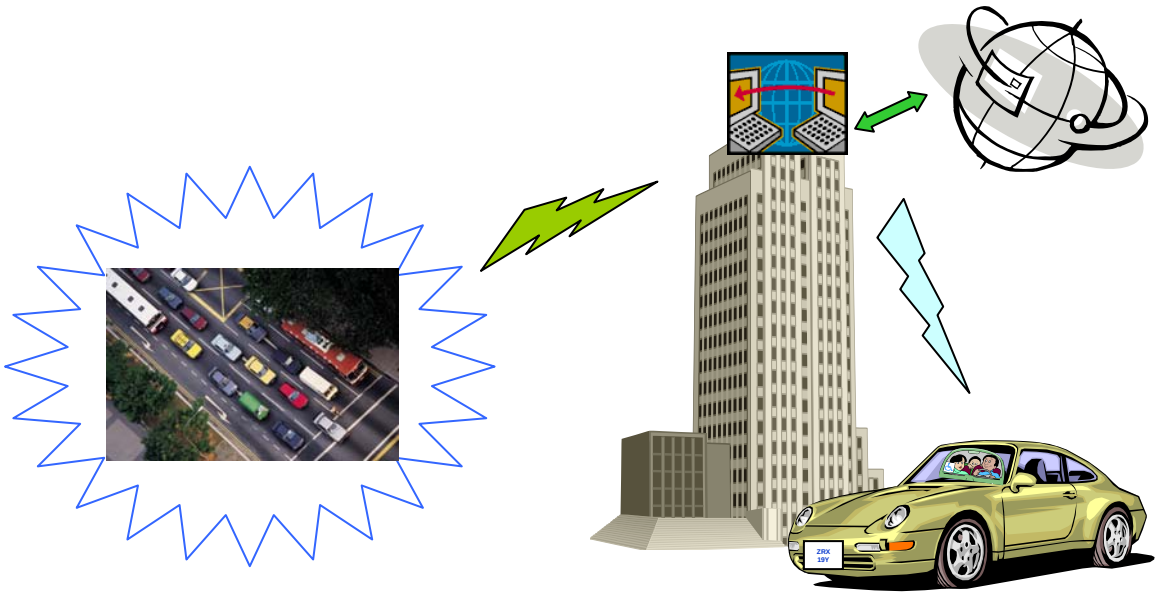
Texas Department of Transportation

Report No. TxDOT 0-4054-1

USING REAL-TIME TRAFFIC DATA FOR TRANSPORTATION PLANNING

October 2002



By

Lei Yu, Ph.D., P.E.,

Fengxiang Qiao, Ph.D., Xin Wang, and Li Xu

Department of Transportation Studies
Texas Southern University

Texas Southern University College of Science and Technology Department of Transportation Studies

1. Report No. TX-02/0-4054-1	2. Government Accession No.	Recipient's Catalog No.
4. Title and Subtitle Using Real-Time Traffic Data for Transportation Planning		5. Report Date October 2002
		6. Performing Organization Code
7. Author(s) Lei Yu, Fengxiang Qiao, Xin Wang, and Li Xu		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Transportation Studies Texas Southern University 3100 Cleburne Avenue Houston, TX 77004		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 0-4054
12. Sponsoring Agency Name and Address TxDOT Department of Transportation Research and Technology Implementation Office P.O. Box 5080 Austin, Texas 78763-5080		13. Type of Report and Period Covered Final Report 2000-2002
		14. Sponsoring Agency Code

15. Supplementary Notes

Project conducted in cooperation with
Research project title: Using Real-Time Traffic Data for Transportation Planning

16. Abstract

In this research report, the state-of-the-art and the state-of-the-practice related to using real-time traffic data are reviewed first and the needs of ITS data for planning purposes are identified. Then, an optimization process that can provide the optimized aggregation level of ITS data for different applications is developed. To illustrate the wavelet algorithm based technique, ITS data archived by the TransGuide® center in San Antonio is used for case study. Aggregation levels for different days of a week and different time periods over the whole year of 2001 are obtained by the proposed approach. Subsequently, an optimization-based sampling approach for data archiving is presented. This data archiving approach can identify the best representative samples of the raw ITS data based on either sum squire error (SSE) or cross validation (CV) while minimizing the required storage size. The sampling approach is realized through a data processing procedure, which is designed to archive real-time/raw data, aggregated data, sampled data, as well as extension factors which can be generated from the raw data. It was tested also in the case study of TransGuide of San Antonio, Texas, where real-time data were collected from 527 loop detectors. After the proposed sampling approach was applied in the case study, only one tenth of the original data is needed to be stored, while the resulting optimal samples contain the maximum information of the raw data, which are able to meet the potential uses of various transportation purposes.

17. Key Words	18. Distribution Statement No restriction. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161		
19. Security Classify (of this report) Unclassified	20. Security Classify (of this page) Unclassified	21. No. of Pages 81	22. Price

Using Real-Time Traffic Data for Transportation Planning

By

Lei Yu, Ph.D., P.E., Fengxiang Qiao, Ph.D., Xin Wang, and Li Xu

Final Report Number 0-4054-1
Research Project Number 0-4054
Project Title: Using Real-Time Traffic Data for Transportation Planning

Sponsored by the Texas Department of Transportation
In Cooperation with
U.S. Department of Transportation
Federal Highway Administration

October 2002

DEPARTMENT OF TRANSPORTATION STUDIES
Texas Southern University
3100 Cleburne Avenue
Houston, Texas 77004

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the Texas Department of Transportation. This report does not constitute a standard, specification, or regulation.

Notice

The United States Governments and the states of Texas do not endorse products or manufacturers. Trade or manufactures' names appear herein solely because they are considered essential to the object of this report.

Acknowledgments

The authors would like to express their sincere gratitude for the support and valuable comments that they received from Project Director, Texas Department of Transportation through the course in conducting this project. The authors would like to express their thanks to the Project Monitoring Committee, other TxDOT personnel for any direct or indirect assistance that they received.

The authors would like to thank the research assistants of Department of Transportation Studies at Texas Southern University (TSU) and University of Texas Austin. Also the authors would like to express thanks to all personnel at the two universities who have directly or indirectly contributed to this project or have provided various assistances.

Credits for sponsor

Research performed in cooperation with the Texas Department of Transportation and the U.S. Department of Transportation.

Table of contents

List of figures

List of tables

Summary

Intelligent Transportation Systems have been defined as the application of advanced sensor, computer, electronics, and communication technologies and management strategies in an integrated manner to increase the safety and efficiency of the surface transportation system. The benefits of the ITS system have been achieved in the areas of safety, traffic operations, transit operations, emergency management and other areas of the transportation practice. One of the essential components that are important to the successful implementation of any ITS system is the real-time ITS data that are collected from various sources. More and more transportation professionals have realized that the ITS data collection and surveillance systems used to operate transportation systems on a day-to-day basis can be used as rich sources of data for various transportation purposes. For example, traditionally, transportation planners use census data supplemented by small sample surveys for building travel demand models and for conducting other specialized analyses. Census data, of course, are comprehensive but are only collected every 10 or more years. ITS applications and their respective sensors and detectors are a rich source of data about transportation system characteristics and performance on real-time basis. Once this huge amount of data is archived, processed, and turned into the format, which could be easily used, it is possible to generate performance measures to aid in the planning, design and operation of transportation systems.

The collection of real-time data in the field and transformation of those data into a package of information for decision-making are important thrusts of ITS systems. The data must be processed to make them consistent before they are used for any algorithms of prediction, control, and decision-making. On one hand, transportation professionals need sufficient and reliable data for their planning, design, management, operation, and other purposes. On the other hand, large amounts of real-time data are generated from ITS Systems, but are not effectively archived for practical use and much of the data is wasted. In reality, it is very expensive and difficult to store and manage ITS data in a way that can meet a variety of needs and applications. Some current data archiving systems have been developed to deal with the massive amount of data collected from Traffic Management Centers (TMCs). However, in most of times, the data are too huge to be well organized, managed, archived, and smoothly used by the end users. For example, some archiving systems simply log raw field data into a single text file, which can reach several million records and more than 70 megabytes per day. It is very difficult, if not impossible, to access and utilize the archived data of this size and format. At another extreme, data archives may consist of several thousands of text files per year, perhaps one for each day for each detector.

It is ideal to convert the large quantities of ITS data into easy-to-understand and well-organized information. For applications to transportation planning, it is necessary to

analyze trends in freeway corridors or systems for long periods of time, which could mean converting gigabytes of data into one page of useful information. There are many other potential transportation applications of this useful ITS data, including bypass analysis, gauging the impact of a new, major employer on the transportation network, other types of traffic impact analyses, air quality analysis, estimation of ITS benefits, computer model calibration, congestion monitoring, transportation planning, or even pavement design.

In response to the above problems, this report presents an optimization-based sampling approach for data archiving. This approach intends to identify the best representative samples of the raw ITS data based on either sum squire error (SSE) or cross validation (CV) while minimizing the required storage size. The approach is realized through a data processing procedure, which is designed to archive real-time/raw data, aggregated data, sampled data, as well as extension factors which can be generated from the raw data. The proposed approach is tested in the case study of TransGuide of San Antonio, Texas, in which real-time data are collected from 527 loop detectors. After the proposed sampling approach is applied in the case study, only one tenth of the original data are needed to be stored, while the resulting optimal samples contain the maximum information of the raw data, which are able to meet the potential uses of various transportation purposes.

In addition to the data archiving problems, appropriate aggregation levels and sampling frames of real-time data in Intelligent Transportation System (ITS) are indispensable to transportation planners and engineers. Conventional techniques obtaining aggregation levels are normally based on the statistical comparison between the original and the aggregated data sets. However, it is not guaranteed that errors and noises will not be transmitted to the aggregated ones, and that the desired information is reserved.

This research develops a technique analyzing real-time data from frequency domain for obtaining aggregation level. ITS data is to be decomposed by wavelet transformation and measuring noises as well as the various useful signal components will then be identified. Aggregation level, which is capable of capturing the required components and eliminating unnecessary ones, can be derived from well-designed sampling frequency. The proposed method will be applied to 20-second volume data archived by the TransGuide® center in San Antonio. Ease of access to information, cost influence and space reduction has been presented. At the end, it is indicated that this technique has the potential to effectively support ITS data archiving.

CHAPTER 1 INTRODUCTION

1.1 Background of Research

The transportation system of the United States consists of more than 6.3 million kilometers of highways and roads, and 503 public transit operators. More than 258 million people, 6 million businesses, and 86 thousand federal, state, and local government agencies produce more than 6.3 trillion kilometers of travel and 4.8 trillion ton kilometers of domestic freight each year. Over the next decade, the rate in building new roads cannot match the increase of travel demand. A key feature of Intelligent Transportation Systems (ITS) is the use of information about transportation system conditions to improve overall system performance.

Intelligent Transportation Systems (ITS) aim to improve efficiency and safety of the transportation system in the area through deployment of advanced technologies and systems management techniques. ITS technologies offer benefits ranging from improved safety on the existing transportation infrastructure to enhanced travel information to users of the transportation facilities. ITS technologies also provide managers of the transportation systems the ability to squeeze more out of existing infrastructure by using the information provided from ITS solutions.

The following data can reflect the benefits coming from ITS system in different areas, such as: Accidents Reduction, Travel Time Savings, Throughput Improvements, and Reduced Operating Costs and Increased Productivity etc. For example, ramp metering, speed enforcement cameras, and collision warning systems can yield from 20% to 80% reduction of accident. Travel time could be saved about 4%-48% due to incident management systems, ramp metering, traffic signal systems, signal priority systems, or in-vehicle navigation systems. 200%-300% and 30%-60% increase are reached in throughput for electronic toll collection and collision avoidance systems respectively. 34%-91% reduction in operating costs for electronic toll collection, 4%-9% operating cost reduction for AVL (automotive vehicle locator) and computer-aided dispatching systems, and 5%-25% reduction in operating costs for fleet management systems also can be seen. Not only the benefits mentioned above, ITS also can provide substantial benefits to other public agencies, including public safety officials, medical emergency officials, and the managers of individual jurisdictions' streets and roads.

The benefits of ITS System have been achieved in the areas of safety, traffic operations, transit operations, emergency management and other areas of the transportation practice. However, the development of these innovations has not always been connected to a transportation problem or need. The purpose of defining user services was to relate ITS strategies and technologies to specific user needs. More and

more transportation professionals realized that the data collection and surveillance systems used to operate transportation systems on a day-to-day basis can be used as rich sources of data that can be useful for thousands of purposes. Once these huge amount data are archived, processed, and turn into the format which could be easily used, it is possible to generate performance measures to aid in the planning, design and operation of transportation systems.

It is noteworthy that significant beneficiaries of ITS data are Traffic Management and Transit Operation, whose systems collect the data in the first place. The use of archived data will move ITS system to the next level. ITS generated data have no counterparts, and offer potential applications. Therefore, archiving ITS generated data after they have been used in ITS operations can provide a valuable resource for a variety of uses. These applications other than immediate uses are: Advanced Traveler Information Systems and user services, incident detection and management, dynamic traffic assignment, commercial vehicle fleet routing and scheduling, research and modeling, transportation planning, demand forecasting and travel demand management, to name a few.

1.2 Objectives of Research

The transportation planning process provides a forum for coordination, communication, and decision making by planners, system operators and managers, State and local governments and elected officials. Through the planning process, capital, operating and management strategies are identified to improve the performance of the transportation system. Planning is a continuing process that responds to changing conditions and new opportunities as they arise. Planning involves a broad array of stakeholders including the public and provides opportunities for inclusive participation in transportation decision making. The incorporation of ITS technologies into transportation system will naturally bring some benefits, which might include:

- Meet the mobility and access needs of the growing population
- Ensure that transportation investments are cost effective
- Protect the environment and neighborhoods
- Promote energy efficiency and enhance the quality of life
- Serve transportation needs in a safe, reliable and economical way
- Develop regional transportation solutions that complement the needs of cities and communities
- Promote innovative and market-based transportation strategies
- Encourage new technologies and support the economy

Thus, the goal of this research is to maximize the usefulness of ITS-generated data, in the context of other available data sources, to facilitate and improve transportation planning methods, processes and activities performed in practice. This goal entails the following objectives:

a. Identify in detail and characterize the real-time data available from advanced traffic management centers (TMC's) such as TRANSGUIDE and TRANSTAR

and establish agreement with TMC's for the consistent provision of archived ITS data;

b. Identify what, and in which way, real-time traffic data may be needed/useful in various transportation planning activities by planning agencies in the selected study areas in Texas;
c. Specify the protocols for transforming ITS-generated data into planning-friendly formats, to reside in database management system for easy retrieval by transportation planning entities;
d. Perform selected applications in conjunction with one or two MPO's in Texas to demonstrate use of real-time data for travel demand model calibration and forecasting;
e. Develop an ITS Data for planning guidebook.

1.3 Outline of This Report

The next chapter of this report presents the review of the state-of-the-art/practice on the use of real-time ITS data. Chapter 3 describes the Intelligent Transportation System and transportation planning. Real matches between ITS data and transportation planning uses are presented. Chapter 4 gives the procedures for determining aggregation level of real-time ITS data for planning purposes. ITS data from TransGuide® center in San Antonio was used for case study. Chapter 5 presents an optimization-based sampling approach for designing of architecture and protocols for data archiving and retrieval. Finally, Chapter 6 gives the conclusions for this report.

CHAPTER 2 SYNTHESIZE STATE-OF-THE-ART AND STATE-OF-THE-PRACTICE

This chapter intends to review state-of-the-art and state-of-the-practice on the use of real-time ITS data.

2.1 TMC's That Archive ITS Data

In order to effectively utilize the ITS data, many TMCs have practically used different methods to archive the data. The following is a brief review of these practices.

Phoenix, Arizona

The Traffic Operations Center in Phoenix saves all of the 20-second data including volume, occupancy, and an estimate of speed that comes into the Center. Five-minute summaries of the data are also created and saved. 15-minute volume data are recorded from the video data. One year of 5-minute traffic volume data are obtained from the Arizona Department of Transportation (ADOT) center.

Los Angeles, California

Prior to the development of the new archiving system, this TMC saves three days of 30-second data and four days of 5-minute summaries into temporary storage. The new system will provide on-line access to 13 months of data through a relational database for personnel in the Center. It is their intent to save not only the 30-second data, but also 5-minute, hourly, daily, and monthly data on CD.

Atlanta, Georgia

Volume and occupancy, along with an estimate of speed, are collected by the ATMS (Advanced Traffic Management System) every 20 seconds. The data are not currently saved.

Chicago, Illinois

The Center is currently obtaining volume and occupancy data and estimating speeds of freeways in the area. The 20-second data are aggregated to the 5-minute level, and only the travel time and occupancy data are saved at this level from 5 to 10 am and 2 to 7 pm. Both the 5-minute and hourly data are permanently archived. It is anticipated that about one year of data is desired on-line.

Montgomery County, Maryland

This TMC collects one-minute volume and speed data from loop detectors, and then aggregates to the 5-minute level before saving. They are currently working on a data archiving effort to identify users and uses of the data.

Detroit, Michigan

The older system data are aggregated to hourly lane volumes, which are saved to magnetic tape. The new system data are aggregated up to one-minute and the volume, occupancy, and speed are saved. 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. The Michigan ITS (MITS) Center is developing standards on an aggregation level and determining an on-line data access strategy.

Minneapolis-St. Paul, Minnesota

The volume, occupancy, and estimated speed data are obtained every 30 seconds and every 5 minutes. The 5-minute data are aggregated and saved. 30-second station data is stored in a binary day file and is available upon request.

TRANSCOM, New York/New Jersey/Connecticut

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

Seattle, Washington

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 5-minute aggregation level.

Toronto, Ontario, Canada

The loop detectors report volume, occupancy, and an average speed every 20 seconds. The 20-second data are aggregated to 5-minute, 15-minute, one hour, daily, and monthly time periods. The TMC archives speed, volume, and occupancy data for 20-second and 5-minute time increments. For data summaries of 15 minutes or more, only volume data are saved.

Houston, Texas

The automatic vehicle identification (AVI) travel time data are stored in 15-minute summaries for future use. One terabyte of data storage is being planned. A project is focused on archiving loop detector data from selected locations in Houston.

San Antonio, Texas

They collect data every 20 seconds, and put the data over the Internet. 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.

Table 1 is a summary of TMCs' data archiving practices based on the above review. It is shown that existing practices are diverse and there exist no systematic and scientific approaches in practices that are being used in dealing with the data archiving issue.

TABLE 1 Summary of TMC Loop Detector Data Archiving Practices

	Real-Time Data	Aggregated Data	Data On-Line Service	Save the Data on CD
Phoenix, Arizona	Yes (all)	Save 5-min	Yes	Intend to do
Los Angeles, California	Yes	Save 5-min and intend to save more	Yes	Intend to change from magnetic tape cartridges
Atlanta, Georgia	No	No	No	No
Chicago, Illinois	No	Save 5-min of certain hours	Yes	Yes
Montgomery County, Maryland	No	Save 5-min	No information	No information
Detroit, Michigan	No	Hourly data of the old system; 1-min data of the new system	Yes (one week)	No (on magnetic tape cartridges)
Minneapolis-St. Paul, Minnesota	Upon request	Save 5-min	No information	Yes
TRANSCOM, New York/New Jersey/Connecticut	No information	Save 15-min	No information	No information
Seattle, Washington	No	Save 5-min	No information	Yes
Toronto, Ontario, Canada	Yes (all)	Save 5-min Other types of aggregated data only save volume	No information	Yes
Houston, Texas	No	No	No	No
San Antonio, Texas	Yes	Save 15-min	Yes	No information

2.2 Existing Applications of ITS Traffic Data

Compared with traditional data, ITS traffic data are collected continuously and at a very detailed level. A wide range of stakeholder functions can be supported with data from ITS system. For example, roadway surveillance data can be used in many stakeholder applications, including development and calibration of transportation models; congestion monitoring; transit planning; intermodal planning; and air quality analysis. Some existing applications of ITS traffic data are listed below:

- Freeway Performance Evaluation in Puget Sound Region, Washington. Loop detector data have been used to monitor congestion patterns, including variability in speeds and travel times.
- Evaluation of HOV Lanes in Houston, Texas. Probe vehicle data were used to compare travel times for HOV and non-HOV lanes.
- Traffic Statistics in Chicago, Illinois. Loop detector data has been used to produce an "atlas" of traffic statistics.
- Traffic Statistics and Congestion Monitoring, Montgomery County, Maryland. Arterial loop detector used to develop traffic statistics.
- Ramp Metering Evaluation, Minneapolis-St. Paul, Minnesota. Freeway and ramp loop detector data used to evaluate cycle lengths on ramps.
- Development of Travel Time Prediction Models, Texas. Data from probe vehicles being used to develop short-term travel time prediction models for traffic operations.

Both the amount and the content of the ITS traffic data could be available for planning purposes of transportation systems and facilities. These data can be very valuable for a variety of transportation planning applications as well, both directly and indirectly, in these activities, which include development and calibration of travel demand forecasting and simulation models, congestion monitoring, transit route and schedule planning, intermodal facilities planning, and air quality modeling.

For example, travel demand forecasting (TDF) models used by transportation planners to forecast future demand for transportation facilities can provide the results include forecasted vehicle flows on a schematic network of the area's highway system and the demand for transit ridership between origin and destination points. ITS data is the rich source of input data for this kind of models, and the real-time traffic data also could be used to calibrate such models.

Highway capacity manual (HCM) is a documented set of procedures used by transportation analysts to determine maximum traffic flow rates and vehicle speeds and delays for a given set of traffic and highway geometric conditions. Loop detector data and/or AVI-equipped probe vehicles can do better jobs than traditional traffic count data. The ITS data is the benchmark to evaluate the output results too.

2.3 Stakeholder and Archived Data User Service (ADUS)

Stakeholders of ITS data should include but are not limited to the following:

- MPO and State transportation planners;
- ITS operators and transportation engineers;
- Transit operators;
- Air quality analysts;
- MPO/State freight and intermodal planners;
- Safety planners and administrators;
- Maintenance personnel;
- Commercial vehicle enforcement personnel;
- Emergency management services (local police, fire, and emergency medical);
- Transportation researchers; and
- Private sector users.

ITS-generated data are collected and can be used in many stakeholder applications. In addition to the applications mentioned in section 2.2, several uses and benefits of ITS data can be reached.

- The continuous features of ITS data can remove the bias from traditional data
- ITS system can provide detailed data to meet the input needs of new modeling procedures
- The multiple uses of ITS data can stimulate the support from other stakeholders
- Promoting the use of archived data for multiple purposes
- Uses of ITS data other than real-time operations and control is a value-added work
- ITS data can support the system performance for better management of existing facilities

The basic function for Archive Data User Service (ADUS) is to provide an ITS Historical Data Archive and to integrate user function needs for data. In the near term, ADUS should be incorporated into the National ITS Architecture to provide all stakeholder agencies with ITS data. The system to support the ADUS should be based on, but not limited to, existing data flows within the National ITS Architecture.

An example of the newly established ITS Archive Data User Service is to utilize information technologies to archive incident data and other traffic data obtained by NAVIGATOR, Georgia's ITS system. Thus, trying to eliminate the secondary accidents (caused by the initial incident) to improve traffic conditions.

CHAPTER 3 ITS DATA AND TRANSPORTATION PLANNING

3.1 Intelligent Transportation System

ITS intends to reduce congestion, enhance safety, mitigate the environmental impacts of transportation systems, enhance energy performance, and improve productivity. A broad range of diverse technologies, known collectively as intelligent transportation systems (ITS), holds the answer to many of our transportation problems, such as improve safety and Efficiency of Surface Transportation. ITS is comprised of a number of technologies, including information processing, communications, control, and electronics. Joining these technologies to our transportation system will save lives, save time, and save money.

According to Hideo Tokuyama, there are three stages in the history of ITS. Stage 1 is the beginning of ITS research in the 1960s and 1970s including CACS (Japan), the Electronic Route Guidance System (ERGS, United States), and the other similar systems. All of these systems shared a common emphasis on route guidance and were based on central processing systems with huge central computers and communications systems, never resulted in practical application.

Conditions for ITS development had improved by the 1980s in stage 2. The U.S. Intelligent Vehicle-Highway Systems (IVHS) project was also progressing. Two projects were going on in Europe at the same time: the Program for a European Traffic System with Higher Efficiency and Unprecedented Safety (PROMETHEUS), and the Dedicated Road Infrastructure for Vehicle Safety in Europe (DRIVE), set up by the European Community. In Japan, work on the Road/Automobile Communication System (RACS) project, which formed the basis for our current car navigation system, began in 1984.

At stage 3, we are at last beginning to understand the full potential of ITS, and intelligent transportation systems are being thought of in intermodal terms rather than simply in terms of automobile traffic. Intelligent transportation systems have started to gain recognition as critical elements in the national and international overall information technology hierarchy.

The National ITS Architecture has proven to be an important tool for promoting both technical and institutional integration of ITS. The architecture has provided a basis for developing ITS standards and for developing local and regional architectures, promoting use of the systems engineering process in ITS. A growing acceptance of the need for regional architectures is occurring within the United States. In addition, the National ITS Architecture effort is being used as a model for national architecture work in many foreign countries.

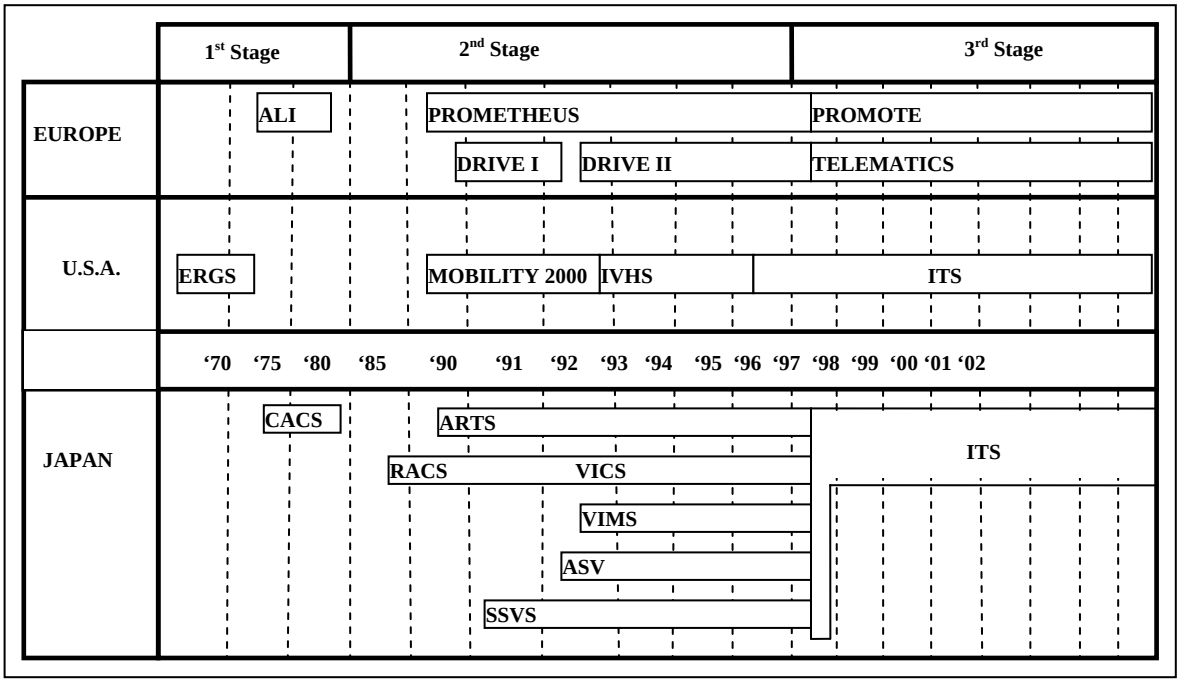


FIGURE 1 History of ITS development.

An architecture maintenance program is needed to keep the National ITS Architecture current and meaningful to increasing numbers of ITS deployers throughout the country who are seeking integrated ITS solutions and system interoperability.

Many areas are either developing or starting to develop regional architectures, with a growing acceptance of the need for such work to provide a framework for project coordination and integration. Most of these areas are using the National ITS Architecture as a tool to reduce the time and effort required to develop their tailored regional architecture. U.S. DOT has developed a software tool, "Turbo Architecture," which is available through McTransTM Center for Microcomputers in Transportation at the University of Florida. This tool has the potential to greatly assist production of regional and project architectures for those using the National ITS Architecture as a basis for the communications and interface framework.

The federal program of providing funding to expedite and facilitate development of ITS standards has had many successes. In particular, work on the ITS Data Bus (IDB) was accelerated and handed off to the private sector, and development of the National Transportation Communications for ITS Protocol (NTCIP) family of standards is being accelerated. The standards program is undergoing a critical period, as the first generations of standards-based products are being deployed.

With some important exceptions, ITS systems use standard off-the-shelf communications media and applicable general purpose data communications standards. Standards unique to ITS are developed primarily for application messages exchanged.

ITS standards can allow equipment from multiple vendors to interoperate, reducing lock-in to single vendors and allowing easier upgrades or expansion of systems. ITS standards also make it easier to add new capabilities and, ultimately, reduce cost.

The ITS standards program is based on a unique partnership, with U.S. DOT providing partial funding to facilitate and expedite he development of ITS-specific standards. The concept is to use the voluntary standards development process normally used in the United States, but with additional funding for specific types of support that facilitate and speed standards development.

In particular, federal support has expedited the development of several important standards, including the in-vehicle ITS Data Bus and NTCIP family of standards, particularly the center-to-roadside subset.

At the same time, availability of Federal support has generated a large number of concurrent efforts, which has made it difficult to track and coordinate efforts and ensure proper focus on critical and timely areas. In addition, some of these efforts, while useful, do not address critical needs and lack the necessary push for successful development and adoption of standards. Some standards efforts would not merely slow down, but would evaporate without Federal funding, which indicates the standard is not essential. Having a large number of concurrent efforts also dilutes available resources, attention, and oversight. Some standards may not be getting sufficient evaluation and critical review from a broad enough base of product developers and users before being adopted.

A few unsuccessful efforts to develop standards have occurred where a perception held that proprietary interests of individual product developers were greater than the benefits of uniform standards.

On a more minor note, the time from start of a standards development effort to availability of standards-conforming products was often underestimated. ITS standards typically take several years or more to develop, and at least another year before products that implement the standard become commercially available.

Despite these problems, a large number of useful ITS standards have been completed, and products that comply with the standards are being developed and deployed. ITS standards are undergoing a critical period within the ITS community, which has expressed skepticism about their value. Successful projects will greatly aid in overcoming the hesitation to move to new, standards-based products, while well-publicized failures will set the program back, regardless of reason for the failure. Guidance and assistance must be available to ensure that first adopters are successful in deploying standards-based systems.

Some critics have cited the need for independent testing of products for standards conformance, including participants at the Institute of Transportation Engineers (ITE) 2000 International Conference in April 2000. Others have stated that vendor self-testing and warranties would be adequate. Both groups have urged standardardized testing procedures.

The metropolitan ITS, rural ITS, commercial vehicle operations, and the Intelligent Vehicle Initiative portions of the National ITS Program each have a specific deployment goal. For example, the metropolitan ITS goal – "Operation TimeSaver" – is to deploy an integrated ITS infrastructure in the nation's 75 largest metropolitan areas before 2006. The commercial vehicle operations goal is to deploy initial operating systems and Commercial Vehicle Information Systems and Networks (CVISN) capabilities in a majority of states by 2003.

The ITS Joint Program Office (JPO) of the Federal Highway Administration (FHWA) of the U.S. Department of Transportation (U.S. DOT) funds the development of several databases that were used to judge various ITS technologies and applications. Those include the following:

- Metropolitan ITS Deployment Tracking Database, maintained by the Oak Ridge National Laboratory.
- Commercial Vehicle Information Systems Network (CVISN) Deployment Tracking Database, maintained by the John A. Volpe National Transportation Systems Center (Volpe Center).
- 1998 Survey of Transit Agencies conducted by the Volpe Center.
- ITS Cost Database, maintained by Mitretek Systems.

Taking Metropolitan ITS Deployment Tracking Database as an example, it contains the results of surveys of states and metropolitan areas to measure progress toward the metropolitan and commercial vehicle operations deployment goals. It contains of surveys of metropolitan areas taken in FY96, FY97, FY99, and FY00 regarding how much ITS equipment they have actually deployed, and surveys of states taken in FY96 and FY98 regarding their CVISN capabilities.

3.2 Select Tested Study Area

San Antonio TransGuide is selected as the tested study area. The city of San Antonio is one of the cities selected for the national deployment of the intelligent transportation infrastructure (ITI). The TransGuide Advanced Traffic Management System (ATMS) currently covers 85 km (53 miles) of freeways with loop detectors placed at 0.8 km (1/2 mile) spacing. The loop detector stations collect volume, lane occupancy, and speed data every 20 seconds. The TransGuide loop detector system collects about 150 megabytes of loop detector data daily. They are archiving this data and making it available on the Internet, but they do not have the resources to effectively manage this data.

TransGuide collects the loop detector data every 20 seconds, and mainly include volume, occupancy, and speed. The Automatic Vehicle Identification (AVI) system will produce vehicle travel time and average speeds data. The TMC save the loop detector data in disaggregate format and make it available through the Internet of TransGuide web site. The TMC also stores the loop data in 15-minute summaries on their Internet site.

3.3 Technologies of Data Collection

Normally, the flow of real-time traffic data can be divided into three parts: sensor technology, data recording and transfer, and data sampling and analysis. For the field of data collection (sensor technology), many devices have been developed to obtain traffic data. These involve radar detectors, video detectors, single or double loop detectors, Automated Vehicle Identification (AVI) systems, and etc. Loop detector and AVI are two major equipment been used to collect real-time traffic data.

The loop detector is wired equipment connected to a power source and placed under the surface of the pavement. This equipment can generate an electromagnetic field, which could be disturbed by inductance when vehicle passing by. The inductance can make the circuit of the loop detector generate an electronic signal for traffic information count. Loop detectors can be installed as single or double. The single detector collects volume and occupancy data, and an estimated speed could be calculated based on the assumption of vehicle length or by flow relationship. Double loop detectors can also get the volume and occupancy data by the first detector, and both are used to calculate the speed data through different arrival time. Data collected from loop detectors is sent to TMCs.

The whole AVI system includes probe vehicles equipped with electronic transponders (tag), detection antennas, roadside data readers, and data collection and processing equipment. The vehicle with AVI tag could be detected by antennas when it is in the read zone. Then the data is transferred to the computer to be processed and stored. AVI is very useful to calculate travel time directly, and it is real accurate. However, it can only be used to count the vehicles with AVI tags, so AVI cannot count volume. The installation and maintenance costs of AVI are very high.

3.4 Data Functions and Transportation Activities

In order to fully use the real-time traffic data, all possible functions of the real-time traffic data need to be listed. The identification of natural matches between the real-time traffic data from the TMC's and the data requirements of various transportation planning activities is very important. Table 2 shows the possible functions of real-time ITS data.

TABLE 2 Matches Between Data Functions and Transportation Activities

Stakeholder Group	Primary Transportation-Related Function	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 and truck percents for TDF and air quality models • 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 • Dynamic traffic assignment • Incident management • Congestion pricing operations
Transit operators	Day-to-day transit operations and short-range planning: 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 (routes and fares) • 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
Construction and maintenance personnel	Planning for the rehabilitation and replacement of	• Pavement design (loadings based on ESALs)

	pavements, bridges, and roadside appurtenances; scheduling of maintenance activities	• Bridge design (loadings from the "bridge formula") • Pavement and bridge performance models
Commercial vehicle enforcement personnel	Accident investigations; enforcement of commercial vehicle regulations	• HazMet response and enforcement • Congestion management • Intermodal access • Truck route designation and maintenance • Truck safety mitigation • Economic development
Emergency management services personnel (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 system monitoring personnel	Data collection related to system conditions and performance	Provide data for other stakeholders: • Traffic counts and travel times for planners (AADT, K- and D-factor estimation; temporal distributions) • Truck weights for maintenance personnel • Performance metrics for administrators
Land use regulation and growth management planners	Development and monitoring of ordinances related to land development	• Zoning regulations • Comprehensive plan development • Impact fees • Taxation policies
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)	
Federal government	Maintain National database related to traffic operations, safety, transit, freight/CVO, etc.	• HPMS • FARS • NTB

3.5 Data Needs and Opportunities for Transportation Planning

There is a broad spectrum of users who must rely on any and all available sources of data to feed the applicable planning models, simulations, and control

strategies. The users' needs for the data are outlined below with their transportation-related tasks:

Metropolitan Planning Organization (MPO) and State Transportation Planning Short- and long-range identification of transportation improvements, congestion management, air quality planning, airport access planning, and the development and maintenance of travel demand forecasting and traffic simulation models. Operation and management of multimodal transportation systems is becoming an important aspect of the transportation planning function.

Transportation System Monitoring Collection and analysis of transportation data for use by policy-making at all levels of government and other customers for policy analysis, performance monitoring, and investment analysis. An example is the Highway Performance Monitoring System (HPMS) which provides data for reporting to Congress on the condition, performance, and future investment requirements of the nation's highway system.

Traffic Management Day-to-day operations of deployed ITS; e.g., operation of traffic signal control systems.

Air Quality Analysis Regional air quality monitoring, and transportation plan conformity with air quality standards and goals.

MPO/State Freight and Intermodal Planning Planning for intermodal freight transfer, goods movement, and port facilities.

Safety Identifying countermeasures for general safety problems or hotspots; automated collision notification; delivery of emergency medical services; automated crash investigation data entry; deployment planning for incident response; hazardous site identification.

Design, Construction, & Maintenance Planning for the rehabilitation and replacement of pavements, bridges, and roadside appurtenances and the scheduling of maintenance activities.

Transportation Research Development of forecasting and simulation models and other analytic methods and improvements in data collection practices. Transportation research encompasses many of the stakeholder functions.

Emergency Management (local police, fire, and emergency medical). Response to transportation incidents, crash investigations, and patrol planning.

Land Use Regulation and Growth Management Development of land use plans and zoning regulations; establishment of growth impact policies; and community economic development.

All ITS historical and nonreal-time data should be capable of being stored, disseminated, and/or manipulated to support users with pre-defined data products. These data include, but are not limited to the following: 1) freeway data, 2) toll data, 3) arterial (nonfreeway) data, 4) parking management data, 5) transit and ridesharing data, 6) incident management data, 7) safety-related data, 8) commercial vehicle operations data, 9) environmental and weather data, 10) vehicle and passenger information data, and 11) intermodal operations data.

3.6 Real Matches between ITS Data and Transportation Planning Uses

It is 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. The archived ITS data is a rich source of historical information and non real-time use. Many of the uses of archived ITS data to date have been for transportation planning or research. These data can be very valuable for a variety of transportation planning applications as well, both directly (as measures of actual demand volumes and indicators of system performance), and indirectly, in the calibration of both demand and supply models for travel demand forecasting and evaluation purposes. These planning uses have primarily consisted of congestion management systems, performance measurement, model calibration/validation, and basic traffic counts and factors.

Application to Travel Demand Forecasting Model Development and Calibration: travel demand forecasting models currently used by Texas MPO's include EMME/2, TRANPLAN and TransCAD. QRS 2 is also used for traffic impact studies in some local areas. These models require huge amounts of network and traffic data as input. Extensive base-year traffic data are also needed for model calibration. Till now, short-duration traffic counts, infrequent or outdated household surveys, and census information are used for model input and calibration. Many fixed and default values are implemented without the recognition of special features in the study area. However, loop detector based data from TMC can provide detailed and continuous information streams of volume, speed, and occupancy etc. AVI equipped vehicles can provide travel times and speed, and OD patterns also could be estimated without the survey. The availability of these real-time data meet the requirements of travel demand forecasting models very well, and enhance the development, validation and calibration processes of these models. Combined with continuous and detailed nature of ITS data, advanced transportation models have been designed.

Application in Traffic Simulation Models: traffic simulation models inputs have consisted of short-duration traffic counts and turning movements at intersections. Other input data needed are obtained through special efforts. Few performance data, such as incidents and speeds, are available for model calibration. TMC installations could provide most of the input data required to run a traffic simulation model. Models could be calibrated through actual conditions directly. Thus, the performance measures generated from a traffic simulation will be more accurate and useful. Intersection turning movements will also be detected and available in the future.

Application to Congestion Monitoring: traditionally, travel times are collected by "floating cars" with usually s few runs on selected routes for the purpose of congestion monitoring. Speeds and travel times are synthesized using analytic methods and traffic simulation. Incidents information is hard to get. Loop detectors provide continuous volume and speed data. Variability of network state descriptors can be directly assessed. AVI-equipped vehicles provide travel times based on larger sample sizes and greater area coverage. Incidents could be reflected through real-time traffic data, and TMC provide detailed incident conditions through their incident management functions. Therefore, congestion monitoring and other system management functions could be implemented more effectively with ITS traffic data.

Application in Air Quality Analysis: mobile source vehicle emission analysis is becoming very important in Texas due to designated non-attainment areas in the state based on the stringent air quality standards issued by the Environmental Protection Agency (EPA). MOBILE is the EPA approved emission factor model and a conventional approach for performing mobile source vehicle emission analysis. The input speed data for MOBILE are obtained from a travel demand forecasting model. Vehicle Miles Traveled (VMT) and vehicle classifications are derived by short counts. Now, with the real-time traffic data, actual speeds, volumes, and even the truck mix will be available by time of day, all of which could be directly input into models such as MOBILE. ITS also changed the basic driving behavior and cycles, which formed the basis for the MOBILE emission factor model. In order to estimate the mobile source vehicle emission in a more accountable and accurate way, new generation of modal emission models are being developed, which require a high level of details of traffic data input. In this context, the ITS traffic data are indeed necessary in performing mobile source emission analysis using the new generation of modal emission models.

Application for MPOs and State Planners: users, as MPOs and state transportation planners, 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 do not require the real-time data, but the data at a relatively aggregated level. For example, hourly or daily summary data are needed. Many planning applications need the data from various segments and locations of a study area. Normally, daily summary results are required, however, peak hours and peak periods data are interested too. For example, the system performance and congestion management system of Washington State can generate the valuable information about when and where congestion is occurring through monitoring traffic volumes of the network and reporting vehicle speeds. 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. Planners can obtain valuable information such as annual average daily traffic (AADT) volumes, monthly seasonal factors, holiday travel rates, hourly variation, and estimated travel times and speeds etc.

Application in Operations: the day-to-day operations of TMCs are mainly concerned by traffic management operators. Archived data applications interested to

traffic management operators include developing traffic control strategies, performing highway capacity analyses, incident management, and performance evaluation and monitoring. Based on the work of personnel of Minnesota Department of Transportation, the archived ITS data has proved valuable for improving ramp metering operations.

Application in maintenance and construction: the planning of maintenance and construction also needs ITS data. It is very important to determine the lane closure schedule in order to cause the smallest traffic impact to motorists. Hourly traffic volume and/or speed data can tell when lanes should be closed and re-opened. Hourly traffic volumes are also of value for pavement and bridge maintenance to determine the amount of infrastructure use to estimate maintenance needs.

Application for Researchers: transportation researchers typically use ITS data available to develop forecasting and simulation models, analyze highway capacity, and develop incident detection modeling. For example, the research team from University of California at Berkeley collected speed, occupancy, and volume data from loop detectors to evaluate traffic features of freeway bottlenecks on two freeways. The research evaluated several freeway bottleneck factors such as 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. The National Institute of Statistical Sciences (NISS) and Bell Laboratories gathered traffic volume and lane occupancy data from single detector to predict congestion in real-time. It is very useful to develop congestion prediction models so that traffic operators can make traffic management changes. Archived AVI data also can be used to develop Travel Time Prediction Models and to estimate Origin-Destination Matrices.

The above examples demonstrate that there are many valuable applications of real-time traffic data in transportation planning, which have not been extensively investigated in the state-of-the-art and implemented in the state-of-the-practice. Combining with the development and deployment of ITS technologies, the ITS traffic data will provide the full advantages to improve both effectiveness and efficiency in transportation planning. The archived data user service will also prompt the utilization of ITS data by transportation planning agencies through a convenient and efficient way.

CHAPTER 4 DETERMINE AGGREGATION LEVEL OF REAL-TIME DATA FOR PLANNING PURPOSES

Appropriate aggregation levels and sampling frames of real-time data in Intelligent Transportation System (ITS) are indispensable to transportation planners and engineers. Conventional techniques obtaining aggregation levels are normally based on the statistical comparison between the original and the aggregated data sets. However, it is not guaranteed that errors and noises will not be transmitted to the aggregated ones, and that the desired information is reserved. Wavelet decomposition is a new technique that can be applied to the determination of aggregation level. The research team develops an optimization process that can provide the optimized aggregation level of ITS data for different applications. To illustrate the proposed technique, ITS data archived by the TransGuide® center in San Antonio was used for case study. Aggregation levels for different days of a week and different time periods over the whole year of 2001 were obtained by the proposed approach.

4.1 Background of Data Aggregation

The potential use of ITS data often exceeds the immediate application for which they were collected, *i.e*. real time traffic control. In addition to their intended use for traffic operational purposes, it is possible and desirable to use these ITS data for various transportation planning purposes. The time intervals for archiving data vary considerably though collecting ITS data from local controllers is fairly consistent among many of the Traffic Management Centers (TMCs). Because of the detailed nature of ITS detector data, data aggregation is often a consideration when archiving ITS data. Aggregation refers to the time interval at which data are summarized (Turner, 2001). The robustness of different traffic measurements defined over different aggregation temporal intervals often depends on the duration of that interval.

At most of TMCs, detector data for either 20 or 30 seconds are polled at the same time interval from the roadside controllers. Wide fluctuations are displayed while traffic Volume, Speed and Occupancy are retrieved over extremely small time intervals (20 or 30 seconds). This kind of fluctuations may be useful for real-time traffic control, but not for planning applications. Aggregation of raw ITS data is essential for the purposes of capacity analysis, congestion monitoring, air quality analysis, general planning, etc. Besides, to provide maximum flexibility for data exploration and mining is another target for data aggregation.

4.2 Current Approaches

Existing guidance for traffic data programs suggests minimum intervals of 15 minutes for urban areas, with 1-hour summaries being acceptable for general purposes

(AASHTO Guidelines for Traffic Data Programs, 1992; Traffic Monitoring Guide, 1995). This kind of guidance is necessary in real application, but lacking of theoretical explanations, yet.

In the recent years, two branches of approaches conducting aggregation levels theoretically are the statistical based approach and the wavelet decomposition approach. Gajewski *et al* (2000) used two statistical techniques that can be used to determine optimal aggregation levels for archiving ITS traffic data: the Cross-Validated Mean Square Error (CVMSE) and F-Statistic Algorithm. Based on results from both statistical techniques, the optimal aggregation levels are 60 minutes or more during periods of low traffic variability, and 1 minute or less during periods of high traffic variability (e.g., congestion). The criteria designed for this kind of optimization process is the similarity between the original ITS data sets and the aggregated sets.

Another attempt to obtain the aggregation level comes from the other point of view. ITS data sets were treated as a kind of signal, which can be decomposed into several different levels. Noises and unnecessary components can be removed depending on the purpose of applications. Conceptual description and theoretical mechanisms had been revealed, with a preliminary process illustrated (Qiao, Yu and Wang, 2002). However, no optimization criteria and practical process has been established.

4.3 Discussion on "Best" Aggregation Level

Select the "best" aggregation level is a local decision that is best informed by data needs/requirements as well as available data management resources (Turner, 2001). In some areas, the data user needs/requirements may not be clearly defined, or someone may suggest that the user needs will change or evolve as more detailed ITS data becomes available. In situations like these, the "best" data aggregation level can be obtained based on some technical approaches.

From common sense, it seems that the aggregated data sets should match the data sets achieved from local detector. This is the base for the conventional statistical approaches to determine the optimal aggregation levels.

Nevertheless, it should be noticed that the achieved data set is already not exactly the same as the original ITS signal existed on road. ITS Data is archived through the process of signal detection, transmission, and even modulation and demodulation. For example, regardless of wireless transmission or optical fiber communication, errors in the detection mechanism can arise from various noises and disturbances, and can be then inevitably introduced when data arrive at TMCs. The noise sources can be either external to the system (e.g., atmospheric noise, equipment-generated noise, etc.) or internal to the system (e.g., shot noise, thermal noise, etc.) (Keiser, 2000). Internal noise is present in every communication system and represents a basic limitation on the transmission or detection of signals (Keiser, 2000). Even though many ITS deployments have traffic management software or field controllers with basic error detection (mainly focusing on suspect, erroneous or missing data by

examining the data set itself), additional advanced error detection capabilities may be desirable for data archiving systems (Turner, 2001).

Figure 2 is an illustration of the basic process of ITS data archiving. Ideally, the archived data set S_e should be the discrete version S_d of the original ITS signal S on road. However, owing to the existence of the noise source, the data set S_e archived on hand contains various kinds of noises e. The statistical techniques for optimal aggregation level try to find the aggregated set S_a so that it is close enough to the noised one S_n. This is risky because obviously, aggregated set S_a obtained in this way also aggregate the undesired noise e, which cannot be separated by the conventional statistical approaches.

Therefore, no matter what need the users may require, the aggregated data set should at least be de-noised. The "best" aggregation level should mean that the aggregated ones contain all or part, if not all are desired, of the components of the original ITS signal S on road. This is what the wavelet approach aims to be (Qiao, Yu and Wang, 2002).

4.4 Methodology

The proposed methodology for optimizing ITS data aggregation level will at first decompose the archived data set into several different levels based on the technique of wavelet decomposition. Then similarities among archived data sets with some kinds of common characters (i.e. from same day of weeks) will be compared. A selection criterion will be formed mathematically to automatically umpire which components should be maintained and which should be discarded. Finally, based on the famous Shannon Theorem, the optimal aggregation level can be determined.

4.4.1 ITS Data Decomposition

ITS data is a kind of signal that is pulled from roadside detector in electronic, optical or magnetic manners. Although the physical background of ITS data is different, its noise reduction, which is one of the important steps in processing different types of signals and images, should be the same as the other sources of data like in medical, geophysical, and synthetic aperture radar signals. De-noising is a well developed technique in the field of Digital Signal Processing (DSP).

According to the theory of DSP, all signals can be decomposed into a series of Sino signals, each of which corresponds to a unique frequency. The famous Fast Fourier Transformation (FFT) is a good tool to transfer the signal from time domain to frequency domain (Porat, 1997). In the frequency domain, one Sino form will only be represented as one single bar line corresponding to its oscillation frequency. Suppose there is a noised signal like ITS signal S_e, which is the combination of an original signal S and the noise e as shown in Figure 3(a). Obviously, it is impossible to distinguish the original single S from the unexpected noise e in time domain like Figure 3(a), where all horizontal axes are time.

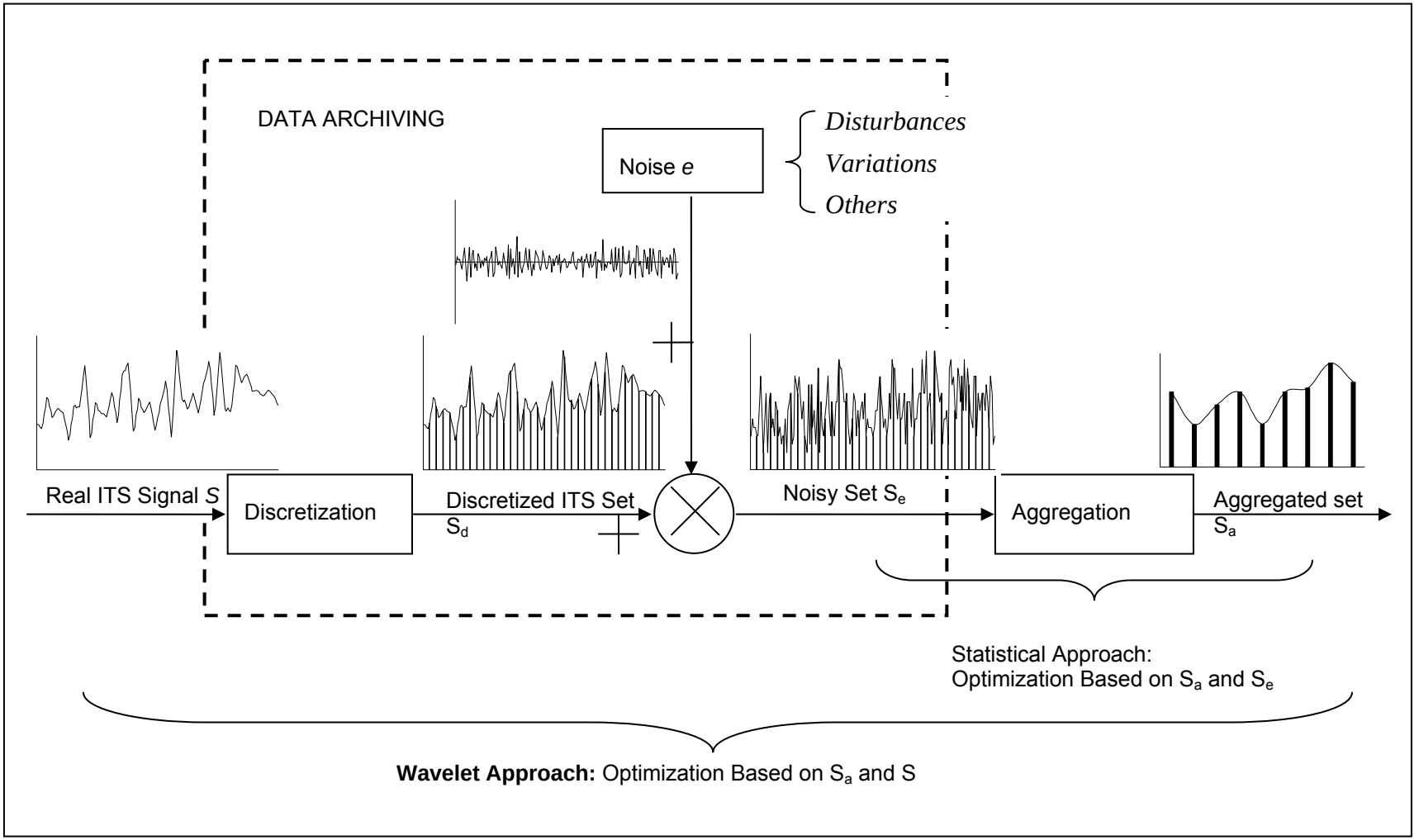
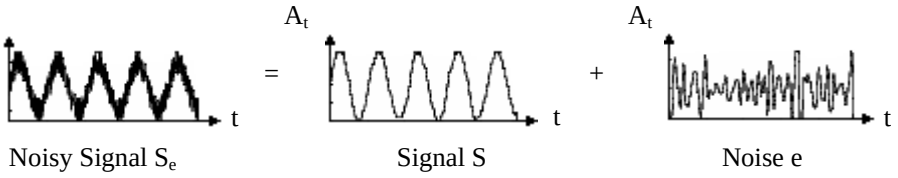
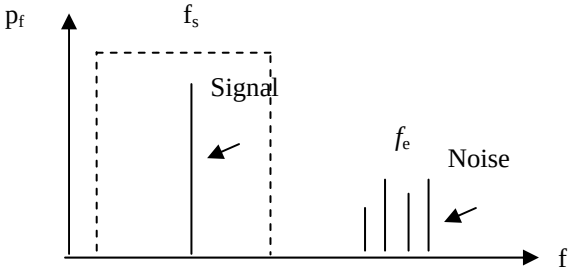


FIGURE 2 ITS data archiving process.

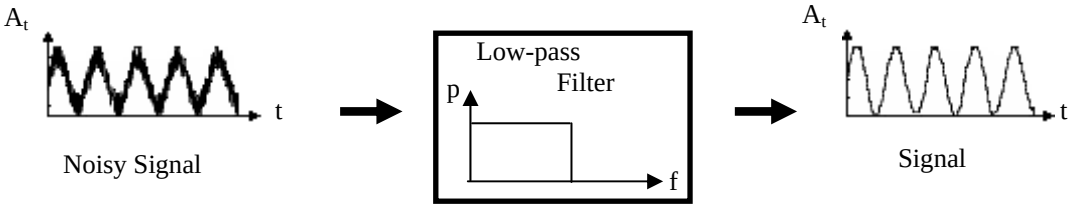
However if all of the frequencies of the noised signal S_e are known, it is very clear to see what is the signal frequency f_s and what is the noise frequencies f_e as in Figure 3(b). Normally, the noise frequencies are higher than the useful signal. By using some kinds of lower pass filter or similar techniques, it is easy to abstract the original signal and eliminate the noise, which is illustrated in Figure 3(c).



(a) Signal decomposition in time domain



(b) Signal decomposition in frequency domain



(c) De-noising of signal

FIGURE 3 Illustration of signal decomposition and de-noising.

Wavelet transformation is a new technique in DSP, which expresses a signal as a linear combination of shifted and scaled versions of a finite duration waveform: the mother wavelet. According to whether the scale can be continuously taken, there are two kinds of wavelet transform: the Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). CWT is defined as the sum over all time of the signal multiplied by scaled and shifted versions of the wavelet function (Misiti *et al*. 2001).

In wavelet analysis, the combination of all the high-scale low-frequency components of the signal is called approximation, while the low-scale high-frequency components of the signal are called details. Wavelet decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower resolution components. This is called the wavelet decomposition tree (Misiti *et al*. 2001) as can be illustrated as:

$$S = A_1 + D_1 = A_2 + D_2 + D_1 = ... = A_n + \sum_{i=1}^{n} D_i \quad (1)$$

where, A_n is the n^{th} approximation and D_i is the i^{th} details.

Unlike conventional techniques, wavelet decomposition produces a family of hierarchically organized decompositions. The selection of a suitable level for the hierarchy will depend on the signal and experience. Often the level is chosen based on a desired low-pass cutoff frequency (The readers may refer to Percival and Walden, 2000 for more detailed information about wavelet).

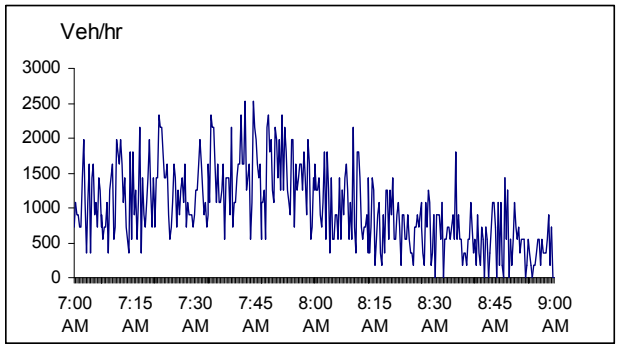
Figure 4 is an illustration of the decomposition of archived ITS data set. The archived ITS data set is plotted in Figure 4(a). By applying the wavelet transformation, two detailed components and one approximation are generated in case the aggregation level is chosen as 2. Figure 4(b) and 4(c) are the two details D_1 and D_2, while Figure 4(d) is the resulted approximation A_2.

It is easy to see from Figure 4 that after decomposition, the two details D_1 and D_2 have much higher fluctuations (*i.e*. higher frequencies) than the resulted approximation A_2. Some of the details higher frequency components normally correspond to some kinds of information that may not be interested by traffic engineers.

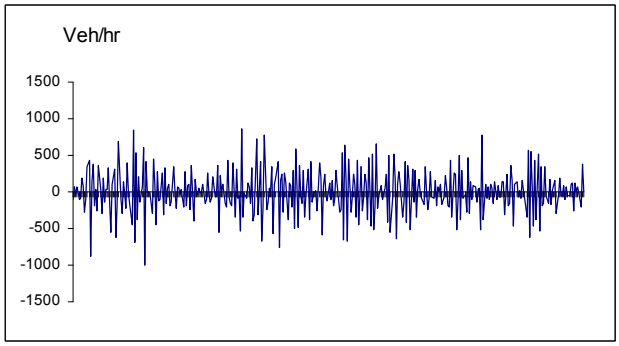
In decomposing the ITS data set and trying to get results like in Figure 4, specific scale needs to be chosen beforehand. The relationship between scale and frequency can only be given in a broad sense, and it is better to speak about the pseudo-frequency corresponding to a scale. A way to do it is to compute the center frequency F_c of the wavelet and to use the following relationship (Abry, 1997).

$$F_a = \frac{\Delta \cdot F_c}{a} \quad (2)$$

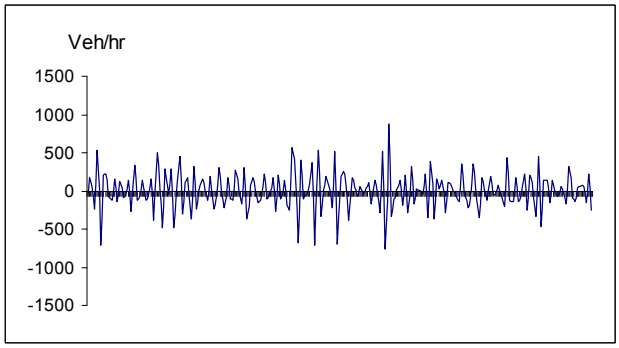
where, a is a scale, Δ is the sampling period, F_c is the center frequency of a wavelet in Hz, and F_c is the pseudo-frequency corresponding to the scale a in H_Z.



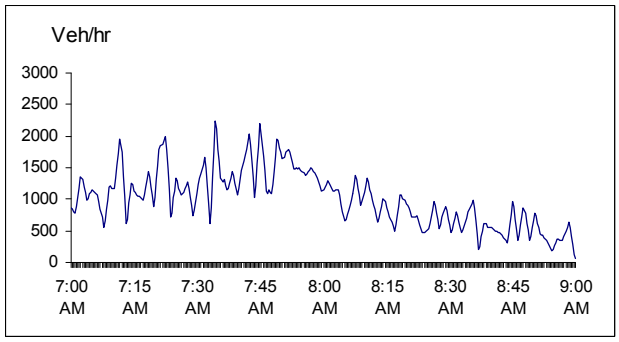
(a) Original Signal = $A_2 + D_2 + D_1$



(b) D_1 – Detail Decomposition Level 1



(c) D_2 – Detail Decomposition Level 2



(d) A_2 – Approximation Decomposition Level 2

FIGURE 4 Wavelet decomposition for archived ITS data set.

Calculation of Continuous Wavelet Transformation is based on Scales, whereas Discrete Wavelet Transformation is based on Levels. Level 1 equals to 21 Scale (Scale 2), Level 2 means 22 Scale (Scale 4), etc. Scales are evenly distributed along time axis, but Levels are not. Therefore, more detailed decomposition levels can be obtained by Continuous Wavelet Transformation.

Both CWT and DWT analyses have their own advantages. DWT analysis ensures space-saving coding and is sufficient for exacting reconstruction, while the CWT analysis is often easier to interpret, since its redundancy tends to reinforce the traits and makes all information more visible. This is especially true for very subtle information.

4.4.2 Similarity Analysis of ITS Data

If more than one day's ITS data are decomposed into the same scales, similarity analysis can be conducted among them. Those scales on which coefficients are similar could be served as the main frequency bandwidth that represents common characteristic of the group of data sets that come from same weekday or weekend. The frequencies being higher than that can then be considered as unnecessary components.

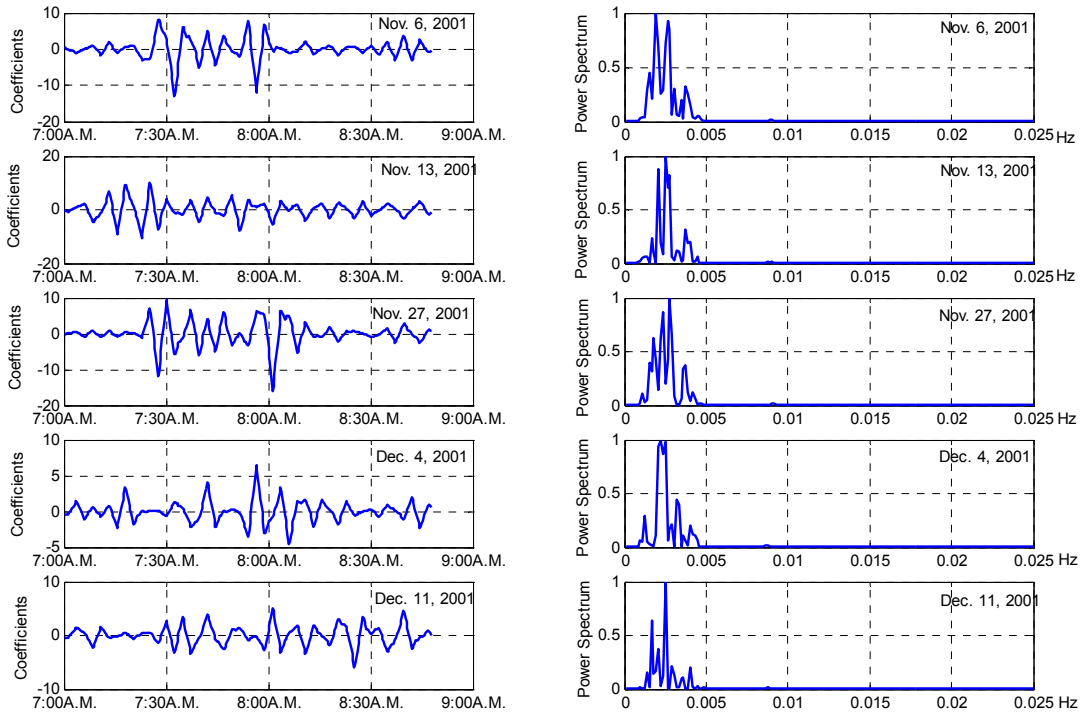
To compare the similarities at a certain decomposition level among several data sets, all the waveforms for that level should be firstly transformed from time domain into frequency domain by Fast Fourier Transform (FFT). This is because frequency is the most important property that distinguishes one signal from the other. The readers may refer to reference like (Porat, 1997) for details of this transform. Similarity can be calculated through the following equation (Turner, 2001), where a dissimilarity index L_i is set for comparison at decomposition level t.

$$L_i = \sum_{t}^{n_t} \sum_{j=1}^{n_i} (p_{ij}^t - \overline{p}_{ij})^2, \quad \text{with } \overline{p}_{ij} = \frac{1}{n_t} \sum_{t=1}^{n_t} p_{ij}^t \tag{3}$$

where L_i is the dissimilarity index for data set t at decomposition level i; p_{ij}^t is the power spectrum, a measurement of the power at various frequencies, at point j for data set t at decomposition level i; $\overline{p}_{ij}$ is the mean of p_{ij}^t along all ITS data set t with the total number of sets as n_t.

By (3) the smaller value L_i stands for a higher coefficients' similarity of that scale.

Figure 5 plots the wavelet coefficients and corresponding FFT for 5 Tuesdays' ITS data via CWT at decomposition level 4. While on the left of Figure 5 the wavelet coefficients were plotted in time domain, the comparison among them should be conducted based on their frequency properties from FFT, which is illustrated on the right of Figure 5.



Discrete Wavelet Transform (DWT) Fast Fourier Transform (FFT)

Data source: 20 second ITS data from TransGuide® at detector L1-0010W-566.641

FIGURE 5 Similarity comparisons among some 5 Tuesday morning peak data via CWT and FFT at decomposition level 4.

4.4.3 Criteria for Selecting Similar Components

The dissimilarity index L_i obtained in (3) has no absolute meaning. However if all the dissimilarity indices for a set of ITS data at different decomposition levels are put together, there should exist a minimum dissimilarity which can be captured by some kind of mathematical algorithm. Note that all the FFT values for different decomposition levels should be unified and standardized before FFT and calculation of dissimilarities for the consistentancy of comparison. According to the definition of dissimilarity index, the decomposition level where the minimum dissimilarity is located should correspond to the most similar components of ITS data.

The next problem is how to obtain the optimal decomposition level. The one with the minimum dissimilarity index cannot be directly used as the optimal one owing to the variations of the relationships between the dissimilarity index and the decomposition level.

For example, when the similarity for Tuesday is to be compared, several groups of Tuesday ITS data can be formed. For each group, one pair of relationship between

the dissimilarities and the decomposition levels can be obtained. Therefore several pairs of relationships for the all the groups can be obtained. Probably, each pair of relationship has its own minimum decomposition level. The answer will not be unique if no further countermeasures are taken.

Since the lower decomposition levels correspond to the higher frequencies, dissimilarities at the lower decomposition level range should be relatively higher; while the similarities in the higher decomposition level range should not vary too much with some lower common frequency information contained, e.g. the monthly change, seasonal change and yearly change of traffic characteristics.

In forming the mathematical representation of the relationship between dissimilarity index and decomposition level, it is easy to think of the attenuation functions that can decrease rapidly at the beginning and change slightly afterwards. For example, the exponential function is one of the good candidates to meet this requirement.

Suppose Φ is the set of this kind of attenuation functions. The best one is the function $f(t) = f^*(t) \in \Phi$ that can meet the following requirement through regression.

$$f^*(t) = \arg\min_f \left(f(t_g) - L_{t_g}^g \right) \quad \forall \text{all } t \text{ and } g \tag{4}$$

where, g is the group number. L_t^g is the dissimilarity index for group g at decomposition level t.

After the proper function is chosen, the optimal decomposition level can be selected based on $f(t)$. Theoretically, the minimum value of attenuation function $f(t)$ occurs when $t \to \infty$. However, it is a common practice that a cut-off value of t be chosen at the point where the maximum value (or initial value) of $f(t)$ already decreased to a certain critical parameter α, where $0 \le \alpha \le 1.00$. The exponential function (5) is normally used as the attenuation function where α is chosen between 0.05 and 0.10 in some Engineering areas such as Electric Engineering.

$$y = ke^{\lambda t} \tag{5}$$

In (5), y is the value at time t, while k and λ are parameters to be calibrated from data. Figure 6 illustrates the one of the pair of relationships between the dissimilarity index and the decomposition level, together with the fitted exponential function.

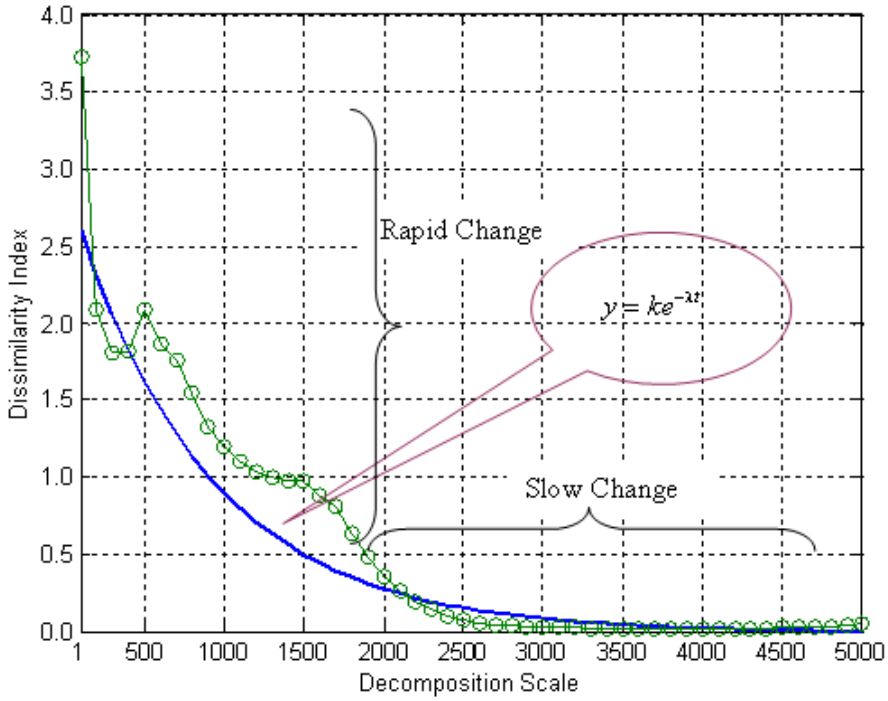


FIGURE 6 Selection of optimal decomposition scale based on regression of attenuation function.

4.4.4 Sampling Interval and Optimal Aggregation Level

Recall that the optimal decomposition level is the one with the most similarity. All the components with the decomposition level less than the optimal one contain more dissimilarity. These lower decomposition levels, especially the lowest level where noises are contained, can not represent the common traffic characteristics of groups of ITS data, and therefore should be discarded.

It is needed to have a way to keep all the information at the higher decomposition levels and discard the lower ones. The famous Shannon's Sampling Theorem can meet this kind of requirement. According to Shannon's Sampling Theorem, in order to accurately represent an analog signal, the minimum sampling frequency must be equal to or greater than twice of the highest frequency component of the original signal. In brief, sampling frequency should be equal to or greater than twice of the interested frequency of the signal (Marven and Ewers, 1996).

Figure 7 shows the result of taking different sampling frequency f_p of signal with a cut-off frequency f_a. It is easy to see that more similarity can be obtained if $f_p >> 2f_a$. However, the sampled signal can basically match the original one when $f_p = 2f_a$; while aliasing occurs when $f_p < 2f_a$.

This sampling frequency f_p can serve as the optimal aggregation level that is able to capture the required frequency component and eliminate the other unnecessary ones and noises.

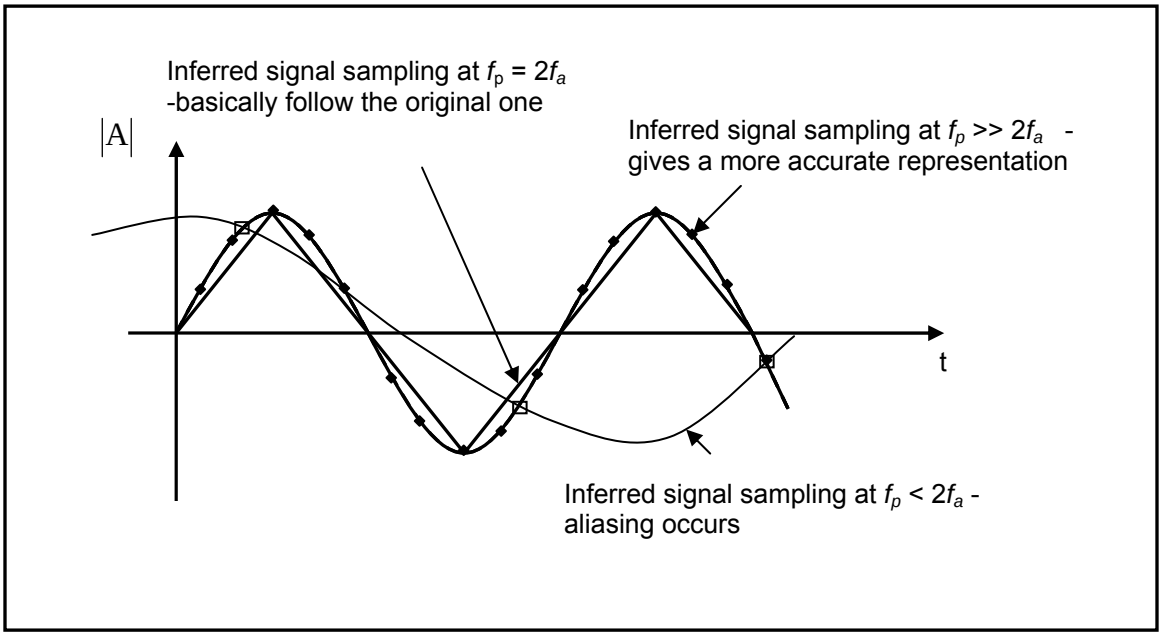


FIGURE 7 Illustration of comparison between sampled and original signal under different sampling frequency f_p.

4.4.5 Procedures Obtaining Optimal Aggregation Level

The whole procedures for obtaining optimal ITS data aggregation level can be summarized into the following steps:

Step 1: Decompose the interested ITS data sets into details and approximations;
Step 2: Compute the dissimilarity indices for each decomposition levels by (3);
Step 3: Find the best attenuation function based on regression by (4);
Step 4: Set the critical parameter α and obtain the optimal decomposition level;
Step 5: Convert the decomposition level into corresponding sampling frequency;
Step 6: Obtain the optimal aggregation level based on the sampling frequency.

4.5 Case Study

To illustrate the proposed methodologies, case study was carried out for getting the optimal aggregation levels for TransGuide® San Antonio ITS data. Parts of ITS data sets on the year of 2001 were analyzed. To meet the various needs from different possible users, optimal aggregation levels for different time of day, different days of week, different weeks, and different months in 2001 were determined.

4.5.1 ITS Data Source and Data Selection

Archived ITS data for the whole year of 2001 (January to December, 2001) were downloaded from San Antonio TransGuide® (ftp://www.transquide.dot.state.tx.us/lanedata) for this study. The data were mainly from detectors: L-0010W-566.641 and L-0010W-572.973 (TransGuide® ID). Both of them are in busy locations (on IH-10, Milepost 566.641 and 572.973) and have three lanes in each direction.

4.5.2 Optimal Aggregation Level for Different Time of Day and Day of Week

It is well known that traffic patterns on most urban roadways vary with time of day (peak hour or non-peak hours) and day of week (weekdays vs. weekend). For example, peak-hours traffic usually appears periodically during weekdays, but not on weekends.

The optimal aggregation level for hourly data (0:00am ~ 11:00pm) from 6 randomly sampled Tuesdays throughout the year 2001 are determined, the results of which are shown in Table 3.

In Table 3, the first several rows list the results by wavelet for Volume, Occupancy and Speed, respectively. In the last row, the results by CVMSE were listed for comparison.

From Table 3 it is shown that different optimal aggregation levels will be obtained based on different traffic variables used as well as the detector where the ITS data were retrieved. Figure 8 plots the relationships between the optimal aggregation levels and the time of day for different variables (left three) and different lanes (right three). It seems that in this case, the optimal aggregation levels for speed at morning/afternoon peaks are observably smaller than other time periods.

For example for speed, the aggregation levels are 6 min and 6 min at 7:00am and 5:00pm, respectively, while become 23 min at midnight. The aggregation levels for volume and occupancy have similar trends although they are not as obvious as for speed. For changes with different lanes as shown by speed on the right column of Figure 8, the aggregation levels for all lanes are apparently similar, especially for lane 2 and lane 3. Therefore, optimal aggregation levels for different variables and lanes are different. A weighted average is suggested if the user needs the optimal aggregation levels for two or more variables and/or lanes.

TABLE 3 Optimal Aggregation Level for Different Time Period

Time Period [4]		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	20	22
Volume [1]	L1[3]	26	25	19	18	13	13	11	13	11	14	17	16	10	12	14	11	14	9	15	19	20	19	16
	L2	18	17	22	19	19	11	13	12	12	14	16	12	12	12	13	12	10	9	13	12	12	15	17
	L3	10	32	28	24	26	14	17	13	13	16	11	13	15	12	13	12	11	13	19	16	15	14	17
Occupancy [1]	L1	25	23	16	18	9	12	13	13	15	15	16	17	11	13	17	7	13	9	14	19	18	22	16
	L2	18	20	20	25	18	13	14	14	19	15	17	14	14	13	14	11	10	12	10	18	14	15	15
	L3	11	22	34	26	24	13	18	13	12	20	15	15	16	14	14	14	14	13	19	16	15	16	18
Speed [1]	L1	23	20	23	18	14	11	11	6	9	13	14	15	10	12	11	10	7	6	8	12	17	18	15
	L2	15	17	18	24	22	13	8	5	6	9	7	7	7	7	6	5	5	5	5	8	11	12	14
	L3	10	23	26	23	17	13	11	7	9	13	9	11	10	11	10	10	8	7	11	14	13	15	17
Speed [2]	L1	60^{+}	60^{+}	60^{+}	60^{+}	60^{+}	30	1	1^{-}	1^{-}	15	15	30	30	30	30	1^{-}	1	1^{-}	1	30	15	30	30
	L2	30	60^{+}	60^{+}	60^{+}	60^{+}	30	5	1^{-}	1^{-}	1	15	60^{+}	30	30	5	1^{-}	1^{-}	1^{-}	1	30	15	15	30
	L3	60^{+}	60^{+}	60^{+}	60^{+}	60^{+}	30	15	1^{-}	1	30	30	30	30	30	30	1^{-}	1	1^{-}	5	1	5	15	15

Note:
[1] Results by wavelet Transformation for ITS data on 6 Tuesday randomly sampled throughout the year 2001

[2] Results by CVMSE for ITS data on Tuesday May 4, 1999 (Source: Gajewski, *et al*. 2000)

[3] L1, L2 and L3 represent lane 1, lane 2 and lane 3

[4] Time periods counted in hours starting from 12:00 am to 10:00 pm

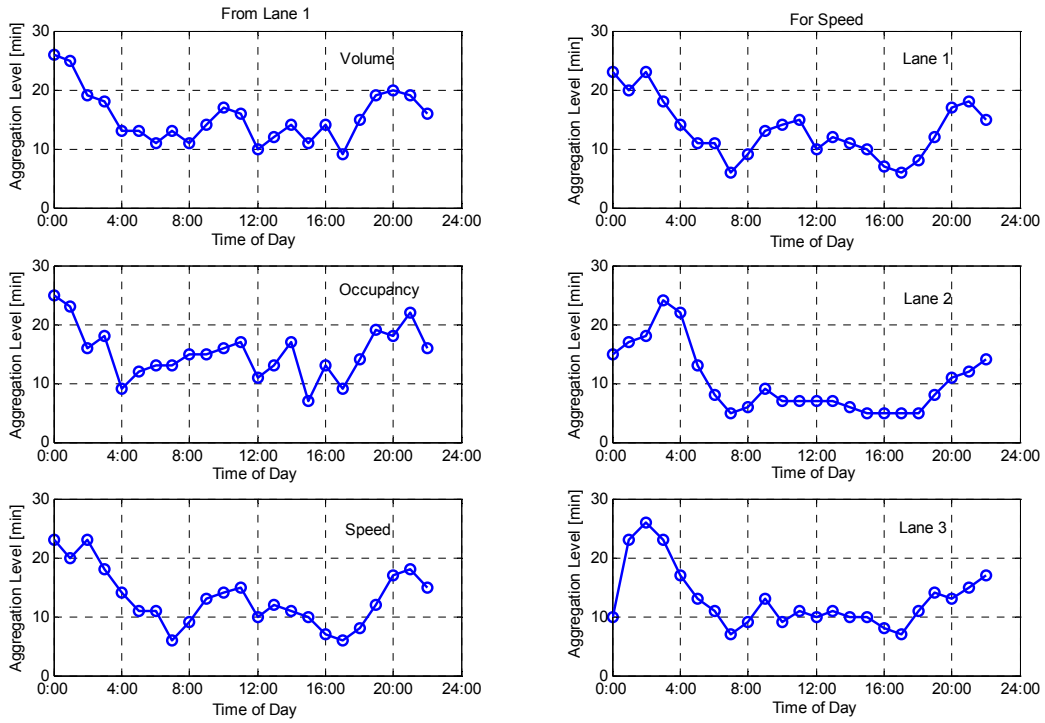


FIGURE 8 Relationships between optimal aggregation level and time of day according to different lanes and different traffic variables.

To determine the optimal aggregation level for different day of week, ITS data were picked up from the whole year's data base. Starting from the first week of the year 2001, Mondays for the $(1+7(n-1))^{\text{th}}$ week, Tuesdays for the $(2+7(n-1))^{\text{th}}$ week, …, Sundays for the $7+7(n-1)=7n^{\text{th}}$ week were chosen for analysis.

Table 4 summarized the results for volume, speed and occupancy at peak hours and entire 24 hours. The peak hours were set as 7:00am~9:00am and 5:00pm~7:00pm, respectively. From Table 5 it is shown that the differences among different days of week are bigger than those among the different traffic variables. For weekdays, the peaks can be divided into two categories. The first category contains Monday and Friday, where the optimal aggregation levels are bigger than those for the second category containing Tuesday, Wednesday and Thursday. This makes senses since normally more uncertain factors will affect the travel behaviors of commuters, which caused the difference of travel characteristics and therefore the optimal aggregation levels between the two categories.

For weekend, the aggregation levels are calculated in 24 hours since normally no obvious peak occurs then. The resulted aggregation levels are around 1~2 hours as shown in the last two columns in Table 4.

TABLE 4 Summary of the Optimal Aggregation Levels for Different Day of Week

Day of week		Morning-peak (min)			Afternoon-peak (min)			24-hour (hr)		
		V	S	O	V	S	O	V	S	O
Weekday	Monday	15	15	16	20	12	24			
	Tuesday	10	9	11	12	11	15			
	Wednesday	16	12	12	12	10	12			
	Thursday	8	13	9	13	11	10			
	Friday	16	11	17	18	15	25			
Weekend	Saturday							1.95	1.95	1.95
	Sunday							1.38	1.38	1.67

Note: V, S, O stand for Volume , Speed and Occupancy, respectively.

4.5.3 Optimal Aggregation Level for Weekly Data and Monthly Data

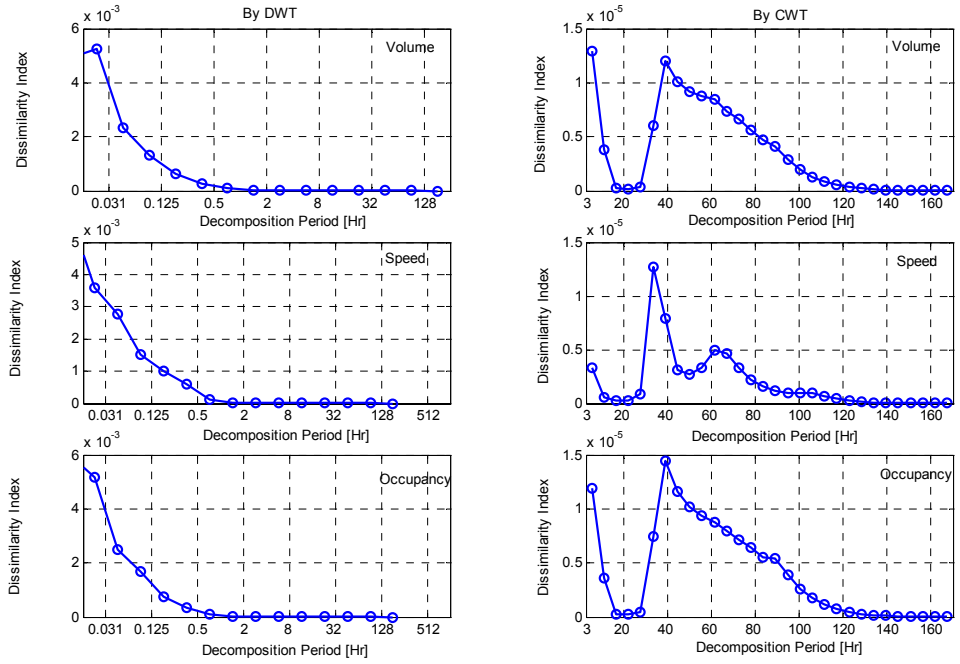
Sometimes the users may be interested in the properties of the relatively long term ITS data, therefore it is necessary to analyze ITS data for a long term. Figure 9 illustrates the relationships between Dissimilarity Index (DI) and Decomposition Period for long term ITS data. The data used for Figure 9(a) were six groups of continuous weekly data starting from September 30, 2001 (Sunday) to November 10, 2001 (Saturday); while the data used for Figure 9(b) were three groups of monthly data, where each four-week was treated as one month, starting from September 30, 2001 (Sunday) to December 22, 2001 (Saturday). All the data were 20-second raw ITS data and start from 12:00 am to 23:00 pm for 23 hours each day.

In the left part of Figure 9(a) it is shown that for the six groups of continuous weekly data, the Dissimilarity Index (DI) obtained by DWT varies smoothly with the Decomposition Period. For all the variables (Volume, Speed and Occupancy), the rapid change part disappeared starting from Decomposition Period as almost 0.5 hour. To find out the detailed variation in the relatively smaller change part, an analysis based on CWT were conducted and the results were shown on the right part of Figure 9(a), where valleys for all the three variables occur when Decomposition Period is around 23 hours, which reflects the daily change of ITS data. So, the aggregation level can be set as 11.5 hours (i.e. half of 23 hours) if the daily change is needed.

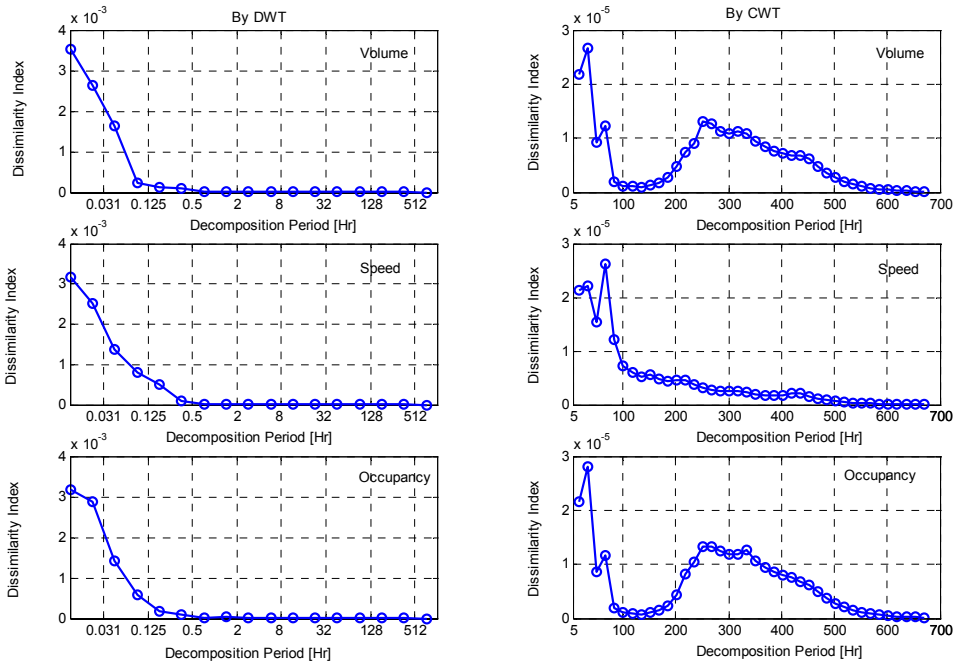
Since the ITS data in each group only contain the weekly data (7 days), the variation period bigger than 3.5 days (84 hours) cannot be caught according to the Shannon Theory, therefore the part of Decomposition Period larger than 84 hours has no meaning.

Figure 9(b) shows the results for other groups of longer term data, which are the three groups of continuous four-week data and can be called as monthly data. The DWT results listed in the left part of Figure 9(b) show the whole pictures of smooth relationships between DI and Decomposition Period for each of the variables. Again, the CWT analyses were conducted for finding out the details in the small change parts, which are listed on the right hand of Figure 9(b). While there is no meaning when Decomposition Period is larger than 322 hours which is the half of four-week span, the valleys that still appear for Decomposition Periods for both Volume and Occupancy are around 160 hours, which is almost the length of a week. Therefore, to keep the information for weekly change from the monthly ITS data, it is suggested to select the aggregation level less than half of 160 hours. However there is no such a valley around 160 hour for Speed; and thus its weekly change is not obvious.

While it is a common sense that the daily and weekly changes exist in the ITS data, the fact that wavelet decomposition can catch these changes cross-validates the effectiveness of the proposed approach.



(a) Analysis on weekly data



(b) Analysis on monthly data

FIGURE 9 Dissimilarity Indices for Volume, Speed and Occupancy by DWT and CWT analysis.

4.5.4 Result Comparison between Wavelet and Statistic Approach

Optimal aggregation levels (in minutes) determined by wavelet analysis and statistics approach CVMSE were compared in Table 3. The rows marked as Speed(1) and Speed(2) are comparable by the both approaches providing hourly optimal aggregation level for Tuesday Speed data. But the dates when ITS data were picked up are different. For CVMSE approach, data were picked up on May 4, 1999 (Tuesday), while data for wavelet approach were from 6 randomly sampled Tuesdays throughout the year 2001.

ITS data Aggregation levels for peak hours are different between these two approaches. In peak hour, optimal aggregation levels by wavelet are bigger than CVMSE, while in non-peak hour the situation reversed. This might because that wavelet has the function to de-noise the signal, which will make the signal more flat to close to the real one. That is why the optimal aggregation level by wavelet is not so small. On the other hand, the wavelet approach is conducted among 6 Tuesdays in 2001 and it has the feature to capture the common parts. Even during the non-peak periods, there may still be some common parts that need to be reserved. This caused the optimal aggregation levels for non-peak hour are less than CVMSE.

CHAPTER 5 DESIGN ARCHITECTURE AND PROTOCOLS FOR DATA ARCHIVING AND RETRIEVAL

This chapter presents an optimization-based sampling approach for data archiving. This approach intends to retain the maximum information of the raw ITS data in the sampled one while minimize the required storage size. The approach is realized through a data processing procedure that is designed to archive real-time/raw data, aggregated data, sampled data, and extension factors which can be generated from the raw data. The proposed approach is tested in the case study of TransGuide of San Antonio, Texas, in which real-time data are collected from 527 loop detectors. After the proposed sampling approach is applied, only one tenth of the original data are needed to be stored, which at the same time is able to meet the potential uses of various transportation purposes.

5.1 Review of Methodologies

A review of existing practices in archiving ITS data by TMCs has identified three major problems: data size, data format, and data quality. The ways that most of TMCs use to store the real-time data are too simple and arbitrary, and most of the data are not archived in a manner that is appropriate for future use by different purposes. The following is a review of more specific methodologies that have been used in archiving the real-time data.

Austin District of Texas stores its detector data to a computer server and can provide data CD to end users. However, analysis of the raw data is very difficult and time consuming. Texas Transportation Institution (TTI) developed a simple approach for dealing with its data: the 1-minute data collected from roadside are archived on CD, aggregated data (*i.e*. 5-, 15-, and 60-minute) can be further processed by using spreadsheets, and data after quality control and re-format could be imported into most spreadsheets.

Seattle, Washington also has a CD-based data archive system for at least past five years operation data. The 20-second field data are summarized to a 5-minute level, and quality control has been done before archiving. One CD can hold three months 5-minute summary data and the data are available upon request.

A sampling approach is used in the case study of TransGuide. The first and the most obvious reason for sampling is to save time, money, and effort. For example, in opinion polls before a general election it would not be practicable to ask the opinion of the whole electorate on how it intends to vote. The second and less obvious reason for sampling is that, even though we have only part of all the information about the data, the sampled data can be useful in drawing conclusions about the whole data, provided that we use an appropriate sampling method and choose an appropriate sampling size. The third reason for sampling applies to the special case in the destructive testing of explosives. Clearly, testing a whole batch of explosives would be ridiculous. The first and second reasons above also apply to the case real-time ITS data.

The sampling approach 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). Data sampling is applicable when it may not be affordable or desirable to archive all disaggregate data. Two examples of data sampling are provided: 1) sampling every "n^{th}" day, and 2) sampling "n" concentrated weeks per "x" months. This approach can save more space in storing the data, and dramatically reduce data size, which makes it easy to manage the data. Approximately only one week's data need to be kept for every two months. While the above sampled method is a very good method, it did not pay attention to the 5% to 25% data losing per day, about half of which are due to detectors failure. In addition, there is no technical guarantee to the quality of the data sampled. It is obvious that quality control is very essential in the process of data sampling, which is not considered by the TransGuide method.

5.2 Proposed Methodology

The proposed optimization-based sampling methodology can be expressed as following. The day-by-day raw data covering the sampling time period (maybe 2-3 months, or any time period that users select) will be first retrieved from the TMC. Then, the quality of these data sets will be examined and a systematic approach will be applied to repair the missing and error data. Subsequently, an approach, based on either sum square error (SSE) or cross validation (CV), will be used to perform the optimization to select the best sampled data. The objective in this whole process is to select the sample of a particular day (e.g. Monday, Tuesday, etc.) of the week that can best represent that particular day of the week for the entire sampling period.

The intended selection of samples could be a particular day, all weekdays, weekends, or even a particular time periods at users' choice. This process will keep rolling when new data flow in, during which the oldest data will be discarded. This process will be able to archive the sampled raw data that best represent the entire data streams, and significantly reduce the required storage spaces. Please note that this approach is essentially different from the aggregation-based archiving in that aggregation will always lose some information in the aggregation process that can never be recovered. However, the proposed approach always retains the original raw data streams, but at selected points. Considering potential needs of raw data for advanced transportation applications in the future, the archiving of the raw data or sampled raw data is definitely necessary.

5.2.1 Data Preparation/Quality Control

The real-time data obtained from the field have significant noises because of imperfections in the collection devices or the failure of communication. Missing and errors in the data could always exist. Hence, it is necessary to prepare the data into a consistent format so that the proposed sampling approach can be implemented. This data preparation procedure, which could also be called Data Quality Control (DQC), can be realized through the following steps:

Step 1: Eliminating the out-of-range data: the incorrect data could be substituted by the threshold value, which is provided by the end user.

Step 2: Filling the missing or incorrect data based on traffic flow curves: for example, if the speed data is missing or incorrect but volume and occupancy data are available, the new speed value will be estimated through the standard traffic flow relationship based on the volume and occupancy.

Step 3: Filling the missing data with the average historic data: if the records of volume, occupancy, and speed are all missing or incorrect for a certain period of time, the mean value of historical data for the same time period will be used to fill the blanks.

Step 4: Filling the missing data using the interpolation method: most of the data should have been fixed after one of the above three steps are completed. In case there are still missing data, interpolation will be implemented to fill the rest of the blanks.

5.2.2 Notation

In order to understand the mathematical expressions in the following sections, a list of important notations for this chapter is provided in the following:

t: record number

n_t: total number of records per day after quality control

w: week day ($w = 1,2,...,n_w$)

k: week number

n_k: total number of weeks in consideration

d: detector number

n_d: total number of detectors

A_i: i^{th} data in consideration for sampling

$\overline{A}_{SSE}$: mean of data in consideration for sampling based on SSE

$\overline{A}_{CV}^{i}$: mean of data in consideration for sampling based on CV for i^{th} data comparison

A_{tdp}^{wk}: t^{th} data record on w^{th} day of k^{th} week for detector d and p^{th} variable (p represents volume, speed or occupancy)

SSE_{dp}^{wk}: total error based on SSE for w^{th} day of k^{th} week for detector d and p^{th} variable (p represents volume, speed or occupancy)

CV_{dp}^{wk}: total error based on CV for w^{th} day of k^{th} week for detector d and p^{th} variable (p represents volume, speed or occupancy)

I_p: evaluation score for the p^{th} variable (p represents volume, speed or occupancy)

α_p: evaluation weight for the p^{th} variable (p represents volume, speed or occupancy)

Err: weighted total score for evaluation

ε_{dp}^{wk}: SSE_{dp}^{wk} or CV_{dp}^{wk}

k_w^{opt}: optimal sample

5.2.3 Description of Sum Square Error (SSE) and Cross Validation (CV)

Mathematically, SSE is to find the target that has the smallest deviation within the sample size by comparing each target with the mean value of the whole data set. Assume A_1, A_2, … , and A_n are the targets to be selected, and $\overline{A}_{SSE}$ is the mean value of these targets. Thus,

$$\overline{A}_{SSE} = \frac{1}{n}\sum_{i=1}^{n} A_i \tag{6}$$

where n is the number of targets. For a single value, the deviation between each target and the mean value will be $A_i - \overline{A}_{SSE}$. For a data set, the square is used to ensure that each value of the deviation is positive. Thus, the trade-off between positive deviation and negative deviation can be avoided when all deviations of targets are added together. Then, the sum of deviations is

$$Error = \sum_{i=1}^{n} (A_i - \overline{A}_{SSE})^2 \tag{7}$$

The comparison of deviations or errors can be conducted to find the best target.

Another methodology for picking the target is called cross validation (CV). The most important difference between CV and SSE is that, in CV, the target will not compare with the mean of all targets. The target is taken out to compare with the mean value of remain targets. Again, assume A_1, A_2, … , and A_n are the targets to be selected. If A_1 will be compared with the mean,

$$\overline{A}_{CV}^{1} = \frac{1}{n-1}\sum_{j=2}^{n} A_j \tag{8}$$

where n is the number of targets. In other word, when A_i is compared, A_i is not included in the mean. So,

$$\overline{A}_{CV}^{i} = \frac{1}{n-1}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} A_j\right) \tag{9}$$

Then, the deviation can be expressed as $A_i - \overline{A}_{CV}^{i}$ (for single value), and the sum of deviations is

$$Error = \sum_{i=1}^{n} (A_i - \overline{A}_{CV}^{i})^2 \tag{10}$$

Similar to SSE, the best target should have the smallest deviation.

To represent the whole data by using part of the data, an appropriate sampling method and sampling size are very important. Another important thing that needs to be mentioned is that the sampling results of SSE and CV could be different. Clearly, CV is a better approach as the bad data has been effectively removed in calculating the mean.

5.2.4 Optimization Based on SSE

Assume the entire sample period for ITS data is n_k weeks, and during each week the weekday w will appear once. The raw data can be reformatted and classified by detectors, and thus d files will be generated. The information within each file could include date, time, detector ID, and the values of volume, speed, and occupancy (a single record is expressed by A_{tdp}^{wk}, where p is the volume, speed, or occupancy). For the method of Sum Square Error (SSE), three different mean values for volume, speed, and occupancy are computed by the following formula.

$$\overline{A}_{tdp}^{w} = \frac{1}{n_k}\sum_{k=1}^{n_k} A_{tdp}^{wk} \quad \forall\ t = 1,2,...,n_t\ ;\ w = 1,2,...,n_w \tag{11}$$

where p *=1, 2 or 3* to represent volume, speed or occupancy. Then, the values of volume, speed, and occupancy of each day at each data record point will be compared with three mean values. The total errors of SSE is as following:

$$SSE_{dp}^{wk} = \sum_{t=1}^{n_t} (A_{tdp}^{wk} - \overline{A}_{tdp}^{wk})^2 \tag{12}$$

The day with the minimum value of SSE_{dp}^{wk} is selected as the best representation of all n_k days.

5.2.5 Optimization Based on CV

The basic idea of cross-validation is to take one value out, and try to compare this value with the mean of other data. A similar technique has been used for choosing the optimal aggregation level by Texas Transportation Institute.

The most important difference between CV and SSE has been discussed earlier. For CV, the mean is not for the data of n_k days. The target day with which the data will be compared will be taken out in calculating the mean. The total errors of CV is expressed as follows:

$$CV_{dp}^{wk} = \sum_{t=1}^{n_t}\left[A_{tdp}^{wk} - \frac{1}{n_k - 1}\left(\sum_{\substack{j=1 \\ j\neq k}}^{n_k} A_{tdp}^{wj} \right) \right]^2 \tag{13}$$

The day with the minimum value of CV_{dp}^{wk} is selected as the best representation of all n_k days.

5.2.6 Selection Approach

During the sampling process, through either SSE or CV, three total errors are provided for volume, speed and occupancy. In order to make the final selection of best samples be made based on any weighted combination of these three variables, an evaluation scoring system is proposed in this sampling approach.

The total errors for different traffic variables (volume, speed, and occupancy) have different scales, it is necessary to unify the real values of the errors into a common scoring system. An evaluation score is used to substitute the real value of the total errors. The weighted total score is expressed as

$$Err = \sum_{p=1}^{3} \alpha_p I_p \tag{14}$$

where I_p is the score for p^{th} variable (could be volume, speed, or occupancy). I_p is determined by the following:

$$I_p = 1 + \left[\frac{\max\left(\varepsilon_{dp}^{wk}\right) - \min\left(\varepsilon_{dp}^{wk}\right)}{(I_{\max} - I_{\min})} \times \left(\varepsilon_{dp}^{wk} - \min\left(\varepsilon_{dp}^{wk}\right)\right) \right] \tag{15}$$

where, $\max\left(\varepsilon_{dp}^{wk}\right)$ is the maximum error within the whole sample group and ε_{dp}^{wk} = SSE_{dp}^{wk} or CV_{dp}^{wk}, and $\min\left(\varepsilon_{dp}^{wk}\right)$ is the minimum error within the whole sample group. ε_{dp}^{wk} is any error within the group (SSE_i^{wk} or CV_i^{wk}) that will be transferred into the score, and p could represents volume, speed, or occupancy. $I_{\max}$ and $I_{\min}$ can be set as 10 and 1 respectively in the 1-10 scale evaluation system.

As mentioned earlier, α_p is the weight for each traffic variable, and $\sum_{p=1}^{3} \alpha_p = 1$, where $0 \le \alpha_p \le 1$. Thus, the weight is used to reflect the level of importance of each traffic variable in selecting the best sampled data. It is noted that when α_p=1, the best sampled data for a single variable will be selected. Now, the final selection of the best samples can be made as:

$$k_w^{opt} = \arg\min\left(\sum_{p=1}^{3} \alpha_p I_p\right) \tag{16}$$

where, k_w^{opt} is the best day whose data can represent the entire sampling period.

5.3 Implementation

A computer program is developed to test the proposed optimization-based sampling approach by using MATLAB. The program consists of two major stages. The first stage performs the initial processing, which formats the data and generates the output file for each detector. Step 1 and Step 2 of quality control are also implemented in this program. The output from the first stage becomes the input for the second stage. Once all data in the sampling period are processed, the second stage is entered. Within this stage, Step 3 and Step 4 of data quality control will be implemented. Then a sampling method, either sum square error or cross validation, is carried out. Figure 10 shows the flowchart of the program.

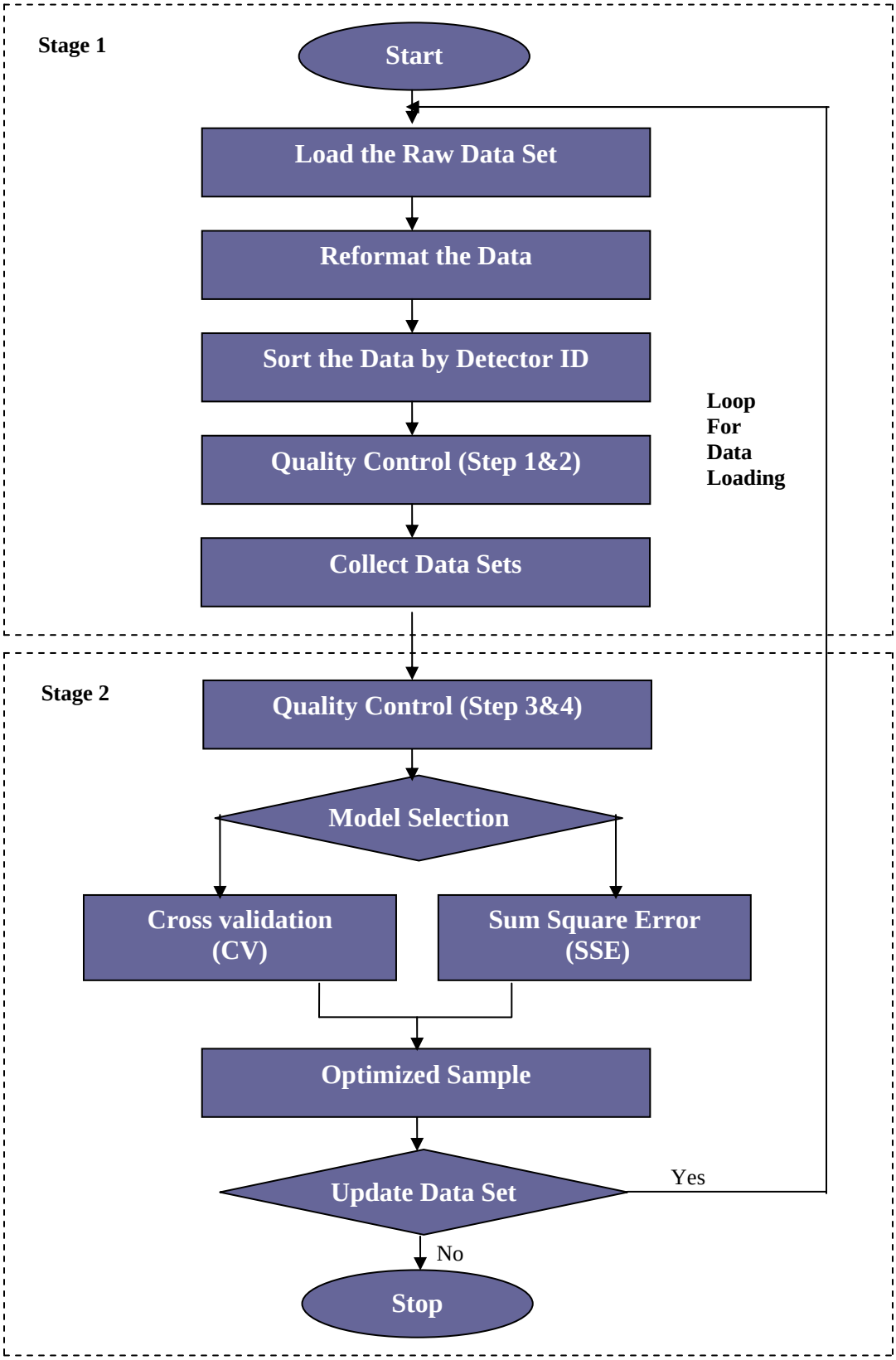


FIGURE 10 Sampling and optimization flowchart.

5.4 Case Study

5.4.1 Data Preparation

In order to examine the proposed sampling approach, a field study was conducted based on the data acquired from San Antonio TransGuide. There are 527 Loop Detectors and about 120 MB data will be generated everyday on the 20-second interval, which means that there will be 4,320 counts per day for each detector. These counts mainly include the traffic data about volume, speed, and occupancy. The data are saved in the text file. About two million counts from all of the 527 Loop Detectors will be generated each day.

The data for the test include ten weeks' data for the period from April/01, 2001 (Sunday) to June/09, 2001 (Saturday). It is noted that the approach does not limit the number of sampling weeks to ten. The objective is to select a particular weekday (*e.g*. Tuesday) during this 10-week period that can best represent all of the particular weekdays during these ten weeks. In this way, only one-tenth of all data is needed to be archived.

Data from two detectors were selected for the test. One detector is with the ID as EN1-0010W-572.059, where volume and occupancy data are available while speed data are all marked as "-1", which is a symbol of the missing data. Another detector used for test is with the ID as L2-0035N-152.005, where data are available for all the three variables.

Following the four steps of Data Quality Control (DQC), all the missing and error data (till 23:00) were substitute by the threshold (Step 1), flow curve fitting (Step 2), historic data (Step 3) and interpolation data (Step 4). After whole DQC process, all the missing and error data were given the suitable value. For detector L2-0035N-152.005, the rates of missing and error data for all variables along the testing period range from 6% to 13%, about 1.9% to 2.4% of the missing and error data were repaired in Step 1 and Step 2, while 4.1% to 10.6% in Step 3. In this case, none was fixed in Step 4. Similar results were obtained for detector EN1-0010W-572.059. The total rates for missing and error data at this detector range also from 6% to 13%. 1.7% to 2.3% can be fixed in Step 1 and Step 2, while all the rests were all fixed in Step 3. Still, none were fixed in Step 4.

5.4.2 Comparison of Results based on SSE and CV

After the data were prepared and passed the quality control, the sampling errors for both SSE and CV were conducted afterwards. Table 5 lists the total sampling errors for volume, occupancy, and speed based on both CV and SSE. Data for volume and occupancy are from detector EN1-0010W-572.059, while data for speed from detector L2-0035N-152.005. In Table 5 it is shown that the errors based on SSE and CV are different.

Also it is shown in Table 5 that the places for minimum error based on SSE may are located in the same column (i.e. the same day) for volume and occupancy. The minimum errors are all in day 7. However, it is not the case for speed data, where the minimum error based on CV is in day 3, while that on SSE is in day 9.

TABLE 5 Results of Errors for Volume, Occupancy, and Speed based on SSE and CV

Day#	Volume (1.0e+009)		Occupancy (1.0e+004)		Speed (1.0e+007)	
	SSE	CV	SSE	CV	SSE	CV
1	1.1416	1.4094	3.0708	3.7911	0.1101	5.5292
2	1.1804	1.4573	3.5830	4.4234	0.0707	5.6106
3	1.2706	1.5687	3.3445	4.1290	0.0707	**5.3457**
4	1.2202	1.5065	3.1896	3.9378	0.0690	5.4974
5	1.2001	1.4816	3.2622	4.0274	0.0725	5.5351
6	1.2405	1.5315	3.2315	3.9895	0.0691	5.3599
7	**1.0257**	**1.2663**	**2.4296**	**2.9996**	0.0616	6.1481
8	1.2317	1.5206	3.2645	4.0302	0.0723	5.7414
9	1.1312	1.3965	2.4330	3.0037	**0.0588**	6.8866
10	1.3076	1.6143	3.1585	3.8994	0.0663	6.3169

Note: Data of volume and occupancy from detector EN1-0010W-572.059
Data of speed from detector L2-0035N-152.005

To better compare the results in a more consistent way, the computed errors were transferred into scores by equation (15). Figure 11 visualizes the results based on both CV and SSE for speed data. Y-axis indicates scores, and X-axis indicates day numbers. In Figure 2, the differences of the scores based on SSE and CV are significant in some days (day 1, 7, 9 and 10), while is very small occasionally (day 2). The optimal day based on the both approaches are quite different in Figure 11, the optimal day is day 9 based on SSE and is day 3 based on CV.

In order to ascertain which result is more reliable, the real speed data for the two days analyzed (day3 and day 9) are plotted in Figure 12. Figure 12 shows the speed dispersion shape from 5:00 AM to 10:00 AM of the two different days. It is obvious that in Figure 12, the variation of speed in day 3 is stronger than that in days 9 during the peak hour (7:00 – 8:00). Actually, the typical speed profile being similar to that for day 3 is widely used in speed analysis. For example in the *guidelines for developing ITS data archiving systems*, this kind of speed profile is illustrated for analyzing the optimal aggregation level. So, it is perceptibly that the variation in day 3 is more acceptable. It seems that the result based on CV is better. Based on theory, cross-validation is one of the main advantages of most automatic schemes of optimization.

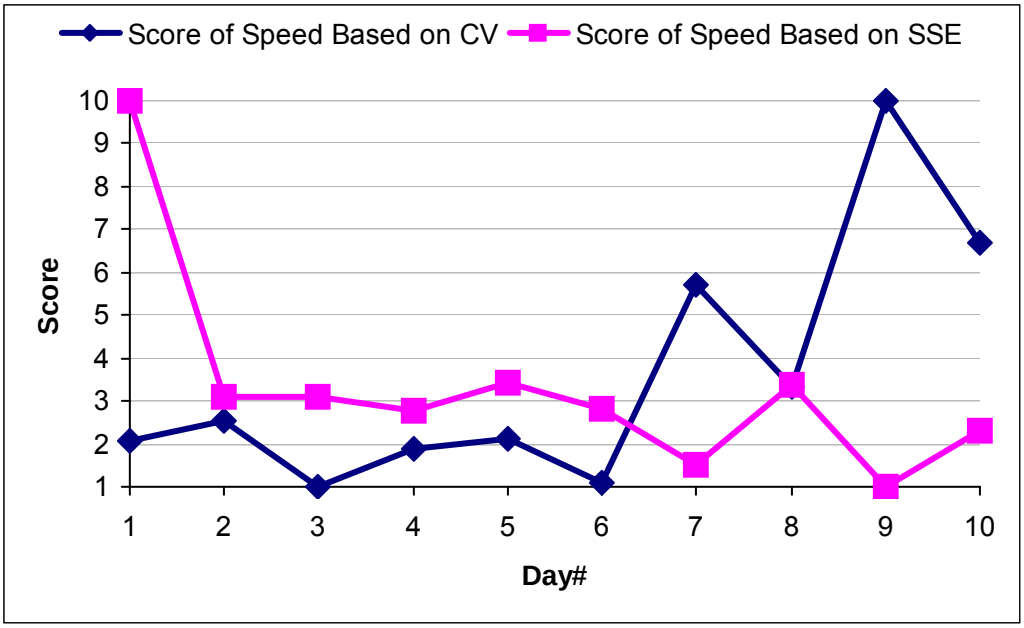
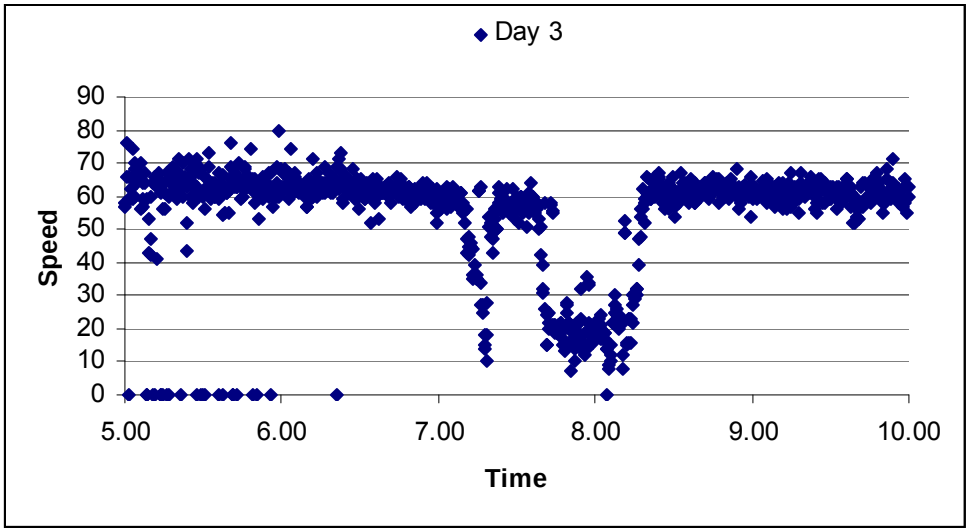
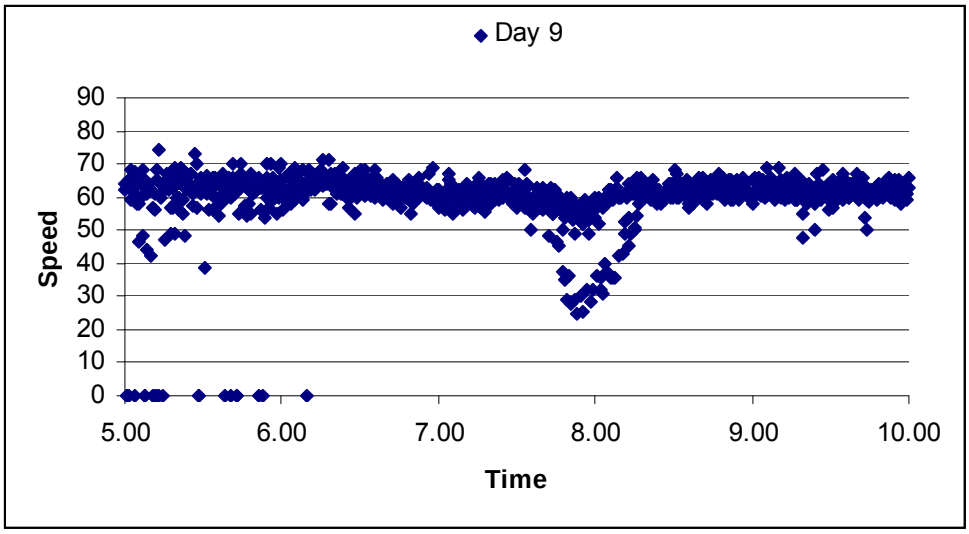


FIGURE 11 Example of different results of speed based on CV and SSE.



(a) For day 3.



(b) For day 9.

FIGURE 12 Real-time speed comparison of day 3 and day 9.

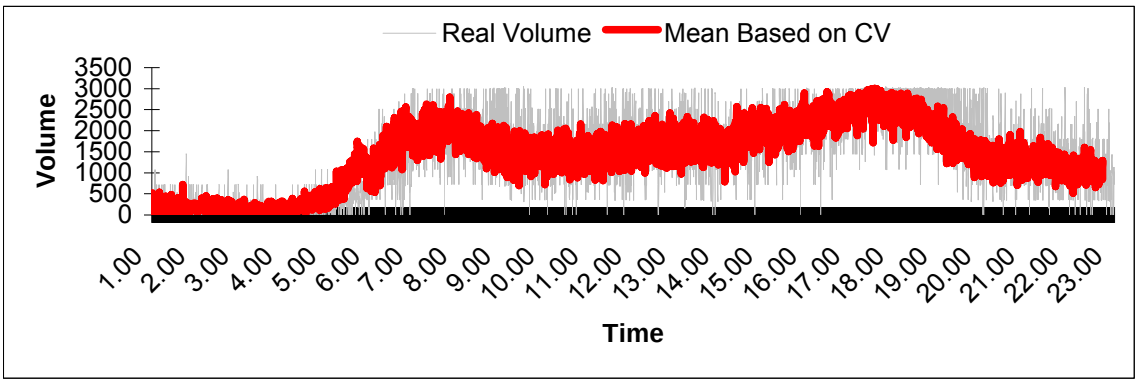
In this case study, results of volume and occupancy obtained from CV and SSE are very close. In other words, the two methods provide similar results regarding which day could be the best representation of other days of the same weekday. As discussed earlier, the results from the two methods could be different. However, the smaller the differences between sample sets, the closer the results from CV and SSE. Therefore, it is conclude that the volume and occupancy data used for the test did not include any bad day.

5.4.3 Selection of optimal date

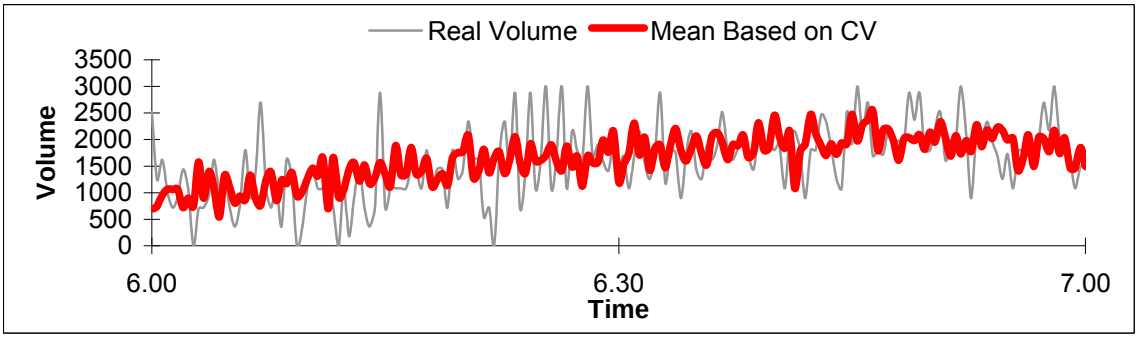
Figure 13 visualizes the cross validation results of one day (May/15, 2001) for detector EN1-0010W-572.059. In Figure 13, the top one displays the whole day real volume and the mean value of other days. The dark line in the middle is the mean value. To give some detailed visual results, the comparisons between real volume and mean of CV for the morning peak (6-7 AM) and non-peak (12-13 PM) are given. In each chart, the dark line is the mean based on CV, and the light line is the real data.

For the weighted combination of traffic variables (volume, speed and occupancy), the optimized sampling results may be different from those based on a single traffic variable. Table 6 and Figure 14 show the sampling results for either single or combined traffic variables based on CV. In this example, volume, speed and occupancy data are used, and the data come from Tuesday.

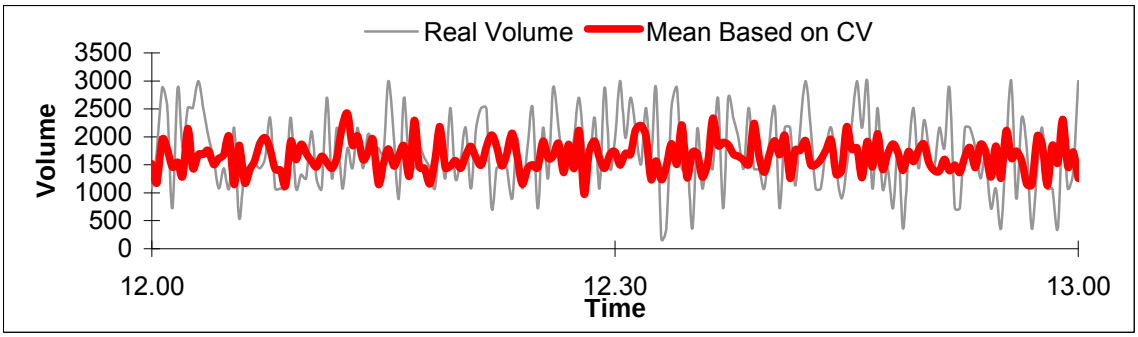
While the best day based on volume and occupancy is the seventh day (May/15, 2001), and the best day based on speed is the third day (Apr/17, 2001). The weighted combination results are also different. When the weights of volume, speed and occupancy are set as 1/3, 1/3, and 1/3 respectively, the weighted combination shows the fourth day is the best sample day; when the weights of volume, speed and occupancy are set as 1/2, 1/4, and 1/4, respectively, the seventh day becomes the best sample day. This example does show the differences that possibly occur when different variables are used.



(a) Data for one day.



(b) Data for morning peak 6:00-7:00.



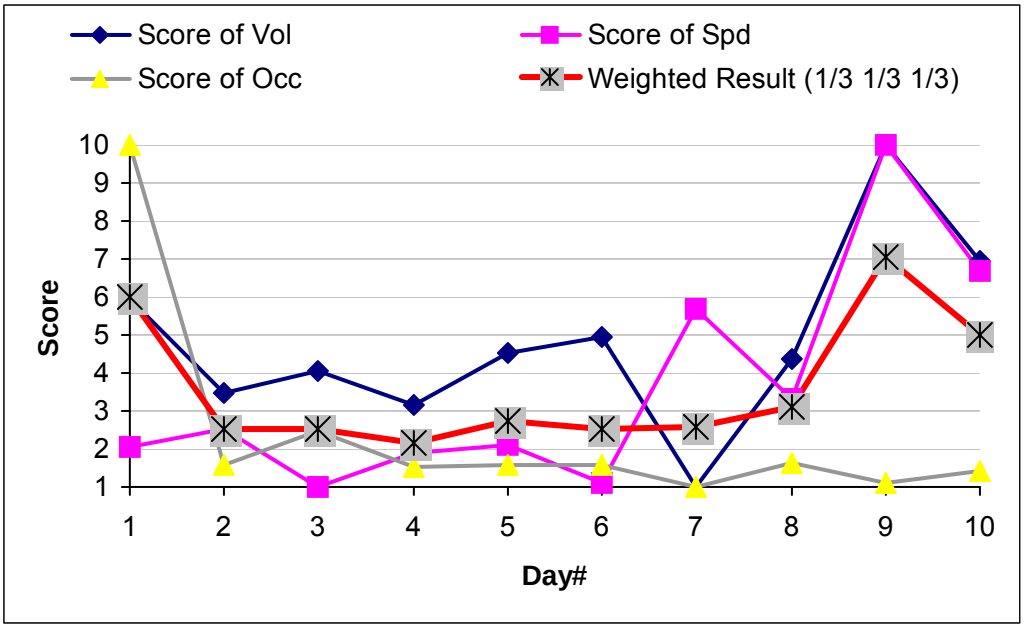
(c) Data for non-peak 12:00-13:00.

FIGURE 13 Volume and the mean value of May 15, 2001 for detector EN1-0010W-572.059.

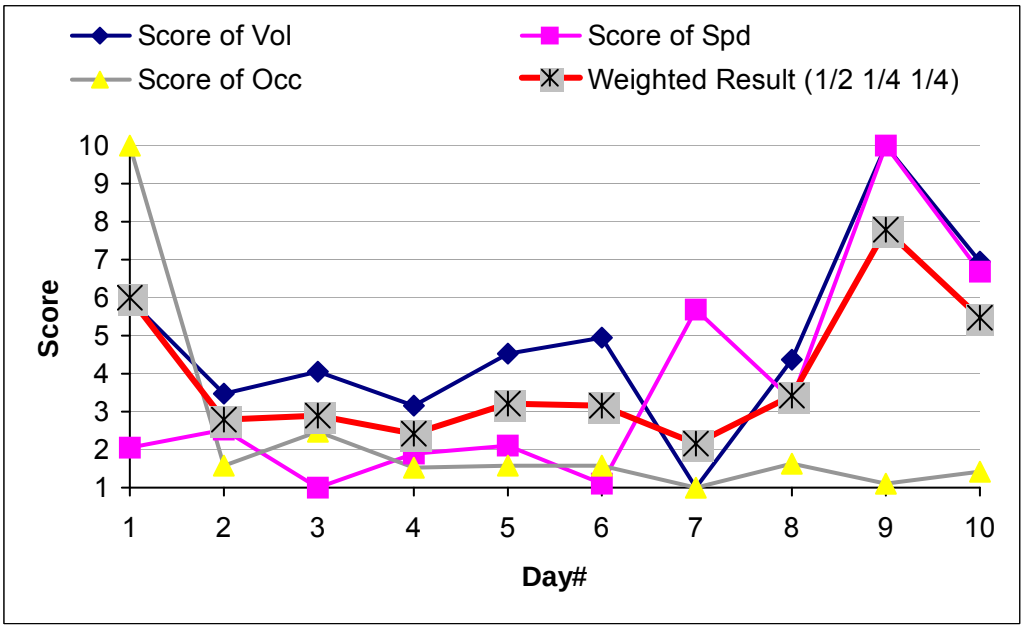
TABLE 6 Weighted Scores for Volume, Speed and Occupancy

Day#	Volume	Occupancy	Speed	Weighted Sum (1/3 1/3 1/3)	Weighted Sum (1/2 1/4 1/4)
1	5.9730	10.0000	2.0718	6.0149	6.0044
2	3.4814	1.5948	2.5476	2.5412	2.7763
3	4.0589	2.4602	1.0000	2.5064	2.8945
4	3.1419	1.5087	1.8863	**2.1789**	2.4197
5	4.5456	1.6048	2.1062	2.7522	3.2005
6	4.9431	1.5966	1.0830	2.5409	3.1414
7	1.0000	1.0000	5.6867	2.5622	**2.1717**
8	4.3594	1.6400	3.3115	3.1036	3.4176
9	10.0000	1.1169	10.0000	7.0390	7.7792
10	6.9614	1.4044	6.6722	5.0127	5.4999

Note: Data are from detector L2-0035N-152.005



(a) For weights as 1/3, 1/3 and 1/3.



(b) For weights as ½, ½, and ¼.

FIGURE 14 Sample results of weighted selection approach.

Table 7 illustrates the suggested sample day for each day in the week within the sample period based on weighted results of all traffic variables of detector L2-0035N-152.005. The best sample day of each weekday is marked by the date and the method(s) used. A virtual week, which is not the nature week, can be made up by

following this table. Data of each weekday within this combined week have good quality and value and the weekly features are also maintained. There are two days, Friday and Saturday, which have same results based on SSE and CV.

TABLE 7 Suggested Sample Day of Each Week Based on Volume, Speed, and Occupancy

Week#	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1				Apr/04 CV			
2							
3		Apr/16 SSE			Apr/19 CV		
4		Apr/23 CV	Apr/24 CV				
5	*						
6				May/09 SSE	May/10 SSE		
7	May/13 SSE		May/15 SSE				May/19 SSE, CV
8	May/20 CV						
9							
10						June/08 SSE, CV	

*Note: Data is not available.
Data from detector L2-0035N-152.005

In a whole, the test of the proposed approach using the TransGuide data demonstrates that the approach is very practical and can be implemented easily and satisfactorily for the real-world data. The approach realizes the objective in archiving ITS data, which is to save as less data as possible, and keep the maximum information in the original data streams.

CHAPTER 6 CONCLUSIONS

In this research report, the state-of-the-art and the state-of-the-practice related to the project were reviewed first. The research in the report has focused on the determination of appropriated aggregation level, and on the developing of the approaches for ITS data archiving.

In this research, the newly developed wavelet transformation was successfully used in the analysis of real time ITS data and in the obtaining of their aggregation level. The result for aggregation level can not only capture the sufficient information in the signal, but also eliminate the undesired components and noise.

Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, such as trends, breakdown points, discontinuities in higher derivatives, and self-similarity. Furthermore, because it affords a different view of data than those presented by traditional techniques, wavelet analysis can often compress or de-noise a signal without appreciable degradation.

In order to let the wavelet technology be better applied to the optimization of aggregation level of ITS data, the optimal technique was developed. This optimization technique is based on the decomposition of archived ITS data sets, and on the comparison of the similarities of interested sets of signals. The dissimilarity index is used for quantify the dissimilarity (and therefore the similarity) among different sets of signals. The optimization criteria were set as the curve fitted for the dissimilarity trends attenuates to a certain predefined critical parameter. This optimization procedure, which was tested by the 2001 San Antonio TransGuide® ITS data, can be applied to any place where ITS real time data is available. As illustrated in the case study, optimal aggregation levels for any time periods can be determined, even in dynamic situation.

The case study for long term ITS data sets cross validates the effectiveness of the proposed approach. However, since the wavelet transformation is a new and novel approach, validations under various conditions and different data sources as well as detailed implementation procedures should be further conducted to put this approach into real applications.

As for the archiving of ITS real-time data, an appropriate sampling approach can ensure the accuracy and safety of the sampling results. This research proposed an optimization-based sampling approach for ITS data, which can dramatically reduce the data size while still be able to archive the best representative raw data.

The four-step Data Quality Control process included in this approach ensures the quality of the archived data. The missing and error data will be substituted by the threshold value, traffic flow curve fit value, average historic data and purely mathematical interpolation data.

Sum Square Error (SSE) and Cross Validation (CV) are two optimization approaches used to determine the best samples while keep as much information of the original data as possible. For SSE optimization ITS data sets are compared with the total mean of all sets, with the sum square errors among them being calculated. For CV optimization, each ITS data set is compared with the mean of the remaining sets, the sum square errors of which is then measured. For both approaches, the day with the minimum errors are considered as the optimal sample.

Since normally more than one traffic variable is provided by Traffic Management Centers, different variables may have different optimal samples. Considering the case that there might be a need to provide a general optimization result among all the different variables, a selection process is developed to transfer all the results in errors into the so-called corresponding scores while the best sample is chosen based on their weighted sum. The transferred scores are marked from 1 (the best case) to 10 (the worst case), and the weights are flexible and subject to change so as to match the various needs from users.

To demonstrate the proposed approach, the TransGuide data were used for testing. Among the 10 weeks 20-second ITS data from 2 of the 527 detectors, including volume, speed and occupancy, all the 6% ~ 13% missing and error data were made up through the four-step process of Data Quality Control. In particular case, 100% of the missing data were repaired in the first three steps with none of them repaired in Step 4 (interpolation).

Optimal results of sample day based on SSE and CV were compared and found that CV might be more reliable if the variances of ITS data sets among different days are strong. Testing results show that different optimal samples could be obtained for different traffic variables and different weight. The users may grant their particular preferences to the optimization process to emphasize which one or two of the three variables are favored.

Through the whole optimal process of sample selection, only one virtual week is selected for storing with nine tenth of the original size was saved.

Therefore, this proposed optimization-based sampling approach is reliable to provide high value data for kinds of transportation purposes. It is not the preferred solution of ITS data archiving, but it is sort of useful. Even though only part of all the data is archived, the sampled data can be useful in drawing conclusions about the whole data.

References

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

Abry, P. (1997), Ondelettes et turbulence. Multirésolutions, algorithmes de décomposition, *invariance d'échelles*, Diderot Editeur, Paris.

Albert, L., P. Irwin, and C. Zimmerman. An Evaluation of the Potential of Public/Private Partnerships for the Management of Archived ITS Data. Prepared for *CVEN 677*, 1999.

Archived Data User Service (ADUS), An Addendum to the ITS Program Plan, September 1998. http://www.itsdocs.fhwa.dot.gov/jpodocs/repts_pr/41401!.HTM. Accessed May 10, 2001.

B. L. Smith, and R.E. Turochy. Trends in Traffic Management Staffing Levels. *Proceedings of the 10th Annual Meeting, Intelligent Transportation Society of America* (CD-ROM), Washington, D.C., 1999.

Booz Allen, and Hamilton. Emissions Management Using ITS Technology. *Prepared for U.S. Department of Transportation Federal Highway Administration* Washington, D.C., September 1998.

Brian L. Smith, and Billy M. Williams. ITS Data-Tapping the Resource for System Operation. *A paper presented and published at the ITS America 9th Annual Meeting,* Washington, D.C., 1999.

Byron Gajewski, Shawn Turner, William Eisele, and Clifford Spiegleman. ITS Data Archiving: Statistical Techniques For Determining Optimal Aggregation Widths For Inductance Loop Detector, Presented At *The 79th Annual Meeting Transportation Research Board,* Washington, D.C. Jan 2000.

C. Marven and G. Ewers, *A Simple Approach to Digital Signal Processing*, John Wiley & Sons, Inc. © 1996.

Cambridge Systematic, Inc. *Strategic Plan for the Development of Adus Standards*. May 5, 2000.

Gajewski B., Turner S., Eisele W. and Spiegleman C., ITS Data Archiving: Statistical Techniques for Determining Optimal Aggregation Widths for Inductance Loop Detector, Presented at the 79th *Annual Meeting Transportation Research Board*, Paper No. 00-1361, Washington, D.C., January 2000.

Ghosh, M., and G. Meeden. *Bayesian Methods for Finite Population Sampling*, Chapman & Hall, 1997.

J. Zietsman, and L.R. Rilett. A Comparison of Aggregate and Disaggregate Based Travel Time Estimation for Sustainability and ATIS Systems Applications. *Presented at the 79th Annual Meeting Transportation Research Board* Washington, D.C., Jan 2000.

Keiser, G., *Optical Fiber Communications*, McGraw-Hill Companies, Inc. 2000.

Kikuchi, S., and D. Miljkovic. Method to Preprocess Observed Traffic Data for Consistency: Application of Fuzzy Optimization Concept. *Transportation Research Record*, No. 1679, 1999, pp. 73-80.

Lockheed Martin Federal Systems, Odetics Intelligent Transportation Systems Division. ITS Mission Definition. *Prepared for Federal Highway Administration US Department of Transportation* Washington, D.C. 20590, December 1999.

Luke P. Albert, Patrick L. Irwin, and Carol A. Zimmerman. AN EVALUATION OF THE POTENTIAL OF PUBLIC/PRIVATE PARTNERSHIPS FOR THE MANAGEMENT OF ARCHIVED ITS DATA. *Prepared for CVEN 677 Advanced Surface Transportation Systems*, August 1999.

Malay Ghosh & Glen Meeden. Bayesian Methods for Finite Population Sampling, Chapman & Hall, 1997.

Margiotta R. ITS AS A DATA RESOURCE—Preliminary Requirements for a User Service. *Prepared for Federal Highway Administration Office of Highway Information Management*, April 1998.

Marven C. and Ewers G., *A Simple Approach to Digital Signal Processing*, John Wiley & Sons, Inc. © 1996.

Misiti M., Misiti Y., Oppenheim G. and Poggi J. - M.. *Wavelet Toolbox for Use with MATLAB: User's Guide Version 2.1*. The Math Works, Inc. Natick, MA. June 2001.

Mitretek Systems Inc. Intelligent Transportation Systems Benefits: 1999 Update. *Under Contract to the Federal Highway Administration United States Department of Transportation* Washington, D.C., 28 May 1999.

Percival, D. B., and Walden A. T.. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge CB2 CRU, UK, 2000.

Porat B., *A Course in Digital Signal Processing*, New York: John Wiley, 1997.

Qiao, F., Yu L., and Wang X.. Determining Aggregation Level for ITS Data via Wavelet Transportation. Proceedings of the 12th Annual Meeting (CD-ROM). Long Beach, CA, April 29 – May 2, 2002.

Rod E. Turochy, and Brian L. Smith. SOME OF THE MANY USES OF INFORMATION GENERATED FROM ARCHIVED TRAFFIC DATA. *Proceedings of the 10th Annual Meeting, Intelligent Transportation Society of America* (CD-ROM), Washington, D.C., 2000.

Sampath, S. *Sampling Theory and Methods*. Narosa Publishing House, 2001.

Smith, S. *Integrating Intelligent Transportation systems within the Planning Process: An Interim Handbook*. Report No. FHWA-SA-98-048, Federal Highway Administration Washington, D.C. 20590, Jan 1998.

Souleyrette, R., D. Plazak, T. Strauss, and S. Andrle. Applications of State Employment Data to Transportation Planning. *Transportation Research Record*, No. 1768, 2001, pp. 26-35.

Southwest Research Institute. Prototype Implementation of ITS Technology for Data Collection and Transportation Planning. SwRI Project No. 10-03997. *Prepared for HGAC*, Houston, Texas, April 12, 2001.

Southwest Research Institute. Recommended Prioritization for Addressing Identified TranStar Data Warehouse User Data Needs. *Prepared for Texas Department of Transportation,* December 7, 2000.

Southwest Research Institute. TranStar Data Warehouse User Data Needs Identification. *For The Texas Department of Transportation TxDOT ITS Statewide Development and Integration Project.*

Tokuyama H. *Intelligent Transportation Systems in Japan*, http://www.tfhrc.gov/ Access: July, 2002.

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

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

Turner S. M., Guidelines for Developing ITS Data Archiving Systems, Report 2127-3, Project Number 0-2127, Sponsored by the Texas Department of Transportation in cooperation with the U.S. Department of Transportation Federal Highway Administration.

Turner, S. A Simple Approach to Archiving Operations Data: Case Study in Austin, Texas. *TRB 2002 Annual Meeting* (CD-ROM), Jan 2002.

Turner, S. *Guidelines for Developing ITS Data Archiving Systems*. Report 2127-3. FHWA, U.S. Department of Transportation, 2001.

Turner, S., L. Albert, B. Gajewski, and W. Eisele. Archived Intelligent Transportation System Data Quality: Preliminary Analyses of San Antonio TransGuide Data. *Transportation Research Record*, No. 1719, 2000, pp. 77-84.

Turner, S., L. Albert, B. Gajewski, R. Benz, and W. Eisele. *ITS Data Archiving: Case Study Analyses of San Antonio TRANSGUIDE Data*. Report No. FHWA-PL-99-024. FHWA, U.S. Department of Transportation, 1999.

U. Barnett. Professor of statistics, University of Sheffield and Rothamsted Experimental Station. Sample Survey Principles & Methods, Edward Arnold a division of Hodder & Stoughton, 1991.

Using ITS-Derived Data for Transportation Planning, Programming, and Operations, Aug 11, 1998. http://www.itsa.org/. Accessed May 8, 2001.