Research Report
Agreement T2695 Task 85
IR Evaluation

DATA QUALITY HANDLING APPROACH OF TRACFLOW SOFTWARE

TECHNICAL REPORT

by

Duane R. Wright
Systems Analyst Programmer

John M. Ishimaru
Senior Research Engineer

Washington State Transportation Center (TRAC)
University of Washington, Box 354802
1107 NE 45th Street, Suite 535
Seattle, Washington 98105-4631

Washington State Department of Transportation Technical Monitor
Ted Trepanier
State Traffic Engineer

Sponsored by

Washington State Transportation Commission
Department of Transportation
and in cooperation with
U.S. Department of Transportation
Federal Highway Administration

May 2007

TECHNICAL REPORT STANDARD TITLE PAGE

WA-RD 679.1	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
4. TITLE AND SUBTITLE DATA QUALITY HANDLING APPROACH OF TRACFLOW SOFTWARE, TECHNICAL REPORT		5. REPORT DATE May 2007 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Duane R. Wright and John M. Ishimaru		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Washington State Transportation Center (TRAC) University of Washington, Box 354802 University District Building; 1107 NE 45th Street, Suite 535 Seattle, Washington 98105-4631		10. WORK UNIT NO. 11. CONTRACT GRANT NO. Agreement T2695 Task 85
12. SPONSORING AGENCY NAME AND ADDRESS Research Office Washington State Department of Transportation Transportation Building, MS 47372 Olympia, Washington 98504-7372 14 Doug Brodin, Project Manager, 360-705-7972		13. TYPE OF REPORT AND PERIOD COVERED Final Research Report 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This study was conducted in cooperation with the University of Washington and the US Department of Transportation

16. ABSTRACT

The TRACFLOW software processes induction loop data to develop performance metrics for freeways in the Seattle area. The loop data are sometimes subject to errors. To find and correct errors, the TRACFLOW system uses a three-step approach to detect and address variations in the quality of the traffic data. Each step can include data replacement if sufficient supporting data are present. This combination of methods is automated whenever feasible to more efficiently handle the large data sets involved.

This report describes the three steps, detailing how each contributes to cleaner and more robust data sets. The objectives of these methods are to detect a higher percentage of anomalous data points, replace them with higher quality values, enable more of the data to be used, and increase overall automation of the process.

17. KEY WORDS Freeways, performance, data quality, loop detectors, data banks, software		18. DISTRIBUTION STATEMENT	
19. SECURITY CLASSIF. (OF THIS REPORT) None	20. SECURITY CLASSIF. (OF THIS PAGE) None	21. NO. OF PAGES	22. PRICE

DISCLAIMER

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

CONTENTS

Section	*Page*
EXECUTIVE SUMMARY	**vii**
INTRODUCTION	**1**
EVOLUTION OF TRACFLOW'S DATA QUALITY PROCESS	**2**
DESCRIPTION OF BASIC FLAG-BASED METHODS (LAYERS 1 AND 2)	**4**
1) Low-Level Individual Data Point Evaluation	5
Overview	5
Detection of Bad Data with Flags	5
Modification of Bad Data with Flags	7
2) Quality Inventory Tables and Analyses	8
Overview	8
Display of Bad Data Frequency in Flag Summary Tables	9
Highlighting Bad Data in Flag Summary Tables	11
Modification of Bad Data by Using Flag Summary Tables	11
DESCRIPTION OF MACROSCOPIC METHODS AND FILTERS (LAYER 3)	**12**
Overview	12
Detecting Bad Data Prior to Imputation	13
Imputing Bad Data: Overview	14
Step One of the Imputation Process: Determine Historical Relationships	15
Step Two of the Imputation Process: Calculate Replacement Candidates	20
Step Three of the Imputation Process: Select Replacement Values	20
SUMMARY	**21**
REFERENCES	**21**

FIGURES

Figure		*Page*
1	The 11 possible lane pairs for the imputation process on a four-lane section of freeway	15
2	Relationship between volumes of a neighboring lane (Lane 2) and volume of the lane being imputed	19

TABLES

Table		*Page*
1	The four error types	14
2	The four day types	16
3	Available good data for generating linear relationships and imputation candidates	17

EXECUTIVE SUMMARY

This report documents research carried out for the Washington State Department of Transportation to enhance the effectiveness of methods to detect and address variations in the quality of traffic data used by the TRACFLOW system. TRACFLOW is a software system that processes induction loop data to develop performance monitoring metrics for freeways in the Seattle area. The TRAFLOW system uses three methods to detect and address variations in the quality of traffic data.

Method one uses automated data scanning to look for patterns of questionable values and replaces those values by using historical relationships with nearby good data. Data are reviewed at the loop-day level. Method two produces automated summary tables of individual data point quality at the 5-minute level; the tables are produced on the basis of 5-minute data quality flags from the WSDOT FLOW raw loop data archive. Data are reviewed at the individual 5-minute level. The resulting summary tables are then used in a manual review process that is based on professional judgment. Method three uses microscopic automated value-by-value review and replacement of individual data points based on data quality flags. Data are reviewed by this method at the individual 5-minute level. Each of these steps specializes in detecting certain types of data inconsistencies, and the three steps usually occur sequentially.

This report describes the three steps, detailing how each contributes to cleaner and more robust data sets. The objectives of these methods are to detect a higher percentage of anomalous data points, replace them with higher quality values, enable more of the data to be used, and increase overall automation of the process.

INTRODUCTION

The TRACFLOW software processes loop data to develop performance metrics for freeways in the Seattle area. The loop data are sometimes subject to errors associated with field sensor malfunctions, aggregation and transmission errors, or interruptions related to construction. In many cases, the erroneous data values are readily detectable by examining the output metrics based on those data. However, some erroneous data have more subtle effects on the computed metrics; in such cases, it may be difficult to detect the presence of questionable data on the sole basis of a review of the analytical output. Therefore, a review of data quality prior to analytical work is important.

This technical working paper describes the TRACFLOW software's methods for handling data quality as of March 2007. The TRACFLOW system uses a three-step approach to detect and address variations in the quality of the traffic data that it uses to compute freeway performance monitoring metrics. The three-step process consists of

1. automated scanning of the data for patterns of questionable values
2. manual review of automatically produced data pattern summaries
3. microscopic automated value-by-value review of individual data points.

Each step can include data replacement if sufficient supporting data are present. This combination of methods is automated whenever feasible to enable the large data sets involved to be scanned more efficiently.

The initial scanning stage (the first step of the three-step process described above) is the most recent addition (March 2007) to the TRACFLOW data review process and reflects a desire to automate the method by which patterns in the data sets are reviewed (supplementing a previous semi-automated process). It is anticipated that this new

feature will more efficiently filter out questionable data points that might otherwise unduly or subtly bias performance estimates.

EVOLUTION OF TRACFLOW'S DATA QUALITY PROCESS

The TRACFLOW system's approach to ensure input data quality has evolved since the mid-1990s when development of the foundation of the system first began. The principal input data set is a multi-year archive of 5-minute volume and occupancy percentage loop data collected in the Seattle area freeway network by the Washington State DOT's FLOW data collection system, and compiled quarterly on CDs. The method that was originally developed to evaluate data quality relies primarily on data quality indicators (flags) that accompany each data point in the archive. These data quality flags assign a quality estimate (good, bad, suspect) to the associated value and also detect incomplete data collection (e.g., as a result of manual shutdown by traffic system operators); these flags enable the point-by-point evaluation and potential replacement of bad or questionable values. Their inclusion in the raw data sets also enables automated processing via software; the software analyzes the flags of each data point and uses that information to replace or skip the associated data value as necessary. This method is still in use today and represents the first layer of data quality review in the TRACFLOW process.

After experience was developed with this methodology, as well as familiarity with the data set, researchers determined that although a review of individual data quality flags is useful, some data quality problems are more easily detected by a higher-level summary of data characteristics over time. This led to the development of tools to construct data quality inventory tables that summarize the flag status on a month-by-month and lane-by-

lane basis; these summaries enable the analyst to review the condition of large sets of data relatively quickly and make some determinations about the temporal and spatial distribution of data quality. The summary values are supplemented with color coding; the colors are used as visual cues to facilitate quicker review of overall quality patterns. This information enables the analyst to determine, for example, that a loop exhibits sufficiently irregular behavior over time that its data should be removed altogether from a corridor analysis, a conclusion that microscopic point-by-point data replacement (without knowledge of the overall data quality trend) might overlook. This method represents the second layer of data quality review in the TRACFLOW process.

The review of broader patterns of data quality using these inventory tables is usually performed manually by an analyst and is based on professional judgment. Because it can be difficult to establish a definitive standard of quality below which a loop is considered "bad," however, this manual review process can be subjective and also limits the ability to automate the process. Furthermore, experience with the data sets revealed that some questionable patterns are difficult to discern by looking only at the flag values, but they might be clear upon visual inspection of the data values themselves. To address this, the tables were enhanced by introducing non-flag-based criteria, such as the percentage of loop counts that exceeds a fixed, very high occupancy (thus potentially indicating loop error); however, this output also requires a manual review.

These limitations of the inventory tables are exacerbated by the fact that the flag values are not only sometimes insufficient, but they are also not always accurate (e.g., on occasion, values have been incorrectly flagged as good). Extensive experience with the

data sets has shown that while such values are not common, their occurrence can subtly but noticeably affect the analytical results.

To overcome these limitations, other data evaluation methods were reviewed to determine how they might a) enhance data quality and b) evaluate broader data quality patterns in a more automated fashion. Two methods were chosen for combined implementation as a third layer of data quality review in the TRACFLOW process. This layer is a pre-processing step performed prior to the existing value-by-value review (the combined layers one and two). One of the methods is a variation of the approach described by Chen et al (2003), which avoids the difficulties of determining the quality of each specific data point by instead looking for day-long patterns of questionable data; questionable values are then replaced by using historical relationships between the loop in question and nearby loops. The second method implements additional automated tests to detect data points that are without question unusable, even when their flag values indicate they are good (for example, lane occupancy values greater than 100 percent). Such values are flagged and sent to the value-by-value replacement process.

The following is a general overview of the methods used in the three layers of data evaluation.

DESCRIPTION OF BASIC FLAG-BASED METHODS (LAYERS 1 AND 2)

The TRACFLOW process includes two methods to use the flags in the raw data to analyze data quality: 1) a value-by-value evaluation of data quality, and 2) a review of aggregated summaries of flag occurrences.

1) Low-Level Individual Data Point Evaluation

The following description is condensed from the *CDR User's Guide* by Ishimaru (1998).

Overview

The TRACFLOW software takes advantage of preexisting data validity flags produced for each 5-minute data point in the WSDOT FLOW loop data archive. The flag values are used by the TRACFLOW software to determine whether a data point can be used or should be replaced (if possible) by nearby good data. This process automatically evaluates individual data points one at a time at a low level (each 5-minute value) and replaces bad data points with nearby good data, both spatially and temporally. If replacement is not possible, the questionable values are not used.

Detection of Bad Data with Flags

The WSDOT FLOW loop data, as compiled on CDs on a quarterly basis, are archived as 5-minute records. A typical portion of a data file is as follows:

Time	Vol	Occ	Flg	nPds
00:00	5	0.4%	1	15
00:05	12	1.1%	1	15
00:10	25	2.1%	1	15

"Time" is the 5-minute period. "Vol" is the volume (number of vehicles passing over the loop during the period). "Occ" is the lane occupancy percentage, the percentage of time a vehicle is detected by the loop. "Flg" is the data quality flag. The flag value is 0, 1, 2, or 3, representing Bad, Good, Suspect or Disabled By Operator, respectively. The data quality is first determined at the 20-second level, then evaluated at the 5-minute level (see below). A second quality indicator, "nPds" (number of periods), indicates the

number of 20-second records available in the 5-minute period (thus nPds normally equals 15, since 15 x 20 seconds = 5 minutes).

Each 5-minute data point validity flag is automatically determined by the WSDOT FLOW system on the basis of a data quality evaluation of each of the 15 20-second values that make up that 5-minute value, as follows:

- **Bad Data**: If the loop is locked in an "on" state for longer than a prescribed time period or the loop's data value is outside a defined volume-occupancy "envelope" (based on WSDOT research regarding the combination of reasonable values in the two-dimensional volume/occupancy space), the corresponding 20-second data value is labeled "Bad Data." The assumption is that a locked "on" state or highly atypical combinations of volume and occupancy values are symptomatic of erroneous data collecting conditions.
- **Disabled**: If the loop has been disabled by a system operator, the corresponding 20-second data value is labeled "Disabled."
- **Good Data**: If the loop does not meet the criteria for the first two conditions, the corresponding 20-second data value is considered "Good Data."

After each of the 15 20-second counts that make up a 5-minute value is evaluated in this way, the associated 5-minute value is given an overall data validity flag according to the following rules:

- If all 15 20-second values are labeled "Disabled," the associated 5-minute data value is considered "**Disabled**" (flag = 3).

- If all 15 20-second values are labeled “Good,” the associated 5-minute data value is considered “**Good**” (flag = 1).
- If all 15 20-second values are labeled “Bad,” the associated 5-minute data value is considered “**Bad**” (flag = 0).
- For all other combinations of 20-second validity flags, the associated 5-minute data value is considered “**Suspect**” (flag = 2).

Note that the suspect category is conservative, in that only one of the 15 20-second values needs to be considered bad or disabled for the entire 5-minute value to be declared suspect. Therefore, the TRACFLOW software gives the user the option to treat suspect data as good if there is reason to believe that the suspect values are actually good.

Modification of Bad Data with Flags

If a data point is determined to be bad, replacement can occur at one of three levels, in the following sequence:

Within a lane: First, a replacement value is located by searching for the closest (in time) good data within that lane, from the same location (loop) on the same day, subject to a maximum threshold (e.g., the closest good data point no more than 15 minutes from the bad data point). The program first looks to the previous record (previous 5-minute value), then to the following record (the 5 minutes after), then to the second previous record (10 minutes prior), then to the second following record (10 minutes after), and so on. As the time window increases, the likelihood of a reasonable replacement value diminishes; hence the use of a search limit. The user can specify the allowable search limit. (This method is derived from an approach WSDOT uses in its analytical software (Ishimaru 1998).)

Across Lanes: Next, bad lane data that cannot be temporally replaced within its lane can be spatially replaced by using an average of data for the same 5-minute time interval from adjacent lanes. It is not necessary for every lane to have good data in order to derive a cross-lane average. The user can specify the minimum percentage of lanes that must have good data in order for a cross-lane average to be computable.

Data Replacement via Interpolation or Extrapolation: Finally, if replacement cannot be accomplished at the location level, and a corridor or trip is being processed, data can be replaced along the corridor or trip route by interpolation or extrapolation from nearby loops (upstream/downstream). The user can specify the maximum search distance for good data.

For each method, the replacement of questionable data can be performed automatically by software. If good replacement data are not available by any of the above methods, the original data point will not be used.

The user can adjust the process by which the detection and replacement of questionable data takes place by activating options to 1) use or replace data labeled as suspect, 2) modify the search parameters when looking for nearby good replacement data, or 3) change the threshold for the minimum amount of good data required to use values in subsequent computations.

2) Quality Inventory Tables and Analyses

Overview

The flag occurrences in the raw data can be summarized as tables, to enable the analyst to more quickly and easily evaluate broader temporal and spatial patterns of data quality and detect questionable patterns. These summaries are generated with the CD DQ

software utility, and color-coding is used to highlight potential quality issues. Unlike the flag-based method described above, the resulting summary tables are reviewed manually.

Display of Bad Data Frequency in Flag Summary Tables

The automated, flag-based system described earlier as flag approach 1 will identify many, but not all, individual cases of questionable data. For situations in which there is a high percentage of questionable data, however, it may be preferable to simply not use a data set (e.g., from a particular loop), rather than rely on replacement of individual data points from a potentially sparse collection of nearby good data. Those situations can be more readily detectable if they are viewed in a broader context, e.g., aggregated over time or space, so that longer-term patterns can be revealed.

For that reason, the TRACFLOW process includes a software utility that produces flag-based breakout summaries of data quality for different time and location intervals (month by month, lane by lane, cabinet by cabinet, over a calendar year). These summaries provide an overview of data quality for each location, and they help the analyst determine which data sets should and should not be used and to what extent the quality patterns indicate possible seasonal or lane biases that might affect analytical results. The user can then determine the best response to these patterns

CD DQ (Data Quality) is the program TRAC developed to summarize and report on the flag-based data quality of WSDOT FLOW data. Output from the program is in two files: **dq_out_sum.txt**, which summarizes data at the cabinet or location level (i.e., summarized across all lanes), and **dq_out.txt**, which summarizes the same flag data, but at the individual lane level.

Details of the contents of these files follow.

dq_out_sum.txt

The following values are presented for a given combination of location (summed across all lanes), direction of travel (e.g., northbound), and lane type (e.g., general purpose or HOV).

%Flags Available by Month

This is the percentage of expected data that is actually available.

Bad, Good, Suspect, Disabled

This is a summary of the percentage of data points in each data quality flag category.

High Occ (>35%) (flag 1 only)

This is the percentage of data, with a flag = 1, that has lane occupancy above 35 percent.

Stuck (>12) (flag 1 only)

This is the percentage of data, with a flag = 1, that comprises 12 or more consecutive repetitive volumes or occupancies.

Missing

This is the percentage of expected data NOT available. Note that the "Missing" percentage plus the "%Flags Available by Month" percentage should equal 100.

Bad Lane (<70% 1 flags)

This is the lane number of the worst of the lanes with less than 70 percent good (flag = 1) values (i.e., the lane with the lowest percentage of flag 1 data). A list of the months with more than one bad lane is displayed as well.

dq_out.txt

This file is a lane-by-lane version of the same data shown in **dq_out_sum.txt**. The values are presented for a given combination of location, lane, direction of travel, and lane type (e.g., general purpose or HOV).

Highlighting Bad Data in Flag Summary Tables

To assist the analyst in the review process, values in the DQ_OUT_SUM.TXT table can be color-coded according to a fixed scale of quality (e.g., Green = 75 to 100 percent, Yellow = 50 to 75 percent, Red = less than 50 percent). Note that the category colors vary on the basis of the type of value being displayed; for example, if the value were "% Good Data," Green would represent 75 to 100 percent good data, whereas if the value were % Bad Data, Green would represent 0 to 25 percent bad data (which would be the same as 75 to 100 percent non-Bad data). The color-coding is performed automatically with an Excel macro, ColorDQ.

Modification of Bad Data by Using Flag Summary Tables

As noted earlier, unlike the automated flag-based approach 1, the DQ_OUT_SUM.TXT and DQ_OUT.TXT summary tables require manual review by an analyst. Typical temporal and spatial patterns to look for include the following:

- Long-term data quality problems at a given cabinet location. This is indicated if there are a significant number of months of bad (red), or at least non-good (red + yellow), data.
- Long-term data quality problems at a given loop. This is indicated if there are a significant number of months of bad, or at least non-good, data at a given lane. Note that this requires looking at the data from DQ_OUT.TXT, which shows lane-by-lane information.
- Large data quality gaps along a corridor or trip route. This is indicated if there is a specific month or months of bad, or at least non-good, data at a series of

consecutive cabinet locations. Note that this may require looking at the data from DQ_OUT.TXT, which shows lane-by-lane information.

Once a pattern has been noticed, the analyst must determine what course of action to take. This requires first determining the threshold of acceptable data quality to use. For example, if a cabinet has bad data for several months of the year, one must determine a) what constitutes a month of bad data (e.g., red or just yellow), and b) what constitutes a minimum acceptable number of good months of data (e.g., at least three-quarters of the year and/or sampling from every season of the year).

Potential responses to bad data patterns include one or more of the following:

- Skip bad time periods (months)
- Skip bad locations
- Skip bad lanes at specific locations
- Skip bad locations or lanes along a corridor or trip route

These steps usually require editing a batch file that describes the trip or location being analyzed. Because the changes can be small (e.g., skipping one lane at one location), it is important to document the change in the file name or some other notation in the file or the analytical output.

DESCRIPTION OF MACROSCOPIC METHODS AND FILTERS (LAYER 3)

Overview

The use of summary tables in Layer 2, described above, suggests the benefits of looking at aggregated results over time. However, the table-based method requires manual inspection and still relies on only the flags for quality evaluation. Therefore in

Layer 3, the most recent update to TRACFLOW's data quality approach, an automated evaluation of aggregated results uses the actual values rather than the flags. This method evaluates a day of data for each loop, looking for unusual patterns. When such a pattern is located, the entire day is replaced with new data. The replacement data are based on historical relationships between nearby loop data and the loop data being replaced.

In addition, filters have been put in place to detect data points that are considered good by the flag system but should not be considered for use in the imputation process. Examples include occupancies of greater than 100 percent, or nPds values of other than 15; values collected while the loop has been disabled by the operator (flag = 3) are also detected. These values are not only removed from the imputation process but are also flagged for subsequent replacement using the Layer 1 and Layer 2 processes.

The resulting hybrid approach to data enhancement combines the microscopic, point-by-point flag-based approach (Layer 1) and the flag summary tables (Layer 2) with a day-by-day detection and enhancement process (Layer 3) that produces broader automated quality enhancement throughout the dataset prior to any analytical task, thus reducing the need to perform a time-consuming exploration of questionable data points.

Detecting Bad Data Prior to Imputation

The detection method, prior to imputation, reviews the data values of each loop on a day by day basis. The values of each day of each loop's data are evaluated for the likely occurrence of several types of data errors, on the basis of the presence of unusual values or data patterns. This process is an adaptation of the approach described by Chen et al.(2003). Table 1 shows the types of data values or patterns that are tracked.

Table 1. The four error types (from Chen et al. 2003)

Error Type	Measurement	Type of Error Check
Type I	Occupancy = 0	Loop is stuck off
Type II	Occupancy > 0 when Volume = 0	Loop is hanging on
Type III	Occupancy > 35%	Loop is hanging on
Type IV	Entropy (of Occupancy)	Loop is stuck off or on

If at least one of these conditions occurs with sufficient frequency in a single day, all the data from that loop for that one day are considered "bad," and are chosen for replacement. The threshold of sufficient frequency for each error type is chosen by the user and based in part on the observed distribution of the error type in the FLOW data. In particular, when the distribution of data shows a bifurcated pattern, with bad and good data clearly separated by a large midrange that has few data points, this allows flexibility in setting the threshold in the midrange.

Only the day's data from 05:00 to 22:00 are evaluated. Late night data are not processed for evaluation purposes because of the relative difficulty of detecting unusual data quality patterns.

Imputing Bad Data: Overview

For each loop-day that has been determined to require imputation, the data of the entire 24-hour day are replaced by using historical relationships between nearby surrounding loops and the imputed loop. Each surrounding loop (including loops at the closest upstream and downstream cabinets as well as those at the same cabinet location) is evaluated to determine the relationship between values at that loop and those at the

loop being imputed. The evaluation uses recent past data, or, when not enough recent past data are available, near-term future historical data and a linear regression process. Once the relationships with each nearby loop have been established, each associated function is used to then contribute a candidate value for replacement for each 5-minute interval during the day. In the example in Figure 1, lane 2 of the middle location is the loop being imputed. The surrounding 11 loops are each eligible to potentially contribute a candidate, one of which might eventually become the replacement value.

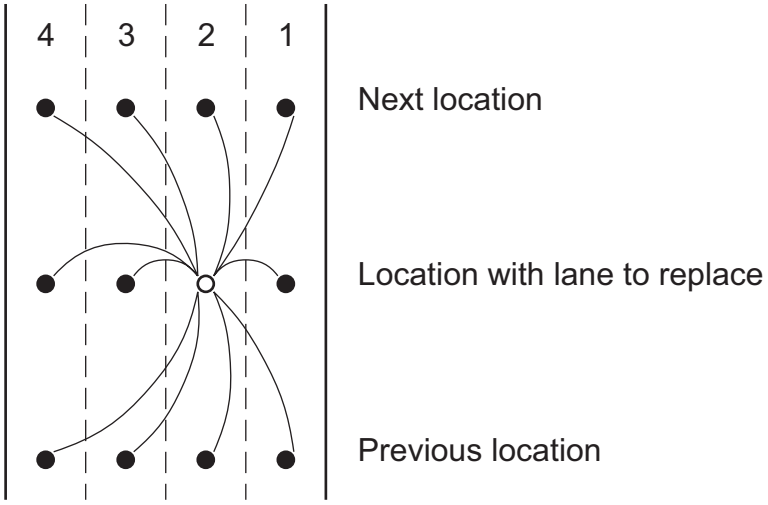

Figure 1. The 11 possible lane pairs for the imputation process on a four-lane section of freeway (from Chen et al. 2003, based on Figure 4)

Step One of the Imputation Process: Determine Historical Relationships

Replacement values for a bad loop-day are computed by first determining relationships between values at the bad loop and values at nearby good loops. Each pairwise relationship uses the bad loop as one member of the pair; each of the other loops at the same (cabinet) location, and each loop at each of the two adjacent (i.e., upstream and downstream) locations, is then used as the other member of the pair.

To develop these relationships, data are collected for each nearby loop, X, and the loop being imputed, Y. Historical data are taken from up to five days of recent data. To better address any time-of-year variations or day-of-week differences, only data for the five previous days of the same day-of-week type (see Table 2) are considered. For example, if the day of the loop data being imputed is a Monday, only Monday historical data are used. All 24 hours of data are used to develop the relationships. Only loops with good data, based on the diagnostic tests in Table 1 and flag settings, are used.

If five previous days of good data are not available, up to five days in the future can be checked (bringing the total to five days) for data that meet the two criteria. This process is restricted to locations in the same freeway facility. (Each corridor is processed separately, and every available loop with good data is used.)

Table 2. The four day types

• **Saturday, Sunday**
• **Monday**
• **Tuesday, Wednesday, Thursday**
• **Friday**

Table 3 gives an example of the lanes and dates involved in a typical imputation process. In this example, the bad loop location being processed is labeled "Current." The adjacent locations are "Upstream" and "Downstream." The "Good" and "Bad" labels are the loop quality evaluations from the previous diagnostic step, determined as the data were first read.

Table 3. Available good data for generating linear relationships and imputation candidates

Day	Date	Lane	Location		
			Upstream	Current	Downstream
Thursday	26	1	GOOD	GOOD	GOOD
		2	BAD	IMPUTE	BAD
		3	BAD	GOOD	GOOD
		4	GOOD	GOOD	BAD
Wednesday	25	1	GOOD	GOOD	GOOD
		2	GOOD	GOOD	BAD
		3	GOOD	BAD	BAD
		4	BAD	GOOD	GOOD
Tuesday	24	1	GOOD	GOOD	BAD
		2	GOOD	GOOD	BAD
		3	GOOD	GOOD	GOOD
		4	BAD	BAD	GOOD
Thursday	19	1	GOOD	BAD	GOOD
		2	BAD	GOOD	GOOD
		3	GOOD	GOOD	GOOD
		4	BAD	GOOD	GOOD
Wednesday	18	1	GOOD	BAD	GOOD
		2	GOOD	BAD	GOOD
		3	BAD	GOOD	GOOD
		4	GOOD	BAD	BAD
Tuesday	17	1	GOOD	BAD	GOOD
		2	GOOD	GOOD	BAD
		3	GOOD	GOOD	GOOD
		4	BAD	BAD	GOOD

The loop-day being imputed is Thursday the 26th, Lane 2. The five previous days of this day type (see Table 3) are Wednesday (25th), Tuesday (24th), Thursday (19th), Wednesday (18th), and Tuesday (17th). For each of the possible pairs (a nearby loop matched with Lane 2 of the current location), these five previous days will be sources of data pairs to use in creating the historical relationship. For example, for the combination of Lane 1 in the upstream location and the loop of interest (Lane 2 at the current

location), we combine all the good data points from each of the five eligible days, as long as both loops are good on those days. On Wednesday the 25th, for example, Lane 1 of the upstream location and Lane 2 from the Current location are both good, so values from that loop (on that day) are matched with those from the loop being imputed (on the same day). Those values then contribute to the collection of data points used to develop the linear fit for that specific combination of loops.

Note how the data quality of a given loop on a given day affects the usability of those data for this process. In the case of Wednesday the 18th, for example, no data can be contributed for any nearby loop X because Lane 2 of that day, for the Current location, is bad. But other pairs are valid for the four other dates, so ample data are available to create the historical relationships.

After the eligible days of data for a given pair of X and Y loops have been determined, a trend line is fitted, by using a linear regression process, to the combined data of the pair. This relationship is shown in a typical scatter plot in Figure 2. There is a separate trend line process for volume and occupancy. For cases in which the trend line goes below zero, negative imputed values are set to zero. Occupancy is capped at 100 percent.

The result is a series of functions

$$Y_i = f_i(x_i)$$

where

Y_i = the replacement value from the relationship between Y and the i^{th} nearby loop

f_i = the linear least-squares function describing the relationship between historical data values at loop Y and the i^{th} nearby loop

x_i = the data value from the i^{th} nearby loop

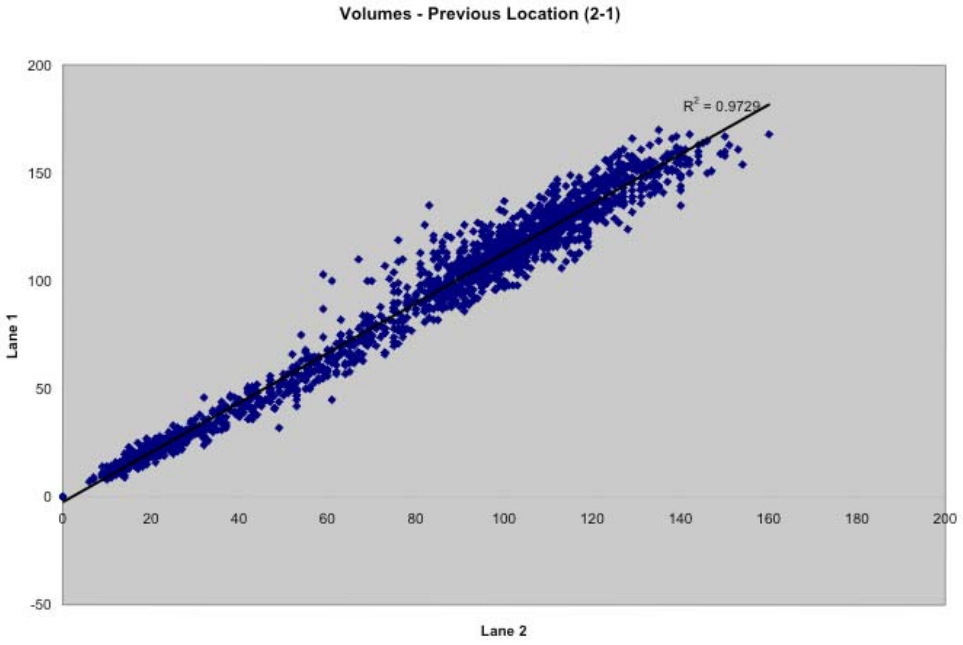

Figure 2. Relationship between volumes of a neighboring lane (Lane 2) and volume of the lane being imputed

Once the relationships (for each lane pair) have been developed, each relationship is used to determine a potential replacement value for the bad loop. This is done by using the value from the matching nearby loop, on the day of the data being imputed, as the input to each good loop's associated fit formula.

Step Two of the Imputation Process: Calculate Replacement Candidates

For each 5-minute period of the 24-hour loop-day being replaced, a pool of candidate replacement values is produced; each formula from step 1 contributes a candidate. Neighboring good data, for the same day as the loop data **being replaced**, are used as the input values to the formulas to produce the replacement candidate values. For each 5-minute period, the appropriate formula from step 1 for a particular nearby loop and data type (volume or occupancy) is used. The input value at the nearby loop X must meet these conditions: the value a) must be from a day that meets all error tests, b) must have a flag value of 1, and c) must not have an occupancy of > 100 percent or an npds of other than 15. If the X value does not meet all these conditions, that loop X does **not** contribute a candidate value. Note that each input data value must come from a loop that is good. In our example of four lanes, a maximum of 11 relationships can be developed from each of the 11 nearby loops (assuming good data are available for all lanes). However, a review of Thursday the 26th (the date of the data **being replaced**) shows that a maximum of seven of the 11 relationships can actually be used to produce candidates for the replacement values because the other loops on that day do not have good data.

This process is performed for volumes and occupancies independently.

Step Three of the Imputation Process: Select Replacement Values

For each 5-minute pool of candidates, the candidates are sorted in ascending order, and the median candidate becomes the replacement value. In the event of an even number of candidates, the average of the middle two candidates is used. The endpoint locations on a corridor can also be imputed, even though they can not have a location before and after them. For this process, the two closest locations (upstream, or

downstream, as the case may be) are used in the regression and generate replacement value candidates.

SUMMARY

The result of these enhancements is a data quality evaluation system that detects a higher percentage of anomalous data points, replaces them with higher quality replacement values, enables more of the data to be used (instead of simply being skipped), and does so through a more automated process. These enhancements do not yet produce a perfectly clean data set that no longer requires scrutiny of its analytical products, but they do address some of the more nagging issues that in the past have required significant levels of analytical effort to resolve. As such, they provide practical time-saving benefits. In addition, these enhancements form the foundation for further data quality filters that can fill even more of the quality "holes" in the data set.

REFERENCES

Chen, et al. Detecting Errors And Imputing Missing Data For Single Loop Surveillance Systems. 82nd Annual Meeting of the Transportation Research Board. Washington, D.C. January 2003.

Ishimaru, J.M. CDR User's Guide, Washington State Transportation Center (TRAC), University of Washington, Seattle, WA, March 1998.