ORNL/TM-2001/247

Proof of Concept of ITS as An Alternative Data Resource: A Demonstration Project of Florida and New York Data

Final Report

September 31, 2001

Prepared for
Federal Highway Administration
U.S. Department of Transportation
Washington, DC 20590

P. Hu
R. Goeltz
R. Schmoyer
Center for Transportation Analysis
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831-6073
managed by
UT-Battelle, LIC
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The authors wish to thank the following individuals for their inspirational vision, constructive suggestions, and careful review of this research:

- Harshad Desai, Lap Hoang, Rick Reel, and Trey Tillander of Florida State Department of Transportation,
- Rick Zabinski, Emilio Sosa, and Ron Tweedie (retired) of New York State Department of Transportation, and
- Tony Esteve, Frank Jarema (retired), and David McElhaney (retired) of Federal Highway Administration.

Finally, the authors express their gratitude to Ralph Gillmann of Federal Highway Administration for his guidance in spearheading this research.

SECTION 1. BACKGROUND

Congestion, inefficiency, traffic accidents, and pollution are some of the effects that partially offset the benefits of today's transportation infrastructure. It is estimated that traffic congestion costs the American people about $100 billion each year in the form of lost productivity. Also, in 2000, highway traffic accidents claimed almost 42 thousand lives and injured an additional 3.2 million people. Vehicle emissions are a major cause of air pollution. Trucks, buses, and automobiles idling in traffic emit tons of pollutants each year and waste billions of gallons of fuel.

For many years, these problems have been solved by merely building more highways. However, a strategy that relies solely on building more roads to increase capacity, without addressing the underlying problems of our transportation system, is no longer adequate and acceptable. In response, U. S. Congress passed the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA). ISTEA calls for the creation of an economically efficient

and environmentally sound transportation system that will move people and goods in an energy efficient manner, and that will provide the foundation for a competitive American transportation industry.

A broad range of information processing, communications, control, and electronics technologies, known collectively as Intelligent Transportation Systems (ITS), offers an innovative solution to many of our transportation problems, and is one of the tools that are expected to help meet the ISTEA's goal. In essence, ITS provides an intelligent link between travelers, vehicles, and infrastructure. Applying these technologies to our transportation system is anticipated to save lives, save time, and save money. ITS' goal ". . . is to apply modern computer and communications technologies in our transportation systems, resulting in improved mobility, safety, air quality, and productivity."

In order to provide safer, more environmentally friendly and more efficient transportation systems, one of the ITS technologies is to assist drivers to reach a desired destination with navigation systems enhanced with pathfinding or route guidance. Another is to collect and transmit real-time information on traffic conditions for travelers before and during their trips. For this intelligent link between travelers, vehicles, and infrastructure to be truly effective, it will need to be supported in part by an enormous amount of real-time data.

Providing real-time information on traffic conditions requires collecting data on traffic volume, vehicle speeds, lane occupancy, congestion, and incidents. These data are collected at substantial expense and in an extremely short time increment (e.g., 30 seconds). Lack of appropriate means of storing this immense volume of data is one of the reasons that forces

this wealth of valuable information to be "recycled" (or "purged") before it can be utilized to help fulfill transportation planning purposes.

On the other hand, considerable resources (both in time and money) are committed to collecting traffic volume, vehicle classification and other data to meet transportation planning needs and other reporting requirements. In fact, traffic data collection is often one of the largest expenditure items in most annual planning budgets. However, due to equipment failure and a host of other factors, the reliability and representativeness of traffic data are sometimes in question. This proposal is motivated by the desire to improve the quality of traffic data in a cost-effective way. Specifically, the proposal suggests that traffic data requirements be fulfilled with data currently (or soon to be) collected by ITS systems. Unfortunately, administrative or technical barriers have thus far hindered utilizing ITS data to fulfill these other transportation data needs. If linkages can be established for ITS data to be shared with other needs, then scarce resources that are obligated for collecting the same data twice will be reduced, and concerns about data quality will be lessened. In addition, ITS data can improve the timeliness and frequency of traffic data.

This type of proposal, which is to develop ways of using ITS data, will lead to win-win situations. It will not only benefit the transportation planning community by allowing it to access freely more and better data, but will also enhance the appeal of ITS deployment by significantly broadening its originally intended benefits. The notion of using ITS data as an alternative data resource is reflected in the Archived Data User Services (ADUS)[1].

[1] http://www.itsa.org/resources.nsf/urls/adusr.html

SECTION 2. OBJECTIVES

The use of ITS-generated data as a data resource is a multifaceted challenge. The most effective way of confronting this challenge is to focus early efforts on localized areas with well-defined parameters. With the idea of starting with a well-defined problem, this research demonstrates the feasibility of using ITS-generated data to meet traffic information needs. Specifically, this study focused on two crucial traffic parameters: (1) **total traffic volume**, and (2) **total VMT** — basically, the information collected from the Traffic Monitoring Program. Traffic data collected from Florida and New York ITS deployments were used to test the communications and estimation procedures.

This challenge has three main components:

1. **Data archiving**. How are ITS-generated data transmitted from the field? How should the data be archived? What are the communications and other technological barriers?

2. **Cost of using ITS-generated data**. The cost will be measured in terms of effort needed to re-format the data, re-vamp the software, and address data quality issues. For example, once data are archived, are data readily "usable" with respect to data format and data quality? What efforts are needed to incorporate ITS-generated data into the mainstream, traditional data? One example of such effort might be the revamping of existing software. In addition, what are the institutional and logistical barriers?

3. **Benefits of using ITS-generated data.** After all the "costs" of using ITS-generated data are addressed, what is the value-added which the ITS-generated data contribute? What is the cost-benefit ratio for using ITS-generated data? Questions specific to this project are the following: Are ITS data able to provide or estimate needed information? How close are the ITS-based estimates to those based on traditional data collection efforts?

Data used in this research are described in Section 3, followed by a discussion in Section 4 on data and data quality issues in the original data. Statistical procedures to identify and correct unacceptable data are reported in Section 5. Aggregating 5-minute or 30-second traffic counts to hourly counts is described in Section 6. Section 7 discusses procedures to impute missing data. The benefits of using ITS-generated data are summarized in Section 8. An ADUS prototype for the sensor data is reported in Section 9. Barriers to archiving and using ITS-generated data, and recommendations to overcome these barriers, conclude this report in Section 10.

SECTION 3. DESCRIPTION OF DEMONSTRATION SITES AND DATA

This section describes the ITS-generated data and the facilities from which the data are generated. The description focuses on data resources and information technology.

3.1 INFORM in Long Island

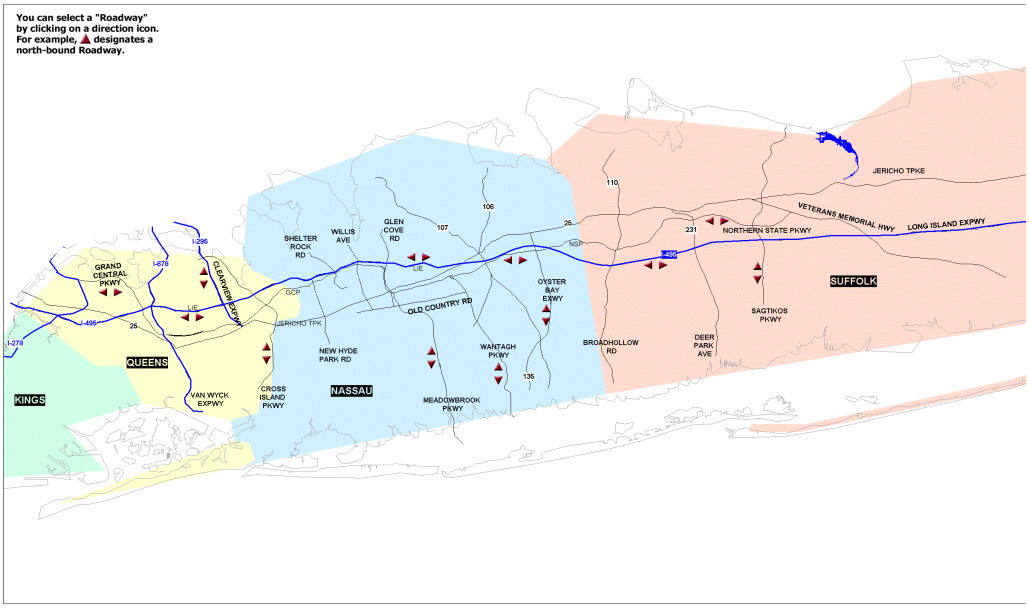
INFORM (INformation FOR Motorists)[2] is one of the largest traffic information systems in the nation. It covers Long Island's 45-mile central corridor, including the main east-west highways and the busiest north-south connecting roads (Figure 3.1). At the time of the study, there were twelve continuous count stations installed in Long Island. However, only one of these counters (Station 798) is in the proximity of the INFORM. Thus, data from this counter were used for comparison purposes.

The Traffic Management Center (TMC) is located at New York State Department of Transportation's Long Island Regional Headquarters in Hauppauge, New York. The major information source for INFORM comes from some 2,800 electronic sensors embedded in the roadways at half-mile intervals. When a vehicle moves over a sensor it sends a signal to the TMC. Based on this information, computers continuously calculate the volume and speed of traffic on different sectors of the highways. Specially, system components include:

- 2,800 loops at 0.5-mile intervals,
- 125 traffic message signs, and traffic light controllers at 170 intersections,
- 80 video cameras, and

[2] http://www.dot.state.ny.us/reg/r10/inform.html.

Figure 3.1 INformation FOR Motorists in Long Island

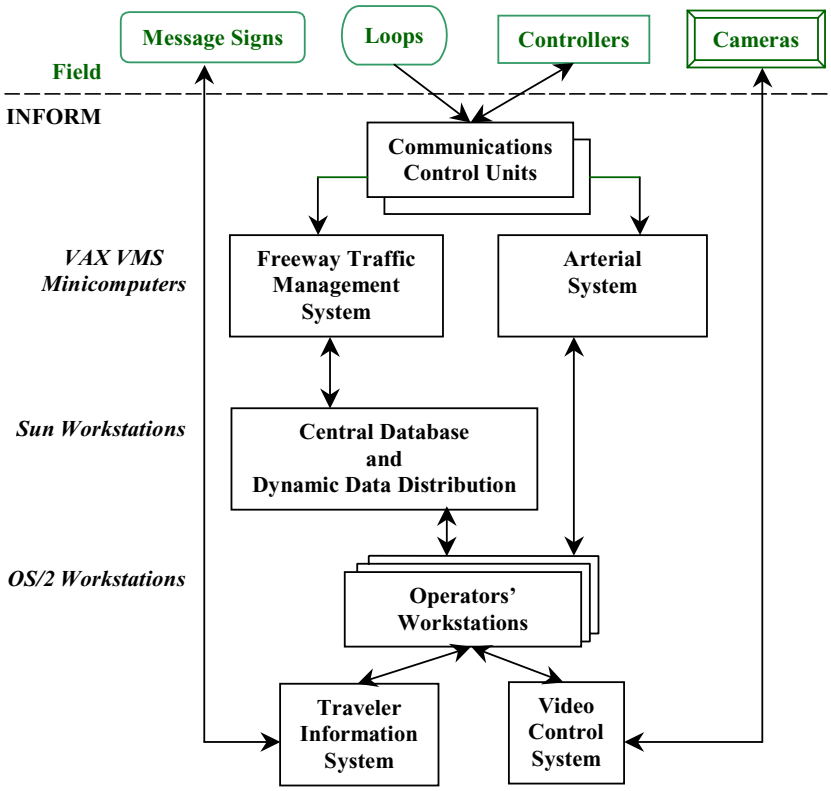


- about 75 merge lights at entrance ramps.

TMC computers display traffic conditions throughout the system on three 8-foot rear-projection liquid crystal display (LCD) monitors. Traffic Operations Coordinators can immediately spot delays and maintain traffic message signs. Furthermore, coordinators provide advice to police, emergency, maintenance, and news organizations. Major upgrades were planned, at the time of our study, to instrument additional highway segments (e.g., Southern State Parkway) and replace coax with fiber optic cable.

A data-centric view of the INFORM information system contains the following data elements: loops, video, messages, controls, incident logs, communications, weather, and equipment status (Figure 3.2).

Figure 3.2 Information Components of the New York INFORM



Loop data are transmitted to the TMC via coax and fiber optic cable. New construction and upgrades use fiber, which improves system reliability and data quality. There is considerable new construction and upgrade activity; therefore the information on the loop configuration (e.g., location, identification, type of the loop detector) is constantly in flux.

Loop sensor outputs are gathered in the Communications Control Unit (CCU) and presented to the two VAX minicomputers, one for the Freeway Traffic Management System (i.e., the mainline) and one for the arterials. Two Sun workstations provide database

management for the operators' OS/2 workstations. Loops can be identified by service type, including:

- Mainline;
- Speed trap (i.e., a dual loop);
- Ramp, including demand, queue, and passage;
- Exit;
- HOV mainline and trap; and
- Doubly defined (i.e., a connector exits one roadway and enters another).

All real-time applications in the TMC use 1-minute data. In the field, loops are "polled" 60 times per second by roadside controllers (Calstart models 170e and 2070). The field accumulator performs 15 "heart beats" per second. These data are read from the CCU by the VAX, which creates 1-minute interval data. *Volume*, *lane occupancy*, and *speed* (VOS) are calculated for each loop detector.

Traffic signal data include failure logs and changes to the control plans (e.g., phasing and timing). These records are kept 6 days and then deleted.

Variable Message Sign (VMS) messages are archived, including text, sign ID, and time of the day. Messages can be generated automatically by an operator. Some messages are blank, though public service messages are usually displayed rather than blank signs. The operators can verify that a message was successfully transmitted and received in the field, but they cannot verify that the correct message is displayed. Logic checks are set up at the remote sites to "void" invalid signals (e.g., loss of communications).

An incident log is maintained but not archived. The information that is recorded in the incident logs is very complex. The large number of overlapping enforcement jurisdictions (e.g., police departments, fire departments, emergency response agencies) makes it difficult to identify the owner of the incident data.

Data Archiving

The INFORM archive of VOS is created on the main VAX computer using a program that continuously collects VOS using a 1-minute timer. The 1-minute data on VOS are aggregated into five-minute intervals and are written/archived to a local file system — one record every five minutes. At midnight of each day, the daily archive file is closed, and some summary calculations are performed. At this point, a new daily archive file is initialized. In sum, three files are created for each day:

(1) a binary file that contains VOS for all loops,
(2) a similar binary file for the speed traps (i.e., dual loops), and
(3) an ASCII file that contains loop identification data with pointers to the binary data (henceforward referred to as the CONFIG file).

Not all of the data from all loops are archived. Currently, data from about 2,000 loops are archived, amounting to about 3.3 MB per day (in an uncompressed format). These daily files are organized into monthly folders with data for recent months kept online and data for not-so-recent months written to tape. The binary files are extremely efficient, using 8-bit integers to store *speed* and *lane occupancy* data and 16-bit integers for traffic *volume*. All fields are "packed" together and the record begins with a 16-bit time-of-day field. Although an efficient and compact way to store data, the 16-bit protocol can perceivably present

technical challenges to users. This is because the 16-bit binary data might be recorded differently on different computers (e.g., little-endian vs big-endian, which indicates which of the 2 bytes comes first).

Description of Data

This feasibility study used INFORM data from February to August, 1999. An example of raw five-minute counts is in Table 3.1. Note that the four-minute partial counts at 80 and 100 minutes. Although all of the 2,800 sensors were operational, only data from about 2,000 sensors were archived. Since the objective of this feasibility study was to determine the utility of ITS-generated data to meet traffic monitoring information needs, data for each lane were summarized to directional-data.

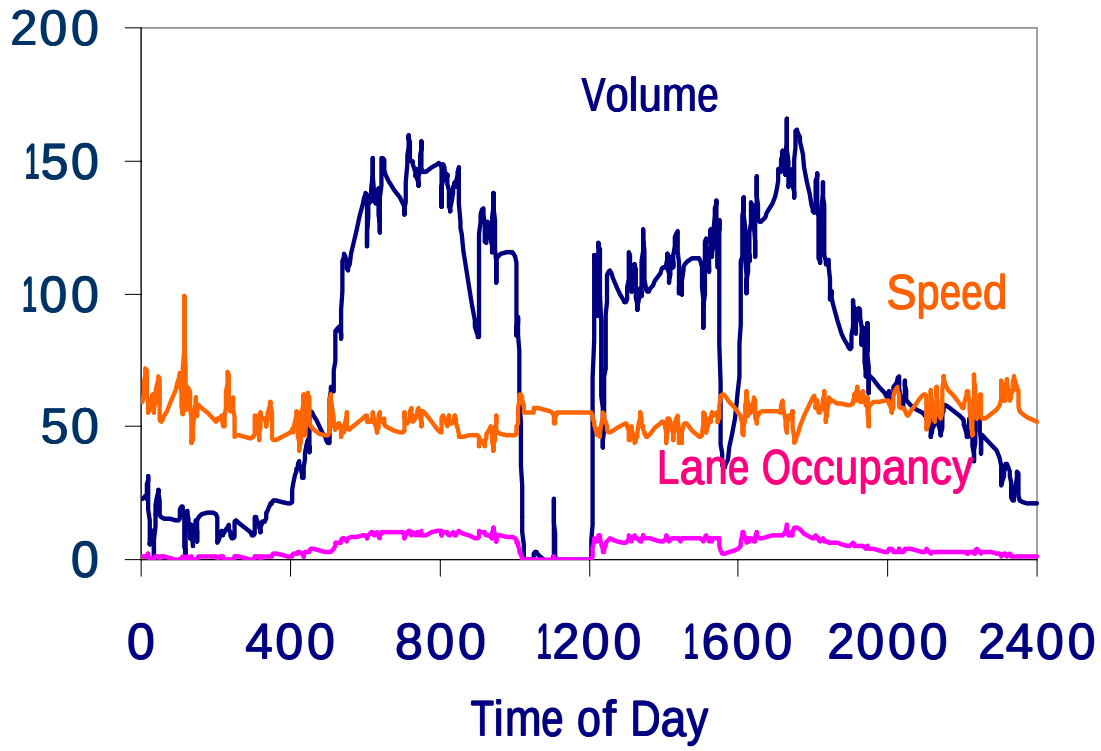
Raw data were not without data quality challenges. The lack of documentation made the deciphering of the raw data demanding. An additional challenge were changes in sensor identification numbers (sensor ID) in the INFORM data. For example, Sensor No. 123 had been continually labeled "123." On March 11, at 11:48 am, Sensor 123 was "re-sequenced" to be Sensor 124. This type of change typically occurred when a sensor was added to the system, all of the "down-stream" sensors were automatically re-numbered. Although this type of change was reported in the CONFIG file, the CONFIG file was not integrated with the binary data file. The implication of not having this type of information integrated with the traffic data is that it is almost impossible to develop a traffic profile over time. To rectify this problem, information documented in the CONFIG file needs to be integrated with the data file so that the data user can easily recognize the ID changes.

Table 3.1 Example of New York INFORM Five-Minute Counts (From a site on the Northern State Parkway (Zone 321) February 1, 1999

Time Stamp in Minutes	Counts in Lane #1	Number of Minutes Counted in Lane #1	Counts in Lane #2	Number of Minutes Counted in Lane #2
5	16	5	23	5
10	21	5	27	5
15	23	5	29	5
20	18	5	29	5
25	8	5	17	5
30	11	5	22	5
35	18	5	22	5
40	14	5	22	5
45	17	5	22	5
50	10	5	10	5
55	11	5	13	5
60	13	5	10	5
65	11	5	13	5
70	8	5	11	5
75	5	5	13	5
80	8	4	6	4
85	6	5	11	5
90	7	5	12	5
95	2	5	8	5
100	9	4	8	4

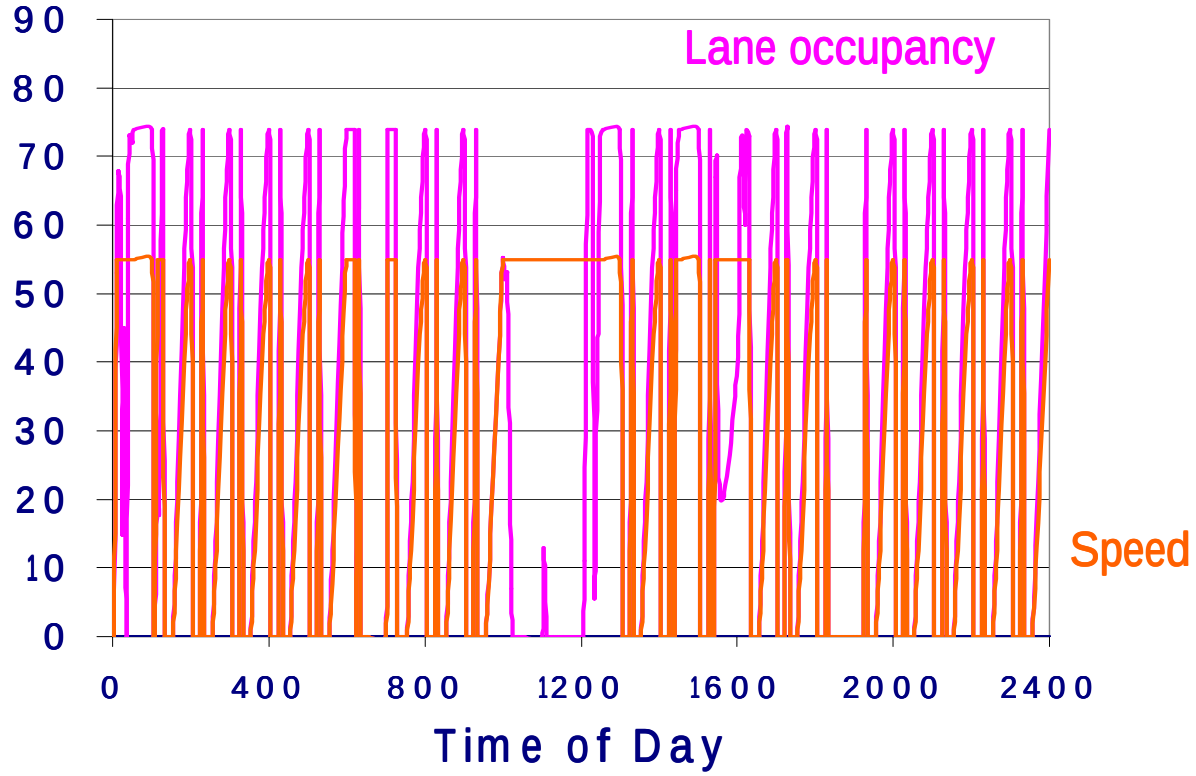
Simply having ITS-generated data recorded and archived by no means implies that data are "ready-to-use." Figure 3.3 illustrates an example of acceptable data. Even largely acceptable data, as depicted in Figure 3.3, are not completely free from error. On the other hand, Figure 3.4 presents archived but totally unacceptable data.

Figure 3.3 An Example of Acceptable Loop Detector Data



Missing data are common in traffic data, and frequently are inappropriately coded as "0." Coding missing data as "0" presents another difficulty. It is likely that some less traveled roads have no traffic at certain times of the day, for example at three o'clock in the morning. Therefore, a traffic count of "0" is quite legitimate. To distinguish legitimate 0's from missing data that are coded as "0" adds a new dimension to the data quality problem.

Figure 3.4 An Example of Unacceptable Loop Detector Data



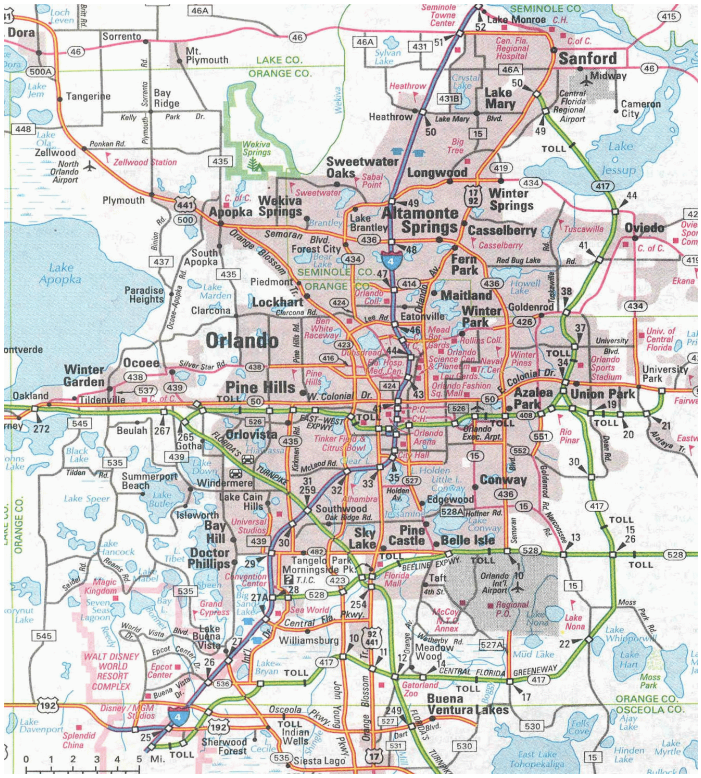
3.2. SMIS in Orlando

To meet the transportation management needs of the I-4 corridor of the Metro Orlando area, the Florida State Department of Transportation implemented the Surveillance and

Motorist Information System (SMIS)[3] (Figure 3.5). The SMIS information system instruments 39 miles of highways. Its components include the following:

- 845 loops (covering a 38-mile section of I-4),
- 69 170-type controllers (stations),
- 50 CCTV color cameras at approximately one-mile intervals with pan-tilt-zoom and focus controls,

Figure 3.5 Surveillance and Motorist Information System (SMIS) in Orlando



[3] http://www.dot.state.fl.us/planning/systems/sm/its/its.html

- 24 changeable message signs,
- 39 miles of single and multimode fiber optic cable, and
- 5 weather stations.

There were two continuous-count sites (Sites 130 and 303) installed in the proximity of the 39 SMIS miles. Data from these sites were used in the subsequent analyses that compare ITS-generated data to those collected from traditional counters.

The Freeway Management Center (FMC) is co-located with the Florida Highway Patrol in DeLand, Florida (within FDOT's District 5). FMC operations include monitoring, incident detection, sign control, camera control, and emergency response. A data-centric view of the FMC contains the following data elements: dual loops, video, messages, incident logs, weather, and equipment status. The central system consists of three main computers: (1) the SQL Server database, (2) a communications server (CommServ), and (3) an application called MIST which performs incident detection and reporting. This is a second generation system — a network of PCs has replaced the original VAX.

Data Collection and Archiving

All loop locations use dual loops spaced about 15 feet apart. The 170 controllers are "polled" every 30 seconds, but the controllers do not perform data quality checks. Information on traffic volume, speed and lane occupancy is computed and transmitted to the FMC over fiber. The 170 controllers collect information useful for vehicle classification, including vehicle length. Unfortunately, these data are not transmitted to the FMC. Thus, they are not archived. The Transportation Systems Institute of the University of Central

Florida (UCF) has been archiving SMIS loop detector data since September 1997 as part of ongoing research projects.

The data are acquired in real-time using a phone connection between a PC at UCF and one at the FMC, called a Dial-up Server. Within the FMC, the Dial-up Server has an Ethernet connection to the data source, CommServ. When the FMC receives electronic requests for archived data, CommServ generates a file containing the protocols for file transfer (typically referred to as an FTP) which is used to transmit data. A file is transferred every 30 seconds. The data file is stored as ASCII on UCF's client PC and contains data on traffic volume, speed, and lane occupancy for all of the loops.

UCF uses Microsoft Access software to manage the archive. Daily loop detector data are stored in one database while the monthly data are archived in another. The names of the table column indicate the direction, the lane, and the measurement (i.e., volume, lane occupancy, or speed). Also, there are columns for time-of-day and station ID. Therefore, there is one record for each time interval and each station.

The FMC has a 35' x 8' video wall with a variety of configurations and controls. Video data are managed and transmitted by the same hubs as the loop detector data. The FMC has the ability to save selected video images to VCRs. However, these images are not saved.

The texts of standard message sign messages come from a database and are indexed by message identification number (ID). The operator has the ability to edit the message before sending it out. The edited messages are not saved.

Incident data are recorded and saved. The current archive covers data for more than 2 years. Information on weather conditions (i.e., wet vs. dry) is collected at the 5 hubs, and transmitted in the VOS data stream. Weather data are archived. A LOTUS application is used to document equipment status, including the quality of the video. This "log" is kept for a short time and disposed.

Description of Data Used

Florida SMIS data for April through October 1998 were compiled and provided by the University of Central Florida[4]. These thirty-second traffic data were collected from sixty-nine "sites" along I-4 in or near Orlando. Each site has either two or three lanes in each direction. Although some of the sites have ramps, ramp data were not received. Some of the data for some of the months were incomplete in terms of missing data. For example, about 70% of the data were missing on thirty-second traffic counts collected at Station 16W (east bound on the intersection of I-4 and Highway 528) on an August day.

Although SMIS loop detector data are all thirty-second counts, they are not sent synchronously[5]. In fact, the "time stamps" tend to drift. For example, rather than 30 seconds apart, the time stamps shift from 0:00:10 to 0:00:41 to 0:01:13 to 0:01:45, etc. The implication of this time-stamp shift is irrelevant to traffic management and operations. However, the implications for using ITS-generated data for other purposes might be significant. For example, to determine the relevance of ITS-generated data in meeting traffic

[4] Data were kindly provided by Sherif Ishak, University of Central Florida.

[5] "Controllers are not synchronized to send data at the same time but they all do in 30 seconds."—Sherif Ishak, University of Central Florida, personal communication.

monitoring needs, one approach is to compare traffic count data collected from traditional traffic count programs to those generated by ITS deployments. The standard data protocol of traditional traffic count programs is on a hourly basis. Therefore, the 30-second traffic count data generated by ITS need to be aggregated to an hourly interval. The shift in time-stamp complicates this task.

SECTION 4. DATA AND DATA QUALITY ISSUES

To determine the feasibility of using ITS-generated traffic data to meet traffic monitoring information needs, data and data quality issues must be overcome. One of the major barriers to integrating ITS-generated traffic counts with traditional traffic counts is that the former are aggregated into a different time scheme than that used by the traditional traffic counts program.

The traditional traffic count program consists of two major components: continuous counters and coverage counters (short-term counters). Continuous counters collect traffic data on a continuous basis — 24 hours a day and 356 days a year. They are expensive to operate, and thus there are few of them. By providing information on seasonal variability and on the variability by time-of-day and day-of-week, data collected from these counters are used to adjust data collected from short-term counters that collect data for 48 or 72 hours. As previously mentioned, continuous counters are expensive. A very limited number are installed. This limitation results in adjustment factors that are not as representative of the road network as one would desire. Because of the limited installation of continuous counters, ITS loop detectors, which perform functions similar to continuous counters, can serve two purposes. First, they can provide additional traffic data to improve the quality of the adjustment factors. Second, they can replace traditional counters installed in nearby roads.

In order to address this goal, the first task is to determine whether ITS counts are equivalent to traditional counts by comparing count data from the two different sources. Whereas traditional traffic counts are recorded on an hourly basis, ITS traffic counts are

recorded more frequently (typically, every 30 seconds, 1 minute, or 5 minutes). The logical next step then is to aggregate ITS data into hourly counts.

If ITS traffic counts consisted simply of a continuous stream of counts, with no counts missing and with no counts outside of a reasonable range or otherwise spurious, then it would be straightforward to aggregate them into hourly totals. Unfortunately, the ITS and other traffic volume data are rarely flawless. Data anomalies embedded in the ITS data make data aggregation a non-trivial task. Before reliable estimates can be computed and before hourly counts can be aggregated, questionable data must be identified and corrected.

Because data from New York INFORM and Florida SMIS are recorded at different time intervals, 5-minute and 30-second, respectively, the procedures developed for the two states are different. These procedures identify questionable data and aggregate the data into hourly counts. The rationale for developing different procedures is that more data quality checks are possible for thirty-second data than for 5-minute data. Figure 4.1 presents steps used to "prepare" the 5-minute traffic counts from New York INFORM. The more complex steps to prepare the 30-second Florida counts are in Figure 4.2.

Figure 4.1
Data Flow Schematic for New York INFORM Five-Minute Traffic Volume Counts from Long Island.

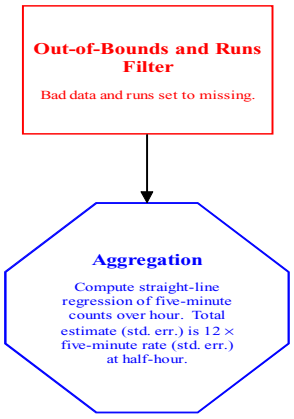
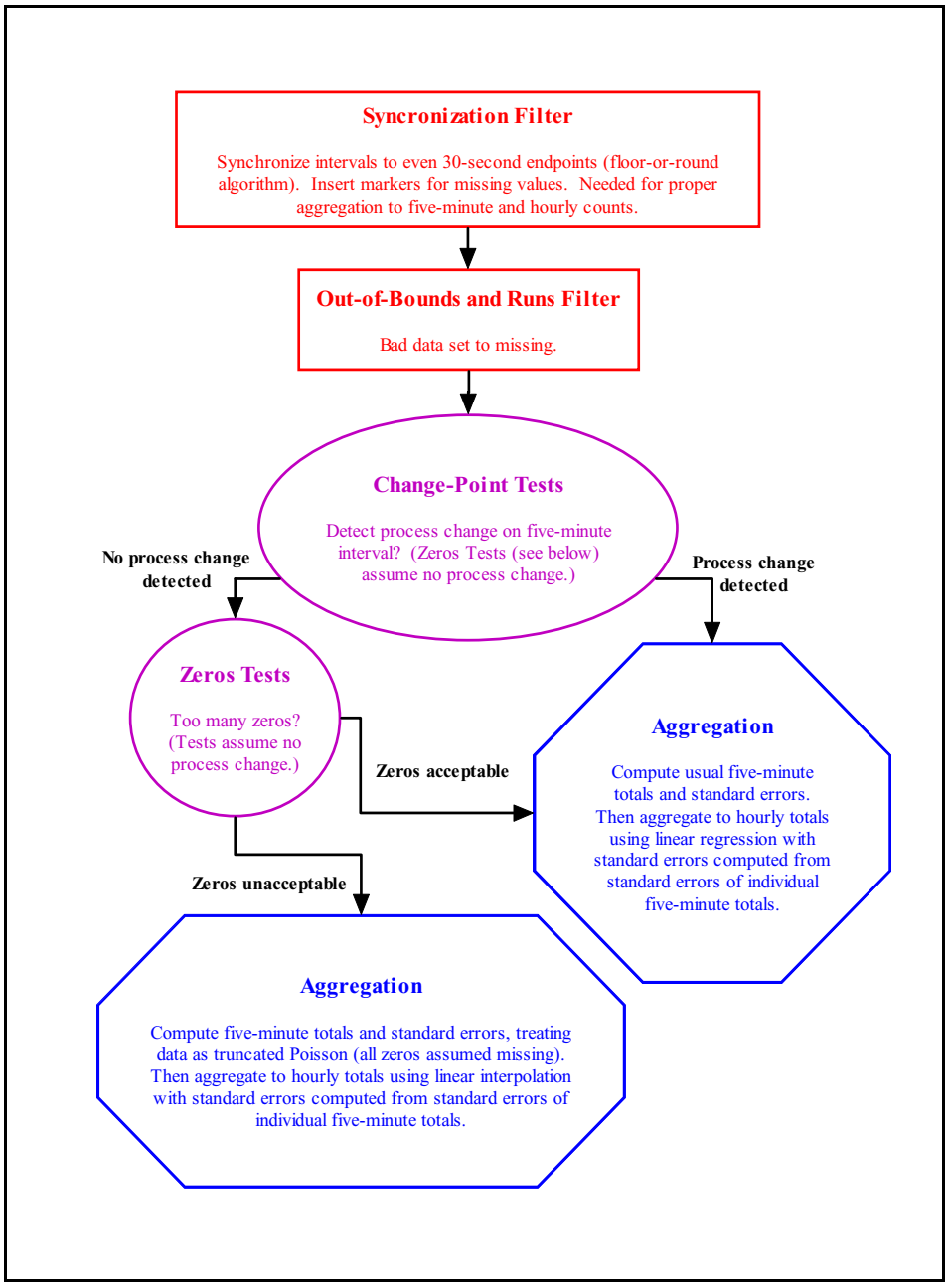


Figure 4.2
Data Flow Schematic for Orlando Florida Thirty-Second Traffic Counts



Questionable traffic volume data in INFORM and SMIS can be summarized into four categories:

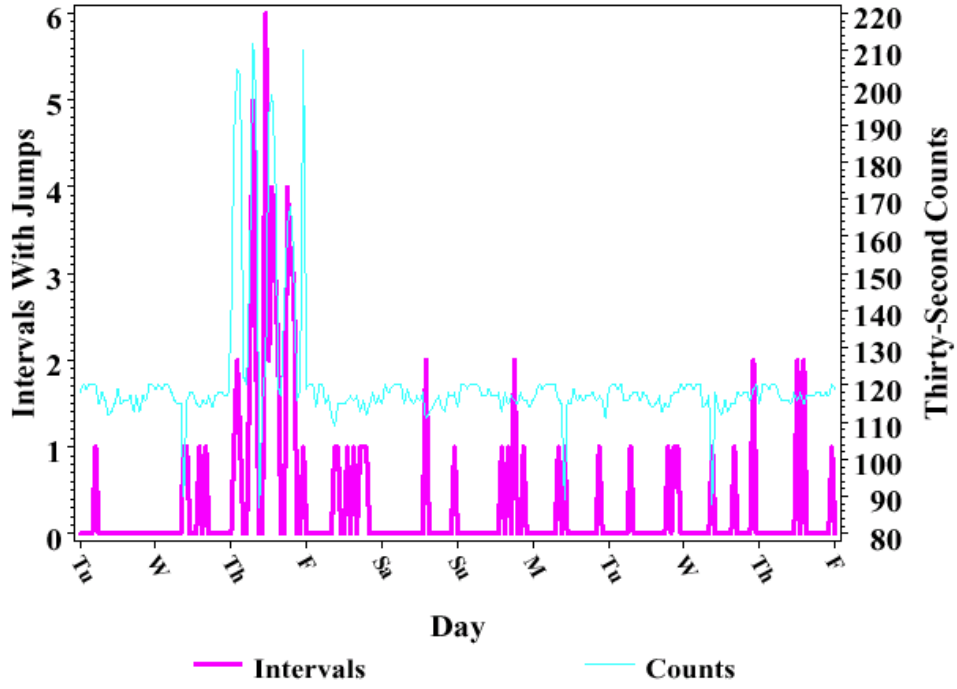
1. **An unreasonable value for the traffic count.**

   Questionable data in this category are counts that are either too high or too low. One criterion used to determine the reasonableness of the count data is to compare the count data in question with counts from adjacent lanes, counts from the previous period, or counts from the following period. Figure 4.3 depicts 30-second raw count data recorded during a ten-day period in September. Also included in Figure 4.3 are the process changes (jumps) in the thirty-second counts. If the traffic count in a thirty-second interval is statistically different from the traffic counts in the preceding and succeeding thirty-second intervals, then a change in the process (or a "jump" in the data) is identified. Process changes can be attributable to a number of factors such as incidents or equipment failure. Note the anomalous data on the third day (Figure 4.3).

2. **Consecutive runs of an identical value, including 0.**

   An example of this type of anomaly occurs when a traffic counter consecutively records 30-second traffic counts that have the same value of, for example, 17 vehicles throughout a period of 5 minutes. These are successions of the same value that are too long to be credible. Long stretches of the same value are called "runs" of data. When runs of the same value occur, the value is most often zero. However, runs of values other than zero do occur. Sensors may experience similar mechanical errors or, for reasons we have not determined, data may just be

Figure 4.3 Raw Thirty-Second Traffic Count Data



repeating themselves. Runs of both small numbers (e.g., a run of twelve 1's) and larger numbers (e.g., a run of fifteen 27's) were observed.

3. **Legitimate zero traffic counts vs. ambiguous zero traffic counts.**

Most loop detector units report "0" when either (a) no vehicles are detected, or (b) the software detects errors. Zero values in traffic count can also occur at later stages of the data processing. It is likely to record no traffic on less traveled roads at certain times of the day (for example, three o'clock in the morning). Thus, some zeros in the data stream are legitimate zeros, indicating that there is no

traffic. However, zeros are often used to denote malfunctions. These ambiguous zeros occur both intermittently and in runs.

For ITS applications that are performed in conjunction with video, it might be possible to resolve the two different kinds of situations with zeros. If the data are processed automatically, then occurrences of zeros present a problem. If zeros that denote malfunctions are instead treated as valid values, then the count estimates would be biased downward. On the other hand, if valid zeros are interpreted as missing data and otherwise ignored, then count estimates would be biased upward. The challenge, then, is to distinguish legitimate zeros from ambiguous zeros.

4. Missing traffic count data.

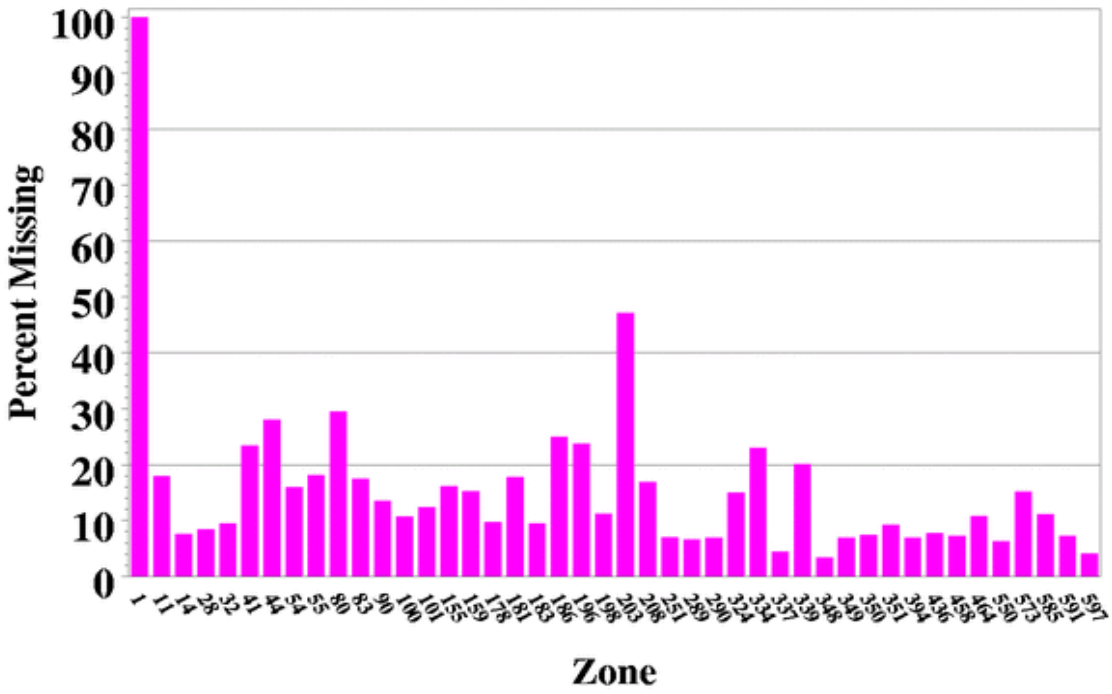
Missing values (including ambiguous zeros interpreted as missing data) pose a significant problem in ITS traffic counts. Table 4.1 illustrates the missing data problem. It is obvious from Table 4.1 that as far as data collection is concerned, some stations are essentially nonexistent, and almost all stations exhibit an appreciable amount of missing data. Hours with no counts recorded at all are not uncommon. Although data in Table 4.1 are for August 1998, data for other months reflect similar problems. In fact, missing data are not unique to ITS data collection systems. Traditional continuous count data are also often incomplete. For example, two nearby Florida DOT continuous count sites (Sites 130 and 303) report the following percentages of non-missing hourly counts for 1998 (all lanes): 93% (on Highway 130 eastbound), 86% (on Highway 130 westbound), 88% (on Highway 303 eastbound), 98% (on Highway 303 westbound).

Table 4.1 Percentages of Available Hourly Counts August 1998

Station ID	East Left	East Center	East Right	West Left	West Center	West Right
60	83.8	.	0.0	83.8	.	83.8
61	0.0	.	0.0	0.0	.	0.0
62	0.0	.	0.0	0.0	.	0.0
63	0.0	.	0.0	0.0	.	0.0
64	83.8	.	83.8	83.8	.	83.8
65	83.8	.	83.8	83.8	.	83.8
66	83.8	83.8	0.0	83.8	83.8	0.0
67	0.0	83.8	0.0	83.8	83.8	0.0
68	83.2	83.8	83.6	83.3	83.8	82.8
69	80.3	80.4	80.4	80.4	.	80.4
70	80.3	.	80.4	80.4	.	80.4
71	83.8	.	83.8	83.8	.	83.8

Another example of the missing data problem is in Figure 4.4. Figure 4.4 depicts the percentage of missing values from 44 stations. These sites were selected after a preliminary examination as sites that have fewer missing data points than other sites. Note that even the best sites among these "good" sites are plagued by missing data problems. For example, about 3% of the five-minute counts were missing at Site 348. This finding indicates that addressing missing traffic counts is a considerable challenge in using ITS traffic count data.

Figure 4.4 Percent of Missing Five-Minute Total Counts (All Lanes) for 44 Selected Sites, February-April, 1999



Data Aggregation

As previously mentioned, traditional traffic counts are recorded on an hourly basis whereas ITS data are recorded on significantly shorter intervals – typically 30 seconds, 1 minute or 5 minutes. In order to determine how ITS-generated data are related to traditional traffic counts, time stamps of both data sources need to be aligned on an hourly basis. If ITS counts are in a continuous stream, with no counts missing, no counts outside reasonable ranges, or otherwise spurious, then it would be straightforward to aggregate them into hourly totals. Unfortunately, ITS data, like other traffic volume data, are rarely flawless.

Consequently, aggregating ITS data into hourly counts becomes a challenge. And, data aggregation cannot be performed until all data quality issues are fully addressed.

Florida's 30-second traffic counts will first be aggregated to 5-minute counts, then into hourly counts.

SECTION 5. STATISTICAL PROCEDURES TO IDENTIFY AND CORRECT UNACCEPTABLE DATA

This section describes steps developed to identify questionable data. Procedures to address each of the four aforementioned data quality issues are also reported. Although the results are for traffic volume data and are not for vehicle class counts, the methods could apply equally well to vehicle class data when ITS-generated vehicle classification counts become available. Since the data quality check procedures can be applied periodically (for example, whenever data are archived), a by-product of these procedures is a method to monitor automatically the quality of the data as well as the operational status of the hardware.

In order to determine objectively the quality of the data in the context of the data's natural variability, it is important to make certain statistical assumptions about the traffic count process. For example, traffic counts exhibit seasonal, day-of-week, and hourly trends. The changes in traffic within a one-hour time interval are difficult to characterize. As the time scale decreases from hours to minutes, erratic count patterns are likely to predominate. A model based on three assumptions was used. This model seems suited for the purpose of data quality checks. It computes other measures of data quality as well as standard errors. The three assumptions are:

1. Means of the thirty-second or one-minute counts are approximately constant on a five-minute interval. This assumption implies that traffic counts recorded in a 30-second interval (or on a one-minute interval) do not vary drastically from each other within a five-minute interval.

This assumption seems reasonable because a five-minute interval is relatively short.

2. Differences between the average counts and the actual counts during five-minute intervals are statistically independent. This is an essential assumption for subsequent data quality checks. Because the averages of traffic counts can change every five minutes (under Assumption 1), **differences** between the five-minute averages and the actual thirty-second counts are, approximately, independent random noise. The rationale is that because the model is flexible enough that averages can change every five minutes, the differences between the thirty-second counts and five-minute averages are, approximately, independent random noise. This is despite that fact that traffic counts observed in a period are highly related to traffic counts observed in the previous or the latter periods (positive serial correlation). The Durbin-Watson test for serial correlation only rarely rejected this assumption.

3. In any given hour, the mean values of the total counts in five-minute intervals are linearly related to time[6]. The assumption assumes that an hour is short enough that the five-minute count totals are, on average,

---

[6] Specifically, for any given hour, let C_i denote the mean of the total count for the i^{th} five-minute interval $(i=1,...,12)$ in the hour. Let t_i denote the midpoint time for that interval. We assume that the C_i are approximately linear in the t_i: $C_i \approx a+bt_i$ for some intercept a and slope b.

approximately linear in time. This is a judgment call, but note that: (1) a peak or dip in traffic counts does NOT contradict this approximation, and (2) the piecewise linear approximation can hold even if the traffic counts change rapidly (i.e., the line is steep). This assumption is used later to identify "outliers" and to adjust missing data.

5.1 Identify Unreasonable Values for the Traffic Counts

Five-minute counts that are over 300 were set to missing. The basis for setting the threshold at 300 was the assumption that it is highly unlikely that vehicles traverse traffic counters more frequently than one vehicle per second for five consecutive minutes (= 60 vehicles/minute × 5 minutes). The 300 threshold was adjusted proportionately for partial counts (counts measured in fewer than 5 minutes). At the low end of the scale, counts less than zero were also set to missing.

If Florida's thirty-second counts are over 30, then they were set to missing. This criterion was based on the same rationale used for the 5-minute traffic count data. At a vehicle speed of 88 ft. per second, a traffic count of 30 would mean an average traffic flow rate of one eighteen-foot passenger car per second for the entire thirty-second period. Counts less than zero were also set to missing.

5.2. Determine the Validity of Runs of Identical Value

Procedures were developed to detect runs (or successions) of an identical value that are too long to be credible. When runs of identical values occur, the value is most often zero. However, runs of values other than zero also occur. The procedures we established to address runs of zero were different from those for runs of non-zero.

To decide whether a run is "credible," three assumptions were made. First, the probability of a run should decrease with run length, suggesting that the likelihood of observing a run of sixteen 8s should be greater than the likelihood of observing a run of twenty-six 8's. Second, runs are more likely when the average count rate is steady. Thirdly, the probabilities of runs should decrease with the amount of traffic flow. The third assumption is that when counts are small, runs are more likely because fewer different values are likely to occur. For example, six consecutive counts of exactly 20 are improbable simply because, even at an average flow of exactly 20, it would be unlikely to see six consecutive counts of **exactly** 21.

The Poisson distribution satisfies these aforementioned assumptions[7]. This distribution was used to calculate the greatest probability of observing a run of length N for a non-zero value X. Based on Assumptions (1) and (2) on Pages 28 and 29, the maximum probability that the next n five-minute counts are all x is $[e^{-\lambda}\lambda^x/x!]^n$ with $\lambda = x/n$, where $x > 0$. Table 5.1 contains the maximum probabilities for runs of X's from 1 to 10 of length 1 to 10. For example, the maximum possible probability of recording 3 consecutive 2's is 0.0198 while the maximum possible probability of recording 3 consecutive 4's decreases to 0.0075.

When this probability becomes "too small," then the run is considered highly unlikely and thus unacceptable as valid data. Values of such runs are set as being missing. A threshold of "too small" is subjectively set to be less than .0005. Based on this table, the largest values whose significance levels exceed .0005 gives the sequence 7, 5, 5, 4, 4, 4, 3, 3, 3, 3 of maximum acceptable run lengths for runs of values 1 to 10, respectively. Figure 5.1

[7] See Chapter 8.1 of Snedecor and Cochran 1989.

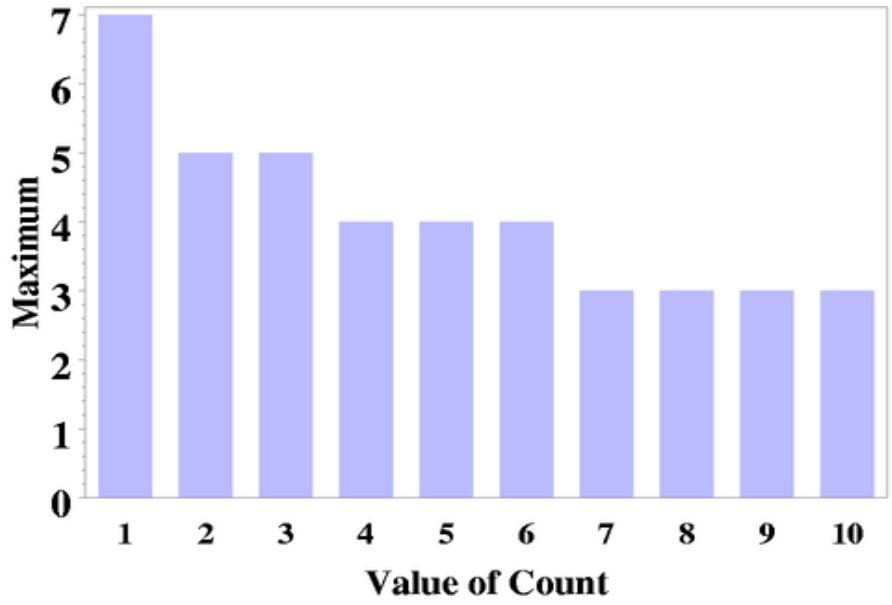
Table 5.1
Poisson Probabilities of Runs of Length *N* for Non-Zero Values *X**

Count Value (*X*)	Length of Run (*N*)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	**0.0003**	**0.0001**	**0.0000**
2	0.2707	0.0733	0.0198	0.0054	0.0015	**0.0004**	**0.0001**	**0.0000**	**0.0000**	**0.0000**
3	0.2240	0.0502	0.0113	0.0025	0.0006	**0.0001**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
4	0.1954	0.0382	0.0075	0.0015	**0.0003**	**0.0001**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
5	0.1755	0.0308	0.0054	0.0010	**0.0002**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
6	0.1606	0.0258	0.0041	0.0007	**0.0001**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
7	0.1490	0.0222	0.0033	**0.0005**	**0.0001**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
8	0.1396	0.0195	0.0027	**0.0004**	**0.0001**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
9	0.1318	0.0174	0.0023	**0.0003**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
10	0.1251	0.0157	0.0020	**0.0003**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**

*Probabilities . Red (bold) indicate probabilities < .0005. For X > 10, we always accept runs of 3.

illustrates the most likely length of runs for traffic counts 1 through 10. For example, given the acceptance level of 0.0005, observing 7 consecutive 1's is somewhat conceivable while observing 8 consecutive 1's is highly unlikely with a probability of 0.0003. Similarly, observing 4 consecutive 6's is probable while observing 5 consecutive 6's becomes highly unlikely with a probability of 0.0001. For values greater than 10, a length of 3 is considered credible for this value. For example, 3 consecutive 13's are considered valid, and 4

Figure 5.1 Maximum Acceptable Run Lengths for Count Values Greater than Zero



*Based on Poisson distribution. Maximum of three taken for all count values greater than ten.

consecutive 13's are considered valid. This process applies to Florida's 30-second traffic counts as well.

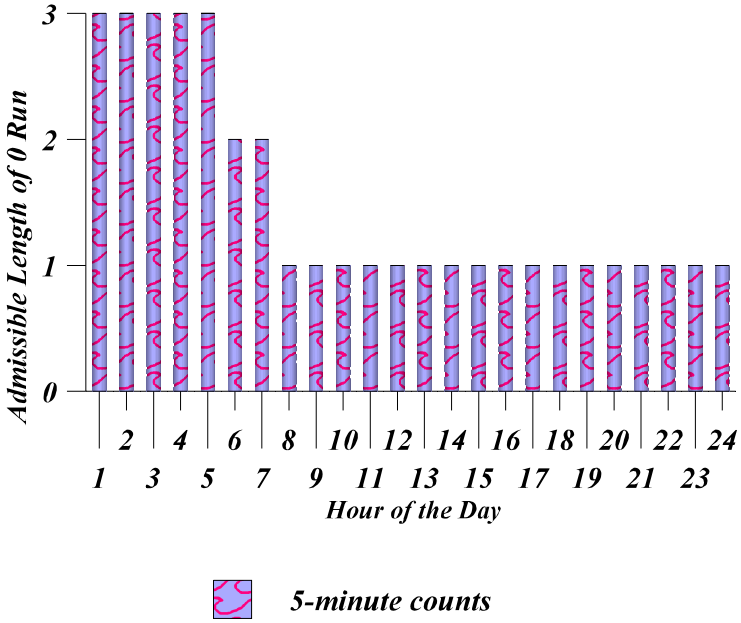
5.3 Determine the Validity of Zero Traffic Counts

The approach used to determine the validity of runs of zeros is different from that discussed above. This is because when the average traffic flow is zero, runs of zeros are valid and are exactly what **should** occur. This is the issue of legitimate zeros vs. ambiguous zeros.

Given a zero recorded traffic count, resolving whether it represents a valid zero or a detector malfunction can often be accomplished by exploring the context: if nearby counters record high traffic volume, a zero count most likely indicates a malfunction. However, if nearby counters record low traffic volume, a zero count most likely indicates a valid zero count. This approach is not foolproof, however, as detectors can malfunction when traffic flow is very low, and accidents or congestion can cause the true flow to decrease to zero very suddenly. Furthermore, implementing an automatic resolution (on a computer) in a large database is much harder than applying the resolution by simply looking at a few cases of data. Of course, the best solution to this ambiguity is to use a symbol other than zero to denote data collected during malfunction.

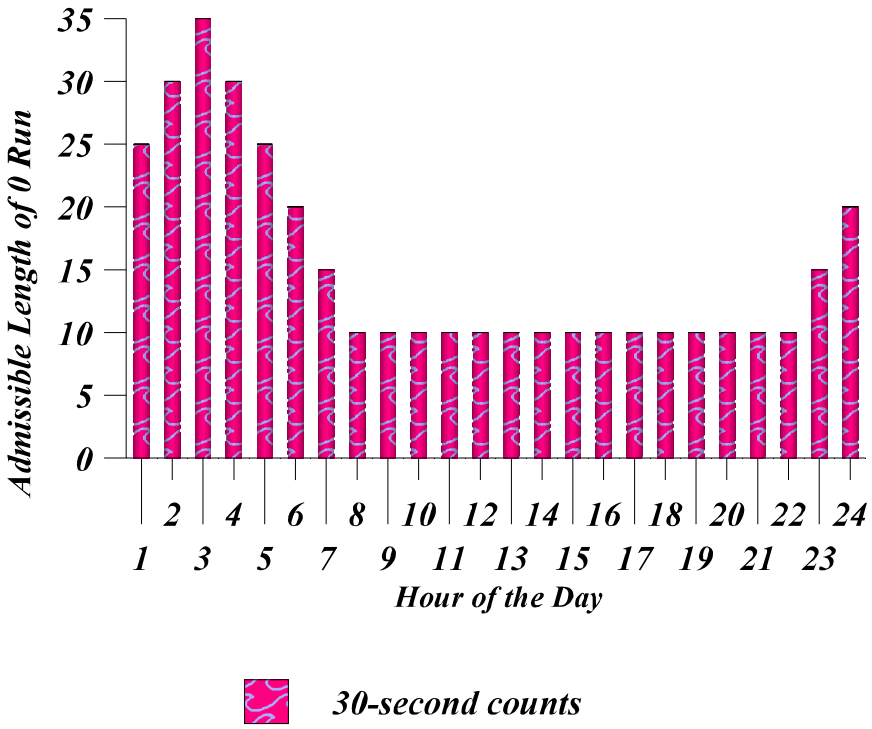
The probability of a run of consecutive "true" zero counts depends on the location of the loop detector and on the hour of the day, as well as other factors such as the season and day-of-the-week. As an approximation, thresholds were subjectively set for the number of valid consecutive five-minute zero counts (Figure 5.2). Note that these thresholds are specific to the time of the day. For example, on most roads it is reasonable to assume that there is no traffic for as much as three consecutive 5-minute intervals at three o'clock in the morning. On the other hand, it is improbable to observe no traffic for more than one 5-minute interval at three o'clock in the afternoon. Data with consecutive zeros longer than these thresholds were set to be "missing." Similarly, thresholds for the number of valid consecutive 30-second zeros were subjectively set (Figure 5.3), and runs of zero that were outside these ranges were set to be missing.

Figure 5.2
Maximum Admissible Number of Consecutive Zeros
New York's Five-Minute Counts



We recommend that people **not** use the imputed traffic counts we developed in this study for any "official" traffic counts or traffic monitoring. The focus of this study was on assessing the potential for using ITS-generated data for such purposes, rather than to actually develop a revised set of traffic counts. In particular, we recommend that the question of setting thresholds for legitimate strings of 0's be addressed by field study, as well as by further scrutiny of the data. Field study would involve direct observation of traffic at different locations at certain times of the day to determine reasonable ranges for traffic at these locations and times. Scrutiny of data would consider the spatial correlation in traffic data and

Figure 5.3
Maximum Admissible Number of Consecutive Zeros
Florida's Thirty-Second Counts



identify ways of correcting inadmissable strings of 0's when adjacent or nearby sites record non-zero traffic.

In addition to runs of consecutive zeros, the thirty-second Florida data are also plagued with an excessive number of intermittent zeros. To test whether the frequency of intermittent zeros was indeed excessive, we considered the surrounding context — other thirty-second

counts that were near in time to the ones in question. This test depends, however, on these counts having the same statistical distribution. Although this situation of identical statistical distributions is an implication of Assumptions (1) and (2), the assumption could certainly not be true, especially when there are traffic incidents or other congestion that cause traffic patterns to deviate from what is expected. In this case, a statistical test about intermittent zeros that assumes that the underlying statistical process is homogeneous (i.e., identical distributions) would likely be invalid. Therefore, before performing a test on intermittent zeros, a test was performed to decide whether there were local changes in the traffic flow (or "process"). This test compared, for each thirty-second time interval, the traffic counts for the five minutes before and for the five minutes after that point.[8]

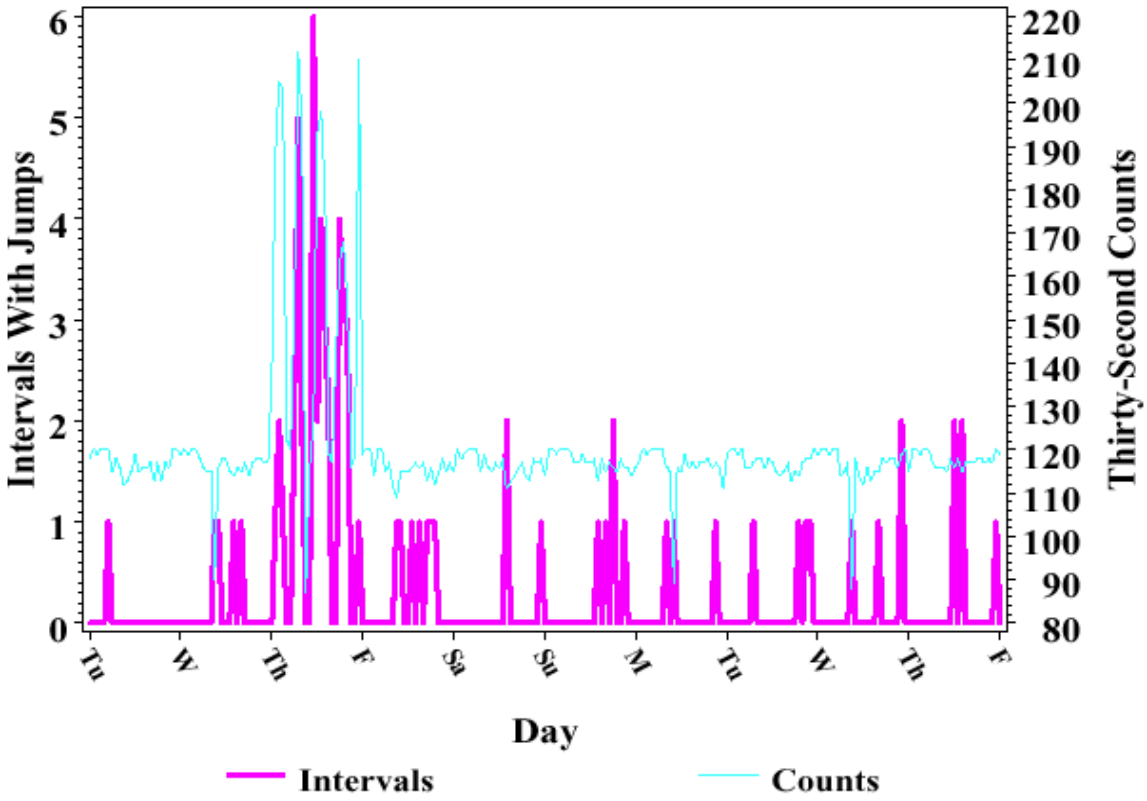
Figure 5.4 shows the number of thirty-second intervals found to have process changes (or "jumps" in data) for the Station 18, left lane, eastbound in I-4, during September 1-10,

[8] Actually two tests were considered for this process-change test. The first was a test based on the Poisson distribution, comparing the "before" and "after" counts during the five-minutes before and five minutes after each thirty-second time interval. The second test is a pooled t-test, comparing the average values of the "before" and "after" five-minute counts. Both tests are based on Assumption (1), that the "before" and "after" process mean values are each constant in their respective five-minute intervals. (The test is to determine whether the two mean values differ.) Both tests are based on Assumption (2), that the observations are independent. In the Poisson-based test, given the sum of the counts before and after the interval when potential change might take place, the "before"and "after" counts have a conditional binomial distribution with a "success" probability of $n_1/(n_1+n_2)$, where n_1 and n_2 are the numbers of thirty-second intervals before the half-minute interval. Which of the two tests is better depends on the statistical distribution of the underlying counts. The closer to Poisson that distribution is, the better the Poisson-based test tends to be. Although no definitive conclusion was reached about which test was better, the Poisson test was selected because it seemed more reliable when counts were small.

1998. Figure 5.4 also shows the number of thirty-second count each hour. Obviously, there is a lot of "chatter" (erratic fluctuation) in the count process on September 3.

For each five-minute interval, if the process-change tests indicated a change in traffic counts might have taken place in the interval, then no test about the intermittent zeros was performed for that interval. This limitation implies that the validity of the intermittent zeros cannot be deciphered simply by the techniques discussed here.

Figure 5.4
Intervals with Jumps in Florida Thirty-Second Counts
Station 18, Left Lane Eastbound, I-4
September 1-10, 1999



SECTION 6. AGGREGATION TO HOURLY COUNTS

As pointed out in Figures 4.1 and 4.2, five-minute or 30-second traffic data were aggregated into hourly counts after questionable data were corrected. Next, if any of the hourly total counts was missing, then procedures were developed to impute missing hourly counts.

Because the patterns of the 30-second traffic counts are more variable than those of the 5-minute counts, aggregating data into hourly counts is more complex for 30-second data. Separate aggregation procedures were developed for five-minute data and for 30-second data. This section describes those procedures.

6.1 Aggregating Five-Minute New York INFORM Traffic Counts

If none of the five-minute traffic counts was missing within a one-hour interval, then aggregating to hourly counts is straightforward — simply by adding together 12 five-minutes counts. Unfortunately, this is not the case. One way to resolve this problem would be to use the mean value of the available five-minute counts to impute the missing five-minute counts. For example, if 4 of the 12 five-minute counts were missing, then the four missing points would be estimated by the average value of the 8 five-minute counts.

However, the average traffic flow and the individual 12 five-minute counts within an hour may differ appreciably, especially from the beginning to the end of the hour. Thus, if five-minute counts are missing at the beginning or at the end of an hour, simple averaging the available counts might lead to a biased estimate of the total counts for the hour. To avoid

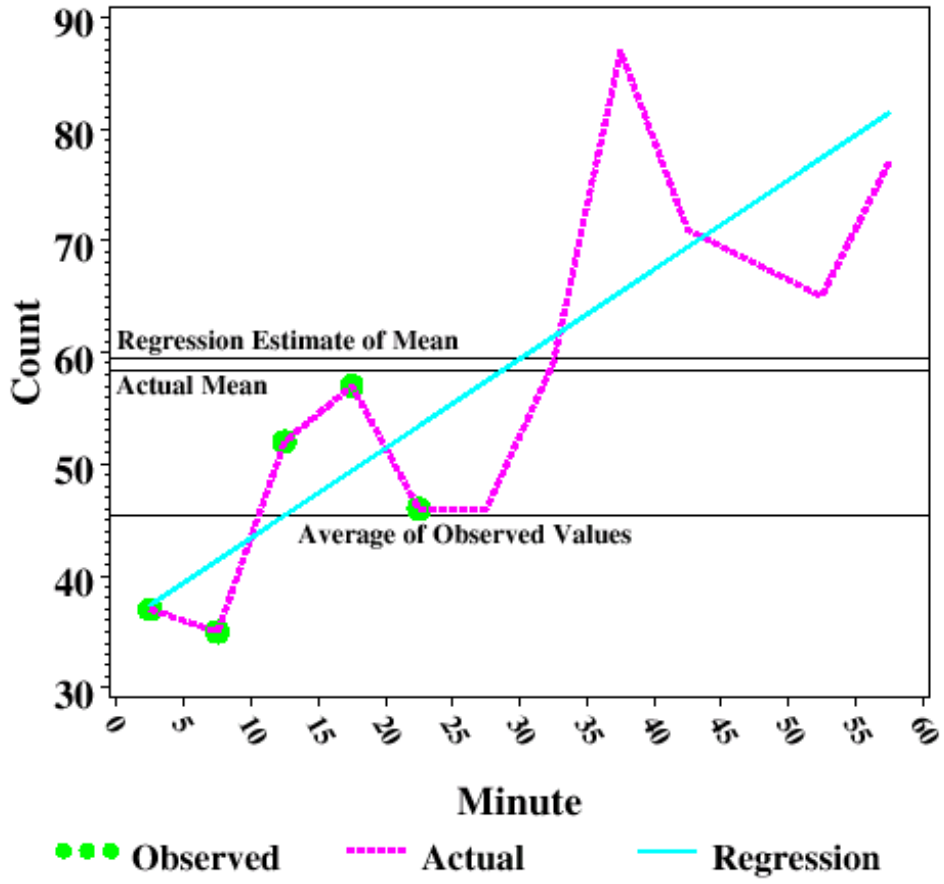
biases, adjustments were made with linear (i.e., straight-line) regressions. This is a reasonably approach based on Assumption 3. When a regression line is fit to the five-minute counts, the "fitted" value at the thirty-minute point provides an estimate of the mean value for the hour that is adjusted for missing five-minute counts. This mean traffic count, when multiplied by twelve (twelve five-minute intervals per hour), is a trend-adjusted estimate of the hourly total.

When some of the twelve five-minute counts are missing in a one-hour period, the benefit of using the regression adjustment to estimate the hourly total counts is illustrated in Figure 6.1. In this illustration all of the twelve five-minute counts in an hour were actually "recorded," as represented by the dotted magenta line. The actual hourly count is 696 vehicles, and the average five-minute count for this hour is 58 (= 696/12) vehicles, as represented by the line labeled "Actual Mean."

Now, assume that only the first five of the twelve five-minute counts were "recorded," as represented by the green dots, and that the remaining seven of the twelve five-minute counts were "not-recorded" or missing. The average of the five "recorded" traffic points yields an estimated average of 46 vehicles for every five-minute interval, as represented by the line labeled "Average of Observed Data." The total traffic counts for this hour can be estimated as 552 (= 46 × 12).

Alternatively, a regression line can be fitted to the five "recorded" traffic counts, as represented by the blue solid line. The fitted regression line suggests that during this hour there are, on average, about 59 vehicles for every five-minute interval. The regression-based average yields an hourly traffic count of 708 (= 59 × 12) vehicles. The bias introduced by not using regression adjustments is substantial – an under-estimate of about (58-46) × 12 =

Figure 6.1
Regression Adjustments for Estimating Hourly Traffic Counts



144 vehicles per hour. This conclusion confirms Assumption 3 that says that the regression estimate of the mean tends to be closer to the true mean than the average of the observed counts, and likewise for their corresponding estimates of the hourly total.

Occasionally, all twelve of the five-minute counts are missing except few. In such cases the regression line cannot be fitted to the five-minute counts in a one-hour period, and the simple un-weighted average is used (i.e., there is no adjustment for time trend).

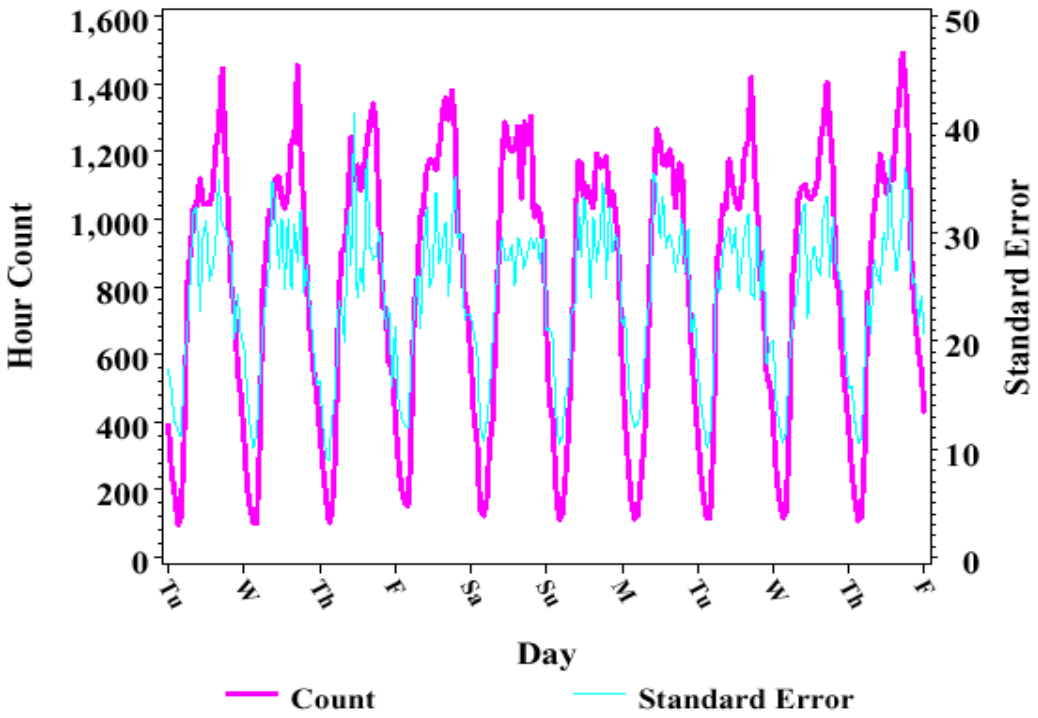
Note that occasionally traffic counts in a five-minute interval are actually recorded based on traffic counts recorded for fewer than five minutes (e.g., four minutes, see Table 3.1). A restriction is subjectively imposed to require that if there are fewer than *two* one-minute counts available in a one-hour period, then the hourly traffic count estimate is set to be missing. Otherwise, hourly traffic count estimates were estimated using the regression adjustments. In practice, this criterion of *two* one-minute counts does not seem critical because missing values tend to occur (though not always) in succession. If an hourly traffic count estimate can be computed, then usually a majority of the counts are available to compute it.

The total traffic count estimates and the associated standard errors were computed for each hour of a day (Figure 6.2). Furthermore, the number of five-minute counts, the number of one-minute counts, and the number of five-minute intervals with zero counts were computed. These are useful quality control statistics because they embody many characteristics of the data, with the standard error being probably a better overall measure of data quality.

6.2 Aggregating Thirty-Second Florida SMIS Traffic Counts

In aggregating the 30-second SMIS traffic counts into hourly totals, an intermediate step was added. The thirty-second counts were first aggregated to five-minute totals. The five-minute totals were then aggregated into the hourly totals. The procedure of first

Figure 6.2 Hourly Traffic Counts and the Associated Standard Errors
ITS Station 18, Left Lane, I-4 Eastbound
September 1-10, 1998



aggregating to five-minute counts and then to hourly totals was done to take advantage of the approximate independent and identical distributed (IID) nature of the thirty-second counts over short spans of time (e.g., five minutes), while at the same time adjusting for trends within a one-hour period.

The choice of five minutes for the interval length was subjective, and was made because five-minute intervals comprise a reasonably number of thirty second counts (ten), and because there are also a reasonable number of five-minute intervals in an hour (twelve). The choice was also made for convenience because five minutes is the minimum recording-time interval for the New York INFORM counts. Once the thirty-second traffic counts were aggregated into five-minute intervals, the aggregation of the five-minute traffic counts into the hourly totals followed the procedure used for the INFORM five-minute traffic counts.

Based on Assumptions (1) and (2), there are no appreciable trends of thirty-second traffic counts within any given five-minute interval. Thus, if any of ten thirty-second counts is missing within a five-minute interval, then the traffic counts for that five-minute interval was estimated by "scaling up" the total of the non-missing thirty-second counts. For example, if two of the ten thirty-second counts are missing, then the total five-minute counts was estimated as the sum of the eight non-missing counts multiplied by 10/8.

Furthermore, the standard error of the five-minute counts was estimated using the usual formula for the standard error. This was not possible with the New York INFORM counts, because only the five-minute totals or partial five-minute totals, and the number of minutes recorded within each five-minute interval, were recorded in the INFORM data.

The process of aggregating these five-minute traffic counts into hourly totals basically follows the procedure used for the INFORM data - the weighted regression adjustment approach as depicted in Figure 6.1. The details of this approach are presented in Appendix 1. If there are fewer than *four* thirty-second counts available in an hour, then the hourly total estimate is set to be missing. This restriction was subjectively imposed. If there are at least

two five-minute intervals with count estimates, then hourly totals are estimated using the weighted regression approach. If only one five-minute count is available, then that count (and its standard error) is "scaled up" to an estimate for the hour (i.e., multiplied by 12).

Figure 6.3 illustrates the hourly traffic totals for Orlando ITS Station 18 on I-4 westbound. Station 18 is effectively coincident with Florida telemetering continuous count Site 130 (near State Route 482, Orange County). The figure also shows the corresponding traffic counts from Site 130[9]. Although the counts from these two sources are not identical, the percentage differences are very small and they are essentially the same.

In addition to the hourly count estimates, for each station, direction, lane, day and hour, we computed standard errors of the hourly total estimates and of the following four "quality control" statistics:

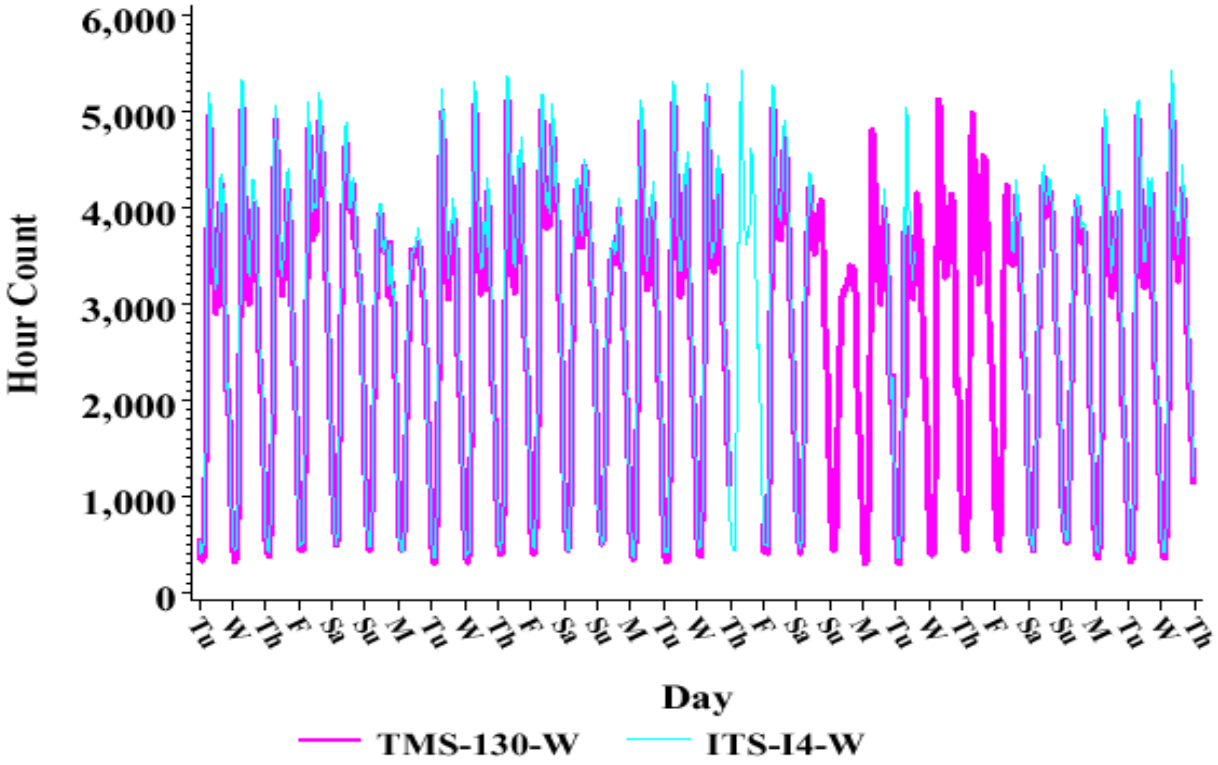
(1) the number of thirty-second counts observed within an hour,
(2) the usable number of five-minute counts within an hour,
(3) the number of thirty-second counts within an hour that are zero, and
(4) the number of five-minute intervals that are considered to have a valid change in traffic volume (from the process-change test).

The standard errors reflect the variability in the hourly total estimate, including the natural statistical variability traffic volume.

---

[9] Data were kindly provided by Mr. Harshad Desai, Florida Department of Transportation.

Figure 6.3 Comparison of Hourly Traffic Counts from Two Different Sources Conventional Counter and ITS Counter



Although the restrictions on the minimum number of data points were relatively permissive, there remained a substantial amount of missing hourly data. Table 6.1 lists the proportions of available hourly counts in August 1998 for Stations 60-71 on Florida I-4. These are the proportions of hours in a month that have a sufficient number of thirty-second counts (at least four) and a sufficient number of five-minute counts (at least one) to estimate an hourly total. Excluded from this table are statistics for Stations 1 through 10, which recorded no traffic volume data for the entire month of August.

Table 6.1
Percentages of Available Hourly Counts for August 1998 Orlando ITS Traffic Monitoring Stations

	Eastbound			Westbound		
Station ID	Left lane	Center lane	Right lane	Left lane	Center lane	Right lane
60	83.8	.	0.0	83.8	.	83.8
61	0.0	.	0.0	0.0	.	0.0
62	0.0	.	0.0	0.0	.	0.0
63	0.0	.	0.0	0.0	.	0.0
64	83.8	.	83.8	83.8	.	83.8
65	83.8	.	83.8	83.8	.	83.8
66	83.8	83.8	0.0	83.8	83.8	0.0
67	0.0	83.8	0.0	83.8	83.8	0.0
68	83.2	83.8	83.6	83.3	83.8	82.8
69	80.3	80.4	80.4	80.4	.	80.4
70	80.3	.	80.4	80.4	.	80.4
71	83.8	.	83.8	83.8	.	83.8

It is obvious from Table 6.1 that as far as data collection is concerned, some stations are essentially nonexistent, and all stations have a considerable amount of missing data. There are hours when no traffic volume is recorded at all. Furthermore, missing data are not unique to the ITS collection system. Traditional continuous count data are also often incomplete. The next section describes procedures developed to impute missing data.

SECTION 7. IMPUTATION OF MISSING HOURLY COUNTS

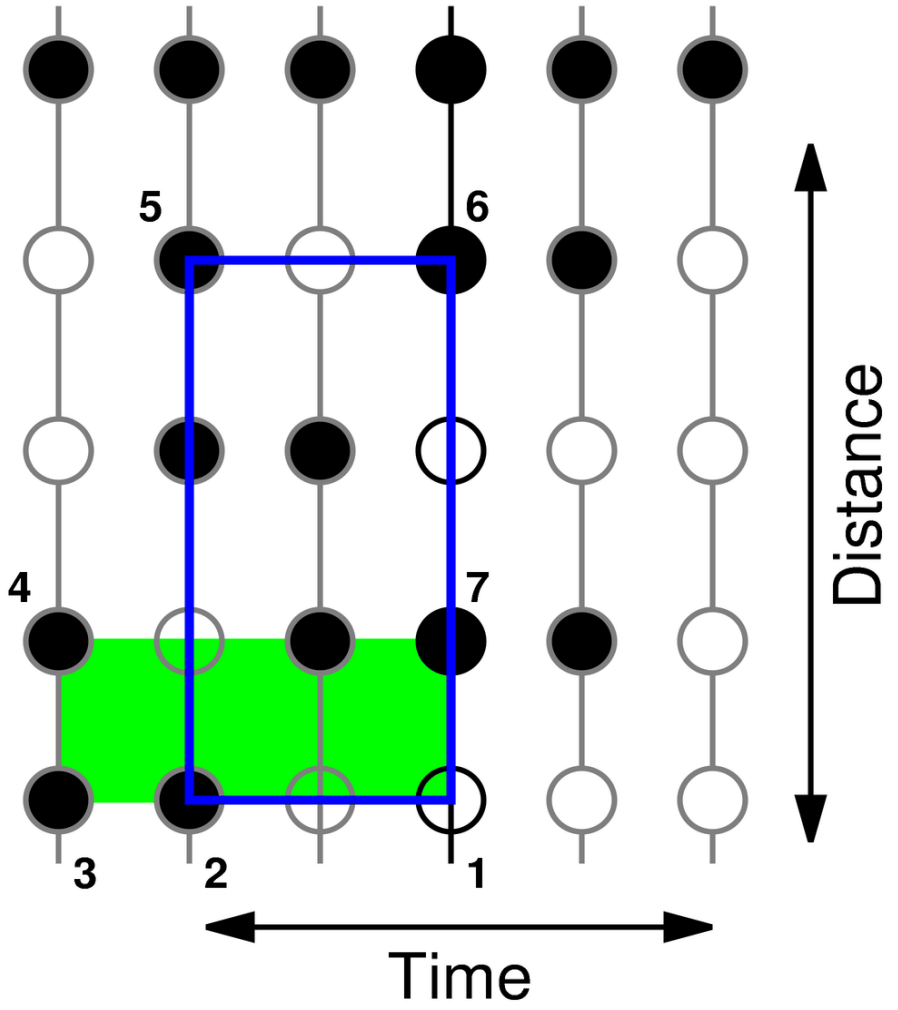
The concept of imputation, or imputing missing data, refers to the process of replacing missing values with estimated values that are based on data collected from neighboring counts – neighboring in terms of either location or time. Because ITS counters are spaced at short (e.g., half-mile) intervals, they are particularly suitable for providing data to develop imputed estimates. In this section an approach is developed to impute missing hourly counts.

The basic idea of the imputation approach is to assume that neighboring count stations will experience similar changes in traffic volume from one part of the day to the next, and that the traffic volume at one station will not be dramatically different from those at nearby stations at a given time. For example, if Station 1 experiences a traffic volume decrease of 2% at time point *t* from the time two periods prior, *t-2*, then the stations nearby will also experience about a 2% decrease from *t-2* to *t*. In this approach, traffic count data that are collected from count stations further away, or from the same station but a long time apart, carry less weight in the imputation approach than those closer in time and location. Figure 7.1 is used to explain this idea.

Solid points in Figure 7.1 denote *available* traffic counts, whereas hollow points denote *missing* counts. Points 1, 7, and 6 denote count stations at the "current" time point, at stations upstream or downstream from each other. Assume that Point 1 denotes Station X at the "current" time. The hollow point in Point 1 indicates that traffic volume is missing at Station X at the "current" time. The objective here is to impute this missing data. Let Point 2 denote the most recently available traffic count data collected from Station X, and let Point 3 denote

the second most recently available traffic count data collected from Station X. Point 7 denotes the traffic volume collected from the station immediately upstream from Station X at the

Figure 7.1
Schema of Data Imputation Approach



"current" time. Point 6 denotes the first station before Point 7 at which there is a count available at the current time.

The idea is to identify a "rectangle" where Point 1 is one of the four points and the remaining three points have available traffic count data (solid points). In this example, there are two possible rectangles, one defined by points 1, 3, 4 and 7(the green rectangle), and the other defined by 1, 2, 5, and 6 (the rectangle outlined in blue). The available counts at either 3, 4, and 7, or 2, 5, and 6 can be used to impute the missing count at Point 1. These estimates use points upstream in location and back in time. Similar estimates and weights can be computed for points future in time or downstream in location. All of the estimates are combined into a weighted average, using reciprocal areas as weights.

7.1. Imputation Algorithm

Consider two count stations at two time periods, as indicated in Figure 7.1 by Points 1, 2, 5, and 6. Let $C(i)$ denote a count at point i. A basis for an imputation approach is that

$$\frac{C(1)}{C(2)} \approx \frac{C(6)}{C(5)}, \tag{7.1}$$

and hence

$$C(1) \approx \frac{C(2)\,C(6)}{C(5)}. \tag{7.2}$$

Note that this is essentially equation 7.1. Thus, if $C(1)$ is missing, but $C(2)$, $C(5)$, and $C(6)$ are each available, then $C(1)$ can be imputed using $C(2)$, $C(5)$, and $C(6)$.

To impute a missing count at Point 1 in Figure 7.1, Points 2, 5, and 6 were selected by taking Point 2 to be the closest point back in time from Point 1, and by taking Points 5 and 6 to be the closest upstream pair of points having available counts at the times of both Point 2 and Point 1.

Another approach is illustrated in the figure with Points 1, 3, 4, and 7. Point 7 is taken to be the closest point upstream from Point 1. Points 3 and 4 are taken as the closest pair of points back in time from both Point 1 and Point 7. It is possible that the rectangles determined by these two approaches coincide. It is also possible that one or both of the rectangles do not exist at all—for example there might not be a point satisfying the conditions for Point 2 or Point 3.

Lower and upper bounds of 0.75 and 1.25, respectively, were arbitrarily set on the ratios of traffic counts at different locations nearby, at a given time (e.g., *C(2)/C(5)*). These bounds imply that the difference in traffic volume at a given time between two nearby stations should not exceed 25%. In practice, we found that the ratios were occasionally too big to be credible. Using these bounds seemed to produce more intuitively reasonable results. That said, different data might suggest different bounds, and different bounds certainly might be appropriate for a different network of stations. The algorithm and the reasonableness of these bounds may require further examination in the future.

In addition to selecting "triples" of points previous in time and upstream in direction, points can be selected that are either subsequent in time and downstream in direction. Each such "triple" leads to a different imputation of the missing count. Thus, the question arises

as to which three points should be chosen, or if multiple selections are made, how the multiple imputations should be combined.

Though neither temporal (trends in time) nor physical proximity (trends in location) implies a correlation between counts, it is reasonable to choose points that are nearby in both time and location. Thus, we chose points that minimize either *d*, the roadway distance from Point 1, or ΔT, the time span from Point 1. We did this separately for upstream, downstream, time-backward, and time-forward selections. This procedure leads to the selection of up to eight different sets of "triples" and thus eight different imputation results.

To combine these eight imputation results into one estimate, a weighted averaged

$$\frac{\sum_{i=1}^{8} w_i C(i)}{\sum_{i=1}^{8} w_i}, \tag{7.3}$$

can be used, where

$$w_i = \frac{1}{d_i \Delta T_i}, \tag{7.4}$$

and where for the i^{th} selected "triple" of points, d_i denotes the distance between Point 1 and the point in the triple closest to Point 1, and ΔT_i denotes the time span between Point 1 and the point in the triple closest in time to Point 1. The greater the distance in location and time

between Point 1 and the "triple" *i*, the smaller the weight for that "triple" in the weighted average. If for point selection *i* (*i* = 1, ..., 8), there is no point to select, we take $w_i = 0$.

The quality of the imputed estimates depends on:

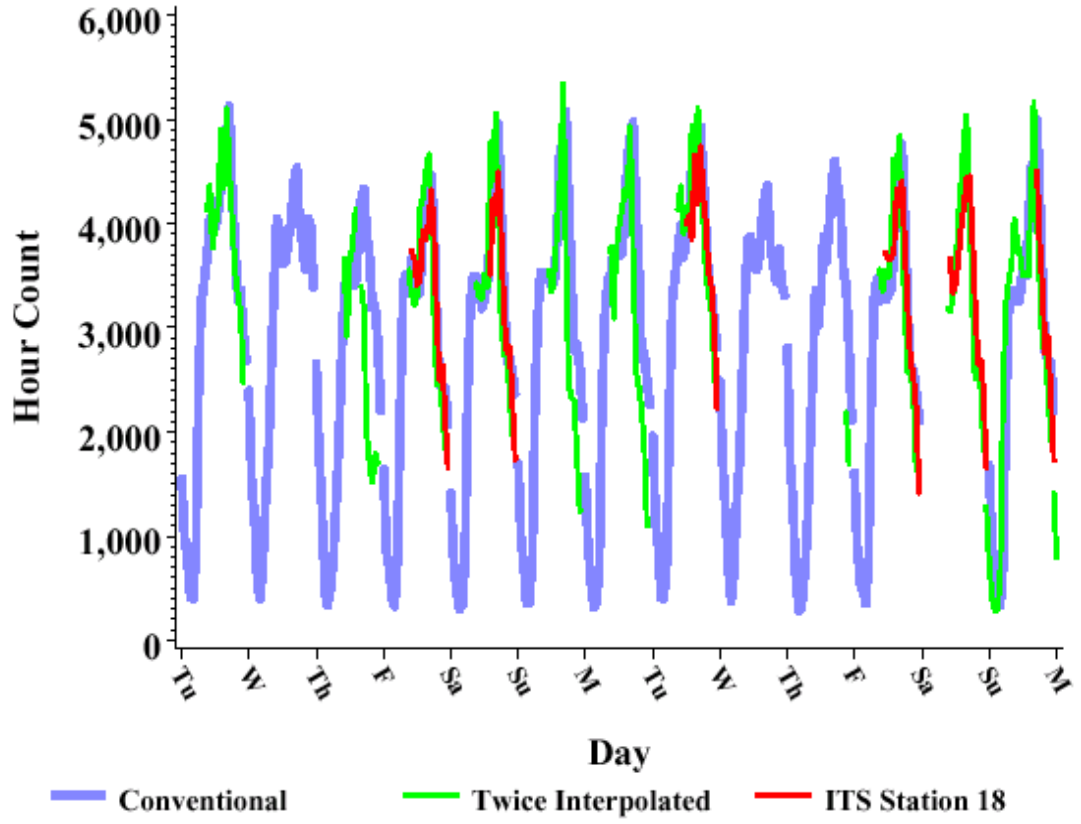
1. the quality (e.g., standard errors) of the count "triples" used in the imputation, and
2. The accuracy of the approximation.

It seems reasonable to assume that the accuracy of the approximation depends on the proximity of the triples in location and time to the original Point 1. Since the counts *C(2)*, *C(5)*, and *C(6)* are jointly correlated, however, we cannot ascertain precisely how the variability of the triples affects the precision of the estimated imputation.

Another approach to assess the quality of the imputations is to simulate the imputations by imputing the missing counts when counts (the *C* (*1*)'s) are actually available, and to compare the imputed counts to the actual counts. A statistic such as the root mean squared error can be used to characterize broadly the difference between the imputations and the actual values. Unfortunately such a characterization is biased because when the actual traffic counts are available, counts from the nearby stations are also more likely to be available. As such, imputations tend to be better when they are simulated when *C(1)*'s are actually recorded, compared to imputations that must be estimated when *C(1)*'s are actually missing. Therefore, we assessed the imputation procedure by comparing the imputations to an independent source, namely, count data from a nearby State continuous counter.

Figure 7.2 shows the original and imputed ITS hourly counts. Also shown in the figure are hourly counts from a continuous count station – a station essentially coincides with the ITS station. The figure demonstrates that the actual and imputed traffic counts from the ITS station agree closely with those collected from a nearby conventional traffic count station.

Figure 7.2 Comparing Three Sets of Hourly Counts:
Conventional Source, Imputed ITS, and Recorded ITS



When a count was missing, but the counts immediately before and after either in time or location were both available, we used the average of the "before" and the "after" counts to impute the missing count.

SECTION 8. BENEFITS OF USING ITS DATA

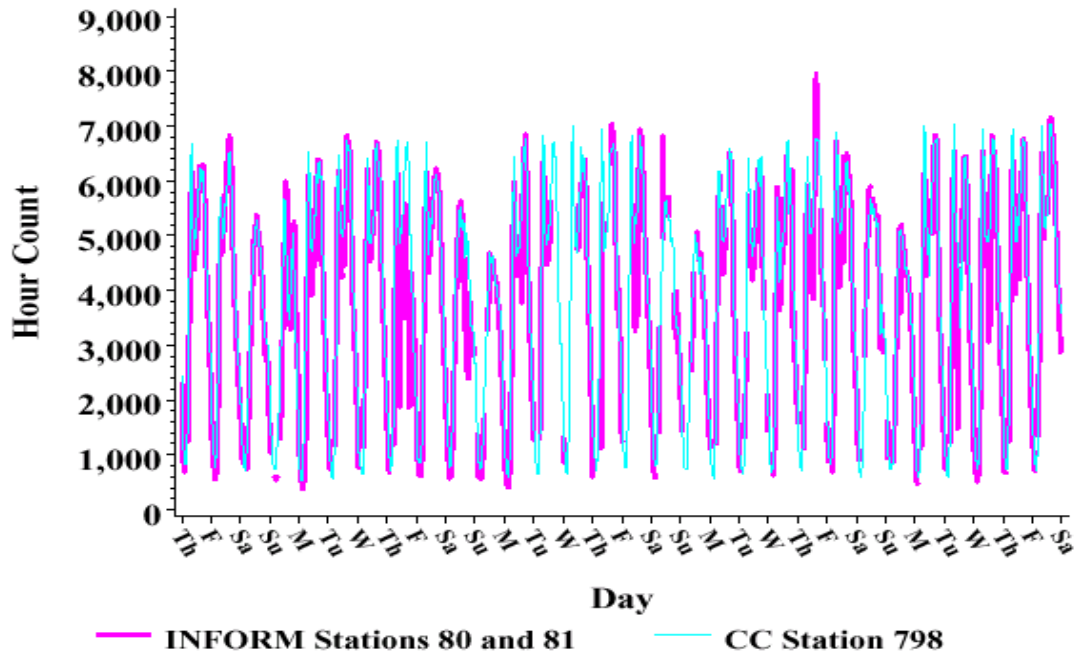
In this feasibility study, the specific benefits of using ITS-generated traffic count data for purposes other than the original intent (i.e., traffic management and operations) can be gauged in three ways. First, can ITS data replace data collected from nearby traditional count stations? If so, this benefit can in fact be measured in monetary terms. Second, can ITS data supplement the traditional data so that more reliable estimates can be developed? Third, can ITS data be used to calculate adjustment factors that are more reliable than those calculated based on a limited number of continuous counters?

Not included in this study is the feasibility of using ITS-generated data to meet traffic monitoring needs in urban areas. Monitoring traffic in urban areas has been a demanding challenge because of the traffic conditions in these areas. It is conceivable that ITS-generated traffic volume data can significantly improve traffic monitoring programs in urban areas as long as the following questions are addressed. First, how representative are the traffic conditions monitored by the ITS sensors of condition in the other parts of the urban area? This question would be less relevant when more ITS sensors are installed in the area. Second, if the conditions are not representative, then what is the relationship between traffic conditions monitored by the ITS sensors to those on the rest of the network within the area? An understanding of these questions is crucial to "expand" the traffic conditions monitored by the ITS sensors to the entire network. More research is needed to determine the feasibility of using ITS-generated data to monitor traffic conditions in urban areas.

8.1 Supplement or Replace Conventional Traffic Count Data

ITS loop detectors not only provide traffic counts continuously, but they provide a redundancy of counts in that counts from nearby stations tend to be similar. This notion is clearly demonstrated in Figure 8.1. In this figure, the red line represents the combined traffic counts of INFORM Stations 80 (main lanes) and 81 (ramp) on Long Island Expressway East at Round Swamp. Superimposed on the INFORM traffic counts are traffic counts (in blue) from the essentially coincident, conventional continuous count station 798. Although there

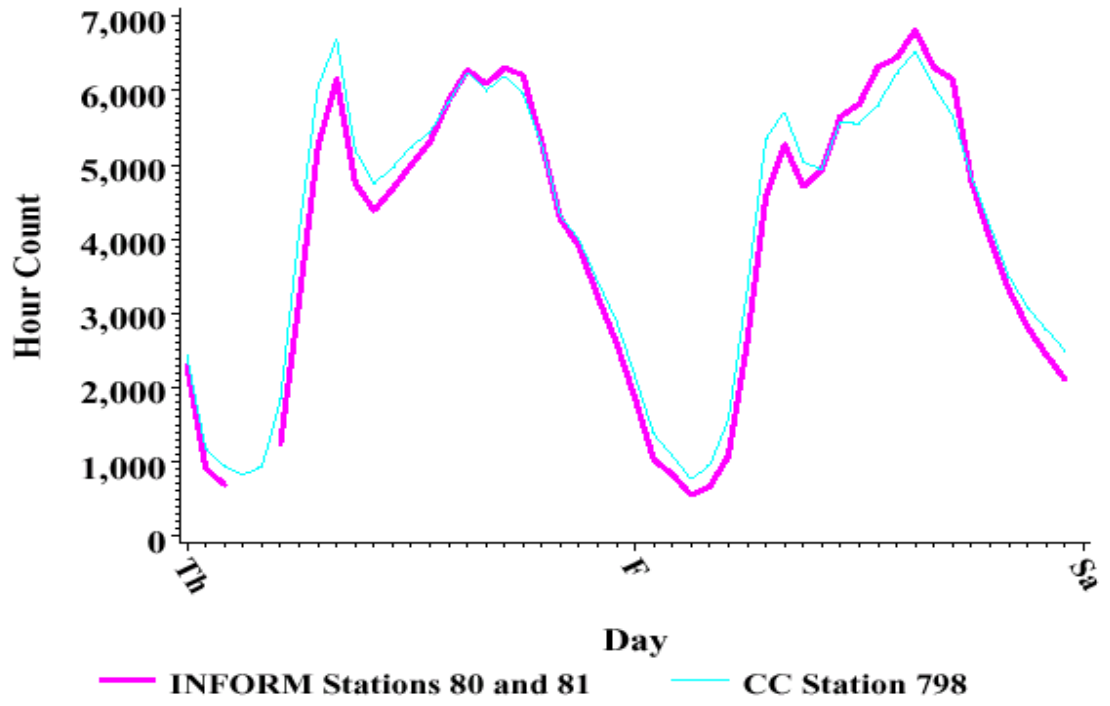
Figure 8.1 Traffic Count Comparison Between ITS and Conventional Counter



are slight discrepancies, the two sets of counts agree well despite the slight differences (Figure 8.2).

The redundancy afforded by ITS data is exploited in the imputation algorithm discussed in the last section. The similarity between ITS loop detector data and traditional

Figure 8.2 A Two-Day Comparison between ITS and Conventional Counters
INFORM Stations 80 and 81 vs. Conventional Counter 798
April 1 and 2, 1999



count data suggests that, from the data-sharing perspective, ITS counters could replace the nearby conventional counters. However, from the operations perspective, it might not be appropriate to suggest such replacement. "Turning-off" an existing continuous counter probably might be more involved than continuing its operation. Thus, the answer to the first question – Can ITS data replace data collected from nearby traditional count stations? – is "perhaps, but only after a more complete assessment is made of other reasons for having these stations in operation and of the cost of discontinuing their use."

That said, the similarity between the two data sources offers a different opportunity. Like ITS counters, continuous counters experience significant problems in terms of missing and questionable data. Data from nearby ITS counters can be used to supplement missing or questionable traffic counts resulting from equipment malfunction or other reasons. Given the size of the missing data problem experienced by conventional counters, this use of ITS data would probably be more valuable than the notion of replacing the conventional counters. With missing and questionable data corrected, traffic estimates (such as the estimated annual average daily travel (AADT)) that are derived from conventional counters will be significantly more reliable. Thus, the answer to the second question – Can ITS data supplement traditional data so that more reliable estimates can be achieved? – is "yes."

8.2 Development of Adjustment Factors

The primary application of the continuous traffic counts was to adjust short-term counts (typically collected during a 48- or 72-hour period) for temporal effects on traffic volume. The underlining assumption of this approach is that a continuous site and a short-term site are close enough and functionally similar enough that their relative changes in traffic volume are about the same. That is,

$$F = \frac{\text{AADT at continuous site}}{\text{Short count at continuous site}} \approx \frac{\text{AADT at short-count site}}{\text{Short count at short-count site}}. \qquad (8.1)$$

F can be computed using continuous count data, and can serve as an adjustment factor for estimating AADT at the short-count site:

$$\text{AADT estimate for short-count site} = F \times (\text{short-count}). \qquad (8.2)$$

The validity of this approach hinges on what is "close enough" and what is "functionally similar." Continuous counts are expensive to install and maintain — that is the reason for installing numerous short-term counts. The expense of the continuous counts has caused their deployment to be considerably reduced.

To demonstrate the value-added of ITS data in developing adjustment factors, data from 27 ITS stations on I-4 were used – with each ITS station emulating a continuous count station. The adjustment factors were calculated to "adjust" the average traffic counts for the week of September 28, 1998[10]. The X-axis in Figure 8.3 represents the spatial relationship among all ITS stations. The distance (in miles) is plotted from a given count station to the beginning of the ITS deployment (denoted by point 0 on the x axis). The Y-axis plots the adjustment factors calculated using count data from ITS stations. The magenta dots denote adjustment factors calculated from ITS count data while the blue dot denotes the adjustment factor calculated from conventional count station 3069.

[10] The approach Florida Department of Transportation takes to compute adjustment factors is actually based on weekly average rather than daily counts.

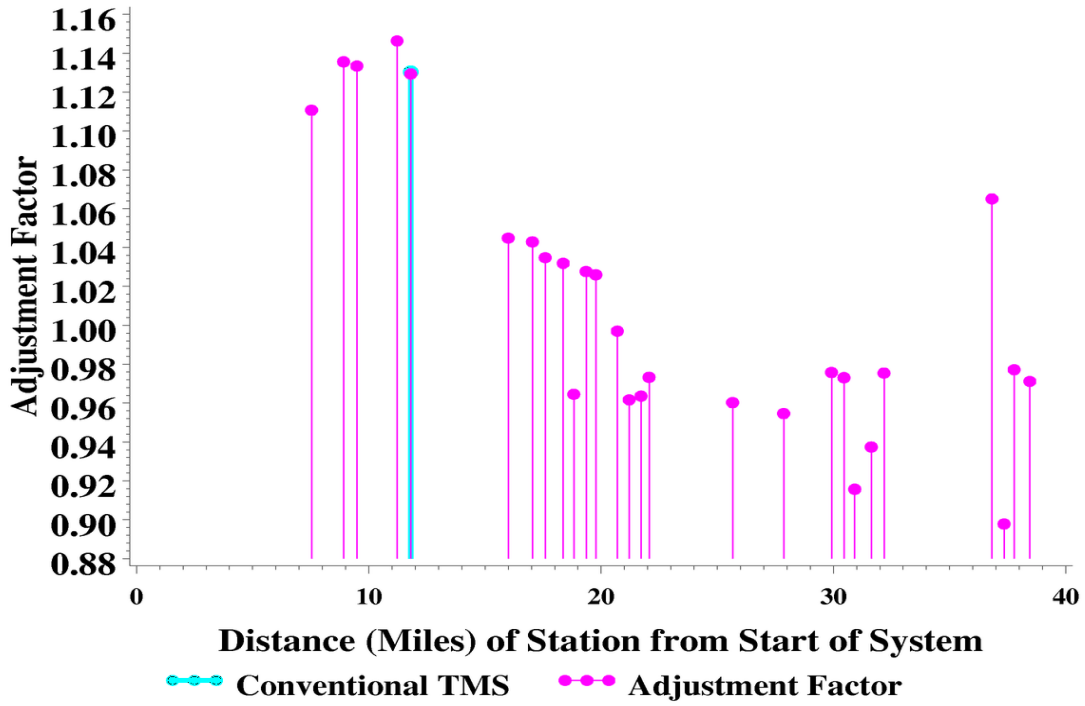
It is obvious from the figure that there is considerable variation in the adjustment factors. However, adjustment factors calculated from nearby counts are nevertheless correlated. Data from conventional count station 3069 yields an adjustment factor of 1.13. If this adjustment factor had been used to adjust the short-term counts collected from stations in the 39 instrumented miles, then the AADT for this 39 mile stretch would probably have been over-estimated by at least 10%.

Note that rather than use the entire year of count data to calculate these adjustment factors, they are computed from counts taken from an 85-day period. This is because data for only about 85 days during April - October of 1998 (the period for which we received traffic data) were sufficiently complete to provide hourly counts for an entire day, without any data interpolation. Despite the short time span, the adjustment factors in Figure 8.3 are true in the sense that they are the correct value for "scaling up" short-term counts.

8.3 Estimate Vehicle Miles Traveled (VMT)

Short-term counts and AADT estimates are usually converted to estimates of VMT. To do this, the traffic counts are multiplied by the length of the roadway segment or link with which the count site is associated. To make a conversion of ITS counts to VMT, we must therefore define links and align them with ITS stations. The information on ITS station location was used for this purpose, though there is not a unique way to do it. With ITS loop detectors installed every half mile apart, the conversion from traffic counts to VMT is straightforward. The benefit of using ITS count data to estimate VMTs is implicitly implied by the benefits of using ITS count data to calculate adjustment factors.

Figure 8.3
A Comparison of Adjustment Factors Calculated from ITS Data and Conventional Count Data



SECTION 9. DATA ARCHIVE AND DISSEMINATION

In previous sections of this report, we described how we assessed ITS and traditional count data to determine whether they can be used in conjunction with, or in addition to, each other. Comparisons of ITS-generated loop detector data and data from traditional continuous counters demonstrated that it is feasible to use ITS-generated loop detector data to either replace or supplement data collected from nearby traditional traffic counters. It was also demonstrated that ITS data can be used to develop adjustment factors that more accurately reflect the spatial variability in traffic volume. However, it was also clear that significant effort is generally required to "prepare" **archived** ITS-generated traffic count data, in terms of both data format and data quality.

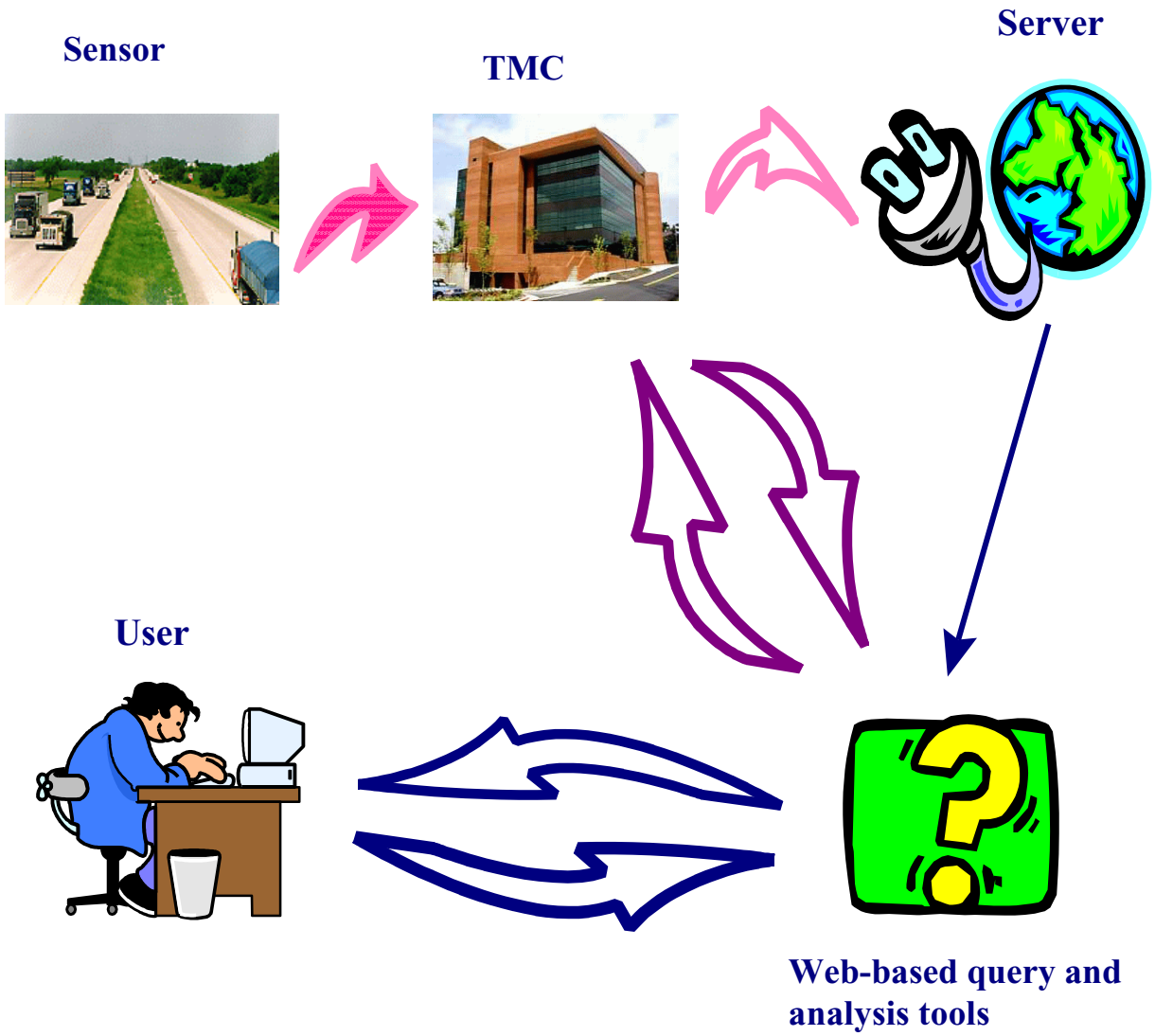
With these fundamental points established, the discussion now turns to the mechanics of acquiring, storing and disseminating these data. Ideally, an ADUS system will acquire the field data, store these data in a "warehouse," conduct quality checks, correct erroneous data, derive desired information from the data, and deliver the information to the user (Figure 9.1). And, ideally, these functions should be seamless and automatic.

Both INFORM and SMIS data are archived on a regular basis. An ADUS prototype was developed to integrate most of the functions downstream from archiving and transmitting data. The architecture of this prototype is in Appendix 2. The URL of this prototype is:

http://www-cta.ornl.gov:8000/adus/

Although data archiving and transmission are not operational in our prototype, methods to do so are proposed.

Figure 9.1. Seamless Flow of Information in ADUS



Proposed Methods to Acquire ITS Data

In this project, data transmission was accomplished by data being sent to ORNL on magnetic tapes (in the case of New York INFORM) or through file-transfer protocols (FTP) (in the case of Florida SMIS). New York INFORM data were received on VAX VMS backup tapes. Each tape contains one month of INFORM data and is about 100 MB in size. The tape media (the TK50 tape) are so old that it was difficult to find a drive that is capable of reading these tapes. If ADUS is to be fully operational, then this method of data distribution would impose a significant technical barrier to users, and is unacceptable.

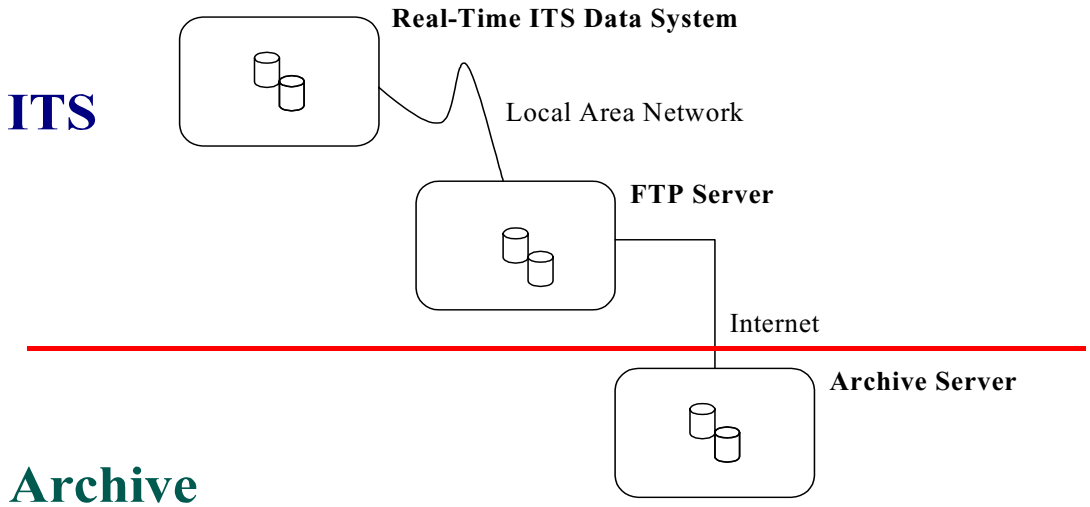
The Florida data are in Microsoft Access format with monthly data archived in an individual database. The compressed databases were transferred from UCF to ORNL using the File Transfer Protocol (FTP) over the Internet. This was accomplished with significant human intervention. Ideally, an ADUS system should have a built-in function to automatically "transmitted" archived data to users without any human intervention - a "pull" strategy whereby archived data are "pulled by" rather than "pushed to" users.

To solve these data-transfer problems, we propose a methodology and implementation that can automatically transfer data on a daily basis using File Transfer Protocol (FTP) service and/or the Internet (Figure 9.2).

In this method, data acquisition is performed with the following steps:

- Real-time data are saved in daily files that are closed at midnight of each day
- At 1:00 am, a second computer "fetches" a copy of the daily file for the previous day from the first computer, saves it on a local disk, and compresses it. This second computer is connected to the Internet and provides a **secured** FTP service.
- At 2:00 am, the off-site archive server (a third computer) uses FTP to download the compressed data, decompresses it, performs data quality checks, corrects erroneous data, and develops a meta file of the data – "data about data."

Figure 9.2. Proposed ADUS Data Acquisition



We also propose that data acquisition be implemented in a comprehensive integrated module called the Data Acquisition Service (DAS). The DAS should have the following operational characteristics:

- All data management is under program control and is performed automatically, requiring no human intervention.
- Logs are maintained on data acquisition status and are reviewed on a regular basis.
- Data acquisition is fault-tolerant – if the FTP transfer fails, it will be repeated.
- DAS actions are logged/documented.

- The archive administrator is immediately notified of severe problems.
- New data will be automatically integrated into the "archive."

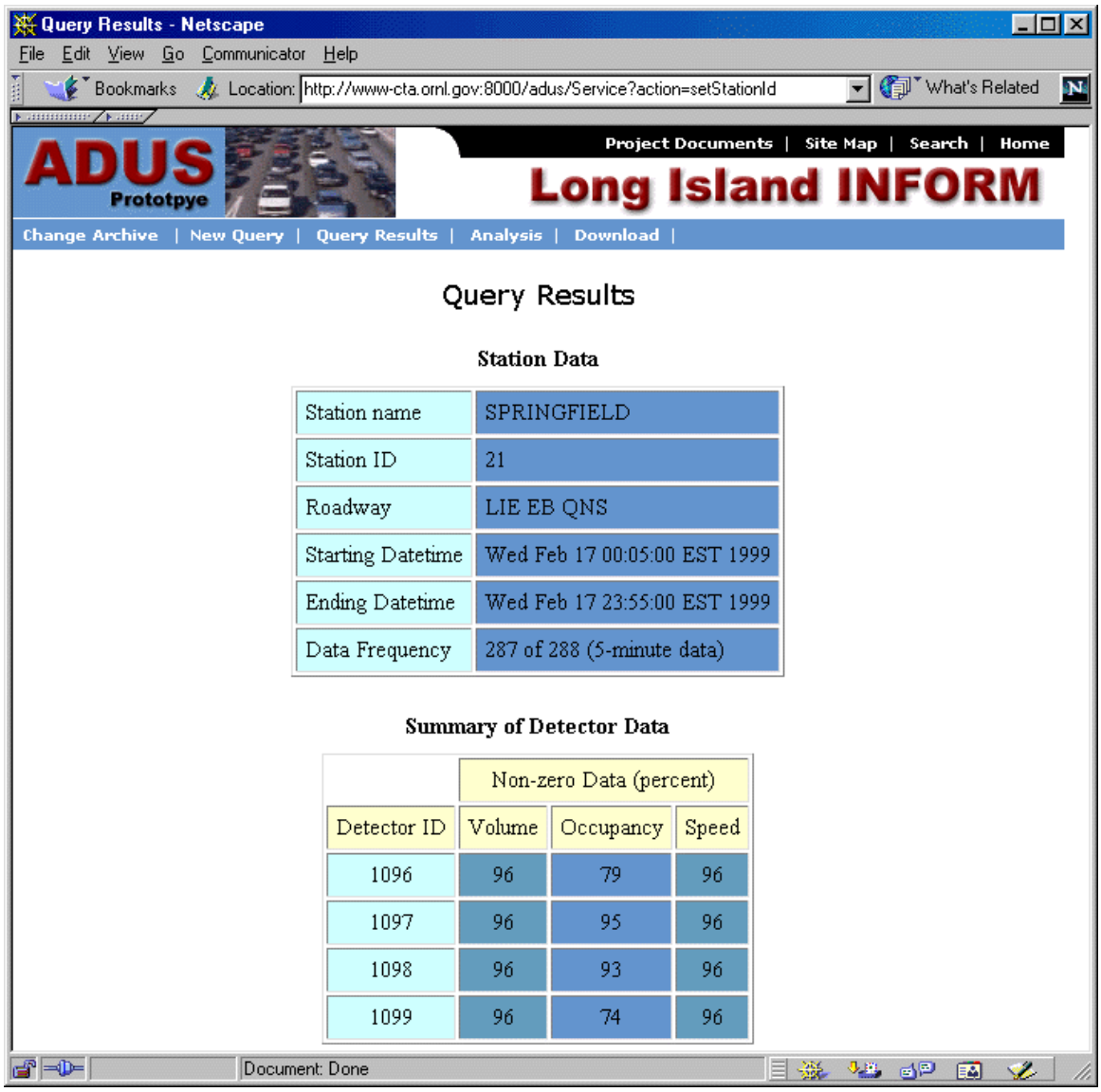
ORNL's ADUS Prototype System

As previously mentioned, most of the functions downstream from data acquisition, as depicted in Figure 9.1, are included in an ADUS prototype system. The ADUS prototype system is on a web site that allows users to query, analyze, display, and download the data. Procedures needed to prepare the ITS data for analysis are "invisible" to the user.

After data acquisition, data verification is performed automatically by the Data Import and Verification (DIV) function. Again using the INFORM data as an example, the incoming data are stored in its original format, which is very efficient and organized into a daily file. The prototype executes a limited number of quality-checks to verify the integrity and consistency of the archived data. Although only simple checks are included in this prototype, all of the procedures described from Sections 5 through 8 should be incorporated in a fully-operational system.

Results from the data quality checks are summarized in the "meta-data" base. A meta-data base contains data about data. It contains descriptive information on a station such as station location, time intervals for data collected, and the results of the data quality checks. Figure 9.3 illustrates an example of the "meta data" of Station 21 on the Long Island Express eastbound for Wednesday, February 17, 1999. This stations has four loop detectors, one for each of the four lanes. Four percent of the volume data collected from Detector 1096 were non-zero while 21% of the lane occupancy data from the same detector were non-zero.

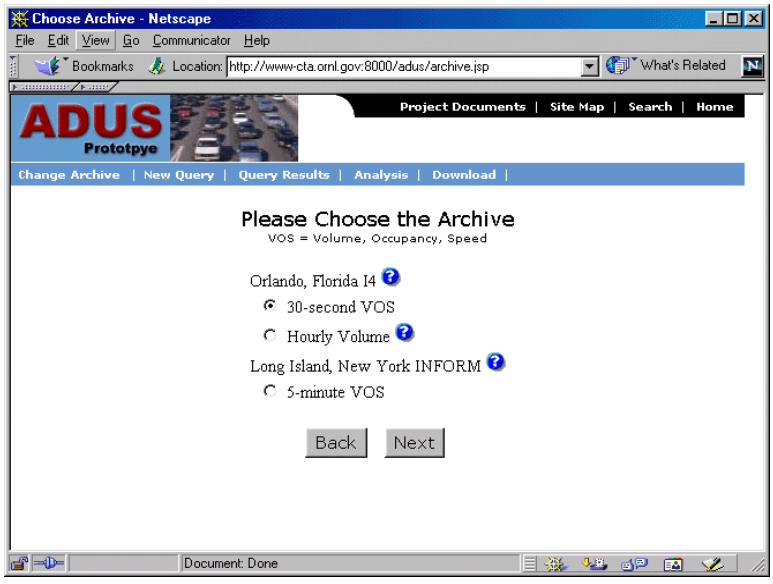
Figure 9.3. An Example of the Meta-Data Screen



Web-Based Tool to Access and Analyze Archived Data

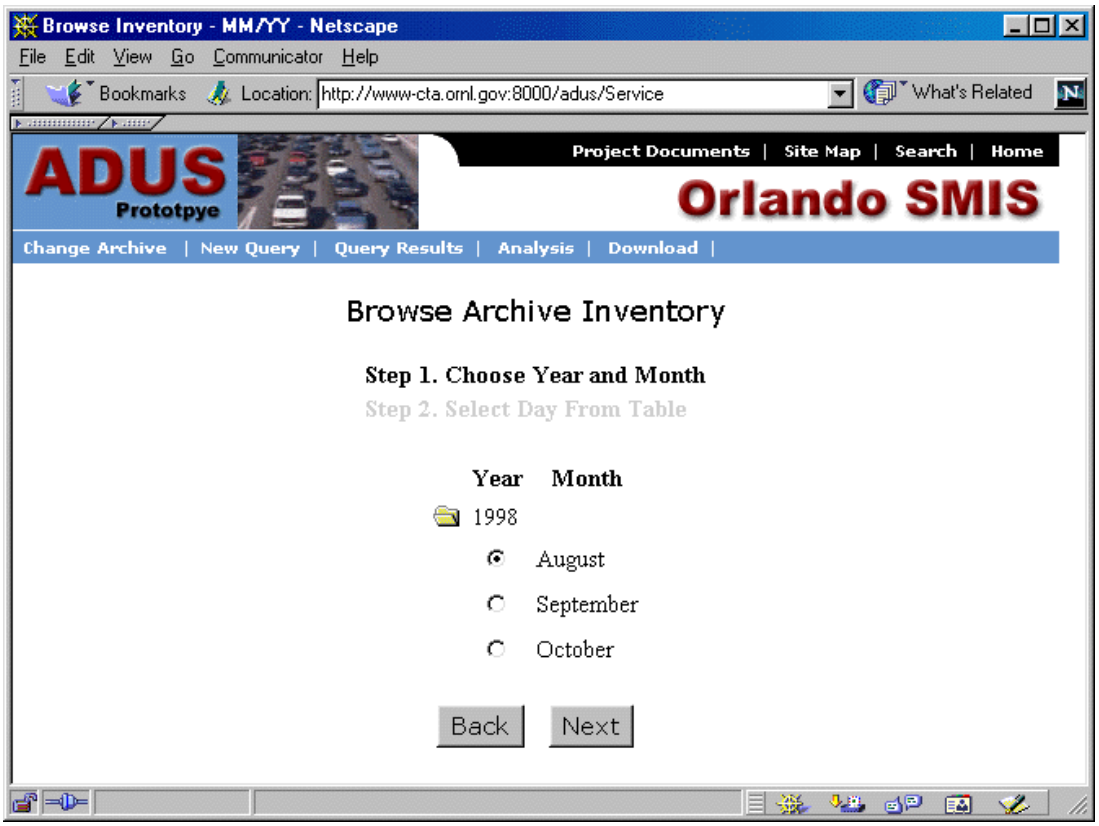
A web-based tool was developed to allow users to access the archived ITS data. There are two parts to the system. The first part has a description of this study, presents the two demonstration sites, and provides links to pertinent sites. The archived ITS data and the web-based tool are in the second part – labeled "Archives." Note that this system is a prototype developed for demonstration purposes and should not be viewed as a complete system. Five tools are available on this web site:

1. **Change Archive**. One can choose data from any one of the three data archives, New York's INFORM, or Florida's SMIS (30-second counts or hourly counts)
2. **New Query**. This allows the user to query the archived data.
3. **Query Results.** Query results are stored in a "shopping cart."
4. **Analysis.** Once the user identifies the data of interest, two plotting functions are available in this module.
5. **Download.** Once the user identifies the data of interest, these data can be downloaded into the user's computer.

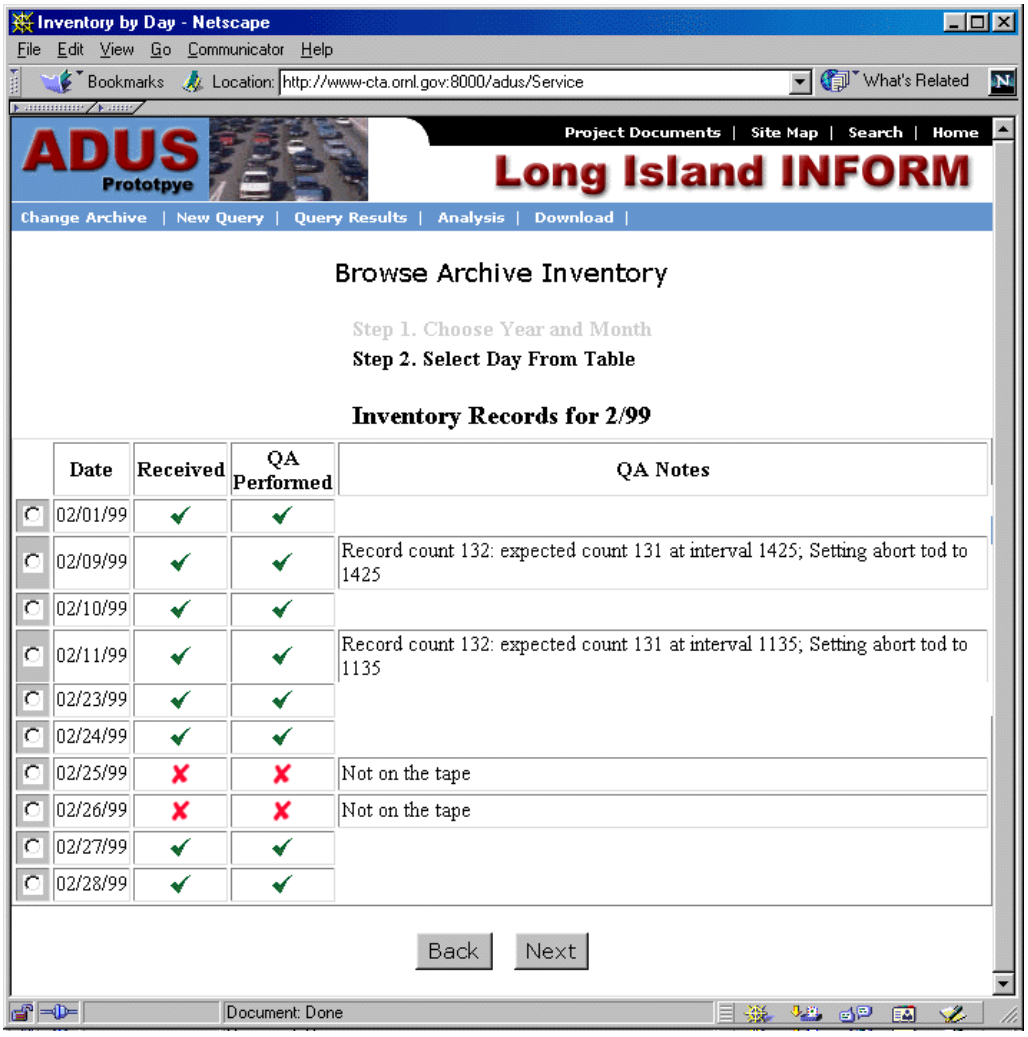


The user can choose from three data archives: 30-second or hourly traffic counts from Florida's SMIS, or 5-minute data from New York's INFORM. Once one of the data archives is chosen, the user is asked to select the time frame for the archived data. Two options are given to do that. One can browse the data inventory or specify the time frame in the format of mm/dd/yy.

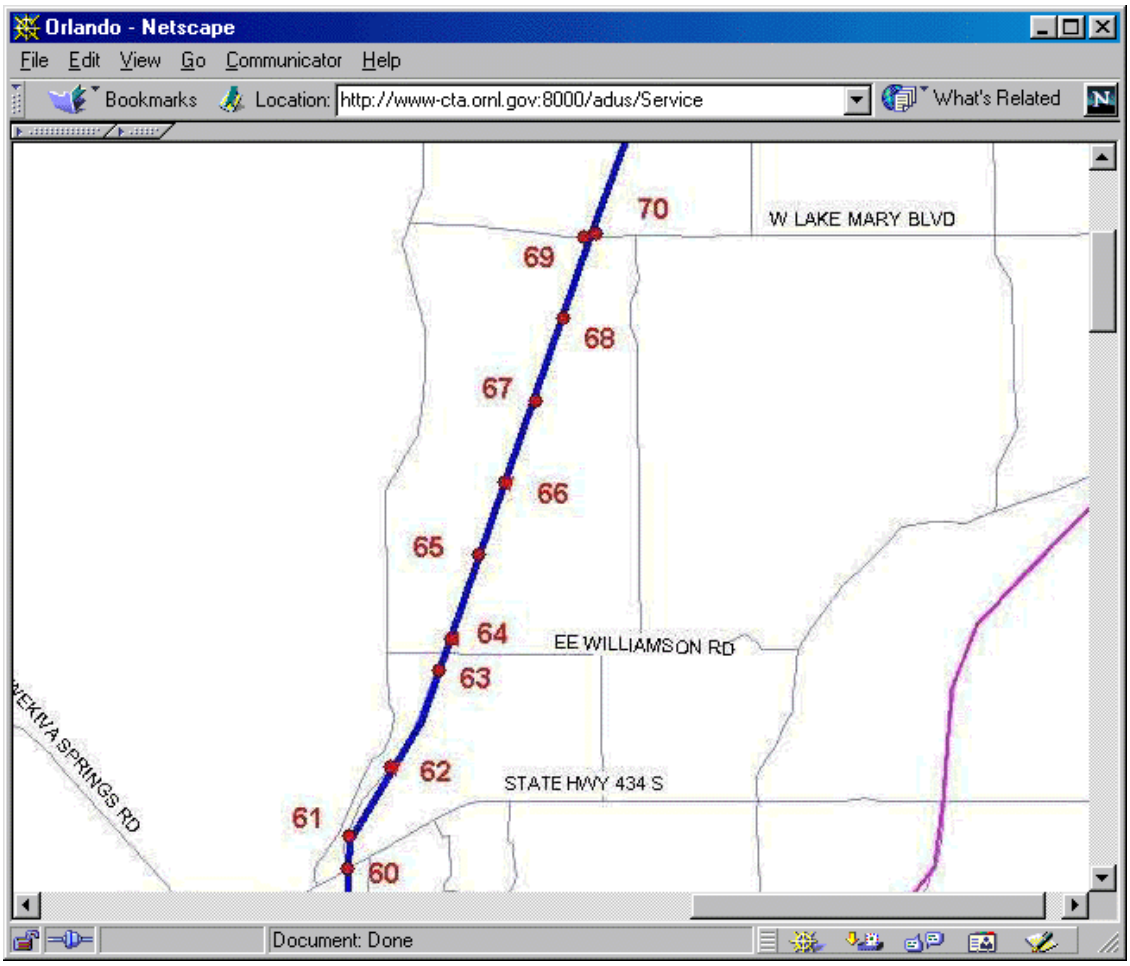
Should one choose to browse the data inventory, the system will display the time frame for which archived traffic count data are available. In the case of Florida's SMIS, data for only three months are archived.



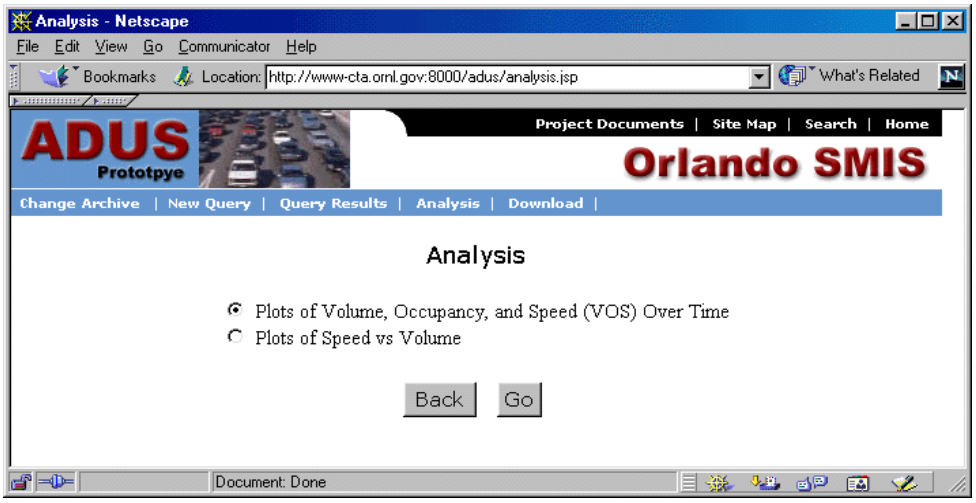
Once a time period is selected, the system summarizes an inventory report of the selected time period. The inventory report includes the information on the day of the month for which archived data are available, whether QA procedures have been performed on the data, and the results of the QA procedures. For example, the data check procedure found that no traffic count data were available on February 25, 1999.

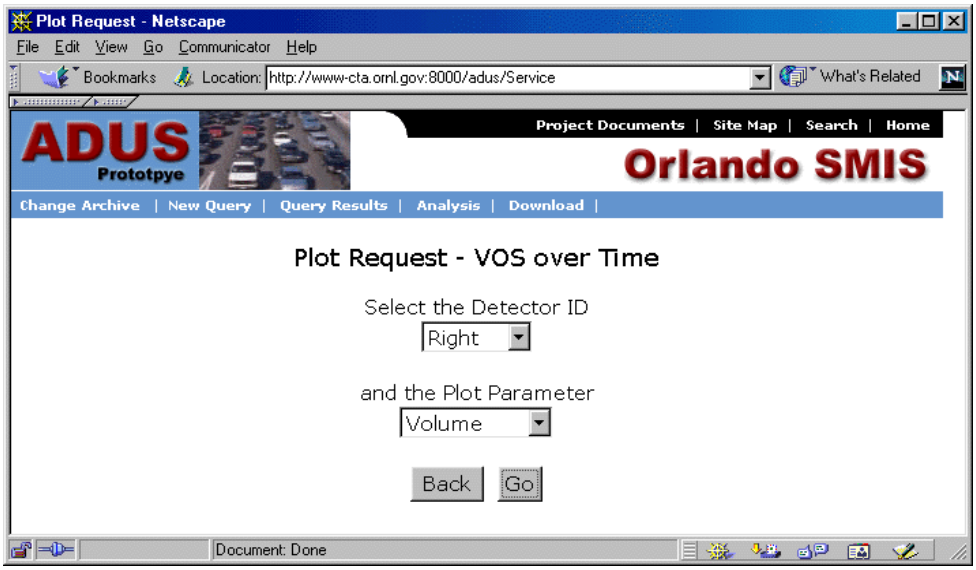


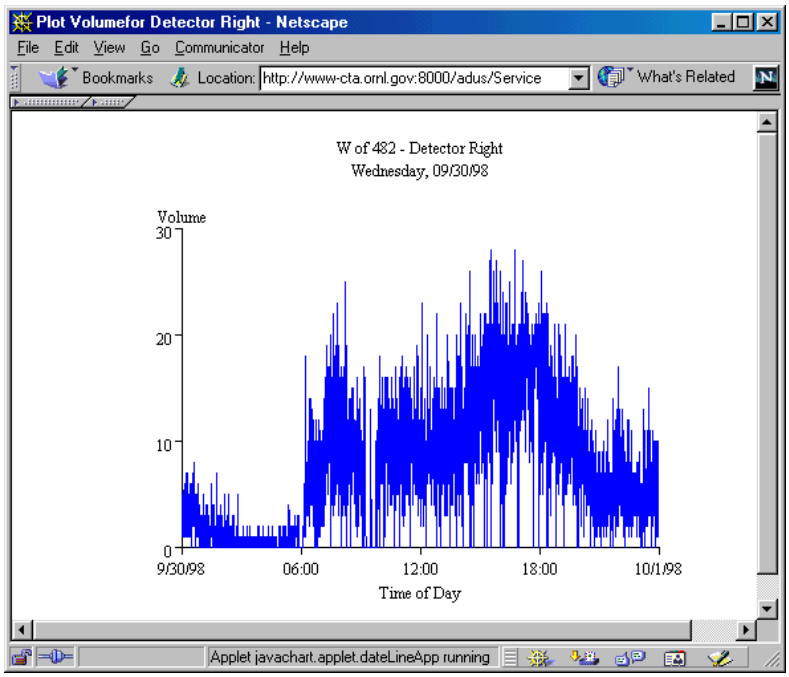
After selecting a date on the archive inventory, the user is prompted to select a location where archived data are available. Three options are available to select a location: (1) browse station data, (2) "point and click" from a map, and (3) identify station ID. The last option can only be useful to those who are extremely familiar with the ITS deployment. Should one decide to choose a station from the map, a screen depicting station locations will appear. This following example illustrates a concentration of stations on Florida's SMIS network. Each of the dots is "clickable." If one clicks on a station, a summary report (meta data) will appear on the archived data for the station (Figure 9.3).



At this point in the example, the user has requested the archived data for a specific station during a specified time frame. These identified data can be graphically display and/or downloaded by clicking either the "*Analysis*" or the "*Download*" button on the button bar. The "Analysis" function allows one to plot one of the three data elements (i.e., Volume, Lane Occupancy, and Speed) and a cross plot of volume and speed. Plots can also be requested for different lanes (denoted by Detector ID).

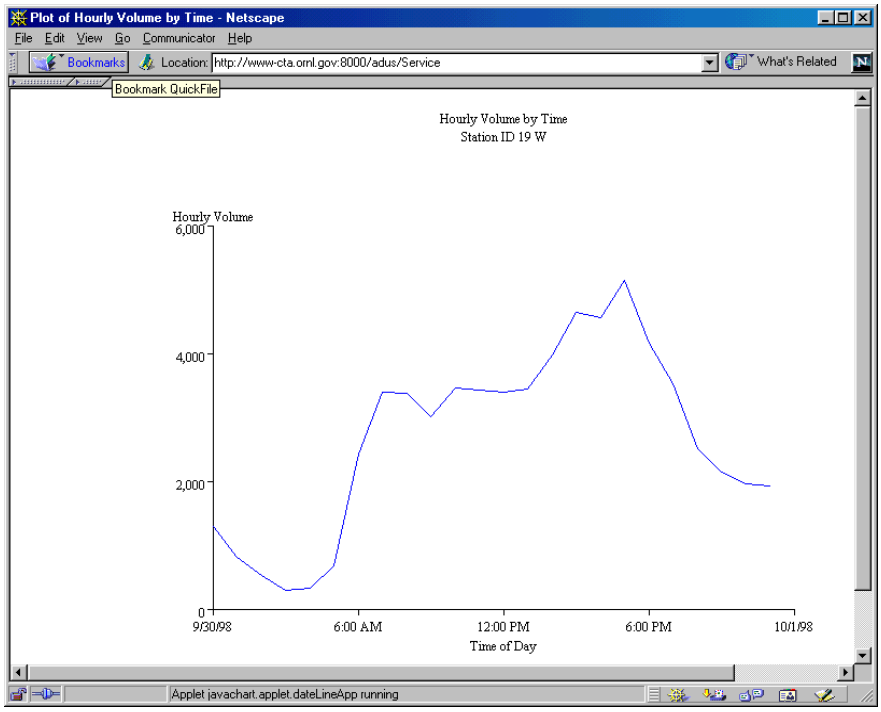




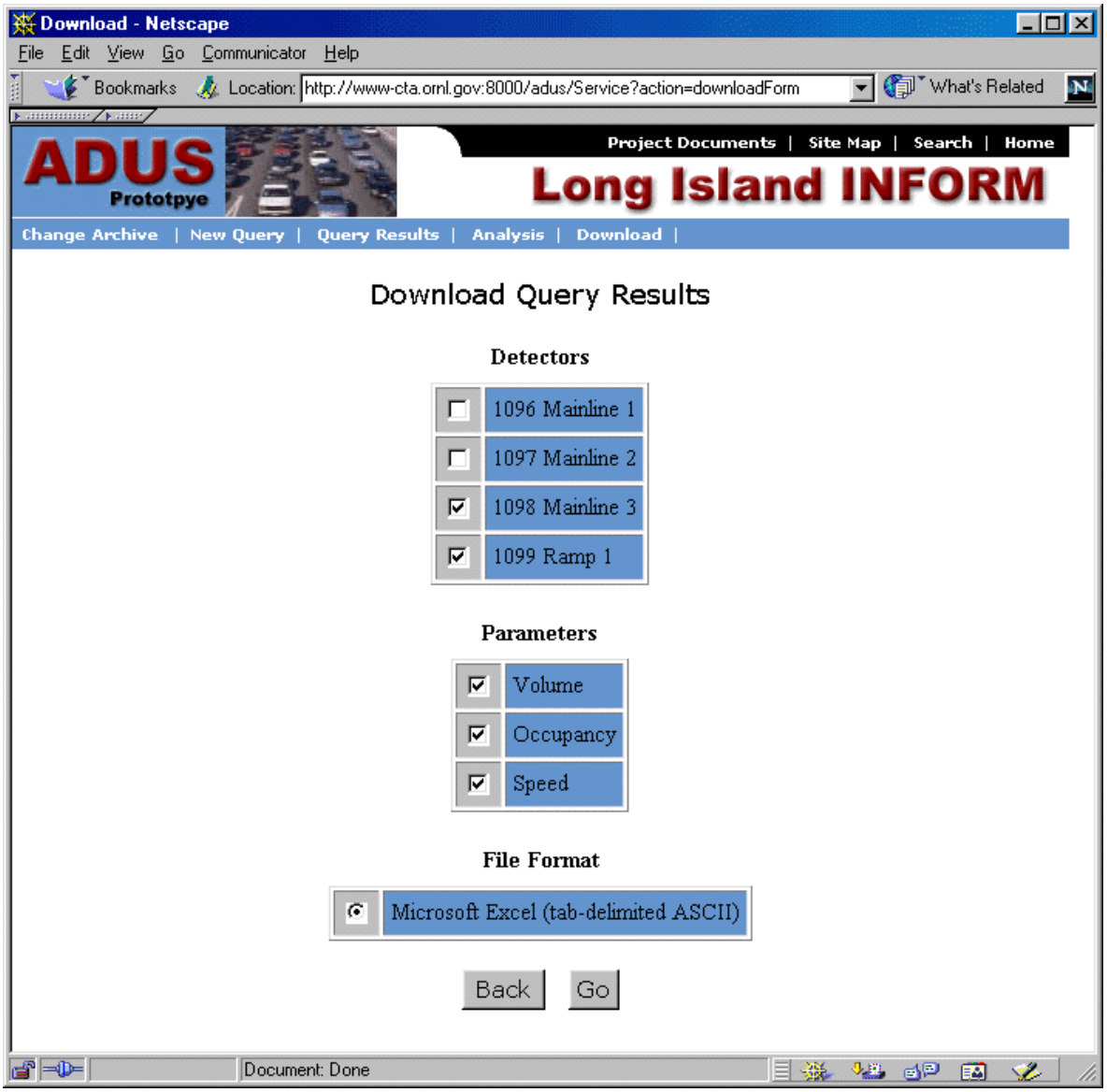


A Plot of the INFORM **5-minute** Traffic Volume Data

A Plot of the INFORM **Hourly** Traffic Volume Data



The specified data can be downloaded to the user's computer in Microsoft Excel format, more specifically, in tab-delimited ASCII format. An example of a downloaded file is included.



An example of a downloaded file (in Microsoft Excel format), with column headings of Station ID, Date, Time, Detector ID and Data Elements (V, O or S)

Microsoft Excel - Service.xls

File Edit View Insert Format Tools Data Window Help

A1 = Station ID

	A	B	C	D	E	F	G	H
1	Station ID	Date	Time	1096-v	1096-s	1097-v	1097-s	
2	21	2/17/99	0:05:00	91	83	84	67	
3	21	2/17/99	0:10:00	82	87	78	67	
4	21	2/17/99	0:15:00	79	80	92	68	
5	21	2/17/99	0:20:00	69	93	86	67	
6	21	2/17/99	0:25:00	40	93	49	81	
7	21	2/17/99	0:30:00	55	86	60	69	
8	21	2/17/99	0:35:00	51	86	76	74	
9	21	2/17/99	0:40:00	9	59	16	57	
10	21	2/17/99	0:45:00	32	93	48	71	
11	21	2/17/99	0:50:00	42	90	56	69	
12	21	2/17/99	0:55:00	31	90	52	73	
13	21	2/17/99	1:00:00	20	72	30	75	
14	21	2/17/99	1:05:00	16	92	35	99	
15	21	2/17/99	1:10:00	24	81	47	71	
16	21	2/17/99	1:15:00	30	90	55	61	
17	21	2/17/99	1:20:00	29	92	41	67	
18	21	2/17/99	1:25:00	14	65	22	59	
19	21	2/17/99	1:30:00	6	64	8	60	
20	21	2/17/99	1:35:00	4	62	21	78	
21	21	2/17/99	1:40:00	24	86	41	65	

SECTION 10. SUMMARY AND RECOMMENDATIONS FOR FUTURE RESEARCH

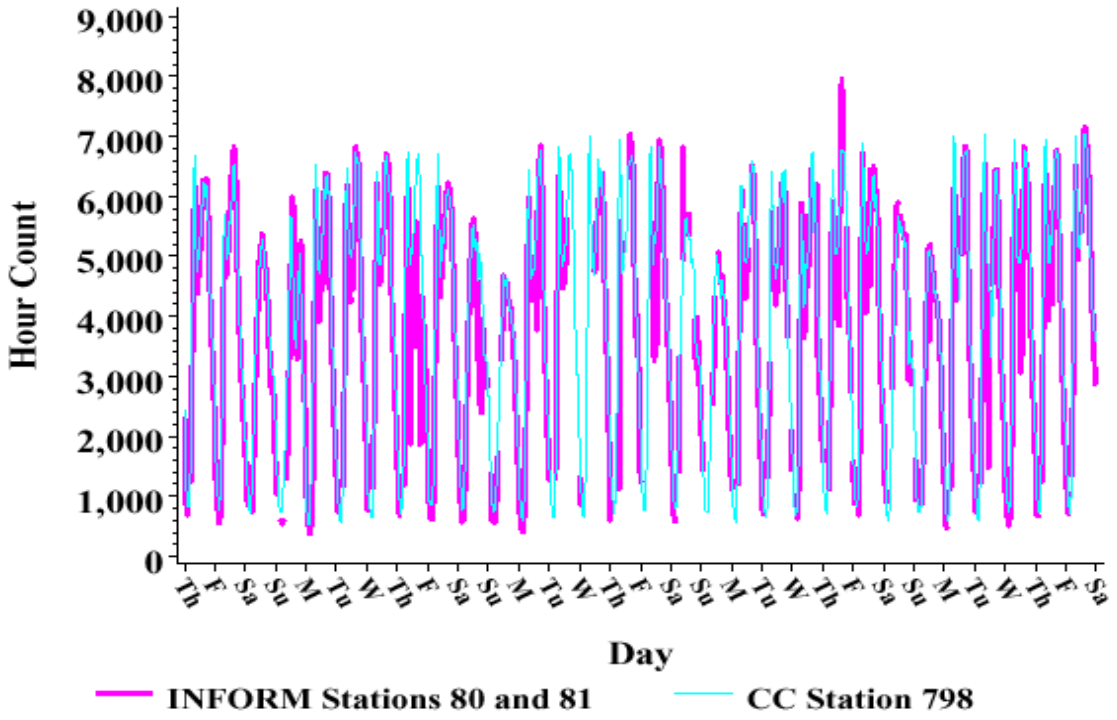
10.1 Benefits in Using ITS-Generated Data

If ITS-generated data are archived, then our analysis has clearly demonstrated the potential of these archived data. Specifically, we demonstrated the benefits of using ITS loop detector data in meeting the information needs of traffic monitoring programs. The specific benefits can be gauged in three ways. First, can ITS data replace data collected from nearby traditional count stations? Second, can ITS data supplement the traditional data so that more reliable estimates can be developed? Third, can ITS data be used to calculate adjustment factors that are more reliable than those calculated based on a limited number of continuous counters? Conceivably, the ITS-generated traffic counts data can significantly improve the traffic monitoring programs in urban areas . This question was not addressed in this study.

The similarity between ITS loop detector data and traditional count data suggests that, from a data-sharing perspective, ITS counters could *replace* the nearby conventional counters (as illustrated in Figure 10.1). However, from an operations perspective, it might not be appropriate to suggest such replacement. "Turning off" an existing continuous counter probably might be more involved than continuing its operation. Thus, the answer to the question – Can ITS data replace data collected from nearby traditional count stations? – the answer is "perhaps, but only after a more complete assessment is made of other reasons for having these stations in operation and of the cost of discontinuing their use."

That said, the similarity between the two data sources offers an opportunity for using ITS-generated data. Data from nearby ITS counters can be used to supplement missing or questionable traffic counts resulting from equipment malfunction or other reasons. Like ITS counters, continuous counters experience significant problems in terms of missing and questionable data. Given the size of the missing data problem experienced by conventional counters, this use of ITS data would

Figure 10.1 Traffic Count Comparison Between ITS and Conventional Counter



probably be more valuable than the notion of replacing the conventional counters. After their missing and questionable data are corrected, traffic estimates (such as the estimated annual average daily travel (AADT)) derived from conventional counters will be significantly more reliable. Thus, the answer to the second question – Can ITS data supplement traditional data so that more reliable estimates can be achieved? – is "yes."

Another value-added aspect of ITS data is their use to develop adjustment factors that are more representative of traffic patterns in the surrounding roads. Figure 10.2 shows that there is considerable variation in the adjustment factors. Adjustment factors calculated from nearby counts are nevertheless correlated. Data from the one conventional count station yielded an adjustment factor of 1.13. If this adjustment factor had been used to adjust the short-term counts collected from stations in the 39 instrumented miles, then the AADT for this 39 mile stretch would probably have been over-estimated by at least 10%. Thus the answer to the third question – Can ITS data be used to calculate adjustment factors that are more reliable than those calculated based on a limited number of continuous counters? – is "yes."

In another stretch of the network, the single adjustment factor could very well under-estimate AADT. In general, ITS-generated data provide a greater number, and thus a more precise and accurate set of adjustment factors.

One can envision numerous other benefits of using ITS-generated traffic data. For example, these data offer the opportunity to more accurately monitor traffic patterns in urban streets.

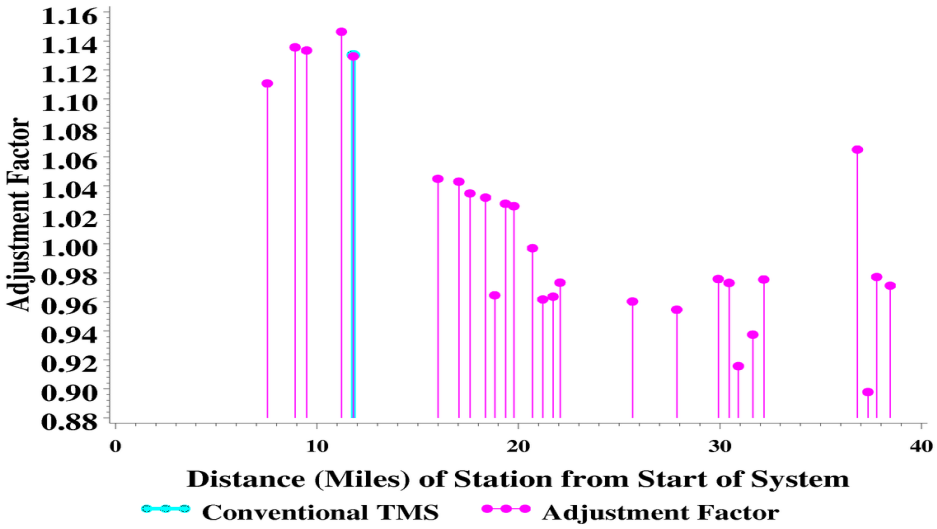
10.2 Barriers to Using ITS-Generated Data

However, there are a number of barriers that impede the realization of these benefits. They are: (1) the institutional considerations that deter data sharing, (2) the lack of standardization in archiving data that reduces the ease of using the archived data, (3) data quality issues that multiply the complexity of using the data, and (4) legacy systems that are intolerant to accepting new data.

Institutional Considerations

Although the benefits of ADUS are recognized, data sharing among jurisdictions will continue to hinder full ADUS implementation. This concern can be overcome if the "mechanics" of transmitting the data to data *users* become more "transparent" to data *providers* (in this case, the

Figure 10.2
A Comparison of Adjustment Factors Calculated from ITS Data and Conventional Count Data



TMCs). Many TMCs archive their data to some extent, but it will be incumbent on data users to develop systems to "pull" or "fetch" the data, rather than the data providers "push" the data to users.

The institutional concerns become a completely different challenge when data sharing carries the risk of violating individuals' privacy. This is a topic outside the scope of this project and will not be elaborated here.

Lack of Standardization

The lack of standardization has hindered progress in many areas. A few examples of such barriers were observed in this project. Although an efficient and compact way to store data, the 16-bit protocol can present technical challenges to users. This is because the 16-bit binary data might

be recorded differently on different computers (e.g., little-endian vs big-endian, which indicates which of the 2 bytes comes first).

Another example is the changes in sensor identification numbers (sensor ID) as described in Section 3. Although this type of change is documented, the information is not integrated with the data file. The implication of not having this type of information integrated with the traffic data is that it is almost impossible to develop a traffic profile over time.

Recognizing the need to develop standards, an effort is underway[11]. This effort has identified the following as the most crucial elements for initial development: (1) General guidelines and principles for archiving all forms of ITS-generated data, and (2) Specifications for archiving ITS-generated travel (or traffic) monitoring data.

From the perspective of standardization, this project is rather straightforward in that it only focuses on traffic monitoring data. When ADUS advances to the point of integrating different information from multiple sources, then it will be all the more important to develop standards so as to realize the benefits of archiving and sharing ITS-generated data.

Data and Data Quality Concerns

It is obvious from our analysis that the cost of using ITS data to meet the information needs of traffic monitoring is significant, particularly in terms of data "preparation." This data preparation is extensive and includes checking data quality, identifying and correcting questionable data, imputing missing data, and formatting data to a format that can be "plugged" into the existing

[11] *"ITS Standards Development Support Project Plan - Archived Data User Service: Guidelines for Archiving ITS-Generated Data and Specifications for Archiving Travel Monitoring Data*." American Society for Testing and Materials. July 2000.

software. That said, in the long run, these data preparation systems can be developed more efficiently and the benefits of using ITS data are likely to outweigh the costs needed to overcome these issues. Sharing lessons learned and best practices is probably the fastest way to reach that point where the benefits surpass the costs. The ADUS five-year strategic plan[12] aims at that goal by encouraging, among other activities, demonstration projects.

Legacy Systems

Typically, it is often difficult and not recommended to "retrofit" legacy systems so that data can be archived in accessible media and transmitted via the internet. In this project, INFORM data were transmitted through mail on magnetic tapes which were difficult to read because a compatible tape drive was difficult to find. Presumably, obstacles that are the result of legacy systems will decrease over time as more and more obsolete systems are replaced by current technologies.

10.3 Summary

Although the costs are high at this point to use ITS-generated data to "*improve transportation decisions through the archiving and sharing of ITS generated data*," this project has proven the concept that ITS-generated data can indeed improve transportation decisions by, in this particular case, improving traffic estimates. As more and more ITS is deployed in the future, ITS-generated data can no doubt replace many of the current data collection processes. However, before then, the greatest contribution that ITS data can offer is probably when they are integrated with, or used to supplement, traditional non-ITS data to address gaps in these data.

Despite the promise of ITS-generated data, it is imperative to caution the user community that:

1. ITS-generated data should not be viewed as a "silver bullet" for addressing data gaps,
2. It can be extremely misleading to assume that ITS data are "ready-to-use," and

---

[12] *"ITS Data Archiving: Five-Year Program Description*." U. S. Department of Transportation. March 2000.

3. Extensive and thorough data quality checks are essential and can be extremely demanding.

It is important to point out that the costs of using ITS-generated data should decrease considerably as their use increase. More ADUS activities will be cultivated by fostering different types of demonstration projects, sharing best practices and lessons learned, making data "preparation" software and tools accessible and easy to use, and quantifying and demonstrating the benefits of ADUS. At that point, the costs could very well become inconsequential compared to the benefits.

APPENDIX 1. DATA AGGREGATION PROCEDURE FOR TRUNCATED DISTRIBUTION

When the process-change test indicated that there was **not** a process change on the five-minute interval, the test about the frequency of intermittent zeros was performed. If the intermittent zeros test indicated that the number of zeros is acceptable, then, in accordance with assumptions (1) and (2), counts were aggregated to five-minute totals by multiplying the usual mean and standard error by ten. Because there is no evidence of a process change, these standard errors should be approximately valid. If, however, the intermittent zeros test indicated that some of the zeros are likely to be codes for malfunctions, then all of the zeros on the five-minute interval were set to missing. Because, in this procedure, any valid zeros were also set to missing, this implies that the remaining nonzero observations are from a distribution truncated to exclude zeros.

For observations from a distribution truncated to exclude zeros, simply taking their average and scaling to a five-minute estimate would ignore that zeros were missing and would lead to an estimate that is biased high. Although this bias may or may not have an appreciable effect on statistics computed from the counts (e.g., hour totals), trying to correct for it is more reasonable than trying to analyze it. To correct for the bias, we use an adjustment, again based on the Poisson distribution: If the underlying traffic counts are from a Poisson distribution, then, after removing the zeros, the observed counts are from a **truncated** Poisson distribution:

$$P(X=x) = \frac{1}{e^{\lambda}-1}\frac{\lambda^{x}}{x!} \qquad x=1,\ 2,\ 3...,$$

where $\lambda>0$ is the mean of the corresponding complete (i.e., not truncated) Poisson distribution. It is the mean of the **complete** distribution that we would like to know. By making the assumption that the counts are Poisson (and by Assumptions (1) and (2)), the method of maximum likelihood can be used to estimate the mean of the complete distribution, even when the observed counts are truncated. For $n \leq 10$ truncated Poisson thirty-second counts, $x_1,...,x_n (x_i \geq 1)$, on a five-minute interval, the log of the likelihood is

$$-n\log(e^{\lambda}-1)+\sum_{i=1}^{n} x_i \log(\lambda)-\sum_{i=1}^{n} \log(x_i!),$$

and the derivative with respect to λ of the log likelihood is

$$\frac{1}{\lambda\,(e^{\lambda}-1)}\,[-n\lambda e^{\lambda}+(e^{\lambda}-1)\sum_{i=1}^{n} x_i\,]. \qquad (1)$$

The sign of the derivative is determined by the expression inside the square brackets in (1). The expression inside the brackets is 0 at $\lambda = 0$, approaches $-\infty$ as λ approaches ∞, and has derivative which is negative if and only if $\lambda > \sum x_i / n - 1$. Note that $\sum x_i \geq n$. If $\sum x_i = n$, then (2) is negative for all $\lambda > 0$, and the maximum likelihood estimate (MLE) is 0. If $\sum x_i > n$, then (2) is positive for $\lambda < \sum x_i / n - 1$ and negative for $\lambda > \sum x_i / n - 1$. Therefore the expression inside the square brackets in (1) is, as λ increases, positive and then negative, and the same is true about (1) itself. This shows that the log of the likelihood is pseudoconcave, and hence that the MLE occurs where its derivative (1) is zero. Thus, by using a numerical

procedure such as Newton's method (what we used) to solve for a λ for which (1) is zero, the MLE can be determined.

When the zeros test indicates that zeros likely denote malfunctions, we used this approach to compute an adjusted estimate of the mean thirty-second count for the five-minute interval. This truncated Poisson mean estimate was then multiplied by ten to estimate the total count for the five-minute interval.

APPENDIX 2. THE ADUS PROTOTYPE ARCHITECTURE

The first step in developing the prototype archive is data acquisition. Data acquisition is often a multi-step process. For example, for INFORM, the Operational Data Control (ODC) activities described in the architecture occur on VAX where volume, land occupancy, and speed data *(VOS)* are written to disk every 5 minutes. Once a month, the data that are two months old are written to magnetic tape. Copies of these tapes were sent to ORNL. This appendix describes these transfer procedures and formats.

An automated procedure was developed by ORNL to perform data transfers from ITS facilities to the archive. This procedure provides the high levels of integration and management required by the National ITS Architecture (e.g., administrator interface and data integrity). Note that to transfer archive data from the ODC to users using magnetic tape and customized formats may present formidable technical barriers to use. For example, the user must have a compatible tape drive, have expertise in its operation, and write custom software to read and extract the desired data.

After acquisition, the Data Import and Verification (DIV) function is performed automatically by the archive. Again using the INFORM as an example, the incoming data is stored in its original format, which is very efficient and has a daily organization. An "ingest" program is executed that verifies the integrity and consistency of the daily files; metadata describing these observations are written into a database table. Inventory records are created to describe data locations, time intervals, and quality. Next, programs that compute derived archive data are executed (e.g., hourly totals of volume).

There are several implications of the fact that the ITS system configuration is defined on a daily basis and will change over time (e.g., new loop construction). The Operational Data Control (ODC) must incorporate into their archives and distributions the configuration information. There has to be a method of a change notification and a protocol for communicating the new configuration or a change in configuration. For some facilities, this may be a new responsibility (and potentially an administrative barrier). Consider, for example, the New York INFORM: while the daily data set contains the configuration file, more information than that is needed. The configuration file contains only the station and loop identifiers. Another file, the station-level metadata (e.g., lane configuration) is also needed. This aforementioned data has to be maintained, tracked, and distributed. Ideally, the configuration will change at midnight when the daily archive file is initialized. However, it is likely that operations staff will prefer to install and test a configuration during working hours. Therefore, configuration tracking must be able to accommodate multiple configurations per day; a messy task for archive users.

Currently, INFORM configuration tracking does not have the granularity that is needed. To compensate, the Data Import and Verification (DIV) function must look for evidence of configuration change and, if found, dispose of the data from change to midnight; this is recorded in the metadata and used by the query engine. Another implication of configuration changes is that user queries and other data manipulation routines must correctly manage the data, including potentially complex user interface to the queries. A user interface that helps the user browse the inventory and construct a query must deal with the complexity of configuration management. The University of Washington has developed the Self Describing Data (SDD), a data transfer method that puts the data dictionary on the wire with the data in a way that is portable and extendable. The SDD was used for real-time data

transfers of raw archive data over a leased line (Dailey, 1999). Similar methodologies are being developed for e-commerce and other data-sharing initiatives by using the Extensible Markup Language (XML).

The prototype ADUS implements the Automatic Data Historical Archive (ADHA) using a relational database management system (RDBMS). As described above, the raw loop data is stored in its native format. The RDBMS is used to store metadata and derived products (e.g., hourly volumes). The RDBMS has an application programming interface (API) that allows applications to query and manipulate the data. In addition, the RDBMS provides the data administration methods described above as industry best practives (e.g., backup, privacy, and security).

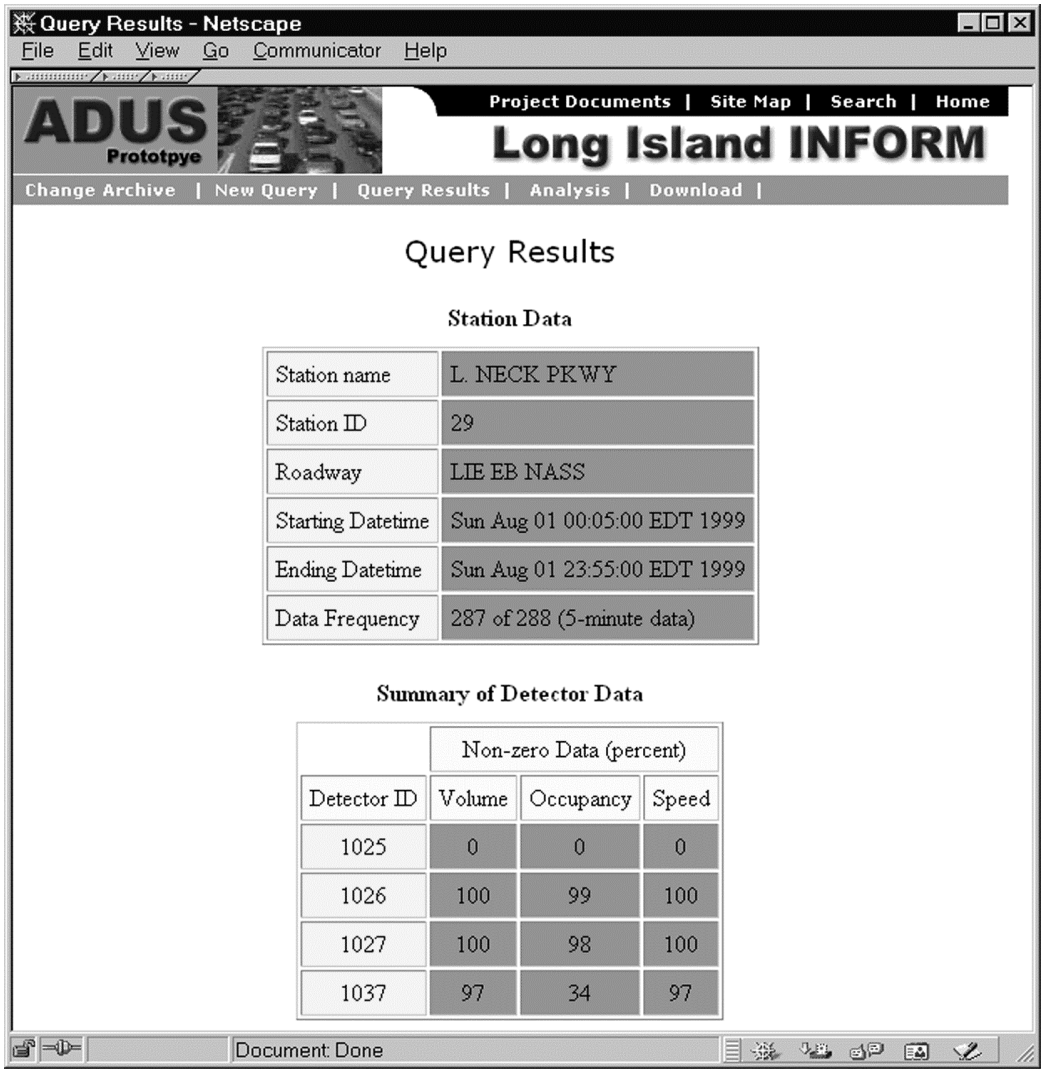
The next section describes the Data Warehouse Distribution (DWD) function and ITS Community Interface of the ADUS prototype.

Archive Products and User Interfaces

Figure A2.1 shows an example screen of the ADUS prototype web site with buttons for the following archive functions: change archive, new query, view query results, query analysis, and download. In addition, the web page provides links to archive documentation and site navigation aids including web site search and site map.

The example web page shows the results of a query of the Long Island INFORM loop data for a selected day and location. The second table in the results screen shows that station 29 at Long Neck Parkway has four lanes; on this particular day, one of the lanes has no data while the other three have most of the data for volume, occupancy, and speed.

Figure A2.1 An Example of the ADUS Prototype System Web Page



Development and execution of a query is achieved by working with a sequence of screens, sometimes called a "wizard" or a multi-step form. A query consists of the following steps and options:

1. Select Archive
    1. Long Island 5-minute VOS
    2. Orlando 30-second VOS
    3. Orlando hourly volume
2. Select Time Interval
    1. Browse inventory (year / month / date drill-down)
    2. Enter date
3. Select Location
    1. Browse station data (roadway / station drill-down)
    2. Map
    3. Enter station ID
4. Summary of Query Results
    1. Station description
    2. Number of records
    3. Number of records with non-zero data
5. Working with Query Results
    1. Plots
    2. Download

Note that the configuration of loops and stations changes over time (e.g., due to road construction); therefore, the query determines the time interval of interest, then the location(s). Data on the archive's inventory holdings allows the user to browse the database to determine what is available. Also, a map is provided that allows the user to select the roadway; the map is an image generated with a geographical information system (GIS) containing vector data for the road coverage. On subsequent queries, the system remembers query parameters so the user does not have to re-enter them.

Queries execute quickly and the results are available for interactive analysis. Query results are remembered by the archive system; much like the concept of a "shopping cart." To quickly visualize the characteristics of the data, the archive provides plot services, such as loop data over time and volume versus speed. These are delivered to the user as a Java applets; the same code, after a small change, could generate the image on the server and then send that to the user's browser.

A very flexible download service is provided which allows the user to choose individual lanes and parameters (i.e., VOS). The data is delivered as a Microsoft Excel spreadsheet.

The same user interface is used for all archives and data types. The archive software uses metadata to guide operations and the web pages are dynamically generated.

Hardware and Software Components of the Prototype

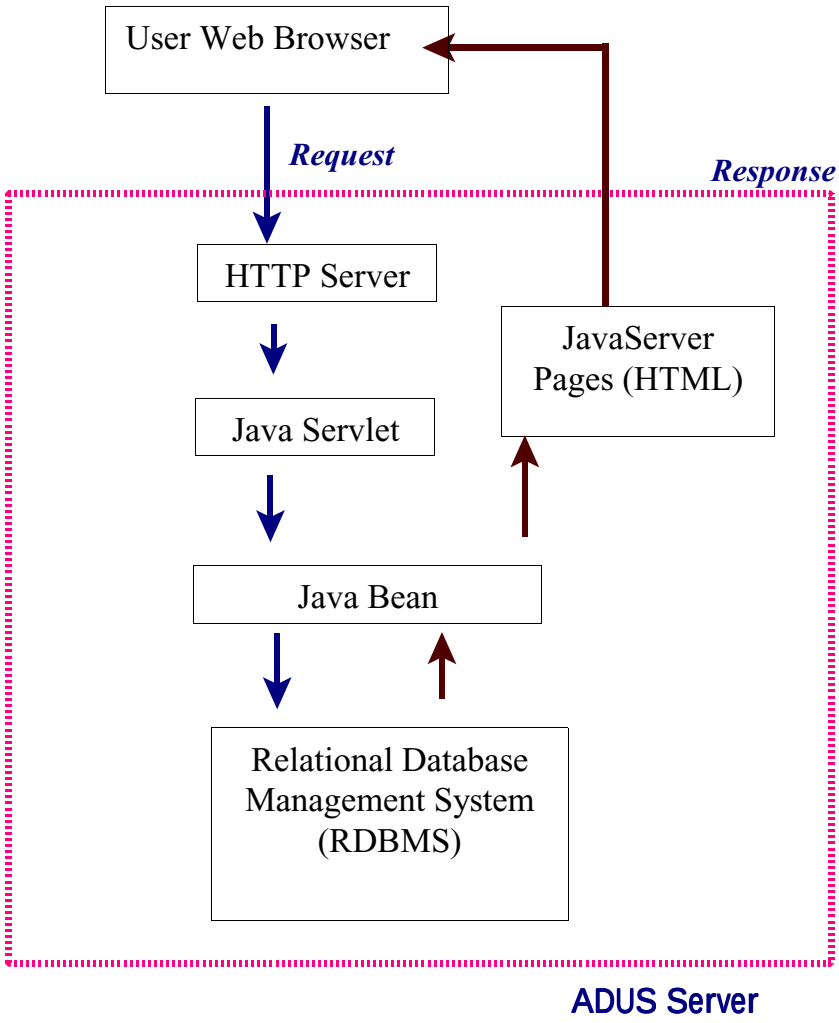
The prototype ADUS server was developed using the following components:

- Standard PC hardware,
- Microsoft Windows NT operating system,
- Apache web server,
- Apache Tomcat Java servlet engine,
- MySQL database engine,
- Statistical Analysis System (SAS), and
- A Java plotting package for server-side and client-side generation of visualizations.

These system components are available at a reasonable cost, are reliable, and provide good performance for mid-sized archives.

Figure A2.2 shows the sequence of operations for a web browser request and response. Java servlets are used for server-side programming. JavaBeans and JavaServer Pages (JSP) technologies are used to separate the logic of archive operation (e.g., execution of queries) from the presentation (HTML). Well-designed system components are easier to extend to other tasks and are easier to maintain. For example, the same module could be used to provide a service to a browser and to assist the archive administrator with supporting operations (e.g., QA/QC). All of the Java codes use Java Database Connectivity (JDBC) for database functions; therefore, the Java codes are portable and database independent.

Figure A2.2 Sequence of ADUS Prototype Operations



Obstacles to Development of the Prototype and Lessons Learned

- Lack of documentation at the ITS facility. Counter intuitive for real-time data systems. Very difficult and/or expensive to develop an understanding of the nature of the data and what to do.
- Shortage of staff time at the ITS facility. Probably due to financial realities, new construction, inherent difficulties of managing a high-visibility data resource. New construction and upgrades.
- Volume of data. Only in transfer, storage and processing is not a problem, even for a low-cost ADUS.
- Data quality. See Section 4. Communications and other hardware problems. Improvements due to upgrading from coax to fiber. Coding missing data as zero.
- ITS software may be one-of-a-kind and poorly documented.
- Loss of staff may result in loss of "corporate knowledge."
- Binary data is hard to work with.
- System configuration changes over time.
- Trust and acceptance of the data.
- Integration with existing data management and analysis operations.