

PORTAL Data Quality Analysis

October 2008

Kristin A. Tufte, Robert L. Bertini

Portland State University

Understanding the data and data quality in the PORTAL transportation data archive is important to maintaining a high quality data archive. This report presents an initial study of the quality of the Oregon Department of Transportation (ODOT) data that is received by PORTAL. We examine the data to attempt to identify trends in the data that may point to issues with the freeway loop detectors and the ATMS system.

Summary of Results

This report presents an initial high-level analysis of quality of the loop detector data received by PORTAL, the regional transportation data archive which is housed at Portland State University (PSU). The report analyzes data received by PORTAL during the month of September 2008 to identify high-level data quality concerns. The purpose of this analysis is to detect large scale problems with invalid readings or incorrect data reporting; more detailed analysis is required to fully understand detailed data quality issues. This report focuses on four categories of readings: No Traffic, Zero Occupancy, Very High Speeds and Low Overnight Speeds.

We reach the following conclusions:

- **No Traffic.** In September 2008, 16% of the readings indicated no traffic was observed at the detector. The analysis indicates that the majority of the no traffic readings likely reflect actual no traffic conditions. However, it does appear that the activation of the SWARM system is causing a slight increase in no traffic readings during the afternoon; these afternoon no traffic readings are likely communication failures due to increased network congestion from SWARM.
 - *Action:* PSU will leave no traffic readings as is and assume those readings are valid. ODOT may wish to investigate the impact of SWARM on the afternoon no traffic readings.
- **Zero Occupancy.** In September 2008, 6% of the readings had a zero occupancy value along with speed and volume greater than zero. Such a reading is technically invalid as occupancy should be greater than zero if there is traffic at the detector (as would be indicated by the greater than zero speed and volume). The analysis indicates that this problem may be due to occupancy readings being truncated to integers. More detailed data than is currently available is needed to completely answer this question.

- *Action:* PSU will assume that the zero occupancy readings are valid readings for which an occupancy value of between 0 and 1 has been truncated to 0. The data quality project at PSU led by Dr. Tufte may investigate this issue further.
- **Very High Speeds.** The vast majority of high speed readings in September 2008 were due to high speed readings from two detectors.
 - *Action:* These two detectors should be investigated by ODOT staff.
- **Low Overnight Speeds.** Low overnight speed readings are a known problem for Portland area loop detectors. Related work at PSU indicates a potential relationship between the type of loop amplifier card installed at a detector and the severity of the low overnight speed reading problem. This analysis suggests that for detectors with large numbers of low overnight speeds, very low speeds (0-10mph) dominate and also that the averaging such very low speeds may be a potential cause of the low speed readings in the 10-40 mph range.
 - *Action:* PSU will consider filtering low overnight speeds when producing filtered or imputed aggregations.

Data Description

PORTAL receives volume, speed, occupancy, and status readings every 20 seconds from the loop detectors on the Portland-area freeways; there are almost 500 detectors on the Portland-area freeways. This data is received by PSU in real time over the TATII network and is inserted into the PORTAL database. The data is provided as an XML file which is parsed by scripts on the PSU server before the data is inserted into the database. Volume, speed, occupancy and status are provided as integer values. PORTAL also receives data from freeway detectors in Vancouver, WA; the Vancouver data is archived, but is not yet included in PORTAL analysis reports or web pages and is not addressed in this report.

General Data Classification

Table 1 shows a breakdown of the data readings received by PORTAL from the ODOT ATMS system during the month of September 2008. The table is broken into three sections: valid readings, readings converted by PORTAL, and theoretically invalid readings. Note that in Table 1, a '*' indicates any (all) status value(s), so the first line includes all readings with volume, speed and occupancy all > 0 and any status code.

Valid Readings: The first section shows the percent (68.5%) of theoretically valid readings where volume, speed and occupancy are all > 0. We note that some of these readings may be suspicious or unreasonable (i.e. speeds > 100 mph); such suspicious readings will be investigated in the future. For the purpose of this report, such readings are considered valid.

Converted Readings: In the next section, we consider readings which are converted on input into the PORTAL system. The ODOT ATMS uses patterns of -1s and 0s to indicate particular traffic conditions or errors including: no controller communications (communication failure) and no traffic at a detector. The patterns of -1s and 0s are converted to NULL values as appropriate on import into the PORTAL system. Of the three rows in this section, we observe that ODOT reports communications failures 5.4% of the

time and no traffic at the detector 16.3% of the time. Both number seem reasonable, the 16.3% no traffic will be examined in more detail below.

Invalid Readings: The last seven rows of the table report on data readings that are theoretically invalid, such as a speed > 0 with zero volume or volume > 0 and zero occupancy and so on. For most of the invalid reading categories (i.e. rows in the table), there are a negligible number of readings in that category. However, we observe that 5.8% of the readings have non-zero speed and volume, but zero occupancy. It is suspected that such readings may be due to rounding or truncating of the occupancy value. The ODOT ATMS system reports occupancy as an integer value. Certain valid volume and speed combinations are expected to generate occupancies of $< 1\%$; thus, many of these 5.8% readings may indeed be valid readings where an occupancy value was truncated or rounded. This issue is investigated in more detail below.

Table 1 Breakdown of Data Readings Received by PORTAL from ODOT ATMS for September 2008

What PORTAL Receives				PORTAL Actions/Comments	What is inserted into the PORTAL Database					
Volume	Speed	Occupancy	Status		Volume	Speed	Occupancy	Status	Percent	Count
>0	>0	>0	*	Valid Readings. No conversion, data inserted as is.	>0	>0	>0	*	68.5%	34,104,393
-1	-1	-1	0	ODOT says this reading indicates no controller communications (no data received from detector); PORTAL converts the -1's to NULLs	NULL	NULL	NULL	0	5.4%	2,685,113
0	0	0	0	ODOT says this reading indicates detectors reporting no traffic; PORTAL sets speed to NULL and volume and occupancy to 0	0	NULL	0	0	16.3%	8,134,996
-1	-1	-1	1	PORTAL also receives -1's with status 1; as status 1 indicates disable, PORTAL converts these -1's to NULLS	NULL	NULL	NULL	1	1.3%	642,300
>0	0	0	*	Invalid Reading. No conversion, data inserted as is.	>0	0	0	*	0.9%	455,043
>0	0	>0	*	"	>0	0	>0	*	0.7%	346,445
0	>0	0	*	"	0	>0	0	*	0.0%	11
0	>0	>0	*	"	0	>0	>0	*	0.0%	18
0	0	>0	*	"	0	0	>0	*	1.1%	523,039
>0	>0	0	*	"	>0	>0	0	*	5.8%	2,903,280
0	0	0	*	Not applicable. See row 2.						
				Total					100%	49,794,638

No Traffic

The initial breakdown indicated 16.3% of all readings were 'no traffic' readings. This percentage seems reasonable in general; however, we examine whether the patterns in the no traffic readings match expected patterns of low traffic. In particular, we expect higher numbers of no traffic readings when flow is low and lower numbers when flow is high and we expect a difference in percent of no traffic readings between the 'fast' (left-hand) lane and the 'slow' (right-hand lane).

Figure 1 shows a breakdown of no traffic readings by hour of day. This figure shows for each hour the percentage of readings in that hour that were 'no traffic' readings. The grey box indicates the time periods during which the SWARM system is active; 6AM-10AM and 1PM-7PM. The no traffic readings follow a reasonable pattern leading us to believe that most of the reported no traffic readings may indeed be valid 'no traffic' readings. The apparent slight increase in no traffic readings in the afternoon hours (1PM-5PM) is somewhat suspicious. It is known that in the past the activation of the SWARM system caused congestion in the data transfer network between the loop detectors and the ATMS system. Further, the ATMS often generated a 'no traffic' reading, when in fact the problem was actually a communications error. We question whether the afternoon increase in no traffic readings could be due to SWARM system activation.

Figure 2 shows the same plot as in Figure 1, but with the addition of a box plot showing average flow plus or minus one standard deviation. As with the previous figure, this figure supports the idea that most of the no traffic readings are likely due to no traffic conditions, with the possible exception of the slight afternoon increase.

Figure 4 breaks down the no traffic readings by hour of day and by lane. In this figure, lane 1 is the left-most (high speed) lane. In general, the pattern in this plot is as expected. During most hours, the majority of the no traffic readings come from the low-traffic, left-most lanes. Also, for most hours there are significant differences between the number of no traffic readings in lane 1 versus the other lanes. However, we note that in the afternoon period (1PM-5PM), the difference between number of no traffic readings between lanes lessens – the graph shows similar number of no traffic readings in lanes 1 through 3. This similarity again suggests a problem in this afternoon peak period, possibly due to activation of the SWARM system. We note that freeways in Portland have from one to four lanes, with most freeways having two or three lanes. For the purposes of this figure, lane 1 is always the left-most lane, lane 2 the next lane over, and data is counted for lanes 3 and 4 if the freeway segment has 3 or 4 lanes.

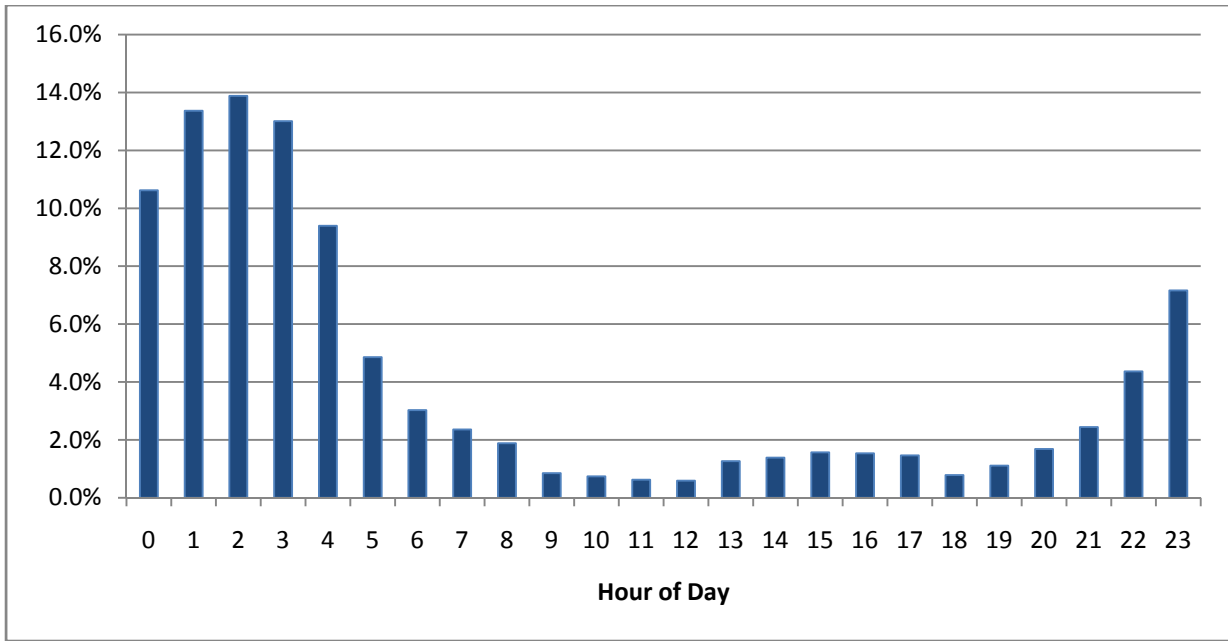


Figure 1 Percent of No Traffic Readings by Hour of Day (September 2008)

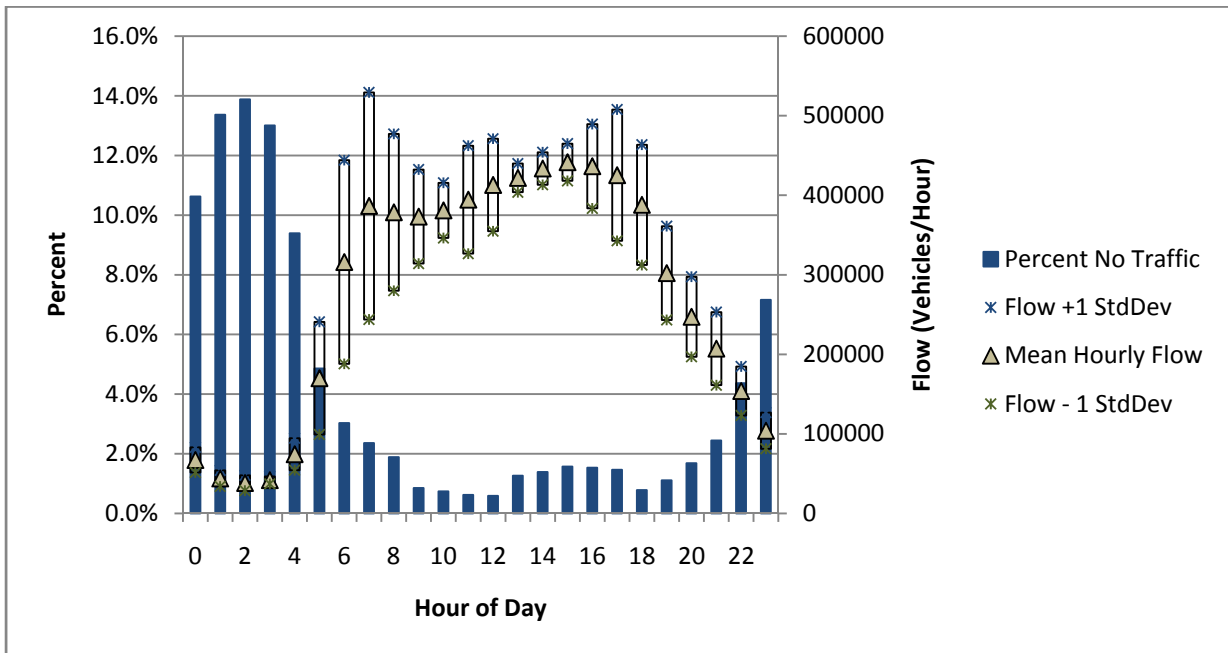


Figure 2 No Traffic Readings and Flow by Hour of Day (September 2008)

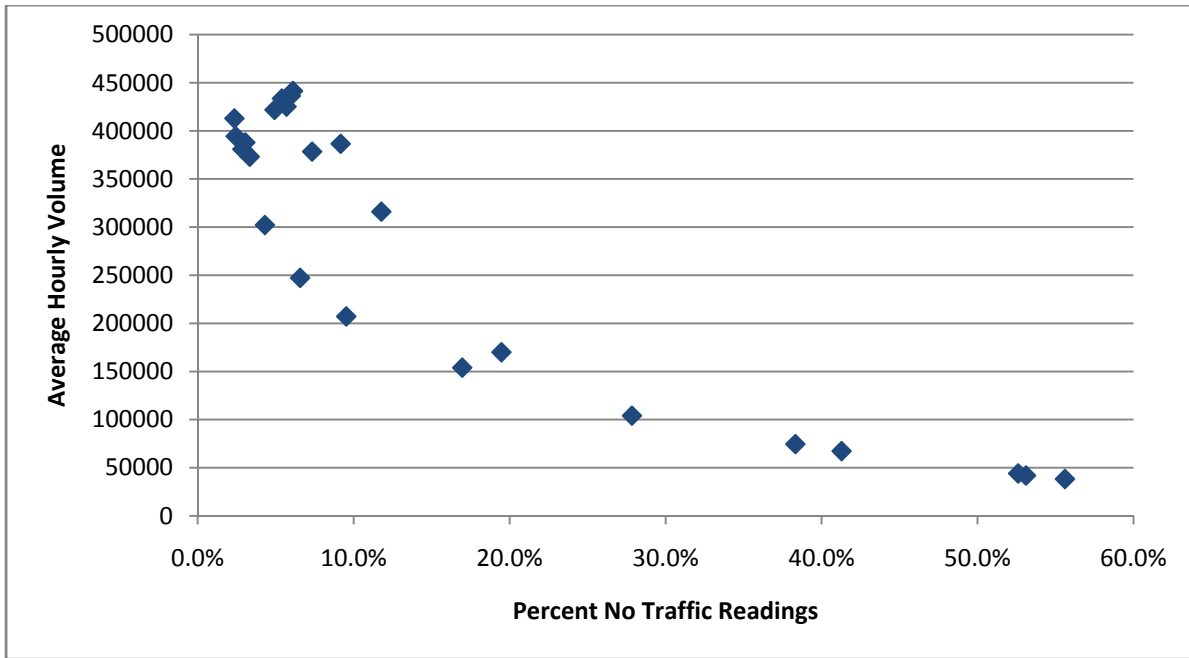


Figure 3 Percent No Traffic Readings vs. Average Hourly Flow (September 2008)

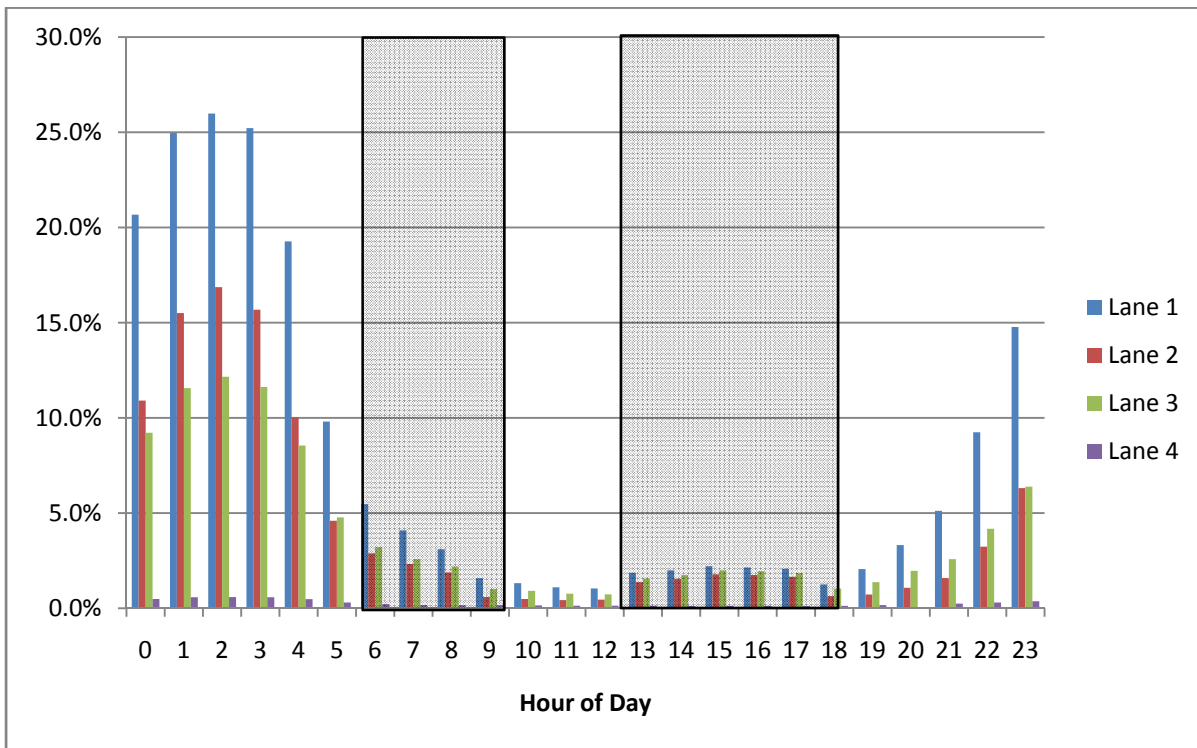


Figure 4 Percent of No Traffic Readings by Hour of Day and Lane (Pct of Hour Total) (Sept 2008)

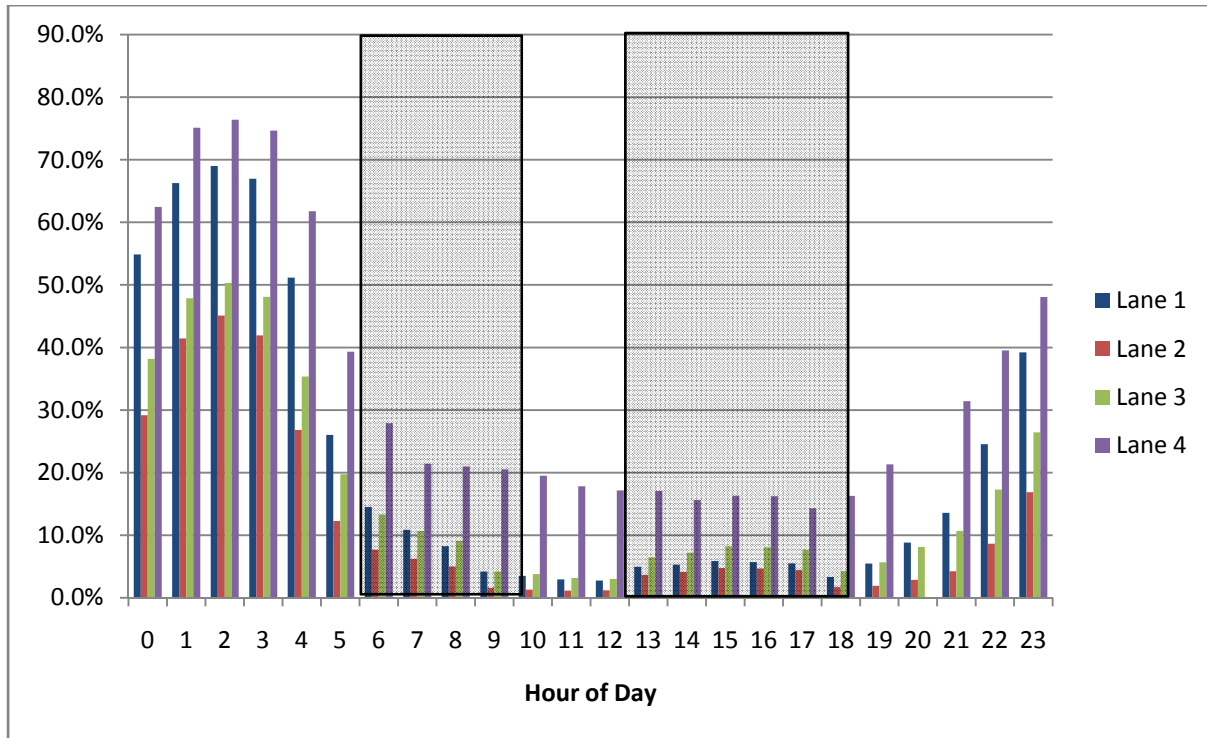


Figure 5 Percent of No Traffic Readings by Hour of Day and Lane (Pct of Hour, Lane Total) (Sept 2008)

Zero Occupancy

The initial breakdown indicated that 5.8% of the readings had a theoretically invalid combination of zero occupancy with non-zero speed and volume. We suspect that a large percentage of these readings may be due to truncated or rounded occupancies, as occupancy is provided as an integer value. The following analysis investigates that hypothesis.

Occupancy Percentage (O) can be expressed as a function of Count (C), Vehicle Length (L) and Vehicle Speed (S) as follows:

$$O = ((C*(L+6))/S)*(0.0341) \quad (1)$$

The derivation of Equation (1) can be found in the Derivations section of this report. Using this equation, we can calculate occupancies for various speed and length values. Table 2 shows calculated occupancies (using the formula above) for a wide range of counts, lengths and speeds. Cells highlighted in grey have occupancy less than 1%. We observe that a single 15-foot vehicle going 80 mph will generate an occupancy reading of 0.9%. In comparison, two 15-foot vehicles going 80 mph will generate an occupancy reading of 1.8%. In general, it appears that only readings with a count of one could legitimately generate occupancy readings of less than 1% as the vehicle lengths required to generate occupancy readings of less than 1% for counts of two or more are too short. Thus we assume all readings with zero occupancy and count of two or more are likely to be invalid and focus on the readings with count equal to one.

Table 2 Occupancies for Various Counts, Lengths and Speeds

		Speed									
Count	Length	10	20	30	40	50	60	70	80	90	100
1	0	2.0%	1.0%	0.7%	0.5%	0.4%	0.3%	0.3%	0.3%	0.2%	0.2%
1	5	3.8%	1.9%	1.3%	0.9%	0.8%	0.6%	0.5%	0.5%	0.4%	0.4%
1	10	5.5%	2.7%	1.8%	1.4%	1.1%	0.9%	0.8%	0.7%	0.6%	0.5%
1	15	7.2%	3.6%	2.4%	1.8%	1.4%	1.2%	1.0%	0.9%	0.8%	0.7%
1	20	8.9%	4.4%	3.0%	2.2%	1.8%	1.5%	1.3%	1.1%	1.0%	0.9%
2	0	4.1%	2.0%	1.4%	1.0%	0.8%	0.7%	0.6%	0.5%	0.5%	0.4%
2	5	7.5%	3.8%	2.5%	1.9%	1.5%	1.3%	1.1%	0.9%	0.8%	0.8%
2	10	10.9%	5.5%	3.6%	2.7%	2.2%	1.8%	1.6%	1.4%	1.2%	1.1%
2	15	14.3%	7.2%	4.8%	3.6%	2.9%	2.4%	2.0%	1.8%	1.6%	1.4%
2	20	17.7%	8.9%	5.9%	4.4%	3.5%	3.0%	2.5%	2.2%	2.0%	1.8%
3	0	6.1%	3.1%	2.0%	1.5%	1.2%	1.0%	0.9%	0.8%	0.7%	0.6%
3	5	11.3%	5.6%	3.8%	2.8%	2.3%	1.9%	1.6%	1.4%	1.3%	1.1%
3	10	16.4%	8.2%	5.5%	4.1%	3.3%	2.7%	2.3%	2.0%	1.8%	1.6%
3	15	21.5%	10.7%	7.2%	5.4%	4.3%	3.6%	3.1%	2.7%	2.4%	2.1%
3	20	26.6%	13.3%	8.9%	6.6%	5.3%	4.4%	3.8%	3.3%	3.0%	2.7%

Table 3 shows a breakdown of zero occupancy readings by count. We observe that almost 85% of the readings have count = 1 and are potentially valid; the other 15% of the readings can be assumed to be invalid as discussed above.

Table 3 Breakdown of Zero Occupancy Readings by Count (September 2008)

Count	Percent
1	84.8%
2	11.8%
3	1.9%
4	0.6%
5	0.3%
>6	0.6%

We further investigate the readings with count equal to one to determine if they are likely to be valid or invalid. From equation (1), it is clear that occupancy is dependent on vehicle length. Given an assumed occupancy reading, we can calculate the vehicle length that would be expected to be associated with that occupancy. Thus we assume occupancy values and analyze the associated expected vehicle lengths.

Solving equation (1) for length, produces the following equation:

$$L = ((O * 29.3 * S)/C) - 6 \quad (2)$$

Further, assuming that count is one and assume that occupancy is 0.5%, then we have: $L = (.1465 * S) - 6$. Using this formula, we can derive calculated vehicle lengths based on the assumed count and occupancies. Figure 6 shows a histogram of such calculated vehicle lengths for 0.5% occupancy. From this histogram, we observe that the estimated vehicle lengths typically range from 0 feet to 10 feet. Most passenger vehicles are at least 15 feet in length, so this suggests that perhaps many of the zero occupancy readings are invalid. (For comparison, a Honda Civic is about 15 ft in length, while a Chevrolet Suburban is about 18 ft long.) We note that if we assumed a higher occupancy (i.e. 1%) the calculated vehicle lengths would be longer and more reasonable. Figure 7 shows the same plot, but adds a histogram for occupancy = 1%.

Based on the analysis in this section, we conclude that the zero occupancy readings are likely to be valid. We accept the fact that there will be error due to vehicles changing lanes and other similar issues. The readings with count > 1 are within the tolerance of what one might expect for a loop detector that is functioning properly. The readings with count = 1 follow a reasonable pattern. The range of calculated vehicle lengths is somewhat lower than would be expected, but the shape of the histograms of calculated vehicle lengths appear valid. Based on the available data, which is limited in its accuracy and level of detail, we believe that the most likely situation is that the zero occupancy readings are mostly valid occupancy readings with values between 0 and 1 that have been truncated to 0. More detailed data would be required to answer this question fully; more detailed examination of the existing data is unlikely to produce any additional information.

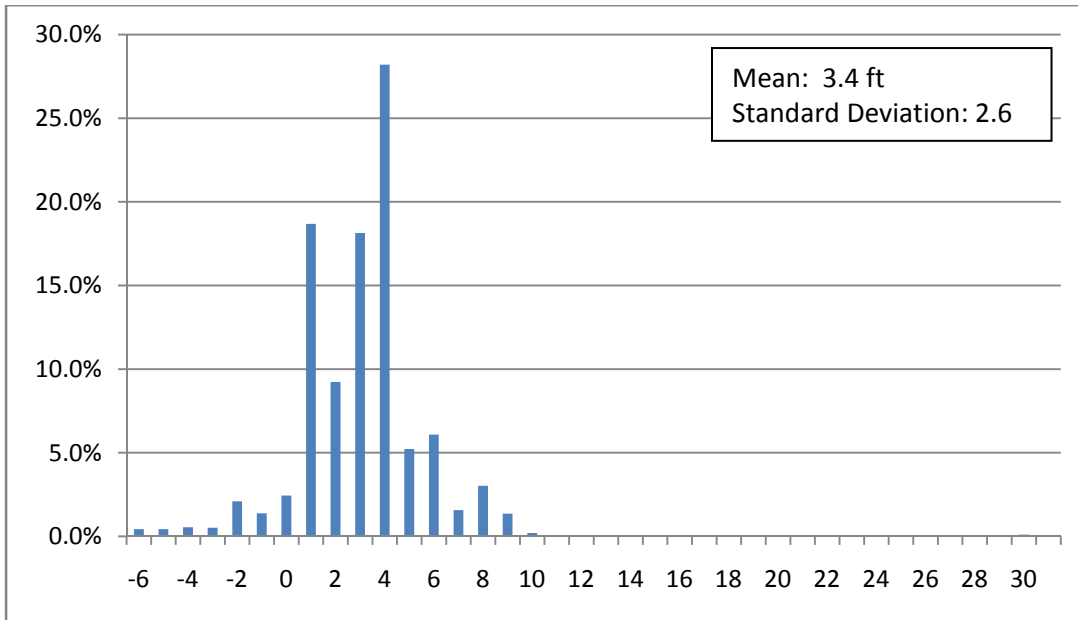


Figure 6 Calculated Vehicle Lengths; Count = 1; Occupancy = 0.5% (September 2008)

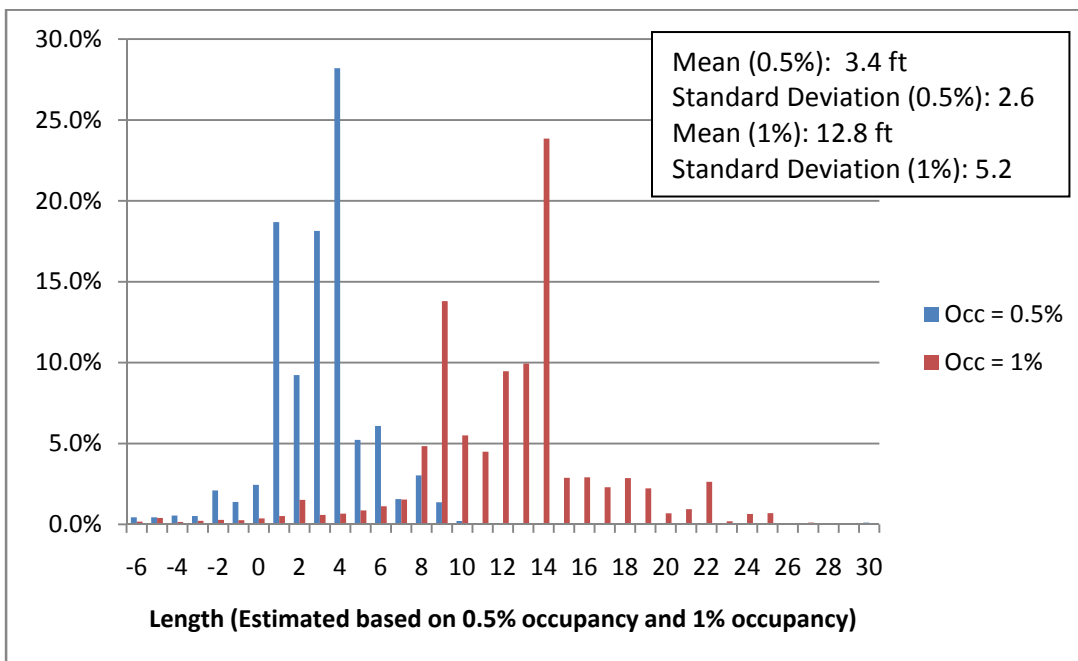


Figure 7 Calculated Vehicle Lengths; Count = 1; Occupancy = 0.5% and 1% (September 2008)

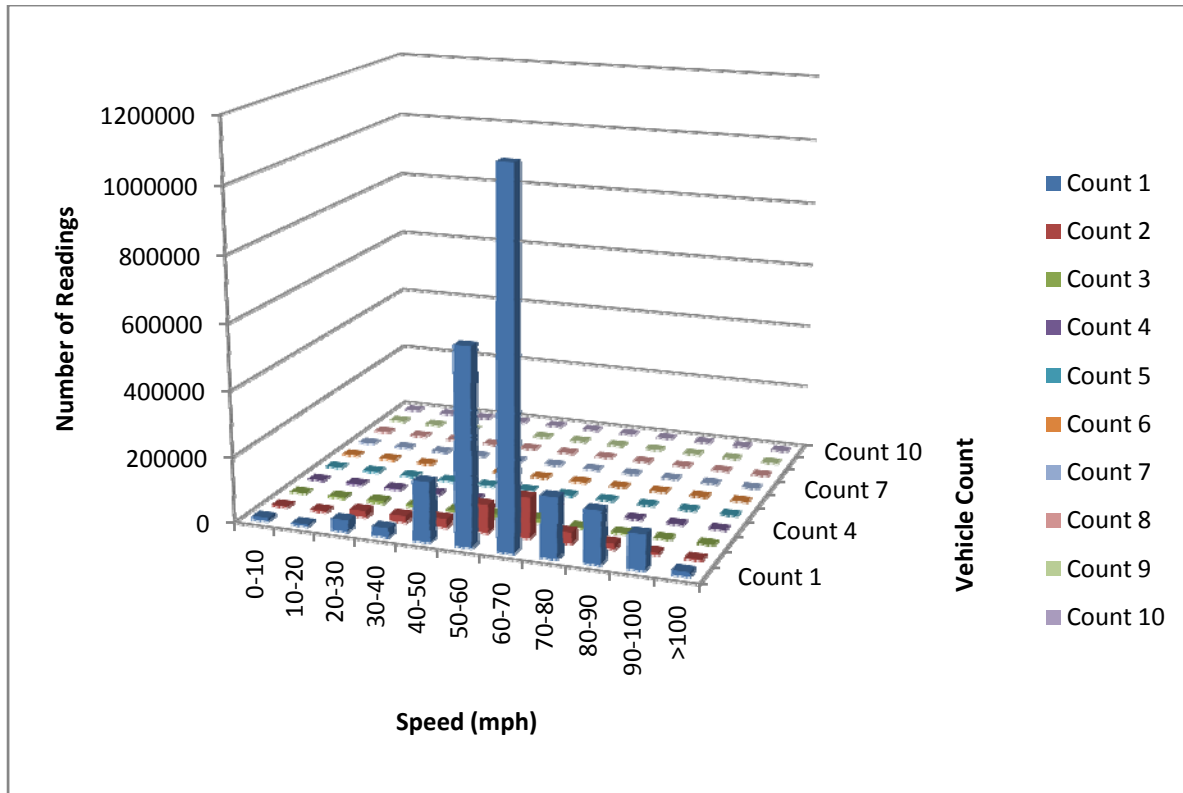


Figure 8 Breakdown of Zero Occupancy Readings by Speed and Count (Sept 2008)

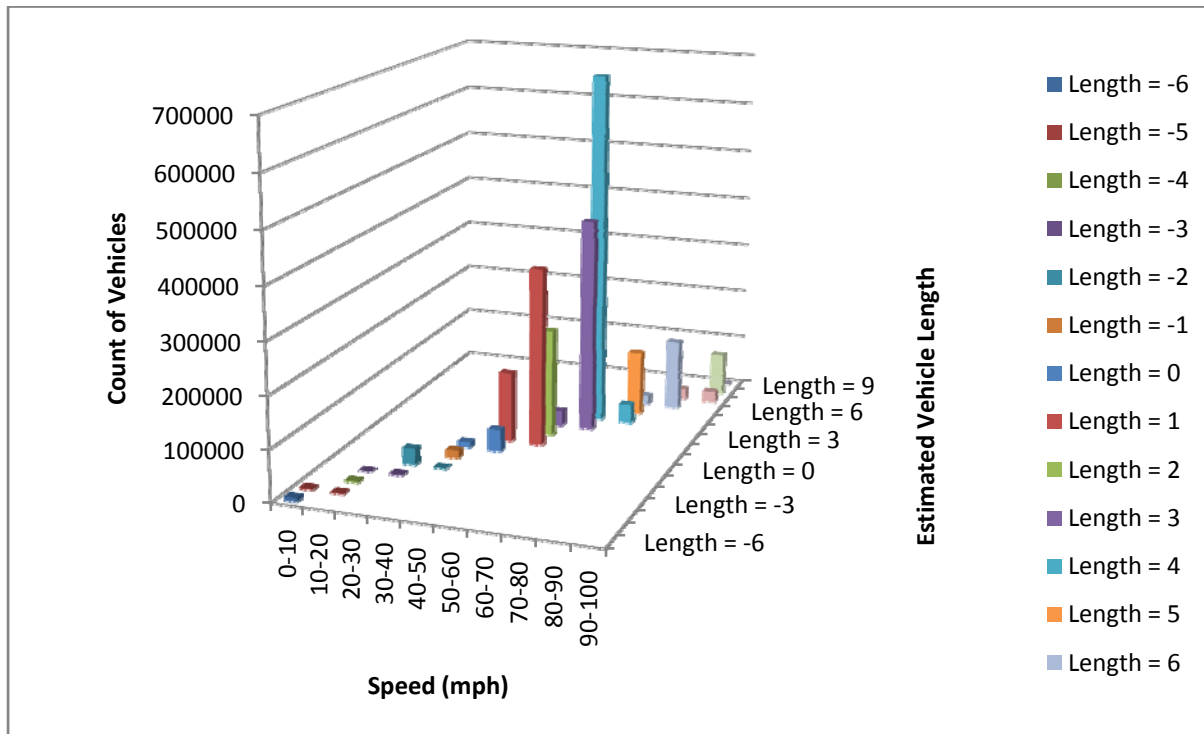


Figure 9 Breakdown of Zero Occupancy Readings by Speed and Estimated Vehicle Length (Count = 1, Occ = 0.5%) (Sept 2008)

Very High Speeds

In September 2008, 0.6% of the readings had speeds in excess of 100 mph.

Table 4 breaks down these high speed readings. We observe that 81.3% of the readings are theoretically valid (speed, volume and occupancy all > 0); while the rest of the high speed readings have a zero occupancy value.

Of all the readings with speed > 100 mph, 80% come from the three detectors at US 26 WB at milepost 73.33 (Jefferson to Sunset WB); the detectorids are 1793, 1794, and 1795 and the station is 1112. 94.2% of the valid readings and 18.6% of the invalid readings. Thus these detectors appear to be the primary source of high speed readings and should be investigated. It is not clear that high speed readings need to be investigated as a separate source of error from zero occupancy readings.

Table 4 Breakdown of Readings with Speed > 100 mph (September 2008)

Description	Count	Percent
Total Speed > 100	295,001	100%
“Valid” readings (vol >0, occ > 0, speed > 100)	239,822	81.3%
Zero Occupancy Readings (occ=0, vol > 0, speed > 100)	55,174	18.7%
Readings from station 1112	236039 225,800 (valid) 10,239 (zero occ)	80% 94.2% of “valid” readings 18.6 % of zero occupancy readings

Low Overnight Speeds

During the nighttime hours, traffic flow is low and speeds are expected to be high. However, many freeway loop detectors in the Portland region report suspiciously high rates of low speeds during the overnight time period. For the purposes of the analysis in this report, we define the overnight time period as being midnight-4AM and we define a “low” speed as a speed of less than 40 mph. Table 5 shows a list of all detectors that had more than 50% of their speed readings less than 40 mph for the overnight hours (12-4AM) in September 2008. For the rest of this section, we refer to the detectors listed in Table 5 as “low-speed detectors”.

We note that there are two detectors for which every speed reading in September 2008 was either a zero speed or a null speed. These two detectors are listed in Table 6, but are excluded from Table 5 and the analysis in the rest of this section.

Table 5 Detectors with High Rates of Low Overnight Speeds (September 2008)

Highway	Milepost	Lane	Location	Detector Id	Percent of Speed Readings < 40 mph; 12-4AM
I-5 SB	304.85	1	Portland Blvd SB	1275	85.6%
I-205 NB	21.12	1	I-205 NB at Glisan	1949	62.2%
I-205 SB	16.24	2	Johnson Creek SB	1731	58.4%
I-205 SB	10.24	1	Park Place SB	1754	57.6%
I-205 NB	21.12	2	I-205 NB at Glisan	1950	57.5%
I-205 SB	16.24	1	Johnson Creek SB	1730	52.3%
I-84 WB	2.1	2	33 rd WB	1481	51.4%
I-205 SB	10.24	3	Park Place SB	1756	51.0%

Table 6 Detectors Reporting Zero or Null Speed Readings for September 2008

Highway	Milepost	Lane	Location	Detector Id	Percent of Speed Readings < 40 mph; 12-4AM
I-5 NB	283.93	3	Wilsonville to I-5 NB	1886	95.9%
US 26 WB	71.07	3	Skyline Rd WB	1695	92.8%

Figure 10 and Figure 11 show breakdowns of speed readings by volume for the “low-speed” detectors listed in Table 5. Figure 10 shows the percent of readings for low-speed detectors that fall into each speed range for counts between 1 and 5. For example, of the overnight readings for the selected detectors 34% of the readings showing a count of 1 also had a zero speed, while 48% of the readings with count of 1 had a speed reading between 1 mph and 10 mph. From Figure 10, we see that zero speeds and speeds between 1 and 10 mph dominate the overnight speed readings for the low-speed detectors. Figure 11 is similar to Figure 10 except that the zero speed and 1 to 10 mph speed ranges have been eliminated. An interesting pattern can be observed. When count = 1, there are a limited

number of speed readings in the range 10-50 mph; for count = 2, low speeds in the range 30-40 mph dominate, for count 3, the range 20-40 mph dominates and for counts of 4 and 5, speeds in the range or 10-30 mph dominate. Thus, there appears to be an association between count and low speed readings. Such an association could be explained by the averaging of zero speed readings. For example, the average of one 70 mph speed and one 0 speed would be 35 mph – right in the center of the 30-40 mph range observed for count = 2. Further, the average of a 70 mph speed and two zero speeds is about 23 mph. One possible explanation for the low speed readings is the averaging in of incorrect zero speed readings. In addition, previous work at PSU has found an association between high percents of low overnight speed readings and the type of loop amplifier card installed at the detector station. Upgrading the loop amplifier card appeared to reduce, but not eliminate, the low overnight speed problem.

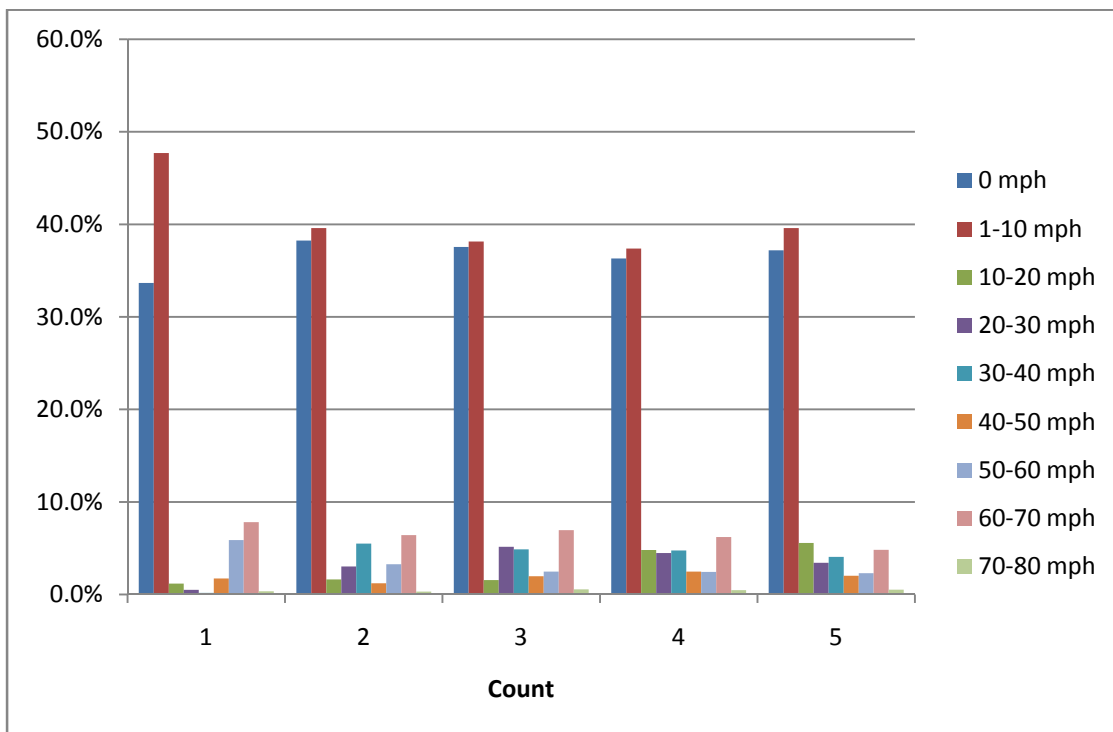


Figure 10 Breakdown of Overnight Speed Readings for Low-Speed Detectors by Volume (Sept 2008)

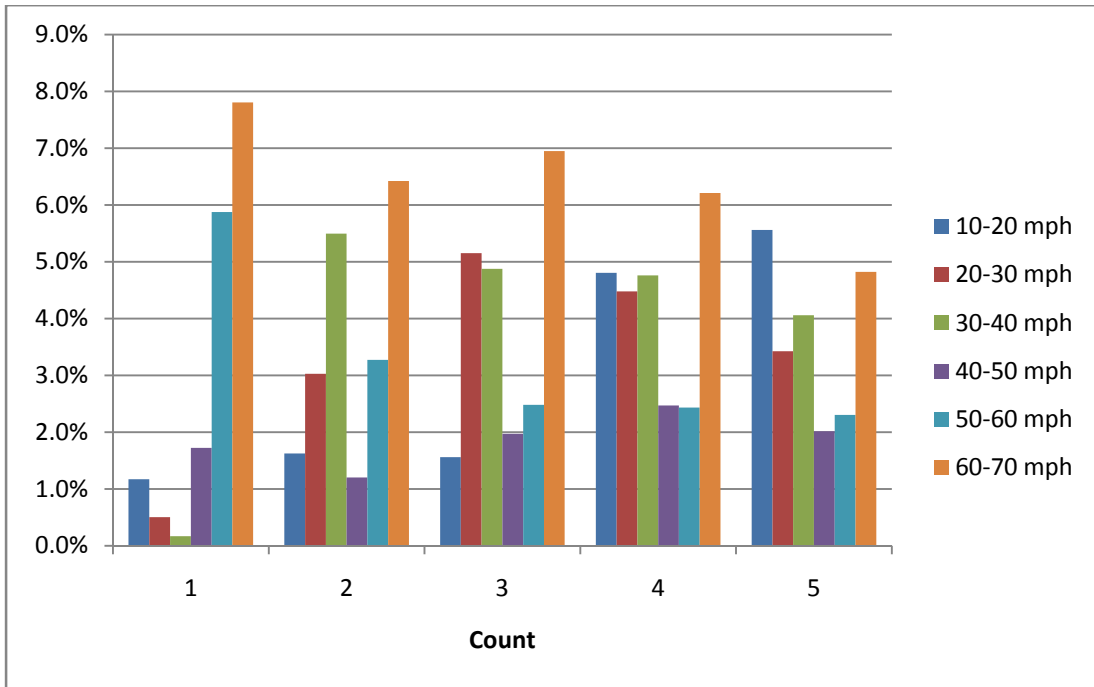


Figure 11 Zoomed In Breakdown of Speed Readings for Low-Speed Detectors (Sept 2008)

Derivations

This section contains the derivations of formulas used in previous sections.

The general formula relating occupancy to count, speed and vehicle length is below:

$$\text{Seconds of Occupancy} = (\text{count} * (\text{length})) / \text{speed} \quad (1D)$$

We define a set of variables to be used in the derivation along with specifying their units.

O = occupancy (pct)

O(s) = occupancy (seconds)

L = length (ft)

S = Speed (mph)

C = count

We proceed to derive a detailed formula for occupancy percent as a function of count, length and speed and then solve for length.

$$O(s) = (C*(L+6))/S \quad (2D) \text{ Eqn 1D, plus add 6 to vehicle length because we assume 6 foot loops}$$

$$O = ((C*(L+6))/S)*(1/20) \quad (3D) \text{ Convert to occupancy as a percentage (assume 20 sec readings)}$$

$$O = ((C*(L+6))/S)*(1/20)*(1/5280) \quad (4D) \text{ Convert miles (in mph) to feet to match vehicle length}$$

$$O = ((C*(L+6))/S)*(1/20)*(3600/5280) \quad (5D) \text{ Convert hours (in mph) to seconds (to match occupancy)}$$

$$O = ((C*(L+6))/S)*(0.0341) \quad (6D) \text{ Calculate constant}$$

Solve for Length:

$$((C*(L+6))/S)*(1/20)*(3600/5280) = O$$

$$((C*(L+6))/S) = O * (5280/3600) * 20$$

$$C*(L+6) = O * (5280/3600) * 20 * S$$

$$L+6 = (O * (5280/3600) * 20 * S)/C$$

$$L = ((O * (5280/3600) * 20 * S)/C) - 6$$

$$L = ((O * 29.3 * S)/C) - 6$$

If $C = 1$

$$L = (O * S * 29.3) - 6$$

Assume $O = 0.005$

$$L = (.1465 * S) - 6$$

Solve for Occupancy assuming a vehicle length.

$$L = (O * S * 29.3) - 6$$

$$(O * S * 29.3) - 6 = L$$

$$O = (L + 6)/(S * 29.3)$$