U.S. Department
of Transportation

**Federal Highway
Administration**

ITS Data Archiving: Case Study Analyses of San Antonio TransGuide® Data

Report No. FHWA-PL-99-024 Final Report August 1999

Prepared for

ITS Joint Program Office and the
Office of Highway Policy Information

Prepared by

Texas Transportation Institute
The Texas A&M University System

Technical Report Documentation Page

1. Report No. FHWA-PL-99-024	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle ITS Data Archiving: Case Study Analyses of San Antonio TransGuide® Data		5. Report Date August 1999
		6. Performing Organization Code
7. Author(s) Shawn M. Turner, William L. Eisele, Byron J. Gajewski, Luke P. Albert, and Robert J. Benz		8. Performing Organization Report No.
9. Performing Organization Name and Address Texas Transportation Institute The Texas A&M University System College Station, Texas 77843-3135		10. Work Unit No. (TRAIS)
		11. Contract or Grant No.
12. Sponsoring Agency Name and Address Federal Highway Administration ITS Joint Program Office (HOIT) and the Office of Highway Policy Information (HPPI) 400 Seventh Street, SW Washington, DC 20590		13. Type of Report and Period Covered Final: March 1998 - August 1999
		14. Sponsoring Agency Code FHWA CO6489605 (Amend. 1)

15. Supplementary Notes

Research performed in cooperation with the Federal Highway Administration and the Texas Department of Transportation. Project COTR: Mr. Ralph Gillmann, HPPI
Research Project Title: Retention, Management, and Sharing of ITS Data

16. Abstract

This report documents: 1) the state-of-the-practice in retaining and using archived ITS data, and 2) lessons learned from case study analyses of archived ITS data from TxDOT's TransGuide® center in San Antonio. The case study analyses of TransGuide® data were focused on data aggregation and data quality (i.e., data errors, missing data, and data accuracy).

The state-of-the-practice findings indicate that many operations centers are becoming more interested in archiving ITS data. Several concerns have been expressed by potential users about ITS data quality and location referencing of ITS detector locations. Analyses of TransGuide® data resulted in statistical procedures that can be used to help guide decisions about appropriate data aggregation levels. The researchers also present innovative data archiving approaches that can help to address different data aggregation needs, including capabilities such as data sampling, on-demand archiving, and data broadcasts. In other analyses, the researchers found mixed results in comparing traffic volumes from TransGuide® detectors to other nearby TxDOT traffic counting stations and ground truth video. At one location, traffic volumes from both detectors were within 5 percent of ground truth, whereas detectors at the second location ranged between 12 to 38 percent less than ground truth. In analyses of a full month of data, the researchers also found that the level of missing data was nearly 20 percent, whereas basic quality control checks flagged only 1 percent of the data as suspect or erroneous.

17. Key Words archived ITS data, data aggregation, quality control, ITS data accuracy, ITS data management	18. Distribution Statement No restrictions. This document is available to the public through NTIS: National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report) Unclassified	20. Security Classif.(of this page) Unclassified	21. No. of Pages 146	22. Price

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

DISCLAIMER

ACKNOWLEDGMENTS

The authors wish to acknowledge the assistance and support of the following persons:

- **Dr. Joseph Peters** in the ITS Joint Program Office and **Mr. Ralph Gillmann** in the Office of Highway Policy Information of FHWA, who supported this research effort;

- **Mr. Pat Irwin** and his **staff at TransGuide®**, who have cooperated with and supported TTI research (including ITS data archiving) for many years;

- **Mr. Ken Mora** in the Transportation Planning and Programming Division at TxDOT, who served as the TxDOT liaison for this FHWA-funded study;

- **Mr. Mark Hodges** and **Mr. Jeff Reding** in the Transportation Planning and Programming Division at TxDOT, who provided ATR traffic volume data and detector locations;

- **Mr. Christopher Poe**, director of the TransLink® ITS research program at TTI, who has supported ITS data research through the TransLink® program since 1996;

- **Mr. Darrell Borchardt** in the Houston office of TTI, who helped support this and other research by providing ITS-generated data from Houston;

- **Dr. Clifford Spiegelman** in the Statistics Department of Texas A&M University, who provided advice and statistical expertise on ITS data archiving issues;

- the numerous transportation professionals who took time out of busy schedules to share their ITS data archiving experiences with TTI researchers, as well as those who helped to improve this research report by providing valuable review comments.

TABLE OF CONTENTS

PAGE

LIST OF FIGURES vii

LIST OF TABLES ix

LIST OF ACRONYMS xi

EXECUTIVE SUMMARY xiii

CHAPTER 1. INTRODUCTION 1
- HISTORICAL PERSPECTIVE 2
- ITS DATA ARCHIVING SYSTEM IMPLEMENTATION ISSUES 2
- STUDY OBJECTIVE 4
- REPORT OVERVIEW 4

CHAPTER 2. BACKGROUND AND STATE-OF-THE-PRACTICE 5
- CASE STUDIES OF EXISTING ITS DATA ARCHIVING PRACTICES 5
  - Phoenix, Arizona 8
  - Los Angeles, California 9
  - Orange County, California 10
  - San Francisco, California 10
  - Atlanta, Georgia 10
  - Chicago, Illinois 10
  - Montgomery County, Maryland 11
  - Detroit, Michigan 11
  - Minneapolis-St. Paul, Minnesota 12
  - TRANSCOM, New York/New Jersey/Connecticut 12
  - Ft. Worth, Texas 13
  - Houston, Texas 14
  - San Antonio, Texas 16
  - Seattle, Washington 17
  - Toronto, Ontario, Canada 17
  - State of Virginia 18
  - North America Pre-clearance and Safety Systems 19
- STATE-OF-THE-PRACTICE IN USING ARCHIVED ITS DATA 21
  - Metropolitan Planning Organizations / State Transportation Planners 23
  - Traffic Management Operators 34
  - Maintenance and Construction 35
  - Transportation Researchers 35
  - Private Sector 39

TABLE OF CONTENTS (Continued)

PAGE

ADVANCING THE STATE-OF-THE-PRACTICE IN ITS DATA ARCHIVING 40
Development of the ADUS Architecture 41
Data Standards Development 44
ITS Data Registry 46
Volpe Center Case Studies 47
Oak Ridge National Lab/Florida DOT/New York DOT Demonstration Project . 47
Washington State 48
TransLink® ITS Research Program, Texas Transportation Institute 51
SUMMARY OF FINDINGS AND LESSONS LEARNED 53

CHAPTER 3. EXAMINATION OF ITS DATA ARCHIVING ISSUES 55
INCORPORATING DATA NEEDS INTO ITS DATA ARCHIVING 56
Process of Defining or Establishing Data Needs 56
Previous Work on Data Needs and User Requirements 60
Summary of Factors to Consider in Developing ITS Data Archiving User Requirements 67
AGGREGATION OF ITS TRAFFIC MONITORING DATA 68
Approach 1: Save Aggregated Data 68
Approach 2: Save All Data 68
Approach 3: Focus on Strategic Archiving, Build Advanced Capabilities 70
Data Sampling 71
Aggregation Based on Statistical Variability 72
Archiving on Demand 88
Data Broadcast 89
ITS DATA QUALITY CONTROL 90
Erroneous or Suspicious Data 90
Nature and Extent of Missing Data 96
Accuracy and Comparability of ITS Data 102
DATA STORAGE TOOLS AND ISSUES 114
Archived ITS Data Storage Tools 114
Privacy Issues 118

CHAPTER 4. FINDINGS AND CONCLUSIONS 119
STATE-OF-THE-PRACTICE IN RETAINING AND USING ARCHIVED ITS DATA 119
EXAMINATION OF ITS DATA ARCHIVING ISSUES 120

REFERENCES 123

LIST OF FIGURES

FIGURE		PAGE
1	Summary of Steps in Implementing ADMS	3
2	Proposed AZTech Data Server	8
3	Proposed Data Management System in Ft. Worth's TransVISION	14
4	Press Release Describing Advantage CVO Data Policy	20
5	Traffic Profile of SR-520 General Purpose Lanes	24
6	Estimated Average Weekday Travel Time for Eastbound SR-520	25
7	Estimated Weekday Volume, Speed, and Reliability Conditions for Westbound SR-520, 76th Avenue N.E.	26
8	Average Morning Peak Hour Travel Times for Katy Freeway, 1994	27
9	Morning Peak Hour Travel Time Reliability	28
10	Average Weekday Traffic Volume Patterns Derived from ITS Data, Phoenix Urban Area	30
11	Monthly Traffic Variations for Freeways Using ITS Data, Phoenix Urban Area	31
12	Traffic Volume-Speed Relationship Using ITS Data	31
13	Illustration of Hypothesis that ITS Strategies are Most Effective in High-Variability Systems	32
14	Accident and Weather Scenario Mapping by Demand Level Using Archived ITS Data and Other Data Sources	33
15	National ITS Architecture	43
16	ITS Data Registry Functional Concepts	46
17	Main Interface of CDR Software	49
18	Schematic of TDAD System	50
19	Query Interface of DataLink System (http://vixen.cs.tamu.edu)	52
20	Process that Starts with Enumeration of Available ITS Data	57
21	Process that Starts with Stakeholder Data Needs	58
22	Iterative Process that Establishes Dialogue between Data Stakeholder Groups and Data Provider	59
23	Cross-Validated Mean Square Error Technique	74
24	F-Statistic Technique	78
25	Typical Speed Profile and Results of CVMSE Analysis for Optimal Aggregation Level	81
26	Typical Speed Profile and Results of F-Statistic Analysis for Optimal Aggregation Level	81
27	Illustration of Loss Profiles for the Cross-Validated Mean Square Error Technique	84
28	Range of Possible ITS Data Aggregation Solutions	85
29	Example of Statistical Process Control Guidelines	94
30	Identifying Detector-Specific Missing Data	98

LIST OF FIGURES (Continued)

FIGURE **PAGE**

31 Identifying the Extent of Detector-Specific Missing Data 99
32 Identifying Server-Specific and System-Wide Missing Data 100
33 Location of ATR S184 and Nearby TransGuide® Loops 103
33 Location of ATR S185 and Nearby TransGuide® Loops 103
35 Comparison of Hourly Volumes from ATR S184 EB and Equivalent Upstream TransGuide® Detectors 107
36 Comparison of Hourly Volumes from ATR S184 EB and Equivalent Downstream TransGuide® Detectors 107
37 Comparison of Hourly Volumes from ATR S184 WB and Equivalent Upstream TransGuide® Detectors 108
38 Comparison of Hourly Volumes from ATR S184 WB and Equivalent Downstream TransGuide® Detectors 108
39 Comparison of Hourly Volumes from ATR S185 SB and Equivalent Upstream TransGuide® Detectors 109
40 Comparison of Hourly Volumes from ATR S185 SB and Equivalent Downstream TransGuide® Detectors 109
41 Comparison of Hourly Volumes from ATR S185 NB and Equivalent Upstream TransGuide® Detectors 110
42 Comparison of Hourly Volumes from ATR S185 NB and Equivalent Downstream TransGuide® Detectors 110
43 Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-10 East 111
44 Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-10 West 111
45 Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-37 South 112
46 Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-37 North 112

LIST OF TABLES

TABLE PAGE

1 Contacts for Existing ITS Data Archiving Practices 6
2 TMCs and Agencies Contacted in First and Second Telephone Surveys 7
3 Example of User Requirements Matrix for Houston TranStar 15
4 Stakeholders and Example Applications for Archived ITS Data 21
5 Efforts to Advance the State-of-the-Practice in Archived ITS Data 40
6 Incorporation of the ADUS into the National ITS Architecture 42
7 Attributes of a Vehicle Detector Count Using ITE's TMDD 44
8 Selected ITS Standards Relevant to ITS Data Archiving 45
9 Documentation of Data Needs in Transportation Decision Making 61
10 ITS Data Relevant for Archiving 62
11 Comparison of Approaches for Aggregating Archived ITS Data 69
12 Illustration of Every Nth Day Sampling 71
13 Illustration of Concentrated Weeks Sampling 72
14 CVMSE and F-Statistic Algorithm Results, IH-10 West, Milepost 566.641 80
15 Example Matrix: Level of Aggregation for ITS Data Archiving 86
16 Quality Control Screening Scenarios and Corresponding Flag Values 92
17 Summary of Suspect or Erroneous TransGuide® Data, October 1998 93
18 Extent of Suspect or Erroneous TransGuide® Data for Individual
Scenarios, October 1998 95
19 Typical Causes and Characteristics of Missing ITS Traffic Monitoring Data 96
20 Identification of Missing Data for San Antonio's TransGuide® 97
21 Equations Used to Compare ATR Stations to TransGuide® Detectors 104
22 Results of Hourly Traffic Volume Comparisons Between ATR and
TransGuide® Detectors, May 1998 106
23 Results of Ground-Truth Traffic Volume Comparisons, December 1998 113
24 Comparisons of Computer Storage Media and Devices 115

LIST OF ACRONYMS

AADT	annual average daily traffic volume
AASHTO	American Association of State Highway and Transportation Officials
ADS	AZTech data server (Maricopa County, Arizona)
ADMS	archived data management subsystem
ADUS	archived data user service
ANN	artificial neural network
ANOVA	analysis of variance
ASTM	American Society for Testing and Materials
ATMS	advanced transportation management system
ATR	automatic traffic recorder
AVI	automatic vehicle identification
AWDT	average weekday traffic volume
BTS	Bureau of Transportation Statistics
CAAA	Clean Air Act Amendments
CAD	computer-aided dispatch
Caltrans	California Department of Transportation
CATS	Chicago Area Transportation Study (Illinois)
CCTV	closed circuit television
CD	compact disc
CMAQ	congestion mitigation/air quality
CMS	changeable message sign
CVMSE	cross-validated mean square error
CVO	commercial vehicle operations
DMS	dynamic message sign
DOT	department of transportation
DVD	digital video/versatile disk
FHWA	Federal Highway Administration
FOT	field operational test
GIS	geographic information system
HAR	highway advisory radio
HGAC	Houston-Galveston Area Council (Texas)
HOV	high-occupancy vehicle
IEEE	Institute of Electrical and Electronic Engineers
ITE	Institute of Transportation Engineers
ITS	intelligent transportation system
ITS	Institute of Transportation Studies (University of California at Berkeley)
KTC	Kentucky Transportation Center (University of Kentucky)
LCS	lane control signal

LIST OF ACRONYMS (Continued)

LCU	local controller unit
MDOT	Michigan Department of Transportation
M-NCPPC	Maryland-National Capital Park and Planning Commission
MITS	Michigan Intelligent Transportation System Center (Detroit)
Mn/DOT	Minnesota Department of Transportation
MOE	measure of effectiveness
MPO	metropolitan planning organization
NISS	National Institute of Statistical Sciences
NORPASS	North American Pre-Clearance and Safety System
OLAP	on-line analytical processing
OLTP	on-line transaction processing
RAID	redundant array of independent disks
RCOC	Road Commission of Oakland County (Michigan)
SDO	standard development organization
SNN	spectral-basis neural network
SQL	structured query language
SwRI	Southwest Research Institute (San Antonio, Texas)
TDAD	Traffic Data Acquisition and Distribution (University of Washington)
TLFD	trip length frequency distribution
TMC	transportation management center
TMDD	Traffic Management Data Dictionary
TP&P	Transportation Planning and Programming Division (TxDOT)
TRAC	Washington State Transportation Center (University of Washington)
TRANSCOM	Transportation Operations Coordinating Committee (New York/New Jersey/Connecticut)
TransGuide	Transportation Guidance System (San Antonio, Texas)
TRANSMIT	TRANSCOM's System for Managing Incidents and Traffic
TransVISION	Transportation Vehicular Information and Surveillance on an Intelligent highway Optical Network (Ft. Worth, Texas)
TSC	Traffic Systems Center (Chicago, Illinois)
TSM	transportation system management
TSMC	Transportation Systems Management Center (Seattle, Washington)
TTI	Texas Transportation Institute (Texas A&M University System)
TxDOT	Texas Department of Transportation
VDOT	Virginia Department of Transportation
VMT	vehicle-miles of travel
WIM	weigh-in-motion
WSDOT	Washington State Department of Transportation

EXECUTIVE SUMMARY

The primary objective of this research study was to develop guidance for implementing (intelligent transportation system) ITS data archiving systems by identifying and examining selected implementation issues. To accomplish this objective, Texas Transportation Institute (TTI) researchers performed case study analyses using archived loop detector data from the TransGuide® system in San Antonio, Texas. The selected implementation issues and associated questions examined in this study include:

- **data aggregation -** What is the desired aggregation level at which to retain traffic monitoring data? What are the tradeoffs in saving data at different levels? Are there other data archiving strategies besides simply saving all data at a single aggregation level?

- **data quality -** What types of data quality should I be concerned about (i.e., erroneous or suspect data, missing data, and data accuracy)? How do I address these concerns with quality control procedures?

The researchers also gathered and synthesized information about existing ITS data archiving practices, the most common applications of archived ITS data, and data storage tools and issues.

ITS DATA ARCHIVING PRACTICES

The research team surveyed fifteen traffic management centers (TMCs) in the spring of 1999 to gather information on current ITS data archiving practices. Information on the current uses of archived ITS data was also gathered from transportation agencies in these same areas. The surveys indicated that nearly all TMCs were saving ITS-generated data, but most were at different stages in making the archived data accessible to a wide variety of users. Several of the TMCs interviewed were developing compact disc (CD) or Internet-based archived data management systems to reduce the problems associated with archived data on magnetic tape cartridges. User requirement studies were being conducted in several areas to identify the users and uses of the archived data. Several interviewees mentioned data quality as a major issue in ITS data archiving.

Many of the uses of archived ITS data to date have been for transportation planning or research. These planning uses have primarily consisted of congestion management systems, performance measurement, model calibration/validation, and basic traffic counts and factors. The applications for research have focused on model development and traffic flow theory. A few of the traffic management operators are using archived ITS data to refine operating procedures and perform annual evaluations.

AGGREGATION OF ITS TRAFFIC MONITORING DATA

Because ITS data are collected at frequent intervals (typically 20 to 30 seconds) at many locations, data aggregation (to 5 or 15 minutes) is seen as a way to reduce data storage and make ITS data less cumbersome for ordinary data users. The research team developed two statistical methods for determining optimal data aggregation intervals based upon statistical variability. The two statistical methods yielded similar results: little or no aggregation (one minute or less) during congested periods, hourly aggregation during low traffic periods, and moderate aggregation (five or fifteen minutes) during transition periods.

The authors acknowledge that there are a wide range of possible aggregation solutions based upon these statistical techniques (Figure S-1). Ultimately, the aggregation solutions may be driven by non-statistical parameters such as cost (e.g., "how much do we/the market value the data?"), ease of implementation, system interface requirements, and other constraints. Some may only desire a simple ITS data archiving solution for existing data needs, thereby negating the need for using statistical techniques. Others may desire a single fixed aggregation level throughout the day, in which the statistical techniques or their results could be used to examine the tradeoffs between optimal aggregation levels and existing constraints. Still others may wish to implement dynamic aggregation levels in ITS data archiving, in which case the statistical techniques described here could be programmed and automated within the data archiving system.

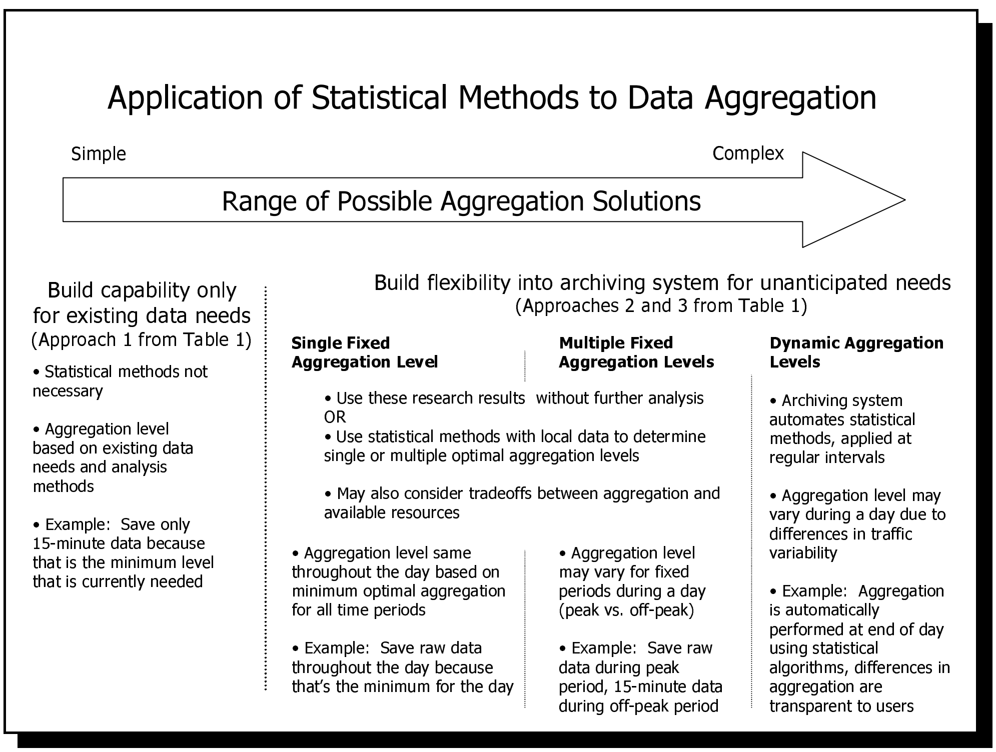


Figure S-1. Range of Possible ITS Data Aggregation Solutions

QUALITY CONTROL FOR ITS DATA

The quality control efforts in this study were focused on a month of ITS traffic monitoring data (collected by inductance loop detectors) from San Antonio's TransGuide® system, and analyzed three basic attributes of quality control:

- suspicious or erroneous data;
- nature and extent of missing data; and
- accuracy and comparability of ITS data to similar data sources.

Suspicious or Erroneous Data

The analysis of suspicious or erroneous data found that about 1 percent of the archived ITS data could be labeled as "suspect" according to current practices for identifying ITS data errors. The algorithms used to identify suspect or erroneous data were aimed at identifying blatant data errors, so more stringent quality control rules would result in a higher percent of the data being labeled as "suspect." The research team recommended the development of more advanced quality control procedures for archived ITS data. These advanced quality control procedures should examine each data record in relation to other data records in close proximity in both time and space. Traffic patterns and profiles from adjacent lanes can also be compared to ensure consistency and reliability. More advanced quality control procedures could also compare each data record to historical patterns and trends for average values and typical ranges of variability at that location.

Nature and Extent of Missing Data

Although the percent of archived data labeled as "suspect" was quite small, missing data records did account for a significant portion of the data (about 22.5 percent). Missing ITS data is a result of roadway construction activities, equipment failures, disruptions in communications, or computer hardware and software failures. Given the continuous operation of most ITS traffic monitoring devices, missing data is almost inevitable. In San Antonio, this missing data typically ranged from 5 to 25 percent for most days. Because of excellent loop maintenance programs in San Antonio, only about 5 to 15 percent of this missing data can be accounted for by loop detector failure. The remaining 10 or more percent can most likely be explained by central traffic management or data archiving system failures.

Accuracy of ITS Data

In initial conversations with TxDOT's Transportation Planning and Programming (TP&P) Division about using San Antonio's TransGuide® data, statewide transportation planners named data quality as their most pressing concern. More specifically, the statewide planners were most interested in how TransGuide® data compared to data collected at nearby automated traffic recorder (ATR) stations. To answer these questions about data comparability, TTI compared a full month of TransGuide® traffic volume data to hourly traffic volume data collected at two

ATR stations that were located along freeways instrumented in Phase One of TransGuide®. Aside from the comparability of TransGuide® traffic volumes to ATR traffic volumes, TTI researchers were concerned about the accuracy of both types of traffic monitoring equipment to "ground truth." To answer questions about how well these two data sources compared to "ground truth," TTI researchers used the TransGuide® surveillance cameras to record actual traffic conditions.

The analysis of TransGuide®, ATR, and ground truth traffic volume data found a wide range of comparability between the ATR and TransGuide® data. At the IH-10 location, the ATR and TransGuide® traffic volumes compared favorably, typically falling with 10 percent of one another. However, the IH-37 location showed large differences in traffic volumes, ranging from 18 to 39 percent in directly comparing ATR and TransGuide® detectors. The researchers also found mixed results in comparing the collected traffic volumes to ground truth. At the IH-10 location, both the ATR and TransGuide® traffic volumes were within 5 percent on average of ground truth volumes, within no clear difference between ATR and TransGuide® detectors. At the IH-37 location, however, the ATR detectors were slightly more accurate (within 4 and 11 percent on average) than the TransGuide® detectors (within 13 to 38 percent on average).

CONCLUSIONS AND RECOMMENDATIONS

Most Centers Have ITS Data Archiving on Their "Radar" - The research team found that most operations and management centers contacted were at least considering ITS data archiving, if not actively developing plans for archived data systems. Because of this widespread interest in ITS data archiving, ADUS deployment guidance will be imperative in the further development of integrated data systems. A common assertion is that many centers have not been archiving ITS data; however, the research team found in 1997 that 12 of 15 centers were archiving data, but that the data were either not readily accessible or usable (stored on proprietary magnetic tape cartridges) or not widely distributed within the metropolitan area.

Access to Archived ITS Data Creates Opportunities - The research team found the most widespread use of archived ITS data was in locations where the data were easily or publicly accessible. In these locations, user groups outside of the operations center were able to develop data extraction and analysis tools because the archived ITS data were easily accessible. The logical conclusion is that in areas where the operations center may not be able to develop systems that support ADUS, these centers should at least provide easy access to the data so that other user groups can develop systems that support ADUS functions. Examples of easy access to archived ITS data include distribution via the Internet or CDs.

Planners and Researchers Have Been Primary Archived Data Users to Date - The research team found that the two most common user groups have been researchers and planners, with applications ranging from basic traffic statistics to advanced model and algorithm development. In the few locations where traffic management operators were using archived ITS data, they were realizing benefits in improving traffic operations and management, as well as an ability to quantify these benefits. Private sector users that add value to archived ITS data are emerging users, although the business models are not yet clearly defined.

EXAMINATION OF ITS DATA ARCHIVING ISSUES

Need to Establish Ongoing Dialogue Between Data Providers and Data Users - The best synergy for archiving ITS data occurs when there is an ongoing dialogue between data providers (e.g., operations centers) and data users. An ongoing dialogue will help in the following ways: 1) data users may better understand the available ITS data and its intricacies; 2) data providers may better understand the needs of data users; and 3) data providers and data users may work together in establishing ITS detector designs that meet the data needs of many groups.

Consider Innovative Archiving Approaches to Address Different Data Aggregation Needs - To date, there have been two basic suggested approaches to archiving ITS data: 1) save aggregated data to meet existing needs in a cost-effective manner; 2) save raw, disaggregate data to maintain flexibility of analyses and data exploration. The research team suggests a third approach, in that basic aggregated data are saved for current needs, yet innovative archiving capabilities exist that can provide advanced data users access to raw, disaggregate data. The innovative capabilities discussed in this report include raw data sampling, aggregation based upon statistical variability, on-demand archiving, and data broadcasts.

Need Improved Quality Control Procedures - These advanced (and ideally automated) procedures could compare current data at a given location to upstream and downstream detectors, as well as examining historical values, trends, and patterns. Quality control procedures to establish absolute data accuracy ("ground truthing") should be also be included. The current consensus is that archived ITS data not passing quality control checks should be flagged as such, with users making the choice of imputed or replacement values appropriate for their analyses.

Account for Missing Data in Data Archiving System Designs - It is critical that ITS data archiving systems be able to reflect missing data in summary or aggregate statistics. Cumulative statistics, such as vehicle-miles of travel (VMT), are most affected by missing data and will likely require data users and system designers to collaborate on how to account for the missing data. As with quality control checks, the consensus is that missing data be flagged as such, with users making the choice of appropriate imputed or replacement values.

CHAPTER 1.
INTRODUCTION

🕮 CHAPTER OVERVIEW

☞	**Historical Perspective**	Provides an historical perspective on archiving of operational data, summarizes problems noted in previous work.
☞	**ITS Data Archiving System Implementation Issues**	Lists implementation issues associated with ITS data archiving, suggests steps in implementing data archiving.
☞	**Study Objective**	Describes objectives of research study.
☞	**Report Organization**	Summarizes report organization.

With the advent of the archived data user service (ADUS) in 1998, the National ITS Architecture officially embraced the concept of saving (a.k.a. retaining, archiving) real-time, operational data for other non-real-time applications, such as transportation planning, safety analyses, and research. Intelligent transportation system (ITS) applications and their respective sensors and detectors are a potentially rich source of data about transportation system performance and characteristics. Examples of the most common data elements potentially available from ITS include:

- **traffic monitoring and detection systems -** vehicle volume, speed, travel time, classification, weight, and trajectories;
- **traffic control systems -** time and location of traffic control actions (e.g., ramp metering, traffic signal control, lane control signals, message board content);
- **incident and emergency management systems -** location, cause, extent, and time history of roadway incident/emergency detection and clearance; and
- **advanced public transit systems -** transit vehicle boardings by time and location, vehicle trajectories, origins and destinations.

In this report, ITS data refers to data that is typically collected by and/or generated for ITS applications. ITS data archiving refers to the systematic retention of data that is generated by ITS applications. The philosophy of ADUS encompasses more than just ITS data archiving, however, and can include data extraction, manipulation and analysis tools as well.

HISTORICAL PERSPECTIVE

Although the practice of saving or archiving operational data from traffic control sensors and detectors is not a new concept, national interest in ITS data archiving has grown significantly since 1997, due mostly to the formation of ADUS. As early as the 1970s, the Illinois Department of Transportation (DOT) was saving aggregated loop detector data in Chicago to report "minute-miles" of congestion (1). Similarly in 1968, the Texas Highway Department and the Texas Transportation Institute (TTI) were using an IBM 1800 computer to save and analyze loop detector data along the Gulf Freeway in Houston (2). The archived Houston data were used to support level of service and merging research studies, as well as to demonstrate and quantify the effects of incidents on the freeway corridor. The Washington State DOT (WSDOT) and the University of Washington have been archiving loop detector data from Seattle's freeway traffic management system since 1981, with researchers and planning agencies being primary users (3). Loop detector data from Highway 401 in Toronto, Ontario (Canada), also has been used extensively since the 1980s for traffic flow theory and capacity research.

To date, however, ITS and operational data archiving has mostly been informal arrangements with a single data user group (e.g., researchers, planners) who were able to work with these large data sets. Numerous problems surfaced in this early work with archived data:

- proprietary data formats and data storage devices (e.g., magnetic tape cartridges) hindered archived data distribution;
- distributing archived data to users sometimes placed an unreasonable burden on operations personnel;
- processing and analyzing the large quantities of data in a mainframe computer environment was not user-friendly; and
- missing and erroneous data were evident throughout the archived data.

ITS DATA ARCHIVING SYSTEM IMPLEMENTATION ISSUES

The ADUS addition to the National ITS Architecture, set to be released later in 1999, will provide an overall framework for designing archived data management subsystems (ADMS) and archived data user systems. Some of the problems described earlier may be lessened by adoption of the ADMS architecture and associated data standards. However, there are numerous technical and institutional issues that agencies may still have to address in implementing ADMS (4):

1. development, operation, and maintenance costs
2. system access
3. ownership

4. data quality
5. data management
6. data and communications standards
7. privacy concerns
8. data analysis
9. coordination with other data collection efforts
10. liability
11. confidentiality of privately collected data
12. incremental and uncoordinated ITS deployments
13. retrofitting vs. new development of systems
14. data not defined by the National ITS Architecture
15. conformance with metric conversion standards
16. training and outreach

Figure 1 shows important steps in the implementation of ADMS. The figure illustrates that both institutional arrangements (i.e., "Data Partnerships") and technical details (i.e., "System Design") are necessary in deploying an effective ADMS.

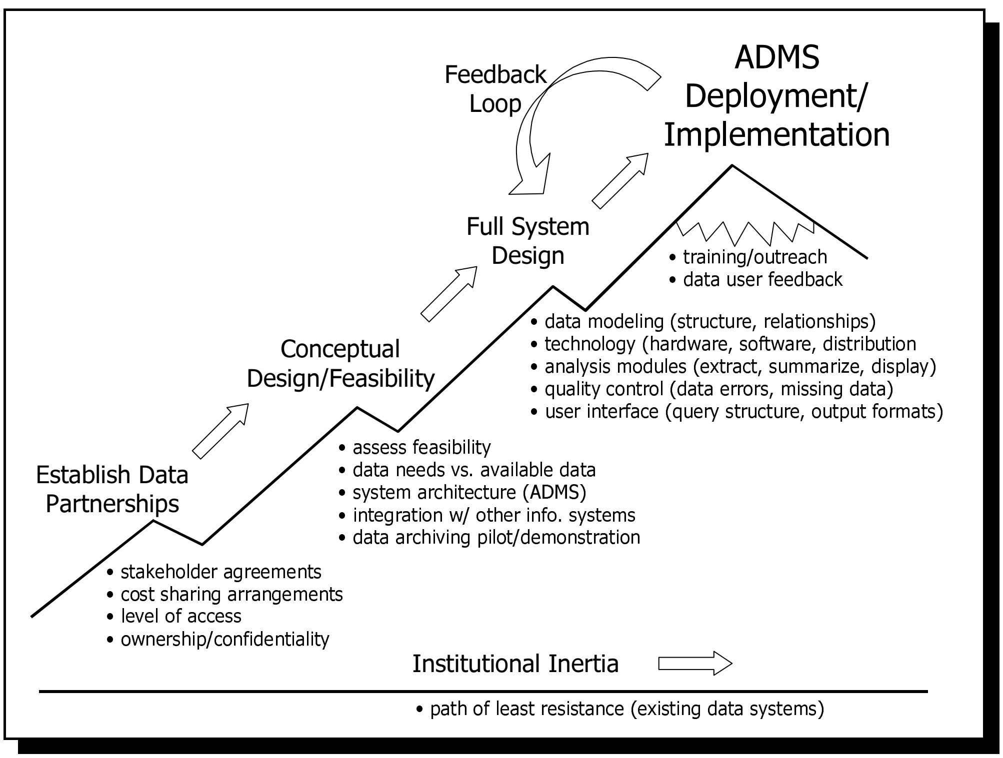


Figure 1. Summary of Steps in Implementing ADMS

STUDY OBJECTIVE

The primary objective of this research study was to develop guidance for implementing ITS data archiving systems by identifying and examining selected implementation issues. To accomplish this objective, TTI researchers performed case study analyses using archived loop detector data from the TransGuide® system in San Antonio, Texas. The selected implementation issues and associated questions examined in this study include:

- **data management -** What is the desired aggregation level at which to retain traffic monitoring data? What are the tradeoffs in saving data at different levels? Are there other data archiving strategies besides simply saving all data at a single aggregation level?

- **data quality -** What types of data quality should I be concerned about (i.e., erroneous or suspect data, missing data, and data accuracy)? How do I address these concerns with quality control procedures?

The researchers also gathered and synthesized information about existing ITS data archiving practices, the most common applications of archived ITS data, and data storage tools and issues.

REPORT OVERVIEW

This report is organized into the following four chapters:

- **Chapter 1. Introduction -** introduces ITS data archiving, implementation issues, and study objective;

- **Chapter 2. Background and State-of-the-Practice -** summarizes existing ITS data archiving practices, describes and provides examples of common applications for archived ITS data, and discusses efforts to advance the state-of-the-practice;

- **Chapter 3. Examination of ITS Data Archiving Issues -** describes exploratory analyses conducted to examine implementation issues such as data management and aggregation, data quality, and data storage tools; and.

- **Chapter 4. Findings and Conclusions -** summarizes major findings and lessons learned related to selected ITS data archiving issues.

CHAPTER 2.
BACKGROUND AND STATE-OF-THE-PRACTICE

CHAPTER OVERVIEW

☞	**Case Studies of Existing ITS Data Archiving Practices**	Describes data archiving practices at several operations or transportation management centers.
☞	**State-of-the-Practice in Using Archived ITS Data**	Summarizes and provides examples of the most common applications of archived ITS data.
☞	**Advancing the State-of-the-Practice in ITS Data Archiving**	Presents National ITS Architecture development as well as research efforts aimed at advancing ITS data archiving practices.
☞	**Summary of Findings and Lessons Learned**	Summarizes findings in regard to state-of-the-practice in ITS data archiving.

CASE STUDIES OF EXISTING ITS DATA ARCHIVING PRACTICES

This section contains information on the state-of-the-practice in ITS data archiving. These experiences are drawn from transportation management centers (TMCs) or other operations centers (e.g., commercial vehicle operations (CVO) at weigh stations) where real-time traffic monitoring data are collected. Table 1 contains the contacts from which this ITS data archiving information was obtained.

Previous TTI research documents the existing practices of retaining ITS data for transportation analyses (5,6). The research found that the uses and users of archived ITS data appeared to be relatively consistent at the TMCs that were surveyed. They include requests for traffic counts to estimate construction impacts, requests for planning and modeling uses, and researcher requests. While the different TMCs have similar uses and user requests, the data storage and aggregation practices were not consistent. Of the 15 TMCs interviewed in the spring of 1997, three were not saving any of the ITS data. Nine of the TMCs saved traffic data at intervals of one minute or less, and the remaining three TMCs saved the data at an aggregation level between 5 and 15 minutes.

As part of this study, the research team re-visited the list of TMCs from the previous research to identify any changes in ITS data archiving strategies. Table 2 summarizes the locations that were contacted in the first and second set of telephone surveys. In this second follow-up survey, it was found that the TMC operators and managers are beginning to recognize the value of ITS data archiving for secondary uses and are interested in saving, formatting, and aggregating data to facilitate other uses beyond the traditional real-time needs. There appears to be a stronger interest in interdisciplinary coordination (i.e., operations staff working more closely with planning personnel), although financial allocations are often the leading hindrance to such efforts. The following paragraphs describe ITS data archiving and retention practices at these selected TMCs.

Table 1. Contacts for Existing ITS Data Archiving Practices

Location	Agency	Contact Person and Phone Number
Phoenix, Arizona	Maricopa County DOT	David Wolfson, (602) 506-6950
	Maricopa Association of Governments	Mark Schlappi, (602) 254-6300
Los Angeles, California	Caltrans	David Lau, (213) 897-0391
Orange County, California	Caltrans	Omid Segal, (949) 724-2467
San Francisco, California	Caltrans	Dick Fahey, (510) 286-5761
Atlanta, Georgia	Georgia DOT	Dennis Reynolds, (404) 635-8027
Chicago, Illinois	Illinois DOT	Tony Cioffi, (708) 524-2145
Montgomery County, Maryland	Montgomery County DOT	Bruce Mangum, (301) 217-2197
	Maryland National Capital Park and Planning Commission	Don Ostrander, (301) 495-2184
Detroit, Michigan	Michigan DOT	Arvyd Satraitis, (313) 256-9800
Minneapolis-St. Paul, Minnesota	Minnesota DOT	James Aswegan, (612) 373-2761
Jersey City, New Jersey	TRANSCOM	Sanjay Patel, (201) 963-4033
Fort Worth, Texas	Texas DOT	Abed Abukar, (817) 370-6621
Houston, Texas	Texas DOT	Cindy Gloyna, (713) 802-5147
San Antonio, Texas	Texas DOT	Pat Irwin, (210) 731-5249
Seattle, Washington	Washington DOT	Mahrokh Arefi, (206) 440-4462
Toronto, Ontario, Canada	Ontario Ministry of Transportation	David Tsui, (416) 235-3538
Lexington, Kentucky	NORPASS, Kentucky Transportation Center	Jennifer Walton, (606) 257-4513

Table 2. TMCs and Agencies Contacted in First and Second Telephone Surveys

Geographic Location	Contacted in First Survey (Spring 1997)	Contacted in Second Survey (Spring 1999)
Phoenix, Arizona	Arizona DOT Traffic Operations Center	Maricopa County DOT Maricopa Association of Governments
Los Angeles, California	Los Angeles (Caltrans District 7) Transportation Management Center	Los Angeles (Caltrans District 7) Transportation Management Center
Orange County, California	—	Orange County (Caltrans District 12) Transportation Management Center
San Francisco, California	San Francisco Bay Area TMC	San Francisco Bay Area TMC
Atlanta, Georgia	Georgia DOT ATMS	Georgia DOT ATMS
Chicago, Illinois	Traffic Systems Center	Traffic Systems Center
Montgomery County, Maryland	Montgomery County TMC	Montgomery County TMC Maryland National Capital Park and Planning Commission
Detroit, Michigan	Michigan ITS Center	Michigan ITS Center
Minneapolis-St. Paul, Minnesota	Mn/DOT TMC	Mn/DOT TMC
Jersey City, New Jersey	TRANSCOM	TRANSCOM
Long Island, New York	INFORM	—
New York City, New York	MetroCommute	—
Ft. Worth, Texas	—	TransVISION
Houston, Texas	TranStar	TranStar
San Antonio, Texas	TransGuide	TransGuide
Seattle, Washington	WSDOT TSMC	WSDOT TSMC
Toronto, Ontario, Canada	COMPASS	COMPASS
Lexington, Kentucky	—	NORPASS (formerly Advantage CVO) (Kentucky Transportation Center)

Note: “—” indicates that personnel at this location were not contacted during the particular survey.

Phoenix, Arizona

The Arizona DOT (ADOT) Traffic Operations Center in Phoenix began operation in September 1995. All of the 20-second loop detector data that comes into the Center are saved. The loop data includes volume, occupancy, and an estimate of speed from the single-loop detectors in the field. Five minute summaries of the data are also created and saved. Previous telephone interviews with TMC personnel indicated that data storage may become a significant concern, and due to the large size of data files used for analysis, it is often difficult to manage the data sets with most desktop spreadsheet and database software.

Efforts at the Maricopa County DOT, in cooperation with the Arizona DOT, are developing a system to meet data archiving and access needs (7). The system is called the AZTech Data Server (ADS), and is shown in Figure 2. The formatted records at the top of the figure represent data inputs from different sources such as incident data or loop detector data. The data spooler/logger provides the quality control and screening of the data as it is received from the different sources and then logs the data. Data are then provided to ETAK, which is currently developing their traveler information service for the public, the ADS, and to other agencies that have a data feed. As shown in Figure 2, it is within the ADS that user-defined queries of the data can be performed for both on-line and off-line applications.

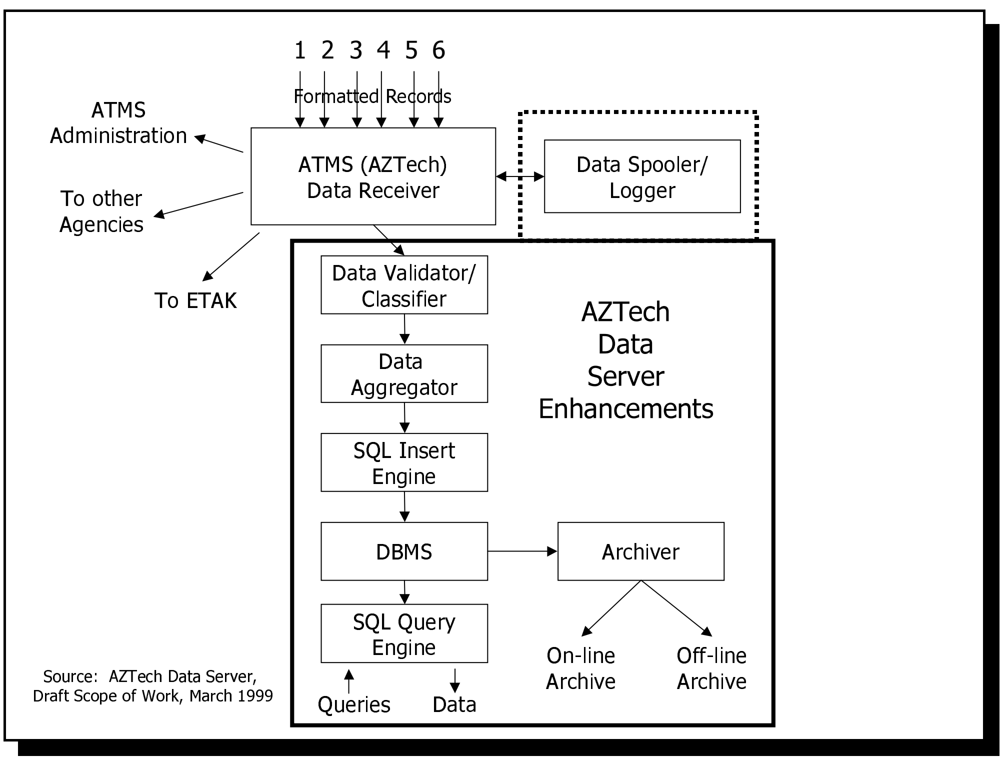


Figure 2. Proposed AZTech Data Server

While loop detector and incident data are currently the primary data inputs, it is anticipated that many other sources will also become available. The objective of the project is to "enhance the ADS to provide and maintain valid, classified ITS-derived data for use in planning, modeling, and real-time operation applications." Ultimately, the ADS will provide an interface for all data that enter the system. The project is funded at $500,000 through congestion mitigation and air quality (CMAQ) program funds and is scheduled for completion in September 2000.

The first steps of the project are underway. These include evaluating available user services, data screening and formatting issues, and appropriate validation and quality control checks. It is anticipated that different users will accept differing levels of data validity, and the system developers are considering rule-based methods (with thresholds defined by users) for screening the loop detector data. These data validity rules may be based upon Boolean operators on volume, occupancy, and speed for the data of interest to the user, or they may also allow the user to select statistical validity thresholds (e.g., 30 percent of the weekly average).

The Maricopa Association of Governments is performing a congestion study in the region utilizing data from the ADOT traffic operations center. Speed, volume, and vehicle classification data were desired on freeways and arterial streets. Aerial photographs were used to obtain vehicle density on freeways and queue lengths on arterial streets. One year of 5-minute traffic volume data were obtained from the ADOT center. Video was used to obtain classification data. Fifteen-minute volumes were recorded from the video data, and it was compared to the ADOT volume data. It was found that sometimes the data matched well, and other times the ADOT loop detectors were not operating or the detector volumes were low. The Association has been in contact with ADOT about these data quality issues and was informed that ADOT would take efforts to maintain the loop detectors.

Los Angeles, California

The Los Angeles (Caltrans District 7) TMC is in the process of developing a new system for archiving and accessing ITS data. The TMC currently polls their loop detectors every 30 seconds. Prior to the development of the new archiving system, the TMC saved three days of 30-second data and four days of five-minute summaries into temporary storage. The 30-second data have been permanently archived to magnetic tape cartridges since the TMC was opened, and Caltrans had developed special software that was necessary to retrieve this 30-second data from their mainframe computer.

The new system will provide on-line access to 13 months of data through a relational database for personnel in the Center. Beyond the 13-month time period, the data are planned to be archived to magnetic tape cartridges. In the future, as part of the new system development, it is anticipated that archiving to magnetic tape cartridges will be replaced by archiving the data to compact disc (CD). Future plans also include evaluating different aggregation levels for a given CD (e.g., having a CD for daily summaries, one for 15-minute data, etc.). This CD system would act as a CD data clearinghouse for interested users. If successful, the system being developed in Los Angeles will serve as the prototype for eventual implementation in all Caltrans TMCs.

Orange County, California

TMC personnel in Los Angeles are developing a new system that will be used at all TMC locations in California, including Orange County. It is their intent to save not only the 30-second data, but also 5-minute, hourly, daily, and monthly data on CD. In Orange County, volume and occupancy data are directly measured from the single loop detectors and speed is estimated.

San Francisco, California

There were not significant changes in the data archiving and management strategies at the San Francisco Bay Area TMC since the surveys were performed in the Spring of 1997. The TMC in San Francisco does not archive data at this time. Personnel at the TMC do recognize the importance of the data for future transportation applications, and they are interested in data archiving activities. They have recently submitted a proposal to use State Planning and Research Funds to evaluate archiving activities and automated techniques.

Atlanta, Georgia

Data archiving practices at the Geogia DOT Advanced Traffic Management System (ATMS) have not changed since the original survey was performed in the Spring of 1997. Volume and occupancy, along with an estimate of speed, are sent from video detection cameras to the ATMS every 20 seconds. The data are not currently saved. Evaluating data archiving issues is of interest to personnel at the ATMS, and efforts to evaluate these considerations will be investigated at a later date.

Chicago, Illinois

The Illinois Traffic Systems Center (TSC) is another TMC within which archiving and data retrieval issues are being strongly re-evaluated. The Center is currently obtaining volume and occupancy data and estimating speeds from approximately 2,000 single-loop detectors that cover 210 km (130 miles) of freeways in the area. The detectors are located on all lanes every 4.8 km (3 miles) and at the entrance and exit ramp locations. At approximately 0.8 km (0.5 mile) intervals, there are detectors in the center lane from which lane occupancy is obtained and travel time is estimated. The 20-second data are aggregated to the five-minute level, and only the travel time and occupancy data are saved at this level from 5 to 10 a.m. and 2 to 7 p.m. Hourly volume data across all lanes are saved from the detector stations spaced at three miles. Both the five-minute and hourly data are permanently archived to magnetic tape cartridges.

The Center personnel have an interest in using the available data for many secondary uses and want to be able to support the demand. Primary users of the data include in-house requests for lane closures and performance evaluation, the local metropolitan planning organization (MPO), media, and researchers. They have recently hired a consultant to evaluate the data retention and archiving efforts in the Center. Many issues are being considered in this effort including determining what data should be saved, at what level it should be saved, what medium should be used for storage, and what type of automation should be developed. It is anticipated that about one year of data is desired on-line, and a database management system that saves the data to CD, rather than tape cartridges, is also desired. One of the first tasks for the contracted consultant is to evaluate the potential uses and users of the data.

Montgomery County, Maryland

The TMC in Montgomery County, Maryland collects one-minute volume and speed data from loop detectors, then aggregates to the five-minute level before saving. The loop detectors are located on arterial streets downstream of signalized intersections. This information is used to assist in determining offsets and splits for signal timings along the arterial streets.

The Center personnel also have an interest in developing a more strategic ITS data archiving plan. They are currently working on a data archiving effort funded at $100,000 to identify users and uses of the data, a least common denominator of data aggregation to satisfy these uses, and a method or system to query data of interest about a particular roadway link. Work is being performed to supply data to the local Maryland-National Capital Park and Planning Commission (M-NCPPC). This data application is discussed on page 29 of this report.

Detroit, Michigan

The Michigan ITS (MITS) Center is another TMC that is developing standards on an aggregation level and determining an on-line data access strategy. The Center contains two loop detector systems. The older system covers 52 km (32 miles) of freeway with single loops located at nominal 0.54 km (0.33 mile) spacings. The detectors are scanned by the central traffic control system every 10 milliseconds and the data are aggregated to hourly lane volumes, which are saved to magnetic tape. The newer system covers 240 km (150 miles) of freeway with double-loop detectors (i.e., double loops provide spot speed estimates by using the time difference of arrival) at nominal 3.2 km (2 mile) spacings. These data are aggregated in the field at 20-second intervals and then sent to the Center. The data are then aggregated up to one-minute and the volume, occupancy, and speed are saved. Personnel at the Center hope to eventually fill in additional dual-loop detectors between the existing loops in the newer system. The Center keeps one week of data on-line at the one-minute aggregation level. Data older than one week are saved to magnetic tape cartridges.

Personnel at the MITS Center are interested in developing better access to archived data for a variety of users and uses. They are evaluating the possibility of a common aggregation level to satisfy all users as well as trying to develop a more efficient on-line data access strategy. Center personnel commented that this is difficult since different users inevitably require different types of data at different aggregation levels. They would also like to incorporate their changeable message sign (CMS) log into this data set. The CMS log includes data such as the time a message was posted or removed, as well as the message content for a given incident. Personnel at the Center also expressed the desire to integrate information that comes into the Center through public telephone and the highway advisory radio (HAR) into the database. They will likely consider CD or digital video/versatile disk (DVD) for archiving rather than the magnetic tape cartridges that are currently used.

Loop detector failures are another concern at MITS. In the hourly volume summaries that are saved, a measure of the "lane operability" is also saved. The measure is the loop-minutes of failure, which is calculated as the number of loops failing times the number of minutes failing. When making temporal or spatial comparisons, this measure provides some insight into the extent to which a loop detector(s) may not be reporting. For example, if the loop operability equals nine, this could mean that three loops are out for three minutes, or one loop is out for nine minutes. This does not evaluate the extent that volumes may be reading too high or too low, and these data quality issues are also a concern at the Center. To date, the primary users of the archived data are in-house for freeway management.

Minneapolis-St. Paul, Minnesota

The Mn/DOT TMC monitors over 85 percent of the freeways in the Minneapolis-St. Paul metropolitan area. The most significant source of ITS data is from 3,800 loop detectors. The detectors provide volume and occupancy. The speeds can be estimated based on an assumed vehicle length. These data are obtained every 30 seconds and every five minutes. The five-minute data are aggregated and saved. Thirty-second station data is stored in a binary day file and is available upon request. Daily five-minute information is stored to a CD and about six months of data can be saved to one CD. The process is automated and the data have been saved since a computer conversion in 1993. To date, only the loop data have been archived; however, there is interest in eventually saving ramp metering data (e.g., vehicle release rates) as well.

Most requests for the archived data are from the DOT for traffic analysis, construction impact determination, and planning applications. Researchers from local universities often request the data as well. Another significant use of the data is for the evaluation of ramp metering strategies. This use has fostered an interest in a more detailed study and calibration of the loops. This effort is explained in greater detail on page 34 of this report.

TRANSCOM, New York/New Jersey/Connecticut

TRANSCOM (the Transportation Operations Coordinating Committee) opened in 1986 to coordinate traffic management in New York, New Jersey, and Connecticut. TRANSCOM's System for Managing Incidents and Traffic (TRANSMIT) utilizes vehicles with automatic vehicle identification (AVI) transponders on the vehicles. Data archiving practices at TRANSCOM had not changed since they were contacted as part of the original survey in 1997.

Raw tag reads for each vehicle are not saved, and software at the TMC saves the data into 15-minute periods.

Ft. Worth, Texas

The TransVISION TMC in Fort Worth, Texas is scheduled to open by September 1999. The TransVISION system is currently under development, and many questions are still being addressed. These questions include (8):

- What type of data is available?
- Do we need to save it all?
- Will it add value to our TMC system operations?
- In what form should that data be saved?
- How do we share this data with other agencies in our area, including the City of Arlington, the Dallas District, and others?

The ultimate system design for TransVISION includes a relational database for on-line storage of archived ITS data; however, the funds are not yet available for building this data archiving component. The data anticipated in the center include traffic incident information, ramp meter timing, dynamic message sign (DMS) stored messages, lane control signal (LCS) patterns, geographic information system (GIS) data, traffic signal timing, closed circuit television (CCTV) control functions, electronic work orders, detection system data, road construction, maintenance, and special events data. These data will be organized and logged around four areas including incident data, closure data, operator data, and detector data.

There are five major subsystems to the TransVISION center: communications, video, dynamic traffic management, data management, and the operating platform. The data management subsystem architecture is shown in Figure 3. The subsystem includes both an on-line transaction processing (OLTP) server and an on-line analytical processing (OLAP) database server. The OLTP server provides data storage for the most recent month. At periodic intervals, data will be replicated (i.e., duplicated and verified) from the OLTP server to the OLAP server. After about one year, the data from the OLAP server will be archived to permanent storage media, such as a CD. As additional financial resources become available, TransVISION personnel would like to expand the system beyond the 30-day limitation within the OLTP server.

The OLTP server manages the relational database and the real-time operations. The most recent information is held in the OLTP, and this information is updated in real-time. The OLTP server will provide short-term storage of about one month of data. With more than 2,000 loops, 100 bytes per loop, and 1 record every 50 seconds, the data load in the OLTP is about 10.3 gigabytes per month. When implemented, the OLAP server will interact automatically with the OLTP server to perform data replication (i.e., duplication and verification). The OLAP database will also provide for queries of the data that are available for the last year, and the database will permanently archive data to either CD or magnetic tape cartridges.

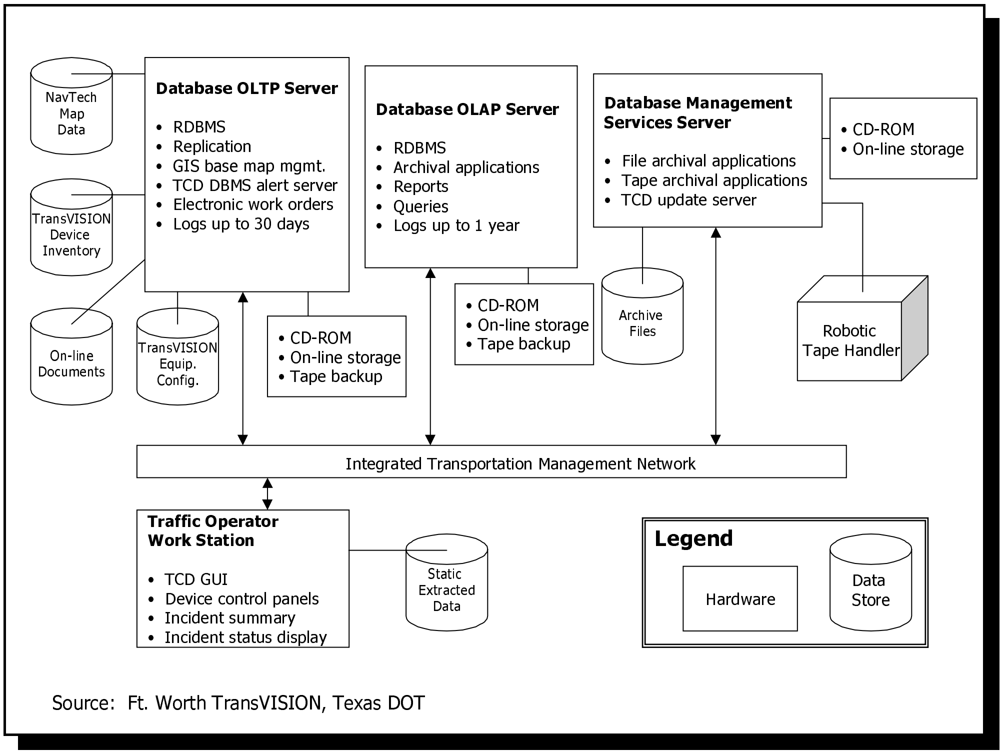


Figure 3. Proposed Data Management System in Ft. Worth's TransVISION

Many issues are still being considered in the development of the TransVISION data archiving system. These include:

- Is the current design direction adequate?
- Does new technology cause a change in direction for archived storage?
- How much do we aggregate the data and how long do we keep it?
- Which specific data will other agencies be interested in?
- How can others connect to this data considering security, system load, etc.?

Houston, Texas

The Houston TranStar Center provides traffic management for the Houston region's freeway system. Inductance loops and video detection units are used on a portion of the freeway system, but the data from these devices are not archived. Houston also has an extensive AVI system in operation. Over 300,000 AVI tags are in circulation in the metropolitan area, and tag reads are sent to the Center in real-time. The AVI travel time data are stored in 15-minute summaries for future use. In addition, AVI data have been used for quantifying the benefits of ramp metering and high-occupancy vehicle (HOV) lanes.

In a recent ITS data archiving workshop in Texas, information resource personnel for TxDOT's Houston District and TranStar discussed current data warehousing development activities (9). The information resource manager discussed how the statewide ITS integrator, Southwest Research Institute (SwRI), can help TranStar define user requirements for the development of an enterprise data management system. Approximately $1.2 million is currently estimated for this data warehousing system, and one terabyte of data storage is being planned.

The information resource administrator noted that some data warehouse efforts have failed because the system was not designed to meet end user requirements. Therefore, the determination of end user requirements by SwRI is an important component of the project effort. Table 3 illustrates an example of a matrix of user requirements that will be developed for TranStar's data warehouse.

Table 3. Example of User Requirements Matrix for Houston TranStar

User Requirements	**Functions**	**Agency**	**Type of Data**
• List of data elements based on user requirements	• Planning • Construction • Operation • Maintenance	• City • County • TxDOT • Transit • MPO	• Planned or Proposed
User Requirements Questions: • Who is the end user? • Where are they? • What data do they need? • How are they using it? • Does it meet their needs? • Is it still needed?			

Source: adapted from Wegmann and Gloyna 1998 (9)

As a precursor to the enterprise-wide data warehousing plans, the Houston-Galveston Area Council (HGAC) is working with Houston TranStar on a pilot project that will provide HGAC with archived ITS data for their planning and modeling applications. The project, which was funded through the ITS Priority Corridors program, is focused on archiving loop detector data from selected locations in Houston. This pilot effort began in late 1998.

San Antonio, Texas

Phase One of the TransGuide Center, complete since 1995, has double-loop detector coverage over 42 km (26 miles) of freeways at nominal 0.8 km (0.5 mile) spacings. TransGuide also has an AVI system that covers about 98 miles of freeways and limited access highways.

TransGuide computer servers poll the local loop controllers every 20 seconds, and this 20-second data is made available over the Internet (http://www.transguide.dot.state.tx.us/statistics.html) at the end of every day. Providing the data via the Internet reduces the time spent distributing data that is requested. The Internet site typically contains the most recent month of loop detector data. Data are archived to tape, but it is anticipated that storage will be moving to CD in the near future.

Data that are collected at TransGuide can be categorized into five types (10) as described below, with the operational data being the most common type made publicly available:

- **operational data -** real-time traffic data from the loop detector and AVI systems, scenario execution logs, incident information, and the database of nominal travel times on non-instrumented routes.

- **real-time procedural data -** user access logs, system accounting records, and resource utilization logs.

- **performance data -** system error logs, system access logs, and database performance monitoring logs.

- **system data -** system data include the source code, executable programs, design documents, equipment configurations, and network management databases.

- **administrative data -** operator logs, equipment maintenance records, and system configuration data.

TransGuide is committed to providing the data that are available through the TMC in an accessible and efficient manner as encouraged in TxDOT's Information Sharing Policy.

Data from the TransGuide system have been used by TTI researchers for the development of an ITS data management system (DataLink) (5,6). DataLink provides a user-friendly interface capable of querying the TransGuide data for different roadway facilities, time periods, and aggregation levels. Such a system provides a common repository of information for individuals that may want to utilize the information for different applications (e.g., operations, incident management, planning). Researchers at TTI are also using incident detection and response logs from the TransGuide system to evaluate their incident management program.

Seattle, Washington

The data archiving strategies at the WSDOT TMC in the Puget Sound region have not changed since personnel were contacted for the first survey in 1997. Volume, occupancy, and an estimate of speed are collected and sent to the ATMS every 20 seconds. The loop detector data are stored at the five-minute aggregation level and it currently requires four CDs to save one year of five-minute data.

Toronto, Ontario, Canada

COMPASS is a traffic management system that monitors portions of Highway 401 in Toronto, Canada. The center monitors over 2,800 loops as well as numerous video surveillance cameras. The loop detectors are polled every 20 seconds and report volume, occupancy, and an average speed from an assumed vehicle length for the single-loop detector stations. Whenever maintenance is required of the highway pavement, double-loop detectors are being installed to replace the single-loop detectors.

The 20-second data are aggregated to five-minute, 15-minute, one hour, daily, and monthly time periods. The TMC archives speed, volume, and occupancy data for 20-second and five-minute time increments. For data summaries of 15 minutes or more, only volume data are saved. The five-minute time increment was selected because it appeared to provide a convenient time increment for many users. Common archived data users and uses include in-house requests and researchers, who desire 20-second data for simulation and algorithms. In-house requests account for about 60 percent of all data requests (e.g., traffic forecasts, roadway impact analyses).

There are other systems in operation is the Toronto area as well. The COMPASS system is along Highway 401 as described above. Two additional COMPASS systems are along the Queen Elizabeth Way. One TMC is located in Mississauga and the other is in Burlington that monitors the skyway bridge. The later two systems collect data at 30-seconds and the data are archived at the same aggregation levels as the Highway 401 COMPASS system described above.

The data have traditionally been archived on 8-mm magnetic tape cartridges. The tapes require special hardware and software to download the data, and data users found it relatively difficult to obtain historical data from this system. Further, over extended periods of time (i.e., greater than about five years), the tapes begin to lose reliability and potentially data.

Since about 1997, the COMPASS system has been saving the data to CD, and they have been developing software to extracting data from a CD. The CD itself is being developed to contain software that will create and display tabular summaries of interest to the user. COMPASS personnel are currently assessing the appropriate and most useful format and aggregation level for the CDs for a variety of users and uses (i.e., should the CD contain only 20-second or hourly data?). This decision will likely be made depending upon the demand for the CDs that are being developed at the different aggregation levels. One other difficulty of the CD application is that it operates slowly.

The sharing of archived ITS data to numerous widespread users is a relatively new effort at COMPASS and they are in the process of developing information use and sharing policies. The experience at COMPASS illustrates the interest in designing systems that promote more efficient storage and retrieval of archived ITS data.

State of Virginia

The Virginia DOT (VDOT) is developing an enterprise data warehouse system that will integrate the nine legacy systems within the Department (11). One of the many data sources being considered in this integration are archived ITS data. The intent of the data integration is to bring all of the different systems together into one unified data system. The nine legacy systems of the Department operate on four different database systems. Adding to the complexity, there are three hardware platforms and three operating systems in use as well. The system under development at VDOT will allow for the screening, quality control, and summarization of the data as it is input into the data warehouse. The enterprise data model describes the planning data that are the most common throughout the agency and serves as the foundation for the data integration effort. Fully understanding the data uses prior to system design is important.

VDOT stresses the importance of integrating the databases that are currently available for different uses. Without full integration, and appropriate connections between the different databases, “data islands” may occur. These are databases that cannot be accessed through the integrated system and are left out of the larger archived data system. This is especially important to ensure that database systems provide regional integration and can share information.

Another observation that is clear with the VDOT experience is that the Department has assigned significant resources for the data warehouse project. This demonstrates the importance of these growing concerns and the recognition that providing archived ITS data and information through an integrated system is a necessary role of transportation agencies. VDOT has assembled a collective group of representatives collectively know as the MIS 2000 committee that is assigned to act as a functional enterprise level authority for different functional areas and different districts. By functioning across different divisions and districts, it can also help in strategic systems planning efforts. In addition, a configuration control board has been established that is composed of personnel from both VDOT and TransCore to provide input into technical aspects of the data warehouse. Data custodians have also been assigned to monitor data for the data warehouse that is specific to a transportation subject matter. Finally, a data management division has been staffed with data management professionals to provide leadership on the technical aspects of the data management. These groups all provide valuable insight into the development and continued success of a data management system that incorporates data from many sources into one comprehensive data warehouse.

North America Pre-clearance and Safety Systems

The NORPASS (North America Pre-clearance and Safety Systems) program has been recently established to assist in the interoperability of safe and efficient pre-clearance of trucks at weigh/inspection stations,. NORPASS merges the CVO project based at the Kentucky Transportation Center (Advantage CVO) and the Multi-jurisdictional Automated Preclearance System (MAPS) programs as well as other states and provinces. CVO provides a unique ITS data source that could have many secondary applications (12). However, there are considerable privacy issues that arise when using WIM and AVI transponders. Truckers are concerned that the data can potentially be saved with such a system and used against them for enforcement. Therefore, privacy concerns are high in the use of CVO data. In fact, one of the largest issues facing the electronic screening programs is ensuring that the operations do not provide stricter enforcement on vehicles equipped with transponders and to ensure that the carriers are kept anonymous. To assist in relieving such concerns, Advantage CVO prepared a data policy, which is illustrated in Figure 4.

Available CVO data typically include the carrier and truck identification, trip history, weight, and vehicle classification. From adjacent weigh-in-motion (WIM) detectors, data such as the date and time stamp, axle and gross weight, vehicle classification, and speed are collected. As stated in the policy, Advantage CVO does not retain any carrier-specific or vehicle-specific data for over 24 hours. One common concern among truckers is that the date/time stamp will not be consistent with their log books. The second concern is that the speed measured at consecutive WIM sites will be used for speed enforcement. However, these data are not saved for over 24 hours or used for enforcement purposes. There is not as much concern over the actual weight data that are recorded because that is what is traditionally recorded at the weigh stations and is understood by the truckers. The ITS America CVO Policy Subcommittee has adopted “Fair Information Principles for ITS/CVO” and the “ITS/CVO Interoperability Guiding Principles” that describe their position on secondary uses of CVO data (available at http://www.itsa.org/cvo.html).

PRESS RELEASE
Tuesday, August 25, 1998

Advantage CVO adopts data policy to ensure continued protection of motor carrier privacy.

On July 29, 1998, the Advantage CVO Policy Committee approved an **Operations Center Data Policy** which specifies procedures for the collection, handling and distribution of electronic data. Under this policy, no carrier-specific or vehicle-specific data will be retained any longer than 24 hours, either at weigh stations or at the Advantage Operations Center. Participating state agencies will be provided with summary data that is equivalent in nature to the data collected manually at that state's weigh stations. No additional data will be collected as a result of a carrier participating in electronic screening.

Since Advantage CVO does not charge transaction fees for participating carriers, there is no need to collect and retain transaction data for billing purposes. This allows the Operations Center to discard data on a daily basis, or even more frequently if it becomes necessary.

In the four years since Advantage CVO first began operations, the system has granted over half a million green lights (i.e., weigh station bypasses) to participating motor carriers on Interstate-75 and Canadian Route 401. During that time, there has never been even a single instance of transaction data collected by Advantage CVO being used for enforcement purposes. Even with this outstanding track record, the Advantage CVO Policy Committee chose to implement a formal policy to reassure participating carriers that data will not be used inappropriately in the future.

The Advantage CVO Policy Committee includes representatives from State/Provincial agencies in Ontario, Michigan, Ohio, Kentucky, Georgia, and Florida, as well as from the National Private Truck Council and several major trucking companies.

For more information contact:
Joe Crabtree, Kentucky Transportation Center
606-257-4508, fax 606-257-1815, e-mail crabtree@engr.uky.edu

Figure 4. Press Release Describing Advantage CVO Data Policy

STATE-OF-THE-PRACTICE IN USING ARCHIVED ITS DATA

The previous section described existing ITS data archiving practices at selected TMCs and other operations centers. Personnel from many of the TMCs indicated that they receive data requests for a variety of transportation applications. There are numerous stakeholder groups for ITS data archiving because of the utility and variety of ITS-generated data items. Table 4 summarizes the key stakeholder groups and their primary applications for archived ITS data. These stakeholders and their applications may help in determining appropriate designs for ITS data archiving systems. This section describes several of the stakeholders shown in Table 4 in further detail, including discussion of the following:

- typical data elements of interest;
- spatial/temporal aggregation levels of interest;
- additional data characteristics of importance; and
- example applications.

The applications highlighted in this section represent the most common applications of archived ITS data to date; therefore, example applications are illustrated for a few of many several stakeholder groups.

Table 4. Stakeholders and Example Applications for Archived ITS Data

Stakeholder Group	Primary Transportation-Related Functions	Example Applications
MPO and state transportation planners	Identifying multimodal passenger transportation improvements (long- and short-range); congestion management; air quality planning; develop and maintain forecasting and simulation models	• congestion monitoring • link speeds for TDF and air quality models • AADT, K- and D-factor estimation • temporal traffic distributions • truck travel estimation by time of day • macroscopic traffic simulation • parking utilization and facility planning • HOV, paratransit, and multimodal demand estimation • congestion pricing policy
Traffic management operators	Day-to-day operations of deployed ITS (e.g., Traffic Management Centers, Incident Management Programs)	• pre-planned control strategies (ramp metering and signal timing) • highway capacity analysis • saturation flow rate determination • microscopic traffic simulation -- historical -- short-term prediction of traffic conditions • dynamic traffic assignment • incident management • congestion pricing operations • evaluation and performance monitoring

Table 4. Stakeholders and Example Applications for Archived ITS Data (Cont.)

Stakeholder Group	Primary Transportation-Related Functions	Example Applications
Transit operators	Day-to-day transit operations: scheduling, route delineation, fare pricing, vehicle maintenance; transit management systems; evaluation and planning	• capital planning and budgeting • corridor analysis planning • financial planning • maintenance planning • market research • operations/service planning • performance analysis planning • strategic/business planning
Air quality analysts	Regional air quality monitoring; transportation plan conformity with air quality standards and goals	• emission rate modeling • urban airshed modeling
MPO/state freight and intermodal planners	Planning for intermodal freight transfer and port facilities	• truck flow patterns (demand by origins and destinations) • HazMat and other commodity flow patterns
Safety planners and administrators	Identifying countermeasures for general safety problems or hotspots	• safety reviews of proposed projects • high crash location analysis • generalized safety relationships for vehicle and highway design • countermeasure effectiveness (specific geometric and vehicle strategies) • safety policy effectiveness
Maintenance personnel	Planning for the rehabilitation and replacement of pavements, bridges, and roadside appurtenances; scheduling of maintenance activities	• pavement design (loadings based on ESALs) • bridge design (loadings from the "bridge formula") • pavement and bridge performance models • construction and maintenance scheduling
Commercial vehicle enforcement personnel	Accident investigations; enforcement of commercial vehicle regulations	• HazMat response and enforcement • congestion management • intermodal access • truck route designation and maintenance • truck safety mitigation • economic development
Emergency management services (local police, fire, and emergency medical)	Response to transportation incidents; accident investigations	• labor and patrol planning • route planning for emergency response • emergency response time planning • crash data collection
Transportation researchers	Development of forecasting and simulation models and other analytic methods; improvements in data collection practices	• car-following and traffic flow theory development • urban travel activity analysis
Private sector users	Provision of traffic condition data and route guidance (Information Service Providers); commercial trip planning to avoid congestion (carriers)	

Source: Margiotta 1998 (4), pp. 4-5.

Metropolitan Planning Organizations / State Transportation Planners

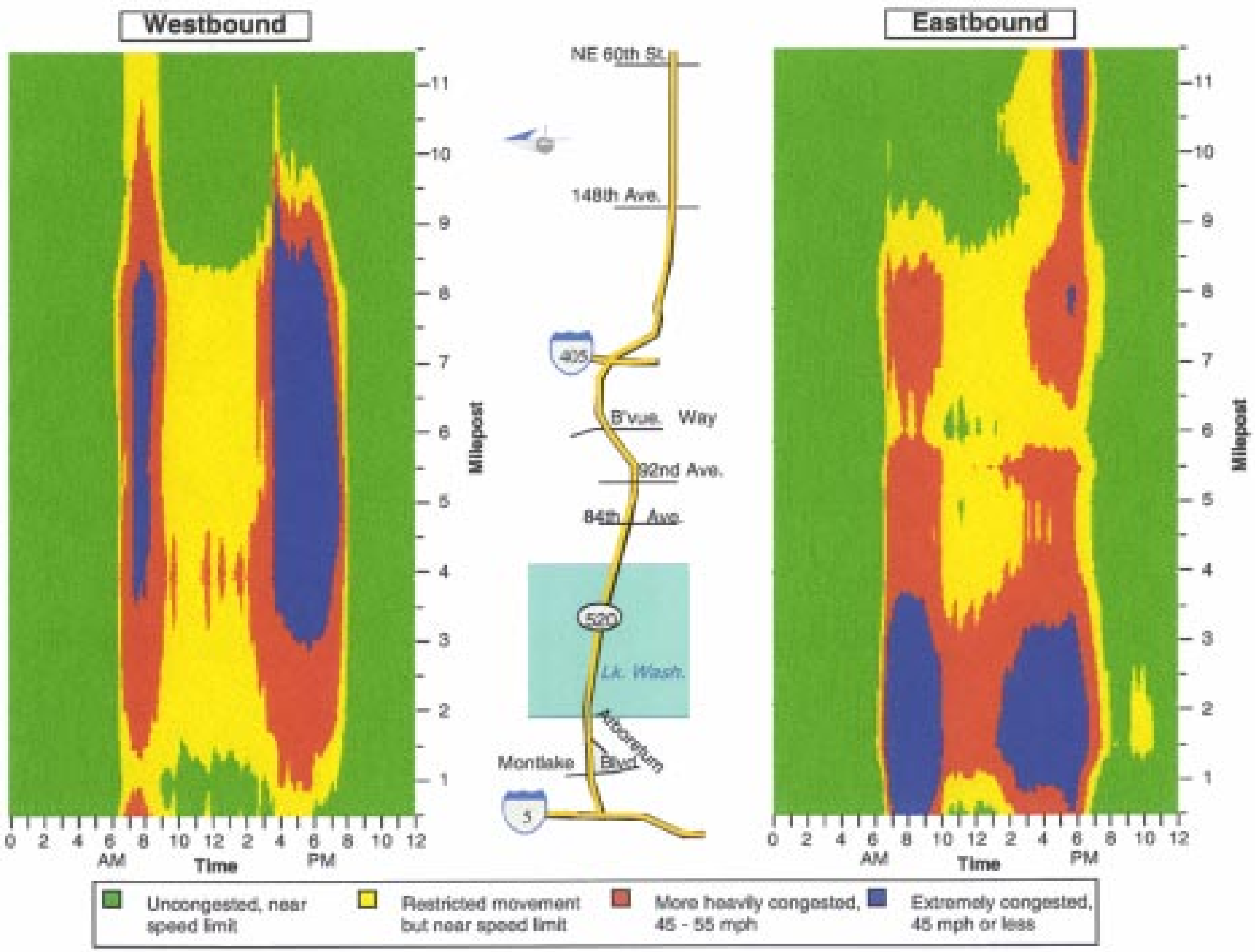
The first stakeholder group identified in Table 4 are MPOs and state transportation planners. These users commonly require ITS data for uses such as congestion monitoring, system performance evaluation, determination of average traffic statistics, and multi-modal demand estimation. Transportation planning applications often require data at a relatively aggregate level. For example, hourly or even daily summaries of volume data and hourly summaries of speed data are used for long-range planning, system performance evaluation, or congestion monitoring. Most planning applications typically do not require data at a more disaggregate level than 15 minutes. Many planning applications (e.g., congestion management, system performance evaluation) require data from several different segments and locations throughout a region. These applications generally require daily summary statistics; however, peak hours and peak periods are also often of interest. The remainder of this section will describe applications of archived ITS data for planning applications.

Example: Freeway Performance Evaluation in Puget Sound, Washington

Researchers at the Washington State Transportation Center (TRAC), in cooperation with WSDOT, have prepared a report on facility usage and freeway performance in the central Puget Sound area using 1997 archived ITS data (13). Data for the study were obtained from the WSDOT FLOW system inductance loop detectors. These detectors are placed at approximately 0.8 km (0.5 mile) spacings and obtain volume and occupancy data. Speeds are estimated from the single detectors. These data are aggregated at five-minute intervals and archived to CD.

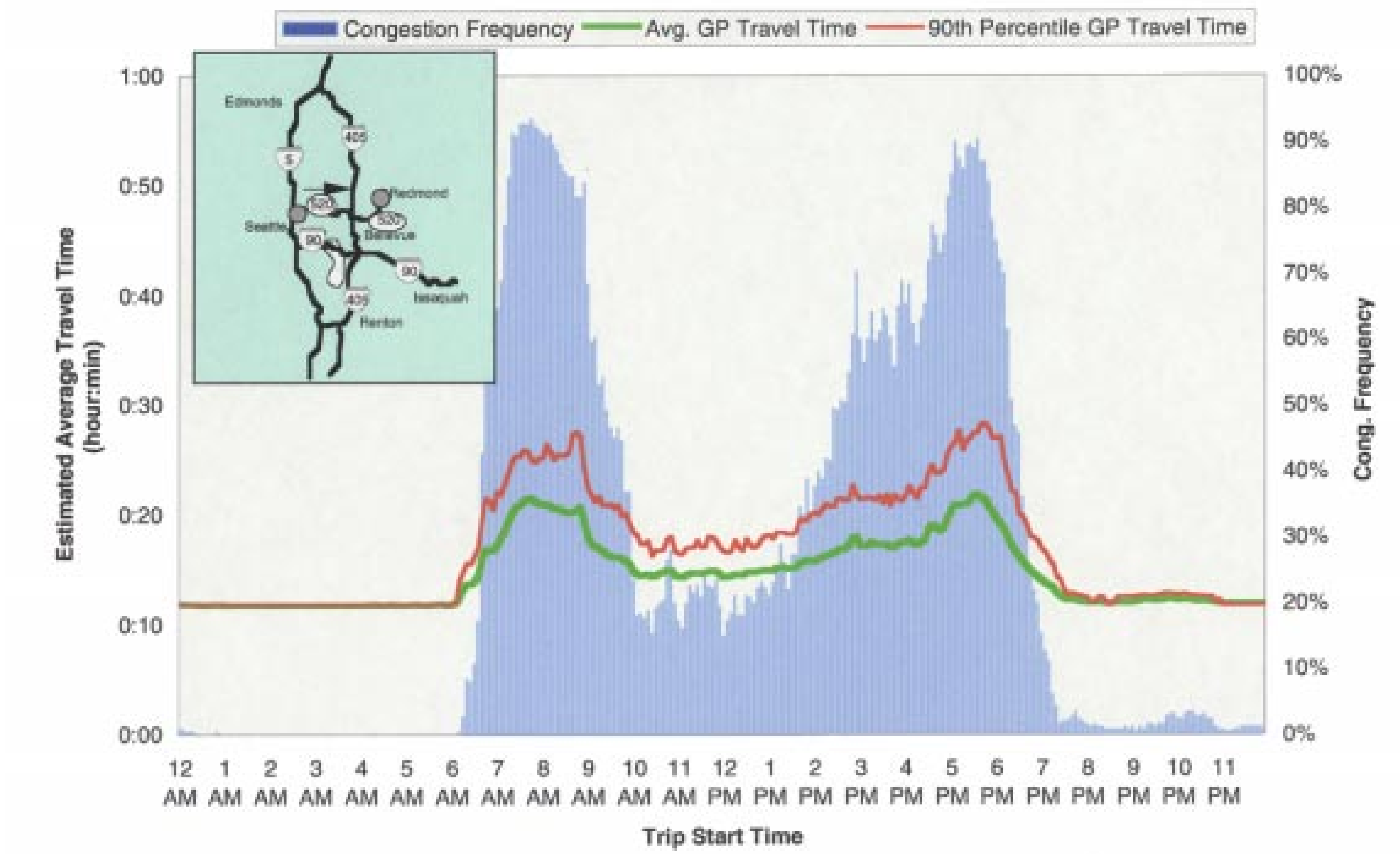
The results of the WsDOT FLOW evaluation produce valuable information for system performance and congestion management by monitoring traffic volumes throughout the network and reporting vehicle speeds and when and where congestion is occurring. Facility usage data are presented for various facilities for the peak period and peak hour. The presentation methods in the evaluation were designed to illustrate temporal and spatial aspects of congestion for non-technical audiences (e.g., policy makers, legislators). For example, Figure 5 presents the congestion level in each direction of travel along SR-520.

Figure 6 presents the results of average facility travel times for a given hypothetical travel route on the general-purpose lanes. The figure shows the average general-purpose travel time, the 90th percentile general-purpose travel time, and the percent of time the speed is less than 45 mph (66 km/hr) for a given trip start time (congestion frequency). Figure 7 shows a similar graphic that presents the congestion frequency by time of day along with the corresponding traffic volumes. The line representing the volume is presented in colors that indicate the speed of the traffic as well.



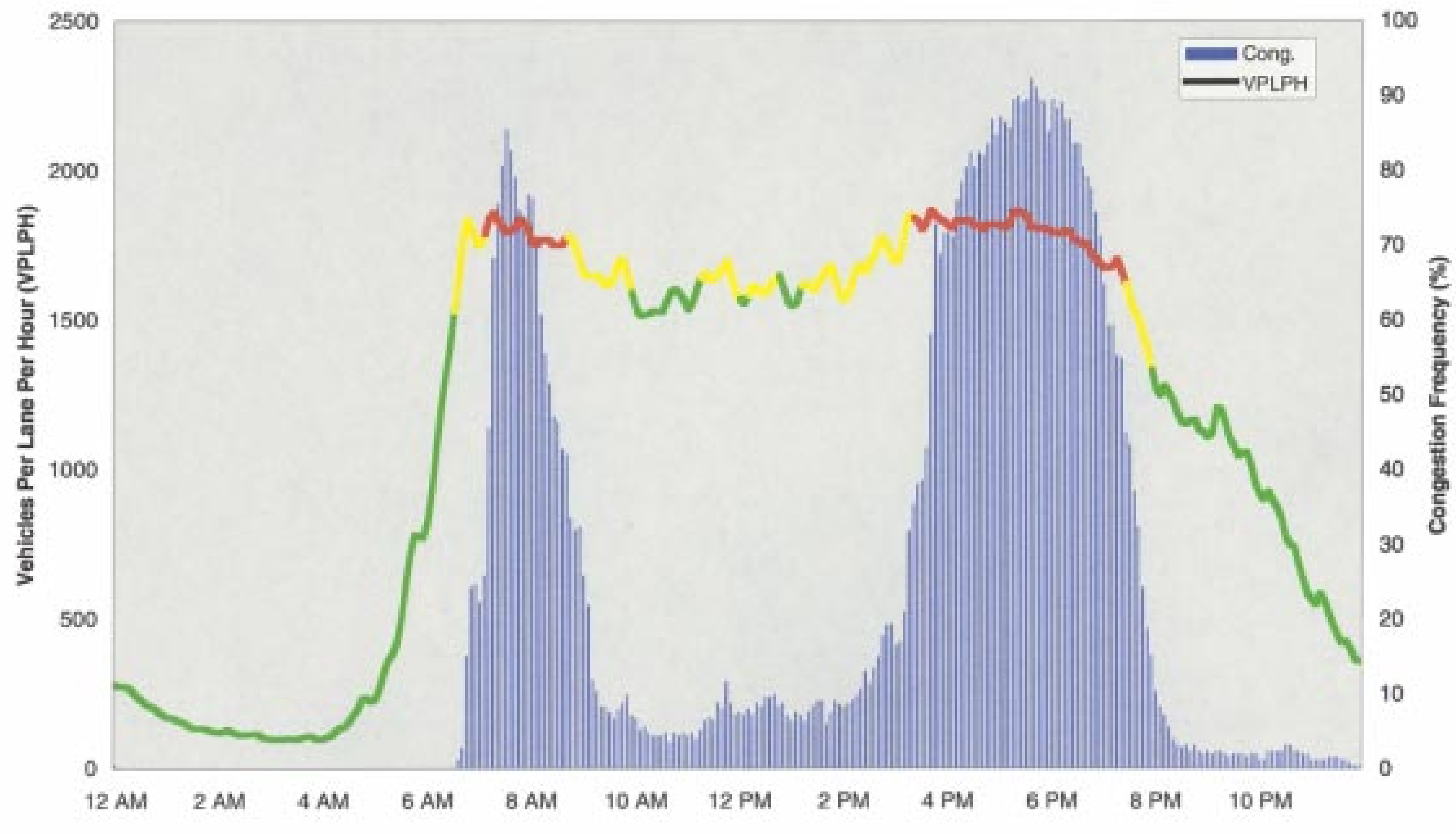
Source: Ishimaru and Hallenbeck 1999 (13)

Figure 5. Traffic Profile of SR-520 General Purpose Lanes



Source: Ishimaru and Hallenbeck 1999 (13)

Figure 6. Estimated Average Weekday Travel Time for Eastbound SR-520



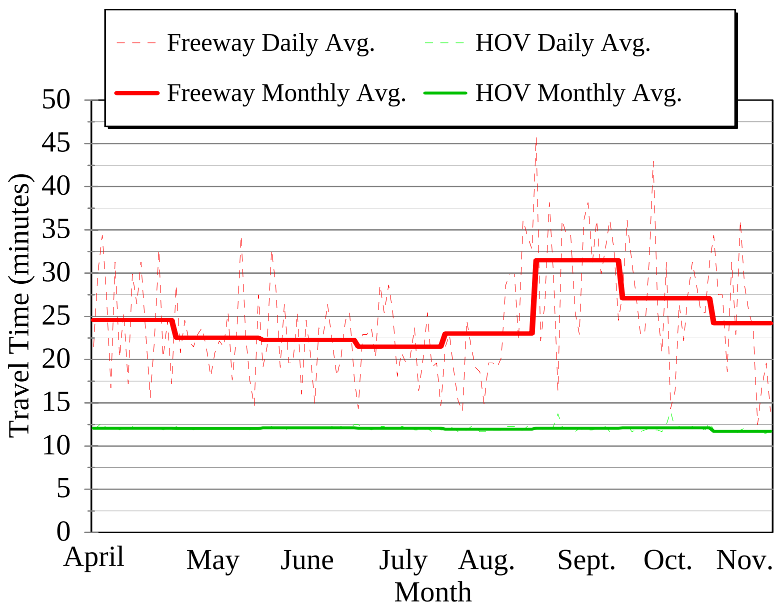
Source: Ishimaru and Hallenbeck 1999 (13)

Figure 7. Estimated Weekday Volume, Speed, and Reliability Conditions for Westbound SR-520, 76th Avenue N.E.

Example: Evaluation of Houston HOV Lanes

TTI researchers evaluated three HOV corridors in the Houston area using archived AVI data from 1994 (14,15). Eight months of the AVI probe vehicle data were used to determine travel time savings and reliability along three major freeway and HOV corridors.

The results of the travel time analysis indicated that the HOV lanes provided a substantial travel time savings over the adjacent freeway general-purpose lanes. Savings were proportional to the level of congestion being experienced in the adjacent freeway general-purpose lanes. The extensive data set available with the AVI traffic monitoring system provided detailed information about day-to-day speeds and variability of speeds. For example, Figure 8 shows large variations in daily peak hour travel times for the freeway general-purpose lanes with small variations on the HOV lane. Travel time variability was evaluated with measures of standard deviation and confidence intervals. For example, Figure 9 supports one of the asserted benefits of HOV lanes, in that they provide higher levels of travel time reliability than freeway general-purpose lanes. The results of this study provided valuable information for HOV system performance measurement.



Source: Turner 1997 (15)

Figure 8. Average Morning Peak Hour Travel Times for Katy Freeway, 1994

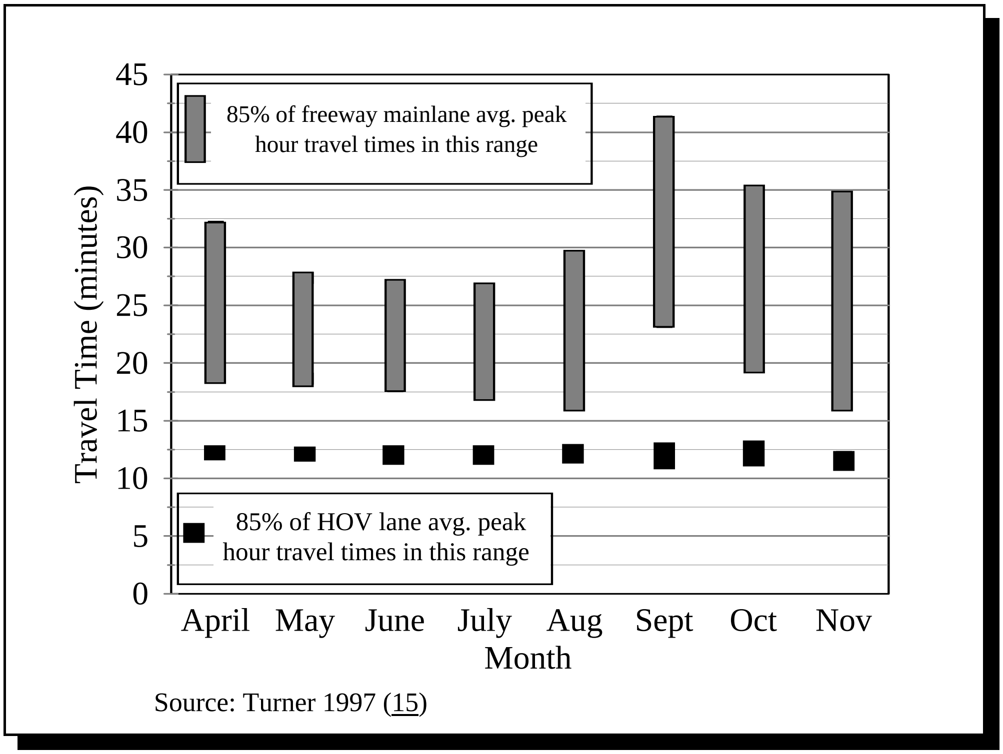


Figure 9. Morning Peak Hour Travel Time Reliability for Katy Freeway, 1994

Example: National Traffic Performance Indicators

Research sponsored by the Bureau of Transportation Statistics (BTS) is being performed at the Oak Ridge National Laboratory to develop real-time traffic indicators using data collected by ITS sensors (16). An Internet site documents this research effort at http://micron4.ed.ornl.gov/. This research effort is exploring the possibility of using data available from TMCs throughout the country to develop national performance indicators in real-time. The research is using total vehicle miles of travel (VMT) and total vehicle-hours of delay as performance measures. An automated process has been developed to obtain and process ITS traffic condition data from four cities in the U.S.: San Diego, California; Milwaukee, Wisconsin; San Antonio, Texas; and Seattle, Washington. The authors have noted many useful results (16):

- The minute-by-minute data for individual cities reveal that delay is highly sensitive to small changes in system average speeds;
- The individual cities' system-wide travel and congestion indices show well defined daily and weekly patterns;

- There is a need for greater consistency in definitions, methods, and conventions for ITS data;

- There is a need for further addressing the following technical issues: error handling, data capture, statistical inference, and data reduction and storage; and

- There are many possible uses of a developed national set of indicators, such as quantifying congestion trends and monitoring other congestion attributes like delay per VMT or throughput per lane-mile.

Example: Traffic Statistics in Chicago, Illinois

Planners at the Chicago Area Transportation Study (CATS) have been working with personnel at the Illinois TSC to use 1995 archived ITS data for planning applications (17). CATS has used the data available from the TSC to develop a travel atlas for 1995 for the Chicago metropolitan area. The document provides a valuable resource to planners for information such as annual average daily traffic (AADT) volumes. Additional information is illustrated in the document, including monthly seasonal factors, day-of-the-week traffic variation, holiday travel rates, hourly variation of traffic, and estimated travel times and speeds.

Example: Traffic Statistics and Congestion Monitoring, Montgomery County, Maryland

Personnel at the Montgomery County TMC collect arterial street traffic volume and speed data and are working with staff at the Maryland-National Capital Park and Planning Commission (M-NCPPC) to supply them with traffic volume data. The driving force for such an effort is limited budget allocations for data collection programs. Planning personnel see the use of archived ITS data as an opportunity to save money in the data collection effort. Increasingly, there appears to be more interest from operations personnel in developing working relationships to facilitate the coordination of such efforts.

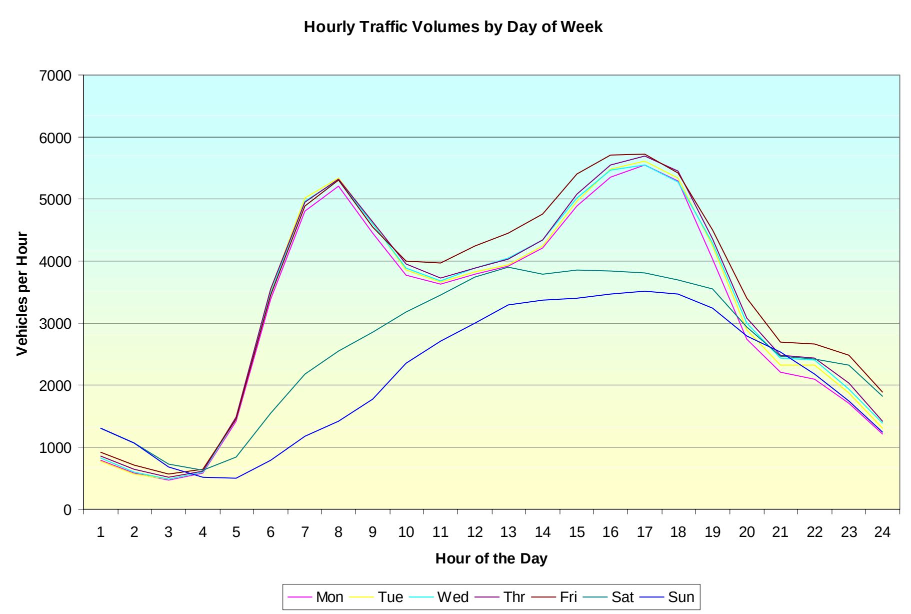
The M-NCPPC has worked closely with TMC staff that are performing data archiving activities. M-NCPPC planners view the archived ITS data as an opportunity to obtain traffic volume data for primary arterial facilities in the area at a reduced cost and increased safety. Funding has been secured and personnel are developing a prototype interface for displaying volumes and speeds near intersections. The prototype interface is anticipated to be completed in April 1999. Obtaining traffic volume data from the Center has saved considerable financial resources for M-NCPPC.

One drawback that was noted is the accuracy of the speed data from the arterial street detectors. Since the loop detectors are located downstream of signalized intersections, there is concern about the accuracy of the speed estimates. The intent is to use the hourly and daily data summaries to evaluate daily and seasonal variations in traffic characteristics. There is also interest in evaluating more disaggregate data levels since the data are available for closer study of peaking characteristics. Although the prototype interface is anticipated in April 1999, this effort is relatively new and there appears to be a rather large learning curve for the relational database used to store the data.

Another concern is precise geographic location references or the Montgomery County TMC's loop detectors. For example, the Center has the detectors and intersections numbered, but to match this information with a map, some type of referencing must be performed. It is anticipated that software will need to be developed to determine the grid coordinates on a map to refer to the detector and intersection locations.

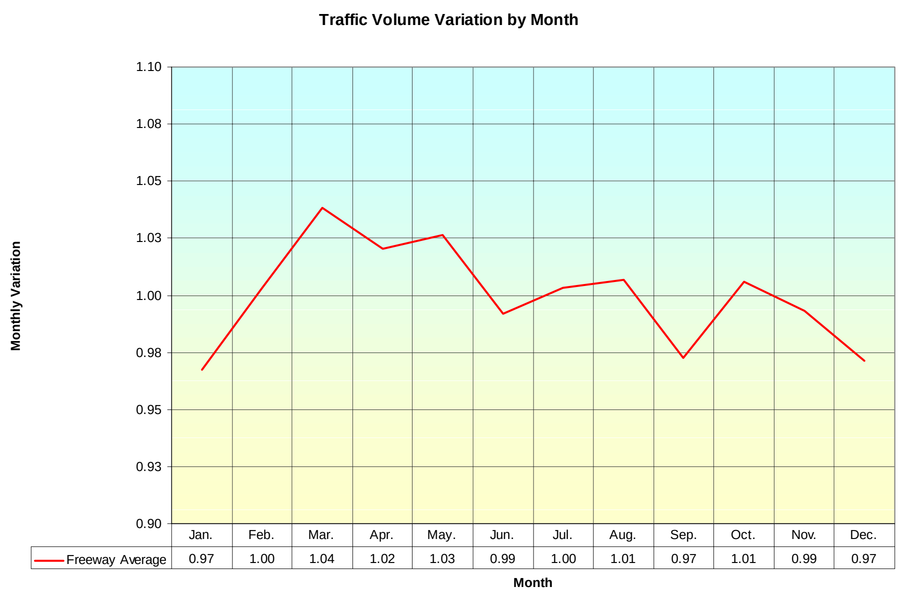
Example: Traffic Statistics and Patterns, Phoenix, Arizona

Transportation planners at the Maricopa Association of Governments are using archived ITS data provided by the ADOT Traffic Operations Center. The planners are using the archived ITS data to study traffic patterns, characteristics, and trends in the Phoenix area. For example, Figure 10 illustrates average weekday traffic patterns by hour of the day. Figure 11 illustrates monthly traffic variation for freeways using ITS data. Figure 12 displays a traffic speed-volume relationship at a given freeway detector location.



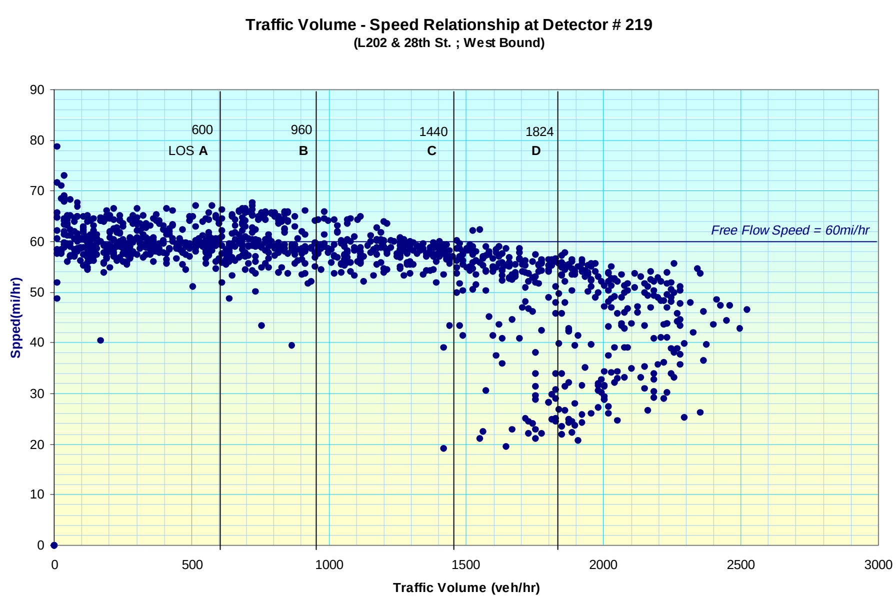
Source: Maricopa Association of Governments

Figure 10. Average Weekday Traffic Volume Patterns Derived from ITS Data, Phoenix Urban Area



Source: Maricopa Association of Governments

Figure 11. Monthly Traffic Variations for Freeways Using ITS Data, Phoenix Urban Area



Source: Maricopa Association of Governments

Figure 12. Traffic Volume-Speed Relationship Using ITS Data

Example: Regional ITS Planning, Puget Sound, Washington

In regional ITS planning efforts in the Seattle area, researchers at Mitretek Systems have developed scenarios to produce probability maps of the variability in roadway capacity from incidents and extreme weather for a range of demands (18). The hypothesis of the work is that ITS strategies are most effective during high-variability situations (Figure 13). Several data sources were used in the case study analysis in the Seattle area, including archived ITS data from the regional TMC. Additional data on accidents, weather, and incidents were also gathered. A scenario hierarchy was developed that began by separating a condition into an event/non-event, then by the weather condition, incident condition, volume ratio, and then accidents. This approach provides a mapping of incident and weather scenarios by demand level as shown in Figure 14. This approach also provides more insight than traditional travel demand forecasting techniques by considering how ITS may affect travel.

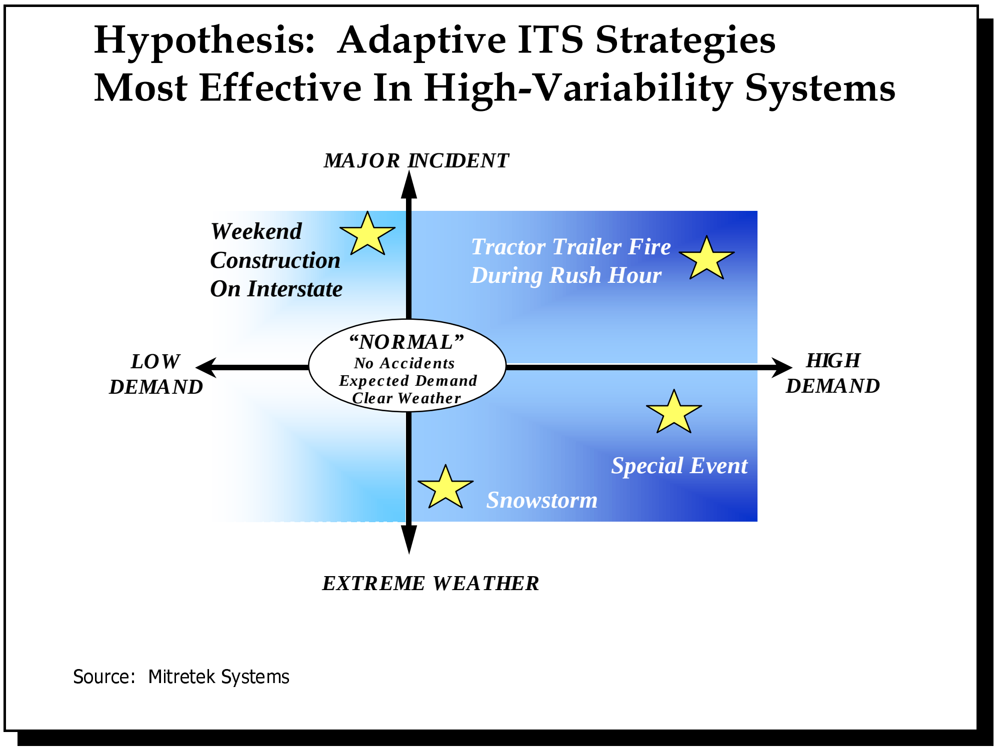


Figure 13. Illustration of Hypothesis that ITS Strategies are Most Effective in High-Variability Systems

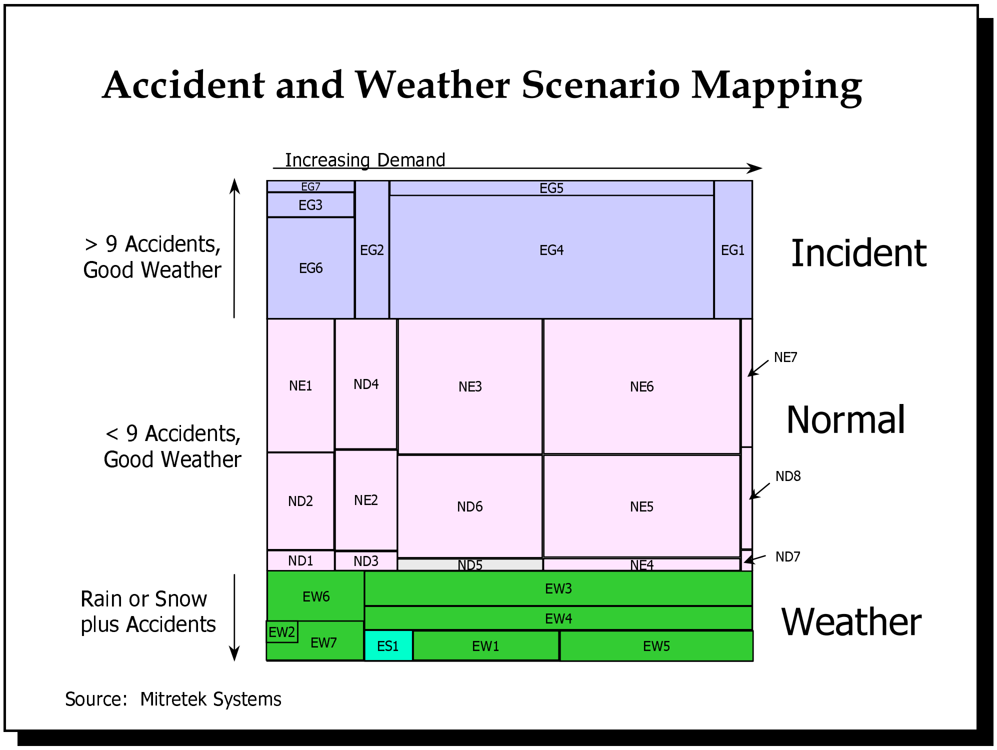


Figure 14. Accident and Weather Scenario Mapping by Demand Level Using Archived ITS Data and Other Data Sources

Example: Alternatives Analysis and Model Calibration, Houston, Texas

Research at TTI has evaluated the benefits of HOV lanes relative to other transportation improvements through a range of congestion levels (14). Five different alternatives were evaluated: 1) no-build; 2) add an HOV lane; 3) add a general-purpose lane in each direction; 4) add two general-purpose lanes in each direction; and 5) implement transportation system management (TSM) practices to increase the freeway capacity by ten percent. The measures of effectiveness (MOE) that were evaluated included speed, delay, emissions, and fuel consumption. Archived AVI probe vehicle data were aggregated into 15-minute time periods for calibration of baseline FREQ10PL models. Once the baseline models were calibrated, the various alternatives were analyzed.

Additionaly, TTI researchers used the same 1994 AVI travel time data set to analyze variability to determine required sample sizes (19). The study found that existing sample sizes, or the number of AVI-equipped vehicles, varied by location and proximity to Houston's toll facilities. In most 15-minute periods during the peak hour, about 15 to 20 probe vehicles reported travel times on the freeway general-purpose lanes. On the HOV lanes, about 5 to 10 probe vehicles reported travel times during most 15-minute periods in the peak hour. The study concluded that the existing probe vehicle system was providing reliable peak period travel speed information.

Traffic Management Operators

Traffic management operators are the second stakeholder group identified in Table 4. These stakeholders are concerned primarily with the day-to-day operations of TMCs. Archived data applications of interest to traffic management operators include developing traffic control strategies, performing highway capacity analyses, incident management, and performance evaluation and monitoring. Data requirements for these analyses are often at a more disaggregate level than the transportation planning applications discussed in the previous section. Data are often desired at aggregation levels of 1, 5, or 15 minutes. Some applications that benefit traffic management operators (e.g., incident detection algorithm development) may require data at intervals of less than 1 minute. Typical applications evaluating highway capacity or ramp-metering strategies may utilize data at the 5-minute aggregation level. These applications often require data for various segments and time periods.

Example: Ramp Metering Evaluation, Mn/DOT TMC

Personnel at the Mn/DOT TMC monitor an extensive ramp metering system that is an integral component of traffic operations. TMC personnel had a growing concern that the ramp detectors in the system were not providing reliable data. Therefore, calibration of the loops was performed to ensure that accurate data were being used as input to the ramp metering system. The occupancy and volume of detector stations were plotted to evaluate the resulting speed estimate provided by the loop data from an estimate of vehicle length. The data were evaluated during relatively free-flow traffic conditions when there are more accurate estimates of the speed along the particular segment. The loop detector data can then be calibrated accordingly with the resulting calibration factor for the particular detector.

The archived ITS data has proved valuable for improving ramp metering operations. Due to the importance of the ramp metering system and the accuracy of the loop data as input to such a system, the work is being further pursued. After reviewing the archived ITS data, TMC staff have realized that even the addition of as little as 5 to 10 vehicles in a 5-minute period, or a change of as little as 2 percent occupancy, can have a dramatic impact on traffic operations.

An issue that was noted during conversations was the geographic location referencing for the detectors. They are in the process of developing a GIS database to establish more accurate, easily-defined detector station locations.

Example: Ramp Metering and Incident Management Evaluation in Texas

TTI researchers assisted TxDOT in the evaluation of ramp metering on the Katy Freeway (I-10) in Houston, Texas through the use of 15-minute summaries of archived AVI data. The study evaluated travel times along several freeway corridors that contain ramp metering. This work provided insight to traffic management operators that the ramp-metering strategies were successful and the ramp metering system was expanded (20).

Researchers at TTI are also using the incident detection logs from the TransGuide® system in San Antonio to evaluate incident time events such as when certain types of incidents/events occur, how long it takes to execute a response, when incidents are responded to, and how long it takes to clear the incident. The archived incident log data provide time stamps and location references of all incident-related events. The results of this research will be valuable in assisting transportation management operators in effectively responding to incidents on the San Antonio freeway system.

Maintenance and Construction

Maintenance and construction stakeholders require data for pavement maintenance, roadside maintenance, or lane closure scheduling. Lane closure scheduling was one of the applications most frequently mentioned in the previously discussed surveys. Hourly traffic volume and/or speed data are often required to determine when lanes should be closed and re-opened to cause the smallest impact to motorists. Hourly traffic volumes are also of value for pavement and bridge maintenance to determine the amount of infrastructure use to estimate maintenance needs.

Transportation Researchers

Transportation researchers typically desire the most disaggregate data of all the stakeholders identified in Table 4. They will often use data at the most disaggregate form available (i.e., 20-second or individual probe vehicle data). Sample applications include forecast and simulation model development, theoretical analysis of highway capacity, and incident detection modeling. Advanced data users such as transportation researchers generally have a significant amount of computing power and data manipulation tools at their disposal to assist them in analysis of large data sets. Data are often desired over many segments and different time periods for research purposes for close study of temporal and spatial changes.

Example: Traffic Flow Characteristics Research, California

Research performed at the Institute of Transportation Studies (ITS) at the University of California at Berkeley has used ITS data supplied by loop detectors to evaluate traffic features of freeway bottlenecks on two freeways in Toronto, Ontario, Canada (21). Speed, occupancy, and volume data were collected every 30 seconds along the Queen Elizabeth Way and the same data were collected every 20 seconds along the Gardiner Expressway. The study evaluated several freeway bottleneck measures including the flow immediately prior to queue, discharge rate immediately following the queue, recovery discharge rate, average discharge rate, and the percent difference between the flow immediately prior to the queue and the average discharge rate. Findings of interest include the following (21):

- Flow can drop substantially following the formation of an upstream queue;

- The discharge flows in active bottlenecks exhibit near-stationary patterns that (slowly) alternate about a constant rate;

- A bottleneck's long-run discharge rates are consistent from day to day, while other flow features exhibit daily variation; and

- Bottlenecks occur at fixed (i.e., reproducible) locations.

The research team further discovered that the difference between the flow immediately prior to the queue formation and the average discharge rate can be as high as 10 percent. The research team indicated that many of the effects of the freeway bottlenecks are presented and that the actual causes are being studied. This includes the reasons for the driver behavior that caused the relatively high vehicle flows to cause the bottlenecks or the reasons the queues were always observed at about one kilometer (0.6 mile) downstream or more from the on-ramp of interest. Understanding the driver behavior will allow for better decision-making for traffic management operations such as ramp metering and changeable message signs. Finally, the research team anticipates the evaluation of freeway segments in the U.S. to determine the similarity between U.S. freeways and those studied in Toronto.

Example: Development of Congestion Prediction Model

The National Institute of Statistical Sciences (NISS) and Bell Laboratories has performed research using loop detector data for predicting congestion in real-time (22). This research is valuable for developing models that may predict congestion so that traffic operators can make traffic management changes (i.e., altering ramp-metering strategies or changeable message signs appropriately). Twenty-second data were obtained from the Seattle area, which include traffic volume and lane occupancy obtained from the single detectors. The data were aggregated to the one-minute level and breakdown was defined as, "the first of five consecutive one-minute intervals in which the estimated speed over the detector at that location is less than 48 km/hr (30 mph). Similarly, a breakdown ends with the first of five consecutive one-minute intervals in which speeds exceed 48 km/hr (30 mph)" (22). The study corridor was an 18.8 km (11.7 mile) section of I-5 in Seattle with 24 detectors in the northbound direction and 23 detectors in the southbound direction. Speeds were estimated from the occupancy and volume data using an assumed vehicle length, and the data were averaged across lanes at each detector station.

A tree-structured CART classification was used in the study to predict a breakdown or no breakdown given the set of predictor variables. The predictor variables are a function of the volume, occupancy, and speed data. The conclusions of the study were as follows (22):

- Data quality is a significant concern and about 40 percent of the data were discarded;

- CART was found to be a useful tool for constructing models to predict breakdown based upon speed and occupancy;

- Using more than one detector station increases the predictability of breakdown further into the future;

- There was a gain in predictability by using the finer detail of the one-minute data over the 5-minute data in both the training data set and the test set with respect to detection rate, although the difference averaged only about three percent.

Example: Development of Travel Time Prediction Models, Texas

Research underway at TTI has been using archived AVI probe vehicle data from Houston TranStar for both travel time forecasting and origin-destination estimation. One research effort evaluated the use of an artificial neural network (ANN) for use in estimating travel times in the near future (23). Five-minute time periods were used for the prediction. The ANN model was used since it allows for the use of non-linear relationships for travel time prediction that include both spatial and temporal components. When predicting one or two periods into the future (i.e., 5 to 10 minutes), the ANN model that considered only the previous travel time information from that link gave the best estimates. When predicting three to five time periods into the future (i.e., 15 to 25 minutes), the ANN model that considered both spatial and temporal inputs including the travel time estimates from the upstream and downstream links, as well as the link of interest, provided the best prediction ability.

Similar research has also compared the ANN modeling with other methods of travel time prediction (24,25). The first step of the modeling methodology was to group the data into similar clusters with an unsupervised clustering technique. The second step was to calibrate and test the ANN model with actual travel time data obtained from the Houston TranStar AVI system. It was found that the ANN provided the best results over several other existing methods of estimating link travel times including a Kalman filtering model, an exponential smoothing model, using historical profiles, or using a real-time profile. The latter methods of using real-time profiles are those often employed by most advanced transportation management systems.

Continued research has also evaluated the use of spectral basis neural networks (SNNs) (26). This research has found that SNN's direct forecasting approach provides better travel time prediction results than the other models studied. Further, this research effort also evaluated the correlation of travel time along a route. It was found, based on both theoretical and empirical results, that the correlation between link travel times affects the accuracy of the resulting route travel time. An interesting result is that the error in link travel times are not additive and the error associated with the route travel times may actually be lower than the error in link travel times.

Example: Estimation of Origin-Destination Matrices, Texas

Archived AVI probe vehicle data available from the Houston TranStar system are also being used for origin-destination estimation (27). The basic premise of this research is that while the AVI data has traditionally been used for real-time link speeds, it also provides a sample of point-to-point trip estimates as well. A computer program has been developed that can trace individual vehicles from the start of their journey on the Houston freeway system to their destination. This information can then be used to derive a sample origin-destination matrix, which is then used to estimate the complete freeway origin-destination matrix. The research found that trip length frequency distributions (TLFDs) vary by time of day and day of the week, although there was similarity among weekdays. The information provided by the research could be used for such applications as estimating parameters for highway-based trip distribution models (e.g., gravity models) or validating planning models with information about link or route travel times as well as freeway trip volumes. Essentially, the matrix of information can provide an inexpensive cordon count of spatial and temporal information in the network. Future research in this area will evaluate using the sampling of route choices, the estimation of link choice proportions, methods for relating the AVI origin-destination methods to the network origin-destination pairs, developing origin-destination matrix data for different time periods, and the real-time application of origin-destination for traffic operations. The potential for combining real-time traffic volume counts with the AVI origin-destination to forecast real-time origin-destination matrices on the I-10 corridor in Houston is also being studied.

Example: Traffic Flow Theory Research, Ontario

Research has been performed at McMaster University in Hamilton, Ontario, Canada to study speed-flow relationships during congested freeway conditions (28). This research is valuable to operations personnel as well as planners that desire a better understanding of the speed-flow relationship under congested conditions. Loop detector data (speed, volume, occupancy) along the Gardiner Expressway and Highway 401 were used in the study. Data were aggregated to the five-minute level for analysis and suspect data were removed from the data set. In addition, data that represented congested conditions was selected from the data available (i.e., data from transitions between congested and uncongested were removed from the data set). The analysis tested many different models including a cubic equation, power function, and an exponential function, and the R^2 values for the equations ranged from 0.75 to 0.78. The research also addressed questions such as whether the equation should be forced through the origin, the influence of construction data, and whether different freeways require their own functions. The research findings were as follows (28):

- It is necessary to have data from a range of flows to fit the curve for the congested portion of the speed-flow curve;

- It is important to remove data from transitions between the congested and uncongested data regimes;

- The relationship for construction periods seems to be different than when construction is not present;

- There may likely be a difference in the speed-flow relationship for different freeways due to different lane widths or other geometric factors;

- Downstream operating conditions affect the speed-flow relationship within congested periods; and

- Several sites may likely need to be combined for development of the speed-flow curves during congested periods, yet care must be taken in determining the sites and conditions that may be combined.

Private Sector

The private sector is also a significant stakeholder for archived ITS data. Many private sector companies are developing business models to add value to the data that are available through ITS and then reselling the value-added data. One business model includes the private sector receiving the raw archived data from the public sector agencies that collect the data from publicly-owned infrastructure. The private sector then reduces the large data sets to useful information for targeted users (e.g., state transportation planners) in return for a fee. Another business model has private sector (e.g., telecommunications companies) collecting data from privately-owned infrastructure, then selling the real-time and historical archived data to various public sector markets. There are many issues related to ownership of the data, intellectual property, and privacy concerns, to name only a few, that must be addressed when considering the private sector stakeholder.

ADVANCING THE STATE-OF-THE-PRACTICE IN ITS DATA ARCHIVING

This section describes activities and efforts focused on advancing the state-of-the-practice in the archiving or use of ITS data (Table 5). Several of these activities are concerned with development of the ADUS in the National ITS Architecture and related data standards. Several researchers are examining issues that affect ITS data archiving practices or data applications. It is clear from the discussion that many technical and institutional issues remain to be addressed before the retention and use of archived ITS data becomes standard operating procedure.

Table 5. Efforts to Advance the State-of-the-Practice in Archived ITS Data

Effort	Anticipated Contribution to ITS Data Archiving	Resources
Development of ADUS Architecture	Provide framework for deploying ADUS applications.	*ITS as a Data Resource: Preliminary Requirements for a User Service* (4) and National ITS Architecture, Version 3.0 (to include ADUS architecture)
Data Standards	Provide data dictionaries and meta attributes that provide capability for unambiguous exchange and storage of ITS data.	Standards documentation at http://www.its.dot.gov/standard/itssum.pdf (also see Table 8)
ITS Data Registry	Provide centralized clearinghouse for all data element descriptions and associated meta attributes.	Preliminary documentation at http://grouper.ieee.org/groups/scc32/datareg/DRFOP8.DOC.
Volpe Center Case Studies	Communicate lessons learned, opportunities, and guidance on the use of ITS data with case studies in four cities (Chicago, Detroit, Minneapolis, and San Antonio).	Projected deliverables include technical memoranda, four case study reports, and a final report. Final report publicly available later in 1999.
Oak Ridge National Lab/Florida DOT/New York DOT Demonstration Project	Identify institutional, technological, communication, data, and logistical issues that need to be addressed if and/or when ITS data is used for traffic monitoring needs.	Projected deliverables include technical memoranda and task reports. Expected study completion date in mid-2000.
Washington State	Develop a methodology, detailed procedures, and tools for conducting ongoing performance evaluations of the FLOW traffic management system.	*Central Puget Sound Freeway Network Usage and Performance* (13)
Texas Transportation Institute, TransLink® ITS Research Program	Development, maintenance, and expansion of DataLink, an on-line prototype ITS data management system; helped sponsor Texas ITS data archiving workshop; investigating performance measures using archived ITS data.	*ITS Data Management System: Year One Activities* (5); DataLink web site (http://vixen.cs.tamu.edu); and *Proceedings from the Texas ITS Data Uses and Archiving Workshop* (29)

Development of the ADUS Architecture

The ADUS is currently being defined and developed as a component of the National ITS Architecture, the purpose of which is to serve as a guide in the development of an integrated, multi-modal intelligent transportation system. It is envisioned that the ADUS architecture will provide a framework in which transportation information collected by ITS could be made available to a wide variety of stakeholders for data analysis and exploration. The following paragraphs describe the development of the ADUS architecture (also see Table 6).

The first organized efforts to create the ADUS began in 1997, subsequent to the first release (Version 1.0) of the National ITS Architecture. At the time, the Architecture did include a planning subsystem (Figure 15), which had been designed mainly as a "place holder" and had no user service to support. Because user services describe ITS functions from a user's perspective, it was incumbent upon the planning subsystem users and stakeholders to formally request the inclusion of a data user service in the Architecture. FHWA's Office of Highway Information Management (now Office of Highway Policy Information) led this effort on behalf of the data stakeholders, with ITS America providing support in assembling stakeholders for early discussions.

Based upon several stakeholder meetings held in 1998, an addendum to the ITS program plan and a set of user service requirements were developed to guide the creation and development of ADUS, the 31st user service. In January 1998, numerous stakeholder groups gathered to discuss the desirability of and issues associated with creating ADUS at the "ITS as a Data Resource" workshop held in Washington, D.C. An ADUS resource document was created from this initial meeting that contained preliminary user requirements (4). Stakeholder comments and input from a subsequent workshop in Seattle in July 1998 were used to finalize the ADUS addendum to the ITS Program Plan and the User Service Requirements. By late 1998, the U.S. DOT had begun work on developing the ADUS component of the National ITS Architecture.

At the time of this writing, the details of the ADUS architecture are being developed by U.S. DOT and its contractors in conjunction with numerous stakeholders at the local, state, and federal levels. The ADUS revision to the Architecture is scheduled for completion in August 1999, and once revised, will be available as part of the National ITS Architecture documentation (Version 3.0 or later). More information on the ADUS revision can be found at the following web sites:

- ITS American ADUS Resource Page: http://www.itsa.org
- U.S DOT National ITS Architecture: http://www.its.dot.gov/arch/
- Odetics Browsable National ITS Architecture: http://www.odetics.com/itsarch/

Table 6. Incorporation of the ADUS into the National ITS Architecture

Action	Date
Champion for ADUS emerges	1997
Initial stakeholder meeting	January 1998
Position paper with preliminary requirements	April 1998
Second stakeholder meeting	July 1998
Addendum to the *ITS Program Plan*	September 1998
User Service Requirements	October 1998
Architecture Revision begins	December 1998
Subsequent stakeholder meetings	January 1999 March 1999 July 1999
Architecture revision complete	approx. September 1999

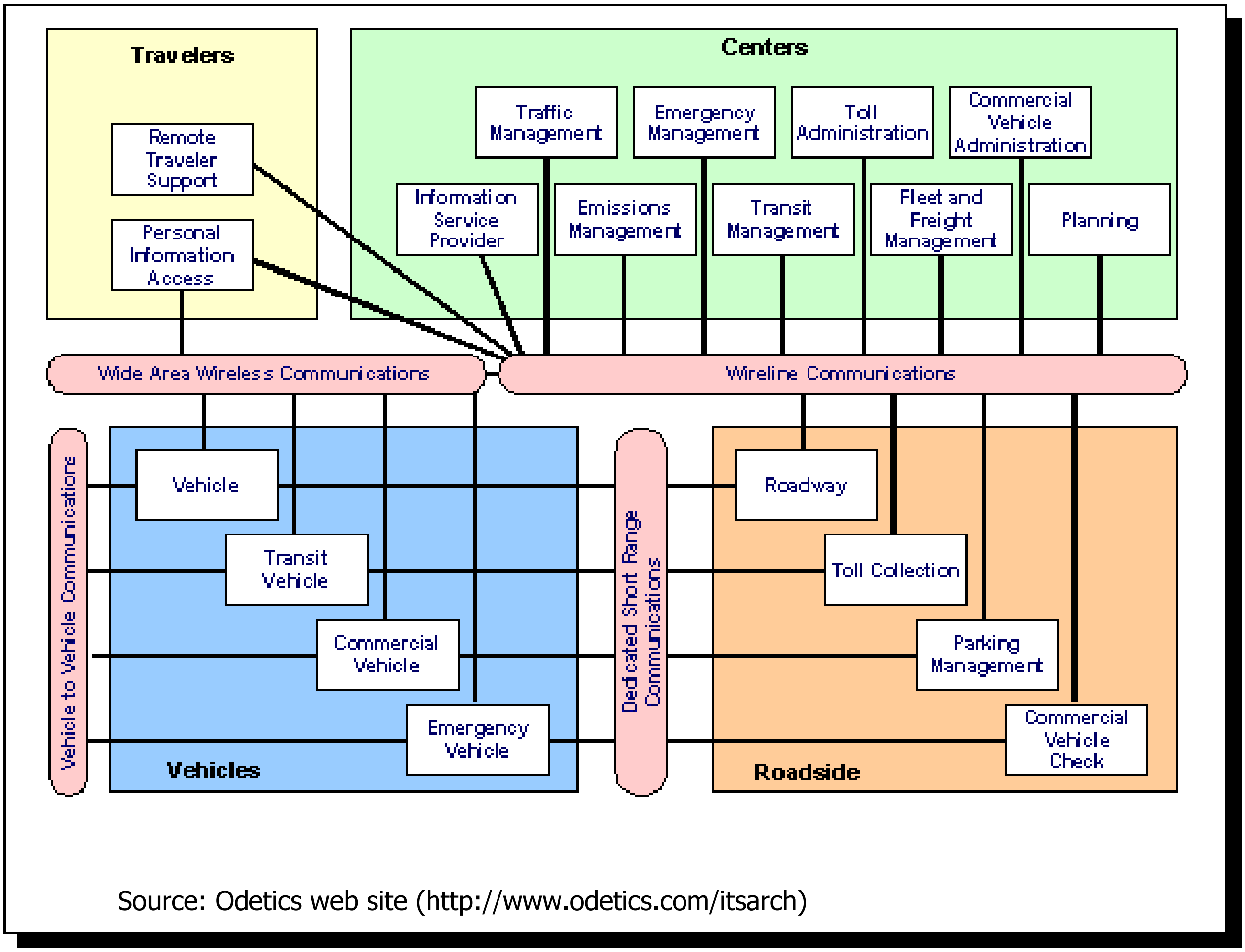


Figure 15. National ITS Architecture

Data Standards Development

Numerous transportation data standards are being developed to ensure interoperability in the implementation of ITS technologies, many of which are relevant to the ADUS. When implemented, these standards prescribe the attributes and format of data elements collected, communicated, or stored by ITS applications or components. Using the Institute of Transportation Engineer's (ITE's) Traffic Management Data Dictionary (TMDD) standards as an example, a vehicle count from a traffic detector has numerous attributes as shown in Table 7.

Table 7. Attributes of a Vehicle Detector Count Using ITE's TMDD

Data Attribute	Data Attribute Value
Descriptive Name:	DETECTOR_VehicleCount_quantity
Descriptive Name Context:	Manage Traffic
Definition:	The number of vehicles detected by a detector during a specific time period.
Class Name:	Traffic Detectors
Classification Scheme Name:	IEEE P1489, Annex B
Classification Scheme:	19971009, V0.0.7
Keywords:	Detector Vehicle count
Related Data Concept:	LINK_Volume_rate, DETECTOR_IdNumber_number
Relationship Type:	Possible identifier for a link volume or a detector count
ASN1 Name:	Detector-vehicle-count
ANS1 Data Type:	Integer
Representation Class Term:	Quantity
Value Domain:	SI 10-1997; vehicles
Valid Value Range:	0 to 100000
Valid Value List:	Not Applicable
Valid Value Rule:	Not Applicable
Internal Representation Layout:	999999
Internal Layout Maximum Size:	SIZE(32)
Internal Layout Minimum Size:	SIZE(32)
Remarks:	V1.1 - Revised Definition and Permissible Data Element Values.
Data Concept Identifier:	3515
Data Concept Version:	V1.3
Submitter Organization Name:	TMDD
Last Change:	19980610

Source: ITE TMDD Section 3, Version 1.3, June 1998

Table 8 lists selected data standards that are most relevant to the development of ADUS applications. The ITS standards setting effort has been industry-driven, with several standards development organizations (SDOs) leading various data standardization efforts. With numerous data standards efforts being led by several different SDOs, there is a potential for overlaps and/or conflicts between the various data standards efforts. The ITS Data Registry has been suggested as a solution to potential overlaps and/or conflicts between different data standards or dictionaries.

Table 8. Selected ITS Standards Relevant to ITS Data Archiving

Title of Standard	Standard Description	Lead SDO	For More Information
Advanced Traffic Management System (ATMS) Data Dictionary (TMDD): Sections 1 and 2 (Links/Nodes/Events) [TM 1.01]	A functional-level data dictionary for traffic management applications. Describes and standardizes roadway links and nodes in accordance with location-referencing message standards data elements for incidents and traffic-disruptive roadway events.	ITE	ITE Web Site: http://www.ite.org/standards/index.html
ATMS Data Dictionary (TMDD): Sections 1 and 2 (DMS/Video/Control/Etc.) [TM 1.02]	A functional-level data dictionary for traffic management applications. Includes data elements for traffic control, traffic detectors, actuated signal controllers, traffic modeling, vehicle probes, ramp metering data, dynamic message signs, video and camera control, parking management and weather stations.	ITE	ITE Web Site: http://www.ite.org/standards/index.html
NTCIP Data Collection and Monitoring Devices [AASHTO TS 3.DCM]	Specifies object definitions that may be supported by data collection and monitoring devices, such as roadway loop detectors.	AASHTO	NTCIP Web Site: http://www.ntcip.org
NTCIP Object Definitions for Transportation Sensor Systems [AASHTO TS 3.TSS]	Object definitions that are specific to and guide the data exchange content between advanced sensors and other devices in an NTCIP network. Advanced sensors may include video-based detection sensors, inductive loop detectors, sonic detectors, infrared detectors, infrared detectors and microwave/radar detectors.	AASHTO	NTCIP Web Site: http://www.ntcip.org
Template for ITS Message Sets [IEEE P1488]	Describes the structure and content of message sets for exchange between traffic centers, emergency management centers, and traveler information systems in a consistent and uniform manner.	IEEE	IEEE Web Site: http://www.ieee.org
Standard for Data Dictionaries for ITS [IEEE P1489]	Specifies a common set of meta entities and meta attributes for ITS data dictionaries, as well as associated conventions and schema, that enable describing, standardizing, and managing all ITS data.	IEEE	IEEE Web Site: http://www.ieee.org

Source: http://www.its.dot.gov/standard/itssum.pdf, Feb. 12, 1999.

ITS Data Registry

Sponsored by the Institute of Electrical and Electronic Engineers (IEEE), the ITS Data Registry is envisioned to be a central repository for all ITS data element definitions (30). The Data Registry will not replace any standard setting efforts, but will supplement these efforts by serving as an arbitrator of overlaps and/or conflicts in data element definitions. The Data Registry will classify all submitted data elements into one of seven quality status categories (Figure 16).

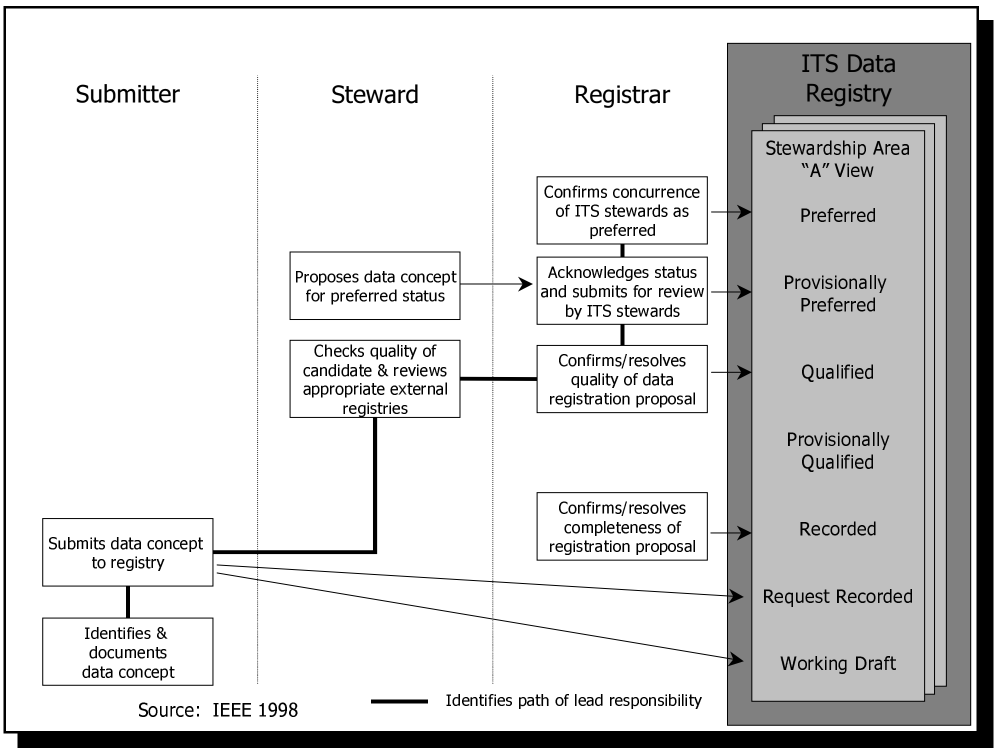


Figure 16. ITS Data Registry Functional Concepts

At the time of this writing, only the functional concepts related to the ITS data registry have been defined (31). The operation of the ITS Data Registry has not officially started, nor has the prototyping of a registry system been initiated. However, small-scale testing of the ITS Data Registry is expected in the near future. Access to data elements in the ITS Data Registry will likely be provided on a per agency basis, in which a single fee covers access to all Registry data elements for an entire agency.

Volpe Center Case Studies

Researchers at the Volpe Center are performing case studies on the use of ITS data for measuring transportation system performance and other planning applications (32). The case studies are being performed in four cities:

- Chicago, Illinois;
- Detroit, Michigan;
- Minneapolis, Minnesota; and
- San Antonio, Texas.

The research study has the following five objectives:

1. Provide an overview and inventory of current uses of data on both passenger and commercial vehicle use of the transportation system;
2. Identify system performance data needs and the requirements to meet those needs;
3. Identify opportunities for the full and effective use of ITS-generated data in transportation planning and system management;
4. Perform cross-cutting analysis to identify how ITS-generated data can contribute to better decision making and can be used to improve the metropolitan planning process nationwide; and
5. Communicate lessons learned, opportunities, and guidance to state and local transportation professionals on the effective use of ITS-generated data.

The case studies were completed in 1998 and a draft project report has been prepared. It is expected that the project report will be publicly available later in 1999.

Oak Ridge National Lab/Florida DOT/New York DOT Demonstration Project

Researchers at the Oak Ridge National Laboratory have initiated a study with the Florida and New York DOTs to assess the feasibility of using ITS data in traditional traffic monitoring programs (33). The demonstration project started in November 1998 and is projected for completion in 18 months. The researchers are attempting to address the following questions:

- Are ITS-generated data able to provide or estimate needed traffic monitoring information?

- If so, how close are the ITS-based estimates to those based on traditional data collection efforts?

- If so, what are the institutional, technological, communication, data, and logistical issues that need to be addressed? And how might they be addressed?

- When fully operational, can ITS-generated data be integrated with the more traditional data collection efforts?

The research study consists of the following three tasks:

1. Identify Data Gaps and Survey Lessons Learned
   - Inventory data collected under ITS deployments
   - Evaluate the adequacy of ITS-generated data
   - Identify barriers in using ITS-generated data
   - Characterize data gaps
   - Survey lessons learned

2. Develop, Instrument, and Test Procedure to Estimate Traffic Parameters
   - Develop schematic to link ITS-generated data and information needs
   - Estimate regional traffic parameters based on ITS-generated data
   - Quantify the precision of ITS-based traffic data estimates
   - Develop a communications procedure
   - Summarize results and document final procedure

3. Recommend Future Activities

Washington State

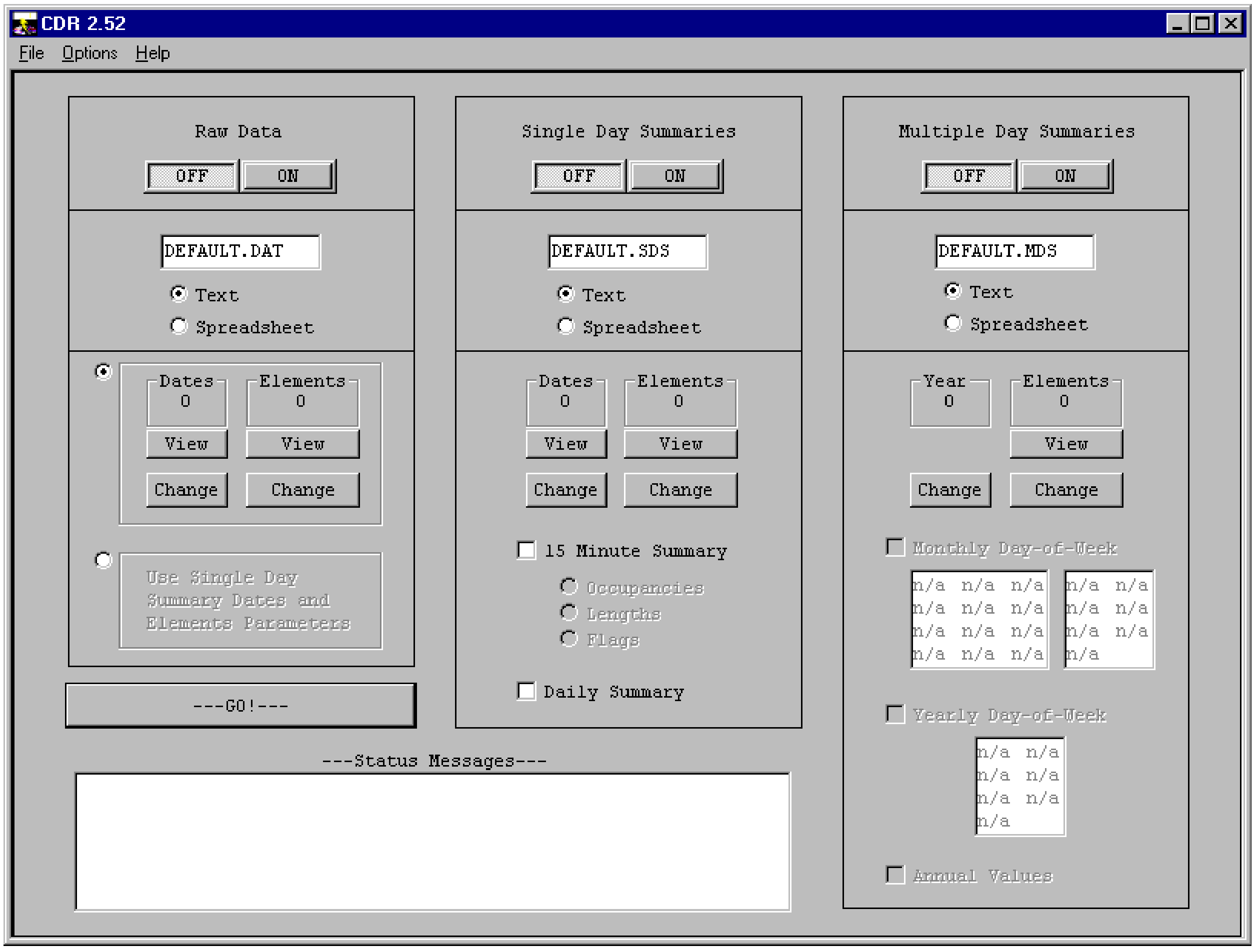
Researchers at TRAC are working with WSDOT in demonstrating the usefulness and extending the applications of archived ITS data. The FLOW Evaluation Design study had the following objectives (13,34):

1. Develop a methodology, framework, and detailed procedures for conducting an ongoing series of evaluations of the performance and effects of the FLOW traffic management system now in operation on Puget Sound area freeways;

2. Create tools for performing those evaluations; and

3. Supplement earlier evaluation data with updated results by using the developed framework to evaluate selected portions of the FLOW system.

The FLOW Evaluation study created several analytical tools that are now being used by TRAC and WSDOT for reducing and analyzing archived ITS data. The regional TMC in WSDOT's Northwest Region (Seattle urban area and surrounding environs) currently archives loop detector data to CD at five-minute intervals (see previous section for discussion of TMC capabilities and practices). TRAC researchers have developed PC-based CD retrieval (CDR) software to retrieve and aggregate the five-minute loop data to local or network disk drives. Figure 17 shows the graphical user interface for the publicly available CDR software. Another software program developed in this study, CD Analyst, uses spreadsheet macros to analyze the loop data as well as develop graphical presentations.

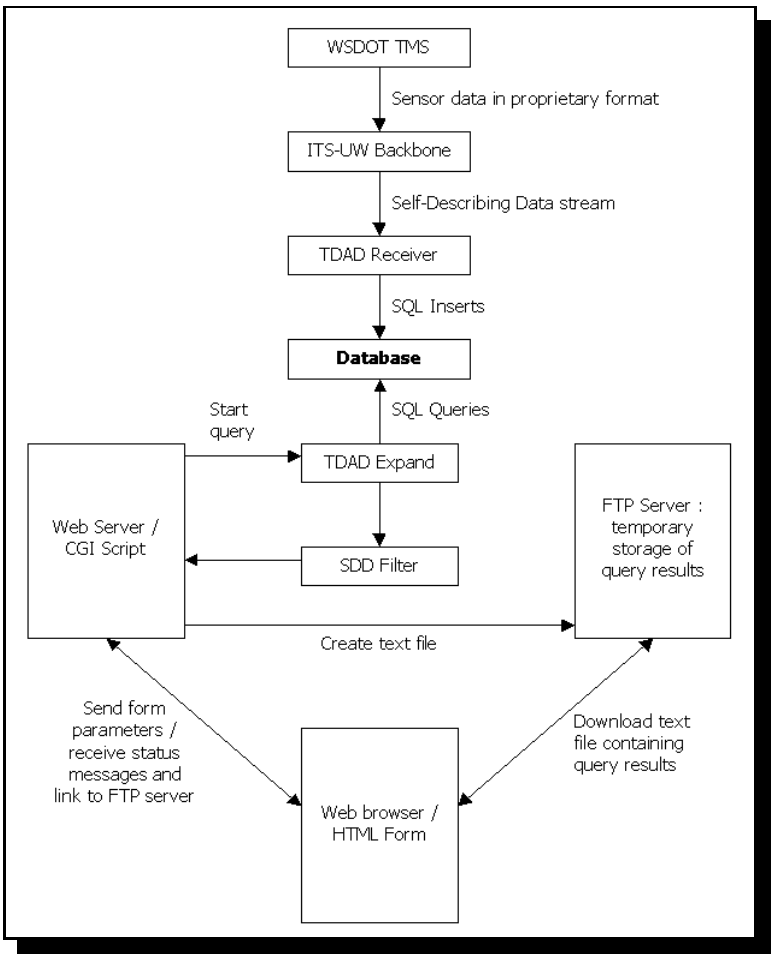


Source: CDR Software v. 2.52, Washington State DOT

Figure 17. Main Interface of CDR Software

TRAC researchers and WSDOT engineers and planners use the CDR and CD Analyst software tools to analyze archived ITS data. TRAC researchers are supplying the extracted archived data for policy development as well as the measurement and evaluation of freeway system performance. WSDOT engineers and planners use the archived data to generate historic traffic volumes and trends for use in numerous planning and design applications.

University of Washington researchers are also developing a web-based ITS data mining application in the Traffic Data Acquisition and Distribution (TDAD) study (35). This prototype application, located at http://www.its.washington.edu/tdad/tdad_top.html, has a map-based interface that is used to select loop detector location(s) of interest. At this time, the raw, 20-second loop data is being used to populate the TDAD database, and approximately six months of data have been accumulated. Figure 18 shows a schematic of the TDAD system database.



Source: http://www.its.washington.edu/tdad/tdad.gif,
University of Washington

Figure 18. Schematic of TDAD System

TransLink® ITS Research Program, Texas Transportation Institute

The TransLink® ITS Research Program at TTI has sponsored an "ITS Data Management System" focus area since 1997. Early work in this area was focused on the development of DataLink, a prototype on-line ITS data management system (5,6). The DataLink system warehouses loop detector data from San Antonio's TransGuide® system in a 40+ gigabyte relational database. The loop detector data are aggregated from 20-second intervals to five-minute increments, mainly because of user data requirements as well as database performance. Users of the DataLink system can perform ad hoc queries within a web browser interface (Figure 19), eliminating the need to know structured query language (SQL). To date, the DataLink system has been storing data on-line since late 1997, and researchers are examining data aging issues related to permanent archiving of a subset of the five-minute data.

Ongoing work in this TransLink® focus area includes the following:

- **DataLink maintenance and refinement -** The DataLink system has been beta tested by TxDOT planners, and this feedback will be used to improve the user interface and query capabilities. The research team is also working on integrating incident information into the loop detector database.

- **Texas ITS data archiving workshop -** In cooperation with TxDOT and ITS Texas, TransLink helped to sponsor an ITS data archiving workshop in Texas (29). Participants included various agency personnel involved in operations, planning, safety, air quality, and research.

- **Performance measures from ITS data -** Researchers are developing a methodology and examples of using ITS data to support a range of transportation system performance measures.

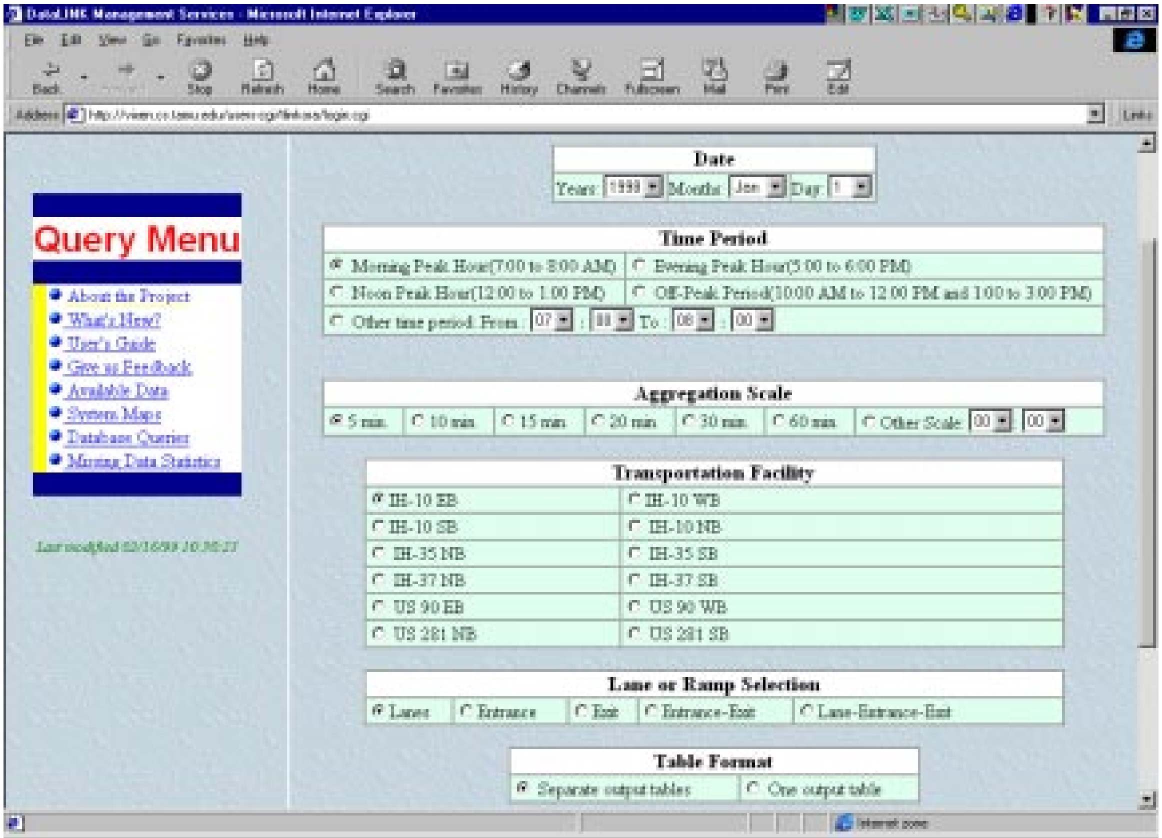


Figure 19. Query Interface of DataLink System (http://vixen.cs.tamu.edu)

SUMMARY OF FINDINGS AND LESSONS LEARNED

The research team summarizes the state-of-the-practice in ITS data archiving as follows:

- **Improved Interdisciplinary Coordination -** It appears that many TMC personnel are more interested in supporting the data needs of secondary users. For example, there appears to be increased support by operations personnel to make data accessible and in the correct format for planning applications. Although many agencies are still investigating appropriate formats, storage media, and interfaces, the coordination has improved.

- **Similar Concerns Among Practitioners -** There was strong interest in and support for the work being performed and the questions being raised during the telephone conversations. Nearly all of the experiences indicate similar interests and concerns in different regions. Common questions include:

    - Who are the users of the data?
    - What are the data uses?
    - What data should be kept?
    - What aggregation level should be selected?
    - What type of interface should be developed?
    - How can the system be automated?
    - What is the appropriate data format?

    In addition, concerns were expressed related to the steep learning curve of data management systems (e.g., relational databases) by both operations personnel and other data users (e.g., planning staff).

- **Data Quality Concerns -** Many individuals noted concern over the quality of the data itself. Some of the concern may be attributed to the fact that these large ITS data sets are new to many data users, thus there is some unfamiliarity with the inherent quality of the data. The concern with data quality also may be because, in some cases, only minimal error detection is performed as the data is being collected. Many expressed that more work needs to be performed in regard to data quality.

- **Need for Geographic Location Referencing -** It was often noted that referencing the actual data obtained from the TMC with a location on a roadway segment was difficult. Secondary users of the data often have to develop equivalency matrices to compare several location referencing schemes.

- **Easy Data Access is Important -** It appears that as easy access to archived ITS data is provided, most data users create or use existing software to manipulate and analyze the large data sets to meet their needs. Although there may be a relatively steep learning curve for developing or using such software and techniques, most dedicated users will often find a way to meet their needs. Many TMCs are considering the distribution of data via the Internet or CD, as opposed to traditional practices of archiving data onto magnetic tape cartridges or off-line storage devices.

- **Operations Personnel Receiving Benefits -** One of the primary driving forces in the improved interdisciplinary coordination is likely the fact that operations personnel are receiving benefits and improvements in how they perform their jobs. For example, when research and other activities provide valuable insight about when or how a changeable message sign or ramp-metering operations should be altered, operators see the benefit for such activities.

- **"One Size Does Not Fit All" Regarding Data Aggregation -** There does not appear to be a least common denominator of data aggregation that is significantly favored. Many of the applications provided here span many stakeholders including planning, operations, and research, and they use many different levels of aggregation. In fact, for many of the research applications, the research teams often express their interest in performing similar analysis at different aggregation levels.

CHAPTER 3.
EXAMINATION OF ITS DATA ARCHIVING ISSUES

📖 CHAPTER OVERVIEW

☞	**Incorporating Data Needs into ITS Data Archiving**	Summarizes information that could be useful in defining data archiving user requirements.
☞	**Aggregation of ITS Traffic Monitoring Data**	Presents results of statistical analyses aimed at determining optimal aggregation level, suggests advanced data archiving capabilities.
☞	**ITS Data Quality Control**	Presents results of case study analyses of San Antonio data, focuses on erroneous data, missing data, and data accuracy.
☞	**Data Storage Tools and Issues**	Summarizes information on data storage and distribution as it relates to ITS data archiving.

As the previous chapter indicated, the ADUS architecture will provide an overall framework within which ITS data archiving, analysis, and user systems may be constructed. Given the general nature of the National ITS Architecture, however, agencies that wish to implement ADUS will still have to adapt the ADUS architecture to regional architectures and make numerous system design decisions. For those agencies considering implementation of ADUS, this chapter provides information and lessons learned on ITS data archiving strategies and practices. The information and lessons learned presented in this chapter are focused on four major areas:

- incorporating data needs into ITS data archiving;
- aggregation of ITS data generated by traffic monitoring devices;
- ITS data quality control; and
- ITS data storage tools and issues.

The information and lessons learned in this chapter are primarily based upon exploratory analyses of San Antonio's TransGuide® TMC data conducted specifically for this study. Additionally, some of the information is based upon TTI experience gained in working with, analyzing, and archiving ITS data from Texas' TMCs since 1995.

INCORPORATING DATA NEEDS INTO ITS DATA ARCHIVING

Data needs (i.e., user requirements) should be a definitive element of ITS data archiving activities, as these needs will affect many of the decisions to be made regarding data storage and management system designs. This section discusses the incorporation of data needs into ITS data archiving and presents information that could be helpful in defining user requirements.

Process of Defining or Establishing Data Needs

This section discusses the process of defining or establishing data needs and user requirements through data stakeholder input. The user requirements are used in designing archived ITS data storage and management systems. The process for defining or establishing data needs will likely vary from region to region, or may not even be necessary if TMCs elect to make all ITS data accessible to data stakeholders.

There appear to be several different processes for defining or establishing data needs in Texas as well as other survey locations:

- Determine availability of ITS data from TMC, then cross-reference TMC data availability to current data needs (Figure 20);

- Establish a comprehensive list of data needs, then cross-reference the data needs to TMC data availability (Figure 21); and

- An iterative process that establishes a dialogue or partnership between data collectors/providers and data users, thereby affecting future ITS detector/sensor designs (Figure 22).

Figure 20 illustrates the first process, whereby ITS data availability is enumerated by data providers first. This process appears to be most common where TMCs are established and data flows within the TMC are well-documented. The advantages to this process are:

- The enumeration of existing ITS data elements is often easier to perform than an exhaustive inventory of all stakeholder's data needs.
- Many data stakeholders may not be aware of available ITS data, and enumeration of available data is necessary for education and awareness.

The primary disadvantage to this process is:

- Data user stakeholders are typically limited to available data only and do not have input into modifying existing ITS data collection practices or sensor designs.

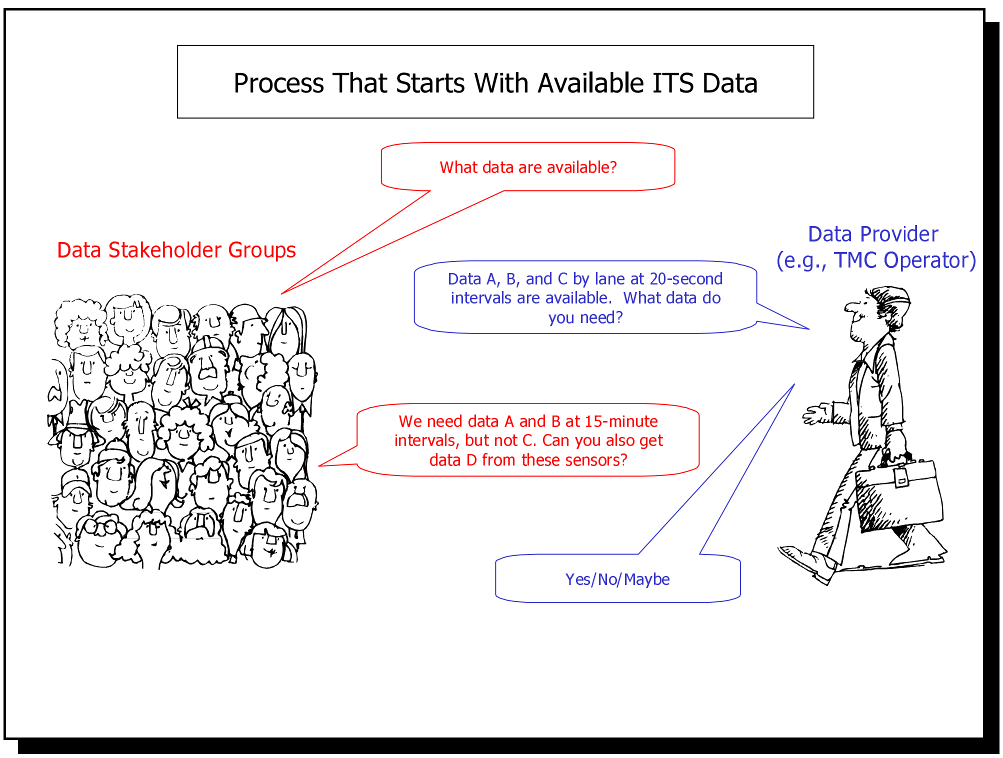


Figure 20. Process that Starts with Enumeration of Available ITS Data

Figure 21 illustrates the process whereby data user stakeholders establish their data needs first, and then reference their data needs to TMC data availability. This process may occur most often when: a) the data flows within a TMC are not well-documented, or b) the TMC may be in early stages of development and therefore flexible in providing data to meet existing needs. The advantage to this process is:

- Provides an opportunity for TMC personnel to modify or adapt data collection practices or sensor designs to meet data needs.

The disadvantages to this process are:

- Establishing a common set of data needs and user requirements may be difficult if many data stakeholders are involved;

- The development of an exhaustive set of data needs may not be efficient if the TMC is only capable of providing a limited number of data elements.

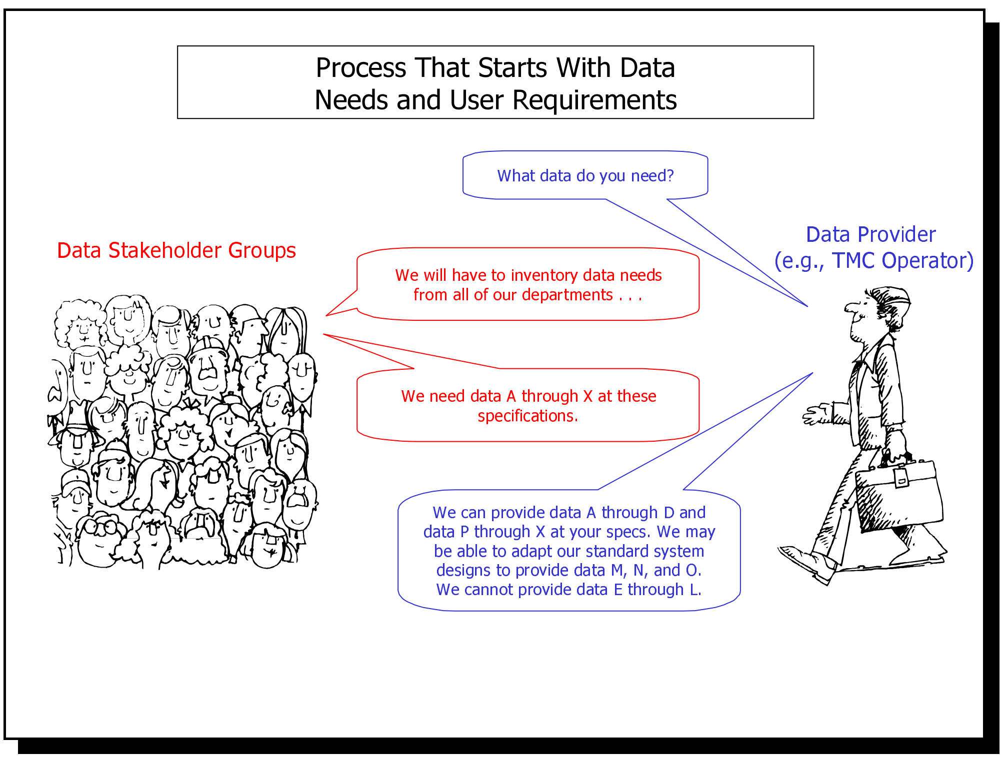


Figure 21. Process that Starts with Stakeholder Data Needs

Figure 22 illustrates an iterative process that cross-references available ITS data with stakeholder data needs, in a continuing dialogue about stakeholder data needs and the TMC's ability to provide certain data elements. The advantages of this process are:

- With a continuing dialogue, the TMC operators may be better able to adapt and/or modify existing data collection practices and sensor designs if they understand stakeholder data needs.

- Cross-referencing of available ITS data to stakeholder data needs may prevent the development of an exhaustive inventory of data needs that cannot be met by TMCs.

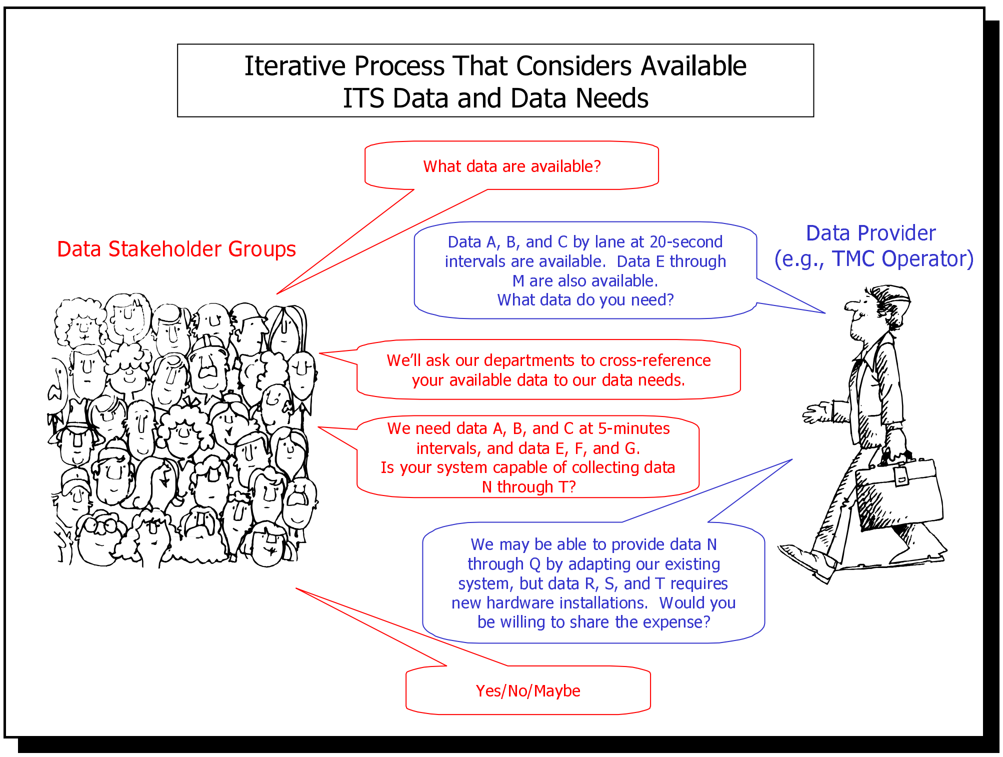


Figure 22. Iterative Process that Establishes Dialogue between Data Stakeholder Groups and Data Provider

The exact process used to incorporate user requirements into ITS data archiving systems will vary among urban areas based upon technical details of the ITS data collection, as well as existing relationships between public and private transportation agencies. Strong existing institutional partnerships and teaming arrangements are most conducive to the process shown in Figure 22, which is essentially the development of a data partnership.

Previous Work on Data Needs and User Requirements

In designing an ITS data archiving system, the definition or establishment of user requirements does not necessarily have to start from "scratch" with a blank sheet of paper. Several reports written in the past five years document user requirements and data needs (Table 9). These comprehensive lists of documented data needs can serve as a starting point to be cross-referenced with actual data needs for existing and anticipated analyses. Other data collection or data needs documents specific to your urban area or state may be available from the transportation planning divisions of your state DOT, city DOT, or regional MPO.

Perhaps the most useful description of data application needs and user requirements to date has been those included in the preliminary requirements for ADUS (4). Because of their utility in developing ITS data archiving systems, these descriptions are excerpted in full on the following pages as Table 10. A tabular summary of common data applications by stakeholder group was presented earlier as Table 4 on pages 21 and 22.

Table 9. Documentation of Data Needs in Transportation Decision Making

Reference	Comments
ITS as a Data Resource: Preliminary Requirements for a User Service (4)	Comprehensive treatment of user requirements and data needs as it specifically relates to archived ITS data. Addresses stakeholder groups and typical applications, as well as potential ITS data sources and relevant applications.
HPMS Field Manual (36)	Defines data collection requirements for federal Highway Performance Monitoring System (HPMS) data submittals.
Traffic Monitoring Guide (37)	Provides guidance for improved traffic counting, vehicle classification, and truck weighing. Also provides statistical procedures to determine the level of traffic monitoring needed to achieve a desired precision level.
AASHTO Guidelines for Traffic Data Programs (38)	Provides recommendations for traffic data programs based on several foundational principles: common traffic monitoring practice; need to provide quality data for decision making; and "truth-in-data", or the disclosure of practice and estimate of variability.
ASTM Standard Practice for Highway-Traffic Monitoring (ASTM E1442-94) (39)	Describes standard practices for traffic monitoring (i.e., traffic volume, classification, and weight), including the principle of "truth-in-data."
Information Needs to Support State and Local Decision Making into the 21st Century (40)	Contains recommendations regarding the data needed to improve decision making in the future. Data issues were categorized into socioeconomic, financial, supply and system characteristics, system operations, and impact and performance data.
Guidance Manual for Managing Transportation Planning Data (41)	Presents guidance for identifying, collecting, organizing, and using data for transportation planning purposes. Suggested data organization framework has following categories: supply, demand, system performance, and system impact attributes.
Identification of Transportation Planning Data Requirements in Federal Legislation (42)	Identifies planning and associated data collection requirements established by the Clean Air Act Amendments and ISTEA.
Data for Decisions: Requirements for National Transportation Policy Making (43)	Provides an assessment of data needed for national decision making.

Table 10. ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	**Spatial coverage**	**Temporal coverage**		
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	• Congestion monitoring • Link speeds for TDF and air quality models • AADT, K- and D-factors • Saturation flow rates • Pre-planned TMC operations
	• vehicle classification • vehicle weight	• loop detectors • WIM equipment • video imaging • acoustic	Usually 50-100 per state; by lane	Usually hourly	Pre-screening for weight enforcement	• Truck percents by time of day for TDF and air quality models • Truck flow patterns • Pavement loadings
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	• Network details for microscopic traffic simulation models (e.g., TRAF, TRANSIMS)
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	• Network details for microscopic traffic simulation models (e.g., TRAF, TRANSIMS) • Pre-planned TMC operations
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	• Congestion monitoring • Car-following and traffic flow theory
Vehicle counts from electronic toll collection	• time • location • vehicle counts	Electronic toll collections equipment	At instrumented toll lanes	Usually full-time	Automatic toll collection	• Traffic counts by time of day
TMC-generated traffic flow metrics (forecasted or transformed data)	• link congestion indices • stops/delay estimates	TMC software	Selected roadway segments	Hours of TMC operation	• incident detection • traveler information • preemptive control strategies	• congestion monitoring • effectiveness of prediction methods

Table 10. ITS Data Relevant for Archiving (Cont.)

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	**Spatial coverage**	**Temporal coverage**		
ARTERIAL AND PARKING MANAGEMENT						
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 20-60 second intervals	• progression setting • congestion/queue identification	• congestion monitoring • link speeds for travel forecasting models (free flow only) • AADT, K- and D-factors
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	Network details for microscopic traffic simulation models (e.g., TRAF, TRANSIMS)
Parking management	• time • lot location • available spaces	Field controllers	Selected parking facilities	Usually day time or special events	Real-time information to travelers on parking availability	Parking utilization and needs studies
TRANSIT AND RIDESHARING						
Transit usage	• vehicle boardings (by time and location) • station origin and destination (O/D) • paratransit O/D	Electronic fare payment systems	Transit routes	Usually full-time	Used for electronic payment of transit fares	• route planning/run-cutting • ridership reporting (e.g., Section 15)
Transit route deviations and advisories	• route number • time of advisory • route segments taken	TMC software	Transit routes	Usually full-time	Transit route revisions	Transit route and schedule planning
Rideshare requests	• time of day • O/D	computer-aided dispatch (CAD)	Usually areawide	Day time, usually peak periods	Dynamic rideshare matching	• travel demand estimation • transit route and service planning

Table 10. ITS Data Relevant for Archiving (Cont.)

ITS data source	Primary data elements	Features of the Data Source: Typical collection equipment	Features of the Data Source: Spatial coverage	Features of the Data Source: Temporal coverage	Real-time uses	Possible multiple uses of ITS-generated data
INCIDENT MANAGEMENT AND SAFETY						
Incident logs	• location • begin, notification, dispatch, arrive, clear, depart times • type • extent (blockage) • HazMat • police accident report reference • cause	• CAD • computer-driven logs	Extent of Incident Management Program	Extent of Incident Management Program	Incident response and clearance	• incident response evaluations (program effectiveness) • congestion monitoring (e.g., % recurring vs. nonrecurring) • safety reviews (change in incident rates)
Train arrivals at highway-rail intersections	• location • begin time • end time	Field controllers	At instrumented HRIs	Usually full-time	• coordination with nearby traffic signals • notification to travelers	Grade crossing safety and operational studies
Emergency vehicle dispatch records	• time • O/D • route • notification, arrive, scene, leave times	CAD	Usually areawide	Usually full-time	Coordination of Emergency Management response	• Emergency management labor and patrol studies • Emergency management route planning
Emergency vehicle locations	• vehicle type • time • location • response type	Automatic Vehicle Identification (AVI) or GPS equipment	Usually areawide	Usually full-time	• tracking vehicle progress • green wave and signal preemption initiation	• Emergency management route planning • Emergency management response time studies
Construction and work zone identification	• location • date • time • lanes/ shoulders blocked	TMC software			Traveler information	Congestion monitoring

Table 10. ITS Data Relevant for Archiving (Cont.)

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	Spatial coverage	Temporal coverage		
COMMERCIAL VEHICLE OPERATIONS						
HazMat cargo identifiers	• type • container/package • route • time	CVO systems	At reader and sensor locations	Usually full-time	• Identifying HazMat in specific incidents • routes for specific shipments	• HazMat flows • HazMat incident studies
Fleet Activity Reports	• carrier • citations • accidents • inspection results	CVO inspections	N/A	Usually summarized annually	May overlap with SAFETYNET functions	
Cargo identification	• cargo type • O/D	CVO systems	At reader and sensor locations	Usually full-time	Clearance activities	Freight movement patterns
Border crossings	• counts by vehicle type • cargo type • O/D	CVO systems	At reader and sensor locations	Usually full-time	Enforcement	Freight movement patterns
On-board safety data	• vehicle type • cumulative mileage • driver log (hrs. of service) • subsystem status (e.g., brakes)	CVO systems	At reader and sensor locations	Usually full-time	Enforcement and inspection	Special safety studies (e.g., driver fatigue, vehicle components)
ENVIRONMENTAL AND WEATHER						
Emissions Management System	• time • location • pollutant concentrations • wind conditions	Specialized sensors	Sensor locations	Usually full-time	Identification of hotspots and subsequent control strategies	• trends in emissions • special Air Quality studies
Weather data	• location • time • precipitation • temperature • wind conditions	Environmental sensors	At sensor locations	Usually full-time	Traveler information	• congestion monitoring (capacity reductions) • freeze/thaw cycles for pavement models

Table 10. ITS Data Relevant for Archiving (Cont.)

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	**Spatial coverage**	**Temporal coverage**		
VEHICLE AND PASSENGER INFORMATION						
Location referencing data	Special case; pertains to all location references in ITS and planning					Need conversion from lat/long to highway distance and location (e.g., milepost references for queue lengths)
Vehicle probe data	• vehicle ID • segment location • travel time	• probe readers and vehicle tags • GPS on vehicles	GPS is areawide; readers restricted to highway locations	Usually full-time	• coordinate traffic control response • congestion/queue identification • incident detection • real-time transit vehicle schedule • adherence • electronic toll collection	• congestion monitoring • link speeds for travel forecasting models • historic transit schedule adherence • traveler response to incidents or traveler information • O/D patterns
VMS messages	• VMS location • time of message • message content	TMC software	VMS locations	Hours of TMC operation	Traveler information	Effects of VMS message content on traveler response
Vehicle trajectories	• location (route) • time • speed • acceleration • headway	• AVI or GPS equipment • advanced video image processing	AVI restricted to reader locations; GPS is areawide	1-10 second intervals	Collected as part of surveillance function	• Traffic simulation model calibration for local conditions (driver type distributions) • Modal emission model calibration • Traffic flow research
TMC and Information Service Provider generated route guidance	• time/date • O/D • route segments • estimated travel time	TMC/Information Service Provider software	Usually areawide	Hours of TMC operation	Traveler information	• O/Ds for TDF model inputs • Interzonal travel times for TDF model calibration
Parking and roadway (congestion) pricing changes	• time/date • route segment/lot ID • new price	TMC software	Facilities subject to variable pricing	Hours of TMC operation	Demand management	• Special studies of traveler response to pricing • Establishment of pricing policies

Source: Margiotta 1998 (4), pp. 10-13.

Summary of Factors to Consider in Developing ITS Data Archiving User Requirements

The definition of ITS data archiving user requirements can appear particularly daunting given the large number of stakeholders and potential uses for the data. There are several considerations summarized below that may help in the process of developing user requirements and designing archived ITS data archiving systems:

- **Establish a dialogue between TMC operators/data providers and data users -** The best synergy for archiving ITS data occurs when there is an ongoing dialogue between TMC operators (or equivalent data providers) and data users. An ongoing dialogue will help in the following ways: 1) data users may better understand the available ITS data and its intricacies; 2) operations personnel may better understand the needs of data users; and 3) operations personnel and data users may work together in establishing ITS detector/sensor designs that efficiently meet the data needs of many groups.

- **Reference existing user requirements as a starting point -** Significant effort has resulted in a generic set of user requirements for archived ITS data (Tables 4 and 10). Local agencies should utilize these existing user requirements as a starting point in defining urban area or agency-specific user requirements for archived ITS data systems.

- **Start simple and build on early success -** This fundamental approach is relevant here based on the documented failures of data management and data warehouse efforts in the business environment. These experiences indicate that many of the unsuccessful attempts at developing data warehouses have been due to failure to meet basic user requirements in a timely fashion. In some cases, excruciating amounts of time and detail go into system design requirements, with the end result being a data management system delivered several years late that no longer meets evolving data user needs. In other cases, basic user requirements are not incorporated into a data management system, yielding an expensive database that does not meet basic user needs.

AGGREGATION OF ITS TRAFFIC MONITORING DATA

To date, there appear to be two basic approaches for aggregating and retaining ITS data: 1) a cost-effective approach that entails archiving aggregated ITS data based on existing or anticipated data needs; and 2) a resource-intensive approach that entails archiving all ITS data to provide maximum flexibility to data users for conducting existing analyses as well as data exploration. The research team suggests a third, alternative approach that combines the philosophy of the two previous approaches: the focus of ITS data archiving is on meeting data needs for existing and anticipated analyses, but ITS data management system designs can provide the capability for advanced data archiving capabilities, such as data sampling, archiving on demand, and data broadcasts. These three basic approaches are summarized in Table 11 and the following paragraphs.

Approach 1: Save Aggregated Data

The cost-effective approach to design archived ITS data storage and management systems is to compare the data needs from existing and anticipated analyses to the data being collected by TMCs. With this method, each data need corresponds to an existing or anticipated analysis in which: 1) manual data collection or estimation procedures have been used in the past to collect the data to support this analysis; or 2) no data have been used in the past to support the analysis. The difficulty and cost of manual data collection or estimation has affected many transportation analyses, in that most existing analyses have been designed or developed to deal with data that can be easily collected or estimated. Data collected from ITS applications, however, typically have a much greater level of detail (in time and space) than data from manual collection efforts, as well as providing data that typically have not or could not be collected manually. Therefore, it is hypothesized that if ITS data are archived at a level necessary only to support existing or anticipated analyses, future analyses may not evolve to fully exploit the detailed nature of ITS data.

Approach 2: Save All Data

The maximum flexibility approach is to archive all ITS data so that data analysts may have flexibility in performing analyses, as well as the flexibility to develop analyses and tools that are capable of fully exploiting the detailed nature of ITS data. With this approach, large ITS data sets are made accessible to all potential data users and analysts. Users with data needs for existing analyses are still able to access the archived ITS data; however, the archived data is very detailed and the corresponding data sets are typically large and cumbersome, requiring powerful desktop or database server computers to process the data into an aggregated format. The advantage of having these large detailed ITS data sets is that it enables data exploration and mining, which can lead to more comprehensive, non-traditional analyses that were not previously possible because of the lack of supporting data.

Table 11. Comparison of Approaches for Aggregating Archived ITS Data

Approach 1: Archive Aggregated ITS Data Based on Existing and Anticipated Analyses	Approach 2: Archive All ITS Data to Provide Flexibility in Unanticipated Analyses	Approach 3: Focus on Strategic ITS Data Archiving but Provide Advanced Capabilities
PHILOSOPHY: • Base design solely on existing and anticipated data needs and analyses. • Although data storage is inexpensive, database management and administration costs can be significant. • Data are stored in medium-size databases, making it moderately easy for regular data users to access and manipulate data.	**PHILOSOPHY:** • Base design on archiving all data elements collected or generated through TMC, then users can select data of interest. • Data storage (e.g., CD, DVD, disk drive) costs are insignificant and should not affect system design. • Data are stored in very large, detailed data sets or log files, making it somewhat difficult to access or manipulate data (with the exception of advanced data users).	**PHILOSOPHY:** • Design is focused on existing and anticipated analyses, but advanced data archiving capabilities provide for analysis flexibility. • Core data elements are provided to regular data users through an accessible database interface, and various archiving capabilities can be adapted to help advanced users meet their needs.
ADVANTAGES: • Many existing analytical tools can be used. • Data management costs are kept to a reasonable level. • Ease of data access and manipulation for regular data users.	**ADVANTAGES:** • Provides complete flexibility for existing analyses as well as developing new analyses. • Process to define data user requirements not necessary.	**ADVANTAGES:** • Meets needs of most users by archiving core data for existing analyses, as well as providing advanced capabilities for advanced data users. • Combines the primary advantages of other two approaches.
DISADVANTAGES: • Data exploration and mining by advanced data users are limited by aggregated data sets.	**DISADVANTAGES:** • May require extensive resources for on-line, easily accessible implementation. • Analytical tools must be developed or adapted to work with large, detailed data sets. • Detailed data sets may require more knowledge/training by inexperienced users.	**DISADVANTAGES:** • Software and application development costs may be higher than other approaches.

Approach 3: Focus on Strategic Archiving, Build Advanced Capabilities

The first two approaches have their benefits and limitations. The third approach is an attempt to combine several benefits of the first two approaches. The third approach focuses on meeting data needs for existing analyses while still providing the advanced capabilities to archive all ITS data. These advanced archiving capabilities, which are described in subsequent sections, include:

- **data sampling -** samples of raw, disaggregate data can be archived periodically (e.g., sampling raw data every "8^{th}" day or one continuous week every three months, etc.);

- **aggregation based on statistical variability -** aggregation of data is based upon statistical variability, such that data in periods with little to no changes in traffic conditions may be aggregated more than data during dynamic, constantly-changing traffic conditions;

- **archiving on demand -** systems enable users to specify time(s) and location(s) where raw, disaggregate data will be archived, and limited temporary space is available for users until they move their "special study" data off-line; and

- **data broadcast -** raw, disaggregate data streams are "broadcast" to users, who are able to save at the aggregation level that best suits their needs.

The third approach retains the cost-effectiveness by primarily storing and managing only data that is needed, but it also provides the capability to archive all ITS data for limited periods of time for data exploration and mining. For example, a TMC may routinely archive system performance data at 15-minute intervals and make this data accessible through an on-line database. However, the TMC may also archive (on demand or as a routine practice) a full week of raw system performance data for every month of the year. Alternatively, the TMC may maintain a fixed amount of temporary disk space for providing archived data on request. The system design using this approach economically meets basic data needs as well as providing flexibility for data analyses.

The approach taken in designing an ITS data archiving system may vary between regions, as a number of local factors will determine the most desirable or realistic approach. These factors include presence and cooperation of archived data users, institutional culture, available resources, and current system design, to name a few. The following sections describe in more detail the advanced data archiving capabilities described above.

Data Sampling

As described earlier, sampling of raw, disaggregate data can be performed in conjunction with aggregated data archival to meet the needs of advanced data users. In many cases, only spatial or temporal samples of disaggregate data are necessary for these advanced users (as opposed to continuous coverage). Of course, any plans for data sampling should be coordinated with the needs of the data users. Data sampling is applicable when it may not be affordable or desirable to archive all disaggregate data. Two examples of data sampling are shown here: 1) sampling every "nth" day, and 2) sampling "n" concentrated weeks per "x" months.

For the first sampling approach, data can be saved at the most disaggregate level for every "nth" day (Table 12 shows data archiving for every 8th day). Aggregated data (e.g., 15-minute, hourly, etc.) that meets basic user needs can also be saved for all days of the week, thereby providing continuous aggregated data. With such an example, detailed data would be available for each day of the week after seven weeks (about two months).

Table 12. Illustration of Every Nth Day Sampling

Week Number	Day of Week						
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	X						
2		X					
3			X				
4				X			
5					X		
6						X	
7							X
8	X						

Note: "X" denotes days when data are archived at the most disaggregate level (e.g., 20 seconds). More aggregate summaries (e.g., 15-minutes, hourly, daily) can be saved for all days throughout the year.

The second example of data sampling is illustrated in Table 13, which shows a "concentrated weeks" sampling approach. Rather than saving detailed data on rotating days as shown in Table 12, this technique concentrates the detailed data archiving effort to a one- or two-week period every three months. Of course, this could be performed as often as necessary depending upon data needs (i.e., every three, four, or even six months). The clear disadvantage of such a technique is that data will not be collected for every month as with the Nth day sampling plan. Also, the concentrated weeks sampling would have to be performed during "typical" weeks, avoiding any holiday or other periods that are not representative of the larger three-month period.

However, this data collection method has the advantage that it can be scheduled on a quarterly basis and personnel can plan for the scheduled data collection effort. This may prove useful in systems that are less automated.

Table 13. Illustration of Concentrated Weeks Sampling

Month 1							Month 2							Month 3						
S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S
							X	X	X	X	X	X	X							
							X	X	X	X	X	X	X							

Note: "X" denotes days when data are archived at the most disaggregate level (e.g., 20 seconds). More aggregate summaries (e.g., 15-minutes, hourly, daily) can be saved for all days throughout the year.

Aggregation Based on Statistical Variability

This section focuses on a specific advanced data archiving capability--data aggregation based upon statistical variability. A general discussion of data summarization from a statistical point of view is presented first. The authors then introduce two statistical techniques that can be used to determine optimal aggregation levels based upon the variability of traffic parameters. The statistical methods are applied to ITS traffic data collected by the TransGuide® traffic management center in San Antonio, Texas. The results of the analyses are presented, as well as an interpretation of the analysis results, including implications for traffic management operators or others interested in archiving ITS data.

Theory of Minimal Sufficient Statistics

There is a theory in statistics known as "minimal sufficient statistics" of the parameters that define a probability density function (44). This theory addresses the issue of whether all the data observed (or collected) should be saved, or if condensing the data and studying the condensed data is equivalent without significant loss to the original data. When data come from the same population, say a sample of size *n*, this theory enables statistics of relatively smaller quantity (rather then all of the data) to be the only information necessary to "reproduce" the data. For example, it is a well-known fact that the mean and variance are the minimal sufficient statistics for data from the same normal distribution. In ITS applications (and other fields) it is desired to summarize speed or travel time data with the mean and the variance over a particular time. The question of "what aggregation width should be selected?" inevitably arises. In current practice, it is rather arbitrary as to what aggregation to use. In the TransGuide® system, loop detectors collect aggregated speeds every 20 seconds. The data collected are averages, and with increasing sample size, the averages become normally distributed according to the Central Limit Theorem.

The likelihood of the speed data conditional on the volume within each 20-second interval is written in Equation 1. From here, the sufficient statistics for the mean and the variance for the normal distribution are determined (where $x_1, x_2, x_3, ..., x_N$ are assumed to be independent such that $x_i \sim N(\mu, \sigma^2 / n_i)$ according to the Central Limit Theorem). Notice that the normal distribution is of the exponential family as shown in Equation 2. Using standard arguments (44), the minimal sufficient statistics can be found.

$$f_{x_1, x_2, \ldots x_N | \mu, \sigma, n_1, n_2, \ldots n_N} = \frac{\exp\left(\sum_{i=}^{n} \left(x_i - \mu\right)^2 / \left(\sigma^2 / n_i\right) \right)}{(2\pi)^{N/2} \sigma^{N/2} \prod_{i=1}^{N} 1/\sqrt{n_i}} \quad (1)$$

which can be expanded to:

$$f_{x_1, \ldots x_N | \mu, \sigma, n_1, \ldots, n_N} = \frac{\exp\left(-\left(\sum_{i=}^{N} x_i^2 n_i^2 / \sigma^2 - 2\sum_{i=1}^{N} x_i n_i \mu / \sigma^2 + \mu^2 / \sigma^2 \sum_{i=1}^{N} n_i^2 \right) / 2 \right)}{(2\pi)^{N/2} \sigma^{N2} \prod_{i=1}^{N} 1/\sqrt{n_i}} \quad (2)$$

Therefore, the minimal sufficient statistics for the mean and variance (μ,σ^2) and conditional on volumes is $\sum_{i=1}^{N} x_i n_i$, $\sum_{i=1}^{N} x_i^2 n_i$. This argument justifies the above statistics as being minimal sufficient statistics based on standard arguments found in texts like Silvey. This indicates it is sufficient to save the weighted mean and the weighted variance.

It is rather difficult (maybe impossible) to use the same argument for the 20-second volumes observed and incorporate this into the likelihood argument, because the structure of the likelihood is rather complex. To keep the argument simple, the speed distribution is conditioned on the volumes as well as the true mean and variance. What the researchers propose is to save the mean and the variance of the volume data (justified by Equations 1 and 2). It is assumed that the optimal aggregation methods suggested in this paper will be driven by the mean speeds.

It is well known that most urban roadways have a different speed distribution during the peak period relative to the off-peak period. Therefore, for a particular part of a highway on a particular day, the data do not come from the same normal distribution. At some point there is a "change point" in the system. These techniques will group data into their own normal distributions and suggest summary statistics, thus searching for an optimal aggregation width for each hour of the day for each lane. The following paragraphs discuss two statistical techniques for determining optimal aggregation width for ITS data archiving.

Algorithm 1: Cross-Validated Mean-Squared Error

Optimal aggregation widths within each hour of the day are determined with this algorithm. Therefore, the maximum aggregation width is 3,600 seconds and the minimum is 60 seconds. The problem of finding optimal aggregation widths is similar to bandwidth selection of a kernel in non-parametric function estimation (45). The underlying principle of the cross-validated mean-squared error (CVMSE) is to take one value out, attempt to predict this value with an aggregated mean and then take this difference squared and average this for all values within the time period of interest (Figure 23).

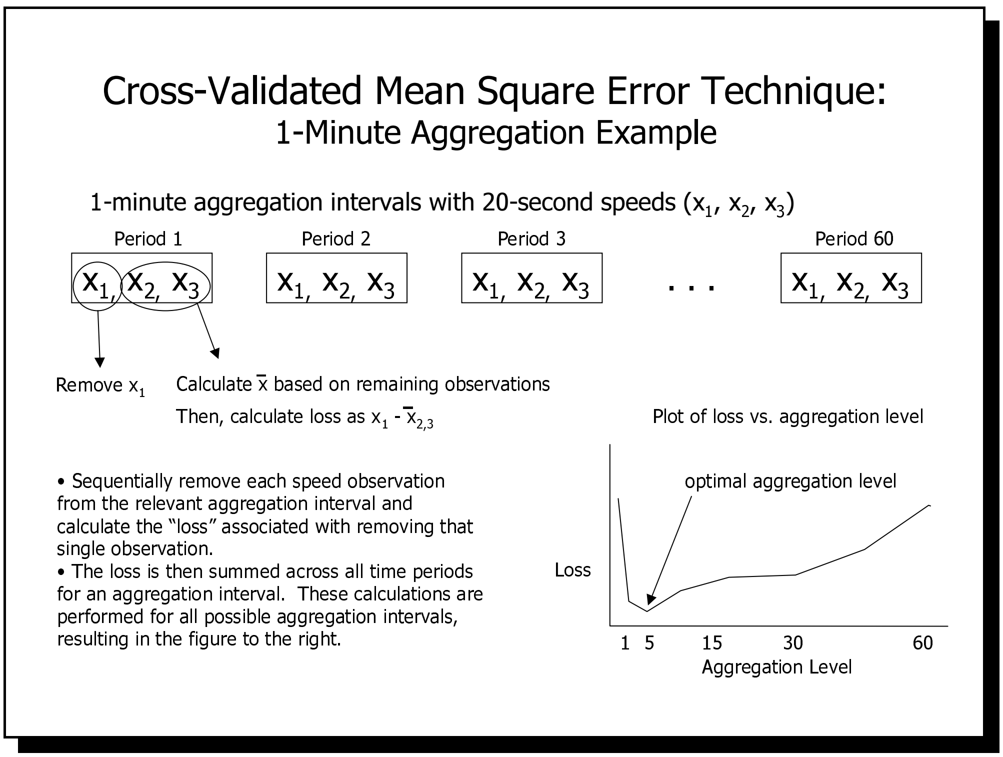


Figure 23. Cross-Validated Mean Square Error Technique

The following notation is necessary for the discussion that follows.

x_{ij} the i^{th} 20-second average speed within aggregation interval j
n_{ij} the volume for the i^{th} 20-second average speed within aggregation interval j
m_a number of 20-second intervals within each aggregation width for aggregation a
(Example: m_{60} = 3, the number of 20-second intervals within 60 seconds)

$$N_j = \sum_{i=1}^{m_a} n_{ij} = n_{\cdot j}$$

$\bar{x}_{\cdot j}$ the weighted mean within aggregation width j
$\bar{x}^{(i)}_{\cdot j}$ the weighted mean within aggregation width j, without value x_{ij}
L_j^a the CVMSE loss within aggregate j with aggregate width a
L_a the weighted sum of L_j^a, which represents the CVMSE for 3,600 seconds

First, a test is performed to determine whether the x_{ij} within a 3,600-second interval come from the same distribution. Since x_{ij} are averages, $x_i \sim N(\mu, \sigma^2 / n_i)$, if they come from the same distribution within each 20-second interval (this can also be assumed by the Central Limit Theorem). Therefore, the individual speeds can be standardized in the following way . If the data do come from the same normal distribution, then saving the weighted mean and variance of the speeds and the mean and the variance of the volumes are suggested. The asymptotic Kolmogorov-Smirnov goodness-of-fit test (46) is used to test whether the data come from the same normal distribution.

If the above test is rejected, one evaluates the CVMSE. For now, consider a fixed aggregation width, within one time interval.

$$L_j^a = \sum_{i=1}^{m_a} n_{ij}\left(x_{ij} - \bar{x}^{(i)}_{\bullet j}\right)^2 \tag{3}$$

Therefore,

$$L_a = \frac{\sum_{j=1}^{3,600/a} L_j^a}{\left(\sum_{j=1}^{3,600/a} N_j - 3,600/a\right)} = \frac{\sum_{j=1}^{3,600/a} \sum_{i=1}^{m_a} n_{ij}\left(x_{ij} - \bar{x}^{(i)}_{\cdot j}\right)^2}{\left(\sum_{j=1}^{3,600/a} N_j - 3,600/a\right)} \tag{4}$$

Notice that the loss function is weighted by the volume of vehicles within each second interval, and is penalized (in the denominator) for having a lot of aggregation widths. Now, the above loss function is rewritten.

$$L_j^a = \sum_{i=1}^{m_a} n_{ij}\left(x_{ij} - \bar{x}_{\bullet j}^{(i)}\right)^2 = \sum_{i=1}^{m_a} n_{ij}\left[\left(N_j - n_{ij}\right)x_{ij} - N_j\bar{x}_{\bullet j} + n_{ij}x_{ij}\right]^2$$
$$= \frac{\sum_{i=1}^{m_a} n_{ij}\left[N_j^2\left(x_{ij} - \bar{x}_{\bullet j}\right)^2\right]}{\left(N_j - n_{ij}\right)^2} \qquad (5)$$

There are two reasons for writing the CVMSE component in this form: the first is to notice that taking out one value at a time is not necessary to compute L_j^a. The second is to notice how this criteria works. The aggregation width with the smallest L is chosen. Summarizing data within intervals is desired. Notice that if $N_j \approx n_{ij}$, L will be large. If $N_j >> n_{ij}$, then the individual volumes are negligible, which prevents $a=20$ as being an optimal width. This objective function is minimized when $\hat{a} = \arg\min L_a$. That is, the objective function is minimized with respect to a and $\hat{a}$ is the estimated optimal aggregate width. One obtains plots as shown in the top portion of Figure 23.

Now consider $n = n_{ij}$ (all the volumes within a 20-second interval are the same). Therefore,

$N_j = \sum_{i=1}^{m_a} n_{ij} = m_a n$ and

$$L_j^a = \sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}^{(i)}\right)^2 = \sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j} + \bar{x}_{\bullet j} - \bar{x}_{\bullet j}^{(i)}\right)^2$$
$$= \sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}\right)^2 + \sum_{i=1}^{m_a} n\left(\bar{x}_{\bullet j} - \bar{x}_{\bullet j}^{(i)}\right)^2 + 2\sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}\right)\left(\bar{x}_{\bullet j} - \bar{x}_{\bullet j}^{(i)}\right)$$
$$= \sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}\right)^2 + \sum_{i=1}^{m_a} n\left(\bar{x}_{\bullet j} - \bar{x}_{\bullet j}^{(i)}\right)^2 + 2\sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}\right)\left(\bar{x}_{\bullet j} - \left(m_a n\bar{x}_{\bullet j} - nx_{ij}\right)/\left(m_a n - n\right)\right) \qquad (6)$$
$$= \sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}\right)^2 + \left(m_a + 1\right)/\left(m_a - 1\right)\sum_{i=1}^{m_a} n\left(\bar{x}_{\bullet j} - \bar{x}_{\bullet j}^{(i)}\right)^2$$

And so,

$$L = \frac{\left(\sum_{i=1}^{3,600/a} \sum_{i=1}^{m_a} n\left(x_{ij} - \bar{x}_{\bullet j}\right)^2 + \frac{(m_a+1)}{(m_a-1)} \sum_{i=1}^{3,600/a} \sum_{i=1}^{m_a} n\left(\bar{x}_{\bullet j} - \bar{x}_{\bullet j}^{(i)}\right)^2 \right)}{(180n - 3,600/a)}$$

$$= \frac{\left(\sum_{i=1}^{3,600/a} \sum_{i=1}^{m_a} \left(x_{ij} - \bar{x}_{\bullet j}\right)^2 + \frac{(m_a+1)}{(m_a-1)} 3,600/a \sum_{i=1}^{m_a} \left(\bar{x}_{\bullet j}^{(i)}\right) 2 \right)}{(180n - 3,600/a)} \tag{7}$$

Therefore, the first term is an average variance and the second is an average bias.

Algorithm 2: F-Statistic

The second algorithm developed uses the F-statistic from a one-way analysis of variance (ANOVA). Suppose the question is whether one can aggregate from 1 minute to 5 minutes. The question can be viewed as: are the 20-second realizations from the five treatments (1-minute widths within 5 minutes) from different populations or the same population? Statistically the hypothesis is H_o: $\mu_1 = \mu_2 = ... = \mu_5$ vs. H_a: *The means are not the same*. The F-statistic is calculated and aggregating at 5 minutes is recommended if H_o is not rejected. In this paper, one-hour periods are considered, so if one-minute vs. five-minutes, are tested, there are 12 test statistics (see Figure 24). To pool the overall error, the $min\ \{p_1,p_2,p_3,...,p_{12}\} < \alpha/12$ is suggested as the rejection region. That is, if one of the p-values is less than $\alpha/12$, aggregation is done at one minute not five minutes. The test is done sequentially and if this initial test is not rejected, one looks at five vs. 15 minutes, 15 vs. 30 minutes, and 30 minutes vs. one hour.

There are some advantages and disadvantages to this method over the CVMSE. The advantage is that the method uses a justifiable test between groups to rely on a cut-off value, if the differences between groups are significant at the α level. This can lead to an application-dependent algorithm. What this means, statistically, is the α level can be adjusted to the power which the user is willing to accept. For some transportation applications, the probability of accepting H_o when in fact H_a is true need not be as high as others (e.g., planning versus operational applications).

The disadvantage of this method is a fundamental issue. For the F-test to be valid, the within group statistics must be identically distributed. As shown in the previous section, this is not true since the 20-second speeds have different volumes, which result in different distributions. However, if the volumes within each 20-second interval are approximately equal the algorithm is valid.

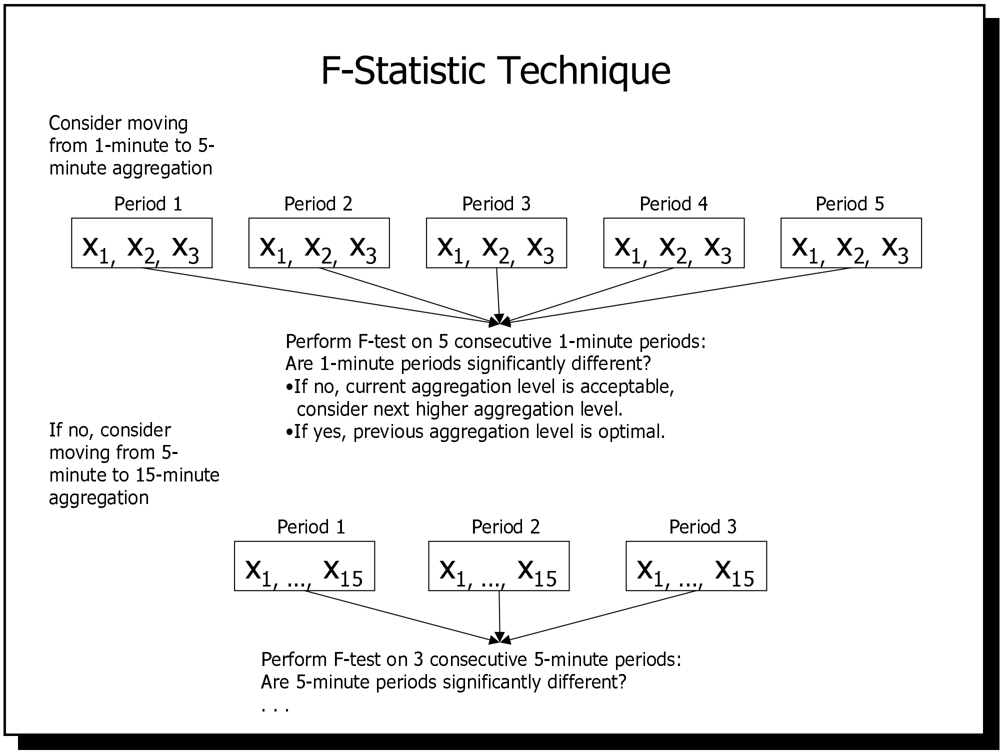


Figure 24. F-Statistic Technique

ITS Data Source

The TransGuide® advanced traffic management center (TMC) in San Antonio collects traffic data using point detectors (e.g., inductive loops) and probe vehicles (e.g., automatic vehicle identification (AVI) transponders). For this study, the authors focused on analyzing the variability of the loop detector data. The TransGuide® loop detectors are typically located in every lane and nominally spaced every 0.8 km (0.5 mi) (loop detectors are also located on all entrance and exit ramps). Each loop detector station on the main freeway lanes is located in a trap, or double-loop configuration, where two loops are spaced about 10 m (30 ft) apart. The first loop detector collects vehicle counts and lane occupancy (e.g., percent of time that the loop is occupied by vehicles). The arrival time difference between consecutive loops is used to calculate a spot speed for each lane loop detector. Local controller units (LCUs) in the field accumulate the collected information, and three computer servers at the TransGuide® center poll, or retrieve, the aggregated data from the LCUs in a sequential pattern. The system gathers the following from each lane loop detector station every 20 seconds:

- average spot speed (mph);
- vehicle volume (number of vehicles); and,
- lane occupancy (percent of time loop is occupied).

The data selected for this analysis come from three of the most congested locations at which TransGuide® currently collects traffic data:

- IH-410 North, Milepost 13.117 (Device ID: L-0410N-013.117)
- US 281 North, Milepost 145.396 (Device ID: L-0281N-145.396)
- IH-10 West, Milepost 566.641 (Device ID: L-0010W-566.641)

Each location has three lanes in each direction and experiences recurring congestion on a daily basis. A full week of archived ITS data (May 3 to 7, 1999) was used to test the statistical methods described in the following sections. The findings below show results from one location, although the results were consistent across all three locations.

Findings

Table 14 and Figures 25 and 26 show typical results for both statistical techniques by lane and day of week. For the F-statistic algorithm, five possible aggregation levels were considered in minutes (1, 5, 15, 30, and 60). Table 14 shows the hourly optimal aggregation widths in minutes for each algorithm by day and lane number. Note that a "1^-" indicates data could be kept at 1 minute or less. A "60^+" indicates that data could be saved at 60 minutes or greater. The analysis for this paper was performed with 60-minute intervals and to determine if aggregation at levels greater than 60 minutes is possible, the analysis must be performed at these levels. Casual observation of these results indicate that it is much easier to obtain higher aggregation widths with the F-statistic than the CVMSE. This is because the F-statistic is generally less sensitive than the CVMSE when the F-statistic is computed at the Type I error level (α) of 5 percent. This is especially evident in the afternoon hours.

Several key advantages and disadvantages of each of the algorithms are worth noting. Each of the techniques are relatively easy to program with low computing needs. While more theoretical and perhaps more foreign to first-time users, the CVMSE concept provides aggregation estimates that are application independent. When faced with the question of what aggregation width is ideal, TMC operators and managers are often torn between different applications and their aggregation data needs. This method provides an indication of when it may be ideally possible to aggregate the data, and at what level, from which the disaggregate data can be replicated without significant loss in the data set. Conversely, the F-statistic, while generally more familiar to engineers and scientists, is a function of the Type I error (α) considered acceptable. Since there is likely a relationship between the Type I error someone is willing to accept and their application, this method still provides a link to applications. However, this link is not fully defined as transportation professionals generally accept a 5 to 10 percent Type I error level for many different applications.

Table 14. CVMSE and F-Statistic Algorithm Results, IH-10 West, Milepost 566.641

CVMSE Results for Five Consecutive Days (May 3-7, 1999)

	Optimal Aggregation Widths (Minutes) for Hour Starting:																							
Lane	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	60^+	60^+	60^+	60^+	60^+	30	1	1^-	1	30	30	15	5	30	1	1^-	1^-	1^-	1^-	30	5	60^+	30	60^+
2	60^+	60^+	60^+	60^+	60^+	30	1	1^-	1^-	5	30	30	5	30	30	1^-	1^-	1^-	1^-	60^+	5	60^+	30	60^+
3	60^+	60^+	60^+	60^+	60^+	30	1	1^-	1	15	30	5	5	5	15	1^-	1^-	1^-	1	30	5	30	5	30
1	60^+	60^+	60^+	60^+	60^+	30	1	1^-	1^-	15	15	30	30	30	30	1^-	1	1^-	1	30	15	30	30	60^+
2	30	60^+	60^+	60^+	60^+	30	5	1^-	1^-	1	15	60^+	30	30	5	1^-	1^-	1^-	1	30	15	15	30	30
3	60^+	60^+	60^+	60^+	60^+	30	15	1^-	1	30	30	30	30	30	30	1^-	1	1^-	5	1	5	15	15	30
1	60^+	60^+	60^+	60^+	60^+	30	30	1^-	1^-	30	5	30	30	1^-	30	1	1^-	1^-	1	1	15	15	5	60^+
2	30	60^+	60^+	60^+	60^+	15	30	1	1^-	30	15	30	30	30	30	5	1^-	1^-	1	1	15	1	5	15
3	60^+	60^+	60^+	60^+	60^+	30	15	1^-	1^-	5	30	30	5	30	5	1	1^-	1^-	1	5	5	5	30	60^+
1	60^+	60^+	60^+	60^+	60^+	30	30	1^-	5	5	5	15	30	60^+	30	1	1^-	1^-	1^-	15	15	30	30	30
2	60^+	60^+	30	60^+	60^+	15	60^+	1^-	5	5	5	15	1^-	15	1	1	1^-	1^-	5	15	30	15	5	30
3	30	60^+	60^+	60^+	60^+	60^+	1	1^-	15	5	5	30	30	30	30	15	1^-	1^-	5	15	5	30	5	30
1	60^+	60^+	60^+	60^+	60^+	30	5	1^-	1^-	30	30	1^-	-	60^+	30	5	1^-	1^-	1	15	15	5	1	30
2	30	30	60^+	60^+	60^+	30	30	1^-	1^-	30	30	60^+	-	1^-	1^-	30	1^-	1^-	1^-	1	5	5	5	5
3	30	60^+	60^+	60^+	60^+	5	60^+	1	1	30	15	60^+	-	60^+	30	5	1^-	1^-	1^-	1	5	5	30	30

F-Statistic Results for Five Consecutive Days (May 3-7, 1999)

	Optimal Aggregation Widths (Minutes) for Hour Starting:																							
Lane	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	60^+	60^+	60^+	60^+	60^+	60^+	30	1^-	1^-	60^+	60^+	60^+	60^+	60^+	5	1^-	1^-	1^-	1^-	1^-	30	60^+	60^+	60^+
2	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	60^+	60^+	60^+	15	60^+	60^+	1^-	1^-	1^-	1^-	60^+	15	60^+	5	60^+
3	60^+	60^+	60^+	60^+	60^+	60^+	15	1^-	1^-	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	5	1^-	60^+	15	60^+	5	60^+
1	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	1^-	30	5	60^+	60^+	60^+	60^+	1^-	1^-	1^-	5	60^+	30	60^+	60^+	60^+
2	60^+	60^+	60^+	60^+	60^+	60^+	30	1^-	1^-	60^+	60^+	60^+	60^+	60^+	30	1^-	1^-	1^-	1^-	60^+	30	60^+	60^+	5
3	60^+	60^+	60^+	60^+	60^+	60^+	15	1^-	1^-	60^+	60^+	60^+	60^+	60^+	60^+	5	1^-	1^-	1^-	60^+	15	60^+	60^+	60^+
1	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	60^+	60^+	60^+	60^+	60^+	60^+	15	1^-	1^-	5	5	30	15	5	60^+
2	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	1^-	1^-	5	30	15	5	60^+
3	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	60^+	60^+	60^+	30	60^+	60^+	15	1^-	1^-	5	15	30	15	60^+	60^+
1	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	30	5	60^+	60^+	60^+	60^+	60^+	15	1^-	1^-	1^-	30	30	60^+	60^+	60^+
2	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	15	5	60^+	60^+	60^+	60^+	60^+	15	1^-	1^-	1^-	30	30	30	30	60^+
3	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	15	60^+	1^-	60^+	60^+	60^+	60^+	15	1^-	1^-	5	60^+	30	60^+	60^+	60^+
1	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	60^+	60^+	5	-	15	60^+	60^+	1^-	1^-	1^-	60^+	30	30	30	60^+
2	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	1^-	60^+	60^+	5	-	15	30	60^+	1^-	1^-	1^-	1^-	1^-	30	60^+	60^+
3	60^+	60^+	60^+	60^+	60^+	60^+	60^+	1^-	5	1^-	60^+	60^+	-	15	60^+	60^+	1^-	1^-	1^-	1^-	5	60^+	60^+	60^+

Note: "-" indicates insufficient data.

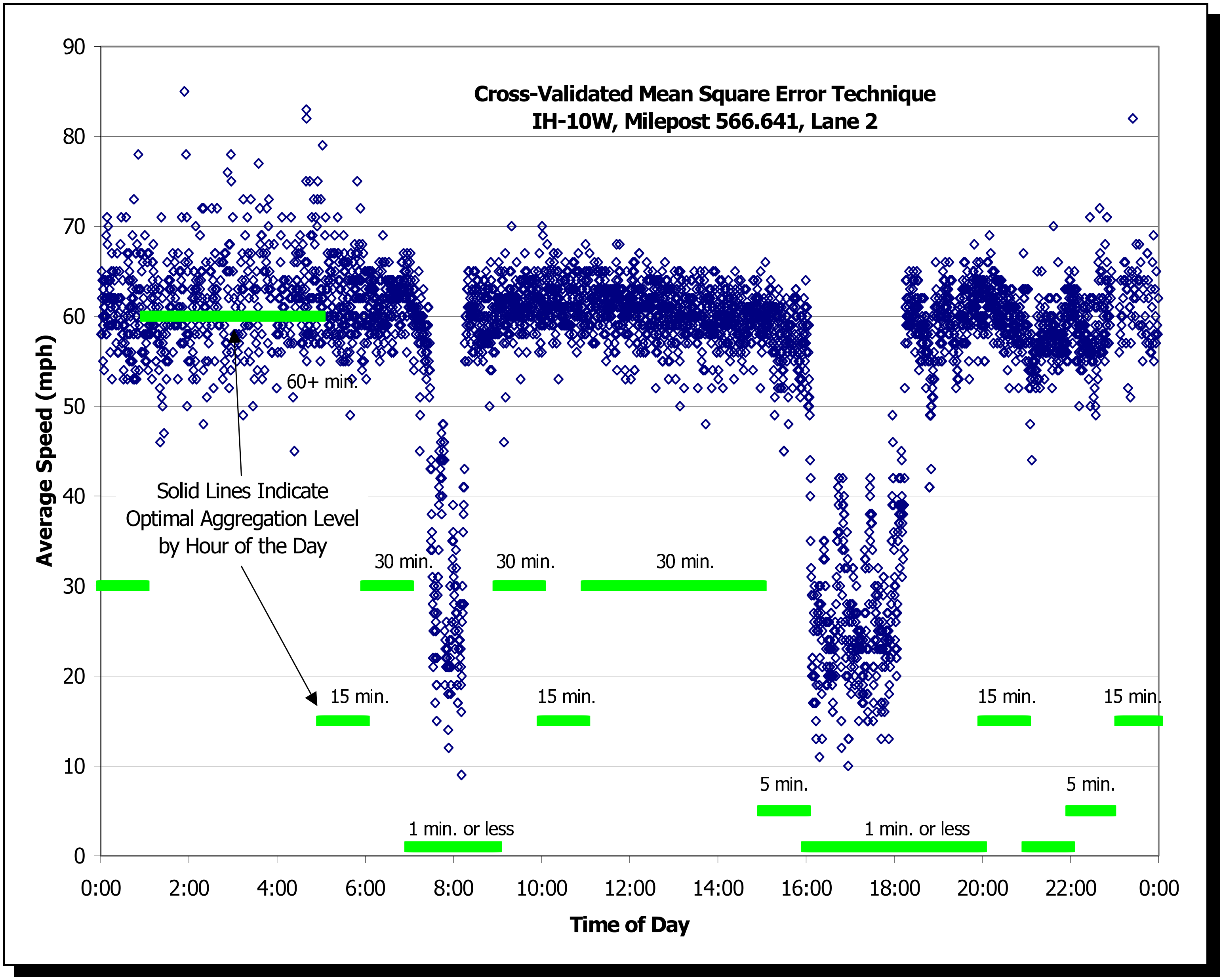


Figure 25. Typical Speed Profile and Results of CVMSE Analysis for Optimal Aggregation Level

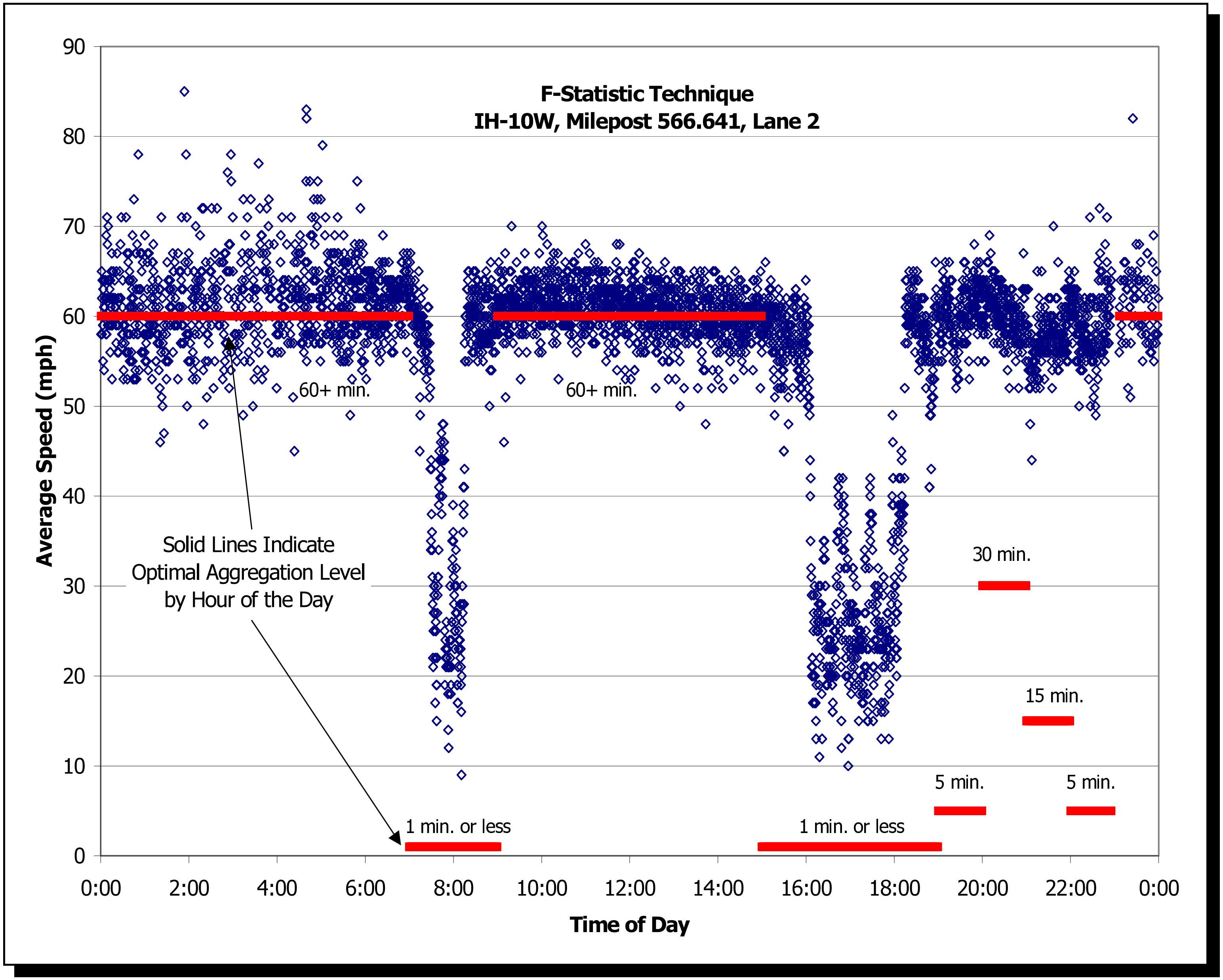


Figure 26. Typical Speed Profile and Results of F-Statistic Analysis for Optimal Aggregation Level

Summary of Findings and Conclusions

In this section, the research team presented two statistical techniques for use in determining optimal aggregation levels for archiving ITS traffic monitoring data. Although the two statistical techniques are methodically distinctive, they both seek to determine the minimal sufficient statistics necessary to characterize the full set of information in the traffic parameters distribution. The application or implementation of these statistical techniques may vary considerably based upon numerous local conditions; thus, the authors distinguish their conclusions about the statistical methods from how the methods can or should be applied to ITS data archiving.

The authors conclude the following related to the statistical techniques described in this section:

Results of the Statistical Techniques Appear Reasonable and Intuitive - Both algorithms produced intuitive results for optimal aggregation levels. For instance, both algorithms calculated optimal aggregation levels of 60 minutes or more during off-peak periods and other times with low speed variability. Both algorithms also calculated optimal aggregation levels of 1 minute or less during peak traffic periods or other periods with significant changes in traffic speeds. Although the authors consider the cross-validated mean square error algorithm to be the more robust of the two, it produced optimal aggregation levels that were more dynamic throughout the day than the F-statistic algorithm. In some cases, relatively low aggregation levels were calculated during periods of apparently low variability. The authors hypothesize that these results are due to the nature of the loss profiles during these two time periods (Figure 27). That is a flat slope on the loss profile during low variability periods still produces short aggregation time periods.

Importance of Saving Both Mean and Variance - The theory of minimal sufficient statistics suggests that, given certain circumstances, the mean and variance can be used to adequately describe a traffic parameter distribution. However, current practice is focused only on archiving mean parameter values. The authors recommend that ITS data archiving efforts strongly consider saving variance in addition to mean values for summary statistics.

Techniques Most Directly Related to Data Variability, Indirectly to Data Uses - One approach (as shown in the first column in Table 11) for determining adequate aggregation levels is to select a "minimum common denominator" based on existing data uses. The aggregation levels produced by the statistical techniques and illustrated in Figures 25 and 26 are most directly related to the variability of the traffic parameter, not the data application. For example, the cross-validated mean square error algorithm is based upon statistical theory of loss, not an acceptable error level (although presumably one could consider acceptable loss at less than optimal aggregation levels). The F-statistic algorithm is based on selecting the level of Type I error, which could potentially be related to the types of data uses. The researchers consider this attribute of application independence desirable given the evolving nature of archived data uses.

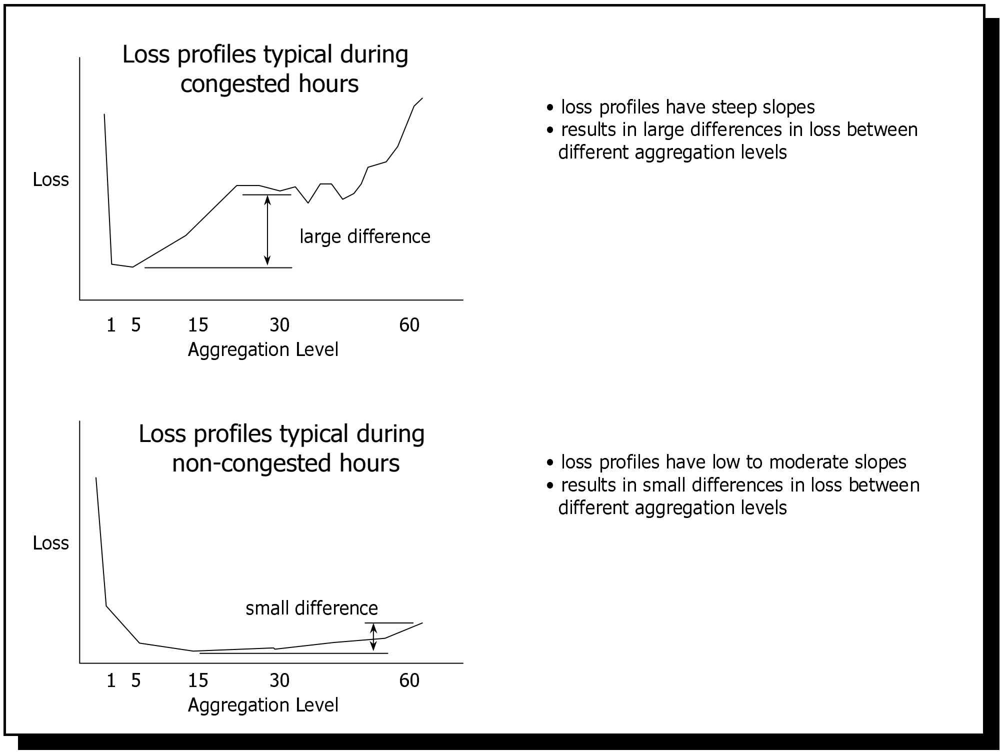


Figure 27. Illustration of Loss Profiles for the Cross-Validated Mean Square Error Technique

Techniques Are Applicable to Other Traffic Parameters - The techniques described in this paper were applied only to the traffic speed parameter, primarily because the traffic speeds were of most interest for further research applications. However, the two statistical algorithms could also be applied to any other traffic parameter of interest, such as vehicle volume, lane occupancy, vehicle weights, or vehicle classifications.

The following conclusions are provided relating to the application of the statistical techniques and results:

Wide Range of Possible Aggregation Solutions - The authors acknowledge that, although the statistical techniques described here may not be disputed, there is a wide range of possible aggregation solutions based upon these statistical techniques (Figure 28). Ultimately, the aggregation solutions may be driven by non-statistical parameters such as cost (e.g., "how much do we/the market value the data?"), ease of implementation, system requirements, and other constraints. Some may only desire a simple ITS data archiving solution for existing data needs, thereby negating the need for using statistical techniques or their results. For example, these users may construct a table of data uses such as that shown in Table 15, and use this table as a basis for determining the minimum aggregation level desired. Others may desire a single fixed aggregation level throughout the day, in which the statistical techniques or their results could be used to examine the tradeoffs between optimal aggregation levels and existing constraints. Still others may savor the challenge of implementing dynamic aggregation levels in ITS data archiving, in which case the statistical techniques could be programmed and automated within the data archiving and management system.

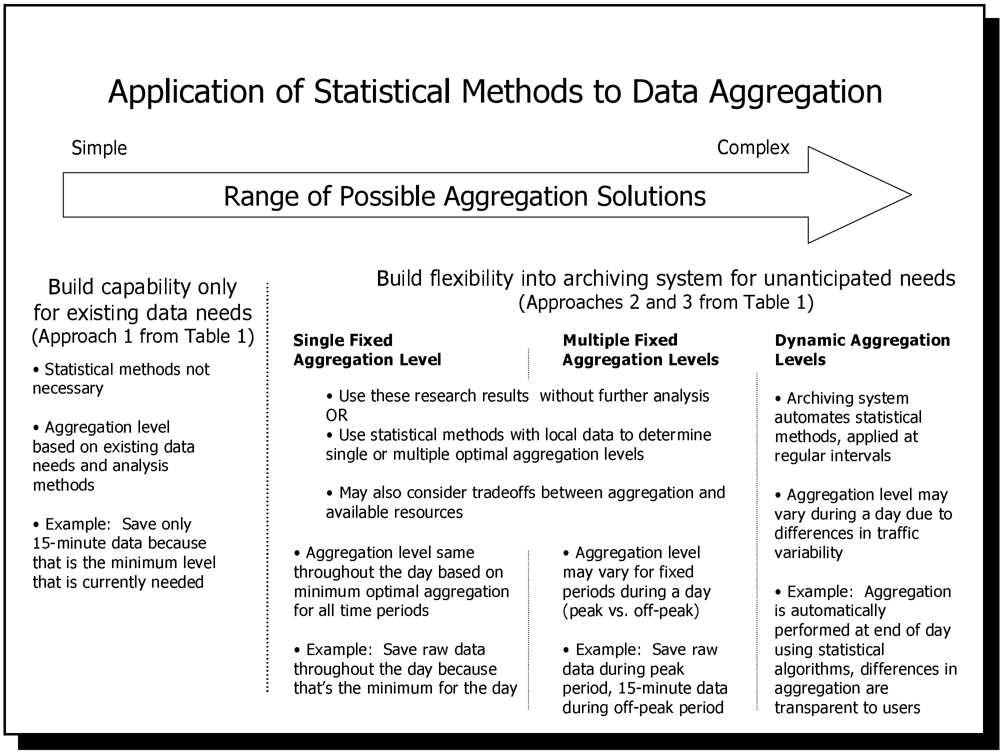


Figure 28. Range of Possible ITS Data Aggregation Solutions

Adequate Data Sets Necessary for Single Fixed Aggregation Levels - If the desired solution (or existing constraints) requires a single fixed aggregation level, the statistical techniques described here can be used to examine an adequate sample of local data. Typically, however, the minimum optimal aggregation level will always be 1 minute or less because of high variability conditions during peak hours. At this point, practitioners may begin to weigh the tradeoffs between less than optimal aggregation levels and reduced cost and/or other constraints.

Fixed Aggregation by Time-of-Day not Endorsed - Although fixed multiple aggregation levels during a day (e.g., 1-hour summaries off-peak, 5-minute summaries during peak period) is one of many possible solutions, it is not particularly endorsed by the authors as the most desirable solution. The primary drawback is that this practice assumes that high-variability traffic conditions only occur during peak periods, thus requiring a lower aggregation such as 5 minutes. This assumption often does not hold due to the nature of incidents and other traffic disruptive events that occur outside the peak period.

Table 15. Example Matrix: Level of Aggregation for ITS Data Archiving

ITS Data Applications	Level of Aggregation								
	Time					Space			
	No Aggregation (Ind. Veh.)	Less than 5 minutes	5 to 15 minutes	Hourly or Multi-Hour	Daily	Point (by Lane or Screenline)	Segment	Corridor	Sub-area or Region
Design and Operations									
Design future ITS components		D,W,M	D,W,M						
Develop historical travel time database			D,W	D,W					
Input/calibration for traffic models (traffic, emissions, fuel consumption)	D,W	D,W	D,W						
Real-time freeway and arterial street traffic control	D,W	D,W	D,W						
Route guidance and navigation	D,W	D,W	D,W						
Traveler information	D	D							
Incident detection	D	D							
Congestion pricing			D,W	D,W					
Planning									
Develop transportation policies and programs				M,Y	M,Y				
Perform needs studies/assessments				M,Y	M,Y				
Rank and prioritize transportation improvement projects for funding				M,Y	M,Y				
Evaluate project-specific transportation improvement strategies				M,Y	M,Y				
Input/calibration for mobile source emission models	D			M,Y	M,Y				

Table 15. Example Matrix: Level of Aggregation for ITS Data Archiving (Continued)

ITS Data Applications	Level of Aggregation								
	Time					Space			
	No Aggregation (Ind. Veh.)	Less than 5 minutes	5 to 15 minutes	Hourly or Multi-Hour	Daily	Point (by Lane or Screenline)	Segment	Corridor	Sub-area or Region
Input/calibration for travel demand forecasting models				M,Y	M,Y				
Calculate road user costs for economic analyses				M,Y	M,Y				
Evaluation									
Congestion management system/performance measurement				M,Y	M,Y				
Establish and monitor congestion trends (extent, intensity, duration, reliability)			M,Y	M,Y	M,Y				
Identify congested locations and bottlenecks			D,M,Y	D,M,Y					
Measure effectiveness and benefits of improvements (before-and-after studies)			M,Y	M,Y					
Communicate information about transportation problems and solutions				M,Y	M,Y				
Input/calibration for traffic models (traffic, emissions, fuel consumption)	D,W	D,W	D,W	D,W					

Notes: Shaded cells of the table represent applicable aggregation levels.
D = Daily; W = Weekly; M = Monthly; Y = Yearly (Annual).

Archiving on Demand

Another data archiving capability that can be provided for advanced users is archiving on demand. This capability permits data users to submit automated requests for raw or aggregated ITS data to be archived at specified times and locations. Archiving on demand is analogous to special studies using current terminology, in that specific data are collected in response to a periodic or infrequent study need. Temporary storage may be provided at the TMC or location of the data provider; however, permanent storage of the on-demand archiving would likely be the responsibility of the advanced data user.

An on-demand archiving capability could offer data users the following choices in archiving ITS data:

- time(s) and duration of archiving;
- location(s) of interest, specified by system, corridor, link, or point location;
- facilities of interest (e.g., freeway mainlanes, ramps, arterial streets, HOV lanes, etc.)
- data items of interest (e.g., speed, volume, travel time, etc.); and
- desired aggregation level (presumably raw for most advanced data users).

For example, a transportation modeler may wish to have a full week of raw, disaggregate data to use in calibrating a simulation model. With the on-demand archiving capability, the modeler could use a web browser to access the main request page for on-demand archiving (see Figures 17 and 19 for analogous user interfaces). At this main page, the modeler may be required to type in a name and password since the on-demand archiving system may be only available to public and private sector partners that have helped to fund system development. Once in the on-demand archiving system, the modeler could specify the needs of the "special study" by selecting from a menu of available choices (see above list). The results of the on-demand archiving requests are stored temporarily until the modeler downloads the data to the agency's computer storage system.

Data Broadcast

The concept of a data broadcast is analogous to that of an FM radio signal: a raw data stream is "broadcast" to users through a receiver, which is capable of playing the data stream or archiving the data stream in its entirety. With an FM radio signal, the receiver is a radio and the data stream is archived by recording the radio signals to magnetic audio tape. With ITS data, the receiver could be a web browser or other desktop client software, and the data stream could be archived to any computer storage media. Using a data broadcast, a TMC or other data provider could provide (or "push") a raw data stream to subscribers, and the data subscribers could then archive or aggregate as necessary. The advantage of the data broadcast method moves the burden of data storage and management from the data provider to the actual data users; however, in our case, the data provider may also be an archived data user. The disadvantage of a data broadcast is that it is bandwidth-intensive and may require a high-speed communications infrastructure.

Data broadcasts can be performed in real-time if adequate communications infrastructure exists, or the data broadcast can be delayed, in the sense that data is buffered temporarily, then streamed during low bandwidth traffic hours. As an example of real-time broadcasts, the University of Washington has developed client software that can be used to obtain real-time Seattle area freeway sensor information over the Internet (47). The "loop_client" software enables users to receive the freeway sensor data every 20 seconds, and modifications to the software could be made that enable the data to be archived as it is received. This real-time data broadcast method is analagous to "subscribe-push" Internet technologies that are currently being developed. This real-time data broadcast is also referred to as an "IP multicast" (48), which is the technology used to deliver up-to-the-minute news and stock market quotes to computer desktops.

As an example of a delayed data broadcast, consider ITS data archiving efforts at the TransGuide® system in San Antonio. At the end of every day (after midnight), the TransGuide® system moves the most recent day of raw, disaggregate freeway sensor data to temporary storage on an FTP site (ftp://www.transguide.dot.state.tx.us/lanedata/). TransLink® researchers at TTI automatically retrieve that data over the Internet, then load aggregated, 5-minute data into the DataLink system (see page 51). In this case, the data broadcast from the TransGuide® system is buffered in temporary data storage for a full day. Data users are then able to retrieve or "pull" this data from the Internet site for a short period of time thereafter. Although not as sophisticated as real-time data broadcasts, this delayed method may be more suitable where high communications bandwidth is not available.

ITS DATA QUALITY CONTROL

As with any data collection or analysis effort, data quality should be an important consideration in designing ITS data archiving and/or analysis systems. Quality control procedures are especially critical with ITS data for several reasons: 1) the large potential volume of ITS data makes it difficult to detect errors using traditional manual techniques; 2) the continuous monitoring nature of ITS data implies that equipment errors and malfunctions are more likely during operation than periodic data collection efforts; and 3) archived data users may have different (and potentially more stringent) data quality requirements than real-time users.

It is important to note that archived data users represent an excellent source of feedback on ITS data quality. However, a quality control feedback loop or mechanism is necessary to get this information about data quality (or lack thereof) from the archived data users to the data providers (e.g., TMCs). In some cases, archived data users scrutinize ITS data in greater detail than most real-time users in an operations or management center. Because of this greater scrutiny afforded by off-line analysis, archived data users may discover errors or data quality problems not immediately obvious to real-time users or center operators.

The quality control efforts conducted in this study were focused on ITS traffic monitoring data (collected by inductance loop detectors) from San Antonio's TransGuide® system, and analyzed three basic attributes of quality control:

- suspicious or erroneous data;
- nature and extent of missing data; and
- accuracy and comparability of ITS data to similar data sources.

The analyses and findings for these quality control areas are presented in the following sections.

Erroneous or Suspicious Data

Error detection capabilities are a critical component of archived ITS data management systems. Even though many ITS deployments have traffic management software or field controllers with basic error detection (49,50,51), additional advanced error detection capabilities may be desirable for data archiving systems.

Most data screening techniques used to detect such errors at TMCs are based on comparing reported volume, occupancy, and speed values to minimum or maximum threshold values. These minimum or maximum thresholds are typically defined as the lower or upper limit of plausible values. These data screening techniques in place at many TMCs have been criticized as providing only a ". . . minimal examination of credibility" (51).

It appears that data quality control procedures for planning applications are more rigorous than those for data collected through operational centers. Data quality control screening procedures for planning applications are the focus of an ongoing pooled-fund research study, which is developing an expert system for detecting errors in vehicle volume, classification, and weigh-in-motion data (52). An existing software application developed for traffic volume data at automatic traffic recorder (ATR) stations utilizes historical traffic patterns to identify anomalous traffic volume data (53). The rules in this expert system are based on inductive logic and the collective experience of personnel at three state DOTs.

Most quality control screening techniques at TMCs or operational centers either flag or remove erroneous and/or suspicious values. Several of the data quality procedures developed specifically for planning applications provide guidance for replacing erroneous or suspicious data (a.k.a. imputation); however, this is not the recommended practice of the *AASHTO Guidelines for Traffic Data Programs.*

As an example, the TTI research team analyzed October 1998 traffic monitoring (i.e., loop detector) data from Phase One of the TransGuide® TMC in San Antonio for erroneous or suspicious (potentially erroneous) data. Because the TransGuide® system and the corresponding field equipment have minimal error checking procedures implemented, the scope of the analysis was focused on identifying the nature and extent of erroneous or suspicious data using a variety of quality control techniques.

Aside from basic data format validity checks performed by the Statistical Analysis Software (SAS) (e.g., are all data in the expected format? is text in a numeric data field?), the following statistics were reviewed for each traffic monitoring variable (e.g., speed, volume, occupancy) for each day of data:

- total number of data records/observations;
- minimum value;
- maximum value;
- 5^{th}, 10^{th}, 95^{th}, and 99^{th} percentile values; and
- normality and skewness statistics.

Additionally, the researchers examined all three variables at the 20-second record level for minimum and maximum values of speed, volume, and occupancy. This examination at the 20-second record level enabled the researchers to identify illogical or physically impossible combinations of speed, volume, and occupancy at the individual record level.

Table 16, which shows various combinations of speed, volume, and occupancy, was developed by examining and analyzing the types of data errors observed in the TransGuide® loop detector data. The table also shows examples of quality control flag values, which can be attached to individual data records as a way to alert data users of actual or suspected data errors.

Table 16. Quality Control Screening Scenarios and Corresponding Flag Values

Scenario Number	20-Second Record Value			Condition	Quality Control Flag Values		
	Speed (mph)	Volume (vehicles)	Occupancy (percent)		Speed	Volume	Occupancy
Single-Loop Detectors (speed not available (e.g., “-1"), typically only exit and entrance ramps)							
1	-1	0	0	**good data** - no vehicle(s) present	NA	NV	NV
2	-1	0	OCC>95	**good data** - vehicle stopped over loop	NA	GD	GD
3	-1	1<VOL<17	>0	**good data** - vehicle(s) present	NA	GD	GD
4	-1	0	1<OCC<95	suspect combination - cause unknown	NA	SD	SD
5	-1	>0	0	suspect combination - cause unknown	NA	SD	SD
6	-1	VOL>17	>0	suspect data - high volume count could be loop chatter	NA	SD	SD
Double-Loop Detectors (typically only freeway mainlanes)							
7	0	0	0	**good data** - no vehicle(s) present	NV	NV	NV
8	0	0	OCC>95	**good data** - vehicle stopped over loop	SV	SD	GD
9	>0	1<VOL<17	>0	**good data** - vehicle(s) present	GD	GD	GD
10	0	0	1<OCC<95	suspect combination - cause unknown	SD	SD	SD
11	0	>0	0	suspect combination - cause unknown	SD	SD	SD
12	0	>0	>0	suspect combination - single vehicle between loops in double-loop configuration	SD	SD	SD
13	>0	0	0	suspect combination - cause unknown	SD	SD	SD
14	>0	>0	0	suspect combination - cause unknown	SD	SD	SD
15	>0	0	>0	suspect combination - cause unknown	SD	SD	SD
16	>0	VOL>17	>0	suspect data - high volume count could be loop chatter	SD	SD	SD
17	no values reported			missing data - see next section	MD	MD	MD

Key to Quality Control Flag Values: GD = good data; MD = missing data; NA = not available or disabled function; NV = no vehicles present; SD = suspect data; SV = stopped vehicle.

The quality control rules shown in Table 16 were applied to October 1998 loop detector data from Phase One of TransGuide® to determine the extent of suspect data. The results of this analysis are shown in Tables 17 and 18. The analysis summarized in these tables indicate that the "suspect" data is a small problem overall, accounting for only one percent of all data records. On average, data records classified as "good" accounted for 76.5 percent of all data records for October 1998. Conversely, missing data accounted for 22.5 percent of all data records, indicating that more than one in five data records were missing throughout the month of October. The next section discusses the causes and extent of missing data in more detail.

Table 17. Summary of Suspect or Erroneous TransGuide® Data, October 1998

Data Classification	Data Records within a Classification	
	Average Values	**Median Values**
"Good" - scenarios 1-3 and 7-9 in Table 18	76.5 % (1,741,079 records per day)	82.9 % (1,886,919 records per day)
"Suspect" - scenarios 4-6 and 10-16 in Table 18	1.0 % (22,239 records per day)	1.0 % (23,170 records per day)
Missing - scenario 17 in Table 18	22.5 % (513,349 records per day)	16.1 % (365,567 records per day)

In summary, the following lessons were learned from the development of basic quality control screening rules and the analysis of a typical month of TransGuide® data:

- **Important to examine ranges of data values to identify suspect data -** The research team conducted basic validity checks of the TransGuide® data by examining the distribution and statistics for each data variable. This step was important in identifying specific data problems and developing appropriate data screening and flagging procedures for suspect data.

- **Consensus is to flag suspect data and not impute replacement values -** The consensus of many traffic data professionals is to have suspect ITS data flagged appropriately in data archives. The AASHTO traffic data guidelines specifically recommend against imputing replacement values for suspect data. Data archives may suggest alternative replacement values, but in most cases the data archives should retain the original suspect data, as data users may interpret the suspect nature differently depending upon their specific application. A good example of this is provided in the CDR data archival software developed at the Washington State Transportation Center (54). Suspect data that have been determined to be erroneous (e.g., from examination of equipment) need not be archived.

- **Missing data may account for significantly more data records than suspect data -** In the analysis of TransGuide® data, missing data accounted for 22.5

percent of all data records, whereas "suspect" data accounted for only one percent. The percent of suspect data could, however, increase if stricter quality control screening rules are used.

- **Development and testing of more advanced quality control procedures -** The quality control procedures developed in this study are considered basic, as they only examine extreme outlying or illogical values for various combinations of volume, occupancy, and speed. With these basic quality control procedures, each data record is evaluated in isolation of other data records. More advanced quality control procedures should examine each data record in relation to other data records in close proximity in both time and space. For example, traffic volumes from a loop detector can be compared to traffic volumes at upstream and downstream detectors to ensure that reported traffic volumes follow basic traffic flow continuity principles. Traffic patterns and profiles from adjacent lanes can also be compared to ensure consistency and reliability. More advanced quality control procedures could also compare each data record to historical patterns and trends for average values and typical ranges of variability at that location, similar to the expert rules developed as part of the ATR expert system software discussed earlier. Statistical process control methods, similar to those shown in Figure 29, could also be used to detect outlying data for additional scrutiny or examination.

```
                        Sigma
UCL     --------------- +3
        Zone A
        --------------- +2
        Zone B
        --------------- +1
        Zone C
Center ========          0
        Zone C
        ---------------  -1
        Zone B
        ---------------  -2
        Zone A
LCL     ---------------  -3

UCL = upper control limit
LCL = lower control limit
sigma = standard deviation
```

Eight Tests for "Special Causes"

1. One point beyond zone A.
2. Nine points in a row in Zone C or beyond (on one side of center line).
3. Six points in a row, all increasing or decreasing.
4. Fourteen points in a row, alternating up or down.
5. Two out of three points in a row in zone A or beyond.
6. Four out of five points in a row in zone B or beyond (on one side of center line).
7. Fifteen points in a row in zone C (above and below center line).
8. Eight points in a row beyond zone C (above and below center line).

Source: Minitab Statistical Software

Figure 29. Example of Statistical Process Control Guidelines

Table 18. Extent of Suspect or Erroneous TransGuide® Data for Individual Scenarios, October 1998

Scenario Number	20-Second Record Value			Condition	Classification of Data Records	
	Speed (mph)	Volume (vehicles)	Occupancy (percent)		Average Percent	Average Records per Day
Single-Loop Detectors (typically only exit and entrance ramps)						
1	-1	0	0	**good data** - no vehicle(s) present	9.51	216,433
2	-1	0	OCC>95	**good data** - vehicle stopped over loop	0.00	3
3	-1	1<VOL<17	>0	**good data** - vehicle(s) present	15.84	360,590
4	-1	0	1<OCC<95	suspect combination - cause unknown	0.06	1,444
5	-1	>0	0	suspect combination - cause unknown	0.06	1,284
6	-1	VOL>17	>0	suspect data - high volume count could be loop chatter	0.08	1,810
Double-Loop Detectors (typically only freeway mainlanes)						
7	0	0	0	**good data** - no vehicle(s) present	11.31	257,563
8	0	0	OCC>95	**good data** - vehicle stopped over loop	0.00	25
9	>0	1<VOL<17	>0	**good data** - vehicle(s) present	39.82	906,465
10	0	0	1<OCC<95	suspect combination - cause unknown	0.06	1,320
11	0	>0	0	suspect combination - cause unknown	0.08	1,726
12	0	>0	>0	suspect combination - single vehicle between loops in double-loop configuration	0.46	10,402
13	>0	0	0	suspect combination - cause unknown	0.01	164
14	>0	>0	0	suspect combination - cause unknown	0.05	1,180
15	>0	0	>0	suspect combination - cause unknown	0.00	15
16	>0	VOL>17	>0	suspect data - high volume count could be loop chatter	0.13	2,869
17	no values reported			missing data - see next report section	22.55	513,349

Nature and Extent of Missing Data

Missing data has been noted in several reports as a common attribute of ITS traffic monitoring data because of the continuous operation of the traffic monitoring equipment (4,13,55). The typical causes of missing data, as well as how the causes affect missing data, is shown in Table 19. The characteristics of missing data may vary considerably depending upon the type of traffic monitoring equipment, field controllers, and central traffic management systems.

Table 19. Typical Causes and Characteristics of Missing ITS Traffic Monitoring Data

Cause of Missing Data	Characteristics of Missing Data	
	Spatial Attributes	**Temporal Attributes**
Construction activity that disrupts the traffic monitoring installation	data missing at a single location or several consecutive locations along a corridor	data typically missing for extended periods of time (i.e., several months, but depends upon type of construction activity)
Failure of traffic monitoring equipment (could include the inductance loop hardware or the field controller software)	data missing at a single or several isolated locations	data typically missing for short or long periods of time (i.e., several minutes to several weeks)
Disruption of communications between field controllers and central traffic management system	data missing at a single or several isolated locations	data typically missing for short periods of time (i.e., less than several minutes)
Failure of central traffic management system or data archiving system (hardware or software) failure	data missing at all locations (or all locations on a given data server)	data typically missing for short periods of time (i.e., several hours to less than one day)

The nature and extent of missing ITS traffic monitoring data should be identified and reflected in the design of an ITS data archiving and/or analysis system. Missing data is nearly inevitable; therefore, knowing the characteristics of the missing data will help in identifying how best to handle the missing data in aggregation, summarization, or analysis algorithms. Regardless of the methods or algorithms used, it is important that data users be informed of missing data when summary or analysis results are presented. For example, when an average hourly speed is calculated and presented, an additional missing data statistic (e.g., percent complete value, Equation 8) should also be calculated. The percent complete value assists users in determining the reliability of average or summary statistics. For example, if the percent complete value is 50 percent, then a summary statistic has included data from only 50 percent of all possible time periods or locations.

$$Percent\ Complete\ Value = \frac{actual\ number\ of\ records/observations}{total\ possible\ number\ of\ records/observations} \quad (8)$$

Because the requirements of various data analyses and applications can vary significantly, a common practice is to simply flag missing data with no edited or replacement values. Data users with specific application needs can then edit or replace missing data values as appropriate to their individual analysis.

As an example, the researchers identified the nature and extent of missing traffic monitoring data from San Antonio's TransGuide® system for the month of October 1998. The calculations for identifying missing ITS data are quite simple (Table 20); however, the implications for ignoring the effects of missing data can be substantial. The results of this missing data identification are presented on the following pages.

Table 20. Identification of Missing Data for San Antonio's TransGuide®

Steps in Identifying Missing Data	Calculation for San Antonio TransGuide®
1. Determine the frequency of observations (e.g., polling cycle) at each location and lane. Use missing data score at this step to identify location-specific missing data problems.	Loop detectors polled every 20 seconds, producing **4,320 possible records** at each lane and location every day.
2. Determine the number of unique traffic monitoring locations for each computer server.	Two computer servers: Poll Server A and B Server A: 297 unique lane detectors Server B: 230 unique lane detectors
3. Determine the total possible number of records per computer server. Use missing data score to identify server-specific missing data problems.	Server A = 4,320 records × 297 detectors Server A = 1,283,040 records per day Server B = 4,320 records × 230 detectors Server B = 993,600 records per day
4. Determine the total possible number of records per day for the entire system. Use missing data score to identify overall missing data problems.	Entire System (Phase One) = Server A + B Entire System = 2,276,640 records per day Missing Data Records = Total Possible Records - Observed Records

Charts, graphs, or data tables can be used to identify the nature and extent of missing data once the basic calculations shown in Table 20 are performed. For example, Figure 30 shows the percent of non-missing data for several detectors and the overall computer poll server over the entire month of October 1998. By glancing at this or similar charts, the extent of missing data can be quickly discerned. Figure 30 indicates the following about the detectors and poll server shown:

- For most days, the percent complete values for the two lane detector locations are above 80 percent, whereas the percent complete values for Server B typically ranges between 70 and 90 percent for the month.

- There are several days in which significant missing data is detector-specific (i.e., Oct. 11, Oct. 28). The missing data for these days are most likely associated with specific locations, such as detector or controller failure.

- There are several days in which missing data is systematic across all detectors on a computer poll server (i.e., Oct. 9, Oct. 18). The missing data for these days are most likely associated with traffic management or data archiving software failure.

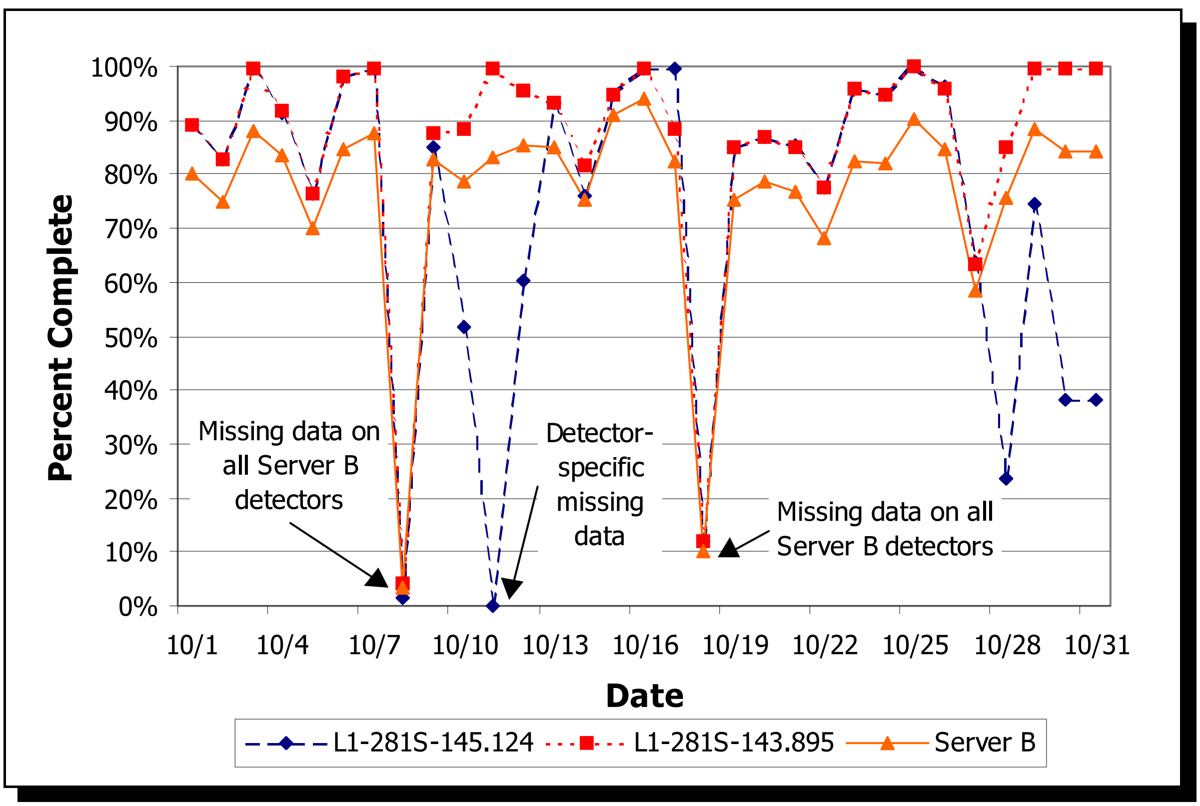


Figure 30. Identifying Detector-Specific Missing Data

Many traffic management systems have detector locations that number in the thousands, so it may be difficult to chart percent complete values for each specific lane detector location. Figure 31 shows another example of how to determine whether missing data is limited to a few locations over the full day, or whether the missing data is spread over all detectors for part of the day. For example, Figure 31 shows the following:

- For most days, about 90 percent of the detectors (approximately 270 of the 297 total) have at least 75 percent complete data. The percentage of detectors with at least 95 percent complete rarely exceeds 80 to 90 percent.

- There are several days in which most detectors have very little missing data (i.e., Oct. 7, 25, 29, 30, 31). For these days, there are most likely a small number of detectors that are missing data for the entire day.

- There are several days in which nearly all detectors have significant missing data (i.e., Oct. 8, 18, 19, 26). For these days, it is most likely that many detectors will be missing part of the day's data.

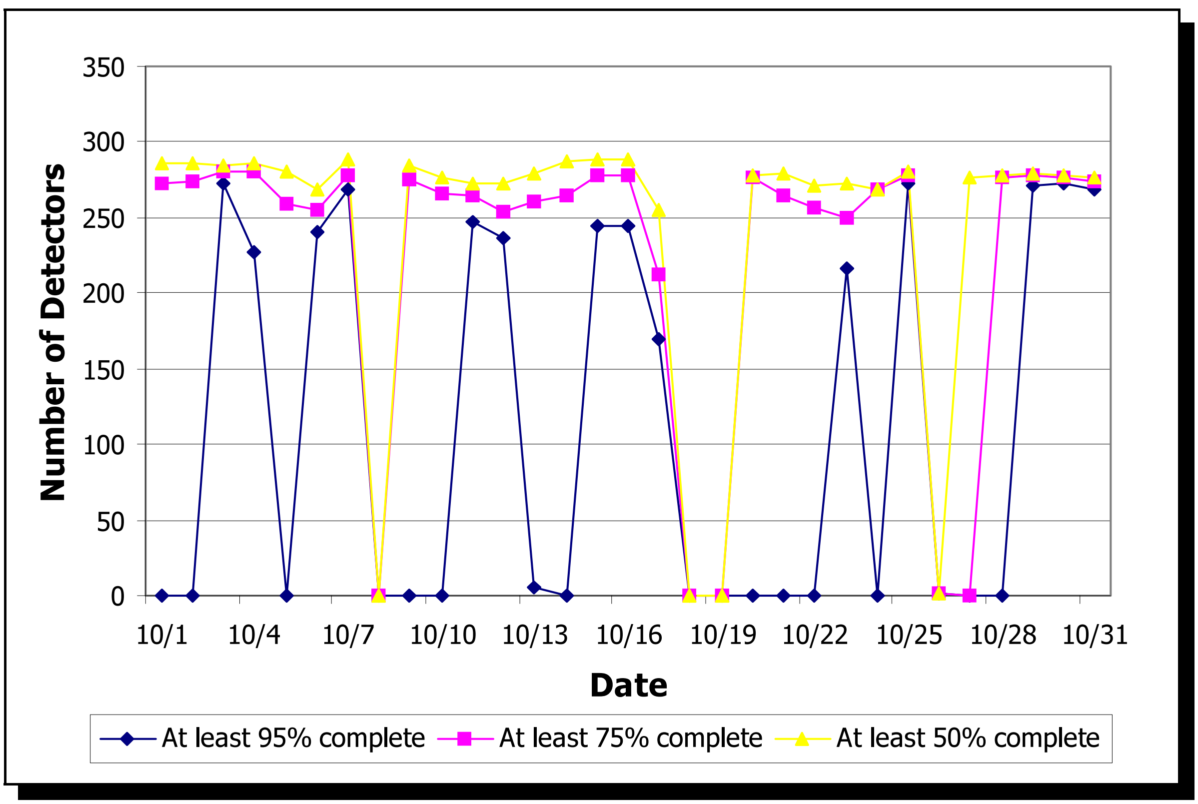


Figure 31. Identifying the Extent of Detector-Specific Missing Data

Figures 30 and 31 have enabled the identification of specific or systematic missing data at detector locations. These figures also enable data analysts or users to discern whether data is missing from a few detectors over the entire day, or from nearly all detectors over part of the day. Figure 32 provides a macroscopic view of missing data, in that it shows the percent complete values at the computer server level. For Phase One of TransGuide®, there are two computer servers (i.e., "A" and "B"). This figure shows the following:

- For most days, both computer servers have between 70 and 95 percent data. There are only three days during the month in which both computer servers drop below 70 percent complete data.

- The percent complete value drops to below 10 percent on two days: Oct. 8 and 18. For these days, there were significant data missing system-wide.

- In some cases, there may be computer server-specific problems with missing data. For example, Oct. 19 shows Server A with 32 percent complete data and Server B with 75 percent complete data.

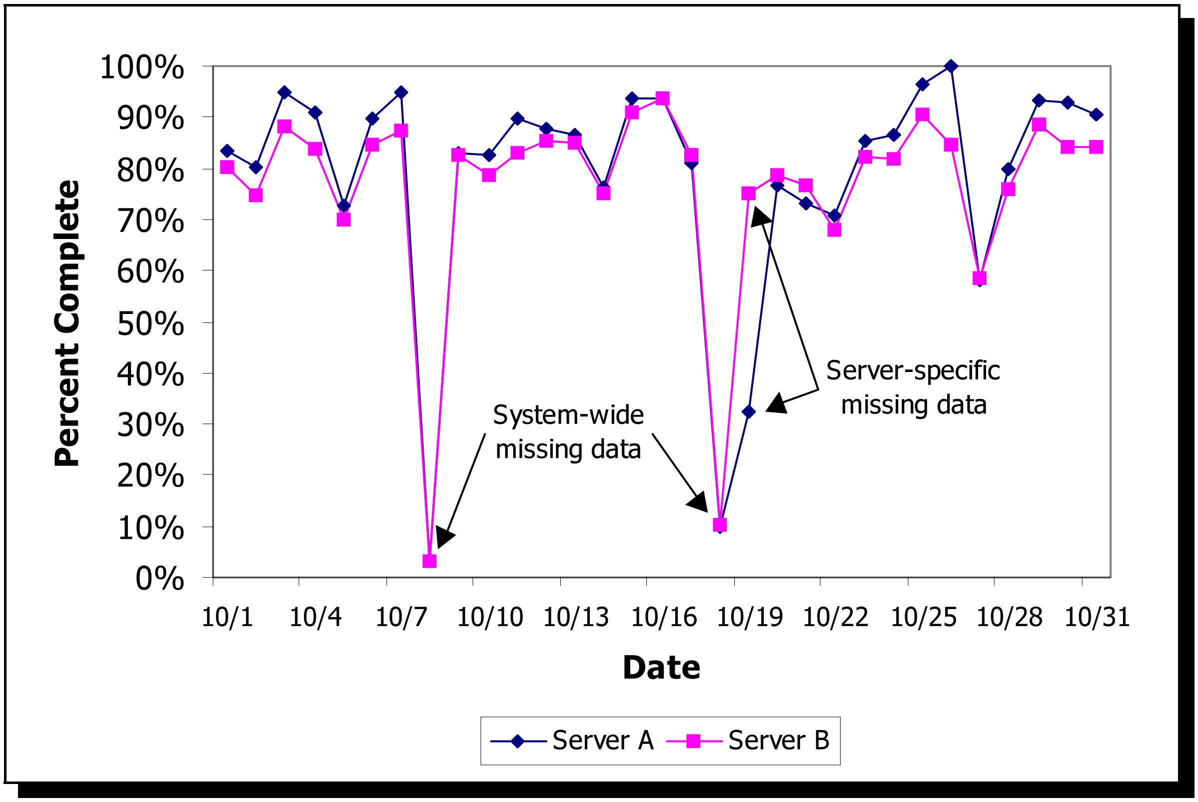


Figure 32. Identifying Server-specific and System-wide Missing Data

In summary, the following lessons were learned in analyzing the nature and extent of missing data from Phase One of San Antonio's TransGuide®:

- **Missing Data is Inevitable -** Given the continuous operation of most ITS traffic monitoring devices, missing data is almost inevitable. In San Antonio, this missing data typically ranged from 5 to 25 percent for most days. Because of excellent loop maintenance programs in San Antonio, only about 5 to 15 percent of this missing data can be accounted for by loop detector failure. The remaining 10 or more percent can most likely be explained by central traffic management system failures.

- **Identify the Nature and Extent of Missing Data for System Design -** Archived data management system designers should perform basic analyses to determine the nature and extent of missing data. By performing these simple analyses, system designers will be able to minimize the adverse effects of missing data on databases and analysis tools. The information from this exercise also serves as useful feedback to traffic management operators and maintenance personnel.

- **Flag Missing Data for Users -** At this time, the recommended practice for data archiving is to simply flag missing data as such, without providing edited or replacement values. Because analysis requirement for archived data can vary dramatically, it is prudent to allow users the flexibility in editing or replacing missing data values. The reason for missing data should also be noted. For example, missing data due to construction will affect traffic patterns, whereas missing data from equipment failure will not affect traffic patterns. The reason for missing data may also affect the interpretation of other suspect data.

- **Account for Missing Data in Statistics -** Summary or aggregated statistics should reflect the amount of missing data by reporting a missing data statistic (percent complete values). Cumulative statistics, such as vehicle-miles of travel (VMT), are most affected by missing data and will likely require data users and system designers to collaborate on how to account for the missing data.

Accuracy and Comparability of ITS Data

Accuracy is another attribute of data quality that is often a concern for archived ITS data users. In this section, accuracy refers to the traffic monitoring equipment's ability to truly reflect actual traffic conditions (e.g., reported vehicle counts closely approximate actual number of vehicles). Accuracy typically is a concern with archived ITS traffic monitoring data because the primary data collectors (e.g., TMCs) may have different accuracy requirements than the majority of archived data users. For example, TMCs may only require vehicle speeds to the nearest 5 or 10 mph for congestion or incident detection, whereas simulation model validation may require archived speed data to the nearest 1 mph for accurate calibration. Additionally, the primary data collectors' accuracy requirements may be for shorter periods of time (i.e., less than 15 minutes for real-time operations and management), whereas archived data users typically have accuracy requirements for much longer periods of time (i.e, one hour to a full year). A small bias or calibration error at the five-minute level can accumulate significant error in aggregated statistics, such as in average weekday traffic (AWDT) or average annual daily traffic (AADT).

In initial conversations with TxDOT's Transportation Planning and Programming (TP&P) Division about using San Antonio's TransGuide® data, statewide transportation planners named data quality as their most pressing concern. More specifically, the statewide planners were most interested in how TransGuide® data compared to data collected at nearby automated traffic recorder (ATR) stations. To answer these questions about data comparability, TTI compared a full month of TransGuide® traffic volume data to hourly traffic volume data collected at two ATR stations that were located along freeways instrumented in Phase One of TransGuide®. Aside from the comparability of TransGuide® traffic volumes to ATR traffic volumes, TTI researchers were concerned about the accuracy of both types of traffic monitoring equipment to "ground truth." To answer questions about how well these two data sources compared to "ground truth," TTI researchers used the TransGuide® surveillance cameras to record actual traffic conditions. The following sections describe these analyses.

Accuracy Study Design

Two TxDOT ATR stations (S184 and S185) are located on freeways instrumented by TransGuide® loop detectors (Figures 33 and 34), and were used in this analysis of accuracy and comparability. In most cases, the two ATR stations were separated from the upstream and downstream mainlane loops by an entrance or exit ramp. Since the entrance and exit ramps are also instrumented, simple equivalency equations were used to correlate both upstream and downstream mainlane detectors to the ATR stations (Table 21).

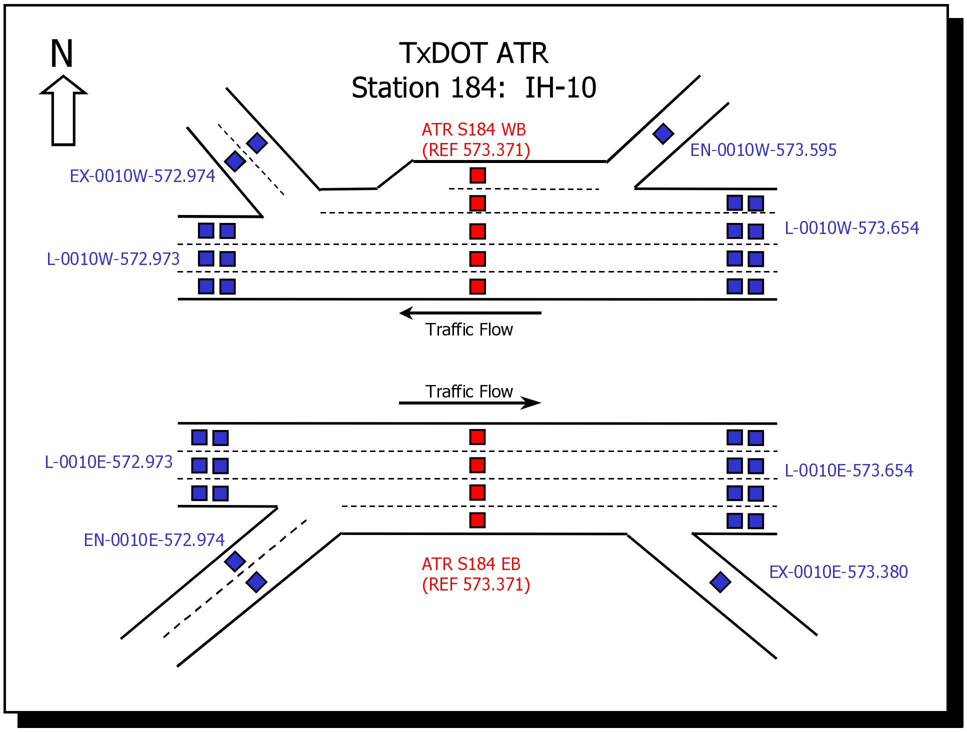


Figure 33. Location of ATR S184 and Nearby TransGuide® Loops

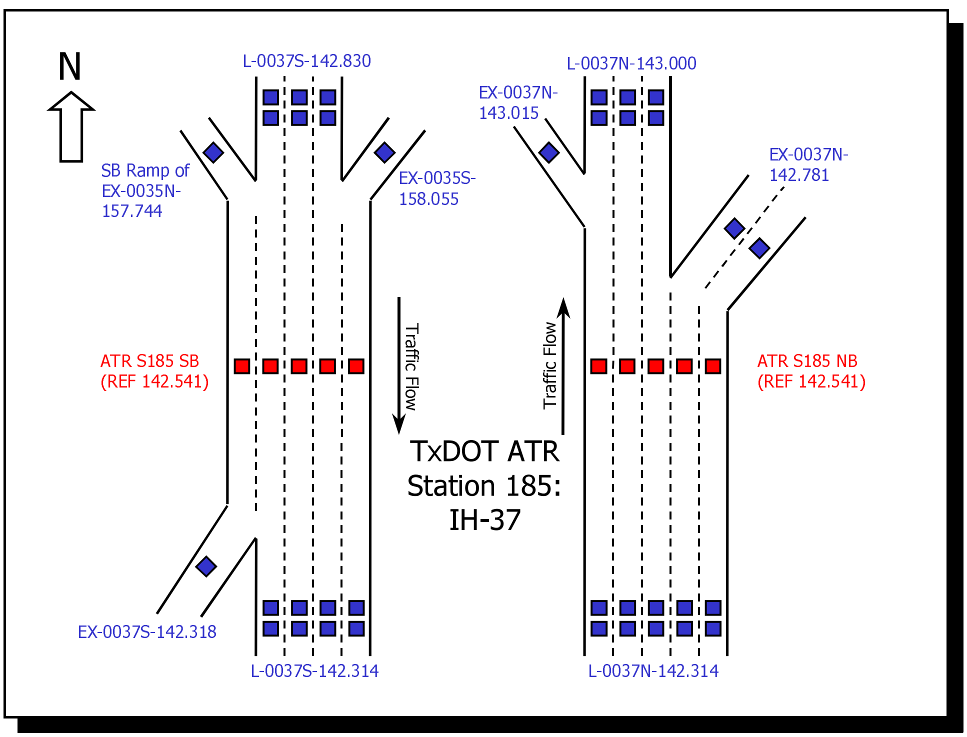


Figure 34. Location of ATR S185 and Nearby TransGuide® Loops

Table 21. Equations Used to Compare ATR Stations to TransGuide® Detectors

ATR Station	Equivalency Equations	
	Upstream TransGuide® Detectors	**Downstream TransGuide® Detectors**
S184, IH-10 Westbound	L-0010W-573.654 + EN-0010W-573.595	L-0010W-572.973 + EX-0010W-572.974
S184, IH-10 Eastbound	L-0010E-572.973 + EN-0010E-572.974	L-0010E-573.654 + EX-0010E-573.380
S185, IH-37 Southbound	L-0037S-142.830 + EX-0035S-158.055 + EX-0035N-157.744 (SB ramp only)	L-0037S-142.314 + EX-0037S-142.318
S185, IH-37 Northbound	L-0037N-142.314	L-0037N-143.000 + EX-0037N-142.781 + EX-0037N-143.015

The edited ATR hourly traffic volumes were provided by TP&P for the entire month of May 1998. TTI researchers obtained the May 1998 TransGuide® data from their FTP site (ftp://www.transguide.dot.state.tx.us/lanedata/) and archived it to CD for this and future analyses. The TransGuide® traffic volume data were aggregated from 20-second increments by lane to hourly traffic volumes by direction to match the hourly ATR traffic volumes.

Completely missing TransGuide® data prevented comparisons for several hours on several days. In other cases where only a small portion of TransGuide® data were missing (e.g., a single lane or part of an hour), the calculated traffic volumes were factored up by the percent complete values (Equation 9). Most of the TransGuide® hourly traffic volumes were factored up by less than 5 percent. An analysis of the missing data expansion factors indicated that the these factors did not significantly contribute to the variability of the traffic volumes.

$$
\begin{aligned}
\begin{array}{c}\textit{Adjusted TransGuide}\\ \textit{Hourly Traffic Volume}\\ (\textit{adjusted for missing data})\end{aligned} &= \frac{\begin{array}{c}\textit{calculated hourly traffic volume}\\ (\textit{with missing data})\end{array}}{\textit{percent complete value}} \\
&= \begin{array}{c}\textit{calculated hourly traffic volume}\\ (\textit{with missing data})\end{array} \times \frac{\textit{total possible records/observations}}{\textit{actual records/observations}}
\end{aligned} \tag{9}
$$

Accuracy Study Findings

Table 22 and Figures 35 through 42 show the hourly traffic volume comparisons between the upstream and downstream TransGuide® detectors and the two ATR stations. Figures 43 through 46 show comparisons between just the upstream and downstream TransGuide® detectors. This comparison was made to check the continuity in traffic volumes between TransGuide® detector locations. Table 23 shows the comparisons of the ATR station and TransGuide® detector volumes ground truth volume data, which was collected using TransGuide® surveillance cameras and manual "double-blind" counting from video with ±2 percent accepted tolerance between manual counts.

The analysis of TransGuide®, ATR, and ground truth traffic volume data yielded the following findings:

Large Range in ATR to TransGuide® Comparisons (Table 22) - At the IH-10 location, the ATR and TransGuide® traffic volumes compared favorably, typically falling with 10 percent of one another. However, the IH-37 location showed large differences in traffic volumes, ranging from 18 to 39 percent in directly comparing ATR and TransGuide® detectors. The authors hypothesize that discrepancies in traffic volumes at this location may be due to its location in a major freeway weaving section with left-hand exit ramps. In nearly all cases, the TransGuide® detectors consistently counted less than the ATRs. The differences between ATR and TransGuide® detectors were consistent, with R^2 values typically above 95 percent.

Mixed Results in Comparisons to Ground Truth (Table 23) - At the IH-10 location, both the ATR and TransGuide® traffic volumes were within 5 percent on average of ground truth volumes, within no clear difference between ATR and TransGuide® detectors. At the IH-37 location, however, the ATR detectors were slightly more accurate (within 4 and 11 percent on average) than the TransGuide® detectors (within 13 to 38 percent on average). As with the comparisons in Table 22, the IH-37 location again showed larger differences in traffic volumes than was seen on IH-10.

Table 22. Results of Hourly Traffic Volume Comparisons Between ATR and TransGuide® Detectors, May 1998

Detector 1	Detector 2	Average Percent Difference (%)	R^2	Comments
ATR S184 EB	TG I-10 EB Upstream	4.36	0.9966	See Figures 35 to 38.
	TG I-10 EB Downstream	19.91	0.9938	
ATR S184 WB	TG I-10 WB Upstream	9.45	0.9964	
	TG I-10 WB Downstream	2.39	0.9971	
ATR S185 SB	TG I-37 SB Upstream	20.9	0.8637	See Figures 39 to 42. ATR S185 (both directions) was located in a major weaving section.
	TG I-37 SB Downstream	32.76	0.9424	
ATR S185 NB	TG I-37 NB Upstream	39.39	0.9508	
	TG I-37 NB Downstream	17.85	0.9756	
TG I-10 EB Upstream	TG I-10 EB Downstream	18.6	0.9954	See Figures 43 and 44.
TG I-10 WB Upstream	TG I-10 WB Downstream	8.57	0.9970	
TG I-37 SB Upstream	TG I-37 SB Downstream	51.73	0.9028	See Figures 45 and 46.
TG I-37 NB Upstream	TG I-37 NB Downstream	31.91	0.9546	

Note: See Table 23 for comparison to ground truth.
ATRs are inductance loop detectors operated and maintained by TxDOT's Transportation Planning and Programming (TP&P) Division.
TG represents inductance loop detectors operated and maintained by TransGuide® in TxDOT's San Antonio District.

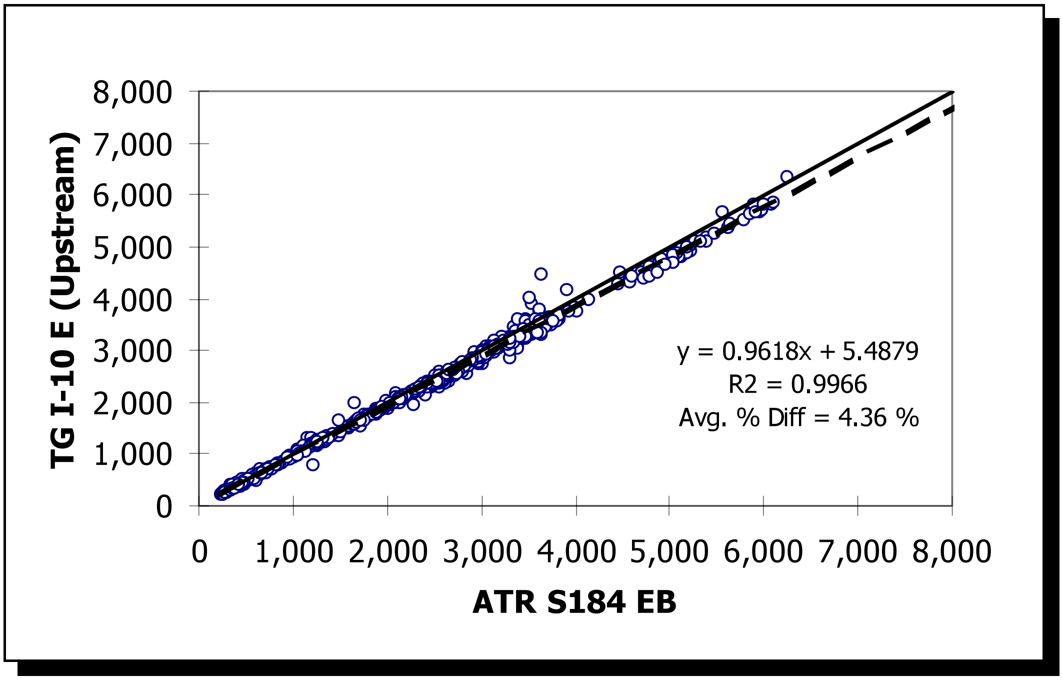


Figure 35. Comparison of Hourly Volumes from ATR S184 EB and Equivalent Upstream TransGuide® Detectors

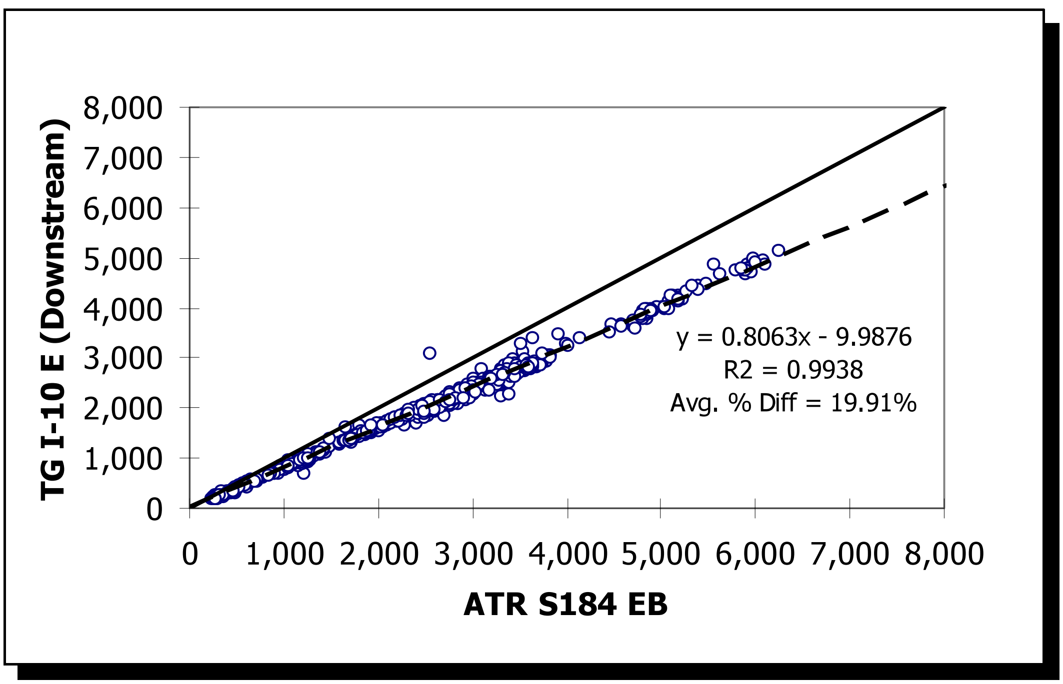


Figure 36. Comparison of Hourly Volumes from ATR S184 EB and Equivalent Downstream TransGuide® Detectors

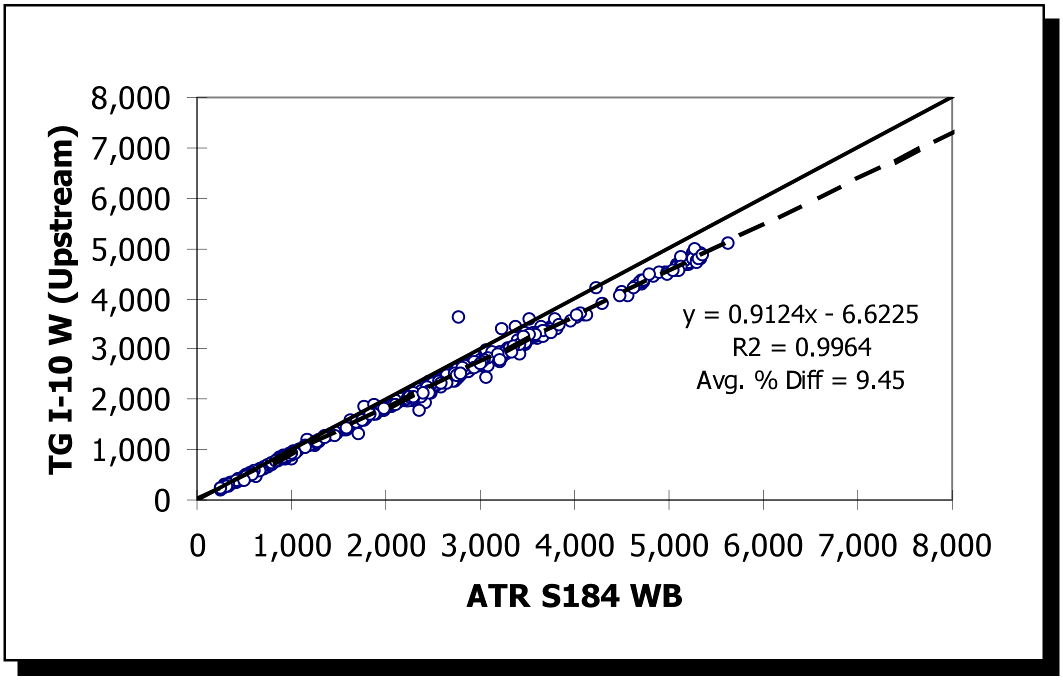


Figure 37. Comparison of Hourly Volumes from ATR S184 WB and Equivalent Upstream TransGuide® Detectors

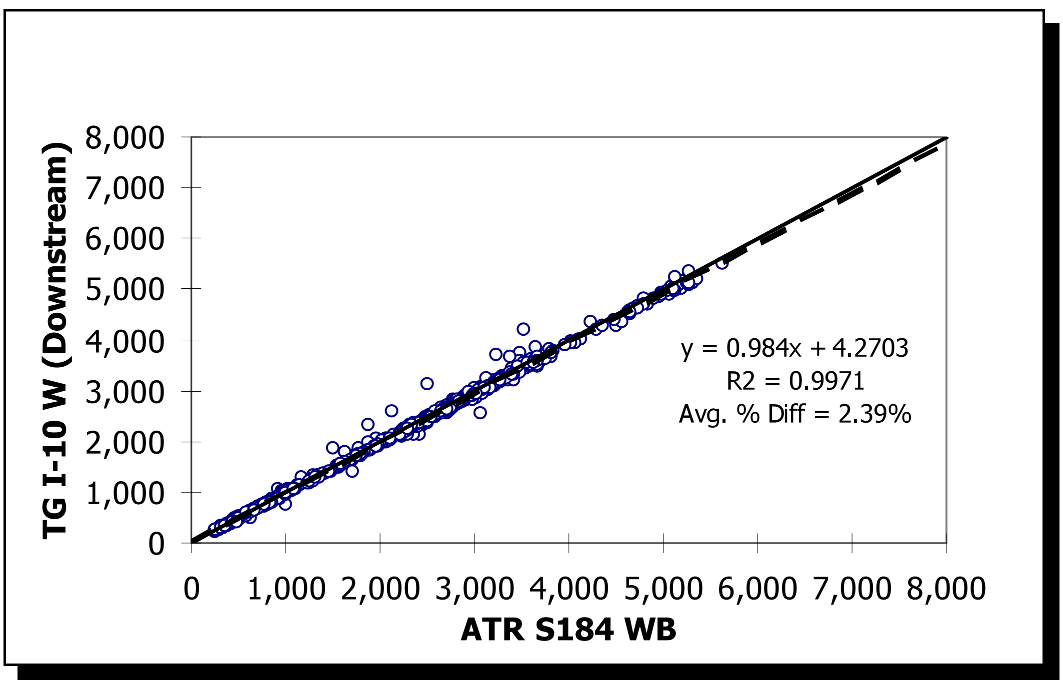


Figure 38. Comparison of Hourly Volumes from ATR S184 WB and Equivalent Downstream TransGuide® Detectors

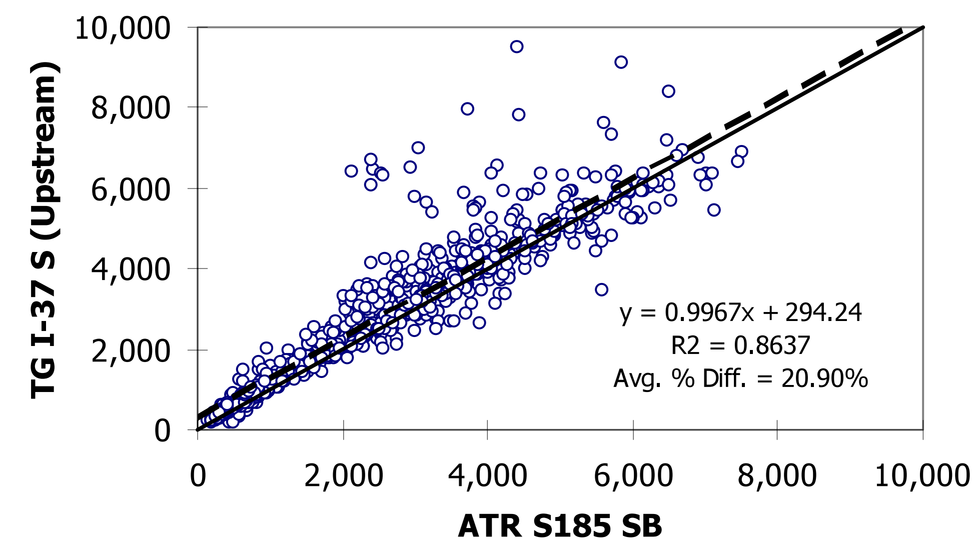


Figure 39. Comparison of Hourly Volumes from ATR S185 SB and Equivalent Upstream TransGuide® Detectors

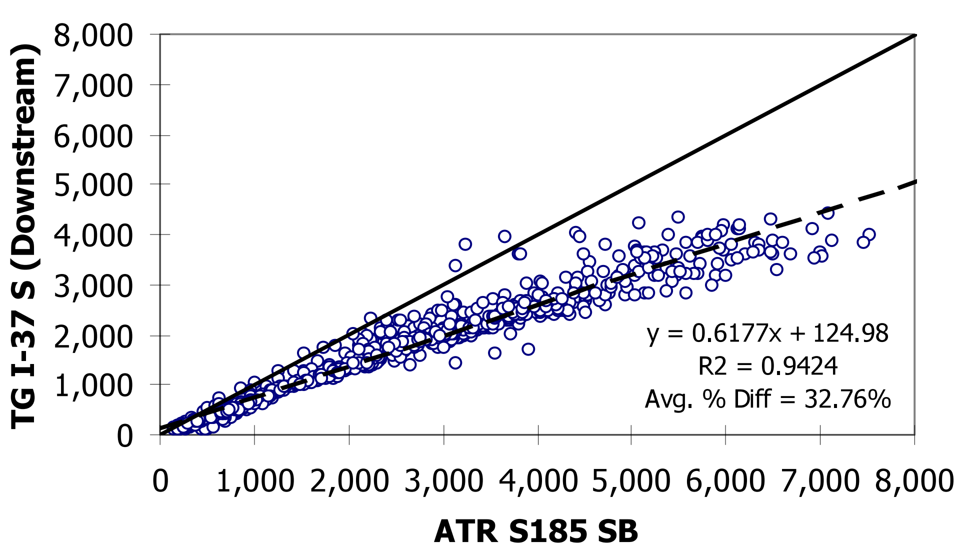


Figure 40. Comparison of Hourly Volumes from ATR S185 SB and Equivalent Downstream TransGuide® Detectors

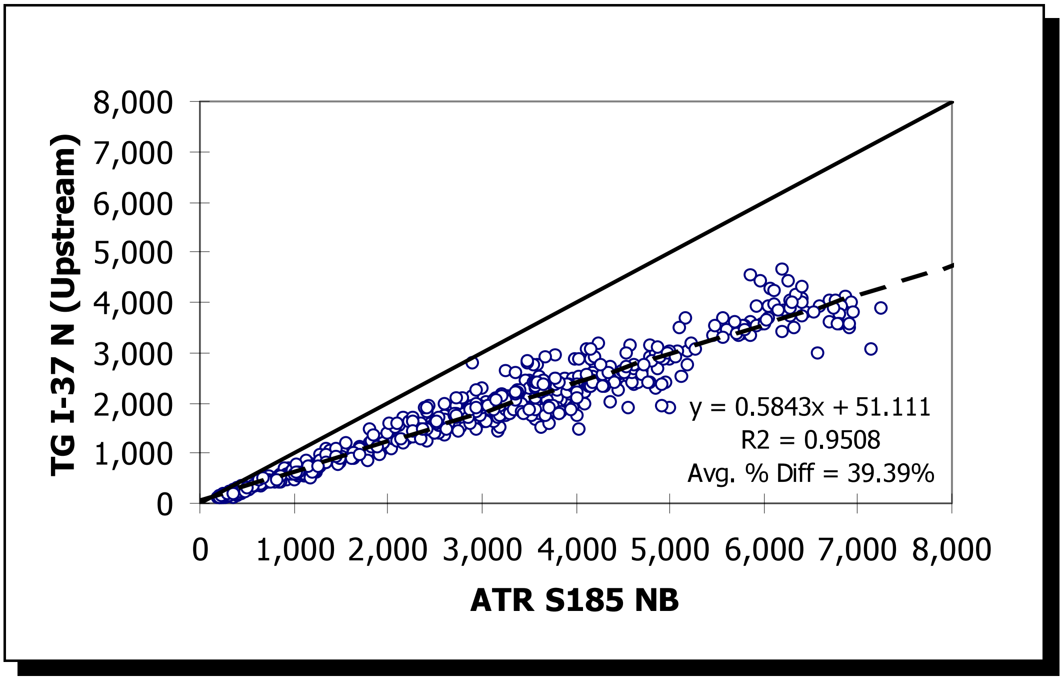


Figure 41. Comparison of Hourly Volumes from ATR S185 NB and Equivalent Upstream TransGuide® Detectors

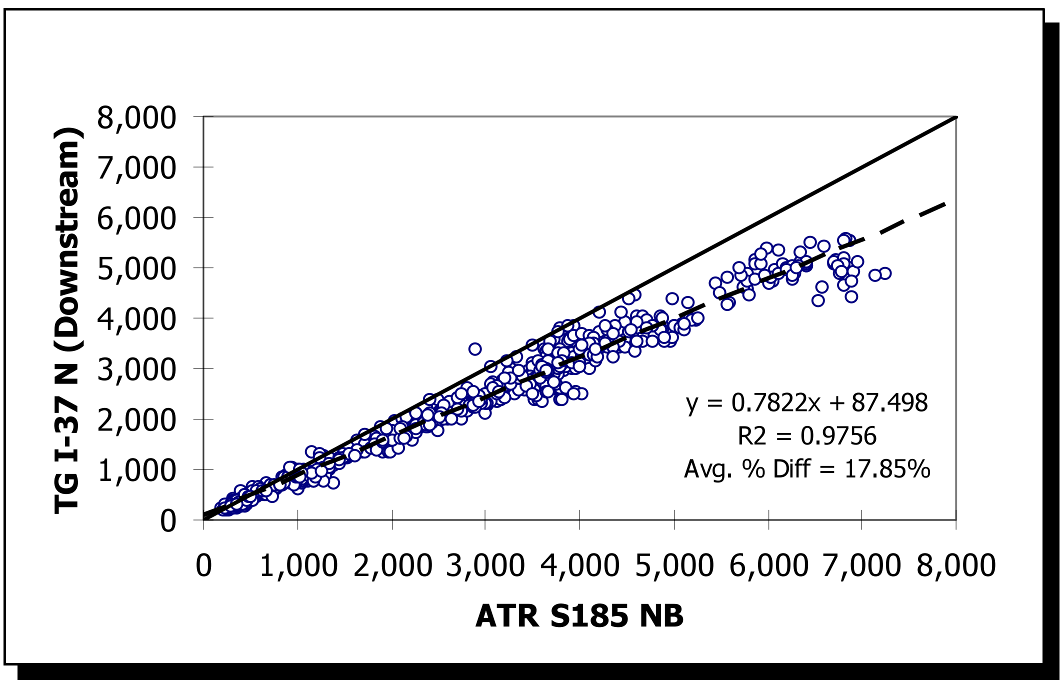


Figure 42. Comparison of Hourly Volumes from ATR S185 NB and Equivalent Downstream TransGuide® Detectors

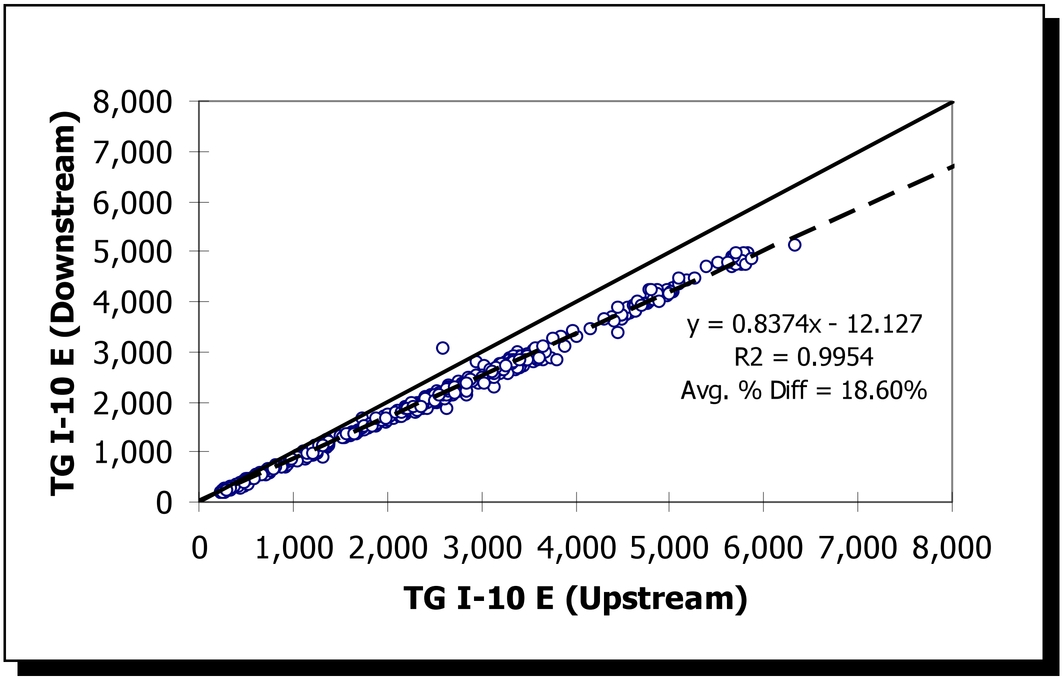


Figure 43. Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-10 East

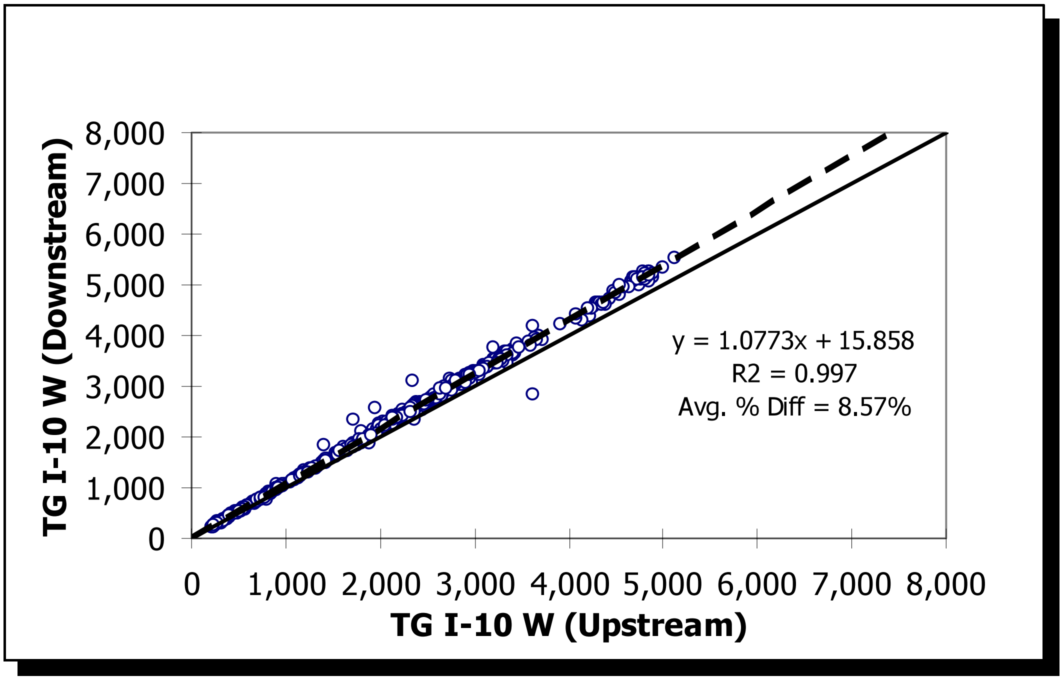


Figure 44. Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-10 West

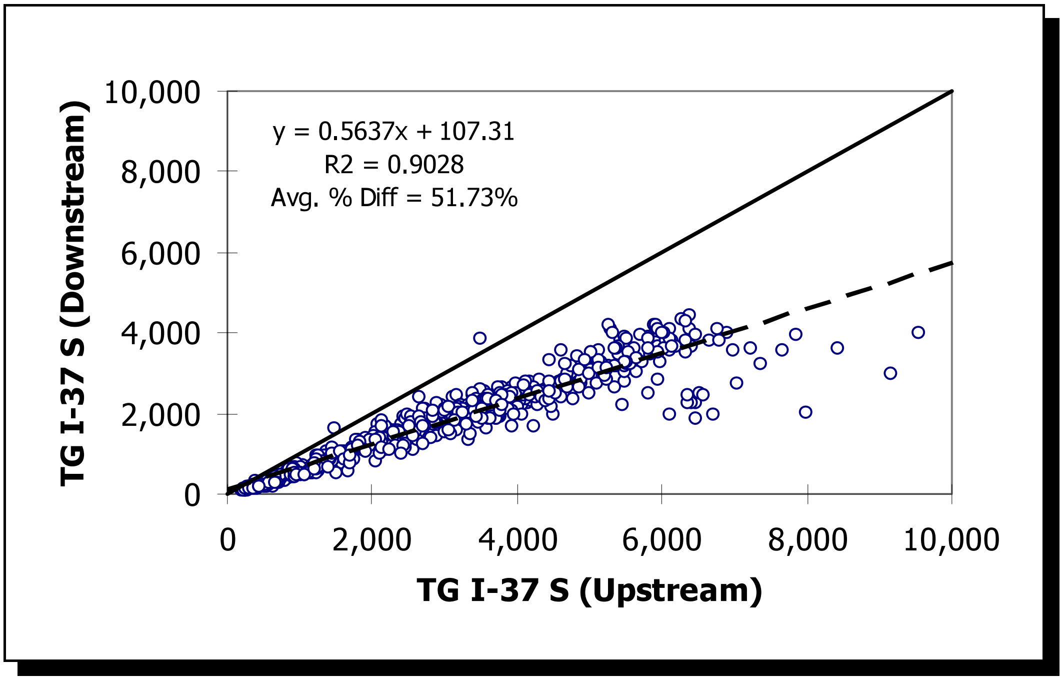


Figure 45. Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-37 South

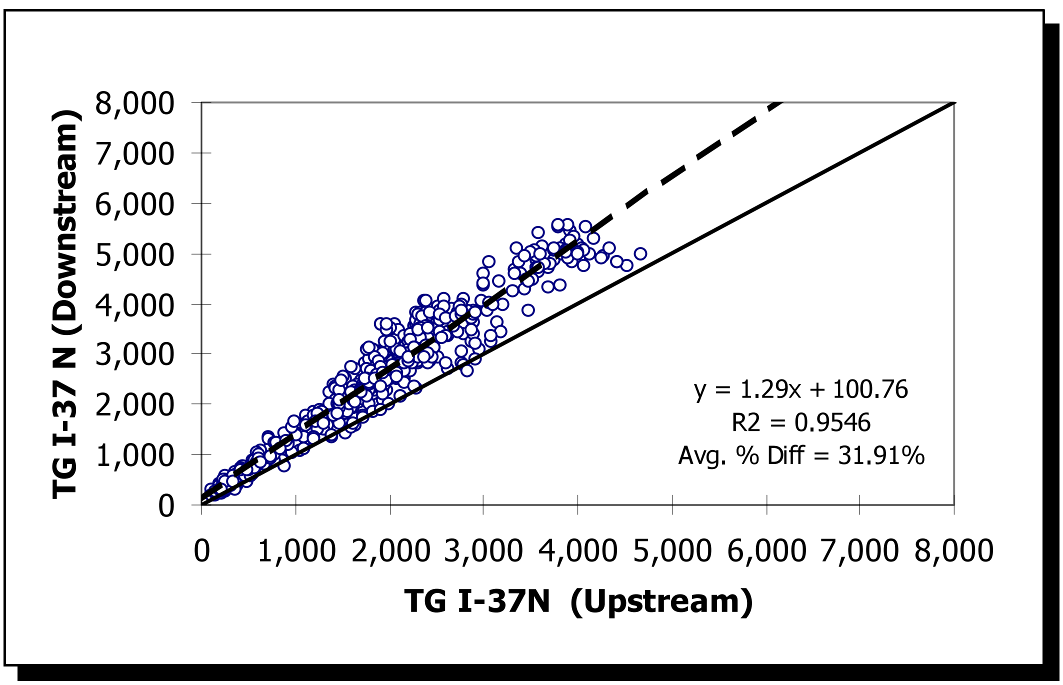


Figure 46. Comparison of Hourly Volumes from Upstream and Equivalent Downstream TransGuide® Detectors on I-37 North

Table 23. Results of Ground-Truth Traffic Volume Comparisons, December 1998

Date	Time	Percent Difference (%) from Ground Truth Traffic Volumes					
I-10		**EB TG Upstream**	**EB TG Downstream**	**EB ATR 184**	**WB TG Upstream**	**WB TG Downstream**	**WB ATR 184**
12/17/98	9 to 10 am	-6.05	no data	0.39	-10.33	-1.78	2.49
	10 to 11 am	-5.67	no data	0.35	-9.46	-1.74	1.65
	11 am to 12 pm	-6.27	no data	1.33	7.05	2.02	3.31
	12 to 1 pm	-3.11	no data	1.56	11.96	3.67	1.83
	3 to 4 pm	0.74	no data	4.40	-4.52	2.94	3.28
	4 to 5 pm	0.62	no data	3.32	-5.04	2.63	4.37
12/21/98	10 to 11 am	-3.38	no data	1.98	no data	0.85	2.25
	11 am to 12 pm	-2.94	no data	1.07	no data	1.91	0.40
	12 to 1 pm	-3.04	no data	1.73	no data	-3.66	-1.62
	1 to 2 pm	-2.72	no data	1.09	no data	-2.29	0.32
	2 to 3 pm	-3.25	no data	1.10	no data	-3.24	-0.45
	3 to 4 pm	-1.35	no data	1.88	no data	1.98	3.24
Average, I-10		**-3.03**	**-**	**1.68**	**-1.72**	**0.27**	**1.76**
I-37		**SB TG Upstream**	**SB TG Downstream**	**SB ATR 185**	**NB TG Upstream**	**NB TG Downstream**	**NB ATR 185**
12/18/98	10 to 11 am	24.90	-16.17	14.47	-42.11	-15.28	0.25
	12 to 1 pm	21.85	-15.89	12.01	-40.67	-16.43	1.75
	1 to 2 pm	18.41	-15.68	9.37	-38.34	-21.56	1.72
	4 to 5 pm	9.18	-13.07	14.00	-36.12	-17.53	1.55
	5 to 6 pm	7.33	-14.98	11.07	no data	no data	no data
12/22/98	11 am to 12 pm	14.61	-0.43	9.30	-41.83	-25.94	3.08
	12 to 1 pm	13.35	-0.02	17.51	-38.31	-16.92	2.30
	1 to 2 pm	19.32	3.48	10.84	-35.94	-21.10	2.44
12/28/98	4 to 5 pm	20.25	-21.23	14.96	-36.58	-7.01	10.59
	5 to 6 pm	14.27	-24.12	13.22	-35.18	-4.54	9.70
	6 to 7 pm	27.12	-14.11	13.16	-32.44	2.45	7.28
12/30/98	11 am to 12 pm	29.54	-16.99	3.73	-41.51	-6.47	2.31
	12 to 1 pm	33.19	-20.26	1.81	-38.66	-1.59	-0.14
	2 to 3 pm	27.81	-17.42	2.84	-39.45	-1.58	4.16
Average, I-37		**20.08**	**-13.35**	**10.59**	**-38.24**	**-12.18**	**3.64**

DATA STORAGE TOOLS AND ISSUES

The following sections provide basic information about computer data storage devices and privacy issues. The information about data storage devices in this section is oriented toward those individuals with a basic understanding of computer storage technology. As such, this data storage information is intended for generating discussion among transportation engineers and planners about archived data management system plans and designs. Growing advances in the computer industry will make some of this data storage information obsolete within several years of publication; however, some of the principles related to privacy will be valid independent of changes in technology. Because of the rapidly evolving nature of computer and data storage technology, the research team strongly recommends that information systems personnel be involved in archived data management system design efforts as early as possible.

Archived ITS Data Storage Tools

Once an agency or group of agencies decide to archive or store ITS data, some very fundamental questions need to be answered, such as:

- **User Requirements -** Who will use the data, what type of data and at what level of detail is required for the desired applications?
- **User Interfaces -** How will the data be accessed and how often will the data be accessed?
- **Interfaces to other Systems -** How will the archived data relate to other agency information systems?
- **Data Storage and Retention Policies -** How much data will be stored and for how long?
- **Privacy Issues -** Does the archived data potentially contain unique identifiers about an individual person or vehicle, and what steps are necessary to anonymize these unique identifiers?
- **Security Issues -** What type of security is required for the host network and the data?

The answers to these questions will help in developing computer system and data storage designs for archived data management systems. This section provides an overview of the various data storage devices and media. Table 24 summarizes the advantages and disadvantages of different types of data storage devices.

Table 24. Comparison of Computer Storage Media and Devices

Storage Media or Device	Typical Storage Range	Typical Storage Cost per GB (1999 $)	Typical Applications	Advantages (+) and Disadvantages (-) for Archived ITS Data
Portable Storage Media				
Magnetic Tape Cartridges (e.g., 4-mm,8-mm, 1/2-in. data cartridge)	up to 70 GB	$2 to $5/GB	permanent data archival or storage for PC desktop or enterprise systems	+ low storage cost per GB - sequential (slow) data access - low to moderate reliability over time - unstructured data storage (mostly "flat" files)
Magnetic Disk (e.g., Zip®, Jaz®, Super Disk®)	100 MB to 4.0 GB	$75 to $150/GB	PC desktop storage	- moderate storage cost per GB for portable media - proprietary formats
Magneto Optical Disk	up to 5.2 GB/disk	$25 to $75/GB	PC desktop or enterprise storage	- proprietary formats
CD-R or CD-RW	650 MB/disk	$3/GB for CD-R $23/GB for CD-RW	data and application distribution, PC desktop storage	+ low storage cost per GB + industry standard, excellent reliability - requires "jukebox" for mass storage - will be replaced by DVD in future
DVD	2.6 to 17 GB/disk	$6 to $8/GB for DVD-RAM	currently consumer-driven applications (home theater, etc.)	+ low storage cost per GB + becoming industry standard, will replace CD - emerging technology, tool sets and applications not mature
Semi-Permanent Storage and Management				
Desktop PC or Workstation Data Storage	up to 18 GB/drive (125+ GB total)	$20 to $150/GB	individual database to workgroup data storage	+ structured data storage and analysis - may require database administrator(s)
Dedicated Data Server (includes concepts such as data mart or data warehouse)	up to 1,000+ GB	can vary significantly depending upon size and structure	enterprise data storage and management	+ structured data storage and analysis - requires database administrator(s)

Magnetic Tape Cartridges

Magnetic tape cartridges come in many forms (e.g., 4-mm, 8-mm, 1/2-in data cartridge) and are primarily used for computer system backups. This type of storage is slow due to the sequential nature of access (i.e., if you need something at the end of the tape, you must fast forward through the beginning of the tape). Magnetic tape storage is typically used for long term storage of raw data in a flat file format and would not be appropriate for data analysis unless the data was downloaded to a hard drive to be processed. The advantage to this type of storage is that it is very inexpensive (i.e., less than $200 for the tape drive and tapes cost about $2 to $5 per gigabyte). However, the reliability of this storage media is also questionable, as tapes can break, stretch, or experience other failures. This type of storage would be useful if the data would only be rarely used or needed.

Magnetic and Magneto Optical Disks

Zip®, Jazz®, and Super Disk® are brand names of high-speed, high-capacity data storage devices with read-write capabilities. These can be internal or external similar to a computer hard drive, but the magnetic medium is removable. Extra cartridges with varying capacity can be purchased separately. These cartridges are random access but are magnetic in nature and similar to a very high capacity floppy disk. The advantage is that when the drive fills up you just buy another cartridge. The disadvantages are the expense and the ability to store only files or databases limited to the size of each individual storage disk.

Compact Disk (CD)

CDs have a moderate storage capacity of 650 megabytes but are very durable with an estimated shelf life of 100 years. Quick random access is provided, which is the ability to access the location on the CD where the data is stored without going through the data that comes before the required data on the storage device. The prices range considerably, from CD-ROMs which are relatively cheap and very common, less than $100 for a 32X player and about $300 to $500 for a CD-R or CD-RW with 8X read and 2X write capabilities. The writeable disks are typically $1 to $3 for each disk. CDs can be combined with jukeboxes (multiple CD readers) that allow a large amount of data to be stored and retrieved quickly and cheaply. This type of storage device can be used to store flat files that could be read by user-created programs to get the data from a single CD or a CD jukebox. The disadvantage is that the data must be in 650 MB groups, and if large amounts of data are required, then a CD jukebox and custom applications would be required to develop any new queries of the data. CD technology has all but stopped since the development of the DVD. However, CDs still have a very valuable role in the portable storage and distribution of data.

Digital Video/Versatile Disks (DVD)

DVD are similar to CD but have much greater storage capacities. These disks are more expensive but each disk can hold from 4 to 17 gigabytes which is about 7 to 26 times the storage capacity of CDs. DVD drives are also capable of reading a CD. For these reasons, DVDs are steadily replacing the CD for computer applications because of the vast storage capacity. These types of disks also work with jukebox systems but are more expensive. Like the CD storage media, DVD lends itself to flat file storage which would require programs to be developed to analyze the data.

Desktop PC or Workstation

Standard hard drives on desktop personal computers (PCs) or workstations are very affordable when used with a storage device such as CD or DVD for permanent backup and storage. Hard drives come in a variety of sizes and can be chained together or used as a network system. This device can store both flat files and databases. The cost is about $100/GB for a SCSI (Small Computer System Interface) which are more expensive but faster than the IDE (Integrated Drive Electronics) or EIDE (Enhanced Integrated Drive Electronics). Another advantage of a hard drive is that flat files or a data base can be used which allow easy access to the end user with more security features.

Summary

Generally speaking, CDs provide the most inexpensive long-term storage and can be quite useful if only pre-programed types of information are required. The portable data storage industry is moving toward DVD, which makes this type of storage device more attractive due to its large storage capacity and high speed access. Hard drives would be suitable for small to moderate size data storage. Redundancy and durability are a concern with such a large investment of time, data and money. Dedicated data storage servers with a redundant array of independent disks (RAID) is a good starting point for most agencies. A dedicated server combined with a relational database provides moderate storage capacity that is scalable and flexible at a moderate cost.

The range of combinations that can be used together is almost unlimited. Users will need to determine the value of accessing the data while weighing the cost of security. The user-based criteria must be defined up front for any sort of data storage system to work as intended. Once the user requirements, data accessability, and quantity of data are defined companies can be contacted and prices compared to find a means for storing the information. Based on the application of the warehouse, these companies can point the customer towards other pieces of the data warehouse puzzle. Before anything is purchased, the customer should discuss with each vendor the compatibility of each part of the system with other parts that are being purchased or may be purchased in the future.

Privacy Issues

Privacy concerns are an issue with archived ITS data that can be used to distinguish individual vehicles or persons. For example, automatic vehicle identification (AVI) transponders used for traffic speed monitoring or commercial vehicle applications report unique identification numbers that potentially could be used to track an individual vehicle or person's movements. With real-time data, privacy was not as large as issue because the real-time data was typically discarded soon after its use. In the development of ADUS, the consensus has been that the collecting agency has a responsibility to keep the vehicle or individual information secure. If individual vehicle or person information is to be archived, the consensus is that the collecting agency has the responsibility to "anonymize" the vehicle or person records so that the original unique identification number can not be determined or deciphered.

In Texas, these "anonymous" approaches are being used in the archival of individual travel time records from AVI transponders. In Houston, TTI has worked with TranStar in archiving data from their extensive freeway-based AVI traffic monitoring system. The unique tag identification numbers are scrambled such that these "anonymized" numbers can not be traced back to any transponder registration information at the Harris County Toll Road Authority. Similarly in San Antonio, Southwest Research Institute worked with TxDOT in developing a transponder scrambling algorithm that changes on a regular basis (e.g., every day or week). In this way, the same "anonymous" transponder number can be tracked throughout a day or a week, but the "anonymous" number can not be traced back to any transponder or vehicle registration records. The anonymized transponder data is then made available to data users through an FTP site (ftp://www.transguide.dot.state.tx.us/avidata).

CHAPTER 4.
FINDINGS AND CONCLUSIONS

CHAPTER OVERVIEW

State-of-the-Practice in Retaining and Using Archived ITS Data	Summarizes findings and conclusions related to existing ITS data archiving practices and uses of archived ITS data.
Examination of ITS Data Archiving Issues	Summarizes findings and conclusions from case study analyses of San Antonio TransGuide® data.

STATE-OF-THE-PRACTICE IN RETAINING AND USING ARCHIVED ITS DATA

Most Centers Have ITS Data Archiving on Their "Radar" - The research team found that most operations and management centers contacted were at least considering ITS data archiving, if not actively developing plans for archived data systems. Because of this widespread interest in ITS data archiving, ADUS deployment guidance will be imperative in the further development of integrated data systems. A common assertion is that many centers have not been archiving ITS data; however, the research team found in 1997 that 12 of 15 centers were archiving data, but that the data were either not readily accessible or usable (stored on proprietary magnetic tape cartridges) or not widely distributed within the metropolitan area. It is important to note here that ADUS encompasses more than just simple data storage, but also data extraction, manipulation, and analysis tools.

Access to Archived ITS Data Creates Opportunities - The research team found the most widespread use of archived ITS data was in locations where the data were easily or publicly accessible (e.g., Seattle, San Antonio). In these locations, user groups outside of the operations center were able to develop data extraction and analysis tools because the archived ITS data were easily accessible. The logical conclusion is that in areas where the operations center may not be able to develop systems that support ADUS, these centers should at least provide easy access to the data so that other user groups can develop systems that support ADUS functions. Examples of easy access to archived ITS data include distribution via the Internet or CDs.

Planners and Researchers Have Been Primary Archived Data Users to Date - The research team found that the two most common user groups have been researchers and planners, with applications ranging from basic traffic statistics to advanced model and algorithm development. In the few locations where traffic management operators were using archived ITS data, they were realizing benefits in improving traffic operations and management, as well as an ability to quantify these benefits. Private sector users that add value to archived ITS data are emerging users, although the business models are not yet clearly defined.

Concerns about Data Quality and Location Referencing - The most common concerns voiced from the archived ITS data users in this study were data quality and location referencing. In some cases, adequate quality control may not be performed by the operations centers, or many of the archived ITS data users may simply not be familiar with quality control procedures already in place. Location referencing was a concern because ITS detector locations are either not referenced to a defined location or they use a different location referencing scheme than roadway inventory or other traffic databases.

No Clear Consensus on Data Aggregation Levels - There does not appear to be a least common denominator of data aggregation that is significantly favored, other than simply saving raw, disaggregate data. Many of the applications provided throughout this report cover many stakeholders including planning, operations, and research, and they use many different levels of aggregation. In fact, several researchers expressed their interest in performing similar analyses at different aggregation levels.

EXAMINATION OF ITS DATA ARCHIVING ISSUES

Need to Establish Ongoing Dialogue Between Data Providers and Data Users - The best synergy for archiving ITS data occurs when there is an ongoing dialogue between data providers (e.g., operations centers) and data users. An ongoing dialogue will help in the following ways: 1) data users may better understand the available ITS data and its intricacies; 2) data providers may better understand the needs of data users; and 3) data providers and data users may work together in establishing ITS detector designs that meet the data needs of many groups.

Consider Innovative Archiving Approaches to Address Different Data Aggregation Needs - To date, there have been two basic suggested approaches to archiving ITS data: 1) save aggregated data to meet existing needs in a cost-effective manner; 2) save raw, disaggregate data to maintain flexibility of analyses and data exploration. The research team suggests a third approach, in that basic aggregated data are saved for current needs, yet innovative archiving capabilities exist that can provide advanced data users access to raw, disaggregate data. The innovative capabilities discussed in this report include raw data sampling, aggregation based upon statistical variability, on-demand archiving, and data broadcasts.

Need Improved Quality Control Procedures - Basic quality control checks for loop detector data are used in several operations centers; however, many of these checks can only detect illogical combinations of speed, volume, and occupancy for a single data record at a single location. TTI researchers used these and other basic quality control checks to analyze TransGuide® loop detector data from San Antonio, and found that about 1 percent of the data were flagged as suspect. The research team concluded that more advanced quality control procedures are needed for use with archived ITS data. These advanced (and ideally automated) procedures could compare current data at a given location to upstream and downstream detectors, as well as examining historical values, trends, and patterns. Quality control procedures to establish absolute data accuracy ("ground truthing") should be also be included. The current consensus is that archived ITS data not passing quality control checks should be flagged as such, with users making the choice of imputed or replacement values appropriate for their analyses.

Account for Missing Data in Data Archiving System Designs - In analyses of TransGuide® loop detector data from San Antonio, the research found that, on average, nearly 20 percent of all possible data records were missing for various reasons (e.g., local or central hardware and software failures). Because the TransGuide® center's detector maintenance program is considered above average by many (about 10 percent of detectors malfunctioning at any given time), this estimate of missing data may be considerably higher at other locations with less maintenance. Therefore, it is critical that the design of archived ITS data be able to reflect missing data in summary or aggregate statistics. As with quality control checks, the consensus is that missing data be flagged as such, with users making the choice of appropriate imputed or replacement values.

REFERENCES

1. McDermott, J.M. "Update: Chicago Area Freeway Operations." Illinois Department of Transportation, June 4, 1991.

2. Goolsby, M.E. "Digital Computer Serves Effectively in Traffic System Study." *Texas Transportation Researcher*, Vol. 4, No. 2, College Station, Texas, April 1968.

3. Nihan, N., L.N. Jacobson, J.D. Bender, and G. Davis. *Detector Data Validity*. Report WA-RD 208.1. Washington State DOT, Washington State Transportation Center, Seattle, Washington, March 1990.

4. Margiotta, R. *ITS as a Data Resource: Preliminary Requirements for a User Service*. Report FHWA-PL-98-031. Federal Highway Administration, Washington, DC, April 1998, also available at http://www.fhwa.dot.gov/ohim/its/itspage.htm.

5. S.M. Turner, R.E. Brydia, J.C. Liu, and W.L. Eisele. *ITS Data Management System: Year One Activities.* Research Report 1752-2. Texas Department of Transportation, Texas Transportation Institute, College Station, Texas, August 1997, also available at http://ttibiker.tamu.edu/reports/rpt17522.pdf.

6. R.E. Brydia, S.M. Turner, W.L. Eisele, and J.C. Liu. Development of Intelligent Transportation System Data Management. In *Transportation Research Record 1625.* Transportation Research Board, Washington, DC, 1998.

7. Wolfson, D. *Scalable AZTech Data Server Enhancements for Planning and Operations.* Draft Scope of Work. Maricopa County DOT, Arizona, March 1999.

8. Abukar, A. Ft. Worth TransVISION Data Archiving. In *Proceedings from the Texas ITS Data Uses and Archiving Workshop.* Texas Department of Transportation, Texas Transportation Institute, College Station, Texas, March 1999.

9. Wegmann, S. and C. Gloyna. Houston TransStar Data Archiving. In *Proceedings from the Texas ITS Data Uses and Archiving Workshop.* Texas Department of Transportation, Texas Transportation Institute, College Station, Texas, March 1999.

10. Irwin, P. San Antonio TransGuide® Data Archiving. In *Proceedings from the Texas ITS Data Uses and Archiving Workshop.* Texas Department of Transportation, Texas Transportation Institute, College Station, Texas, March 1999.

11. Morris, W.H. and J.B. Robinson. "Data Integration for Design, Construction, and Operations: Effects of Data Integration." TRB Preprint. *Transportation Research Board, 78th Annual Meeting CD-ROM.* Transportation Research Board, Washington, DC, 1999.

12. Walton, J. “Data Collection Through ITS/CVO Initiatives: The Advantage CVO Program.” Presented at the National American Travel Monitoring Exhibition and Conference (NATMEC). Charlotte, North Carolina, May 1998.

13. Ishimaru, John M. and Mark E. Hallenbeck. *Central Puget Sound Freeway Network Usage and Performance*. Draft Research Report. Washington State Department of Transportation, Washington State Transportation Center, Seattle, Washington, 1999, also available at http://weber.u.washington.edu/~trac/floweval/FlowResults.pdf.

14. Turner, S.M., G.A. Carlin and R.H. Henk. *Quantifying the Benefits of High-Occupancy Vehicle Facilities Using Automatic Vehicle Identification Technology.* Report No. SWUTC/95/465020-1, Southwest Region University Transportation Center, Texas Transportation Institute. College Station, Texas, November 1995, also available at http://ttibiker.tamu.edu/reports/rpt650201.pdf.

15. Turner, Shawn. Using ITS Data for Transportation System Performance Measurement. In *Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and Opportunities*. Proceedings of the ASCE Conference, American Society of Civil Engineers, Chicago, Illinois, June 1997, pp. 340-346.

16. Chin, S.M., D.L. Greene, J. Hopson, H.L. Hwang, and B. Thompson. “Towards National Indicators of VMT and Congestion Based on Real-Time Traffic Data.” TRB Preprint. *Transportation Research Board, 78th Annual Meeting CD-ROM*, Washington, DC, 1999.

17. *1995 Travel Atlas for the Northeastern Illinois Expressway System.* Chicago Area Transportation Study. Chicago, Illinois, 1998.

18. Bunch, J., K. Wunderlich, and G. Nelson. *Development of Representative Day Scenarios to Capture the Impacts of Variation in Conditions for Transportation and ITS Planning.* Presented at the National American Travel Monitoring Exhibition and Conference (NATMEC). Charlotte, North Carolina, May 1998.

19. Turner, S.M. and D.J. Holdener. “Probe Vehicle Sample Sizes for Real-Time Information: The Houston Experience.” In *1995 Vehicle Navigation & Information Systems Conference Proceedings*. Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey. 1995.

20. Gaynor, J.M. and D.W. Borchardt. “A Successful Return to Ramp Metering: The Houston Experience.” Presented at the Seventh Annual Meeting of ITS America, Washington, DC, June 1997.

21. Cassidy, M.J. and R.L. Bertini. “Some Traffic Features of Freeway Bottlenecks.” In *Transportation Research Part B 33.* 1999, pp. 25-42.

22. Graves, T.L., A.F. Karr, N.M. Rouphail, and P. Thakuriah. *Real Time Prediction of Incipient Congestion on Freeways from Detector Data.* National Institute of Statistical Sciences Technical Report. March 1998.

23. Park, D. and L.R. Rilett. "Forecasting Freeway Link Travel Times with a Multilayer Feed-Forward Neural Network." Computer-Aided Civil and Infrastructure Engineering, Special Issue on Advanced Computer Technologies in Transportation Engineering. 1999

24. Park, D. and L.R. Rilett. Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Networks. In *Transportation Research Record 1617.* Transportation Research Board. National Research Council. Washington, DC, 1998.

25. Park, D., L.R. Rilett, and G. Han. "Forecasting Multiple-Period Freeway Link Travel Times Using Neural Networks with Expanded Input Nodes." Proceedings of the 5th International Conference on Applications of Advanced Technologies in Transportation Engineering. Newport Beach, California.. April 1998.

26. Rilett, L.R. and D. Park. "Direct Forecasting of Freeway Corridor Travel Times Using Spectral Basis Neural Networks." TRB Preprint 990950. Transportation Research Board, 78th Annual Meeting. Washington, DC, 1999.

27. Dixon, M. and L.R. Rilett. "Evaluation of ITS Automatic Vehicle Identification for Origin-Destination Estimation." TRB Preprint 991459. Transportation Research Board, 78th Annual Meeting. Washington, DC, 1999.

28. Zhou, M. and F.L. Hall. "Investigation of the Speed-Flow Relationship Under Congested Conditions on a Freeway." Presented at the 78th Meeting of the Transportation Research Board. Washington, DC, 1999.

29. Turner, S.M. and W.L. Eisele, Editors. *Proceedings from the Texas ITS Data Uses and Archiving Workshop*. Texas Department of Transportation, Texas Transportation Institute, College Station, Texas, March 1999.

30. Intelligent Transportation Systems Data Registry Functional Operating Procedures. IEEE, http://grouper.ieee.org/groups/scc32/datareg/DRFOP8.DOC, September 28, 1998.

31. Seymour, Ed. Development of a National Data Registry. In *Proceedings from the Texas ITS Data Uses and Archiving Workshop*. Texas Department of Transportation, Texas Transportation Institute, College Station, Texas, March 1999.

32. Mergel, Joseph. *The Capability of ITS to Provide Systems Performance Data: The Use of ITS Data in Planning in the San Antonio Area*. Draft Report. Volpe National Transportation Systems Center, Cambridge, Massachusetts, August 1998.

33. Oak Ridge National Laboratory. *Proof of Concept of ITS as an Alternative Data Resource: A Demonstration Project of Florida and New York Data*. Statement of Work. Oak Ridge, Tennessee, September 28, 1998.

34. Puget Sound Freeway Performance Monitoring Using ATMS Data: A Status Report. Washington State Transportation Center, http://weber.u.washington.edu/~trac/floweval/, January 28, 1998.

35. TDAD: Traffic Data Acquisition and Distribution. University of Washington, http://www.its.washington.edu/tdad/, March 18, 1999.

36. *Highway Performance Monitoring System Field Manual: For the Continuing Analytical and Statistical Data Base*. FHWA Order M5600.1B. Office of Highway Information Management, Federal Highway Administration, Washington, DC, August 1993.

37. *Traffic Monitoring Guide*. Third Edition. Federal Highway Administration, U.S. Department of Transportation, Washington, DC, February 1995.

38. *AASHTO Guidelines for Traffic Data Programs*. American Association of State Highway and Transportation Officials, Washington, DC, 1992.

39. American Society for Testing and Materials. E1442-94 Standard Practice for Highway-Traffic Monitoring. http://www.astm.org. Accessed August 10, 1999.

40. *Information Needs to Support State and Local Transportation Decision Making into the 21st Century*. TRB Conference Proceedings 14. Transportation Research Board, Washington, DC, 1997.

41. *Guidance Manual for Managing Transportation Planning Data*. NCHRP Report 401. National Cooperative Highway Research Program, Transportation Research Board, Washington, DC, 1997.

42. Karash, Karla H. and Carol Schweiger. *Identification of Transportation Planning Data Requirements in Federal Legislation*. Report DOT-T-94-21. John A. Volpe National Transportation Systems Center, U.S. Department of Transportation, Washington, DC, July 1994.

43. *Data for Decisions: Requirements for National Transportation Policy Making*. TRB Special Report 234. Transportation Research Board, Washington, DC, 1992.

44. Silvey, S. *Statistical Inference*. Chapman and Hall, New York, 1993.

45. Eubank, R. *Nonparametric Regression and Spline Smoothing*. Marcell Dekker, New York, 1999.

46. Stephens, M.A. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*. Vol. 69, No. 341, 1974, pp. 730-737.

47. Loop_Client: Real-Time Freeway Sensor Information Over the Internet, http://www.its.washington.edu/software/loop_cli.html, updated April 8, 1997.

48. Zhou, Tao. "IP Multicast and Your Network." *Windows NT Magazine*, April 1999, pp. 109-115.

49. Chen, L. and A.D. May. Traffic Detector Errors and Diagnostics. In *Transportation Research Record 1132*. Transportation Research Board, Washington, DC, 1987, pp. 82-93.

50. Jacobson, L.N., N.L. Nihan, and J.D. Bender. Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System. In *Transportation Research Record 1287*. Transportation Research Board, Washington, DC, 1990, pp. 151-166.

51. Cleghorn, D., F.L. Hall, and D. Garbuio. Improved Data Screening Techniques for Freeway Traffic Management Systems. In *Transportation Research Record 1320*. Transportation Research Board, Washington, DC, 1991, pp. 17-23.

52. Flinner, M. "Traffic Data Quality Procedures." The Pooled Fund Project Newsletter, Issue 1, Minnesota Department of Transportation, St. Paul, Minnesota, December 1997.

53. "The Traffic Analysis Expert System User Guide, Version 3.0." Intelligent Decision Technologies, Longmont, Colorado, 1998.

54. Ishimaru, John M. *CDR User's Guide*. Version 2.52. Washington State Transportation Center, University of Washington, Seattle, Washington, March 1998.

55. Wright, T., P.S. Hu, J. Young, and A. Lu. *Variability in Traffic Monitoring Data: Final Summary Report*. Oak Ridge National Laboratory, University of Tennessee-Knoxville, Oak Ridge Tennessee, August 1997.