# ANALYSIS OF QUALITY METRICS IN THE I-95 VEHICLE PROBE PROJECT

**Stanley E. Young. P.E., Ph.D.**
**Research Engineer, University of Maryland**
**2200 Technology Ventures Building**
**University of Maryland, College Park, MD 20742-3021**
**+1 301 403 4593, seyoung@umd.edu**

**Nayel Urena Serulle, Ph.D. Student**
**Graduate Research Assistant**
**2200 Technology Ventures Building**
**University of Maryland, College Park, MD 20742-3021**
**+1 301 403 4623, seyoung@umd.edu**

## ABSTRACT

The I-95 Vehicle Probe project (VPP) has been in production since July 2008 providing estimates of speed and travel time on network of roadways that has expanded to over 5,000 freeway miles. Each estimate of speed is accompanied by two metrics that provide an indication of the level of confidence and/or quality of the traffic data: Score and C-value. This paper statistically characterizes the distribution of Score and C-Value in the VPP data archive in order to better understand the quality of the VPP and reveal any improvements over the life of the project. Both Score and C-Value are tested to determine their effectiveness as a data filter to improve real-time traffic data quality.

## INTRODUCTION

The I-95 Vehicle Probe project (VPP) has been in production since July 2008 providing estimates of speed and travel time on network of roadways that has expanded to over 5,000 freeway miles. Each estimate of speed is accompanied by a quality metric of "10", "20", or "30", designated the 'Score' attribute within the VPP data feed. The intent of the Score attribute is to provide an indication of confidence in the reported data. The three discrete values correspond to [1]:

"30" – high confidence, based on real-time time data for that specific segment
"20" – medium confidence, based on real-time data across multiple segments and/or based on a combination of expected and real-time data
"10" – low confidence, based primarily on historical data

The percentage of data with Score equal to 30 is an indication of the amount of real-time data (as opposed to reliance on historical data) in the VPP. Tracking the proportion of Score equal 30 over time indicates increasing or decreasing data quality as the VPP progresses.

In December of 2009, the VPP began including a C-Value, or confidence value in addition to the Score in the VPP data feed. The C-Value is separate from the Score attribute. The C-Value is designed to provide supplemental information to the Score to best identify the confidence of the data being sent by INRIX [1]. C-Value is only defined when Score equals 30. When Score is equal to 10 or 20, the C-Value is not reported. The C-Value ranges from 0 to 100. A C-Value equal to 100 represents the combination of (1) high data density, (2) the current data is very consistent with the data over the past 45 minutes, and (3) the current data is very consistent with the historical data. Data with a C-Value equal to 100 is referred to as high confidence data in this paper. As each of the three criteria degrade, the C-Value will decrease until it reaches a minimum C-Value of zero, corresponding to the lowest possible confidence in the estimate of traffic conditions.

The statistical analysis on Score and C-Value from the VPP archives are restricted to VPP data from the state of Maryland's freeway network for sake of efficiency. The analysis may be expanded to all states, though past research has indicated that summary results are typically consistent from one state to the next when analyzing overall population trends.

The final analysis determines the utility of filtering data based on the Score and C-Value. A data validation set from VPP validation effort in Pennsylvania is used to show the impact of omitting data whose Score and C-Value falls below specified thresholds.

## RESULTS

Figure 1 illustrates the overall distribution of Score for each year in the state of Maryland. The percentage of Score equal to 30 has increased steadily from 55 percent in 2008 to 70 percent in 2010. Likewise the percent of Score equal to 20 has decreased from 23 percent in 2008 to 12 percent in 2010 and the percent of Score equal to 10 has also decreased, but to a lesser extent.

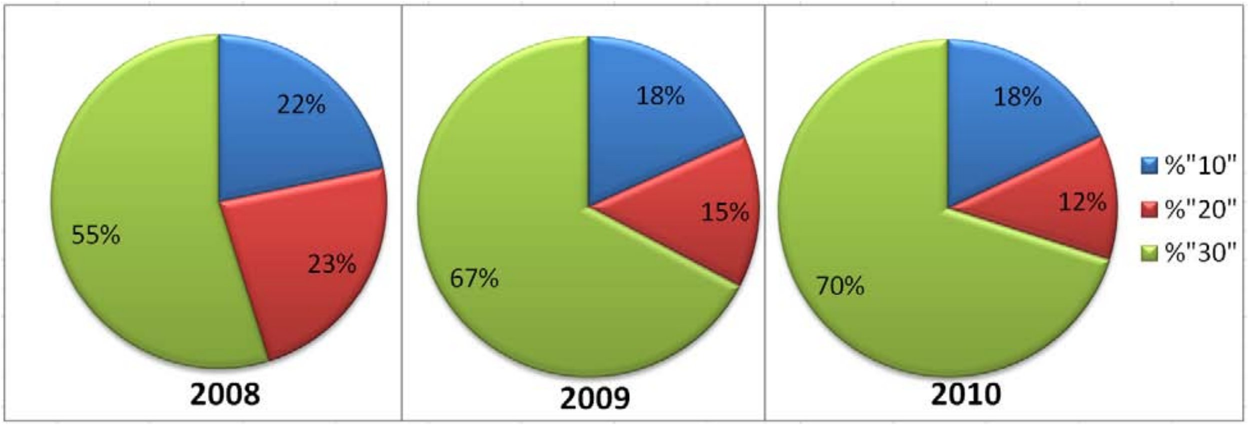



*Figure 1*. Yearly Average Percentage of Scores equal to "10", "20", and "30" in the state of Maryland.

Figure 2 shows in more detail the percentage of Scores equal to 10, 20, and 30 as a function of the time of day based on data from October 2008 to October 2010. Figure 2 follows expected volume trends for most freeways. Periods of low volume, from roughly 10 PM until 5 AM, correspond to lower percentage of Score equal 30, and periods of higher volume, between the hours of 5 AM to 10 PM, correspond to a higher likelihood of a Score equal to 30. As the VPP is a probe-based system in which data is sampled from the traffic stream, this trend agrees with expected operation. During daytime hours when data density is insufficient to support real-time data (as evidenced by Score less than 30), the system uses historical data based on time of day and day of week and reports a Score of 20 as illustrated. During nighttime hours if data density is insufficinet, a free-flow speed (also referred to as a reference speed is reported) and Score is set equal to 10 as illustrated.

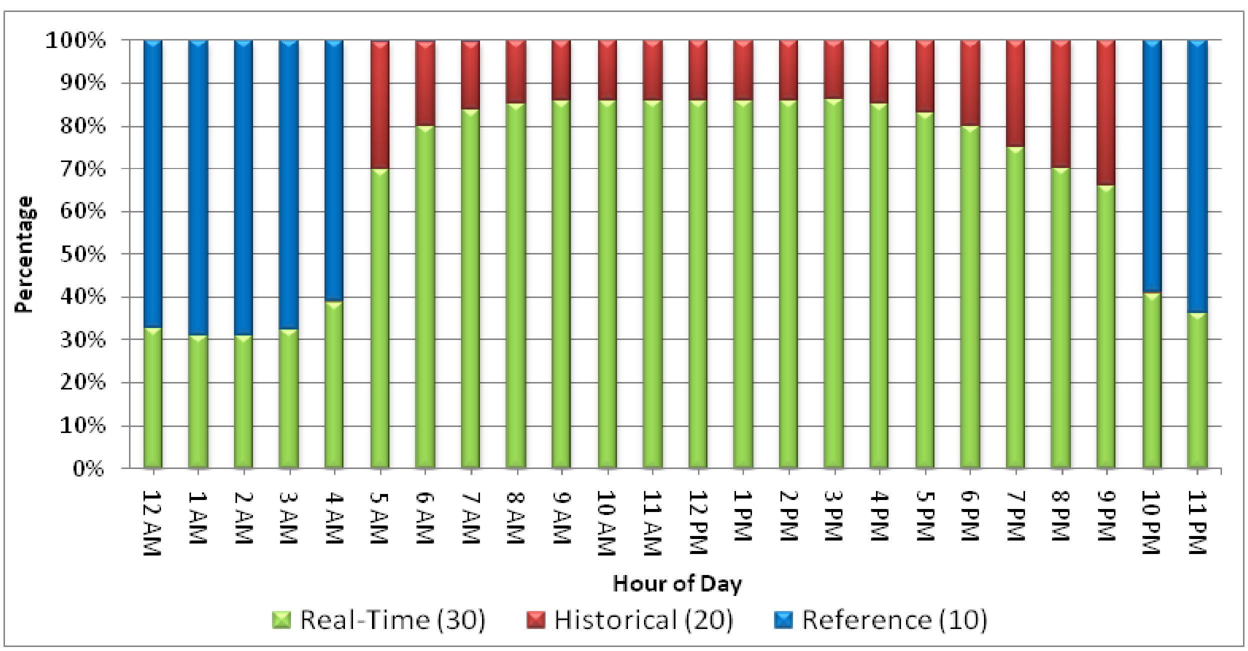



*Figure 2*. Hourly Average Percentage of all Scores from October 2008 to October 2010 for the State of Maryland.

Figure 3 illustrates the Score and C-Value distribution from April 2010 to October 2010 as obtained from the Center for Advanced Transportation Technology Laboratory (CATT Lab) VPP archive for the state of Maryland. 71.3 percent of all data has a Score of "30", as shown in the leftmost pie-chart in Figure 1. Of the 71.3 percent with Score equal 30, 76.6 percent of the data has C-Value equal to 100, as shown in the middle pie-chart in Figure 3. Recall that C-Value is only defined when Score equals "30", when Score is less than "30", C-Value is not reported. Roughly 55 percent of all the data (71.3% * 76.6% = 54.7 %) has a Score of 30 and a C-Value of 100, the highest indicator of data quality. Of the remaining 23.4 percent of data when Score = 30, the breakout into the various ranges of C-Value is shown in the furthest right pie-chart in Figure 3.

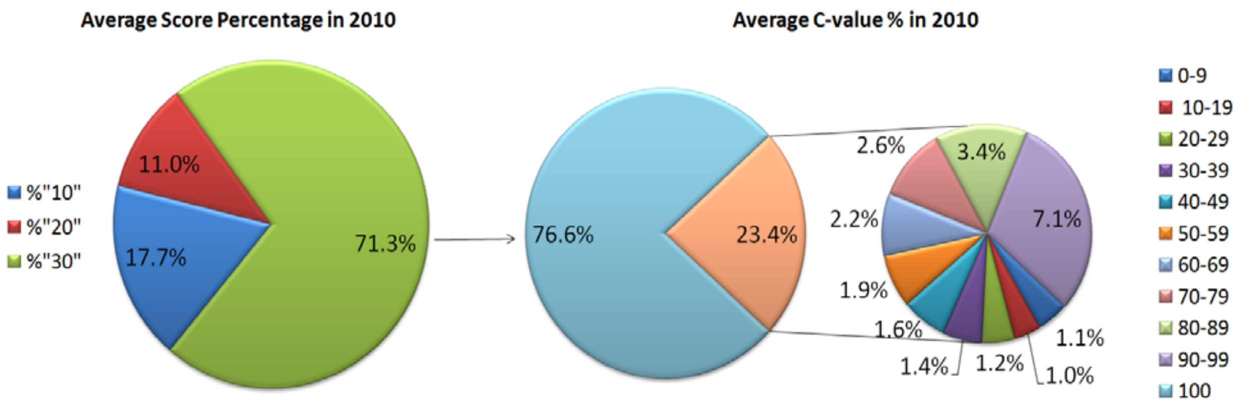



*Figure 3.* Score and C-value distribution for the state of Maryland from April to October 2010.

Figure 4 depicts the percentage of C-Value equal to 100 (given that Score = 30) on an hourly basis for VPP in the State of Maryland from April 2010 through October 2010. The highest percent C-Value is obtained between 5AM and 10PM (referred to as daytime hours), during which traffic volumes are typically higher. Also, in recent discussion with INRIX Corporation, the vendor providing data to the VPP, it was indicated that the algorithm that generates the VPP data operates differently at night (10 PM to 5 AM) than during daytime hours. The algorithm is more stringent at night, requiring more data to achieve the same level of confidence as during the day. The abrupt changes in the graph in Figure 4 at 5 AM and 10 PM illustrate such functioning of the algorithm. On average, 85 percent of the VPP data gathered during daytime hours has a C-Value equal to 100 when Score equals 30.

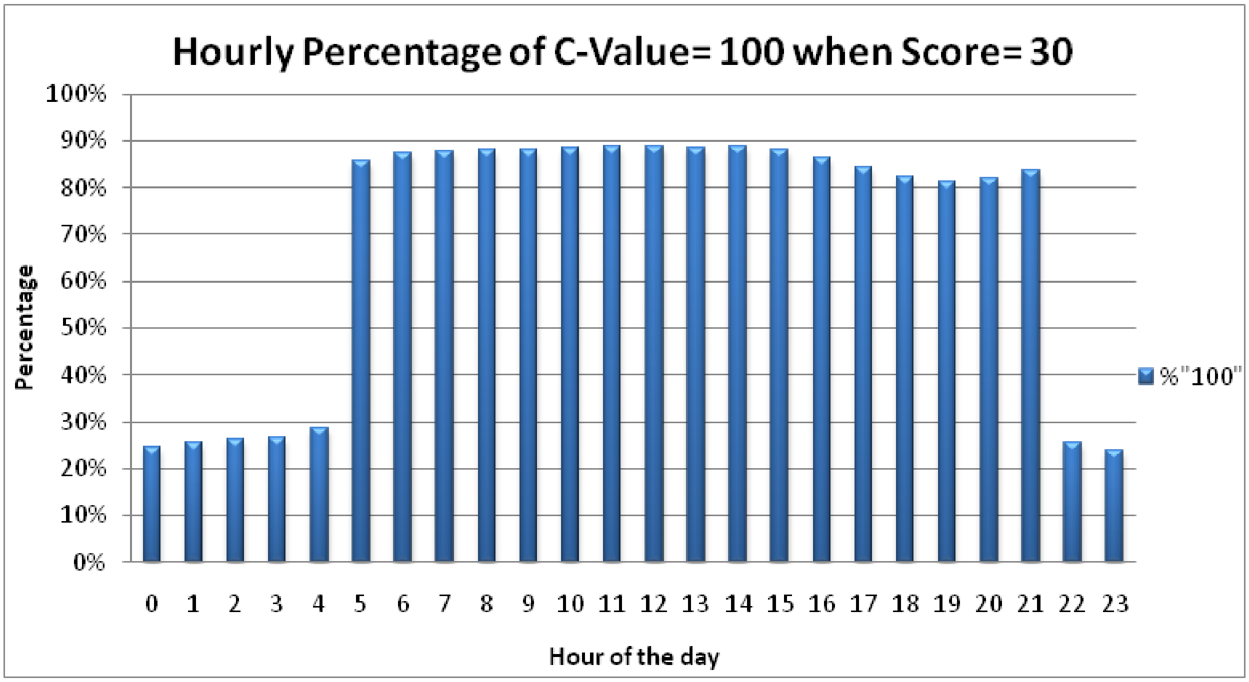



*Figure 4.* Hourly Performace of the C-value for data with SCORE = 30. Souce: MD VPP data from April to October 2010.

Figure 5 illustrates a percentage of data in which C-Value = 100 for each month of data collection. On average, 75 percent of the data gathered in each month has C-Value equal to 100 when Score equals 30. The uniform distribution across each month suggests that seasonal characteristics have little effect on the determination of C-Value. The analysis only takes into consideration data points with Score equal to 30.

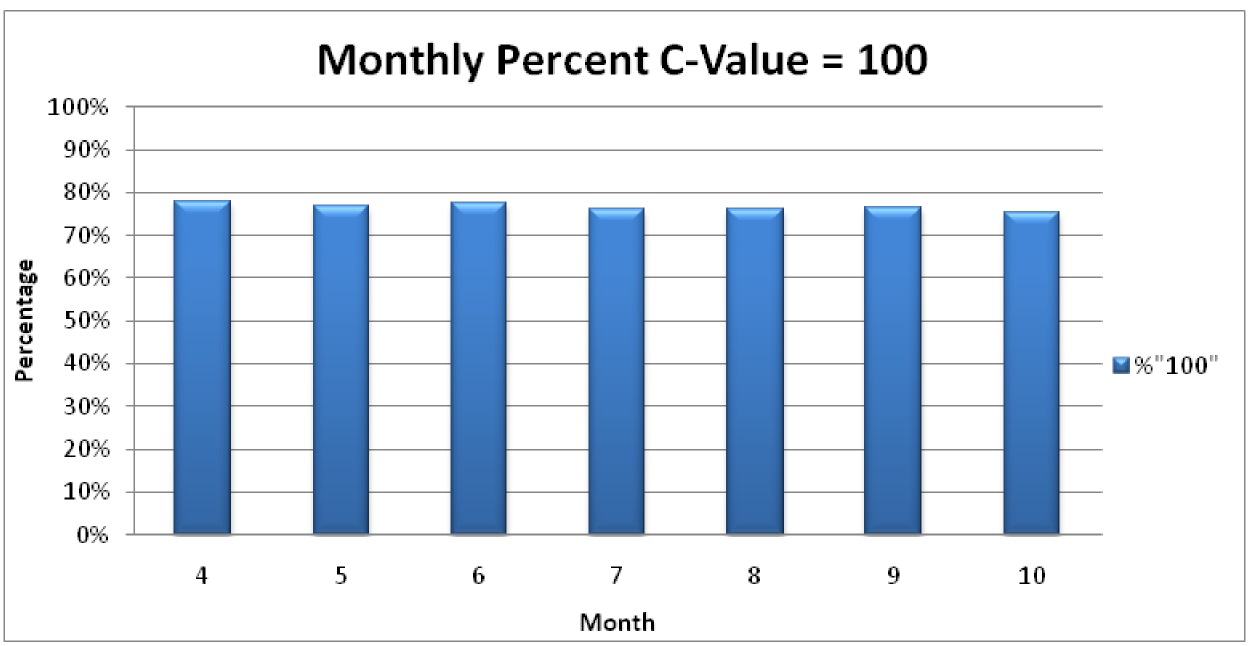



*Figure 5.* Monthly Performace of the C-Value. Souce: MD VPP data from April to October 2010.

## SCORE AND C-VALUE AS DATA FILTERS

The objective of this analysis is to characterize the effectiveness of Score and C-Value as a filter to distinguish good quality data from outliers. This analysis investigated the impact to data quality when Score and C-Value are used to identify and remove outliers. Validation data collected in Pennsylvania (PA) on August 20, 2010 is used to illustrate the analysis. Ground truth data obtained using Bluetooth™ traffic monitoring (BTM) technology and its corresponding VPP data are shown in Figure 6. In this figure, the VPP data points within the grey area represent outliers that would be desirable to filter because they disagree with the BTM ground truth data. These points are referred to as outliers in the following discussion.

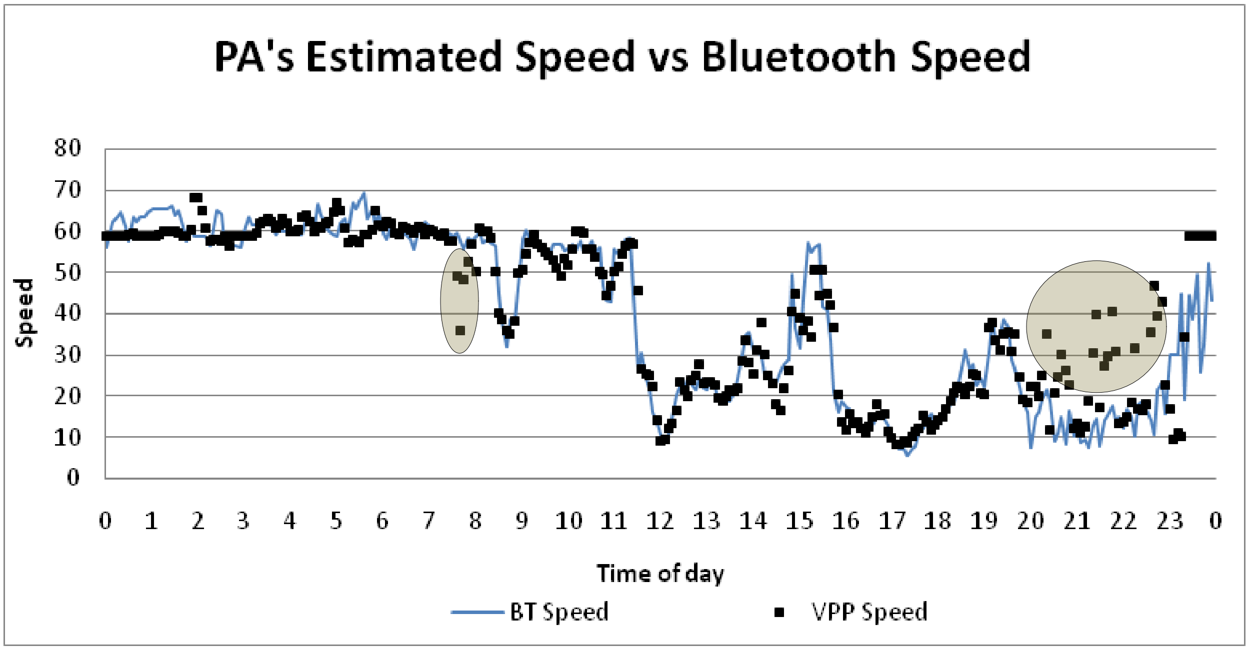
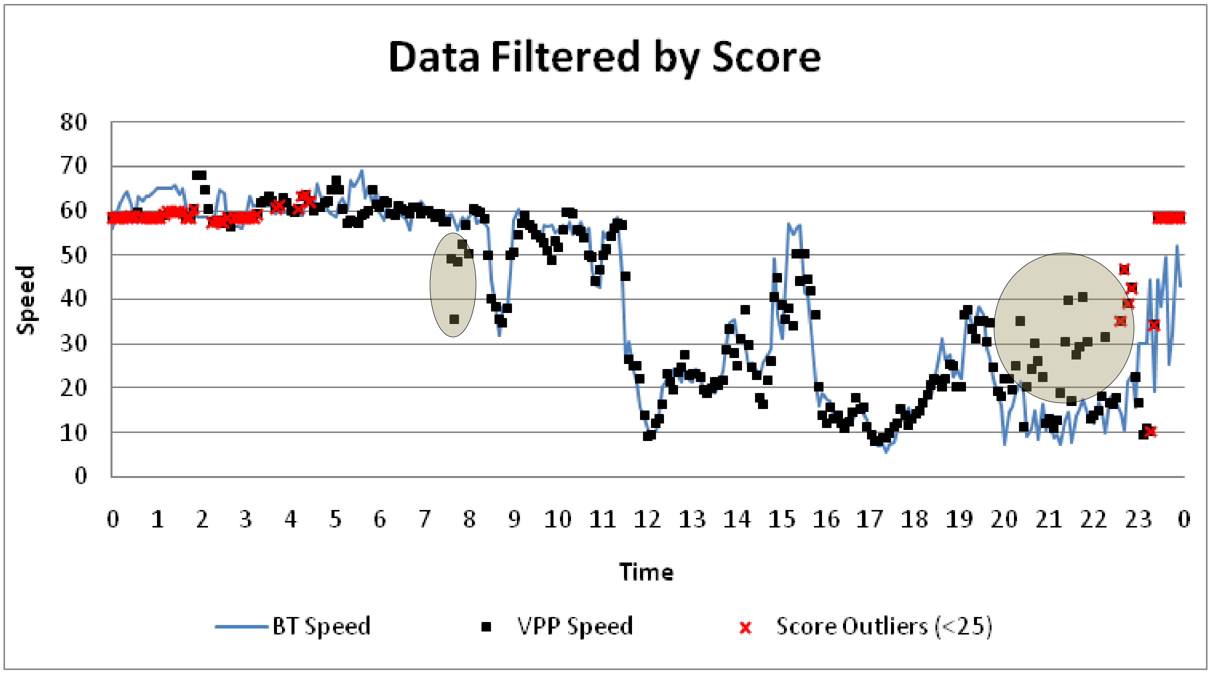
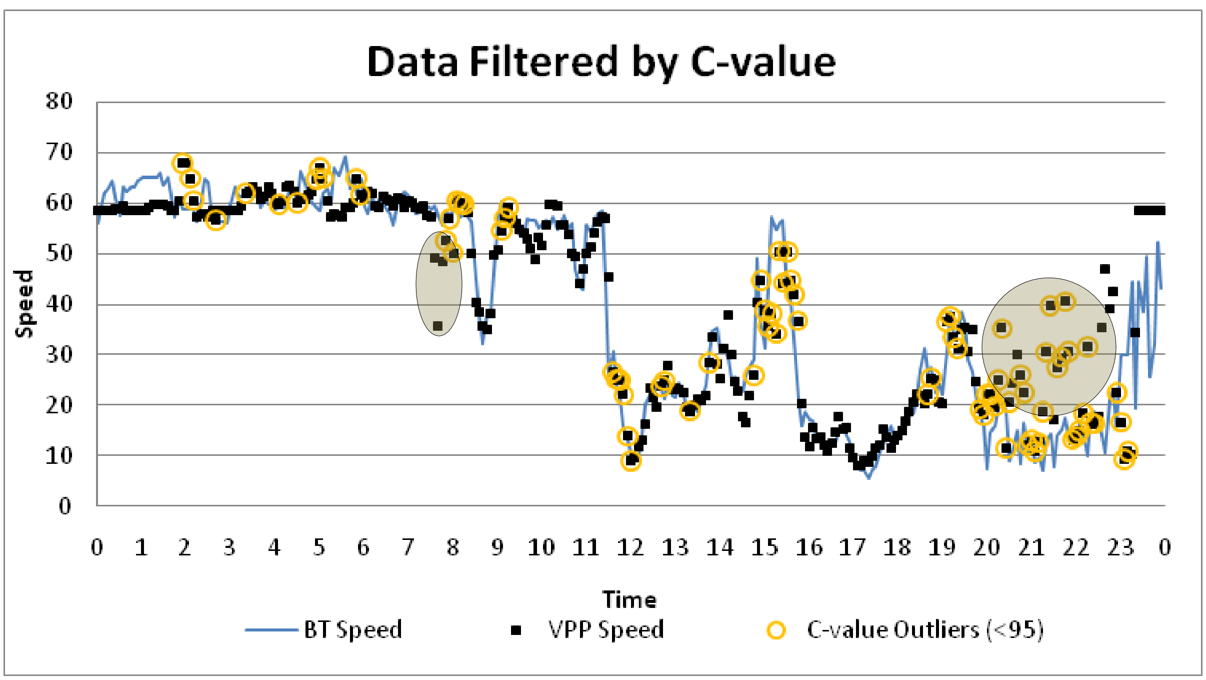



*Figure 6.* Bluetooth data (Ground Truth) and VPP data for a stretch of I-95 in PA on August 20, 2010.

Figures 7a and 7b illustrate the assessment of VPP data based on Score and C-Value. Filtering any data with Score less than 25 eliminates the data marked with red x's, which include much of the nighttime data (10 PM to 5 AM) when volumes are low, see Figure 7a. A few of the suspect outliers (shown in gray shaded areas) near 10 PM are eliminated, but none of the suspected outliers are filtered during daytime hours. Aggregating VPP data from short base-level segments into more meaningful routes and averaging across multiple time periods may result in continuous rather than discrete Score values, thus a threshold of 25 indicates more than half of the base level data had a Score of 30. Figure 7b illustrates the result of additionally filtering data when C-Value is less than 95.




a)




b)

*Figure 7.* Data filterd by: a) Score < 25, and b) C-Value < 95. Source: PA's VPP data from August 20, 2010.

Figure 8 provides a more detailed review of the performance of the filtering based on C-Value < 95. The gray regions, labeled A through E in Figure 8, highlight groups of data of interest. Data points in region E can be visually classified as outlier data which are being filtered by the C-Value criteria. On the other hand, regions A and C reflect groups of outliers not filtered. Finally, regions B and D stress a cluster of "good" data points that are being filtered using the C-Value criteria.

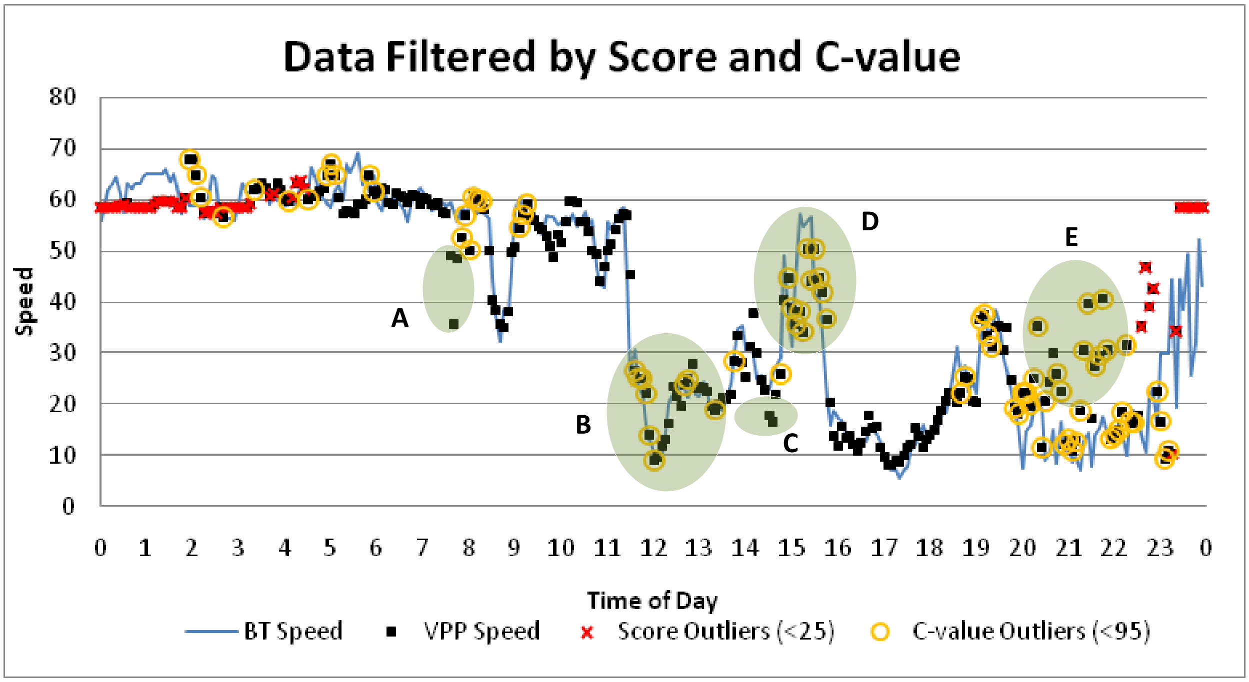



*Figure 8.* C-Value filter process assestment.

Table 1 illustrates the impact on Average Absolute Speed Error (AASE) resulting from filtering the entire Pennsylvania validation data set by Score and C-Value as illustrated in Figures 7a and 7b. The analysis presented in Table 1 compiles Score data from August 19, 2010 to August 29, 2010 (day and night) and C-value data for August 20, 23, and 25, 2010 (daytime only - recall that Score equal 30 is less common during nightime hours, see Figure 2). This discrepancy in dates, and therefore number of observations, is due to lack of available C-value data for the complete time period and segments.

Eliminating VPP data with Score less than 25 removes 22.5 percent of the data overall. The 0-30 mile per hour (mph) speed bin exhibits the most improvement where the AASE is reduced from 4.7 to 3.8 mph, a significant reduction. The AASE of the 112 data points with Score < 25 in this lowest speed bin was 23.12 mph. The latter indicates that filtering based on Score < 25 removes data points containing significant error.

Eliminating data with C-Value less than 95 removes 11.3 percent of the data. None of the speed bins showed significant improvement in AASE when filtered by C-Value. In contrast to filtering by Score, the data points in the 0-30 mph speed bin removed by filtering based on C-Value had an average AASE of 4.53 indicating that filtering based on C-Value had no appreciable effect.

*Table 1.* Effect of Filtering Based on Score and C-value by Speed Bin.

| | Filter based on Score < 25 | | | | Filter based on C-value < 95 | | | |
|---|---|---|---|---|---|---|---|---|
| **Speed Bin** | **Data Set** | **AASE** | **# Observ.** | **%** | **Data Set** | **AASE** | **# Observ.** | **%** |
| **0-30** | Complete | 4.70 | 2414 | 100% | Complete | 3.33 | 796 | 100% |
| | >25 Only | 3.80 | 2302 | 95.4% | >95 Only | 3.21 | 721 | 90.6% |
| | < 25 Only | 23.12 | 112 | 4.6% | < 95 Only | 4.53 | 75 | 9.4% |
| **30-45** | Complete | 4.19 | 926 | 100% | Complete | 6.57 | 477 | 100% |
| | >25 Only | 4.05 | 857 | 92.5% | >95 Only | 6.41 | 404 | 84.7% |
| | < 25 Only | 5.85 | 69 | 7.5% | < 95 Only | 7.43 | 73 | 15.3% |
| **45-60** | Complete | 2.33 | 6947 | 100% | Complete | 2.07 | 2168 | 100% |
| | >25 Only | 2.52 | 5087 | 73.2% | >95 Only | 2.13 | 1929 | 89.0% |
| | < 25 Only | 1.83 | 1860 | 26.8% | < 95 Only | 1.62 | 239 | 11.0% |
| **>60** | Complete | 2.35 | 1284 | 100% | Complete | 2.23 | 216 | 100% |
| | >25 Only | 2.40 | 719 | 56.0% | >95 Only | 2.29 | 190 | 88.0% |
| | < 25 Only | 2.29 | 565 | 44.0% | < 95 Only | 1.79 | 26 | 12.0% |
| **TOTAL** | Complete | | 11571 | 100% | Complete | | 3657 | 100% |
| | >25 Only | | 8965 | 77.5% | >95 Only | | 3244 | 88.7% |
| | < 25 Only | | 2606 | 22.5% | < 95 Only | | 413 | 11.3% |

The issues discussed from Figure 8 and the data from Table 1 indicate that C-Value does not reliably discern accurate data points from outliers. A more detailed analysis was conducted to confirm this as explained in the next section.

## ASSESSING C-VALUE FILTER USING SEM BAND THRESHOLDS

The standard error of the mean (SEM) band is used during the validation process to place confidence limits on the ground truth data. SEM is based on variance of the ground truth data collected using Bluetooth™ traffic monitoring (BTM) technology. It is calculated using the following formula:

where:

μ is the mean of the BTM data

σ is the standard deviation of the BTM data

#BT is the number of BTM records

All calculations are based on data received in a five minute interval

In order to further analyze the impact of filtering by C-Value, additional criteria were established based on the measured SEM band of the validation data set. Three sets of VPP data were defined based on specified ranges of the SEM band:  quality data points, bad data points (also known as outliers), and indeterminate data points. Quality data includes all VPP data points that fall within five mph of the validation SEM band. Note that quality data are defined as data points that falls within five mph of the SEM band for the respective five minute reporting interval, as well as the succeeding five minute time interval. The latter was added to account for acceptable time lag in traffic reporting. Bad data (or outliers) are defined as VPP data points that fall ten mph or more outside the SEM band for the respective five minute time interval, as well as the succeeding five minute time interval. The third set of data points encompass all remaining data not classified as either quality data or bad data. These data points fall five-ten mph outside the SEM band of the validation data, and are labeled indeterminate data in the tables below. A graphical illustration of the procedure can be found in the appendix.

The C-Value filter was scored in terms of the ability to allow good data points to pass, and its ability to correctly identify bad data points as outliers. This is a classical type I / II classification error analysis where a type I error is defined as a quality data point incorrectly classified as an outlier by the filter; and a type II error is a bad data point that is not classified as an outlier by the C-value filter. Undetermined data points could be classified as either quality data points or as outliers with no penalty. The analysis was conducted for the validation data set during daytime hours (5 AM to 10 PM) in Pennsylvania. The data was collected on six different segments of I-95 for 3 days. Table 2 provides a characterization of the test data set.

*Table 2.* Characterization of Vehicle Probe Data from PA validation data used in filter analysis

| | # Data Points | % |
|---|---|---|
| **Total data** | 3690 | 100% |
| **Data with Score < 25** | 47 | 1.3% |
| **Data with C-Value < 95** | 813 | 22.0% |
| **Quality data (within ±5mph of SEM Band)** | 3391 | 91.9% |
| **Indeterminate data (5 to 10mph from SEM band)** | 211 | 5.7% |
| **Bad data (outside ±10mph of SEM band)** | 88 | 2.4% |

Table 3 shows the results of the C-value filtering process applied to the sample data set from Pennsylvania. Of the 3,391 quality data points, the filter incorrectly identified 704 as outliers, creating a type I error rate of 20.8 percent. Of the 88 bad data points, the filter correctly identified 40 as outliers, and incorrectly classifying the remaining 44 for a type II error rate of 54.5 percent.

*Table 3.* C-Value filtering results.

| Data Category | # Data Points | Filtered by C-Value | % |
|---|---|---|---|
| Quality Data ( < 5 mph of SEM band) | 3391 | 704 | 20.8% |
| 5-10 mph of SEM (Average) | 211 | 69 | 32.7% |
| >10 mph of SEM (Bad) | 88 | 40 | 45.5% |

The results from Table 3 indicate that the C-Value filter eliminates excessive quality data (type I error of 20.8 percent) while only correctly characterizing roughly half of the bad data as outliers. A type I error rate of less than ten percent is desirable, and less than five percent is preferred, though no specifcations for such filtering are specified with the vehicle probe project. One of the factors influencing the C-Value is whether the current data is consistent with data over the last 45 minutes. Any sudden drop in speed, whether as a result of the onset of congestion or as a result of extraneous data reports, may be reflected as a descreased C-Value. As a result many of the type I errors may occur at the onset of congested conditions, making the penalty for a type I error a delay in the detection of traffic slowdowns. Acceptable percentage of type II errors are much larger than type I errors in this application. The approximate 50 percent type II error rate is acceptable, indicating that about half of the bad data points are correctly identified as outliers.

Figure 9 provides a graphical illustration of the C-Value fileter with the defined categories based on the  SEM band of the validation data.

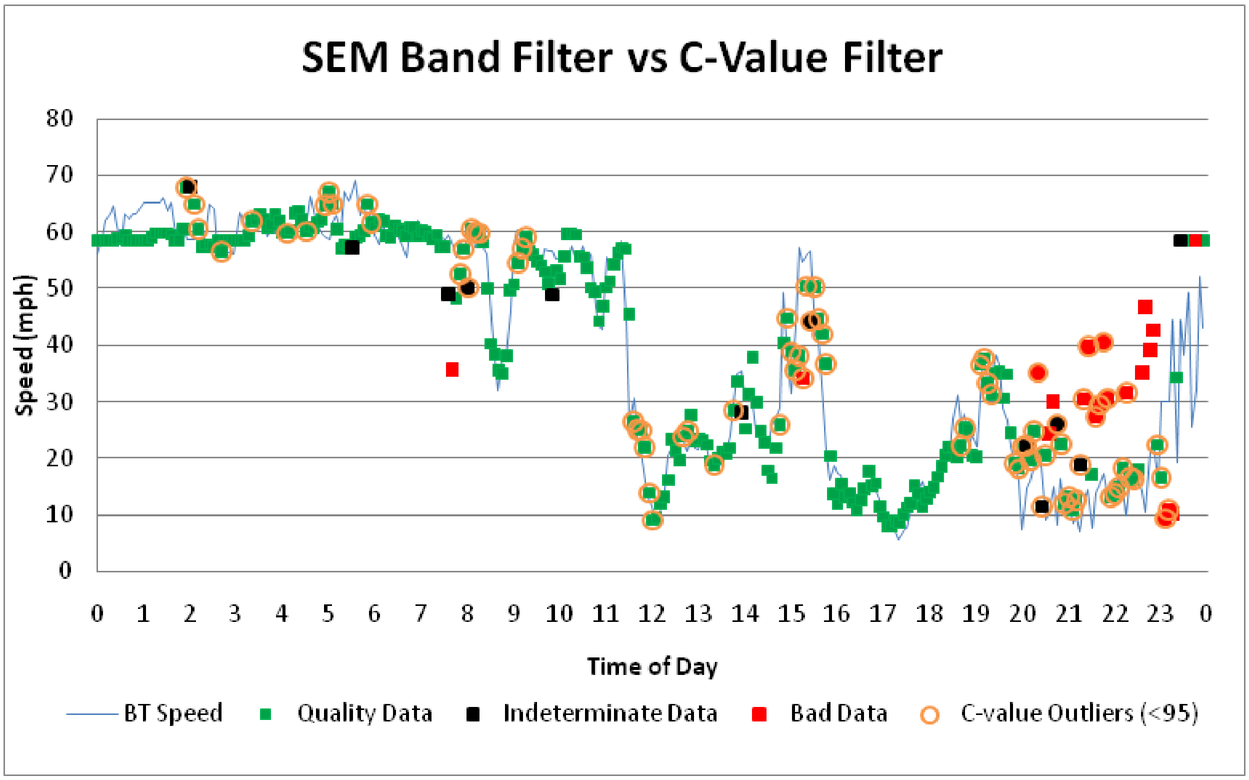



*Figure 9.* Illustration of the C-Value filter performance based on categories defined by the SEM band of the validation data. Source: PA's VPP data from August 20,  2010.

## CONCLUSION AND RECOMMENDATIONS

The VPP includes two metrics to convey anticipated data quality and/or data confidence in the form of the Score and C-Value metrics. Score has been part of the VPP data feed since its initiation in July 2008. The statistical distribution of Score indicates a trend of increased percentage of Score equal 30 starting from 55 percent in 2008 to 70 percent in 2010 within the state of Maryland. The higher percentage of Score equal 30 reflects increasing base data over time resulting in higher quality traffic information. The official validation results have also exhibited more consistent results over time, confirming the trend shown in the Score data.  The percent of real-time data during daytime hours (defined as 5 AM to 10 PM) averages well over

70 percent, while the percent of real-time data at night (defined as 10 PM to 5 AM) averages less than 40 percent.

In December 2009, the VPP began including a C-Value, or confidence value in addition to Score in the data feed. The C-Value is designed to provide supplemental information to the Score to better identify the confidence of the data being sent by INRIX. C-Value is reported only if Score equals 30. Analysis on the distribution of C-Value from April to October 2010 in the state of Maryland revealed that C-Value equal 100 in 76.6 percent of data when Score equal 30. There has been no significant trend (either increasing or decreasing) in this percentage over the seven months of the analysis. Roughly 85 percent of the data gathered during daytime hours (5 AM to 10 PM) has a C-Value equal 100 when Score equal 30.

The impact of filtering based on Score and C-Value were assessed using the monthly validation methodology to determine the improvement in the primary quality metrics of Average Absolute Speed Error (AASE). Eliminating VPP data with Score less than 25 removes 22.5 percent of the data overall, the majority of which were data points during nighttime hours when volumes were low and the corresponding data density was insufficient to support real-time reporting. The impact of removing data points with Score less than 25 was significant. The AASE in the 0-30 miles per hour (mph) speed bin was reduced from 4.7 to 3.8 mph. The AASE of the eliminated data points in this band was 23.1 mph. These results indicate that Score is a reliable indicator of data quality. Although a Score equal 30 does not quarantee data quality, a Score less than 30 does indicate that data density is insufficient to detect major slowdowns. Based on the results, it is recommended to filter VPP by Score for any applications sensitive to real-time data availability.

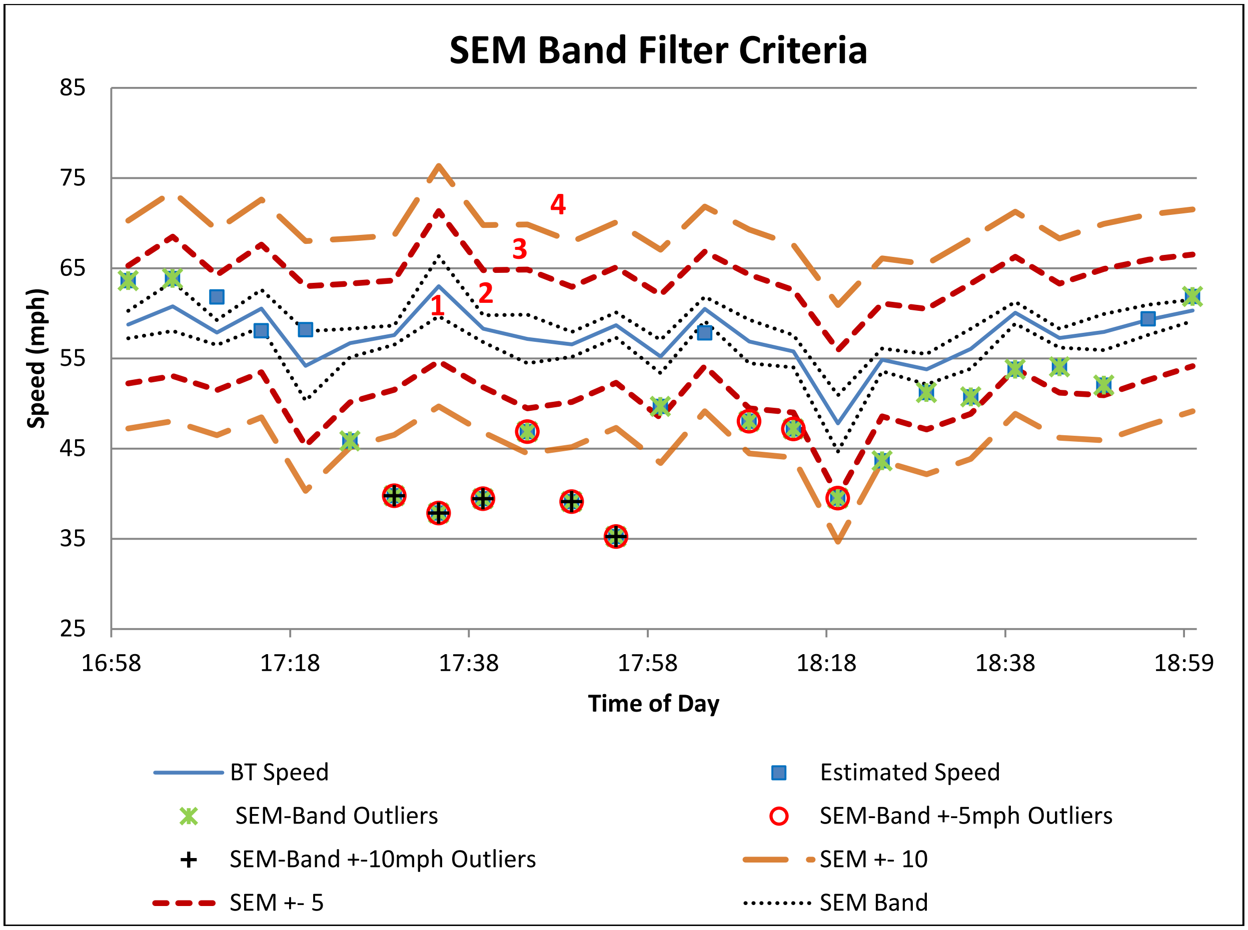
The results of filtering based on C-Value were less conclusive. Note that any filtering by C-Value implies pre-filtering based on Score, because C-Value is only reported if Score equal 30. As such the benefits of filtering based on Score are inherited, and only the additional benefits of excluding additional data based on a C-Value threshold are under consideration. The analysis indicated that filtering based on a C-Value threshold of 95 did remove additional outliers, but at the expense of eliminating a significant percentage of quality data. The C-Value filter correctly identified roughly 45 percent of the bad data points but at the expense of losing approximately 20 percent of quality data. Because C-Value is sensitive to short-term fluctuations in traffic data, many of the quality data points that are eliminated (type I classification errors) occur at the onset of congestion, thus increasing the penalty of type I errors. It is recommended not to filter by C-Value at this time.

The analysis methodology based on the SEM band provides an effective framework to analyze the performance of filters based on standard error classifications. Additional research is needed to determine acceptable errors rates for filters, penalties for misclassification, and other uses of C-Value apart from data filtering.

## REFERENCE

[1] I95 Vehicle Probe: I95 Interface Guide, updated June 2010, version 3.3

## APPENDIX




*Appendix 1.* SEM Band filter process assestment and zones.

Four zones were created for this analysis. Zone 1 accounts for all data points that are within the SEM band range. Zone 2 clusters all data points within the "SEM band ± five" range, excluding zone one. Zone 3 typifies data points within the "SEM band ± ten" range, excluding zones one and two. Zone 4 accounts for everything outside zone three. These zones determine the quality of the data based on speed, evaluating the difference between the data point and the ground truth. Based on these zones, three types of data are proposed:

1) Quality data, if they are within the "SEM band ± five" range (Zone 1 and 2).
2) Indeterminate data, if they are between "SEM band ± five" and "SEM band ± ten" range (Zone 3).
3) Bad data, if they are outside the "SEM band ± ten" range (Zone 4).