

Toward the Systematic Improvement of Data Quality in the Portland, Oregon Regional Transportation Archive Listing (PORTAL)

Kristin A. Tufte
Department of Computer Science
Department of Civil & Environmental Engineering
Portland State University
P.O. Box 751
Portland, OR 97207-0751
Phone: 503-725-3312, Fax: 503-725-5950
E-mail: tufte@cecs.pdx.edu

Soyoung Ahn
Department of Civil & Environmental Engineering
Portland State University
P.O. Box 751
Portland, OR 97207-0751
Phone: 503-725-3312, Fax: 503-725-5950
E-mail: sahn@pdx.edu

Robert L. Bertini
Department of Civil & Environmental Engineering
Portland State University
P.O. Box 751
Portland, OR 97207-0751
Phone: 503-725-4249, Fax: 503-725-5950
E-mail: bertini@pdx.edu

Ben Auffray
Department of Civil & Environmental Engineering
Portland State University
P.O. Box 751
Portland, OR 97207-0751
Phone: 503-725-3312, Fax: 503-725-5950
E-mail: bauffray@cecs.pdx.edu

James Rucker
Department of Computer Science
Portland State University
P.O. Box 751
Portland, OR 97207-0751
Phone: 503-725-3312, Fax: 503-725-5950
E-mail: jgrucker@cs.pdx.edu

Submitted for presentation and publication to the
86th Annual Meeting of the Transportation Research Board
January 21–25, 2007
Nov 15, 2006

Word Count: 5249 + 4 Figures + 5 Tables = 7499 words

ABSTRACT

The Portland Oregon Regional Transportation Archive Listing (PORTAL) is the official intelligent transportation systems data archive for the Portland metropolitan region. PORTAL has been archiving 20-second speed, count, and occupancy data from the approximately 500 inductive loop detectors in the Portland metropolitan region since July 2004. With a web-based interface, PORTAL provides performance metrics designed to assist practitioners and researchers. In order to provide accurate and useful information, it is essential that PORTAL process and maintain high-quality data, and further it is critical that users understand exactly how the data are processed. Users should also have a clear understanding of the quality of the data that they are viewing or extracting. There are several elements involved in maintaining quality data: effective allocation of maintenance resources for the inductive loop detectors, identification of suspect data values, and proper data aggregation. This paper presents a methodology for automatically detecting malfunctioning inductive loop detectors; describes how suspect data values are detected, handled, and reported in PORTAL; and finally describes PORTAL's data aggregation procedures. Examples of tools used to communicate data quality to PORTAL users are also provided.

INTRODUCTION

Data collection in any form always involves known and unknown errors. Understanding, managing and dealing with data quality is crucial in an Intelligent Transportation Systems (ITS) data archive. The ability to collect high-quality data and to track how its quality changes over time helps ensure that performance metrics and results produced by the archive are accurate and useful. In addition, a data quality monitoring system can help establish proactive improvement strategies and allocate ITS hardware maintenance resources more efficiently. Inductive loop detectors are a primary source of ITS data; however, inductive loop detectors may produce planned or unplanned poor quality data for several reasons such as detector or controller misconfiguration, communication failures, construction activities, and sporadic data errors. Monitoring and maintaining data quality has several facets. Loop detectors require proper installation and periodic maintenance to function properly; effective maintenance requires efficient allocation of limited maintenance resources and knowledge of the detectors most in need of maintenance. Further, maintenance itself will not eliminate poor-quality data; therefore, data suspected of being invalid must be flagged and eliminated from certain aspects of aggregation and analysis. Finally, aggregation must be performed correctly and data quality information must be communicated to the users as it is important for an ITS archive to have the ability to indicate quickly and graphically to a user how a given detector or series of detectors in a corridor performed over a particular time interval.

To address these issues, a data quality monitoring and management process has been developed for the Portland Oregon Regional Transportation Archive Listing (PORTAL). PORTAL is the ITS data archive for the Portland metropolitan region and has been archiving loop detector data from freeways in the Portland area since July 2004. The PORTAL system has a policy to retain all data in its most raw form and make it available to users. In addition, PORTAL produces an array of aggregated data and performance measures using well-documented aggregation procedures. The PORTAL data quality process includes automatic detection of malfunctioning inductive loop detectors using data stored in the PORTAL database, which addresses both detector configuration issues and communication problems. Data cleaning and aggregation in PORTAL are described herein as well as data quality display in PORTAL.

LITERATURE REVIEW

Various research has been done on data quality for ITS data archives and inductive loop detector data. The Texas Transportation Institute (TTI) has published guidelines for ITS data archives (1); this document provides suggestions for identifying suspect values including standard threshold tests as well as sequential, historical and spatial tests. Further research built on the original TTI research by analyzing San Antonio Trans Guide data to optimize thresholds for particular controller types and to analyze spatial and temporal trends in data quality failures (2). In addition, the FHWA has published a document on Traffic Data Quality Management (cite) which provide guidelines for calculating data quality measures and points out the importance of identifying the users and uses of the data (3). Early work in data quality used a series of threshold tests on 20-second data to eliminate samples that fell outside of a valid region in the count-occupancy plane (4). Turochy, et al. provide a method for detecting erroneous data including threshold tests and a test based on a calculation of effective vehicle length; this final test requires a real speed measurement (5). An algorithm that uses reference stations to detect and correct errors due to poorly calibrated sensors was proposed by Wall, et al.; high-quality reference stations are first located and then those reference stations are used to test and correct samples from adjacent stations (6). The PeMS project has proposed an algorithm that analyzes a day's worth of detector data to determine if the detector is performing well; the motivation

of this algorithm is to use multiple samples to evaluate a detector as opposed to tests on individual samples (7). A downside of this algorithm is that it includes or excludes detectors at the granularity of one day; which is too coarse of a granularity for the Portland region. Rilett and Vanajakshi propose data quality tests using conservation of vehicles (8); these techniques may be difficult to apply to PORTAL data as off-ramp counts are not available. Finally, there is a body of work that uses items such as loop signatures and event-level data to improve data quality (9); however this work is not directly applicable to PORTAL as the loop detectors in the Portland area report 20-second aggregated data. A further area of study is imputation. The PeMS project has proposed using linear regression for imputation (7). Smith, et al. provide an overview of imputation techniques; including discussions of types of data problems (missing data, erroneous data) and methods for evaluating imputation techniques (10).

PORTAL SYSTEM AND DATA

PORTAL, the ITS data archive for the Portland metropolitan area, was developed in accordance with the Archived Data User Service (ADUS) framework (11) developed by the U.S. Department of Transportation (DOT) as part of the National ITS Architecture. PORTAL archives a variety of ITS data including data from inductive loop detectors on Portland-area freeways, weather data, and incident data, and will soon store automatic vehicle location (AVL) data from buses in the Portland transit system. As suggested in FHWA ITS Guidelines (1), user access to the PORTAL data archive is through a web-based interface providing easy access to both raw data sets and a wide range of common summary data and performance measures. PORTAL's users include transportation planners, metropolitan planning organizations (MPOs), traffic management operators, transit operators, and transportation researchers.

PORTAL receives a live stream of freeway loop detector data from the Oregon Department of Transportation (ODOT). This stream consists of 20-second count, occupancy and speed measurements from the freeway and ramp detectors in the metropolitan area. PORTAL has been archiving the loop detector data since July 2004. Loop detectors in Portland are positioned in each mainline lane (typically just upstream from on-ramp locations) and on 138 metered on-ramps (ramp detectors only report count); dual mainline loops are used. Count and occupancy are measured; speed is estimated from the count and occupancy. In Portland, the mainline loops are dual-loop configurations; however, the loop controllers are not capable of measuring speed directly at this time. The data are stored in a PostgreSQL relational database management system (RDBMS) and data and performance measures are accessed through PORTAL's web-based front end (12).

AUTOMATIC DETECTION OF MALFUNCTIONING DETECTORS

Properly functioning loop detectors are crucial to quality data; further, inductive loop detectors require regular maintenance to function properly. In order to allocate their loop detector maintenance resources effectively, it is important for ODOT to know that detectors have been installed correctly and to understand which detectors are most in need of maintenance. Manually inspecting the loops or even all the data from the nearly 500 loop detectors in the Portland area is prohibitive in terms of time; therefore, a system for automatically detecting which detectors require maintenance is necessary. Two primary categories of loop detector malfunction are detector or controller configuration errors and communication failures. The next two sections present a methodology for automatically detecting both types of problems.

CONFIGURATION ERRORS

Configuration errors consist of problems occurring at detectors or controllers including miscalibration of a detector, incorrect detector location or spacing, and misconfiguration of the controller. Figure 1 shows a set of flow-occupancy plots for various detectors in the Portland area; the plots are scatter plots of 20-second samples over the period June 26, 2006 – June 30, 2006 (M-F) and are automatically generated by PORTAL. In all plots the *x*-axis is occupancy and the *y*-axis is flow; the range of the axes is 0-100% for occupancy and 0-4000 vehicles/hour for flow. The plots demonstrate various types of errors exhibited by detectors in the Portland area. In plots (a), (b), and (c) for detectors 1011, 1204, and 1212, respectively, the reported occupancy never reaches above around 20%. This is particularly strange for detectors 1204 (b) and 1212 (c), since they are in highly-congested areas. Detector 1244 (d) reports a flow of 0 always, with varying occupancies; while detector 1481 (e) consistently reports an occupancy of 0 with varying flows. Other detectors show other types of abnormalities; detectors 1695 (f), 1123 (g) 1637 (h) show unusual scatter in the flow-occupancy curve, particularly in the high-occupancy region of the curve. Detector 1787

(i) shows an unusually low maximum occupancy and unusual scatter. As will be shown below, all of these detectors are flagged as malfunctioning by our detector configuration tests.

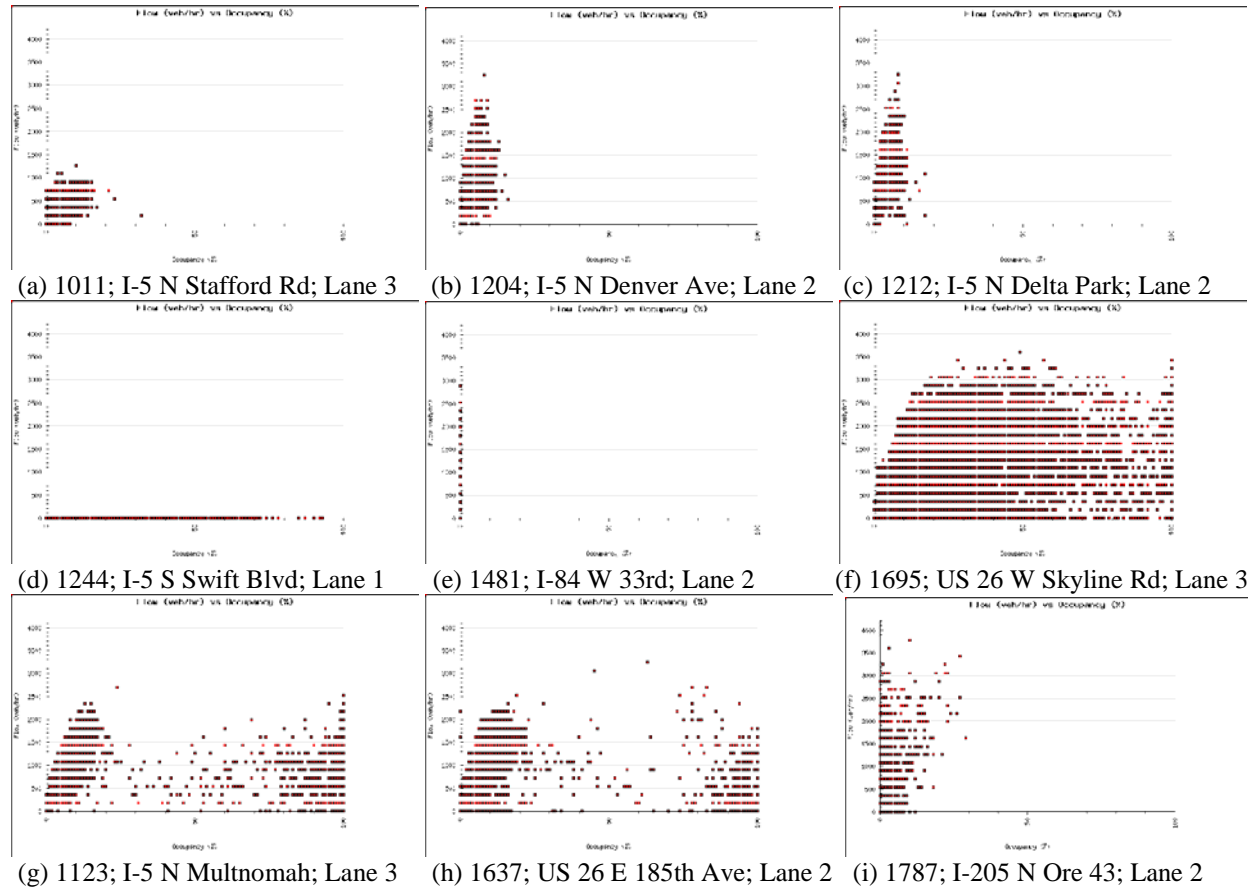


FIGURE 1 Flow-occupancy plots for sample detectors; July 26, 2006–July, 30 2006 (M-F).

In PORTAL, detector configuration errors are detected using a set of tests adopted from a publication on monitoring urban freeways (13). The tests apply to 20-second aggregated data. Each test includes a condition; data samples that satisfy that condition are considered improbable and may indicate a malfunction. For example, a report of a 20-second count greater than 17, which corresponds to a flow of over 3000 vehicles/hour, is considered improbable. Similarly, a 20-second speed reading of greater than 100MPH, while possible, is unlikely. Thus “count > 17” and “speed > 100 MPH” are the conditions of two of the PORTAL data quality tests.

All detectors periodically report unlikely values; therefore, a detector should only be considered as a maintenance priority when it reports higher-than-average occurrences of suspicious values. Therefore, the PORTAL quality tests consist of a condition (i.e. “count > 17”) and a threshold. The threshold is generally specified as a percentage of samples for which a detector fails the condition. For example, a detector producing speeds of greater than 100MPH for more than 5% of its samples is likely to be malfunctioning. For each test, there is a condition and a set of thresholds as shown in Table 1. The use of a set of thresholds is motivated by the fact that one can not easily make a clear distinction between a malfunctioning detector and a properly-functioning detector; rather there is a spectrum of malfunction. Setting of thresholds is discussed below.

TABLE 1 Detector Configuration Tests: Conditions and Thresholds

<i>Condition</i>	<i>Threshold</i>		
	Low	Medium	High
Count > 17	0.2% of samples	0.3% of samples	1% of samples
Occupancy > 95%	0.25% of samples	0.5% of samples	5% of samples
Speed > 100 MPH	0.5% of samples	1% of samples	5% of samples
Speed < 5 MPH	5% of samples	10% of samples	30% of samples

Low max occupancy	Max occupancy < 25	Max occupancy < 20	Max occupancy = 0
Low average occupancy peaks	Average occupancy < 6%	Average occupancy < 5%	Average occupancy < 4%

In PORTAL, the above tests are used to create a monthly report that lists detectors in need of maintenance and which is given to ODOT. The report is created as follows. The tests are run over weekdays in a one-month period (see below for discussion on the selection of a one-month report period). For each condition, for each detector, the number of times the detector “fails” the condition is counted. If the percentage of times the detector fails the condition is greater than a threshold, the detector is considered to fail the test at that threshold. For example, a detector is said to fail the “Count > 17” test at the High threshold if more than 1% of that detector’s samples over the evaluation period show a count greater than 17. To limit the number of detectors listed as “failing,” and to begin a systematic detector improvement program, the high threshold is currently used. Future work will include tuning of the thresholds so that detectors passing the tests are ensured to produce a certain quality of data. Table 2 shows the number of mainline detectors failing the tests at the different thresholds for weekdays in May and June 2006; percentage of detectors failing is shown in parentheses. There are approximately 350 mainline detectors in the Portland area. Peak periods are defined to be 7-9 AM and 4-6 PM.

TABLE 2 Number (Percent) of Detectors Failing Tests at Various Threshold Levels

Condition	Threshold					
	Low		Medium		High	
	May 06	June 06	May 06	June 06	May 06	June 06
Count > 17	14 (4%)	18 (5%)	5 (1%)	12 (3%)	2 (1%)	0 (0%)
Occupancy > 95%	26 (7%)	26 (7%)	15 (4%)	17 (5%)	7 (2%)	7 (2%)
Speed > 100 MPH	15 (4%)	15 (4%)	7 (2%)	7 (2%)	1 (0%)	1 (0%)
Speed < 5 MPH	36 (10%)	36 (10%)	27 (8%)	22 (6%)	9 (3%)	11 (3%)
Low max occupancy	25 (7%)	28 (8%)	11 (3%)	10 (3%)	8 (2%)	7 (2%)
Low average occupancy peak periods	34 (10%)	36 (10%)	22 (6%)	22 (6%)	14 (4%)	14 (4%)

Table 3 shows a sample of a monthly report provided to ODOT for a portion of June 2006. The report contains a list of detectors with the poorest performance over the most recent month based on the above tests. The report includes detector number, detector location, milepost, lane number, station number, a list of tests failed by that detector, and a place for ODOT to provide feedback to Portland State University (PSU). ODOT is physically inspecting the detectors identified in the report and is reporting back to PSU what problems were found and what measures were taken to resolve the issues. PSU will monitor the maintained detectors for improvements in data quality and will use the feedback to understand the relationship between the types of tests a detector fails and the maintenance required on those detectors. In the future, the report will be web-based; an email link to the report will be automatically emailed to ODOT maintenance and Transit Monitoring Operations Center (TMOC) personnel and PORTAL staff. ODOT maintenance personnel will enter their feedback using the web-based interface.

TABLE 3 Sample Monthly Configuration Errors Report Provided to ODOT (Partial Report for June 2006)

DetectorId	Location	Milepost	Lane Num	StationId	Tests Failed	Action Taken
1011	I-5 N Stafford Rd WB to NB	286.3	3	1002	Low avg occ	
1244	I-5 S Swift Blvd/Marine Dr SB	307.35	1	1027	Speed < 5	
1123	I-5 N Multnomah Blvd NB	296.6	3	1012	Occ > 95, Speed < 5	
1204	I-5 N Denver Ave NB	306.51	2	1022	Low max occ, Low avg occ	
1205	I-5 N Denver Ave NB	306.51	3	1022	Low avg occ	
1206	I-5 N Denver Ave NB HOV	306.51	1	1022	Low avg occ	
1212	I-5 N Delta Park NB	306.51	2	1023	Low max occ, Low avg occ	
1695	US 26 West Skyline Rd WB	71.07	3	1092	Occ > 95	
1637	US 26 East 185th Ave SB to EB	64.6	2	1084	Occ > 95	

1787	I-205 North Ore 43 NB-NB	9.03	2	1110	Low avg occ
1481	I-84 West 33rd WB	2.1	2	1062	Speed < 5

Report Period

A one-month report period over weekday data has been chosen for the PORTAL monthly configuration errors report. The one-month time period was selected to provide a long enough baseline to flatten out day-to-day variances without being overly long. Observations of data indicate that periods longer than one month may be overly influenced by transient events such as construction. To evaluate the validity of the report period, Figure 2 plots the number of samples from detectors on Highway I-5 failing selected conditions over weekdays in the period January-June 2006. Highway I-5 had limited to no construction during this period and thus provides a reasonable sample. Over this time range, the number of samples failing the configuration error conditions ranges from approximately 600,000 to slightly over 700,000; the total samples per month ranges from 10-12.5 million. As the number of samples failing in each category remains relatively stable, Figure 2 indicates that one month is a reasonable report period.

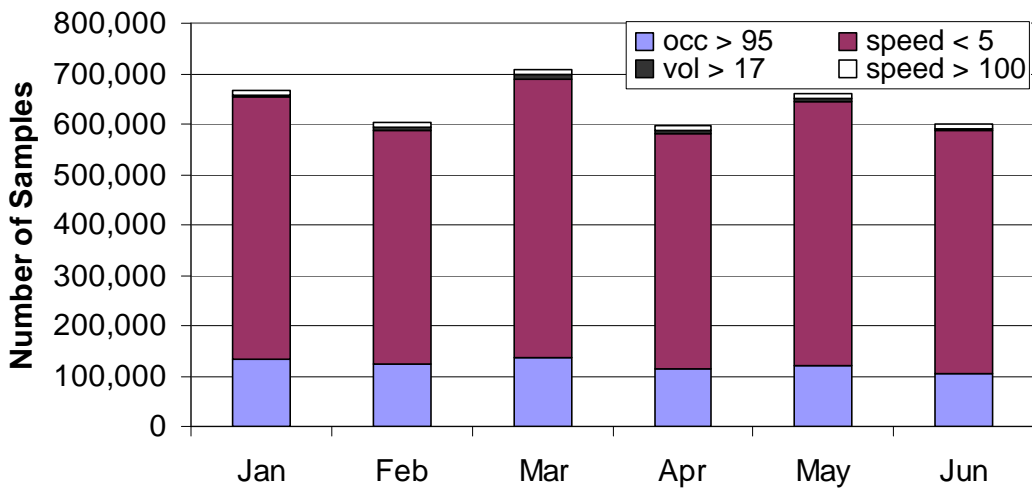


FIGURE 2 Number of Samples Failing Selected Conditions; January-June 2006; highway I-5.

Setting Thresholds

As shown in Table 1, each test condition is associated with a set of thresholds. While the conditions were adopted straightforwardly from the TTI and Cambridge Systematics report (13), setting the thresholds appropriately was more complicated. While it seems clear that it would be unusual for a detector to report speeds greater than 100MPH for more than 5% of its samples, it is not clear what percentage of samples with speeds greater than 100 MPH would be received from a normally-functioning detector. In fact, observations indicate that there is not a precise delineation between malfunctioning and properly-functioning detectors. Observations reveal that detectors reporting very high percentages of failures of various conditions (i.e. 15% of speed samples > 100MPH or 40% of occupancy samples > 95%) produce high percentages of probably-invalid data; detectors producing lower, but still moderately high, percentages of failures, appear to still produce probably-invalid data, but at a lower rate. Even properly-functioning detectors produce probably-invalid values, but at a low rate. Therefore, it may not be possible to classify detectors as “malfunctioning” or “OK”, but rather that there is a spectrum of malfunction and the goal should be to identify the detectors which are worst on the failure spectrum.

Several techniques were used to help determine the thresholds. First, for each condition, for each detector, the percentage of times the detector failed each condition was determined and detectors were ranked by their percentages. These data were inspected along with flow-occupancy plots (as in Figure 1). The threshold levels were thus set by inspecting the failure rates of various detectors and inspecting how those failure rates related to abnormalities seen in flow-occupancy plots.

COMMUNICATION FAILURES

Communication failure is a second major cause of loop detector data errors. The data that PORTAL receives from ODOT is coded to indicate various conditions. For example a record with speed = -1, count = -1, occupancy = -1 and status = 0 indicates a communication failure. On the other hand, a record with speed = 0, count = 0, occupancy = 0 and status = 0 indicates that no traffic passed over the detector in the associated 20-second time interval. A reading of no traffic is theoretically a valid reading; however, it appears that the ODOT system periodically reports a reading of no traffic during highly-congested periods; it has been theorized that this may be a communication failure.

To identify detectors with potential communication problems, detectors reporting communication failures a high percentage of time and detectors reporting no traffic a high percentage of time are flagged; both tests are conducted only during peak periods (7-9AM and 4-6PM). Similar to the configuration error tests, each communication failure test consists of a condition and a set of thresholds. Sporadic communications failures are expected even on the best links; detectors with the highest failure rates will be addressed first followed by detectors exhibiting lower failure rates. Table 4 shows the communication failure tests and associated thresholds as well as the number of detectors (and percentage of detectors) failing the conditions at the various thresholds for weekdays in May and June 2006.

TABLE 4 Communication Failure Tests: Conditions, and Thresholds, and Number (Percent) Failing Detectors

<i>Condition</i>	<i>Threshold</i>					
	<i>Low</i>		<i>Medium</i>		<i>High</i>	
High percentage communication failure during peaks	5% of samples		15% of samples		30% of samples	
	May 06	June 06	May 06	June 06	May 06	June 06
Number (percent) of Failing Detectors	19 (5%)	27 (8%)	12 (3%)	22 (6%)	9 (3%)	8 (2%)
High percentage no traffic samples during peaks	10% of samples		20% of samples		55% of samples	
	May 06	June 06	May 06	June 06	May 06	June 06
Number (percent) of Failing Detectors	46 (13%)	49 (14%)	38 (11%)	36 (10%)	6 (2%)	6 (2%)

Information on communication failures is also included in the monthly detector status report sent to ODOT, in the form of a map and a report. The communication failure map for June 2006 is shown in Figure 3; this map graphically shows all detectors failing either the Communication Failure or No Traffic condition at the medium threshold for weekdays in June 2006. Table 5 shows a portion of the June 2006 Communication Failures report provided to ODOT. This report provides a list of stations that appear to have communication issues. The report is in terms of stations instead of detectors since communications failures occur at the station level, not the detector level. A station is a set of detectors in one location on the highway. In the Tests Failed column, "Comm Failure" indicates that the detector had a high percentage of samples indicating a communication failure; "No Traffic" indicates the detector had a high percentage of samples indicating no traffic. COULD DELETE THIS TABLE TO SAVE SPACE

TABLE 5 Sample Communication Failures Report Provided to ODOT (Partial Report for June 2006)

<i>StationId</i>	<i>Location</i>	<i>Milepost</i>	<i>Tests Failed</i>	<i>Action Taken</i>
1020	I-5 N 1188 Alberta St NB	304.4	Comm Failure	
1038	I-5 S Upper Boones SB	291.25	Comm Failure	
1094	US 26 West Cedar Hills SB to WB	68.2	Comm Failure	
1013	I-5 N Terwilliger Blvd NB	297.33	Comm Failure	
1135	Broadway to I-405 SB	1.08	No Traffic	
1110	I-205 N Ore 43 NB-NB	9.03	No Traffic	
1134	I-405 S 5th to I-405 SB	0.97	No Traffic	
1130	I-5 S Lombard EB to SB	305.4	No Traffic	

stations. Free-flow speed is defined as the minimum of: assumed free-flow speed and actual five-minute speed. Assumed free-flow speed is 60MPH. Free-flow speed is computed in this way in order to avoid negative delay.

Five-Minute Count: Valid count samples are summed and then rescaled to represent a five-min count. When all the data is present, there are fifteen data samples in a five-minute period. If fewer than fifteen samples are present, then the count is rescaled by the number of missing samples, so as to indicate a five-minute count.

Five-Minute Occupancy and Speed: All valid occupancy (speed) samples in a five-minute period are averaged, using a count-weighted average, to provide five-minute occupancies (speeds).

Five-Minute Vehicle Miles Traveled (VMT):

Five-Minute VMT = five-minute count * length of influence area.

Five-Minute Vehicle Hours Traveled (VHT):

Five-Minute VHT = five-minute count * (length of influence area / five-minute speed)

Five-Minute Travel Time:

Five-Minute Travel Time = length of influence area / (five-minute speed)

Delay (in minutes): Five-minute delay = travel time – (influence area / free-flow speed)

The above formulas and descriptions describe how all five-minute performance measures are calculated in PORTAL. The results of the five-minute performance measure calculations are stored in tables in the PORTAL database and these tables are used to produce per-detector timeseries plots of any performance measure and to calculate additional per-detector performance measures such as congestion frequency, 95% travel time, and average travel time at the daily and monthly level. While it is useful to have these five-minute performance measures at the detector level, PORTAL must also provide those performance measures over larger freeway segments as well as larger time intervals.

AGGREGATION OVER LARGER INTERVALS, LANES AND STATIONS

PORTAL provides analysis to its users at the detector, station, highway and freeway system levels. The above-described five-minute aggregates support analysis and graphs for individual detectors; to provide graphs and analysis at a coarser granularity, PORTAL data must be further aggregated over time, lanes and stations. For example, if we wish to calculate the VMT for I-5 Northbound over a particular time interval t , we must aggregate data over the time interval t , then we must aggregate across the lanes for each detector station in I-5 North, and finally, we must aggregate over all detector stations on I-5 Northbound. The following describes how aggregation over larger time intervals, lanes and stations is performed in PORTAL.

Aggregation over Larger Time Intervals

The PORTAL database stores pre-computed fifteen-minute and one-hour aggregates for the seven PORTAL performance measures. These aggregates are calculated as follows:

Count: Five-minute counts are summed to computer larger-interval count. To create a fifteen-minute count, three five-minute counts are summed; 12 five-minute counts are summed to create a one-hour count;

Occupancy and Speed: Larger-interval occupancy (speed) is a count-weighted average of the appropriate five-minute occupancies (speeds).

Vehicle Miles Traveled (VMT): Larger-interval VMT is the sum of the appropriate five-minute VMTs.

Vehicle Hours Traveled (VHT): Larger-interval VHT is the sum of the appropriate five-minute VHTs.

Travel Time: Larger-interval Travel Time is a count-weighted average of the appropriate five-minute Travel Times.

Delay (in minutes): Average: Larger-interval Delay is the count-weighted average of five-minute delays.

Aggregation over Lanes and Stations

PORTAL does not store pre-computed lane or station aggregations; these aggregations are performed on-the-fly as necessary and are computed as below. Note that aggregations over lanes in a station are performed to obtain station-level data; aggregations over station-level data are then performed to obtain corridor-level or system-wide data.

Count: Sum counts from appropriate lanes. (Aggregation over stations is generally not necessary.)

Occupancy: Count-weighted average of occupancies from appropriate lanes. (Aggregation over stations is generally not necessary.)

Speed: Count-weighted average of occupancies (speeds) from appropriate lanes/stations.

Vehicle Miles Traveled (VMT): Sum VMTs from appropriate lanes/stations.

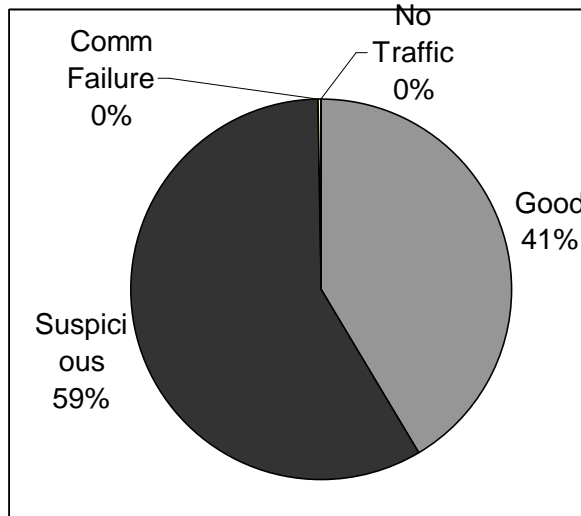
Vehicle Hours Traveled (VHT): Sum VHTs from appropriate lanes/stations.

Travel Time: Lanes: Count-weighted average of Travel Times from appropriate lanes. Stations: Sum Travel Times over appropriate stations.

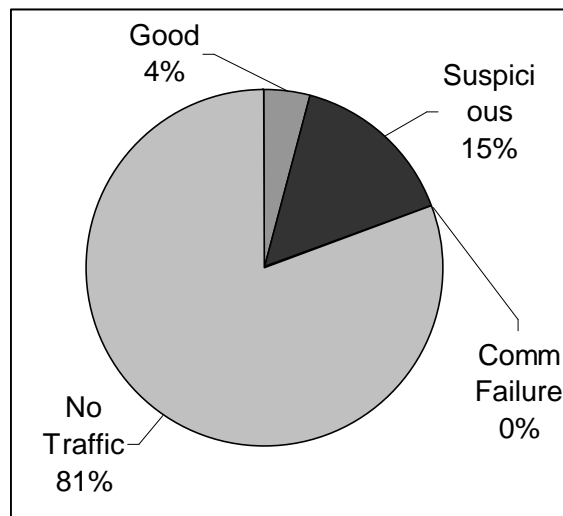
Average Delay: Lanes: Count-weighted average of Delays from appropriate lanes. Stations: Sum Delays over appropriate stations.

DATA QUALITY DISPLAY IN PORTAL

While it is important to detect invalid or suspicious data, it is also important to communicate to the user the quality of the data at a given detector at a given time. PORTAL uses pie charts and time-space corridor plots to represent data quality. Data samples are broken into four categories: Good, Suspicious (fails one or more data quality conditions), No Traffic, and Communication Failure. PORTAL users can request a data quality pie chart for a detector, or for a highway corridor over any period of time. Figure 4A shows the data quality pie chart for detector 1123 (I-5 N Multnomah Blvd NB; Lane 3) for June 26, 2006. As shown in Table 3, Detector 1123 failed several configuration error data quality tests in June 2006. Figure 4B shows a data quality pie chart for the detector in lane 2 at station 1110 (I-205 N Ore 43 NB-NB; Lane 2) for June 26, 2006. As presented in Table 5; this station is exhibiting significant communication issues; its communication issues manifest as an unreasonably high percentage of No Traffic samples.



(a) Detector: 1123; I-5 N Multnomah Blvd NB; Lane 3



(b) Station 1110; Hwy: I-5 N; MP 286.3; Lane 2

FIGURE 4 Display of monthly data quality information in PORTAL (June 2006).

CONCLUSION

This paper describes the measures PORTAL uses to improve and maintain the quality of the data in its archive. Automatic detection of malfunctioning detectors and a monthly report to ODOT listing the detectors most in need of maintenance have been described. Detection of malfunctioning detectors focuses on detector configuration errors and communication errors. Initial reports on detector malfunctions have been provided to ODOT and PSU is awaiting feedback from ODOT. This feedback will be used to improve the data quality tests. Methods for detecting suspicious data values as well as aggregation and performance measure calculations in PORTAL are described in detail as well as the availability of new data quality information to PORTAL users. The reports and procedures described in this paper form the basis for the PORTAL data quality process; this process will evolve and improve as the effects of the detector maintenance and data quality process are observed.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Foundation for funding this research. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The National Science Foundation assumes no liability for the contents or use thereof. The contents do not necessarily reflect the official views or policies of the National Science Foundation. This report does not constitute a

standard, specification, or regulation. The authors gratefully acknowledge the Oregon Department of Transportation and, in particular, would like to thank Dennis Mitchell and Jack Marchant for their contributions to the project.

REFERENCES

1. S. Turner. *Guidelines for Developing ITS Data Archiving Systems*. Report 2127-3. FHWA, U.S. Department of Transportation, Texas Department of Transportation and Texas Transportation Institute, 2001.
2. Quiroga, C., Hamad, K., Park, E. S. *Incident Detection Optimization and Data Quality Control*. Report FHWA/TX-06/0-4745-3. Texas Transportation Institute. <http://tti.tamu.edu/documents/0-4745-3.pdf>
3. *Traffic Data Quality Measurement, Final Report*. Federal Highway Administration, Washington, D.C., Sept. 2004. Available at http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/14058.htm.
4. Jacobson, L., Nihan, N., and J. Bender, Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System. In *Transportation Research Record 1287, Transportation Research Board*, Washington, D.C., 1990.
5. Turochy, R.E. and B.L. Smith. A New Procedure for Detector Data Screening in Traffic Management Systems. In *Transportation Research Record 1727: Journal of the Transportation Research Board*, Washington, D.C., 2000.
6. Wall, Z.R., D.J. Daily. Algorithm for Detecting and Correcting Errors in Archived Traffic Data. In *Transportation Research Record 1855: Journal of the Transportation Research Board*, Washington, D.C., 2003.
7. Chen, C., Kwon, J., Rice, J., Skabardonis, A., Varaiya, P. Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems. In *Transportation Research Record: Journal of the Transportation Research Board*, Washington, D.C., 2003.
8. Vanajakshi, L. and Rilett, L.R. Loop Detector Data Diagnostics Based on Conservation-of-Vehicles Principle. In *Transportation Research Record 1870: Journal of the Transportation Research Board*, 2004, pp. 162-169.
9. Coifman, Benjamin; Dhoorjaty, Sudha. Event Data Based Traffic Detector Validation Tests. *ASCE Journal of Transportation Engineering*, Vol 130, No 3, 2004, pp 313-321.
10. Smith, BL, W.T. Scherer, J.H. Conklin. Exploring Imputation Techniques for Missing Data in Transportation Management Systems. In *Transportation Research Record 1836: Journal of the Transportation Research Board*, Washington, D.C. , 2003.
11. *Archived Data User Service (ADUS): An Addendum to the ITS Program Plan*. FHWA, U.S. Department of Transportation, September 1998.
12. Bertini, Robert L; Hansen, Steve; Byrd, Andrew; Yin, Thareth. Experience Implementing a User Service for Archived Intelligent Transportation Systems Data. In *Transportation Research Record: Journal of the Transportation Research Board*, Washington, D.C., 2005.
13. *Monitoring Urban Freeways in 2003: Current Conditions and Trends from Archived Operations Data*. Texas Transportation Institute and Cambridge Systematics, December 2003.
14. Cleghorn, D., F. Hall, and D. Garbuio. Improved Data Screening Techniques for Freeway Traffic Management Systems. In *Transportation Research Record 1320, Transportation Research Board*, Washington, D.C., 1991.