WORK ORDER NUMBER BAT 03-007

TRAFFIC DATA QUALITY MEASUREMENT

FINAL REPORT

To

Office of Highway Policy Information
Federal Highway Administration
U.S. Department of Transportation
Washington, D.C.

September 15, 2004

Battelle
The Business of Innovation
505 King Avenue
Columbus, Ohio 43201

In Association with

Cambridge Systematics Inc.
Texas Transportation Institute

Disclaimer

Acknowledgments

The authors gratefully acknowledge the support and guidance of Mr. Ralph Gillmann of the Office of Highway Policy Information of Federal Highway Administration throughout this project. His valuable review comments of the concepts and methodologies are acknowledged.

The authors also acknowledge the valuable review comments by the following individuals and offices of the Federal Highway Administration (FHWA):

- Mr. Richard Reel, Florida Department of Transportation
- Mr. Mark Demidovich, Georgia Department of Transportation
- Mr. John Rosen, Washington State Department of Transportation
- Ms. Jonette Kreideweis, Minnesota Department of Transportation
- Mr. Joe Avis, California Department of Transportation
- Mr. David Gardner, Ohio Department of Transportation
- Ms. Raelene Viste, Idaho Department of Transportation
- Dr. Mark Hallenbeck, University of Washington
- FHWA’s Traffic Monitoring and Standards Division
- FHWA’s HPMS Division
- FHWA’s Office of Safety on Highway Safety

The authors also acknowledge the assistance of Mr. Brian Burk of Texas Department of Transportation, Ms. Terri Johnson and Mr. John Collins of Mobility Technologies, Inc, and Mr. David Garner of Ohio Department of Transportation in providing data for the case studies included in this report.

Table of Contents

Table of Contents (Continued)

List of Appendices

Table of Contents (Continued)

Page

List of Tables

Table of Contents (Continued)

Page

List of Figures

List of Acronyms

AADT	Average Annual Daily Traffic
AASHTO	American Association of State Highway Transportation Officials
ADMS	Archived Data Management Systems
ADT	Average Daily Traffic
ASTM	American Society for Testing and Materials
ATIS	Advanced Traveler Information System
ATMS	Advanced Traffic Management Systems
ATR	Automatic Traffic Recorder
AVC	Automatic Vehicle Classifier
Caltrans	California Department of Transportation
CAMPO	Capital Area Metropolitan Planning Organization
CMS	Congestion Management System
CORSIM	Corridor Simulation
CSV	Comma-separated value
DOT	Department of Transportation
DVMT	Daily Vehicle Miles Traveled
EDL	Electronic Document Library
EPA	Environmental Protection Agency
FARS	Fatal Accident Reporting System
FDOT	Florida Department of Transportation
FGDC	Federal Geographic Data Committee
FHWA	Federal Highway Administration
FTP	File Transfer Protocol
GDOT	Georgia Department of Transportation
GIS	Geographic Information System
GVW	Gross Vehicle Weight
HERS	Highway Economics Requirements System
HPMS	Highway Performance Monitoring System
ISDMI	Integrated Surveillance and Data Management Infrastructure
ISO	International Standards Organization
ISP	Information Service Provider
ITS	Intelligent Transportation System
LRTP	Long-range transportation plan
LVACTS	Las Vegas Area Commuter Traffic System
MAPE	Mean absolute percent error
MnDOT	Minnesota Department of Transportation
MPO	Metropolitan Planning Organization
MT	Mobility Technologies
MTC	Metropolitan Transportation Commission
MVMT	Million miles of travel
NAAQS	National Ambient Air Quality Standard
NHS	National Highway System
ODOT	Ohio Department of Transportation
PAS	Principal Arterial System

QA/QC	Quality Assurance/Quality Control
RDBMS	Relational Database Management System
RDD	Registered Data Disseminator
RMSE	Root mean squared error
RR	Route designation
RTMS	Remote Traffic Microwave Sensor
SCU	System Controller Unit
SUV	Sport Utility Vehicle
TDQ	Traffic Data Quality
TDQM	Traffic Data Quality Management
TIP	Transportation Improvement Program
TMC	Traffic Management Center
TMG	Traffic Monitoring Guide
TTI	Texas Transportation Institute
TxDOT	Texas Department of Transportation
VDOT	Virginia Department of Transportation
VMT	Vehicle-miles of travel
V/SF	Volume to service flow ratio
WIM	Weigh-in-Motion
WsDOT	Washington State Department of Transportation

Executive Summary

Introduction

One of the foremost recommendations from the FHWA sponsored workshops on Traffic Data Quality (TDQ) in 2003 was a call for "guidelines and standards for calculating data quality measures." These guidelines and standards are expected to contain methods to calculate and report the data quality measures for various applications and levels of aggregation.

The objective of this project is to develop methods and tools to enable traffic data collectors and users to determine the quality of traffic data they are providing, sharing, and using. This report presents the framework that provides methodologies for calculating the data quality metrics for different applications and illustrated with case study examples. The report also presents guidelines and standards for calculating data quality measures that are intended to address the following key traffic data quality issues:

- Defining and measuring traffic data quality
- Quantitative and qualitative metrics of traffic data quality
- Acceptable levels of quality
- Methodology for assessing traffic data quality.

Framework for Data Quality Measurement

The framework is developed based on the six recommended fundamental measures of traffic data quality. These are defined below:

- **Accuracy –** The measure or degree of agreement between a data value or set of values and a source assumed to be correct. It is also defined as a qualitative assessment of freedom from error, with a high assessment corresponding to a small error.

- **Completeness** (also referred to as availability) – The degree to which data values are present in the attributes (e.g., volume and speed are attributes of traffic) that require them. Completeness is typically described in terms of percentages or number of data values. Completeness can refer to both the temporal and spatial aspect of data quality, in the sense that completeness measures how much data is available compared to how much data should be available.

- **Validity** – The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values. Data validity can be expressed in numerous ways. One common way is to indicate the percentage of data values that either pass or fail data validity checks.

- **Timeliness –** The degree to which data values or a set of values are provided at the time required or specified. Timeliness can be expressed in absolute or relative terms.

- **Coverage –** The degree to which data values in a sample accurately represent the whole of that which is to be measured. As with other measures, coverage can be expressed in absolute or relative units.

- **Accessibility** (also referred to as usability) – The relative ease with which data can be retrieved and manipulated by data consumers to meet their needs. Accessibility can be expressed in qualitative or quantitative terms.

The framework takes into account the facts that there are different types of traffic data and different customers and users. The framework also recognizes that traffic data is used for different applications. As such, the needs and quality requirements are different for the different data customers and applications. Table ES-1 shows the range of data consumers, types of data, and possible applications that are considered in developing the framework.

Table ES-1. Types of Data Consumers and Applications

Data Consumers or Users	Type of Data	Applications or Uses
Traffic operators (of all stripes)	Original source data Archived source data	Traffic management Incident management
Archived data administrators	Original source data	Database administration
Archived data users (Planners and others)	Original source data Archived source data, Archived processed data	Analysis Planning Modeling (development and calibration)
Traffic data collectors	Original source data Archived source data	Traffic monitoring Equipment calibration Data collection planning
Information Service Providers	Original source data (real time)	Dissemination of traveler information
Travelers	Traveler information	Pre-trip planning

The framework is structured as a sequence of steps in calculating the data quality measures and assessing the quality as shown in Figure ES-1. The first step in assessing the quality of data is to determine the potential data consumers or users of the data. This is important because the type of data consumer or application determines the type of data and thus the methods of calculating the quality measures and the thresholds for evaluating the quality of data. The other steps include methods for calculating each data quality measure to allow quantitative assessment of quality, establishing acceptable quality targets, and reporting of data quality.

The application of the methods in the framework is illustrated with three case studies. The case studies are intended to only illustrate the application of the methodologies in evaluating traffic data quality. These case studies are not intended to and do not represent a review of the quality of data of the agencies providing the data for this case studies.

Step 1. Know your customers: Enumerate the consumers of your data and the type of the data they are using i.e., original source, archived, or traveler information type data).

Step 2. Select measures: Ensure that each of the 6 data quality measures are meaningful and relevant for each data consumer and the version of data they are using. Develop a matrix that shows which data consumers use which version of data.

Step 3. Set data quality targets: Set targets for each measure based on the data consumers' needs and applications.

Step 4. Calculate data quality for unique data: Calculate the data quality measures using the procedures shown for each unique version of the data in the matrix. That is, data quality changes for different version of data, so one must calculate data quality for each significantly unique data version.

Step 5. Identify data quality deficiencies: Compare the data quality results to targets and identify deficiencies in data quality. Identify and program resources to improve data quality or lower targets to be financially constrained.

Step 6. Assign responsibility and automate reporting: Automate data quality reporting and include it with metadata. Assign data quality responsibilities to data steward(s) who ensure that data problems get fixed at the root cause and not simply "scrap and rework". Provide performance incentives based on data quality levels.

Step 7. Complete the feedback cycle: Periodically reassess your data consumers, how they use your data, and the quality targets for their applications.

Figure ES-1. Structure of Data Quality Assessment Framework

Guidelines for Data Quality Measurement

The guidelines address technical issues related to the data quality standards, data sharing, estimates of the level of effort required for measuring and reporting data quality, and specifies procedures for using metadata. The guidelines include the following essential elements

Establishing Acceptable Data Quality Targets

Estimated data quality targets are provided for different applications. These targets are defined for the six data quality measures. These targets reflect the acceptable quality based on the data user's needs and applications. Depending on the user and application, data quality measures falling outside the thresholds could be unacceptable for the intended application or it could be an indication that the data ought to be used with caution. Table ES-2 shows the estimated data quality targets for some representative transportation applications. These estimates were based on experience and validated through beta testing with several FHWA offices. With regards to the accessibility measure, it is estimated that, all applications can be adequately serviced with access times in the 5 to 10 minute range.

Level of Effort Required for Traffic Data Quality Assessment

The extra costs associated with assessing and reporting data quality was considered an important issue at the regional TDQ workshops. Estimates of the level of effort are expressed in hours of labor required to implement a data quality assessment program. The estimated levels of effort do not account for the level of effort required to maintain or improve data quality. These estimates represent the level of effort required to assess the quality of existing data. Table ES-3 shows estimated levels of effort in developing and maintaining a data quality assessment system within an agency.

Specifications and Procedures for Using Metadata for Reporting Data Quality

Metadata is an extremely important consideration for data sharing in general, and especially for communicating data quality. Commonly referred to as "data about data," metadata is typically thought of as dataset descriptions. It is recommended that the ASTM standard, once approved, be used for documenting traffic data quality.

Table ES-2. Data Quality Targets

Applications		Data	Data Quality Measure[1]				
			Accuracy[2]	Completeness	Validity	Timeliness	Typical Coverage
Transportation Planning Applications	Standard demand forecasting for Long Range Planning	Daily traffic volumes	Freeways: 7% Principal Arterials: 15% Minor Arterials: 20% Collectors: 25%	At a given location 25% – 12 consecutive hours out of 48-hour count	Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations	Within three years of model validation year	55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors
	Highway Performance Monitoring System	AADT	5-10% Urban Interstate 10% Other urban 8% Rural Interstate 10% Other Rural Mean Absolute Error	80% continuous counts 70-80% for portable machine counts (24-/48-hour counts)	Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations	Data one years old or less	55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors
Transportation Operations	Traveler Information	Travel times for entire trips or portions of trips over multiple links	10-15% RMSE	95-100% valid data	Less than 10% failure rate	Data required close to real-time	100% area coverage
Highway Safety	Exposure for safety analysis	AADT and VMT by segment	5-10% Urban Interstate 10% Other urban 8% Rural Interstate 10% Other Rural Mean Absolute Error	80% continuous count data 50% for portable machine counts (24-/48-hour counts)	Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations	Data one years old or less	55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors
Pavement Management	Historical and forecasted loadings	Link vehicle class	20% Combination unit 12% Single unit	80% continuous count data 50% for portable machine counts (24-/48-hour counts)	Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations	Data three years old or less	55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors

Notes: [1] "Accessibility" for all applications is discussed in the text.
[2] Percentage figures correspond to estimate of Mean Absolute Percent Error (MAPE).

Table ES-3. Level of Effort Estimates for Traffic Data Quality Assessment and Reporting

Task	Action item	Assumed Units	Level of effort	Frequency
General				
Develop mechanism/ system for data quality assessment	Develop data reduction software or procedures	Per program	40 hours	One time
	Design and implement input data procedures	Per program	40 hours	One time
	Test, refine, and update systems and software	Per program	40 hours	Periodic
Develop data quality reporting system	Design/develop reporting procedures and metadata templates	Per program	40 hours	One time
Accuracy				
Develop reference or ground truth data	Design and collect sample baseline data	Per site or data source	8 hours	As required
Assess accuracy of original source field data using independent equipment; and archived data	Download/process review data. Implement framework/software to calculate accuracy measures	Per site or data source	1 hour	As required
	Review results compared to targets	Per site or data source	15 mins	As required
Completeness, validity, timeliness				
Assess quality of original source and archived data	Download, process, and review data. Implement framework to calculate quality measures	Per site or data source	1 hour	As required
	Review results compared to targets	Per site or data source	15 mins	As required
Coverage, and accessibility				
Assess coverage and accessibility qualities of data for the program	Review coverage, accessibility requirements for the program	Per program	1 hour	As required
	Download and review data. Implement framework to evaluate data	Per program	1 hour	As required
Data Quality Reporting and Improvements				
Summarize and report data qualities to potential users (Metadata).	Compile and report data quality to users (Metadata)	Per program	8 hours	Periodic/ as required
Identify improvement and communicate quality problems.	Communicate quality problems to field personnel; schedule maintenance	Per site or data source	4 hours	Periodic/ as required

Note: As required – based on need and time scales e.g., annual, semi-annual, monthly, weekly, daily, or per request.

Guidelines for Data Sharing Agreements

Data sharing agreements codify the roles, expectations and responsibilities among the parties providing and using traffic data. Such agreements can conceivably occur between public entities, entirely between private entities or between private and public entities. Table ES-4 below presents suggested minimum data acceptance standards for the incorporation of ITS-generated traffic data into traffic monitoring programs for planning and engineering purposes.

Table ES-4. Standards for Data Transfer Agreements

Type of Location		Proposed Minimum Quantity Standard	Proposed Quality Standard
Roadway sections	Single location	Seven consecutive days per month	
	Single corridor	100 percent coverage one day per month	Daily count within 10 percent of machine or manual count within 15 percent of hourly count as measured once per year. Twenty percent sample of locations.
	Areawide	75 percent coverage one day per month	Daily count within 10 percent of machine or manual count within 15 percent of hourly count as measured once per year. Five percent sample of locations.
Intersections	Single location	Seven consecutive days per month	N/A
	Single Corridor	100 percent coverage one day per month	Five and 10 percent standard applied every five miles in corridor once a per year. Five percent sample of intersection locations.
	Areawide	75 percent coverage one day per month	Five and 10 percent standard applied to one location per corridor per year. One percent sample of locations.

Conclusions and Recommendations

Data quality is directly based on the extent to which a data set satisfies the needs of the person judging it. A better understanding and means to assess the quality of data offers various benefits including confidence and efficacy in decisions based on data. This project developed a framework and guidelines for measuring and assessing the quality of traffic data for different applications. The case studies used to illustrate the application of the framework are selected to represent a diverse range of data sources and applications. The guidelines include guidance on quality targets, levels of effort required to establish a data quality assessment system within an agency, approaches for including metadata with data quality, and standards for data sharing agreements. The examples for metadata and proposed standards for data sharing agreements provide useful guidance in those areas.

The beta testing although limited, has provided the opportunity to validate the concepts and methodologies presented in the framework and also validate some straw man estimates of data quality targets and estimates of the levels of effort. It is recommended that the estimated levels of effort and quality targets need to be tested and validated based on actual experiences in the use of the framework and guidelines.

1.0 Introduction

The FHWA sponsored workshops on Traffic Data Quality (TDQ) in Columbus, Ohio, and Salt Lake City, Utah in March 2003. The primary objective of the workshops was to define an action plan to address traffic data quality issues. One of the foremost recommendations that came out of the workshops was the call for "guidelines and standards for calculating data quality measures.[1]" These guidelines and standards are expected to contain methods to calculate and report the data quality measures for various applications and levels of aggregation.

The focus of this project is to demonstrate how the concepts for defining and measuring traffic data quality can be implemented in practice. This includes the development of a framework to enable traffic data collectors and users to determine the quality of traffic data they are providing, sharing, and using. This report presents the framework and outlines methods for calculating the data quality metrics for different applications and illustrated with case study examples. The structure of the framework is summarized in Figure 1 which shows the steps to follow in assessing the quality of traffic data.

The intent is to develop a framework that is applicable to a broad spectrum of application areas such as ATIS; advanced traffic management systems; advanced public transportation systems; archived data management; traffic monitoring systems; and other applications dependent on accessibility to timely traffic information.

1.1 Background

Traffic data provide key indicators for a variety of transportation operations and planning purposes and in order to fulfill these purposes, the quality of traffic data must be assured. Traffic data for transportation performance measures must be sufficiently timely and accurate for decision-makers to use with confidence. Various groups within State Departments of Transportation (DOTs) have been collecting traffic data for generations for their own decision-making as well as for Federal Highway Administration (FHWA) programs, such as the Highway Performance Monitoring System (HPMS). Advances in traffic detection systems and the growth of Intelligent Transportation Systems (ITS) infrastructure have provided new sources and new challenges for traffic data collection. ITS infrastructure provides the ability to collect large amounts of traffic data for immediate use in operations as well as data for analytical applications through Archived Data Management Systems (ADMS). The increasing amounts and types of traffic data available from ITS enable new applications but also raise concerns about data quality.

Recent research and analysis have identified several issues regarding the quality of traffic data available from ITS for transportation operations, planning, or other functions. For example, the Advanced Traveler Information Systems (ATIS) Data Gaps Workshop in 2000 identified information accuracy, reliability, and timeliness as critical to ATIS. The key findings of the

[1] Report to FHWA, Traffic Data Quality Workshop Proceedings and Action Plan, Battelle, 2003.

workshop, which are included in a document titled "Closing the Data Gap: Guidelines for Quality Advanced Traveler Information System (ATIS) Data", are the following:

- Guidelines for quality ATIS data are desirable
- Need for further refinement in classifying types of data, quality attributes for each type of data, and quality levels for each attribute
- Guidelines for quality data go beyond ATIS.

These concerns prompted the FHWA sponsored workshops on Traffic Data Quality (TDQ) in Columbus, Ohio, and Salt Lake City, Utah in March 2003. The primary objective of the workshops was to define an action plan to address traffic data quality issues. The action plan included work items that can be executed through the U.S. Department of Transportation (DOT), stakeholder organizations (e.g., American Association of State Highway Transportation Officials [AASHTO], ITS America), and state DOTs. The action plan builds upon the findings in the white papers and inputs obtained from the regional workshops. The action plan provides a blueprint for specific actions to address traffic data quality issues. Implementation of the plan will require collaboration among both public and private partners with the FHWA and state DOTs playing leading roles. The plan identifies 10 priority action items. One of the foremost recommendations that came out of the workshops was the call for "guidelines and standards for calculating data quality measures." These guidelines and standards are expected to contain methods to calculate and report the data quality measures for various applications and levels of aggregation. In addition, the guidelines should also include:

- Examples or case studies of application of data quality methods
- Recommendations for targets for the various quality measures – these data quality goals represent what state agencies can strive to achieve in their operations
- Guidance on how to construct and store quality measures
- Specifications and procedures for reporting data quality metadata
- Costs to calculate and report quality measures.

1.2 Project Objectives and Scope

In order to move the state of the art and practice in traffic monitoring forward, these new methods and tools are needed to meet the challenge of delivering quality data. These methods will enable agencies involved in traffic data collection to formally assess and report the quality of data to the various users. The objective of this project is to develop methods and tools to enable traffic data collectors and users to determine the quality of traffic data they are providing, sharing, and using. The guidelines and standards for calculating data quality measures are intended to addresses the following key traffic data quality issues:

- Defining and measuring traffic data quality
- Quantitative and qualitative metrics of data quality
- Acceptable levels of quality
- Methodology for assessing traffic data quality.

As such, the guidelines will include the following essential elements:

- Definition of qualitative and quantitative metrics for expressing traffic data quality
- Framework for assessing traffic data quality
- Guidelines for considering data quality measures in data collection efforts and data sharing.

The focus of this project is to demonstrate how the concepts for defining and measuring traffic data quality can be implemented in practice. The concepts of data quality measurement are sufficiently developed and adequate data exist to allow such a demonstration to take place. Therefore, it is foreseen that most of the effort will be devoted to developing examples and case studies to apply data quality framework and documenting the framework in the form of guidelines.

1.3 Organization of Report

The remainder of the report is organized as follows:

- Chapter 2 presents the research approach.

- Chapter 3 presents the framework for measuring traffic data quality. This includes concepts and methodologies for calculating the data quality measures. These are illustrated with case studies.

- Chapter 4 presents the guidelines for developing a data quality assessment system within an agency. This chapter also includes estimates of the level of effort required for reporting data quality and specifies procedures for using metadata

- Chapter 5 presents the results of beta test of the framework and guidelines.

- Chapter 6 presents the concluding remarks.

2.0 Research Approach

In developing the framework and guidelines for calculating data quality measures a work plan was prepared that describes the technical approach for developing the various elements of the framework and guidelines. The following are the major steps in developing the framework and guidelines

2.1 Define Traffic Data Quality Metrics

Six fundamental data quality measures were identified in the regional Traffic Data Quality workshops namely accuracy, completeness, validity, timeliness, coverage, and accessibility. The various types of data (e.g., original source, archived data), data consumers or users, and broad groups of applications or uses were identified.

2.2 Prepare a Framework for Assessing Traffic Data Quality

Having defined the quality measures, a framework was developed that is applicable to a broad spectrum of application areas. The framework includes the steps and methods to calculate and report and data quality measures taking into consideration (i) data from different sources (e.g., original source, archived database), (ii) different users and customers and users of traffic data, (iii) various applications and levels of aggregation. The methods are illustrated with three specific case study examples.

2.3 Prepare Guidance on Data Quality Assessment

The guidelines on assessing and measuring data quality include guidance on how data quality interests can be considered in data collection, reporting, and sharing agreements. The following are the main elements of the guidelines.

- Acceptable levels of quality for different applications – the guidelines include estimated acceptable targets for each of the traffic data quality measures developed for each possible application. For each application, estimates of "acceptable" levels for each data quality measure were established based on experience and professional judgment. These were then validated through beta testing with state DOTs and several offices of FHWA.

- Level of effort for assessing and reporting data quality measures – the guidelines include estimates of the levels of effort required to calculate and report the data quality measures were generated. The initial estimates were validated through beta testing with several offices of FHWA and a few state DOTs.

- Specifications and Procedures for using Metadata – the guidelines provide guidance on metadata with data quality measures. The guidelines also include recommendations on the use of ASTM standard for documenting traffic data quality.

- Guidelines for Data Sharing Agreements – This section of the guidelines focuses on how data quality interests can be considered in data sharing agreements. The guidelines for data sharing are based on information derived from three existing data sharing agreements.

2.4 Beta Testing and Review of Guidelines

The draft framework and guidelines were reviewed by a few state DOT representatives and offices of FHWA for review and comments. The state DOTs were expected to apply the framework to the data from actual projects to test the concepts and methodologies contained in the framework. The beta testing was intended to address issues including ease of use, practicality, applicability, costs involved, integration with other state DOT functions, and required enhancements. The beta testing was also intended to validate the estimates of the levels of effort required to establish the data quality assessment system and also validate the straw man estimates of the acceptable data quality targets. Based on results of the beta test, the estimates were revised.

The following chapters present the framework and the guidelines for measuring and reporting traffic data quality as well as the results of the beta testing of the concepts and methodologies.

3.0 Framework for Data Quality Measurement

This section presents the framework for data quality assessment. The framework is structured as a sequence of steps in calculating the data quality measures and assessing the quality (Figure 3-1). The framework takes into account the facts that there are different types of traffic data and different customers and users. The data quality assessment approach is determined by the type of application and the type or source of traffic data. The framework identifies three main types of traffic data for which to calculate data quality:

- **Original source data –** refers to original data (this could be real time or archived) collected from various traffic data collection devices.

- **Archive data –** refers to data stored in an archive database. This dataset is derived from original source data and can be processed or in its original raw state.

- **Traveler information –** refers to data provided as information to travelers. This is usually real time information (i.e., aggregated or processed) and is derived from the original source data.

The framework also recognizes that traffic data is used for different applications. As such, the needs and quality requirements are different for the different data customers and applications. Table 3-1 shows the range of data consumers, types of data, and possible applications.

Table 3-1. Types of Data Consumers and Applications

Data Consumers or Users	Types of Data	Applications or Uses
Traffic operators (of all stripes)	Original source data Archived source data	Traffic management Incident management
Archived data administrators	Original source data	Database administration
Archived data users (Planners and others)	Original source data Archived source data, Archived processed data	Analysis Planning Modeling (development and calibration)
Traffic data collectors	Original source data Archived source data	Traffic monitoring Equipment calibration Data collection planning
Information Service Providers	Original source data (real time)	Dissemination of traveler information
Travelers	Traveler information	Pre-trip planning

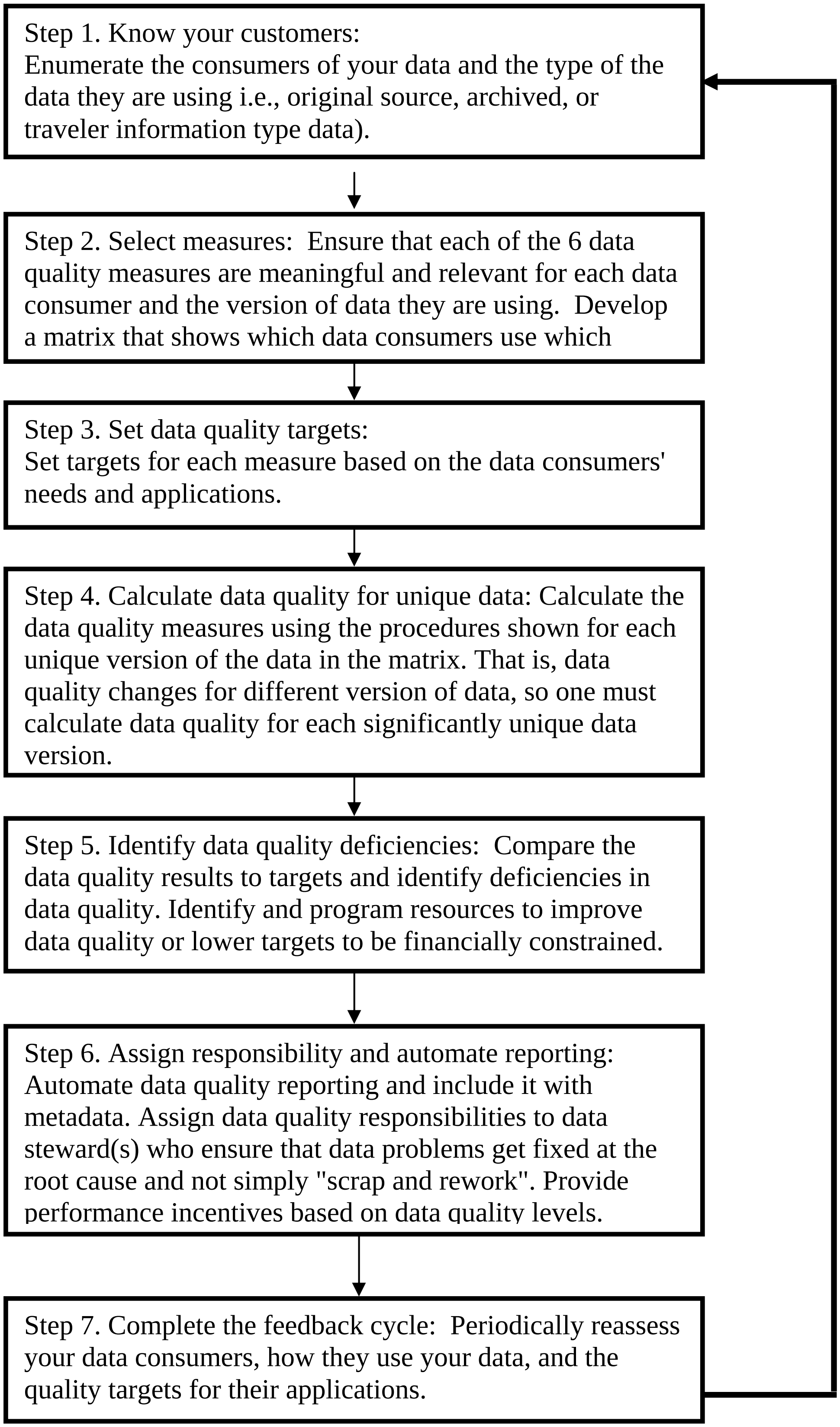

Figure 3-1. Structure of Framework

The following sections present descriptions of the various components of the data quality assessment framework shown in Figure 1.

3.1 Define Data Consumers

The first step in assessing the quality of data is to determine the type of application or data consumer for which the data is intended. This is important because the type of application or data consumer determines the type of data and thus the methods of calculating the quality measures and the thresholds for evaluating the quality of data. Therefore, each agency measuring data quality will have to know their customers. The following are typical primary data consumers or customers whose perspectives should be represented in calculating data quality measures. The terms data consumers and customers are used interchangeably throughout this document.

- **Traffic operations personnel –** This category of data customers typically use original source data for a number of applications related to traffic operations. Examples of applications include traffic monitoring.

- **Archived data administrator –** This category of data customers are responsible for typical administration activities associated with large-scale databases, including managing the archive schema, managing the submission of data into the system, and monitoring system metrics. Archived data administrators use original source data and may assess the quality of such data by applying certain quality checks.

- **Archived data users –** This category of data customers are any users or systems representing users of archived data and/or reports derived from archived data. The archived traffic data may be processed or in the original form. Processed data can have several qualifiers such as the type of processing, the level of aggregation (if applicable) and type or level of summary (if applicable).

- **Traffic data collectors –** This category of data customers represent any users involved in collecting traffic data for statewide, county or local traffic monitoring programs. Examples of applications include, model calibration, traffic monitoring.

- **Information Service Providers (ISPs) –** This category of data customers uses original source data to provide traffic information to various users or customers including travelers.

- **Travelers –** This category of data customers uses traveler information derived from real time original source data. This information is used for pre-trip planning applications.

3.2 Define Quality Measures

As part of the Traffic Data Quality Workshop project, a white paper titled "Defining and Measuring Traffic Data Quality"[2] was developed. This paper reviews current data quality measurement practices in traffic data collection and monitoring; introduces data quality approaches and measures from other disciplines; and recommends approaches to define and measure traffic data quality. The six recommended fundamental measures of traffic data quality are defined below:

- **Accuracy –** The measure or degree of agreement between a data value or set of values and a source assumed to be correct. It is also defined as a qualitative assessment of freedom from error, with a high assessment corresponding to a small error.

- **Completeness** (also referred to as availability) – The degree to which data values are present in the attributes (e.g., volume and speed are attributes of traffic) that require them. Completeness is typically described in terms of percentages or number of data values. Completeness can refer to both the temporal and spatial aspect of data quality, in the sense that completeness measures how much data is available compared to how much data should be available.

- **Validity –** The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values. Data validity can be expressed in numerous ways. One common way is to indicate the percentage of data values that either pass or fail data validity checks.

- **Timeliness –** The degree to which data values or a set of values are provided at the time required or specified. Timeliness can be expressed in absolute or relative terms.

- **Coverage –** The degree to which data values in a sample accurately represent the whole of that which is to be measured. As with other measures, coverage can be expressed in absolute or relative units.

- **Accessibility** (also referred to as usability) **–** The relative ease with which data can be retrieved and manipulated by data consumers to meet their needs. Accessibility can be expressed in qualitative or quantitative terms.

These six (6) data quality measures constitute reasonable "categories" but the actual definition or calculation of the measures could vary by application or data user. It is acceptable (and even desirable) to have slightly different measure calculation procedures for different application or groups of users, as the original source traffic data will likely undergo numerous transformations or other changes as it goes from field data collection equipment to data/information consumer. Thus, the original source data changes as it is collected, transformed, and disseminated, and consequently the data quality is also likely to change on its way to the end consumer.

---

[2] Available on the ITS/JPO Electronic Document Library –http://www.its.dot.gov/itsweb/welcome.htm, EDL # 13767

3.3 Establish Acceptable Data Quality Targets

The next step is to set the threshold values for the data quality measures of interest. It is expected that there will be different threshold values for the same measure depending on the application or the data consumer. These thresholds should reflect the acceptable quality based on the data user's needs and applications. Depending on the user and application, data quality measures falling outside the thresholds could be unacceptable for intended application or indication that the data ought to be used with caution.

3.4 Calculate Data Quality Measures

This section presents the methods for calculating the six data quality measures.

3.4.1 Accuracy

Accuracy is defined as "the measure or degree of agreement between a data value or set of values and a source assumed to be correct." Accuracy can be expressed using one of the following three error quantities. Note that in each of these error formulations, the error is the difference between the observed value(s) and the reference (i.e., ground truth) value, and percent error is the error divided by the reference value.

1. Mean absolute percent error (MAPE) (see Equation 1)
2. Signed percent error (see Equation 2)
3. Root mean squared error (RMSE) (see Equation 3)

$$Mean\ Absolute\ Percent\ Error,\ MAPE\ (\%) = \left(\frac{1}{n}\right) \times \left(\sum_{i=1}^{n} \left| \frac{x_i - x_{reference}}{x_{reference}} \right| \right) \qquad \text{….. Eqn. 1}$$

where: x_i = the observed data value
$x_{reference}$ = the reference value
n = the total number of observed data values

$$Signed\ Error\ (\%) = \left(\frac{1}{n}\right) \times \left(\sum_{i=1}^{n} \frac{x_i - x_{reference}}{x_{reference}} \right) \qquad \text{….. Eqn. 2}$$

where: x_i = the observed data value
$x_{reference}$ = the reference value
n = the total number of observed data values

$$Root\ Mean\ Squared\ Error,\ RMSE = \sqrt{\left(\frac{1}{n}\right) \times \left(\sum_{i=1}^{n} (x_i - x_{reference})^2 \right)} \qquad \text{… Eqn. 3}$$

where: x_i = the observed data value
$x_{reference}$ = the reference value
n = the total number of observed data values

The RMSE can also be expressed as a percentage value (e.g., % RMSE). When so specified, the % RMSE is the RMSE divided by the average of all reference data values.

These different error formulations are all valid measures of accuracy but may reveal slightly different interpretations. The mean absolute percent error (MAPE) and signed error are expressed as percentages; thus, these formulations may be used to compare the relative accuracy of different attributes (e.g., traffic volume count and speed measurement accuracy). Because the signed error does not use absolute error values (as MAPE does), the signed error formulation may reveal whether there is a consistent bias in measurements. The root mean squared error (RMSE) is an error formulation that is commonly available in many statistical software applications.

As its definition indicates, accuracy requires "…a source (of data) assumed to be correct." This correct source of data is typically referred to as ground truth, reference, or baseline measurements. Ground truth data can be collected in several different ways for each traffic data element. In many cases, ground truth data are collected from specialized equipment and reduced in a rigorous manner that minimizes error. For example, consider the case of collecting ground truth data for traffic volume counts from inductance loop detectors. For the ground truth data, one could record video of the same traffic flows measured by the loop detectors, and then have two different human observers count the number of vehicles during the specified test period. If the ground truth vehicle counts from both human observers are within a specified tolerance (e.g., ±1% to ±3%), one could assume that the average of these two manual counts represents the ground truth vehicle count.

Another common method for establishing ground truth is to perform rigorous and routine calibration of data collection equipment, and then assume that the data from calibrated equipment represents ground truth. For example, one might calibrate an inductance loop detector on a weekly basis, and then use this loop detector data as ground truth to evaluate other non-intrusive detection devices. However, it should be noted that calibration is specific to type and model of the equipment. Comparison across different types (such as microwave radar detectors versus loops, microloops versus loops) can distort results.

The following reports document obtaining ground truth or reference measurements for traffic data:

- *Evaluation of Non-Intrusive Technologies for Traffic Detection*, available at http://projects.dot.state.mn.us/nit/index.html.

- *Travel Time Data Collection for Measurement of Advanced Traveler Information Systems Accuracy*, available at http://www.its.dot.gov/itsweb/welcome.htm, EDL Document No. 13867.

Accuracy tests should be performed on usable data from working sensors. In addition to the suggested accuracy measures, quick response or qualitative measures are also needed by data consumers such as TMCs to monitor the performance of detectors. These quick response methods could be graphs showing performance of the detector over time which would indicate any systematic data biases and suggest a need for calibration.

3.4.2 Completeness

Completeness is defined as "the degree to which data values are present in the attributes that require them." Completeness can be expressed using a percentage (see Equation 4). The equation expresses the available number of data values as a percent of the number of total expected data values.

$$\textit{Percent Complete}\ (\%) = \frac{n_{\textit{available values}}}{n_{\textit{total expected}}} \times 100 \qquad \text{….. Eqn. 4}$$

where: $n_{\textit{available values}}$ = the number of records or rows with available values present
$n_{\textit{total expected}}$ = the total number of records or rows expected

The number of data records expected is a function of the application. For example, state DOTs need at least two weeks worth of data in a month to calculate AADT from automatic traffic recorders. The same DOT might require 30 days of data from ATRs for seasonal adjustment factor calculation. However, from a TMC standpoint, while some data losses can be acceptable, a whole day's worth of incomplete data can be problematic.

The percent complete statistic is defined to include all "values present". In this respect, completeness is defined as including both valid and invalid data values (validity is discussed in Section 3.4.3), as long as both types of data values are present in the version of data being evaluated. However, if a particular data process removes invalid data values from a database instead of flagging them as invalid and permanently storing them, then these purged invalid data values would not be included in the completeness statistic because they are not "present".

The quantities in the percent complete equation can be further specified beyond the example shown here. For example, consider that data analysts may wish to know the percent of data that has actually been measured versus the percent of data that has been estimated. In such a case, one could specify two separate completeness measures: percent complete as defined in Equation 4, and a modified percent complete that counts only directly measured data in the numerator. For example, consider that a particular dataset is 80 percent complete, but only 20 percent complete when counting only directly measured data. These statistics would indicate that 60 percent (80 percent complete minus 20 percent measured data) of the expected dataset contains estimated values.

3.4.3 Validity

Validity is defined as "the degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values." Validity can be expressed as the percentage of data passing validity criteria (see Equation 5 below).

$$Percent\ Valid\ (\%) = \frac{n_{valid}}{n_{total}} \times 100 \qquad \text{…. Eqn. 5}$$

where: n_{valid} = the number of records or rows with values meeting validity criteria
n_{total} = the total number of records or rows subjected to validity criteria

Validity criteria (also referred to as business rules or validation checks) are defined in many data management applications and can range from a single simple rule to several levels of complex rules. A simple rule might specify that traffic volume counts cannot exceed a maximum value associated with road capacity (such as 2,600 vehicles per hour per lane) or that traffic speeds can not exceed a reasonable threshold (such as 100 mph). Other validity criteria for traffic data could include the following:

- Compare multiple data elements (i.e., volume, occupancy, and speed) to check for inconsistency among traffic data values;
- Compare traffic data to historical averages or trends from previous days, months, or years;
- Compare traffic data to similar nearby locations (upstream or downstream) to check for continuity of traffic flow; and
- Compare traffic data in consecutive time periods to identify rapid fluctuations.

Validity criteria are often based on "expert opinion" and are generally viewed as "rules of thumb," although some validity criteria may be based on established theory (e.g., road capacity) or scientific fact (e.g., cannot record a zero volume and non-zero speed). The specific validity criteria will likely vary from place to place, as each traffic data collector or manager brings experience with certain roadway locations, traffic data collection equipment, or collection hardware and software.

The difference between completeness and validity is best represented in Figure 3.2. As seen in this figure, the pie represents the total amount of data that is expected to be collected (based on data collection plan or data polling rates). The percent complete statistic includes both valid (slice #3) and invalid (slice #2) values, divided by the total expected number of values (entire pie). The percent valid is the valid values (slice #3) divided by the total values checked (slice #2 and #3).

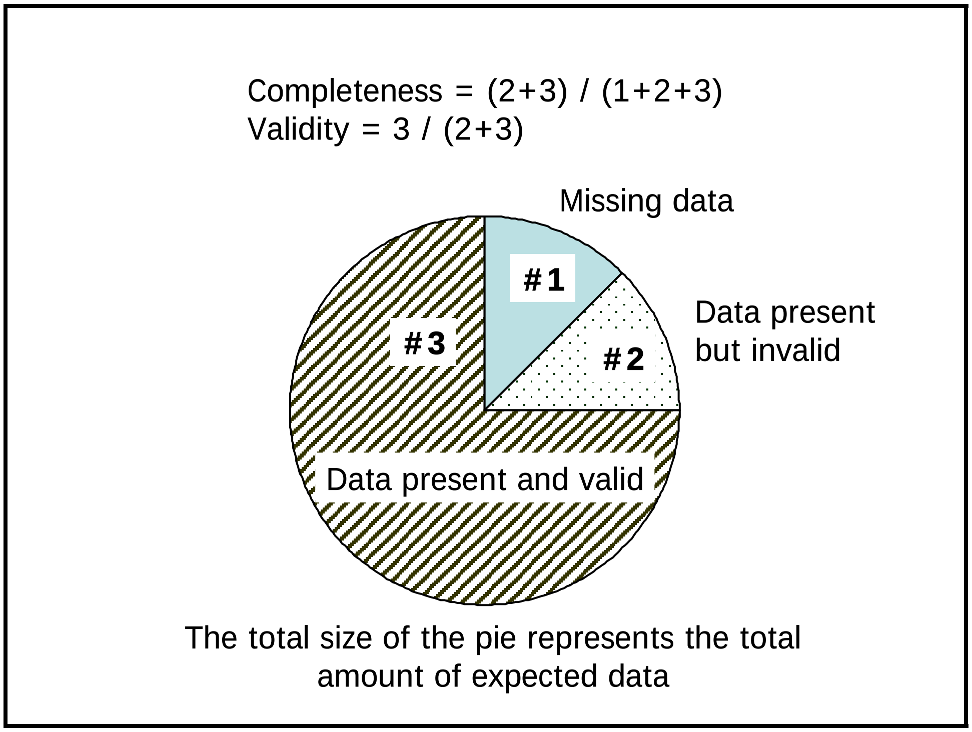

Figure 3-2. Illustration of Completeness and Validity Measures

3.4.4 Timeliness

Timeliness is defined as "the degree to which data values or a set of values are provided at the time required or specified." Timeliness can be expressed as one of these measures (see Equations 6 and 7 below):

- Percentage of data received within acceptable time limits; or
- Average delay for late data.

$$Percent\ Timely\ Data\ (\%) = \frac{n_{on-time}}{n_{total}} \times 100 \qquad \text{… Eqn. 6}$$

where: $n_{on\text{-}time}$ = the number of data messages or packets received within acceptable time limits
n_{total} = the total number of data messages or packets received

Eqn. 6 applies to both device-to-TMC communications and for TMC-to-end user applications. The percent timely data indicates the number of submissions or reports delivered on time.

$$Average\ Delay\ \ for\ \ Late\ Data = \left(\frac{1}{n_{late}}\right) \times \left(\sum (t_{late}) - (t_{expected})\right) \qquad \text{…….. Eqn. 7}$$

where: n_{late} = the number of data messages or packets received outside acceptable time limits
t_{late} = the actual arrival time of a late data message or packet
$t_{expected}$ = the expected arrival time of a late data message or packet

3.4.5 Coverage

Coverage is defined as "the degree to which data values in a sample accurately represent the whole of that which is to be measured." Coverage can be expressed as the percent of roadways (or the transportation system) represented by traffic data. Separate coverage statistics should be calculated for different functional classes of roadway.

The definition of coverage leaves several quantities open for interpretation. For example, how much of a sample is needed to "…accurately represent the whole…"? Or in the case of traffic detectors that are placed at several points along a roadway, what spacing between detectors is necessary to "…accurately represent the whole…"? In addition, coverage can also vary with time as detectors are taken off-line or new detectors are added. Ultimately these issues of interpretation are left to those who calculate the coverage statistics. However, additional information should be provided with coverage statistics to indicate the total sample size or nominal/average detector spacing during a particular time period.

Data quality reports should include the coverage measure because it helps to interpret the other data quality measures. The percent coverage statistic essential tells analysts what portion of the system is being measured, and could explain fluctuations in other data quality measures. For example, the completeness could drop for a particular month. If the sensor coverage remained constant, then clearly some problem has arisen in the existing sensor system. If the sensor coverage recently increased and the completeness has dropped, a likely cause could be the new sensors that are not providing as complete data as the previous sensors.

3.4.6 Accessibility

Accessibility is defined as "the relative ease with which data can be retrieved and manipulated by data consumers to meet their needs." Of the six recommended data quality measures, accessibility is the only measure that is best described in both qualitative and quantitative terms:

- **Qualitative**: A listing or description of the mechanisms or media in which data can be obtained for use. Features like user-friendliness of interface, options for online-queries, and ability to download data can be qualitatively scored under this measure.

- **Quantitative**: The average time required for data consumers to perform specified data retrieval or manipulation tasks. This time should be determined under realistic conditions (i.e., not when no one else is accessing the data system).

3.4.7 Composite Data Quality Measures

Some analysts may wish to have a "composite data quality score" that represents two or more data quality attributes in a single number. For example, suppose that you want a single data quality score that captures the results calculated for each of the 6 data quality attributes. Calculating a composite score could be accomplished by assigning a grading scale (say 1 to 10) to the range of expected results for each data quality attribute. For example, an 85 percent value of completeness is graded as an 8.5 value and an accuracy value of 6 percent is graded as a 9.2 (according to a grading scale). These two data quality attributes could then be combined into a composite score of 8.85 by averaging (or weighting by importance) the "grades" of 8.5 and 9.2. A composite data quality score in this sense would be most useful in relative comparisons or rankings; the composite score could be difficult to interpret as a dimensionless number and no insight could be gained as to possible causes or solutions.

A composite data quality score that may be useful for performance monitoring applications combines the completeness and the coverage attributes to create a "composite system completeness" or "composite system availability" measure. For example, the coverage measure is typically used to represent the portion of the total road network represented by collected traffic data. The completeness measure represents the amount of available data on this subset of "covered" roads, and the validity measure captures the percent of all available data values that are valid. In calculating the amount of valid and usable data as a sample percentage of the entire road system in an urban area, the composite system completeness is the percent coverage multiplied by the percent complete and percent valid (see Equation 8).

$$Composite\ System\ Completeness\ (\%) = Percent\ Coverage \times Percent\ Complete \times Percent\ Valid \qquad \text{……….Eqn. 8}$$

For example, assume that the coverage for freeway operations sensors is 90 percent (i.e., 90 percent of the urban area freeway road mileage is represented by the collected traffic data). Further assume that the data archive from these sensors is 75 percent complete for the year 2003, and the validity is 80 percent. The composite system completeness for 2003 is 54 percent (i.e., 90 percent × 75 percent × 80 percent). This means that the traffic data archive represents 54 percent of the total data that could possibly be collected for the areawide freeway network. This example is further illustrated in Table 3-2.

3.4.8 Level of Detail for Data Quality Measures

Each of the six data quality measures can be calculated at many different levels of detail, from a statistic for a single traffic sensor location for a short time period, to a traffic data archive system that spans multiple years. Certainly, several of the data quality measures are most meaningful at certain levels of detail. For example, accessibility is best used at a system level, whereas accuracy may typically be measured at a few locations that represent the system.

Table 3-2. Illustration of Composite Data Quality

		System	Instrumented (expected data)	Data Available	Valid data
Data values (sites)		200	180	135	108
Individual measures	Coverage		(180/200)*100 = 90%		
	Completeness			(135/180)*100 = 75%	
	Validity				(108/135)*100 = 80%
Composite measures	Coverage completeness			(0.90 *0.75)*100 = 67.5%	
	Valid completeness				(0.75*0.80)*100 = 60%
	Complete valid coverage				(0.90*0.75*0.80)*100 = 54%

It should be recognized that different data consumers will want data quality information available at different levels of detail. For example, a maintenance technician will probably require significantly more detail than an information systems manager. The technician needs detailed data quality information to diagnose and solve problems, whereas the manager may wish to track data quality at a system level to assign resources when needed.

Information systems that report data quality should have the capability to do so at different levels to support the different users of data quality information. "Drilldown" capabilities, which are common in many data warehousing tools, support this presentation and analysis of data at aggregate and disaggregate levels. For example, consider a computer interface that shows a single value for percent complete for the entire data collection system for the entire year. By clicking on the single completeness value, users can "drilldown" to the next level of detail to see completeness values by freeway corridor. Clicking on a completeness value for a freeway corridor "drills down" to a single sensor location, then clicking on that sensor location could provide a day-by-day summary for that location. As such, an information system with this "drilldown" capability easily supports a wide variety of data consumers that would like data quality information at different levels of detail.

3.5 Identify Data Quality Deficiencies

This section identifies data quality deficiencies by comparing the data quality results to targets. Based on the results of the comparisons, identify and program resources to improve data quality or lower targets due to resource constraints. Table 3-3 shows the structure of the data quality statistics.

Table 3-3. Traffic Data Quality Summary

Data Quality Measures	Data Consumer: Original Source Data e.g., • Traffic operations personnel • Archived data administrator • ISPs	Data Consumer: Archive Database e.g., Archived data users	Data Consumer: Traveller Information e.g., travellers
Accuracy • MAPE • RMSE	X (T)	X (T)	X (T)
Completeness • Percent Complete	X (T)	X (T)	X (T)
Validity • Percent Valid	X (T)	X (T)	X (T)
Timeliness • Percent Timely Data • Average Data Delay	X (T)	X (T)	X (T)
Coverage • Percent Coverage	X (T)	X (T)	X (T)
Accessibility • Avg. Access Time	X (T)	X (T)	X (T)

Note: X – calculated value
(T) – threshold value

3.6 Assign Responsibility and Automate Reporting

To facilitate reporting of data quality it is important to document data quality. The ASTM Committee E17.54 is currently developing metadata standards for archiving ITS-generated data. Once the ASTM standard is approved, it should be used for documenting traffic data quality.

The next step in the data quality assessment framework is to assign data quality responsibilities to data steward(s) who would ensure that data problems get fixed at the root cause and not simply “scrap and rework”. Having assigned the responsibility then data quality reporting can be automated. Data stewards could be anybody charged with the responsible for collecting, accessing and retrieving data and reporting such data to users within an organization as well as to users outside the organization. These could be archived database administrators, or heads of traffic collection and monitoring programs. Data stewards would generate and use data quality measures to track system performance and address problems as they occur through either policy, institutional, or technological decisions.

Data quality reporting includes metadata. Several existing standards provide a framework for using metadata to document data quality. For example, FGDC-STD-001-1998[3] is an existing American standard for digital geospatial data. The FGDC standard is used by numerous public

[3] http://www.fgdc.gov/metadata/contstan.html

agencies and private software companies in the U.S. and does support the reporting of data quality measures. It is noted however, the metadata standards community in the U.S. is beginning to move toward eventual adoption of ISO 19115, an international metadata standard maintained by the International Standards Organization.

3.7 Perform Periodic Assessment

The final next step in the data quality assessment framework is to periodically reassess the data consumers, how they use your data, and the quality targets for their applications. The results of the periodic assessment should guide revisions to data collection protocols including data collection equipment selection, calibration, and maintenance as well as review of acceptance targets. This information would also be useful in reviewing cost implications of data quality assessments and the impacts of decisions based on such data.

3.8 Case Studies

The following three case studies demonstrate how the data quality measures can be calculated at three different primary groups of data consumers:

- Traffic managers (recipient of the field-to-center data flow);
- Information service providers (recipient of center-to-center data flow); and
- Data archive managers (recipient of center-to-center data flow, although in some legacy systems the data archives actually get field-to-center data flows).

Table 3-4. Case Studies

Austin, Texas Case Study	Based on a single day of data collected by the Texas Department of Transportation (TxDOT)
Pittsburgh, Pennsylvania Case Study	Based on data from Mobility Technologies Inc
Ohio Department of Transportation Case Study	Based on data collected by traditional methods at various locations in Ohio

The application of the framework is illustrated with these case studies and presented in Appendices A through C. The case studies are intended to only illustrate the application of the methodologies in evaluating traffic data quality. These case studies are not intended to and do not represent a review of the quality of data of the agencies providing the data for this case studies. Note that while most of the data used in these case studies are provided by agencies, some hypothetical data is also used in the illustration.

4.0 Guidelines for Data Quality Measurement

4.1 Introduction

While traditional methods have been used to collect traffic data for generations, intelligent transportation systems (ITS) provide new sources and new challenges for traffic data collection. The ITS data includes large amounts of traffic data for immediate use in operations as well as data for analytical applications through archived data management systems (ADMS). The increasing amounts and types of traffic data available from ITS enable new applications but raise concerns about data quality. The potential for ITS data to fulfill data requirements for transportation planning, engineering, and operations applications has only begun to be realized. Institutional, technical and possibly financial issues remain to be resolved before these data are adopted into widespread use for mainstream applications. This section of the report addresses technical issues related to the data quality standards users require, discusses and describes the salient features of existing and future data sharing agreements, estimates the level of effort required for reporting data quality and specifies procedures for using metadata. Each topic is discussed in the following sections.

4.2 Establishing Acceptable Data Quality Targets

While the planning, engineering and operations disciplines all require transportation data for their analytical procedures and applications, their spatial and temporal requirements differ considerably, with planning applications generally associated with the least stringent requirements and operations applications associated with the most stringent. Traffic data are also variously important as inputs to analyses and applications, as some applications are more sensitive to variations in input traffic values than others. Traffic data providers can benefit from understanding the data requirements of their customers, either in setting their pricing policies, developing truth-in-data statements or in responding to data requests that do not include clear direction concerning the quality needs of the application. By understanding and being responsive to the data quality needs of secondary users, the traffic data collection community can develop a demand for its services and integrate its business operations with those of the rest of the transportation community. In this way, revenue streams or other types of non-monetary support for ITS related and other traffic operations data can be developed and grown.

The following sections discuss the data quality requirements for several planning, operations, and engineering applications. A description of the application and its data requirements, and the significance of traffic data as a source of error for each of the applications are discussed. For purposes of these discussions, the accuracy measure is used to illustrate the importance of data quality in the various applications.

4.2.1 Travel Demand Modeling

Municipal governments, metropolitan planning organizations (MPOs) and state DOTs develop and apply travel demand models to determine infrastructure needs and to set land use and transportation policies. Model analyses are integral to the development of air quality conformity analyses and long-range transportation plans by MPOs. State-of-the-practice transportation models provide estimates of annual daily traffic or AADT by direction. State-of-the-art models may provide a finer grain of temporal and spatial coverage, may account for a larger number of travel markets and, correspondingly, require more and better data. The models often cover large geographic areas, including entire states or metropolitan statistical areas. A typical regional model includes all freeways, expressways and major arterials and most minor arterials in its description of the highway network; relatively few collectors and local roads are included. For sub area and corridor studies requiring more precise results, additional network and zonal detail are added, and additional traffic counts are used in the calibration. The Environmental Protection Agency and the Federal Highway Administration have formulated guidelines for acceptable model practice in model formulations and have provided guidance on measures of performance.[4]

In order to provide reliable forecasts, models are developed to be robust, sensitive and accurate. There are no definitive standards for these qualities. A robust model is capable of providing useful guidance on issues of interest to local policy makers, while sensitivity refers to the model's ability to predict changes in travel behavior resulting from changes in demand (e.g., demographic variables) and supply (e.g., level of infrastructure) characteristics. Accuracy is measured as the level of agreement with observed data in a base-year model whose demand and supply attributes will be modified to reflect alternative future conditions. These observed data range from household trip generation rates and distribution patterns obtained from travel surveys to vehicle and passenger counts.

Traffic counts are the single most important source of observed data used in the calibration of the traffic assignment. Traffic count screen lines demarcate major areas of the model region, and provide one measure of how well the model replicates travel between adjoining regions. Percentage deviations from each crossing location, across the entire screen lines and across all screen lines are major outputs of the typical screen line report. Matches within 5 to 10 percent of observed daily volumes across all screen lines are generally considered adequate. Traffic count on individual links is a second source of assignment calibration data. A measure of average variation between observed and modeled data is often used to measure the quality of the traffic assignment calibration, using percentage deviation, root mean square error (RMSE) and percent RMSE. Percent RMSE is reported by facility type or by volume grouping; in general error tolerances are lower for high-volume facilities than for lower-volume facilities. FHWA-recommended targets for traffic count matches range from seven percent RMSE for freeways to 25 percent for collectors.

Models with transit assignment capabilities utilize station boarding and screen line ridership data for calibration. Time-of-day data are often more critical for transit assignment calibrations, since many assignments cover the morning or afternoon peak period only. More advance modeling practices perform multiple assignments by time of day. This is a considerable effort, because the

---

[4] See the *Model Validation and Reasonableness Checking Manual*, Barton Aschman and Cambridge Systematics, Federal Highway Administration, Travel Model Improvement Program, February 1997.

service characteristics – routes, headways and fares – differ between the peak and off-peak periods.

Traffic count data are only one of several sources of error in a traffic model. Travel behavior is inherently complex and beyond the ability of the relatively simple formulations used in current state-of-the-practice models to predict with a high degree of accuracy. Understanding these limitations, many transportation agencies use the models to predict daily travel patterns, use summary statistics cast over broad areas and round results to an order-of-magnitude estimates, rather than roadway section-specific volumes. Model results are often used in a relative sense to evaluate the differences between two alternative scenarios.

Errors in calibration traffic count datasets may occur and cause temporal and spatial inconsistencies with the underlying network. Neighborhoods and other activity centers are represented as one or more points of access to the street system, making for very "lumpy" traffic distributions, in which modeled traffic volumes change sharply on either side of the traffic loading/unloading points. Traffic counts cannot be reconciled with these loadings very easily. In some cases the count must be moved to one side or another of the actual count location to avoid errors caused by the spatial aggregation of the activity centers. Temporal inconsistencies may arise as well. The model is supposed to represent a snapshot of travel behavior on an average day, when in fact the traffic counts are taken during different years or at different points in time during the year. The application of seasonal, growth and day-of-week factors does not guarantee a consistent distribution of the average day's travel. Counts are sometimes manually smoothed to reduce such inconsistencies.

Overall, the error tolerances of state-of-the-practice travel demand models are relatively high. The traditional threshold for error is one lane of hourly capacity, which can range from 700 for a local road to 2200 for a freeway or expressway. As more sophisticated techniques are adopted to address issues beyond roadway capacity needs, error tolerances will lessen correspondingly.

4.2.2 Air Quality Conformity Analysis

The Clean Air Act Amendments of 1990 stipulate that designated planning organizations ensure that the transportation projects identified in long-range plans contribute to air quality improvement goals for the region. The Act created air quality planning procedures that require the use of mobile source emissions estimates using vehicle miles of travel (VMT) derived from travel demand forecasting methods and other sources.

Emissions modeling uses VMT and emissions rates, which are developed from an emissions factor model, such as MOBILE 6.0, to estimate total emissions. Emissions of carbon monoxide, volatile organic compounds, sulfur dioxide and oxides of nitrogen are modeled using these inputs. The emissions conformity analysis requires the development of VMT distributions by 15 speed categories by vehicle class, hour and four facility types. In most cases, travel demand models are used for the VMT estimates while traffic count data, existing vehicle classification data and vehicle registration data are used to complete these distributions as inputs to the emissions factor model. Current year vehicle-miles of travel (VMT) are adjusted to match Highway Performance Monitoring System (HPMS) database totals by functional classification.

HPMS data are also used for calibration and validation of the model in areas that perform air quality conformity analysis. Observed speeds and VMT are two critical data elements for model validation and calibration. Post-processing programs calibrated to match existing speed data from travel time surveys or dual loop count locations. Modeled VMT is adjusted to match total base year VMT from the HPMS.

Some transportation professionals believe that current state-of-the-art methods can forecast emissions with an accuracy of plus or minus 15 percent to 30 percent.[5] Total regional VMT for the base year, which is dependent on accurate HPMS data, is an essential and critical input to the model calibration and thus to emissions estimates. EPA and FHWA have sought to improve modeling practices for air quality conformity analyses less through insisting on improved input data than in providing guidance on improved modeling procedures, such as the introduction of travel time feedback into trip distribution and the development of modeling estimates by time period.[6]

Air quality conformity analysis requires more detailed model and data than traditional transportation demand modeling analyses. Therefore we conclude that the coverage and accuracy needs for such application would be slightly more stringent than those for state of the practice modeling.

4.2.3 Congestion Management Systems

Federal rules require transportation management areas with populations over 200,000 to develop and implement Congestion Management Systems (CMS). The CMS is intended to be a systematic approach for monitoring and measuring transportation system performance and of diagnosing safety, mobility or congestion issues. The CMS is also used as the basis of evaluating and recommending alternative strategies to manage or mitigate regional congestion and to improve regional air quality. CMS findings may be used to inform project selections in the formulations of transportation improvement programs (TIPs) or constrained long-range transportation plans (LRTPs).

System performance measures based on travel time are generally preferred for CMS reports. Many areas routinely conduct floating car travel time studies to identify and monitor congestion in key metropolitan corridors. Real time traffic data from ITS systems are increasingly used to provide the data. For example, a contractor in Virginia (AirSage) recently began collecting cellular phone positional data in the Hampton Roads area from Sprint for the Virginia Department of Transportation (VDOT) and the regional MPO. Typically, the travel time data represented peak travel conditions. In some areas, travel demand models are used to meet CMS reporting requirements. Highway Capacity Manual techniques may be used to translate travel times or volumes to level of service estimates.

---

[5] Chatterjee, A., et. al. Improving Transportation Data for Mobile-Source Emissions Estimates (NCHRP 25-7). Washington, D.C.: National Cooperative Highway Research Program, 1995.

[6] United States Environmental Protection Agency, Procedures for Emission Inventory Preparation, Volume VI: Mobile Sources, December 1992.

The CMS measures mobility trends at identical or similar locations over time. Consistency of data collection procedures and data analysis techniques is one of the major requirements for the CMS.

4.2.4 Highway Performance Monitoring System (HPMS)

The Highway Performance Monitoring System (HPMS) is a federally sponsored highway database containing data on the extent, condition, and use of the nation's highway system. The HPMS is used for estimating highway needs, apportioning Federal highway funds to states, and reporting on highway condition and performance at the national level. Urban areas designated as National Ambient Air Quality Standard (NAAQS) non-attainment areas use the HPMS to report total vehicle miles of travel and other statistics for air quality conformity analysis. The HPMS is the data source for the Highway Economics Requirements System (HERS), which is an analytical tool used to estimate long-range national highway infrastructure needs and to set funding levels for Federal transportation appropriations bills. At the most detailed levels of application, states use HPMS to evaluate long-range funding needs in their own for statewide needs analysis.

States provide data for the HPMS annually on a valid sample of roadways, excluding local roads and minor collectors (for urban sections). Among the critical data items provided are average annual daily traffic (AADTs), percentage single unit and combination unit trucks on these sample sections. AADTs are reported for the current reporting year and for a forecast year, which usually corresponds to a 20-year forecast. Various geometric and operational characteristics of the sample roadway segments are reported as well. The HPMS is not used for analyzing individual corridors, roadway segments or sub areas. FHWA advises that HPMS traffic data be updated on a three-year basis, and that all counts are factored to represent current year AADTs, i.e., the appropriate growth, seasonal and axle correction factors be applied.

For the most part, AADT estimates on sample segments are derived from permanent count stations and short counts. Forecast AADT may be generated from travel demand models, or linear regression models which relate traffic growth to growth in population and jobs, or an extrapolation of growth trends exhibited in past traffic count data.

The sample sections are randomly selected from a list of highway sections belonging to one of a number of volume groups. Sample sections are fixed, that is to say the same sections are inventoried and updated on a regular, cyclical basis. Volume groups are established for each functional classification, and are defined by urban area size, air quality conformity status, and AADT volume ranges. The number of traffic count samples needed for each volume group is determined by the level of precision needed for the volume group, the variability of AADT in the group and the size of the universe of available sample sections. In general, the sampling target for most volume groups is associated with an error tolerance of 10 percent and a confidence interval of 90 percent. This means that 90 percent of the time, the data collected for any sample section in a volume group will be within 10 percent of its "true" AADT. Sample sections may be assigned to a different volume group if traffic growth warrants such a change.

FHWA provides HPMS submittal software with internal auditing and validation procedures to state DOTs. FHWA performs its own audit on the HPMS data as well. Audit procedures include screening AADT entries across multiple years to isolate and identify large deviations and abnormally high volume to service flow ratios (V/SF). FHWA field offices also perform HPMS process reviews with DOTs. One of the data items with the largest uncertainty is the truck percentages. Many HPMS segments use truck percentages from permanent count stations or similar functional classification locations.

Given the multitude of uses for the HPMS, accuracy, completeness and timeliness are essential. The data are only as accurate as the sampling methods, traffic data and the factoring procedures that underlie them.

4.2.5 Permanent Count Station Reports

The FHWA asks state DOTs to provide copies of continuous traffic volume data collected monthly by permanent count stations within 20 days after the close of the month for which data are collected. While providing volume data only is acceptable, FHWA encourages the provision of vehicle classification data whenever possible. Hourly traffic volumes are reported for each day that data are available. An acceptable submittal contains a minimum of seven days of data covering all days of the week, not necessarily from consecutive days.

Permanent count station data are the bedrock of a transportation agency's traffic count program. This data are used for the various factors used in a traffic count program, including seasonal, day of week, axle correction and growth factors. Data from count station sites are used as default values for time-of-day factors and for vehicle class distributions. Some agencies use these sites to identify speed enforcement needs.

4.2.6 Safety Studies

Transportation agencies conduct safety studies to identify high-probability accident locations, and to identify and treat the cause of the accidents. Traffic data provide information on the relative exposure of travelers to accidents. Exposure is typically expressed in terms of accidents per million miles of travel (MVMT). Desktop safety studies may lead to field reconnaissance to gather additional information on traffic control measures, geometric characteristics or to perform speed studies.

Safety studies report to and use several databases. The Fatal Accident Reporting System (FARS) provides information on traffic fatalities nationwide, with state DOTs contributing most of the data. Additionally, many states maintain a safety management system, which is used to identify safety issues, document the testing and evaluation of potential safety enhancements, and finally, to implement solutions.

Safety studies are hampered by a lack of vehicle classification data, and particularly data on single unit and combination trucks, SUVs and other vehicles. In keeping with the recommendations of the 2001 Traffic Monitoring Guide, state DOTs are beginning to create factor groups for trucks.

The VMT estimates used in safety analyses are subject to the same factoring errors as daily counts used for other analyses. Safety studies would appear to have a relatively high tolerance for systematic bias, since the candidate sites are evaluated in comparison to one another. Likewise, because of the use of the accident per million vehicle miles as a metric, the statistic will not be as adversely affected by errors in the DVMT estimate as other types of analysis.

4.2.7 Traffic Simulation

Traffic simulations mimic the real-time movement of vehicles through intersections, roadway corridors or small areas. Unlike most regional travel demand assignment software, simulation packages take into account most or all of the geometric and operational characteristics of the facility being simulated. These packages can produce second-by-second turning movement data by signal phase, weaving movements across lanes and the delay caused by the buildup and dissipation of queues in the traffic system. Traffic simulations are used for operations and design studies, and are essential in assessing whether a particular geometric configuration will accommodate the anticipated traffic demand. A freeway to arterial interchange design is a typical application of a simulation program. Examples of software packages in use today include Synchro and CORSIM.

Several of the packages produce striking visualizations of the projected motion of vehicles in the traffic stream, as well as detailed statistics such as stopped delay, speed by small increments, gap and headway statistics. Studies using these packages analyze relatively small increments of time such as peak-hour conditions. Relatively small areas such as intersections, portions of roadway corridors or small sub areas are analyzed.

Simulation packages are data intensive, often requiring detailed information about the operational and geometric characteristics of the roadway being simulated. This limits their application for planning purposes. Traffic data are a critical input to the simulation packages since the facility will be engineered to accommodate the traffic demand, recognizing right-of-way and other constraints. Most frequently, the most recent traffic counts available are used for the simulations, although forecast model data are sometimes used as well. For signal timing applications, turning movement data for morning, evening and off-peak are generally required.

There is a high level of confidence in the algorithms that are used to simulate traffic at the microscopic and mesoscopic levels. The largest source of error comes not from the algorithms themselves but from the traffic data inputs. There is a considerable though unquantified uncertainty over whether the input data are representative of the likely variability in the magnitude, temporal and spatial distribution of traffic. Another uncertainty is the degree to which the traffic count input is representative of peak demand, for which a facility is typically designed.

4.2.8 Program and Technology Evaluation

FHWA and many state DOTs perform field evaluations of new technologies in advance of large scale procurements of third-party products. These evaluations are often large, expensive and multidisciplinary, and consider the broader economic and institutional implications of the technology, as well as the narrow questions of effectiveness and efficiency of the technology itself. These technology evaluations assess the potential for success of the technology in large scale deployment, help determine their most appropriate applications and identify the critical external factors which are likely to contribute to the technology's success or failure. These evaluations vary widely in geographic scope, but corridor-level studies are not uncommon. In 2000, for example, the FHWA initiated a multi-year study on the use of wireless technologies for monitoring travel speeds on the Capitol Beltway around Washington, D.C.

The technology evaluations often develop detailed data collection plans as part of the overall evaluation plan. Data needs are specific to the evaluation and can vary from one application to another, but in general site-specific, finer grained data, temporally and spatially, is required for these evaluations than for other types of planning applications. An ideal data collection plan for such a study might include speed, volume and vehicle classification data at less than five-minute increments and between or at the approach to all roadway junctions covered by the study. Most studies fall short of this ideal due to resource constraints. The quality of the traffic data being collected must be monitored almost in real time, since the reliability of the results and findings depend so heavily on accurate, valid and reliable data.

Obviously, the reliability of these program evaluations depends greatly on the amount and the quality of the data collected. Relative to other types of applications, the need for valid, reliable and accurate traffic data is high.

4.2.9 Ramp Signal Coordination

Ramp signals at inbound freeway interchanges meter inbound traffic, allowing vehicles to enter the mainline traffic stream as acceptable gaps appear. Ramp signals have been installed in radial freeway corridors in many North American cities. The signals are designed to minimize disruptions to mainline freeway traffic flow and to maintain steady speeds on the freeway, as even minor, sudden reductions in speed can have major upstream ripple effects. The more advanced systems include algorithms that balance the objectives of smoothing freeway flow, with those of minimizing signal delay and the potential for spillover traffic into adjoining neighborhoods. Most systems are set not to exceed a maximum amount of maximum delay at the ramps regardless of main line conditions.

More advanced ramp signal systems are coordinated over an entire corridor and utilize real-time traffic information from the mainline and at the ramp approaches. These systems are able to adjust their signal timings automatically as conditions change, or be overridden by an operator. Older systems which are not demand responsive, however, rely on fixed timing schemes based on available traffic counts. Optimally, traffic volume data at two- to five-minute increments would be a minimum data requirement for adequate operation of the ramp signals.

Whether governed by fixed or demand-responsive timing schemes, the effectiveness of ramp signals is directly related to the timeliness and accuracy of the traffic volumes data received. There is a low tolerance for delay among travelers at the ramp signals, and the need for reliable and accurate data is very high.

4.2.10 Traveler Information

Advanced traveler information systems (ATIS) alert travelers to unusual traffic conditions, allowing travelers to adjust their departure time, route or mode of travel so as to reduce or avoid travel delay. Sources of traveler information include radio and television-based traffic reports derived from monitored police, fire and rescue transmissions, information provided by transportation management centers (TMCs) or helicopter and video surveillance, 511 phone systems, web sites and freeway variable message signs. Many metropolitan travelers can access web sites that provide region wide color-coded maps of current traffic conditions, along with information about incident and accident locations. As of 2003, there were at least 11 metropolitan areas that offered travel time estimates on major freeways.[7] A recent study[8] estimated that the minimum ATIS accuracy requirements for freeway travelers in Los Angeles to be in the 13 to 15 percent error range. En-route information accessible from in-vehicle systems still lacks an attractive business model to entice widespread private sector participation and a demonstrated willingness to pay by the traveling public.

The most commonly available sources of traveler information are ubiquitous and free, but have not advanced in quality significantly over the past 20 years. The available data are neither timely nor of sufficient spatial coverage to provide reliable route-choice options for individual travelers. According to some studies, widespread availability of accurate, detailed and timely traveler information could improve the efficiency of highway operations by five to 10 percent, albeit at a significant cost.[9]

4.2.11 Pavement Management Systems

Pavement management systems use pavement condition data and sophisticated deterioration models to estimate future reconstruction, rehabilitation and overlay needs and costs. Pavement maintenance needs are a function of several factors, including the composition and condition of the surface and base, the geometric design of the roadway and the composition and magnitude of existing and anticipated traffic.

Pavement design requires information about vehicles and the loads they exert on the pavement beneath them. The 1986 AASHTO roadway design equations used 18,000-pound equivalent

---

[7] Wunderlich, et al., Urban Congestion Reporting. Ongoing task for the Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

[8] Toppen, A., and Wunderlich, K., "Travel Time Data Collection for Measurement of Advanced Traveler Information Systems Accuracy, Federal Highway Administration, June 2003.

[9] Wunderlich, K., et al., "On-Time Reliability Impacts of Advanced Traveler Information Services: Washington, DC Case Study", Federal Highway Administration, January 2001.

single-axle loads as the measure of load. The 2002 AASHTO pavement design equations use load spectra, which characterize traffic loads in terms of the distribution of single-, tandem-, tridem- and quad-axle configurations within each of a number of weight classifications. Volume, vehicle classification and weight data are required to develop load spectra estimates. Typically, weights by vehicle type are developed using data at static weigh stations or weigh-in-motion stations, and these data are applied to vehicle classification data derived from permanent count station and other count locations where classification count data are collected. Vehicle distribution factors, growth factors and seasonal factors are also used to develop volume estimates. Techniques for converting traffic counts to load spectra are under development through work sponsored by the Transportation Research Board (TRB, 2004 [NCHRP 1-37-A]).

Variability in traffic data and especially truck weight data is a significant issue in pavement design. To account for variability, the 1986 AASHTO Design equation included terms for standard deviation and the standard error for truck weight. The 1992 *AASHTO Guidelines for Traffic Data Programs*[10] cites studies suggesting that the standard deviations for WIM data range from 0.55 to 0.80.

The 1992 AASHTO Guidelines demonstrated the relationship between traffic volume errors and overlay thickness. Because the error in overlay thickness increases non-linearly as traffic volumes increase, errors in vehicle classification can have a substantial impact on pavement design estimates. The *Guidelines* notes that traffic monitoring systems that can achieve traffic data accuracies representative of a 50 percent confidence interval result in pavement overlays (+/-) one-quarter inch to one-half inch of the true pavement thickness needed compared to counts representative of the 80 percent confidence interval,[10] for roadway sections experiencing 2.5 million design-equivalent axle loads over the life of a roadway section. Errors of such magnitude can arise, for example, when system-level defaults for vehicle distributions are used for entire functional classifications of roadways, rather than using factors that reflect the prevailing traffic patterns for the roadway sections being analyzed.

4.3 Quantifying Data Quality Targets

The previous section described several typical planning, operations, and engineering applications, discussed various sources of error common to the application and assessed the application's tolerance for error in the types of traffic data ITS systems can provide. Table 4.1 presents a summary of estimated data quality targets for the different applications discussed above. These targets are defined for the six data quality measures:

[10] Joint Task Force on Traffic Monitoring Standards of the AASHTO Highway Subcommittee on Traffic Engineering, AASHTO Guidelines for Traffic Data Programs, American Association of State Highway and Transportation Officials, 1992.

Table 4.1. Draft Data Quality Requirements for Planning, Engineering, and Operations Applications

			Data Quality Attribute[1]				
			Accuracy[2]	*Completeness*	*Validity*	*Timeliness*	*Typical Coverage*
Transportation Planning Applications	*Air Quality Conformity Analysis*	*VMT by vehicle class, hour and functional classification*	*10%*	*At a given location 50% – Two weeks per month, 24 hours*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Within three years of model validation year*	*75% Freeways/Expressways* *25% principal and minor arterials* *10% collectors*
		VMT by hour and vehicle classification (Distribution of VMT by speed)	*+- 2.5 mph*	*At a given location 25% – one week per month, 24 hours*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent counts*	*Within three years of model validation year*	*75% Freeways/Expressways* *25% principal and minor arterials* *10% collectors*
	Standard demand forecasting for Long Range Planning	*Daily traffic volumes*	*Freeways: 7%* *Principal Arterials: 15%* *Minor Arterials: 20%* *Collectors: 25%*	*At a given location 25% – 12 consecutive hours out of 48-hour count*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Within three years of model validation year*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		Hourly traffic volumes	*Freeways: 7%* *Principal Arterials: 15%* *Minor Arterials: 20%* *Collectors: 25%*	*At a given location 25% – 12 consecutive hours out of 48-hour count*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent counts*	*Within three years of model validation year*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		Vehicle occupancy	*10-15%*	*At a given location 25% – 12 consecutive hours out of 48-hour count*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent counts*	*Within three years of model validation year*	*1-5% of total population (from surveys)*
		Percentage single unit trucks *Percentage combination trucks*	*7-10%* *3-5%*	*Minimum 25% – 12 consecutive hours out of 48-hour count* *Minimum 50% – 12 consecutive hours out of 24-hour count*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent counts*	*Within three years of model validation year*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		Transit boardings and alightings by station and/or stop	*15-20%* *7-10% (Transit Planning)*	*75% of annual data collection*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent counts*	*Within three years of model validation year*	*100% of rail boardings* *10% of bus route ridership from screen line data*
		Transit vehicle speeds by analysis time period	*15-20%*	*<5% – one peak and one off-peak route*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent counts*	*Within three years of model validation year*	*100%*
		Free Flow link speeds	*15-20%*	*90-100% validity for instrumented floating car data collection*	*90-100% validity for instrumented floating car data collection*	*Within three years of model validation year*	*100% Freeway mileage* *100% Major arterial mileage* *80-100% Collectors mileage* *10% Local road mileage*
		Congested link speeds	*At V/C < 1.0, 10 mph* *At V/C >1.0, 2.5 mph*	*90-100% validity for instrumented floating car data collection*	*90-100% validity for instrumented floating car data collection*	*Within three years of model validation year*	*100% Freeway mileage* *100% Major arterial mileage* *80-100% Collectors mileage* *10% Local road mileage*

Notes: [1] "Accessibility" for all applications is discussed in the text.
[2] Percentage figures correspond to estimate of Mean Absolute Percent Error (MAPE).

Table 4.1. Draft Data Quality Requirements for Planning, Engineering, and Operations Applications (Continued)

			Data Quality Attribute[1]				
			Accuracy[2]	***Completeness***	***Validity***	***Timeliness***	***Typical Coverage***
Transportation Planning Applications	*Traffic simulation*	*Traffic volumes by minute or sub-minute*	*2.5%*	*90% validity*	*Up to 15% failure rate – portable traffic counts*	*Within one year of study*	*100% of study area*
		Turning movements by 15 minutes	*5-10% error rate*	*95% validity – manual traffic counts*	*0% failure – manual traffic counts*	*Within one year of study*	*100% of study area*
		Free Flow link speeds	*5.0%*	*90-100% validity for instrumented floating car data collection*	*90-100% validity for instrumented floating car data collection*	*Within one year of study*	*100% of study area*
		Congested link speeds and delay statistics	*2.5%*	*90-100% validity for instrumented floating car data collection*	*90-100% validity for instrumented floating car data collection*	*Within one year of study*	*100% of study area*
		Queue length		*95% validity – manual count*	*100% validity – manual count*	*Within one year of study*	*100% of study area*
	Congestion management	*Corridor-level vehicle speeds and/or travel times by hour*	*5%*	*90-100% validity for instrumented floating car data collection*	*90-100% validity for instrumented floating car data collection*	*Within six months of study*	*100% of study area*
		Origin-Destination travel times by hour	*5%*	*90-100% validity for instrumented floating car data collection*	*90-100% validity for instrumented floating car data collection*	*Within six months of study*	*1-5% of study area (from surveys)*
	Highway Performance Monitoring System	*AADT*	*5-10% Urban Interstate* *10% Other urban* *8% Rural Interstate* *10% Other Rural* *Mean Absolute Error*	*80% continuous count data* *70-80% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		K factor *D factor*	*5-10% RMSE (relative)* *1% RMSE (relative)*	*80% continuous count data* *50% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		Percent combination and single-unit trucks – Daily	*20% RMSE* *15% RMSE*	*80% continuous count data* *50% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		VMT	*5-10% RMSE* *Downward bias*	*80% continuous count data* *50% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Data one year old or less*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*
		Percent combination and single-unit trucks – Peak	*25% RMSE* *20% RMSE*	*80% continuous count data* *50% for portable machine counts*	*Up to 15% failure rate – 48-hour counts* *Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*55-60% of freeway mileage* *25% of principal arterials* *15% of minor arterials* *10-15% of collectors*

Notes: [1] "Accessibility" for all applications is discussed in the text.
[2] Percentage figures correspond to estimate of Mean Absolute Percent Error (MAPE).

Table 4.1. Draft Data Quality Requirements for Planning, Engineering, and Operations Applications (Continued)

			Data Quality Attribute[1]				
			Accuracy[2]	*Completeness*	*Validity*	*Timeliness*	*Typical Coverage*
	Monthly count station volume reports	*Hourly volumes for seven consecutive days each month*	*2% RMSE*	*100% valid data*	*100% valid data required*	*Data one month old or less*	*<1% of total roadway mileage*
		AVC stations: Hourly volumes by vehicle class category	*15% Single-Unit Truck Classification Error*	*100% valid data*	*100% valid data required*	*Data one month old or less*	*<1% of total roadway mileage*
Transportation Operations	*Program and Technology Evaluations*	*Link and corridor volumes*	*2% RMSE*	*90% valid data*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Less than six months old*	*75-80% coverage of corridor needed*
		Link and corridor delay statistics	*2% RMSE*	*90% valid data*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Less than six months old*	*75-80% coverage of corridor needed*
	Pre-Determined Ramp and Signal Coordination	*Link and corridor volumes*	*2% RMSE*	*90% valid data*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Less than three months old*	*75-80% coverage of corridor needed*
		Link and corridor and delay statistics	*2% RMSE*	*90% valid data*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Less than three months old*	*75-80% coverage of corridor needed*
	Traveler Information	*Travel times for entire trips or portions of trips over multiple links (e.g., travel time to popular destinations from a point)*	*10-15% RMSE*	*95-100% valid data*	*Less than 10% failure rate*	*Data required close to real-time*	*100% area coverage*
	Predictive traffic flow methods (still under research	*Link volumes*	*2% RMSE*	*90% valid data*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*100% area coverage*
		Link delay statistics	*2% RMSE*	*90% valid data*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*100% area coverage*
Highway Safety	*Exposure for safety analysis*	*AADT and VMT by segment*	*5-10% Urban Interstate 10% Other urban 8% Rural Interstate 10% Other Rural Mean Absolute Error*	*80% continuous count data 50% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Data one year old or less*	*55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors*
		Traffic volumes and flow characteristics at times of specific crashes	*25%*	*80% continuous count data 50% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Data one yeas old or less*	*2-5% of total roadway segments*
Pavement Management	*Historical and forecasted loadings*	*Link volumes*	*5-10% Urban Interstate 10% Other urban 8% Rural Interstate 10% Other Rural Mean Absolute Error*	*80% continuous count data 70-80% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors*
		Link vehicle class	*20% Combination unit 12% Single unit*	*80% continuous count data 50% for portable machine counts (24-/48-hour counts)*	*Up to 15% failure rate – 48-hour counts Up to 10% failure rate – permanent count stations*	*Data three years old or less*	*55-60% of freeway mileage 25% of principal arterials 15% of minor arterials 10-15% of collectors*

Notes: [1] "Accessibility" for all applications is discussed in the text.
[2] Percentage figures correspond to estimate of Mean Absolute Percent Error (MAPE).

- **Accuracy –** The measure or degree of agreement between a data value or set of values and a source assumed to be correct. Also, a qualitative assessment of freedom from error, with a high assessment corresponding to a small error.

- **Completeness** (also referred to as availability) – The degree to which data values are present in the attributes that require them.

- **Validity –** The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values.

- **Timeliness –** How current the data are with regard to their collection time.

- **Coverage –** The degree to which data values in a sample accurately represent the whole of that which is to be measured.

- **Accessibility** (also referred to as usability) – The relative ease with which data can be retrieved and manipulated by data consumers to meet their needs.

Note that assessments of accessibility by application are not included in Table 4.1. This is because, with one exception, the applications are not extremely sensitive, i.e., they do not typically require short access times. The exception is predictive traffic flow methods, which would require archive access time less than 30 seconds. The remainder of the applications can be adequately serviced with access times in the 5-10 minute range.

4.4 Level of Effort Required for Traffic Data Quality Assessment

Sufficient temporal coverage and minimal data quality standards should be in place in advance of the transfer of data to the traffic monitoring system managers. System managers would initiate application specific QA/QC procedures for integrating other data sources into their systems. The data would then be transferred on request to users for applications.

It is clear that maintaining data quality levels requires additional effort on the part of transportation agencies to:

- ensure that field equipment and communication systems are working properly
- develop and implement software to perform data quality checks; and
- review data quality and communicate problems to field personnel.

The extra costs associated with assessing and reporting data quality was considered an important issue at the regional TDQ workshops.

Table 4.2 presents estimates of the level of effort, expressed in hours of labor, required to implement a data quality assessment program. These estimates include the time required to calculate and report each of the measures. These are crude estimates that have not been validated in a real situation.

Table 4.2. Level of Effort Estimates for Traffic Data Quality Assessment and Reporting

Task	Action item	Assumed Units	Level of effort	Frequency
General				
Develop mechanism/ system for data quality assessment	Develop data reduction software or procedures	Per program	40 hours	One time
	Design and implement input data procedures	Per program	40 hours	One time
	Test, refine, and update systems and software	Per program	40 hours	Periodic
Develop data quality reporting system	Design/develop reporting procedures and metadata templates	Per program	40 hours	One time
Accuracy				
Develop reference or ground truth data	Design and collect sample baseline data	Per site or data source	8 hours	As required
Assess accuracy of original source field data using independent equipment; and archived data	Download/process review data. Implement framework/software to calculate accuracy measures	Per site or data source	1 hour	As required
	Review results compared to targets	Per site or data source	15 mins	As required
Completeness, validity, timeliness				
Assess quality of original source and archived data	Download, process, and review data. Implement framework to calculate quality measures	Per site or data source	1 hour	As required
	Review results compared to targets	Per site or data source	15 mins	As required
Coverage, and accessibility				
Assess coverage and accessibility qualities of data for the program	Review coverage, accessibility requirements for the program	Per program	1 hour	As required
	Download and review data. Implement framework to evaluate data	Per program	1 hour	As required
Data Quality Reporting and Improvements				
Summarize and report data qualities to potential users.	Compile and report data quality to users (Metadata)	Per program	8 hours	Periodic/ as required
Identify improvement and communicate quality problems.	Communicate quality problems to field personnel; schedule maintenance	Per site or data source	4 hours	Periodic/ as required

Note: As required – based on need and time scales e.g., annual, monthly, weekly, daily, or per request.

These levels of effort estimates are based on experienced data archive administrators who are familiar with the data collection and archiving protocols. Level of effort estimates could be significantly higher in other scenarios.

It is important to note that the estimates presented in Table 4.2 do not account for the level of effort required to maintain or improve data quality. These estimates represent the level of effort required to assess the quality of existing data. Since the labor rates for individuals who would be responsible for function may vary by agency and type of application, it is more appropriate to give guidance on the approximate duration required to perform these data quality calculations. It is also acknowledged that experience in performing these tasks will be reflected in the time and therefore of costs. It is also assumed that the time (cost) will also be a function of the type or source of data and the application. These variables are taken into account in developing the guidelines for costs associated with assessing and reporting data quality measures.

In estimating the level of effort, it is recognized that there are two components of time (cost) involved. First, an initial one time cost will be incurred in establishing the mechanism for assessing the quality of data. While the framework for assessing data quality developed in this project establishes that mechanism to some extent, some extra effort will be required to familiarize with the application of the framework and develop software programs or procedures based on the framework. Second, recurrent cost associated with the application of the framework to assess the quality of any new data. The information presented in Table 4.2 distinguishes between these two cost components.

4.5 Specifications and Procedures for Using Metadata for Reporting Data Quality

Metadata is an extremely important consideration for data sharing in general, and especially for communicating data quality. While data users may be several degrees of separation away from data collection, knowledge about what the data represent and their collection conditions is key to their use.

Commonly referred to as "data about data," metadata is typically thought of as dataset descriptions. Metadata are analogous to a library card catalog that contains information about books: accession number, place of printing, author, etc. In this analogy, the books themselves are the "data". The descriptions typically found in a data dictionary (e.g., definition, size, source) are also metadata. Metadata has several purposes:[11]

- **Summary –** to summarize the meaning of the data.
- **Finding –** to allow users to search for the data.
- **Advisement –** to allow users to determine if the data is what they want.
- **Selection –** to help decide which instance of the data should be retrieved (if multiple formats are provided).

[11] Hodgson, Katrina, *Metadata: Foundations, Potential and Applications,* School of Library and Information Studies, University of Alberta, March 1998

- **Retrieval –** to retrieve and use a copy of the data (i.e., where does one go to get the data).

- **Restriction** – to prevent some users from accessing data.

- **Interpretation –** to instruct on how to interpret the data (e.g., format, encoding, encryption).

- **Specifications –** to give information that affects the use of data (e.g., legal conditions on use, its size, or age); terms and conditions for use of an object (an access list of who can view the object, a conditions of use statement; a schedule of fees for use of the object; or a definition of permitted uses of an object).

- **History –** to describe the history or provenance of data, such its original source and any subsequent transformations (filtering, decimation, etc.).

- **Data administration –** to give specifications for the management of an object within a server or repository (date of last modification, date of creation, and the administrator's identity).

- **Data linkages or relationships** – to give specifications about the relationship between objects. (For example, linkages between a set of articles and a containing journal, between a translation and the work in the original language, between a subsequent edition and the original work, and between the components of a multimedia work.)

- **Data structure –** to list the logical components of complex or compound objects and how to access those components (table of contents; the list of components of a software suite).

Several existing standards provide a framework for using metadata to document data quality. For example, FGDC-STD-001-1998[12] is an existing American standard for digital geospatial data. The FGDC standard is used by numerous public agencies and private software companies in the United States and does support the reporting of data quality measures; however, the metadata standards community in the U.S. is beginning to move toward eventual adoption of ISO 19115[13], an international metadata standard maintained by the International Standards Organization.

---

[12] *Content Standard for Digital Geospatial Metadata*, Metadata Ad Hoc Working Group Federal Geographic Data Committee, 590 National Center Reston, Virginia 20192.

[13] DRAFT INTERNATIONAL STANDARD ISO/DIS 19115, ISO Central Secretariat 1 rue de Varembé 1211 Geneva 20 Switzerland.

ASTM Committee E17.54 is currently developing metadata standards for archiving ITS-generated data. ASTM distinguishes several types of metadata that must be considered:

(a) *Archive Structure Metadata*, descriptive data about the structure of the data archive itself and of the data and information in the archive that facilitate use of the archive. This form is for metadata that does not change often. Coverage is the data quality attribute best suited to this form. Also, descriptions of the tests used to define the remaining data quality attributes are best documented here. Both the ISO and FGDC standards are limited to this form of metadata.

(b) *Processing Documentation Metadata*, information that describes the processes applied to data from original source data through to storage in an archive. The results of completeness, validity, and timeliness tests are examples of this form of metadata. Note that the metadata itself is probably stored as data elements in a data dictionary rather than as traditional metadata.

(c) *Data Collection System Metadata*, data about the conditions and procedures under which original source data were observed, surveyed, measured, gathered, or collected as well as about the equipment that was used. The reporting of accuracy results is in this category. As with processing documentation metadata, the metadata itself is probably stored as data elements in a data dictionary rather than as traditional metadata.

It is recommended that the ASTM standard, once approved, be used for documenting traffic data quality. This standard borrows heavily from the FGDC standard for general types of metadata (archive structure metadata) and is developing detailed data elements and record structures for processing documentation and data collection system metadata. An example of how the ISO 19115 standard can be used to document archive structure metadata is shown below.

Example Data Quality Documentation Using ISO 19115

This example is provided in a tabbed-outline format (Figure 4.1). Element values are underlined and role names are denoted with a "+". Underlines indicate entered data. Not all potential forms of metadata are entered since the focus here is on data quality.

This data archive contains traffic data summaries for several different granularity levels in time and space. For example, the available data granularity levels include both 15 and 60 minutes, as well as by lane or all directional lanes combined. The data in this archive have been organized in comma-separated value (csv) ASCII-text files in a way that supports easy import and use in desktop computer spreadsheet or database programs such as Microsoft Excel or Access. Alternatively, the data can also be batch-imported into a relational database management system (RDBMS) such as Oracle or Sybase.

MD_Metadata
fileIdentifier: AUSTIN_FREEWAY_2002
language: en
characterSet: 001
contact:
CI_ResonsibleParty
organisationName: Texas Department of Transportation
role: 002
dateStamp: 20030803
metadataStandardName: ISO 19115
metadataStandardVersion: DIS
+identificationInfo
MD_DataIdentificaiton
citation:
.CI_Citation
. title: ITS Traffic Data for Austin
. date:
. CI_Date
. date: 193001
. dateType: 001
abstract: This dataset contains archived traffic data that were collected during 2001 on select Austin area freeways by the Texas Department of Transportation (TxDOT). The data were originally collected by the Operations Group of the Austin District of TxDOT for the purposes of traffic management and traveler information. The data were provided to the Texas Transportation Institute (TTI), who performed additional quality assurance, summarized and re-organized the original source data for eventual use and distribution.

Figure 4.1. Example of Data Quality Documentation Using ISO 1915

The data archive also includes a sensor inventory spreadsheet that describes approximate sensor locations, sensor location groupings, and other descriptive information. The sensor inventory spreadsheet was developed by TTI with basic sensor information provided by TxDOT.

A shortcoming of the TxDOT ATMS filename convention is that it indicates only the day of the week, not the date. The date stamp on the file itself typically reveals the actual date since it is not contained in the filename. To add date stamps to the filename, we un-zip these files into 52 separate folders that correspond to the week of the year. The file "aus_unzip.xls" was used to create a *.bat file for batch processing. We then use a batch renaming program (CKRename) to substitute a date stamp (YYYYMMDD) for the weekday name, treating separately the files in each individual weekly folder. The renamed files have the filename convention "RR #### SCU YYYYMMDD HHMM.det" where RR=the route designation (e.g., IH, US, etc), ####=the route number (e.g., 0035, 0290, etc). These "SCU date stamp added" text files are then compressed for long-term storage. Note that there are probably more efficient solutions to getting the date stamps from these files into SAS (instead of including in the filename).

spatialRepresentationType: 001
spatialResolution:
geographicBox:
..EX_GeoBoundingBox
..westBoundLongitude: -97.82832
..eastBoundLongitude: -97.66088
..southBoundLatitude: 30.51693
..northBoundLatitude: 30.21198
geographicDescription: Central Texas
+resourceConstraints
.MD_Constraints
.useLimitation: This dataset is provided as unofficial traffic data collected by TxDOT and further processed by TTI. While efforts have been made to improve the quality of the data since its original collection, no warranty--express or implied--is made by TTI or TxDOT as to the accuracy or completeness of this data. Nor shall the fact of distribution constitute any such warranty, and no responsibility is assumed by TTI or TxDOT in connection herewith.
+dataQualityInformation
DQ_DataQuality
scope:
. DQ_Scope
. level: dataset
+lineage
.LI_Lineage
.statement: Source Data History: The Austin District of TxDOT sends compressed comma-separated value (csv) files that are organized into different folders by freeway corridor or system controller unit (SCU). Within each freeway corridor folder, there should be a *.zip file for each day of the year, with the filename convention "mmddyy.zip". Within each *.zip file, there should be 24 files (one for each hour of the day) that contain detector data for that corridor/SCU. Each hourly file has a descriptive long-format name, consisting of the SCU location name, the day of week, and the hour. The filename extension is ".DET" for detector. For example, "IH 0035 SCU Wednesday 1300.DET" contains detector data for the IH-35 SCU for the "1300" hour (13:00-13:59) on a Wednesday.

Figure 4.1 (contd.). Example of Data Quality Documentation Using ISO 1915

Once date stamps have been added to the filename, we can then use SAS to import the CSV text files. We have developed "aus_reformat.sas" for this purpose. The SAS program "aus_reformat.sas" uses a csv template (e.g., "aus_2001_US0183.csv") for each corridor that contains the hourly files to be processed and the corresponding dates. This program combines all original source data (1-minute) for each corridor for the entire year into a single SAS dataset. Thus for 2001 we have 4 SAS datasets, with the filename convention "aus_2001_RR####". These 4 datasets are then compressed for long-term storage. The data are then ready for the next process step. In summary, the pre-processing is as follows:

+ unzip original files to folder corresponding to week number of the year using “aus_unzip.xls”
+ use batch processing and CKRename to change the weekday name to a date stamp, then compress and store these “date stamp added” text files
+ use “aus_reformat.sas” to import the text files into SAS datasets by freeway corridor/SCU
+report
. DQ_Completeness
. nameOfMeasure:
. DQ_Percent_Complete
. measureDescription:
. value: the degree to which data values are present in the attributes (e.g., volume and speed are attributes of traffic) that require them (also referred to as availability); defined as: (the number of records or rows with valid values present) divided by the (total number of records or rows that require data values)
. evaluationMethodType:
. value: statistical quality control
. evaluationMethodDescription:
. value: computed automatically by data quality software
. evaluationProcedure: “Traffic Data Quality Measurement, Final Report, 2004”
. result:
. DQ_QuantitativeResult
. value: Volume and occupancy data are 99% complete. Speed data are 98% complete
. dateTime:
. value: All of calendar year 2003

. DQ_Accuracy
. nameOfMeasure:
. DQ_Accuracy_RMSE
. measureDescription:
. value: the measure or degree of agreement between a data value or set of values and a source assumed to be correct, as measured by the root mean square error

$$Root\ Mean\ Squared\ Error,\ RMSE\ (\%) = \sqrt{\left(\frac{1}{n}\right) \times \left(\sum_{i=1}^{n} (x_i - x_{reference})^2\right)}$$

. evaluationMethodType:
. value: statistical quality control
. evaluationMethodDescription:
. value: the accuracy of traffic volume values from Sensor 111A was compared to a nearby permanent traffic recorder (Station 075000) that was calibrated on week before the test. Hourly volumes are the basis of comparison.
. evaluationProcedure: “Traffic Data Quality Measurement, Final Report, 2004”
. result:
. DQ_QuantitativeResult
. value: the root mean squared error was calculated as 131 vehicles
. dateTime:
. value: tests were conducted from June 24 through June 27 for all hours of the day.

. DQ_Validity
. nameOfMeasure:
. DQ_Percent_Validity
. measureDescription:
. value: the degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable value; defined as the percent passing a series of quality control checks
. evaluationMethodType:
. value: statistical quality control
. evaluationMethodDescription:
. value: computed automatically from data quality software
. evaluationProcedure:
. value: 14 data quality control checks performed; see Exhibit 3-5 of "Monitoring Urban Roadways in 2001: Examining Reliability and Mobility with Archived Data
. result:
. DQ_QuantitativeResult
. value: Volume and occupancy data are 100% valid. Speed data are 99% complete
. dateTime:
. value: All of calendar year 2003

. DQ_Timeliness
. nameOfMeasure:
. DQ_Percent_Timely_Data
. measureDescription:
. value: the degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values; defined as: (the number of records or rows with values meeting validity criteria) divided by (the total number of records or rows subjected to validity criteria)
. evaluationMethodType:
. value: statistical quality control
. evaluationMethodDescription:
. value: computed automatically by data quality software
. evaluationProcedure: "Traffic Data Quality Measurement, Final Report, 2004"
. result:
. DQ_QuantitativeResult
. value: Volume and occupancy data are 100% timely. Speed data are 99% complete
. dateTime:
. value: All of calendar year 2003

. DQ_Coverage
. nameOfMeasure:
. DQ_Electronic_Surveillance_Percent_Coverage
. measureDescription:
. value: the degree to which data values in a sample accurately represent the whole of that which is to be measured; defined as the percent of centerline miles under electronic surveillance
. evaluationMethodType:
. value: statistical quality control
. evaluationMethodDescription:
. value: computed automatically by data quality software
. evaluationProcedure: "Traffic Data Quality Measurement, Final Report, 2004"

. result:
. DQ_QuantitativeResult
. value: the percent of Austin-area freeways covered is 13 percent
. dateTime:
. value: All of calendar year 2003
. DQ_Coverage
. nameOfMeasure:
. DQ_Detector_Spacing
. measureDescription:
. value: the average spacing of mainline roadway-based detectors for monitoring traffic flow; calculated as the (the total directional mileage) divided by (total number of directional "stations")
. evaluationMethodType:
. value: statistical quality control
. evaluationMethodDescription:
. value: computed automatically by data quality software
. evaluationProcedure: "Traffic Data Quality Measurement, Final Report, 2004"
. result:
. DQ_QuantitativeResult
. value: the detector spacing in Austin is 0.4 miles
. dateTime:
. value: All of calendar year 2003

4.6 Guidelines for Data Sharing Agreements

4.6.1 Review of Data Sharing Agreements

Data sharing agreements codify the roles, expectations and responsibilities among the parties providing and using traffic data. Such agreements can conceivably occur between public entities, entirely between private entities or between private and public entities. In developing the guidelines for data sharing, three existing agreements were reviewed. A summary of these three data sharing agreements is presented below.

SMART Roads

The Virginia Department of Transportation (VDOT) has developed a set of "guidelines for access" to data from the five electronic traffic monitoring sites VDOT operates, under its SMART Roads system. The guidelines apply to new public/private partnerships between distribution providers (VDPs) and VDOT. The VDPs gain access to the traffic management centers and can resell the images collected to third parties, such as television stations. They can also install new equipment within the highway right-of-way. In return, the VDP must advance and support VDOT's goals for improved mobility and, more specifically, must provide free access to the video images through a web site. The only requirement relating to data quality is that the video images be refreshed at a rate of more than one frame per second. This document states that separate contracts will be entered into with individual firms who succeed in their bids to become partners with VDOT.

TRAVinfo

The San Francisco-based TravInfo provides basic ATIS services through a telephone traveler advisory system, which alerts users to incidents, accidents and congestion on the freeway system. Callers are also able to receive up-to-the-minute route-specific information, and are able to connect to all Bay Area transit and ride-share providers. Registered private sector entities are allowed to access TravInfo's open architecture database to provide value-added information on web pages, in-vehicle map displays, or personal digital assistants.

The engineering firm, PB/Farradyne (PBF), is under contract to manage the current ATIS system. The TravInfo contract with PB/Farradyne details "basic" and "enhanced" functional requirements for all aspects of the ATIS operation. Basic data requirements describe the types of data collected and the level of detail and accuracy required. Link speeds for example, are required to be accurate to within 25 percent of actual speeds. Incident data must be posted within one minute of accident verification. Basic data fusion requirements include quality controls for accuracy, timeliness, reliability and usefulness. Enhanced data requirements specify the extent of the data collection effort. Interestingly, these data quality requirements are not extended to third party data consumers.

PBF is responsible for entering into and managing data sharing agreements with third party users, known as registered data disseminators (RDDs). The RDDs are entitled to redistribute, enhance, repackage, or otherwise add value to the data they receive. The data sharing agreement goes to great length to indemnify the public sector data providers and PBF from responsibility for the quality of the data delivered and in fact warns the RDD that "information availability and data accuracy are all subject to change."

Las Vegas

The Las Vegas Area Computer Traffic System (LVACTS) developed a closed circuit video surveillance system for congestion management and accident and signal failure identification on the arterial roadway system 1993. LVACTS's data sharing agreement sets the broad terms for access to the live video images from the system to third parties. The video images are made available for the cost of the access connection; the agreement also states that a monthly subscription fee to defray the operating cost of the traffic management center may be applied. In the subscription agreement, LVACT agrees to provide the same video feed to all subscribers and retains control over the operation of the cameras, the traffic management center and the transmission equipment. The agreement also sets the specific terms of the permitted data uses and the actual charge. The subscribers are responsible for installing and operating any equipment needed for accessing the video feed, which cannot be resold to anyone who is not a party to the subscriber agreement. Finally, the agreement makes no mention of who is responsible for the quality of the data being transmitted nor are data quality standards specified. However, the agreement does contain a broad disclaimer indemnifying LVACTS from misuse or negligent use of the data.

Prior to any agency or company initiating a data sharing program, an agreement between the two parties must be negotiated and signed. This agreement is needed to define the expectations of both parties, a description of the information to be shared, the responsibilities of each party in the transaction, the limits of use or reuse of the data, any required procedures to send or receive the data and liability responsibilities.

Summary

Three themes emerge from a review of three data sharing agreements:

- Preservation of privacy rights and control of the flow of data by the public sector is very important. Many traffic management centers have an explicit policy prohibiting their operators from using the video surveillance system to obtain detailed images of vehicles or individuals, unless explicitly ordered to do so by law enforcement officers. In the past, some traffic management centers have prohibited the distribution of video data from its system due to concerns over privacy rights and the potential for litigation over infringement of those rights.

- The agreements seek to maintain some level of free public access to the data.

- Traffic data sharing agreements in use today offer no guarantees about data quality and offer the data on an "as-is" basis.

A review of data sharing agreements conducted for this project found that most existing agreements concerned the sharing of video images. Two agreements were reviewed that specifically address data other than video images. These are agreements developed by Virginia DOT and the Metropolitan Transportation Commission (MTC) in the San Francisco Bay Area. An excerpt from the MTC agreement makes the following statement concerning data quality:

> *"PBF, MTC, Caltrans, and CHP and their suppliers make, and Registered Data Disseminator receives, no warranty regarding Provided Data, whether express or implied, and all warranties of merchantability and fitness of provided data for any particular purpose are expressly disclaimed. PBF, MTC, Caltrans, and CHP and their suppliers make no warranty that the information will be provided in an uninterrupted manner or that the Provided Data will be free of errors. Provided Data is provided on an "as is" and "with all faults" basis, with the entire risk as to quality and performance with Registered Data Disseminator."*

The VDOT agreement does not address data quality. The agreement does make the following statement about video image quality:

> *"VDOT makes no warranty that the imagery will be provided in an uninterrupted manner. Imagery will be provided on an "as is" and "with all faults" basis."*

Data quality can be addressed in data sharing agreements by including clauses that provide one of several levels of guarantee, including the following:

1. The provider does not warrant quality of data stating that data is provided "as is" and "with all faults," which seems to be the current state-of-the-practice in the ITS industry. The "as is" approach can (and should) include descriptions of the provider's quality control and quality assurance procedures.

2. The provider will provide the user with a set of data quality indicators, i.e., accuracy, completeness, validity and coverage as described in the previous section of this report. These indicators can be furnished with each data file for periodic downloads or on a daily basis for continuous data flows.

3. Provider agrees to meet certain data quality standards. As providers gain experience with data management they may become comfortable with data checking and quality assurance techniques and be willing to provide data with an assurance that it meets a specific standard for any or all of the data quality attributes described above.

The data quality clause included in the agreement should apply to both public-public agreements and public-private agreements.

4.6.2 Data Quality Provisions in Data Sharing Agreements

As noted above, data quality specifications rarely appear in data sharing agreements between the end user and the data provider. Data sharing agreements typically discuss such items as security and confidentiality, liability, frequency of data transmittals, to whom the data may be disseminated, and fees. However, public sector end users are unlikely to adopt ITS data for their applications on a widespread basis without some assurances that the data meet some minimum standards consistent with current expectations. This section offers guidance on how data quality provisions can be added to data sharing agreements; the entirety of data sharing agreements is not discussed here.

4.6.3 Model Data Quality Sections of Data Quality Agreements

Data providers in data sharing agreements can be either public or private agencies. The same goes for data recipients. Thus, four types of agreements are possible: public-to-public-to-private, private-to-public, and private-to-private. Ignoring other terms of data sharing agreements (such as liability, restrictions on use) and focusing strictly on data quality, there is not much difference in how data quality would incorporated into any of these arrangements. The key decision in structuring data quality clauses is to what extent minimum acceptable data quality criteria are established and enforced. Conceptually, three levels exist for this type of specification:

1. ***Level 1: Reporting/documenting the quality of the data***. At this level, the six quality attributes (defined in this report) are transmitted with the actual data. Examples of how this can be achieved are presented in the "Metadata" section later in this chapter. Quality-related metadata provided with the data files will indicate to the data user whether the data meets quality standards necessary for that application and will assist the user in determining any additional data processing or manipulation needed. At the same time the data generating agency is not required to conduct data processing that may not be needed by the specific user or application. In the future, after the ITS industry has more experience with data sharing and archiving, quality metadata standards may be adopted and ITS data files may be required to meet those standards.

2. ***Level 2: Specifying what the quality of the data must be.*** Acceptance criteria for ITS data should coincide with existing criteria used by traffic monitoring systems. It is reasonable to treat at least a sampling of data collection points as a permanent count stations, and to apply the minimum standards used by FHWA for permanent count station reports. Table 4.3 below presents some suggested minimum data acceptance standards for the incorporation of ITS-generated traffic data into traffic monitoring programs for planning and engineering purposes. However, specifications of hard standards (minimum acceptable quality levels) may or may not be desirable, based on the application and the entities involved. Because ITS systems offer much more comprehensive temporal and spatial coverage, entire corridors or routes can be analyzed. These acceptance thresholds cover roadway segment and intersection approach locations according to the amount of data that should be collected and the accuracy of the data.

   The specification of the actual tests to be conducted to determine data quality is extremely important if minimum quality criteria are established. This needs to be done for all six quality attributes, and is particularly important for accuracy and validity. The frequency of the testing also needs to be specified. Figure 4.3 shows an example of a how this may be done in a data sharing agreement. This is a proposal that has not been tested nor validated.

Table 4.3. Standards for Data Transfer Agreements

Type of Location		Proposed Minimum Quantity Standard	Proposed Quality Standard
Roadway sections	Single location	Seven consecutive days per month	
	Single corridor	100 percent coverage one day per month	Daily count within 10 percent of machine or manual count within 15 percent of hourly count as measured once per year. Twenty percent sample of locations.
	Areawide	75 percent coverage one day per month	Daily count within 10 percent of machine or manual count within 15 percent of hourly count as measured once per year. Five percent sample of locations.
Intersections	Single location	Seven consecutive days per month	N/A
	Single Corridor	100 percent coverage one day per month	Five and 10 percent standard applied every five miles in corridor once time per year. Five percent sample of intersection locations.
	Areawide	75 percent coverage one day per month	Five and 10 percent standard applied to one location per corridor per year. One percent sample of locations.

3. ***Level 3: Structuring payment schedules based on amount of data passing minimum criteria.*** In some cases, such as when the private sector is the data provider, it may be desirable to structure payment clauses based on the amount of data that meet or exceed minimum quality criteria. Such an arrangement provides incentives to the provision of quality data. Two options are available: (1) "all-or-nothing", in which data must meet all quality criteria or payment is not rendered and (2) "sliding scale" or "award fee", where payment is based on the amount of data at different quality levels. For example, extending the information in Figure 4.2, the following could be potential graduated payment schedule.

3. DATA QUALITY FOR ITS-GENERATED VOLUMES AND SPEEDS (Note: text in italics indicate options)

3.1 Reporting Data Quality*. The data to be supplied under this agreement shall be reported using the latest metadata standard developed for archived ITS data by the American Society for Testing and Materials.*

3.2 Minimum Data Quality Criteria. *All tolerances refer to the testing methods in Section 3.3. The definitions of these attributes appear in "Traffic Data Quality Measurement, Final Report, 2004".*
3.2.1 Accuracy. *Volumes shall be certified to be within a tolerance of +/- 10%. Speeds shall be certified accurate to within 5 mph.*
3.2.2 Completeness. *Volume and speed data shall be at least 90% complete as received from the field prior to any* post hoc *error checking.*
3.2.3 Validity. *At least 85% of volume and speed data shall pass validity checks.*
3.2.4 Timeliness. *Data shall be submitted no less than seven days after they are collected. Timeliness statistics as defined in "Traffic Data Quality Measurement, Final Report, 2004" shall be developed.*
3.2.5 Coverage. *The volume and speed data shall be collected on the following corridors:* {list corridors with beginning and ending mile points or cross streets}

3.3 Tests to Determine Data Quality and Frequency of Reporting
3.3.1 Accuracy. *Each field measurement device shall be tested and reported for accuracy every six months. Tests will be run for 15-minute time intervals for a weekday peak period and a daylight off-peak period. Volumes shall be collected using a video device and vehicles shall be manually counted at a later time. Speeds will be collected using a portable RTMS, sonic, or video image device, or other Department-approved that has been calibrated in accordance with Department standards.*
3.3.2 Completeness. *See "Traffic Data Quality Measurement, Final Report, 2004". Completeness statistics shall be reported on the 5th of every month for the previous month.*
3.3.3 Validity. *Data will be subjected to the following quality control tests* {list specific tests; examples include those developed for the Mobility Monitoring Program and ADMS-Virginia}. *Validity information will be submitted for each data record received in accordance with* {the latest ASTM standard on metadata}.
3.3.4 Timeliness. *Refer to section 3.2.4. Monthly reports indicating timeliness statistics shall be submitted.*
3.3.5 Coverage. *Refer to section 3.2.5.*

Figure 4.2. Example Language for Specifying Minimum Data Quality Criteria in a Data Sharing Agreement

Payment of the contract amount shall be determined based on the percentage of volume data that annually pass a composite accuracy, completeness, and validity score as follows. The combined score is calculated as the product of the accuracy, completeness, and validity tests:

Composite Score	*% of Contract Amount*
75-100%	*100%*
50-74%	*75%*
30-49%	*50%*
15-29%	*25%*
< 15%	*0%*

Note that other quality measures can be used in computing the composite score. The choice of measures could be driven by the application or the source of the data. Also note that the graduated scale presented above is for illustration purposes only. This concept has not been tested.

5.0 Beta Testing of Framework and Review of Guidelines

The purpose of the beta testing phase of the project is to test the concepts, framework, including the methodologies, as well as the guidelines for assessing traffic data quality. It was expected that data from actual projects of State DOTs would be used for testing the applicability of the data quality assessment methodologies. In this regard, the framework was sent out to selected individuals in DOTs of the states of Florida, Georgia, Illinois, Maryland, Utah, Virginia, and Washington to review and where possible apply the framework using their individual local data. The original intent of testing the framework with state data was abandoned when it became clear that a fully fledged beta testing is quite an unreasonable demand on the public agencies. It became necessary to request only reviews of the concepts presented in the framework. A few review comments were received and these are summarized below.

In developing the guidelines for implementing the framework, estimates of the level of effort required to establish a data quality assessment system and straw man estimates of data quality acceptability levels were developed. Review comments on the guidelines and reality checks on the straw man estimates were sought from various offices of the U.S. DOT, FHWA and a few state DOTs. The review comments are also presented below. The estimates of the levels of effort and acceptability levels were revised based on recommendations by the reviewers.

5.1 Framework Review Comments

Out of the seven state DOTs included in the beta testing, four provided written comments on the framework as whole. None of the states actually applied the framework to their data. The following are detailed review comments on the framework.

Florida DOT

Florida DOT (FDOT) agrees that the quality measures are useful but noted that some may be difficult to calculate. Completeness and validity measures can be easily calculated for continuous counters. FDOT believes that it would not be practical to compute the accuracy of counters by routinely performing manual counts and comparing with machine counts. This is because (i) it is very time consuming to count from a video, (ii) manual counts are very error prone, and (iii) it is very difficult to synchronize the times on the video to the times in the permanent counter.

As far as timeliness measure is concerned, as long as the data resides in the database at the time it is extracted for processing, it is considered to be timely.

FDOT estimates 100 percent coverage of the state highway system every year, because those few roads that cannot be counted due to construction are estimated by applying a growth factor to the latest measured year.

Georgia DOT

Georgia DOT (GDOT) agrees with all of the quality measures and noted that the framework is well-presented and easy to follow. GDOT is currently undertaking a similar exercise in assessing data quality and determining archiving processes and needs. The results to date are generally similar to those presented by the case studies in the framework.

Washington State DOT

Washington DOT (WsDOT) noted that timeliness definition works for their ATR sites. However, for short counts (AADT - archived data), AADTs are calculated only at the end of year (usually May-June timeframe to coincide with HPMS submittal). This is because factors based on ATR sites are not available until February/March. WsDOT observed that coverage should also mention group factors (sufficient number of stations to accurately develop group factors).

Washington state has started using Data Stewards, data dictionaries, and providing a "data mart" for ISP's to access the data they need. WsDOT is pleased to see this concept in the framework.

California DOT

California DOT (Caltrans) has just deployed a quality check program for their WIM data that incorporates the accuracy, completeness and validity measures outlined in the framework.

5.2 Guidelines Review Comments

Beta testing of the guidelines focused on validating the straw man estimates of the acceptability levels of the quality measures and the estimated level of effort to implement a data quality assessment system within an agency. The following are comments on the draft guidelines.

Minnesota Department of Transportation (MnDOT) noted that data quality and anticipated variance in data quality are related to so many variables that, establishing parameters and targets is extremely difficult. The Office of Transportation Data and Analysis at the MnDOT have traditionally approached data quality from at least three perspectives:

- **Data Inputs:** Certainly, primary factors in data quality center on the accuracy, completeness, validity, timeliness, coverage and accessibility of data inputs. The variability of traffic count data and traffic forecasting results are challenged and influenced by a host of factors including the type of roadway and its current AADT; the reliability of traffic data collection equipment; the ability to capture and incorporate information on trip making, demographic, land use and traffic generator changes; the precision of hourly, weekly and monthly adjustment factors; the accuracy of axle correction factors; and the robustness of the program to indicate when recounts are warranted and the availability of staff to perform recounts. Even when acceptance criteria are established for individual sites or types of roadways, there is no guarantee that the criteria will continue to be applicable over time as travel behavior at individual sites change.

MnDOT have established a "census" cycle for collecting traffic data on what is called "uniform traffic data segments" in Minnesota. The counted system consists of Interstates, U.S. and Minnesota highways, County Highways and Municipal State Aid roadways. To help assure data quality MnDOT has established "customer driven" screening criteria based upon AADT ranges to assist us in determining when recounts of traffic volumes are desirable. The attached chart shows a "general" guide to assist MnDOT's traffic count program administrators and field personnel who collect traffic count data. Comparisons are made between new, adjusted traffic count volumes and earlier annual estimates of AADT when individual sites were actually counted. Recounts are taken as program priorities, budget and time allow.

- **Data Uses:** Another approach for considering data quality is related to how well the data match the needs of the users. Traffic data are used throughout the transportation community for planning, investment analysis, project development, environmental analysis, pavement design, operations and maintenance. MNDOT believes that data quality targets should be tied at some level to the sensitivity of the decisions they support. As a transportation data community, there is the need to work to ensure that these sensitivities are more clearly defined, articulated and universally understood by all of the stakeholders involved in providing and using traffic data.

- **Performance of Data Results Over Time**: A third factor for considering the quality of traffic data has to do with how well what data collected and especially how the forecast matches actual trends and future volumes. MnDOT believes that most transportation agencies monitor their traffic data collection efforts over time and perform various trend analyses. Trend data outside expected parameters typically receive more scrutiny. In respect to travel demand modeling and forecasting, Mn/DOT Metro District Office in the Minneapolis-St. Paul metropolitan area uses a confidence interval to assist in evaluating the accuracy of their traffic forecasting and traffic modeling efforts. Currently, they are using a confidence interval of plus or minus 15%. The 15% range was based in part on an analysis of how well 20-year traffic forecasts done in the 1960's and 1970's for Twin Cities freeways compare to actual volumes in the forecast year. MnDOT believes that one may wish to consider adding an attribute to the list that deals with how well the data performs over time.

Acceptable Percent Change

Past AADT	*Decrease*	*Increase*
0-19	*-100%*	*+400%*
20-49	*-40%*	*+50%*
50-99	*-30%*	*+40%*
100-299	*-25%*	*+30%*
300-999	*-20%*	*+25%*
1000-4999	*-15%*	*+20%*
5000-49999	*-10%*	*+15%*
50000+	*-10%*	*+10%*

** A recount is requested if adjusted volumes exceed acceptable percent change margins.*

FHWA Offices

The following office of FHWA reviewed the straw man estimates of acceptable levels for the quality measures and levels of effort. The suggested changes to the initial estimates were incorporated.

- Traffic Monitoring and Surveys Division: commented on the acceptable levels for the completeness measure and offered suggestions.

- HPMS Division: provided guidance on the acceptable levels for accuracy measure for traffic counts on rural and urban highways, completeness for portable machine counts, and VMT.

- Office of Safety on Highway Safety section: noted that annual VMT is used to calculate the highway safety rates (e.g., fatality and crash rates) and the data is needed in 1-year or less for Federal, State and local statistical and program purposes. It was recommended that this data and time be added to the Highway Safety section of the guidelines. It was also noted that the Daily VMT (i.e., DVMT) is not a relevant static for safety analysis. Several other changes were recommended.

5.3 Other Comments

There was some discussion about the data completeness measure. It was observed that a single statistic completeness measure does not distinguish completeness over time from completeness at the same time. For example, consider two extreme cases in which 95 percent of the possible readings are provided i.e., completeness (or availability is 95 percent). In the first case, data from 95 percent of the sensors is always available but is never available from the other 5 percent of the sensors. In the second case, data from all the sensors is available 95 percent of the time. The two cases differ significantly from each other, but both of them provide an overall availability of 95 percent. By this example, it would appear that a single statistic combines spatial and temporal completeness and can be misleading. It was noted that completeness needs to be addressed in greater detail i.e., a user needs a report of when and where data are available.

It is important to note that a single completeness statistic as illustrated above is not misleading. A single completeness statistic provides an indication of the magnitude of data completeness in the same way a national congestion statistic does not provide any details on congestion at specific locations. It is desirable to have more detailed completeness statistics and this can also be achieved by showing several levels of detail for completeness measure. It is conceivable that operations managers and other mid-level managers would appreciate a single completeness statistic (and not having to run calculations on their own to get a single statistic). Based on the single statistic, one can determine whether further details are required to characterize the quality of the data. Similarly, they could be links to other levels of detail e.g., “completeness by day” or “completeness by road” or something similar that permits detailed analysis by the desired category.

The framework is developed to provide a single statistic of the completeness measure. The single measure is intended to provide an “overall” measure of completeness of data. Detailed information can then obtained about the spatial and temporal variability of completeness if desired.

6.0 Conclusions and Recommendations

Data quality is directly based on the extent to which a data set satisfies the needs of the person judging it. A better understanding and means to assess the quality of data offers various benefits including confidence and efficacy in decisions based on data. This project developed a framework and guidelines for measuring and assessing the quality of traffic data for different applications. The framework is comprehensive in providing alternative methods and tools for calculating the six fundamental data quality measures that would allow traffic data collectors and users to determine the quality of traffic data they are using, providing, sharing. The case studies used to illustrate the application of the framework are selected to represent a diverse range of data sources and applications. The guidelines include guidance on quality targets, levels effort required to establish a data quality assessment system within an agency, approaches for including metadata with data quality, and standards for data sharing agreements. The examples for metadata and proposed standards for data sharing agreements provide useful guidance in those areas.

The beta testing although limited has provided the opportunity to validate the concepts and methodologies presented in the framework and also validate some straw man estimates of data quality targets and estimates of the levels of effort. Overall, feedback from the beta test indicates that data quality assessment is important and that the framework provides necessary and useful tools to measure and assess the quality of traffic data.

The estimated levels of effort and quality targets need to be tested and validated based on actual experiences in the use of the framework and guidelines. Even though these have been validated through limited beta testing, more extensive validation is recommended.

7.0 References

Barton Aschman and Cambridge Systematics, *Model Validation and Reasonableness Checking Manual*, Report to, FHWA, Travel Model Improvement Program, February 1997.

Chatterjee, A., et al. *Improving Transportation Data for Mobile-Source Emissions Estimates* (NCHRP 25-7). Washington, D.C.: National Cooperative Highway Research Program, 1995.

Content Standard for Digital Geospatial Metadata, Metadata Ad Hoc Working Group Federal Geographic Data Committee, 590 National Center Reston, Virginia 20192.

Draft International Standard ISO/DIS 19115, ISO Central Secretariat 1 rue de Varembé 1211 Geneva 20 Switzerland.

Evaluation of Non-Intrusive Technologies for Traffic Detection, available at http://projects.dot.state.mn.us/nit/index.html.

Fekpe, E. and Gopalakrishna, D. *Traffic Data Quality Workshop Proceedings and Action Plan* Report to FHWA, U.S. Department of Transportation, Battelle, June 2003.

FGDC-STD-001-1998 available at http://www.fgdc.gov/metadata/contstan.html

Hodgson, Katrina, *Metadata: Foundations, Potential and Applications*, School of Library and Information Studies, University of Alberta, March 1998 available at http://web.archive.org/web/20020814115740re_/www.slis.ualberta.ca/538/khodgson/metadata.htm.

Joint Task Force on Traffic Monitoring Standards of the AASHTO Highway Subcommittee on Traffic Engineering, *AASHTO Guidelines for Traffic Data Programs*, American Association of State Highway and Transportation Officials, 1992.

Toppen, A. and Wunderlich, K. *Travel Time Data Collection for Measurement of Advanced Traveler Information Systems Accuracy*, FHWA, U.S. Department of Transportation, June 2003. Available at http://www.its.dot.gov/itsweb/welcome.htm, EDL Document # 13867.

Transportation Research Board, *Development of the 2002 Guide for the Design of New and Rehabilitated Pavement Structures*, NCHRP 1-37-A, National Cooperative Highway Research Program, , National Research Council, Washington, D.C., March 2004

Turner, S. *Defining and Measuring Traffic Data Quality*. Available at the ITS/JPO Electronic Document Library http://www.its.dot.gov/itsweb/welcome.htm, EDL Document # 13767.

U.S. Environmental Protection Agency, *Procedures for Emission Inventory Preparation*, Volume VI: Mobile Sources, December 1992.

Wunderlich, et al., *Urban Congestion Reporting*. Ongoing task for the FHWA, U.S. Department of Transportation, Washington, D.C.

Wunderlich, K., et al., *On-Time Reliability Impacts of Advanced Traveler Information Services: Washington, DC Case Study*, FHWA, U.S. Department of Transportation, January 2001.

APPENDIX A:

AUSTIN, TEXAS CASE STUDY

AUSTIN, TEXAS CASE STUDY

Introduction

The following sections describe procedures for calculating these data quality measures in a specific setting: traffic data collection, dissemination and archiving in Austin, Texas. Because the exact calculation of the data quality measures may vary depending upon the specific data consumer, it is useful to first identify who will be primary data consumers in the traffic data flows (via the National ITS Architecture at http://itsarch.iteris.com/itsarch/) and thus whose perspective should be represented in calculating data quality. Readers should note that most of the information and details in this case study example are accurate and true; however, some details and results have been embellished or simplified for the purposes of the example.

Traffic Data Flows: Identifying the Data Consumers

Figure A.1 shows a simplified version of the physical entities (as defined in the National ITS Architecture) involved in traffic data collection, dissemination, and archiving in Austin, Texas. Figure A.2 illustrates the data flows from another perspective, with additional detail related to the specific context of Austin traffic data. In this example, there are 5 primary data consumers whose perspectives should be represented in calculating data quality measures:

- Traffic operations personnel: use original source data
- Archived data administrator: uses original source data
- Information Service Providers (ISPs): use original source data
- Archived data users: use archive database
- Travelers: use traveler information

In this example, the archived data administrator and the ISP use the exact same data stream (i.e., original source data) as the traffic operations personnel. Thus, these three data consumers should share a common definition for data quality measures since their data is identical (they may, however, have different views of what quality level is acceptable for the six measures). From Figure A.2, it can be seen that there are three different types of traffic data for which we should calculate data quality:

- **Original source data –** refers to original data (this could be real time or archived) collected from various traffic data collection devices. Typical source data includes volumes, speeds, occupancy, vehicle classification, and travel times.

- **Archive database –** refers to data stored in an archive database. This dataset is derived from original source data and can be processed or in its original raw state. For example, original source hourly volume data is checked for quality and then converted to AADTs using adjustment factors and archived.

- **Traveler information –** refers to data provided as information to travelers. This is usually real time information and is derived from the original source data. In some cases, historical archived data may also be used to generate traveler information.

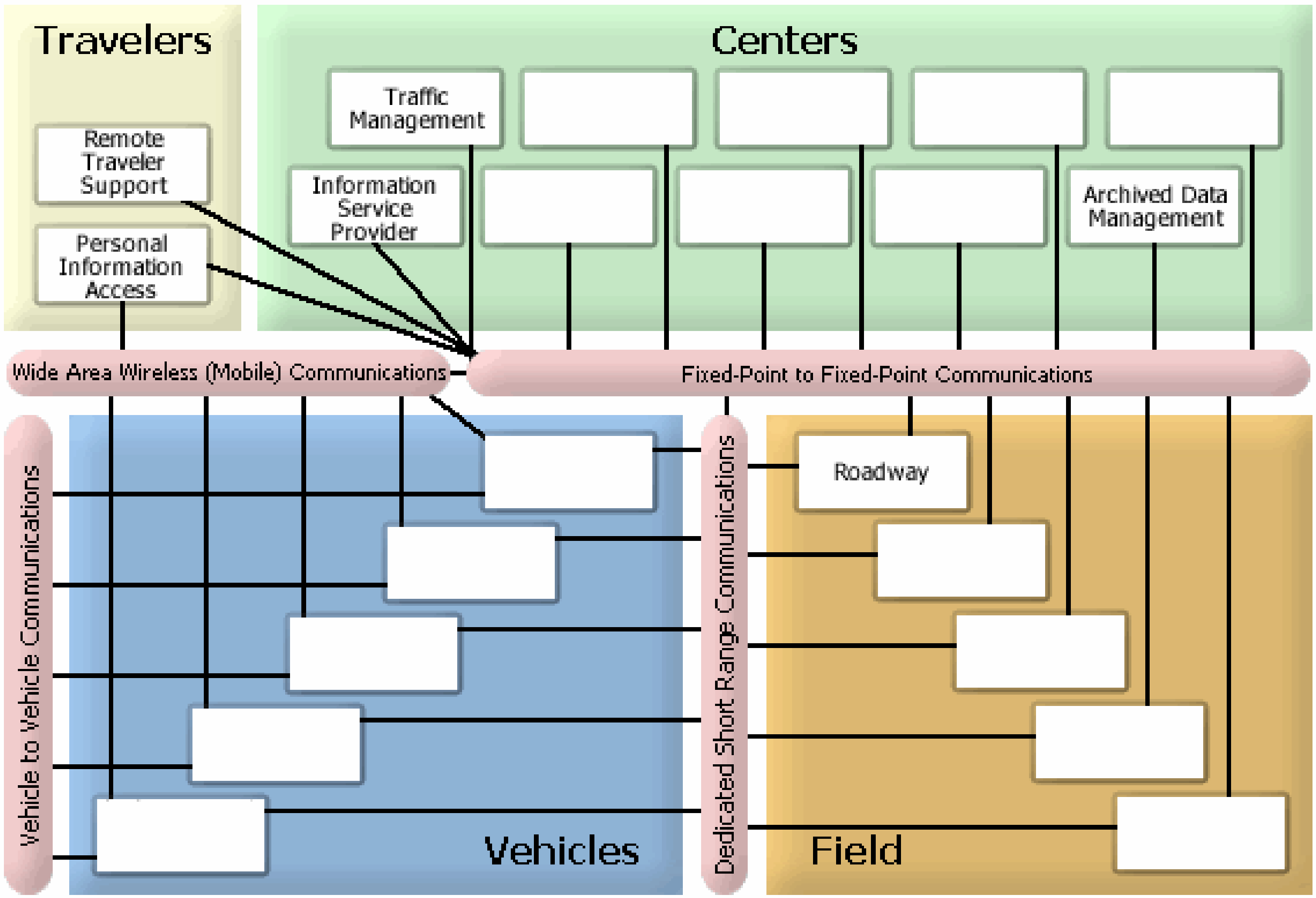

Figure A.1. Simplified Austin Case Study Mapped to National ITS Architecture

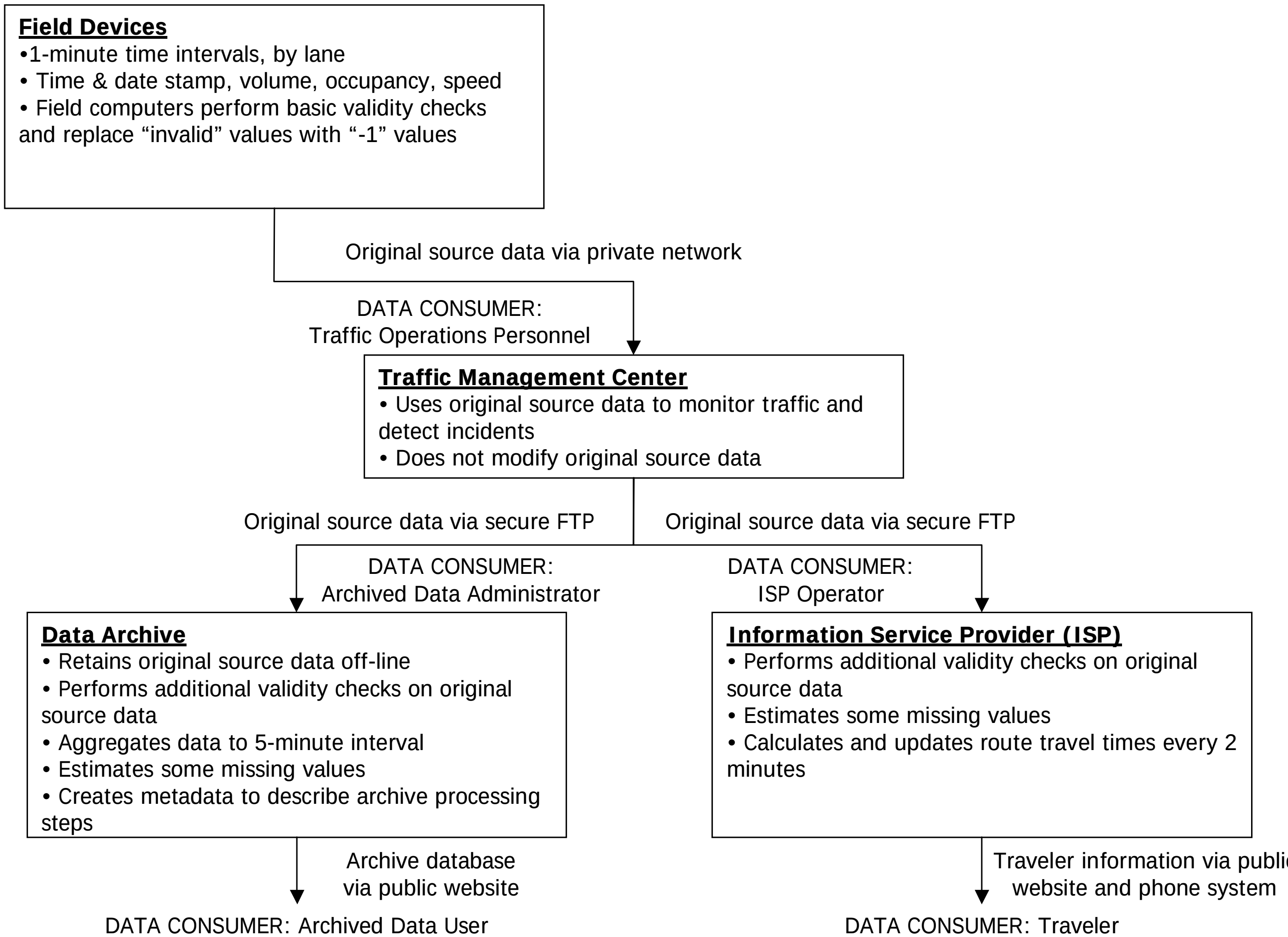

Figure A.2. Data Flows and Data Consumers in Austin Case Study

An inherent principle in this methodology is the need to re-calculate data quality when the data has undergone significant change or transformation. Thus, the data quality results at different points in the data flow may be slightly different because the data itself has been modified as it flows from field devices to various data consumers. The following sections describe specific calculation procedures for the six data quality measures for these three different types of data

Calculation of Data Quality Measures

For the Austin case study, we consider a single day of data (i.e., August 29, 2003) collected by the Texas Department of Transportation (TxDOT) as an example. Readers should note that data quality could also be reported for other time scales, such as every hour, week, month, or year. For this particular example day, there were 654 unique single- and double-loop detectors (in which a "detector" measures traffic data for a lane) configured to report lane-by-lane traffic data (i.e., volume, occupancy, speed) at 1-minute intervals. Each 1-minute reading from each detector is considered to represent one record. For example, Figure A.3 shows a sample of the original source data for Austin.

DET_ID	DATE	END_TIME	VOLUME	OCC	SPEED
6009921	08/29/2003	7:00:24	22	8	67
6009921	08/29/2003	7:01:23	27	10	63
6009921	08/29/2003	7:02:23	23	9	68
6009921	08/29/2003	7:03:23	29	10	68
6009921	08/29/2003	7:04:23	19	11	67
6009921	08/29/2003	7:05:23	34	12	68
6009921	08/29/2003	7:06:23	22	12	67
6009921	08/29/2003	7:07:23	28	11	63
6009921	08/29/2003	7:08:23	29	11	67
6009921	08/29/2003	7:09:23	22	8	63
6009921	08/29/2003	7:10:23	18	7	68
6009921	08/29/2003	7:11:23	28	10	66
6009921	08/29/2003	7:12:23	21	12	66
6009921	08/29/2003	7:13:23	29	11	65
6009921	08/29/2003	7:14:23	34	13	66
6009921	08/29/2003	7:15:23	20	7	69

Figure A.3. Sample of Original Source Data for Austin

The Austin data was used in illustrating the calculation of the six data quality measures are described below.

Accuracy

For the purposes of this example, we assume that reference measurements are available for the three different versions of data: original source data, archive database, and traveler information.

Original Source Data

In this example, consider that we would like to know the accuracy of the speed values in the original source data. Ground truth data have been obtained from a calibrated portable reference sensor that was temporarily installed at a representative sample of detector locations. To compute the accuracy of the original source data, we will summarize the reference data to 1-minute time periods to match those of the detector data being tested.

For visual reference, a chart is created that compares the actual speed measurements to the reference measurements (Figure A.4). The mean absolute percent error was calculated as 12.0 percent using Equation 1, and the root mean squared error was calculated as 11 mph using Equation 2.

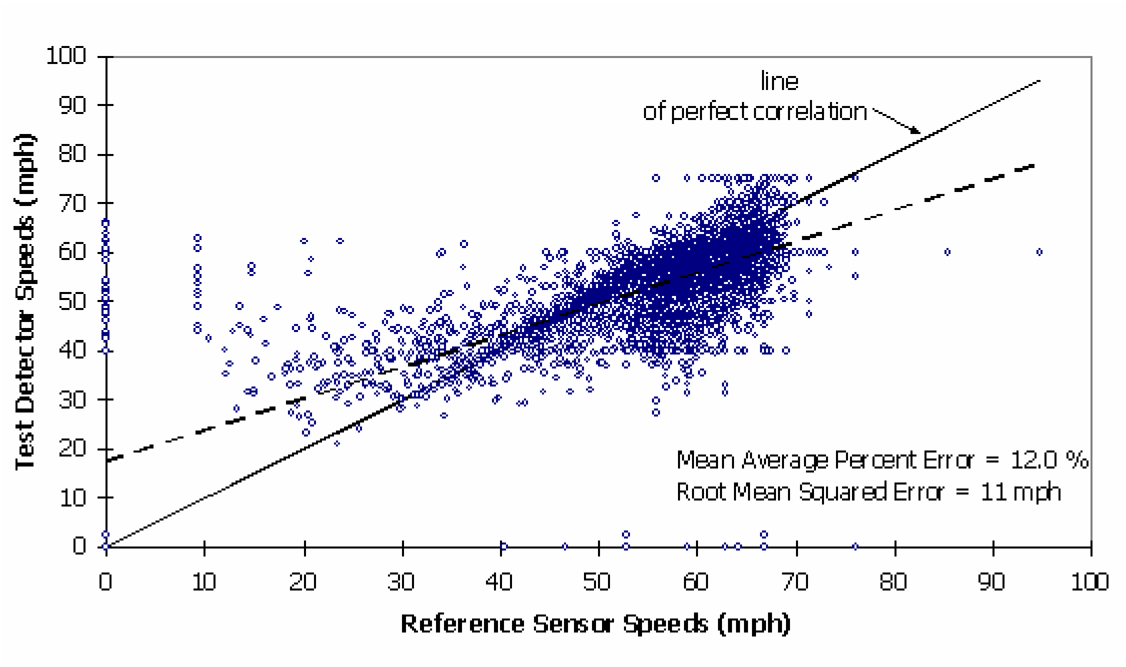

Figure A.4. Accuracy of Speed Values in Original Source Data

Archive Database

In this example, consider that we wish to compare the accuracy of traffic volume values from an operations-based sensor to a nearby automatic traffic recorder (ATR) that has recently been calibrated. One of the many data products available through the data archive are hourly traffic volumes; therefore, the reference measurements are also summed to match the exact date and time of the hourly traffic volumes in the data archive.

For visual reference, a chart is created that compares the volume counts from the archive database to the reference counts from the ATR (Figure A.5). The mean absolute percent error was calculated as 4 percent using Equation 1, and the root mean squared error was calculated as 131 vehicles using Equation 2.

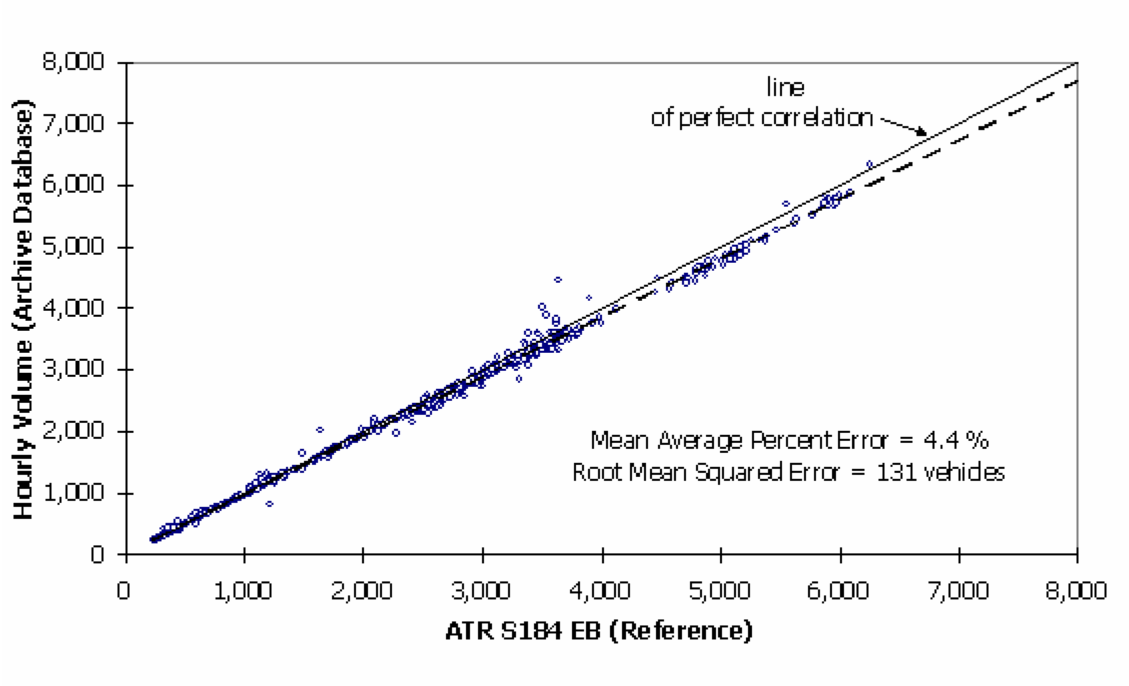

Figure A.5. Accuracy of Hourly Traffic Volumes in Archive Database

In this example, one can see that the accuracy of the hourly traffic volumes in the archive database is fairly good, with the mean absolute percent error being less than 5 percent.

Traveler Information

In this example, the ISP provides route-based speed and travel time reports on its website and through a voice-responsive phone system. The route speeds and travel times are updated every minute in both systems, while the speed and travel time values are based on a rolling 2-minute average. The ISPs also estimates route speeds and travel times if some of the original source data are missing.

As a means to ensure a quality product, the ISP arranges for reference travel time measurements to be obtained along selected Austin routes for various times of the day. The ISP uses the travel time accuracy procedures described in an FHWA report (*Travel Time Data Collection for Measurement of Advanced Traveler Information Systems Accuracy,* EDL Document No. 13867).

The ISP travel times are visually compared to the reference travel times using similar charts (see Figure A.6). The mean absolute percent error was calculated using Equation 1, and the root mean squared error was calculated using Equation 2.

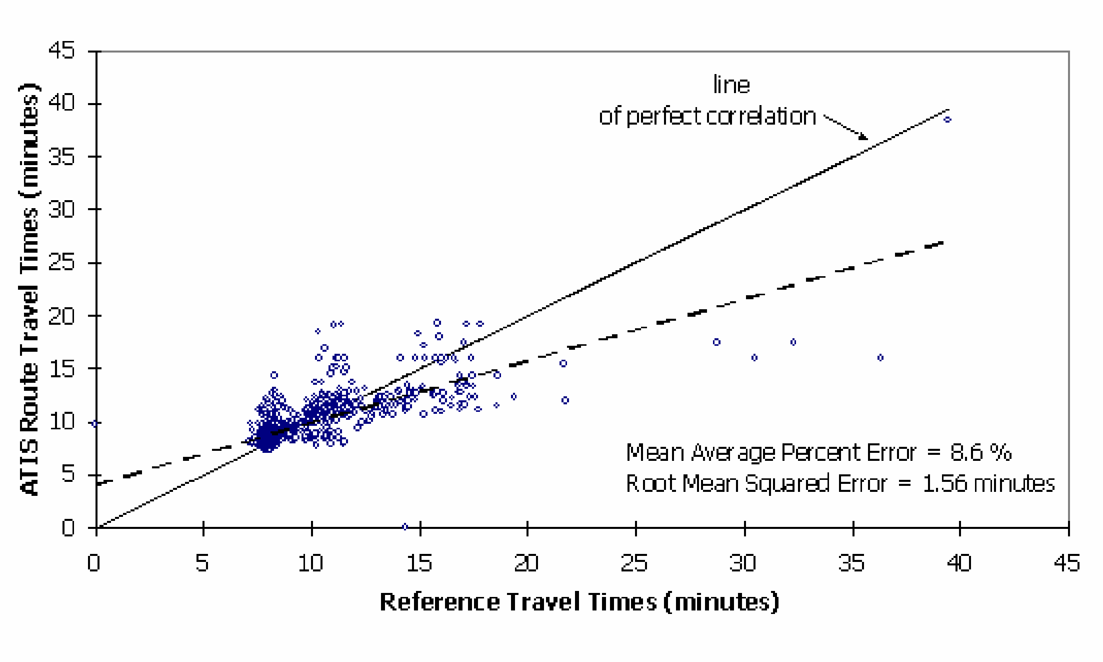

Figure A 6. Accuracy of Route Travel Time Values in Traveler Information

Completeness

In the Austin example, we calculated data completeness for the three different versions of data: original source data, archive database, and traveler information. In this particular example, the data process includes the flagging and eventual purging of invalid data values. Therefore, the completeness statistics will only include valid data values. The potential contribution of invalid data values to the completeness measure can be determined by combining the completeness and validity statistics.

Original Source Data

In Austin, there are 513 on-line detectors that should report a data record every minute for the entire day. Another 141 detectors are configured but "off-line" for acceptance or evaluation testing—these detectors are not counted in the completeness measure because they are not malfunctioning. Of the 513 detectors, 78 are non-trap or single-loop detectors (mostly on ramps and service roads) that only report volume and occupancy values. The remaining 435 detectors are trap or double-loop detectors that report volume, occupancy and speed values. Thus, we expect to have 738,720 valid volume and occupancy records per day (513 total detectors × 1,440 records per day) but only 626,400 valid speed records per day (435 trap detectors × 1,440 records per day). The Austin field computers perform basic validity tests on 20-second data and replace "invalid" values with "-1" values. When computing completeness, however, we only consider valid values and the "-1" values are not considered valid. Table 1 contains the completeness statistics and data used in the calculations. The completeness statistics in Table A.1 indicate that the original source data in Austin is almost fully complete, with only 1 to 2 percent of the data being incomplete (i.e., missing or invalid).

Table A.1. Completeness Statistics for Original Source Data

	Volume	Occupancy	Speed
Number of records with valid values	731,787	731,787	616,991
Number of records that require valid values	738,720 (513 total detectors)	738,720 (513 total detectors)	626,400 (435 trap detectors)
Percent Complete	**99%**	**99%**	**98%**

Incomplete data can be caused by 1) large amounts of invalid data; or 2) missing data due to communication, hardware, or software failures. The completeness statistics must be viewed in combination with validity statistics to pinpoint the most likely cause of missing data.

Data quality reports should fully specify or disclose information related to the amount of expected data (the denominator of percent complete), especially for the completeness measure. Note that in this example, we have observed that 141 detectors are off-line for acceptance or evaluation testing. Malfunctioning detectors should not be discounted from expected data counts simply because device owners are aware of their malfunction but have not been able to repair the devices. The practice of listing malfunctioning detectors by considering them "off-line" is not recommended as it obscures the true device failure rate and data quality results.

Archive Database

As shown in Figure A.2, the archived data administrator retrieves the original source data from the traffic management center. The archive administrator performs several data processing steps in preparation for loading into the data archive:

1. Additional validation checking (beyond what is done by field computers);
2. Aggregation of 1-minute data to 5-minute intervals; and
3. Estimation or imputation of data values for 5-minute intervals with incomplete data.

The additional validation rules (step 1 above) used by the archive administrator are described later when discussing data validity. If a data value from the original source data fails these additional validation rules, the original value is flagged as invalid and not included in subsequent processing steps. Data values of "-1" that were marked as invalid by field computers are also flagged and are not included in subsequent processing steps.

The aggregation step (step 2) combines all 1-minute records within a 5-minute period (e.g., 12:00 to 12:05 am, 12:05 to 12:10 am, etc.) and computes total volume, average occupancy and average speed. Additional attributes are appended to the record to indicate how many valid 1-minute records were included in the 5-minute summary statistics.

Once the preliminary 5-minute subtotals are calculated, a factored volume estimate is calculated based on the number of valid 1-minute volume values in the 5-minute subtotal. For example, consider a 5-minute volume subtotal of 125 vehicles that is based on 4 valid 1-minute volume values. Because one of the 1-minute volume values is missing, the archive administrator calculates a volume estimate for the full 5-minute period based on the 4 minutes of data as follows: 125 vehicles × 5 values expected ⁄ 4 valid values = 156 vehicles. This estimated volume count is marked as an estimate and the estimation method is documented in the archive metadata. Five-minute average occupancy and speed values with less than 5 valid values are not factored up since they are averages and not sums (as is the case with volumes). Five-minute time periods with no valid values are left as missing or null and no estimates are provided.

After the archive administrator has performed these processing steps, completeness statistics are computed by counting the valid data values in the data archive. With 5-minute subtotals, the data archive should have 288 records per day for each detector. There should be 147,744 records with valid volume and occupancy values (513 total detectors × 288 records per day). Similarly, there should be 125,280 records with valid speed values (435 trap detectors × 288 records per day). Note that missing or null data values are not counted as valid data values for the purposes of the following completeness statistics. Table 2 contains the completeness statistics and data used in the calculations. Table A.2 indicates that the completeness of the archive database is still nearly fully complete.

Table A.2. Completeness Statistics for Data Archive

	Volume	Occupancy	Speed
Number of records with valid values	146,729	146,925	124,420
Number of records that require valid values	147,744 (513 total detectors)	147,744 (513 total detectors)	125,280 (435 trap detectors)
Percent Complete	**99%**	**99%**	**99%**

Readers should note that in processing the speed data, it was determined that about 17 percent of the reported speed values were “missing” because no vehicles were recorded during a 5-minute period (VOLUME=0 and OCCUPANCY=0). In the original source data, the value of SPEED=0 was replaced with a null or missing data value to better represent the traffic being recorded by the detectors. Even though 17 percent of the speed values are missing, in this example we do not count that against the percent complete measure since the missing speed values were presumably not caused by a detector malfunction, and thus should not lower data quality. However, data users should be cautious when datasets contain many time periods where no vehicles were recorded (VOLUME=OCCUPANCY=SPEED=0), as this may indicate detector failures.

Traveler Information

In this example, the ISP provides route-based speed and travel time reports as traveler information on its website and through a voice-responsive phone system. The route speeds and travel times are updated every minute in both systems, while the speed and travel time values are based on a rolling 2-minute average. There are a total of 6 routes being monitored, thus one would expect to have a total of 8,640 reported travel times during the day (6 routes × 1,440 updates per day, or one update per minute). The ISP estimates route speeds and travel times if some of the original source data are missing. The ISPs policy is to provide their best estimate of travel time, even if that travel time is based on historical data or speed limits instead of real-time data.

The ISP has automated a quality control process that monitors the availability of its website and voice-responsive phone system at periodic times throughout the day. Because of the ISP's policy of estimating travel times even when the original source data is incomplete, the main factor affecting completeness will be website and phone system availability. For this example, consider that a hardware failure in the phone system caused 60 minutes of downtime during this particular day. Travel time reports via the website were available at all sampled times of the day. Table A.3 contains the completeness statistics for the traveler information.

Table A.3. Completeness (Availability) Statistics for Traveler Information

	Travel Times on Website	Travel Times on Voice-Responsive Phone System
Number of records with valid values	8,640	8,280
Number of records that require valid values	8,640 (6 routes, updated every minute)	8,640 (6 routes, updated every minute)
Percent Complete	**100%**	**96%**

Table A.3 indicates that the completeness or availability of the traveler information was relatively high for both traveler information products. Because of the common ISP practice of estimating values when original source data are missing, the availability of traveler information can be more affected by hardware or software failures associated with ISP operations. In cases where the ISP does not estimate missing values, the availability may also reflect missing values in original source data.

Validity

For the Austin example, we calculate data validity for the three different datasets: original source data, archive database, and traveler information.

Original Source Data

In Austin, the field computers perform these basic validity checks on the original source data before it is sent to the traffic management center:

- If 20-second VOLUME > 15 vehicles, then invalid and VOLUME=-1
- If 20-second OCCUPANCY > 25 percent, then invalid and OCCUPANCY=-1
- If 20-second SPEED > 75 mph, then invalid and SPEED=-1
- If the 20-second VOLUME or OCCUPANCY in either upstream or downstream loop in a trap detector is found to be invalid, then VOLUME=-1, OCCUPANCY=-1, and SPEED=-1 for that trap detector.
- If there is a communication failure to the field controller, then VOLUME=-1, OCCUPANCY=-1, and SPEED=-1 for that detector.
- Non-trap detectors are assigned a SPEED=-1 (it is not capable of measuring speed), but the -1 value is disregarded since a SPEED value is not required for non-trap detectors.

Unfortunately, the Austin field computers use the same error code of "-1" for both invalid data as well as communication failures. Ideally, different error codes would be used so that missing data problems could be diagnosed.

To calculate validity of the original source data, we simply count the number of 1-minute data values that have been marked as valid values (i.e., those without "-1" values), and then divide by the total number of data values. Table A.4 contains the validity statistics and data used in the calculations.

Table A.4. Validity Statistics for Original Source Data

	Volume	Occupancy	Speed
Number of records meeting validity criteria	731,787	731,787	616,991
Number of records subjected to validity criteria	731,886	731,886	620,616
Percent Valid	**99.9%**	**99.9%**	**99%**

Table A.4 indicates that the validity of the original source data was very high, as less than 1 percent of the data failed the validity checks. This could be due to several reasons: 1) the data could be legitimately valid; 2) the validation checks could be too few or not rigorous enough.

Archive Database

The archive administrator uses the following additional validation rules:

- If DATE ={valid date value}, then valid data
- If TIME={valid time value}, then valid data
- If DET_ID={valid detector location value}, then valid data
- If VOLUME > {maximum volume threshold}, then invalid data
- If there are more than 8 consecutive identical VOLUME values, then invalid data
- If there are more than 8 consecutive identical OCCUPANCY values, then invalid data

- If there are more than 8 consecutive identical SPEED values, then invalid data
- If OCCUPANCY > {maximum occupancy threshold}, then invalid data
- If SPEED < {minimum speed threshold}, then invalid data
- If SPEED > {maximum speed threshold}, then invalid data
- If SPEED = 0 and VOLUME > 0 (and OCC > 0), then invalid data
- If VOLUME = 0 and SPEED > 0, then invalid data
- If SPEED = 0 and VOLUME = 0 and OCC > 0, then invalid data
- If OCC = 0 and VOLUME > {maximum volume threshold for truncated/rounded occupancy values}, then invalid
- If {estimated density using occupancy, volume and speed} > {maximum density threshold}, then invalid data

Note that these additional validation rules are applied to the original source data before it is aggregated into 5-minute periods. In some cases, validation rules may be applied at several different points in the data flow between original source data and the archive database. Table A.5 contains the validity statistics and data used in the calculations.

Table A.5. Validity Statistics for Archive Database

	Volume	Occupancy	Speed
Number of records meeting validity criteria	712,828	713,809	599,518
Number of records subjected to validity criteria	731,886	731,886	620,616
Percent Valid	**97%**	**98%**	**97%**

Table A.5 indicates that the validity of the archive database is still quite high, as less than 3 percent of the data failed the additional validity checks. Because of the number of additional validation checks, we can be reasonably assured that there are no major data validity problems with either the original source data or the archive database.

Traveler Information

In this example, consider that the ISP applies its validity criteria after the original source data has already been processed into route travel times. Thus the ISP uses a different set of validation rules than the archive administrator:

- If TRAVEL TIME < {minimum travel time based on free-flow traffic}, then invalid data
- If TRAVEL TIME > {maximum travel time threshold based on historical information}, then invalid data

Table A.6 contains the validity statistics for the ISP route travel times and data used in the calculations.

Table A.6. Validity Statistics for Traveler Information

	Route Travel Time
Number of records meeting validity criteria	8,380
Number of records subjected to validity criteria	8,640
Percent Valid	**97%**

Timeliness

Original Source Data

In measuring the timeliness of the original source data, we examine the data flow between the field computers and the traffic management center. There are four field computers that are expected to supply the traffic management center computer with data messages every minute, where a data message consists of the volume, occupancy and speed values for the previous minute. By examining the timestamps of the data messages, we can calculate the timeliness of this data flow. Note that in this example, the timestamps represent the time the data messages arrived at the traffic management center, not the time the data messages departed the field computers. This data timestamp convention should be confirmed when calculating timeliness, as it could dramatically affect the results.

The traffic operations personnel have decided that data messages received up to 5 seconds later than when they are expected are acceptable. In analyzing the timestamps on the 1-minute data messages, we find that 5,699 of the 5,707 data message were received at the traffic management center within 65 seconds of the previous message. Therefore, timeliness is calculated as:

$$\%\ timely\ data = \frac{5{,}699\ on-time\ messages}{5{,}707\ total\ messages\ received} = 99.8\%$$

By further analyzing the timestamps, we calculate that the **average delay for the 8 late messages is 28 seconds**. This means that, when a data messages were received late, on average it was received 28 seconds later than expected.

Archive Database

The archive administrator has a scheduled secure FTP download of the previous day's original source data from the traffic management center at 3 a.m. the following morning. The administrator also has a scheduled script that automatically transforms and loads the aggregated data into the archive database at 6 a.m. that same morning. In the Austin example, the original source data were collected August 29th, and the data were downloaded from the center's FTP site at 3 a.m. on August 30th, then loaded into the archive database by 6 a.m. One of the data archive users (e.g., the traffic operations personnel) expects the previous day's archived data to be available by 6 a.m. every morning since they have traffic management applications that rely on the data archive. Thus, if any portion of the previous day's archived data is loaded after 6 a.m. that data are considered to be late.

In this example, assume that a software malfunction prevented 10 percent of the archived data records from being loaded into the archive database. The archive administrator arrived at work, fixed the malfunction, and had the remaining 10 percent of the archived data loaded by 9 a.m. In this example, the timeliness is as follows:

- **90 percent timely data; and**
- **3 hours average delay for late data.**

Traveler Information

In this example, consider that the ISP would like to evaluate the timeliness of the updates to the route travel times on its website and voice-responsive phone system. For both systems, the ISP has a goal of providing condition updates every minute, based on the original source data. Now consider the hardware failure in the phone system that was discussed in the completeness example. The phone system was not available for 60 minutes, and the phone system provides travel times for 6 different routes. Thus, the timeliness is as follows:

$$\%\ timely\ data = \frac{8{,}280\ on-time\ updates}{8{,}640\ route\ travel\ time\ updates} = 96\%$$

In this example, average delay for late data is not calculated because the travel time updates in the phone system were not available at all for the entire 60 minutes.

Coverage

Original Source Data

The traffic operations personnel at TxDOT have focused their real-time data collection and traffic management activities on the freeways in the Austin area. Therefore, their goal is to monitor the entire freeway network in the Austin area with real-time traffic data. They have chosen to focus initial deployments on the most congested parts of the freeway network, with later deployments covering less congested freeway locations. As standard practice, TxDOT has installed detectors on the freeway main lanes between every major entrance or exit ramp, which results in an average detector spacing of 0.4 miles. Therefore, they consider this sample to adequately represent the freeway locations between point detectors. Additionally, they have placed detectors in every freeway lane and on all entrance and exit ramps.

Because of their emphasis, the traffic operations personnel only consider the functional class of freeways. In the Austin metropolitan planning area, there are a total of 174 centerline-miles of freeway. TxDOT has installed traffic detectors along 23 freeway centerline-miles. **Therefore, the percent of freeway coverage is 23/174 = 13 percent, with an average detector spacing of 0.4 miles.**

Archive Database

The archive administrator has also chosen to focus the coverage statistics on the freeway network only. Therefore, the coverage statistics in the archive database are exactly the same as in the original source data. **Therefore, the percent of freeway coverage is 23/174 = 13 percent, with an average detector spacing of 0.4 miles.**

Traveler Information

Because the ISP is attempting to provide traveler information for all major roadways, they consider arterial streets in reporting coverage statistics. Because the freeway routes for which travel times are provided correspond with freeway detector locations, the freeway coverage statistics for traveler information are the same as in the original source data and archive database. **Therefore, the percent of freeway coverage is 23/174 = 13 percent, with an average detector spacing of 0.4 miles. However, the arterial street coverage is 0 percent since no arterial street data is available from TxDOT or the City of Austin.**

Accessibility

Original Source Data

The accessibility of the original source data is first described in qualitative terms:

The original source data is accessible through a private computer network to the traffic operations personnel, who provide the same original source data to the archived data administrator and the ISP through periodic secure file transfer protocols. The archived original source data is also available on CD-ROM upon written request.

The archive administrator and the ISP have different software scripts they use to import, validate, and load the data into their system. The archive administrator is using customized software with advanced features that enables relatively fast data imports. The ISP is using commercial-off-the-shelf software that performs slightly slower than the archive software. The accessibility of the original source data in quantitative terms is as follows:

The archive administrator is able to retrieve and import a full day of original source data in 8 minutes (actual clock time). Over the course of a day, the ISP has tracked and recorded its data retrieval and import time as 10 minutes (actual clock time).

From this example, we can see that the original source data is easily accessible only to a limited number of data consumers (e.g., traffic operations personnel, archive administrator and ISP). We also note that the original source data is more accessible to the archive administrator than the ISP (e.g., 8 minutes vs. 10 minutes) because of customized software.

Archive Database

The accessibility of the archive database will be of interest to archived data users, who wish to retrieve and manipulate data products from the archive. The accessibility is first defined in qualitative terms:

The data archive is available to all data consumers through a public website.

In this example, consider that the archive administrator would like to measure how accessible the planning data products are to data consumers from the Capital Area Metropolitan Planning Organization (CAMPO). The archive administrator devises a simple exercise that asks users to retrieve average annual daily traffic (AADT) volumes for specific locations, then records how long it takes a sample of users to retrieve the desired data.

The accessibility of planning data in the data archive is such that it requires data consumers an average of 12 minutes to retrieve the desired data (e.g., AADT values).

In this example, the accessibility exercise is modeled after website usability tests since the primary access to the data archive is through a website. This example illustrates a test for a single data retrieval function. In most data archives, however, the accessibility exercise will most likely be performed for several of the most common data retrieval functions.

Traveler Information

The accessibility of traveler information will be of interest to travelers, who wish to make more informed travel decisions. The accessibility of the traveler information is as follows:

Route-based traveler information is available through a public website and a voice-responsive phone system.

Consider that the ISP would like to measure how accessible traveler information is on its website and phone system. The ISP recruits a sample of 50 travelers to do usability tests by offering a small incentive (e.g., three free months of personalized travel information). The usability tests measure how long it takes travelers to obtain current travel conditions for a specified route.

The accessibility of traveler conditions on the public website is such that it requires an average user 20 seconds to obtain data for the specified route. The accessibility of the phone system is such that is requires 60 seconds to obtain data for the specified route.

In this example, we can see that the traveler information is relatively accessible to many users. However, there are numerous other information outlets (e.g., changeable message signs, radio reports, etc.) that might improve accessibility for those without Internet access or mobile phones. We also note that the website appears to be much more accessible than the phone system. After viewing these statistics, the ISP might decide to upgrade the voice recognition software because the usability tests revealed that the delay was associated with poor voice recognition. Or the ISP

might note through the usability tests that, despite the longer access time, travelers using the phone system had comparable satisfaction ratings as those travelers using the website.
In most cases, data accessibility may not be as dynamic as the other data quality measures. The most appropriate time(s) to measure data accessibility is after major system interface or design changes. Measuring accessibility or usability at this time will allow system designers to see whether their interface or design changes have improved accessibility to data consumers.

Interpretation of Data Quality Statistics

The data quality statistics for the Austin case study are summarized in Table A.7.

Table A.7. Traffic Data Quality "Scorecard" for Austin Case Study

Data Quality Measures	Original Source Data	Archive Database	Traveler Information
Accuracy • MAPE • RMSE	One-minute speeds: 12.0% 11 mph	Hourly volumes: 4.4% 131 vehicles	Travel times: 8.6% 1.56 minutes
Completeness • Percent Complete	Volume: 99% Occupancy: 99% Speed: 98%	Volume: 99% Occupancy: 99% Speed: 99%	Website: 100% Phone: 96%
Validity • Percent Valid	Volume: 99.9% Occupancy: 99.9% Speed: 99%	Volume: 97% Occupancy: 98% Speed: 99%	Route travel times: 97%
Timeliness • Percent Timely Data • Average Data Delay	99.8% 28 seconds	90% 3 hours	96% n.a.
Coverage • Percent Coverage	Freeways: 13% with 0.4 mile spacing	Freeways: 13% with 0.4 mile spacing	Freeways: 13% with 0.4 mile spacing; Arterials: 0%
Accessibility • Avg. Access Time	Archive admin.: 8 minutes; ISP: 10 minutes	Retrieve AADT values: 12 minutes avg. access time	Website: 20 second avg. access time Phone: 60 second avg. access time

The results in Table A.7 indicate that, in general, the quality of traffic detector data is reasonably high for most data consumers. The quality measure with perhaps the lowest score was percent coverage, which can be expected since Austin is in the process of deploying their freeway detector system. The accuracy of speed data as collected from the field could be improved, as the mean absolute percent error was 12 percent. The traffic operations personnel might decide to devote additional resources to calibrate the double-loop detector for speed measurement, or perhaps they might decide that the existing accuracy is adequate to detect incidents. The phone access time for traveler information is 3 times as long as the website, so perhaps the ISP might decide to fine-tune the voice recognition software to decrease phone access times.

Although all six data quality measures are recommended for each data consumer, it should be evident that some data consumers will value certain aspects of data quality more than others. For example, traffic operations personnel and ISPs may consider timeliness a critical measure, whereas archived data users may be less concerned about timeliness.

The project team considered the calculation of a composite data quality score but did not further develop the concept for a number of reasons. This does not preclude data consumers from constructing their own composite score based on their priorities. A single data quality score would be difficult to interpret unless some value judgments were used with the measures that are not reported as percentages, such as timeliness or accessibility. Different data consumers may wish to weight each data quality measure differently according to their own priorities.

APPENDIX B:

PITTSBURGH, PENNSYLVANIA CASE STUDY

PITTSBURGH, PENNSYLVANIA CASE STUDY

Introduction

This case study is based on data from Mobility Technologies Inc. that operates and maintains a traffic information system to provide an Integrated Surveillance and Data Management Infrastructure (ISDMI). This case study presents procedures for calculating these data quality measures in a specific setting: traffic data collection, dissemination and archiving by Mobility Technologies, Inc. in Pittsburgh, Pennsylvania. The Pennsylvania Department of Transportation also collects and archives data in Pittsburgh; however, we will focus on the data collected by Mobility Technologies, Inc. to keep the example uncomplicated. The same principles used for the Mobility Technologies, Inc. data source can be applied to other data sources. Readers should note that some of the information and details in this case study example are accurate and true representations of actual data measures. However, some details and results are hypothetical and have been embellished or simplified for the purposes of the example. The embellishments are for illustration only and are not intended as criticisms of the data quality or suggested requirement for future data quality measures.

Traffic Data Flows: Identifying the Data Consumers

Figure B.1 illustrates the data flows involved in traffic data collection, dissemination, and archiving showing details related to the specific context of Pittsburgh traffic data. In this example, there are 3 different types of the data whose quality should be represented in the data quality measures:

1. **Original source data**: used by the information service provider (ISP), data warehouse administrator, and data warehouse users
2. **Archived data** (in data warehouse): used by archived data users
3. **Traveler information products**: used by travelers and other media outlets.

Calculation of Data Quality Measures

For the Pittsburgh case study, we consider a month of data (i.e., December 2002) collected by Mobility Technologies, Inc. as an example. Note that data quality could also be reported for other time scales, such as every hour, week, quarter, or year. For this particular example month, there were 103 unique stations (in which a “station” measures traffic data for a logical grouping of lanes, typically all functionally similar lanes in a direction) configured to report traffic data (i.e., volume, occupancy, speed, and two vehicle classes) at 1-minute intervals. Each 1-minute traffic reading from each station represents one record. The following sections describe specific calculation procedures for the six data quality measures for the above three different types of data.

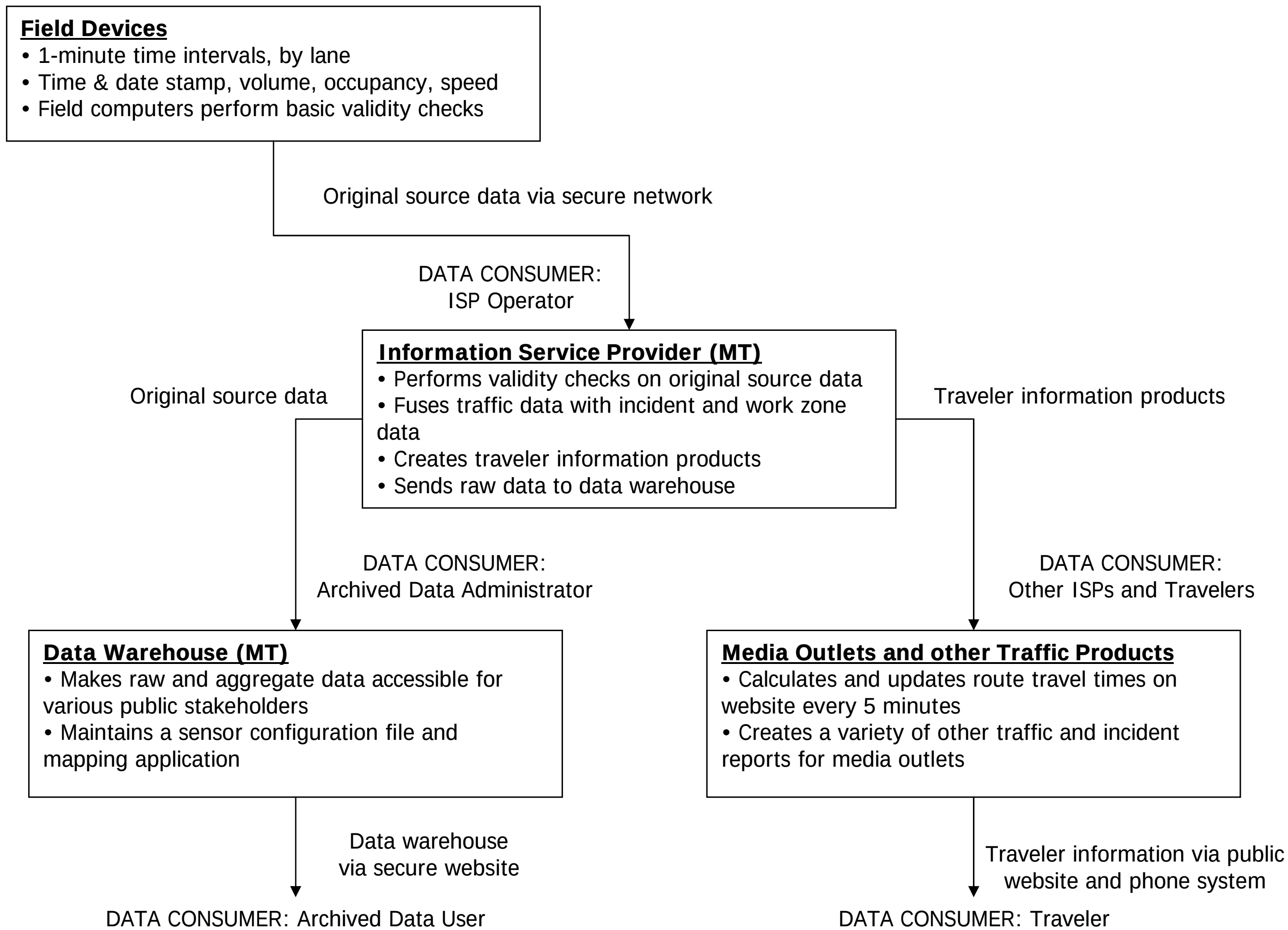

Figure B.1. Data Flows and Data Consumers in Pittsburgh Case Study

Accuracy

For the purposes of this example, we assume that reference measurements are available for two of the three different versions of data: data warehouse and traveler information.

Archived Data / Data Warehouse

In this example, consider that we wish to compare the accuracy of traffic volume values from an operations-based sensor to a nearby permanent traffic recorder (ATR). One of the many data products available through the data warehouse is hourly traffic volumes; therefore, the reference measurements are also summed to match the exact date and time of the hourly traffic volumes in the data warehouse.

Note that in many cases, it may be difficult to get ground truth or extremely accurate traffic measurements for an extended period of time. In many cases, an acceptable (or the only) substitute is traffic data from another trusted or familiar source. For planning groups, this is most commonly their continuous vehicle counts from ATR stations. If ATR stations are used as a comparison, one should recognize the limitations of such an approach.

For visual reference, a chart is created that compares the daily volume counts from the ATR to the percent difference from the data warehouse data value (Figure B.2). The mean absolute percent error is calculated by averaging together all of the percentage values in Figure B.2. Thus, the **MAPE for this comparison was calculated as 4 percent** using Equation 1, and the **root mean squared error (RMSE) was calculated as 1,280 vehicles** using Equation 2.

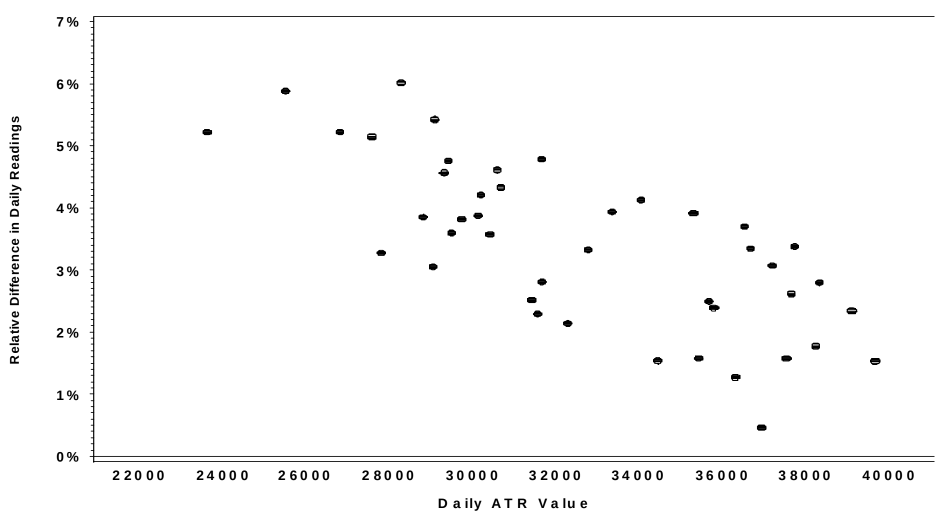

Figure B.2. Accuracy of Hourly Traffic Volumes in Archive Database

The comparison described above is actually from the *Evaluation of the Integrated Surveillance and Data Management Infrastructure (ISDMI) Program in Pittsburgh and Philadelphia, Pennsylvania* (prepared by Battelle, September 5, 2002). Their evaluation found that the daily traffic volume counts at a particular location varied from 0.4 to 6 percent from a nearby ATR station. This accuracy level is considered to be in reasonably good agreement. However, the possibility does exist that both the ISDMI and the ATR data could both be under or over-counting true traffic volumes. Only extensive calibration of a reference sensor will yield a "ground truth" measurement that has a high probability of being very accurate and useful for comparisons.

Traveler Information

In this example, the ISP operator provides route-based speed and travel time reports on its website and through other media outlets. The route speeds and travel times are updated every 5 minutes. Assume for the sake of example that the ISP operator arranges for reference travel time measurements to be obtained along selected Pittsburgh routes for various times of the day. The ISP travel times are visually compared to the reference travel times using similar charts (see Figure B.3). The mean absolute percent error was calculated using Equation 1, and the root mean squared error was calculated using Equation 2.

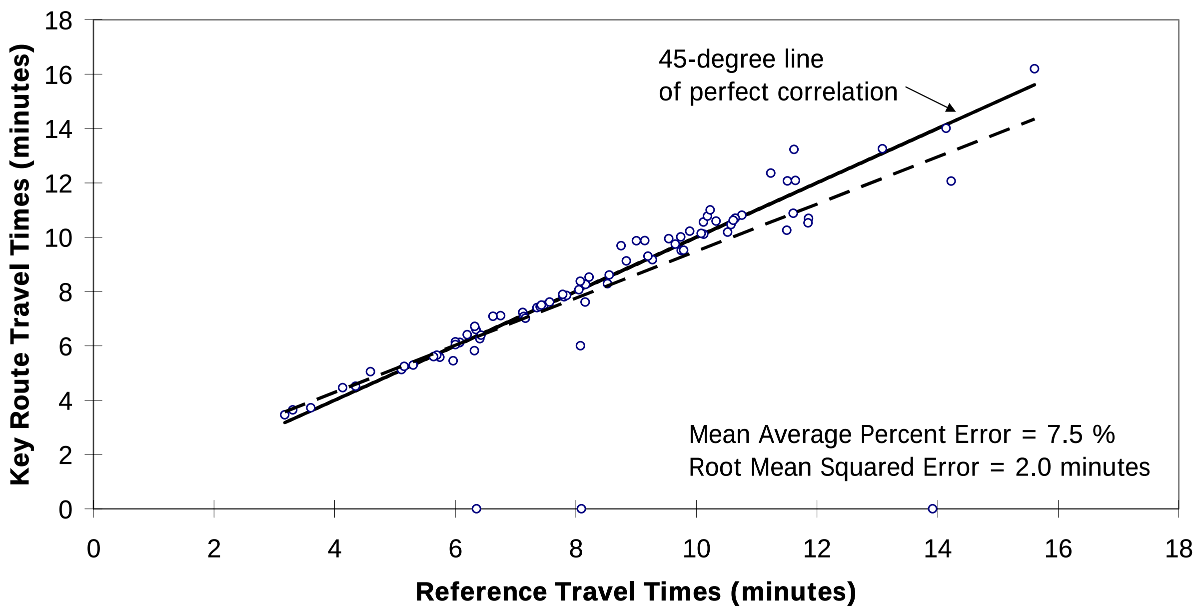

Figure B.3. Accuracy of Route Travel Time Values in Traveler Information

Completeness

In the Pittsburgh example, we calculate data completeness for the three different versions of data: original source data, data warehouse, and traveler information. In this particular example, the data process includes the flagging and eventual purging of invalid data values. Therefore, the completeness statistics will only include valid data values. The potential contribution of invalid data values to the completeness measure can be determined by combining the completeness and validity statistics.

Original Source Data

In Pittsburgh, there are 103 on-line stations that should report a data record every minute for the entire day. Thus, we expect to have 4,597,920 valid volume and occupancy records per day (103 total stations × 1,440 records per day × 31 days in December 2002). The Pittsburgh field computers perform basic validity tests on 1-minute data and remove invalid values; thus, this invalid data is considered missing since it is removed. Table B.1 contains the completeness statistics and data used in the calculations.

Table B.1. Completeness Statistics for 1-Minute Original Source Data

	Volume, Occupancy, Speed and Vehicle Classification Data
Number of records with non-missing values	4,352,000
Number of records that require non-missing values	4,597,920 (103 stations)
Percent Complete	**95%**

The completeness statistic in Table B.1 indicates that the original source data in Pittsburgh is 95 percent complete, with 5 percent of the data being incomplete (i.e., missing or invalid). Incomplete data can be caused by 1) large amounts of invalid data; or 2) missing data due to communication, hardware, or software failures. Note that the completeness statistics must be viewed in combination with validity statistics to pinpoint the most likely cause of missing data.

Data quality reports should fully specify or disclose information related to the amount of expected data (the denominator of percent complete), especially for the completeness measure. Malfunctioning detectors should not be discounted from expected data counts simply because device owners are aware of their malfunction but have not been able to repair the devices. The practice of listing malfunctioning detectors by considering them “off-line” is not recommended as it obscures the true device failure rate and data quality results.

Archived Data / Data Warehouse

As shown in Figure B.1, the archived data administrator retrieves the original source data from the ISP operator. Note that in this example, these functions are both done by the Mobility Technologies company. The archive administrator performs several data processing steps in preparation for loading into the data warehouse:

1. Additional data validation checks (beyond what is done by field computers);
2. Aggregation of 1-minute data to 5-minute intervals; and
3. Addition of metadata to describe number of valid 1-minute samples in 5-minute statistics.

After the archive administrator has performed these processing steps, completeness statistics are computed by counting the valid data values in the data archive. With 5-minute subtotals, the data archive should have 288 records per day for each station. Thus there should be 919,584 records with valid volume and occupancy values (103 total stations × 288 records per day × 31 days in December 2002). Note that missing or null data values are not counted as valid data values for the purposes of the following completeness statistics. Table B.2 contains the completeness statistics and data used in the calculations.

Table B.2. Completeness Statistics for 5-Minute Summary Data in Data Warehouse

	Volume	Occupancy	Speed
Number of records with valid values	867,606	868,898	835,855
Number of records that require valid values	919,584 (103 total stations)	919,584 (103 total stations)	919,584 (103 total stations)
Percent complete	**94%**	**94%**	**91%**

Table B.2 indicates that the completeness of the archive database is still fairly complete. In this example, the completeness of the archived data it slightly less than that of the original source data because it has undergone additional validation criteria before being stored in the data warehouse. Also note that different traffic attributes (e.g., volume, occupancy, speed) have different completeness statistics because several of the validation checks only removed invalid data for a particular attribute. For example, a high speed value (greater than 100 mph, for example) may have been found to be invalid, but the corresponding volume and occupancy values were kept as valid values.

Traveler Information

In this example, the ISP operator provides route-based speed and travel time reports as traveler information on their website. The route speeds and travel times are updated every 5 minutes on the website. There are a total of 10 key routes being reported, thus one would expect to have a total of 89,280 reported travel times during the day (10 key routes × 288 updates per day, or one update every 5 minutes × 31 days in December 2002). If data is not available for a key route, it is the ISP operator's policy to not provide an estimate of travel time.

Assume for this example that the ISP operator has automated a quality control process that monitors the availability of key route travel times on their website at all times throughout the day. For this example, consider that an intermittent communications failure interrupted data transmittals along one of the ten key routes for 12 days. Thus, there were 12 days of travel time updates for one route that were not available (1 route × 12 updates per hour × 24 hours per day × 12 days of downtime). Table B.3 contains the completeness statistics for the traveler information.

Table B.3 indicates that the completeness or availability of the traveler information was relatively high for the key route travel times. In this example, the ISP operator does not estimate missing travel time values, thus the availability may also reflect missing values in original source data. Where ISPs estimate missing data values when original source data are missing, the availability of traveler information is more affected by hardware or software failures associated with ISP operations.

Table B.3. Completeness (Availability) Statistics for Key Route Travel Times

	Key Route Travel Times on Public Website
Number of records with valid values	85,824
Number of records that require valid values	89,280 (10 routes, updated every 5 minutes, 31 days in December 2002)
Percent Complete	**96%**

Validity

For the Pittsburgh example, we calculate data validity for the three different datasets: original source data, data warehouse, and traveler information.

Original Source Data

In Pittsburgh, the field computers perform some very basic validity checks on the original source data before it is sent to the ISP operator. Then assume that the field computers remove invalid data and replace it with an error code that distinguishes invalid data from missing data. Having different error codes for different data problems helps to diagnose the root cause of missing data.

To calculate validity of the original source data, we simply count the number of 1-minute data values that have been marked as valid values (i.e., those without "invalid" error codes), and then divide by the total number of data values subjected to the validity criteria. Table B.4 contains the validity statistics and data used in the calculations.

Table B.4. Validity Statistics for 1-Minute Original Source Data

	Volume	Occupancy	Speed
Number of records meeting validity criteria	4,337,983	4,343,876	4,287,885
Number of records subjected to validity criteria	4,352,000	4,352,000	4,352,000
Percent Valid	**99.7%**	**99.8%**	**98.5%**

Table B.4 indicates that the validity of the original source data was very high, as less than 2 percent of all data failed the validity checks. The speed data had slightly lower validity—this could have been due to an improperly calibrated sensor that was reporting speeds outside of an acceptance criteria threshold.

Archived Data / Data Warehouse

The archive administrator uses several other validation rules once the original source data arrives at the data warehouse. In most real-time data processing (as on field computers), validation criteria are kept simple because processing time must be minimized. In a data warehouse environment, there is less time restriction and more advanced validation criteria can be applied.

Note that these additional validation rules are applied to the original source data before it is aggregated into 5-minute periods. In some cases, validation rules may be applied at several different points in the data flow between original source data and the archive database. Table B.5 contains the validity statistics and data used in the calculations.

Table B.5. Validity Statistics for 5-Minute Summary Data in Data Warehouse

	Volume	Occupancy	Speed
Number of records meeting validity criteria	855,603	853,862	849,510
Number of records subjected to validity criteria	870,400	870,400	870,400
Percent Valid	**98.3%**	**98.1%**	**97.6%**

Table B.5 indicates that the validity of the archive database is still quite high, as less than 3 percent of the data failed the additional validity checks. Because of the number of additional validation checks, we can be reasonably assured that there are no major data validity problems with either the original source data or the archive database.

Traveler Information

In this example, consider that the ISP operator and the other media outlets do not apply any additional validity criteria to the route travel times beyond what is applied to the original source data. Because no additional criteria are applied, all reported route travel time values are valid (as there are no criteria by which to reject a route travel time as invalid). By using this practice, the ISP operator is assuming that all invalid data is being addressed in an "upstream" data process (i.e., a data process that occurs before route travel times are computed). Table B.6 contains the validity statistics for the ISP route travel times and data used in the calculations.

Table B.6. Validity Statistics for Key Route Travel Times

	Route Travel Times
Number of records meeting validity criteria	85,824
Number of records subjected to validity criteria	85,824
Percent Valid	**100%**

Timeliness

Original Source Data

In measuring the timeliness of the original source data, we examine the data flow between the field computers and the traffic management center. There are four field computers that are expected to supply the traffic management center computer with data messages every minute, where a data message consists of the volume, occupancy and speed values for the previous minute. By examining the timestamps of the data messages, we can calculate the timeliness of this data flow. Note that in this example, the timestamps represent the time the data messages arrived at the traffic management center, not the time the data messages departed the field computers. This data timestamp convention should be confirmed when calculating timeliness, as it could dramatically affect the results.

The ISP operator has decided that data messages received up to 30 seconds later than when they are expected are acceptable. In analyzing the timestamps on the 1-minute data messages, we find that 4,347,648 of the 4,352,000 data messages were received at the traffic management center within 90 seconds of the previous message. Therefore, timeliness is calculated as:

$$\%\ timely\ data = \frac{4{,}347{,}648\ on-time\ messages}{4{,}352{,}000\ total\ messages\ received} = 99.9\%$$

By further analyzing the timestamps, we calculate that the **average delay for the 4,352 late messages is 48 seconds**. This means that, when a data message was received late, on average it was received 48 seconds later than expected.

Archived Data / Data Warehouse

Immediately after collection, the original source data is replicated and copied to a staging area in the data warehouse. The data then go through an automated validation and loading process at a scheduled time during off-peak hours. The goal of the ISP operator is to have the previous day of archived data available through the data warehouse by 8 a.m. of the next day. Thus, any data not loaded by 8 a.m. of the following data is considered late and not on-time.

Assume for this example that a problem in the data warehouse software caused the data to be loaded later than expected for two separate days. Assume that, on each day, the problem was diagnosed and the data were loaded by 6 p.m. later that day. For all other days in December 2002 (29 of 31 days), the data were loaded in the data warehouse and were available by 8 a.m. Thus, the timeliness is as follows:

$$\% \ timely \ data = \frac{29 \ days \ of \ \ timely \ data \ uploads}{31 \ days \ of \ \ planned \ service} = 94\%$$

In this example, average delay is 10 hours, which is the average amount of time between when the data was expected and when it actually became available.

Traveler Information

In this example, consider that the ISP operator would like to evaluate the timeliness of the updates to the key route travel times on their website. For this example, assume that the ISP operator has a goal of providing condition updates every 5 minutes. Now consider hypothetically that the ISP operators' web servers have a series of crashes and problems that prevents users from accessing the website for a period of 24 hours. Thus, the timeliness is as follows:

$$\% \ timely \ data = \frac{30 \ days \ of \ \ "uptime"}{31 \ days \ of \ \ planned \ service} = 97\%$$

In this example, average delay for late data is not calculated because the travel time updates in on the website were not available at all for the entire 24 hours.

Coverage

Original Source Data

The ISP operator has focused their real-time flow data collection on the freeways in the Pittsburgh area (Note that MT collects real time incident and event data on all roads). Therefore, their goal is to monitor the most important freeway routes in the Pittsburgh area with real-time traffic data. They have chosen to focus initial deployments on the most congested parts of the freeway network. Because the ISP operator is using the data primarily for traveler information (and not for traffic management or incident detection), they have installed sensors on the freeway main lanes only, with an average sensor of about 1.5 miles. Therefore, they consider this sample to adequately represent the freeway locations between point detectors for the purposes of traveler information.

Because of their emphasis, the ISP operator only considers flow data on the functional class of freeways. In the Pittsburgh metropolitan planning area, there are a total of 284 centerline-miles of freeway. The ISP operator has installed traffic detectors along 78 freeway centerline-miles. **Therefore, the percent of freeway coverage is 78/284 = 27 percent, with an average detector spacing of 1.5 miles**.

Archived Data / Data Warehouse

The archive administrator has also chosen to focus the flow coverage statistics on the freeway network as well. Therefore, the coverage statistics in the archive database are exactly the same as in the original source data. **Therefore, the percent of freeway coverage is 78/284 = 27 percent, with an average detector spacing of 1.5 miles**.

Traveler Information

The traveler information flow data is also focused on the freeway network only. Therefore, the coverage statistics for traveler information as the same as the archive database and the original source data. **Therefore, the percent of freeway coverage is 78/284 = 27 percent, with an average detector spacing of 1.5 miles**.

Accessibility

In this example, we will describe the accessibility of traffic data using only qualitative terms.

Original Source Data

The accessibility of the original source data is described in these qualitative terms:

- The original source data are accessible in real-time to only the ISP operator. The lane by lane data for a sensor is available for stakeholder agencies on a real time and sensor by sensor basis through the ISP's secure website. After it is loaded into the data warehouse, the original source data are accessible through to certain public agency stakeholders through a secure website.

Archived Data / Data Warehouse

The accessibility of the archive database will be of interest to archived data users, who wish to retrieve and manipulate data products from the archive. The accessibility is described in qualitative terms as follows:

- The archived data are accessible to certain public agency stakeholders through a secure website. The archived data may also be made available on a request-by-request basis.

Traveler Information

The accessibility of traveler information will be of interest to travelers, who wish to make more informed travel decisions. The accessibility of the traveler information is as follows:

- Route-based speeds are available through a public website which requires free registration. More advanced traveler information products are also accessible through other public media outlets such as media websites, local television, and local/satellite radio).

In most cases, data accessibility may not be as dynamic as the other data quality measures. The most appropriate time(s) to measure data accessibility is after major system interface or design changes. Measuring accessibility or usability at this time will allow system designers to see whether their interface or design changes have improved accessibility to data consumers.

Interpretation of Data Quality Statistics

The data quality statistics for the Pittsburgh case study are summarized in Table B.7.

Table B.7. Traffic Data Quality Summary for Pittsburgh Case Study

Data Quality Measures	Original Source Data	Archived Database	Traveler Information
Accuracy • MAPE • RMSE	n.a. n.a.	Daily counts: 4% 1,280 vehicles	Travel times: 7.5% 2.0 min.
Completeness • Percent Complete	All data: 95%	Volume: 94% Occupancy: 94% Speed: 91%	Website key route travel times: 96%
Validity • Percent Valid	Volume: 99.7% Occupancy: 99.8% Speed: 98.5%	Volume: 98.3% Occupancy: 98.1% Speed: 97.6%	Key route travel times: 100%
Timeliness • Percent Timely Data • Average Data Delay	99.9% 48 seconds	94%	97% n.a.
Coverage • Percent Coverage	Freeways: 27% with 1.5 mile spacing	Freeways: 27% with 1.5 mile spacing	Freeways: 27% with 1.5 mile spacing
Accessibility • Qualitative access	Accessible in real-time on sensor bias through secure website; historical data accessible to public agency stakeholders through secure website	accessible to public agency stakeholders through secure website	Accessible through public website by registration; also accessible through other public media outlets (e.g., web, TV, radio)

APPENDIX C:

OHIO DEPARTMENT OF TRANSPORTATION CASE STUDY

OHIO DEPARTMENT OF TRANSPORTATION CASE STUDY

Introduction

This case study describes procedures for calculating the data quality measures in a specific setting: statewide traffic data collection and dissemination by a traffic monitoring group. Calculation of quality measures for the traffic monitoring program is different from the ITS-based traffic operations. This case study is based on data provided in part by the Ohio Department of Transportation (ODOT) and partly on hypothetical assumptions.

Traffic Data Flows: Identifying the Data Consumers

Figure C.1 illustrates the data flows involved in traffic data collection, dissemination, and archiving showing details related to the specific context of traffic monitoring perspective. The following are the main sources of data whose quality should be represented in the data quality measures.

- **Original source data from continuous counts –** This includes data from ATRs, classification stations (AVCs), WIM stations etc. These data streams provide data daily to the traffic monitoring groups. The data from these streams maybe downloaded/polled daily or weekly and reviewed in the subsequent days.

- **Original source data from short-term counts –** This data stream represents data collected by the traffic monitoring unit over 24-48 hours spans using traffic counters connected to portable or permanent equipment.

- **Archived data –** A combination of business rules and validity checks are used to test the quality of original source data. The processed data are archived. These data includes AADT and summaries, GIS applications, truck data. Users of such archived data include FHWA, MPOs, research organizations, and the general public.

The following sections present the calculation of the six data quality measures for the three data sources described above.

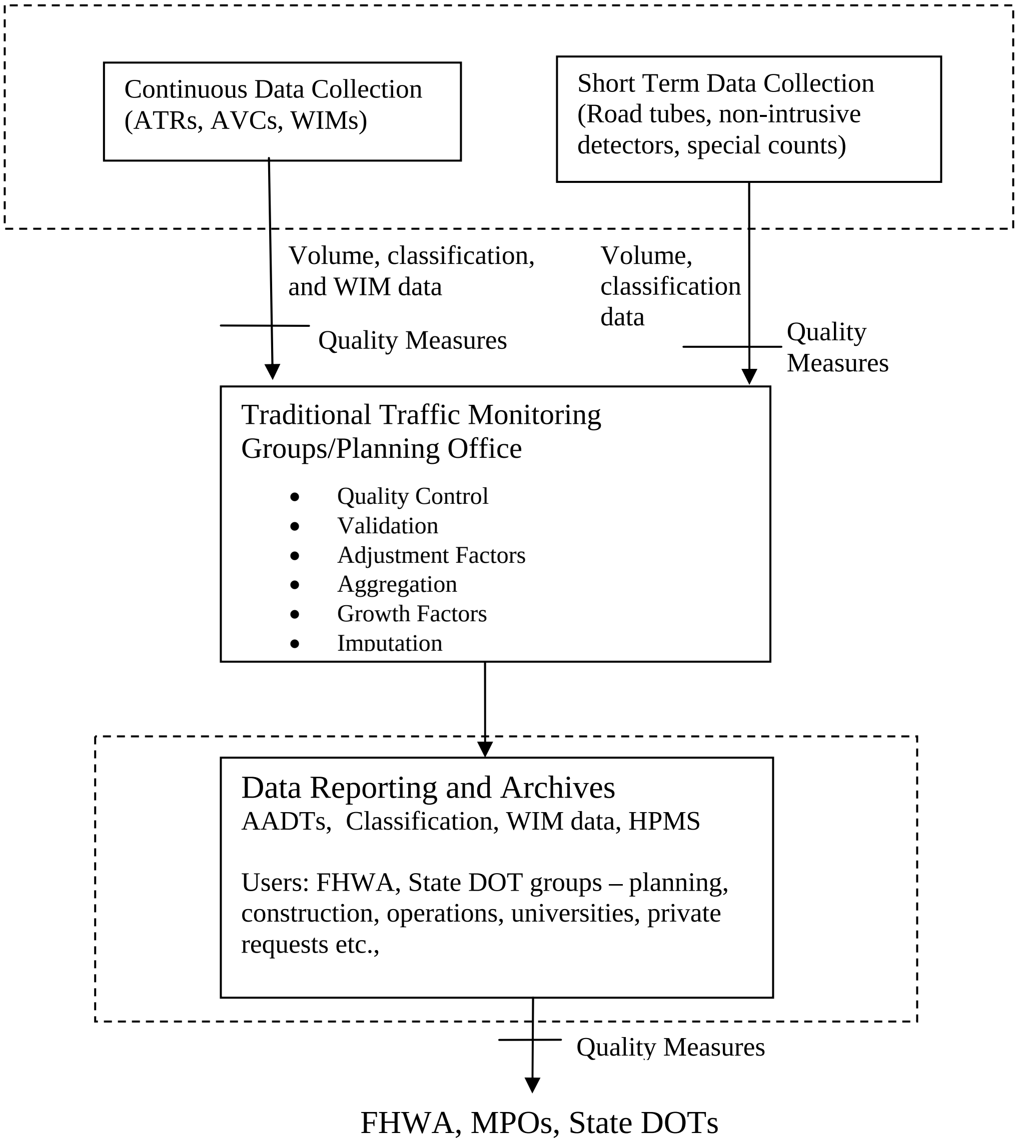

Figure C.1. Data Flows and Consumers – Ohio Case Study

Accuracy

Accuracy is defined as "the measure or degree of agreement between a data value or set of values and a source assumed to be correct." As its definition indicates, accuracy requires "…a source (of data) assumed to be correct." This correct source of data is typically referred to as ground truth, reference, or baseline measurements. Ground truth data can be collected in several different ways for each type of traffic data. Calculation of accuracy in a traditional traffic monitoring context is a very difficult task. Often times, the only version of the ground-truth or the reference values are from manual counts, which are expensive to perform and are also not error-free.

Original Source Data – Continuous Counts

ATRs and AVCs

In this example, four hours of data from manual counts and vehicle classification is compared to the data reported by the traffic counter as shown in Table C.1. The manual count is assumed to represent the ground truth or baseline. The accuracy of ATR total volume counts compared very well with the total volume from manual counts. The RMSE for the total volume counts is about 4 vehicles with MAPE of only 0.89%. However, vehicle classification accuracy varies depending on the class of vehicle and quite significant for some vehicle classes. This could be due to the programmed class tables or the classification bins or a pointer to the need for recalibration of equipment.

This example is based on data from on tests performed on new equipment. The performance of the equipment in the field might be significantly different. It is recommended that the accuracy measure is calculated at random or periodic intervals on equipment installed in the field. While this is desirable, the cost implications should be taken into account. Accuracy tests can be performed during routine or periodic maintenance or calibration visits.

Table C.1. Comparison of Manual and ATR Counts for Vehicle Classes and Volumes

Vehicle Class	Hour 1		Hour 2		Hour 3		Hour 4		Accuracy Measures	
	Manual Count	Detector Count	Manual Count	Detector Count	Manual Count	Detector Count	Manual Count	Detector Count	RMSE	MAPE
Class 1	0	0	0	0	0	0	0	0	0	
Class 2	153	146	139	127	129	123	166	147	12.1	7.33%
Class 3	47	52	53	61	44	57	54	61	8.8	17.06%
Class 4	1	5	0	4	0	8	0	14	8.5	100.00%
Class 5	16	14	18	15	26	13	18	7	8.7	35.07%
Class 6	5	5	14	15	5	4	9	6	1.7	15.12%
Class 7	0	0	0	3	0	0	0	0	1.5	
Class 8	15	19	9	13	15	17	14	23	5.4	37.18%
Class 9	187	186	190	191	197	196	191	191	0.9	0.39%
Class 10	3	3	6	4	7	7	2	2	1	8.33%
Class 11	9	10	17	19	16	17	15	15	1.2	7.28%
Class 12	3	3	2	2	1	1	1	1	0	0.00%
Class 13	0	0	0	0	0	0	0	0	0	0.00%
TOTAL	439	443	448	454	440	443	470	467	4.2	0.89%

WIM Data

This example illustrates the calculation of accuracy measure using data collected during WIM station calibration. Table C.2 describes three runs over a WIM station with a known truck configuration (i.e., baseline). As shown in Table C.2, the WIM data corresponds well with actual configurations with vehicle length being the parameter with highest error. For the three runs during this calibration, the accuracy measures for length and weight measurements at the WIM station can be calculated as follows.

Accuracy of vehicle length measurement
- MAPE – 7.26%
- RMSE – 1.89 feet (over the allowable tolerance)

Accuracy of gross vehicle weight measurement
- MAPE – 4.17%
- RMSE – 1,760 pounds (under allowable tolerance of 15% of GVW)

Table C.2. WIM Calibration Report from DOT

Measurements	Baseline	Run 1		Run 2		Run 3	
		Value	% Error	Value	% Error	Value	% Error
Dimensions (ft)							
Axle 1-2	12.6	12.6	0.00%	12.5	0.79%	12.7	-0.79%
Axle 2-3	4.5	4.6	-2.22%	4.5	0.00%	4.5	0.00%
Total Length	20.2	21.2	-4.95%	23.3	-15.35%	19.9	1.49%
Weight (kips)							
Steering axle	13.46	13.1	2.67%	15.3	-13.67%	12.9	4.16%
Drive tandem axle	28.56	27	5.46%	25	12.46%	27.6	3.36%
GVW	42.02	40.1	4.57%	40.2	4.33%	40.5	3.62%
Speed (mph)	55, 53, 55	55	0	53	0	55	0

Note: Percent errors are calculated relative to the baseline value.
Negative error indicates that the baseline value is lower.
Allowable tolerances (source, ODOT)
- Vehicle length = ± 12 inches
- Axle spacing = ± 6 inches
- Steering axle weight = ± 30%
- Single axle weight = ± 30%
- Tandem axle weight = ± 20%
- GVW = ± 15%
- Speed = ± 2 mph

The MAPE indicates the mean absolute percent error of total vehicle length over the three runs for the WIM station. The percent error was 7.26 percent. The RMSE for vehicle length was about 1.89 feet which is over the tolerance of 12 inches. Similarly, for GVW, the detector was in error by 4.17 percent (and RMSE of 1.76 kips) which is less than the 15 percent allowable tolerance for the truck used in the test.

Original Source Data – Short-Term Counts

For short-term volume and classification data collection using portable equipment – the accuracy measure is calculated at the time of procuring the traffic counters. A manual count is performed and compared with data reported from the counter in each of the vehicle classes. The manual count is assumed to be the ground truth. Since this is a single count, the MAPE and RMSE cannot be calculated for the particular detector. A more useful measure of accuracy in this case is the percent or the actual error in each of the 13 vehicle classes and in the total volume. (Figure C.2). The figure indicates that there was a positive error (undercount) of 2 vehicles for Class 2 and Class 4, whereas a negative error (over count) of 4 vehicles was reported for class 3. All other vehicle classes were correctly classified and counted with no errors. Also, it is noted that while there are small classification errors, the total volume is accurate as the errors cancel each other out.

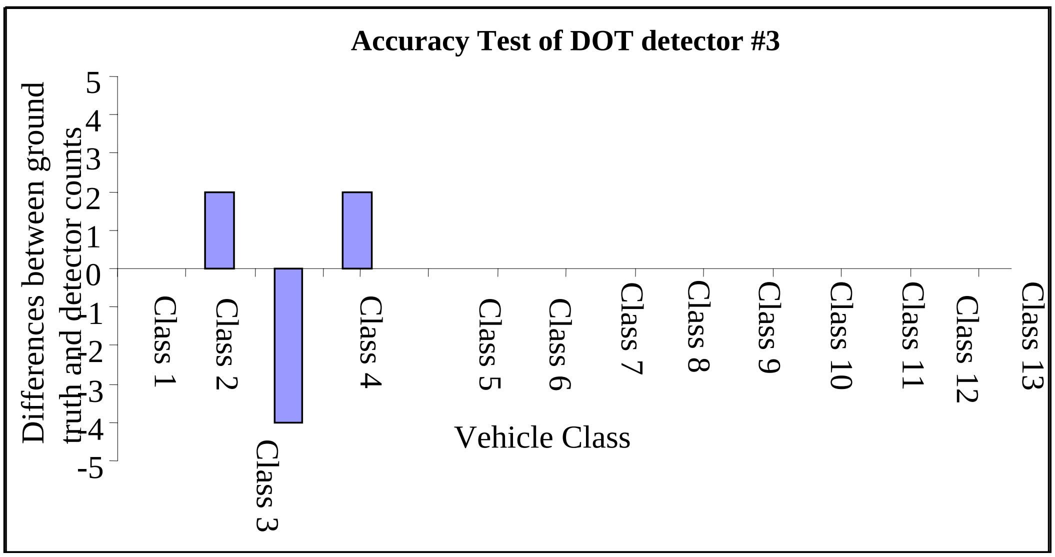

Figure C.2. Accuracy Test of Detector #3
(Difference between manual and detector counts)

Table C.3 above shows the manual counts compared to five detectors from the same vendor. In this case MAPE and RMSE can be calculated across detectors for each vehicle class and total volumes. These measures represent the detection capability of the set of detectors. The accuracy measures for two classes vehicles with the highest errors compared to total volume are shown below.

Table C.3. Accuracy Tests of Five Detectors from a Vendor

Vehicle Class	Baseline	Detectors									
		Detector 1		Detector 2		Detector 3		Detector 4		Detector 5	
		Count	Diff.	Count	Diff.	Count	Diff.	Count	Diff.	Count	Diff.
1	1	1	0	1	0	1	0	1	0	1	0
2	399	399	0	399	0	397	2	401	-2	403	-4
3	138	138	0	140	-2	142	-4	138	0	137	1
4	4	4	0	4	0	2	2	4	0	4	0
5	39	39	0	39	0	39	0	39	0	39	0
6	35	35	0	35	0	35	0	35	0	35	0
7	2	2	0	2	0	2	0	2	0	2	0
8	37	37	0	37	0	37	0	37	0	37	0
9	647	643	4	644	3	647	0	644	3	642	5
10	5	5	0	5	0	5	0	5	0	5	0
11	36	36	0	36	0	36	0	36	0	36	0
12	1	1	0	1	0	1	0	1	0	1	0
13	1	1	0	1	0	1	0	1	0	1	0
Total	**1345**	**1341**	**4**	**1344**	**1**	**1345**	**0**	**1344**	**1**	**1343**	**2**

Class three volume
- **MAPE – 1.01%,**
- **RMSE – 2 vehicles**

Class nine volume
- **MAPE – 0.46%**
- **RMSE – 3.43 vehicles**

Total volume
- **MAPE – 0.11%,**
- **RMSE – 2 vehicles**

These accuracy measures indicate that the volumes for class three vehicles measured by the detectors have an error of 1.01 percent or 2 vehicles associated with them. However the total volume has a much lower error of about 0.11 percent and a root mean square error of 2 vehicles.

As with continuous counts, these accuracy numbers can be misleading because the tests may not necessarily represent long term field conditions. It may be useful, however informally, to collect at least 5 minutes (or 100 vehicles) worth of data manually at the beginning of each short term count. The data from the manual counts and the data collected by the traffic counter can then be compared to establish count and classification accuracy in a more realistic manner.

Archived Data

Accuracy for archived data is a function of the processes used to generate some of the outputs of traffic data including AADTs, classification and WIM data. AADT especially from short-counts is adjusted for weekly and seasonal variations. While these processes may be estimates, it is often difficult to determine the reduction in accuracy due to these processes for the lack of reference values. Nonetheless, it is possible to generate some accuracy estimates. For example, by using historical data, forecasted AADTs, and current estimates, it is possible to determine if the data follow existing trends. A problem with trend analyses for such data is that while it identifies anomalies, there is no way of determining if the anomaly is an error or an unusual but true value (for example, spikes due to incidents, construction etc.).

In addition, it is possible to have an idea about the quality of AADT estimates if the accuracy measures of the ATRs and the short-count equipment are available. Since AADT = ADT (from short counts) * Adjustment factors (from ATRs), data derived from high-quality ATRs and accurate short-counts is expected to result in better AADT estimates as shown.

Completeness

Original Source Data - Continuous Counts

For continuous counts, the completeness measure used by the DOT is the number of complete days of data in a month. Complete data is characterized by hourly records for each day of the month containing volume, and classification data for each lane being monitored.

ODOT has a total of 220 ATRs statewide. Data is aggregated in 60 minute intervals. Completeness can be calculated for traffic volume data in two ways. In the first approach, the calculation of completeness assumes a perfect data collection situation where all ATRs record data in all the days of a given month (30 days average). Hypothetically, assume only 140 ATRs have no missing records. In that case, completeness will be calculated as follows:

> **Total Expected Records – 24 (hours)*30 (days)*220 = 158,400**
> **Records with no data missing – 24*30 (days)* 140 = 100,800**
> **Completeness = (100,800/158,400)*100 = 63.6 %**

The second case, which is more realistic, the DOT uses data from a particular ATR if there are at least 14 days of useable data in a given month. On an average, there are 190 sites with sufficient data to generate a monthly ADT (i.e., at least 14 days of 24 hour worth of data). In this case, completeness for traffic data volume can be calculated as shown below:

> **Minimum Expected Records – 24 (hours)* 14 (days) * 220 = 73,920**
> **Available Records – 24 (hours) * 14 (days) * 190 = 63,840**
> **Completeness = (63,840/73,920) * 100 = 86.4%**

It is important to note that the completeness measure calculated above is a good indicator of completeness for monthly ADT calculations only. These measures will need to be recalculated if the agency requires more than 14 days of data as a minimum.

Original Source Data - Short-Term Counts

For short-term counts, usually for 24-48 hours aggregated in 60 minute intervals, the completeness measure is slightly different because the agency has the option of resetting the count and collecting data again. The DOT has a goal of 4,200 short-term counts annually. Incomplete counts are not used. If the count is not complete, they are reset in the field. There is no available statistics on this number. All counts are 24-48 hours in duration and are distributed as follows:

- 42% of the counts are 48-hour classification and volume counts,
- 35% are 24-hour classification and volume counts,
- 1% are 48-hour volume only counts, and
- 22% are for 24-hour volume only.

The completeness measure for short-term data collection, as defined, is 100%.

Archived Data

The completeness measure for data users is determined by their applications. As a hypothetical example, FHWA could define completeness of DOT data based on HPMS submittals. The State DOT submits data for 3,900 segments annually of which 3,600 records are deemed complete based on FHWA review. The completeness measure then is **(3600/3900) * 100 = 92%**

The completeness measures are summarized in Table C.4.

Table C.4. Summary of Completeness Measures

Categories	**Completeness Measure**
Original Source Data	
Continuous Count	86.4% for ADT generation
Short-Term Count	100%
Processed Data	
Archived Data	92% based on HPMS submittals (hypothetical)

Validity

Original Source Data – Short-Term Counts

The DOT uses a mainframe based program to analyze data from continuous count stations. Data is downloaded daily and processed through the software. Questionable data records are flagged for review by a manual operator. At this point, the manual operator makes a decision on whether to accept the data or to delete it.

According data provided for this example, 190 of the 220 ATRs (86.4%) on average have complete data as presented above. Out of the 190 ATRs, only 180, on average, record valid data. Valid data is defined as data that is verified using high/low range checks and historical ADT trends data. Thus the validity measure, as calculated as

Expected records – 24 (hours)* 14 (days) * 220
Complete records for validity criteria – 24 (hours)*14 (days) * 190
Valid records – 24 (hours) * 14 (days) * 180
Validity = (180/190) * 100 = 94.7%

Original Source Data - Short-Term Counts

The validity of short count data is intrinsically related to completeness. Since it is possible for the DOT to reset the counter, the key indicator remains how many of these counts are usable for ADT data and how many have to be reset. Historical data indicates that approximately 5 percent of the 4,200 (i.e., 210 counters) annual short term counts need to be reset. Resets are based upon reviews of hourly data, high/low ADT and historical AADT values.

Thus, the **validity of short-term counts is (100 – 5) = 95%. Note that invalid counts are reset ensuring that all the short-counts are valid.**

Archived Data

Validity as perceived by archived data users is dependent on the application. Frequently, the users assume the data provided to them as valid. The data reported by the DOT is considered "official". Several users might have their own validity criteria which are applied to the data from the DOT. For example, HPMS administrators at FHWA might check the validity of the dataset submitted by ODOT for sample size adequacy, inventory errors, pavement information errors etc.

As a hypothetical example, FHWA could define validity of DOT data based on HPMS submittals. The state DOT submits data for 3,900 segments annually and 3,600 records are complete. Assume that only 3,200 records pass the validity checks. **Thus the validity measure can be calculated as (3200/3600) * 100 = 88%**

Table C.5. Summarizes the Validity Measures

Categories	**Validity Measure**
Original Source Data	
Continuous counts	94.7% for ADT generation
Short-Term counts	95.0 %
Processed Data	
Archived Data	88% for HPMS submittal (hypothetical)

Timeliness

Original Source Data – Continuous Counts

Due to the archival nature of traffic monitoring, timeliness is not as critical measure as in traffic operations and management. Data from about 180 sites with telemetry are polled every weekday by the DOT with the remaining 40 sites manually polled monthly. All data is received within time period required to process and submit to FHWA before the 20th of the month deadline.

Timeliness of continuous data collection is not as critical as for ITS applications like traveler information. However, the time elapsed between the download of the data to the review and approval of the data is important. Long review times can result in delayed identification of detector problems and consequently loss of data. According to information provided by the state DOT, the average review time (time from download to approval) for ATR data is about a week.

Original Source Data – Short-Term Counts

Data is collected by district crews. It is sent to central office for processing every 1-2 weeks. Data is typically sent by email from the count crews, however some of the data is relayed through central office personnel via Take Away Memory (TAM) cards.

Timeliness of data collection is important to short-term counts in a similar manner to continuous counts. Once the data is collected, the time to review, approve and upload to the database is critical. In this case, the average time for data to be sent to the central office from the time it was collected is 2 weeks. The time to review and approve data is another 2 weeks.

Totally, 4 weeks are required for short-count data to be collected, processed, reviewed and uploaded.

Archived Data

The timeliness measure for data users indicates the availability of data when they require it. For example, FHWA drives the 20th of every month deadline for submittal of permanent count data for the DOT. As a hypothetical example, let us assume the DOT is able to provide data to FHWA on the 20th of every month for 8 months in a year and by the 30th for the four remaining months. The timeliness measure for FHWA can be based on the number of timely submittals from the DOT (say 8 months of the 12 = 75%) and an average delay of 10 days.

$$\%\ timely\ data = \frac{8\ months - ontime\ submittal}{12\ months\ required} = 75\%$$

Coverage

Coverage for the DOT traffic monitoring program is driven by federal requirements and guidelines. The DOT bases the program on the Traffic Monitoring Guide 2001 and HPMS requirements. Table C.6 below compares the coverage requires with actual coverage in the state for continuous counts and Table C.7 shows the comparisons for short term counts.

Table C.6. Coverage for Continuous Counts

Recommended by TMG and HPMS	Actual Coverage in State
TMG –Traffic Volume – minimum of 5 Factor Groups with 5 to 8 ATRs per group. Vehicle Classification – Determine appropriate number of factor groups and assign at 6 continuous counters within each group.	*TMG* – 8 factor groupings based on functional classification. More than 6 permanent count stations per factor group. Higher functionally classified groupings have the greatest number of sites.
HPMS - At least one continuous counter on each major PAS/NHS highway route.	*HPMS* – there is at least one continuous count station on each "major" PAS/NHS route.

Table C.7. Coverage for Short-Counts

Recommended by TMG and HPMS	Actual Coverage State
TMG – Traffic Volume – roadway segment-specific data traffic count information on a cyclical basis. Vehicle Classification – Count all arterial and major collector roadways. *HPMS* – All HPMS universe, standard sample, and donut area sample sections.	*TMG* – Over 12,000 counts spotted statewide. One vehicle classification count spotted between each interchange of the entire Interstate system. At least one count spotted between all State and US routes statewide. *HPMS* – Standard samples are collected by state DOT using 24-hour vehicle classification counts.

Based on the two tables, it can be concluded that the DOT has adequate coverage with respect to the coverage requirements of the Traffic Monitoring Guide and HPMS. Thus, coverage is quantified as 100%.

Accessibility

Continuous data is combined and made available to public in an annual Traffic Survey Report. Internally, permanent data made available in a GIS/Web application on the intranet. Short term count hourly and AADTs made available to the public via the Traffic Survey Report and PKZIP ASCII files by individual count. AADT data is compiled and used in DOT's Traffic Survey report.

Qualitative information on the times required for data consumers to perform specified tasks is not available.

Interpretation of Data Quality Statistics

The data quality statistics for the Ohio case study are summarized in Table C.8.

Table C.8. Data Quality Summary

<table>
<tr><th rowspan="2">Quality Measure</th><th colspan="2">Original Source Data</th><th rowspan="2">Archived Data</th></tr>
<tr><th>Continuous Count</th><th>Short-Term Counts</th></tr>
<tr><td rowspan="2">Accuracy</td><td>ATR
Volume
• MAPE – 0.89%
• RMSE – 4 vehicles</td><td rowspan="2">Volume and classification
• MAPE -0.11%,
• RMSE – 2 vehicles</td><td rowspan="2">Estimates of AADTs are derived from ATR adjustment factors and ADT data from short-term counts. Accuracy of AADT estimates depends on the underlying accuracy of ATRs and short-term counts.</td></tr>
<tr><td>WIM
Vehicle Length
• MAPE – 7.26%,
• RMSE – 1.89 feet
Gross Vehicle Weight
• MAPE – 4.17%
• RMSE – 1,760 pounds</td></tr>
<tr><td>Completeness</td><td>86.3%</td><td>100% (Only complete counts are accepted)</td><td>92% based on HPMS submittals</td></tr>
<tr><td>Validity</td><td>94.7%</td><td>95%</td><td>88.8% based on HPMS submittals</td></tr>
<tr><td>Timeliness</td><td>Average review time for ATR data is one week.</td><td>Average time from data collection to final approval is 4 weeks.</td><td>20th of every month deadline met 75% of the time
Average delay when deadline was not met – 10 days.</td></tr>
<tr><td>Coverage</td><td colspan="3">100% of TMG and HPMS requirements</td></tr>
<tr><td>Accessibility</td><td colspan="3">Data is readily accessible</td></tr>
</table>