr	1019	309	9229	3.35%
San Angelo	686	38	2255	1.67%
San Antonio	1425	266	8919	2.99%
Tyler	1697	81	4970	1.63%
Waco	1392	216	6551	3.29%
Wichita Falls	962	75	3438	2.18%
Yoakum	1511	68	3873	1.75%

To make better use of location information, this study used ESRI's ArcGIS ® software to obtain some road information from street map (provided by TxDOT's Michael Chamberlain) by joining road links (vector layer) to the point layer of SPTC sites. Each site obtained attributes of the road section that is closest to it. However, only road "class" is available in the street map file. Some basic visual checks suggest that the "class" variable indicates roads' functional class, but because the documentation for street map is not available, the precise definition of the variable "class" could not be determined. Table 4.10 provides count information by class. Over 85% of sites are coded Type 1, so this class variable would not prove so useful in a universal Kriging approach.

Table 4.10: Traffic Counts Changing Patterns for Sites at Different Classes

Class	Number of SPTC Sites	Change in 24 Hour Traffic Count/Year (Slope)	Average 24 Hour Traffic Count	Slope / Mean, (%)
1	24031	80	3539	2.26
2	3746	515	19986	2.58
3	74	336	17252	1.95
4	11	74	25660	0.29
5	5	147	13075	1.12
6	18	311	13247	2.35
7	8	745	18271	4.08
8	83	37	2504	1.48
9	219	58	8598	0.68
10	13	184	4032	4.57
11	1	299	19691	1.52
13	24	97	7718	1.25
19	1	585	13456	4.35

However, one can estimate traffic counts for each class separately. Hence, traffic counts on segments of the same class were spatially interpolated using Kriging. While using network distances interpolating spatial dependencies between locations along a network is behaviorally most reasonable, it is far more computationally intensive at the outset (E.g., for Class 1 road segments, the distances would have to be calculated 24,031x24,031 times. Fortunately, several existing studies (e.g., Hoef et al., 2006; and Kruvoruchko and Gribov, 2004) on small networks imply that using Euclidean distances can yield satisfactorily results, even when the dependencies arise over network. To ensure computational tractability, the Kriging method used here relies on Euclidean distances. The following two examples show how Class 1 and Class 2 road segments' AADT are estimated. They both use exponential form for the semi-variogram (Equation (4.7)). For Class 1 segments, range (a) is 0.254, partial sill (c_1) is 3.772E7 and nugget value (c_0) is 1.743E7. For Class 2 segments, range (a) is 0.147, partial sill (c_1) is 2.382E8 and nugget (c_0) is 5.228E8. The fitted prediction lines are shown in Figures 4.30 and 4.31.

Based on the calibrated semivariograms, traffic counts on all segments of these two classes can be predicted, as in Figure 4.32. The color shades indicate AADT levels, with the darkest spots representing AADT estimates above 19,918 vehicles per day and the lightest spots representing AADT lower than 679. It should be noted that there is probably something of an "edge effect" at the Texas border, due to a lack of count data. In other words, count prediction for areas outside the net of data points is less reliable.

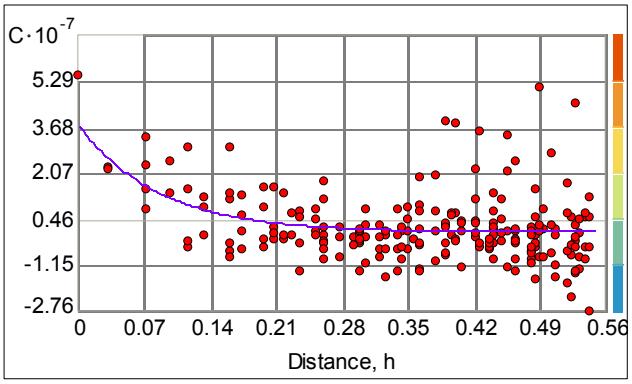


Figure 4.30: Semivariogram Fitting for AADT on Class 1 Segments

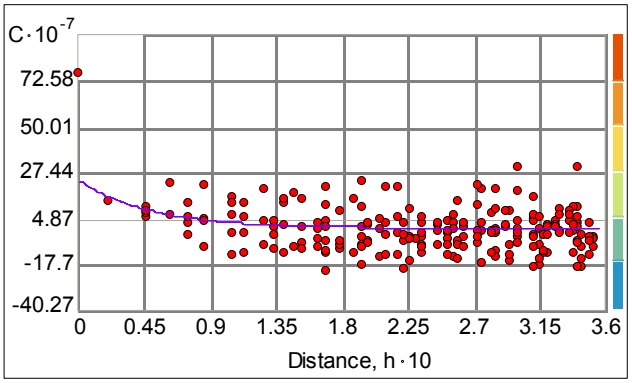
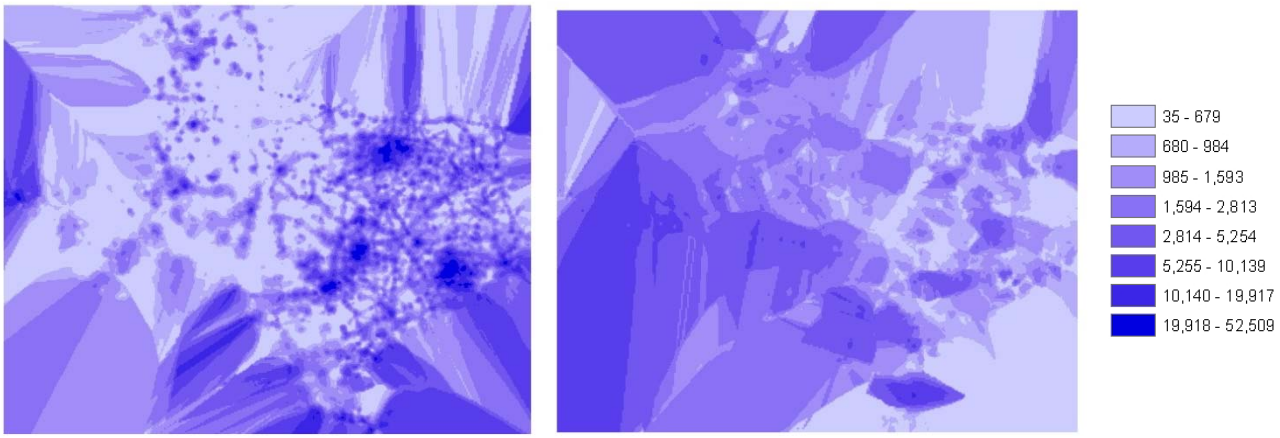


Figure 4.31: Semivariogram Fitting for AADT on Class 2 Segments



(a) Estimates of Class 1 Locations *(b) Estimates of Class 2 Locations*

Figure 4.32: Kriging-based Estimates of Traffic Counts for Year 2006 (Vehicles/day)

4.5 Assessing Goodness of Fit

In order to validate the Kriging method, the study used 80% of the Class 1 observations to interpolate AADT values and compare these estimates to the actual AADT values for the remaining 20% of observations. Differences in these values were evaluated using the error ratio indicator, where

$$Error_i = \frac{AADT_i - AADT_{est,i}}{AADT_i} \tag{4.12}$$

The spatial distribution of these error ratios are shown in Figure 4.33, and their histogram is shown in Figure 4.34. The mean of these errors is -0.16 (or -16%) with a standard deviation 0.68. Apparently, the spatial interpolated values tend to over-estimate AADT by about 16%, and in locations with very low or very high AADT values (e.g., the various 99,000 values, which may be unknown), the estimation does not perform very well.

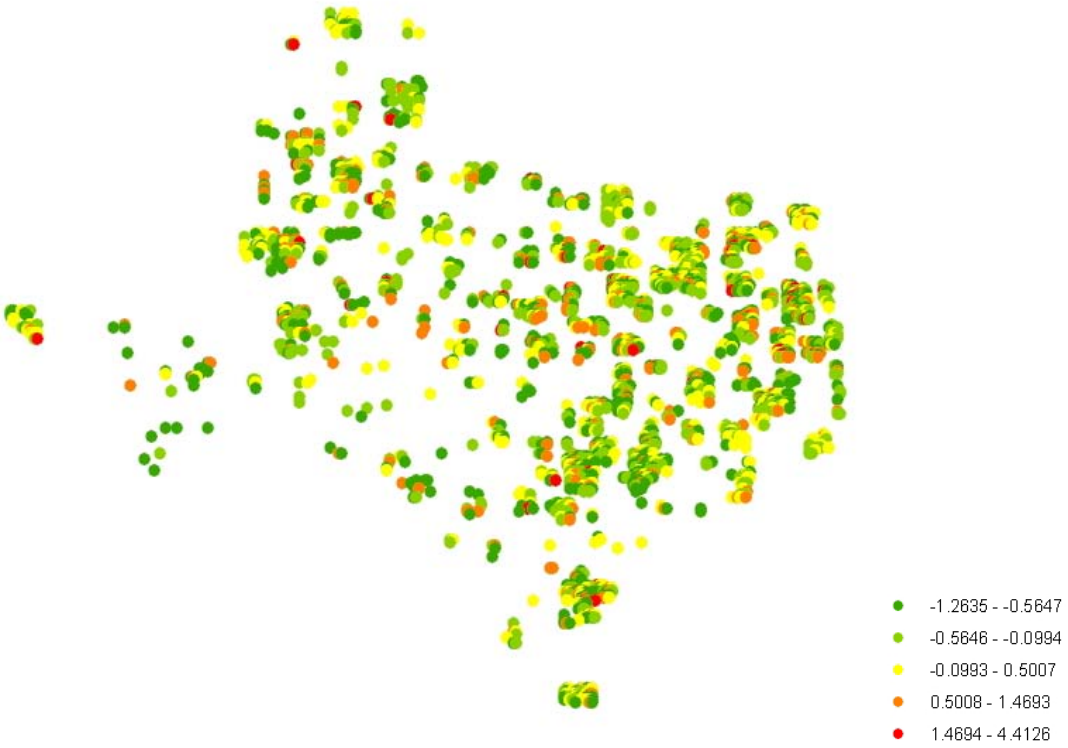


Figure 4.33: Differences between Kriging Estimates and Observed Traffic Counts

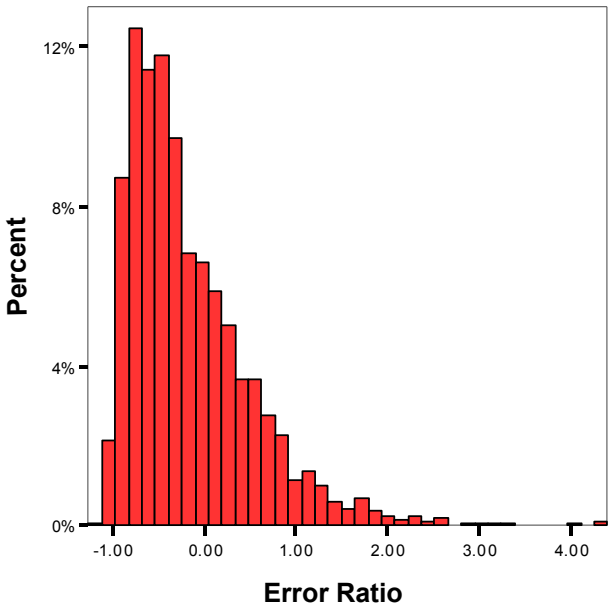


Figure 4.34: Histogram of Differences between Kriging Estimates and Observed Traffic Counts

4.5.2 Summary

This section describes a process that can be used to estimate real-time and longer-term traffic counts throughout a network based on limited information, using ArcGIS tools. Though such methods can be further refined (for example, by including more influential factors and using network distances between all count locations), these predictions make effective use of temporal and spatial information in existing data sets. These predicted values can be used as estimates of traffic conditions at unmonitored sites in any year, facilitating system management, data analysis, and investment decisions.

Further, kriging can be applied over large geographic scales (such as an entire state) even with sparse detector coverage; of course, accuracy is improved with greater coverage. At the same time, kriging has several drawbacks: it is at best a coarse approximation of traffic volumes, and does not account for the actual transportation infrastructure or demand patterns, and can only be applied at aggregate time-scales. Still, when only sparse data exist, kriging is the best interpolation method.

Chapter 5. Prototype System Test

Chapters 3 and 4 studied and developed individual components of an integrated data archive, emphasizing organizational, methodological, and technical aspects of such a system. In particular, the following challenges were addressed:

- A **common data format** was identified, allowing any type of traffic detector (including innovative technologies) to provide input to the archive at any reporting frequency.
- A novel **data reliability algorithm**, based in continuous set theory, was developed to flag suspicious data according to three criteria: the degree to which data respect fundamental traffic laws, spatial consistency with nearby detectors, and temporal consistency with previously observed data at the same location.
- Several **error correction and imputation** procedures were studied and developed, using a variety of statistical techniques. While imputed data are not suitable for all applications of a data archive, they are highly useful for others, and the accuracy of these systems was studied.

This chapter describes a test of these using field data from northwest Houston, collected in October 2007. Three detectors are used for this test: two side-fire radar detectors, and one ATR. Data from each of these detectors was converted into a common format, and entered into a flat-file database.

The analysis in this task is driven by one of the primary motivations for implementing such an archive: using other ITS detectors to augment ATRs in AADT counts. Thus, the focus is on the suitability of using neighboring ITS detectors to estimate missing or unreliable ATR data.

Of course, it is impossible to compare truly missing ATR data to the actual volumes on that day; for this reason, missing data was simulated by randomly deleting selected observations, applying the imputation procedure, and comparing the imputed value to the actually-observed value. Both sporadic communication failures and longer power outages were simulated, by either deleting individual observations, or by deleting a block of observations.

The remainder of this technical memo describes this test in greater detail. First, the data are described, followed by the results of the reliability analysis on the ATR data. Two different imputation techniques are then compared, and the results discussed. The memo concludes with a summary of the key findings from this experiment.

5.1 Data Acquisition and Processing

Traffic data was obtained from three detectors in northwest Houston. Two of these detectors are side-fire radar (IDs 1083 and 3989, respectively located on US-59 at SH 288 and on US-290 at Telge) operated by the Houston Traffic Management Center, and the other is an ATR operated by the Transportation Planning and Programming division of TxDOT (Station 3, ID 66, located 4.5 miles west of FM 1960). Thus, the ATR and detector 3989 are closely located in northwest Houston, while detector 1083 is farther away, located downtown (Figure 5.1). All available data were obtained for October 2007; this month was chosen for its lack of public holidays, and to avoid demand fluctuations that occur in the summer. The ATR provides hourly

volumes for each lane, while the side-fire detectors provide per-lane volume, speed, occupancy, and vehicle classifications at a 30-second resolution.

October has a total of 744 hours; of these, the ATR reported data for all but 72 hours, indicating 90.3% data availability. Likewise, this month has a total of 89,280 thirty-second intervals. Detectors 3989 and 1083 respectively report data for 88,967 and 80,268 of these, indicating 99.6% and 89.9% data availability for this month, respectively.

The ATR and radar detectors report this information in considerably different formats: the ATR uses a text spreadsheet, while the radar detectors use a flat-file database. For further analysis, software was written to post-process the data, converting these data into a common format, allowing direct comparison.

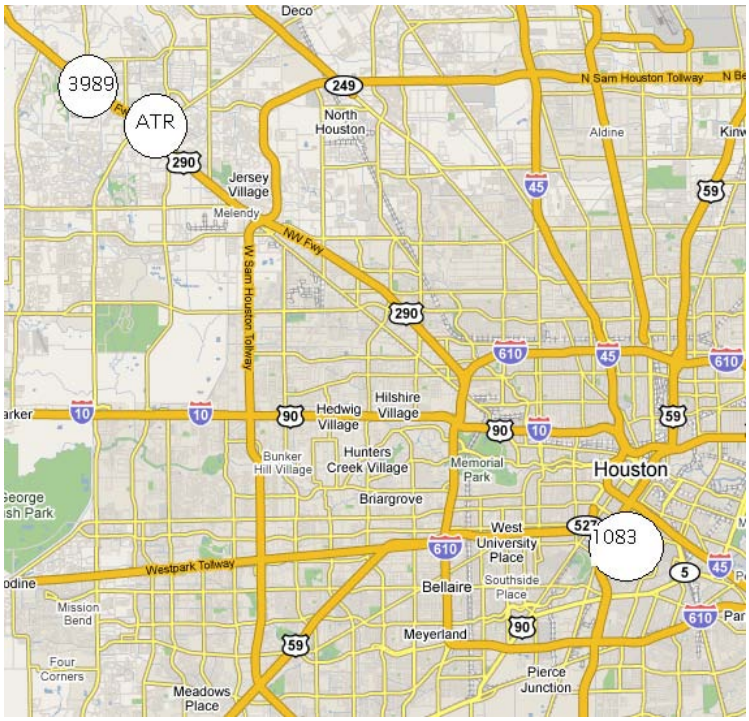


Figure 5.1: Location of detectors used in this study.

5.2 Reliability Analysis

The reliability index (RI) calculation procedure, described in Section 4.2, was applied to the ATR data. Because data was only available for one nearby detector (3989) and no intermediate onramps, the "network reliability" calculation had to be modified as follows: first, the average ratio between the 3989 and ATR volumes was calculated; second, the 3989 volume was divided by this ratio; and third, this value was input as the only upstream detector for the network reliability procedure. Essentially, we are assuming that the on- and off-ramp volumes vary in direct proportion with the volume at station 3989. Although not entirely accurate, this provides a reasonable approximation to the missing data at these locations.

The results of this analysis are shown in Table 5.1 and Figure 5.2, providing a histogram of reliability indices. In Table 5.1, the column marked "Cumulative" indicates the percentage of the data attaining *at least* a certain reliability level; that is, 51% of the data have a RI of 9 or higher, 60% have an RI of 8 or higher, and so on. Note that over half of the data have an RI of

either 9 or 10, and nearly 80% of the data has an RI of 5 or higher, indicating that a significant majority of the data is highly reliable.

Table 5.1: Distribution of reliability indices

Reliability Index	*Proportion*	*Cumulative*
0	0%	100%
1	10%	100%
2	0%	90%
3	1%	90%
4	10%	89%
5	13%	79%
6	2%	66%
7	4%	64%
8	10%	60%
9	24%	51%
10	27%	27%

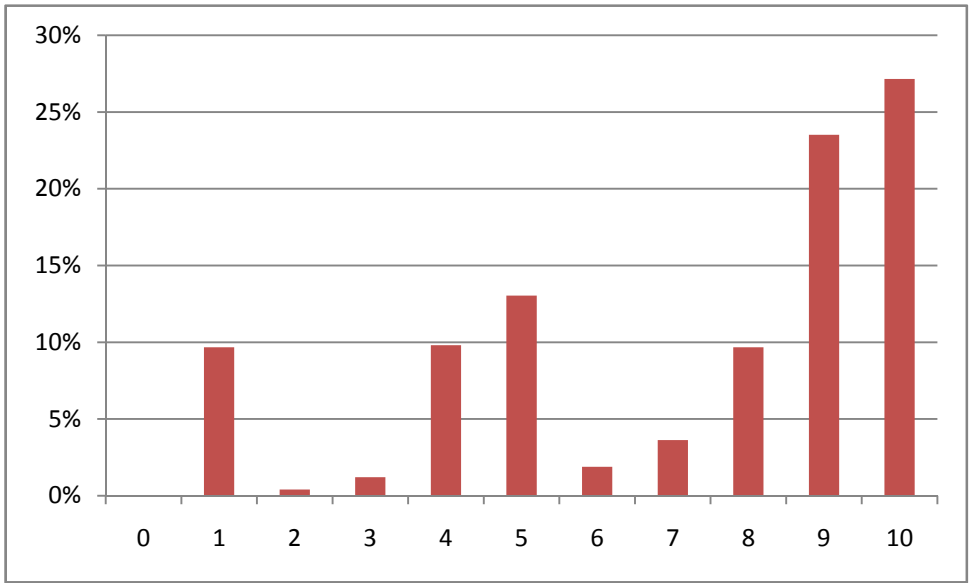


Figure 5.2: Histogram of reliability indices.

5.3 Simulation Experiments

Field data are found missing for a variety of reasons; sporadic communication errors result in occasional "dropped" data at randomly distributed times, while routine system maintenance creates a regular pattern of missing data. Other events, such as power outages, result in blocks of data missing, but at unscheduled intervals.

Using the data described above, simulation experiments were performed to examine the performance of data imputation algorithms for different types of missing data. In order to study the accuracy of these algorithms, the following procedure was adopted: first, selected observations were deleted (forming the "missing" data). Next, two imputation procedures were applied: double linear regression, and historical imputation. Finally, the imputed data was compared to the actual observation (the one which was deleted), allowing the accuracy to be assessed. Note that the missing data must be created artificially, by deleting from actual observations, because it is impossible to measure the accuracy of the imputed data unless the actual observation is known for comparison.

The first experiment is intended to represent irregular communication failures, and is simulated by randomly deleting ten percent of the (hourly) ATR data observations from October 2007. The remaining ninety percent of the data is used to calibrate two models:

Double linear regression relates the hourly volumes measured at the two radar detectors (3989 and 1083) to the hourly volume measured at the ATR, that is,

$$q_{ATR} = \beta_0 + \beta_{1083} q_{1083} + \beta_{3898} x_{3989}$$

where q_{ATR}, q_{1083}, and q_{3989} are the volume readings at the ATR, 1083, and 3989, respectively, while β_0, β_{1083}, and β_{3989}, are regression parameters. Note that q_{1083} and q_{3989} are actually observed, while q_{ATR} is the imputed value. A "least squares" estimate of the regression parameters is made; that is, β_0, β_{1083}, and β_{3989} collectively minimize the average squared difference between the actual ATR volume reading and $\beta_0 + \beta_{1083} q_{1083} + \beta_{3898} x_{3989}$, according to the calibration data set.

When applied to the observations from October 2007, the following relation was established:

$$q_{ATR} = -144.2 + 0.4576 q_{1083} - 0.00472 x_{3989}$$

Regression analysis also provides an indication of confidence in the equation, called the *R*-squared value, which ranges between zero (no relation between radar detectors and missing data) and one (exact correspondence). The *R*-squared value for this equation is 0.989, which is extremely good, and indicates a very close fit between the calibrated equation and the observed data. Another measure of the equation's validity is the *t*-statistic, which shows how well each of the input data (side-fire radar volumes) predicts the missing ATR data. The *t*-statistics for detectors 1083 and 3989 are 0.674 and 106.8, respectively, indicating the detector 3989 is extremely useful in predicting the missing ATR data, while detector 1083 is somewhat less useful. This is in accordance with intuition, since detector 3989 is located very close to the ATR station, while detector 1083 is located much farther away, on a different facility.

Historical imputation adopts a different method for predicting missing data: rather than deriving an equation based on detectors at other locations, historical imputation is accomplished by replacing the missing observation with a "historical average." For this experiment, the historical average is defined as the mean volume corresponding to the hour of the day that is missing. This method has the advantages of being simpler and not relying on the presence other data (for example, double linear regression cannot be applied if either of the input detectors is also missing data); the clear disadvantage is that a purely historical method cannot model any deviation from past patterns.

After calibrating both of these models using ninety percent of the data, the remaining ten percent were used to test their accuracy. Double linear regression proved much more accurate: the average absolute error was 7.55%, as compared to 23.8% with historical imputation. This difference can also be seen graphically: Figures 5.3 and 5.4 plot the actual observations (horizontal axis) against the imputed values (vertical axis); the accuracy of the linear regression model is clearly evident in these figures. Notice that historical imputation can only predict one of twenty-four possible values (one for each hour), providing less flexibility for modeling actual field conditions.

A second experiment was performed in order to test the suitability of these imputation methods for generating AADT counts, one of the main motivations for this project. Current best practices have stringent data requirements: if an hour's observation is missing, that day cannot be used in the calculation of AADT. Factoring approaches exist in the literature, but suffer from the same limitation as historical imputation, namely, assuming that past conditions are replicated perfectly in the present. However, the regression approaches have the potential to bypass this limitation to some extent, by using concurrent data from other locations.

The following experimental setup was adopted: twenty-four consecutive hours of data were chosen, corresponding to one of the days in October. For each of these hours, volume at the ATR location was imputed using the double linear regression equation calibrated in the first experiment. The total daily volume estimated in this way is then compared to the actual total daily volume measured at the ATR site. Table 5.2 shows the results of this comparison (note that October 23-26 data are missing, since ATR data are missing for these days, and cannot be used for comparison.

The average error in calculating daily volumes in this way is 3.43%, which is smaller than that used when predicting individual hourly volumes. (This is in accordance with statistical results such as the Central Limit Theorem.) This suggests that regression can be a powerful tool for estimating daily volumes when missing data are present.

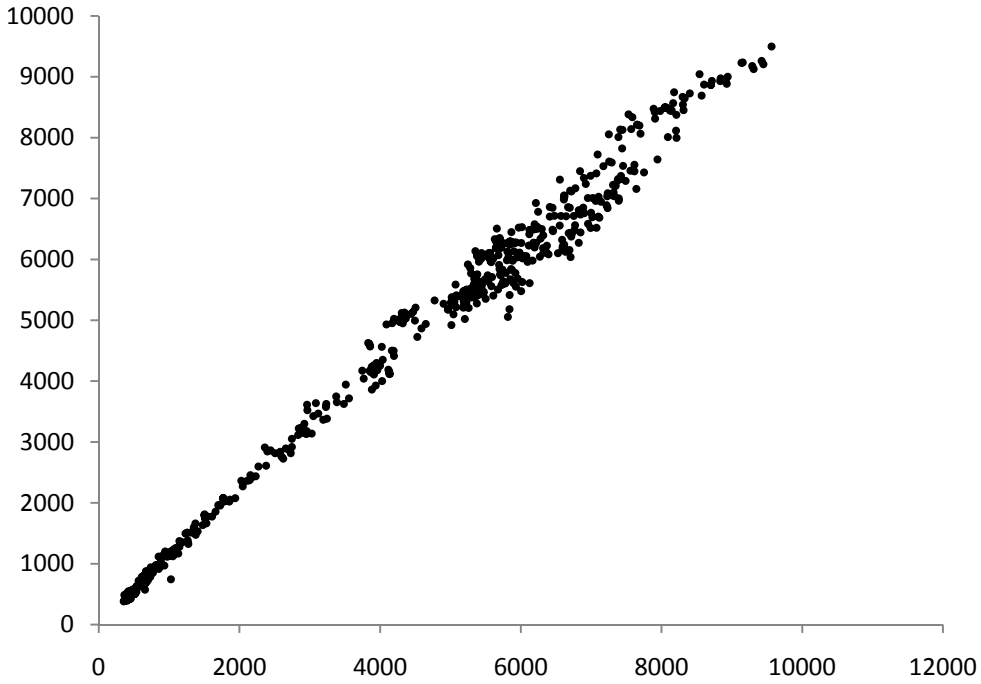


Figure 5.3: Imputed vs. actual observations, double linear regression

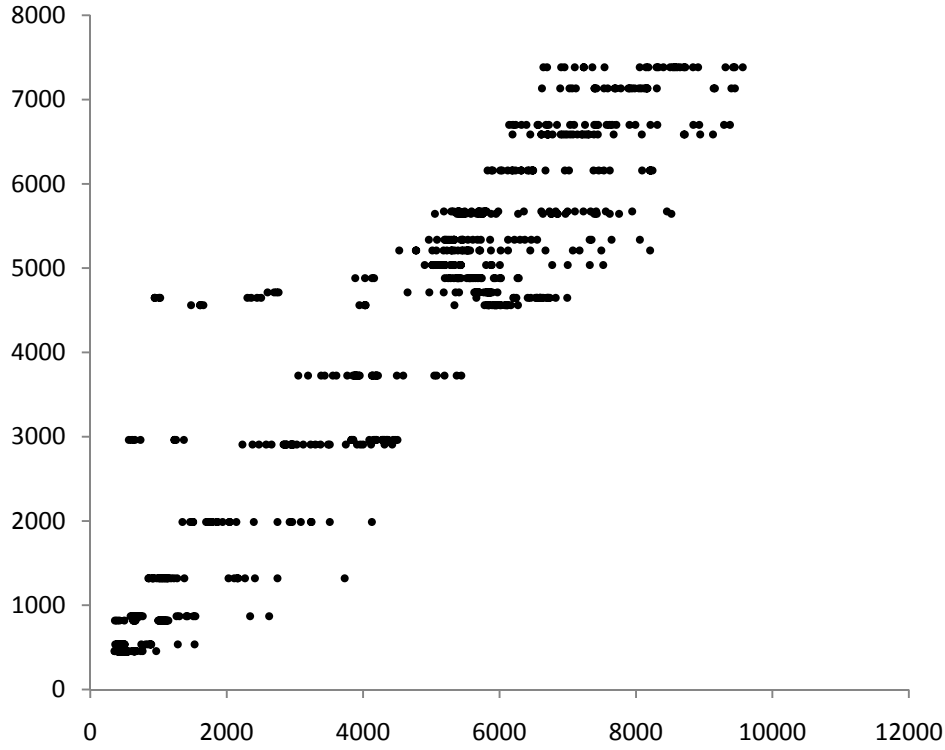


Figure 5.4: Imputed vs. actual observations, historical imputation

Table 5.2: Observed vs. imputed volumes for daily traffic counts

Day	*Observed volume*	*Imputed volume*	*Error*
1	101905	106232	+4.07%
2	103435	108158	+4.37%
3	105106	110689	+5.04%
4	108477	113539	+4.46%
5	123713	127162	+2.71%
6	105033	107918	+2.67%
7	86796	87206	+0.47%
8	100991	108445	+6.87%
9	102613	108862	+5.74%
10	105311	110859	+5.00%
11	110641	113633	+2.63%
12	124126	126811	+2.11%
13	106645	109447	+2.56%
14	89082	87356	–1.97%
15	98464	101924	+3.39%
16	104179	109658	+5.00%
17	103401	110089	+6.08%
18	109041	114358	+4.65%
19	123007	127213	+3.31%
20	107282	110100	+2.56%
21	89285	88848	–0.49%
22	98441	103314	+4.72%
26	127322	129209	+1.46%
27	119568	118932	–0.53%
28	95107	93286	–1.95%
29	104402	108589	+3.86%
30	107249	111343	+3.68%
31	105179	109116	+3.61%

Chapter 6. Conclusions

This research has addressed a number of issues related to the storing and archiving of data collected by ITS equipment, and its suitability for use in a variety of applications. Both the research literature and agency experience indicate that data quality is one of the most significant barriers preventing widespread adoption of this type of data sharing. To this end, a new technique for quantifying confidence in ITS data has been developed, evaluating data according to its fundamental consistency (consistency with basic physical constraints), network consistency (consistency with nearby detectors), and historical consistency (consistency with past data at the same detector). Missing data is a major factor as well, and a comparison of imputation methods was undertaken. Our results indicate that the linear regression-based models are the most accurate, although they have higher input data requirements. While imputing missing values based on historical data is less accurate, this approach is almost always usable.

A prototype system was constructed, incorporating these routines, and receiving data from an ATR and a side-fire radar detector. This system is flexible with regard to input data type, and is capable of receiving data from virtually any traffic detector once a small program is written to translate the detector data into the standard format. A web interface was constructed, allowing any user to access stored data and generate reports. Implementation guidelines are also given to show how such a system can be constructed incrementally.

Appendices to the main text illustrate other interesting findings, including an example application to transportation planning, and a demonstration suggesting that reducing the amount of stored data need not significantly degrade the quality of stored data – statistical time-series analysis techniques can be used to recover omitted data with a high degree of accuracy.

Given the technical feasibility of implementing a data archive, as illustrated by the prototype system, and given the potential benefits of a rich data source to transportation planners, operations personnel, and others, the advantages of implementing a shared archive should be evident. Although institutional barriers still exist, the future outlook for this type of data sharing is bright.

References

ADUS. Archived Data User Service (ADUS). (1998) An Addendum to the ITS Program Plan. ADUS Program, USDOT.

Al-Deek, H. M. and C. V. S. R. Chandra. (2004) New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Transportation Research Record* 1867, 116-126.

von Altrock, C.(1995). *Fuzzy logic and NeuroFuzzy applications explained*. Upper Saddle River, NJ: Prentice Hall PTR

Ashok K. and Ben-Akiva M. (1993) Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. Proc., 12th International Symposium on Transportation and Traffic Theory, 519-540.

Bertini, R.L. , S. Hansen, A. Byrd and T. Yin. (2005) Experience Implementing a User Service for Archived Intelligent Transportation Systems Data. *Transportation Research Record* 1917, 90–99.

Brockwell, P. J. and Davis, R. A., 2002. *Introduction to Time Series and Forecasting*, 2nd. ed., Springer-Verlang.

Bureau of Public Roads. (1964) *Traffic Assignment Manual*. Urban Planning Division, U.S. Department of Commerce.

Casella, G. and Berger, R.L., 2001. *Statistical Inference*, 2nd ed. Belmont, CA: Duxbury Press.

Casey F., Labell L., Carpenter E., LoVecchio J., Moniz L., Ow R., Royal J., Schwenk J., Schweiger C., and Marks B. (1998) Advanced Public Transportation Systems: The State of the Art. FHWA report FTA-MA-26-7007-98-1.

Cassidy, M. J., and R. L. Bertini. (1999) Some Traffic Features at Freeway Bottlenecks. *Transportation Research Part B* 33, 25–42.

Cathey F. and Dailey D. (2003) A prescription for transit arrival/departure prediction using automated vehicle location data. *Transportation Research C* 11, 241-264.

Chen, C., J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya. (2003) Detecting errors and imputing missing data for single-loop surveillance systems. *Transportation Research Record* 1855, 160-167.

Chen, L. and A. D. May. (1987) Traffic detector errors and diagnostics. *Transportation Research Record* 1132, 82-93.

Codd, E.F. (1970) "A relational model of data for large shared data banks", Communications of the ACM, Volume 13, Issue 6, pp. 377-387.

Codd, E.F. (1971) “Further normalization of the data base relational model” Courant Computer Science Symposia.

Coifman, B. (1999) Using dual loop speed traps to identify detector errors. *Transportation Research Record* 1683, 47-58.

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley Interscience: New York.

Dahlgren, J., Garcia R., and Turner S. Completing the Circle: Using Archived Operations Data to Better Link Decisions to Performance. PATH report UCB-ITS-PRR-2001-23

Dowling R., A. Skabardonis, J. Halkias, G. McHale, and G. Zammit, (2004). Guidelines for Calibration of Microsimulation Models. *Transportation Research Record* 1876,. 1–9.

FHWA. (1999) ITS Data Archiving: Case Study Analyses of San Antonio TransGuide Data. FHWA Report FHWA-PL-99-024.

FHWA. (2005a) Archived Data Management Systems: A Cross-Cutting Study.

FHWA. (2005b) Planning Analysis Tools for Operations/ITS Evaluation: Gap Study

Gadda, S., Magoon, A. and Kockelman, K. (2007) Estimates of AADT: Quantifying the Uncertainty. Presented at the World Conference on Transportation Research, Berkeley, California, and under review by the *Journal of Transportation Engineering*. (Available at http://www.ce.utexas.edu/prof/kockelman/public_html/TRB07AADTUncertainty.pdf.)

Gold, D. L., S. M. Turner, and B. J. Gajewski, and C. Spiegelman. (2000) Imputing missing values in ITS data archives for intervals under 5 minutes. Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Hall J. (2003) Data Partnerships: Making Connections for Effective Transportation Planning. Transportation Research Circular E-C061.

Hamilton, J. D. (1994) *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Hoef, J., Peterson, E., Theobald, D. (2006) Spatial Statistical Models that Use Flow and Stream Distance, *Environmental and Ecological Statistics* 16: 449-464.

Ishimaru J. and Hallenbeck M. (1999) FLOW Evaluation Design Technical Report. TRAC technical report, Project T9903, Task 62.

Jack Faucett Associates (1997) Guidance Manual for Managing Transportation Planning Data. NCHRP report 401.

Kruvoruchko, K. and Gribov, A. (2004) Geostatistical Interpolation and Simulation with Non-Euclidean Distances. *geoENV IV*, eds. Xavier Sánchez-Vila, Jesús Carrera, and J. Jaime Gómez-Hernandez. Kluwer Academic Publishers.

Lafleur, C. (1998) MATLAB Kriging Toolbox. Available online: http://globec.whoi.edu/software/kriging/V3/english.html. Accessed February 8, 2008.

Lomax T., Turner S., and Margiotta R. (2001) Monitoring Urban Roadways in 2000: Using Archived Operations Data for Reliability and Mobility Measurement. FHWA report FHWA-OP-02-029.

Margiotta R. (2002) State of the Practice for Traffic Data Quality. White paper, prepared for Traffic Data Quality Workshop, Work Order BAT-02-006.

Nguyen, L. H. and W. T. Scherer. (2003) Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications. Technical report UVACTS-13-0-78, University of Virginia, Center for Transportation Studies.

Nihan, N., X. Zhang, and Y. Wang. (2002) Evaluation of dual-loop error using video ground truth data. Technical report, Joint Report TNW02-02, WA-RD 535.1, Transportation Northwest/Washington State Department of Transportation

Nihan, N., L. N. Jacobson, J. D. Bender, and G. Davis. (1990) Detector data validity. Technical report WA-RD 208.1, Washington State Transportation Center.

Payne, H. J., E. D. Helfenbein, and H. C. Knobel. (1976) Development and testing of incident detection algorithms. Technical report FHWA-RD-76-20, Federal Highway Administration, US Department of Transportation, 1976.

Taylor C. and Meldrum D. (2000) A Programmer's Guide to the Fuzzy Logic Ramp Metering Algorithm: Software Design, Integration, Testing, and evaluation. TRAC technical report, Project T9903, Task 84.

Transportation Research Board. (2000) *Highway Capacity Manual*. National Research Council, Washington, D.C.

Turner S., Brydia R., Liu J., Eisele W. (1997) ITS Data Management System: Year One Activities, Transportation Institute Technical Report 1752-2

Turner, S. (2001) Guidelines for Developing ITS Data Archiving Systems. Texas Transportation Institute Technical Report 2127-3.

Turner, S. (2002) Defining and Measuring Traffic Data Quality. White paper, prepared for FHWA Traffic Data Quality Workshop, Work Order BAT-02-006

Turner, S., L. Albert, B. Gajewski, and W. Eisele. (2000) Archived intelligent transportation system data quality: preliminary analyses of San Antonio TransGuide data. *Transportation Research Record* 1719, 77-84.

Turner, S.M., Eisele, W.L., Gajewski, B.J., Albert , L.P. and Benz, R.J., August 1999. *ITS Data Archiving: Case Study Analyses of San Antonio TransGuide® Data*. Report No. FHWA-

PL-99-024. Federal Highway Administration, Texas Transportation Institute, College Station, Texas.

USDOT (2000). Incorporating ITS Solutions into the Metropolitan Transportation Planning Process: Overcoming Institutional Barriers

Vanajakshi, L. and L. R. Rillett. (2004) Loop detector data diagnostics based on conservation-of-vehicles principle. *Transportation Research Record* 1870, 162-169.

Varaiya P. (2002) California's performance measurement system: improving freeway efficiency through transportation intelligence. *TR News* 218, Jan.-Feb. 2002.

Wall Z. and Dailey D. (1999). An algorithm for predicting the arrival time of mass transit vehicles using automated vehicle location data. Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, DC.

WSDOT (2006). Ramp & Roadway 2006: Traffic Management Center Summary Report.< This is an example of the "Reference Text" predefined style for references. This is an example of the "Reference Text" predefined style for references. >

Appendix A: Equipment Guidebook

A data archive intended for long-term use should be able to accommodate new and innovative detector technologies introduced at a later date. As described in Chapter 3, this is accomplished on the database side by converting all recorded data into a common format, allowing all data to be treated identically by the archive.

In the course of this research, an assessment was made of both standard and innovative detector technologies. Strengths and weaknesses of each technology were identified, relative to implementation and suitability for data archiving (e.g., data quality). This guidebook summarizes the technical research performed in this manner, describing the following technologies in this order:

- Video detection technology
- Wireless location technology
- Laser detection
- Infrared technology
- Radar and acoustic traffic sensors
- Inductive loop detectors
- Weigh-in-motion
- Wireless magnetic technology
- Intelligent road studs (IRS)
- Aerial image analysis

A list of references is provided along with each technology.

Video Detection Technology

Description: Video cameras placed alongside a roadway can be used for traffic detection and monitoring, providing real-time images of traffic conditions (Figure A.1).

(Source: City of Overland Park, Kansas)

Figure A.1: Video detection technology

How it Functions: Cameras continuously record images of traffic conditions, which can be analyzed to extract a variety of information.

Data Provided: Image analysis algorithms can be applied to video data to obtain traffic volumes, vehicle classification, occupancy, or speed by identifying and isolating individual vehicles.

Advantages: At least in theory, nearly all desired traffic data can be extracted from video footage. Further, these systems are non-intrusive, and do not require roadway disruption for installation or maintenance.

Disadvantages: The performance of image analysis algorithms can suffer in poor weather, darkness, glare, or shadows; these can be partially mitigated by installation of lighting near the camera, or by installing multiple cameras to improve resolution and accuracy.

Additional Information:

Chang, S., Chen, L., Chung, Y., and Chen, S. "Automatic License plate Readers." IEEE Transactions of ITS, Volume 5, Issue 1 pp 42-53, 2004.

Tseng, B.L., Lin, C., and Smith, J.R. "Real Time Video Surveillance for Traffic Monitoring Using Virtual Line Analysis" 2002 IEEE International Conference on Multimedia and Expo, Vol.2 pp 541-544

Douret, J., Benosman, R., "A multi-cameras 3D volumetric method for outdoor scenes: a road traffic monitoring application." Proceedings of the 17th International Conference on Pattern Recognition, 2004.

Lee, H., Daehwan, K., Daijin, K., Bang, S. Y., "Real-Time Automatic Vehicle Management System Using Vehicle Tracking and Car Plate Number Identification" Proceedings. 2003 International Conference on Multimedia and Expo Volume 2, Issue , 6-9 July 2003 Page(s): II - 353-6 vol.2

Wireless Location Technology

Description: Wireless devices, such as cellular phones, can serve as traffic probes that can collect data based on direct observations (Figure A.2).

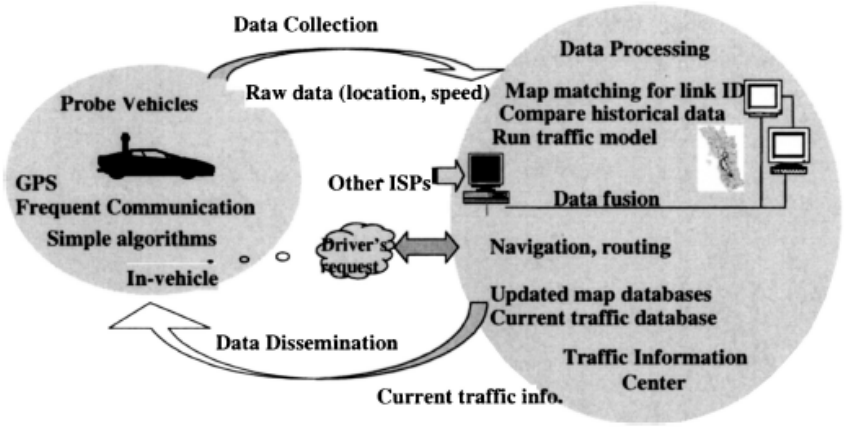


Figure A.2: Wireless location technology

How it Functions: The locations of cellular phones or other devices can often be triangulated using either ground infrastructure (such as phone towers) or global positioning satellites (GPS). (While the former is less accurate, power requirements are smaller.) By observing this information over time, vehicle trajectories and speeds can be obtained.

Data Provided: Real-time speed and travel times can be directly observed. Further, volume can be indirectly observed, if the approximate penetration of wireless probes in the driving population is known.

Advantages: Agencies are not required to deploy or maintain any infrastructure; rather, the probe devices are privately owned. Thus, this technology has very low deployment and operational costs.

Disadvantages: The accuracy of the triangulation can be problematic, especially with parallel facilities that are closely located (for instance, a freeway and a frontage road). Further, privacy issues are extremely important, and users must either be assured of the anonymity of the data, or receive some benefit (such as free real-time travel information) in return for the use of their wireless device as a traffic probe**.**

Additional Information:

Fontaine, M. D. , Smith, B. L. "Improving the effectiveness of traffic monitoring based on wireless technology" Virginia Transportation Research Council 05-R17 , 2004.

Fontaine, M. D., Smith, B. L, " Investigation of the Performance of Wireless Location Technology-Based Traffic Monitoring Systems" Journal of Transportation Engineering , March 2007

Wunnava, S., Yen, K., Babij, T., Zavaleta, R., Romero, R., Archilla, C., "Travel Time Estimates Using Cell Phones on Highways and Roads" Final Report Prepared for the Florida Department of Transportation, Florida International University, January 29, 2007

Laser Detection

Description: A laser is installed above the roadway, emitting a beam aimed at a photodiode array placed on the pavement (Figure A.3).

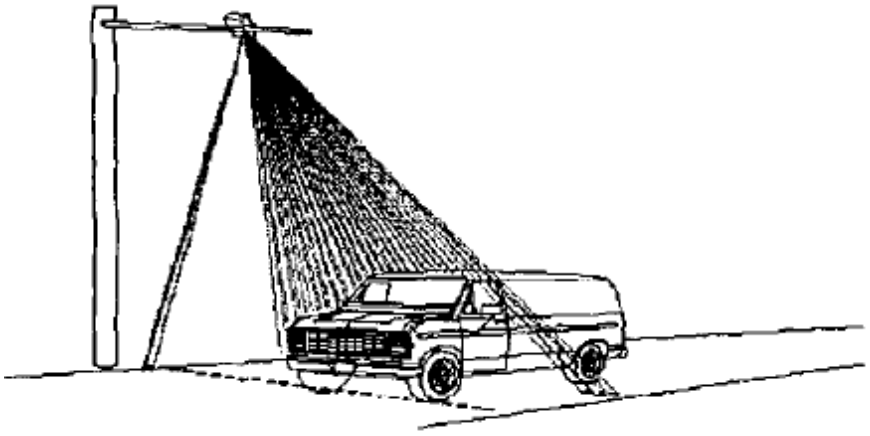


Figure A.3: Laser detection technology

How it Functions: Vehicles passing underneath the laser break the beam, allowing the photodiode array to detect its presence.

Data Provided: Volume is directly measured; two closely-spaced detectors can also provide speed information. Vehicle classification can also be attempted, based on vehicle length.

Advantages: Lasers have been found to outperform loop detectors as well as video detection in conditions of rain and fog. Further, the system does not distract drivers and poses no safety risk.

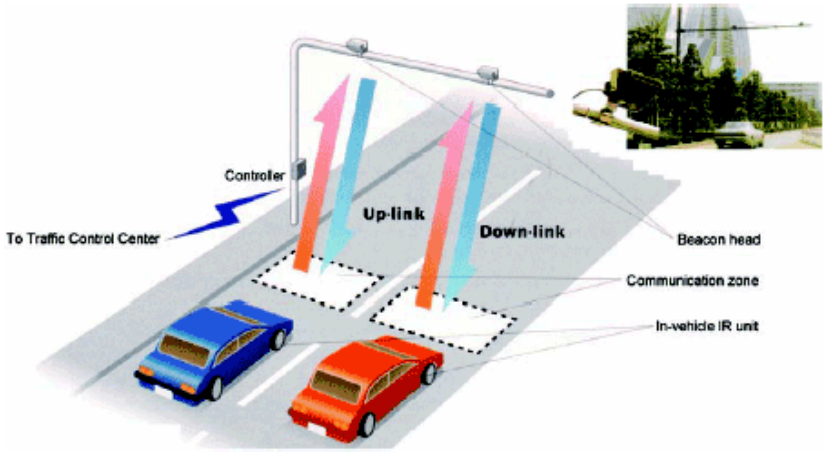
Disadvantages: Detection accuracy is diminished in stop-and-go traffic. Also, some research indicates that laser detectors perform less well in extreme temperatures.

Additional Information:

Cheng, H. H., Shaw, B. D., Palen, J., Lin, B., Chen, B., and Wang, Z. "Development and Field Test of a Laser-Based Nonintrusive Detection System for Identification of Vehicles on the Highway." IEEE Transactions on Intelligent Transportation Systems, Vol. 6, No. 2, June 2005

Infrared Technology

Description: Infrared transmitters and receivers are located on the roadside, and detect passing vehicles (Figure A.4).



(Source: Sumitomo Electric USA, Inc.)

Figure A.4: Infrared technology

How it Functions: Passing vehicles can be observed through disturbances in the infrared beam.

Data Provided: Volume, speed, vehicle classification, and lane position are all recorded by infrared detectors**.**

Advantages: Infrared systems are highly accurate, and a single installation can provide data on up to twenty lanes. Further, installation and maintenance do not disturb traffic flow, since the devices are located on the side of the road.

Disadvantages: Infrared detection can be expensive, and the signal can be scattered due to rain or snow. Further, the roadside location makes equipment vulnerable to collisions and vandalism.

Additional Information:

Kotzenmacher, J., Minge, E., and Hao, B., "Evaluation of Portable non-intrusive traffic detection system," Minnesota Department of Transportation, St. Paul, Minnesota. IMSA Journal, 2004.

Radar and Acoustic Traffic Sensors

Description: A radar or acoustic device is installed either on the side of the roadway, or above one or more travel lanes (Figure A.5).

(Source: SmarTek Systems)

Figure A.5: Radar/acoustic transportation technology

How it Functions: Radar or acoustic signals are emitted by the device, and the resulting reflection used to collect traffic data.

Data Provided: Speed, volume, length-based classification, and lane position are all directly obtained from the radar system.

Advantages: Radar systems are reliable under a variety of weather conditions, and have only a moderate cost.

Disadvantages: Accuracy is reduced for slow-moving traffic, and large vehicles (such as trucks) create both "shadowing" effects that mask vehicles in other lanes, as well as an overcounting effect if the truck is mistakenly counted as multiple vehicles in the same lane; this latter effect is most common if the truck passes very close to the detector.

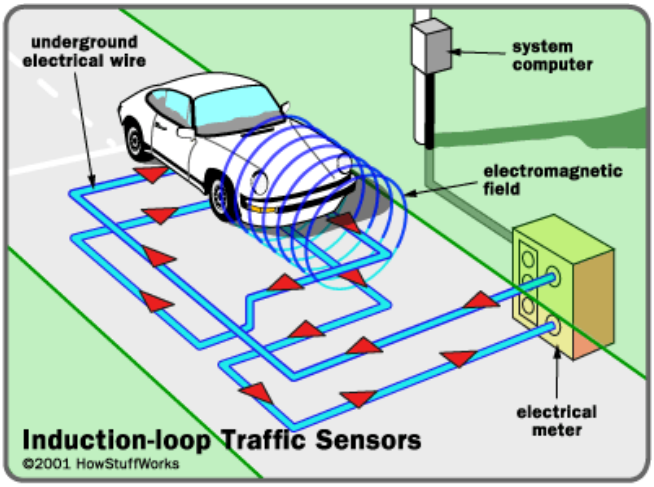
Additional Information:

Kotzenmacher, J., Minge, E., and Hao, B., "Evaluation of Portable non-intrusive traffic detection system," Minnesota Department of Transportation, St. Paul, Minnesota. IMSA Journal, 2004.

Michalopoulos, P., Hourdakis, J., "Review of Non-Intrusive Advanced Traffic Sensor devices for advanced traffic management systems and recent advances in video detection" Proceedings of the Institute of Mechanical Engineers, Vol. 215 Part 1, 2001

Inductive Loop Detectors

Description: A circular loop is placed in the pavement, and connected to an electronics box on the side of the road (Figure A.6).



(Source: HowStuffWorks)

Figure A.6: Inductive loop detector technology

How it Functions: Large metallic objects passing over loops induce a current, which is then observed to detect the presence of a vehicle.

Data Provided: Volume and occupancy are directly measured. Two detectors located in close sequence can be used to obtain speed data. Attempts at vehicle classification can also be made, based on speed and occupancy.

Advantages: Loop detectors are a well-known and well-studied technology in place throughout the world. Installation is relatively inexpensive, and power requirements are low.

Disadvantages: Being located within the pavement, installation, and maintenance requires traffic disruption; as a result, malfunctioning detectors are often not repaired until the next construction project in that area. Loop detectors are also subject to pavement damage due to vehicles or thermal stresses.

Additional Information:

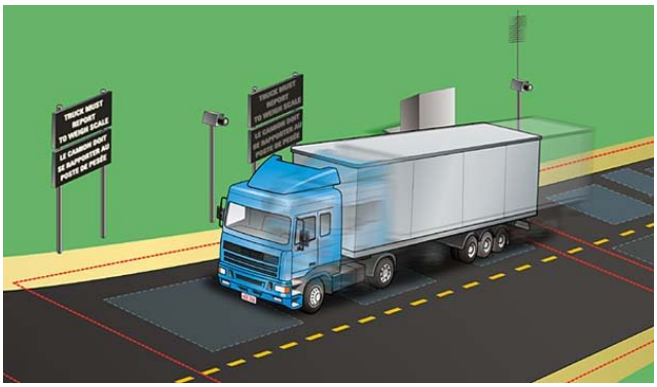
Lin, L., Han, X. B., Ding, R., Li, G., C-Y Lu, S., Hong, Q., "A New Rechargeable Intelligent Vehicle Detection Sensor." Journal of Physics: Conference Series 13 (2005) 102-106.

Coifman, B., Krishnamurthy, S., "Vehicle re-identification and travel time measurements across freeway junctions using the existing detector infrastructure" Transportation Research Part C 15 (2007) 135-153

Oh, C., Ritchie, S. G., "Recognizing Vehicle Classification Using Blade Sensors" Pattern Recognition Letters 28 (2007) 1041 -1049

Weigh-in-Motion

Description: Additional equipment is added to current or newly-constructed weigh-in-motion stations, allowing additional information to be recorded (Figure A.7).



(Source: New Brunswick Department of Transportation)

Figure A.7: Weigh-in-motion technology

How it Functions: Weigh-in-motion technology often uses bending plates, piezoelectric sensors, or fiber-optic load sensors to detect vehicle weight, and transmit this information to a nearby station, often wirelessly. By equipping such stations with additional sensors to detect vehicle speed, additional information can be collected while taking advantage of the existing communications infrastructure used to transmit this data.

Data Provided: Weigh-in-motion stations already record vehicle weight, which can be an excellent proxy for vehicle classification. Obtaining volume from these sensors is not difficult and, as mentioned above, speed data can also be collected given installation of the proper sensor.

Advantages: Measuring weight is extremely useful for vehicle classification, and is rarely collected with other types of detector. Furthermore, weigh-in-motion stations are already equipped with communication devices, indicating that recording other data at these locations can be done with a lower installation cost.

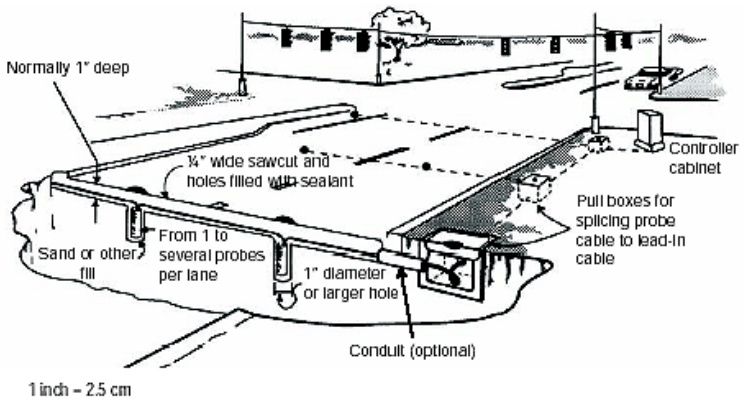
Disadvantages: Although technology has improved, weigh-in-motion technology still has substantial error when measuring vehicle weight. Furthermore, as they are located within the pavement, lane closures are needed, even to install additional sensors in the first place.

Additional Information:

Li, Z., Xiao-Ming Yang; and Zongjin Li "Application of Cement-Based Piezoelectric Sensors for monitoring Traffic Flows" Journal of Transportation Engineering, July 2007 Pages 565 – 573

Wireless Magnetic Technology

Description: Magnetic equipment (such as giant magneto-resistive technology, magnetometers, or magnetic impedance sensors), a microcontroller, a fast semiconductor memory, and a radio transmitter are installed on or near the roadway (Figure A.8).



(Source: Federal Highway Administration)

Figure A.8: Wireless magnetic technology

How it Functions: Moving vehicles generate perturbations in the magnetic field. The microcontroller captures the magnetic signal, and performs an analog-digital conversion to process it; the results are then transmitted to a remote computer using radio wavelengths. with A/D conversion and processes signal. Number, speed, and length of the car are stored in ROM and transmitted through Radio frequency to a PC.

Data Provided: Volume and speed are measured directly

Advantages: Magnetic technology requires relatively little power to operate, since they can utilize the change in magnetic field to assist with the power requirements, and the detectors can operate in a "sleep" mode, greatly reducing power consumption when no vehicles are present.

Disadvantages: Magnetometers are less effective at determining the exact vehicle perimeter, making calculations of vehicle length and occupancy less accurate unless used in combination with other sensors.

Additional Information:

Coleri, S., Cheung, S., and Varaiya, P., "Sensor Networks for Monitoring Traffic" 2005

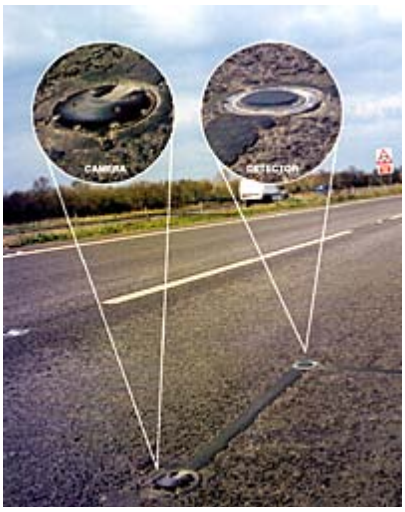
Cheung, S. Y., S. C. Ergen and P. Varaiya "Traffic Surveillance with Wireless Magnetic Sensors," ITS World Congress, November 2005.

Nishibe, Y., Ohta, N., Tsukada, K., Yamadera, H., Nonomura, Y., Mohri, K., Uchiyama, T. "Sensing of a Passing Vehicle Using a Lane Marker on a Road with Built In Thin Film MI Sensor and Power Source" IEEE Transactions on Vehicular Technology. Vol. 23, Issue 6, pp 1827–1834 , Nov. 2004

Doran, V. P. A., Crawford, C. B., "Trial and Evaluation of Intelligent Road Studs for Hazard Warning." The Institute of Electrical Engineers, 2004

Intelligent Road Studs

Description: Optical sensors are placed inside road studs located on the ground (Figure A9).



(Source: Better Roads for the Government/Contractor Project Team)

Figure A.9: Intelligent road stud technology

How it Functions: Two light-activated optical sensors detect passing vehicles within each lane. Solar-powered batteries provide power, needing to be replaced approximately every three years.

Data Provided: Volume, speed, and vehicle classification can be obtained from the optical sensors.

Advantages: Location inside a road stud provides some protection against vehicle impact, while also easing installation and maintenance.

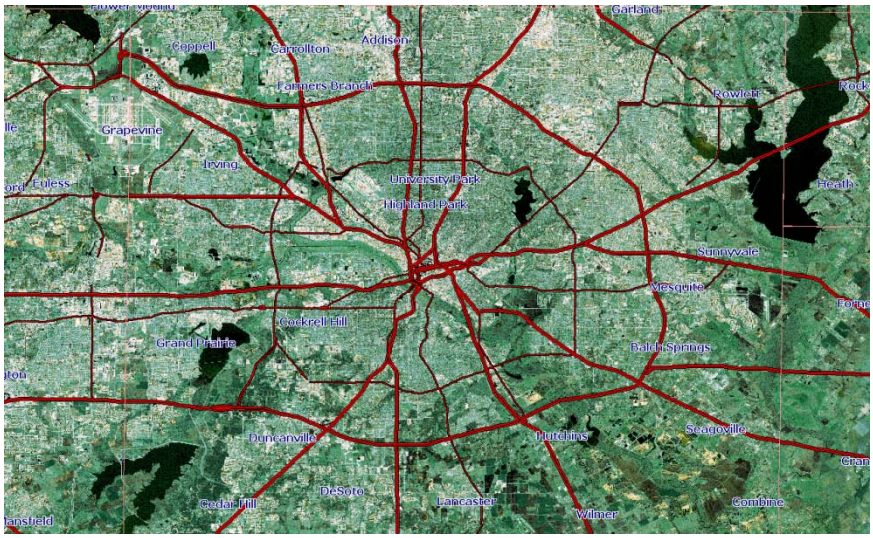
Disadvantages: This technology is new, and conclusive results on accuracy and ease of implementation are not yet available.

Additional Information:

Doran, V. P. A., Crawford, C. B., "Trial and Evaluation of Intelligent Road Studs for Hazard Warning." The Institute of Electrical Engineers, 2004

Aerial Image Analysis

Description: Aerial images are analyzed to obtain traffic data (Figure A.10).



(Source: MapMart)

Figure A.10: Aerial image technology

How it Functions: Aerial imagery (such as from a satellite or aircraft) is combined with a computer map and image analysis software, allowing traffic density and queue lengths to be observed and measured.

Data Provided: Density, queue lengths

Advantages: Aerial imagery provides data for all roadways in a given region, which may not be possible using traditional detectors that only provide data at the point of installation. Further, since all locations are observed simultaneously, the data is always internally consistent.

Disadvantages: As a static observation, aerial photos cannot provide direct information on volume, speed, or other dynamic phenomena. Further, it is costly to obtain such data on a regular basis, and continuous observation is most likely prohibitive.

Additional Information:

Leitloff, J., S. Hinz, U. Stilla "Automatic Vehicle Detection in Space Images Supported by Digital Map Data" CMRT05. IAPRS, Vol. XXXVI, Part 3/W24 2005

Appendix B: Survey Distributed to Texas TMCs

The following survey was distributed to nine traffic management centers in Texas; five responses were received.

Utilizing the Data Collected at Traffic Management Centers for Planning Purposes Through Non-Traditional Sources and Improved Equipment

Thank you for your participation in this project! Due to your experience in traffic operations, you are in the best position to offer us guidance on this project, and we greatly appreciate your help.

1. What type of sensors or detectors does your TMC use? (e.g., loop detectors, CCTV cameras, video-based detection)

2. Approximately how many of these sensors or detectors do you control?

3. What type of data is recorded by these sensors or detectors?

4. What procedures (if any) are performed on the data from the detectors to verify that they are accurate and functioning properly?

5. Is this data stored or archived in any way? (If “no”, skip to question 9)

6. Briefly describe how and where this data is stored.

7. How does one access this data? (e.g., is there a software program that helps you retrieve it?)

8. How often is this data used?

9. What are the most common uses for this data?

10. What is your opinion about how your TMC archives its data? Do you have any suggestions for how this might be improved?

11. There is interest in using archived ITS data for planning purposes as well. What do you foresee as the biggest obstacle in this type of data sharing?

12. A variety of innovative detector technologies have been proposed in recent years, such as in-vehicle transponders, license plate readers, digital aerial image processing, telematics, and so on. Are you familiar with these technologies? Do you feel such technologies would be useful for your TMC? If so, what gaps in current detector coverage would these help fill?

Appendix C: Analysis of Variability in Count Data

To assist in developing measures of data reliability, ATR data has been obtained from Texas, and its variability has been analyzed. Due to a lack of information on variables such as urban or rural classification, number of lanes, or functional class for each ATR site, the analysis is limited to the effects of day of week (DOW), month of year (MOY) and year. Using Florida and Minnesota's ATR data sets, Gadda et al. (2007) provide a sense of the impact of other variables on AADT estimate uncertainty and error.

According to the FHWA, AADT across large-scale networks can be estimated by taking short-period traffic counts (SPTCs) and adjusting for year-to-year trends, month of year (MOY) and day-of-week (DOW) factors developed using count data obtained from permanent automatic traffic recorder (ATR) stations.

Of course, all estimates are only estimates. Uncertainty is involved. This analysis seeks to quantify this uncertainty by essentially treating ATR sites like SPTC sites and examining the mis-prediction that accompanies one or more day's data. AADT can be determined precisely at sites having permanent ATRs that are accurately recording traffic flows throughout the year. There are totally 208 ATRs' data available in Texas, but a certain portion of these 208 sites are not functioning properly in any given year. Thus, of the 7 years (1999-2005), there are totally 900 year-long records (or roughly 130 per year) with adequate data. Based on these 900 year-long records, month-of-year and day-of-week factors could be created expressly and precisely for each location. Thus a year's AADT can also be estimated from each day's SPTC using the following formula:

$$AADT_{est,i} = VOL_i \times M_i \times D_i \times A_i \times G_i \quad \text{(C.1)}$$

where $AADT_{est,i}$ is the estimate of annual average daily traffic count at location i, VOL_i is the actual 24-hour axle volume, M_i is the applicable "seasonal" (MOY) factor, D_i is the applicable DOW factor, A_i is an axle-correction factor for location i, and G_i is a traffic growth factor. In this study, vehicle counts (rather than axle counts) were given and traffic growth through inter-sample years is not considered, so A_i and G_i are both equal 1.0, and the equation reduces to:

$$AADT_{est,i} = VOL_i \times M_i \times D_i \quad \text{(C.2)}$$

Quantifying Errors in Count Estimation

M_i and D_i can be calculated as the ratio of the average daily traffic for the applicable month (for M_i, e.g., all days in January) or day (for D_i, e.g., all Mondays in the year) in question. Since both actual and estimated AADT values were available for all ATR sites in Texas, percentage errors in AADT estimation were calculated as follows:

$$\%Error_i = \frac{\left| AADT_i - AADT_{est,i} \right|}{AADT_i} \quad \text{(C.3)}$$

These are computed as absolute errors, for purposes of averaging, and to achieve a sense of the overall magnitude of uncertainty. It should be noted here that at present, the actual vehicle

counts provided as the *average* of a specific day of week in each month of the year (e.g., the average count for all Mondays in January). Thus, the $AADT_{est,i}$ here is already an average of 4 or 5 days' AADT estimates. Thus, the values effectively represent a 96- or 120- hour count, not a 24-hour count, and the resulting errors are expected to be 50% lower than an error calculated using actual 24-hour daily vehicle counts.

As Figure C.1 shows, most factoring errors ($\%error_i$) are around 10%, with the highest at 31%. In addition, based on the information from these 900 records, Figures C.2 through C.4 provide factoring errors that emerge in different months, different days of the week, and different years are provided in Figures C.2 through C.4. It appears that factoring errors are highest on Sundays (19% average) and Mondays (13%), and lowest on Fridays (6%), when travel patterns may be most stable. In terms of season, average uncertainty appears greatest in September (almost 14%), but the differences across months of the year is not so great (ranging from over 9% to just under 14%). Errors appear to peak in 2004, but no time trends are clearly visible.

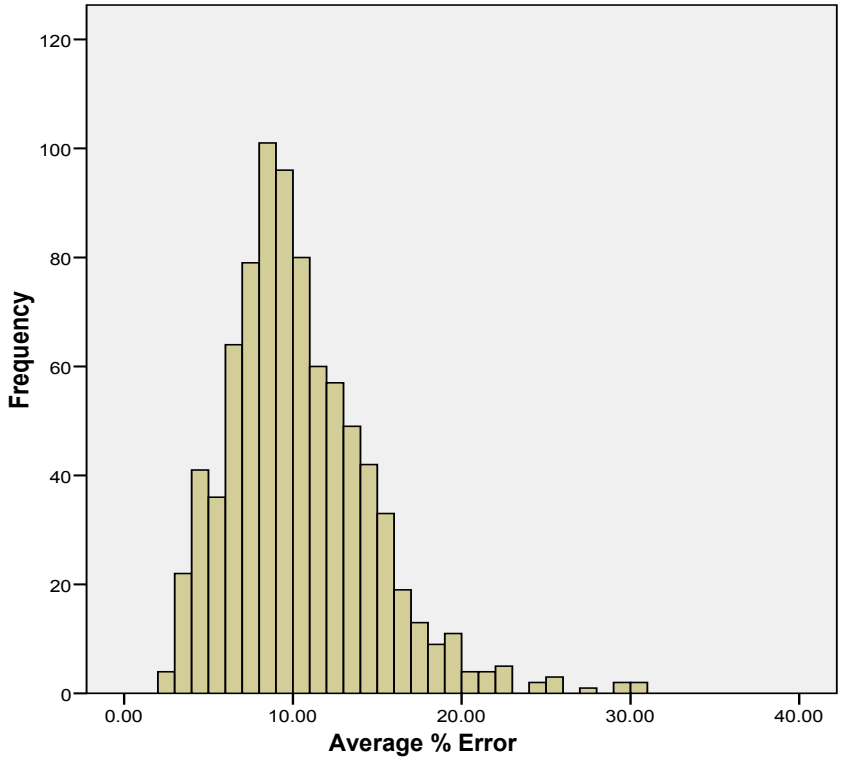


Figure C.1 Histogram of Errors in Predicting AADT from A Single Count Record

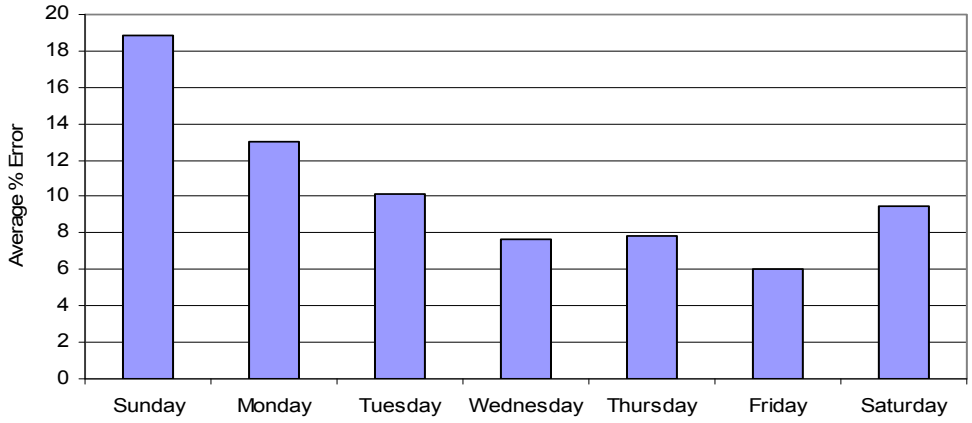


Figure C.2 Average Errors in AADT Estimation Errors by Day of Week

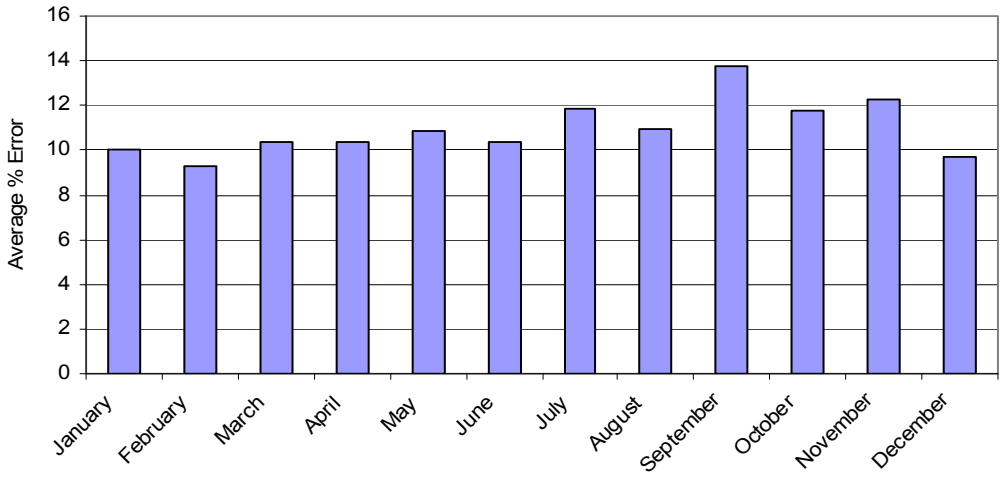


Figure C.3 Variation in AADT Estimation Errors by Month of Year

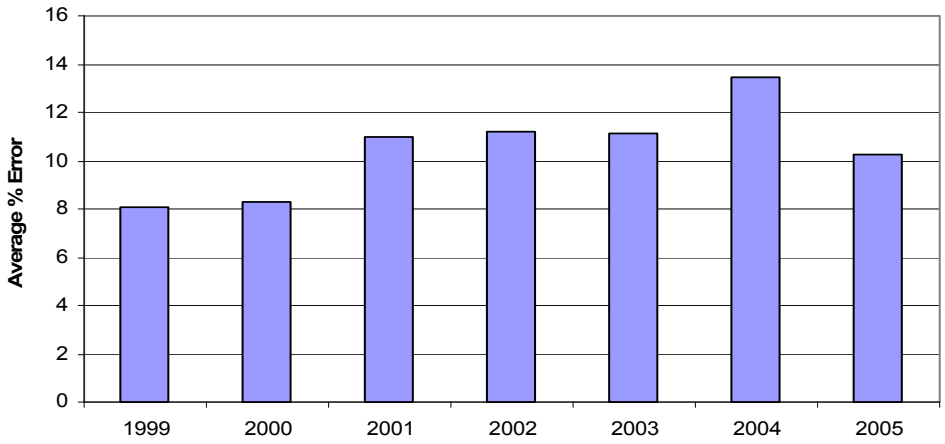


Figure C.4 Variation in AADT Estimation Errors by Year

Summary of Count Estimation Errors

This analysis uses TxDOT's ATR dataset to evaluate AADT estimation errors. In general, if one is using a 4- to 5-day count to estimate AADT, typical errors or mis-prediction are expected to range from 6% to 19%, and average about 10%. And Fridays and February appear to offer best chance at prediction. Such values are lower than Gadda et al.'s (2007) prediction errors for Florida (where average error is 17.5% when using weekend counts, and 12.8% when using weekday counts) and Minnesota (17.8% and 11.3%, respectively), largely because of the 96- to 120-hour count data that TxDOT provided (which reduces variation, relative to Florida and Minnesota's 24-hour counts). Errors in actual AADT estimates are expected to be more than twice as high for estimates based on 24-hour counts and using proxy ATR MOY and DOW factors (rather than relying on the SPTC site's own factors, as done here, which requires "perfect [365-day] information").

Appendix D: Example Application—VDF Calibration

As described in this document, there are many planning applications that can benefit from ITS data. One example is described in this appendix: calibration of volume-delay functions (VDFs) used in trip assignment.

It is well-known that the travel time on a roadway depends on the traffic volume (or demand) on that segment. A VDF specifies the relationship between the roadway volume and average travel time needed to traverse that segment. VDFs are crucially important in trip assignment, the final step of four step planning process that assigns vehicles to the links of a network.

Numerous research studies (TRB, 2000; Bureau of Public Roads, 1964; Bertini et al., 2005)) have investigated the exact functional form of the travel time and volume dependency. The functional form of the VDF depends on the underlying traffic model used in the derivation. In this study, we derive VDFs from the simple Greenshields model, and also from the speed volume relationship specified in Highway Capacity Manual (HCM). VDFs were also obtained from calibration using detector data, and the CORSIM simulator was also used to generate travel time for different volume of traffic. CORSIM was also used to estimate the queuing delay due to traffic entering and exiting the freeway.

The volume delay function proposed by Bureau of Public Roads (BPR) is one of the commonly used volume delay function. The generic BPR volume delay function has the following function form.

$$t(v) = t_0 * \left(1 + \alpha \left(\frac{v}{c}\right)^{\beta}\right)$$

In the above equation *t(v)* denote the travel time on link with volume *v*, capacity *c* and free flow travel time t_0 . The commonly used α and β values are 0.15 and 4, respectively.

The BPR function can be derived from the fundamental Greenshields traffic flow model, offering insight into the assumption of the BPR volume delay function. Let *v* denote flow on link, *s* space mean speed and *k* is the density. s_0 is the free flow speed on the link, *c* is the capacity of the link, and k_j is the jam density. We know from "speed-volume" relationship of Greenshields model that

$$v = \frac{sk_j}{s_0} * (s_0 - s)$$

$$\frac{vs_0}{k_j} = \frac{s_0^2}{4} - \left(s - \frac{s_0}{2}\right)^2$$

$$s - \frac{s_0}{2} = \sqrt{\left(\frac{s_0^2}{4} - \frac{vs_0}{k_j}\right)}$$

$$s = \frac{s_0}{2}\left(1 + \sqrt{\left(1 - \frac{v}{\frac{k_j s_0}{4}}\right)}\right)$$

$$s = \frac{s_0}{2}\left(1 + \sqrt{\left(1 - \frac{v}{c}\right)}\right) \; with\; c = \frac{k_j s_0}{4}$$

We assume that drivers are moving with constant speed *s* and *t* is the time taken to cover the distance *x*. Let t_0 be the time taken to cover the same distance *x* at free flow speed s_0.

$$st = s_0 t_0 = x$$
$$\frac{s}{s_0} = \frac{t_0}{t} = 0.5 * \left(1 + \sqrt{1 - \frac{v}{c}}\right)$$
$$t = \frac{t_0}{0.5 * \left(1 + \sqrt{1 - \frac{v}{c}}\right)}$$

By Taylor series expansion,

$$\sqrt{1 - \frac{v}{c}} = 1 - \frac{v}{2c} - \frac{1}{2^3} * \left(\frac{v}{c}\right)^2 - \frac{1}{2^4} * \left(\frac{v}{c}\right)^3 - \frac{5}{2^7} * \left(\frac{v}{c}\right)^4 - \frac{7}{2^8} * \left(\frac{v}{c}\right)^5 - \ldots..$$
$$\sqrt{1 - \frac{v}{c}} = 1 - \left(\frac{v}{2c} + \frac{1}{2^3} * \left(\frac{v}{c}\right)^2 + \frac{1}{2^4} * \left(\frac{v}{c}\right)^3 + \frac{5}{2^7} * \left(\frac{v}{c}\right)^4 + \frac{7}{2^8} * \left(\frac{v}{c}\right)^5 + \cdots\right)$$

Based on the field observations, the above term is approximated as

$$\sqrt{1 - \frac{v}{c}} = 1 - 0.30 * \left(\frac{v}{c}\right)^4$$
$$t = \frac{t_0}{0.5 * \left(2 - 0.30 * \left(\frac{v}{c}\right)^4\right)} = \frac{t_0}{\left(1 - 0.15 * \left(\frac{v}{c}\right)^4\right)}$$

By Taylor series approximation,

$$t = t_0\left(1 + 0.15 * \left(\frac{v}{c}\right)^4\right)$$

The two main assumptions used in the above derivation are that road users travel with constant speed, and that there is no variability across drivers.

The Greenshields model assumes a simple linear relationship between speed and density. So using the flow conservation equation we get a parabolic relationship between speed and volume. However, the observation on field suggests that the speed and flow relationship is not exactly parabolic. So Highway Capacity Manual (HCM) uses a piecewise non-linear curve to describe the speed –flow relationship: (The notation described in the previous section is used in this section as well.)

For $70 \le s_0 \le 75\ mi/hr$

$$s = s_0 - \left[\left(s_0 - \frac{160}{3}\right) * \left(\frac{v + 30s_0 - 3400}{30s_0 - 1000}\right)^{2.6}\right] \quad 3400 - 30s_0 < v \le 2400\ vph$$
$$s = s_0 \quad v \le 3400 - 30s_0$$

For $55 \le s_0 \le 70\ mi/hr$

$$s = s_0 - \left[\left(\frac{7s_0 - 340}{9}\right) * \left(\frac{v + 30s_0 - 3400}{40s_0 - 1700}\right)^{2.6}\right] \quad 3400 - 30s_0 < v \leq 2400$$
$$s = s_0 \quad v \leq 3400 - 30s_0$$

The freeway section analyzed in this study has a free flow speed of 70 mi/hr. So the HCM speed volume relationship for this section can be simplified as shown below.

$$s = 70 - \left[16.67 * \left(\frac{v - 1300}{1100}\right)^{2.6}\right] = 70 - 2.06 * 10^{-7} (v - 1300)^{2.6} \quad 1300 < v \leq 2400$$
$$s = 70 \quad v \leq 1300$$

A new volume delay equation can be derived using the above speed volume relationship. The new speed volume relationship is referred to as HCM volume delay equation in this study. In this derivation too we assume that all users travel with constant speed and that there is no variability across road users.

$$\frac{s}{s_0} = \frac{t_0}{t}$$

$$\frac{t}{t_0} = \frac{70}{70 - 2.06 * 10^{-7} (v - 1300)^{2.6}} \quad 1300 < v \leq 2400$$
$$= 1 \quad v \leq 1300$$

$$t = \frac{t_0}{1 - 2.943 * 10^{-9} (v - 1300)^{2.6}} \quad 1300 < v \leq 2400$$
$$= t_0 \quad v \leq 1300$$

By using Taylor series approximation,

$$t = t_0(1 + 2.943 * 10^{-9} (v - 1300)^{2.6}) \quad 1300 < v \leq 2400$$
$$= t_0 \quad v \leq 1300$$

It can be noted that the above volume delay function is convex and hence is suited for solving optimization problems, such as traffic assignment.

Next, we compare expected travel time predicted by BPR volume delay equation and HCM volume delay equation to traverse a 10 mile-long link under different levels of congestion. The expected time to traverse the segment under free flow condition (*volume* = 0) is 8.57 minutes. The BPR function predicts a smooth and continuous increase in travel time from the free flow condition with increasing congestion. The HCM delay function, on the other hand, predicts the travel time to be close to the free flow travel time until volume exceeds 1500 vph. At that point, the travel time is predicted to increase very steeply until capacity is attained. For volumes not exceeding 800 vph, the BPR and HCM functions predict similar travel time, after which the BPR function predict higher travel times, until volumes exceed 2000 vph, at which point the HCM function predicts a higher travel time. At capacity, the BPR volume delay function predicts the travel time to be 9.9 minutes, which is 15% more than the free flow

condition. By contrast, The HCM volume delay function predicts a travel time of 10.6 minutes at capacity, which is about 24% more than free flow case.

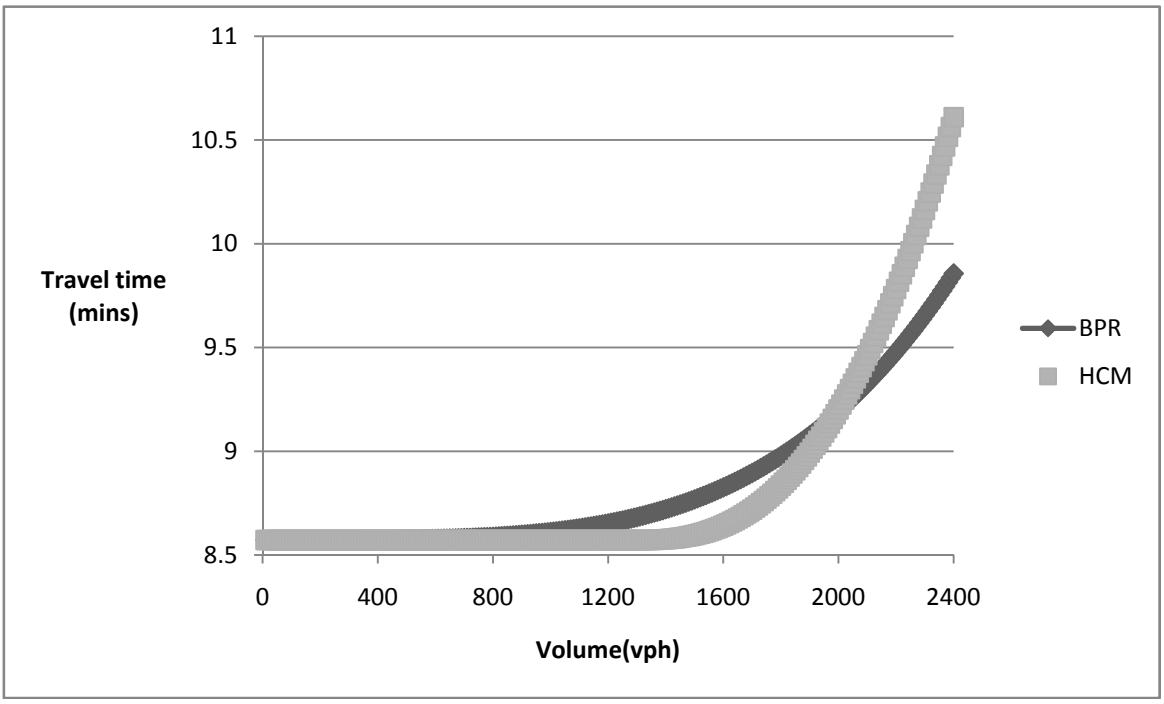


Figure D.1: BPR and HCM Volume Delay Function

Calibration of generic BPR Volume Delay Function

Calibration is the adjustment of model parameters to improve model's ability to reproduce the field observation (5). As discussed earlier, the functional form of generic volume delay function is:

$$t(v) = t_0 * \left(1 + \alpha \left(\frac{v}{c}\right)^{\beta}\right)$$

In the above equation, the parameters α and β are parameters can be determined by calibration. The calibration process typically involves solving an optimization formulation to find the model parameters that generate predictions closest to the field observation. The goodness-of-fit measure used in this study is sum of squared differences (SSD) between the model prediction and field observation. We seek to find the values of α and β that minimize the SSD between travel time predicted by BPR volume delay function, and travel time measured in the field. The desired values of α and β can be determined by solving the following optimization formulation.

$$\underset{\alpha,\beta}{Minimize} \sum_i \left(t(v_i) - t_m(v_i)\right)^2$$

$$s.t\ \ t(v_i) = t_0 * \left(1 + \alpha \left(\frac{v_i}{c}\right)^{\beta}\right)\ \ i = 1,2,3 \ldots$$

$t_m(v_i)$: Observed travel time on the field corresponding to volume v_i

The data required for calibration was obtained from an autoscope detector located close to the intersection of IH35E and Regal Road in Dallas (Figure D.2). The detector reports average speed and volume data at every 2 minutes. The time taken to cover the link at speed measured by the detector for different volumes is denoted as the measured travel time. The data reported by detector is aggregated for 16 minutes, and then used for calibration. Twenty-four-hour data from 6 PM on Monday, Jan 31st, 2007, to 6 PM on Tuesday Feb 1st, 2007, was used for calibration. The section capacity was assumed to be 2400 vph per lane.

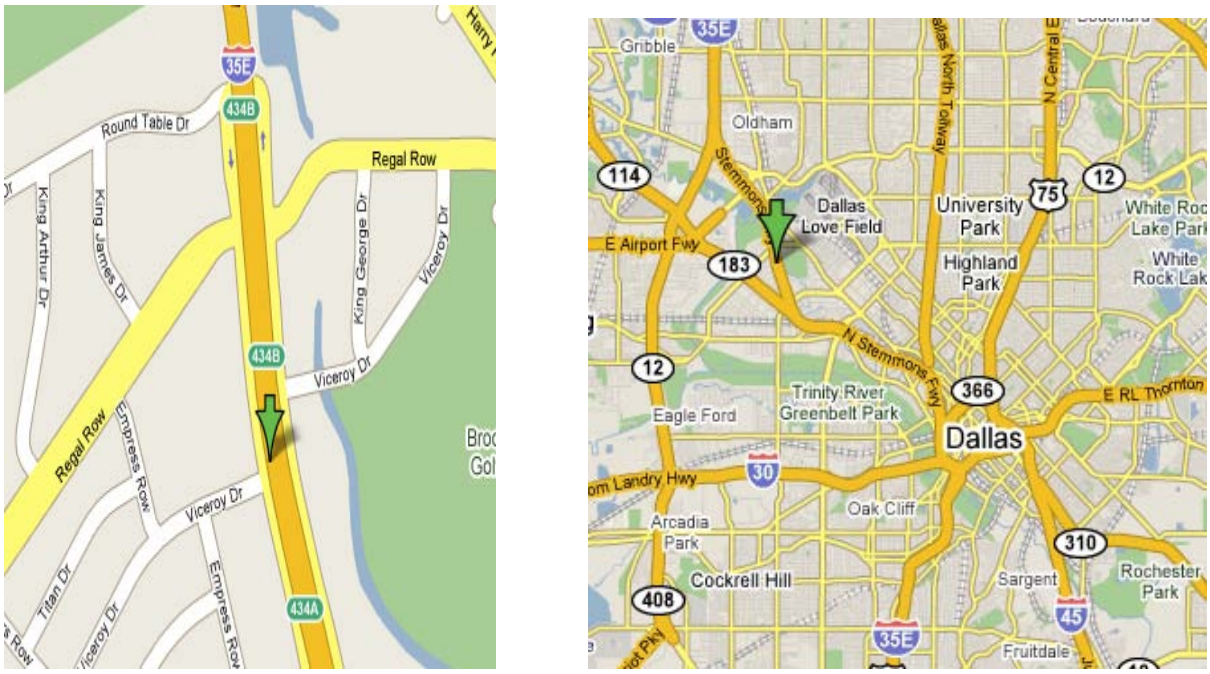


Figure D.2: Location of the detector used for calibration.

The optimization formulation was solved using an inbuilt Matlab optimization program based on the "Nelder- Mead" method. The optimal solutions obtained after solving the calibration problem are:

$$\alpha = 0.856$$
$$\beta = 0.9156$$

A comparison of travel time predicted using calibrated BPR volume delay function and measured travel time is shown in Figure D.3.

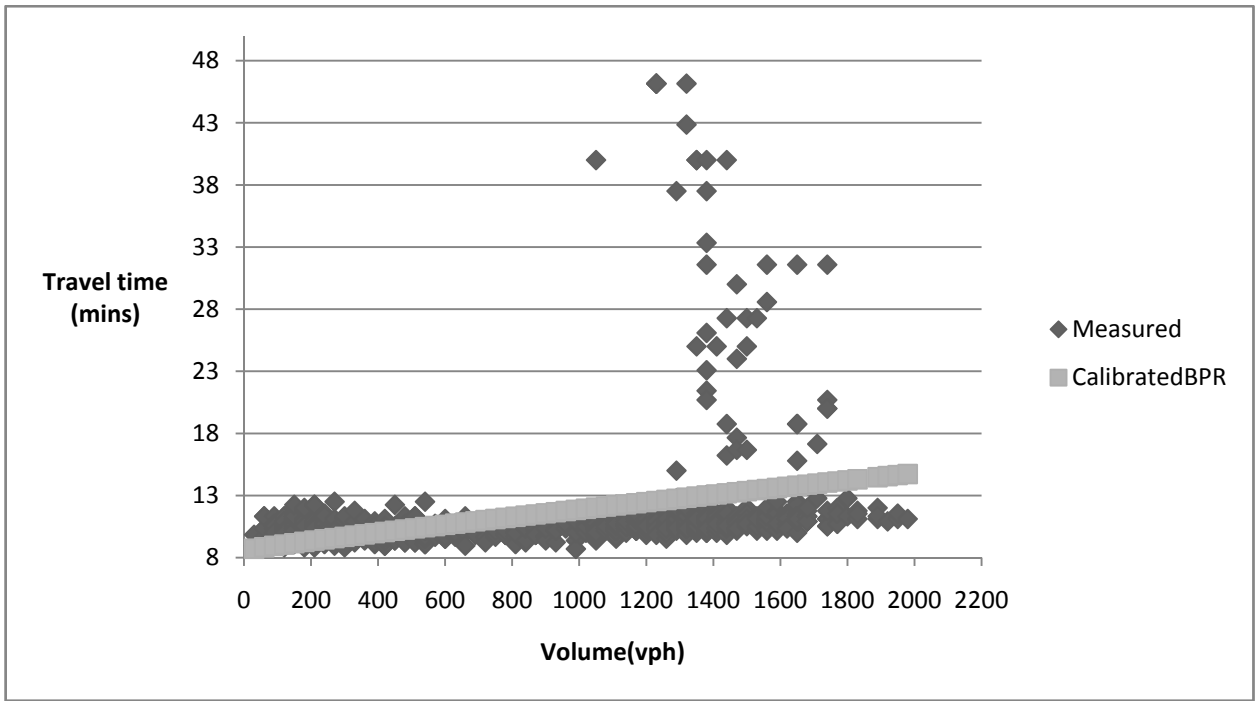


Figure D.3: Calibrated BPR and Measured travel time

The calibrated BPR volume delay function returns a unique expected travel time for a given volume. On the other hand, measured travel time incorporates the effect of variability across the drivers. At the same volume, different drivers tend to travel at different speeds and hence experience different travel time. When volume exceeds 1300 vph, we can see that some drivers are experiencing very high travel times. These conditions point to the existence of an unstable situation. If the demand exceeds the capacity of the freeway, then it would create a bottleneck and thus resulting in a start and stop condition. It must be noted that the BPR function models only an average situation and hence does not capture these extreme situations. On the whole the calibrated BPR function prediction closely matches the average measured travel time for a given volume.

Figure D.4 shows a comparison of expected travel time predictions by Calibrated BPR, BPR, and HCM volume delay function for different volumes. The expected travel time prediction of the calibrated BPR function is significantly higher than the BPR and HCM travel time predictions. The calibrated BPR function predicts a continuous and monotonic increase in travel time with increasing volume, whereas the BPR and HCM volume delay predicts the travel time to be close to free flow condition till a volume 1200 vph, and a marginal increase thereafter. At the maximum volume of 2000 vph, calibrated BPR predicts the travel time to be 70% more than the free flow condition while BPR and HCM volume delay function predicts the travel time to be only about 7% more than free flow condition.

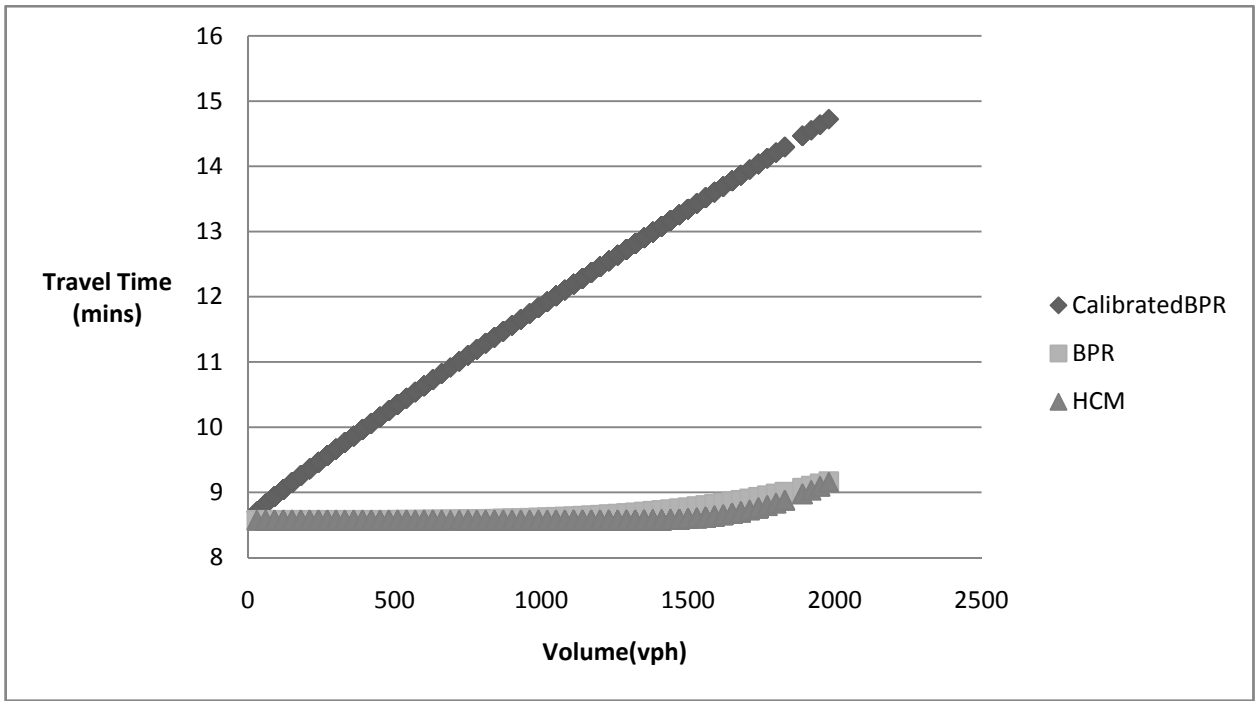


Figure D.4: Calibrated BPR, BPR, and HCM volume delay functions

CORSIM Volume Delay Function

One of the main assumptions made in the derivation of BPR and HCM volume delay function is that all road users are travelling with constant speed. In reality, drivers tend to accelerate and decelerate frequently. The constant speed assumption is relaxed by measuring the travel time using CORSIM simulator. The car following model is the underlying traffic propagation model used by the CORSIM simulator. In car following models, drivers are allowed to accelerate and decelerate while traversing the distance. CORSIM also explicitly models the variability of drivers using the roadway.

A hypothetical roadway 10 mile long was constructed in CORSIM and the time taken to travel the roadway was measured for different volume. Ten types of drivers were considered for simulation. Each driver class is characterized by a free flow speed. The average free-flow speed across all the driver class was set to 65 mph (maximum free-flow allowed in CORSIM). The travel time variation with volume predicted by BPR and HCM for this hypothetical link is also computed. The CORSIM, BPR and HCM volume delay function for the roadway considered is shown in Figure D.5.

The average travel time calculated by CORSIM simulation is more than the travel time predicted by BPR and HCM volume delay functions for all possible flows on the link. CORSIM simulations predict that the travel time increases almost uniformly with increasing volume. BPR and HCM volume delay function predict the travel time to be close to free-flow case till a volume of 1200 vph and from the volume of 1200 vph to 2400vph, travel time increases very steeply with increasing volume. It is interesting to note that at capacity of 2400 vph, the average travel time obtained from CORSIM simulation is close to expected travel time predicted by HCM volume delay function.

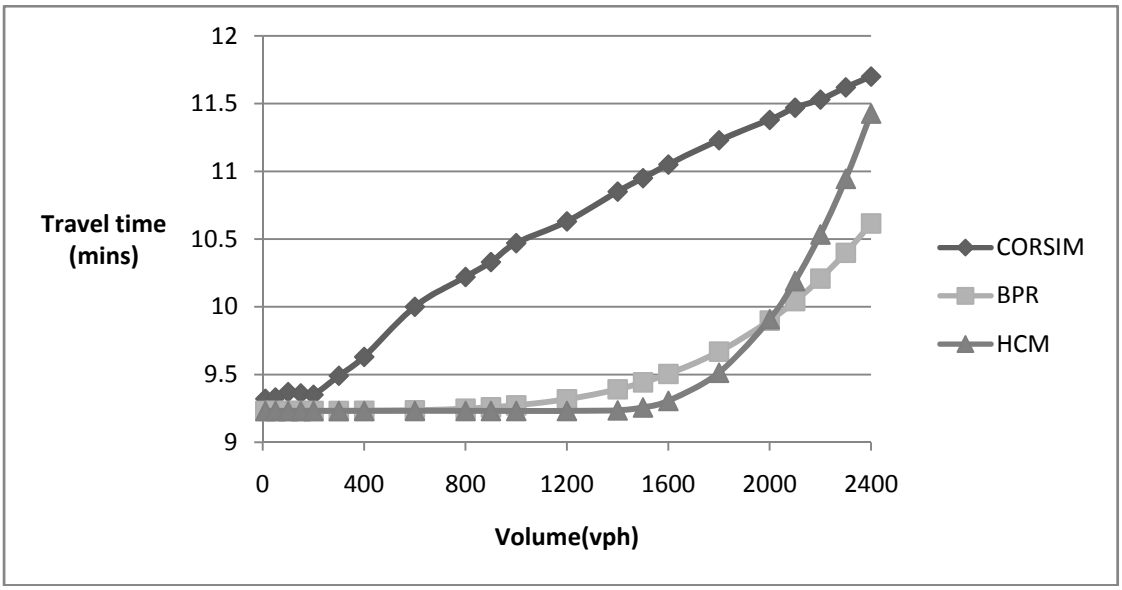


Figure D.5: Comparison of CORSIM, BPR and HCM Volume Delay Function

Queuing Delays

The entry and exit of traffic through ramps induce delay on the free flow of traffic on the freeways. If large volume of traffic is trying to enter or exit the freeway then it could result in formation of queues. The queuing delay is a function of the through traffic and volume of entering or exiting traffic. The CORSIM simulator is used to calculate the queuing delay for different volumes of through and entering or exiting traffic. Figure D.6 plots the queuing delay for different combinations of through volume and entry volume.

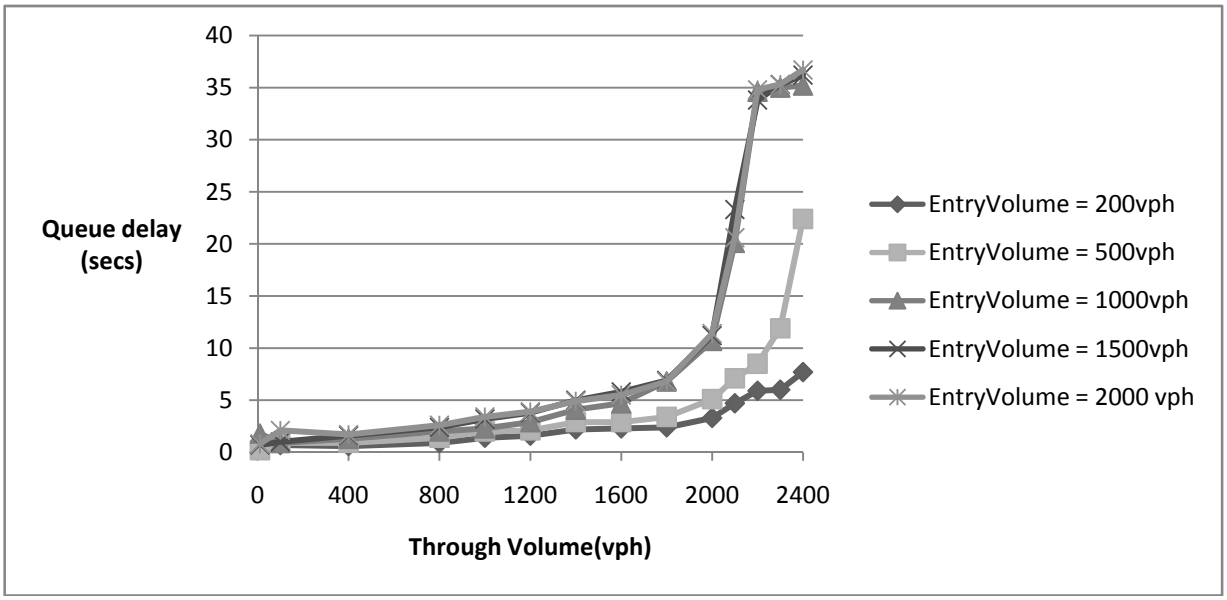


Figure D.6: Queuing delay for different entry volume at entry ramps

It can be seen that the queuing delay increases with increasing through volume for a given constant volume of traffic entering through the ramps. The magnitude of increase in queuing delay becomes steeper as the through volume increases. The queuing delay also increases with increasing entry volumes. There is significant increase in queuing delay as entry volume increases from 200 vph to 1000 vph. However, CORSIM predicts only small increase in queuing delay as the entry volume increases from 1000 vph to 2000 vph.

The queuing delay for different combination of through volume and percentage of through volume traffic exiting the freeway is depicted in Figure D.7.

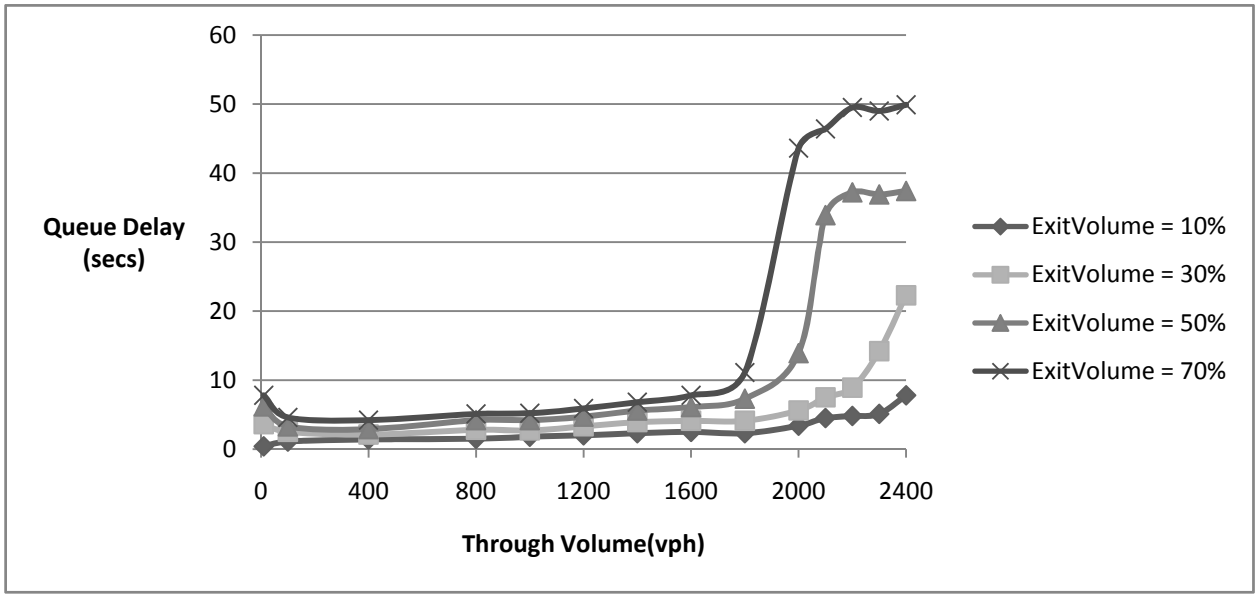


Figure D.7: Queuing delay for different exit ramp volumes

As can be seen from the figure, the queuing delay increases with increasing through volume of traffic for a given exit volume percentage. The slope of the queuing delay increases with increasing through volume for a given percentage of exit volume. The queuing volume also clearly increases with increasing percentage of traffic exiting the freeway. At capacity of 2400 vph, the queuing delay is 10 seconds when only 10% of through traffic is exiting, while the delay is about 50 seconds when 70 % of through traffic exits.

Conclusion

The VDF is used to calculate expected travel time of a roadway and is commonly used in trip assignment phase of the four step transportation planning process. The widely used BPR volume delay function is derived from the fundamental Greenshields model to offer insights into the assumption used in the derivation. A new HCM volume delay function is obtained from the speed-volume relationship specified in the HCM User Manual. The BPR volume delay function was found to predict higher travel time for low volume of traffic while the HCM volume delay function predicts higher travel time for high volume of traffic.

The generic BPR contain parameters that can be determined by calibration. BPR volume delay function was calibrated using the data obtained from a detector located close to the intersection of I35E and regal road in Dallas. Calibration process involves solving an optimization formulation to find the BPR parameters so that BPR travel time prediction closely

matches the detector data. The calibrated BPR travel time predictions were found to be significantly higher than the travel time predictions of commonly used BPR and HCM volume delay functions.

Traversing the roadway with constant speed and lack of variability across drivers are the two main assumptions used in BPR and HCM volume delay functions. These assumptions are relaxed by calculating the time taken to traverse the roadway using CORSIM. CORSIM simulation is based on car following models where drivers are allowed to accelerate and decelerate. CORSIM simulations predict the travel time to increase uniformly with increasing volume. On the other hand, BPR and HCM volume delay function predict the travel time to be the same as free flow condition till a certain threshold volume and the travel time increases rapidly when the volume exceeds this threshold volume.

Large volume of entering and exiting traffic could result in the formation of queues in freeway. CORSIM simulator is used to calculate the queuing delay for different combination of through traffic and entering or exiting traffic on a freeway. As expected, the queuing delay was found to increase with increasing through volume. The slope of queuing delay was not uniform but was also increasing with increasing through volume. An increase in the entering\exiting volume was also found to increase the queuing delay.

Appendix E: Data Reduction

In Chapter 2, several obstacles to ITS data archiving were noted, including the management of very large amounts of data. In this section we will present a promising technique to reduce the amount of data to be stored, while at the same time we can recover—to a certain accuracy—the data discarded. Furthermore, this technique also enables us to make judicious statements about the level of aggregation. Numerical experiments illustrate its effectiveness. This routine is not considered a key part of the prototype system because storage space is not likely to be a concern; however, the findings are of sufficient interest to report in this appendix.

Motivation

Loop detectors typically collect data almost continuously. Hence it is quite likely that subsequent observations are closely related to each other, which leads to the hypothesis that it is not really necessary to store all the data recorded by the detectors. This is especially true from the perspective of transportation planners. However, at the same time, it is also imaginable that the operations division does want to have access to the highest resolution data. Therefore, we have a trade-off to make about the level of aggregation. The choice of aggregation level has already received some attention in literature (see, for instance, Turner et al., 1999). However, to the best of our knowledge, no attempt has been made to recover the discarded data in the process of aggregation. Next we will present a model that both makes statements about the level of aggregation and possesses the ability to recover discarded data to a certain extent. The crucial observation that underlies this model is that subsequent recorded data values are somehow dependent.

The Proposed Method

In order to illustrate the proposed approach, we consider the DalTrans data on May 9-10, 2007[1]. A plot of the 5-minutes speed data used is given in Figure E.1. As we have noticed above, subsequent data values might show a certain dependence structure. However, the modeling of dependence is not straightforward. Instead we will examine a weaker and simpler type of dependence, i.e. correlation. Loosely speaking, correlation is a measure of a linear relationship between subsequent random observations. Mathematically, correlation is defined as follows. Given the random variables X and Y, with expectations EX and EY, and standard deviations σ_X and σ_Y, respectively, the correlation (coefficient) of X and Y (ρ_{XY}) is given by

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_y},$$

where

$$Cov(X,Y) = E\big((X - EX)(Y - EY)\big)$$

denotes the covariance of X and Y.

[1] http://ttidallas.tamu.edu/detectordataarchive/archive/DalTrans/

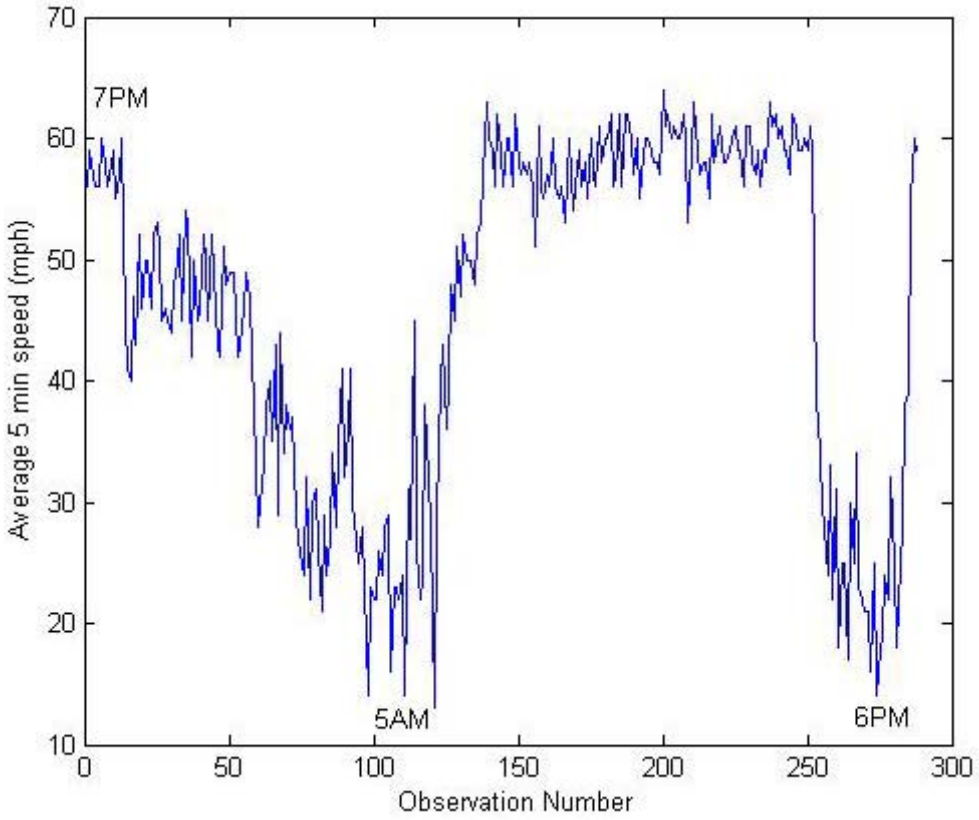


Figure E.1: Speed data for sensor with ID 10043 1084

One of the fundamental properties of the correlation coefficient is that $|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $Pr(Y=aX+b) =1$. Furthermore, if $\rho_{XY} = 1$, then $a > 0$, and if $a < 0$, then $a < 0$. A more detailed discussion on correlation coefficients can be found in Casella and Berger (2002).

Since we are dealing with time series, we will change notation accordingly. Instead of X and Y, we will use the indexed collection of random variables $\{X_t\}$, which simply denotes the following time series of length N: $X_1, X_2, \ldots, X_{t-1}, X_t, \ldots, X_N$. When correlations are computed within observations from a single time series, it is customary to refer to correlation as autocorrelation since correlations are computed with the other values in the same time series. Furthermore, autocorrelations can be computed at different lags. For instance, if autocorrelations are computed for X_{t-1} and X_t, then we have lag 1 values. On the other hand, if we use X_{t-10} and X_t, then we will compute lag 10 autocorrelations values. One can imagine that in such a way, an autocorrelation function arises, where the independent variable is the time lag. When we estimate the autocorrelation function, we call the resulting autocorrelation function the sample autocorrelation function. More details on this estimation procedure can be found in Brockwell and Davis (2004).

The sample autocorrelation for our speed data is given in Figure E.2. As can be seen from the figure, the autocorrelation is the highest at lag 1 (0.92), i.e. the linear relationship is strongest between values $X_{t-1}$and X_t. Therefore, we will investigate the following. Since subsequent speed observations are related in a linear fashion (at least, to some degree), one might consider not storing every other speed measurement in the central database. Moreover, after estimating the

linear relationship between $X_{t-1}$and X_t, we will be able to predict X_t once we know X_{t-1}. More specifically, in our case we will choose to discard the even numbered observations (thereby reducing the storage cost by one half). The linear functional relationship between $X_{t-1}$and X_t is obtained via the least squares method, which resulted in

$$X_t = 0.92X_{t-1} + 3.57$$

where *t* is an even number greater than or equal to 2. These steps are pictorially summarized in Figure E.3, where we have shown a lagplot (a plot of X_{t-1} versus X_t) and the estimated linear relationship.

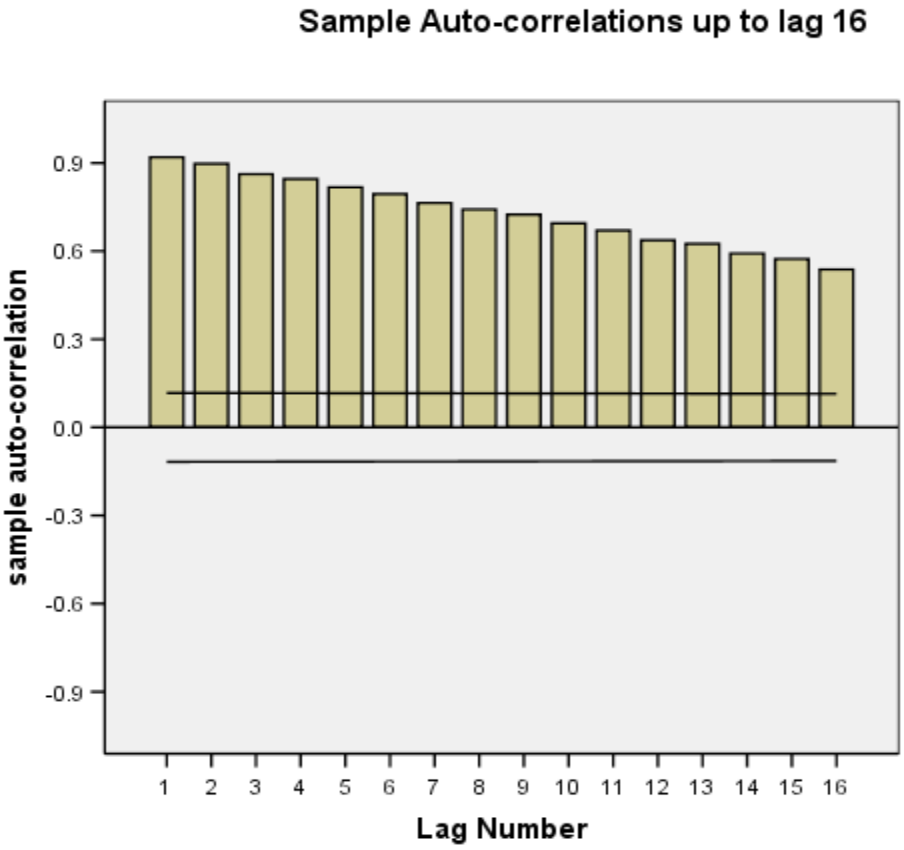


Figure E.2: Sample auto-correlation function for speed data

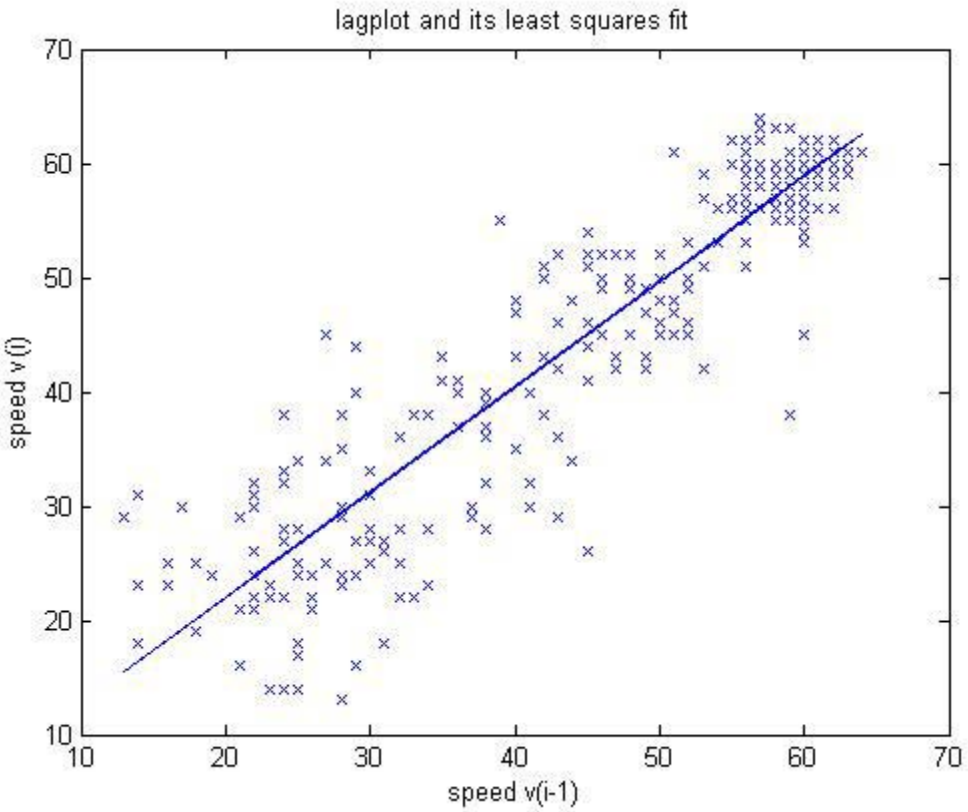


Figure E.3: Lagplot and the least squares line

In Figure E.4, we have shown the result of the recovery operation: based on the odd numbered observations and the above estimated linear relationship, we predict the even numbered observations. The resulting mean absolute error (MAE) between the predicted and observed values was found to be 3.9 mph. Further, suppose we are interested in 15-minutes averages as is often the case in planning applications. There are two ways obtaining these. The first is based on the 5 minutes speed data, the second is based on the predicted data. Figure E.5 depicts the errors (defined as observed 15 minutes averages minus predicted 15 minutes averages). The MAE in this case is 1.66 mph.

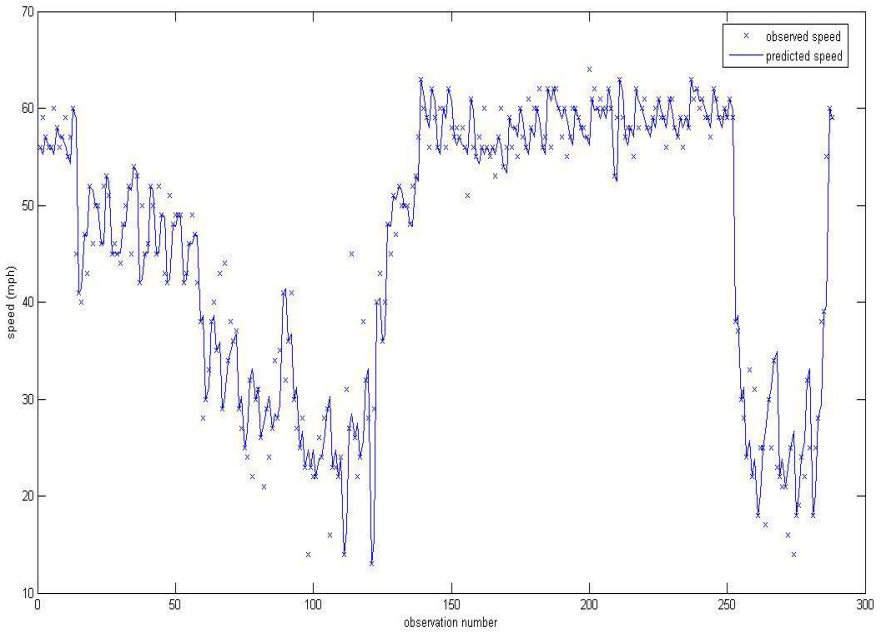


Figure E.4: Predicted and observed speeds

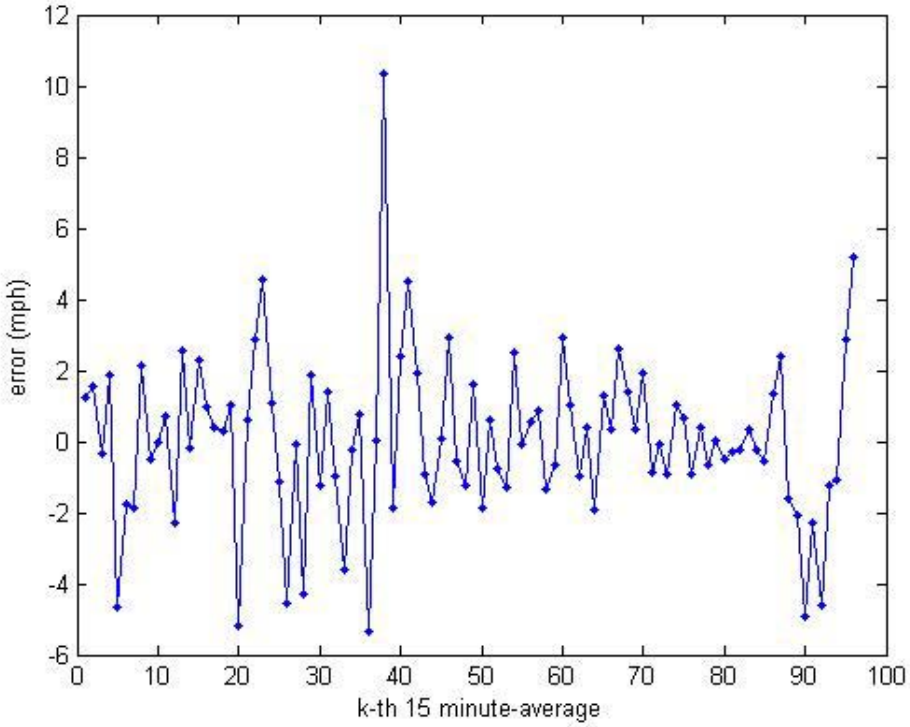


Figure E.5: Errors when calculating 15-minute averages

Conclusions

In this section we have developed a new technique to determine the data aggregation level. The distinguishing feature is that it is able to recover data that has been discarded, at least to a certain degree. Here we illustrated the method via data that is reported every 5 minutes, which are aggregates themselves. It is conjectured that the proposed method gives smaller errors when higher resolution data is used, e.g. data recorded every 20 seconds.

Recall that the objective of this technique is to facilitate data sharing. As we have seen, many data sharing architectures store both the highest resolution data as well as 5 minutes aggregates, or they store only the aggregated data. The proposed technique opens a way for intermediate storage: There is no need to store all recorded data since we are able to recover the discarded data. With a judiciously chosen aggregation level, there is less need to store averages since it is easier to compute these when needed.

Appendix F: Training Workshop

A workshop has been developed in order to disseminate the findings of this project. Including breaks and time for questions, this workshop is intended to require three to four hours of time. The material is divided into seven modules, each focusing on one aspect of the research:

Module 1: Introduction
Module 2: Basic Features
Module 3: Implementation Plan
Module 4: Measuring Data Reliability
Module 5: Corridor-Level Imputation
Module 6: Regional Imputation
Module 7: Conclusion

This appendix includes the slides developed for this workshop; these may be reproduced as handouts for workshop attendees.

Utilizing the Data Collected at Traffic Management Centers for Planning Purposes through Non-Traditional Sources and Improved Equipment

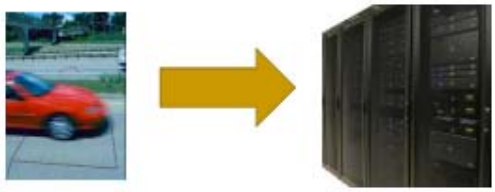
Workshop prepared for Texas Department of Transportation
0-5686 Deliverable P3

MODULE 1

Introduction

Introduction to ITS Data

Vast amounts of data are automatically recorded for traffic operations. Can this data be used by others as well?



Basic Detector Types

Traffic data is collected in many ways:

- Inductive loops
- Video detection
- Wireless location technologies
- Laser detection
- Infrared technology
- Radar and acoustic traffic sensors
- Weigh-in-motion
- Wireless magnetic technology
- Intelligent road studs
- Aerial image analysis

Different locations use different technologies to collect data: how can one plan for all of these options?

TMC Operations

Traffic management centers (TMCs) use this data for real-time congestion monitoring, incident management, and operational studies.

Example Applications

How else might this data be useful?

- Annual average daily traffic (AADT) counts
- Automatic incident location
- Before-and-after policy evaluation
- Calibration of delay functions for planning
- Calibration of travel demand models
- Signal timing and warrant analysis

Any application which can benefit from large amounts of traffic data can gain from a common interface.

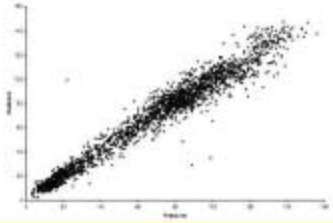
Specific Opportunities for Planning

Duplication of effort can be minimized – if ITS data is available at a location, manual or tube counts may not be necessary.

This saves time, money, and eliminates safety issues of placing tube counts in busy areas.

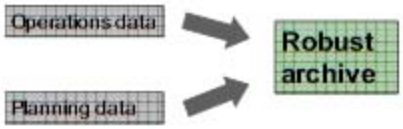
Specific Opportunities for Planning

If data is missing at a specific location, nearby ITS sensors offer redundancy and enable statistical inference.



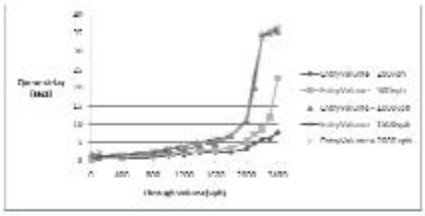
Specific Opportunities for Planning

Including *both* operations and planning traffic data provides redundancy and greater coverage area.



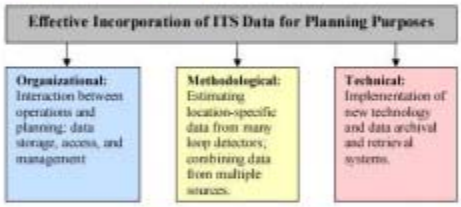
Specific Opportunities for Planning

Parameters in volume-delay functions can be better calibrated using the large amounts of ITS data



Three Major Considerations

To effectively share data, **organizational**, **methodological**, and **technical** concerns must all be addressed.



Organizational

- How should we store data?
- What data should we store?
- How do users access this data?
- Who controls the data?
- Who is responsible for maintaining this data?
- How should access be granted?

Methodological

- Quality control issues
  - What is the data being used for?
  - Can we infer data for areas not covered?
- Data fusion
  - Combining data from multiple sources
  - Innovative technologies
- Can planning models be enhanced by using operations-specific data?

Technical

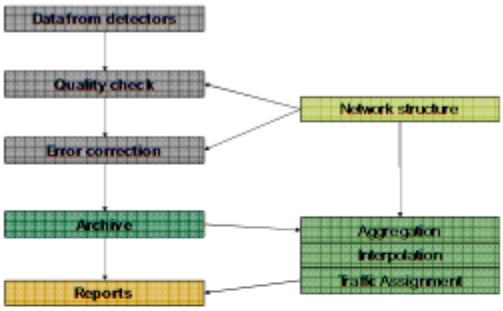
- How can innovative ITS technologies be introduced to address both operations and planning needs?
- How do these technologies function as part of an integrated system?

A Unified Data Archive System

The remainder of this workshop describes an archive system which

- Can use data from all types of detectors;
- Can be accessed by many users;
- Modular (more sources can be added later);
- Quantifies reliability of observed data; and
- Allows several imputation methods for missing data

A Unified Data Archive System



Workshop Structure

The remaining workshop modules discuss elements of the unified data archiving system in detail:

- **Module 2**: Basic features
- **Module 3**: Implementation plan
- **Module 4**: Measuring data reliability
- **Module 5**: Corridor-level data imputation
- **Module 6**: Regional-level data imputation
- **Module 7**: Conclusion

END OF MODULE 1

Questions?

MODULE 2

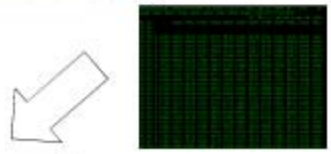
Basic Features

Data Storage

Detector data is first transmitted to TMCs, and then sent to the central archive

Data Storage

Next, the archive converts incoming data into a common format:



$$(x, t, \mathbf{s}, RI)$$

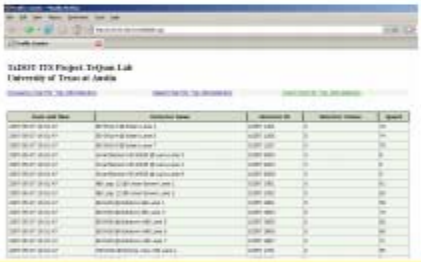
Location — Time Interval — State: Volume, speed, etc. — Reliability index (calculated later)

Data Storage

Once converted into the general format, data is stored in a PostgreSQL database, where it can be accessed from other locations:

Data Storage

Data can also be accessed using a web interface to generate reports:



Data Storage

PostgreSQL is very flexible, and custom interfaces and reports can be generated for particular applications.

END OF MODULE 2

Questions?

MODULE 3

Implementation Plan

Implementation Plan

Such an archive can be smoothly implemented in three phases:

Phase I: Establish policies and standards
Phase II: Implement central archive
Phase III: Connect individual TMCs to archive

Phase III is repeated for each TMC that is connected. This allows flexibility in connecting additional TMCs as a later date.

Phase I: Establish Policies and Standards

The goal of Phase I is to determine the scope of the archive, the control hierarchy, and the infrastructure requirements:

Task 1.1: Establish desired functionality
Task 1.2: Determine equipment needs
Task 1.3: Determine leadership roles
Task 1.4: Identify physical location(s)

Task 1.1: Establish Desired Functionality

What functions does the data archive need to perform?

- Which TMCs will provide data?
- Who will be allowed to access the archive?
- What data must be collected at a minimum?
- What reporting frequency will be used?

Task 1.2: Determine Equipment Needs

Based on desired functionality, what communication and equipment are needed?

- How much bandwidth is needed?
- Wired or wireless communication?
- What hardware is required for the archive?
- Backup and redundancy options?

Task 1.3: Determine Leadership Roles

Responsibility must be assigned for different roles regarding the archive.

- Who determines which TMCs participate?
- Who is responsible for archive maintenance?
- Who will provide space for the hardware?
- Who is the operational point of contact?

Task 1.4: Identify Physical Location

A suitable location must be chosen for the archive itself, along with any off-site space needed for backup and redundancy.

- Location depends on communication needs and leadership roles
- Backup systems may require additional space

Phase II: Implement Central Archive

The goal of Phase II is to make the central database operational:

Task 2.1: Install needed equipment
Task 2.2: Implement database and interface
Task 2.3: Enable remote access

Task 2.1: Install Needed Equipment

As determined in Phase I, the equipment and infrastructure needed for the central archive must be implemented and tested:

- Computing hardware
- Data storage (including any off-site backup)
- Communications infrastructure

Task 2.2: Implement Database

The PostgreSQL database must be installed and initialized, and additional routines must be written and integrated into the system:

- Reliability assessment algorithms
- Custom report generation procedures
- Imputation or data correction procedures

Task 2.3: Enable Remote Access

Once the database is operational, communications infrastructure can be brought online.

Phase III: Integrate TMCs

The goal of Phase III is to connect TMCs to the central archive. This phase is repeated for each TMC, and whenever new TMCs are added.

Task 3.1: Define needed parameters
Task 3.2: Implement conversion procedures
Task 3.3: Establish communication link

Task 3.1: Define Needed Parameters

For each TMC, the archive needs additional data to perform the reliability and correction procedures:

- Detector types, IDs, and locations
- Facility capacity and jam density
- Identification of upstream and downstream detectors

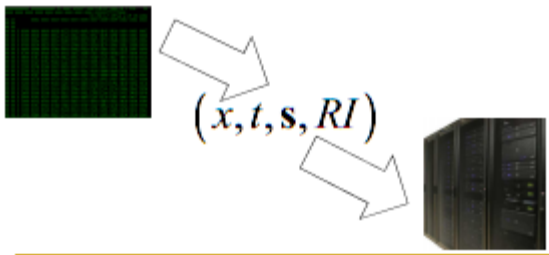
Task 3.2: Implement Conversion Procedures

Procedures must be coded to convert the detector data format into the standard form:

$$(x, t, \mathbf{s}, RI)$$

Task 3.3: Establish Communications Link

At this point, the TMC can be connected to the central archive.



END OF MODULE 3

Questions?

MODULE 4

Measuring Data Reliability

Quality Control

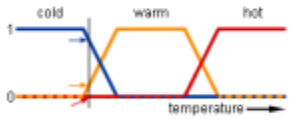
Error checking is accomplished by looking at three different types of "reality checks":

- **Fundamental**: Is the state physically possible?
- **Network**: Is the state consistent with nearby locations?
- **Historical**: Is the state reasonable, given past data?

A single **reliability index** (0-10) is used to capture the overall confidence in the data.

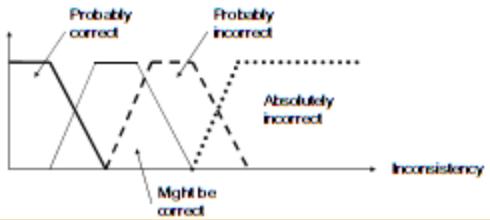
CST Classifiers

Continuous set theory (CST) is used to rigorously combine assessments that are **qualitatively different** and **inherently imprecise**.



CST Classifiers

Continuous set theory (CST) is used to rigorously combine assessments that are **qualitatively different** and **inherently imprecise**.



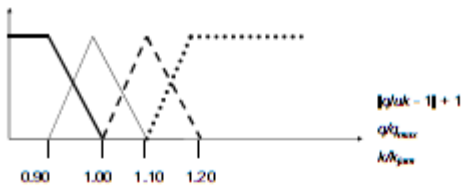
Fundamental Consistency

Is the data consistent with basic traffic relationships?

- Volume = Speed x Density
- Volume cannot exceed facility capacity
- Density cannot exceed jam density

Fundamental Consistency

For each of these relations, we have defined suitable continuous intervals:



Fundamental Consistency

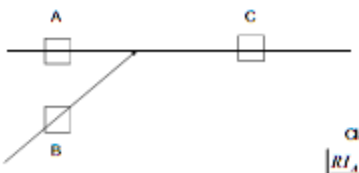
Example: Detector reads volume of 1800 veh/hr, density of 50 veh/mi, speed of 38 mi/hr

- $|q/uk - 1| + 1 = 1.05$ → 0.5 maybe correct, 0.5 probably incorrect
- $q/q_{max} < 0.9$ → 1.0 probably correct
- $k/k_{jam} < 0.9$ → 1.0 probably correct

The overall "fundamental consistency" is the **least** reliable of these: 0.5 MC, 0.5 PI

Network Consistency

Data for upstream and downstream detectors should respect flow conservation:



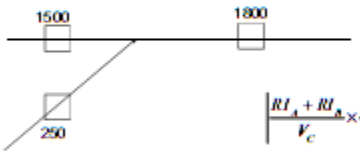
Volume at detector C should be the sum of detectors A and B

Classify based on:

$$\left| \frac{RI_A + RI_B}{V_C} \times \frac{V_A + V_B - V_C}{2\,RI_{max}} \right|$$

Network Consistency

Example: C records 1800 veh/hr, A records 1500 veh/hr with RI of 8, B records 250 veh/hr with RI of 5

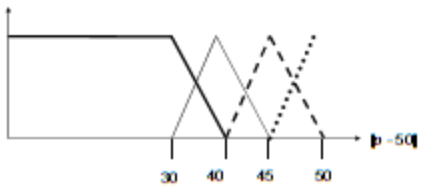


$$\left| \frac{RI_A + RI_B}{V_C} \times \frac{V_A + V_B - V_C}{2\,RI_{max}} \right| = 0.04$$

0.6 Probably Correct, 0.4 Maybe Correct

Historical Consistency

Is the data consistent with past measurements at this site?



Percentile should be measured with respect to comparable data in the past (same day of week, time of day, etc.)

Historical Consistency

Example: Detector volume of 1800 veh/hr is the 70th percentile of comparable times.

|70 – 50| = 20 → 1.0 Probably Correct

Reliability Resolution

How do we combine these separate assessments together?

Fundamental: 0.5 maybe correct, 0.5 probably incorrect
Network: 0.6 probably correct, 0.4 maybe correct
Historical: 1.0 probably correct

Reliability Resolution involves a set of decision rules that map back to "crisp" quantities (RIs)

Reliability Resolution

First, a set of 64 "aggregate states" is defined based on each possible combination of fundamental, network, and historical reliability

For example, these are all the states corresponding to data which is "probably correct" according to the "fundamental" criterion:"

Network ↓ Historical →	Probably correct	Maybe correct	Probably incorrect	Absolutely incorrect
Probably correct	(F_{PC}, N_{PC}, H_{PC})	(F_{PC}, N_{PC}, H_{MC})	(F_{PC}, N_{PC}, H_{PI})	(F_{PC}, N_{PC}, H_{AI})
Maybe correct	(F_{PC}, N_{MC}, H_{PC})	(F_{PC}, N_{MC}, H_{MC})	(F_{PC}, N_{MC}, H_{PI})	(F_{PC}, N_{MC}, H_{AI})
Probably incorrect	(F_{PC}, N_{PI}, H_{PC})	(F_{PC}, N_{PI}, H_{MC})	(F_{PC}, N_{PI}, H_{PI})	(F_{PC}, N_{PI}, H_{AI})
Absolutely incorrect	(F_{PC}, N_{AI}, H_{PC})	(F_{PC}, N_{AI}, H_{MC})	(F_{PC}, N_{AI}, H_{PI})	(F_{PC}, N_{AI}, H_{AI})

Reliability Resolution

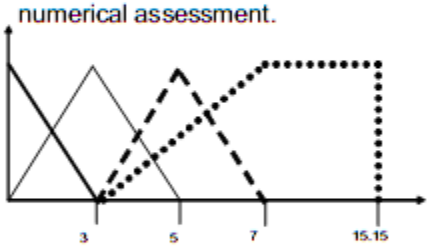
Next, a decision table converts each "aggregate state" into an assessment about the overall reliability:

For example, this is the decison table corresponding to data which is "probably correct" according to the "fundamental" criterion:"

E_{PC}	H_{PC}	H_{MC}	H_{PI}	H_{AI}
N_{PC}	PC	PC	PC	PC
N_{MC}	PC	PC	MC	MC
N_{PI}	MC	MC	MC	PI
N_{AI}	MC	MC	PI	AI

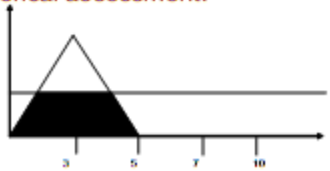
Reliability Resolution

Next, another continuous classifier is used to translate each of these decisions into a numerical assessment.



Reliability Resolution

Next, another continuous classifier is used to translate each of these decisions into a numerical assessment.



Typically, this number is the geometric centroid of the area below the relevant continuous shape and the horizontal membership line

Reliability Resolution

Finally, a weighted average of these centroids determines the final reliability index.

State	μ	Decision	Area	Centroid
$\mu(F_{PC}, N_{MC}, H_{PC})$	0.1	PC	0.49	9.174
$\mu(F_{PC}, N_{MC}, H_{MC})$	0.1	PC	0.49	9.174
$\mu(F_{PC}, N_{MC}, H_{PI})$	0.1	MC	0.38	5
$\mu(F_{PC}, N_{PI}, H_{PC})$	0.1	MC	0.38	5
$\mu(F_{PC}, N_{PI}, H_{MC})$	0.1	MC	0.38	5
$\mu(F_{PC}, N_{PI}, H_{PI})$	0.1	MC	0.38	5
$\mu(F_{MC}, N_{MC}, H_{PC})$	0.5	PC	2.25	9.56
$\mu(F_{MC}, N_{MC}, H_{MC})$	0.8	MC	1.92	5
$\mu(F_{MC}, N_{MC}, H_{PI})$	0.333	PI	1.111	3
$\mu(F_{MC}, N_{PI}, H_{PC})$	0.2	MC	1.92	5
$\mu(F_{MC}, N_{PI}, H_{MC})$	0.2	MC	1.92	5
$\mu(F_{MC}, N_{PI}, H_{PI})$	0.2	PI	0.72	3

RI = 5.9

END OF MODULE 4

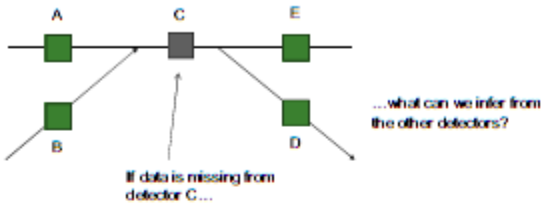
Questions?

MODULE 5

Corridor-Level Imputation

Introduction

Corridor-level imputation is used to estimate missing (or suspicious) data when other, nearby detectors are available.



Introduction

This is a significant problem which must be addressed.

It is common for 25-30% of data to be "missing" or "suspicious"

Imputed data must be flagged, since they are not suitable for all applications.

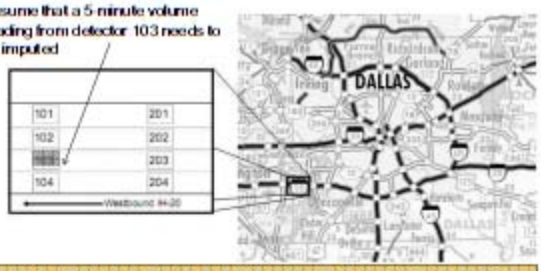
Introduction

Many techniques exist to impute data:

- Simple linear regression
- Multiple linear regression
- Local and global regression
- Historical imputation

Example Application

Assume that a 5-minute volume reading from detector 103 needs to be imputed



Each of these methods is evaluated using real data from the Dallas region. Delete 10% of volume readings, compare imputed values with original observations.

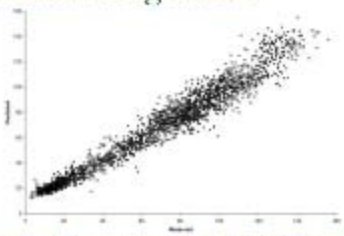
Simple Linear Regression

Linear relationships are found between detector 103 and all nearby detectors, using past data.

The average of these seven estimates is then used.

$$\hat{v}_1^{103} = \beta_0^{101} + \beta_1^{101} v^{101}$$
$$\hat{v}_2^{103} = \beta_0^{102} + \beta_1^{102} v^{102}$$
$$\hat{v}_3^{103} = \beta_0^{104} + \beta_1^{104} v^{104}$$
$$\hat{v}_4^{103} = \beta_0^{201} + \beta_1^{201} v^{201}$$
$$\hat{v}_5^{103} = \beta_0^{202} + \beta_1^{202} v^{202}$$
$$\hat{v}_6^{103} = \beta_0^{203} + \beta_1^{203} v^{203}$$
$$\hat{v}_7^{103} = \beta_0^{204} + \beta_1^{204} v^{204}$$

Simple Linear Regression



Success rate	2457/2457 (100%)
R^2	0.9508
Root mean square error	8.33
Bias	+0.08

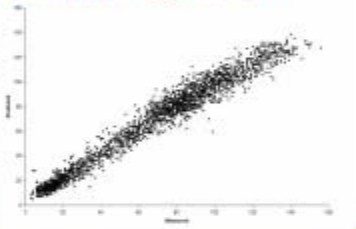
Multiple Linear Regression

Regressions can also account for speed and occupancy data, and quadratic relations:

$$\hat{v}_i^{103} = \beta_0^i + \beta_1^i v^i + \beta_2^i o^i + \beta_3^i s^i + \beta_4^i (v^i)^2 + \beta_5^i (o^i)^2 + \beta_6^i (s^i)^2 + \beta_7^i v^i o^i + \beta_8^i v^i s^i + \beta_9^i o^i s^i$$

As before, the average of these seven estimates is used.

Multiple Linear Regression

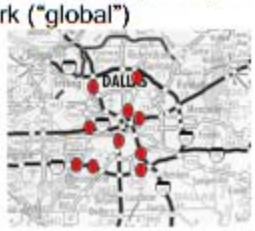


Success rate	2457/2457 (100%)
R^2	0.9625
Root mean square error	7.29
Bias	+0.06

Local and Global Regression

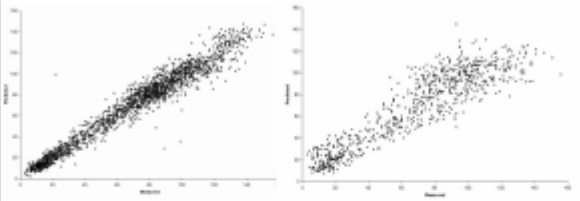
Regressions can be based on multiple (~10) detectors from the same corridor ("local") or throughout the network ("global")

Local



Global

Local and Global Regression



	Local	Global
Success rate	2150/2457 (88%)	755/2457 (31%)
R^2	0.9574	0.8392
Root mean square error	7.77	15.0
Bias	+0.06	+0.58

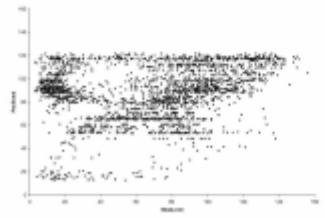
Local and Global Regression

Although local regression generally performs better than global regression, it is less robust to localized power and communication failures.

	Local	Global
Success rate	2150/2457 (88%)	755/2457 (31%)
R^2	0.9574	0.8392
Root mean square error	7.77	15.0
Bias	+0.06	+0.58

Historical Imputation

If regression models are unavailable, imputation based on recent and/or historical data at the same detector can be used.



However, this **severely** reduces accuracy (RMSE ≈ 45)

Conclusion

- Regression models closely fit volume counts
- Regression models can have high data requirements
- Local regression is more accurate than global
- Global regression is robust to certain types of failures
- Historical imputation can almost always be used, but is far less accurate

END OF MODULE 5

Questions?

MODULE 6

Regional Imputation

Regional Imputation

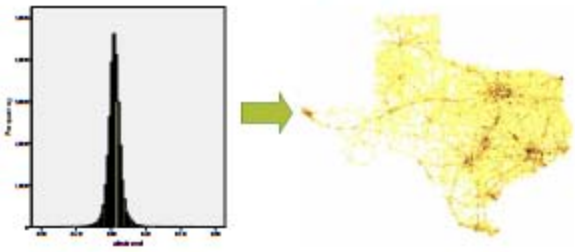
How can missing values be estimated if there are **no** other nearby detectors?

Three options are:

- Historical factoring
- Spatial kriging
- Calibration to traffic assignment

Historical Scaling

Historical data can be scaled using endogenous factors:

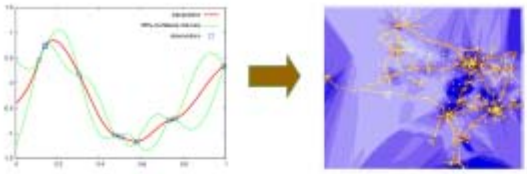


Historical Scaling

Advantages	Disadvantages
Easy to apply	Not very accurate
Requires no other detectors	Subject to sampling bias
Imputed data are locally feasible	Requires past data at a location

Spatial Kriging

An alternative method is kriging, which is more suitable for estimating in areas with little data present:



Kriging is a statistical technique for estimating values of a continuous function, given only a few data points

Spatial Kriging

Advantages	Disadvantages
Very broad scope	Boundary effects can be problematic
Does not create outliers	No network structure

Calibration to Assignment

An alternative method is to use the results of a traffic assignment to estimate volumes:

$$V_a = x_a \left(\sum_{(a,b)\in r} d_a^w \frac{V_w}{x_w} \Big/ \sum_{(a,b)\in r} d_a^w \right)$$

This is most useful in urban areas with calibrated planning models.

Calibration to Assignment

Advantages	Disadvantages
More accurate	Requires calibrated assignment model
Spatial consistency	No guarantee of fundamental consistency

END OF MODULE 6

Questions?

MODULE 7

Conclusion

Recap: Module 1

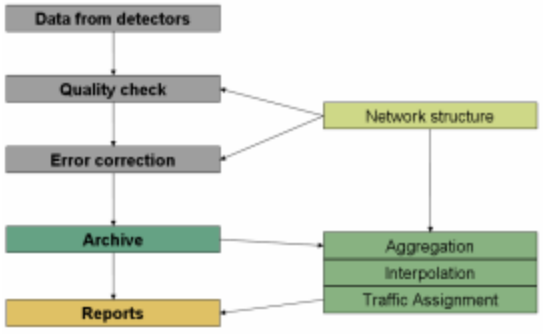
Given the extent of ITS data, it is highly useful to develop a centralized system to archive this data and allow others to access it.

As shown in this workshop, it is possible to design a system which is flexible, modular, and automatically calculates data reliability.

Recap: Module 2

- Detector data is sent to TMCs, and then to the archive
- Reliability checking and imputation occur at the archive
- Data can be accessed remotely
- Custom report generation possible

Recap: Module 2



Recap: Module 3

A central archive can be smoothly implemented in three phases:

Phase I: Establish policies and standards

Phase II: Implement central archive

Phase III: Connect individual TMCs to archive

Phase III is repeated for each TMC that is connected. This allows flexibility in connecting additional TMCs as a later date.

Recap: Module 4

Data reliability is based on three criteria:

- **Fundamental**: Is the state physically possible?
- **Network**: Is the state consistent with nearby locations?
- **Historical**: Is the state reasonable, given past data?

Continuous set theory allows these three assessments to be combined into one number (0-10)

Recap: Module 5

- Missing/suspicious data is a problem (25%)
- Regression models closely fit volume counts
- Regression models can have high data requirements
- Local regression is more accurate than global
- Global regression is robust to certain types of failures
- Historical imputation can almost always be used, but is far less accurate

Recap: Module 6

Data can be imputed even when no nearby detector is available:

Historical Scaling: Use past data and apply factors

Spatial Kriging: Estimate continuous function from all detectors

Calibration to Assignment: Scale an existing traffic assignment

Future Directions and Next Steps

- Include weather, construction, and special event data
- Use to generate public travel information
- Incorporate innovative detector technology
- Adapt error correction procedures as data reporting policies change

END OF WORKSHOP

Questions?