1. Report No. **428**	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle **Development of a Statewide Transportation Data Warehousing and Mining System under the Louisiana Transportation Information System (LATIS) Program**	5. Report Date **June 2008** 6. Performing Organization Code	
7. Author(s) **Bill P. Buckles, Sherif Ishak, and Stephanie Smith**	8. Performing Organization Report No.	
9. Performing Organization Name and Address **Dept. of Electrical Engineering & Computer Science** **Tulane University** **New Orleans, LA 70118**	10. Work Unit No. 11. Contract or Grant No. **04-1SS**	
12. Sponsoring Agency Name and Address **Louisiana Transportation Research Center** **4101 Gourrier** **Baton Rouge, Louisiana 70808**	13. Type of Report and Period Covered **Final Report, 10/01/03 – 08/30/06** 14. Sponsoring Agency Code	

15. Supplementary Notes

16. Abstract

More jurisdictions including states and metropolitan areas are establishing traffic management centers to assist in reducing congestion. To a lesser extent, these centers are helpful in providing information that assists engineers in making such adjustments as signal synchronization or road improvements. However, the main traffic management center function is real-time decision making for freeways and surface streets.

Planning and modification of a traffic network is best pursued in a context that includes large historical data. Understanding traffic "behavioral" properties is the domain of numerous technologies such as data mining which depends on the presence of large amounts of historical data. To these ends, the Louisiana Transportation Research Center (LTRC) commissioned this study for the design of a data warehousing/data mining system that, while limited to Baton Rouge, will serve as a statewide model.
Few traffic-oriented data warehouses exist in the U.S. The methodology employed in designing the Baton Rouge, Louisiana warehouse included visiting many of them and collecting sample data from Baton Rouge sensors. Advisory and stakeholder committees were formed to give advice on the base applications. Base applications are the ones recommended for the initial inclusion in the warehouse. The applications were traced back to the data, resulting in some of them being dropped or modified to suit the data and its quality that was available. From that juncture, the data was tracked forward again to the applications, modeling the transformations necessary. This transformation set constituted the design. Chosen applications, in addition to data mining, included several variations of performance measuring and hydrowatch. The latter is unique among traffic warehouses and is particularly appropriate for the region and State.

The data warehouse design consists of a system with three stages – extraction/transformation/loading (ETL) of source data, main storage of the warehouse, and client workstation software. ETL consists of acquiring, cleansing, formatting, merging, and purging of the source data. Much of this stage entails data quality checking and the report addresses this aspect. Main storage is organized around a star schema also called a multidimensional data cube. This separates the static data such as sensor location from dynamic data such as lane occupancy. This design assumes a one-way data flow, input from the sensors and output to client workstations or other media. This approach is codified commercially as online analytical processing (OLAP). Many warehouses stop at the main storage phase (the "dump and run" model). Here, the solution to key client phase issues, interfaces to GIS and linear referencing are given.

Infrastructure and a marketing plan are given. The key infrastructure decision is determining the warehouse's physical location – at LTRC, at a university, or at a private/public concern. All three variations were discovered in site surveys. Marketing consists of addressing various segments beginning with those who are inclined to add value to the system. These are the engineers and planners but at some stage university researchers and the general public must be given access and be convinced of the value that can be obtained.

17. Key Words	18. Distribution Statement
Traffic data warehouse, Intelligent transportation systems, Data mining, Databases, Archived data management systems	**Unrestricted. This document is available through the National Technical Information Service, Springfield, VA 21161.**

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages **140**	22. Price
Not Applicable			

Project Review Committee

Each research project will have an advisory committee appointed by the LTRC Director. The Project Review Committee is responsible for assisting the LTRC Administrator or Manager in the development of acceptable research problem statements, requests for proposals, review of research proposals, oversight of approved research projects, and implementation of findings.

LTRC appreciates the dedication of the following Project Review Committee Members in guiding this research study to fruition.

LTRC Administrator/ Manager
Chester Wilmot
Planning/Intermodal Research Manager

Members
Peter Allain
Michael Boudreaux
Dom Cali
John Collins
Huey Dugas
Stephen Glascock
Kay Henderson
James E. Mitchell

Directorate Implementation Sponsor
William B. Temple

Development of a Statewide Transportation Data Warehousing and Mining System under the Louisiana Transportation Information System (LATIS) Program

Bill P. Buckles* and Stephanie Smith*
Sherif Ishak**

*Dept. of Electrical Engineering and Computer Science
Tulane University
New Orleans, LA 70118

**Dept. of Civil and Environmental Engineering
Louisiana State University
Baton Rouge, LA 70808

LTRC Project No. 04-1SS
State Project No. 736-99-1219

conducted for
Louisiana Department of Transportation and Development
Louisiana Transportation Research Center

June 2008

ABSTRACT

More jurisdictions including states and metropolitan areas are establishing traffic management centers to assist in reducing congestion. To a lesser extent, these centers are helpful in providing information that assists engineers in making such adjustments as signal synchronization or road improvements. However, the main traffic management center function is real-time decision making for freeways and surface streets.

Planning and modification of a traffic network is best pursued in a context that includes large historical data. Understanding traffic "behavioral" properties is the domain of numerous technologies such as data mining which depends on the presence of large amounts of historical data. To these ends, the Louisiana Transportation Research Center (LTRC) commissioned this study for the design of a data warehousing/data mining system that, while limited to Baton Rouge, will serve as a statewide model.

Few traffic-oriented data warehouses exist in the U.S. The methodology employed in designing the Baton Rouge, Louisiana warehouse included visiting many of them and collecting sample data from Baton Rouge sensors. Advisory and stakeholder committees were formed to give advice on the base applications. Base applications are the ones recommended for the initial inclusion in the warehouse. The applications were traced back to the data, resulting in some of them being dropped or modified to suit the data and its quality that was available. From that juncture, the data was tracked forward again to the applications, modeling the transformations necessary. This transformation set constituted the design. Chosen applications, in addition to data mining, included several variations of performance measuring and hydrowatch. The latter is unique among traffic warehouses and is particularly appropriate for the region and State.

The data warehouse design consists of a system with three stages – extraction/transformation/loading (ETL) of source data, main storage of the warehouse, and client workstation software. ETL consists of acquiring, cleansing, formatting, merging, and purging of the source data. Much of this stage entails data quality checking and the report addresses this aspect. Main storage is organized around a star schema also called a multidimensional data cube. This separates the static data such as sensor location from dynamic data such as lane occupancy. This design assumes a one-way data flow, input from the sensors and output to client workstations or other media. This approach is codified commercially as online analytical processing (OLAP). Many warehouses stop at the main storage phase (the "dump and run" model). Here, the solution to key client phase issues, interfaces to GIS and linear referencing are given.

Infrastructure and a marketing plan are given. The key infrastructure decision is determining the warehouse's physical location – at LTRC, at a university, or at a private/public concern. All three variations were discovered in site surveys. Marketing consists of addressing various segments beginning with those who are inclined to add value to the system. These are the engineers and planners but at some stage university researchers and the general public must be given access and be convinced of the value that can be obtained.

ACKNOWLEDGMENTS

We are indebted to Chester Wilmot of the Louisiana Transportation Research Center for his guidance throughout the performance of this study. Likewise, Huey Dugas, Chief of Planning Capital Region Planning Commission, was always available with documentation and introductions to key personnel responsible for traffic management in Baton Rouge. It was Mr. Dugas who directed us toward performance measures as a valuable application and arranged interviews with originators of key traffic congestion metrics. Stephen Glascock freely provided the services of those people of the Traffic Management Center including Carryn Zeagler, Lucy Kimberly, Ingolf Partenheimer, George Gele, and Peter Allain.

People from other geographic regions were more than helpful. These included Johnny Bordelon of the New Orleans Regional Planning Commission, Tim Lomax of the Texas Transportation Institute, Brian Smith of Virginia's Smart Travel Lab, and Mark Hallenbeck of TRAC in the Seattle area.

Finally, we thank Kun Zhang and Marco Carvalho, Tulane graduate students, who not only did most of the experiments over sample data but a considerable part of the writing of this report.

IMPLEMENTATION STATEMENT

The first step in implementing the recommendations of this report is to determine the owner and location of the Baton Rouge traffic data warehouse. Options include a university such as LSU, directly within the Louisiana Transportation Research Center, another entity of LADOT, the Baton Rouge Traffic Management Center, or another entity of the State or metropolitan government. The next step should be the appointment of an internal committee empowered to make financial decisions but augmented with representatives from various traffic planning and management constituencies.

From this point, there are two possible directions. One is the preparation of a bid document based on this report. Following the bidding process, or perhaps concurrent with it, this document should be provided. The second approach is to procure the main storage component from an existing traffic management center. The principal candidate would be California's Performance Measurement Systems (PeMS). In this latter case, the main storage structures must be modified to reflect the applications envisioned in this document. Following such procurement, the acquisition stage and client station applications could be developed separately. In either case, the performance measures should be given the highest implementation priority.

The staffing of the warehouse should be carefully considered. It is estimated that only four or fewer persons are needed on a permanent basis. These should include database analyst and two computer programmers together with a technician. While traffic engineers have a significant presence, they are infrequently directly employed by the unit that administers the warehouse. Additional applications should be managed one at a time by external contracts.

Some final caveats are:

- Start with the simplest applications with visible, short-term payoffs; these are the performance measures
- Keep in mind that the warehouse is not a static system but grows application by application; avoid trying to be all things to all users at the outset
- Maintain the archival, read-only nature of the data; this means resist the temptation to make it an operational database with disparate users supplying updates spontaneously

A discussion of origins of three key ITS/data warehouse systems is in the section entitled "Current Practice."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Real time traffic management has been underway in Baton Rouge, Louisiana for many years. What is now missing is a central resource for tracking trends and responding quantitatively to planned changes and the effects of incidents. Looking to the future, the Louisiana Transportation Research Center (LTRC) requested a study leading to the design of a data warehousing/data mining component to its Intelligent Transportation System (ITS). While limited in geographic scope to the freeways in Baton Rouge, the study is intended as a statewide model. The key results are a detailed design of the three stages of a data warehouse

- Extraction, transformation and loading of source data,
- Main database based on star schema principles, and
- Client-end processing

Plus research reports on data mining applications based on the recommended design.

This report is based in part on an extensive examination of several prominent Travel Labs, how they function and the benefits they gain from using an Archived Data Management System (ADMS) for ITS data. The ADMS complements a data warehouse system. As noted in several documents transportation professionals are becoming increasingly aware of the benefits of real-time and archive data generated by ITS systems. Particular to Louisiana, the use of ITS data by personnel with the Regional Planning Commission proved that the evacuation counter-flow was implemented later than the most effective time for efficient evacuation of New Orleans during Hurricane Ivan. Some of the benefits noted in this report are freeway performance measures, analysis, and new custom applications for end users. In addition, this report documents the results of lengthy interviews and surveys with key ITS data users and managers to provide our customers with implementation challenges and lessons learned from practitioners experienced in Archive Data Management Systems (ADMS)/ITS data warehouse deployment or uses. For the purpose of this report we will substitute data warehouse with ADMS. ADMS provides the crucial link between the sources of real-time ITS data and archived data users. This has become the accepted term by the transportation industry. The FHWA study "Cross-Cutting Studies and State-of-the-Practice Reviews: Archive and Use of ITS Generated Data" notes that the benefits of using ITS-generated data vary from one application to the next and are driven by end-user requirements and input.

Report Outline

Aspects of both a design document and research report are incorporated within this report. Two sections, one on requirements and one on database design directly follow and constitute the majority of the design information. Those are followed by a section of research reports, each based on original experiments performed using either data collected on Baton Rouge freeways or those collected in Minneapolis, Minnesota[1]. The first of these reports concerns advanced methods for quality assurance during source data extraction. The latter two describe potential data mining applications – incident prediction using traffic conditions and "not now" travel time prediction.

[1] In some cases, the Baton Rouge data obtained was either incomplete or various components not concurrent in time. In these cases, the Minnesota data was used.

The final section is a survey of existing ITS/WAREHOUSE systems such as California's PeMS Arizona's Macapia County RADS, and Washington state's TRAC. It is noteworthy that the majority of applications observed were performance measurement and, apart from visualization, little in the direction of data mining was encountered.

Methodology

The tasks leading to this report are shown in Figure 1. There are four fundamental objectives (O1, O2, O3, and O4) illustrated in the diagram that are the focus of this study. These encompass nine of the eleven tasks (labeled T1, T2, …, T9) in the figure. Arrows between tasks represent information flow. The objectives and tasks are:

O1. Determine Client Needs
- T1/T3. Current Practice Study and Site Survey
- T2. Users & Uses
- T4. Requirements

O2. Determine Data Sources
- T5. Data Source Compilation
- T6. Sample Data Evaluation
- T7. Quality Metrics

O3. Specify Functional Design
- T8. Functional Specifications

O4. Specify Physical Design
- T9. System Design

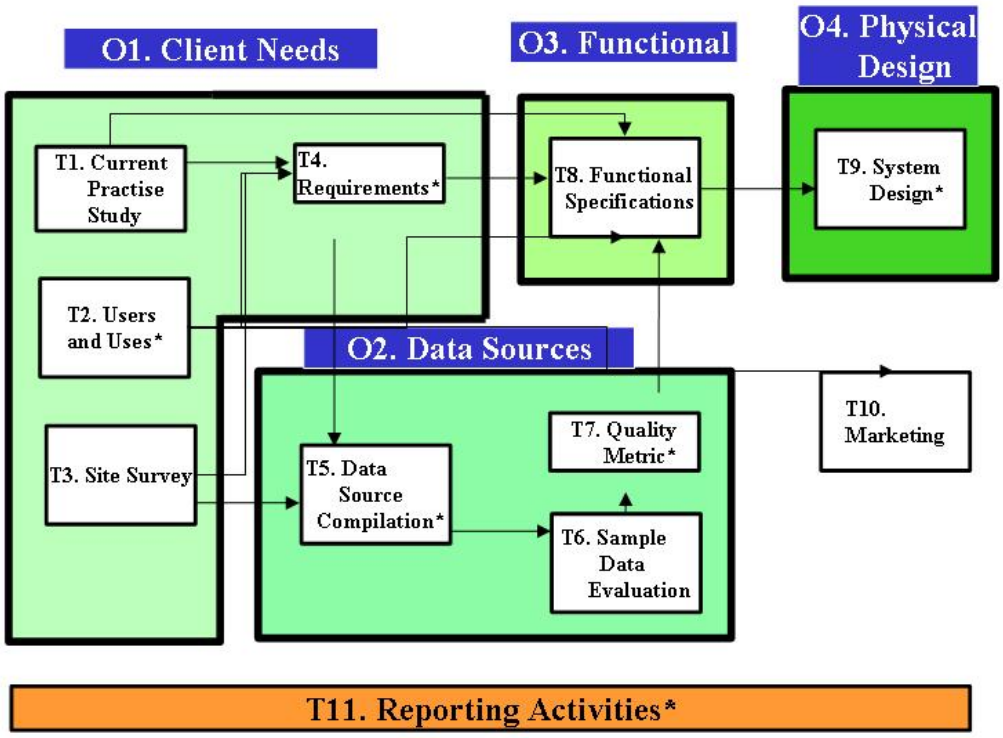
Two additional tasks are not specific to one objective:

- T10. Marketing
- T11. Reporting Activities

Examining the tasks above from a different perspective,

- We first examined the application side by visiting operational WAREHOUSE sites with frequent conferences in Baton Rouge with stakeholders;
- We next examined the source data by collecting a month of data (July 2004) from potential sources in order to determine its format, completeness, and consistency;
- We then constructed data models leading from the source to the application;
- Finally, we used the data models to construct a design of the warehouse.

Overlaying this activity, we developed prototype software to experiment with the feasibility of specific visualizations and specific data mining applications.



* Significant LSU support

Figure 1. Task sequence for project

DATA WAREHOUSE REQUIREMENTS

The requirements can be summarized as

1) determine what data must be available in the data warehouse based on users and uses,
2) how it is organized, and
3) how often it is updated.

The background includes the geographic coverage, the initial clients, and applicable standards.

Coverage Area

The actual coverage of the current sensor set is shown in the upper left corner of Figure 2. Future geographic coverage includes only freeway segments in East Baton Rouge, West Baton Rouge, and Livingston parishes illustrated in the main part of Figure 2. The warehouse design is intended to be a statewide model. As will be explained in a later portion of this report, it is recommended that future extensions be in the form of distributing the data warehouse, i.e., replicating the design in other jurisdictions.

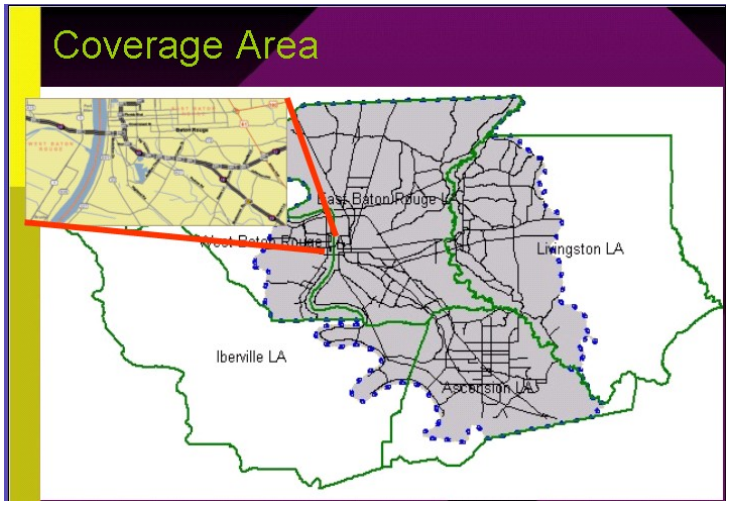


Figure 2. Warehouse geographic coverage

Users

Through several meetings with the DOTD ITS staff, the MPO and EBR TMC a core group of clients were identified. The following table contains the primary users for the case study operation test WAREHOUSE for the Baton Rouge capital region.

Once operational but limited to the capital region, the core group should expand to include:

- Police and fire departments

- Other emergency service providers
- Transit Agencies
- The media
- Event venues
- The general traveling public

Once expanded across the State or significant portions of it, the user group will expand to include the Louisiana Highway Patrol, military bases, and undoubtedly others.

Table 1. Contributing committees

Group 1. Data Holders/Source/EndUsers Working Committee

Organization	Title	Contact
PBF	Consultant	Elizabeth Delaney
City- Baton Rouge	Traffic Engineer	Ingolf Partenheimer
DOTD	Architect	George Gele
City- Baton Rouge	Chief, PCRPC	Huey Dugas
DOTD	ITS	Carryn Zeagler
DOTD-District 61	Traffic Engineer	Ronnie Carter
DOTD	State counts	Bob Smith
DOTD	Traffic Engineer	Peter Allain

Group 2 Advisory End Users

City- Baton Rouge	Emergency Services	Ralph Ladnier
DOTD-District 61	Maintenance Engineer	Terri Hammack
DOTD	Safety	Dan Magri
DOTD	ITS	Stephen Glascock
DOTD	GIS	Jim Mitchell
DOTD	IT	Dom Cali
*FHWA		John Broemmelsiek
*Homeland Security	Emergency Operations	Matt Farlow
*Louisiana State Police		Michael Edmonson

In Table 1 there is an expansion of the core group. These are the stakeholders in the WAREHOUSE for the Baton Rouge region. It represents a two level stakeholder effort for the development and implementation of the regions ITS WAREHOUSE. Group 1 represents the local East Baton Rouge region and is the potential data owners/warehouse contributors. Group 2 represents a statewide advisory committee with expanded interests and uses.

This report recommends meeting among the members to determine their participation in the project, prior to any operational implementation. It has been noted in other warehouse development projects that stakeholders' involvement is critical to the success. Meeting schedules range from one to four meetings annually.

Uses

A traffic data warehouse is a component of an ITS. An ITS consists of roadside infrastructure, communications, traffic centers, and software systems. The four function as an integrated system with the high level goals of making transportation systems run more smoothly and more safely. The warehouse is one software component and has the special requirement of serving as a platform for information relevant to planning.

The other major software components are the real-time traffic management system and task-specific operational database.[2] Before examining uses, it is beneficial to examine how we believe the warehouse should be embedded within the larger software framework. This is shown in Figure 3.

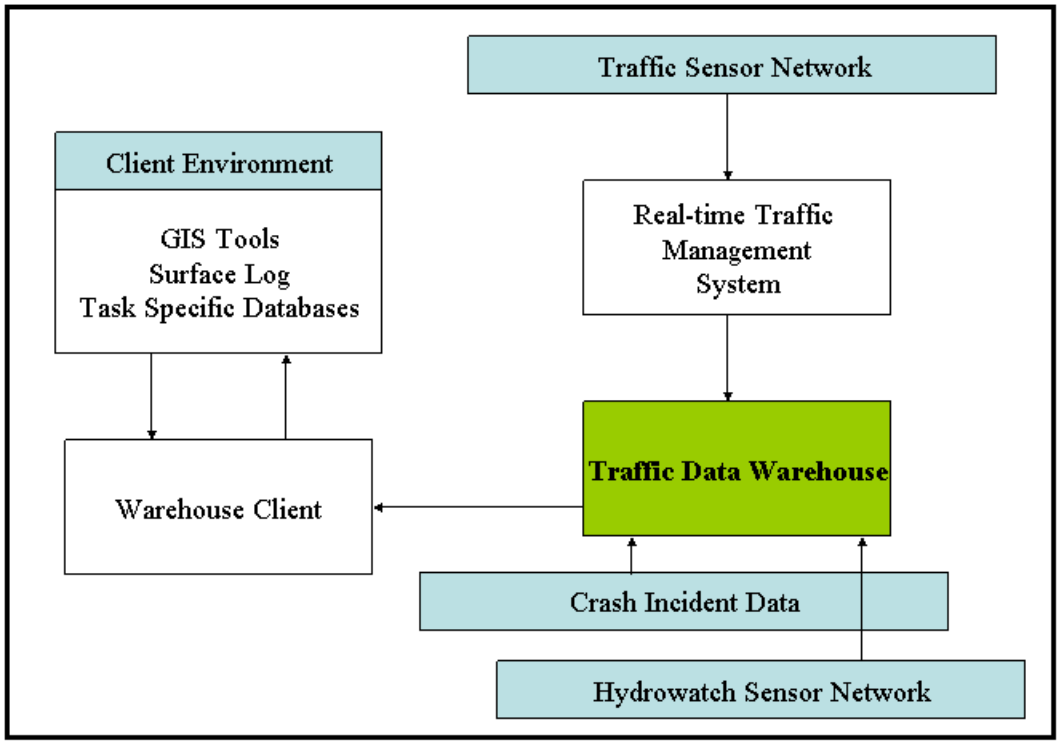


Figure 3. The warehouse embedded within the ITS software components

Noteworthy in our vision of the warehouse is that the input can be largely automated with the exception of the crash incidents. The clients are consumers, not producers. This is consistent with state-of-the-art approaches.

It is best to contrast the uses of the warehouse with that of a Real-time Traffic Management System (RTMS) since the two are so closely linked within an ITS. A scan of Table 2 shows that

[2] An operational database is one designed with the goal of allowing clients to both retrieve and update. A data warehouse is a database for which the clients are restricted to retrieval and the update functions are tightly restricted to a second, generally, automated, group.

an RTMS is focused on the future while a data warehouse is focused on the past. Nevertheless, while the warehouse does not make decisions affecting the future, it is designed to provide information that assists in making choices.

Table 2. RTMS and data warehouse applications

RTMS	Data Warehouse/Clients
Manage traffic flow to ensure the highest utilization of the transportation infrastructure	Track changes in infrastructure utilization at different levels of aggregation in both time and geography
Reduce response time for incidents, keep roads clear of obstructions, minimize secondary incidents	Assist in identifying recurring problems such as excessive congestion or unusually high incident counts
Dispense traffic information with respect to congestion, incidents, road conditions, and maintenance activities – this includes travel times, dynamic message signs, and the 511 systems	Provide information to managers and planners relevant to making decisions that effect utilization – this includes planning for events, planning for infrastructure maintenance, and planning for infrastructure additions
Adjust dynamic traffic control devices such as signal timing, camera directions, and ramp meters	Track the performance of infrastructure components such as sensors and report anomalies that occur consistently

The design we produce here includes applications that support the first three items in Table 3. Sensor health is not directly included but we do include a research report that is relevant to adding it in the future.

Table 3. Capital Area Traffic Data Warehouse Applications

Application	Description
Mobility and performance measures	Straightforward measures include averaging speed, occupancy, and volume; they also include vehicle miles traveled, volume-to-capacity ratios, and others.
Reliability measures	Chief among the reliability measures is the *buffer index*, a measure of congestion recognized by the FHWA but there are others; the buffer index expresses the "extra buffer" needed to be on time for 95% of the trips between points A and B.
Hydrowatch	Correlation of road conditions, particularly open vs. closed, with conditions of local hydrological features – particularly streams at bridges.
Visualization	The above information can be viewed as a two- or three-dimensional graph; many of the more informative visualizations are in conjunction with the surface network which is obtained by combining the warehouse data with the surface log and a Geographic Information System (GIS).

The mobility and reliability measures are common to each of the operational warehouses that we examined including PeMS (California) and Smart Travel Lab (Virginia). The visualization

application is also present in most warehouses. It is not, however, directly a function of the warehouse but of the client stations that the warehouse supports. The hydrowatch application is unique to the Capital Area/Statewide design. Its inclusion recognizes the unique effect climate has on traffic in our State.

Applicable Standards

The warehouse should be based upon industry and ITS standards to insure national integration and statewide implementation. While no standard currently exists explicitly for a transportation archival database, there does exist ITS standards which can be reviewed prior and during deployment to ensure that all appropriate data points defined by those standards are accounted for in the archival database and that the data is stored in a manner that is consistent with the metadata defined in the standards. Companies such as Open Roads Consulting, Berkeley Transportation Systems, TTI, etc. have extensive experience in complying with industry and national ITS standards.

VDOT has assembled a comprehensive (but not complete) list of standards for their "TMC Applications of Archived Data Operational Test" document. This study is by the Virginia Department of Transportation Prime, GMU Subcontractor, and its purpose is to document how transportation management center (TMC) operational practices and procedures can benefit through the applied use of archived data from highway-based and/or transit-based ITS sources. This effort will consider how specific TMC functions can be enhanced through performance measures and analytical techniques enabled through archived data. GMU's role will be to assist the local planning organization in the development of models that will feed from ITS archived data and at the same time, support their needs as a planning organization.

The standards examined for this study include those in Table 4.

Table 4. Standards applicable to ITS data warehousing

Standard	Significance to Project	Use on Project
NTCIP 1201 - Global Object Definitions	This standard was published in November 1997 and later amended (Amendment 1) in December 1998. Current work is being done on a Version 2 of this standard. Version 2 is currently available for User Comment under NTCIP bulletin B0071. This standard defines data object that are common to all (or most) field controllers used in ITS systems.	The data object defined by this standard will be evaluated for inclusion into the ADMS, especially the data pertaining to the controller's identification, type, and location.
NTCIP 1206 - Object Definitions for Data Collection	This standard was released in February 2002 as a User Comment Draft under NTCIP bulletin B0072. This standard defines the data object for Data Collection systems. These systems include those that measure and log traffic volumes and classifications.	The data object defined by this standard will be evaluated for inclusion into the ADMS, especially the data pertaining to the measured traffic data.
SAE J2353 ATIS Data Dictionary	This standard defines a minimum set of data elements needed by potential information service providers to deploy ATIS services.	The data object defined by this standard will be evaluated for inclusion into the ADMS.
SAE J2354 ATIS Core Message List And Data Dictionary	This standard defines a basic message set using the data elements from the ATIS data dictionary needed by potential information service providers to deploy ATIS services.	The data object defined by this standard will be evaluated for inclusion into the ADMS.
ITE TM 1.03 - Standard for Functional Level Traffic Management Data Dictionary (TMDD)	This standard defines the data elements for roadway links and for incidents and traffic-disruptive roadway events. It also includes data elements for traffic control, ramp metering, traffic modeling, video camera control traffic, parking management and weather forecasting, as well as data elements related to detectors, actuated signal controllers, vehicle probes, and DMSs.	The data object defined by this standard will be evaluated for inclusion into the ADMS, especially the data pertaining to specific device state, measured traffic data, recorded incidents, and actions taken.
ITE TM 2.01 - Message Sets for External TMC Communication (MS/ETMCC	This standard defines a message set for communication between traffic management centers and other ITS centers, including information service providers, emergency management systems, missions management systems, and transit management systems.	The data object defined by this standard will be evaluated for inclusion into the ADMS.

THE DATA WAREHOUSE DESIGN

In this section, following the design overview, the three warehouse stages are described in successive subsections. Sample source data was requested for July 2004 and a mixture from the two months of July and August was obtained. This data was examined and its form together with an evaluation of the quality is in the penultimate subsection. The last subsection is reserved for further discussion of difficult design issues.

Design Overview

An overview of the design is shown in Figure 4. The two principal points to be noted from the figure are:

- There are three stages or layers consisting of the ETL, the main storage, and clients;
- Within the stages, the information flow is strictly left to right, that is, direct updating of the archived data is not allowed.

The left-to-right flow distinguishes a data warehouse from an operational database. In an operational database, the clients are permitted to update data directly. The restriction is circumvented by the feedback loop which allows client-generated information to be inserted from a source file on the next update cycle.

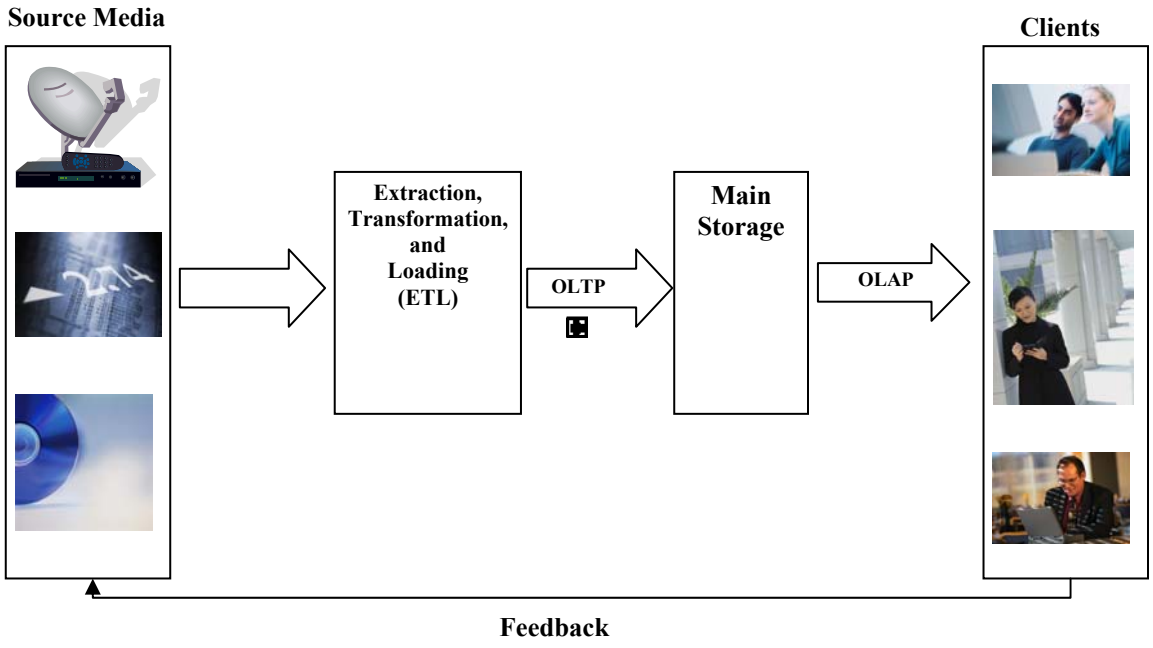


Figure 4. Overview of DW design

In the figure, OLTP stands for online transaction processing and is the set of procedures for ordinary operational databases. That is to say, the source media inputs are staged in an ordinary

relational database prior to loading into the main storage tables. OLAP stands for online analytical processing and is a set of procedures, including an SQL-like query language, for accessing data in the specialized schema of data warehouses.

The Extraction, Transformation, and Loading Stage

One research paper [63] presents a review of data failure screening methods and proposes its own methodology for detecting potentially erroneous observations. According to the paper, tests for screening traffic data can be divided into two categories:

- Threshold value tests
- Traffic flow theory principles tests

The paper makes an argument for using traffic flow theory tests, and claims that they can detect a wide range of failure modes.

Archived data management systems (ADMSs) are data warehouses created to support analyses based on data collected by transportation operations systems. While many believe that an ADMS can be created by simply exporting data from an operations system, experience in developing the Virginia ADMS illustrates that the creation of an effective ADMS requires careful attention to the extraction, transformation, and loading (ETL) process. This process refers to the activities conducted when creating a data warehouse from an operational data store.

This paper addresses four critical elements of an ADMS ETL process: data aggregation, data quality assessment, data imputation, and data characterization. For each element, the purpose and need are documented, a review of available alternative implementation methods is presented, and, finally, a description of the approach used in the Virginia ADMS is detailed. Transportation professionals are becoming increasingly aware that traffic data collected to support operations, such as those collected by signal control systems and freeway management systems, hold considerable value for use in a variety of planning and analysis applications . This awareness has led to the development of the archived data user service (ADUS) element of the National Intelligent Transportation Systems (ITS) architecture. Specific implementations of ADUS, referred to as archived data management systems (ADMS) can best be classified as data warehouses, an emerging area in the broader arena of information technology. A formal definition of a data warehouse is a subject-oriented, point-in-time, inquiry-only collection of operational data. Thus, one can see that the development of an ADMS involves creating a system that supports analysis based on operational data.

A common misperception concerning ADUS is that implementation involves simply providing query access to the operational database used in a traffic operations center. This approach does not meet the requirements of the service for two primary reasons: (1) it increases risk of system failure by allowing additional access to the information technology subsystem of a traffic operations center, and (2) the operational database is not designed to support query and analysis. Thus, effective warehouse systems involve creating data warehouses that are separate from the operations center's database. The process of populating this data warehouse, or preparing and moving data from an operational database to the warehouse, is known as Extraction, Transformation, and Loading (ETL). The quality of the ETL process dictates the effectiveness of a data warehouse in meeting user needs.

The purpose of this research was to investigate the ETL process in the context of ADUS applications. The results of the research indicate that effective warehouse ETL is a complex process that requires careful investigation of statistical, traffic flow theory, and information technology concepts. This paper presents the research team's findings using the ETL architecture developed for the Virginia warehouse, a data warehouse intended to support the use of traffic operations data collected by the Virginia Department of Transportation (VDOT).

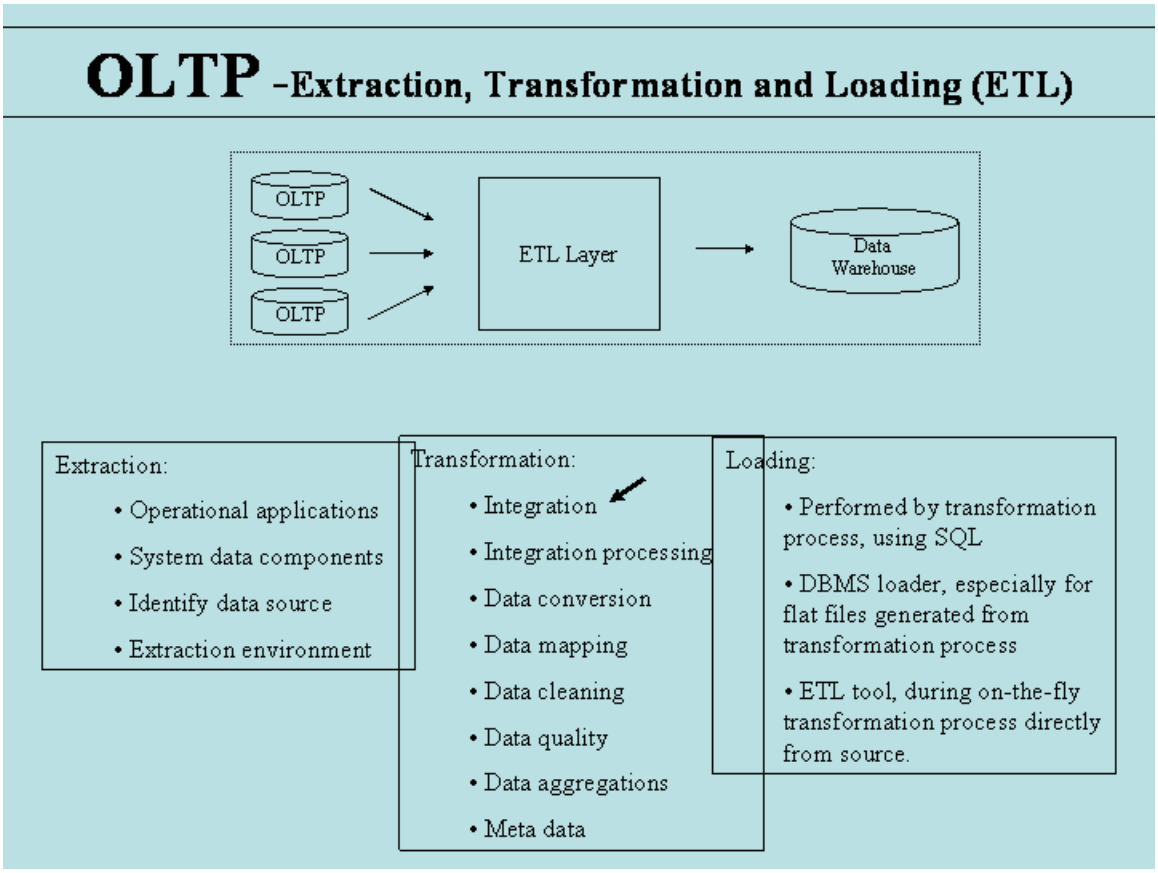


Figure 5. Warehouse backend – extraction, transformation, and loading

The data warehousing literature refers to databases created to support operations as operational data stores (ODSs) (*1*). Examples of ODSs in transportation operations consist of databases incorporated in the management and control software of operations centers, such as transportation management systems. ODSs are intended to support real-time insertion of data in a reliable, available manner. They are not intended to support query and analysis – this is the role of a data warehouse. The process of "moving" data from an ODS to a data warehouse is known commonly as Extraction, Transformation, and Loading (ETL) and is depicted in Figure 5, with ADUS terminology included in the appropriate locations. The following sections provide detail on each element of the ETL process [64].

Extraction

Extraction is simply the process of selecting and obtaining data from the ODS server to use within the data warehouse. This activity must be designed in such a way to minimize risk to the ODS. Extraction can occur on varying time scales. In "real-time" extraction, data is pulled from the ODS to populate the warehouse on a nearly continual basis. On the other hand, extraction can

be designed to occur infrequently, such as once a day, in which all of the data accumulated over the day at the ODS is pushed to the warehouse.

Transformation

This critical aspect of the ETL process is frequently overlooked in transportation data warehousing applications. The purpose of transformation is to prepare data from them ODS to best support analysis. Transformation includes such activities as data screening (or validation of values), combining data from multiple ODS sources, and building aggregates to improve query performance (*1*). It is generally accepted that the transformation step must be tailored to the specific industry/application in order to create an effective data warehouse. For this reason, this paper focuses on transformation activities in the ETL process of the Virginia warehouse.

Loading

The key aspect of loading is quality control. The main issues in quality control are data screening, data imputation, and abnormality testing. These are described below.

> *Data Screening* – Certainly, an important prerequisite of any effective warehouse is high quality data. A number of detector screening algorithms should be developed. Industry standards such as those published by the Virginia Smart Travel Lab or Texas Transportation Institute can be used as a guide. These standards have been published in the *Transportation Research Record* and are currently in use in TMC's throughout the country. The first research report in Appendix B describes our contribution to this topic.
>
> *Data Imputation* – Missing data is a fact of life in all ITS deployments. The harsh field conditions, communications frailties, and software bugs result in a missing data rate that often averages 30 percent of all detectors at any given time. In the past, systems usually "gave-up" on missing data, and did not try to estimate conditions based on adjacent detectors and/or archived data. However, as illustrated by some of the work that Oak Ridge National Laboratories conducted on ADUS, the spatial and temporal patterns of traffic data allow for effective data imputation. Through interviews with local officials it was determine that no data imputation should be used in the Louisiana warehouse.
>
> *"Abnormality" Testing* – Arguably the most important traffic information is a report of traffic conditions that can be classified as out of the "norm." This can explain the widespread popularity of CCTV cameras. Operators can quickly glance at a camera and visually ascertain the normality of the situation. To support a myriad of real-time and archived data, it is necessary to classify conditions as normal or abnormal (and measure the degree of abnormality).

The Texas Transportation Institute publishes a comprehensive list of data validity checks that should be utilized by the proposed warehouse. Below is an example of critical tests. Please refer to publication TTI Report 1752-5 [65] for more detailed information on these types of data checks.

Table 5. Extraction validity checks from Texas Transportation Institute

Exhibit 3-5. 2002 Data Validity Checks in Mobility Monitoring Program

Quality Control Test and Description	Sample Code with Threshold Values	Action
Controller error codes • Special numeric codes that indicate that controller or system software has detected an error or a function has been disabled.	If VOLUME={code} or OCC={code} or SPEED={code} where {code} typically equals "-1" or "255"	• Set values with error codes to missing/null, assign missing value flag/code.
No vehicles present • Speed values of zero when no vehicles present • Indicates that no vehicles passed the detection zone during the detection time period.	If SPEED=0 and VOLUME=0 (and OCC=0)	• Set SPEED to missing/null, assign missing value code • No vehicles passed the detection zone during the time period.
Consistency of elapsed time between records • Polling period length may drift or controllers may accumulate data if polling cycle is missed. • Data collection server may not have stable or fixed communication time with field controllers.	Elapsed time between consecutive records exceeds a predefined limit or is not consistent	• Action varies. If polling period length is inconsistent, volume-based QC rules should use a volume flow rate, not absolute counts.

More sophisticated quality control procedures include:

- Sequential Data Checks – will compare values in consecutive time periods for consistency (e.g., speeds cannot go from 60 mph to 20 mph and back to 60 mph in consecutive 5-minute time periods.
- Corridor Data Checks – will examine the relationship between data along a corridor (e.g., volume into an area should approximately equal volume out).
- Historical Data Checks – will examine the changes from one year to the next for reasonableness (e.g., high increases in volume or drastic changes in speed).

To instantiate these with specific tests used by the Texas Transportation Institute, we have compiled the following list:

- Maximum Volume Threshold – e.g., greater than 250 vehicles per lane for five minutes

- Maximum Occupancy Threshold – e.g., greater than 90 percent for five minutes
- Maximum Speed Threshold – e.g., greater than 80 mph for five minutes
- Minimum Speed Threshold – e.g., less than 3 mph.
- Inconsistency of traffic data values (volume, occupancy, and speed) within the same data record or with traffic flow theory (e.g., occupancy is less than 3 percent but speed is less than 45 mph; speed equals zero but volume is nonzero; occupancy is greater than zero but volume and speed are zero)
- Sequential Volume Test – e.g., if the same volume is reported four or more consecutive time periods, assume that the detector is malfunctioning

A research report on data quality checking is included. The simulations utilized data from outside the Baton Rouge coverage area. Traffic surveillance systems are a key component for providing information on traffic conditions and supporting traffic management functions. A large amount of data is currently collected from inductive loop detector systems in the form of three macroscopic traffic parameters (speed, volume and occupancy). Such information is vital to the successful implementation of transportation data warehouses and decision support systems. The quality of data is, however, affected by erroneous observations that result from malfunctioning or inaccurate calibration of detectors. The open literature shows that little effort has been made to establish procedures for screening traffic observations in real-time. This study presents a realistic approach for modeling and real-time screening of freeway traffic data. The study proposes a simple methodology to capture the probabilistic and dynamic relationships between the three traffic parameters using historical data collected from the I-4 corridor in Orlando, Florida. The developed models are then used to identify the probability that each traffic observation is partially or fully invalid.

The Main Storage Stage

As stated above, the design of the warehouse should take the dimensional approach. It is proposed that this design be further developed in an operational test with stakeholder input. The purpose of this effort would be to test and learn about data management techniques needed to support the warehouse. This new structure could first be developed for East Baton Rouge as a case study, but will be expanded to deal with the TMC and RPC's statewide ITS archive. It is fully understood that the design must contend with multiple locations and allow for the smooth addition of new areas. It is recommended that a separate database segment will be created for each geographic area and all the segments will be tied together in a bus-like manner allowing a warehouse to draw from all of them at once. We refer to these segments as *data marts*, a common term in the data warehousing field. While each data mart is unique, in that it has a distinct fact table, it shares dimensions with other data marts. The bus architecture and the use of common dimensions allow each data mart to have its own characteristics while keeping a consistent format to the entire database. This then would emulate other prominent system designs being developed throughout the nation, such as Virginia's Smart Travel Lab.

Figure 6 illustrates the dimensional approach. The database schema for warehouses follows a pattern quite different from the normalized tables of relational databases. The schema form is sometimes referred to as a "star schema." Examining Figure 6, the center of the "star" is a so-called "fact table" in which is captured the dynamic real time data. Within the fact table are

references to the "dimension tables." The dimension tables contain data that is relatively static over time such as location and calendar data.

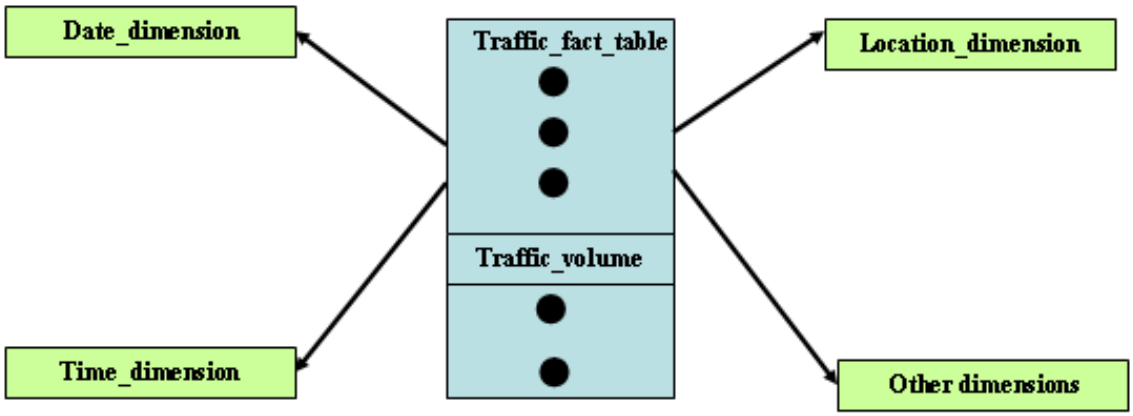


Figure 6. Star Schema Example

Queries use the dimensional information to isolate one or more facts to be extracted from the warehouse. A typical query often has the form:

```
Select average(traffic_volume)
        From Traffic_fact_table
        Where
                <Date_dimension constraint> and
                <Time_dimension constraint> and
                <Location_dimension constraint>
```

The design enables rapid access and transfer of the dynamic, real time data using query selection information (e.g., dates, times, locations) available to all users. Updating the fact table is also straightforward. New sensor data is formed into records and appended to the fact table.

The star schemas for mobility and reliability measures for the Capital Region but extendable Statewide are shown in Figure 7. The fact tables are aligned down the center of the figure while the dimensions are shown to the left and right. The dynamic, real-time data is shaded. The remaining fields in the fact tables are to provide control information and to link the facts with dimensions, e.g., time and place.

Note that there are three fact tables – *DETECTOR_DATA, STATION_DAILY,* and *CRASH_DATA.* Of these, the *STATION_DAILY* is redundant but recommended. Our study has determined that many clients use daily information regularly. Storing precomputed daily volumes takes little space and conserves large amounts of retrieval time (both for the machine and for the client). Construction of a record for the *STATION_DAILY* fact table takes the form of a "roll-up query," which are useful in many other contexts. A roll-up query is one that partitions the original fact table into blocks, then constructs one record per block, that record being the aggregate of the records in the block it represents. For *STATION_DAILY*, the partitioning criterion is "day" and the aggregation operators are sums and averages. Omitting the control

fields, the SQL[3] roll-up query that constructs all the records in *STATION_DAILY* at once has the form:

```
Select sum(Total_Volume), average(Total_Occupancy),
                average(Total_Speed), sum(TypeA_Volume),
                sum(TypeB_Volume,sum(TypeC_Volume)
        From Detector_Data DD, Calendar C
        Where DD.Calendar_Key = C.Calendar_Key
        Group By C.Day_Month_Year
```

A "drill-down query" replaces one or more roll-up records with those from the block from which the aggregate was constructed. Roll-back queries have special syntactical forms in OLAP systems and cannot be illustrated here in SQL.

[3] SQL is the most common database query language in use today. Initially, the letters stood for Sequential Query Language but in recent years it has been treated as a stand-alone technical word.

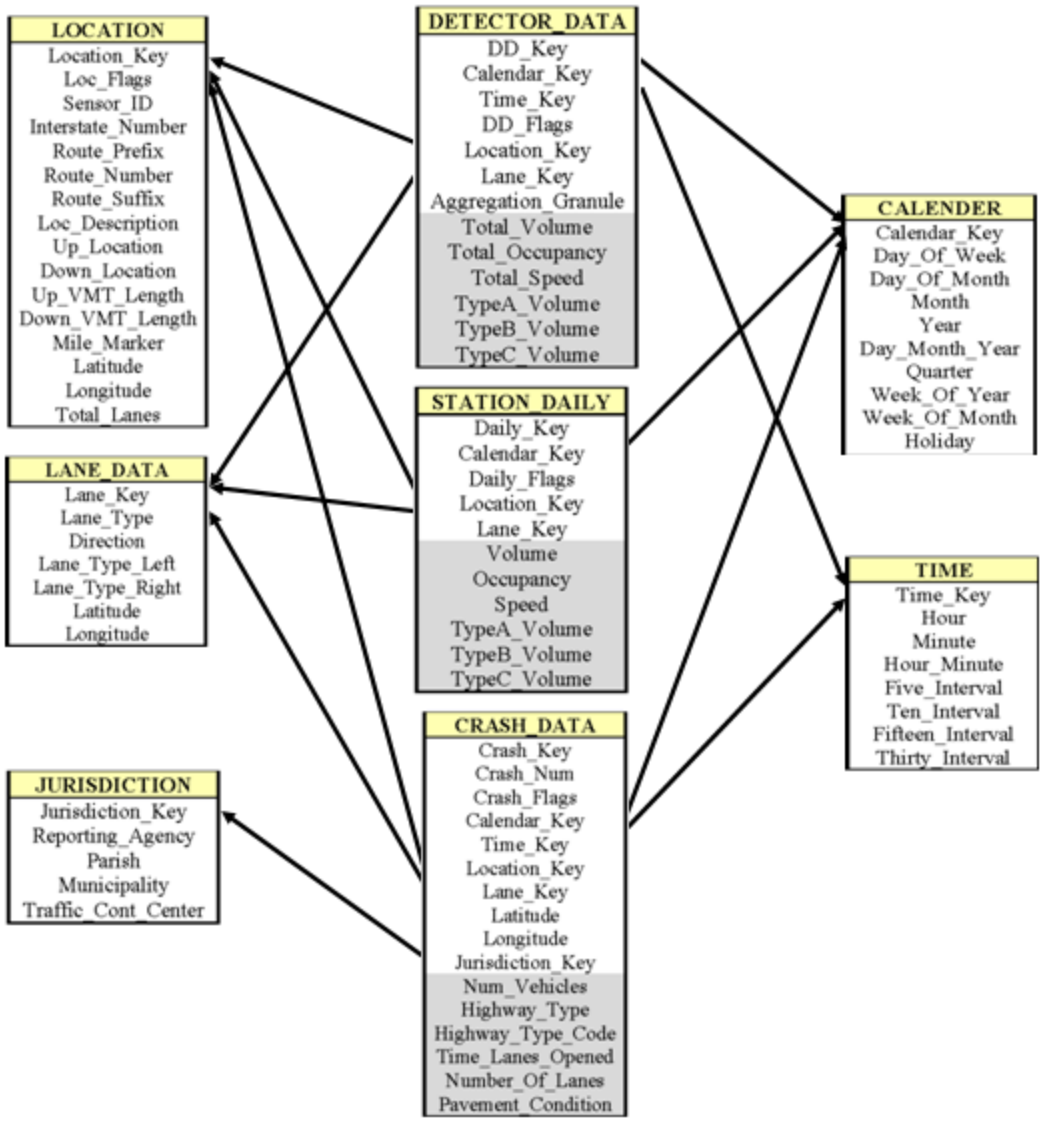


Figure 7. Star Schemas for Mobility and Reliability Measures

The descriptions of the attributes for the fact tables are given in Table 6 and descriptions for the dimensions are given in Table 7. While the descriptions are self-explanatory, this is an appropriate point to indicate that the record keys are internally generated by the warehouse system. It is not good policy to incorporate keys that are in the legacy data directly as keys in the database. Organizations tend to change systems over time and thus change the key designation procedures in the legacy data. Where such keys exist (e.g., *Crash_Num* in *CRASH_DATA*) they are incorporated (but not as database keys) in order to permit records to be correlated with offline files having complementary data.

Table 6. Performance Measure Fact Table Attribute Descriptions

ATTRIBUTE	DESCRIPTION
	DETECTOR_DATA
DD_Key	A unique record key assigned and maintained by the traffic data warehouse system

Calendar_Key	A pointer to a record in the calendar (date) dimension indicating the date on which the data was collected
DD_Flags	A sequence of binary signals (a bit vector) corresponding to a sequence of yes/no variables; here, the field is used to indicate which data verification tests were conducted and to indicate the type of sensor
Time_Key	A pointer to a record in the time dimension indicating the time at which the data was collected
Location_Key	A pointer to a record in the location dimension indicating the position of sensor from whence the data originated
Lane_Key	A pointer to a record in the traffic lane dimension indicating which lane the data describes
Aggregation_Granule	Sensors can aggregate values across a pre-programmed time interval (typically 30 sec. to 15 min.); here the aggregation interval, in seconds, is given
Total_Volume	Total count of vehicular traffic in the lane over the past number of seconds given in *Aggregation_Granule*
Total_Occupancy	Total occupancy in the lane over the past number of seconds given in *Aggregation_Granule* and expressed as a per cent
Total_Speed	Average traffic speed in the lane over the past number of seconds given in *Aggregation_Granule* and expressed in miles/hour
TypeA_Volume	Count of automobiles in the lane over the past number of seconds given in *Aggregation_Granule*
TypeB_Volume	Count of small trucks and vans in the lane over the past number of seconds given in *Aggregation_Granule*
TypeC_Volume	Count of large trucks and buses in the lane over the past number of seconds given in *Aggregation_Granule*
	STATION_DAILY
Daily_Key	A unique record key assigned and maintained by the traffic data warehouse system
Calendar_Key	A pointer to a record in the calendar (date) dimension indicating the date the data in the record was collected
Daily_Flags	A sequence of binary signals (a bit vector) corresponding to a sequence of yes/no variables; here, the field is used to indicate which data verification tests were conducted and to indicate the type of sensor
Location_Key	A pointer to a record in the location dimension indicating the position from which the data was collected
Lane_Key	A pointer to a record in the traffic lane dimension indicating which lane the data describes
Volume	Total of all volume data in *DETECTOR_DATA* records having the same date and location
Occupancy	Average of all Occupancy data in *DETECTOR_DATA* records having the same date and location
Speed	Average of all speed data in *DETECTOR_DATA* records having the same date and location
TypeA_Volume	Total of all volume data for small vehicles in

	DETECTOR_DATA records having the same date and location
TypeB_Volume	Total of all volume data for intermediate size vehicles in *DETECTOR_DATA* records having the same date and location
TypeC_Volume	Total of all volume data for large vehicles in *DETECTOR_DATA* records having the same date and location
	CRASH_DATA
Crash_Key	A unique record key assigned and maintained by the traffic data warehouse system
Crash_Num	Key of external record from which the data in the record was taken; external record keys are assigned by the jurisdiction investigating the accident
Crash_Flags	A sequence of binary signals (a bit vector) corresponding to a sequence of yes/no variables; here it is used to indicate the crash type and the weather which are available from the crash reports
Calendar_Key	A pointer to a record in the calendar (date) dimension indicating the date the crash occurred
Time_Key	A pointer to a record in the time dimension indicating the time the crash occurred; it is also assumed that this is the time that the lane closures occurred
Location_Key	A pointer to a record in the location dimension which indicates where the crash occurred
Lane_Key	A pointer to a record in the traffic lane dimension indicating the primary lane of the crash
Latitude, Longitude	Redundant with the *Location_Key* field but useful for queries for which the purpose is to find accident clusters
Jurisdiction_Key	A pointer to a record in the jurisdiction dimension indicating the originating agency of the source report
Num_Vehicles	Number of vehicles involved in crash
Highway_Type	There are five types: "city street," "parish road," "State highway," "US highway," and "interstate."
Highway_Type_Code	A code letter: A, B, C, D, E; redundant with *Highway_Type* but useful for querying
Time_Lanes_Opened	A pointer to a record in the time dimension indicating when the lanes reopened to normal traffic
Number_Of_Lanes	The number of lanes closed due to the crash
Pavement_Condition	Values include choices from "no effects" through "construction"

Table 7 includes two features that are designed to facilitate GIS interfaces within client applications. First the latitude and longitude coordinates are given in the *LOCATION* dimension and (redundantly) in the *LANE_DATA* dimension. Commercial GIS products work directly with lat/long coordinates. The second feature is the linking together of *LOCATION* dimension records representing consecutive sensor locations on the same traffic corridor (see the attributes

Up_Location and *Down_Location* in the table). This enables client applications that have selected a central focus to establish the boundaries of the GIS view.

Table 7. Performance Measure Dimension Table Attribute Descriptions

ATTRIBUTE	DESCRIPTION
	CALENDAR
Calendar_Key	A unique record key assigned and maintained by the traffic data warehouse system
Day_Of_Week, Day_Of Month, Month	The meaning of these three fields is evident from the field names
Day_Month_Year	For some queries, it is convenient to have the previous three fields combined into a single field
Quarter, Week_Of_Year, Week_Of_Month	The meaning of these three fields is evident from the field names
Holiday	Holiday name
	TIME
Time_Key	A unique record key assigned and maintained by the traffic data warehouse system
Hour, Minute	The hour and minute since midnight, local time
Hour_Minute	For some queries, it is convenient to have the time combined into a single field
Five_Interval	An integer denoting in which of the 288 five-minute intervals from midnight to midnight the current time occurs
Ten_Interval	An integer denoting in which of the 144 ten-minute intervals from midnight to midnight the current time occurs
Fifteen_Interval	An integer denoting in which of the 96 fifteen-minute intervals from midnight to midnight the current time occurs
Thirty_Interval	An integer denoting in which of the 48 thirty-minute intervals from midnight to midnight the current time occurs
	LOCATION
Location_Key	A unique record key assigned and maintained by the traffic data warehouse system
Loc_Flags	A sequence of binary signals (a bit vector) corresponding to a sequence of yes/no variables
Sensor_ID	Sensors have an identity such as "RVD 41" assigned by the Traffic Management Center when they are installed and put online
Interstate_Number	If applicable, the integer designation
Route_Prefix	Such as "I," "US," "LA," etc.
Route_Number	The integer designation
Route_Suffix	If applicable, "S," "N," …, "ALT," etc.
Loc_Description	Many sensors are located at major crosspoints such as "Government St." which provides a useful alternative query option
Up_Location, Down_Location	For convenient GIS processing, the location records for each route are in the form of a doubly-linked list; "up" and "down"

	are in terms of ascending/descending mile markers
Up_VMT_Length	The length of the section in miles between the "up location" and this location
Down_VMT_Length	The length of the section in miles between the "down location" and this location
Mile_Marker	The mile post value
Latitude, Longitude	The Lat/Long values are for interfacing with GIS systems
Total_Lanes	The total number of lanes at this location including both directions and any turn lanes or ramps
	JURISDICTION
Jurisdiction_Key	A unique record key assigned and maintained by the traffic data warehouse system
Reporting_Agency	Law enforcement unit
Parish	Parish name or code
Municipality	Incorporated area name or code
Traffic_Cont_Center	Presently, only the Capital Region Traffic Management Center is involved
	LANE_DATA
Lane_Key	A unique record key assigned and maintained by the traffic data warehouse system
Lane_Type	On-ramp, off-ramp, left-turn lane, through lane, etc.
Direction	North, South, East, or West
Lane_Type_Left	The type of the lane on the left – left determined by the direction of traffic flow
Lane_Type_Right	The type of the lane on the right – right determined by the direction of traffic flow
Latitude, Longitude	Lat/Long here is redundant with the companion *LOCATION* record, but useful for GIS interface

Available data sources for the mobility/performance tables are as follows:

- *CALENDER* and *TIME* – These tables are common to data warehouses of all types and can be purchased as "plug-ins" or computer programs can be written to generate them; the *CALENDER* dimension should span more than a decade
- *LOCATION* and *LANE_DATA* – These must be compiled from the surface log and records added each time there is a new RTMS sensor installed
- *DETECTOR_DATA* and *STATION_DAILY* – The data source for the Capital Region is the Traffic Management Center in Baton Rouge
- *CRASH_DATA* – While one can expect a delay in availability of the electronic records, the legacy data for both the city and parish is collated by the LADOTD IT division; both the city and state police were reluctant to remit data for us to inspect

The hydrowatch application is unique and is intended, over the long term, to assist in predicting flooding events that affect traffic flow. Given the climate and unique weather features of Louisiana, this application has special relevance.

The star schema for the hydrowatch application is shown in Figure 8. The fact table (*HYDRO_DATA)* is in the center and the dimensions are on the left and right. Two dimensions,

CALENDAR and *TIME*, are shared between the hydrowatch application and the performance measure application. Such dimensions are referred to as *conformed dimensions*. Special care must be exercised in designing conformed dimensions because they play an important role in a special type of OLAP query known as the "drill through query."

The drill through query links one application with another. In this case, one would want to link stream conditions with traffic mobility. Unfortunately, the *CALENDAR* and *TIME* conformed dimensions are not adequate vehicles for doing this. Ideally, the *LOCATION* dimension and the *HYDRO_SITE* dimension should form a single, conformed dimension. There is not sufficient commonality in the two to warrant combining them. (However, it should be noted that the two conformed dimensions assure that roll-ups across the two applications will be consistent.)

As an alternative, the lat/long attributes in the two dimensions much be matched to collate a hydro sensor site with a traffic sensor site. That is, a simple drill through query will resemble

```
Select Location_Key, HS_Key
    From LOCATION L, HS_SITE H
    Where (H.Latitude between <L.Latitude condition>
                        and <L.Latitude condition>)
          and
          (H.Longitude between <L.Longitude condition>
                        and <L.Longitude condition>)
```

This is not an ideal solution and we do suggest that additional thought is given to this problem before implementation occurs.

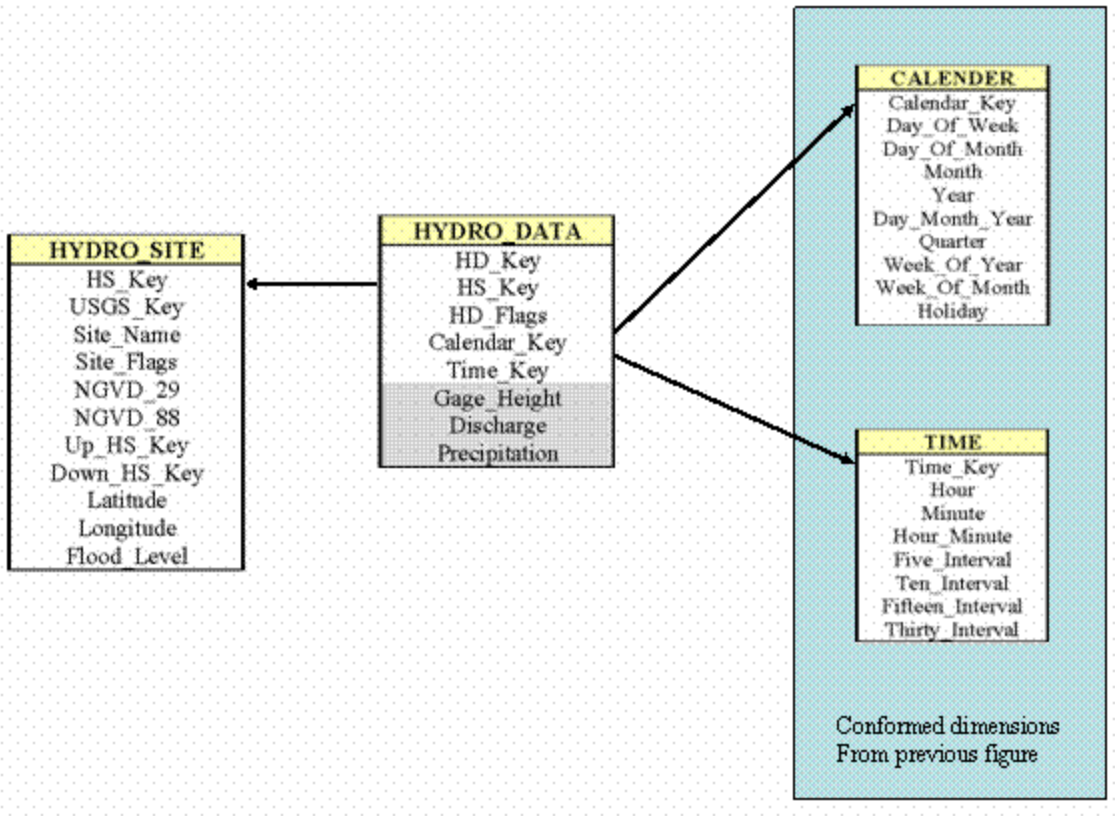


Figure 8. Star Schema for Hydrowatch Application

The attributes for hydrowatch are summarized in Table 8 and Table 9. The dynamic, real time data is shown in the shaded area.

Table 8. Hydrowatch Fact Table Attribute Descriptions

ATTRIBUTE	DESCRIPTION
	HYDRO_DATA
HD_Key	A unique record key assigned and maintained by the traffic data warehouse system
HS_Key	A pointer to the site dimension giving the location of the sensor
HD_Flags	A sequence of binary signals (a bit vector) corresponding to a sequence of yes/no variables; here, the field is used to indicate which screening tests have been applied to the data
Calendar_Key	A pointer to a record in the calendar (date) dimension indicating the day on which the data was collected
Time_Key	A pointer to a record in the time dimension indicating the time at which the data was collected
Gage_Height	Stream height (in feet) where the base height is either NGVD 29 or NGVD 88 survey elevations
Discharge	Discharge volume in cubic feet per second; note that discharge may also be called “stream flow rate” for some sensors
Precipitation	Precipitation in inches since midnight

The solution to the problem of interfacing with GIS products is similar to that employed in the mobility/reliability applications. Table 9 includes two features that are designed to facilitate GIS interfaces within client applications. First the latitude and longitude coordinates are given in the *HYDRO_SITE* dimension. As noted, commercial GIS products work directly with lat/long coordinates. The second feature is the linking together of *HYDRO_SITE* dimension records representing consecutive sensor locations on the same stream flow (see the attributes *Up_HS_Key* and *Down_HS_Key* in the table). This enables client applications that have selected a central focus to establish the boundaries of the GIS view.

Table 9. Hydrowatch Dimension Table Attribute Descriptions

ATTRIBUTE	DESCRIPTION
	HYDRO_SITE
HS_Key	A unique record key assigned and maintained by the traffic data warehouse system
USGS_Key	An eight-digit USGS code for the site
Site_Name	A text string with the USGS name of the site
Site_Flags	A sequence of binary signals (a bit vector) corresponding to a sequence of yes/no variables; here, the field is used to indicate whether elevations are NGVD_29 or NGVD_88 and the type of sensor installed
NGVD_29	The elevation of the site (in feet) using 1929 USGS data; only one of the NGVD fields will be used – some sites use 1929 elevations and others use 1988 values
NGVD_88	The elevation of the site (in feet) using 1988 USGS data; only one of the NGVD fields will be used – some sites use 1929 elevations and others use 1988 values; NGVD 88 values are sometimes called NAVD 88 values
Up_HS_Key	A pointer to the site dimension giving the location of the nearest upstream sensor, if any
Down_HS_Key	A pointer to the site dimension giving the location of the nearest downstream sensor if any
Latitude, Longitude	Lat/Long values are for interfacing with GIS
Flood_Level	The gage level at which flooding occurs

There are 222 hydrowatch stations that produce real-time data in Louisiana which are maintained by the USGS. About a dozen of these are in the Capital Region. They are located mainly at points where bridges cross streams or canals. The data is available at a USGS web site and is maintained there for 30 days following collection. A warehouse front-end "page scrubber" can be written which will automatically collect the data at the any point during the day.

This application has long term payoffs and is unique among the traffic archival systems we studied. Further, the USGS system extends nationwide and the application will serve as a model for similar implementations elsewhere. The notion here is to use historical data to predict imminent road closures based on current gage height and precipitation conditions. This is in contrast to hydrological models which are expensive to develop and applicable to one site. The *Flood_Level* attribute in the *HYDRO_SITE* dimension is the threshold between "safe" and

"unsafe." The application would be more effective still if there were information on road closures due to flooding. We were not able to find centralized, accessible data on this phenomenon.

Much data will need to be collected before the application will be of benefit. We recommend that the previous application (mobility/reliability measures) be given priority.

The Client Stage

Warehouse clients run execute on machines apart from the main data warehouse with appropriate telecommunications connections. Client machines are configured differently for the applications relevant to each group or individual. For example, some individuals will need the surface log and a GIS system to visualize the data on a map. Others might need statistical packages for data mining while others traffic microsimulators which are initialized with data from the warehouse. A client which performs a similar task each day might have the data accessed automatically before he/she arrives so that it is available on the client's computer.

Figure 9 depicts the environment of a client's workstation. All clients will need OLAP tools and a list of leading vendors is given in Table 10. In addition to the OLAP tools, most clients will need little more than commercial off the shelf software including

- Spreadsheet products such as EXCEL
- GIS products such as MAPINFO, ESRI, or InterGraph
- A graphics package capable of producing plots and professional grade illustrations for reports and presentations

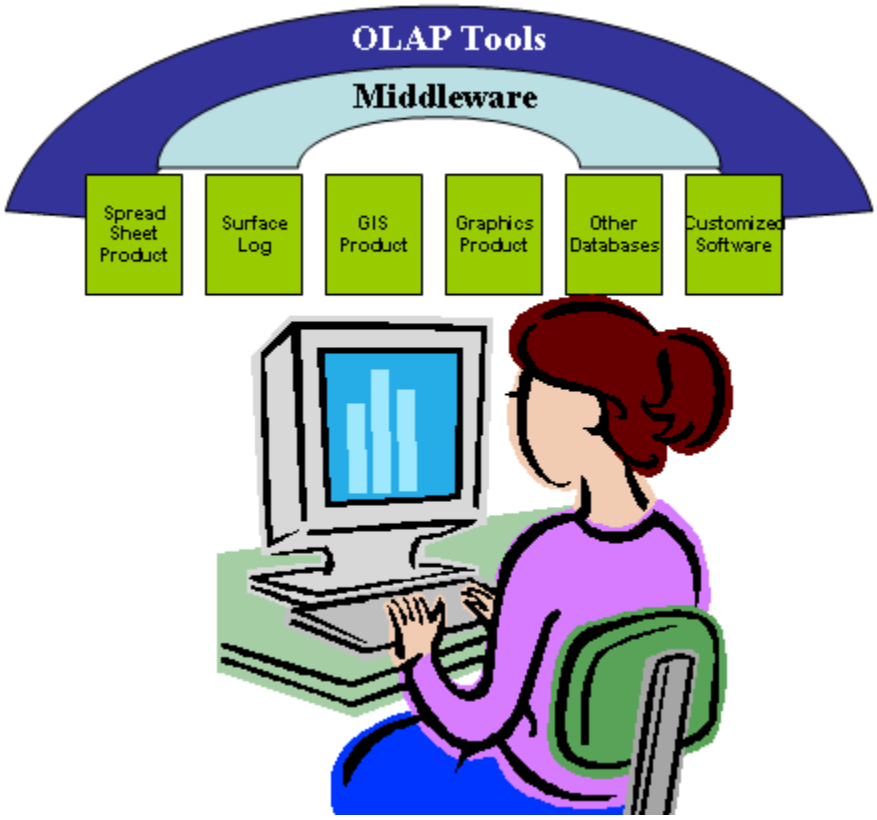


Figure 9. Client Workstation Configuration

Many if not most will need specialized data from other DOTD sources such as the surface log and the complete crash data files. The latter contains information on injuries, vehicle damage, and pedestrians as well as the pre-crash condition of the drivers.

Table 10. OLAP Vendors

Vendor	Market position	Share (%)
Microsoft ecosystem	1	28.0%
Hyperion Solutions	2	19.3%
Cognos	3	14.0%
Business Objects	4	7.4%
MicroStrategy	5	7.3%
SAP	6	5.9%
Cartesis	7	3.8%
Systems Union	8	3.4%
Oracle	9	3.4%
Applix	10	3.2%

A few clients will need extensive commercial and customized systems such as statistical packages and data mining tools as well as very specialized, customized software systems. These clients, typically less than 10 percent of the total, not only acquire value from the warehouse

system but are the most likely to add value to it over the long run. It is worth the effort to invest in them.

We have tested these ideas in a student-based setting by implementing several rudimentary applications. The applications are performed on the client machine just as future applications on the operational warehouse will be. To recapitulate, the four applications and their source datasets are depicted in Figure 10. Three applications are described here. (No sample application relevant to hydrowatch was performed.) These applications were performed using one month's data sample from the summer of 2004 from Baton Rouge agencies (mainly the Traffic Management Center). The applications were performed, in some cases, on earlier versions of the main storage design. In fact, the original purpose was to test concepts of the design. Nevertheless, the results are relevant to the final design – that is to say, the final design is largely a superset of the design used in these applications.

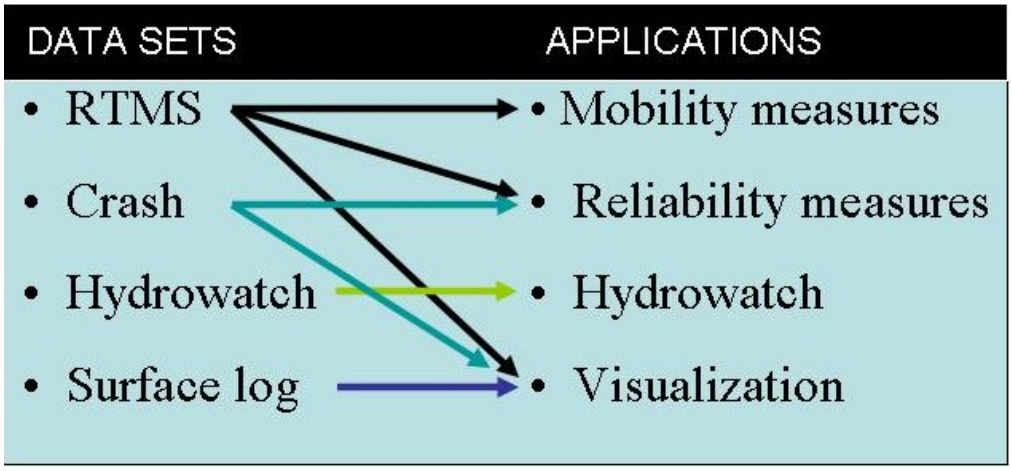


Figure 10. Correlation of Data Sets to Applications

Description of Architecture used for Conceptual Client Design

Precisely, the design used to test concepts for the client interfaces is illustrated in Figure 11. The mapping of table/attribute names to those in Figure 7 is straightforward. The final design is shown in Figure 7 and is a superset of the one shown in Figure 11.

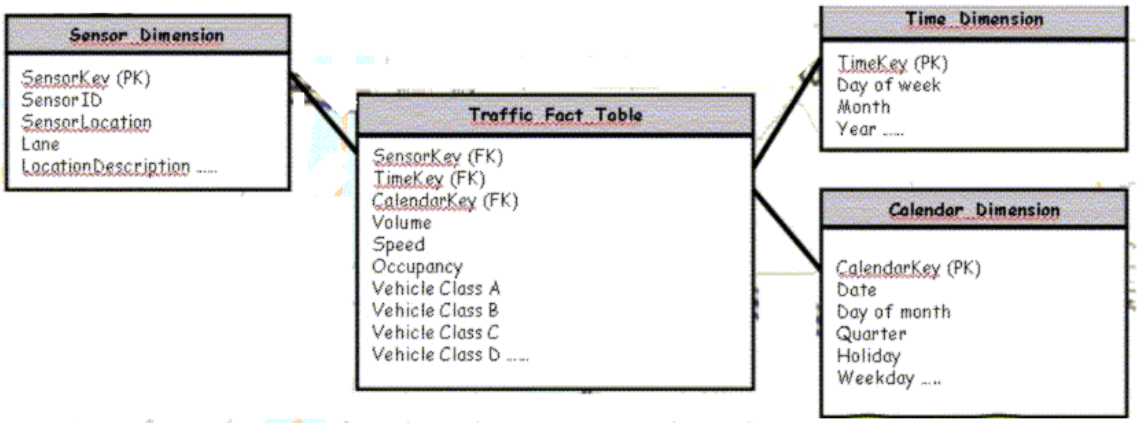


Figure 11. Star Schema Used for Client Experiments

A query over the warehouse of Figure 11 is shown in Figure 12. In fact, the results of this query are used to illustrate an application later in this section.

```
Select SensorID,
       Volume,
       Speed
From   Sensor_Dimension S,
       Traffic_Fact_Table F,
       Time_Dimension T
WhereS.SinsorKeys=F.SensorKey
       AND SensorLocation = 'Government St.'
       AND T.TimeKey=F.TimeKey
       AND T.Month='May'
       AND T.Year=2004
```

Figure 12. Illustrative Slice and Dice Query

Finally a version of the surface log describing road sections was assumed present on the client computer. The version used (and populated from the actual surface log) is shown in Figure 13.

Figure 13. Surface Log Tables

Simple Performance Measures

The first set of applications can simply be performed by a client having nothing more than a spreadsheet product on his/her computer and the skills to construct formulas and graphs using the product. Modify the query in Figure 12 to retrieve only a single day of volume data at each sensor together with the mileage between sensors (see the *LOCATION* table in Figure 7 not *SENSOR_DIMENSION* in Figure 11 for this) and one can easily produce graphics such as that in Figure 14.

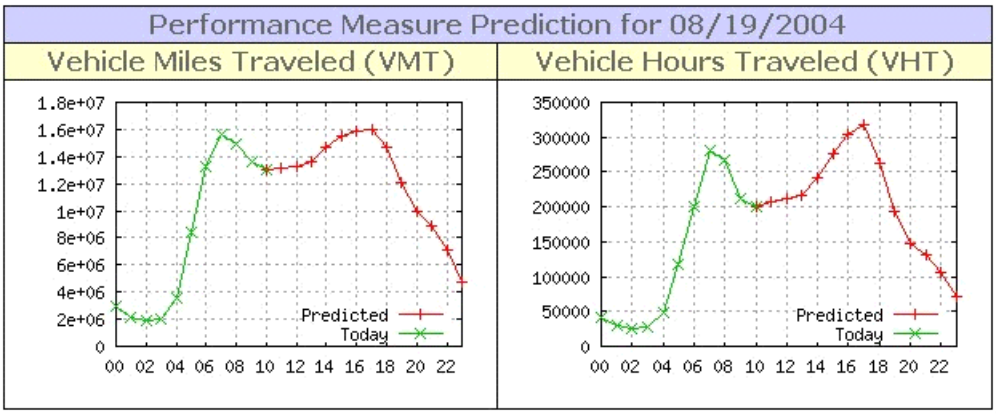


Figure 14. Simple Performance Measures (Virginia's Smart Travel Lab)

As noted, the figure is from Smart Travel Lab documents [60,72], not from the Baton Rouge data.

While the graphic is not ours, we performed this exercise over our design. An intermediate step is determining the volume both entering and leaving a segment (i.e., the number of vehicles completing a segment). Even on closed segments, we found the number entering and exiting seldom to be equal. From consultation with experts at the Minnesota data archival site, we learned that such errors have several causes but that the main one is vehicles being counted twice because the sensor caught them during a lane change.

Reliability Measures

Reliability measures generally require greater calculation ability than spreadsheet products provide. In addition, they sometimes require data (such as vehicle occupancy) that is not represented in the warehouse because sensors do not capture such data.

Using queries similar to that in Figure 12, then applying a measure of congestion,[4] the 10 most congested sites were selected. See Table 11. Based on these results, I-10 at Washington Street and I-10 EB at Perkins Road were identified as the most congested. These locations are approximately two miles apart.

Table 11. Top 10 Congestion Sites from Baton Rouge Sample Data

Rank	Location	Time	RVD Number	Percent Congested Travel	Average Volume
1	I-10 WB AT WASHINGTON ST.	17:15	41	82.5	1077
2	I-10 WB AT WASHINGTON ST.	16:45	41	81.9	1067
3	I-10 EB AT PERKINS ROAD	16:30	46	86.3	**1011**

[4] The measure of congestion is simply the lane occupancy. More sophisticated measures are discussed shortly. Recall that the objective is to test the warehouse design and not to fully implement one.

4	I-10 WB AT WASHINGTON ST.	17:00	41	82.4	**1042**
5	I-10 EB AT PERKINS ROAD	16:45	46	81.7	**1011**
6	I-10 WB WEST OF THE SPLIT	06:45	13	75.7	**1087**
7	I-10 EB AT PERKINS ROAD	15:30	46	79.5	**1018**
8	I-10 WB AT WASHINGTON ST.	16:30	41	78.0	**1030**
9	I-10 EB AT PERKINS ROAD	16:00	46	79.7	**1007**
10	I-10 WB AT WASHINGTON ST.	16:15	41	79.4	**1008**

Given these two sites, solely for the purpose of testing the design, we followed up by finding the time frames for which the combined congestion of the two sites is highest. This led to the discovery that 4:30pm and 4:45pm are the worst congestion times for the two sites combined. The results are shown in Table 12. While this is just a test of capabilities, this form of analysis is useful in determining where to pre-position resources such as police and tow trucks.

Table 12. Time Frame Analysis of Two Consecutive Congestion Points

Rank	Time	Mean Percent Congested Travel	Total Volume
1	16:45	81.8	2078
2	16:30	82.1	2041
3	17:15	79.8	2094
4	17:00	80.1	2043
5	16:15	79.3	2014

Moving from the specific tests we performed to more comprehensive measures, Table 13 contains a list of useful measures gleaned from a nationwide study by the Texas Transportation Institute [73]. The top three are reliability/mobility measures applicable to a segment of a traffic artery. The remaining six pertain to an entire system or significant portion thereof. While the formulas themselves are spreadsheet compatible, the basic measures may not be. For example, travel time is used in several measures but there is no data set being collected in Baton Rouge corresponding to travel times. There is a discussion of travel time estimates based on the data collected in the Capital Region in this report. For the moment, let us say that travel time estimation is possibly the most significant application for intelligent transportation systems but all successful attempts have included the ability to identify the same vehicle at different times within the traffic system.

Table 13. Reliability/Mobility Measures Recommended by Texas Travel Institute

$$\underset{(person-hours)}{Total\ Delay} = \left[\underset{(minutes)}{Actual\ Travel\ Time} - \underset{(minutes)}{FFS\ or\ PSL\ Travel\ Time} \right] \times \underset{(vehicles)}{Vehicle\ Volume} \times \underset{(persons/vehicle)}{Vehicle\ Occupancy} \times \frac{1\ hour}{60\ minutes}$$

$$Travel\ Time\ Index = \frac{\left[\frac{Freeway\ Travel\ Rate}{\begin{matrix} Freeway\ Free\text{-}flow\ or \\ Posted\ Speed\ Limit\ Rate \end{matrix}} \times \begin{matrix} Freeway\ Peak\ Period \\ VMT \end{matrix} \right] + \left[\frac{\begin{matrix} Principal\ Arterial\ Street \\ Travel\ Rate \end{matrix}}{\begin{matrix} Principal\ Arterial\ Street\ Free\text{-}flow\ or \\ Posted\ Speed\ Limit\ Rate \end{matrix}} \times \begin{matrix} Principal\ Arterial\ Street \\ Peak\ Period\ VMT \end{matrix} \right]}{\begin{matrix} Freeway \\ Peak\ Period \\ VMT \end{matrix} + \begin{matrix} Principal \\ Arterial\ Street \\ Peak\ Period\ VMT \end{matrix}}$$

$$Buffer\ Index\ (\%) = \left[\frac{\underset{(minutes)}{95th\ Percentile\ Travel\ Time} - \underset{(minutes)}{Average\ Travel\ Time}}{\underset{(minutes)}{Average\ Travel\ Time}} \right] \times 100\%$$

$$\underset{(vehicle\text{-}miles)}{Congested\ Travel} = \Sigma \left(\underset{(miles)}{Congested\ Segment\ Length} \times \underset{(vehicles)}{Vehicle\ Volume} \right)$$

$$Percent\ of\ Congested\ Travel = \left[\frac{\sum_{i=1}^{m} \left(\left(\underset{(minutes)}{Actual\ Travel\ Time_i} - \underset{(minutes)}{FFS\ or\ PSL\ Travel\ Time_i} \right) \times \left(\underset{(vehicles)}{Vehicle\ Volume_i} \times \underset{(persons/vehicle)}{Vehicle\ Occupancy_i} \right) \right)_{Each\ congested\ segment}}{\sum_{i=1}^{n} \left(\underset{(minutes\ per\ mile)}{Actual\ Travel\ Rate_i} \times \underset{(miles)}{Length_i} \times \underset{(vehicles)}{Vehicle\ Volume_i} \times \underset{(persons/vehicle)}{Vehicle\ Occupancy_i} \right)_{All\ segments}} \right] \times 100$$

$$\underset{(minutes\ per\ mile)}{Travel\ Rate} = \frac{\underset{(minutes)}{Travel\ Time}}{\underset{(miles)}{Segment\ Length}} = \frac{60}{\underset{(mph)}{Average\ Speed}}$$

$$\underset{(PMT)}{Person\text{-}miles\ of\ Travel} = \underset{(people)}{Person\ Volume} \times \underset{(miles)}{Distance}$$

$$\underset{(miles)}{Congested\ Roadway} = \sum \underset{Lengths\ (miles)}{Congested\ Segment}$$

$$\underset{(opportunities)}{Accessibility} = \sum \underset{(e.g.,\ jobs),\ Where\ Travel\ Time \le Target\ Travel\ Time}{Objective\ Fulfillment\ Opportunities}$$

Characteristics worth noting about individual measures in Table 13 include:

- Total Delay – can be computed either in vehicle miles or person-miles and is the sum of time lost on a specific segment due to congestion
- Travel Time Index – is dimensionless compares peak flow to free-flow
- Buffer Index – is a measure of trip reliability that expresses the amount of extra "buffer" time needed to be on time for 95 percent of all trips
- Congested Travel – estimates the extent of the traffic network that is affected by congestion
- Percent of Congested Travel – extends "congested travel" when additional information is available
- Travel Rate – is the rate at which a segment is traversed
- Person Miles – is the magnitude of travel on a section or several sections of a system
- Congested Roadway – is also a measure of the extent of congestion but is limited to one artery
- Accessibility – is a measure of the ability of the roadway system to fulfill a stated common good such as deliver people to jobs within a target travel time

Visualization and GIS Interface

Starting with queries very similar to that of Figure 12 and rolling up by hour (remember, the station dailies roll up by day), one can plot the graph for five sensors shown in the lower part of Figure 15. The vertical axis is in units known as the "travel time index" and is computed as a ratio of free-flow traffic to actual traffic speeds:

Travel Time Index = (60 mph)/(detected average speed)

(More sophisticated versions of the travel time index can be found in the literature [57, 58, 73]. Another version is given in Table 13.) A GIS map can now be attached showing the area of interest. It is possible to integrate the two by allowing a click on the sensor to display the diurnal graph associated with it. However, because this is simply to test the concept, the example was not elaborated further.

This example led to an important insight that affected the final design. The center of the area of interest was determined by the lat/long coordinates for the site. However, determining the limits of the map was done by trial and error. This led to linking the *LOCATION* records together so that the map limits could be automatically extracted.

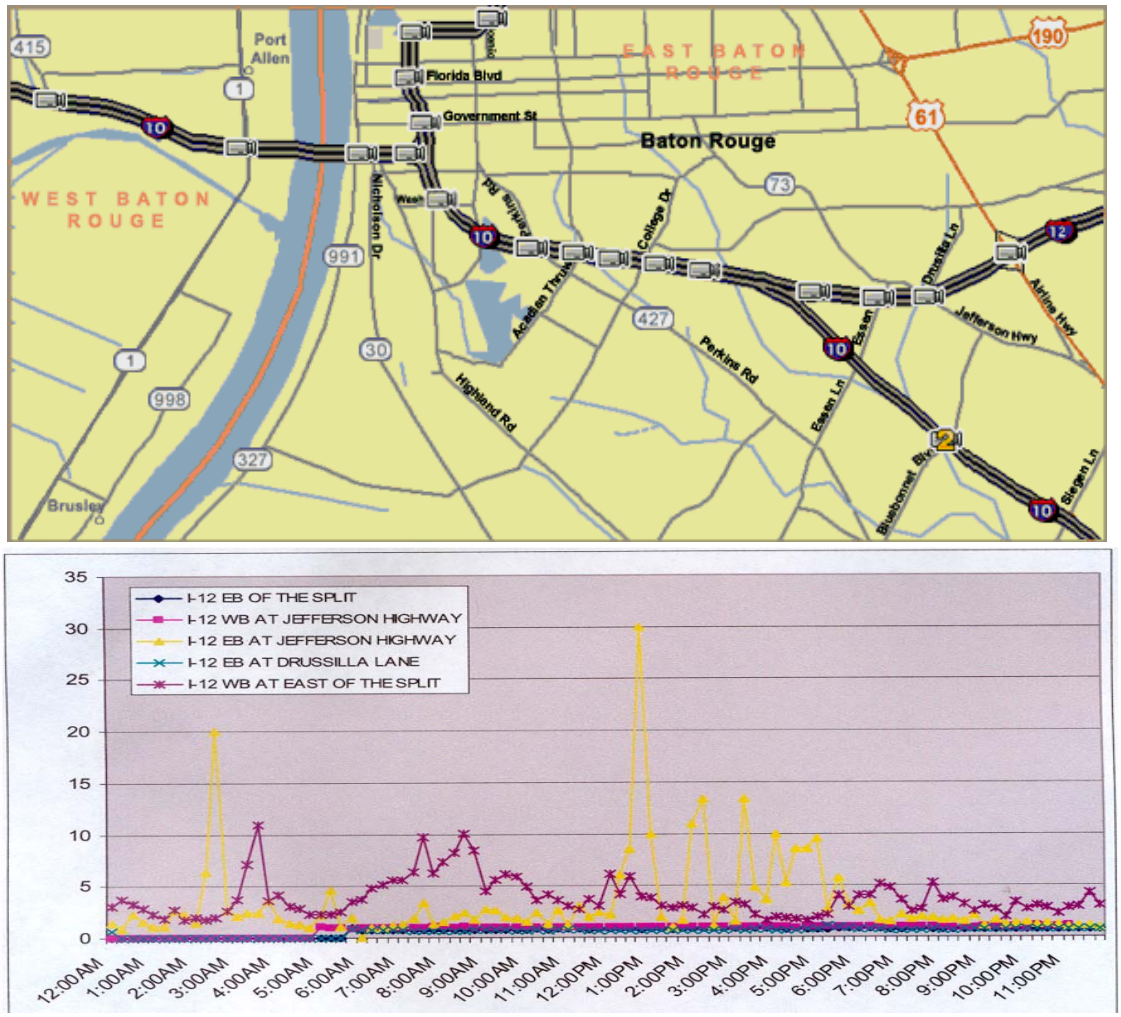


Figure 15. I-10 Congestion Points on a GIS Overlay

Sample Data Evaluation

The data source compilation focused on ITS freeway sensors and complement data from the below area. The following agencies were contacted within this area to inquire about data sources and warehouse inclusion.

1. DOTD Intelligent Transportation System Group at the TMC
2. Baton Rouge MPO
3. PB Farradyne
4. Econolite
5. Lucy Kimberly, Traffic Engineer DOTD
6. Peter Allain, Traffic Engineer DOTD
7. Information Technology, LADOTD, Baton Rouge
8. ITS and MIST, LADOTD, Baton Rouge
9. Baton Rouge 911
10. IBM- Managing Consultant Business Intelligence, CRM for DOTD
11. State Police-
12. USGS- Louisiana District, USGS, WRD

13. DOTD/Hydrowatch
14. City Police

The data acquired for testing and the sources are shown in Table 14.

Table 14. Sources for Baton Rouge Data

Data	Source	Region/Qty
Surface_log	DOTD IT	EBR/month
MPO Infras.	MPO	MPO/area
Crash Data	DOTD IT	EBR/month
ARAN (pave)	DOTD IT	EBR/month
RTMS	PBF/DOTD ITS	EBR/month
USGS Hydro-watch	USGS	EBR/month
MIST Sensor	PBF/DOTD ITS	EBR/month

Data requested but not obtained includes:

- State Police Incident and event data
- Information on 511
- City Police Incident Data
- 911

In addition we requested sample data sets for Average Daily Traffic (ADT) and Congestion Monitoring System (CMS) for the local New Orleans Regional Planning Commission for consideration into the warehouse design.

MARKETING PLAN

Recognize that the client community is divided into segments and that some segments are more likely than others to add value to the warehouse. Recognize also that the initial users must also be the most sophisticated. The segments and an assessment of their roles in the continuing development of the warehouse are listed below.

- Planners and traffic engineers
- University researchers
- General public further divided into
    - Media outlets
    - Motorists/travelers

By media, we include more than traditional broadcasters and news print distributors. In other jurisdictions, we have observed the rise of independent added-value contributors. These entrepreneurs develop web sites, in exchange for advertising, that provide travel information in alternate form.

Table 15. Market Segments and Their Roles

Market Segment	Role
Planners and traffic engineers	**Examining trends in specific corridors; tracking performance measures in general but specifically looking at before, during, and after conditions of an event such as adding a lane or relocating emergency services.**
Researchers	**Extracting data for independent studies such as answering "what if" questions; adding new applications, particularly data mining applications.**
Media outlets	**Providing web sites to customers that respond to needs such as the historical travel times for highly congested corridors; adding value to the collected data by making it available in formats and structures different from that in the "raw" database.**
Motorists and travelers	**Consume information including travel times and trends from secondary sources, primarily media including print, broadcast, and internet.**

Addressing the needs of each segment is not the approach to marketing the warehouse. The suggested approach is to encourage each segment to find the means to address its own needs given the core warehouse capabilities. Planners and traffic engineers are the most likely to add value to the warehouse as well as provide resources to both staff the warehouse and support university researchers. At the other end of the spectrum are the motorists and travelers. While important from the perspective of forming an important political constituency, reaching them should be objective of the media who gain direct benefit in the form of attracting advertising.

The approach to each market segment is described in Table 16.

Table 16. Market Segment Analysis

Market Segment	Analysis
Planners and Traffic Engineers	**Support from this segment must be addressed before the implementation begins, continue during the implementation, and be followed up after the implementation. The research team believes that, to some extent it has already acknowledged this issue via the formation of advisory and stakeholder committees. This structure should be maintained although the personnel involved may change over time. Including the design of the client phase is in part marketing and its implementation should be part of the next development phase. Following implementation, there should be intensive seminars to acquaint traffic engineers with using the warehouse. Always keep in mind that this group will essentially be responsible for the long-term growth of the warehouse.**
Researchers	**LTRC and LADOTD should encourage submission of problem areas that need study using the currently in-place forms and procedures. The agency should then follow up by soliciting proposals for worthy ideas then funding those proposals that meet the evaluation criteria.**
Media outlets	**This is the hardest group to address. One way is to start with the traffic consultants that share the Transportation Management Center. Kickstart the enterprise by using State funds to build an application that will assist them – perhaps a congestion map relating the current day/time to one day and one week earlier. Follow up by sending periodic (daily?) flyers to TV and print media that contains essentially the same information. The research team has no ideas at present concerning motivating entrepreneurs to construct secondary web sites.**
Motorists and travelers	**Reaching this segment is not particularly essential (or even desirable) at the outset. When the warehouse is stable and contains a sufficient history of traffic conditions the warehouse owner should make available a web site that conveys the contents. If there is an active 511 service at that time, it should be considered to integrate useful information from the warehouse.**

CONCLUSIONS AND RECOMMENDATIONS

The DW/DM research team reached the following conclusions and recommendations.

- A data warehouse should be implemented along the lines shown in this report. This includes adherence to the strict one-way data flow implied by the star schema and OLAP processing.
- The warehouse should be located in a central location and have uncontested ownership of the data it collects.
- The initial implementation should extend past the main storage phase and include the configuration of necessary tools, such as GIS, on client workstations.
- The initial implementation should consist of the minimal useful set of applications. These are primarily performance measures and their visualization. The warehouse should be considered a continuing, incremental project in which new applications are added singly. Data mining should be considered an application to be added after the base applications are operational.
- The warehouse should have a permanent staff consisting of at least a technician, a database analyst, and a computer programmer. The database analyst should double as the manager. A transportation engineer needs to be either a full-time or part-time staff member.
- The client community consists of planners/engineers, university researchers, and the general public. The general public is further segmented into media and individuals. There should be different marketing approaches addressed to each segment and subsegment.

REFERENCES

1. Lee, C.; Hellinga, B.; Saccomanno, F., "Real-time Crash Prediction Model for the Application to Crash Prevention in Freeway Traffic," Proc. 82nd Meeting of the Transportation Research Board, Washington D.C., Jan. 2003, pp. 67-78.
2. Lord D.; Persaud, B. N., "Accident Prediction Models with and without Trend: Application of the Generalized Estimating Equations (GEE) Procedure," Proc. 79th Meeting of the Transportation Research Board, Washington D.C., Jan. 2000.
3. Oh, J. S.; Richie, S. G.; Chang, M., "Real-time Estimation of Freeway Accident Likelihood," Tech Report # UCI-ITS-WP-00-21, University of California at Irvine, Dec. 2000.
4. Parrish, A. S.; Dixon, B.; Cordes, D.; Vrbsky, S.; Brown, D., "CARE: An Automobile Crash Data Analysis Tool," IEEE Computer. Vol. 36, No. 6, June 2003, pp. 22-30.
5. Wu; Chun-Hsin; Ho, Jan-Ming; Lee, D.T., "Travel-time Prediction with Support Vector Regression," *IEEE Trans. on Intelligent Transportation Systems*, Vol. 5, Issue 4, Dec. 2004, pp. 276 - 281.
6. Rice, J.; van Zwet, E., "A Simple and Effective Method for Predicting Travel Times on Freeways," *IEEE Trans. on Intelligent Transportation Systems*, Vol. 5, Issue 3, Sept. 2004, pp. 200 - 207.
7. Zhao, Yilin, "Mobile Phone Location Determination and its Impact on Intelligent Transportation Systems," *IEEE Trans. on Intelligent Transportation Systems*, Vol.1, Issue 1, Mar 2000, pp. 55 - 64.
8. Smith, B. L.; Scherer, W. T.; Hauser, T. A., "Data Mining Tools for the Support of Traffic Signal Timing Plan Development," Transportation Research Record 1968, 2001, pp. 141-147.
9. Smith, B. L., "Software Development Cost Estimation for Infrastructure Systems," *ASCE Journal of Management in Engineering*," Vol. 18, No. 3, July 2002, pp. 104-110.
10. Najm W.; Sen B.; Smith J.; Campbell, B., "Analysis of Light Vehicle Crashes and Pre-Crash Scenarios Based on the 2000 General Estimates Systems," U.S. Department of Transportation, Feb. 2003.
11. Jones, W. D., "Keeping Cars from Crashing," IEEE Spectrum, Vol. 38, Sept. 2001, pp.40-45.
12. Chisalita, I.; Shahmehri, N.; Lambrix, P., "Traffic Accidents Modeling and Analysis using Temporal Reasoning," In Proceedings of the 7th IEEE Conference on Intelligent Transportation Systems (ITSC2004), October, 2004, pp. 378-383.
13. Mondelo P.R.; Sorin J., Terres de Ercilla, F., "An Analysis of Traffic Accidents Using Official Records," International Conference on Computer-Aided Ergonomics and Safety, Hawaii, USA, Aug, 2001.
14. NHTSA, U.S. Department of Transportation, National Center for Statistics & Analysis, Crashworthiness Data System (CDS), 2003.

15. Wells, S.; Mullin, B.; Norton, R.; Langley, J; Connor, J.; Jackson, R., "Motorcycle Rider Conspicuity and Crash Related Injury: Case Control Study," *British Medical Journal,* doi: 10.1136/ bjm.37984.574757.EE, April 10, 2004.

16. Davis, G.; Pei, J., Bayesian Networks and Traffic Accident Reconstruction. ICAIL, Edinburgh, Scotland, UK, June 28, 2003.

17. Bonneson, J.A., "A Kinematic Approach to Horizontal Curve Transition Design. Geometric Design and Effects on Traffic Operations," Transportation Research Record 1737, 2000, pp. 1-9.

18. Zegeer, C.V.; Hummer, J.; Herf, L.; Reinfurt, D.; Hunter W., "Safety Effects of Cross-Section Design for Two-Lane Roads," Report No. FHWA-RD-87-008, Federal Highway Administration, Washington, D.C., 1986.

19. Golob, T., Regan, A., "Traffic Conditions and Truck Accidents on Urban Freeways," UCI-ITS-WP-04-03 presented at the Annual Meeting of the Transportation Research Board, July 2005.

20. Scheines, R., "An Introduction to Causal Inference," in V. McKim and S. Turner (eds.), Causality in Crisis? University of Notre Dame Press, 1997, pp. 185–99.

21. Spirtes, P.; Glymour, C.; Scheines, R., *Causation, Prediction, and Search*, Springer Lecture Notes in Statistics, no. 81, New York: Springer-Verlag. 2nd ed., Cambridge, Mass.: MIT Press, 2000.

22. Kockelman, K. M.; Ma, J., "Freeway Speeds and Speed Variations Preceding Crashes, Within and Across Lanes," Proc. Ann. Meeting of the Transportation Research Board, Washington D.C., Jan. 2004.

23. Solomon, D., "Accidents on Main Rural Highways Related to Speed, Driver and Vehicle," US Department of Commerce & Bureau of Public Roads, Washington D.C., 1964.

24. Lave, C., "Speeding, Coordination, and the 55-mph Limit," The American Economic Review," 75 (5), 1985, pp. 1160-1164.

25. Davis, G.A., "Is the Claim that 'Variance Kills' an Ecological Fallacy?" Accident Analysis and Prevention 34 (3), 2002, pp. 343-346.

26. Golob, T., Recker, W., "A Method for Relating Type of Crash to Traffic Characteristics on Urban Freeways," Paper presented at the 82nd Annual Meeting of the Transportation Research Board. Washington, D.C., January 12-16, 2003.

27. Lee, C.; Saccomanno, F.; Hellinga, B., "Analysis of Crash Precursors on Instrumented Freeways," *Transportation Research Record*, Vol. 1768, 2002, pp. 1-8.

28. Dai, H.; Korb, K.B.; Wallace, C.S.; Wu, X. "A Study of Casual Discovery with Weak Links and Small Samples," Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, August 1997, pp. 1304-1309.

29. Bai, X.; Glymour, C.; Padman, R.; Ramsey, J.; Spirtes, P., PCX: Markov Blanket Classification for Large Data Sets with Few Cases, Technical Report CMU-CALD-04-102, 2004.

30. Aliferis C.F.; Tsamardinos, I.; Statnikov A., "HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection," Proc. Annual Symposium Medical Informatics, Washington DC, Nov. 2002, pp. 21-25.

31. Margaritis, D.; Thrun, S., "Bayesian Network Induction via Local Neighborhoods," Advances in Neural Information Processing System (NIPS) 12, 2000, pp. 505-511.

32. Mattison, R., *Data Warehousing, Strategies, Technologies and Techniques.* ISBN 0-07-041034-8, McGraw-Hill, 1996.

33. Han, J.; Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA. 2001.

34. Inmon, W. *Building a Data Warehouse*, John Wiley and Sons, New York, 1996.

35. Minnesota Department of Transportation – MNDOT. Twin Cities Metropolitan Area Traffic Data. <http://www.dot.state.mn.us/tmc/trafficinfo/data/index.html>

36. TDRL – Traffic Data Research Laboratory, University of Minnesota Duluth. <http://tdrl1.d.umn.edu/contact.htm>

37. Turner, S.; Eisele, W.; Benz, R; Holdener, D., "Travel Time Data Collection Handbook,"

    Federal Highway Administration, Report FHWA-PL-98-035 March, 1998.

38. Miwa, T.; Morikawa, T., "Analysis on Route Choice Behavior Based on Probe-Car Data," Proc. of the 11th World Congress on ITS, 2003.

39. Nakata, T.; Takeuchi,J., "Mining Traffic Data from Probe-Car System fro Travel Time Prediction," In Proceedings of the KDD'04 Conference, Seattle, Washington, August 2004.

40. Shuldiner, Paul W.; D'Agostino, Salvatore A.; Woodson, Jeffrey B., "Determining Detailed Origin-Destination and Travel Time Patterns Using Video and Machine License Plate Matching," In Transportation Research Record 1551. TRB, National Research Council, Washington, D.C., 1996, pp. 8-17.

41. Han, Lee; Wegmann, Frederick J.; Chatterjee, Arun, "Using License Plate Recognition Technology for Transportation Data Collection," Urban Transportation Data Committee, TRB Annual Meeting 2001, University of Tennessee, Knoxville

42. Coifman, B.; Cassidy, M. "Vehicle Reidentification and Travel Time Measurement on Congested Freeways," Transportation Research: Part A, 2002, Vol 36, No. 10, 2002, pp. 899-917

43. Zhang, H., Kwon, E., "Travel Time Estimation on Urban Arterials Using Loop Detector Data," ISBN 0-87414-141-9. Transportation Research Board, 1997.

44. Petty, K.; Bickel, P.; Jiang, J.; Ostland, M.; Rice, J.; Ritov, Y.; Schoenberg F., "Accurate Estimation of Travel Times from Single-loop Detectors," *Transportation Research Record* 971043, 1997.

45. Oh, J.; Jayakrishnan, R.; Recker, W., "Section Travel Time Estimation from Point Detection Data," Proceedings of the 82nd Annual Meeting of the Transportation Research Board, Washington, D.C., January, 2003.

46. Bhoite, D., Crouch, C., Crouch, D., Maclin, R., Final Report for the NATSRL EOP Program,

University of Minnesota Duluth, July 2004.

47. Shekhar, S.; Lu, C. T.; Zhang, P.; Liu, R., "CubeView: A System for Traffic Data Visualization," In Proceedings of the Fifth IEEE International Conference on Intelligent Transportation Systems, 2002.

48. Chen, C., Freeway Performance Measurement System (PeMS), UCB-ITS-PRR-2003-22. Ph.D. Dissertation, EECS Department at the University of California, Berkeley. Fall, 2002.

49. Courage, K.; Hammer, F.; Ji, F.; Yu, Q., Feasibility Study for an Integrated Network of Data Sources. Final Report, Contract BC-354-61. University of Florida, Transportation Research Center, January 2004.

50. Sun, C.; Ritchie, S.G.; Tsai, W.; Jayakrishnam, R., "Use of Vehicle Signature Analysis and Lexicographic Optimization for Vehicle Reidentification on Freeways," *Transportation Research Record*, 7C, 1999, pp. 167-185.

51. Coifman, B.; Ergueta, E., "Improved Vehicle Reidentification and Travel Time Measurement on Congested Freeways," *ASCE Journal of Transportation Engineering*, Vol. 129, No 5, 2003, pp 475-483.

52. Smith, B. L.; Smith, B. E. "Configuration Management in Transportation Management Systems," *Transportation Research Record* 1748, 2001, pp. 103-109.

53. Smith, B. L.; Miller, J. S.; Revels, B. M.; Smith, K. W. "Planning for civil engineering applications of information technology," *ASCE Journal of Management in Engineering*, Vol. 17, No. 2, 2001, pp. 86-94.

54. Smith, B. L.; Scherer, W. T., "Developing Complex Integrated Computer Applications and Systems," *ASCE Journal of Computing in Civil Engineering*, Vol. 13, No. 4, October 1999, pp. 238-245.

55. Smith, B. L.; Scherer, W. T. "Development of Integrated Intelligent Transportation Systems," *Transportation Research Record* 1675, 1999, pp. 84-90.

56. Eisele, W. L.; Lomax, T. J.; Gregor, B. J.; Arnold, R. D., "Developing and Implementing Statewide Operations Performance Measures in the State of Oregon: Methodology and Application for Using HERS-ST and Archived Real-time data," 85nd Annual Meeting of Transportation Research Board, Washington D.C., Jan. 2005.

57. Lomax, T. J.; Turner, S. M.; Marigiotta, R. M., "Monitoring Urban Roadways in 2002: Using Archived Operations Data for Reliability and Mobility Measurement," Published as a Federal Highway Administration report. Report No. FHWA-HOP-04-001, 2002.

58. Turner, S. M.; Albert, L. P., "ITS Quality Control and the Calculation of Mobility Performance Measures," Texas Transportation Institute Report 1752-5, Sept. 2000.

59. Oh, S.; Ritchie, S.; Oh, C., Real Time Traffic Measurement from Single Loop Inductive Signatures. UCI-ITS-WP-01-15. Institute of Transportation Studies, University of California, Irvine. <http://www.its.uci.edu/its/publications/papers/WP-01-15.pdf>

60. Babiceanu, S.; Smith, B. L.; Lu, X.; Ngov, T.; Venkatanarayana, R., "Developing Efficient Data Archive Designs for the State of Virginia," *Transportation Research Record* 1727, Transportation Research Board, Washington, D.C., January 2003.

61. Kwon, J.; Varaiya, P.; Skabardonis, A., "Estimation of Truck Traffic Vol. from Single Loop Detector Using Lane-to-Lane Speed Correlation," 83nd Annual Meeting of Transportation Research Board, Washington D.C., Jan. 2003.

62. Wang, Yinhai; Nihan, Nancy L., "Freeway Traffic Speed Estimation Using Single Loop Outputs," accepted for publication by Transportation Research Record, January 2000

63. Avery, R.; Wang, Y.; Nihan, N., "Estimating Freeway Traffic Speeds From Single Loops Using Region Growing," Presented at the TransNow Student Conference at Portland State University, November 2004

64. Chu, T.; Danks, D.; Glymour, C., "Data Driven Methods for Granger Causality and Contemporaneous Causality with Non-Linear Corrections: Climate Teleconnection Mechanisms," Carnegie Mellon University. 2004. <http://www.hss.cmu.edu/philosophy/glymour/chudanksglymour2004.pdf>

65. Heckerman, D., "A Bayesian Approach to Learning Causal Networks," Technical Report MSR-TR-95-04, Microsoft Research. Advanced Technology Division, Microsoft Corporation, March 1995.

66. Freedman, D., "On Specifying Graphical Models for Causation, and the Identification Problem," Technical Report #601, Department of Statistics, UC Berkeley, May 2004.

67. Pearl, J., *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann. 1988.

68. Pearl, J., "Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning," Proc. Cognitive Science Society, Irvine CA, pp.329-334, 1985.

69. Pearl, J., *Causality – Models, Reasoning and Inference*. Cambridge, United Press, 2000.

70. Freedman, D., "Are There Algorithms that Discover Causal Structure?" Department of Statistics, UC Berkeley, June 1998. <www.stanford.edu/class/ed260/freedman514.pdf>

71. Dash, D.; Druzdzel, M., "A Hybrid Anytime Algorithm for the Construction of Causal Models from Sparse Data," Proc. of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99), San Francisco, 1999, pp. 142-149.

72. Turochy, R.E.; B. L. Smith, **"**A New Procedure for Detector Data Screening in Traffic Management Systems**,"** Transportation Research Record 1727, Transportation Research Board, Washington, D.C., Dec. 2000, pp. 127-131.

73. Texas Transportation Institute, "The Keys to Estimating Mobility in Urban Areas: Applying Definitions that Everyone Understands," May 2005.

74. Chen, L.; May A. "Traffic Detector Errors and Diagnostics." In *Transportation Research Record 1132*, TRB, National Research Council, Washington D.C., 1987, pp. 82-93.

75. Coifman, B., "Using Dual Loop Speed Traps to Identify Detector Errors," In *Transportation Research Record 1683,* TRB, National Research Council Washington D.C., 1999, pp. 47-58.

76. Coifman, B.; Dhoorjaty, S., "Event Data Based Traffic Detector Validation Tests," Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.

77. Jacobson, L.N.; Nihan, N.L.; J.D. Bender, "Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System," In *Transportation Research Record* 1287, TRB, National Research Council, Washington D.C., 1990, pp. 151-166.

78. Cleghorn, D.; Hall, F. L.; Garbuio, D., "Improved Data Screening Techniques for Freeway Traffic Management Systems," In *Transportation Research Record 1320*, TRB, National Research Council, Washington D.C., 1991, pp. 17-23.

79. Payne, H. J.; Thompson, S., "Malfunction Detection and Data Repair for Induction-Loop Sensors Using I-880 Data Base," *Transportation Research Record* 1570. Paper No. 971113, January 1997, pp 191-201.

80. Turochy, R.E.; Smith, B., "A New Procedure for Detector Data Screening in Traffic Management Systems," In *Transportation Research Record* 1727, TRB, National Research Council, Washington D.C., 2000, pp. 127-131.

81. Peeta, S.; Anastassopoulos, I., "Automatic Real-Time Detection and Correction of Erroneous Detector Data Using Fourier Transforms for On-Line Traffic Control Architectures," Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.

82. Ishak, S., "Quantifying Uncertainties of Freeway Detector Observations Using Fuzzy Clustering Approach," TRB 03-2056, Transportation Research Board 82nd Annual Meeting Washington, D. C. 2003.

83. Chen, C., "Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems," TRB 03-3507, 82nd Annual Meeting Transportation Research Board, January 2003 Washington, D.C.

84. Wall, Z.; Dailey, D. J., An Algorithm for the Detection and Correction of Errors in Archived Traffic Data. TRB 03-4184, Annual Meeting Transportation Research Board Washington, D.C., 2003.

85. Chilakamarri V.S.R.C.; Al-Deek, H. M., "Revised New Algorithms for Filtering and Imputation of Real Time and Archived Dual-Loop Detector Data," TRB 2004 Annual Meeting, Paper No. 04-3505.

86. Nihan, N., "Aid to Determining Freeway Metering Rates and Detecting Loop Errors," *Journal of Transportation Engineering*, ASCE, Vol. 123, No 6, November/December 1997, pp. 454-458.

87. Klein, L. A., *Data Requirements and Sensor Technologies for ITS*, Norwood, MA, Artech House, 2001

88. Hall, F.; Persaud, B. N., "An Evaluation of Speed Estimates Made with Single-detector Data from Freeway Traffic Management Systems," In *Transportation Research Record*, No. 1232, 1989, pp 9-16.

89. Walpole, R.E.; Myers, Y.E. *Probability & Statistics for engineers and Scientists*, seventh edition, Prentice Hall, New Jersey, 2002.

90. Bertini, Robert L.; Leal, Monica; Lovell, David, "Generating Performance Measures from Portland's Archived Advanced Traffic Management System Data," Transportation Research Board Annual Meeting, November, 2001.

APPENDIX A. CURRENT PRACTICE

Prominent ITS sites, TMCs, and travel labs surveyed include: Washington State Transportation Center (TRAC), Virginia Smart Travel Lab (STL), Houston TranStar, Georgia Department of Transportation, Maricopa County Arizona and California CalTrans. The latter maintains the Performance Measures System (PEMS). Actual site visits occurred to each of the sites with the exception of TRAC. Meetings with TRAC representatives were held at the Transportation Research Board meeting in Washington, D.C. and the NAMTEC conference in San Diego, CA. In addition to visiting the sites, several peripheral meetings, conference calls, and professional society symposia were arranged or attended. These include:

1. LADOTD Intelligent Transportation System Group at the TMC
2. The Transportation Research Board (TRB) meeting Jan 11-15, 2004
3. Chief of Planning Capital Region Planning Commission - Huey Dugas
4. PB consultant - Elizabeth Delaney
5. Econolite - Paul Misticawi
6. The North American Travel Monitoring Exposition and Conference (NAMTEC)
7. Information Technology, LADOTD, Baton Rouge
8. ITS and MIST, LADOTD, Baton Rouge
9. Baton Rouge 911- Ralph Ladnier
10. IBM - Casey Adams, Managing Consultant Business Intelligence, CRM
11. State Police- Captain Jim Mitchell & Colonel Henry L. Whitehorn
12. USGS - David Walters, Data Management Supervisor
    Louisiana District, USGS, WRD
13. DOTD/Hydrowatch - George Gele
14. Regional Planning Commission- Walter Brooks, Lynn Dupont and Johnny Bordelon

Additionally, a Stakeholders group was founded that included:

i. Lucy Kimberly, Traffic Engineer DOTD
ii. Peter Allain, Traffic Engineer DOTD
iii. Stephen Glascock, ITS DOTD
iv. Carryn Zeagler, ITS DOTD
v. George Gele, Hydrowatch DOTD
vi. Huey Dugas, Chief of Planning Capital Region Planning Commission
vii. Ingolf Partenheimer, ATM EBR (Rep. Jason Taylor)

Minutes from these meetings can be viewed on the web site http://129.81.132.174/ITSDW/.

The remainder of this section focuses on the prominent site surveys. The section concludes with a comparative analysis of all labs examined. Each review is structured around the following elements:

- System overview/design
- Warehouse applications/performance measures
- Data collected, stored and formats
- Data metrics
- System operation, management and costs

Essential information is then combined to present the data model and design issues suggested by these sites in the comparative summary. In some cases, a true data warehouse does not exist and the site simply archives data into a traffic management center. The data is used specifically for compiling static traffic reports. When a warehouse is employed, the description provides what it contains, how the data are cleansed for quality control, and which applications and performance measures are derived.

Site Survey Summary and Comparison

Storing and analyzing the data are not free. However, a large number of potential users exist for the information that the surveillance system generates. The key is to work with potential users to fund the modest costs of storing, analyzing, and reporting the data already collected. The agency must also determine *who* will operate the database. As this work gets under way, it is important to recognize that not all surveillance data are reliable. Therefore, analytical procedures must be prescribed that identify and handle "unreliable" data. Mechanisms should also be in place to repair and calibrate unreliable sensors. (After all, unreliable data also hinder the operational control decisions that are based on those data.) Because most traffic management systems have limited equipment maintenance budgets, repair activities have to be prioritized. A key to consider when balancing cost versus data availability is that obtaining useful performance information does not require *all* detectors to be operating. (Does an agency *really* need to report volumes based on continuous data collection at 300 locations in the urban area, or will 12 to 20 sites spread strategically around the region reveal the important facts?) The reality is that necessary data can be obtained with a moderate amount of planning and cooperation.

Origins of TRAC, PeMS, and STL

TRAC (Seattle, Washington), PeMS (California), and STL (Virginia) are three successful traffic data warehouse/data mining systems. Each started with limited resources, each currently serves as a basis for both traffic operations, and each today is supported by competitive funding. Most funding is from the respective state DOTs but presence of the centers enable them to attract some federal funding from the Federal Highway Administration (FHWA) and (in the case of STL) the National Science Foundation.

TRAC began the earliest (1983) and its vision evolved over time. Its form is not that of a traditional data warehouse but it functions as such. It began as a project, guided by a Civil Engineering professor, to rejuvenate traffic research at the University of Washington. Adding one additional professor the next year, the activity began attracting state support, project by project. Both PeMS and STL (Smart Travel Lab) began in the late 1990s with visions that largely reflect what they have become today. PeMS began as a partnership between one post-doctoral researcher in Civil Engineering at the University of California and a private start-up (Berkeley Transportation Systems, Inc.). There was a significant infusion of cash for both theoretical study and implementation, split between the university and the company, at the beginning. Particularly relevant is a large investment in infrastructure such as computers and storage. The first version was operational in 1999. STL began at about the same time with one Civil Engineering professor at the University of Virginia with the part-time participation of a Systems Engineering professor. The state's investment was $55,000 and the university contributed space. Also, several students participated who were supported by various grants in the area of traffic engineering.

Today, PeMS is operated largely independent of the University of California at Berkeley where it began. In the TRAC center, state and university employees work side-by-side. STL is largely a university-based enterprise but one of the two directors is a state employee. Table 17 compares the present staffing of the centers. PeMS is a bit ambiguous because it is a commercial operation that manages traffic warehousing in several states as well as nine California districts.

Table 17. Staffing of TRAC, STL, and PeMS

TRAC	STL	PeMS
Four engineers	Four faculty members	Traffic engineer
"several" staff members	Two post-docs	Network administrator
Receptionist (part-time)	Five Ph.D. students	Database administrator
Accountant	Network administrator	"several" staff members, including programmers
Database administrator	Database administrator	
Editor (part-time)	Programmer	
Graphics artist (part-time)	Traffic engineer	

There are two common threads in the origins and present operations of these three centers. First, each began with one highly dedicated, full-time professional, although each center director said two would have been better. The second common thread is that there is both a "push" and a "pull" to the technology they are developing. On one hand, the state approaches them with problems and they attempt to respond with an application that, if not a solution, provides information to guide decision-makers. On the other hand, each looks for opportunities and unmet needs, prepares a proposal, and then presents the proposal to the state or other agency.

Washington State Transportation Center (TRAC)

System Overview/ Design

The primary purpose of TRAC at Washington State is to encourage research in all aspects of transportation. TRAC is able to marshal the resources of the University of Washington (UW), Washington State University (WSU), and the Washington State Department of Transportation (WSDOT) to tackle the transportation problems of the region and the nation. TRAC maintains an archive of data from the Puget Sound freeway and ramp-metering program and uses the data for a variety of analytical purposes. TRAC monitors regional freeways primarily in the Seattle metropolitan area and extending into the Puget Sound region

The U.S. Department of Transportation, WSDOT, and other Washington State partners have invested in the development of an architecture and infrastructure for a Puget Sound intelligent transportation systems (ITS) backbone. This backbone has been used to obtain traffic data and traveler information from disparate sources, combine those data, and make them available over a standard interface to transportation-related organizations and the public. In this way it supports existing traveler information applications for both traffic and transit information, real-time access to WSDOT data by a variety of public and private groups, research activities within WSDOT and at universities and agencies nationwide, and provides a standard way to include new data sources into the existing traffic management system. The TRAC center supports continuing personnel, equipment, maintenance, software, and communications links for the ITS backbone, as well as associated applications.

One present project, TDAD (Traffic Data Acquisition and Distribution) has the potential to impact future ITS backbone installations in Louisiana. TDAD is funded by the FHWA. Paraphrasing the published project description, the TDAD project has the goal of integrating in a general fashion the wide variety of remote sensors used in Intelligent Transportation Systems (ITS) applications (loops, probe vehicles, radar, cameras, and so on). The variety of sensors has created a need for general methods by which data can be shared among agencies and users who own disparate computer systems. TRAC, via this project, will present a methodology that demonstrates that it is possible to create, encode, and decode a self-describing data stream using:

1. Existing data description language standards
2. Parsers to enforce language compliance
3. A simple content language that flows out of the data description language
4. Architecture neutral encoders and decoders based on ASN.1.

Several organizations in the Seattle, WA area use the traffic data in any of five ways: (1) reporting, (2) long-term planning, (3) project planning, (4) performance monitoring, or (5) research. Specific applications are identified in the paragraphs that follow.

Warehouse Applications/Performance Measures

The ITS Backbone performs several important tasks for the ongoing efforts at WSDOT and UW. For example, the Backbone (1) supports existing traveler information applications for both traffic and transit information, (2) supports real time access to WSDOT data for a variety of public and private groups, (3) off-loads the interaction and support of data users external to WSDOT, (4) provides a standard interface so that all roadway data are available equally to outside agencies/groups, (5) supports research activities within WSDOT - research that is funded by WSDOT at the UW as well as research at universities and agencies nationwide, and (6) provides a standard interface to include new data sources into the existing TMS System. The latter is the Traffic Data Acquisition and Distribution (TDAD) project described above.

Collected data are used to answer questions and address issues such as:

1. Are HOV lanes being used? Do travel time incentives increase HOV use? Should the operational rules for HOV be changed (e.g., should 24/7/365 rules be changed to an appropriate part-time operation)?

2. Use of CVISN tags to compute inter-city travel times.

3. Develop/adjust the ramp metering algorithm. The ramp metering algorithm now used for all of the Puget Sound region's freeway ramp meters is touted as the most advanced in the country. Developed by TRAC researchers, it helps smooth traffic flow and reduce freeway congestion daily.

WSDOT collects vehicle volumes (per lane) by time of day in both the HOV and general purpose lanes. They present this information in a graph of *Vehicle volumes (per lane) by time of day* described below. Center personnel perform this analysis weekly. On the basis of these graphics, WSDOT can determine whether capacity exists in the general purpose lanes, whether sufficient demand exists for HOV lanes, and whether growth in HOV lane use is meeting public policy goals.

The basic volume-by-time-of-day graphic can be extended to illustrate when congestion occurs and its effect on vehicle speed and throughput. First, average speed is color coded to indicate the manner in which conditions routinely change by time of day. Then, because conditions vary considerably from day to day, reliability at selected points in the roadway can be examined by defining "congestion" {in this case, the occurrence of Level of Service "F" Segments (LOS F conditions)} and reporting on the frequency with which that congestion occurs. Graphically, it is possible to lay the "frequency of congestion" over the same graphic that illustrates vehicle volumes and average speeds.

To summarize the visualizations used by TRAC:

1. *Vehicle volumes (per lane)* (y-axis) *by time of day in hours* (x-axis*).* This is a two-line graph. One line is the average across the general purpose lanes and the other depicts HOV activity. This graph shows volume variability over time. The coverage is typically one twenty-four hour period.

2. *Trends in vehicle volumes (per lane)* (y-axis) *by time of day in hours* (x-axis)*.* The four lines in this graph depict two separate years of data for the general purpose lanes and two separate years of data for the HOV lane. The coverage is typically one twenty-four hour period for each year.

3. *Estimated frequency of congestion, volumes and speeds.* The left side y-axis is vehicles per lane per hour (VPLPH), the x-axis is hours of the day and the right side y-axis is congestion frequency. This graph demonstrates peak hours of congestion.

4. *Estimated frequency of congestion for GP lanes, with volumes for both general purpose and HOV lanes.* This graph is similar to number 3 above but overlays the HOV volume to demonstrate that during the peak period, HOV lane vehicle volumes exceed general purpose vehicle volumes (per lane).

5. *Estimated frequency of congestion, volumes and speeds for general purpose and HOV lane.* This graph is similar to number 3 above but overlays the HOV volume to demonstrate that by adding in car occupancy and transit ridership data, it is possible to show relative person throughput which is a key statistic for responding to the public policy debate about the use of HOV lanes.

6. *Person and vehicle throughput per lane, general purpose and HOV lanes.* A bar graph used to compare HOV and general purpose lanes. Statistics are reported for each 3-hour morning peak period.

7. *Total person and vehicle throughput, four general purposes and one HOV lane.* A bar graph similar to number 6 above except each general purpose lane is shown separately. It is used to compare HOV and general purpose lanes. Statistics are reported for each 3-hour morning peak period.

8. *Travel times (by time of day) for a specific route.* The left side y-axis is the estimated average travel time (hour:min) given a trip start time (x-axis). The right side y-axis is congestion frequency (speed < 35 mph). Using vehicle speed data that can be obtained from the freeway surveillance system, it is possible to estimate vehicle travel times throughout the day. These statistics also lead to more informed discussion of the travel conditions that exist (e.g., how bad is off-peak congestion? Is off-peak operation of the service patrol program necessary?)

The various performance measures plotted in the graph are those promulgated by the Texas Transportation Institute (TTI) or are measures directly based upon them.

Data Collected, Data Stored and Data Formats

The Self-Describing Data (SDD) format developed via the TDAD project is TRAC's approach to transmitting and delivering data. Thus incoming data at the ETL (extraction, transformation, and loading) phase is a stream prefixed with metadata that "describes" it. The descriptive data assists users in understanding the incoming data or assists

application code writers in generalizing software. An example data stream is the traffic loop sensor data for which over 2,000 sensors generate new data every 20 seconds. The descriptive data for this stream consists of descriptions of the sensors, their cabinets, their locations, and so on. This preface plus the continuous actual data stream create an SDD stream. SDD is a new approach to data packaging and transmission that requires training on the part of practitioners, and is an approach that is still evolving. SDD, in essence, is the data transmission protocol for the ITS backbone infrastructure.

The surveillance system collects data on vehicle volumes and estimates of lane occupancy by location. These data are then converted into estimates of vehicle speed and travel time. An analysis process developed by TRAC produces facility performance information based on these data. This process also fuses the basic freeway surveillance data with separately collected transit ridership and car occupancy data to estimate person throughput. The Traffic System Management Center (TSMC) collects data from the sensor loops in the freeway every 20 seconds. The output of WSDOT's Traffic Management System (TMS) is prodigious. It contains a dictionary component, which describes the name, position, and type of each of the approximately 5,000 sensors. This dictionary is key to understanding and applying the sensor block component, which contains current measurements taken from each of the 5,000 sensors and is newly generated every 20 seconds. The dictionary is modified as necessary to reflect the effects of road construction or installation of new sensors.

The constituent tables of the database are as follows:

1. COORDINATES - Describes coordinate data types, such as "geodetic"; their measurements, such as "longitude" and "latitude"; and their units of measure, such as "degrees,"
2. MEASURES - This provides additional data on coordinate data types, such as the fact that WSDOT uses NAD23 coordinate referencing, while UW uses NAD89.
3. STATION_FLAGS - Provides flag values as to whether or not data is usable.
4. INCIDENT_DETECT - Provides flag values as to whether an incident has occurred or not.
5. CABINETS - Provides cabinet IDs, freeway names, text descriptions, and whether there is a ramp or not.
6. CABINET_LOCATION - Provides the location of cabinets. It includes the cabinet ID, the coordinate method used (see above), and whether the measure is defined using the WSDOT or the UW methods, and the location.
7. LOOPS - Describes the loop sensors. It contains the loop ID, the cabinet ID, whether or not the loop is metered, the road type, the direction of the traffic, the lane type, the lane number, and the sensor type code (an integer).
8. ALG_DESCRIPT - Provides a complete source listing of Java code that extracts the loop sensor data

Data Metrics
The TRAC warehouse does not impute data. The data metrics include quality assurance (QA) flags for each traffic detector. These flags are stored within the 20-second archive and a composite QA flag is set based on the 20-second data sets aggregated to produce a five-minute data set for each detector. The 20-second error flags refer to specific types of errors (e.g., "the detector is defaulting"). The flags generally indicate whether the detector is good, bad, questionable, or not working.

System Operation, Management and Costs
Twenty-second resolution data is available from the internet via the internet. Five minute data is available on CD-ROM per calendar quarter upon request. TRAC has never had a formal budget to design and build an archive. Therefore, the system has been developed informally in stages. Each phase of the archive development process has been in response to the need to perform some new, desired analytical task. The initial research budget that supported development of the warehouse was $70,000. The CD-ROM approach for the initial archive cost approximately $15,000. Biennium budgets have ranged from $250,000 to $350,000. There is not a cost sharing arrangement with the University of Washington for the development of the warehouse. These costs are borne entirely by WSDOT. TRAC performs analysis and software improvements on a contract basis over and above the biennium budget. These costs include development, maintenance, and support for the ITS backbone and analytical work.

Virginia Smart Travel Lab (STL)

System Overview/ Design
The Virginia Smart Travel Lab (STL) started in 1998. VDOT was one of the first transportation agencies to establish a formal policy on sharing ITS data and video. They anticipated the value of ITS data to a myriad of users, both internal and external. They realized a critical component to a successful warehouse was establishing an ITS data archiving policy and infrastructure. VDOT formally established STL as the facility responsible for archiving ITS data. The University of Virginia (UVA), VDOT's partner in the STL development, along with Open Roads Consulting and George Mason University, collectively support the development of a functional ITS warehouse and the underlying infrastructure. Initially, the only data in the archive was the Hampton Road freeway data (HRSTC). The Hampton Road region was used as the case study for initial testing prior to the implementation of the statewide system. It is presently in operational testing.

The warehouse is now being redesigned and will operate at STL. Access will be provided via an internet browser. The system uses the latest technologies, including XML and SOAP, and allows users to easily construct queries and to access predefined performance measure reports. Furthermore, the open technical design allows for direct integration into the operational software of TMC's.

Currently, the system supports the data from a variety of sources listed below. Detailed information about the data for each of these sources is provided in the following section on data collected.

- Hampton Roads (HRSTC) – freeway data, three video feeds
- Transcription of Hampton Road incident database covering the past five and one-half years
- Richmond – several video channel feeds
- Norfolk Virginia (NOVA) – freeway data including volume and occupancy
- NVSTSS – arterial data

The coverage area is Hampton Roads and Norfolk Virginia. STL uses emerging data warehouse approaches for the operational test of the warehouse as opposed to traditional transactional database design. This best supports targeted extraction of real-time ITS archived data.

Warehouse Applications/Performance Measures
STL's main functions are:

1. Volume aggregation (with minimal or no imputation of aggregated values)
2. Assure data quality
3. Determine normality conditions based upon past data
4. Bench marking – comparing data streams up and down stream of more than one source of data.

Methods for assuring data quality are based upon a paper by Brian Smith [72]. STL uses six rules, compared to 22 used by TTI (Texas Transportation Institute). The granularity of data aggregation is one minute intervals. This is the rate at which it is collected.

During a series of STL stakeholder meetings, a preliminary list of high priority tools was identified for development. Most are not yet functional.

- Quantitative system measures of effectiveness
- ITS data quality monitoring
- Planning and impact studies for detours, evacuations, construction, and transit routes
- Real-time incident management support
- Historical data average reports
- Access to classification data
- Quantitative impact of weather
- Detector system efficiencies
- Regional planning support
- Support to 511/ATIS efforts

The system currently supports only performance measures. These measures are discussed in the subsequent section.

Performance Measures
Many of the performance measures presently implemented are common to most ITS systems and were created by the Texas Transportation Institute. STL researchers have developed a new measure of travel time reliability, named the variability index, using a technique based on quality control. This measure reports the variability for a trip based on the time of day for a specific location. The mobility measures of effectiveness are:

- Throughput
  - Volume
  - Vehicle Miles of Travel (VMT)
- Congestion/Delay
  - Buffer Index
  - Travel Time
  - Travel Rate

Among the various graphical profiles for traffic operation assessment provided by STL include:

- Speed and flow plots for each station
- Loss of capacity – quantification of loss for general classes of freeway incidents
- Transit support - assists in route scheduling to support Flexroute services
- Planning support- developing percentile measures of capacity utilization for planning models

Such data visualization assists in answering the following questions:

1. How do incidents impact traffic?
2. How may incidents occur per area each time period?
3. What are the durations of incidents?

Stakeholder meetings were critical during deployment of the STL applications. They would perform a mock up and then demonstrate it to stakeholders who would then make recommendations. These meetings occurred four times a year during the deployment process.

Data Collected, Data Stored, and Data Formats
Two-minute real time station data (volumes, speeds, occupancies) are archived beginning from July 1998. There are 203 stations and 19 miles of freeway.

- Detector data (including ramps)
- Origin-destination pair estimations
- Classification data (vehicle types)

- Incident data updates that augment transcriptions from the past inclusive from 1997 to July 2002

NOVA (Norfolk) delivers detector data as a flat text file to the lab via ftp every 10 seconds. As part of the extraction, transformation, and loading process, STL parses the 10-second files and aggregates them to one-minute records that are inserted into a staging database. There are approximately 1,000 detectors from which approximately 1.5 million records per day are created.

HRSTC (Hampton Roads freeway data) polls detectors every 20 seconds and aggregates the data to two minutes granularity. STL fetches the HRSTC station data (speed, volume, occupancy, each by lane) as a flat text file via ftp every two minutes. The file is parsed and the records inserted into a staging database. Currently from 114 freeway stations, approximately 82,000 records per day are created.

The NVSTSS arterial data (signal sensors) include one flat file every 15 minutes for 1000+ traffic signals. From these files, approximately 144,000 records are archived each day. Traffic volume, speed, and occupancy are collected using loop detectors. This data is collected by NVSTSS at one minute resolution. There are 440 locations in the region. They also get signal system green light times which are needed for volume and capacity calculation. STL does not believe they will be able to obtain speeds from signal data.

From various sources, traffic incident data are collected. The information archived includes start time, end time, location, incident description, number of vehicles involved, and assisting agency. Also collected is classification data from sensors in selected region which is currently used to assess traffic composition.

To summarize, the types of data and types of sensors include:

1. Weather
2. Incident
3. Loop detectors - The data is collected at one minute resolution for 440 locations in the Norfolk region
4. Classification detectors (PZO detectors giving 13 types of classification)
5. Signal sensors

Data Metrics

STL puts considerable emphasis on ITS data quality evaluation. The quality of the data collected, in many cases, is not measured by the providers. Software has been implemented that continuously monitors the data as it is downloaded, tests it against logical rules and traffic flow theory, compares the data to historical data, and report variations to a system manager. The data quality assessments performed on the real-time data incl.
ude:

- Abnormality checks

- Comparisons against imputed data
- Comparisons against historical data queried at selected levels of aggregation
- Current conditions versus the recent past

The percentage of bad data is sometimes quite high. In total, six screening tests are used. The abnormality checks against "normal" for day of week, time of day, and station versus past averages. On occasion (and it is the only occasion imputed data is used by the STL system) an imputed value is used in place of an actual ("bad") value. Otherwise, a bad value is inserted into the archive but flagged as "unusable."

A data quality report is produced that summarizes the input ingestion results. The report, segmented by station, gives:

- Usable data points - The number of good data points plus the number of imputed data points
- Percentage of usable data - The number of records usable and divided by the number of records available
- Percentage of imputed data - The number of usable records that have been imputed divided by the number of available records that are usable

System Operation, Management, and Costs

The initial ITS data archive was converted from a traditional transactional database to a multidimensional (star schema) database design. Multidimensional data models are designed specifically to support the use and analysis of large archives of data. The operational system is deployed based upon the web service model. Assume that a user sends a request to the warehouse that requires accessing the currently measured traffic conditions. This request is communicated via the internet to the warehouse web services computer. Requests for local web pages or java applets are processed immediately by the web server. Requests for SOAP services (Simple Object Access Protocol)[5], such as forecasting services, incident management services, transit support services, etc. are forwarded to a Java servlet container for dispatch to the appropriate service. The Java servlets and SOAP services access the warehouse database which resides on a separate machine.

The integration of HTML, HTTP, XML and SOAP into an operational system provides a number of benefits. For example, the use of standard protocols ensures that the system may be used by a wide variety of users operating on various hardware, software, and network platforms. The use of HTTP as the core transport protocol eases the communication between client and server when firewalls are in use. HTTP uses the

[5] SOAP is a simple XML based protocol that allows applications to exchange information over HTTP. It builds upon these standards. It provides a simple and lightweight mechanism for exchanging structured and typed information between peers in a decentralized, distributed environment using XML. More simply, SOAP is a protocol for accessing a web service.

TCP/IP port 80 which is normally configured as open by most firewalls. The use of SOAP remote procedure calls allows other applications to easily integrate the services available from the warehouse system. The use of AML as the core data exchange language allows client-side applications (or applets) to easily parse the data for further analysis, filtering, and/or sorting. It also makes it possible for applications to save the results locally for future processing.

VDOT views this system as the foundation of a warehouse that will grow statewide in scope and importance. The project team that has been assembled provides expertise in both institutional and technical aspects of the effort. The project management and staffing for the lab comprises eight faculty, three researchers, three staff, 40+ students and a Smart Travel Van used for research. The disciplines, organizational entities, and head counts are given in Table 18.

Table 18. Staffing Categories for Virginia's Smart Travel Lab

Discipline	Entity	Head-count
Program Managers	VDOT, UVA	2
Team Manager	ORC, GMU	3
Transportation Engineer	VDOT, UVA	4
System Engineer	UVA, ORC	3
Database Engineer	UVA, ORC	2
Software Engineer	UVA, ORC	4
Research Assistance	UVA, GMU	lots

The organizational abbreviations used in Table 18 are:

UVA - University of Virginia
ORC - Open Roads Consulting
GMU - George Mason University
VDOT - Virginia Department of Transportation

Physically, STL resides at UVA. When fully functional, the operation will move to the permanent production site within the VDOT MIS group for continued support. STL will continue to exist at UVA and serve as the development entity. VDOT will be the production entity. The plan is to migrate the system over the 2006-2007 timeframe. Afterwards, the staffing requirements that STL will continue to support at the university through contract funds from VDOT are:

1. Full Time System Analyst
2. ¾ DBA
3. Full time application developer
4. Documentation Guru

5. Project manager
6. Several Graduate students

Direct funding is necessary to support the above personnel and to insure continuity. Approximately, $300,000 each year is allocated by VDOT to STL for development work and to support the production system when necessary.

Naturally, the decision to move the production aspects to VDOT has provoked much discussion within STL. Some of the reasons given for the move include:

1. Residing at university poses a huge security issue,
2. Greater physical access is generally desired by VDOT,
3. Responsiveness may improve if the system is in-house,
4. Perceived sense of loss control by VDOT if production remains at UVA,
5. VDOT has a better communication infrastructure, and
6. At some point, the University will need to think of itself as a production shop rather than fulfilling a mission of teaching and research.

Houston Transtar

System Overview/Design

Officially opened in April 1996, Houston TranStar is a partnership comprised of four government agencies that are responsible for coordinating the planning, design and operations of transportation and emergency management in the greater Houston region. The partners include: The Texas Department of Transportation, Harris County, The Metropolitan Transit Authority of Harris County, and The City of Houston. Additionally, there is an informal relationship with the TTI of Texas A & M. TTI, which is the transportation research arm of Texas A&M University, serves in a variety of capacities for the partnership.

Transtar is engaged in the development and promulgation of performance measures. Additionally, it provides services and applications to assist a number of emergency services and transit system functions. A summary of the areas in which Transtar is active is given in Table 19.

Table 19. Applications Developed Within Houston's TranStar

Traffic Conditions	Emergency Information	Transit Information
· Real-Time Traffic Map	· County Emergency Management	· HOV Lanes
· Traffic Camera Map	· Amber Plan	· Commuter and Bus Services
· Freeway Signs	· School Closings	· Bus Itinerary Request

· Personalized Alerts	· Weather	· QuickRide
· Wireless Web	· Current Ozone Levels	· METRORail
· Incidents/Road Closures	· Homeland Security	
· Roadway Weather Sensors		
· Regional Construction		
· TxDOT Lane Closures		

Data Collected, Data Stored and Data Formats
The Transtar database contains volume, occupancy, and speed by lane for the Houston freeway system. The database also contains rainfall data – unique among the data warehouses we investigated. A significant application is the use of toll tags to identify vehicles and track them at different points within the network. Thus, a time-to-destination prediction is derived. Travel times (but not the identity of the probe vehicles) are maintained in the warehouse.

The data is downloaded, cleansed, and inserted by custom-built software written by Southwest Research.

Data Metrics
No information was collected on data metrics.

System Operation, Management and Costs
No information was collected on system operation, management, and costs.

Georgia Department of Transportation

System Overview/Design
Since the Georgia Department of Transportation (Georgia DOT) Traffic Management Center (TMC) became operational in 1996, data has been archived into 15 minute aggregates per detector. The real-time system has had tremendous impact in the areas of incident management and traveler information, but has not realized its potential as a valuable source of traffic and transportation data for other transportation applications such as the computation of performance measures, benefit analysis, and transportation planning and modeling. Recently, GDOT recognized that archival databases have gained a high level of interest in the transportation community because valuable information can be extracted from commonly collected operations data. Individuals and agencies that reviewed the raw archived TMC data discovered data quality issues that are significant and warrant caution. Data deficiencies (hardware failures, detector malfunctions, server failures and communications problems) are a constant source of concern for those

responsible for maintaining the real-time information system (known as NaviGAtor), as well as for those who wish to find new uses for the archived data.

GDOT personnel are currently defining a methodology for preparing the existing GDOT archived traffic data to support the needs of transportation planning and performance evaluation. The initial phases included interviews with ARC (Atlanta Regional Commission) and GRTA (the transit authority), and GDOT planners to identify potential uses of the archived data for existing and projected responsibilities. One sub-task was analyzing existing data to identify the causes of quality deficiencies. The major data deficiencies come from missing data (gaps in the archive), poor metadata, and poorly maintained/calibrated hardware. Yet another sub-task of GDOT is to identify data aggregation and cleansing methods that lessen the impact of the known errors and prepare the data for use. In summary, GDOT has determined that the existing archived data does have the ability to generate useful information for transportation planners if certain aggregation and analytical techniques continue to be employed. The existing data supports the generation of segment-level, hourly speed, and occupancy data as well as the identification of a variety of temporal factors. Volume, however, cannot be adequately estimated at an acceptable level of accuracy.

Several recommendations have emerged that, once addressed by GDOT, could significantly improve the quality and usefulness of the data (including volume data). These recommendations include the prevention of data loss during the archiving process, the archiving of raw 20-second data instead of 15-minute aggregations, the establishment of a set of validated control stations, and the generation of metadata (including incident and weather log connectivity). Considering the scope and size of the existing infrastructure, and considering the potential value of the data for transportation planning and performance evaluation, the recommendations are minor "tweaks" with large impacts. These recommendations are incorporated into the "NaviGAtor Data Quality Assessment and Mitigation Strategies Final Report" prepared by GeoStats, LP URS Corporation in 2004.

GDOT has a head start in many areas because of the NaviGator system which integrates all of Georgia's TMC ITS data. All centers use the same system. However, the challenge has been to integrate this information in a manner compatible with national standards to ensure compatibility with data from other states. At present, GDOT has developed very high level descriptions of the warehouse.

Operationally, there are only GDOT TMC employees in the control room. They coordinate with the police and fire departments throughout Atlanta. Weather channels and news stations are often displayed simultaneously with the video monitor system. The data is used in real time but, as discussed, GDOT has initiated an effort to develop an ITS warehouse (archive) for the data.

Archived Data Applications/Performance Measures

The Georgia DOT planning group builds travel demand forecasting models for urban areas outside of the Atlanta metropolitan area. These include cities such as Macon,

Augusta, Savannah, and others. The Macon area is not yet operational but cameras are installed along I-475 and are operational. No data from them is currently being archived.

Archiving is possible for Macon and can be started at any time. Because this is a new start-up activity, it is useful to focus on it as being relevant to the Louisiana's current efforts. The Georgia DOT planning group listed the following uses for the archived TMC data relevant to Macon:

- HOV Counts and speeds
- Construction work zone planning
- Identifying peak hour speeds
- I-475 volumes and speeds
- Improving traffic count accuracy
- Incident recovery statistics

The Georgia DOT planning group has expressed interest in additional data regarding:

- Vehicle classification
- HOV vehicle occupancy (in addition to lane occupancy)

Looking beyond Macon, the Association of Metropolitan Planning Organizations conducted a survey of 60 metropolitan areas regarding their current use (or non-use) of archived transportation data. The majority of the MPOs that responded to this survey were using ITS data for model enhancements to their congestion management systems and their travel demand forecasting models. The chart in Figure 16 shows the top responses:

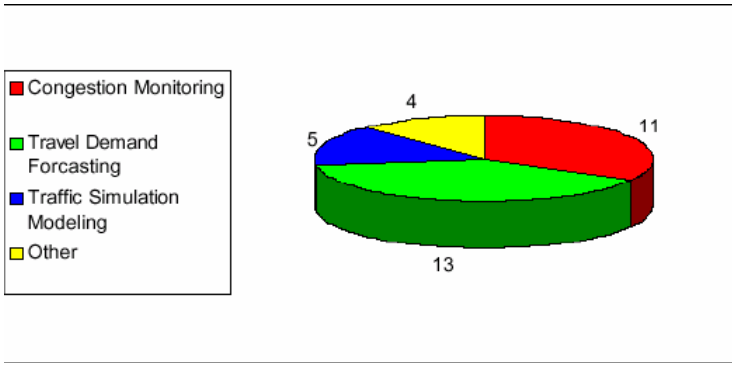


Figure 16. Top Purposes of Archived Traffic Data (Georgia)

A GDOT consultant outlined desired performance measures by citing the Portland Oregon system [90]. This research paper discusses the potential uses of archived data in the development of performance measures. Measures of mobility were calculated using:

- Average Daily Traffic (ADT)
- Average Daily Traffic Per Freeway Lane
- Average Speed

- Travel Time
- Vehicle Miles Traveled (VMT)
- Person Miles Traveled
- Vehicle Hours Traveled
- Person Hours Traveled
- Vehicle Miles Traveled by Congestion Level
- Person Miles Traveled by Congestion Level
- Percent of the freeway not congested during peak hours
- Number and percent of lane-miles congested
- Lost time due to congestion
- Demand vs. capacity
- Delay per VMT
- Reserve capacity

In addition, the data is also used to generate the following measures of economic impact, quality of life, and resource conservation:

- Cost of Delay
- Fuel Cost

Data Collected, Data Stored, and Data Formats
We change the focus again back to Atlanta for which 11 years of archived data exists, presently in the form of compressed XML files. Keep in mind that 15 minute aggregates are being archived although the sensors report five-minute aggregates. Daily reports are generated for each detector. These reports contain a snapshot of the detector's configuration, the raw five-minute aggregates, and pre-calculated higher-order aggregates (15-minute, one-hour, and one-day). Daily reports are also generated for each station from the component detector reports. Like the detector reports, these contain a configuration snapshot and pre-calculated five-min, 15-min, one-hour, and one-day aggregates.

A summary description of both the sensors and data that is archived follows:

- Archived granularity is 15 minutes for each detector, that is, sensor data is averaged over quarter-hour intervals.
- The fields captured are speed, occupancy, volume and in some cases classification (RTMS sensors only).
- Real-time data are used for travel time estimates. However, this is not effective on the major freeways because the detectors are located every third of a mile per road segment.
- Econolite, RTMS, and ATR loop detectors are used. These sensors are capable of collecting data every 20 seconds.
- Currently, three months of data are stored on-line and then burned onto a CD for distribution and storage.

- There are 500 ATR loop detectors throughout the state. In some instances these detectors coincide with other ITS sensors so data comparison is possible.

Data Metrics
As mentioned above, GDOT has serious data validity problems that have not been addressed. A GDOT consultant outlined the current data quality hurdles. A report noted that volumes of observational data are being under-utilized. The consultant enumerated the following immediate needs:

- Accurate volume data for a variety of uses
- Data for improving regional/statewide VMT estimates
- Data for computing level-of-service (LOS)
- Vehicle classification (in conjunction with loop detector data)
- Data for mobility, accessibility, and safety measures
- Free-flow vs. congested volume data
- Event planning
- Express bus performance monitoring
- HOV performance monitoring
- Blue Flyer express bus performance monitoring
- Better incident data to compute traffic impact, time of clearance, etc.
- Better speed data

System Operation, Management, and Costs
No information available.

California Performance Measures System (PeMS)

System Overview/Design
The freeway Performance Measurement System (PeMS) is a joint effort of the California Department of Transportation (CalTrans) and the University of California at Berkeley's (UC Berkeley) Institute for Transportation Studies. Its origins trace to a UC Berkeley white paper in 1997 and a desire by CalTrans to tap into the vast amount of data being generated by the thousands of loop detectors deployed throughout the state. The system was delivered to CalTrans in 2002. PeMS provides CalTrans with a powerful tool for system performance monitoring and congestion management. CalTrans uses the system for performance analysis [48], including congestion monitoring and estimating travel time reliability. CalTrans PeMS travel time estimates are used as the basis for travel time predictions on 20 to 30 routes in the Bay Area. Travel time predictions are posted on dynamic message signs (DMS) in the San Francisco metropolitan area. This information is also provided to value-added resellers (VARs) such as broadcast media and web sites. The system uses volume and lane occupancy data to determine the proportion of travel delay that is based on recurring or non-recurring congestion. The data collected is available to the public. Figure 17 shows the PeMS login page.

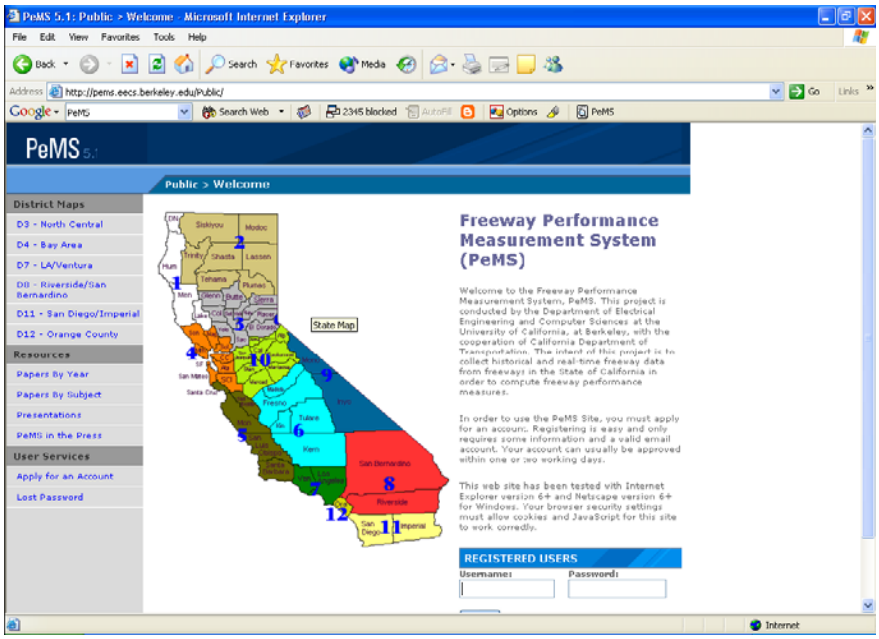


Figure 17. PeMS Login Page

The PeMS system is a data collection, archiving, analysis and display system. It performs detector diagnostics, data filtering, imputation, aggregation, and performance measurement computation. It presents over 140 standard plots, tables and charts. The PeMS system is currently collecting data from nine CalTrans districts. It collects and processes data in real-time and is accessed via a standard internet browser. For employees, i.e., not the public, it supports sensor configuration and management as well as safety statistics and mortality rates. Much analysis is by inter-relating incidents and transportation measures to review the interplay in attempts to relate causes and effects. Real-time applications include estimates of truck volumes, identifying bottlenecks, and tracking shifting peak flows. System users include CalTrans management and operations staff, university researchers, planners from the San Diego Association of Governments, consultants, VARs, the public, and the media.

The data processing involves building pyramids of data over long spatial and temporal scales. “Pyramids” is a term indicating differing degrees of aggregation, where the aggregation is over time intervals or over spatial areas. In the database this is represented by different table objects which contain data rolled up over different spatial and temporal scales. This rollup is done in real-time.

The major components of the system include:

Backend - Oracle;
Servers- Linux and Sun;
Frontend - PeMS (custom built PHP application);

Data processing core- data filtering and roll-up[6];
Diagnostic core- validity checks using filters and thresholds (validity checking is performed daily over the data of the past 24 hours)
ETL (extraction, transformation, and loading) – performed in real time

Much of the programming is custom and is in Perl and PHP. The OLAP is custom programmed. No off-the-shelf products are used. The image/graphics/maps are generated in PNG format.

System design was originally driven by the CalTrans operations community, which realized that improvements in system performance could no longer rely on increased capacity. The operations community wanted to use data collected by ITS to support development of highway congestion reports and to otherwise support the state's Transportation Management System – an integrated system that includes the TMCs, the computer and other automated components, field devices and peripherals, and the communications infrastructure for the transportation network. The research community is also a driving force behind some aspects of application development.

Design and development of PeMS was based on three basic principles:

- Start with simplest measures (e.g., calculate travel time for single highway segment)
- Build more comprehensive measures from the smaller ones
- Use the Internet for data distribution.

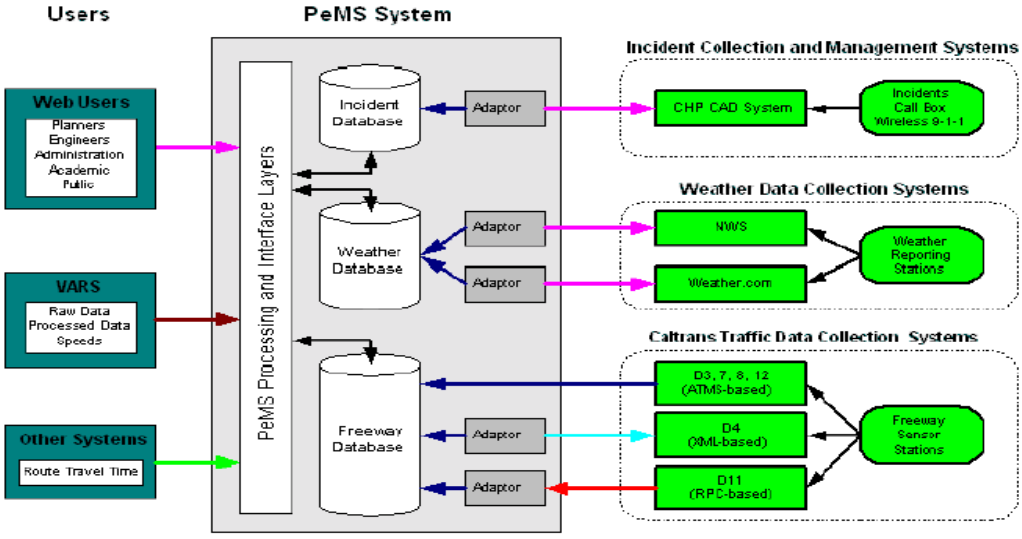


Figure 18. PeMS System design

[6] "Roll-up" is a technical database term referring to aggregation of records. For example, each set of three 20-second sensory data records may be rolled-up into one minute aggregates. Similarly, several segments of a freeway may be rolled-up into a spatial aggregate.

Figure 18. PeMS System design depicts an overview of the PeMS system. The design reflects the fact that CalTrans used professional vendors/software developers (as opposed to graduate students), in order to enhance reliability and responsiveness. Note that the three databases – incident, weather, and traffic – are largely independent.

Warehouse Applications/Performance Measures
As noted above, PeMS is able to produce about 140 standard graphs and tables. These run the gamut of:

- Throughput
  - Volume per segment
  - Vehicle Miles of Travel (VMT)
  - Level of Service (LOS) measures
- Congestion/Delay
  - Buffer index per segment
  - Travel time by route by time of day
  - Velocity averages
  - Effects of ramp metering
  - Bottleneck identification

The developers have noted that users desire more than one view of the data and that managers desire a higher level perspective than the general public. User feedback has been crucial to successful dissemination. The more tailored a data representation is to a specific goal, the more the representation is used. This, of course, has led to the proliferation of standardized graphs and tables.

Developers have noticed that the freeway operations staff often takes advantage of courses in fundamental traffic engineering (such as delay, reliability, and signal timing). These are tools that they can utilize for specific tasks associated with job responsibilities. Training is provided, for example, to CalTrans staff by the San Diego Association of Governments (SANDAG). Tracking these courses and the attendees often presents new opportunities for acquiring ideas on tailoring data views for operations staff members.

Applications that are to be implemented in the near future are:

- Determining optimum time for lane closures and other roadway maintenance,
- Providing information such as alternative route suggestions on selected DMSs in Los Angeles,
- Inclusion of arterial data in Los Angeles, and
- Using electronic toll collection tag data for analyzing origin/destination trip times and the demand generated by special events.

Data Collected, Data Stored, and Data Formats
PeMS collects and stores data in a central database located at UC Berkeley. The sensors/detectors are stable - it is very rare that a detector goes directly from good to bad. When one does, the reason is normally a power outage, water leaks, or due to

communication. Volume and occupancy is captured and stored at a granularity of 30 seconds. Presently, 4.1 terabytes are stored online and two gigabytes are added each day. The primary sensor set is the approximately 23,000 loop detectors that are distributed throughout the state.

PeMS collects traffic data from nine CalTrans districts but the coverage is ever increasing. The data are sent to the system's central database from transportation management centers (TMCs) around the state via the CalTrans wide area network (WAN). PeMS is responsible for cleansing and formatting the data. No district assumes the responsibility of formatting or cleansing the data prior to transmission. Figure 19 shows an overview of the data collection infrastructure.

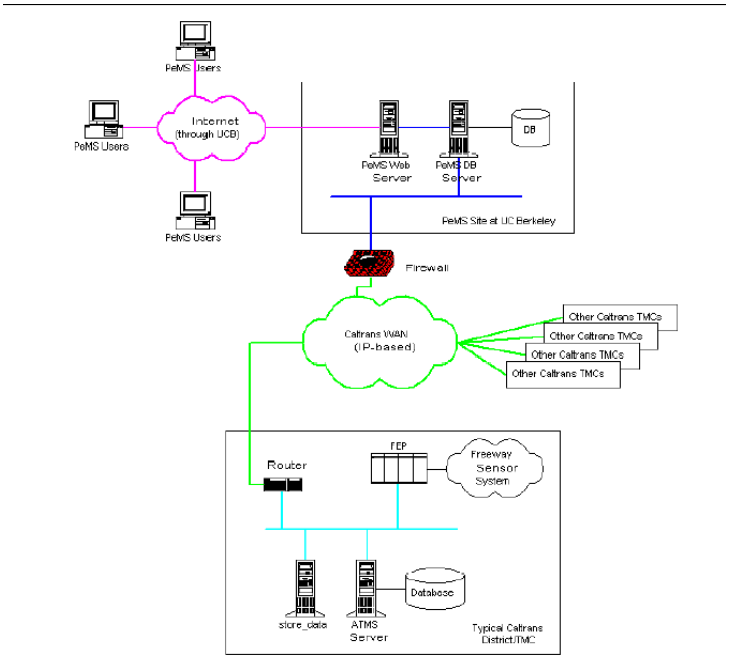


Figure 19. Overview of PeMS Data Collection Infrastructure

Data Metrics

CalTrans PeMS produces a daily diagnostic report that lists loops with problems as well as the likely cause of the problem (e.g., loop malfunction, communications failure, etc.). The system assesses data as they are received and determines if any data are suspect or missing. Missing or suspect data are automatically replaced with a value imputed from adjacent values.

System Operation, Management and Costs

The initial cost to establish PeMS was approximately $8 million and funded entirely by CalTrans. UC Berkeley does host the system at its main campus but the operational aspects are managed by Berkeley Transportation Systems, Inc. Annual maintenance, that is, over and above operational staff and new application development, requires approximately 1.5 full time equivalent (FTE) positions and software upgrades that cost $150,000-$200,000 annually.

Maricopa County Arizona RADS

System Overview/Design

The Maricopa County Regional Archival Data Server (RADS) is the third of the six warehouses profiled in this report that are not yet operational. (The others are VDOT/UVA Smart Travel Lab and the GDOT system in Georgia.) Nevertheless, much can be learned from the ambitious plans. Maricopa County RADS will collect and store data from the various systems in Maricopa County, Arizona, including the Arizona Department of Transportation (ADOT) Freeway Management System (FMS), Highway Closure and Restriction System (HCRS), the AZTech[7] SMART Corridors, Road Condition Reporting System (RCRS), and transit operations. The City of Chandler is the first jurisdiction to provide data to the system, starting in April 2005. Potential data sources include commercial vehicle data, expanded multimodal data, parking and event information, and weather information. Although the system has passed several proof-of-concept milestones such as effective use of HCRS, it is not ready for use as a decision support tool. A prototype is currently in use by ADOT. Full implementation will begin following procurement of a heavy-duty server for the data warehousing function. Approximately, 300 Gb of data from loop detectors (aggregated in 5-minute increments) will be loaded into an online archive. The system will then add about 3.5 Gb each month. Shifting funding priorities have slowed the implementation of the system.

The main system design goal is to take ITS data from systems throughout the Phoenix metropolitan area, store the data in a centralized archive, and then make the data available for a variety of users via a common Web interface. Data to be stored include traffic volumes, speeds, closures, incidents, public transit operations, and data collected by AZTech partner agencies. A key facet of the design approach is the use of Common Object Request Broker Architecture (CORBA) interfaces. Each government agency archives data for its own purposes; data are then "warehoused" via the Maricopa County RADS so that any agency can access it. The source agency can filter data so that only the data they wish to share are accessible. Eventually, all archived data will be available over a public (non-secure) interface.

Warehouse Applications/Performance Measures

Users will access data from the Maricopa County RADS via the Internet. The system will be used to support a variety of analyses. One of the opportunities presented by the system is its capability to blend data from various agencies or sources. The City of Phoenix is expected to develop evacuation planning methods. The Maricopa County RADS will be a reliable source of historical data and near real-time information.

Planned system updates include development of a Web interface to allow the public to access and use the system.

[7] AZTech is a partnership of public and private transportation agencies led by Maricopa County DOT and the Arizona DOT.

Data Collected, Stored and Formats
The Maricopa County RADS will host a variety of database formats to accommodate the range of agencies that will provide data to the system. As part of the requirements development process, the Maricopa County Department of Transportation (MCDOT) conducted extensive user surveys. Based on this work, the main users of the system will be Maricopa Association of Governments (MAG) planners, ADOT staff involved in ITS, local traffic engineers, transit agency staff, commercial vehicle operators, and private sector information providers.

The Maricopa County RADS will depend in large part on open source software to address data collection, given the variety of data types that it will store and process. The system will collect data via the Internet, CD-ROMs, or dedicated landlines, depending upon the agency providing the data. Decisions on which data to archive will be as decentralized as possible, leaving it up to the agencies themselves to determine which data they wish to provide to the system. At a minimum, the system will accommodate freeway data from ADOT and arterial data from the City of Tempe. Phoenix Transit has expressed interest, but has yet to provide data.

Data Metrics
Responsibility for data quality will rest with the agencies providing the data to the Maricopa County RADS.

System Operation, Management, and Costs
The Maricopa County RADS is being funded primarily through Federal Congestion Management and Air Quality (CMAQ) funding with local match and cooperation of Maricopa County, Arizona DOT, and Maricopa Association of Governments. Annual estimated maintenance costs for the system are anticipated to be $150,000, not including hardware and software upgrades.

APPENDIX B. RESEARCH REPORTS

Input Validation: A Probabilistic Approach for Modeling and Real-Time Data Filtering of Freeway Detector Data

Real time traffic information is vital to a variety of advanced operation and management functions undertaken by traffic management centers. The advent of new monitoring technologies has led to nationwide implementation of traffic surveillance systems on major urban freeway segments. Currently, several hundreds of freeway miles are instrumented with traffic surveillance devices such as electro-magnetic detectors, video detectors, radar detectors, and many others, all of which are primarily installed to improve the operation, safety, and productivity of our surface transportation network. These surveillance systems collect large amounts of real-time traffic data, sometimes on the order of a few gigabytes per day, and communicate them to traffic management centers (TMCs) to support critical functions such as incident detection, travel time and delay predictions, congestion management and other emergency services.

Advanced Traffic Information Systems (ATIS) need real time traffic data to disseminate real time traffic information to transportation system users via internet, in-vehicle navigation systems, variable message signs, etc. Such information assists travelers in better pre-trip planning and en-route decisions that affect their departure time, choice of destination, mode, and route. Such decisions can effectively reduce travel costs in terms of travel time and delays. For transportation system providers, traffic information is essential for performance monitoring and decision support systems, which can be greatly influenced by the quality of traffic data. To date, robust data screening methods have not been fully developed to control the quality of data before its archiving, dissemination to the public, or use in relevant applications.

The main goal of this research was to develop a real-time data screening algorithm by considering the stochastic variations in traffic conditions. This can be achieved by accomplishing the following objectives:

- Develop a methodology to examine the probabilistic nature of the three macroscopic traffic parameters (speed, volume, and occupancy), considering the stochastic as well as dynamic changes in the traffic conditions.
- Model the probabilistic nature of the three traffic parameters and evaluate the calibrated model by performance measures.
- Derive a data screening algorithm based upon the consistency of the probabilistic relationships as reflected by the model, and devise a strategy to further identify the partially valid observations (i.e., observations that have one or more invalid parameters).
- Demonstrate, using a sample data set, how the newly developed data screening methodology can be applied to freeway traffic data in real time.

Literature Review

Traffic surveillance systems are primarily used to monitor and collect traffic information from urban freeways. Traffic monitoring equipment can be classified as road-based and vehicle-based. Loop detectors, Closed Circuit Television (CCTV), sensors etc, are examples of road-based detection systems. Vehicle-based traffic surveillance systems include probe vehicles that are equipped with tracking devices, such as transponders, to track the location of vehicles over time. Figure 20 depicts how traffic information is relayed to TMCs from different monitoring equipment.

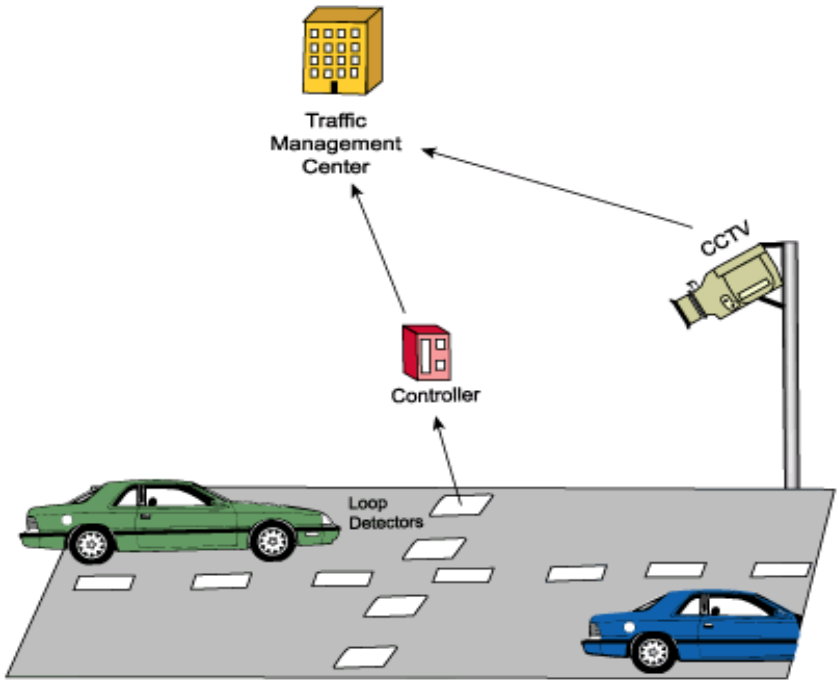


Figure 20. Traffic Surveillance System

The following section presents information about loop detectors, which constitute the majority of real-time traffic monitoring devices.

*Loop Detectors***.** Inductive loop detectors remain the most commonly used device for freeway surveillance and incident detection systems. Inductive loop detectors are constructed by cutting a slot in pavement and placing one or more turns of wire in the slot [87]. The wire is then covered by a sealant. The size of the loop detector ranges from 6-ft x 6-ft (for normal loops) to 6- x 40- to 70-ft (for long rectangular loops). Loop detectors collect vehicle count, lane occupancy and vehicle speed at intervals of 20 to 30 seconds and relay such information to traffic management centers.

Loop detectors operate on the "principle of inductance." Inductance is generated in a loop circuit due to current passing through a loop detector coil buried in the pavement. Loop detectors consist of four parts: a wire loop of one or more turns of wire embedded in the roadway pavement, a lead-in wire running from the wire loop to a pull box, a lead-

in cable connecting the lead-in wire at the pull box to the controller, and an electronics unit housed in the controller cabinet [87]. When a vehicle passes over a loop detector it causes change in the initial inductance and the pulse is transmitted to the controller placed at the side of the pavement indicating the presence of the vehicle. Figure 21 shows the main components of inductive loop detectors.

Single loop detectors are capable of measuring flow and lane occupancy directly, while measurement of vehicle speed requires using dual loops or estimation using traffic flow models. Estimation of speed using traffic flow models is explained below.

Flow = speed times density…….. (1)

where density can be approximated from lane occupancy using:

Density = occupancy times g……..……. (2)

and

$g = k$/(vehicle length + detector length)…….…….(3)

where k is a conversion factor.

Hall and Persaud [88] came up with different values of g for different traffic conditions.

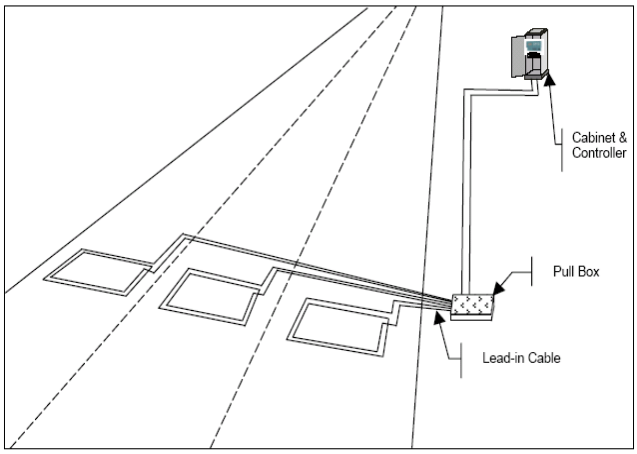


Figure 21: Inductive Loop Detectors

Speed is also estimated using dual loops and can be calculated from the formula given below. Figure 22 represents time-space diagram of the vehicle passing over two closely spaced detectors.

$$S = \frac{D}{[(t_{on})_n]_B - [(t_{on})_n]_A}$$

where,

S is the speed of the vehicle,

D is the distance from upstream edge of detection zone A to the upstream edge of detection zone B (feet),

$[(t_{on})_n]_B$ is the instant that vehicle is detected on detector B,

$[(t_{on})_n]_B$ is the instant that vehicle is detected on detector A.

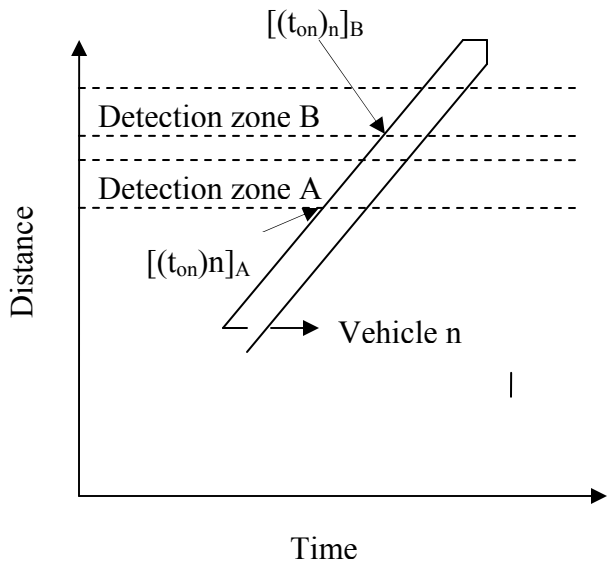


Figure 22 : Vehicle Passing over Two Closely Spaced Detectors

Other Traffic Monitoring Devices. Road-based traffic surveillance systems may also include other types of monitoring devices such as Closed Circuit Television (CCTV), Video Image Detection Systems (VIDS), and sensors such as Remote Traffic Microwave Sensors (RTMS). CCTV's and VIDS systems are more efficient and cost effective traffic monitoring systems that provide real-time traffic information, but are sensitive to all weather conditions. RTMS devices are cost effective and weather resistant. They detect traffic parameters on multiple lanes and have become increasingly popular in the last few years.

Inductive loop detectors form most commonly used traffic surveillance equipment. However, the data collected from these detectors is prone to errors due to loop malfunctions such as cross-talk (interaction of magnetic fields of the closely placed loop detectors), pulse break-up (where a single vehicle registers multiple actuations as the sensor output flickers off and back on) , stuck sensors, etc. With the wide spread of such detectors there appears to be a pressing need to monitor the data quality to ensure better reliability of subsequent applications that rely on loop detector data.

Data Screening Methods. Several research efforts that focused on providing algorithms to screen erroneous data were reviewed and are briefly presented in this section. In general, two basic approaches have been pursued. The first approach involves processing raw signals from the loop detectors, where the sensor on-times are used to compute the volume and occupancy, which are further checked for credibility. The second approach applies reliability checks either directly on the macroscopic parameters (volume occupancy and speed) or on the traffic relationships between these parameters, usually by establishing thresholds beyond which the data represents unrealistic traffic conditions. Examples of each approach are presented next.

Chen and May [74] suggested a methodology for data screening in which the on-time of a detector is compared with the station average to determine the inconsistency in the data.

This approach was criticized as being sensitive to errors such as "pulse breakups" where multiple detections of a single vehicle were registered as sensor output flickers off and back on.

Another study based on the on-time of the detector was conducted by Coifman [75]. This research study emphasized the use of speed traps to identify detector errors and assessed the performance of the speed trap based on the assumption that the on-times should be same for the vehicles at free flow conditions, allowing for hard decelerations, regardless of vehicle length. The proposed study was sensitive to congested conditions as vehicle acceleration from low speeds will cause two on-times to differ and was not applicable for congested traffic conditions.

A later study using on-times was conducted by Coifman and Dhoorjaty [76]. This study presented several detector validation tests that use event data (individual vehicle data) to identify detector errors both at single and dual loop detectors. Detector errors were identified by a series of eight detector validation tests which used event data like head way, vehicle length, number of congested samples, etc. in combination to the on-time, thus making the approach applicable to all traffic flow conditions.

Examples of the studies that used the second approach include a study by Jacobson, et al. [77] to develop a screening algorithm based upon threshold values of occupancy, and occupancy to volume (O/V) ratios. The observations were screened with the thresholds designed to represent different malfunctioning states of the detectors and thus the observations were identified as erroneous.

Another study by Cleghorn, et al [78] suggested a data screening algorithm based upon two strategies: (i) upper bound developed for flow-occupancy data for single loop detector systems and (ii) boundaries for feasible combinations of speed, flow and occupancy data. This study indicated that erroneous observations of the traffic data could lead to deterioration of performance of incident detection algorithm.

A later study by Payne and Thompson [79] presented various types of malfunction identification tests by imposing thresholds on occupancy, speed and volume parameters. The malfunctions were then diagnosed by inspection of aggregate sensor measurements. Data repair of faulty observations were then done by estimating actual traffic conditions and utilizing measurements from adjacent lanes.

Turochy and Smith [80] presented a study emphasizing the development of data screening algorithm based on the combination of threshold value tests and traffic flow theory. The screening procedure devised in this research study was based on four tests. The first two tests were based on maximum volume and occupancy thresholds, while the third test was based on the maximum value of volume that could be observed for zero value of occupancy and the fourth test was based on feasibility of average vehicle length calculated as a function of speed, volume and occupancy.

Peeta and Anastassopoulos [81] conducted a study to detect the errors due to malfunctioning detectors and predict actual data using Fourier transformation based correction heuristic. This approach was capable of detecting abnormalities and distinguishes data faults from incidents and aids the operation of online architectures of real-time route guidance and incident detection.

Ishak [82] presented the concept of fuzzy clustering to measure the level of uncertainties associated with the three traffic parameters: speed, occupancy, and volume. This research study criticized the use of average effective vehicle lengths for identifying detector data errors and devised a data screening algorithm based upon the uncertainty measure derived from membership grade and a decaying function. The uncertainty measure was then compared to a certain threshold limit to screen the observations and identify the erroneous nature of single parameter.

Chen, et al. [83] developed a diagnostic algorithm to identify bad loop detectors from their speed, occupancy and volume measurements using the time series of many samples. About four statistics which represent the summaries of time-series were derived and were used to decide whether the loop is bad or good. Imputation of the missing values was done based upon the linear relationship between neighboring loops.

A study by Wall and Dailey [84] indicated the use of consistency of vehicle counts to judge the validity of the data for an off-line analysis. The study also suggested a methodology to correct the erroneous data by identifying properly calibrated detectors which are used as reference stations to correct the data from poorly calibrated stations.

Chilkamarri and Al-Deek [85] presented a screening algorithm to flag out bad samples using the mathematical relationship between the flow, occupancy, speed and average vehicle length and suggested a pair-wise quadratic regression model to impute the missing data in real time. The study also proposed entropy statistic to identify the detectors which are stuck. Several other studies that used the second approach for data filtering were published. See for instance, Nihan [86].

Summary of Literature Review. Most of the research conducted in this area used the traffic flow relationships or imposed thresholds on the observations to devise data screening strategies. However, no effort was made to model the stochastic relationships between the traffic parameters and develop a real-time data screening algorithm. This study aims to develop a real-time data screening algorithm by considering the probabilistic relationships between the parameters which triggers the online maintenance of the detectors as well.

Data Collection

This section describes the procedure used to collect the data for conducting the research study. In addition, this section includes information about the preliminary screening techniques that were used to remove erroneous observations that result from improper recording of the data.

The data used in this study was collected from a 38-mile freeway segment of the I-4 corridor in Orlando, Florida. Figure 23 shows the map of the study section considered that extends from west of US-192 to east of Lake Mary Blvd. Data was collected using 70 inductive dual loop detector stations that are spaced at nearly .5 miles apart in both directions (east bound and west bound) on the study section considered. Each lane has two 6' x 6' loops embedded in the pavement that are connected to a 170 type controller located in a cabinet adjacent to the road side. Table 20 shows the location and description of each detector station.

Each detector station collects 30 second observations of three traffic parameters (speed, lane occupancy, and volume counts) from all six lanes. Speed and lane occupancies are expressed as average for all vehicles with in each 30 second period, while volume represents the cumulative vehicle counts within each time period (30 seconds). The information collected from each detector station is then transmitted to the Orlando Regional Traffic management center (RTMC). Figure 24 displays the configuration of a typical loop detector station in one direction of travel. The loop detector data is collected in real time via a T1 link between the Orlando RTMC and the ITS lab at the University of Central Florida. Speed, volume counts, and lane occupancies are downloaded and compiled into an Structured Query Language (SQL) server that supports multiple publicly accessible web applications such as real-time and short-term travel time predictions between user-selected on- and off- ramps. Information about the three traffic parameters was extracted from 130 million observations that were compiled in the years 2000 and 2002.

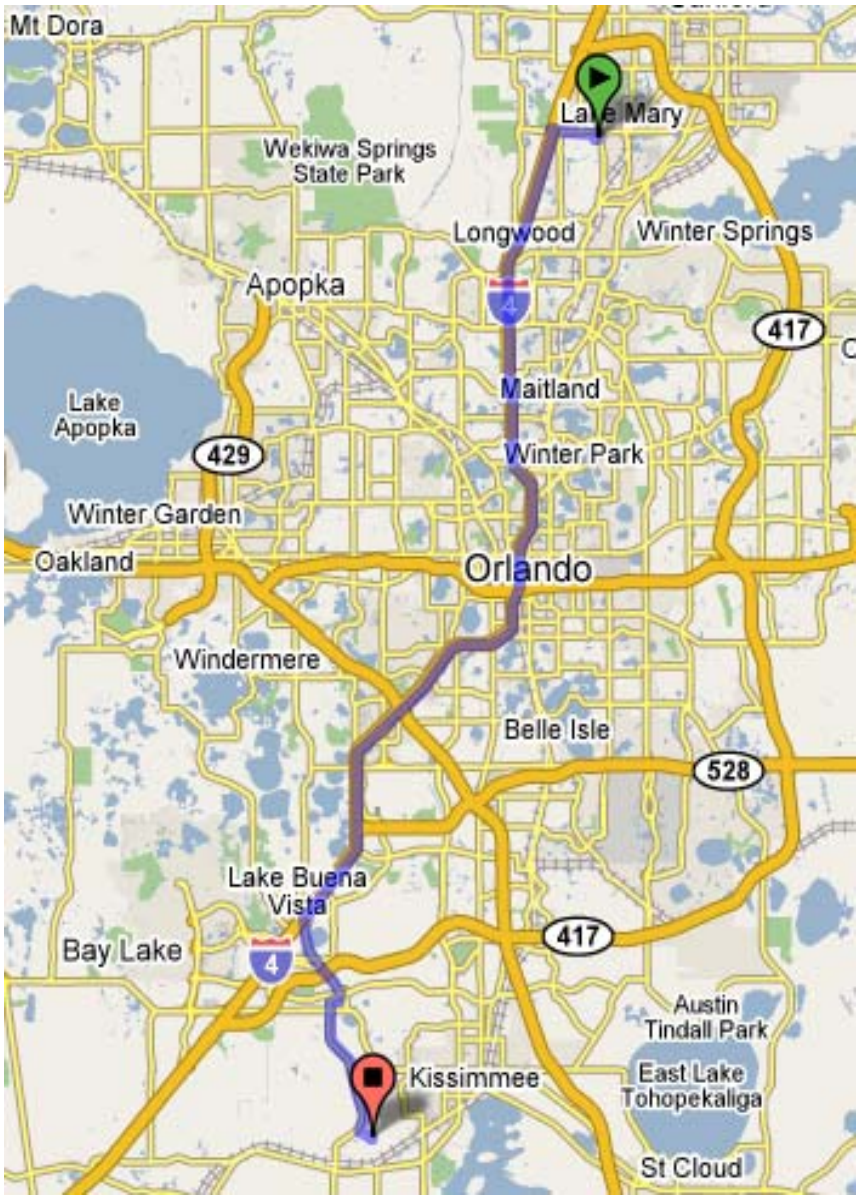


Figure 23. Map of I-4 Study Corridor in Orlando, Florida

Table 20. Location of Loop Detector Stations on I-4 in Orlando, Florida

From Station	To Station	Location	Spacing (feet)
1	2	West of 192	-
2	3	West of 192	2600
3	4	US 192	2470
4	5	West of Osceola	3300
5	6	East of Osceola	3530
6	7	SR 536	3330
7	8	East of SR 536	3370
8	9	West of SR 535	3360
9	10	West of SR 535	3400
10	11	SR 535	3000
11	12	West of Rest Area	3200
12	13	Rest Area	4090
13	14	West of Central Florida Pkwy	3020
14	15	Central Florida Pkwy	2980
15	16	528 EB Ramp	2910
16	17	528 WB Ramp	3250
17	18	West of 482	3100
18	19	West of 482	3450
19	20	SR 482	2000
20	21	West of 435	3100
21	22	West of 435	2600
22	23	SR 435	3000
23	24	435 WB Ramp	2900
24	25	Turnpike	2200
25	26	Turnpike WB Ramp	2900
26	27	Camera 21	2610
27	28	West of John Young Pkwy	2890
28	29	West of John Young Pkwy	2900
29	30	John Young Pkwy	4100
30	31	East of John Young Pkwy	2400
31	32	Rio Grande	2600
32	33	Orange Blossom Trail	2400
33	34	Michigan	2500
34	35	Kaley	2400

Table 20. (Continued)

From Station	To Station	Location	Spacing (feet)
35	36	Camera 28	2700
36	37	Camera29	2700
37	38	Church St	1800
38	40	Robinson	3000
39	41	SR 50	2500
40	42	Ivanhoe	2600
41	43	Princeton	2700
42	44	Winter Pk	2600
43	45	Par Ave	2600
44	46	Minnesota	3000
45	47	SR 426	2200
46	48	Site 1393	2300
47	49	Lee Rd	2600
48	50	East of Lee Rd	1700
49	51	Kennedy	2800
50	52	414 EB Ramp	3000
51	53	East of SR 414	1800
52	54	Wymore	3300
53	55	East of Wymore	2700
54	56	West of SR 436	2900
55	57	SR 436	2400
56	58	West of SR 434	3800
57	59	West of SR 434	2900
58	60	SR 434	3500
59	61	434 Ent Ramp	3400
60	62	434 Ext Ramp	1900
61	63	West of EEWill	2800
62	64	East of EEWill	2600
63	65	Rest Area	3000
64	66	East of Rest Area	2700
65	67	West of Lake Mary Blvd	2100
66	68	West of Lake Mary Blvd	2500
67	69	Lake Mary	2800
68	70	Lake Mary	2300
69	71	East of Lake Mary Blvd	3500

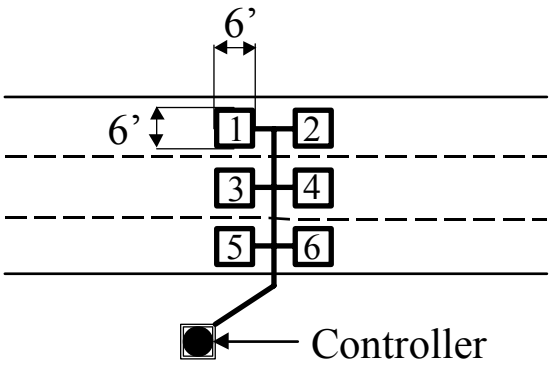


Figure 24: Typical Loop Detector Station

A sample of the data collected from the database is in the form as shown in Table 21. The data contains the following information: station, time, and the three traffic parameters on all the six lanes.

*Initial Data Screening***.** The data obtained from the loop detector could not be directly used for capturing the probabilistic relationships between the parameters. This was due to discrepancies observed in the data such as negative values or errors that resulted from the break down of the detector station or failure of communication infrastructure between detector station and TMC. Hence the data had to be filtered to remove these erroneous observations. Each dual loop detector records two values of occupancy, and volume count that were averaged, while speed is directly calculated using the dual loops. Preliminary filtering techniques that were used to remove invalid observations of the three traffic parameters are listed as follows:

1. Observations with zero or negative values of the parameters were discarded.
2. The errors in the data resulting from the failure of the loop detectors or mis-functioning of the communication between detector and TMC are represented by a –9XX value,-XX value or zero and are filtered out.

*Summary of Data Collection***.** Information in terms of three macroscopic traffic parameters (occupancy, speed and volume was collected from dual loop detection system on the study section. The data was then processed to filter out the preliminary errors (such as negative errors or mis- communication errors) associated with the data. The data can now be used to examine the probabilistic relationships between the parameters

Table 21 Sample of SQL Compiled Data for January 2000[8]

station	time	els	ecs	ers	wls	wcs	wrs	elv	ecv	erv	wlv	wcv	wrv	elo	eco	ero	wlo	wco	wro
2	6:30	0	51	56	0	63	0	0	6	10	0	2	0	0	4	7	0	1	0
2	7:00	57	57	66	66	54	48	4	4	4	1	4	2	4	2	2	0	3	1
2	7:30	9	56	0	60	70	71	2	7	0	4	2	1	0	2	0	1	1	0
2	8:00	59	51	59	57	61	61	1	2	1	2	4	2	0	0	0	0	1	0
2	8:30	58	60	82	77	100	100	5	8	3	2	5	4	3	5	2	1	4	2
2	9:00	57	59	61	61	63	63	3	4	4	1	2	1	2	2	2	0	1	0
2	9:30	56	59	63	0	59	59	8	5	1	0	3	3	3	2	0	0	2	1

Methodology

This research study proposed probabilistic approaches for real-time freeway traffic data screening. The proposed approaches differ from the deterministic approach in that they do not explicitly confirm the validity of an observation but attempt to quantify the likelihood that such observation is valid. Probabilistic relationships between the three traffic parameters (volume, speed, and lane occupancy) were developed to capture the least likely temporal changes in traffic states as well as inconsistencies in traffic conditions expressed by each traffic parameter. The proposed methodology thus primarily investigates the stochastic variation in traffic conditions over time and the probabilistic relationships between the three traffic parameters in order to capture certain characteristics that can be used for data screening purposes.

The methodology is derived from two complementary approaches. The first approach considers the stochastic evolution of traffic conditions over time measured by each of the three parameters independently. The second approach attempts to capture the inherent stochastic variation of traffic conditions measured by each combination of the three traffic parameters. In both approaches models for the conditional probabilities are developed from a vast amount of detector data describing all possible variations of traffic conditions on the study segment considered. Both approaches form the basis for the data screening algorithm and are explained in detail next.

Approach One: Examining Temporal Variations of Traffic Parameters. This approach focuses on capturing possible abrupt changes in traffic conditions that may occur between two successive observations taken over a time span of 30 seconds. These temporal variations, though abrupt, are unlikely to be extreme within the time span considered. For instance, the variation in the value of speed from 90 mph to 0 mph over a time span of 30 seconds is likely to be unrealistic. Thus this approach checks for unrealistic temporal variations that could be used to judge the validity of an observation. In simple terms, an observation is considered valid if the temporal variation from its preceding observation is not unrealistic. The feasible range of temporal variations could be derived by comparing the probabilities of the temporal variations with user-specified thresholds.

[8] els, ecs, ers - speed in the east bound direction on left, center and right lanes. wls, wcs and wrs-speed in west bound direction on left, center and right lanes. elo, eco, ero- occupancy in east bound direction on left, center and right lanes. wlo, wco, wro- occupancy in west bound direction on left, center and right lanes. elv, ecv, erv- volume in east bound on left, center and right lanes. wlv, wcv, wrv- volume in west bound on left, center and right lanes.

The temporal changes observed between the parameters over time are stochastic due to random variations in the traffic conditions. The stochastic variation of the parameters is captured using the conditional probability concept as mentioned earlier. Conditional probability is defined as the probability of an event occurring given that some event has occurred [89].

The methodology used to examine stochastic variations of each traffic parameter over time requires estimation of family of cumulative PDFs. Let X_t and X_{t+1}, represent any of the three parameters (speed, occupancy and volume) observed at time t and t+1. The difference ($X_t - X_{t+1}$), refers to the drop or increase in X over a duration of 30 seconds. Several possible combinations of two successive observations were used to model all possible temporal variations for each variable. The probability of observing the difference ($X_t - X_{t+1}$), given ($X_t = x$) is estimated using the following discrete conditional probability density function:

$$P\{X_t - X_{t+1} = \delta \mid x\} = \frac{N\{X_t - X_{t+1} = \delta \mid x\}}{N\{x\}} \tag{4}$$

where,

δ is the realization of the random variable $X_t - X_{t+1}$,

$P\{X_t - X_{t+1} = \delta \mid x\}$ = the probability of observing a difference of δ given $X_t = x$,

$N\{X_t - X_{t+1} = \delta \mid x\}$ = the number of observations with the difference of δ given, $X_t = x$,

$N\{x\}$ = total number of observations with $X_t = x$.

The difference between the variables ($X_t - X_{t+1}$) of two successive observations may be positive or negative, indicating either a drop or an increase in the value of X over a time interval of 30 seconds. The probability distribution functions for drops or increases in X exhibit different characteristics. Hence, they are studied separately. The probability distribution function for a *drop* in ($_x$) can be found from the cumulative sum of discrete probability mass function as follows:

$$P\{X_t - X_{t+1} \le \delta \mid x\} = \sum_{\forall j \in [0,\delta]} P\{X_t - X_{t+1} = j \mid x\} \quad \forall x \in \{0, X^{\max}\}, \quad \delta \ge 0 \tag{5}$$

where,

$P\{X_t - X_{t+1} \le \delta \mid x\}$ = the cumulative probability of observing a drop in X, given $X_t = x$,

$X^{\max}$ is the maximum feasible value for variable X.

The probability distribution function for *increase* in (X) can be similarly found from the cumulative sum of discrete probability mass function as follows:

$$P\{X_{t+1} - X_t \le \delta \mid x\} = \sum_{\forall j \in [0,\delta]} P\{X_{t+1} - X_t = j \mid x\} \quad \forall x \in \{0, X^{\max}\}, \quad \delta \ge 0 \qquad (6)$$

where,

$P\{X_{t+1} - X_t \le \delta \mid x\}$ = the cumulative probability of observing an increase in X, given $X_t = x$ and $\delta \ge 0$.

Approach Two: Examining Probabilistic Traffic Flow Relationships. This approach aims at checking for inconsistencies in the traffic conditions measured by each of the traffic parameters. Volume count, lane occupancy and speed in any observation are inter-related and should reflect similar traffic conditions. The relationship between the three traffic parameters can be used as a measure to validate an observation. For example, an observation representing a combination of high speed and high occupancy is unlikely under stable flow conditions. Such combinations are inconsistent, and possess less probability. Thus examining the probabilistic relationship between the parameters serves as a source for detecting the inconsistencies in the traffic conditions.

The relationship between the parameters is, however, probabilistic due to the random changes in the traffic conditions. These relationships are examined using a conditional probability concept as mentioned earlier. Estimation of PDFs that represent the probabilistic relationship between the three traffic parameters is described next. Let variables X and Y represent two of the three traffic parameters (speed, occupancy and volume). The probabilistic relationship between X and Y may be approximated by a probability mass function of the form:

$$P\{X = x_i \mid Y = y_j\} = \frac{N\{X = x_i \mid Y = y_j\}}{N\{Y = y_j\}} \qquad \forall\ i \in [0, N],\ j \in [0, M] \qquad (6)$$

where,

N is the number of realizations of X.

M is the number of realizations of Y.

$P\{X = x_i \mid Y = y_j\}$ = the conditional probability of observing $X = x_i$ given $Y = y_j$.

$N\{X = x_i \mid Y = y_j\}$ = the number of observations of $X = x_i$ given $Y = y_j$.

$N\{Y = y_j\}$ = total number of observations with $Y = y_j$.

The discrete probability function is used to calculate the cumulative distribution function as follows:

$$P(X \le x_k \mid Y = y_j) = \sum_{\forall i \in [0,k]} P(X = x_i \mid Y = y_j) \qquad \forall j \in [0, M],\ K \in [0, N]. \qquad (7)$$

where $P(X \le x_k \mid Y = y_j)$ is the probability of observing $X \le x_k$ given $Y = y_j$.

Summary of Methodology. Probability distribution functions that represent the temporal variations of each parameter, and the probabilistic traffic flow relationships were

estimated. These PDFs functions are to be modeled to capture the nature of these relationships. This raises the issue of choosing appropriate modeling tools, which is addressed in the next section.

Probability Distribution Functions

The probability distribution functions derived from the two approaches reflect the random behavior of the traffic conditions, and are non-linear in nature. Hence a non-linear function approximation seems quite appropriate to model the data. This can be best accomplished using Artificial Neural Networks (ANN).This section presents an introduction to Multi-Layer Perceptron (MLP), an Artificial Neural Network (ANN) tool used for function approximation. This section also explains the procedure to train NN models for approximating the probability distribution functions. Finally the performance evaluation of network models built is presented.

Multi-Layer Perceptron (MLP). The MLP is a general static ANN that has been used extensively for nonlinear function approximation. It consists of four layers - an input layer, two hidden layers and an output layer. Figure 25 shows an example of network topology used for the study. The number of neurons in the first hidden layer is double the number of neurons in the second one, as is the general practice in NN topology. The input layer is where the data is fed; the hidden layers extract the features from the input patterns; the output layer which gives the responses to the input. An MLP is trained using the back propagation algorithm, which minimizes the sum of squared errors between the desired and actual output.

Modeling PDFs for Approach One. The process of approximating the discrete probability distribution functions developed from the two approaches is done by training MLP networks with the data.

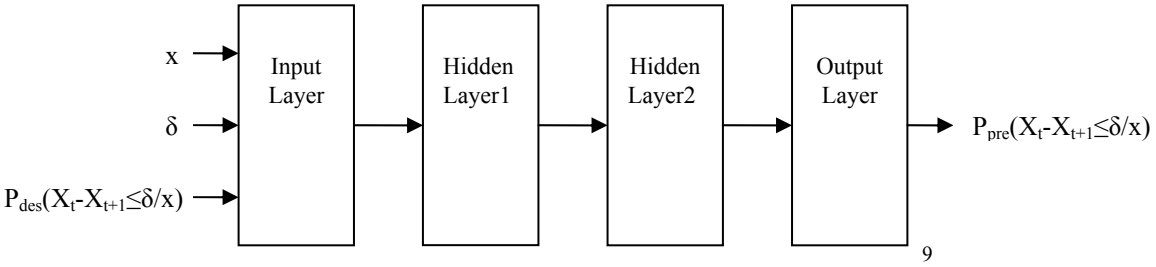


Figure 25. An Example of MLP Network Topology

The probability distribution functions estimated for each traffic parameter (for both drop and increase conditions) were approximated separately using different MLP networks as they possessed different characteristics and probability distributions. The probability distribution function for drop in X was expressed as follows:

$$P\{X_t - X_{t+1} \leq \delta \mid x\} = \sum_{\forall j \in [0,\delta]} P\{X_t - X_{t+1} = j \mid x\} \quad \forall x \in \{0, X^{\max}\}, \quad \delta \geq 0$$

The input data for modeling the PDF representing a drop in X was in the following form:

[9] P_{des}- probability desired P_{pre}- probability predicted

Type 1 Input: $\{X, \delta,\ P\{X_t - X_{t+1} \le \delta \mid x\}\}$

where,

X_t and X_{t+1}, represent any of the three parameters (speed, occupancy and volume) observed at time *t* and *t+1*,

δ is the realization of the random variable $X_t - X_{t+1}$,

$P\{X_t - X_{t+1} \le \delta \mid x\}$ = the cumulative probability of observing a drop in X, given $X_t = x$.

The probability distribution function for increase in X was expressed as follows:

$$P\{X_{t+1} - X_t \le \delta \mid x\} = \sum_{\forall j \in [0,\delta]} P\{X_{t+1} - X_t = j \mid x\} \quad \forall x \in \{0, X^{\max}\}, \quad \delta \ge 0$$

The input data for modeling the PDF representing an increase in X was in the following form:

Type 2 Input: $\{X, \delta,\ P\{X_{t+1} - X_t \le \delta \mid x\}\}$

where,

δ is the realization of the random variable $X_{t+1} - X_t$,

$P\{X_{t+1} - X_t \le \delta \mid x\}$ = the cumulative probability of observing an increase in X, given $X_t = x$ and $\delta \ge 0$.

Separate data sets for modeling the stochastic variations of each traffic parameter were extracted from a large data set complied in the year 2000. These data sets were used to train the MLP networks. Speed and occupancy parameters varying from a range of 0-100 mph and 0-100% were considered to account for the most likely traffic conditions. The maximum value for the volume count was taken to be 20 in compliance with the maximum capacity of 2400 vphpl. Cumulative probabilities representing the stochastic variations for each traffic parameter were calculated using the PDFs. Figure 26 shows a sample of the probability distributions for the stochastic temporal variation of three traffic parameters. These PDFs were now approximated using ANN by inputting the data in the format mentioned earlier as explained next.

Multi Layer Feed-forward networks were trained with the input data using a back propagation algorithm to capture the stochastic variation of the parameters over time. The input data set consists of two independent variables and a dependent variable. The number of neurons in the input layer depends on the number of independent variables considered. Hence the number of neurons in the input layer was fixed to two. The output layer contains a single neuron that represents the dependant variable. The number of neurons in the hidden layers were arbitrarily chosen depending upon the size of the training data. Training process progresses with the aim of reducing the mean square error on the training data and was terminated on the basis of any of the two criteria (i) training error reaching minimum (i.e., MSE is equal to .01) or (ii) training epochs reaching a maximum of 1000.

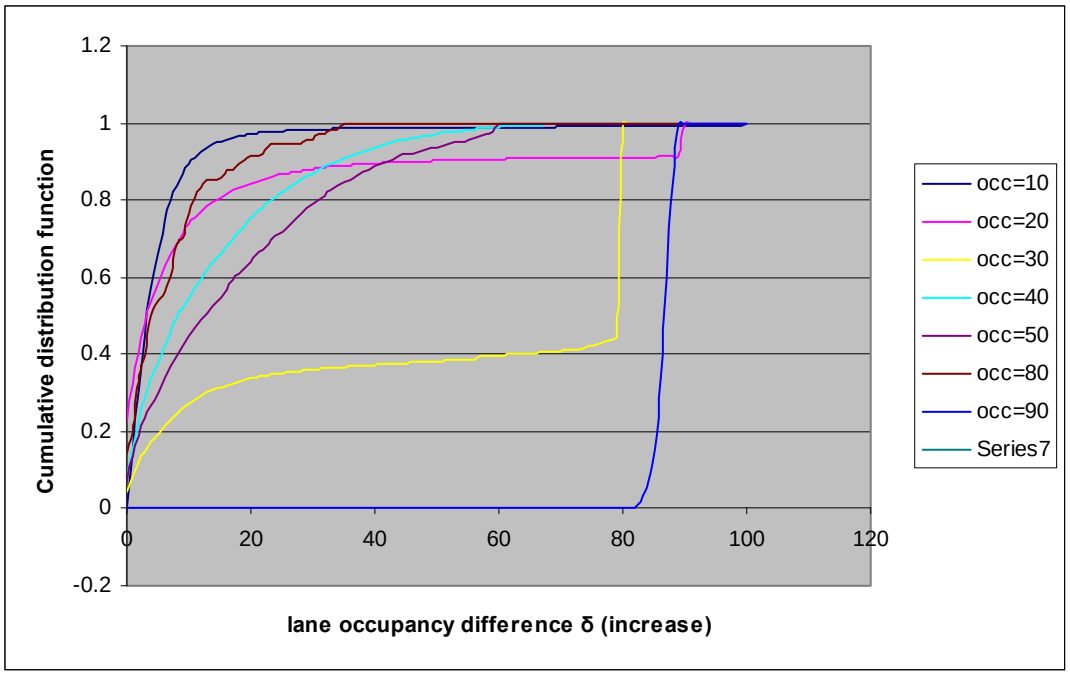


Figure 26. Probability Distribution Functions for Occupancy Parameter

Occupancy and speed parameters ranging from zero to one hundred were split into four uniform intervals of twenty five each, and their corresponding stochastic variations were approximated using separate networks to improve approximation efficiency.

Stochastic variations of volume parameter were modeled directly as the approximation performance achieved was observed to be high. Eight networks (four networks for stochastic variations representing drop and four networks for stochastic variations representing increase) were built to model stochastic variations of either occupancy or speed parameter. Two MLP networks were built to model the stochastic variation of volume parameter.

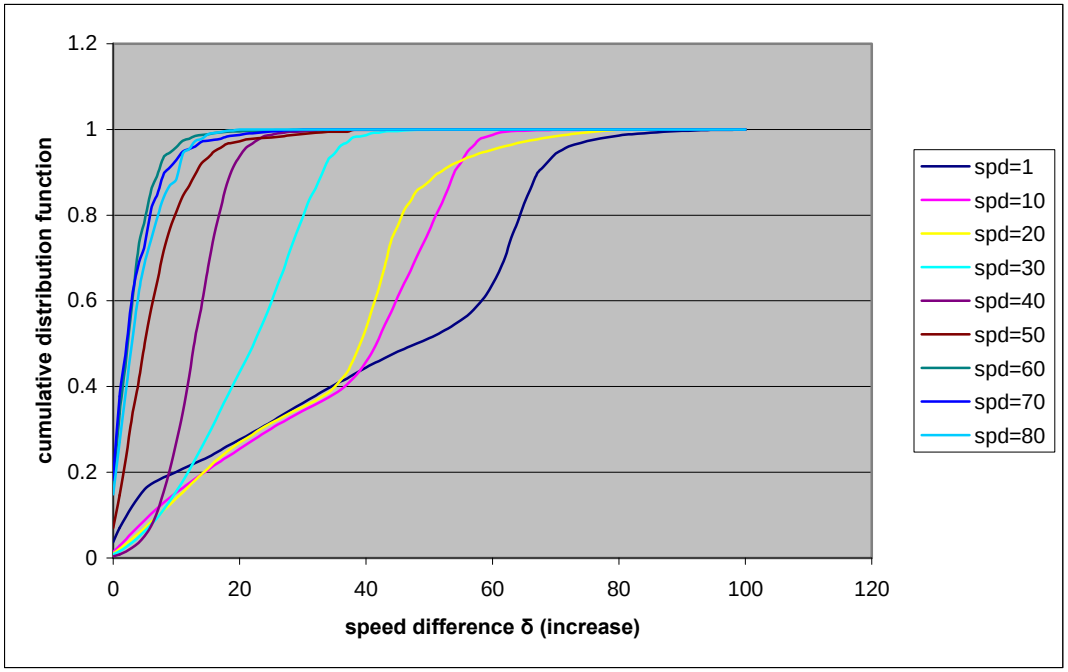


Figure 27. Probability Distribution Functions for Speed Parameter

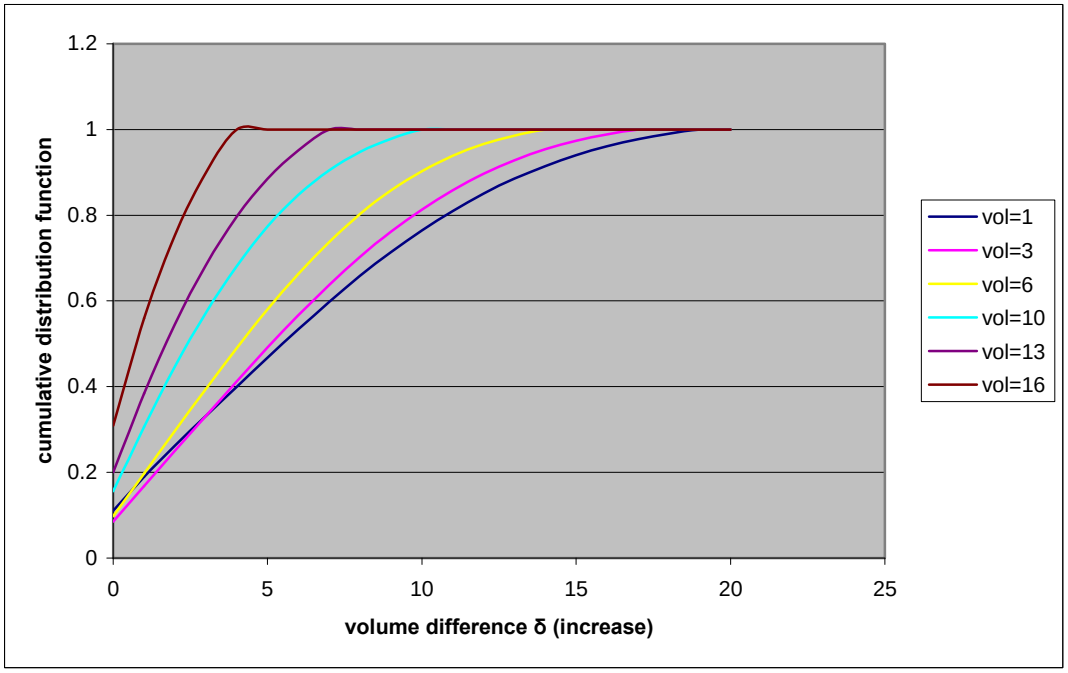


Figure 28. Probability Distribution Functions for Volume Parameter

A total of eighteen MLP networks [(2 x 8) for occupancy and speed parameters plus 2 for volume parameter] were built to approximate the stochastic variation of the three traffic parameters.

*Modeling PDFs for Approach Two***.** Probability distribution functions representing relationship between the parameters were approximated separately using MLP networks. The probability distribution function for a relationship between any two parameters was expressed as follows:

$$P(X \le x_k \mid Y = y_j) = \sum_{\forall i \in [0,k]} P(X = x_i \mid Y = y_j) \qquad \forall j \in [0, M],\ K \in [0, N].$$

The input data for modeling the PDF above was in the following form:
Type 1 input: $\{X, Y, P(X \le x_k \mid Y = y_j)\}$

where,

X and Y represent two of the three traffic parameters,
$P(X \le x_k \mid Y = y_j)$ is the probability of observing $X \le x_k$ given $Y = y_j$,
N is the number of realizations of X,
M is the number of realizations of Y.

Speed and occupancy parameters were divided into bins of five interval size. Separate data sets were extracted for speed conditioned on occupancy, occupancy conditioned on speed, volume conditioned on speed, volume conditioned on occupancy, and the corresponding probabilities were calculated from the probability distribution functions. For the cases of speed conditioned on volume and occupancy conditioned on volume, the data was divided into stable and unstable flows. This was due to the fact that each value of volume corresponded to two values of speed or occupancy, one in the stable flow and other in unstable flow conditions. These observations possessed different probability distributions and thus they were to be modeled separately. Critical speed which separates stable flow and unstable flow conditions was calculated from weighted average method and was found to be varying between 35-40 mph. The data set for speed conditioned on volume was divided using this critical speed and the cumulative probabilities were calculated separately. Similarly critical occupancy (15-20%) which demarcates the stable flow from the unstable flow was calculated using the weighted average method and their cumulative probabilities were calculated separately. Figure 29 through 36 show the probability distributions of several combinations of the parameters.

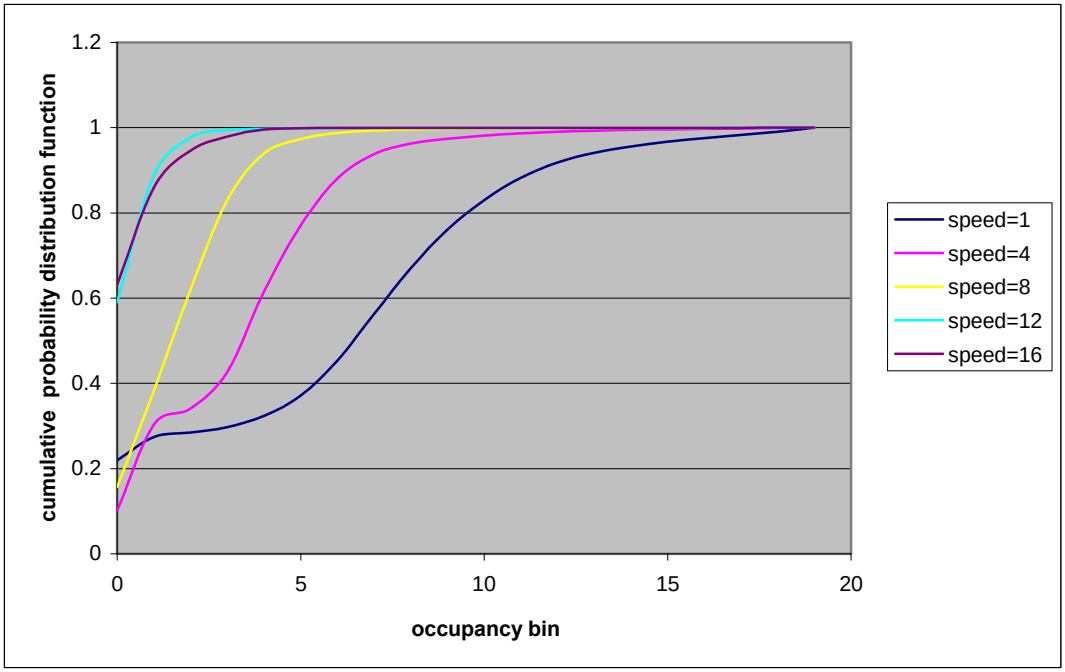


Figure 29. Probability Distribution Functions for Occupancy Conditioned on Speed

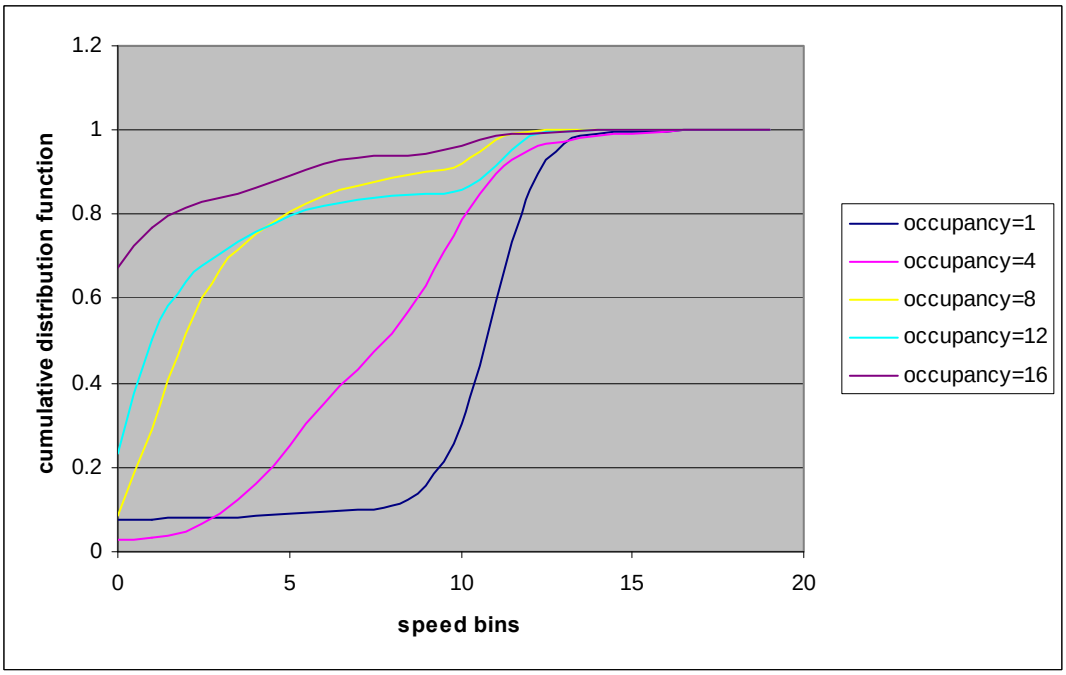


Figure 30. Probability Distribution Functions for Speed Conditioned on Occupancy

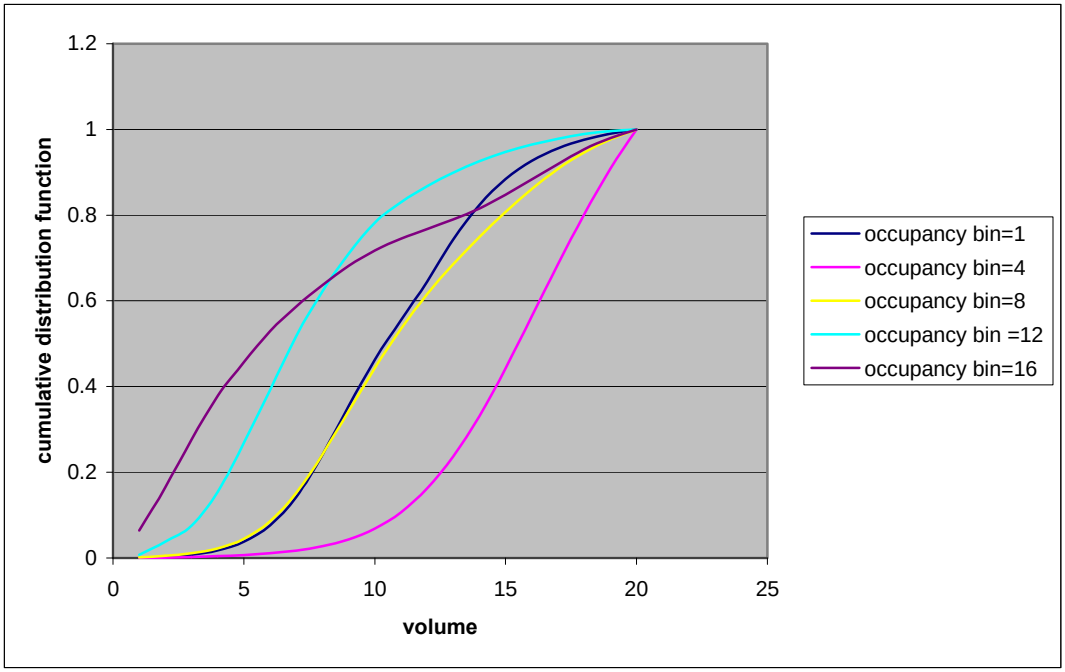


Figure 31. Probability Distribution Functions for Volume Conditioned on Occupancy

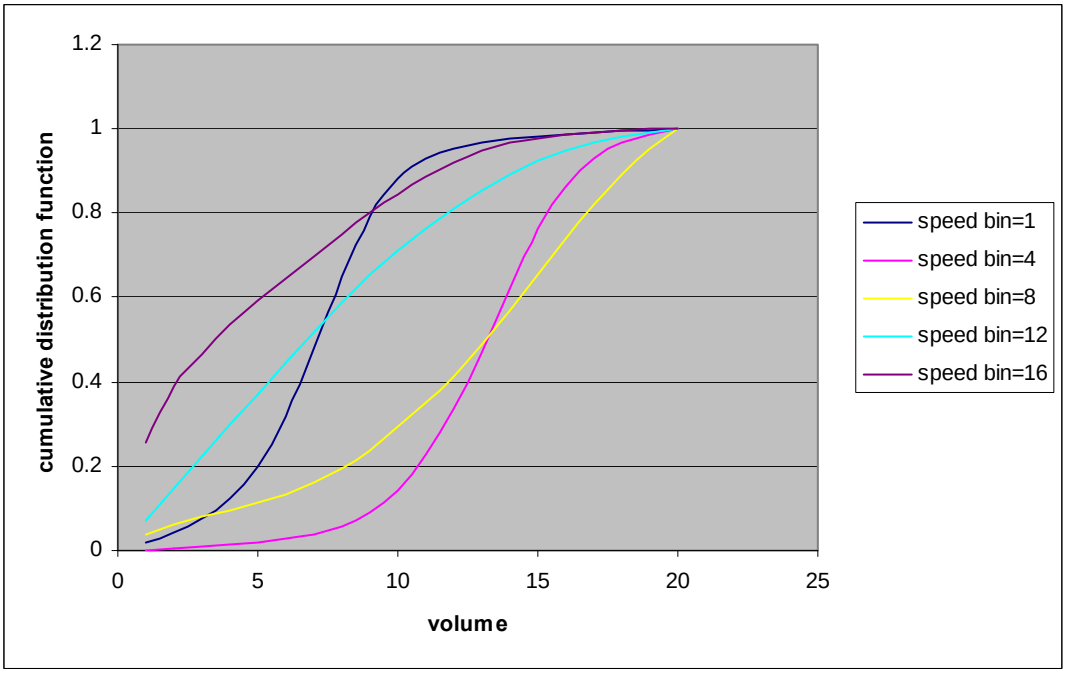


Figure 32. Probability Distribution Functions for Volume Conditioned on Speed

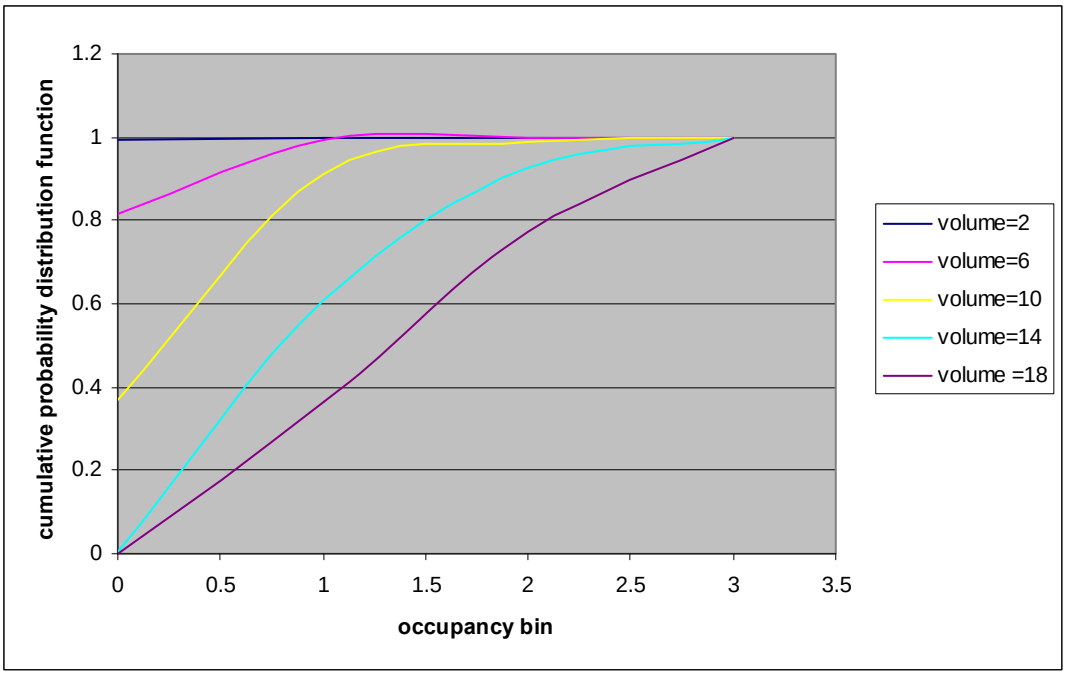


Figure 33. Probability Distribution Functions for Occupancy Conditioned on Volume (Stable flow)

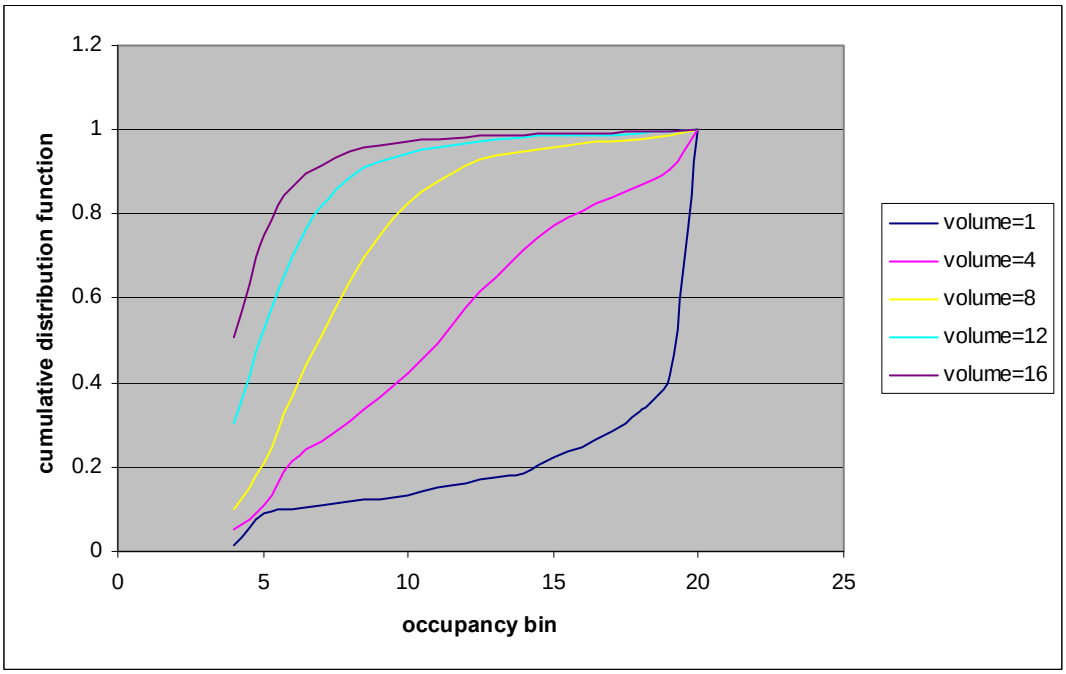


Figure 34. Probability Distribution Functions for Occupancy Conditioned on Volume (Unstable flow)

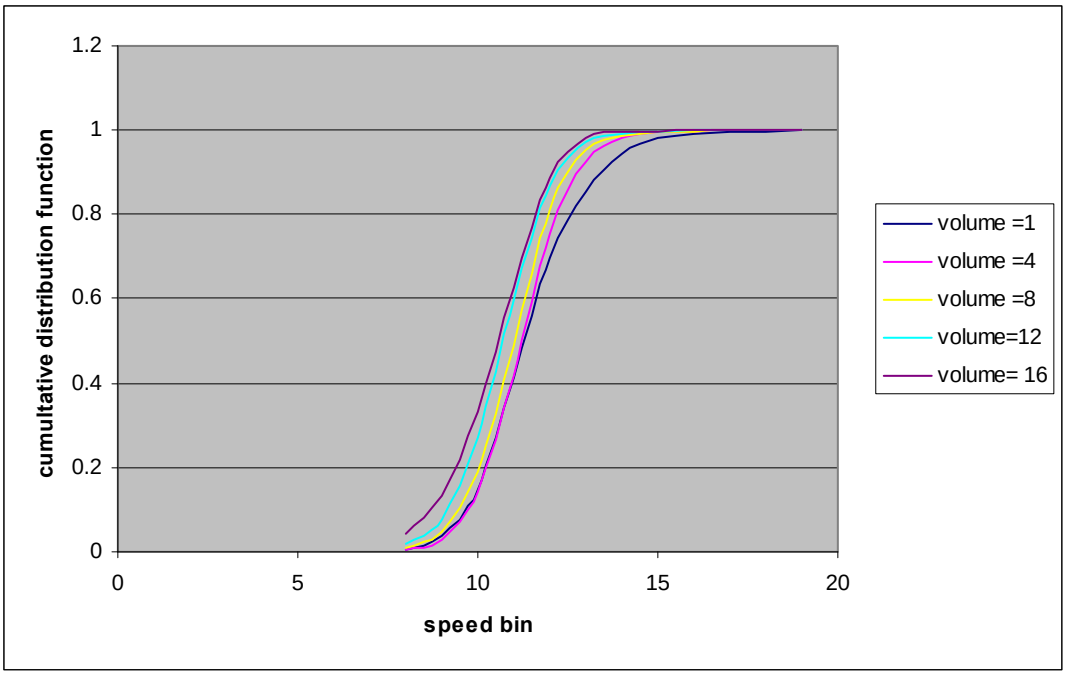


Figure 35. Probability Distribution Functions for Speed Conditioned on Volume (Stable flow)

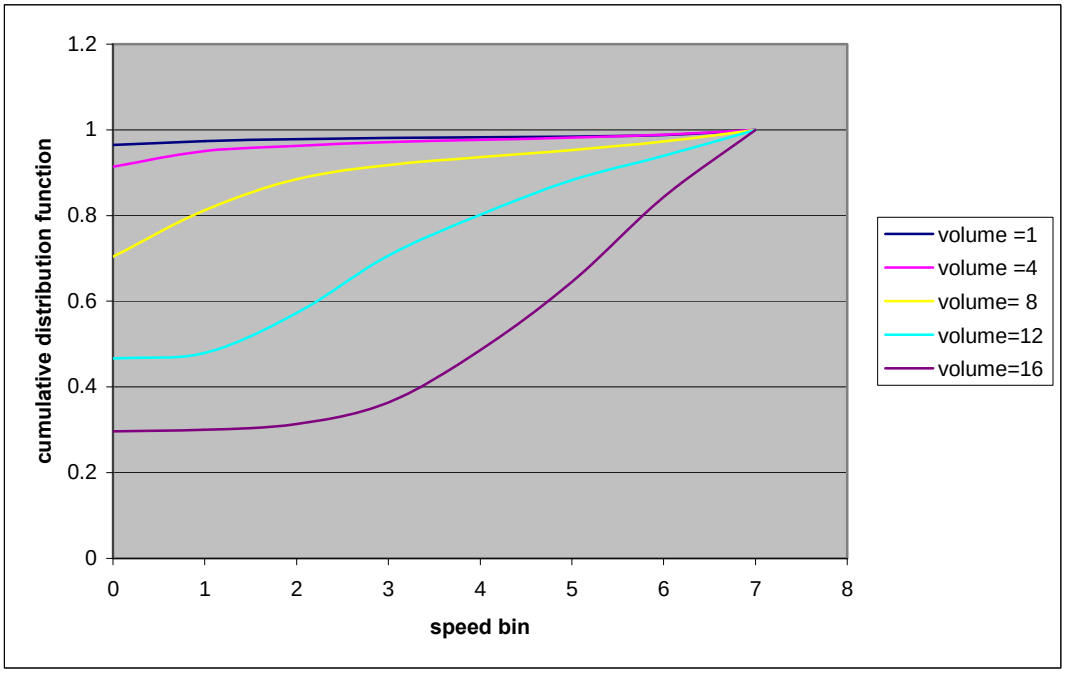


Figure 36. Probability Distribution Functions for Speed Conditioned on Volume (Unstable flow)

Multi-layer feed forward networks were trained with the input data in order to approximate the probabilistic traffic flow relationships. The data sets representing combinations of speed conditioned on occupancy, volume condition on occupancy,

occupancy conditioned on speed and volume conditioned on speed were divided into four uniform intervals and were approximated using different networks. A total of 16 MLP networks (4 for each condition, therefore 16 for all the four combinations) were built to model the probabilistic relationships between the above combinations of parameters. The data sets representing the combinations of speed conditioned on volume and occupancy conditioned on volume was divided into stable flow and unstable flow conditions and were approximated using 4 MLP networks (2 for each combination, therefore 4 for two combinations.) A total of 20 networks were built to approximate the PDFs representing probabilistic traffic flow relationships.

Performance Measures. Trained networks are evaluated using a set of performance measures. Desired probabilities are compared with the predicted probability values generated from the network to calculate three measures of performance: R-square, and Root Mean Square Error (RMSE) and Average Absolute Relative Error (AARE). Each measure of performance is defined as below:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (p(i) - O(i))^2}{N}}$$

and

$$\text{AARE} = \frac{\sum_{i=1}^{n} \left| \frac{P(i) - O(i)}{O(i)} \right|}{N}$$

where

$P(i)$ = predicted value of the parameter for observation i.

$O(i)$ = actual observed value of the parameter for observation i.

N = number of observations.

Performance Evaluation of the ANN Models Developed for Approach One.
The performance measures suggested above were used to evaluate the training efficiency of 18 ANN models built for approach one. Table 22 shows the performance measures for all the networks. From the Table 22 it was observed that RMSE ranges from .072 to .252, AARE varies from .0258 to .477, and R-square varies from .651 to .962. A possible explanation to the variation observed in the performance measures would be variation in the frequencies of observations leading to wide differences in the probability distributions within the interval considered. Failure to approximate such cases accurately also leads to high approximation errors. For instance, an occupancy range of 75-100 implies extreme congested conditions. A wide variation in the probability distributions of different observations could be seen due to unstable flow conditions, thus affecting the approximation efficiency. The variation in the performance measures for the occupancy increase case can be attributed to differences in the probability distributions among the observations and approximation errors.

Performance Evaluation of the ANN Models Developed for Approach Two. Performance of the 20 networks built to model the probabilistic traffic flow relationships was evaluated and tabulated in Table 23. A difference in the performance measures

could be observed from Table 23 (i.e., RMSE varies from .005-.092, AARE varies from .021-.581 and R-square varies from .835-.992). This variation in the performance measures could be attributed to variation in the probability distributions due to insufficient data and approximation errors. The performance measures calculated for evaluating the network models indicated reasonable approximation of the PDFs developed for both the approaches.

Summary of PDFs.

Probability distribution functions for the first and second approach were approximated using 38 Multi-layer Feed-Forward networks. Performance evaluation of the network models conducted showed reasonable approximation of the PDFs. The network models built have the capability to predict the probabilities of real-time data which form the basis for devising a data screening algorithm as explained next.

Table 22. Performance Measures of the ANN Models for Approach One

Network No	No. of observations	Parameter	Type	Interval	RMSE	AARE	R-square	Observations within ±
1	299	Occupancy	Drop	0-25	0.076	0.215	0.946	90.2
2	800	Occupancy	Drop	25-50	0.129	0.393	0.856	75.8
3	481	Occupancy	Drop	50-75	0.078	0.477	0.945	89.6
4	966	Occupancy	Drop	75-100	0.252	0.373	0.651	61.2
5	2162	Occupancy	Increase	0-25	0.102	0.114	0.736	87.6
6	1321	Occupancy	Increase	25-50	0.123	0.18	0.765	83.6
7	286	Occupancy	Increase	50-75	0.131	0.21	0.812	80.9
8	165	Occupancy	Increase	75-100	0.078	0.25	0.962	91.8
9	312	Speed	Drop	0-25	0.092	0.361	0.921	81.2
10	934	Speed	Drop	25-50	0.088	0.331	0.917	86.5
11	1570	Speed	Drop	50-75	0.077	0.124	0.88	92.2
12	1177	Speed	Drop	75-100	0.078	0.23	0.938	84.9
13	2039	Speed	Increase	0-25	0.092	0.416	0.946	83.5
14	1600	Speed	Increase	25-50	0.083	0.279	0.948	88.8
15	1003	Speed	Increase	50-75	0.072	0.113	0.879	91.1
16	252	Speed	Increase	75-100	0.075	0.252	0.921	88.4
17	420	Volume	Drop	0-20	0.076	0.0258	0.94	88.59
18	231	Volume	Increase	0-20	0.091	0.104	0.935	88.59

Table 23. Performance Measures of the ANN Models for Approach Two

Network no.	No. of observations	Type	Interval	RMSE	AARE	R-square	Observations within ± .01
1	100	O/S	0-5	0.092	0.361	0.921	90.1
2	100	O/S	5-10	0.088	0.331	0.917	95.2
3	100	O/S	10-15	0.077	0.124	0.88	95.2
4	100	O/S	15-20	0.078	0.23	0.938	96.5
5	81	O/V (stable flow)	1-3	0.065	0.193	0.976	90.1
6	339	O/V (unstable flow)	4-20	0.005	0.147	0.935	91.4
7	100	S/O	0-5	0.076	0.285	0.964	89.1
8	100	S/O	5-10	0.07	0.259	0.967	91.2
9	100	S/O	10-15	0.065	0.0907	0.907	91.2
10	100	S/O	15-20	0.064	0.021	0.835	99.1
11	160	S/V (unstable flow)	1-7	0.071	0.071	0.939	95.1
12	240	S/V (stable flow)	8-20	0.066	0.285	0.992	94.5
13	100	V/O	0-5	0.079	0.461	0.969	88.3
14	100	V/O	5-10	0.074	0.581	0.982	85.1
15	100	V/O	10-15	0.075	0.345	0.967	86.8
16	100	V/O	15-20	0.077	0.147	0.919	88.8
17	100	V/S	0-5	0.072	0.513	0.973	92.8
18	100	V/S	5-10	0.078	0.502	0.952	90.1
19	100	V/S	10-15	0.061	0.208	0.977	90.9
20	100	V/S	15-20	0.066	0.091	0.954	94

Data Screening Algorithm

This section deals with application of the neural network models for screening of a real-time data set. The process of devising a real-time screening algorithm is carried out in three stages. In the first stage, ANN models developed are used to predict the probabilities for real-time data. In the second stage, the probabilities predicted from the network models are compared with user specific threshold to identify erroneous observations. The third stage deals with further analysis conducted to identify the erroneous parameters in an observation.

Stage One: Prediction of Probabilities for Real-time Data.

The neural network models built were used to predict the probabilities associated with the data presented in real-time format. A 24 hour detector data compiled from 70 detector stations in the year 2002 was considered. A continuous stream of observations was extracted and was inputted into the network models. The probabilities associated with 52,000 continuous observations were predicted from the network models and were further used for screening analysis. Figure 37 shows a snapshot of nine probabilities derived to screen each observation.

Stage Two: Data Screening Algorithm.

The probabilities obtained from the MLP networks were used for deriving data screening strategy to filter the observations. A threshold of 95 percent was considered to demonstrate the implementation of data screening algorithm. The threshold specified is the probability with which the observations could be judged as valid.

				Probabilistic Traffic Flow Relationships						Temporal Variation of Parameters		
Obs	O	S	V	$P(O \le o_k \mid_{S=s_j})$	$P(O \le o_k \mid_{V=v_j})$	$P(S \le s_k \mid_{O=o_j})$	$P(S \le s_k \mid_{V=v_j})$	$P(V \le v_k \mid_{O=o_j})$	$P(V \le v_k \mid_{S=s_j})$	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
1	1	65	3	0.771386	0.896427	0.833872	0.847924	0.267071	0.334452	-	-	-
2	2	63	3	0.730375	0.896427	0.699981	0.682413	0.267071	0.305153	0.800832	0.507005	0.192826
3	3	63	3	0.730375	0.896427	0.699981	0.682413	0.267071	0.305153	0.748359	0.391523	0.192826
4	1	72	2	0.811578	0.912994	0.905528	0.909637	0.184081	0.34027	0.499277	0.680551	0.250393
5	1	63	2	0.730375	0.912994	0.699981	0.669189	0.184081	0.288159	0.426713	0.905027	0.22607
6	2	67	4	0.771386	0.870713	0.833872	0.853674	0.379324	0.372728	0.800832	0.467358	0.81271
7	0	64	1	0.730375	0.923927	0.699981	0.655703	0.129841	0.276449	0.510283	0.58029	0.362439
8	1	63	3	0.730375	0.896427	0.699981	0.682413	0.267071	0.305153	0.838266	0.445506	0.849854

Figure 37. Snapshot of the Nine Probabilities Developed to Test the Validity of an Observation

The observations which had all the nine probabilities less than the threshold specified were identified as valid observations. The observations which had any of the nine probabilities not lying between the thresholds specified were considered invalid. These observations were identified as either partially or totally erroneous, and were further screened to identify the probable erroneous parameters. An observation which had all the three parameters likely to be erroneous was filtered out as totally erroneous observation.

Stage Three: Identification of Erroneous Parameters.

A valid observation was used to derive a strategy for identifying the erroneous parameters. Intentional errors were introduced into the valid observations and the patterns among the set of nine probabilities observed by these changes were used to

identify the erroneous parameters in general. The screening strategy was devised by fixing one parameter of a valid observation and changing the other parameters to erroneous values. This was based on the assumption that for an observation to be partially valid, at least one of the parameters in the observation should be valid.

Separate analysis was conducted for stable and unstable flow observations as they possessed different probability distributions and would likely possess different patterns when the intentional errors were introduced into the observation. Figure 38 and Figure 39 show snapshots of valid observations (representing stable and unstable flows) that were considered to conduct experiment for identifying the erroneous parameters. An experimental analysis was conducted on each of the above observations to derive the patterns for identifying the erroneous parameters.

			Probabilistic Traffic Flow Relationships						Temporal Variation of Parameters		
O	S	V	$P(O \le o_k \mid_{S=s_j})$	$P(O \le o_k \mid_{V=v_j})$	$P(S \le s_k \mid_{O=o_j})$	$P(S \le s_k \mid_{V=v_j})$	$P(V \le v_k \mid_{O=o_j})$	$P(V \le v_k \mid_{S=s_j})$	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
11	60	15	0.732	0.917	0.861	0.81	0.76	0.899	0.2074	0.62305	0.188

Figure 38. Snapshot of a Valid Observation Representing Stable Flow Condition

			Probabilistic Traffic Flow Relationships						Temporal Variation of Parameters		
O	S	V	$P(O \le o_k \mid_{S=s_j})$	$P(O \le o_k \mid_{V=v_j})$	$P(S \le s_k \mid_{O=o_j})$	$P(S \le s_k \mid_{V=v_j})$	$P(V \le v_k \mid_{O=o_j})$	$P(V \le v_k \mid_{S=s_j})$	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
28	11	11	0.203	0.533	0.264	0.68	0.207	0.656	0.0802	0.132	0.203

Figure 39. Snapshot of a Valid Observation Representing Unstable Flow Condition

The process was sequentially carried out in eighteen steps to deduce all the patterns which would capture the nature of the most of the erroneous observations with respect to stochastic and conditional variation of the parameters. A valid observation representing stable flow condition was considered first and errors were introduced into the observation. In first six steps, patterns corresponding to single parameter being erroneous were identified while the next twelve steps dealt with identifying the patterns that reflect the invalidity of two parameters. The experimental design to deduce the patterns corresponding to the erroneous parameters is presented in Table 24.

The probabilistic patterns corresponding to extremely low and high values of parameters were examined separately. This was based upon the reason that the probabilistic nature of the observations differs for the extreme values of the parameters. Intentional errors were first introduced into the valid observation and the temporal variation of the parameters was examined to deduce patterns for identifying the erroneous parameters. Validity of an observation was based on how likely the difference is with its preceding observation as mentioned earlier in methodology. Hence, preceding observational values (O_{t-1}, $S_{t-1,}$ and V_{t-1}) were taken as a reference and the absolute difference of occupancy, speed, and volume parameters were varied in the experimental sequence designed to correspond to the invalid transitions in a time gap of 30 seconds, implying the erroneous nature of the transition. The probabilistic relationship between the parameters was then

examined in response to the intentional errors introduced. The derivation of the patterns to identify erroneous parameters (with reference to both temporal variation and probabilistic traffic flow relationship) for stable flow conditions is presented in Table 25.

Table 24. Patterns Representing Various Erroneous Observations

Pattern	Erroneous parameter	Description
1	Volume-	Only volume parameter is invalid and the value is lower than expected
2	Volume+	Only volume parameter is invalid and the value is higher than expected
3	Speed-	Only speed parameter is invalid and the value is lower than expected
4	Speed+	Only speed parameter is invalid and the value is higher than expected
5	Occupancy-	Only occupancy parameter is invalid and the value is lower than expected
6	Occupancy+	Only occupancy parameter is invalid and the value is higher than expected
7	Speed+, Volume+	Speed and volume parameters are invalid for a combination of high speed and volume values
8	Speed+, Volume-	Speed and volume parameters are invalid for a combination of high speed and low volume values
9	Speed-, Volume+	Speed and volume parameters are invalid for a combination of low speed and high volume values
10	Speed-, Volume-	Speed and volume parameters are invalid for a combination of low speed and volume values
11	Occupancy+, Volume+	Occupancy and volume parameters are invalid for a combination of high occupancy and high volume values
12	Occupancy+, Volume-	Occupancy and volume parameters are invalid for a combination of high occupancy and low volume values
13	Occupancy-, Volume+	Occupancy and volume parameters are invalid for a combination of low occupancy and high volume values
14	Occupancy-, Volume-	Occupancy and volume parameters are invalid for a combination of low occupancy and low volume values
15	Occupancy+, Speed+	Occupancy and speed parameters are invalid for a combination of high occupancy and speed values
16	Occupancy+,Speed-	Occupancy and speed parameters are invalid for a combination of high occupancy and low speed values
17	Occupancy-, Speed+	Occupancy and speed parameters are invalid for a combination of low occupancy and high speed values
18	Occupancy-, Speed-	Occupancy and speed parameters are invalid for a combination of low occupancy and lowspeed values

Table 25. Capturing Probabilistic Patterns of the Erroneous Observations in Stable Flow Conditions (Approach One)

Pattern No	Erroneous parameter	O	δ	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	S	δ	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	V	δ	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
	Default	10	1	0.480	54	6	0.623	14	1	0.460
1	Volume-	10	1	0.480	54	6	0.623	14	13	0.923
2	Volume+	10	1	0.480	54	6	0.623	14	6	0.960
3	Speed-	10	1	0.480	54	54	0.992	14	1	0.460
4	Speed+	10	1	0.480	54	46	1.000	14	1	0.460
5	Occupancy-	10	10	0.910	54	6	0.623	14	1	0.460
6	Occupancy+	10	90	0.972	54	6	0.623	14	1	0.460
7	Speed+, Volume+	10	1	0.480	54	46	1.000	14	6	0.960
8	Speed+, Volume-	10	1	0.480	54	46	1.000	14	13	0.923
9	Speed-, Volume+	10	1	0.480	54	54	0.992	14	6	0.960
10	Speed-, Volume-	10	1	0.480	54	54	0.992	14	13	0.923
11	Occupancy+, volume+	10	90	0.972	54	6	0.623	14	6	0.960
12	Occupancy+, volume-	10	90	0.972	54	6	0.623	14	13	0.923
13	Occupancy-, volume+	10	10	0.910	54	6	0.623	14	6	0.960
14	Occupancy-, volume-	10	10	0.910	54	6	0.623	14	13	0.923
15	Occupancy+, speed+	10	90	0.972	54	46	1.000	14	1	0.460
16	Occupancy+,speed-	10	90	0.972	54	54	0.992	14	1	0.460
17	Occupancy-, speed+	10	10	0.910	54	46	1.000	14	1	0.460
18	Occupancy-, speed-	10	10	0.910	54	54	1.000	14	1	0.460

Table 26. Capturing Probabilistic Patterns of the Erroneous Observations in Stable Flow Conditions (Approach Two)

Pattern No	Erroneous parameter	O	δ	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	S	δ	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	V	δ	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
	Default	2	12	15	0.732	0.916	0.860	0.809	0.759	0.899
1	Volume-	2	12	1	0.732	0.951	0.86	0.657	0.03	0.248
2	Volume+	2	12	20	0.732	0.644	0.86	0.841	0.92	0.923
3	Speed-	2	0	15	0.748	0.916	0.077	0.36387	0.759	0.988
4	Speed+	2	20	15	0.766	0.916	0.97	0.969	0.759	0.942
5	Occupancy-	0	12	15	0.733	0.173	0.694	0.809	0.903	0.809
6	Occupancy+	20	12	15	1	1	0.992	0.809	0.905	0.809
7	Speed+, Volume+	2	20	20	0.766	0.664	0.97	0.969	0.923	0.961
8	Speed+, Volume-	2	20	0	0.791	0.951	0.97	0.966	0.03	0.433
9	Speed-, Volume+	2	0	20	0.791	0.664	0.077	0.342	0.914	1
10	Speed-, Volume-	2	0	0	1	0.951	0.077	0.947	0.33	0.132
11	Occupancy+, volume+	20	12	20	1	1	0.992	0.969	0.914	0.928
12	Occupancy+, volume-	20	12	0	0.733	0.78	0.992	0.36	0.33	0.248
13	Occupancy-, volume+	0	12	20	0.733	0.128	0.694	0.969	0.914	0.928
14	Occupancy-, volume-	0	12	0	1	0.901	0.694	0.36387	0.198	0.248
15	Occupancy+, speed+	20	20	15	0.99	1	0.97	0.635	0.905	0.942
16	Occupancy+,speed-	20	20	15	0.631	1	0.947	0.819	0.905	0.942
17	Occupancy-, speed+	0	20	15	0.67	0.714	0.951	0.809	0.903	0.942
18	Occupancy-, speed-	0	0	15	0.698	0.714	0.08	0.809	0.903	0.998

Similar experimental analysis was conducted on the unstable flow observation and the patterns corresponding to the changes made with respect to stochastic and conditional variation of the parameters were identified. The changes made in accordance with the experimental design and the corresponding probabilistic patterns observed due to these changes are presented in Table 27 and Table 28.

*Results and Interpretation of Stage Three***.** This section presents the results obtained from the six case studies conducted on the stable and unstable flow observations and the process of screening the observation with reference to the patterns derived. Table 27 represents the 18 patterns that were derived form the experimental analysis conducted on a valid stable flow observation considered. From Table 29 it can be seen that pattern five doesn't indicate the erroneous nature of the observation for low values of occupancy. A possible explanation to this would be that the occupancy values for the stable flow conditions are quiet low (0-20%) and a drop from these values is feasible for all the combinations of valid speed and volume parameters.

Similarity among the patterns 2 and 13, 3 and 18, 4 and 17 as shown in the Table 29 could be attributed to the fact that maximum occupancy drop for stable flow (i.e., for low occupancy range conditions is feasible as explained earlier and doesn't have any effect by itself when combined with other erroneous parameters. The patterns developed could be used to identify the erroneous parameters. For instance, if an observation has probabilities matching with the pattern 1 (volume-), a conclusion that the volume parameter is erroneous and the value is less than expected to be valid is reached.

Table 27. Capturing Probabilistic Patterns of the Erroneous Observations in Unstable Flow Conditions (Approach One)

Pattern No	Erroneous parameter	O	δ	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	S	δ	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	V	δ	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
	Default	26	2	0.080	10	1	0.133	11	0	0.204
19	Volume-	26	2	0.080	10	1	0.133	11	10	0.904
20	Volume+	26	2	0.080	10	1	0.133	11	9	0.988
21	Speed-	26	2	0.080	10	10	1.000	11	0	0.204
22	Speed+	26	2	0.080	10	90	1.000	11	0	0.204
23	Occupancy-	26	2	0.080	10	90	1.000	11	9	0.988
24	Occupancy+	26	2	0.080	10	90	1.000	11	10	0.904
25	Speed+, Volume+	26	2	0.080	10	10	1.000	11	9	0.988
26	Speed+, Volume-	26	2	0.080	10	10	1.000	11	10	0.904
27	Speed-, Volume+	26	26	0.984	10	1	0.132	11	0	0.203
28	Speed-, Volume-	26	74	0.991	10	1	0.132	11	0	0.203
29	Occupancy+, volume+	26	74	0.991	10	1	0.132	11	9	0.988
30	Occupancy+, volume-	26	74	0.991	10	1	0.132	11	10	0.905
31	Occupancy-, volume+	26	26	0.984	10	1	0.132	11	9	0.988
32	Occupancy-, volume-	26	26	0.984	10	1	0.132	11	10	0.905
33	Occupancy+, speed+	26	74	0.991	10	90	1.000	14	1	0.189
34	Occupancy+,speed-	26	74	0.991	10	10	1.000	14	1	0.189
35	Occupancy-, speed+	26	26	0.984	10	90	1.000	14	1	0.189
36	Occupancy-, speed-	26	26	0.984	10	10	1.000	14	1	0.189

Table 28. Capturing Probabilistic Patterns of the Erroneous Observations in Unstable Flow Conditions (Approach Two)

Pattern No	Erroneous parameter	O	δ	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	S	δ	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	V	δ	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
	Default	5	2	1	0.2030384	0.085322	0.263805	0.97011	0.104881	0.0986
19	Volume-	5	2	20	0.2030384	0.8738	0.263805	0.3617	0.846	1
20	Volume+	5	0	11	0.872814	0.533433	0.198	0.4351	0.2069	0.91605
21	Speed-	5	20	11	0.957803	0.533433	0.978	0.972	0.2069	0.878
22	Speed+	0	2	11	0.203	0.3146	0.099	0.679	0.8104	0.6561
23	Occupancy-	20	2	11	0.994	1	0.9626	0.679	0.8727	0.6561
24	Occupancy+	5	20	20	0.957803	0.8738	0.978	0.969	0.846	0.96013
25	Speed+, Volume+	5	20	1	0.957803	0.085322	0.978	0.968	0.104881	0.433
26	Speed+, Volume-	5	0	20	0.872814	0.8738	0.198	0.3424	0.846	1
27	Speed-, Volume+	5	0	1	0.872814	0.085322	0.198	0.9471	0.104881	0.1328
28	Speed-, Volume-	20	2	20	0.994	1	0.9626	0.3617	0.9141	1
29	Occupancy+, volume+	20	2	1	0.994	0.784	0.9626	0.97011	0.333	0.098
30	Occupancy+, volume-	0	2	20	0.203	0.128	0.099	0.3617	0.941	1
31	Occupancy-, volume+	0	2	1	0.203	0.901	0.099	0.97011	0.198	0.098
32	Occupancy-, volume-	20	20	11	1	1	0.998	0.97	0.872	0.878
33	Occupancy+, speed+	20	0	11	0.998	1	0.9466	0.453	0.872	0.916
34	Occupancy+,speed-	0	20	11	0.67	0.314	0.96	0.97	0.8104	0.878
35	Occupancy-, speed+	0	0	11	0.698	0.314	0.087	0.453	0.8104	0.916
36	Occupancy-, speed-	5	2	1	0.2030384	0.085322	0.263805	0.97011	0.104881	0.0986

Table 29. Patterns for Screening the Stable Flow Observations[10]

Pattern No	Erroneous parameter	$P(O \le o_k \mid_{S=s_j})$	$P(O \le o_k \mid_{V=v_j})$	$P(S \le s_k \mid_{O=o_j})$	$P(S \le s_k \mid_{V=v_j})$	$P(V \le v_k \mid_{O=o_j})$	$P(V \le v_k \mid_{S=s_j})$	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
1	Volume-	0	1	0	0	0	0	0	0	0
2	Volume+	0	0	0	0	0	0	0	0	1
3	Speed-	0	0	0	0	0	1	0	1	0
4	Speed+	0	0	1	1	0	0	0	1	0
5	Occupancy-	0	0	0	0	0	0	0	0	0
6	Occupancy+	1	1	1	0	0	0	1	0	0
7	Speed+, Volume+	0	0	1	1	0	1	0	1	1
8	Speed+, Volume-	0	1	1	1	0	0	0	1	0
9	Speed-, Volume+	0	0	0	0	0	1	0	1	1
10	Speed-, Volume-	0	1	0	0	0	0	0	1	0
11	Occupancy+, Volume+	1	1	1	0	0	0	1	0	1
12	Occupancy+, Volume-	1	0	1	0	0	0	1	0	0
13	Occupancy-, Volume+	0	0	0	0	0	0	0	0	1
14	Occupancy-, Volume-	0	0	0	0	0	0	0	0	0
15	Occupancy+, Speed+	1	1	1	1	0	0	1	1	0
16	Occupancy+,Speed-	1	1	0	0	0	0	1	1	0
17	Occupancy-, Speed+	0	0	1	1	0	0	0	1	0
18	Occupancy-, Speed-	0	0	0	0	0	1	0	1	0

[10] 1- Cumulative probability >.95, 0-cumulative probability <.95.

In case an observation matches with overlapping patterns (e.g., 2 and 13), it could be concluded that either occupancy or volume parameter is erroneous but a definitive conclusion that both of the parameters are erroneous could not be reached.

The patterns derived from the experimental analysis conducted on the unstable flow observation considered are presented in the Table 30. Similarity between patterns (i.e., 28 and 21) representing speed, and volume, and speed conditions was observed. This could be attributed to the fact that the low value of the volume parameter considered doesn't have an effect on the erroneous nature of the observation when combined with low values of speed. Patterns 22 and 26 overlap, suggesting that either speed or volume parameter is erroneous.

The real-time data set was first screened with a threshold value of .95 as mentioned earlier. The erroneous observations (with probabilities greater than .95) were further matched with the patterns developed to identify erroneous parameters. The implementation of data screening algorithm for the real-time data is presented in Figure 40. Table 31 shows the results of screening process conducted. The results from the table showed that 75 percent of the observations were likely to be valid, while the remaining 20 percent of observations were identified as partially valid observations. The remaining five percent of the observations could be either totally erroneous observations or observations representing conflicting conclusion. These conflicting conclusions could be reached when an observation matches with two or more patterns which indicate contradictory results.

Table 30. Patterns for Screening the Unstable Flow Observations

Pattern No	Erroneous parameter	$P(O \le o_k \mid_{S=s_j})$	$P(O \le o_k \mid_{V=v_j})$	$P(S \le s_k \mid_{O=o_j})$	$P(S \le s_k \mid_{V=v_j})$	$P(V \le v_k \mid_{O=o_j})$	$P(V \le v_k \mid_{S=s_j})$	$\frac{P\{O_t - O_{t+1} \le \delta \mid_o\}}{P\{O_{t+1} - O_t \le \delta \mid_o\}}$	$\frac{P\{S_t - S_{t+1} \le \delta \mid_s\}}{P\{S_{t+1} - S_t \le \delta \mid_s\}}$	$\frac{P\{V_t - V_{t+1} \le \delta \mid_v\}}{P\{V_{t+1} - V_t \le \delta \mid_v\}}$
19	Volume-	0	0	0	1	0	0	0	0	0
20	Volume+	0	0	0	0	0	1	0	0	1
21	Speed-	0	0	0	0	0	0	0	1	0
22	Speed+	1	0	1	1	0	0	0	1	0
23	Occupancy-	0	0	0	0	0	0	1	0	0
24	Occupancy+	1	1	1	0	0	0	1	0	0
25	Speed+, Volume+	1	0	1	1	0	1	0	1	1
26	Speed+, Volume-	1	0	1	1	0	0	0	1	0
27	Speed-, Volume+	0	0	0	0	0	1	0	1	1
28	Speed-, Volume-	0	0	0	0	0	0	0	1	0
29	Occupancy+, Volume+	1	1	1	0	0	1	1	0	1
30	Occupancy+, Volume-	1	0	1	1	0	0	1	0	0
31	Occupancy-, Volume+	0	0	0	0	0	1	1	0	1
32	Occupancy-, Volume-	0	0	0	1	0	0	1	0	0
33	Occupancy+, Speed+	1	1	1	1	0	0	1	1	0
34	Occupancy+,Speed-	1	1	0	0	0	0	1	1	0
35	Occupancy-, Speed+	0	0	1	1	0	0	1	1	0
36	Occupancy-, Speed-	0	0	0	0	0	0	1	1	0

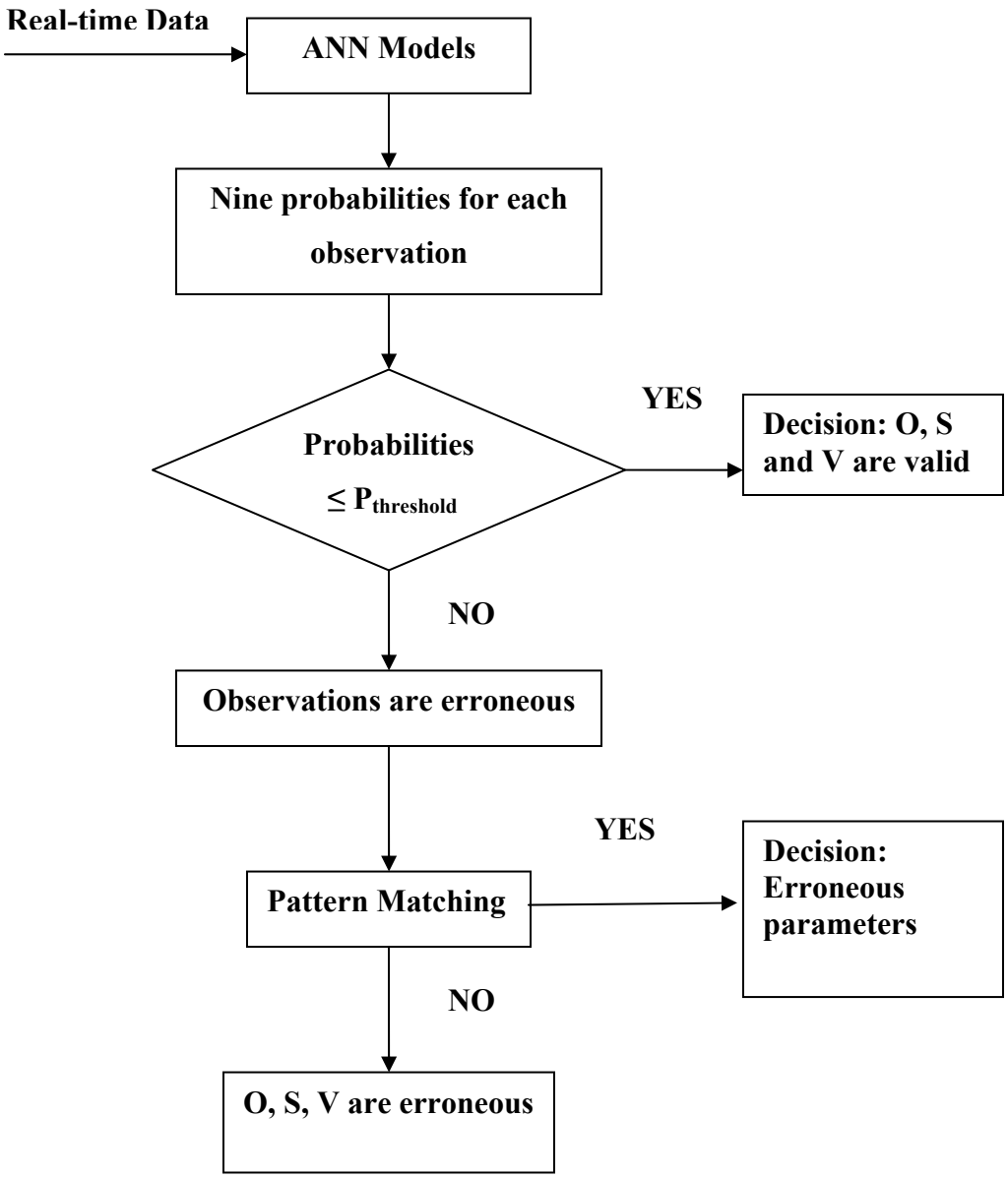


Figure 40. Implementation of Data Screening Algorithm for Real-time Traffic Data

Table 31. Results of Implementation of Data screening Algorithm on Real-time Data

Pattern No	Erroneous parameters	Total number of observations 51,837		Percentage of observations
		Valid observations	Invalid observations	
		38657	-	74.57

1	Volume-	-	3997	7.71
2	Volume +	-	417	0.804
3	Speed -	-	23	0.044
4	Speed +	-	76	0.146
7	Speed +, Volume +	-	7	0.013
8	Speed +, Volume -	-	19	0.036
9	Speed- Volume +	-	40	0.077
15	Occupancy +, speed+	-	3745	7.22
16	Occupancy +, speed -	-	2043	3.94
18	Occupancy -, speed -	-	56	0.108
21	Speed - (unstable flow)	-	19	0.036
22	Speed + (unstable flow)	-	9	0.017
23	Occupancy- (unstable flow)	-	114	0.219
24	Occupancy + (unstable flow)	-	9	0.017
25	Speed +, volume+ (unstable flow)	-	113	0.217
27	Speed -, volume+ (unstable flow)	-	7	0.013

Conclusions

The non-linear nature of the stochastic and conditional relationships between the parameters was reasonably captured using 38 Multi-layer Feed-forward Networks, except for the stochastic variations representing high occupancy conditions. The screening algorithm devised was efficient in judging the validity of the real-time data format with 95% probability. Most of the patterns deduced were capable of identifying the erroneous parameters in the observation thus classifying them into partially valid observations. The following were contributions of this research study.

This approach can be implemented online or offline to screen the observations before or after they are streamed into data warehouses and is user adaptable as it does not impose any restrictions on the thresholds. This approach could be used for imputing the erroneous parameters in the sense that the patterns deduced were capable of identifying the erroneous parameter, and also communicated information about the dimension of the

erroneous parameter (e.g., volume- suggests that volume parameter is erroneous and the value is less than expected). Thus, this study provides preliminary information required for imputation of erroneous parameters. The general approach formulated in this study can be applied to different locations while the transferability of the model needs to be authenticated with more tests using the data from other locations and are not dealt with in this study. The algorithm can also be used to identify the operational status of detectors and detect calibration problems that may call for immediate maintenance due to its real-time screening ability.

Data Mining: Causal Factors of Vehicle Accidents

Background and Analysis

In the context of the problem, the traffic data available in data warehouses (DW) includes:

a) Historical time-series data from sensors (i.e., average speed, volume and occupancy[11]),

b) Incidents, usually provided by police/patrol repots or other types of specialized sensors such as video cameras. Incident reports contain information about the type of the incident, (crash, stall, etc.), location, time, severity, and duration[12].

In general, sensor data is provided in real-time, while incident reports are manually entered into the system immediately after the fact, or in batch mode. Incidents and sensor information can be mapped together by location and time (the time of the incident and the sensor record timestamps). The combined set of historic sensor data and incidents will be referred to as "traffic data." Figure 41 summarizes the main concepts associated with the nature of the data available in traffic data warehouses.

[11] The most common type of sensor in transportation systems is the inductive loop detector.
[12] Of particular interest for this question, are the types of incidents that can be classified as vehicle accidents.

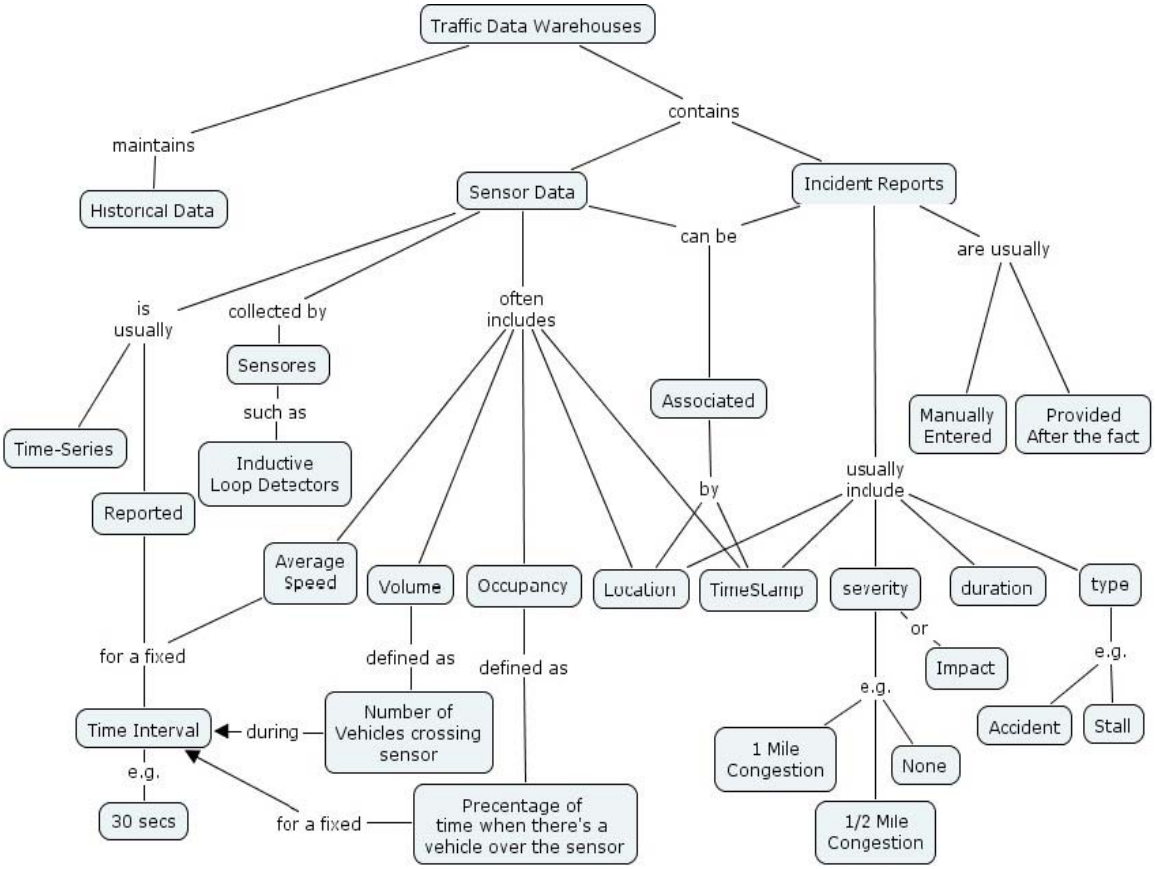


Figure 41. Traffic Data Available in Data Warehouses (DW) for Detecting and Predicting Traffic Accidents

The traffic data, as previously defined, can be used to build inference models to: a) identify changes in traffic that might occur as result of an accident (*accident detection*), b) to detect changes in traffic behaviors that are likely to lead to accidents (*accident prediction*), and c) to provide the necessary information to proactively interfere with traffic conditions (for instance, through dynamic speed control signals, and traffic redirection) in order to avoid the occurrence of traffic accidents (*accident prevention*).

Our goal is to identify causal factors for traffic accidents using only traffic data. In the next subsection, we provide a brief literature review on traffic accident prediction from sensor data; the goal of the literature review is to identify variables that are good predictors for traffic accidents. Good predictors are not necessarily causal factors, but we use these metrics as a starting point for causal discovery algorithms. In the subsection entitled "*A Model for Causal Discovery in Traffic Data*," we introduce and discuss a model for causal discovery in traffic data. In the subsection, "*Expected Results and Limitations,"* we conclude with a few comments about expected results, restrictions and limitations of the proposed approach. Further background information is available in papers by Glymour, Pearl, Heckerman, and others [64] [65] [66] [67] [68] [69] [70]. This review essentially describes the relationships between associational (purely probabilistic Bayesian networks) and causal networks.

Literature Review: Finding the Best Predictors

There's very little literature available on real-time causal analysis for traffic accidents from traffic data. The majority of the literature referring to causality in traffic accidents [4] [8] [10] [13] [16] targets accident forensics and reconstruction (usually for accident prevention and to identify guilty parties) and relies on much more information than just traffic data – usually vehicle size, dynamics, driver profile, etc. Even under these circumstances, some authors have clearly reported the difficulties associated with causal analysis for traffic accidents [11] [12].

There are, however, a relatively larger number of publications associating only traffic data with accidents, both from a perspective of online detection and for prediction of accidents[13] (not causal factors). Some research efforts in that area are primarily focused on capacity-driven measures of traffic flow in freeway segments. Such measures would include, for instance, the Average Annual Daily Traffic (AADT) [17][14]. Other efforts have focused on a more temporal approach to the problem.

In 1964, Solomon [21] published what is usually referenced today as the first work associating vehicle speed (traffic data) and frequency of accidents. Citing their work, Lave [24] in 1985 used aggregate speeds to show that accident rates (more precisely, fatal accident rates) were more dependent on speed variance across vehicles than on average speed. In 2000, Oh [3] proposed a different approach to identify which traffic metrics were better real-time predictors for traffic accidents. Avoiding temporal models, Oh used an accident database to take five-minute traffic samples (averages) much before each accident (30 minutes before) and right before each accident (five minutes before). Each sample set included the average sensor value (occupancy, flow and speed) as well as their standard deviation. Oh's assumption was that traffic right before an accident could be labeled as "disruptive," while the other samples were considered "normal" – a strong assumption. Nevertheless, the author built a Bayesian classifier to identify which of the variables considered were more predictive of the type of traffic (which was directly associated with a traffic accident). Once again, there were significant indications that standard deviation of speed (this time, temporal variation) was the better predictor of "disruptive" traffic.

Most of these results were based on aggregated traffic values (for speed, variations and crashes) and sometimes (as in [24]) were indirect approximations of variations, like differences in percentile values on speed estimation of up to 30 minutes-long intervals. In 2002, Davis [23] criticized the inferences based on data aggregations, noting that aggregate relationships between speed, speed variance, and crash frequency are not necessarily supported by the original data.

More recently, other approaches such as the one proposed by Lee [1] [27], and Kockelman [22] utilized 30-second sensor aggregation data as predictors, with significant improvements. Kockelman augmented the loop-detector based data (volume and

[13] Even though, in some of these works, we sometimes find implicit statements about causal dependencies, it is important to highlight that the fact that predictive models for traffic accidents perform well or not don't necessarily imply causality.

[14] Variations of this approach focused on augmenting aggregated volume metrics (such as AADT) with information about geometric structure variables of the roadway [15] [16], usually not included in the DW.

occupancy) with estimative of average vehicle length to calculate "instantaneous" speed variations. Lee [27] proposed a more detailed model, isolating not one, but a number of what he called "accident precursors" for real-time crash prediction. He identified three main indicators: a) the average variation of speed on each lane, b) the average variation of speed difference across adjacent lanes, and c) traffic density. One year later, in [1] Lee revises his work and concludes that the average variation of speed difference across adjacent lanes (precursor "b") was not a strong indicator of traffic accidents, as initially expected. He proposes instead a new precursor that is essentially based on the differences in speed between upstream and downstream traffic, which was a direct indication of queue formation.

In summary, most of the research in predictive models for traffic accidents from sensor data seem to rely on direct measures of density and average speed, with variations in the way that averages and variations are calculated across lanes and between upstream and downstream traffic.

For our analysis, the volume, occupancy, and speed available in the DW are local to each sensor. We build different levels of data aggregation (as explained in the subsequent item) that will include averages and variances of each metric per lane, as well as averages and variations between lanes and about traffic density.

Causal Discovery from Observational Data

Causal discovery from observational (or non-experimental) data is a topic that has created, for the last two decades, as much of a revolution as it has created polemic. There are a number of philosophical and mathematical arguments used to defend, as well as to criticize the possibility of discovering causal dependencies from observational data. References [20] [21] [65] [66] [67] [68] [69] [70] provide some additional information both in favor and against the concept. Stepping aside from the debate, the goal here is to briefly describe, abdicating from proofs and advanced mathematical arguments, how causal dependencies can be obtained from non-experimental data. In particular, we refer to methods for causal discovery that are based on the notion of Bayesian Networks.

Simply stated, a Bayesian Network (BN) is essentially composed by a Directed Acyclic Graph (DAG) and a joint probability distribution (JPD). Under certain conditions, the structure of the DAG represents the same conditional independence relationships found in the JPD, with nodes in the graph representing random variables and directed arcs (or the lack of them) representing direct probabilistic dependence (or independence) between variables. For a given JPD, there are many DAGs (a class of graphs) that can be constructed to satisfy the probabilistic association between variables. These graphs are known as the "Markov Equivalent" graphs of the distribution.

A causal graph is one instance of the sub-class of Markov Equivalent graphs, that further constrain the DAG by requiring that the direction of the arcs will not only indicate probabilistic dependence but also a causal relation between variables. These instances are sometimes referred to as causal Bayesian Networks. A simple illustration of this difference is shown in Figure 42.

Consider, for example, a dataset containing three variables (S: smoking, Y: yellow fingers and C: lung cancer) with corresponding conditional independence relations shown in Figure 42a. In this example, let's assume that having yellow fingers (Y) is a feature that is independent of having lung cancer (C), given the fact that the individual is a smoker. That is, amongst smokers, the two features are not associated with each other, but they are certainly both associated with the smoking behavior (S).

Under the Markov Condition[15], there are three Bayesian networks, in Figure 42b, that correctly represent the joint probability distribution. That is, in all graphs shown in Figure 42b, given the Markov Condition, (S) is associated to (C) and (Y), and (Y) is independent of (C), given (S). That last conditional independence relation is represented in graphs by the notion of d-separation, which states that a set of edges (A) in a DAG is d-separated by a set of edges (B), given another set (C), if there all paths from (A) to (B) are blocked by the set (C). In this example, there's no directed path from (Y) to (C) unless going trough (S), so (Y) is d-separated from (C), given (S).

However, in order to assume a causal interpretation for the graph (in additional to the probabilistic dependence associations) the edges must be further constrained. Let's consider that smoking causes yellow finger and also causes lung cancer, but having yellow fingers or lung cancer do not lead one to start smoking. Furthermore, let's assume that having a yellow finger (or tar-stained finger), does not lead one to start smoking or to have lung cancer, in the same way that having long cancer does not lead one to start smoking or to develop a yellow finger.

Under these assumptions, the corresponding causal graph for that same distribution is the one shown in Figure 42c. It is important to note that all the graphs shown here are Markov Equivalent graphs for the same joint distribution, that is, they all comply with the independence relations between variables by ensuring the appropriate d-separation between the corresponding nodes.

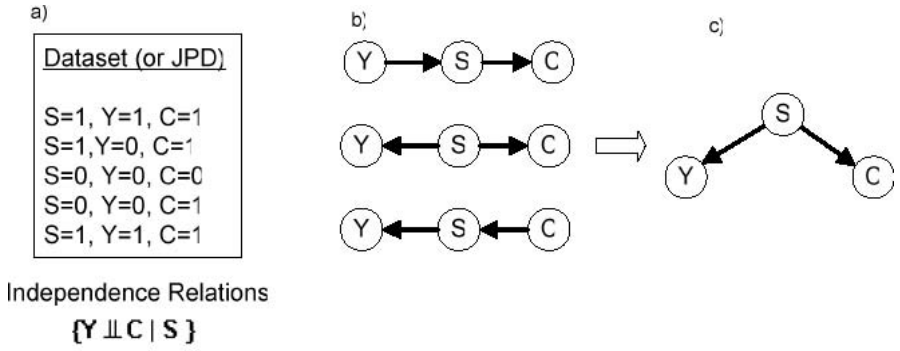


Figure 42. An Example of Causal Interpretation of Bayesian Networks

[15] The Markov Condition in DAGs is equivalent to the concept of d-separation between sets. It states that given a joint probability distribution (JPD), if a set of nodes X is independent for a set of nodes Y given a third set Z, then in a DAG that represents the JDP, all directed paths from X to Y are blocked by Z.

It is also important to note that the constraints between a causal graph and the probability distribution are more restrictive. As described in detail by Scheines [20] [21] these three conditions must hold in order to satisfy the causal interpretation of the graph:

a) *The Causal Markov Condition:* The directed connections of the DAG that satisfy the Markov Condition for a given joint probability distribution represent the causal relations between variables in the correct direction. This essentially means, that out of the graphs that satisfy the Markov Condition, the subset of graphs whose directed arcs indicate causal relations also satisfy the Causal Markov Condition. This is a stronger assumption than the Markov Condition.

b) *Faithfulness*: It essentially states that the graph is complete in the sense that all probability distributions in the data are present in the graph.Considering only the Causal Markov Assumption, a causal graph can generate data that will necessarily contain all the independence relations defined in the graph, but there are no assumptions that (by chance) the data could also contain additional independence relations not initially in the graph. The Faithfulness condition requires that this will not happen, that is, that every independence relation existing in the data is represented in the graph and cannot occur by chance, or by perfect cancellation between competing causal relations over the same target variable.

c) *Causal Sufficiency*: The set of measured variables "M" includes all of the common causes of pairs in M. That is, the causal inference requires conditional independence checks that involve the causal variables. Unobserved variables can not affect pairs of variables in the set in such a way that will make them "look" correlated when they are, in fact, independent conditional to the "unobserved" variable.

Figure 43 shows some of the concepts involved in the process of causal discovery. At some level, all algorithms for Bayesian network induction (causal or not) rely on independence tests between variables. Not discussed in this paper are all the underlying assumptions associated with the independence tests themselves (e.g. partial correlation, chi-square, t-test, etc.). In this section, we consider that conditional independence tests between variables and sets of variables in the dataset can be performed "reliably" by some oracle.

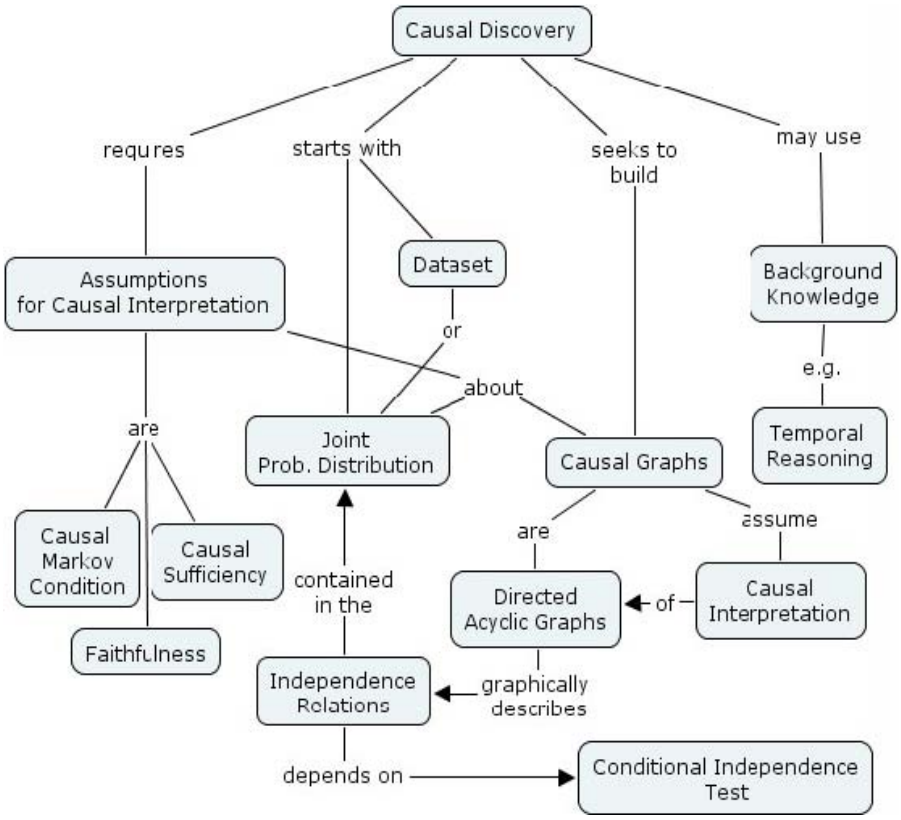


Figure 43. Concept Graph for Causal Discovery

The search for causal graphs from observational data starts with a search for probability independence relations between variables in a set. This is, in fact, very similar to traditional methods to induce Bayesian networks or Markov Random Fields from a given dataset (or a given joint probability distribution).

In general, methods are classified as *global* and *local*, based on the sequence variable selection for test. In the worst case, all possible independence tests between all possible sets of variables should be verified. For instance, a global algorithm search for conditional independence in a set containing three variables (A, B and C), would have to perform the tests for independence $\{A_{\perp\!\!\!\perp} B, A_{\perp\!\!\!\perp} C, B_{\perp\!\!\!\perp} C\}$, and for conditional independence $\{A_{\perp\!\!\!\perp} B|C, A_{\perp\!\!\!\perp} C|B, B_{\perp\!\!\!\perp} A|C, B_{\perp\!\!\!\perp} C|A, C_{\perp\!\!\!\perp} B|A, C_{\perp\!\!\!\perp} B|C\}$[16]. This exponential approach is not practical for larger datasets. There are, however, global algorithms like the graph pattern search available in the PC algorithm [21] that scales better[17] for larger (and sparse) graphs. The graph search in algorithm in PC essentially eliminates edges for positive independence tests at each step. By eliminating lower order edges at the beginning, the algorithm can reduce the complexity of the search.

Local approaches on the other hand, start from a single variable and progressively build the dependence edges through independence tests with the variable neighbors. Most local

[16] The notation $X_{\perp\!\!\!\perp} Y|Z$ represents "X is independent of Y, given Z"

[17] In [28], Dai shows (empirically) that for larger non-sparse graphs, the order of the of the partial independence tests increases and the quality of results tend to drop, for a fixed number of samples.

Markov Blanket (MB) discovery algorithms (like PCX [29], HITON [30] and the Grow-Shrink (GS) algorithm [31]) follow this strategy.

On a Directed Acyclic Graph that satisfies the Markov Condition, a Markov Blanket of a variable X, is composed by all its direct parents, plus its direct children and the parents of its children. If the graph satisfies only the Markov Condition, the Markov Blanket for a target variable T, or MB(T), essentially establishes that T is independent of all variables in the graph, given the set of variables contained in its Markov Blanket.

A complete Bayesian network can be constructed by a successive MB search over all the variables which is often more efficient than global searchers. However, for causal analysis global approaches are recommended, as the orientation of edges around the boundary is further constrained (they must imply causality) and requires d-separation checks with other variables, outside the MB.

A Model for Causal Discovery in Traffic Data

The traffic data is essentially a discrete time-series where some variables (sensor data) are usually sampled (or aggregated) at fixed intervals, while other variables (traffic incidents) might have completely stochastic behavior. The causal (or associational) information that can be extracted from the data will vary greatly, both in terms of content and interpretation, depending on the aggregation strategy used for the data.

For instance, if traffic data is aggregated on a monthly basis, and associated with a corresponding aggregation (frequency) of accidents, the probabilistic association between the variables is likely to present yearly seasonality (maybe due to influences of rains, or freezes, or maybe due to changes in traffic patterns between summer – school vacation – and winter). Such aggregation is unlikely to indicate, for instance, any correspondence between rush hour traffic versus frequency of accidents. Conversely, a five-minute aggregation of variables is likely to provide temporal information between cause and effect, maybe indicating the immediate conditions of traffic that might have lead to an accident, as well as the immediate effects of the accident on traffic, as time progresses.

The choice of aggregation depends on the objective of the search. For this purpose, we have chosen a short-term temporal analysis of causes of traffic accidents (assuming they exist). That is, instead of trying to infer causes of traffic accidents from, let's say, daily average traffic density, and daily average speed, we make the assumption that there are local traffic conditions that lead to the occurrence of the accident. For instance, a sudden variation in traffic density or a progressive increase on speed variation between lanes can be causes that, within reasonable time bounds, will lead to an accident. The object is to analyze the problem from a time-series perspective.

*Building the Dataset***.** In order to illustrate the type of data that would be used for this analysis, consider the example in Figure 44. This example is based on actual sensor data collected and made available by the Minnesota Department of Transportation (MnDOT) for the Twin Cities metropolitan area [35]. Historic records of this data are stored and maintained by the Traffic Data Research Laboratory [36]. This specific example shows a segment of the Interstate 35W. In the police report, the accident was estimated to have occurred at 14:35 hour. (The sensor data includes basically occupancy and volume

information from inductive loop detectors, ILD, embedded in the freeway.) The raw values for volume and occupancy are reported on 30-second intervals as aggregates

Traffic accidents are not integrated with the database but we've contacted the MnDOT and asked for the list of accidents on specific locations. Identifying each of the sensors with a map of the freeway sections, we estimated average vehicle speed (from volume and occupancy) and plotted, in Figure 44, the 15-minute average speed of all lanes before and after the accident. The location of the accident is shown by the solid arrow in the map at the bottom of the image.

The example basically shows the effects of the accident on the average speed. Our goal is to identify the opposite, that is, changes in traffic conditions that could lead to traffic accidents.

This is an investigative procedure so we start with a set of variables per location, with different aggregations and at different time intervals. After the first inference (discussed in the subsection below entitled "*The Search Procedure"*) we might choose to eliminate some of the weekly associated variables and reduce the set for a new search. Consider, for instance, the sample dataset constructed for the segment illustrated in Figure 44. In that figure, the freeway segment has five stations[18], namely S657, S658, S659, S664, and S665 (from left to right).

[18] Each station is a collection of detectors. In the case of IDL sensors, a station over a three-lane freeway might include three detectors, one for each lane.

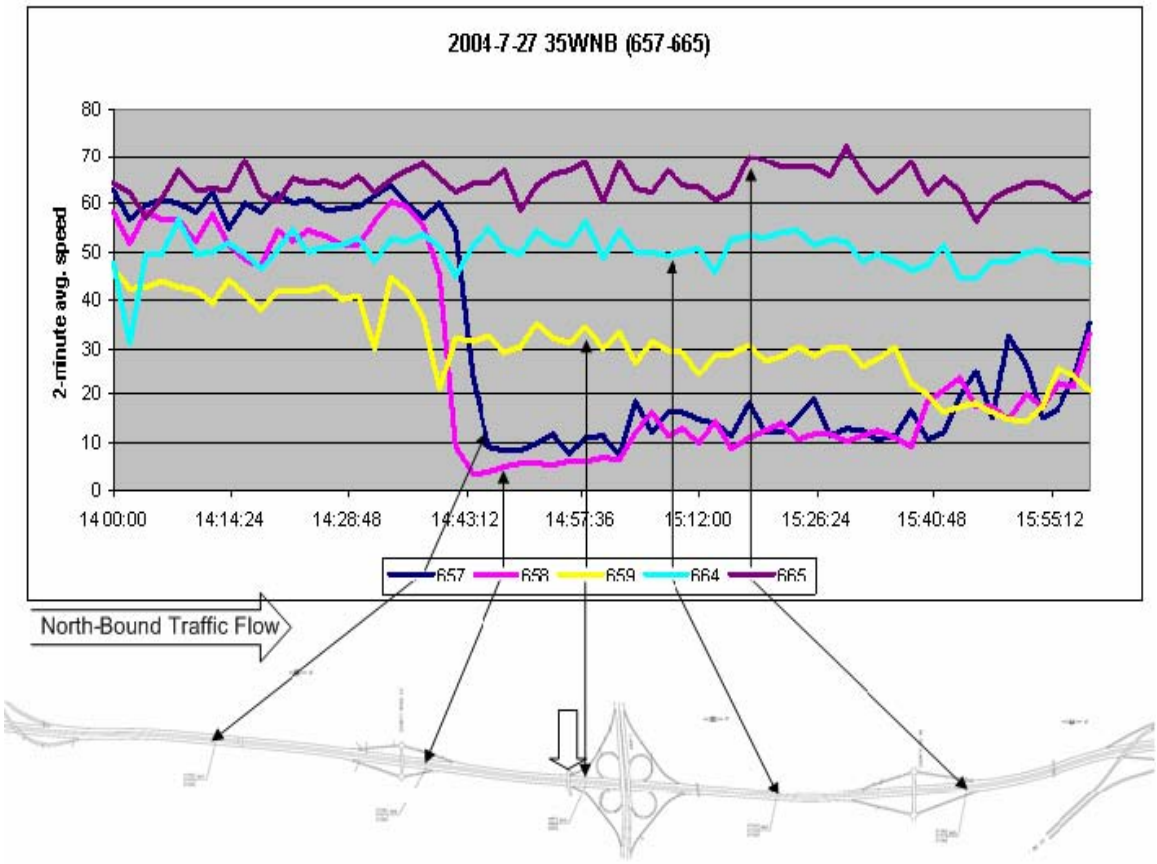


Figure 44. An example of the changes in average speed due to an all-lanes traffic accident. Interstate 35W, North Bound. July 27, 2004. Traffic accident occurred approximately at 14:35h and caused a 3-mile long congestion over all lanes in the freeway. All clear reported at 15:38h.

As a first approach, for instance, we can define the following variables for each station:

- Average speed across all lanes (P)
- Average temporal variation of speed, for all lanes (Q)
- Average variation of speed between lanes (R)
- Occupancy/Volume ratio - or O/C ratio for short (indicative of density) (S)
- Average temporal variation of O/C ratio, for all lanes (T)
- Average variation of O/C ratio between lanes (U)

These metrics can be calculated for different time intervals (five-minute averages, for instance) and for different lags from a current time stamp. An arbitrary letter was assigned to each variable just to simplify notation – note that we assume no background knowledge between variables so their label is unimportant during the discovery process. The traffic accident variable (A) can be binary, as we are not concerned with causal effects for different classes of accidents.

Consider a dataset including the set of variable {P,Q,R,S,T,U} for a different number of time lags (let's say 2) separated by a 10 minute interval, and a pre-defined uniform

(amongst all variables) time aggregation (let's say five-minute averages). As described, a snapshot of one sensor in the freeway would provide the following set: $\{P_2,P_1,P_0,Q_2,Q_1,Q_0,R_2,R_1,R_0,S_2,S_1,S_0,T_2,T_1,T_0,U_2,U_1,U_0\}$, at station D (a set of detectors) and instance T. In this sample, looking at variable P, for instance, we have:

P_0: 5-minute average of P(t) at time t-20 minutes.
P_1: 5-minute average of P(t) at time t-10 minutes.
P_2: 5-minute average of P(t) at t.

Define as *S* the set of variables defined for a specific detector D, at time t, as follows: $S(D,t) = \{P_2,P_1,P_0,Q_2,Q_1,Q_0,R_2,R_1,R_0,S_2,S_1,S_0,T_2,T_1,T_0,U_2,U_1,U_0\}$

In order to construct our dataset we will follow the procedure:

a) Identify all reported accidents in the database that qualify as the types of accidents we're interested in the analysis. We might want to eliminate from this set accidents that seem to be outliers in their characteristics (e.g., involving too many cars, with very severe impact or accidents involving special types of vehicles such as heavy trucks or motorcycles). The more homogeneous the characteristics of the accidents included the higher will be our chances of mitigating the effects of confounded, unmeasured variables.
b) As we're primarily concerned with predictors for accidents, we will collect samples $S(D,t)$ for a time instant immediately before the accident (from the set we've chosen, that will give us two time lags before that). This will create a set of samples that we will identify as precursors, that is, they will be tagged as events that occurred before (given a time window) a known accident. If we choose a single sensor, the set will be <Time, $S(D,t)$, Accident = 1>
c) Based on the time-stamps selected for all samples in $S(D,t)$, we will build our set of negative cases by choosing similar sets from the same location/sensor.

The selection of the data is critical for the approach. There are many external variables that are not being measured, so special care should be taken to try comparing data where these effects are minimized. For instance, by building positive and negative cases from the location we remove the geometric parameters (cross-section, visibility, etc.) from the equation, facilitating the process of finding true associations.

If a single sensor is chosen for each accident the dataset, visualized graphically, will be as illustrated in Figure 45. Where each node represents a random variable measured over many accidents (accident=true) and other "similar" traffic conditions where accidents didn't occur.

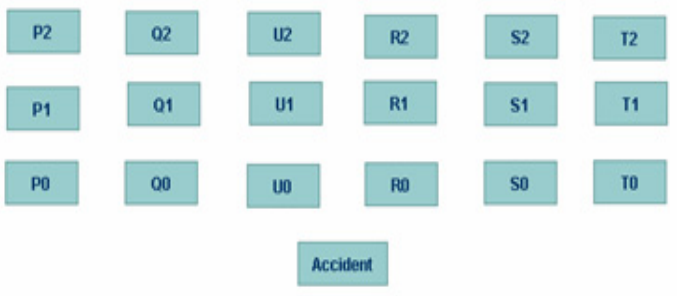


Figure 45. A Sample Data Set Used as Input for the PCX Algorithm

Temporal causal effects will show as causal edges from variables of the type $X_{i-n} \rightarrow X_i$.

For instance, depending on the time interval chosen between lags (Δt), we should expect temporal causal relations between variables of the same time, as illustrated in Figure 46.

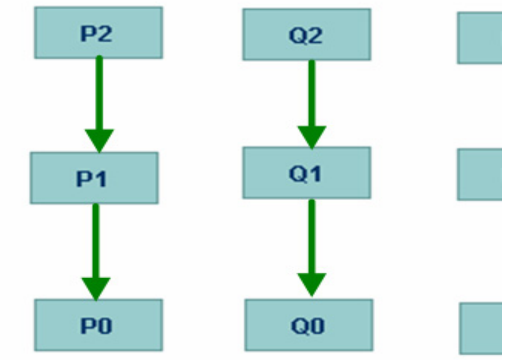


Figure 46. Expected Causal Temporal Relations Between the Same Types of Variables. (Note that this is only a partial view of the graph.)

The same is true for temporal dependencies between variables. For instance, if the algorithm identifies a causal relation between P_2 and Q_1, let's say $P_2 \rightarrow Q_1$ (P_2 causes Q_1), then it would be represented as an edge between P_2 and Q_1. The interpretation is that variable P has a causal effect on variable Q with a Δt lag. If the graph is consistent, the same relations (edges) would be present between $P_1 \rightarrow Q_0$, as illustrated in Figure 47.

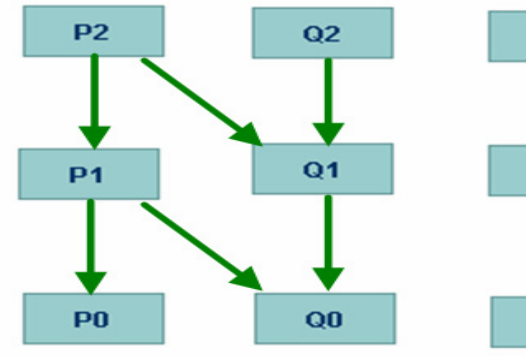


Figure 47. A Temporal Causal Relation Between Variables *P* and *Q*

This technique allows time-series data to be analyzed as static datasets. As we have a priori information about the variables, we can augment the approach by adding background knowledge that will help the search procedure. For instance, we can add a set of "forbidden" edges to the graph. That is, for instance, illustrated as a set of red arrows in Figure 48. The red arrows are causal edges that should not be explored by the search procedure, as they imply causal factors going backward in time (which is impossible).

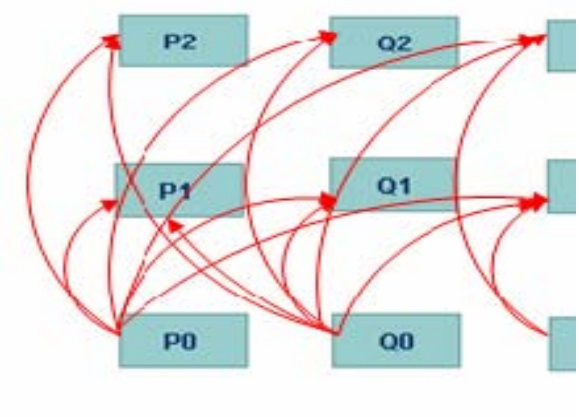


Figure 48. The Sample Dataset with Background Knowledge, a Set of Forbidden Causal Edges. (Note that, for simplicity, not all edges are shown in the picture.)

The final dataset then would include a table of data attributes where each example contains the set of variables $\{P_2,P_1,P_0,Q_2,Q_1,Q_0,R_2,R_1,R_0,S_2,S_1,S_0,T_2,T_1,T_0,U_2,U_1,U_0\}$ shown in Figure 45, and background knowledge about forbidden edges (back in time) and other additional background causal information we might have about the data, in the form of forbidden causal edges, or forced causal edges.

One last point to be made about the data preparation is about nonlinear relationships between variables. Chu [64], showed that, in causal discovery from time-series data, non-linear relations between variables can compromise the validity of the independence tests. In his paper, he proposes to address this issue by building a generalized additive regression model (GAM) between the nonlinear variables and using the residuals for the independence tests. He compares his results with traditional Granger causal analysis for time series (which expects linear relations between variables and between variable-lags) and shows significant improvements. This illustrates that the process of building the dataset is iterative and might require several searches over different lags, aggregation, and variable association strategies, etc.

The Search Procedure. The search procedure for the causal graph can be done in multiple ways. As described in in the background section, the search for a causal graph essentially involves a sequence of conditional independence tests to identify the structure of the DAG that represents the data.

In this report, we will focus our attention on two algorithms that have been widely applied to this type of search, i.e., the SGS and the PC algorithms. Both algorithms are described in [21] and pseudo-code is provided in Figure 49.

SGS Algorithm

A.) Form the complete undirected graph H on the vertex set **V**.
B.) For each pair of vertices A and B, if there exists a subset **S** of **V**\{A,B} such that A and B are d-separated given **S**, remove the edge between A and B from H.
C.) Let K be the undirected graph resulting from step B). For each triple of vertices A, B, and C such that the pair A and B and the pair B and C are each adjacent in K (written as A - B - C) but the pair A and C are not adjacent in K, orient A - B - C as $A \rightarrow B \leftarrow C$ if and only if there is no subset **S** of $\{B\} \cup$ **V**\{A,C} that d-separates A and C.
D.) repeat
If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as $B \rightarrow C$.
If there is a directed path from A to B, and an edge between A and B, then orient A - B as $A \rightarrow B$.
until no more edges can be oriented.

PC Algorithm

A.) Form the complete undirected graph C on the vertex set **V**.
B.)
$n = 0$.
repeat
repeat
select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**(C,X)\{Y} has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**(C,X)\{Y} of cardinality n, and if X and Y are d-separated given **S** delete edge X - Y from C and record **S** in **Sepset**(X,Y) and **Sepset**(Y,X);
until all ordered pairs of adjacent variables X and Y such that **Adjacencies**(C,X)\{Y} has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**(C,X)\{Y} of cardinality n have been tested for d-separation;
$n = n + 1$;
until for each ordered pair of adjacent vertices X, Y, **Adjacencies**(C,X)\{Y} is of cardinality less than n.
C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in **Sepset**(X,Z).
D.) repeat
If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as $B \rightarrow C$.
If there is a *directed* path from A to B, and an edge between A and B, then orient A - B as $A \rightarrow B$.
until no more edges can be oriented.

Figure 49. The SGS and PC Algorithms

In both algorithms, the search procedure starts by building a fully connected undirected graph including all variables. As independence relations are found between variables, edges are progressively removed. In both cases, the search involves two steps: a) reduce the fully connected undirected graph to a sparser undirected graph that contains only the probabilistic dependence relations between variables, and b) based on the d-separation tests, orient the edges to indicate causal relations between variables. At the end of the

search, the algorithms will yield a Markov Equivalent graph that represents the causal structure in the data[19].

The SGS algorithm is a naïve approach to the problem. It essentially performs an exhaustive search of conditional independence checks between all variable pairs, conditional to every possible subset of variables in the graph. Under ideal data conditions the algorithm is very reliable. Its main deficiencies are in the complexity (SGS grows exponentially with the number of variables) and the fact that, for noisy data, the higher order independence tests are likely to produce poor results, compromising the overall reliability of the algorithm. SGS work well, however, for datasets with only a few variables and a large number of samples. In addition to that it provides a very intuitive interpretation of the procedure.

The PC algorithm relies on an optimized search for the causal graph. It essentially removes edges from the graph at each step, which reduces the search space for higher order independence tests. The algorithm is shown to be theoretically unstable (as it depends on the order of the search) but in practice is reliable for causal discovery. It is proven [21] to yield a causal Markov Equivalent graph asymptotically[20]. The complexity of the algorithm is based on the number of vertices (n) and the maximal degree of any vertex (k). The complexity of the algorithm is shown [21] to be:

$$2\binom{n}{2}\sum_{i=0}^{k}\binom{n-1}{i}$$

There are, however, several heuristic optimizations that further reduce the complexity of the algorithm. The IG algorithm (also [21]) is one example. Other algorithms, proposed by different authors could also be applied, for comparison, over the same set. Dash, for instance, in 1999 [71] proposed a hybrid approach to the problem, using a variation of the PC algorithm and a Bayesian search, that could also be tested here for comparison.

In order to illustrate how a final causal graph would look like, we have created a simple example using actual traffic data from the MNDOT [35]. The example includes only a few metrics that were easy to extract from the raw dataset – not exactly the metrics we suggested for our causal analysis in this item. In this particular example, we have compiled only one metric per sensor (average occupancy/volume – which is a factor that indicates local congestion). For each sensor in a freeway segment (as illustrated in Figure 50), we have calculated the five-minute average of occupancy/volume (o/c) for the current instant and for a 15 minute lag. Note that we have two variables (clear and shaded) for each station. The clear measurement occurred at time t, and the shaded measurement at time t-15. For the model, we have also included background knowledge that prevented causal links back in time (from shaded to clear variables).

[19] Recall that a causal structure can be represented by a number of Markov Equivalent graphs. Under a number of assumptions, these algorithms will generate one of the Markov equivalent graphs.
[20] That is, considering an infinite sample size, even over noisy data.

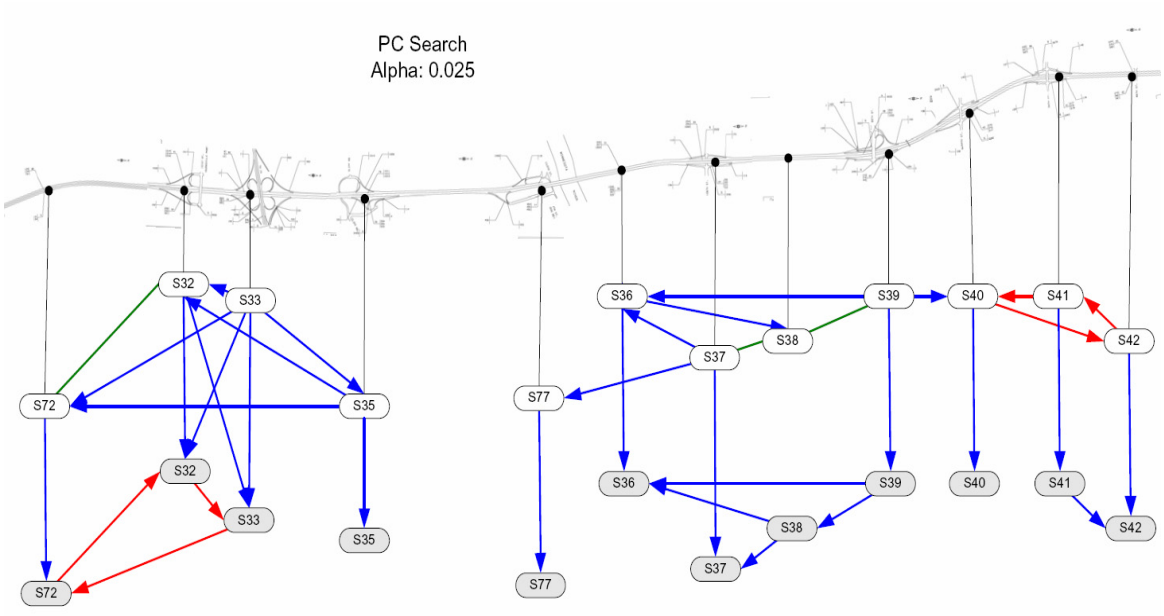


Figure 50. An Example of an Approximate Causal Network, Relating Metrics of Local Congestion in Different Points of a Freeway Segment. (The edges (if correct) should indicate that "congestion in point X causes congestion in point Y," for edges oriented from X to Y.)

For this very simple example, the dataset has less than 1000 data points, which is certainly a small dataset for the number of variables. This is evident in the resulting graph where two loops were found (marked in red) and three edges failed to be oriented (in green). Note, however, that even for this small toy example, we can identify a number of "causal" interactions close to the busy intersection at the left of the freeway, and a lot less interaction in other areas.

In this example, the probabilistic independence tests for the PC algorithm used a 97.5 percent confidence level (alfa=0.025). As shown in the graph, the algorithm failed to orient a few edges (for instance between S37-clear, S38-clear, and S39-clear). That was probably due to insufficient data or the violation of some of the underlying assumptions required by the PC algorithm. However, just by inspection, if we compare the same three variables in the subsequent time (shaded variables), the edges are oriented from S39-shaded to S38-shaded, to S37-shaded. This is (just by inspection) probably the same orientation of the corresponding clear variables, although there was not sufficient data available to verify this claim.

The same inference can be made between the pairs (S42-clear, S41-clear) and (S42--shaded, S41-shaded). The direction of the causal relations should be the same between the lags, so it is likely that the first orientation (the one causing the cycle) could be the orientation in error. Of course, these are just hypotheses that would require further investigation and a better prepared dataset for validation.

This example shows only causal interaction between one metric (o/c) at different location and times. It is just a toy scenario (even though using real data) that was not meant for analysis; it was included here just to show what the output of the algorithm will look like. The cycles at the left and right sides are errors that would require further verification of the source dataset.

To apply the same approach to our dataset, we would probably focus on a single station (where we have sufficient accident data available). Note that, for this example, each station has only two variables (O/C-current, ad O/C-deltaT). Based on our first proposal for the dataset, each station would have the 18 variables shown in Figure 48. A lot more data would be required, with a lot more care on data preparation and validation.

Expected Results and Limitations

Regardless of the specific algorithm used for the search, the resulting product will be a causal graph over all variables involved. From the graph, just by inspection, one should be able to infer the possible causes for the specific variable of interest, "traffic accidents."

It is very important to highlight though, that a number of assumptions about the data can't really hold in practice, so the search procedure is likely to produce something close to a causal graph but with a reasonable number of unexplained edges. These issues are much more related to the characteristics of the data than the algorithms themselves.

The causal interpretation of the graph must be carefully based on the underlying assumption and the imperfections of the data. It is likely that several searches, over different data arrangements will be necessary for better results.

Data Mining: "Not Now" Travel Time Prediction

Real-time travel time prediction is sometimes a function monitored by traffic control centers. It is best performed in an environment in which individual cars can be identified at different points along a route. Houston's traffic monitoring systems allow them to use toll tags to identify vehicles. Once a vehicle completes a route, dynamic message signs are used to display the travel time. Other means of identifying individuals include cell phones, GPS systems, and even imaging of license plates.

That technique concerns "now" travel time prediction. This segment of the report deals with predicting the travel time for next Wednesday or some other date in the future. A solution used, for example, in the CALTRANS systems, consists of collecting historic travel times, averaging them, then using displacements based on standard deviation, providing the client the maximum travel time with 80 percent confidence. But that is not the approach described here.

We propose a data mining solution under much more stringent conditions: There is no record of individual travel times. Instead, there is only the information contained in the star schema described in the design sections.

Background and Analysis

Differing from general purpose and persistent databases, data warehouses are usually designed with a well defined purpose (or purposes) in mind and they are generally optimized to better support these goals [32] [33] [34] [49].

The problem clearly entails the "purpose" for the warehouse, that is, the prediction of travel time. The assumption is that local information from traffic sensors (such as vehicle counts, occupancy, and speed - either measured, or estimated from the previous readings during data acquisition) is available, and the question is how this information can be organized and augmented, if necessary, to better estimate travel times.

We first present (item 2) a brief literature review on travel time estimation from sensor data. The goal of the review is to identify different approaches proposed for travel time prediction and isolate the variables of interest that should be in the DW. We then show a simplified example of sensor data (primarily ILD) in a star-schema and propose extensions to the model that would help improve travel time estimations.

Predicting travel time, a brief literature review

In this context, travel time is the time required to transverse a route between two points of interest. Travel time prediction is of grate importance in transportation systems. It provides drivers an intuitive measure about current (and estimated) traffic conditions for specific routes. Furthermore, it provides the basis for traffic control, infrastructure planning and dynamic route guidance, that is, the a-priori selection, or dynamic change of routes, based on variation of travel time estimates.

As classified by Turner in 1998 [37], there are essentially two broad classes of approaches for travel time prediction: a) link measurements and b) point measurements. Link measurement techniques are direct measurements that are, in general, more accurate but more complex and usually require specialized sensing capabilities. Point Measurement techniques, on the other hand, are indirect measurements that usually rely on local aggregate traffic estimates (such as average speed) for travel time prediction.

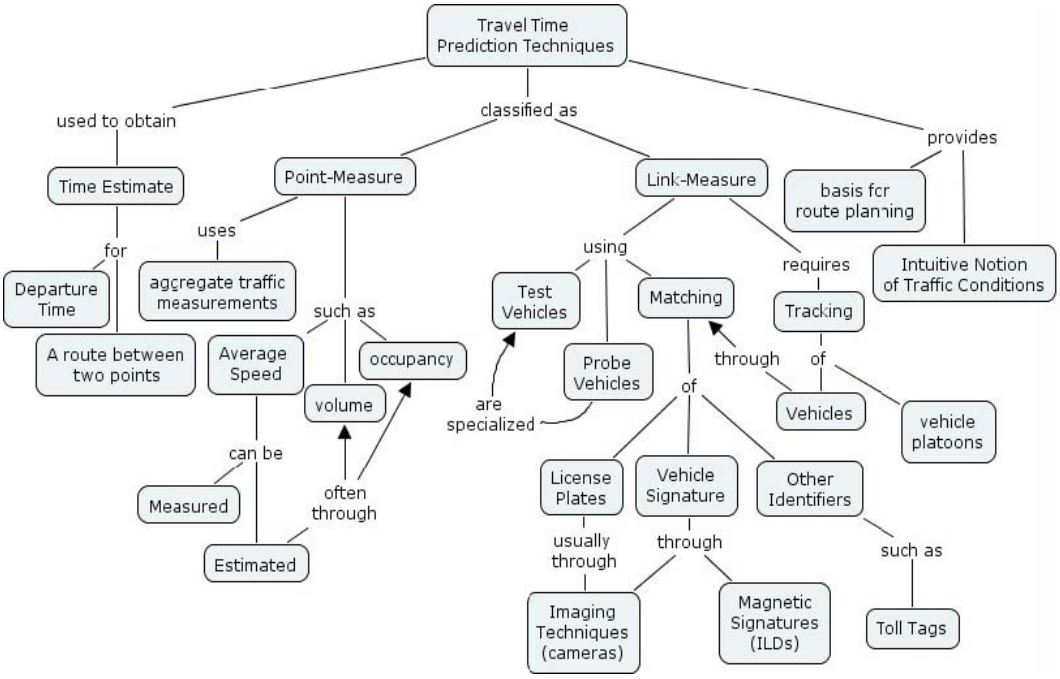


Figure 51. A Brief Overview on Common Travel Time Prediction Techniques.

Link measurements, in general, require the tracking of a specific vehicle or a platoon of vehicles (Figure 51). Such methods directly depend on vehicle re-identification and include, for instance, test vehicles, License Plate matching, vehicle (or platoon) signature matching, and Probe Vehicles.

Point measurements are based on local aggregate traffic metrics for the projection of the overall travel time. Such metrics include, for instance, occupancy, volume and local average speed. A number of approaches in the literature are based on the notion of vehicle tracking. In [7], the use of localization techniques in mobile cellular telephony is proposed for travel time estimation through vehicle tracking. In [38] the author presents experimental data obtained from approximately 1500 instrumented taxi vehicles[21] acting as probes for travel time prediction. In [39], the same data is used in conjunction with local sensor data to build an auto regression model for travel time prediction, which was essentially adjusted with the car-probe experimental results. Other tacking-based method techniques were also presented by [40] [41], who suggested different techniques for travel time estimation based on license plate matching techniques. Sun [50], in 1999 showed that it was possible to re-identify vehicles from their electro-magnetic signature on loop detectors, allowing (to some extent) the use of standard monitoring infrastructure for vehicle tracking.

An interesting variation on tracking-based methods was proposed by Coifman [42] [51], in 2002 and 2003. In his papers, Coifman proposes vehicle re-identification in multiple points in the freeway by augmenting inductive-loop detector (ILD) vehicle signatures with the relative position of the vehicle in its "platoon." Coifman's assumption was that driver's have a tendency to maintain their position within their driving platoon, and that

[21] Tests conducted in the Nagoya (Japan) metropolitan area.

information could be used in addition to the effective length signature estimated from loop-detectors for more accurate re-identification.

These were all link-measurement based approaches. In parallel to these efforts there were also a number of point-measurement techniques such as the ones proposed by [5, 6]. In [5], the author proposes the use of support vector regression to build a predictive model based on local sensor data. In [6], the authors build a linear regression model to estimate future travel time based on current "average-speed" estimation. Although these publications are relatively recent, point-measurement techniques are not a novel concept. In 1997, Petty [44] published travel time estimation results from single loop sensor (loop detector) data. In the same year, the Transportation Research Board published a literature review on the topic [43].

More recently, in 2003, Oh [45] compared conventional ILD methods based on estimated average speed for travel time prediction, with simulations and empirical data. Oh's work is of remarkable importance as he identifies an important limitation on point-based estimates for travel time prediction. In his paper, he shows that point-measurements are good estimators for travel time under free-flow but fail to accurately estimate travel time under congested conditions. He shows that the source of the problem is in the fact that point-based measurements make the assumption that traffic is homogeneous in the segment of the sensor. The assumption is close to true for free-flow traffic but it doesn't hold for congested traffic conditions, where an overall increase of the segment density and traffic "shock waves" are present.

Oh proposes the use of a density-based measured from occupancy and volume that, based on hydrodynamic kinematics traffic model (particularly the mass conservation principle) would improve the time estimation based on local average estimates. The technique yielded results with only five percent from simulated and empirical data, including congested traffic conditions. In [45], the author shows the details of the approach and the principles adopted for his technique.

Leveraging from Oh's approach, we describe, in item 3, a proposal for an augmented DW schema that pre-estimates density for sensor data, improving the efficient of queries that depend on traffic density estimation, such as Oh's approach[22].

An Augmented star-schema for Improved Travel Time Prediction

The start schema is possibly one of the most common schemas used for data warehouses[23]. The schema is based on a large central *facts table* containing the majority of the data and additional *dimension tables* related to the facts table though foreign keys. For traffic data, the facts table will include the sensor data (volume and occupancy, for instance, in the case of ILD sensors). The dimensions for this data are essentially time and space. That is, the location of the sensor and the time when the measurement was taken.

[22] Oh's approach was largely demonstrated on simulations, where traffic density information is readily available. The augmented schema proposed here will improve the efficiency of the density estimation directly from the database, using only sensor data.

[23] Other common schemas include, for instance the snow-flake schema and the fact constellation schema.

The dimensions can be hierarchically organized. For instance, the time dimension in this example can be grouped into "hour," "day," "week," etc. The space dimension can be grouped into "Station" (recalling that a station can contain multiple ILD sensors), "Road," "County," etc. The point in organizing the data this way is to optimize queries and data aggregations. The idea is to quickly and efficiently respond to queries such as, "show the hourly traffic of station 32."

Although, conceptually, we have only two dimensions (time and location), given the types of queries expected for time aggregation, it makes sense to sub-divide the time dimension into multiple tables. Technically, in the star-schema, these are separate dimensions, but in this case, the approach is appropriate as it reduces the size of the time dimension table and it facilitates comparative aggregation queries over different time abstractions. This has been suggested, for instance by Bhoite [46], to better handle queries such as: "compare the traffic of station 32 between rush-hour in weekdays and weekends." One simple version for the schema is illustrated in Figure 52.

In this example, the data granularity in the warehouse is assumed to be "hourly averages." Ideally, the DW would maintain a much finer granularity (in the order of minutes, or seconds) which would require additional dimension tables to be added to Figure 52. In practice, however, most Data Warehouses use a variable granularity level for historic data. That is, they maintain, for instance, 30-second aggregation for the most recent month, five-minutes aggregation up to one year and, after that maybe one-hour aggregation. This is a common practice to ensure a reasonable size for the stored data.

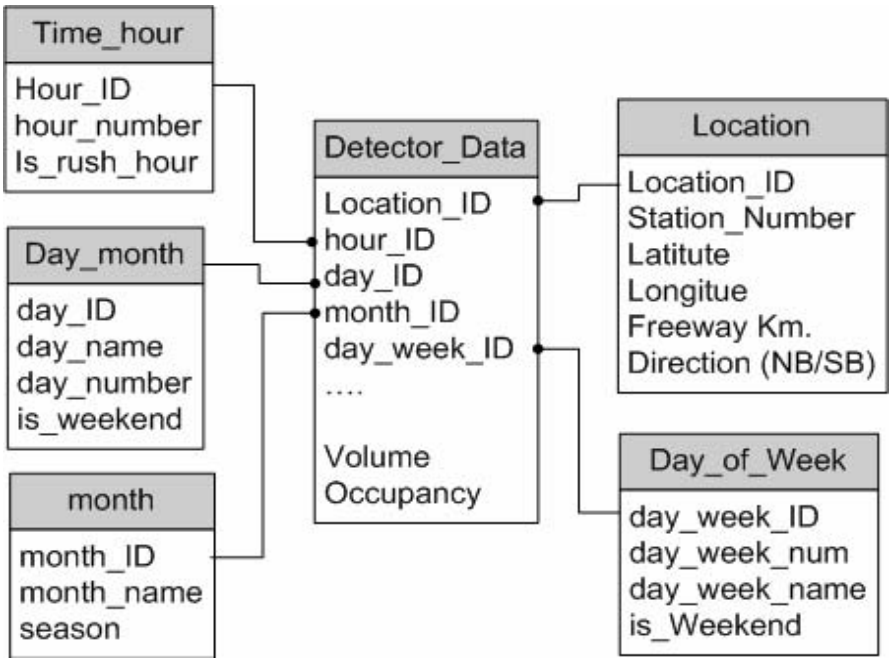


Figure 52. Simple Star Schema for ILD Sensor Data. (Note that only some of the time dimensions are shown. Other time dimensions are omitted for simplicity. Time dimensions are subdivided here for efficiency. Technically, aggregations in the "hour" dimension will be equivalent to a singe value in the "day-month" dimension table.)

Also in this example, because of the choice of a star-schema, there will be some data-redundancy in some of the tables. For instance, in the Location table, there will be multiple entries for a single station, given that a station has multiple detectors. This duplication, however, is usually negligible given the relative size of the facts table in comparison with the dimensions table.

Adding a New Dimension. The star-schema is designed to improve queries[24] to the database. From that perspective, and leveraging from Oh's research, a new dimension can be added to include information about average segment density as part of the sensor data entry. The justification for this is the following: The estimation of traffic density from local sensor data is a direct function of volume and occupancy. As a deterministic function, this new "attribute" provides no additional information that can be used to augment local estimates based in the same sensor data. Oh's approach relied on average density information around the sensor (obtained from sources other than the sensor) to improve local estimates. In Oh's case, the sources for this additional piece of information (average density) could be trivially identified because the validation experiments relied on simulated data (where this information is readily available), or sensor data where freeway segments were defined and validated a priori.

In practice, however, in order to infer the average time density in a freeway segment from information other than the sensor itself and traffic simulation, one must rely on neighboring sensor data. Under certain constraints, a number of surrounding measurement points can be used to identify a "closed" section of the freeway where, based on the principle of mass conservation, variations in density can be estimated with reasonable accuracy. Consider, for instance, the illustration shown in Figure 53.

[24] In practice, Data Warehouses have additional mechanisms (besides the storage schema) to improve queries. Data-cubes, for instance, provide pre-aggregation of the data simultaneously in several dimensions to improve query efficiency. Details on data-cube models for traffic data are provided by Shekhar [43] in CubeView

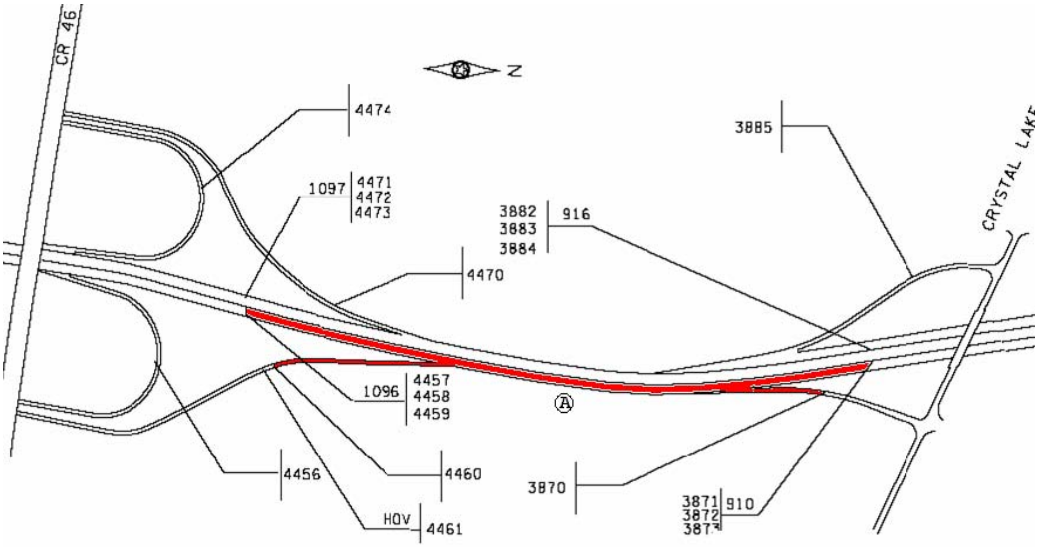


Figure 53. A Closed Freeway Segment as Defined in the Proposed Approach

In this example, the average density variation estimates in the marked area can be calculated based on the flow (vehicles/sec) entering the segment and the flow leaving the second (mass conservation). Assuming no sensor errors, the differences in flow can be directly attributed to changes in density. In this example, the marked area constitutes what we call a "closed segment" of the freeway. That is, a segment that has all its entry and exit points monitored by sensors.

Consider now that we're estimating travel time in a point within the segment above; let's say in point "A" (assuming we had sensors there). The density information obtained by the boundary sensors, as illustrated above, could be very useful to Oh's. In fact, granted the appropriate difference in estimation error, the densities calculated this way would provide the same type of information that Oh obtained via simulations in his paper.

From that perspective, if such density for the closed segment can be calculated, every point-measurement based on "A," in this example, can be augmented with density information. The problem, however, is that the determination of the boundaries of the freeway segment is not as trivial as it seems.

The issue is that, due to variations in traffic conditions, the volume count in some of the boundary sensors can vary greatly. It is possible, through a relatively simple search procedure, to identify a "close segment" around a point "A" that is consistent. That is, a segment that, for a larger time interval (more than a day) will average zero accumulation in traffic count. For instance, if a closed segment is chosen around node "A," we should verify that for a period of one day (or more) the number of vehicle that enter the segment should be close to the number of vehicles leaving the segment, resulting in a net-accumulation approximately equal to zero. This is because, through the course of a day

we expect the density of in the segment to change, but assuming a cyclical behavior, the net-flow should asymptotically approach zero[25].

Now, in practice, if a segment is chosen for monitoring and it accumulates a positive (or negative) number of vehicles through the period of a day or more, then it is reasonable to assume that some sensors, either the ones measuring incoming flow, or the ones measuring outgoing flow (or some of both), are faulty. In fact, there are traffic conditions that might occur during rush-hour that can lead to these types of volume count errors, even if the sensors are not faulty.

The argument is that, through a relatively simple search procedure, multiple closed segments around point "A" can be validated using historical data to identify times of the day when their readings were trusted (net-flow close to zero) or not. There are many reasons whey some segments can be found un-trusted and that might vary arbitrarily through the day based on traffic conditions or sensor faults.

Our proposal is to extend the DW schema to include pre-calculated estimates of segment traffic density for each measurement point. The search for closed segments can occur offline, as opposed to occurring during query time, so density estimates are available a-priori, from the best possible set of neighbor sensors.

The process then, would work as follows: A new dimension called segment density will be added to the database. This dimension will specify bins that constitute different levels of traffic density. As new data is added to the facts table, it is directly mapped to the appropriate segment density bin, based on current density estimates of for that specific segment. Recall that, in order to make such estimates, a search over the best possible boundaries for the segment must be conducted. That search will be done a-priori for each sensor and will be verified periodically as traffic conditions change.

The segment density information now attached to each sensor entry maintains a readily available reasonable estimate of the average segment-density for that sensor at the time of the data. The information can be directly used by algorithms that rely on density correction to improve accuracy.

Figure 54 shows a simple illustration of an augmented star-schema containing the segment density dimension.

[25] Note that it might never be exactly zero, as this would imply knowledge about the initial conditions when density measurements (out – in) started. The initial condition, although unknown, should be irrelevant for if traffic is observed for longer periods of time.

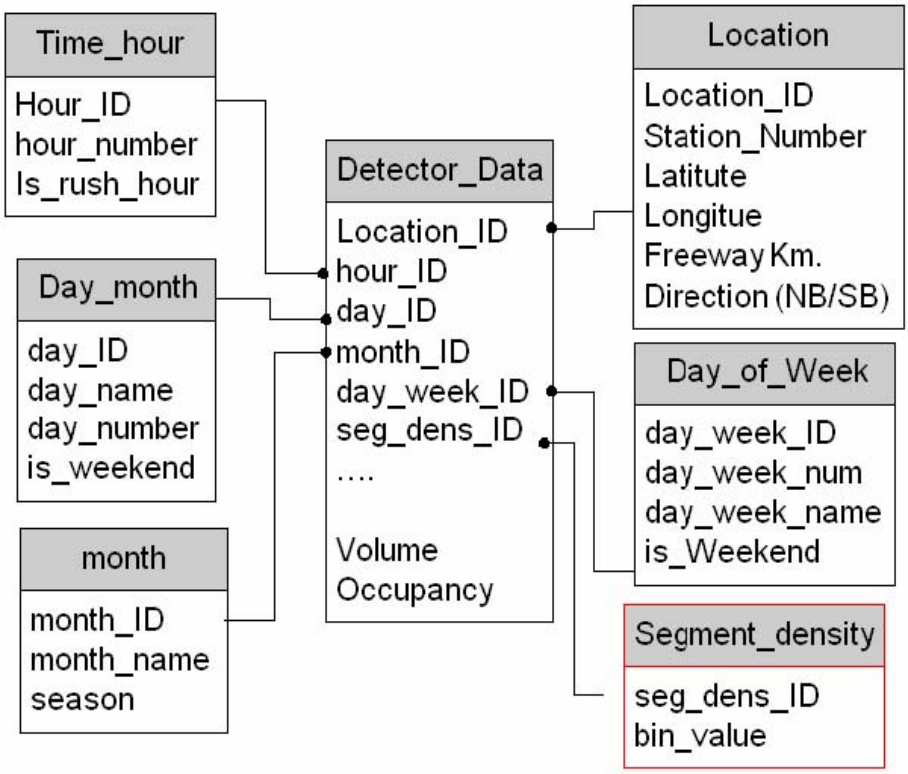


Figure 54. Augmented Star Schema Including Three Segment Dimensions

It is important to note that a segment in this context is not necessarily composed by the minimum "closed volume" (i.e., fully monitored section) of the freeway, around the sensor. Due to the reasons previously stated, the boundaries of a closed segment might change in time, but the density estimates are parameterized and always based on the best set of measurements available at that time.