Final Report

CLEAN-UP OF EXISTING DATA SETS TO SUPPORT DYNAMIC MOBILITY APPLICATIONS DEVELOPMENT

by

Mark E. Hallenbeck
TRAC Director

Yinhai Wang
Professor

John Ishimaru
Senior Research Engineer

Xiaolei Ma
Graduate Research Assistant

Mike Richards
Software Engineer

Jonathan Corey
Graduate Research Assistant

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

Contracting Office Technical Representative
Randy Butler
Transportation Specialist
U.S. Department of Transportation
Office of Freight Management and Operations

Prepared for
United States Department of Transportation
Federal Highway Administration
DTFH61-11-H-00026

May 2012

DISCLAIMER

This material is based upon work supported by the U.S. Department of Transportation under Cooperative Agreement No. DTFH61-11-H-00026. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the Authors and do not necessarily reflect the view of the U.S. Department of Transportation.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

DESCRIPTION OF THE SEATTLE MUTI-MODAL CORRIDOR PERFORMANCE DATA SET

1.0 INTRODUCTION

The Seattle multi-modal corridor data set covers the Interstate 5 freeway corridor from the King/Pierce County line in the south to approximately the City of Everett in the north. The arterial data included in the data set cover the major City of Seattle north--south arterials west of I-5 and east of Puget Sound. The specific geographic areas covered by the data are shown in figures 1 and 2. The data set contains 6 months of data, collected from May 1, 2011, to October 31, 2011.

Figure 1[i]: Geographic Area for I-5 Included in the Data Submittal

Figure 2[ii]: City of Seattle Arterials Included in the Data Submittal

The data set includes data on freeway performance provided by Washington State Department of Transportation (WSDOT) inductance loops that are used to operate the Department's surveillance, control, and driver information system. To allow interesting performance analysis, these basic roadway performance data are accompanied by data collected by WSDOT's incident response program, the Washington Incident Tracking System (WITS) and data on the messages displayed on both WSDOT's Active Traffic Management signs and its conventional dynamic message signs.

While the freeway data represent only I-5, several other types of data that help explain the performance of the freeway are included in the data set. These additional data are select arterial data from City of Seattle arterials that operate next to or connect with the I-5 corridor, a variety of transit availability and performance data sets, weather data, and spot speed data taken from GPS-equipped trucks traveling in the region.

The most detailed roadway performance data available are the "raw" freeway volume and lane occupancy statistics supplied every 20 seconds by each WSDOT inductance loop. These are the base freeway data that WSDOT collects at its traffic management center. In addition, a second copy of the 20-second data is supplied. This second copy differs from the first in that a data quality-based calibration process has been applied to the data. These data provide additional insight into whether the loops functioned as intended. The data quality process has also applied a correction factor to the lane occupancy statistic to compensate for loops that were functioning but were not correctly calibrated in the field due to calibration drift. When selecting data from the supplied interface, users may specify whether to download the "raw" or "processed" 20-second data.

For most planning and performance analyses, the 20-second data are too fine grained. For this reason, the data set also includes volume, occupancy, and speed data (speeds are calculated on the basis of volume and lane occupancy statistics) by loop and by location aggregated to a 5-minute time interval. A different approach to quality assurance was conducted for this level of data aggregation, and only processed data are available for the 5-minute data. The 5-minute data were further processed to compute travel times for specific origins and destinations. Travel times were computed every 5-minutes by using a traditional vehicle trajectory algorithm.

Also included for I-5 are incidents recorded by WSDOT's incident response program. Incident data include the reported time when each incident is first observed, the location of that incident, whether the incident blocked traffic lanes, and the response to each incident. The data set also includes all messages displayed by all of WSDOT's Active Traffic Management and other dynamic message sign displays on I-5. Finally, the freeway data include specific truck vehicle probe points that allow calculation of individual truck speeds and locations on state routes within the study corridor. These data are provided by trucks carrying GPS devices as part of fleet tracking services. The GPS points do not include information that can identify specific trucks over multiple days.

In addition to the freeway performance and incident data, the database also contains a variety of data on a number of arterials in the City of Seattle. Two kinds of arterial data are included: 1) travel times based on matched Automated License Plate

Recognition (ALPR) observations, and 2) traffic volumes from strategically located Sensys data collection devices. The corridors for which travel time data are available are shown in Figure 2 and are listed later in this document. Locations for which traffic volume data are available are also listed later.

To support multi-modal performance analysis and to provide additional data on corridor performance, the data set also includes transit performance information. The available transit data include bus routes, bus stop location identification, and the arrival times of buses at those stops. In addition, the data set provides commuter and light rail schedule information.

Finally, the data set contains weather data (including temperatures, precipitations, and wind speed). Weather data are available from five National Oceanic and Atmospheric Administration (NOAA) weather stations.

All data can be linked by time, date, and location. **Note that all time measurements are listed in current local time.** Each of the data sets is described in more detail below. In addition, for data representing time intervals, the time stamp reports the beginning of that time period unless explicitly noted otherwise (e.g., a travel time listed at 8:00 AM starts the trip at 8:00 AM.)

2.0 SEATTLE FREEWAY PERFORMANCE DATA

Seattle freeway performance data are available in three levels of aggregation:

- loop detector data aggregated at 20-second intervals (the lowest resolution available), which is useful for detailed road simulation model development and calibration
- loop detector data aggregated at 5-minute intervals (data more readily suited for planning-level analysis and model validation activities)
- corridor travel time estimates, computed every 5 minutes by using spot speeds from the 5-minute data archive and a classic vehicle trajectory algorithm.

Data are available on I-5 from the Pierce/King County line in the south (milepost 139.7) to the Stillaguamish River Bridge at milepost 209.3 in Snohomish County, which is located roughly 10 miles north of the City of Everett. Data are present—by individual loop—for all inductance loop detectors WSDOT operates on this stretch of I-5. This includes loops placed in all mainline general purpose and HOV lanes. Loops are also present on collector-distributor roadways (usually part of freeway to freeway interchanges) and on the I-5 reversible lanes in Seattle (which generally operate southbound in the morning and northbound in the evening[1]). The southern end of the

[1] On weekdays the Express Lanes operate southbound from 5:00 AM until 11:00 AM. The facility is then closed to switch operational directions. It then operates northbound from noon until 11:00 PM, at which time it is again closed. On weekends it operates southbound from 7:00 AM to 1:00 PM and then northbound from 2:00 PM to 11:00 PM. The Express lanes are generally closed late at night on both weekdays and weekends. Express Lane operations can be changed from these hours to accommodate special events. The details of the actual Express Lane

northbound I-5 reversible lanes start at the southern edge of downtown Seattle at milepost 165.3 and the northern end is at milepost 172.4 in the Northgate area.

Loops are also present on the ramps leading to and from the mainline. These loops are further designated as queue loops, left lane advanced queue loops, right lane advanced queue loops, intermediate queue loops, stop bar loops (at the ramp meter signal heads), and passage loops; depending on their location on the ramp and function within the WSDOT ramp metering system. Off-ramps generally have only one set of loops, which are placed before a ramp splits into multiple lanes approaching the terminus of that ramp. The data dictionary for the 20-second data set described below will assist users in identifying the kinds of loops they require for a particular research project.[2]

The 5-minute data aggregations cover this same segment of roadway and include the same loops. The only differences between these data sets are that the data have already been aggregated to the 5-minute interval level, and different quality assurance checks have been performed on the data. These data are described in Subsection 2.2 below.

The travel times provided with this data set are the I-5 corridor travel times traditionally computed for use by WSDOT in its freeway performance reporting activities. They are based on the assumptions applied as part of those computations. The supplied travel time trips are divided into southern and northern trips. Travel times are provided every 5 minutes, in both directions of travel, and for both general purpose (GP) lanes and high occupancy vehicle (HOV) lanes. Thus there are eight "trips" for which travel times are computed and stored in the database. The attributes of these trips are described in detail in the Travel Time subsection (Subsection 2.3) below. For trips with other start and end points, users must perform their own travel time computations.

2.1 20-Second Freeway Loop Data

For very detailed freeway performance evaluations, the data set provides freeway performance data for individual loop detectors every 20 seconds. Because this level of detail results in massive data sets (~30 GB per month for the I-5 data), the retrieval of large amounts of these data (e.g., data for the entire corridor for multiple days) can be time consuming and require considerable disk space. We recommend that users think carefully about the data they need before making multi-location, multi-day data extractions at the 20-second level.

---

operational status on a specific day can be determined from the variable message sign data contained in the "Other WSDOT Variable Message Sign Data" table in this database.

[2] Users should also be aware that the accuracy of volumes available from ramp locations is generally more suspect than that of mainline volumes because vehicles using ramps often display poor lane discipline in relation to the loop locations, thus decreasing the count accuracy of the inductance loops. The data cleaning process for this data set did not attempt to determine the accuracy of counts by comparing consecutive mainline volume estimates with on- and off-ramp volumes between those consecutive mainline locations.

2.1.1 Data Description

The 20-second data are divided into four different types of data tables, each of which is transmitted in csv format. The first type of data table (CabinetsInfo) describes the location of the cabinets which contain the loop electronics. While not precise, this is the best available location descriptor for the actual data collection loops. The second table (LoopsInfo) describes each specific loop. The information in this table describes whether the loop is collecting data on a specific general purpose lane (e.g., lane 2), an HOV lane, a ramp, or on some other kind of roadway. A variable (WSDOTCAB) in this table is used to reference the Cabinet table to obtain the physical location of the loop. The actual loop data (e.g., the volumes and scan counts collected by the field equipment), are stored in daily data tables, each submitted as a separate csv file, and identified with the name WSDOT_20Sec-2011-MM-DD, where MM is the month and DD is the day (date). When MM or DD is a 1 digit number there will be a leading zero in the file name. The data in these tables reference (and are referenced by) the LoopsInfo table using the variable LOOPID. The data in these tables have not been sorted into a specific order. Note that because of their large size, if Excel is used to open the 20-second data files, not all records will be loaded. Note also that the field equipment does not directly calculate either lane occupancy or average vehicle speed for the 20-second interval being reported. Instead, the field equipment collects vehicle volume and scan count. Scan count is the number of cycles during each 20-second interval that the loop is occupied. Since the loop detectors operate at 60 Hz, 1200 scans occur for each 20 second interval. Lane occupancy can be directly computed by dividing the scan count by 1200. Average speed for each 20-second interval can then be computed from volume and lane occupancy values using the formula[3]

$$v = q / (o*g)$$

where:

v = speed in miles per hour
o = lane occupancy expressed as a percent
q = volume in vehicles per hour per lane, and
g = a constant that incorporates site characteristics such as average vehicle length and loop detector field strength (and usually assumed by WSDOT to equal the value 2.4)

The last table (file) type contains correction factors developed to account for loop sensitivity issues identified in the quality assurance process. The quality assurance tests and calibration factor computation process are described in the Subsection 2.1.2 of this document. The correction factors are stored in monthly csv files and use the LOOPID variable to link to the actual volume and scan count data. The file naming convention for these files is CF2011MM.csv, where MM represents the month for which correction factor data are stored in that file. As with the 20-second data, a leading zero will be present in the name when the month is represented by a single digit. The specific contents of each of these four types of data tables are described in Table 1, which is subdivided into four sections Table 1a through Table 1d.

[3] See D.J. Dailey A Statistical Algorithm for Estimating Speed from Single Loop Volume and Occupancy Measurements, Transportation Research. Part B: Methodological, 1999.

Table 1: 20-Second Freeway Data Descriptions and Structure

Table 1a: WSDOT_20Sec_YYYY-MM-DD (where YYYY-MM-DD is the date)			
Columns	Data Type	Data Range	Value Description
LOOPID	2byte int	3267-8624	Unique ID number assigned in order of addition to LoopsInfo table
HHMMSS	4byte int	0-240000	24 hour local time in integer format, no leading zeroes (5pm = 170000, 12am =0, 4:35:22pm = 163522) For all volume data reported, the time is the start of the interval being reported.
YYYYMMDD	4byte int	20110501-20111031	Date in year month day format (May, 6, 2011 = 20110506)
Has_data	SMALLINT	0 - 1	An internal WSDOT data check. 0 = no data is present 1 = data were collected by this loop (this does not mean the data are valid.)
Scan_Count	INTEGER	0 - 1200	Loops are scanned every 1/60th of a second. This variable reports the number of times during the 1200 scans (60 x 20 seconds) that the loop is occupied. Lane Occupancy (in percent) is computed as 100 * Scan_Count / 1200 Note: values above 1200 indicate that the detector data comes from a malfunctioning detector
Flag	SMALLINT		0 = good data 1 = short pulse (<1/15th of a second) 2 = Chatter (>3 vehicles observed in 1 sec) 3 = Outside Volume/Occupancy envelope[4] 4 = Indicates a communication failure 5 = Bad Average (for station data, which is not provided in the federal data set) 6 = Operator has disabled this loop 7 = Bad loop
Volume	2byte int	0-~120	Integer volume observed during this 20 second interval

[4] The volume/occupancy ratio boundaries are not constant values, but a function of the occupancy percentage. The values used can be found in Figure 6,on page 25 of the report *Detector Data Validity*, WA.RD 208.1, March 1990.

Table 1b: CabinetsInfo (file: CabinetsInfo.csv)			
WSDOTCAB	10 char	XXXesYYYZZ	Name of roadside equipment cabinet, XXX=route code, es=electronic surveillance, YYY.ZZ=milepost
LAT	decimal	~47N	Latitude to six decimal places
LONG	decimal	~122 W	Longitude to six decimal places
ROUTE	3 char	'005'	Name of State Route or Interstate on which road equipment cabinet is located
MILEPOST	decimal	139.63-209.33	Milepost location of the data collection cabinet to the hundredth of a mile. This is a good approximation of the physical loop location, which can be slightly up- or –downstream of this location.
LOCATION	20 char	Text	20 character description of the loop location, typically the cross street or a land mark (e.g., Olive Way, Green River Bridge, etc) used to help a human locate the loop location
Table 1c: LoopsInfo (file: LoopsInfo.csv)			
WSDOTCAB	10 char	XXXesYYYZZ	Cabinet loop detector is located in
LOOPNAME	18 char	XXXesYYYZZ : AABCDEF	Descriptive name of loop detector, begins with WSDOTCAB, and is followed by additional loop descriptor information after the colon. (e.g., 005es13969:_MN___1) AA = TYPE descriptor, (_M = mainline, _C = collector distributor (CD), _R = reversible (rev), AM = auxiliary (aux) mainline, MM = metered mainline, MC = metered CD, DM = duplicate mainline) B = Direction (N = northbound, S = southbound, E = eastbound, W = westbound) C=HOV or ramp lane identifier (H = HOV lane, L = left lane of advanced queue detector, R = right lane of advanced queue detector, _= not HOV) D = LOOPTYPE descriptor (X = exit ramp, O = on ramp, A = advanced queue, Q = queue, I = inter Q, D = demand, P = passage, _ = mainline) E = if part of dual loop configuration, S = initial loop in dual loop, _ = not part of dual loop configuration) F = LANE Number
LOOPID	2byte int	3267-8624	Unique ID number assigned in order of addition to LoopsInfo table
DIRECTION	1 char	'N', 'S', 'E', 'W'	Main directionality of route (N=North, S=South, E=East, W=West)
LOOPTYPE	7 char		Loop detector specific location.

		'main', 'speed', 'exit', 'onramp', 'left', 'right', 'queue', 'inter q', 'passage', 'demand'	'main'=mainline, 'speed'=speed detector (part of dual loop), 'exit'=exit ramp, 'onramp'=on ramp, 'left'=left advance (ramp metering), 'right'=right advance (ramp metering), 'queue'=queue (ramp metering), 'inter q'=intermediate queue (ramp metering), 'passage'=passage (ramp metering), 'demand'=demand (ramp metering),
isHOV	bit	0 - 1	Bit indication whether loop detector is on an HOV lane (1=HOV, 0=not HOV)
TYPE	5 char	main', 'aux', 'dup', 'dupre', 'mconn', 'conn', 'meter', 'usrdf', 'revrs',	loop detector general usage. 'main'=mainline (data collection), 'aux'=auxillary (loop is on different route from cabinet), 'dup'=duplicate (loop is installed as a dual loop but not used as dual loop), 'dupre'=duplicate on reversible lanes, 'mconn'=metered connector/distributor, 'conn'=connector/distributor, 'meter'=loop used in ramp metering, 'usrdf'=user defined (not named by computer system, 'revrs'=reversible lanes
LANE	1byte int	0-6	Lane number counting from 1 at the outside lane.

Table 1d: CF2011XX (file: CF2011MM.csv) (correction factors for 'main' and 'speed' type loops)			
LOOPID	2byte int	3759-8104	Unique ID number assigned in order of addition to LoopsInfo table
CF	decimal	0 – 1.162	Correction factor to adjust for loop sensitivity errors. Can be used to multiply scan counts found in LD20Sec_2011XX by. It can also be used to directly correct lane occupancy values. DO NOT APPLY Correction factors to loop data identified as invalid by the WSDOT data quality checks. Data identified as "invalid" by the WSDOT QA/QC process can not be "corrected." A value of 1.0 is for good loops, 0 for loops with type 1, 2, and 5 errors. Some decimal value for error types 3 and 4.
ERRORTYPE	1byte int	0,1,2,3,4,5	Type of error discovered. 0= good loop detector, 1=very under-sensitive, chattering or pulse mode loop detector (volume may be compromised), 2=split distribution loop detector (unsuitable for correction), 3=correctible over/under sensitivity, 4=slight split distribution (would be type 3 error when modeled with 2 distributions instead of 3), 5=oversensitive, cross-talking or sticking detector (unsuitable for correction, can overlap with type 1 error when weakly picking up adjacent lanes).

2.1.2 Quality Assurance Testing

In WSDOT's current urban freeway management system, data are collected from loop detectors by 170 or 2070 traffic controllers operating at 60 Hz. The data collected are vehicle volumes (one vehicle is counted each time a loop turns on and off) and scan counts (the number of times a loop is observed as being "on" during the 1200 times the loop is checked for activity during each 20-second period.) Using these basic data as they are observed every 1/60th of a second, the controller performs several simple data quality checks. The checks look for specific types of errors, including short pulses, loop chatter, values outside of allowable volume/occupancy ranges, and having a flag for operator-disabled loops[5]. As a result of those checks, an initial data quality flag accompanies the 20-second aggregation summary data provided via WSDOT's external data feed. (See the FLAG variable in the WSDOT20Sec_2011XX table shown in Table 1a.) In addition, for some conditions, the loop controllers in the field set scan counts to

[5] For a more detailed description of the error checking done at the 170 controller level, see http://www.its.washington.edu/tdad/loop-flags.pdf or Ishimaru, J. M., and M. E. Hallenbeck. 1999. FLOW Evaluation Design Technical Report., WA.RD 466.2 available at: http://depts.washington.edu/trac/bulkdisk/pdf/466.2.pdf

greater than 1200. It will cause the computed lane occupancy value to be greater than 100 percent. This also is an indication of loop failure. However, the basic WSDOT error checking system does not catch all data errors. In particular, it does not always correctly identify "poorly performing, but functioning" loops. Consequently, a secondary data quality process was applied to the 20-second data to examine—and where possible recalibrate—the output from the detectors, as well as identify data that are considered "good" by the WSDOT data feed but that are actually invalid. The result of that process is a second data quality flag, applied to each detector, for each month of data. (That is, one flag is applied for all data at that location for each month.) The remainder of this subsection describes that additional quality assurance process. These additional quality assurance tests are only performed on mainline loops.

Loop Data Error Detection can be accomplished in a variety of ways. The simplest are threshold techniques. Threshold-based techniques are used to identify data above a threshold maximum value or below a threshold minimum value. One example is checks on speed, in which speed readings above a threshold value, such as some maximum expected vehicle speed (e.g., 120 mph) are identified, and used as "sanity checks" on the data collected. Another example is minimum occupancy checks, in which a minimum value of lane occupancy per vehicle is required (based on vehicle speed and length assumptions), and records with occupancies lower than the expected minimum value are identified and flagged as being invalid.

Threshold-based techniques have their uses but have the disadvantage of ignoring systematic errors such as over-sensitivity and under-sensitivity of loops. To identify these systematic errors, a better understanding of loop detectors and the data they produce is required. A loop detector detects the metal of a vehicle as the front bumper passes over it and continues to detect the vehicle until the rear bumper clears the loop detector. This means that the on-time for a given vehicle detection is equal to the vehicle's length plus the length of the loop detector coil divided by the vehicle's speed. Therefore, if the vehicle population's average length and speed are known or can be estimated, one can determine what the average loop detector readings should be. If perfect vehicle data were being collected, corroboration would be a trivial task.

However, loop detector data are often aggregated into 20- or 30-second intervals, with total interval volumes and detection on-time expressed as lane occupancy (the number of scan counts divided by the total number of possible scan counts, and expressed as the percentage of that 20- or 30-second interval.) Because vehicle lengths and speeds can vary wildly depending on traffic conditions, location, and time of day, it can thus be difficult to identify errors.

Therefore, an error identification technique is necessary, and the one used for this project was first used by Corey, et al. (2011)[6]. The process involves inputting the data into a database and then querying all of the intervals during which a single vehicle crossed a given loop detector with no vehicles in the preceding or following intervals. With no vehicles in the preceding or following intervals, one can be sure that the

---

[6] Corey, J, Y Lao, YJ Wu, Y Wang. Detection and Correction of Inductive Loop Detector Sensitivity Errors Using Gaussian Mixture Models. Transportation Research Board Annual Meeting. 2011. Paper #11-2829

occupancy of the selected interval will be equal to the on-time of that vehicle alone. This allows for the collection of pseudo detector on-time data. The majority of these data come from late night time periods, when free flow conditions are expected and speeds are relatively constant, limiting the impacts of speed distribution.

The pseudo on-time data are then fed into a Gaussian Mixture Model (GMM) with three distributions. The GMM determines three parameters for each distribution: weight, mean, and variance. The weight indicates the percentage of the data belonging to that distribution, while the mean and variance describe the position and shape of the normal distribution's fit to the data.

Given an average short vehicle length of 15.2 feet (determined from other research), a 6-foot loop diameter, a measured free flow speed near 65 mph, and low truck percentages, the majority of the data should cluster near 222 ms. Allowing for a 10 percent error tolerance gives a good data window of 200 to 245 ms.

The initial work with this technique identified three error types based on the distribution of lane occupancy values, the vehicle population, and speed characteristics.

- A Type 1 error results from under-sensitive loop detectors detecting only portions of vehicles. A Type 1 error is characterized by a very narrow primary distribution representing more than 70 percent of the data centered at approximately 160 ms with a very low variance.
- A Type 2 error is characterized by split distributions. A Type 2 error typically manifests in splits ranging from a 65-25-10 to 45-45-10, with the two largest distribution means separated by more than 30 ms. Typically, one distribution will be in the "good" zone and one will be either low (160-180 ms) or high (~300 ms). The cause of Type 2 error is still unknown, but the split distributions preclude easy correction.
- a Type 3 error is defined by a large primary distribution of more than 70 percent of the data, and a primary distribution mean of less than 200 ms or greater than 245 ms. There is a mathematical lower bound on Type 3 error of 160 ms, which is discussed below.

The current data cleaning exercise identified two additional error types that were not encountered during the original research. The first is an artifact of the methodology. A Type 4 error occurs because three distributions were chosen for the modeling. If two distributions were chosen for the GMM data fitting, a loop detector with Type 4 data would have one distribution fit to the short vehicle data and one fit to the truck data. However, because three distributions were used, there was an extra distribution for the system to handle, and two closely spaced distributions divided what would be a single distribution under a two-distribution GMM fit. To qualify as a Type 4 error, the sum of the primary and secondary distributions must be more than 75 percent of the data, the means must be less than 10 ms apart, and the average mean must conform to the same conditions as Type 3 errors. A Type 5 error is over-sensitivity. When the mean for the tertiary distribution is greater than 1200 ms, the loop detector suffers from Type 5 error. A heavy truck with a sleeper cab and full trailer will be approximately 75 feet long and at 65 mph will have an expected on-time of 850 ms. For the distribution mean to average 1200 ms or more requires over-sensitivity or cross talk. Loop detectors producing Type 5 errors may also qualify for Type 1 error because of partial detections of vehicles in adjacent lanes.

For the cleaned data in the Seattle data set, each loop detector is identified as suffering from a single error type. The order of precedence is Type 5>Type 1>Type 2>Type 3> Type 4 > Type 0 (no error). Thus, while a loop detector may suffer from Type 1 error, if it also qualifies for Type 5 error, it will be reported as suffering from Type 5 error.

The error detection process was automated for this project, which led to a few issues. The first was the development of Type 4 error, which stemmed from always using three distributions. This choice to use three distributions was based on two factors. First, the original research found that fewer than three distributions would not reliably identify truck data, and the result would be a skewed distribution with very, very large variance. Second, identification of loop detectors suffering from Type 2 error was important, and this was also best done with three distributions because a three-way split in the data was possible.

The second major issue resulting from automation involved the GMM fitting process. Because the pseudo on-time data were derived from 20-second interval occupancy data reported to the tenth of a percent, all of the pseudo on-time values were multiples of 20 ms. Particularly for Type 1 errors, this introduced an error during the fitting process when the variance would drop to zero because all of the values in that cluster for that step had exactly the same value, i.e., a distribution would fit to a data set completely composed of the value 180. This would cause the fitting process to error out because it could no longer cluster and re-cluster the data to improve the distribution fits. This problem was solved by randomly adding or subtracting 1 to 5 ms from each value, guaranteeing that there would always be some variance for the fitting process to use. Several fits were checked for impact, and no appreciable difference was noted.

<u>Development of Error Correction Factors.</u> This is possible only for Type 3 and Type 4 errors. Valid data do not need correction, and data suffering from types 1, 2, or 5 errors cannot be corrected by the applied method.

Loops experiencing Type 1 errors have generally lost the ability to distinguish cars from trucks and may also produce suspect volumes in extreme cases. Type 2 errors cannot be corrected by the proposed method because a shift in the data to make one distribution more correct will result in the other becoming more incorrect. Data from loops experiencing Type 5 errors are suspect because it is probable that a loop detector suffering from a Type 5 error is detecting vehicles from more than one lane. Loops identified as suffering from types 1, 2, and 5 errors are identified in the CF2011XX table for 20-second data, but no correction factor is supplied. Users of the data set can choose to use or avoid using these data, depending on their specific analysis interests.

For loops with Type 3 and Type 4 errors, the methodology is very simple in concept. The correction process changes the mathematical size of the loop detector. If the loop detector is over-sensitive, increasing the size of the loop detector from 6 feet to 6+2d feet will compensate for the over-sensitivity. Similarly, for under-sensitive loop detectors, decreasing the size of the loop detector used in calculations will compensate for the under-sensitivity. Note that this is what necessitates the lower limit on Type 3 and

Type 4 error correction. The loop detector is constrained to stay positive in length, so 6+2d = 0 gives a lower limit of d as -3 feet. For the correction equation, d is calculated as

$$d=1/2(\mu_1 * \upsilon_f - L_V - L_L)$$

where:

μ_1 is the primary distribution mean

υ_f is the free flow speed in feet per second (which is assumed to be 65 mph or 95.33 fps)

L_V is the average short vehicle length (15.2 feet)

L_L is the loop detector coil diameter (6 feet).

Once the new (corrected) loop detector length is known, it is possible to correct the scan count (used to compute occupancy percentage) reported by the loop detector. The new correction factor is the ratio of L_V+L_L to L_V+L_L+2d. The originally computed lane occupancy is multiplied by $(L_V+L_L)/(L_V+L_L+2d)$ to correct for sensitivity errors. Note that the correction factor is based on the average short vehicle length and will be incorrect for longer vehicles that are proportionately less affected by the sensitivity errors.

2.1.3 Meta Data

Figure 3 shows the entity relationship (ER) diagram for the 20-second data as they are included in the four tables provided to FHWA in csv format.

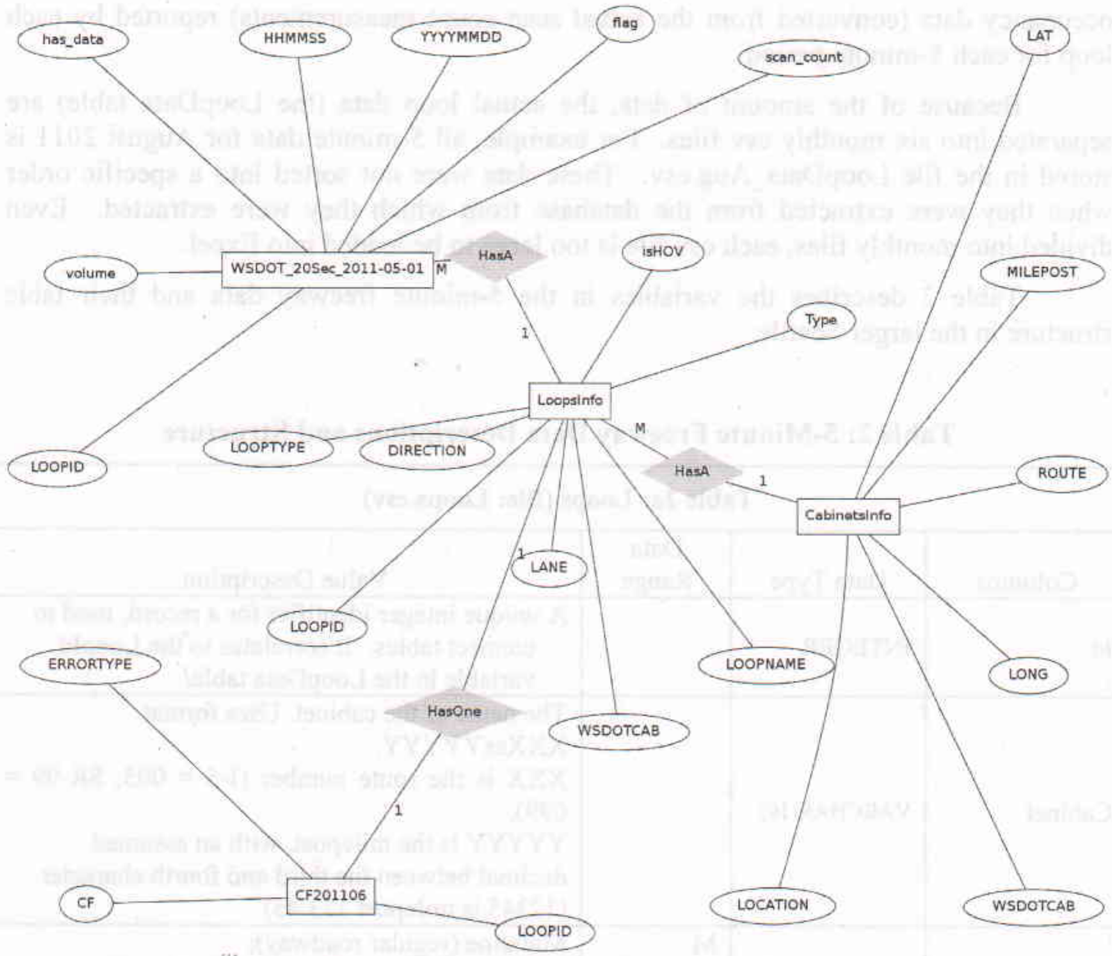

Figure 3[iii]: ER Diagram for WSDOT 20-Second Freeway Loop Data

2.2 5-Minute Freeway Loop Data

The 5-minute aggregation of the freeway loop data is the primary data set WSDOT uses to calculate freeway performance measures. This level of aggregation provides sufficient detail to identify the onset of congestion while limiting the amount of data handling required to develop those performance measures. WSDOT aggregates 5-minute data at its traffic management center and reports them independently from the 20-second data described above.

2.2.1 Data Description

The 5-minute data are organized in three tables; Cabinets, Loops, and LoopData. Each of the tables is provided as a separate csv file. The "Cabinet" table (csv file) describes the location of the data collection cabinets, which is the closest geographic descriptor of the physical location of the loops maintained by WSDOT. The "Loop" table (csv file) describes the specific type of data being collected by a given loop. (That is, whether the loop is collecting data on a specific general purpose lane (e.g., lane 2), HOV lane, or ramp. The "LoopData" table (csv file) contains the actual volume and occupancy data (converted from the actual scan count measurements) reported by each loop for each 5-minute period.

Because of the amount of data, the actual loop data (the LoopData table) are separated into six monthly csv files. For example, all 5-minute data for August 2011 is stored in the file LoopData_Aug.csv. These data were not sorted into a specific order when they were extracted from the database from which they were extracted. Even divided into monthly files, each csv file is too large to be loaded into Excel.

Table 2 describes the variables in the 5-minute freeway data and their table structure in the larger Seattle.

Table 2: 5-Minute Freeway Data Descriptions and Structure

Table 2a: Loops (file: Loops.csv)			
Columns	Data Type	Data Range	Value Description
Id	INTEGER		A unique integer identifier for a record, used to connect tables. It correlates to the LoopId variable in the LoopData table/
Cabinet	VARCHAR(10)		The name of the cabinet. Uses format XXXesYYYYY. XXX is the route number (I-5 = 005, SR-99 = 099). YYYYY is the milepost, with an assumed decimal between the third and fourth character (12345 is milepost 123.45)
RoadwayType	VARCHAR(2)	M C R AM AR AC DM DR DC	Mainline (regular roadway); Collector/Distributor (on-ramp/off-ramp); Reversible (see http://en.wikipedia.org/wiki/Reversible_lane). Optional "A" prefix indicates an auxiliary roadway being served (most common in cases of on and off ramps, where a single cabinet serves both ends of the ramp and thus has loops on two distinct roadways). "D" Also optional, D is a mutually exclusive prefix that indicates a "duplicate" loop.
Direction	CHAR(1)	N S W E	(N)orth; (S)outh; (E)ast; (W)est

LaneType	CHAR(1)	M P D I Q 2 L R O X	(M)ainline; (P)assage; (D)emand (stop bar); (I)ntermediate Queue; (Q)ueue; (2)nd queue (L)eft Advance Queue; (R)ight Advance Queue; (O)n-ramp; Off-ramp (X).
IsHOV	BIT(1)	0 - 1	Bit indication whether loop detector is on an HOV lane (1=HOV, 0=not HOV)
LoopType	VARCHAR(5)	Main Speed	Main = The primary loop in a dual loop installation, or the only loop at a single loop location Speed = is the second loop at a dual loop location, but in this data set, it reports data similar to a conventional loop (if all loops worked and were calibrated perfectly, and no vehicles changed lanes over the detectors, the speed loop data would be exactly the same as main loop data for that same lane and location.)
LaneNumber	SMALLINT	1-9	Lane 1 is the right-most lane when facing the direction of travel, lanes 2-9 (if present) are counted from that right-most lane

Table 2b: LoopData (file: LoopData.csv)

LoopId	INTEGER		A unique integer identifier for a record, used to connect tables. It correlates to the "Id" variable in the "Loops" table.
Date	DATE		The date of the record
Time	TIME		The local time of the record denoting the beginning of the five minute time period
Flag	SMALLINT	0 1 2 3 4 5 6 7 8 9	0: bad data; 1: good data; 2: suspect data; 3: bad data (cabinet turned off); 4: not used for regular loop data; 5: bad data replaced with "good" data; 6: bad data replaced with "suspect" data; 7: bad data with non-bad indicator manually fixed; 8: no data file for day (file missing); 9: loop did not exist that day (not in file)

NumReadings	BIT(4)	0 - 15	The number of 20-second readings incorporated into this 5 minute record (15 is ideal, less than 15 almost always indicates volume data are unusable unless adjusted to account for missing intervals (Volume adjustments are not part of this data set.)
Volume	SMALLINT		Number of cars which passed over the sensor in the time period
OccupancyX10	SMALLINT		Percentage of time a vehicle was physically over the sensor, stored in tenths as units (e.g., 263 = 26.3%)

Table 2c: Cabinets (file: Cabinets.csv)

Cabinet	VARCHAR(10)	The name of the cabinet. Uses format XXXesYYYYY. XXX is the route number (I-5 = 005, SR-99 = 099). YYYYY is the milepost, with an assumed decimal between the third and fourth character (12345 is milepost 123.45)
Milepost	DECIMAL(5,2)	The milepost the cabinet's sensors are located at
CrossStreet	VARCHAR(17)	The nearest cross street to the cabinet
Road	VARCHAR(6)	The name of the road the cabinet is located on
Lat	DECIMAL(9,6)	Latitude of record
Lon	DECIMAL(9,6)	Longitude of record

2.2.2 Quality Assurance Testing

The quality assurance process applied to the 5-minute aggregation data and described below is the process WSDOT currently[7] uses before developing the freeway performance statistics it publishes in the Grey Notebook[8]. The 5-minute data quality process starts with the placement of a simple good/bad/suspect/missing error flag on each 5-minute value for each loop as a result of the aggregation of the 20-second volume and occupancy (scan count) data output by the controllers in the field. An engineer then reviews the 5-minute data with the help of a series of automated data quality checks. The result of that human review is the ability to override the initial WSDOT data quality flag with a new data quality flag indicating that the data for that loop are inappropriate for use in most analyses. These manually produced flags are associated with each loop on a monthly basis (that is, the manually produced flag overrides an entire month of the initial WSDOT flags).

The new error flag indicates whether the data at that location should be used in automated analyses. Not all data within a given month for a location marked as "invalid" may in fact be "invalid." Similarly, not all data in a month marked "good" are always

[7] The quality assurance process continually changes as we identify new data quality issues. The process described in this report is current as of November 2011.

[8] http://www.wsdot.wa.gov/accountability/

valid. For example, for all of May, there are 275,940 5-min periods with lane occupancies greater than or equal to 100% AND have a bad data flag. The remaining 8,951 (3% of the occurrences) do not have bad data error flags. Some of these are valid data (for example, at stop bars for ramp meters. In other cases, they occur on ramps where the manual quality assurance process has not been applied, and where the WSDOT automated QA process has not identified the bad data. However, in the case of the 5-minute data, the fact that an original WSDOT flag has not been replaced indicates that aggregations of the data at that location (e.g., mean weekday statistics) are considered valid for use in performance analyses. These aggregations will still need to interpolate around any small segments of invalid data indicated by the original WSDOT data quality flags.

<u>Background</u>. Errors in sensor data collected at the roadside and transmitted to a central traffic management center can be caused by various types of hardware failures. These failures can be classified as "permanent" or "temporary" failures. "Permanent" failures occur when hardware ceases to function (or function correctly), and no valid data are collected again until the failed piece of hardware is repaired or replaced. Permanent hardware failures most commonly include field sensor malfunctions, equipment calibration issues, and major communications or power failures. "Temporary" errors occur when a hardware begins malfunctioning (and may or may not alternate between "working" and "not working" conditions) or where external factors such as loss of power or communications or other data transmission problems occur. Construction activities can cause both temporary and permanent data errors. For example, construction can cause 1) data stream interruptions during active construction projects due to removal of communications lines, 2) data gaps during replacement of existing sensor equipment, and 3) inaccurate sensor performance due to the effects of lane restriping. (In this last case, the sensor "continues working correctly," but the data it reports are invalid because the sensor is now located incorrectly in relation to traffic.)

<u>Methodology</u>. Because of the potential for these errors, a review of sensor data quality is essential before any analytical task. The review process used for the 5-minute aggregation of WSDOT traffic sensor data sets includes a detection task (identifying bad or questionable data) and a response task (implementing an appropriate method to handle invalid or missing data). The detection and response tasks use automated tools to help make the process more efficient and interactive online query services to enable easy access to results. At the same time, "gray area" cases with borderline or ambiguous data quality sometimes warrant the judicious use of a manual review. The review approach seeks to strike a balance between efficient, timely processing and the appropriate application of professional judgment. The outcome of the manual review is to change error flags from "good" to "bad," allowing the automated software routines to treat the modified data in the same manner they treat other invalid data points.

The detection of questionable 5-minute data relies on a three-step approach. It consists of the following:

1) use of a series of detection filters to automatically scan the data for patterns of questionable values ,
2) supplementary, visual inspection of automatically produced data summary graphics, combined with a manual review of supplementary notes

3) reconciliation of the information from items 1 and 2. Following reconciliation, data quality "scores" are assigned that can then be used to determine which data can be used.

The data quality detection process analyzes data at the "loop group" level. A loop group is a collection of associated loops that are usually combined for the purpose of computing performance metrics. A common example of a loop group is a collection of sensors from all general-purpose (GP) lanes at a given freeway location, in a given direction of travel, for all weekdays of a given calendar year. Such loop groups are typically used to analyze the overall nature of spot GP performance at a given location or to provide GP speed performance data at a location for the purpose of computing a travel time for a route that includes that location. The weekday average for a calendar year reflects the most commonly used metrics, which are based on yearly weekday performance.

The following description expands on the three-step invalid data detection process:

a) The first set of data quality detection filters consists of simple rules, developed on the basis of experience with known patterns of bad data not recognized by the hardware in the field. These rules were designed to help detect bad locations by looking for unrealistic data values or incomplete data sets. The rules are aggregated metrics that are based on yearly or six-month average results. As a result, they do not detect errors that occur on a single day. These rules involve simple yes/no acceptance thresholds. They are sometimes overlapping, in the sense that a given data quality issue can be detected by more than one of the rules. In addition, any one rule may not be conclusive on its own. The filters are used to detect the following:

- prolonged gaps in the data at specific times of day (e.g., a site consistently misses readings at 5:15 AM)
- unusually high minimum speeds (e.g., 60 MPH, meaning the loop never experiences slowing traffic during the analysis period) or unusually low maximum speeds over a typical 24-hour day (e.g., speeds never reach 60 MPH even at 2:00 AM)
- unusually low maximum 5-minute volumes over a typical 24-hour day (because these data come from a high volume urban roadway); this is a visual comparison
- unusually high minimum frequency of congestion over the course of a typical 24-hour day (For a month period, on road segments not affected by construction, late at night, there should not be congestion. Therefore, if the roadway is frequently congested throughout the day, the detector is assumed to have failed. For this analysis "congestion" is defined as lane occupancy greater than 35 percent.)
- unusually low average peak period vehicle volumes at a location (If the maximum observed volume per lane per hour for weekdays at a general purpose mainline location during a month is less than 120 vehicles for a 5-minute period, the location is considered bad.)

- unusually low monthly average 24-hour vehicle volumes in comparison to neighboring detectors.

b) A second set of metrics focuses on the distribution of available data by lane. The objective is to detect unequal distribution of data across lanes, as well as low sample sizes on a per-lane basis, in an effort to locate data that may produce biased analytical results (because not all lanes at a location are being equally sampled). The lane distribution values are represented by the number of days of "good" data per lane relative to the maximum possible number of days of data. (For the purpose of these metrics, data quality is provisionally assigned for each data point on the basis of a quality flag that is built into the WSDOT data collection system. The flag is assigned a score of Good, Bad, Suspect, or Disabled; the scores are assigned on the basis of the magnitudes of the raw data and whether they are considered to be within a reasonable range.[9])

c) The results of the automated filters described above are supplemented with visual inspection of a single variable (24-hour vehicle volume vs. day of the year) in graphical form. These graphs rely on the consistent day-to-day nature of traffic volumes at a given location that have been observed throughout the Seattle-area region; disruptions to that consistent pattern are often associated with data quality or construction issues. While the visual format (a graph) requires a person to interpret the results manually, the daily volume graphs have been demonstrated to help analysts quickly easily detect potential data quality issues in a way that other methods (including the automated filters) do not. These graphs are especially valuable for detecting temporal changes in data quality (e.g., sensors that gradually or suddenly become less accurate over time). The unique benefit these graphs provide is considered a worthwhile trade for the extra review time required. To partly compensate for that extra review time, the production of the volume graphs is fully automated and can be queried online. The review and decision making are still performed manually.

d) Supplementary notes are also reviewed. These notes typically document the presence of construction, unusual traffic diversions (e.g., for construction), or the occurrence of specific events (e.g., snow or the effects of the Navy's Blue Angel's air show, which caused I-90 road closures and diversions on I-5 for several hours each day from August 4 to 7, 2011) that might be relevant in evaluating data quality, and other comments that might not otherwise be documented. As with the other metrics, these notes are accessible on an Internet-based wiki maintained at the University of Washington.

After the aggregate filters, lane distribution information, and daily volume graphs have been reviewed, each type of input is assigned an overall data quality score on a scale of OK/questionable/bad. The result consists of three scores: one for the aggregate metrics, one for the lane distribution data, and one for the daily volume graph.

[9] Ishimaru, J. M., and M. E. Hallenbeck. 1999. FLOW Evaluation Design Technical Report, for WSDOT. WA.RD #466.2

(Supplementary notes are factored into those scores as well.) Those three scores are then reconciled to produce a single overall data quality score. These scores are assigned for each month of data, with aggregate half-year and yearly scores also generated on the basis of the monthly scores. (The daily volume graphs and supplementary notes are used to provide month-by-month information required for the assignment of scores at the monthly level.) In the event of conflicting scores, a conservative "worst score prevails" approach is taken. The monthly scores can be applied as selective filters to advise analysts about which data should and should not be used for analytical purposes.

While this detection process is generally sufficient to determine data quality for the majority of loop groups, review results may still occasionally be ambiguous. For those instances, further review may extend beyond the location's characteristics in isolation to comparisons with "nearby" (both spatial and temporal) data values. Such reviews rely on the assumption that traffic values generally vary smoothly in both time and space, i.e., a given location's attributes (volume, speed, etc.) do not change suddenly over time, nor do adjacent locations have radically different attributes. There are, of course, exceptions to these assumptions (e.g., construction can cause spot volume to change at a location from one year to the next; two adjacent locations can have differing volumes because of intervening on- and off-ramps). However, in general the assumption of smoothly varying values is commonly observed. With that in mind, ambiguous cases can be compared to adjacent locations in time and space to find similarities or differences that may help confirm the appropriate data quality score. Because this review is done manually, it is also possible for the reviewer to factor into that review known geometric or operational conditions that are likely to create exceptions to the assumption of smoothly changing traffic values.

Table 3 summarizes the sequence of the invalid data detection filters used in evaluating the 5-minute data. After the detection task has been completed, the data quality scores are used to filter out data that are considered invalid and are therefore removed from processing in automated analyses—such as the travel time computations.

2.2.3 Meta Data

Figure 4 shows the ER diagram for the data tables that hold freeway performance data at the 5-minute level of aggregation.

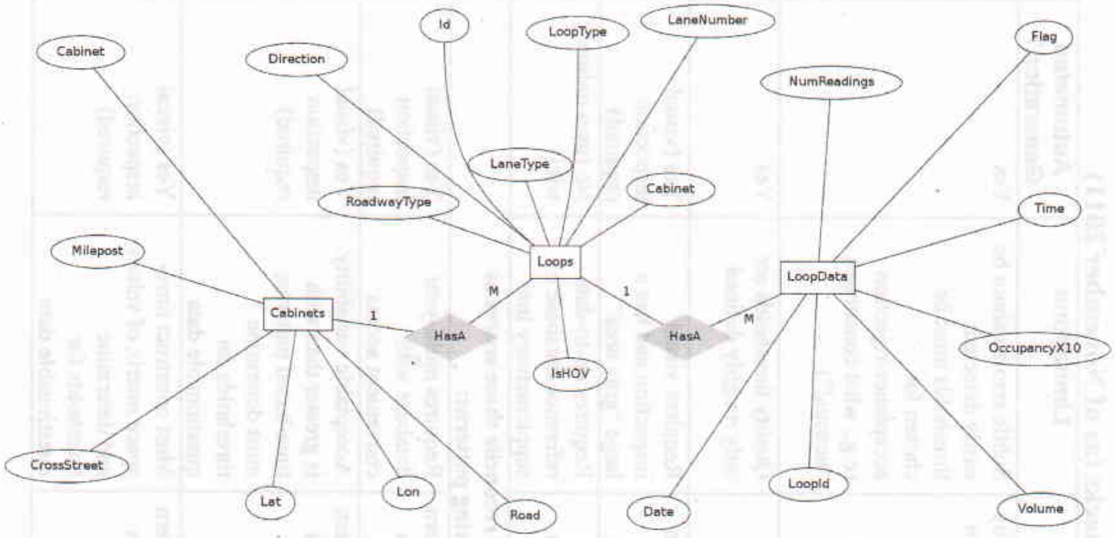

Figure 4[IV]: ER Diagram for WSDOT 5-Minute Traffic Data

Table 3. Data Quality Review Process: Detection Filters and Pattern Checks (as of November 2011)

Filter Type	Description	Objectives	Limitations	Automated Generation?
Aggregate metrics	Use conservative data quality review metrics based on yearly averages: ▪ Gaps in the data ▪ Unusual min or max speed ▪ Unusual max 5-min volume ▪ Unusual min freq. of congestion ▪ Unusually low peak period volume ▪ Unusually low 24 hour volume	Identify bad locations quickly by looking for unrealistic values or incomplete data sets	Subtle errors cannot be easily detected; thresholds must be chosen for acceptance/rejection (e.g., what constitutes "unusual")	Yes
Lane distribution	Review how data are distributed, by lane	Identify potential bias from unequal lane distribution; Identify low sample size	Quality thresholds are only roughly defined	Yes
Daily volume patterns	Review daily volume patterns over time	Detect potential data collection issues relatively quickly	Requires visual inspection and has a large "gray area"	Yes (visual inspection required)
Supplementary information	Cross-reference data clues with construction information or other known data quality issues from other sources	Use information other than sensor data to help clarify data quality	Requires up-to-date reference database of supplementary info	No (uses online wiki)
Develop individual scores for each type of detection filter above, and reconcile those scores as necessary. If results are still ambiguous, review the following patterns:				
Temporal patterns	Review year-to-year changes at a fixed location	Identify discontinuities in pattern that suggest data quality issues	Requires multi-year database with convenient access	Yes (visual inspection required)
Spatial patterns	Review changes in traffic data for a given year, as a function of location (milepost) along a corridor	Identify discontinuities in pattern that suggest data quality issues	Acceptable variability is greater than with time-based patterns; must determine thresholds for questionable data	Yes (visual inspection required)
Time-space patterns	Review changes in traffic data as a function of BOTH time (by year) and location along corridor (by milepost)	Identify discontinuities in pattern that suggest data quality issues	Must construct time-space matrix of values and determine thresholds for questionable data	Yes (visual inspection required)

2.3 5 – Minute Corridor Travel Times

Travel times are computed with the 5-minute data that pass the quality assurance tests described in the previous subsection. Travel times are provided for eight different trips:

- northbound from Federal Way to downtown Seattle in the general purpose lanes
- southbound from downtown Seattle to Federal Way in the general purpose lanes
- northbound from downtown Seattle to SR 526 (just south of Everett) in the general purpose lanes
- southbound from SR 526 to downtown Seattle in the general purpose lanes
- northbound from Federal Way to downtown Seattle primarily in HOV lanes
- southbound from downtown Seattle to Federal Way primarily in HOV lanes
- northbound from downtown Seattle to SR 526 (just south of Everett) in HOV lanes when they are present
- southbound from SR 526 to downtown Seattle in HOV lanes when they are present.

The southern trip in the provided travel time computations travels between S. 320th St. in Federal Way (MP 143.64) and University St. (MP 165.83) in downtown Seattle. The northern trip goes between SR 526 in Everett (MP 189.44) and University Street (MP 165.83) in downtown Seattle. The northern trip uses only the conventional mainline lanes, although the Express Lanes are an option for this trip during some parts of the day. (The Express Lanes are a reversible roadway that exists between just south of downtown Seattle (roughly milepost 165.3) and Northgate (roughly milepost 172.8), part of the way to Everett.[10]) While travel times can be computed with Express Lane loops, these trips are not provided as part of this data set.

The HOV trips have almost the same end points as the GP trips but also have several unusual features. HOV trips moving northbound from Federal Way in the south are terminated at milepost 144.60, where the HOV lane ends. (Travelers are assumed to take an early exit into downtown, rather than continuing to the central downtown exit.) The southbound version of that trip starts at MP 143.98 and continues in the HOV lane until MP 163.19, where travelers are assumed to re-enter the GP lanes (as if vehicles were exiting the freeway); the trip thus incorporates GP travel conditions until it ends at milepost 165.83.

[10] The I-5 Express Lanes contain no southbound exit points between Northgate and downtown Seattle. They do contain several additional entry points for southbound travel (NE 103rd, Lake City Way, and NE 42nd St.), which also serve as exit ramps for northbound travel. Downtown ramps (for southbound travel) and entrances (for northbound travel) are located at Mercer, Stewart/Howell, Pike/Pine, and Cherry/Columbia. The Cherry/Columbia and Pike/Pine ramps are reserved for carpools. The Express Lanes do not allow entry to either SR 520 or I-90, the two freeways that provide access to the east side of the metropolitan region.

For the HOV trips going to and coming from the north, the only HOV facility between downtown Seattle and Northgate (roughly milepost 172.8) is in the reversible Express Lanes. Between Northgate and Everett, a conventional, left-side, open access HOV lane is present on I-5 in both directions. Because the Express Lanes are not accessible to one or the other trips (i.e., northbound or southbound) for part of the day, the HOV trips presented in this database for the northern half of the I-5 corridor do not include the Express Lane HOV facility. Instead, the HOV trips assume that travel occurs in the general purpose lanes between mileposts 165.83 and 172.64.

Note that there is also a short segment of southbound HOV lane in the downtown Seattle section of I-5. The southbound HOV lane starts as the ramp from Mercer St. enters the freeway, and then continues to the southern end of the data set. There is no corresponding northbound HOV segment on I-5 in downtown Seattle. The short downtown section of HOV lane is not included in the Everett to Seattle HOV lane travel time computation, as vehicles are assumed to exit in downtown Seattle, and it is highly unlikely that a vehicle will both use this stretch of I-5 and exit in downtown.

Because travel times from this data set are computed from inductance loop detectors and those detectors do not report speed directionally, it is imperative that anyone computing travel times for trips that include the Express Lanes understand the direction of traffic being served at a specific travel time.

On weekdays the Express Lanes operate southbound from 5:00 AM until 11:00 AM. The facility is then closed to clear the roadway right-of-way of stalled vehicles and debris. The roadway then operates northbound from noon until 11:00 PM, at which time it is again closed. On weekends the Express Lanes operate southbound from 7:00 AM to 1:00 PM and then northbound from 2:00 PM to 11:00 PM. The Express Lanes are generally closed late at night on both weekdays and weekends. Express Lane operations can be changed from these hours to accommodate special events. The details of the actual Express Lane operational status on a specific day can be determined from the variable message sign data contained in the "Other WSDOT Variable Message Sign Data" table in this database.

The assumptions underlying the travel time computation process are major drivers of the output statistics. The most important of these assumptions are part of the data quality tasks previously described for the 5-minute data review. Because the travel times are computed with a vehicle trajectory algorithm, the process of identifying and accounting for invalid data collected along the corridor has significant impacts on the computed travel times. Because WSDOT's sensor spacings are generally[11] ½ mile apart, the assumption that the spot speeds measured at detector locations are valid for the distance halfway to the upstream detector and halfway to the next downstream detector introduces only moderate error into the travel time calculation. However, the presence of invalid data—for example, the loss of an entire loop location because of sensor failure—

[11] WSDOT does not use a fixed distance for its sensor spacing. Instead, the sensors are placed to maximize the utility of the data they provide for the ramp metering system. This generally means that sensors are located up- and downstream of ramp terminals. Because I-5 has closely spaced exits and entrances, the resulting distance between sensors averages about one-half mile.

can create cases in which distances are large enough that estimating speeds by interpolating the up- and down-stream speeds can lead to larger error.

For the travel time computation, if only one or two loops have failed at a location, speeds are assumed to be equal to the average speed of the remaining loops at that location. When all loops at a location must be removed from the travel time computation because of invalid data, speeds from the up- and downstream locations are interpolated to cover the missing roadway section. This data replacement process is conducted on a record by record (i.e., time sensitive) basis. The determination of "invalid" data is based on both the results of the aggregated, yearly data quality process described in the previous section (which marks detectors as "invalid" for one entire month at a time) and the record-by-record data quality flags provided by WSDOT, which are based on the output of the data collection hardware. (See the "Flag" variable in the 5-minute aggregation data table named LoopData. The WSDOT supplied flags are values 0 through 3.)

In addition to issues of data quality, other assumptions, adopted after considerable discussion with WSDOT, also affect the computation of travel times. For the provided corridor travel times, vehicle trajectories are computed by using speed estimates computed from the 5-minute volume and occupancy data, not the 20-second data. In addition, two constraints are placed on the speeds computed from volume and occupancy:

- speeds estimated to be slower than 5 mph are set to 5 mph (that is, for the travel time computation, speeds never drop below 5 mph)
- speeds used in the travel time computation are not allowed to exceed the speed limit.

The lower bound was set because at very slow speeds the computation of space mean speed from volume and occupancy becomes reasonably inaccurate. In addition, minor errors in computed speeds have inordinate effects on travel time estimates. (For a half-mile segment, the travel time difference between 3 mph and 2 mph is 5 minutes, while the same 1 mph speed differential observed between 20 mph and 21 mph results in only a 0.07-minute change in travel time.) Finally, saturated flow (stop and go congestion) is rarely constant over an entire half-mile segment (the approximate distance between most WSDOT loop sensors). Therefore, the decision was made to adopt a 5 mph floor on facility speeds.[12] While this assumption has been shown to work quite well for the vast majority of congestion cases, it does under-estimate travel times in extreme weather conditions (snow in Seattle), when vehicles literally cannot move for very long periods.

A 60 mph limit on facility speeds (the speed limit for I-5) was selected so that free flow speeds would be illustrative of the speed at which a law abiding motorist can make the trip. Because travel times under true free flow conditions (LOS A and B) are a function of driver selection more than traffic flow, it was decided that reporting speeds at the speed limit was appropriate. The primary outcome of this decision is that all travel times under free flow conditions are identical. This simply means that motorists can

---

[12] This is a change from the 10 mph speed floor that WSDOT originally adopted in 1997. Comparisons of floating car data showed that the 5 mph speed floor resulted in more accurate prediction of travel times on the region's most congested corridors, primarily westbound on SR 520 across Lake Washington in the afternoon and southbound on I-405 from Bellevue to Tukwila in the afternoon.

travel the length of the corridor without experiencing congestion. They may travel faster than the speed limit, but travel times are set to the speed limit.

Changing these assumptions will lead to the computation of different travel times for identical corridor end point definitions. Users of the database are able to compute their own location-based speed estimates with volume and occupancy data, select their own speed constraints, develop their own trajectory algorithm, and compare the resulting travel time estimates against the "official" values computed under the assumptions adopted by WSDOT.

2.3.1 Data Description

The 5-minute travel time data are supplied in three separate tables. Each table is supplied to FHWA in a separate csv file. Table 4 describes the specific travel time variables in the Seattle database and their table structure.

Table 4: 5-Minute Freeway Travel Time Data Descriptions and Structure

Table 4a: TravelLocations (file: TravelLocations.csv)		
Id	INTEGER	An Integer ID value uniquely identifying a record. This variable can be ignored.
Name	VARCHAR(30)	The name of a location (a data collection point on I-5 referenced by the nearest cross street) which is a starting or ending point for the named trip.
Milepost	DECIMAL(5,2)	The milepost of the location named in the "Name" variable
Lat	DECIMAL(9,6)	Latitude of the location named
Lon	DECIMAL(9,6)	Longitude of the location named
Table 4b: TravelTimes (file: TravelTimes.csv)		
Id	INTEGER	An Integer ID value uniquely identifying a record. This variable can be ignored.
TripId	INTEGER	An integer value indicating the 'Id' value in the 'TravelTrips' table to be used in matching data across tables
Date	DATE	Date of record
Time	TIME	Time of trip start (local time.)
TravelTime	DECIMAL(5,1)	Travel time at date/time (in seconds). -9999 indicates a missing data point.
Table 4c: TravelTrips (file: TravelTrips.csv)		
Id	INTEGER	An Integer ID value uniquely identifying a record. This variable can be ignored.
Start	SMALLINT	The TravelLocations Id of the start point
End	SMALLINT	The TravelLocations Id of the end point
TripLength	DECIMAL(4,2)	Trip length (in miles)
UsesHOV	1 Bit	A Boolean indicator of whether this trip uses HOV lanes (where available). 1 indicates it did, 0 means it did not

2.3.2 Quality Assurance Testing

The primary quality process for the travel time estimates are the procedures described previously in the quality assurance section for WSDOT's 5-minute data. However, a few additional quality checks are performed after the travel times have been computed. These checks include plotting the mean, 80th, and 95th percentile travel times for the corridor by time of day to look for travel times that are not realistic (for example, if travel times never reach free flow conditions, even late at night). These graphics are also compared to graphics computed for the previous year. When these two graphics create significantly different patterns, additional analysis is performed to determine whether those changes in travel times are the result of actual changes in traffic conditions or the result of unnoticed data errors.

These additional checks involve creating time space diagrams that illustrate the estimated speed at each location along the corridor, by time of day. Review of these time space diagrams allows analysts to determine whether changes in performance at specific locations indicate data sensor failure or can be correlated with specific events. For example, extreme weather conditions (snow falling just before the afternoon commute period) can create very large changes in travel time on one single day. That one very slow day can measurably change the mean travel times computed for annual conditions. Manual review of travel time data—either in terms of time/space diagrams or summary statistics—can identify when significant deviations from past trends occur. The mathematical cause of these deviations is then identified and reconciled with known maintenance activities to determine whether the observed differences are caused by events or sensor malfunction.

2.3.3 Meta Data

Figure 5 shows the ER diagram for the tables holding freeway travel time data at the 5-minute level of aggregation.

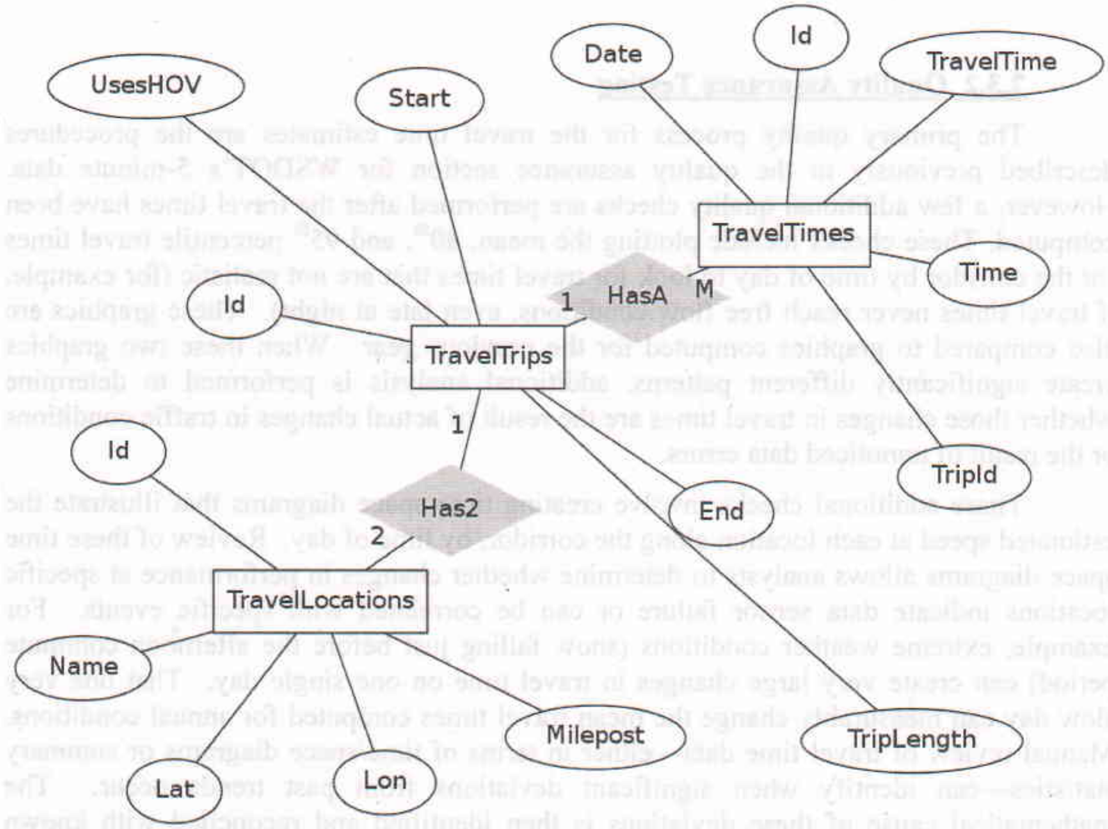

Figure 5[v]: ER Diagram for I-5 Travel Times Based on WSDOT 5-Minute Data

3.0 CITY OF SEATTLE SENSYS DATA

The City of Seattle Department of Transportation (SDOT) has been replacing failing loop detectors at its intersections with Sensys traffic detectors. The data gathered by these sensors are collected by SDOT at 1-minute intervals. The collected data consist of volume, lane occupancy, and vehicle speed.

3.1 Data Description

As with the 5-minute WSDOT freeway data, the Seattle Sensys data are organized in three separate tables. Each table is provided to FHWA as a separate csv file. The organization of the data is similar to the 5-minute freeway data. The table (csv file) "SeattleCabinets" contains data describing the location of specific data collection points. The second table (csv file) "SeattleSensor" provides details about the specific movement being observed by each detector/sensor. Each detector references a cabinet using the "CabinetID" variable. The actual volume data are stored in the "SeattleData" table (csv file), which includes the variable "SensorNum" which is used as a key along with the CabinetID variable to identify a unique data collection sensor.

Table 5 describes the specific variables in the City of Seattle Sensys volume data and their table structure. Traffic data are available at the 23 locations listed below.

Volumes are available for the mainline streets but are often not available for the minor streets crossing at the designated intersections.

Denny Way @ Stewart St. and Yale Ave (three streets that come together in a triangle)
Dexter Ave @ Denny Way
15th Ave NW @ NW 65th St
4th Ave S @ Airport Way S
1st Ave S @ S Hanford St
15th Ave NW @ NW Market St
4th Ave S @ S Lucille St.
The Aurora Bridge (SR 99)
4th Ave S @ S Lander St
Airport Way S @ S Holgate St
1st Ave S @ S Lander St
1st Ave S @ S Holgate St
1st Ave S between S Royal Brougham Way and S Atlantic
1st Ave S @ S Hudson St
4th Ave S @ S Holgate St
4th Ave S @ S Marginal Way
4th Ave S @ S I90 TD
4th Ave S @ S Michigan St
Airport Way S @ S Spokane St
The Fremont Bridge
15th Ave NW @ NW 85th St
1st Ave S @ Denny Way
15th Ave NW @ NW 80th St

Table 5: City of Seattle Sensys Data Descriptions and Structure

Table 5a: SeattleData (file: SeattleData.csv)			
Columns	Data Type	Data Range	Value Description
CabinetID	2byte int	1-41	Unique cabinet ID number (note that intersection IDs are not sequential, as the City will be adding additional intersections later for which data are not currently available.)
SensorNum	2byte int	1-24	Sequential ID number assigned to specific detectors at a location
Occupancy	decimal	0.0-100.0, -1.0	Percentage occupancy to hundredths of a percent. -1 indicates error.
Volume	2byte int	0-140, -1	Volume count. -1 indicates error.
Speed	decimal	0.0-99.0, -1.0	Speed in miles per hour, -1.0 used with zero volume and partial detections

Flag	2byte int	0,1,2	Flags provided by system to indicate bad data. 0=good, 1=volume/occupancy/speed disagreement, 2=bad reading with occupancy set to -1.0 and volume set to -1.
HHMMSS	4byte int	0-240000	24 hour (local) time in integer format, no leading zeroes (5pm = 170000, 12am =0, 4:35:22pm = 163522) The time is the start of the data collection interval.
YYYYMMDD	4byte int	20110501-20111031	Date in year month day format (May, 6, 2011 = 20110506)
GoodBit	1 bit	0 – 1	1 = good data (0 means remove these data, generally due to a bad speed measurement
Table 5b: SeattleSensor (file: SeattleSensor.csv)			
CabinetID	2byte int	1-41	Unique intersection ID number
SensorNum	2byte int	1-24	Sequential ID number assigned to detector at intersection
SensorName	50 char	Text	Text descriptor – for system detectors, the text descriptor is presented in the form of two numbers separated by an underscore. For these cases, the format is X_Y where X = direction of travel and 1 (N), 3 (E), 5 (S) or 7 (W); and Y = the lane number. Numerically described detectors are system detectors 50 feet from the stop bar. For non-system detectors, the text uses abbreviations to describe direction of travel and lane number.
Dir	VARCHAR(2)	N NE E SE S SW W NW	Direction of Travel being observed
DetType	VARCHAR(3)	SB SYS ADV MID LT	SB = Stop bar SYS = System detector ADV = Advanced detector MID = midblock detector LT = Left turn
Lane		1 - 4	Lane 1 = inside (fast) lane, with increasing numbers moving towards the curb lane

Table 5c: SeattleCabinets (file: SeattleCabinets.csv)			
Description	50 char	Text	Text description of intersection (typically cross streets)
CabinetID	2byte int	1-41	Unique intersection ID number
Lat	Decimal	~46N	Latitude
Long	Decimal	~122W	Longitude

3.2 Quality Assurance Testing

The only quality assurance testing performed on the Sensys data occurred within the data collection hardware and software. Error flags produced by the data collection hardware and software are indicated in the "Flag" variable in the SeattleData table.

3.3 Meta Data

Figure 6 shows the ER diagram for the table structure that allows efficient retrieval and use of the csv files that contain the City of Seattle Sensys traffic volume, speed, and lane occupancy data.

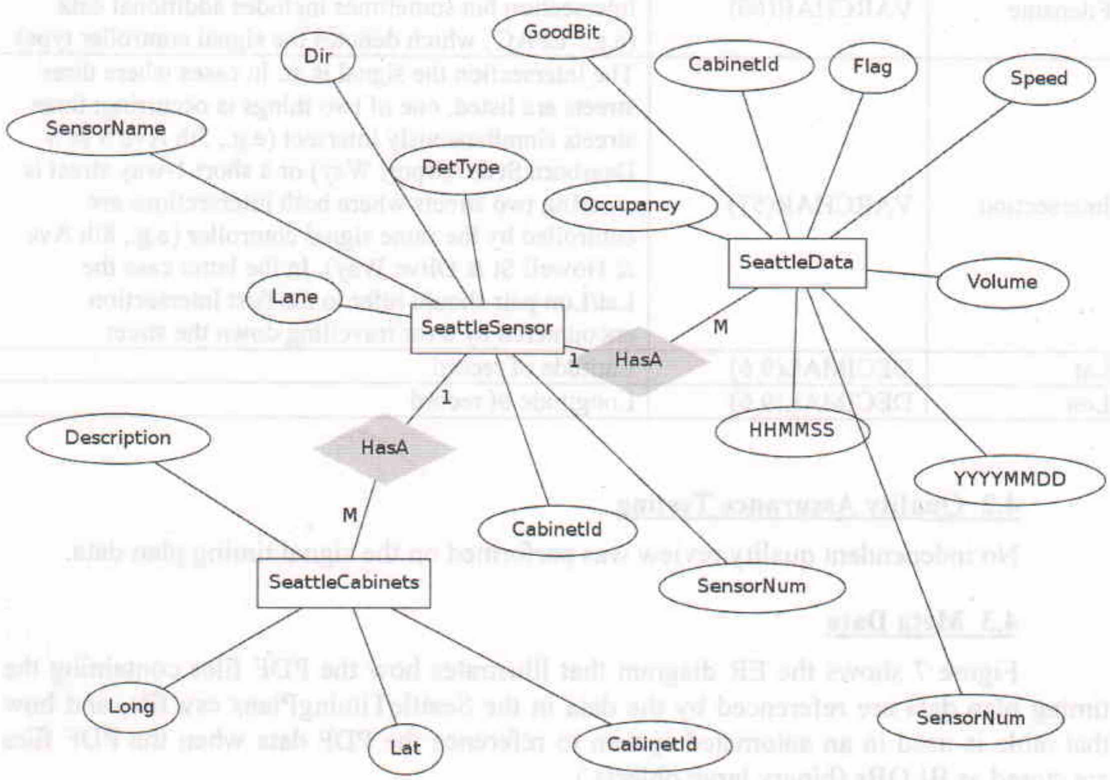

Figure 6[vi]: ER Diagram for City of Seattle Sensys Volume Data

4.0 CITY OF SEATTLE TIMING PLANS

The City of Seattle has provided signal timing plan data (not actual signal operations, but the base timing plans) for roughly 340 signals. Because of the complexity of the signal timing data and the format in which they could be provided to the project team, these data are not converted into detailed records. Instead, timing plans are provided as simple PDF files that researchers can use as desired. Separate PDF files are provided for each signalized intersection.

4.1 Data Description

The data are delivered to FHWA as a single zipped file. When unzipped, a csv file is present (SeattleTimingPlans.csv) that contains the data shown in Table 6 below. In addition a folder is created that contains a series of PDF files. One PDF file is contained in that folder for each intersection. The data in the csv file describe the location of each intersection for which timing plan data are available and the PDF file name that identifies that file.

Table 6: City of Seattle Timing Plan Data Descriptions and Structure

Table: SeattleTimingPlans (file: SeattleTimingPlans.csv)		
Filename	VARCHAR(60)	The original name of the file. This is almost always the intersection but sometimes includes additional data (e.g., 'EPAC', which denotes the signal controller type)
Intersection	VARCHAR(57)	The intersection the signal is at. In cases where three streets are listed, one of two things is occurring: three streets simultaneously intersect (e.g., 5th Ave S & S Dearborn St & Airport Way) or a short 1-way street is crossing two streets where both intersections are controlled by the same signal controller (e.g., 8th Ave & Howell St & Olive Way). In the latter case the Lat/Lon pair should refer to the first intersection encountered by a car travelling down the street
Lat	DECIMAL(9,6)	Latitude of record
Lon	DECIMAL(9,6)	Longitude of record

4.2 Quality Assurance Testing

No independent quality review was performed on the signal timing plan data.

4.3 Meta Data

Figure 7 shows the ER diagram that illustrates how the PDF files containing the timing plan data are referenced by the data in the SeattleTimingPlans csv file, and how that table is used in an automated system to reference the PDF data when the PDF files are stored as BLOBs (binary large objects.)

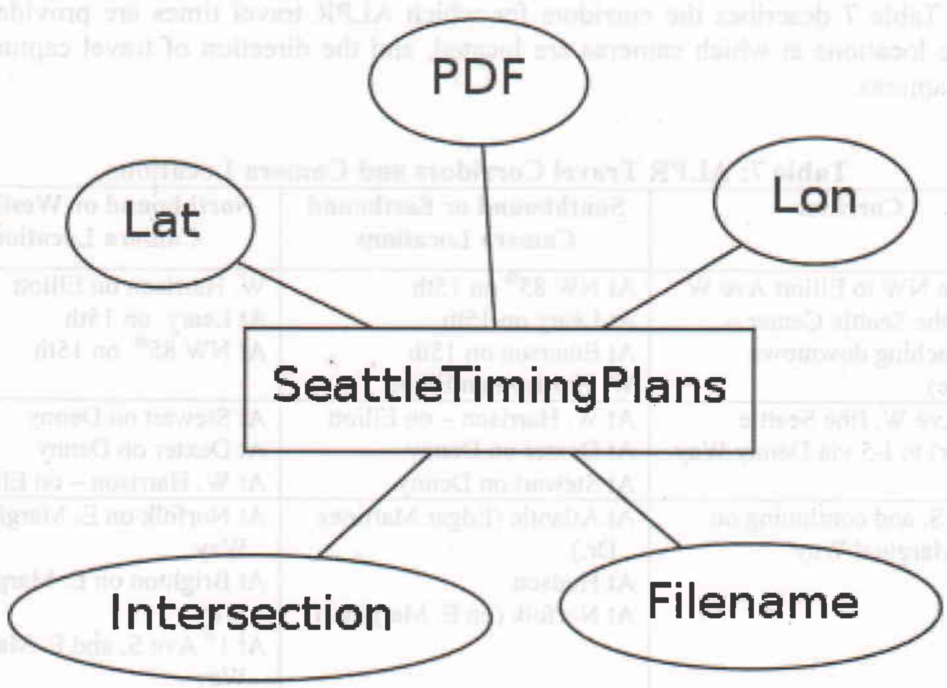

Figure 7[vii]: ER Diagram for City of Seattle Signal Timing Plans

5.0 ARTERIAL TRAVEL TIMES FROM LICENSE PLATE READERS

The City of Seattle and WSDOT have worked together to place a number of automatic license plate readers (ALPRs) at intersections around the city and on some state routes, including SR 522, the arterial that passes around the north end of Lake Washington and serves as one of the diversion routes for the SR 520 and I-90 floating bridges. Matching license plate reads from ALPRs at different intersections allow direct computation of travel times from one intersection to another by subtracting the time of passage at the upstream location from the time of passage at the downstream location.

The data set contains travel times on 96 distinct, directional travel segments. In many, but not all cases, travel is measured in both directions on a specific facility. For example, one travel time segment is on SR 99 from S. Lander St. to Ward St., while a second segment is from Ward St. to S. Lander St on SR 99. These directional movements are two of the 96 distinct segments reported.

In addition, many of the computed segments are partial pieces of other roadway segments. For example, ALPR cameras collect license plates in the westbound direction at eight locations on SR 522. (There are only five eastbound cameras on SR 522.) The eastern-most camera is located at Woodinville Drive. Seven different westbound travel times are computed, with Woodinville Drive as the starting point, one from each of the other seven westbound cameras (e.g., from Woodinville Drive to 68th Ave NE, and from Woodinville Drive to NE 170th St). In addition, travel times are provided between many, but not all, of the mid-point locations (e.g., from 68th Ave NE to NE 170th St along SR 522). While not all sub segments are reported, data are provided on the majority of sub-segments. These are the data provided by WSDOT and the City of Seattle.

Table 7 describes the corridors for which ALPR travel times are provided, the specific locations at which cameras are located, and the direction of travel captured by those cameras.

Table 7: ALPR Travel Corridors and Camera Locations

Corridor	Southbound or Eastbound Camera Locations	Northbound or Westbound Camera Locations
15th Ave NW to Elliott Ave W (near the Seattle Center – approaching downtown Seattle)	At NW 85th on 15th At Leary on 15th At Emerson on 15th W. Harrison on Elliott	W. Harrison on Elliott At Leary on 15th At NW 85th on 15th
Elliot Ave W. (the Seattle Center) to I-5 via Denny Way	At W. Harrison – on Elliott At Dexter on Denny At Stewart on Denny	At Stewart on Denny At Dexter on Denny At W. Harrison – on Elliott
1st Ave S. and continuing on East Marginal Way	At Atlantic (Edgar Martinez Dr.) At Hudson At Norfolk (on E. Marginal)	At Norfolk on E. Marginal Way At Brighton on E. Marginal Way At 1st Ave S. and E. Marginal Way At Lucille At Atlantic (Edgar Martinez Dr.)
4th Ave S and continuing on East Marginal Way	At Atlantic (Edgar Martinez Dr.) At Lucille At ramp to Spokane Street Viaduct At Norfolk (on E. Marginal)	At Norfolk on E. Marginal Way At Brighton on E. Marginal Way At ramp to Spokane Street Viaduct At Lucille on 4th At Atlantic (Edgar Martinez Dr)
Airport Way	At Atlantic At Lander At Lucille At Norfolk	At Norfolk At Lucille At Lander At Atlantic
SR 99 (Aurora Ave/ E. Marginal Way)	At NW 145th St At NW 85th St At Ward At Lander At Lucille	At Norfolk At Lucille At Lander At Ward At NW 85th St. At NW 145th St

SR 522	At I-5 NB Exit Ramp At NE 153rd At NE 170th St At Woodinville Dr	At Woodinville Dr At 83rd Pl. NE At 68th Ave NE At 61st Ave NE At SR 104 At NE 170 St At NE 153rd I-5 SB Ramp
West Seattle Freeway Trips SW Admiral Way – then West Seattle Fwy (aka Spokane St) 35th Ave SW – then Fauntleroy Way SW – then West Seattle Fwy (aka Spokane St.) Fauntleroy Way SW then West Seattle Fwy	 At 34th Ave SW At SR 99 At 1st Ave S. exit At Exit to I-5 NB At SW Snoqualmie St At SR 99 At 1st Ave S. exit At Exit to I-5 NB At 38th Ave SW At SR 99 At 1st Ave S. exit At Exit to I-5 NB	
Misc. trips SR 99 @ Lander to Elliott @ Harrison (a NB trip on SR 99 with an exit at Western, then proceeding north to Elliott) Also Elliott @ Harrison to SR 99 @ Lander (SB - the reverse of the above trip) SR 99 @ Ward to Denny @ Dexter West Seattle Fwy (Spokane St) at the 1st Ave Exit to 1st Ave S @ Atlantic West Seattle Fwy (Spokane St) from the Exit to NB SR 99 @ Lander 15th Ave NW at NW 85th to SR 99 at NW 145th St. (not a direct corridor.)		

Users should be aware that unlike the loop and Sensys detectors, ALPR travel times are not always calculated for a given period, even though the ALPR detectors may have been correctly functioning. Travel times are computed only when license plates are read at both segment end points. Therefore, there are many periods for which travel times are not available for a given segment of interest because the system has not successfully matched license plates at both end points.

To ensure the privacy of travelers, actual license plate numbers are removed before the data are released by WSDOT. Consequently, the data set contains no private identifiers. However, each matched travel time for the study corridors is provided as a unique record. The data cleaning process is then used to indicate which of these travel times is likely to have been affected by non-travel activity. (That is, on the basis of fairly simple statistics, it is relatively easy to identify which "observed trips" are highly likely to have been affected by a stop of some kind, such as a stop for groceries.) These trips remain in the database but are flagged accordingly. Please see the quality assurance testing subsection (Subsection 5.2) below for more information on this process.

The data set does not include the actual ALPR license plate reads. So users cannot perform their own plate matching exercise. Neither does the data set provide data on all possible intersection-to-intersection matches (even though it may be possible to obtain matches from, say, one intersection in the far north of the city with an intersection in the far south because the same vehicle happens to pass both locations at some time during a given day). These travel times are not provided for three reasons; 1) it is unclear what route was taken between the two observations; 2) insufficient matches are made between remote points to perform the statistical analysis necessary to identify travel times containing stops; and therefore 3) the resulting travel times are unlikely to represent valid measures of roadway performance. Consequently, individual observed travel times are only provided for ALPR matches along the primary study corridors.

5.1 Data Description

The ALPR data are stored in two related tables. The first table (the file "AlprCamera.csv") describes the locations of specific ALPR cameras. The second ("AlprTravelTimes.csv") provides the specific travel time segment definitions (based on the starting and ending camera locations), the primary compass direction of travel between those cameras, and the actual observed travel times. Because of their size, the AlprTravelTimes csv files have been loaded into two "zip" files for delivery to FHWA, AlprTravelTimes_Aug-Oct.zip and AlprTravelTimes_May-Jul.zip). **Both contain identical copies of** AlprCameras.csv for convenience. Each also contains an AlprTravelTimes file with name strongly resembling the ZIP name.[13]

Note that the distance between two ALPR cameras is not a variable provided in this data set. If users require that information, a close approximation can be obtained by entering the latitude and longitude of the two cameras into one of the on-line mapping systems (e.g., Bing maps™, GoogleMaps™, Mapquest™). The decision was made to exclude distance information because we do not know the exact field of view of the individual ALPRs. That is, we do not know exactly how far up- or downstream of the reference intersections they are focused. The ALPR cameras—and their output data—are NOT used for law enforcement but only for performance monitoring and traveler information.

Table 8 describes the specific variables in the ALPR-based travel time data and the table structure of those data.

[13] The Aug-Oct file is 109MB and the May-Jul is 95MB. The 109 MB file size slightly exceeds the target maximum file size of 100 MB, but this small overage allowed us to eliminate the need for one more zip file.

Table 8: ALPR Data Descriptions and Structure

Table: AlprCameras (file: AlprCameras.csv)

Column Name	Data Type	Description
CameraID	INTEGER	Unique identifier for a camera
Name	VARCHAR(10)	The name of the camera
Lat	DECIMAL(9,6)	The latitude of the camera
Lon	DECIMAL(9,6)	The longitude of the camera
PrimaryStreet	VARCHAR(30)	The primary roadway the camera is located on
SecondaryStreet	VARCHAR(30)	The secondary roadway (cross street) on which the camera is located

Table: AlprTravelTimes (file: AlprTravelTimes.csv)

Column Name	Data Type	Data Range	Description
ID	INTEGER	0 – 1,000,000	A unique identifier of trips produced by the system that removes personal identifiers. This variable can be ignored.
BeginCamera	INTEGER		The ID of the first (most upstream) camera involved in the travel time calculation
EndCamera	INTEGER		The ID of the last (most downstream) camera involved in the travel time calculation
Date	DATE		Date of record
Time	TIME		Computed local start time of a specific trip ("End Time" is provided with "Travel Time." This is converted to work similarly as the TRAC Travel Times data)
TravelTime	SMALLINT		Travel time at date/time (in seconds)
Direction	CHAR(1)	N S E W	Direction of travel. Please note that this was stored redundantly in AlprTravelTimes table because it simplified the database structure for those users who prefer to work inside applications that favor two dimensions (e.g., SPSS, Excel).
FailedPass	SMALLINT	0 1 2	Which pass (if any) the data point failed the data quality smoothing check. 0 indicates good data which never failed a pass. = first pass, 2 = second pass

5.2 Quality Assurance Testing

This database starts with previously computed individual travel times supplied by WSDOT as an output of its ALPR system. No error checking is performed on the actual ALPR reads. All matched ALPR reads are assumed to be valid observations. The quality assurance process only examines whether the reported travel times are likely to be valid measures of the actual travel time between the two selected ALPR locations.

Erroneous travel times are identified by filtering unnaturally high travel times from all matched license plates by using a "rolling standard deviation" algorithm initially developed and tested by Kevin Mizuta of WSDOT as part of his graduate degree program at the University of Washington.[14]

The testing performed by Mr. Mizuta indicated that license plate read accuracy is extremely good. However, even when reading errors occur, they do not create erroneous travel time estimates, as the likelihood of a reported plate number from an erroneous plate read actually matching the value for a correctly read plate is extremely small. Instead, when plate reading errors occur, plate matches simply do not occur.

Therefore, the primary errors in the ALPR travel time data collection process occur when vehicles either take a circuitous route between two camera locations or when a vehicle stops at a destination located between two ALPR locations. In both cases, the ALPR correctly matches the observed plates, but the computed travel time does not represent actual point-to-point travel on the corridor. These "false travel times" can be identified by the fact that they are "extreme" in relation to other observed travel times with similar start times through the corridor.

By computing a rolling mean and standard deviation, analysts can identify individual outlier travel times that do not reflect corridor conditions. The algorithm is as follows:

- For each record the previous 10 records are retrieved and the mean (X) and standard deviation (σ) for those observations are calculated.
- The current travel time is then compared with the range X +/- σ.
- If the current travel time is within the bounds (one standard deviation), it is considered valid.
- If the current travel time is outside of the bounds, it is considered "invalid."
- However, even if the data are considered "invalid," they are included in the next iteration of the rolling average. This prevents excessive data removal from occurring.
- Once initially invalid data have been removed, the entire process is run a second time. The first pass tends to capture very large outliers, while the second pass captures smaller errors (e.g., stops for coffee or gas).

One additional rule is applied to this algorithm: "The first ten records are never marked as bad." Note that this algorithm runs continuously. It does not "reset" for each new day. Figures 8 through 11 illustrate the application of the algorithm to one of the SR 522 roadway sections. These graphics show the initial ALPR matches as specific data points (red data points), the outcome from the first pass of the QA algorithm through the data (the black line shows the rolling mean), the results of the second pass through the data, and the rolling mean travel time after the second pass has been completed. The variability exhibited in the rolling mean is an excellent illustration of the variability commonly found on arterials because of the combined effects of both traffic volume and signal timing.

---

14 Mizuta, Kevin, Automated License Plate Readers Applied to Real-Time Arterial Performance: A Feasibility Study, report for WSDOT, June 2007

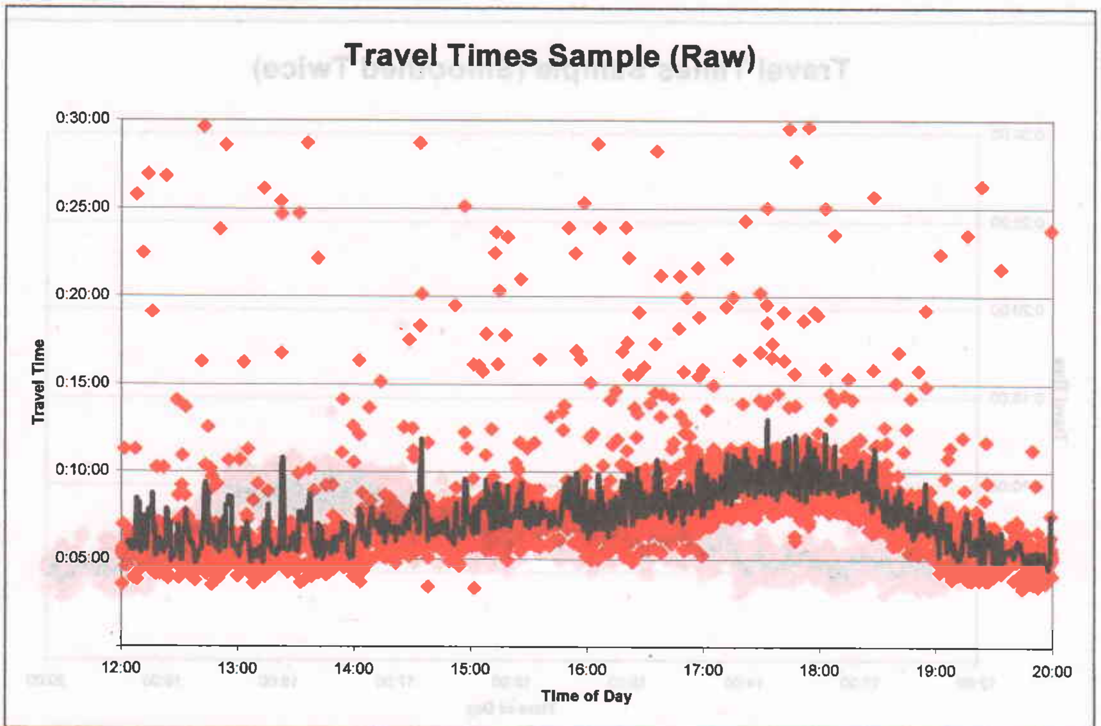

Figure 8[viii]: Initial (Raw) Travel Time Samples

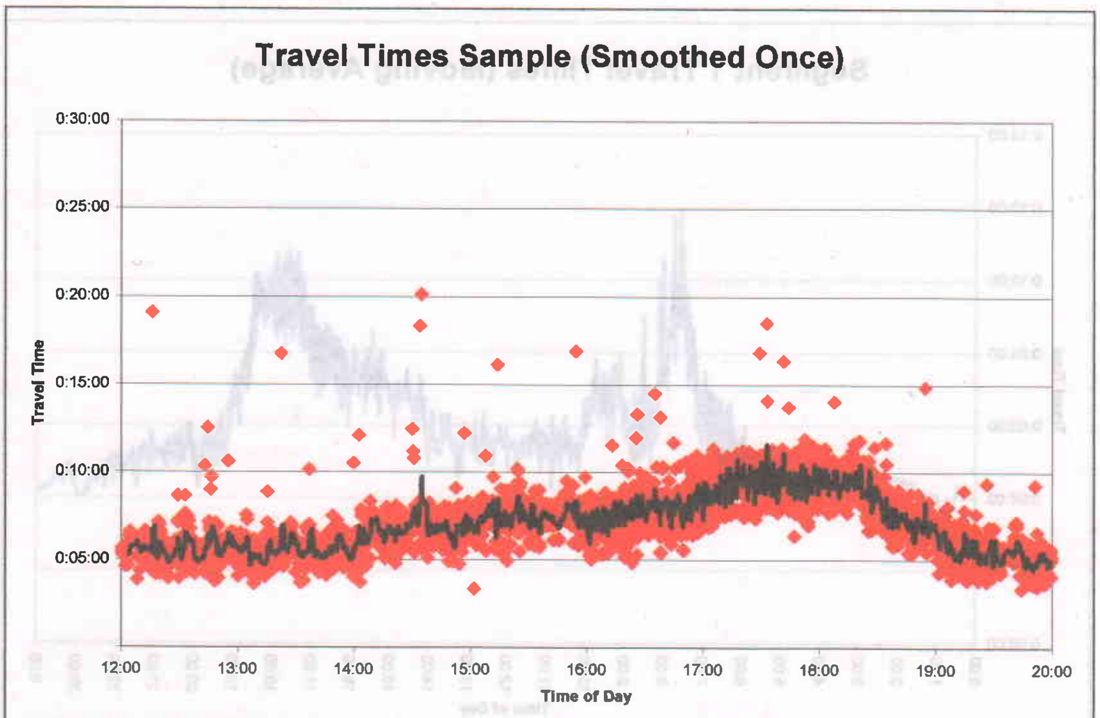

Figure 9[ix]: "Valid" Travel Times after One QA/QC Algorithm Pass

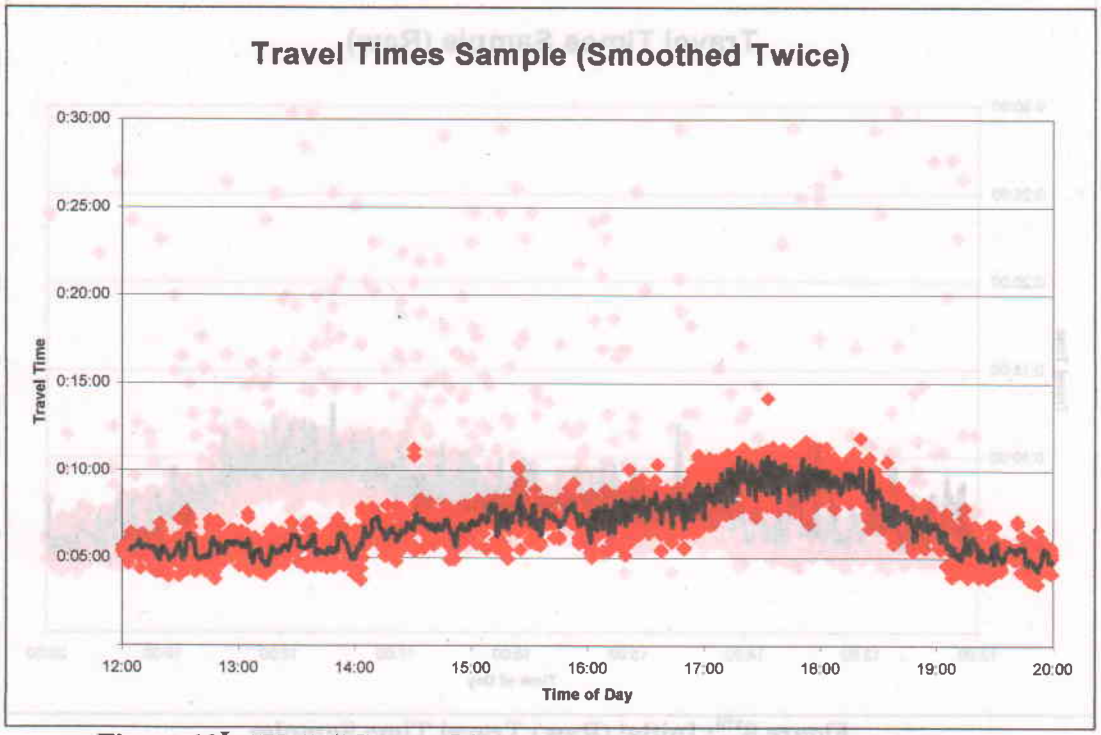

Figure 10[x]: "Valid" Travel Times after Two QA/QC Algorithm Passes

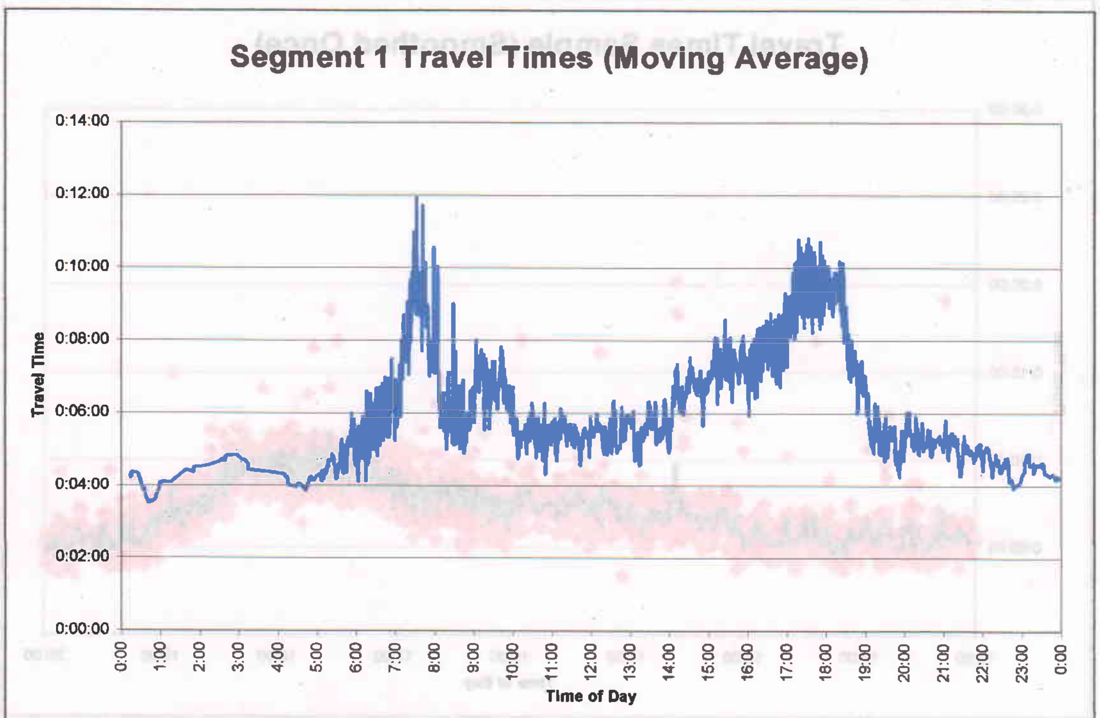

Figure 11[xi]: Moving Average of Travel Times after Two QA/QC Algorithm Passes

FHWA reviewers of the ALPR dataset have noted that a larger percentage of the submitted data are marked as "invalid" as a result of the mean and standard deviation tests described above. Further review of the data indicate that the quality assurance tests are performing as mathematically intended. It is possible that this QA/QC tests does not work as intended on all corridors (its performance has only been thoroughly tested on the SR 522 corridor.) Because of signal delay and poor signal timing (or saturated flow conditions) many arterials can have fairly high travel time standard deviations (hitting two of four traffic signals on a two mile segment may double travel time on that segment.) At the same time, many vehicles using that same arterial may be stopping as part of their travel through the segment. The result on many arterial segments may be that a high percentage of "measured" trips are discarded as a result of that variability. However, it may also be true that these vehicles are actually making stops (some short, some fairly long) within the corridor and therefore do in fact need to be removed. All trips measured on the submitted corridors are included in the database. The users are free to decide for themselves if a different QA/QC algorithm should be used.

5.3 Meta Data

Figure 12 presents the ER diagram which illustrates the relationship of the two ALPR tables submitted to FHWA as csv files.

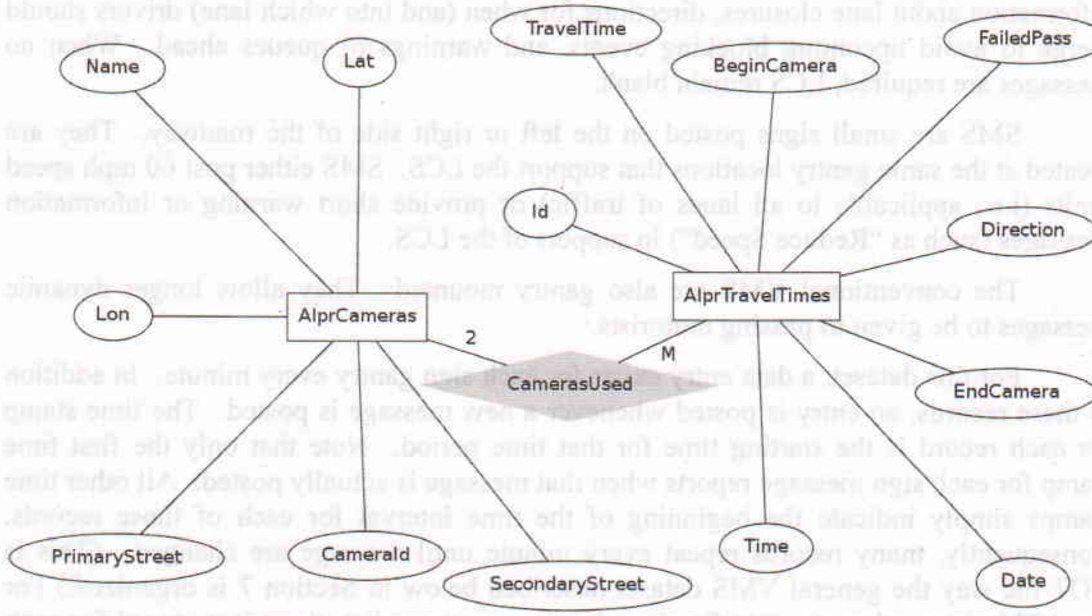

Figure 12[xii]: ER Diagram for Automated License Plate Readers (ALPRs)

6.0 WSDOT ACTIVE TRAFFIC MANAGEMENT SIGNS

The WSDOT is currently conducting a pilot project of active traffic management (ATM) northbound on I-5, south of Seattle. The section of I-5 that contains ATM sign controls starts at roughly milepost 157[15] and ends after the exit to the I-90 interchange but before the southern entrance to the Express (reversible) Lanes. The last ATM sign bridge is located at roughly milepost 164.5. This is just south of the downtown Seattle core exits.

The ATM pilot project currently consists of a series of gantry-mounted, overhead, lane-specific signs that enable adjustable, lane-specific variable speed limits to be posted, as well as the display of selective lane controls (lane closure alerts) and variable length messages that provide notification of incidents and slow traffic conditions downstream. The objectives of this system of dynamic, lane-by-lane traffic management are to improve safety on the most heavily traveled segment of I-5 in the state by reducing congestion-related collisions and to provide modest congestion relief by lowering the number of disruptions to traffic flow. The system has been operating since August 2010 and operated throughout the six-month period covered by the data set.

The ATM communicates to drivers via three kinds of variable messages signs:

- lane control dynamic message signs (LCS)
- side-mounted dynamic message signs (SMS), and
- traditional variable message signs (VMS).

LCS are small signs mounted on gantries, and one LCS sits above each lane of traffic. LCS provide lane-specific, variable speed limit information, as well as information about lane closures, directions for when (and into which lane) drivers should merge to avoid upcoming blocking events, and warnings of queues ahead. When no messages are required, LCS remain blank.

SMS are small signs posted on the left or right side of the roadway. They are located at the same gantry locations that support the LCS. SMS either post 60 mph speed limits (i.e., applicable to all lanes of traffic) or provide short warning or information messages (such as "Reduce Speed") in support of the LCS.

The conventional VMS are also gantry mounted. They allow longer dynamic messages to be given to passing motorists.

For this dataset, a data entry exists for each sign gantry every minute. In addition to these records, an entry is posted whenever a new message is posted. The time stamp for each record is the starting time for that time period. Note that only the first time stamp for each sign message reports when that message is actually posted. All other time stamps simply indicate the beginning of the time interval for each of those records. Consequently, many records repeat every minute until message are changed. (This is NOT the way the general VMS dataset described below in Section 7 is organized.) For the ATM signs, when no specific sign changes are occurring, there is a record for each minute. However if an event occurs which requires a sign change, records occur more frequently than one per minute at the discretion of the system operators.

The data supplied with this data set include all messages posted during the six-month period covered by the Seattle data set.

---

[15] The first ATM sign gantry is located at roughly milepost 157.5. Therefore, vehicles start to be affected by those sign messages around milepost 157.

6.1 Data Description

The ATM data are provided to users in three tables that can be used to create a relational database. Two of these tables provide sensor location and identification information, the third contains the data which describes what the ATM signs actually displayed during the six month data collection period. The table (csv file) "SignGantries" describes the physical location of each gantry that holds the ATM signs. The table (csv file) "Signs" identifies individual signs, indicates what type of sign they are, and includes a simple unique identifier (Id) that is used as the key variable to look up the specific display information for that sign. The actual data displayed on each sign are stored in the table (csv file) "SignDisplays" – which uses a slightly different variable name for this key variable (SignId).

The Signs and SignGantries tables are each provided to FHWA as single csv files. Because the SignDisplays file contains over 25 million rows, it has been split into weekly files. The csv files for this data set are named with the following naming convention: "SignDisplays_WeekBeginning_YYYY-MM-DD.csv" where the year/month/date is the Sunday starting that week. (Sunday is the first day of the week, Saturday is the last day in any given file.)

Table 9 describes the specific variables included in these three tables.

Table 9: WSDOT Active Traffic Management System Actions Data Descriptions and Structure

Table 9a: SignDisplays (file: SignDisplays.csv)			
Column Name	**Data Type**	**Data Range**	**Description**
SignId	INTEGER		The 'Id' value of a record in the 'Signs' table. This variable is used for connecting to the "Signs" table via that table's "Id" variable. (Note that the two tables use a different variable name for this key variable.)
Date	DATE		The date of the record
Time	TIME		The beginning of the time period for the record (provided in local time)
CurrentDisplay	VARCHAR(25)	(multiple)	Current display on the sign ("BLANK" or "blank" indicates a blank sign display.)
DisplayMSGName	VARCHAR(25)	(multiple)	Display Message Name
DisplaySpeed	VARCHAR(2)	<blank> 30 35 40 45 50 55 60	Display Speed (ONLY applies to LCS signs)
ControlMode	VARCHAR(12)	<blank> OVERRIDE LANE CLOSURE TOD ALGORITHM	This is a corridor wide variable that can be: Algorithm, TOD (time of day), Override, Lane Closure
RegulatoryMSGName	VARCHAR(18)	Blank Slow Traffic Ahead Reduced Speed Zone	This is the Message Name used for SMS signs after applying the WSDOT speed sign algorithm
RegulatorySpeed	VARCHAR(2)	<blank> 30 35 40 45 50 55 60	This is display speed in mph for LCS signs after applying the WSDOT speed sign algorithm (can be blank)
SmoothingStatus	VARCHAR(10)	<blank> Override Influenced Smoothed	This indicates the smoothing factor applied. It could be: Smoothed—being in a troupe with adjacent gantry and being smoothed; Influenced—applied WSDOT's standard "3 gantry rule" algorithm for speed reduction; Override—being overridden manually by the operator
LocalSpeed	SMALLINT		Actual local freeway speed that comes from WSDOT's surveillance system

TODSpeed	SMALLINT	60 NULL	Time Of Day speed if it is running on TOD mode
OverrideSpeed	SMALLINT		Override speed manually set by the operator
OverrideDuration	SMALLINT		The duration OverrideSpeed will run
MinSpeed	SMALLINT	30 35 40 NULL	Minimum Speed that can be set on the signs
MaxSpeed	SMALLINT	50 60 NULL	Maximum Speed that can be set on the signs
DefaultSpeed	SMALLINT	60 NULL	Default Speed if there is no local speed
AtRestMSGName	VARCHAR(6)	<blank> AtRest 30 35 40 45 50 55 60	AtRest speed (usually 60 mph) at certain SMS signs if there is no congestion. They serve as regular speed limit signs.

Table 9b: SignGantires (file: SignGantries.csv)

GNName	CHAR(10)		Gantry Name. Used to identify a gantry in both the 'Signs' table and the 'SignGantries' table
Lat	DECIMAL(9,6)		Latitude of record
Lon	DECIMAL(9,6)		Longitude of record

Table 9c: Signs (file: Signs.csv)

Column Name	Data Type	Data Range	Description
Id	INTEGER		The 'Id' value of a record in the 'Signs' table. This variable is for connecting to the SignDisplays table via the SignId variable. (Note that the two tables use a different variable name for this key variable.)

SignName	CHAR(10)		Sign name. Form is XXXYYZZZZZ where XXX is the corridor (005 = I-5) YY is the sign type abbreviation and ZZZZZ is the approximate milepost. Milepost values used in the naming convention are adjusted slightly by WSDOT to avoid duplicate names and to embed positioning data. To determine lane position, arrange the LCS signs from right lane to left lane according to the last 5 digits of the SignName, so that the highest number is located to the left of the lower number. e.g., 005lc16248 is physically left of 005lc16247 is physically left of 005lc16246. The leftmost lane on the road (this includes the HOV lane, if present) has the highest number for that general milepost. There should be exactly as many LCS signs as there are lanes. There is either 0 or 1 VMS signs at a given location. If present, it will be to the right of the LCS signs. There will be either 0 or 2 SMS signs at a given location. Similar to the LCS signs, order these by the last five digits of the sign name. The higher numbered SMS sign will be to the left of the roadway several feet below the leftmost LCS sign. The lower numbered SMS sign will occupy the equivalent position to the right of the roadway. There should be EITHER 1 VMS or 2 SMS signs at a location.
SignType	CHAR(3)	LCS SMS VMS	LCS (Lane Control Dynamic Message Sign), SMS (Side-Mount Dynamic Message sign) VMS (Variable Message Sign)
LaneType	VARCHAR(3)	GP HOV	General purpose High Occupancy Vehicle
GNName	CHAR(10)		Gantry Name
CorridorName	CHAR(10)	005ac15723	Corridor Name

6.2 Quality Assurance Testing

No specific quality assurance tests were applied to the WSDOT ATM data other than simple range and value checks. This process discovered the fact that some data are missing from the file due to holes in the WSDOT ATM sign database. (There are specific chains of events that occur, which force WSDOT to restart its Traffic Management Center computer, and that restart can cause some data collected earlier in the day to be lost.) Consequently, missing time periods in the data set should be interpreted as "missing data" – not as "blank signs."

6.3 Meta Data

Figure 13 shows the ER diagram for the relational tables that contain the WSDOT Active Traffic Management Sign Data.

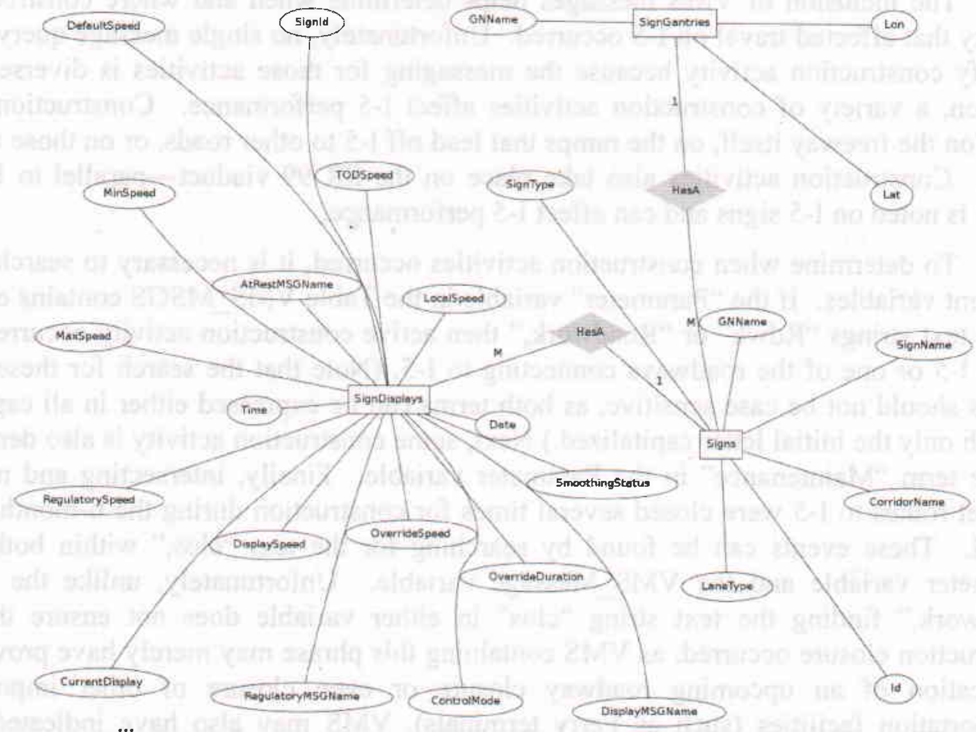

Figure 13[xiii]: ER Diagram for WSDOT Active Traffic Management Sign Data

7.0 OTHER WSDOT VARIABLE MESSAGE SIGN DATA

This data set includes data on all of the information posted on WSDOT dynamic message signs within the I-5 study area boundaries that are not included in the Active Traffic Management data set. There are approximately 33 sign locations, some of which were in use during only part of the 6-month data collection period. This data set indicates when a message was posted on a sign and when that message was either changed or removed. (That is, the basic data are event data. They record the time when the display on each message sign changed.) When a message sign was returned to a blank state, the event is noted as "Blank." However, engineers can shift from one message to another without first blanking the sign. So there is not a one-to-one relationship between a posted message and a record of a "blank" message afterward.

During the 6-month period included in this data set, 277 distinct messages were posted. Some messages are pre-programmed and used routinely (e.g., ROADWORK), while others were unique to one-time activities. During the 6-month period, a number of major road closures occurred, especially on primary intersecting roadways. (The SR 520 bridge and West Seattle freeway viaduct both experienced full closures during some nights and weekends.) While I-5 is not a direct diversion route for these freeways, such closures do affect I-5 traffic patterns. For example, when the SR 520 floating bridge is closed for roadwork, most cross-lake travel diverts to I-90, meaning that additional traffic volumes occur southbound on I-5 from SR 520 to I-90 and northbound on I-5 from I-90 to SR 520. When the West Seattle freeway is closed, traffic diversion occurs onto City of Seattle arterials and other ramps are used to and from I-5.

The inclusion of VMS messages helps determine when and where construction activity that affected travel on I-5 occurred. Unfortunately, no single message query will identify construction activity because the messaging for those activities is diverse. In addition, a variety of construction activities affect I-5 performance. Construction can occur on the freeway itself, on the ramps that lead off I-5 to other roads, or on those other roads. Construction activities also take place on the SR 99 viaduct—parallel to I-5—which is noted on I-5 signs and can affect I-5 performance.

To determine when construction activities occurred, it is necessary to search two different variables. If the "Parameter" variable in the Table VMS_MSGS contains either of the text strings "Rdwk" or "Roadwork," then active construction activity occurred on either I-5 or one of the roadways connecting to I-5. (Note that the search for these text strings should not be case sensitive, as both terms can be expressed either in all capitals or with only the initial letter capitalized.) Next, some construction activity is also denoted by the term "Maintenance" in the Parameter variable. Finally, intersecting and major parallel routes to I-5 were closed several times for construction during the 6-month data period. These events can be found by searching for the text "clos," within both the Parameter variable and the VMS_Message variable. Unfortunately, unlike the term "roadwork," finding the text string "clos" in either variable does not ensure that a construction closure occurred, as VMS containing this phrase may merely have provided notification of an upcoming roadway closure or even closure of other important transportation facilities (such as Ferry terminals). VMS may also have indicated the incident-related closure of a lane. The data set even includes the VMS announcement of a roadway closure on I-5 for the President of the United States. (To observe performance on I-5 during that event, search the "Parameter" variable for the text "POTUS.") Understanding the intent of the "closure" notification currently requires human review of the VMS message. The location of signs posting VMS messages and the times associated with those messages indicate when and where the indicated activities occurred.

7.1 Data Description

Note that the mechanism used by WSDOT to store and VMS data functions differently than the mechanism used to store the ATM sign data described above, despite the fact that they contain many of the same kinds of data. The ATM data tables store sign status (the message being displayed) every minute of the day, making it possible to look up by time of day the message currently being displayed. The VMS data tables store data as a more traditional VMS "event log" in which records are only present when the message on a sign changed.

The non-ATM, VMS data are stored in four tables. Each table is transmitted to FHWA as a separate csv file. The first table (csv file) "VMS_Sign" describes the location of all of these VMS signs. The table "VMS_TT" is used to store data provided on fixed signs that only provide travel time estimates between the sign and specific fixed locations. These signs are not capable of providing text messages. The table "VMS-MSGS" stores text messages that are provided to travelers on fully capable variable message signs. When messages of more importance are not present on these signs (e.g., "Accident Ahead" or "Right Lane Closed Ahead") these same signs can be used to provide travel time estimates from that sign to a designated point. When used in this manner, the travel time estimate and the location associated with that travel time are provided in the table "VMS_MSG_TT."

Given the method used by WSDOT to store these data to determine what message was posted on a sign at a given time of day, an analyst must search temporally backward from the time of interest for the sign location of interest. That is, if a user desires to know what the sign at milepost 155.98 said at 11:00 AM on May 15th, he or she must search the database backward in time, starting at 11:00 AM on May 15th and restricting the query to that sign (ObjectID = 005vm15598). The first record for that sign found with a time prior to 11:00 AM on May 15th indicates the message displayed on that sign at 11:00 AM.

A record can indicate that a sign was blanked (i.e., that no message is being displayed on that sign) in two ways. The first condition is that the Action variable in the VMS_TT or VMS_MSG_TT tables may be set to "Blank Sign." The second is that the Action variable may be set to "Display MSG," with the Parameter variable set to "Blank" or "BLANK." One of these records exists in the database each time the VMS no longer showed a message. If this is the first record found during a backward temporal search, it means that the sign was blank at the time of interest. For example, if roadwork occurred from 8:00 PM until 5:00 AM, a record will appear before 8:00 PM indicating the presence of roadwork. If that work concluded as intended at 5:00 AM, a record indicating that the sign was blanked will be present (perhaps at 4:50 AM, when the road crew called the radio operator to indicate that they were clear of the roadway). In the previous example, the first record found of a search back in time from 11:00 AM, might be the "Blank" record at 4:50 AM, indicating that the sign was blank at 11:00 AM.

Table 10 describes the data available for messages posted on WSDOT's fixed VMS along I-5 during the 6-month study period.

Table 10: WSDOT Variable Message Sign Use Data Descriptions and Structure

Table VMS_Sign (file: VMS_Sign.csv) VMS Metadata			
Columns	**Data Type**	**Data Range**	**Value Description**
ObjectID	Varchar(10)	XXXvmYYYZZ	Name of roadside equipment cabinet, XXX=route code, vm=variable message, YYY.ZZ=milepost
Route	Varchar(3)	'005'	Name of route. I-5 = 005
Milepost	Decimal(5,2)	150-200	Accumulated state route milepost of cabinet
Lat	Decimal(13,10)	~ 47	Latitude
Long	Decimal(13,10)	~ 122	Longitude

Table VMS_TT (file: VMS_TT.csv) Fixed route/destination travel time signs			
ObjectID	Varchar(10)	XXXvmYYYZZ	Name of roadside equipment cabinet, XXX=route code, vm= a variable message sign, YYY.ZZ=milepost
HHMMSS	4byte int	0-240000	24-hour (local) time in integer format, no leading zeroes (5pm = 170000, 12am =0, 4:35:22pm = 163522)
YYYYMMDD	4byte int	20110501-20111031	Date in year/month /day format (May, 6, 2011 = 20110506)
UserID	Varchar(10)	'NG_WebVMS' or user name	Indicates user ID used to authorize data display. NG_WebVMS indicates the automatic travel time calculation system
Action	Varchar(15)	'Display MSG' or 'Blank Sign'	Indicates whether data are being displayed.
Parameter	Varchar(35)	TRAVEL tt YYYZZ	Parameter is a unique ID for each travel time route/destination combination displayed on the VMS system
Line 1	Varchar(35)	Potentially 4 – 119, blank, or N/A	Indicates the actual sign display for the route indicated in TTRoute1. Otherwise it holds a numerical value. (Note that this is essentially a duplicate of the variable TT1, but includes N/A or is blank when travel times can not be computed.)
Line 2	Varchar(35)	Potentially 4 – 119, blank, or N/A	Indicates the actual sign display for the route indicated in TTRoute2. Otherwise it holds a numerical value. (Note that this is essentially a duplicate of the variable TT2, but includes N/A or is blank when travel times can not be computed.)
ToCity	Char(8)	'Bellevue' or 'Seattle'	Indicates destination from sign location
TTRoute1	Char(6)	viaI90 via520 viaI5	Route from sign location to destination
TTRoute2	Char(6)	via405 viaRev	Route from sign location to destination. viaRev indicates I-5 reversible lanes.
TT1	Smallint	0, 4 – 119, 255	Travel time in minutes via route 1. 255 indicates a display of 'NA' or 'N/A' typically used for closures. '0' is typically used when sign is blanked
TT2	Smallint	0, 18 – 97, 255	Travel time in minutes via route 2. Both 0 and 255 indicate a display of 'NA' or 'N/A' typically used for closures. '0' is typically used when sign is blanked

Table VMS_MSG_TT (file: VMS_MSG_TT.csv) Travel time messages on 3 line variable message signs			
ObjectID	Varchar(10)	XXXvmYYYZZ	Name of roadside equipment cabinet, XXX=route code, vm=variable message, YYY.ZZ=milepost
YYYYMMDD	4byte int	20110501-20111031	Date in year/month/day format (May, 6, 2011 = 20110506)
HHMMSS	4byte int	0-240000	24-hour (local) time in integer format, no leading zeroes (5pm = 170000, 12am =0, 4:35:22pm = 163522)
UserID	Varchar(10)	'NG_WebVMS' or a specific WSDOT user's name	Indicates the user ID used to authorize data display. NG_WebVMS indicates the automatic travel time calculation system
Action	Varchar(15)	'Display MSG'	Indicates whether data are being displayed.
Parameter	Varchar (3)	TRAVEL vm YYYZZ	Parameter is a unique ID for each travel time sign, where YYYZZ matches the YYYZZ value of the ObjectID variable
Line1	Varchar(35)	Text	Text line 1 - Typically 'City XX MIN'
Line2	Varchar(35)	Text	Text line 2 - Typically 'City XX MIN'
Line3	Varchar(35)	Text	Text line 3 - Typically 'City XX MIN'

Table VMS_MSGS (file: VMS_MSGS.csv) Non-travel time message displayed on 3 line variable message signs			
ObjectID	Vachar(10)	XXXvmYYYZZ	Name of roadside equipment cabinet, XXX=route code, vm=variable message, YYY.ZZ=milepost
YYYYMMDD	4byte int	20110501-20111031	Date in year/month/day format (May, 6, 2011 = 20110506)
HHMMSS	4byte int	0-240000	24-hour (local) time in integer format, no leading zeroes (5pm = 170000, 12am =0, 4:35:22pm = 163522)
UserID	Varchar(10)	'NG_WebVMS' or user name	Indicates user ID used to authorize data display. NG_WebVMS indicates the automatic travel time calculation system
Action	Varchar(15)	'Display MSG' 'Blank Sign'	Indicates whether data are being displayed. Indicates sign has been blanked
Parameter	Varchar(35)	multiple	Parameter is a general classification of message type, i.e., 'Blank', 'Slow Traffic Ahead', 'Roadwork', etc.
VMS_Message	Varchar(109)	multiple	Descriptive text. Message text is all 3 lines of text concatenated

7.2 Quality Assurance Testing

The quality assurance tests performed for this data set consisted of checking to make sure that all VMS were in fact operational and located on I-5. No other quality assurance tests were performed. Note that "test messages" are present in the database.

7.3 Meta Data

Figure 14 shows the ER diagram for the relational tables that contain the non-active traffic management system variable message sign messages.

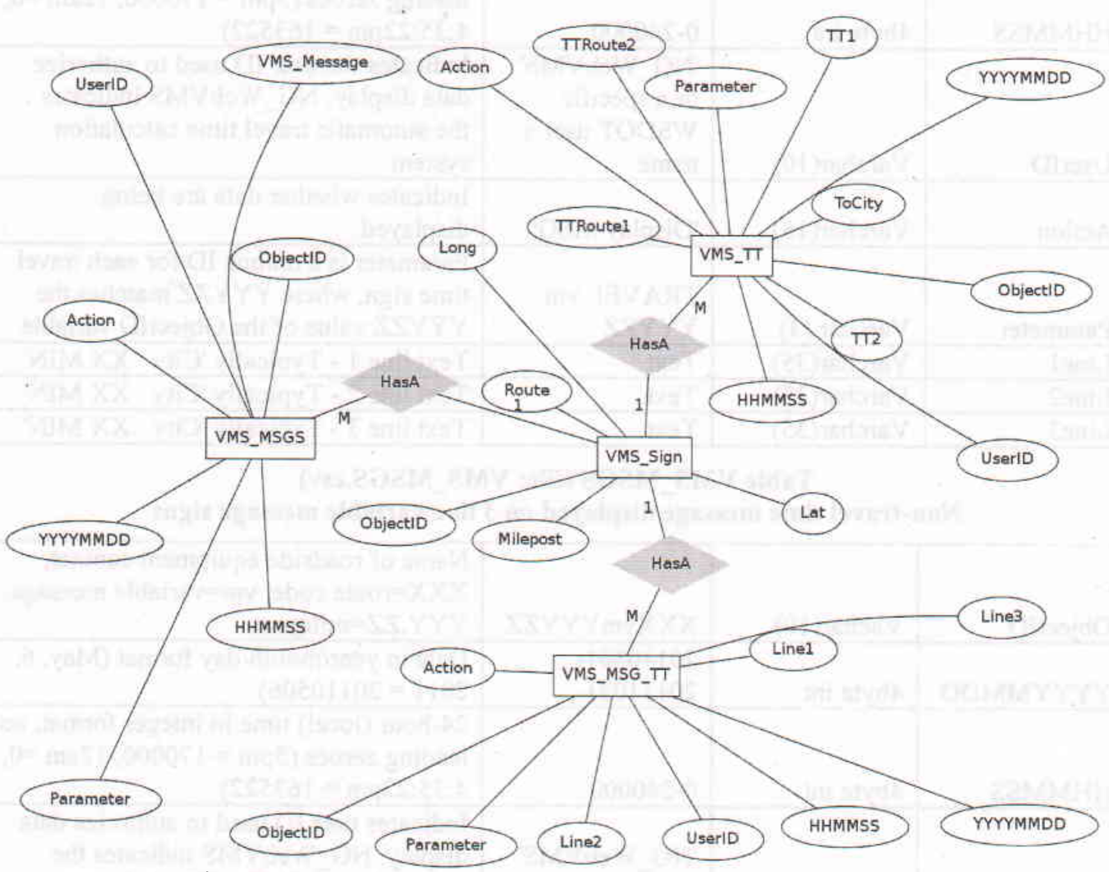

Figure 14[xiv]: ER Diagram for Other WSDOT Variable Message Sign Data

8.0 WEATHER STATION DATA

The weather data provided with the Seattle data set come from a variety of National Oceanic and Atmospheric Administration (NOAA) weather stations in the region. These data give reasonable geographic coverage in the area but do not provide coverage that is sufficiently dense to guarantee that "rain" noted at a given weather station at a specific time also fell on a specific roadway section. Weather data are available from the following locations (the sections of I-5 that are closest to these weather stations are noted):

- SeaTac Airport (southern end of Seattle metropolitan area)

- Boeing Field (just south of downtown Seattle – next to the ATM corridor)
- University of Washington Atmospheric Sciences Department (just north of downtown Seattle)
- Paine Field (Everett, Washington, north of Seattle)
- Renton (south east end of Lake Washington)

Weather data are relatively simple and come in two general forms. Data from the University of Washington Atmospheric Sciences Department (the location is called "University of Washington ATG" – for the Atmospheric Sciences Gauge) are available in readings every minute. All other stations report weather statistics roughly every hour. (Airport stations output records with time intervals that are not always exactly one hour apart. In addition, when sudden weather changes occur (e.g., a rain squall), airport stations will output a record documenting that activity when it occurs and then write another record at the "normal" hourly time. Also note that the hourly records are not output on the hour, but are generally produced at 53 minutes after the hour (e.g., 17:53 Pacific Daylight time).)

8.1 Data Description

Weather data are provided in two tables. Each table is provided to FHWA in a separate csv file. The first table (csv file) "WeatherStations" describes the geographic location where the weather station data are being collected. The second table (csv file) "WeatherReadings" contains the actual weather data. Table 11 describes the specific variables in the weather station data and the table structure used for those data.

Table 11: Weather Data Descriptions and Structure

Table 11a: WeatherStations (file: WeatherStations.csv)

Column Name	Data Type	Data Range	Description
StationId	INTEGER		Unique ID for station
Name	VARCHAR(30)	SeaTac Airport Boeing Field Paine Field Renton Municipal Airport University of Washington ATG	Name of station
Lat	DECIMAL(8,5)		Latitude of station (approx)
Lon	DECIMAL(8,5)		Longitude of station (approx)

Table 11b: WeatherReadings (file: WeatherReadings.csv)		
Column Name	**Data Type**	**Description**
StationId	INTEGER	Unique ID for station
Date	DATE	Date of record
Time	TIME	Time of record – in local time
Temperature	DECIMAL(4,1)	Temperature in Fahrenheit
RelHum	DECIMAL(4,1)	Relative humidity
WindSpeed	DECIMAL(4,1)	Wind Speed (in knots). Peak value for time period measured
WindDirection	SMALLINT	Direction of wind from source in degrees- e.g., 180 = from due South)
Pressure	INTEGER	Sea level pressure (mb)
Solar	SMALLINT	Solar radiation (W/m**2)
Rain	DECIMAL(3,1)	Rain gauge (inches). Measured rainfall since last reported measurement (e.g., since previous record. Note time intervals can vary from record to record.)

8.2 Quality Assurance Testing

Simple range and value checks were performed to detect obviously bad weather data. A total of 16 records were changed due to extreme differences in temperature or humidity readings between consecutive records. For example, 11 temperature readings of zero degrees were noted when the temperature readings on either side of those values was more than 50 degrees different than that zero degree reading. Seattle does not exhibit this kind of temperature swing. In these 16 cases, the bad data was replaced by the average of the values on either side of the bad data. Only values for temperature and humidity were changed. While this is an imperfect solution it will provide a reasonable value in those few cases where a researcher needs such a value.

8.3 Meta Data

Figure 15 shows the ER diagram for the relational tables that contain the data on weather in the Seattle area during the six month data collection period.

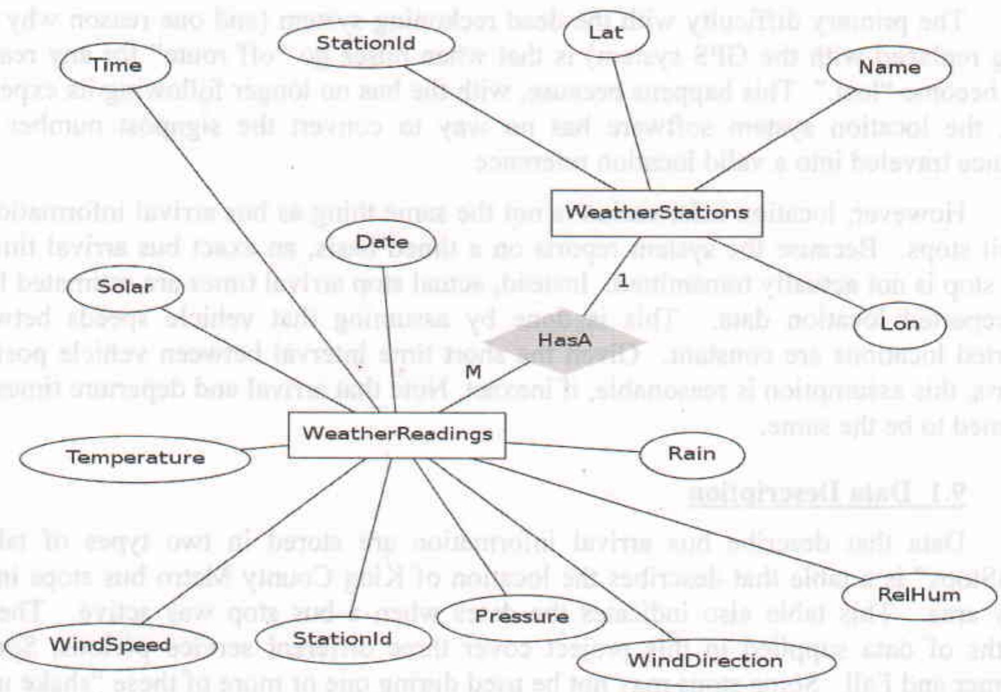

Figure 15[xv]: ER Diagram for Weather Data

9.0 KING COUNTY METRO TRANSIT AVL DATA

The Seattle data set provides estimated arrival times for King County Metro (KCM) buses, as well as Sound Transit buses operated by KCM for Sound Transit.

The bus arrival information has been extracted from the KCM Speed and Reliability Analysis (SandRA) database. These "post processed" data come from two independently operating KCM vehicle location systems. This is because KCM is currently transitioning from an older sign-post and dead reckoning technology to a more modern GPS-based vehicle location system. In both cases, each bus reports its current location to a central server via a data channel on KCM's radio system. The AVL system (whether GPS or sign post plus dead reckoning) reports location on a timed basis. That is, a central polling system requests current location from each bus in turn. This normally occurs once every one to two minutes. When its location is requested, the on-board system replies with its vehicle number and current location.

For the sign post plus dead-reckoning system, the transmitted data include the vehicle number, the time of the data transmission, the last signpost observed, and the distance traveled (from odometer information) since that signpost observation. Once these data are captured by KCM, the vehicle number is compared to a bus assignment table, which indicates the route and trip that vehicle should be on at the specified time (bus route and direction of travel). Given the path followed and the distance traveled since the last signpost, a vehicle location can be accurately estimated. When polled for location data, the newer GPS system simply provides its vehicle number and location in latitude/longitude.

The primary difficulty with the dead reckoning system (and one reason why it is being replaced with the GPS system) is that when buses go "off route" for any reason, they become "lost." This happens because, with the bus no longer following its expected path, the location system software has no way to convert the signpost number and distance traveled into a valid location reference.

However, location information is not the same thing as bus arrival information at transit stops. Because the system reports on a timed basis, an exact bus arrival time at each stop is not actually transmitted. Instead, actual stop arrival times are estimated from the reported location data. This is done by assuming that vehicle speeds between reported locations are constant.. Given the short time interval between vehicle position reports, this assumption is reasonable, if inexact. Note that arrival and departure times are assumed to be the same.

9.1 Data Description

Data that describe bus arrival information are stored in two types of tables. "BusStops" is a table that describes the location of King County Metro bus stops in the study area. This table also indicates the dates when a bus stop was active. The six months of data supplied in this project cover three different service periods, Spring, Summer and Fall. Some stops may not be used during one or more of these "shake ups." This table is supplied to FHWA in a single csv file. (Note, a bus stop End Date of 4000-01-01 is simply an identifier that the bus stop is currently active.)

The second table "BusLocations" lists the scheduled and estimated actual arrival time at each bus stop. The key variable BusStopID is used to link these data to the physical location data in the BusStops table. Because this data table is quite large, it has been broken into six monthly csv files for delivery to FHWA. Each BusLocations csv file contains an alpha numeric identifier indicating the month for which it contains data, for example BusLocations_Aug.csv contains the August data.

Table 12 describes the variables in the data set that describe the performance of the King County Metro transit fleet.

Table 12: King County Metro AVL Data Descriptions and Structure

Table 12a: BusLocations (file: BusLocations.csv)

Column Name	Data Type	Data Range	Description
Date	DATE		Date of record
ScheduledArrival	TIME		The scheduled time for the bus to arrive at the bus stop (given in local time)
EstimatedArrival	TIME		The estimated arrival time of the bus (given in local time)
ServiceRoute	SMALLINT		The route number of the bus (e.g., 72, 44).
Dir	VARCHAR(3)	IN OUT	The direction of travel for the bus: IN[bound] or OUT[bound]. Where IN[bound] refers to whether the trip is headed towards downtown Seattle or away from downtown Seattle.

BusStopId	INTEGER		In this table 'BusStopId' identifies the specific stop that for which data are being presented in this record.
TripId	INTEGER		The ID of the trip. Since a given route may have multiple buses moving in a given direction at a given time, TripId is most useful in tracking a particular bus as it serves its route
Flag	SMALLINT	0 1 2 3 4 5 6 7	The flag value of the data quality check. 0 == Good, non-zero values indicate that data errors have likely occurred and that arrival estimates are likely to be invalid. A detailed description of these values is provided in the Quality Assurance Testing text for this section of the report (page 54)

Table 12b: BusStops (file: BusStops.csv)

Column Name	Data Type	Description
BusStopId	INTEGER	The ID of the bus stop location
Lat	DECIMAL(10,7)	Latitude of station (approx)
Lon	DECIMAL(10,7)	Longitude of station (approx)
BegDate	DATE	The first date a BusStopId was active (Note that the 6 months of data span two service changes. Thus some stops may exist during the Spring "shake up" but not in the Summer or Fall "shake ups" or vice versa.)
EndDate	DATE	The final date a BusStopId was active

9.2 Quality Assurance Testing

Data quality checking guidelines were drawn from King County Metro and from the PhD dissertation of Kari Watkins[16]. The simple data quality process includes three simple analytical tests. If a given record passes all tests, it is assumed to be valid, and the "Flag" variable is set to zero. Otherwise, the flag for the record is set according to the criteria below:

- If the stop is the first or last stop for the bus, set Flag = 1. (Many trips were found to be missing this information because of different driver routing at the beginning and/or end of a trip.)

- If the bus is computed to have arrived 10 or more minutes early at a stop, or 60 or more minutes late to a stop, these are probably flawed or otherwise incorrect data; set Flag = 2.

[16] Watkins, Kari, *Using Technology to Revolutionize Public Transportation*, University of Washington, June 2011.

- If the route pattern (the order in which a bus reports the bus stops to which it arrives) occurs only once per day, the route patterns are almost certainly erroneous; set Flag = 4.

If more than one of these three flags needs to be set, the following "combined" flags are used:

- If both Flag = 1 and Flag = 2 are required, set Flag = 3 (i.e., the bus stop is the first stop AND the bus is more than 10 minutes early or more than 60 minutes late).
- If both Flag = 1 and Flag = 4 are required, set Flag = 5.
- If both Flag = 2 and Flag = 4 are required, set Flag = 6.
- If all three Flags (Flag = 1, and = 2, and =4) are true, set Flag = 7.

9.3 Meta Data

Figure 16 shows the ER diagram for the relational data tables that contain bus location and stop arrival data for King County Metro Transit.

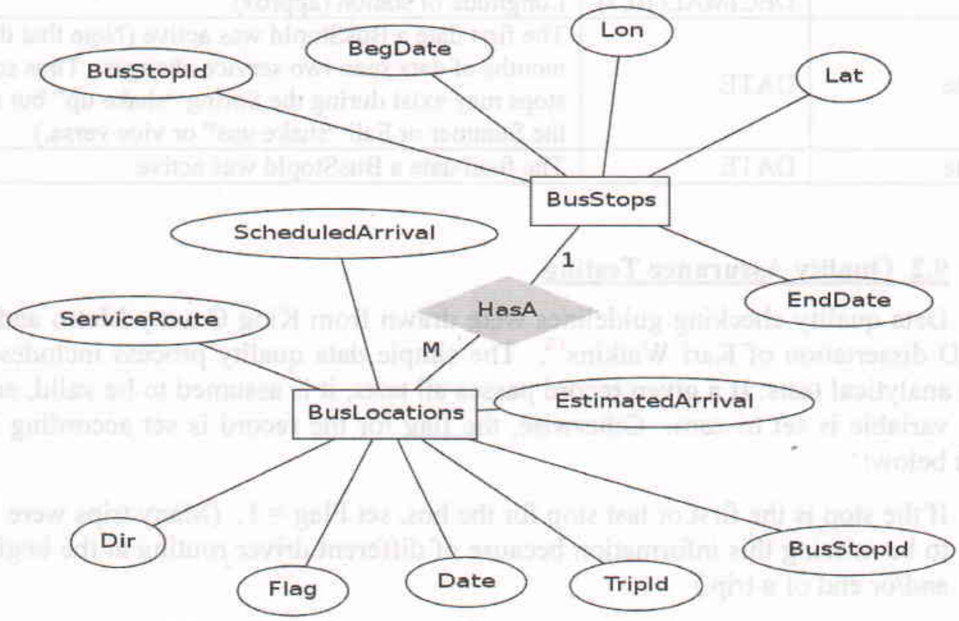

Figure 16[xvi]: ER Diagram for King County Metro Bus Arrival Data

10.0 SOUND TRANSIT RAIL SCHEDULE DATA

Sound Transit is a regional transit authority operating long distance (regional) bus, light rail, and commuter rail services in King, Pierce, and Snohomish counties. The Seattle area data set includes schedule data for the two currently operating train systems that serve the downtown Seattle area. The two train services are the Central Link light rail train (LRT) and the Sounder commuter rail train.

The Central Link LRT service operates from SeaTac airport (located south of downtown) to the Westlake Center Station in the downtown bus tunnel. Westlake Center is located in the heart of Seattle's downtown retail core. The bus tunnel serves both the Link LRT and a variety of bus routes. The tunnel provides high frequency, high speed transit connections throughout Seattle's downtown. The Central Link LRT operates on a headway basis, not a fixed schedule. However, for analytical purposes, train services can be represented as a fixed schedule. This estimated fixed schedule is provided in the data incorporated in this data set. Sound Transit does not currently provide actual vehicle location data for Central Link trains.

Sound Transit operates two commuter rail lines on track shared with Burlington Northern Santa Fe freight trains. One train line operates from the south, starting in Tacoma (in Pierce County) with stops in Puyallup, Sumner, Auburn, Kent, and Tukwila, before ending the trip at the King Street station on the southern edge of downtown Seattle. The second line operates from the north, starting in Everett in Snohomish County with stops in the cities of Mukilteo and Edmonds before also ending at King Street Station. On weekday mornings, the Tacoma–Seattle line operates seven northbound trips and two southbound trips. In the evening, seven southbound trips and two northbound trips occur. For the Everett-Seattle route, four trains operate on weekdays inbounds (from Everett to downtown Seattle) in the morning, with four corresponding outbound trips in the evening.

Currently, Sound Transit does not publish data on actual train arrivals and departures. Therefore, the only data available and provided with this data set are the scheduled train arrivals at each station.

10.1 Central Link Light Rail Schedules

Link runs three different schedules: a weekday schedule, a Saturday schedule and a Sunday schedule. Since the vast majority of Link tracks are grade separated, travel times do not routinely vary significantly from station to station by time of day or day of week. However, headways—and thus arrival/departure times—change among these three schedules.

Sound transit does not differentiate between arrival and departure times at Link stations. It publishes only a single time (assumed for this data set to be a departure time), which is the time provided in this data set. Link does not currently operate additional trains for special events. However, it does operate special schedules on holidays. Three holidays occurred during the 6-month period for the Seattle data set: Memorial Day (May 30, 2011), July 4th, and Labor Day (September 5, 2011). During each of these three holidays, Sound Transit operated Link on a Sunday schedule.

10.2 Sounder Train Schedules

Sound Transit Sounder (heavy rail) and Central Link (light rail) data were acquired from Sound Transit from their developer website. http://developer.soundtransit.org/Developer-Home/Data-Downloads/Download-Data.xml. This site is updated every time Sound Transit updates its schedules (about every four months). The Sound Transit data for all non-light rail is distributed in GTFS (General Transit Feed Specification) format. For more information on GTFS and/or this data set, go to: code.google.com/transit/spec/transit_feed_specification.html. The data obtained from Sound Transit have been simplified and transformed to provide estimates of train arrivals at stations.

In the supplied database, Sounder rail operations data are stored in four tables. The first table describes the arrival and departure times of trains at specific stops and connects those specific arrival and departure times to specific train trips. The second table supplies the detailed stop data needed to geo-locate the train stations. The third table describes specific trip details. The final table identifies specific holiday schedules, which allows handling of exceptions to the normal Sounder schedule on observed holidays.

As with the Link schedule, Sounder train schedules changed from weekday to Sunday schedules on the three holidays that fell during the 6-month period: Memorial Day (May 30, 2011), July 4th, and Labor Day (September 5, 2011). Since no Sounder train service is provided on Sunday, this means that no train service was provided on these three holidays.

10.2.1 Data Description

The same data formats are used for both Sounder commuter rail services and Central Link light rail services. The train schedule data are split into our tables. Each table is provided to FHWA in a separate csv file. The first table (csv file) "Stops" describes the geographic location of the train stations. The second table (csv file) "Trips" defines the specific trips that are made each day. The third table (csv file) "Stop_times" describes the scheduled stop times and stop locations for each trip. The last table (csv file) "Calendar_dates" is used to identify exceptions to the weekday, Saturday, and Sunday schedules. For the six month project period, it identifies specific weekday holidays during which non-weekday schedules were operated, and what schedules (usually a Sunday schedule) were operated instead of the normal Sunday schedule. These data and their table structure are described in Table 13.

Table 13: Sounder (Commuter Rail) and Central Link (LRT) Train Schedules Data Descriptions and Structure

Table 13a: stop_times (file: stop_times.csv)		
trip_id	VARCHAR(6)	The trip ID. Often involves the train number
arrival_time	TIME	Scheduled time of arrival expressed as local time
departure_time	TIME	Scheduled time of departure expressed as local time
stop_id	VARCHAR(10)	Unique stop ID
stop_sequence	INTEGER	Identifies the order of stops. Lower stop_sequences come before higher numbers, but they do not need to be adjacent (e.g., first stop = 1, second stop = 4, third stop = 27).

Table 13b: stops (file: stops.csv)		
stop_id	VARCHAR(10)	Unique stop ID
stop_name	VARCHAR(40)	Verbose form of stop_id: contains the name of the station and the direction of travel
stop_lat	DECIMAL(10,7)	Latitude of stop station
stop_lon	DECIMAL(10,7)	Longitude of stop station

Table 13c: trips (file: trips.csv)

Column Name	Data Type	Data Range	Description
trip_id	VARCHAR(6)		The trip ID. Often involves the train number
route_id	VARCHAR(9)		Unique identifier for a route
service_id	VARCHAR(2)	WD SA SU	WD implies all weekdays (unless the date is explicitly listed in calendar_dates), SA implies all Saturdays, SU all Sundays.

Table 13d: calendar_dates (file: calendar_dates.csv)

Column Name	Data Type	Data Range	Description
service_id	VARCHAR(2)	WD SA SU	WD implies all weekdays (unless the date is explicitly listed in calendar_dates), SA implies all Saturdays, SU all Sundays.
date	INTEGER	20110530 20110704 20110905	The date in question in YYYYMMDD format
exception_type	SMALLINT	1 2	1 indicates adding a service_id, 2 indicates removing it, e.g., to change 20110530 from a regular WD (weekday) schedule to a SU (Sunday) schedule on a holiday use two entries (service_id, date, exception_type): (SU, 20110530, 1), (WD, 20110530, 2). This effectively 'subtracts' WD service from 2011-05-30 and 'adds' SU service

10.2.2 Quality Assurance Testing

Rail schedules were reviewed manually before inclusion in the database. In addition, the schedule tables were compared with Sound Transit's on-line trip planner. When minor differences were observed in these two data sources, the Sound Transit operations staff were asked for clarification of actual trip schedules. These differences only occurred for very late night trips as trains were being taken out of service.

10.2.3 Meta Data

Figure 17 shows the ER diagram for the relational data tables that contain bus location and stop arrival data for Sound Transit rail schedule data.

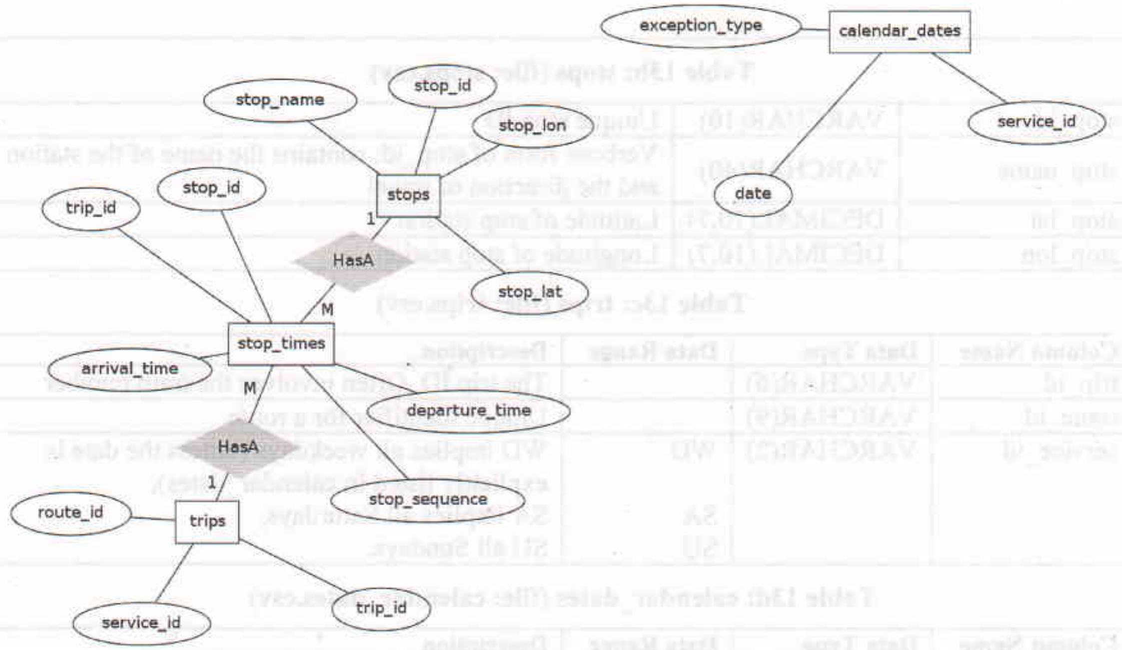

Figure 17[xvii]: ER Diagram for Sound Transit Rail Schedule Data

11.0 TRUCK GPS DATA

The project team currently obtains GPS location (lat/long), speed, and heading reads from roughly 9,200 commercial trucks operating in the Puget Sound metropolitan region. The GPS devices of different vehicles report their positions at a variety of time intervals, ranging from every 30 seconds to every 15 minutes. These data are then snapped to a GIS map representation of the highway network. Only the data points that were successfully "snapped" to one of the study roadway sections are provided in this data set. The map sections included in the data set are all state routes in the study area.

11.1 Data Description

The truck GPS data are provided in two tables. Each table is provided to FHWA as a separate csv file. The first table (csv file), "TruckRecords," includes the GPS data themselves. Each record in the GPS data table includes the vehicle speed, latitude, longitude, the time of the observation (to the minute), and the map segment ID. The second table (csv file), "TruckSegments," describes the roadway segments to which they are "snapped." That is, it is the map segmentation table. This table includes the beginning point of the map segment, the ending point of the map segment, the segment speed limit, the name of the road, and the direction of that road segment. (That is, northbound and southbound directions of travel are represented as different roadway segments.)

Table 14 describes the specific variables in the truck GPS data and the table structure used for those data within the Seattle database.

Table 14: Truck GPS Data Descriptions and Structure

Table: TruckSegments (file: TruckSegments.csv)

Column Name	Data Type	Data Range	Description
SegmentId	INTEGER		Unique ID for section. First three values are the road (e.g., all I-5 sections begin with 005)
SpeedLimit	SMALLINT		Posted speed limit (does not take variable speed limits into account)
Dir	BIT(1)	0 1	0: Direction of increasing milepost values (N/E) 1: Direction of decreasing milepost values (S/W)
BeginMp	FLOAT		The beginning milepost value
EndMp	FLOAT		The ending milepost value
Road	VARCHAR(11)		Name of road (e.g., "5" for I-5). "005RL005EXP" is the reversible Express Lanes. – See the description of those lanes in Section 2.3 of this document.
IsUrban	BIT(1)		Boolean indicator of whether the segment is within an "Urban" area as defined by WSDOT (0 = not urban, 1 = is in an urban area)

Table: TruckRecords (file: TruckRecords.csv)

SegmentId	INTEGER	Unique ID for section. First three values are the road (e.g., all I-5 sections begin with 005)
Dir	BIT(1)	Inc(1): Direction of increasing milepost values (N/E); Dec(0): Direction of decreasing milepost values (S/W)
TruckID	CHAR(16)	The truck identifier, a unique identifier for each vehicle observed on each day
Date	DATE	Date of record
Time	TIME	The local time the record was taken (rounded to the nearest minute)
Lat	DECIMAL(9,6)	Latitude of record
Lon	DECIMAL(9,6)	Longitude of record
Speed	SMALLINT	The observed speed of a truck (mph)

11.2 Quality Assurance Testing

The data cleaning and quality assurance process for these data entails four steps:

1) initially snapping the GPS data to road segments
2) examining heading and speed values to determine whether the GPS is functioning correctly or recently lost communication with satellite signals
3) using vehicle headings to determine whether a vehicle was on a cross-road rather than the primary arterial
4) checking vehicle headings to determine the direction of travel.

The first step uses a 60-foot (30-meter) buffer from the road midpoint to determine whether a specific GPS point is "on the roadway." The initial map snapping algorithm filters out GPS reads that were not traveling along a WSDOT route.

Heading data are then checked to associate the GPS travel bearing with the road segment's bearing. If the vehicle heading is within 15 degrees of the bearing of the roadway, the vehicle is assumed to be on that road. If the heading is more than 15 degrees off from the road's bearing, the vehicle is assumed to be located on an intersecting roadway or on a land parcel adjacent to the road's right-of-way. The roadway bearing used in this analysis is extracted from the WSDOT GIS roadway shape file; it is taken from the portion of the roadway shape file description closest to the GPS location.

GPS data points falling outside of the 15-degree criterion have been removed from the database and are NOT included in the Seattle database. The 15-degree criterion was adopted from previous work performed with the truck GPS data.[17]

Vehicle speed is also used to indicate whether the GPS has "lost signal" (nonsensical speeds such as 170 mph are indications that the GPS location is not reliable), in which case the data point has been removed from the analysis, regardless of whether the GPS data point is located within the 60-foot boundary around a roadway. Note that only extreme high speeds are considered "nonsensical," as any roadway can contain zero speed or near zero-speed traffic. This project did NOT perform a more detailed comparison of truck GPS data with loop-based speed estimates to determine whether additional data points should be removed from the database.

The current data cleaning process errs on the side of data removal, assuring that all data identified as collected on a particular road segment were in fact contained by that segment.

11.3 Meta-data

Figure 18 shows the ER diagram for the relational data tables that contain truck GPS probe data.

[17] Developing a GPS-Based Truck Freight Performance Measures Platform, by McCormack, E.D., Ma, Klocow, Currarei, and Wright, for WSDOT, WA-RD #748.1, 2011.

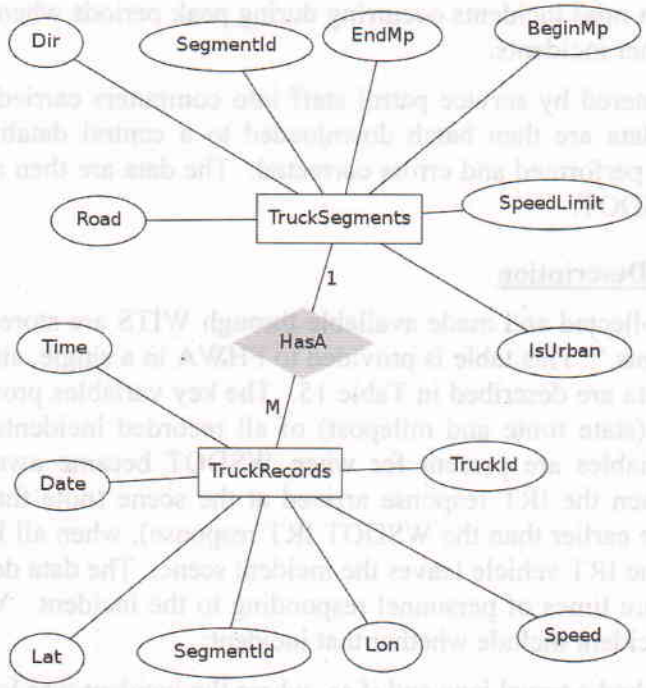

Figure 18[xviii]: ER Diagram for Truck-Based Vehicle Probe GPS Data

12.0 INCIDENT RESPONSE DATA

As is typical for most urban areas, incident response is performed by a combination of agencies in the Seattle metropolitan region. The Washington State Patrol is responsible for law enforcement on urban freeways. Local fire departments are responsible for fire, most emergency medical response (through Medic 1 services), and a significant portion of initial hazardous material response. Additional incident response services are provided by on-call private tow trucks, as well as other additional hazardous material and emergency response service providers (e.g., ambulance companies.)

The WSDOT is responsible for overall roadway operations. To improve roadway operations, provide support for the traveling public, and provide support to other agencies working in the roadway right-of-way, the WSDOT operates a significant freeway service patrol in the Seattle Metropolitan area as part of their formal incident response team (IRT.) The service patrol drives both light trucks equipped with push bumpers and some light duty tow trucks. Service patrol staff rove the metropolitan freeway system during much of the day – with specific emphasis on the peak periods. They collect data on all incidents to which they respond. This data base is called the Washington Incident Tracking System (WITS.)

The WITS data collection program is the basis for the incident data included in this database. It does not capture all incidents. Incidents that are not reported are those for which WSDOT is not directly involved. This includes primarily minor incidents occurring in off-peak hours and even some vehicle crashes occurring in non-peak hours.

WITS data also can miss incidents occurring during peak periods when the service patrol staff are busy at other incidents.

Data are entered by service patrol staff into computers carried in their response vehicles. Those data are then batch downloaded to a central database where quality assurance tests are performed and errors corrected. The data are then made available for analysis within WSDOT.

12.1 Data Description

The data collected and made available through WITS are stored in a single table called "WitsIncidents." This table is provided to FHWA in a single, similarly named csv file. The WITS data are described in Table 15. The key variables provided are the date, time and location (state route and milepost) of all recorded incidents within the study period. Time variables are present for when WSDOT became aware of an incident (notified time), when the IRT response arrived at the scene (note that WSP may have arrived at the scene earlier than the WSDOT IRT response), when all lanes were open to traffic, and when the IRT vehicle leaves the incident scene. The data does not include the arrival and departure times of personnel responding to the incident. Variables available to describe each incident include whether that incident;

- blocked a travel lane and if so, where the incident was located,
- was classified as an emergency,
- involved a fatality
- involved hazardous material clean up,
- involved a fire response,
- occurred within a work zone,
- involved an injury or a medical response,
- indicates whether the IRT staff believe that weather factors contributed to causing the incident.

Other available variables are listed in Table 15.

Table 15: Incident Response Data Descriptions and Structure

Table: WitsIncidents (file: WitsIncidents.csv)			
Column Name	**Data Type**	**Data Range**	**Description**
IncidentId	INTEGER		Unique incident identifier. Six digits
IncidentType	VARCHAR(20)	Abandoned Vehicle Debris Disabled Fatality Collision Injury Collision Non-Injury Collision Other Police Activity	Description of incident

StateRouteId	VARCHAR(11)		State route where incident occurred. 5 == I-5. Uses WSDOT naming convention for special roads (e.g., reversible, spur)
Milepost	DECIMAL(5,2)		Milepost where incident occurred
Landmark	VARCHAR(256)		Text description of where the incident physically occurred on the roadway
Date	DATE		Date of incident notification
NotifiedTime	TIME		The local time at which initial notification took place
ArriveTime	TIME		Time (local) when the IR unit reaches scene - HH:MM:SS
ClearTime	TIME		The local time when IR leaves the scene - HH:MM:SS
AllLanesOpenTime	TIME		The local time when a blocking incident clears and all lanes are again open to traffic. Remains blank if a non-lane-blocking incident
Direction	VARCHAR(10)		Direction of Travel
Blocking	VARCHAR(12)	Blocking Non-Blocking	Indicates whether the incident blocked a lane of travel.
PrimaryLaneClosureType	VARCHAR(16)	All Travel Lanes HOV Multiple Lane Other Shoulder/Median Single Lane SR 167 HOT Lanes Total Closure	Description of lane closure resulting from incident (if any)
SecondaryLaneClosureType	VARCHAR(16)	All Travel Lanes HOV Multiple Lane Other Shoulder/Median Single Lane SR 167 HOT Lanes Total Closure	Description of lane closure resulting from incident (if any). Identical to PrimaryLaneClosureType.
NotificationType	VARCHAR(8)	Call-Out Notified Roving	Type of notification: Roved Upon (discovered by incident response staff while roving), Notified (notified by DOT/WSP agent), Call-Out (requested WSDOT assistance in some manner)
IrUnitType	VARCHAR(12)	Roving IR IR Tow Unit IR Call-Out Maintenance Admin	Type of IR service responding

IrNotifier	VARCHAR(15)	Other WSDOT Roved Upon WSDOT Radio/TMC WSP Public	Defines who called incident response
IsEmergency	BIT(1)	0, 1	Indicates whether incident was flagged as an emergency. 0 = not an emergency, 1 = was an emergency
IsNonTrafficAccident	BIT(1)	0, 1	Indicates whether incident did not involve traffic (e.g., an incident which occurred off of the roadway but ended up effecting the roadway 1 = no traffic was involved
InvolvedHazmat	BIT(1)	0, 1	Indicates whether hazardous materials were involved 1 = hazardous material involved
InvolvedFire	BIT(1)	0, 1	Indicates if a fire was involved 1 = fire involved
InvolvedWeather	BIT(1)	0, 1	Indicates whether the incident appears to be related to extreme weather (e.g., snow, hail) 1 = yes, weather related
InvolvedMedical	BIT(1)	0, 1	Indicates if someone with medical training was summoned 1 = medical assistance called
InsideWorkZone	BIT(1)	0, 1	True indicates if the incident occurred in a work zone 1 = incident in a work zone
InvolvedBikeOrPed	BIT(1)	0, 1	Indicates if incident involved a bicycle or pedestrian 1 = bike or ped involved
MotoristAssisted	BIT(1)	0, 1	Indicates whether there was assistance given from a motorist (e.g., someone's friend delivered gasoline) 1 = assistance given
PoliceAssisted	BIT(1)	0, 1	Indicates whether there was assistance given from the police (e.g., WSP regularly pushes vehicles off of roadway) 1 = police assisted

12.2 Quality Assurance Testing

The quality assurance tests are carried out by WSDOT as part of a routine quarterly review of the WITS data. The QA checks consist of basic range and value checks and a comparison against Washington State Patrol CAD data. These checks allow for the capture of the majority of significant data entry errors, but are not capable of detecting errors such as the input of imprecise locations or incorrect times. Users should note that more than one incident can occur at the same time and same milepost. These are not errors (as best we can determine.) Instead, they are occurrences where more than one reported incident is present. In many, but not all cases, the incident report comes from a roving incident response member that comes upon more than one disabled vehicle sitting on the shoulder within a few hundred feet of each other and lists both vehicles as being at the same location.

12.3 Meta Data

Figure 19 shows the ER diagram for the data table that contains Washington State DOT's incident response data. All incident data are stored in a single table.

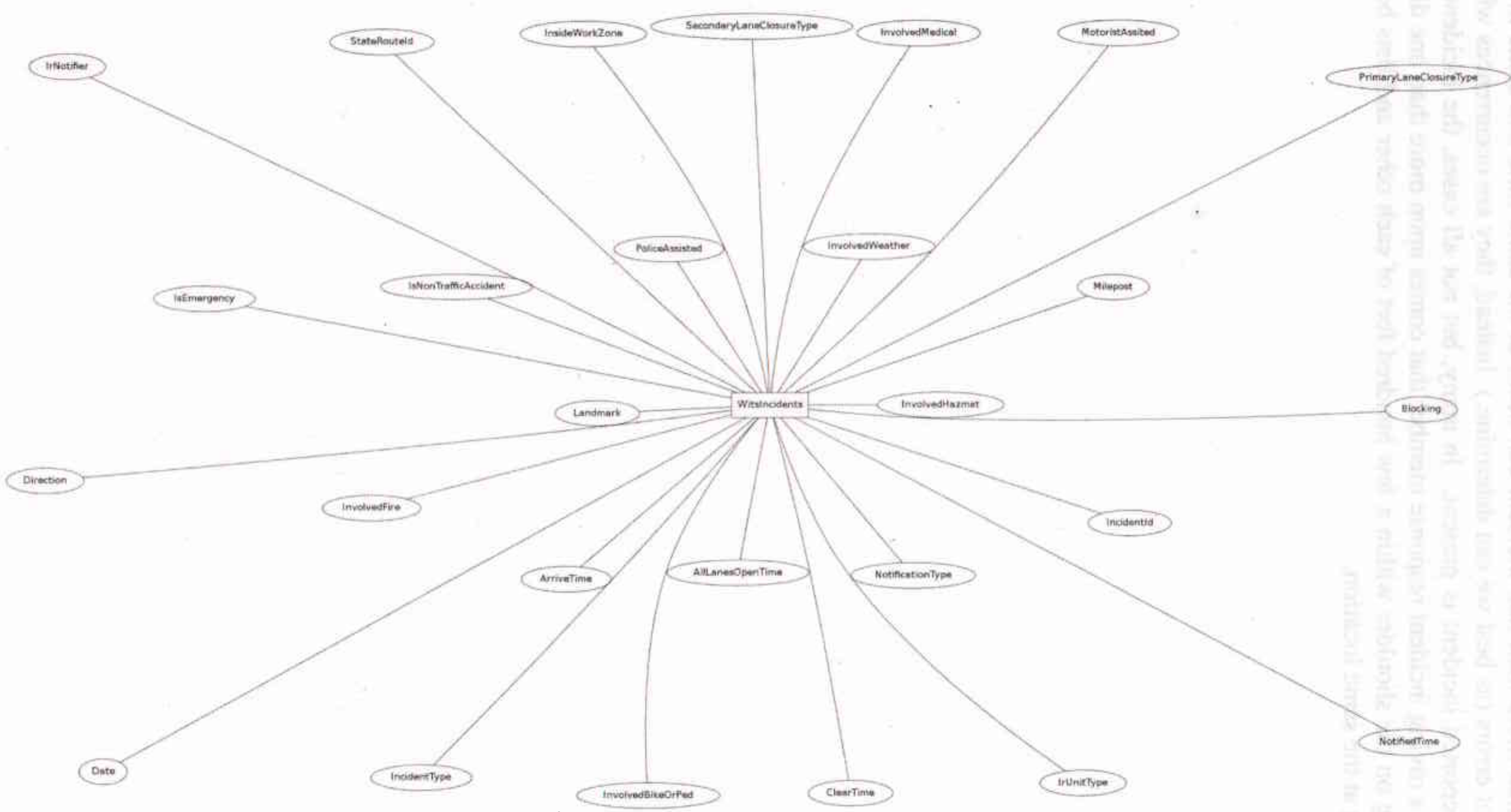

Figure 19[xix]: ER Diagram for Incident Response Data

APPENDIX

SECTON 508 WRITE UPS FOR REPORT FIGURES

[i] Figure 1 (Geographic Area for I-5 Included in the Data Submittal) is a screen capture from WSDOT's Puget Sound congestion website (http://www.wsdot.wa.gov/traffic/seattle/) It shows the general geographic layout of the city, with emphasis on the locations of the major freeway interchanges. The southern edge of the graphic is Federal Way, Washington. The top of the graphic is the southern edge of Everett, Washington. Downtown Seattle is at the center of the map's left side. The map shows Lake Washington in the center of the image, and I-405 on the right side of the image. A large black rectangle overlays the map. The annotation for this rectangle says "Seattle Arterial Fig. 2)." The rectangle refers the reader to Figure 2, which is an illustration of the City of Seattle arterials for which data are available. The rectangle covers much of the City of Seattle, extending in a north/south direction for the length of Lake Washington. On the east/west axis the rectangle extends west from Lake Washington to Puget Sound.

[ii] Figure 2 (City of Seattle Arterials to Be Included in the Data Submittal) illustrates the City of Seattle arterials for which data are available. The image is taken from the City of Seattle's congestion and traveler information website. (http://web5.seattle.gov/travelers/) Figure 2 highlights in green the major north-south arterials for which travel times are collected and included in this data set. The figure does not provide detailed guidance on which arterials are present; however, that information is included in the detailed data set descriptions later in the report.

[iii] Figure 3 (ER Diagram for WSDOT 20-Second Loop Data) shows symbols representing four relational data tables. The tables are named CabinetInfo, LoopsInfo, LD20Sec_2011XX (where XX is the month), and CF2011XX (where again XX is the month.) The diagram shows that the LoopInfo table is the central point of reference. It is linked to the actual loop data (stored in the LD20Sec_2011XX and CF2011XX tables) via the LoopID variable. It is linked to the table that provides details about the equipment cabinets (CabinetInfo), such as the latitude & longitude, via the WSDOTCab variable.

[iv] Figure 4 (ER Diagram for WSDOT 5-Minute Traffic Data) shows that the 5-minute loop detector data is stored in three relational data tables. They are named Cabinets, Loops, and LoopData. The central table is the Loop table, which provides a unique identifier (Id) for each loop. This value exists in all three tables and is used to link the tables. Each record in the Loop table contains a reference to a specific Cabinet, as well as variables which describe the specifics of what data each loop is collecting (e.g., HOV or GP lane, mainline or ramp, northbound or southbound, etc.) These values are used by

researchers to select the loops of interest. These variables are used to look up the LoopID, which then points to the LoopData table which contains the actual data collected at that loop for each 5-minute time period of interest.

[v] Figure 5 (ER Diagram for I-5 Travel Times Based on WSDOT 5-minute Data) shows symbols representing three relational data tables named TravelTrips, TravelTimes, and TravelLocations. TravelTrips names the trip and provides the start and end points, as well as assigning a specific ID to that trip. That ID is used to link to the TravelTimes table which contains the specific travel time data along with date and start time for each trip. The ID is also used to link to the TravelLocations table which lists the latitude and longitude of the state and end points for the trip.

[vi] Figure 6 (ER Diagram for City of Seattle Sensys Volume Data) shows symbols representing three tables, SeattleData, SeattleSensor, and SeattleCabinets. The SeattleCabinets table includes the data which describe the specific location where volume data are collected with Sensys devices. This table is linked by the variable CabinetID to the SeattleSensor table which describes the specific data collection attributes of each sensor (e.g., lane number and direction.) The individual sensors are linked to the actual volume, occupancy, and speed data stored in the SeattleData table by the variable SensorNum that identifies individual detectors and is in both tables.

[vii] Figure 7 (ER Diagram for City of Seattle Signal Timing Plans) is an entity relationship diagram. It shows that the signal timing plan data are stored in a single table called SeattleTimingPlans. (See Table 6 for descriptions of those variables.)

[viii] Figure 8 (Initial (Raw) Travel Time Samples) is an X/Y plot of travel times (on the Y-axis) versus time of day (on the X-axis) for one specific day on the SR 522 corridor. The resulting scatterplot shows individual points in red, with the moving average visible as a black line overlaying the red data points. The scatterplot itself is a thick red line, with a number of random red data points occurring above it. The thick line represents the band of valid travel times. The scattered points above it represent the errors caused by vehicles stopping at some point along the corridor. The X-axis starts at noon (12:00 PM) and goes until 8:00 PM (20:00). At noon, the band of valid trips contains travel times ranging from roughly 4 minutes and 30 seconds, to a little less than 7 minutes, with a mean travel time of just over 5 minutes. (The rolling mean is not very stable, and during the next hour jumps well above 5 minutes because of the inclusion of the invalid data. The invalid data are travel times well above 15 minutes, with some travel times greater than 25 minutes.) The valid travel time band stays fairly consistent until after 2:00 PM, when travel times slowly increase. The band stays roughly the same width. The travel times peak at roughly 6:00 PM, with a mean travel time of roughly 10 minutes, before decreasing back to a mean travel time of roughly 5 minute near 8:00 PM.

[ix] Figure 9 ("Valid" Travel Times after One QA/QC Algorithm Pass) is an X/Y plot of travel times (on the Y-axis) versus time of day (on the X-axis) for one specific day (the same as in Figure 8) on the SR 522 corridor. As in Figure 8, individual data points are

shown in red, with the moving average shown as a black line overlaying the red data points. The scatterplot is a thick red line, with a limited number of random red data points above the thick line. The number of points above the main band of travel time points is much smaller than that observed in Figure 8. The rolling mean travel time is also much more stable in Figure 9 than in Figure 8. The invalid data points above 20 minutes are now gone. Some invalid points (e.g., travel times above 10 minutes in the early afternoon when the mean is about 5 minutes) are still present. However, the same increase in travel times observed in Figure 8 occurs in the evening in this Figure, with the peak in travel times still falling around 6:00 PM. This means that only the travel times above 15 minutes in the heart of the peak period are likely invalid. Several of these exist.

[x] Figure 10 ("Valid" Travel Times after Two QA/QC Algorithm Passes) is also an X/Y plot of travel times (on the Y-axis) versus time of day (on the X-axis) for the same day of data shown in figures 8 and 9. As in Figure 8, individual data points are shown in red, with the moving average shown as a black line overlaying the red data points. Once again, the scatterplot is a thick red line, with a limited number of random red data points above the thick line. Unlike figures 8 and 9. very few data points are visible above the main band of travel time points. In addition, the rolling mean travel time is even more stable than those observed in figures 8 and 9. The increase in travel times is still observed in the evening, with the peak in travel times falling around 6:00 PM.

[xi] Figure 11 (Moving Average of Travel Times after Two QA/QC Algorithm Passes) is also an X/Y plot of travel times (on the Y-axis) versus time of day (on the X-axis) for the same day of data shown in figures 8 through 10. Unlike the previous three figures, this plot shows only the final moving average travel times. All 24 hours of the day are shown, so the first and last values on the X-axis are for midnight. The moving average shows that even late at night (~2:00 AM), travel times on the arterial vary. These variations result from traffic signals, as well as driver behavior. The combined effects of these two factors also produce fairly substantial variation in the moving average travel time throughout the course of the day. The moving average travel times frequently vary by up to 2 minutes from the mean travel time less than 10 minutes earlier. This figure also shows the dual direction congestion present on this test section. Travel times are actually slowest (approaching 12 minutes) at 8:00 AM. In contrast, the average travel time in the PM peak is about 10 minutes. Late night travel times are generally just over 4 minutes.

[xii] Figure 12: (ER Diagram for Automated License Plate Readers (ALPRs) shows that the ALPR data is structured into two tables. One table AlprCameras describes the location of each ALPR camera. This includes the GPS coordinates of those cameras. The second table (ALPRTravelTimes) contains the travel times for specific arterial segments. The segments are defined by the variables BeginCamera and EndCamera. These variables link directly to the ALPRCameras table and are equivalent to the CameraId variable in that table. (Thus the ALPRTravelTimes table links to two different records in the AlprCameras table.)

[xiii] Figure 13 (ER Diagram for WSDOT Active Traffic Management Data) shows that the active traffic management sign data are stored in three tables. The smallest table is called "SignGantires, and contains the geographic location of the ATM sign gantries, eash of which contains several dynamic message signs. The "Signs" table links to the gantry table via the variable GNName, contains the sign and gantry names, and describes the type of sign each sign is (e.g., lane control or one of two types of dynamic message sign.) This information is linked to the SignDisplays table using the sign Id variable. The SignDisplays table contains the actual messages posted on these signs, by time and date.

[xiv] Figure 14 (ER Diagram for Other WSDOT Variable Message Sign Data) shows that four tables are required to maintain the WSDOT's conventional VMS data. The central table of the four is the VMS_Sign table, which creates a single identifying variable (ObjectID) for each sign which serves as the link to the other three tables. It also contains the location identifying information (latitude, longitude, milepost, name, etc.) This table links directly to the VMS_MSGS table, which contains the text messages posted on conventional VMS, including the time and date when each message is posted. The VMS_Sign table also links to two other tables used to describe travel time messages posted by WSDOT. The VMS_TT table lists two specific routes for each sign where "real time" travel times are posted along with the two travel times being posted, along with the date and time for when these travel times change. The other table VMS_MSG_TT includes text messages placed on the dynamic message signs that include a real time travel time estimate as part of the dynamic message.

[xv] Figure 15 (ER Diagram for Weather Data) shows that the weather station data are stored in two tables, WeatherStations and WeatherReadings. These tables are linked by the variable StationID. the WeatherStations table provides the name and location information associated with each of the reported weather stations. The WeatherReadings table provides the reported weather data, along with the time and date for which that data are reported.

[xvi] Figure 16 (ER Diagram for King County Metro Bus Arrival Data) shows that the bus arrival data is divided into two tables, BusStops and BusLocations. These two tables are linked by the BusStopId variable. the BusStops table describes the location of each bus stop, as well as the when that bus stop becomes active, and when it goes out of service (in case a stop moves or is removed at a service change.) The BusLocations table indicates when each bus arrives at a specific bus stop. Associated with each bus arrival is the date and time of arrival and the route and trip being served by that bus.

[xvii] Figure 17 (ER Diagram for Sound Transit Rail Schedule Data) shows that the rail schedule data has been divided into four tables (Trips, Stop_Times, Stops, and Calendar_dates.) One of those tables "Calendar_dates" is linked only to the "trips" table. It identifies when holidays are occurring and describes which schedule (usually a Sunday schedule) is being used for specific weekdays that are state or national holidays. The Trips table includes the data that indicate specific routes and trips being served, whether for the Sounder or LINK rail services. Each trip links to specific bus stops using the bus

stop Id. The arrival data is reported at the stop level for each trip. The stops themselves then link to the Stops table which contains the location information to place each stop on a map.

[xviii] Figure 18 (ER Diagram for Truck Based Vehicle Probe GPS Data) shows that the truck GPS data are stored in two tables, TruckRecords and TruckSegments. TruckRecords contains the specific GPS points (latitude, longitude and heading of the data point observed, the speed of the vehicle, a TruckID and the data and time of the observation. Associated with each GPS point is a road segment ID for state routes. This ID is used to link to the TruckSegment table which contains the description of that road segment, including beginning and ending mileposts, speed limit, and direction of travel on that road.

[xix] Figure 19 (ER Diagram for Incident Response Data) shows that all incident response data are stored in a single table.

The minimum qualifications are required for a Consultant to be eligible to submit a RFP/RFQ response. Responses must clearly show compliance to these minimum qualifications. Those that are not clearly responsive to these minimum qualifications shall be rejected by the City without further consideration:

Minimum Qualification:

The consultant (the individuals performing the analysis) must have a minimum of XXX years continuous experience in providing public sector consulting services, including analysis of transportation management and operations issues.

The consultant should have experience providing transportation consulting services to at least three local government entities and/or jurisdictions in dense urban environments, with populations in excess of 300,000 persons.

The consultant should be able to demonstrate experience providing transportation consulting service to at least three government entities that have multiple lines of business (e.g., street maintenance and paving, parking, traffic signal management, urban forestry, snow removal) and facilitate travel by multiple modes.

Desired Qualifications:

1. The consultant should have a minimum of XXX years continuous experience in providing public sector consulting services, including substantial experience in transportation consulting.

2. The consultant should be familiar with industry standards and best practices in street maintenance, traffic management, parking, urban forestry, and commuter mobility.

3. The consultant should have experience evaluating:

    a. The benefits and drawbacks of different transportation organizational and management structures,
    b. The prioritization of capital transportation projects, many of which are funded through multiple local and federal revenue sources, in times of limited budgets,
    c. The costs and benefits of different methodologies for prioritizing street maintenance projects,
    d. Best practices in transportation infrastructure construction and maintenance (including contracting and construction management),
    e. Weighing these considerations within the context of meeting multiple stated policy goals including fair labor practices, social equity, quality of life, complete streets and environmental goals (including water quality and climate goals).

a) Scope of Work

This work will be completed in two phases. This solicitation will solicitation will select a consultant based on the entire scope of work. The Consultant must respond to and be prepared for both phases of work. The City has provided a very specific schedule and tasks, although the City reserves the right to modify the work as required to fulfill the City's needs, including ending the work after Phase I, if they so choose.

Phase I.
Preliminary Analysis of Seattle Department of Transportation Operations and Management and Identification of Short-term and Long-term Opportunities

September 15, 2012 to January 15, 2013. Funding Available: $XXX,XXX

During Phase I, the consultant will conduct an initial assessment of SDOT's organizational structure and operations to identify operational areas for more in-depth review in 2013.

1. Phase I will last approximately 18 weeks, and the Phase I Report will be due to the City of Seattle on Tuesday, January 15, 2013.

2. Phase I will include the following activities:

   a. A peer review that benchmarks the productivity of SDOT's operations and processes with four to six other similar sized transportation departments across the country. The goal of this review is to identify operations SDOT performs well relative to other jurisdictions, areas that could be improved ("lessons learned"), innovative or promising practices, and areas for further examination in Phase II. The review should include all lines-of-business in the following SDOT Divisions: Street Use and Urban Forestry, Street Maintenance, Capital Projects and Roadway Structures, Traffic Management, Policy and Planning, and Major Projects.
   b. A review of how SDOT measures performance and productivity across all its divisions. The goal of this review is to identify ways to enhance transparency and improve accountability, particularly in regard to how SDOT reports to the public and elected policymakers. The review will also include a comparison of how SDOT measures performance relative to how other jurisdictions measure performance, and recommendations, if any, for changes to how SDOT measures and reports its performance and productivity.

   c. An overview of different transportation management models and a comparison of SDOT's management structure to other jurisdictions. This review should be conducted within the context of peer transportation agencies whose management structures have proven successful based on proven track records and outcome data. The review should also include recommendations for which model(s), if any, warrant further exploration in Phase II.

3. The Phase I Report will identify a broad scope of opportunities for efficiencies, but will not include full analyses of these efficiencies, which will occur in the second phase of this project.

   a. Do we want to differentiate here between short-term and long-term efficiencies? Define the type of efficiencies we are looking for?

   b. Long-term efficiencies are defined as those operational changes, cost reduction items or revenue generation items that are complex in nature, may require labor negotiations, or may result in a different deployment of staff, resources and apparatus for emergency response and prevention services. Long-term efficiencies will be more fully addressed in Phase II of this project.

   c. Long-term efficiencies may include items such as: do we want to provide examples of areas for consideration?

4. The Phase I Report should serve as a high-level workplan for the second phase of this project, and should outline the additional analytical tasks that the consultant expects to undertake in these phases. The Phase I report should include, at a minimum:

   a. A table comparing SDOT's performance to other jurisdictions' performance, for each of SDOT's lines-of-business (see attached example). This comparison should include both output and outcome performance measures.
   b. Recommended performance measures which reflect industry best practice and a recommended one-page dashboard which summarizes the most important measures (for reporting purposes).
   c. A cross-walk/explanation that links the recommended performance measures to SDOT's strategic goals, (I am not sure if this is really of interest or not, but it seems like is should be)

d. Recommendations for short-term efficiencies, if any.
e. Recommendations for areas that should be examined further in Phase II.
f. A comparison of SDOT's management structure to the other jurisdictions, a discussion of the relative merits (pros and cons) of different organizational structures.
g. Recommendations for which model(s) are likely to be the best fit for SDOT, and should be explored further in Phase II.

5. Upon the City's acceptance of the Phase I Report, the City will make a Phase I payment of $XXX,XXX to the consultant.

6. After submission of the Phase I Report, the consultant should be available to conduct briefings to the City Budget Office, the Mayor's Office, the Seattle City Council, the Seattle Department of Transportation, the unions representing Seattle transportation workers, and other stakeholders, as needed, to present the findings of the Report.

Phase II.
Long-term Review of Operational Changes and Efficiencies at the Seattle Department of Transportation Focused on Program Level Improvements and Recommendations

February 15, 2012 to December 1, 2013. Funding Available: $XXX,XXX

During Phase II, the consultant will identify and recommend changes to SDOT's operational approaches and management practices that could yield greater budget efficiencies, more productivity, and/or better outcomes for maintaining and improving the City's transportation infrastructure.

1. Phase II will last no longer than forty-two weeks and the Phase II Report will be due to the City of Seattle on December 1, 2013.

2. Phase II will include the following activities:

    a. Analysis and recommendations on areas identified in Phase I for in-depth review,

    b. An evaluation of the policy and decision-making processes SDOT uses to prioritize transportation maintenance projects and recommendations for how they could be improved. This should include, at a minimum, a risk assessment and a review of how SDOT uses its Asset Management Program, including how they determine life cycle costs. Recommendations should be based on industry best practices related to project level decision-making, and (This was in the draft resolution, I am not sure exactly what it means or how it fits with prioritizing transportation projects.)

    c. An analysis of SDOT's responsibilities to help policymakers narrow its focus and priorities lading up to potential renewal of the Bridging the Gap Levy. This analysis should include a review of SDOT's estimated $1.8 billion deferred maintenance backlog.

3. The Phase II report will include a full description of all recommendations for improving SDOT's policy, practice and operations, with accompanying analysis on:

    a. What exactly would change, and how (staffing, equipment, materials, use of technology, new processes or procedures, etc.),
    b. Evidence from other jurisdictions that supports the value of making the change(s) (see # 4 below),
    c. Fiscal impact (short term cost or savings and long-term savings),
    d. Externalities associated with implementation,
    e. Service impacts,

f. Personnel impacts,
g. Impacts on the Department or City's ability to achieve equity, quality of life, environmental and other stated policy goals or values,
h. A proposed implementation plan,
i. The timeline required to fully implement each recommendation,
j. A proposal for how to measure the effectiveness of the proposed change,
k. An analysis of the expected impact of the change on SDOT's strategic objectives, and
l. Any other factor relevant to the recommendation's policy assessment or implementation.

4. To the extent applicable, the consultant should provide information on other jurisdictions that have implemented similar efficiency measures, the agencies' related decision-making processes and an estimate of the actual financial or operational savings attained.