10-1-2019

Biking and Walking Quality Counts: Using “BikePed Portal” counts to developData Quality checks

Nathan McNeil
Portland State University

Kristin Tufte
Portland State University

Recommended Citation
McNeil, Nathan, and Kristin Tufte, Bike and Walking Quality Counts: Using Bike Ped Portal Count to Develop Data Quality Checks. NITC-RR-1026. Portland, OR: Transportation Research and Education Center (TREC), 2018. https://doi.org/10.15760/trec.236

NITC
NATIONAL INSTITUTE for TRANSPORTATION and COMMUNITIES

Final Report 1026
October 2019

Biking and Walking Quality Counts: Using "BikePed Portal" Counts to Develop Data Quality Checks

Nathan McNeil, M.U.R.P.
Kristin Tufte, Ph.D.
Tammy Lee, Ph.D.
Krista Nordback, Ph.D.

Portland State University

The University of North Carolina Highway Safety Research Center

BIKING AND WALKING QUALITY COUNTS
USING "BIKEPED PORTAL" COUNTS TO DEVELOP DATA QUALITY CHECKS

Final Report

NITC-RR-1026

by

Nathan McNeil
Kristin Tufte
Tammy Lee
Portland State University

Krista Nordback
University of North Carolina
Highway Safety Research Center

for

National Institute for Transportation and Communities (NITC)
P.O. Box 751
Portland, OR 97207

October 2019

Technical Report Documentation Page

1. Report No. NITC-RR-1026	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Biking and Walking Quality Counts: Using "BikePed Portal" counts to develop Data Quality checks		5. Report Date October 2019 6. Performing Organization Code
7. Author(s) Nathan McNeil, and Kristin Tufte		8. Performing Organization Report No.
9. Performing Organization Name and Address Portland State University, PO Box 751, Portland, Oregon 97207		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address National Institute for Transportation and Communities (NITC) P.O. Box 751 Portland, Oregon 97207		13. Type of Report and Period Covered 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Cycling and walking are sustainable modes of transportation which improve community livability, but these modes have not been studied with the quantitative rigor applied to motor vehicle travel. This work aims to change that by improving bicycle and pedestrian traffic monitoring data quality, including how erroneous data can most accurately and efficiently be identified through automated processes. The research approach analyzes continuous bicycle and pedestrian count data stored in BikePed Portal, a data archive of bicycle and pedestrian count data.

A primary goal of the project is to explore tests that could help to identify aberrant and/or erroneous data. A key method deployed in pursuit of this exploration was to comb through a selection of count data (generally – continuous counter locations from 2015 to 2016 with at least 30 days of counts) to identify expected count ranges and patterns, overall and broken down by rough expected volume levels, along with counts on the fringe or tail end of expected ranges or patterns.

In most cases, potentially bad data would need to be manually checked to confirm or reject the basis for the flag. Thus, a method is proposed to flag suspect data with adjustable scrutiny levels. Individual users may wish to apply higher or lower scrutiny based on their knowledge of the dataset or gained experience with previous data flagging (e.g. if most flagged data is determined to be valid data, the user may wish to lower the level of scrutiny, flagging only data further outside the expected range). Data quality check methods, developed based on empirical counts in BikePed Portal, are proposed to identify appropriate flags for repeated zero values, repeated non-zero values, and maximum/excessive count values.

Finally, we developed recommended check thresholds along with an implementation approach and plan to incorporate additional data quality and control checks into BikePed Portal.

17. Key Words Bicycle; Pedestrian; Non-motorized; count; data; quality; check		18. Distribution Statement No restrictions. Copies available from NITC: www.nitc-utc.net	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 61	22. Price

ACKNOWLEDGEMENTS

This project was funded by the National Institute for Transportation and Communities (NITC) under grant number 1026, with additional support from the Oregon Department of Transportation.

DISCLAIMER

The contents of this report reflect the views of the authors, who are solely responsible for the facts and the accuracy of the material and information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation University Transportation Centers Program the National Institute for Transportation and Communities in the interest of information exchange. The U.S. Government the National Institute for Transportation and Communities assume no liability for the contents or use thereof. The contents do not necessarily reflect the official views of the U.S. Government the National Institute for Transportation and Communities. This report does not constitute a standard, specification, or regulation.

RECOMMENDED CITATION

McNeil, Nathan, and Kristin Tufte, *Bike and Walking Quality Counts: Using Bike Ped Portal Count to Develop Data Quality Checks*. NITC-RR-1026. Portland, OR: Transportation Research and Education Center (TREC), 2018.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXCUTIVE SUMMARY

Cycling and walking are sustainable modes of transportation which improve community livability, but these modes have not been studied with the quantitative rigor applied to motor vehicle travel. This work aims to change that by improving the quality of bicycle and pedestrian traffic monitoring data, including understanding how erroneous data can most accurately and efficiently be identified through automated processes. The research approach analyzes continuous bicycle and pedestrian count data stored in BikePed Portal, an archive of bicycle and pedestrian count data with a web-based interface.

A primary goal of the project is to identify data quality tests that both could identify aberrant and/or erroneous data and that could be automated in the BikePed Portal data archive and web site. These two sometimes-contradictory goals guided the development of the tests described in this report. A key method deployed in pursuit of identifying tests that could identify aberrant and/or erroneous data was to comb through a selection of count data (generally – continuous counter locations from 2015 to 2016 with at least 30 days of counts) to identify expected count ranges and patterns, overall and broken down by rough expected volume levels, along with counts on the fringe or tail end of expected ranges or patterns. This method identified a set of potential data quality tests.

The intent of this exploration was to produce a set of tests that could automatically identify potentially erroneous data in BikePed Portal. However, due to the highly varied nature of bicycle and pedestrian data (e.g weather and seasonality impacts on the data values), data which is identified as potentially erroneous through the automated tests must be manually checked by a human to confirm if the data is or is not erroneous. Thus, the data quality tests proposed identified by this project need to: automatically identify potentially erroneous data; be implementable in BikePed Portal with a reasonable amount of development effort; and produce results which can be reviewed by a human.

To avoid overloading a human who is manually checking the data after the automated data quality checks have been run, a method is proposed to identify and flag suspect data with adjustable scrutiny levels. Individual users may wish to apply higher or lower scrutiny based on their knowledge of the dataset or gained experience with previous data flagging (e.g. if most flagged data is determined to be valid data, the user may wish to lower the level of scrutiny, flagging only data further outside the expected range).

Data quality check methods, developed based on empirical counts in BikePed Portal, are proposed to identify appropriate flags for repeated zero values, repeated non-zero values, and maximum/excessive count values. To summarize the findings, the research found that runs of more than 100 zero-counts (over 24 hours of zero counts for 15-minute bins) are suspicious. If checking for non-zeros, runs of nine or longer should be flagged regardless of traffic volume at the site. If over 1000 bicyclists or pedestrians are counted in a 15-minute time, this should generally be flagged as suspicious.

Finally, we developed recommended check thresholds along with an implementation approach and plan to incorporate additional data quality and control checks into BikePed Portal.

1.0 BACKGROUND

Motorized traffic data is regularly collected and stored, providing many opportunities to analyze robust data sets. Nonmotorized counts, primarily pedestrian and bicycle counts, have historically been collected much less commonly. Further, pedestrian and bicycle counts are often inconsistent in terms of duration and equipment used.

With limited consistent collection and reporting requirements or protocols, nonmotorized count data may be siloed within a specific agency or even a single staff member. Bicycle and pedestrian counts, when they are collected, are often collected by local jurisdictions, and may be used for local planning efforts, but are not usually shared beyond the immediate jurisdiction or region. Budgets for collecting nonmotorized count data are usually quite limited, or non-existent, leaving jurisdictions limited data collection options. Manual counts by volunteers for periods for as little as one to two hours are relatively common.

In addition to the limited data collection options, there is little exploration in the academic literature of acceptable automatic quality control checks for automated bicycle and pedestrian counts and almost no investigation of pedestrian data specifically. For this reason, this study includes an unprecedented number of pedestrian and bicycle count locations and data records.

BikePed Portal was created with the goal of providing a national repository for bicycle and pedestrian data. This project builds on BikePed Portal by exploring data quality checks that can be automated in BikePed Portal.

1.1 DEVELOPMENT OF THE BIKEPED PORTAL

In 2014, a Portland State University team began development of BikePed Portal, with the goal of creating a consistent schema and repository for non-motorized count data. A key component of that project was the development of a schema that would allow for the efficient and consistent upload and storage of counts.

BikePed Portal seeks to combine the best elements of the TMAS and NBPDP data protocol elements with data input, output and visualization measures, while accepting data from all over the United States.

The initial focus of BikePed Portal was on continuous data and automated counters, in part because these rich data sources provide more opportunity for complex data analysis. However, manual counts can also be accommodated. BikePed Portal built off knowledge gained during the development of PORTAL – which “provides a centralized, electronic database that facilitates the collection, archiving, and sharing of data and information for public agencies within the region" (http://portal.its.pdx.edu/).

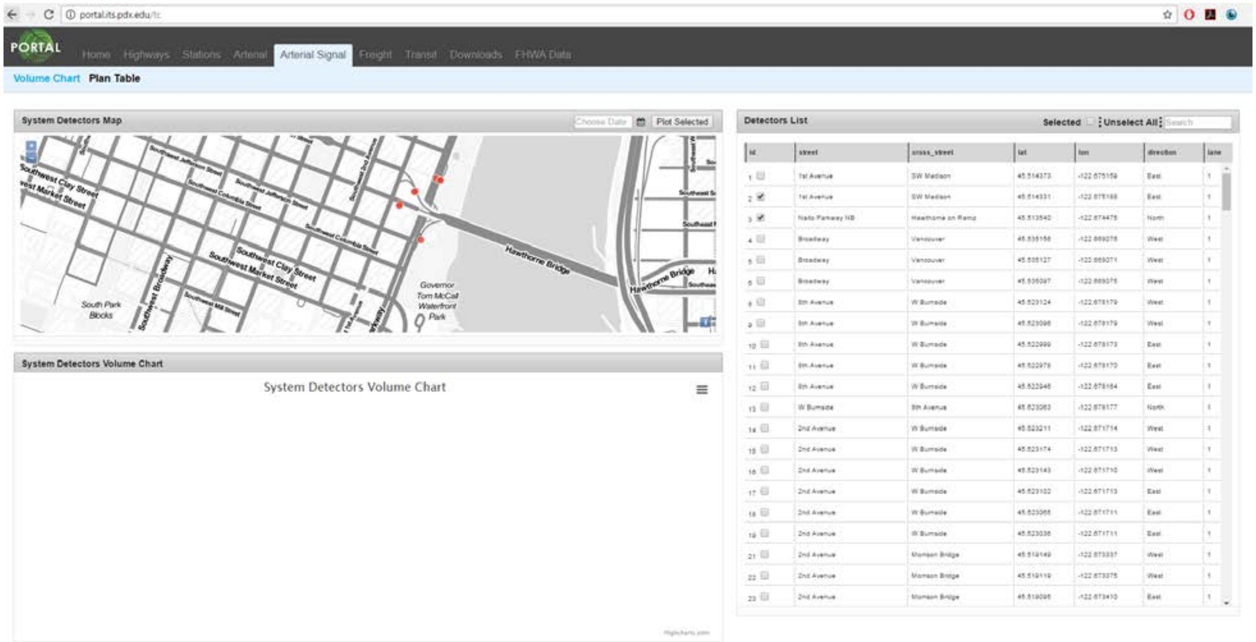

Figure 1-1 PORTAL's Arterial Signal Page

The BikePed Portal development process is detailed in the report "BikePed Portal: Development of an Online Nonmotorized Traffic Count Archive" (2016). Figure 1-2 demonstrates the count data storage schema for the Bike Ped Portal database. As of 2017, the BikePed Portal had over 400 locations in 8 states, with 33,000,000 count records and over 200,000,000 trips. Plans are to continue expand beyond these numbers.

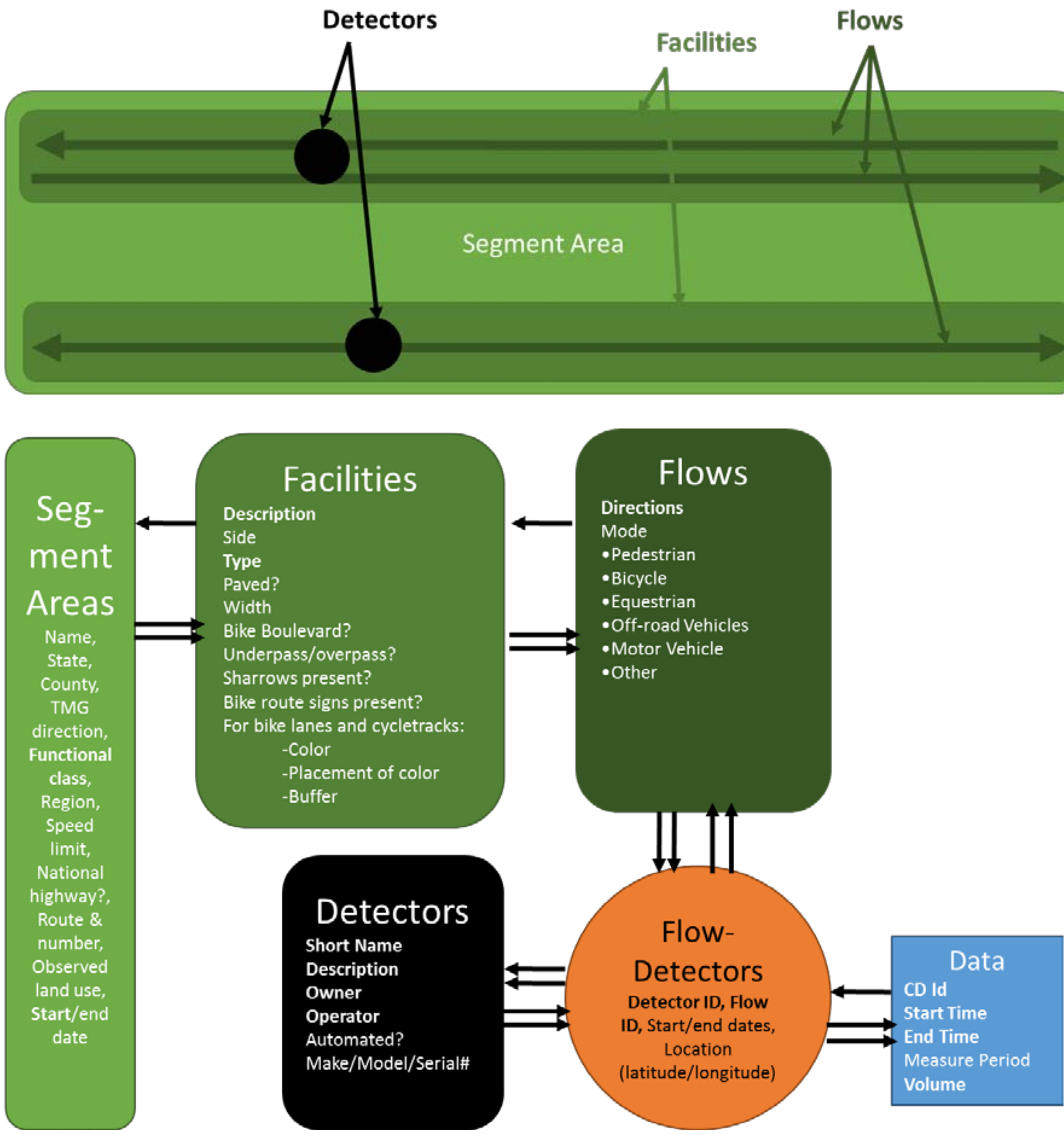

Figure 1-2 BikePed Portal Archive Data Structure (Nordback et al., 2016)

1.2 OTHER BICYCLE AND PEDESTRIAN COUNT DATABASES

Existing bicycle and pedestrian count databases or websites have often been developed by or in coordination with local or regional entities as repositories for local count data. Among other existing online archives for bicycle and pedestrian count data are:

- Bike Count Data Clearinghouse at the University of California, Los Angeles (http://www.bikecounts.luskin.ucla.edu/);

- Central Lane Metropolitan Planning Organization (http://www.thempo.org/356/Bicycle-Counts - see Figure 1-3 for web page screen shot);
- Delaware Valley Regional Planning Commission (https://www.dvrpc.org/webmaps/pedbikecounts/ - see Figure 1-4 for webpage screen shot); and,
- FHWA has also developed a protocol for providing bicycle data into the Travel Monitoring and Analysis System (TMAS).

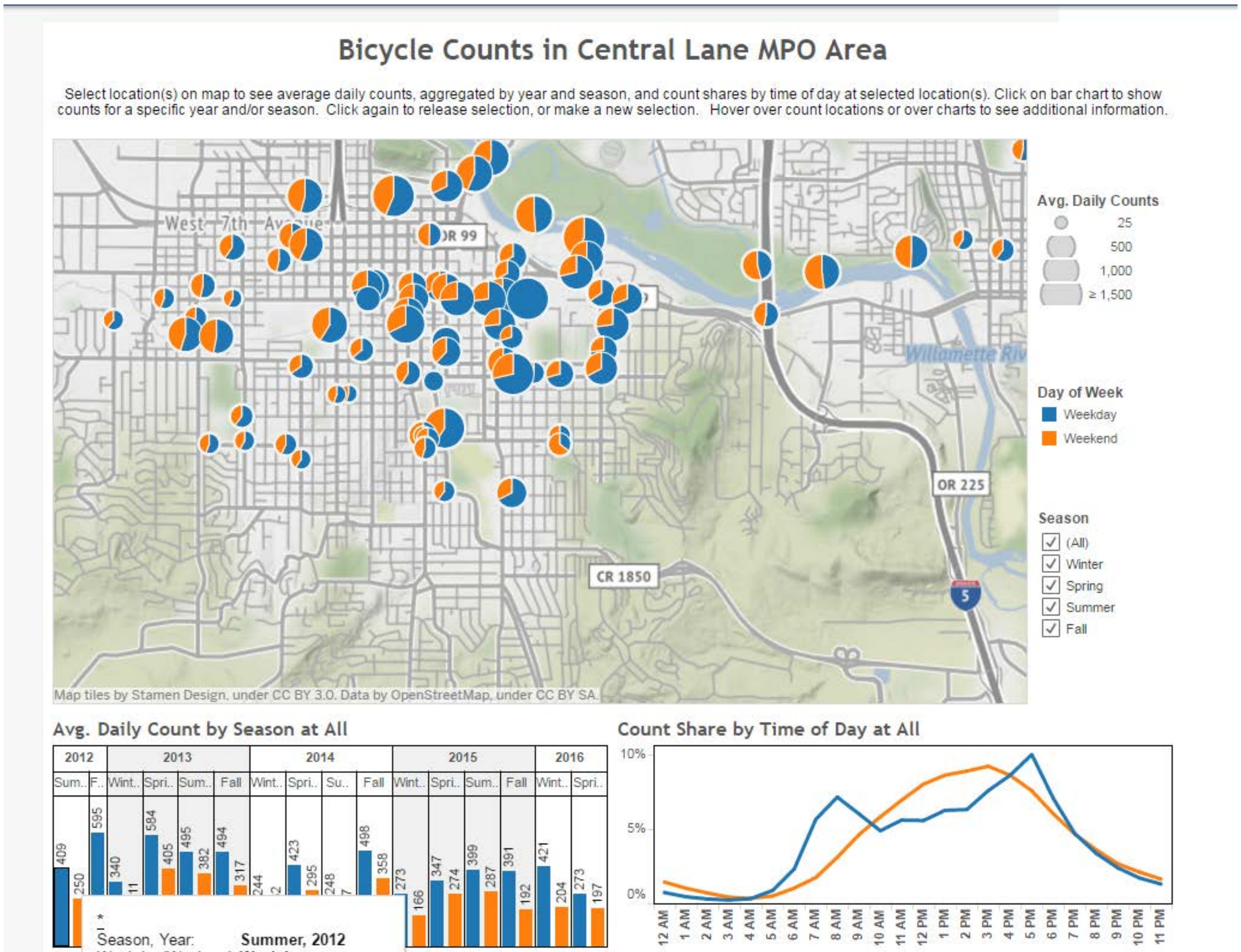

Figure 1-3 Central Lane Metropolitan Planning Organization Bicycle Counts Webpage

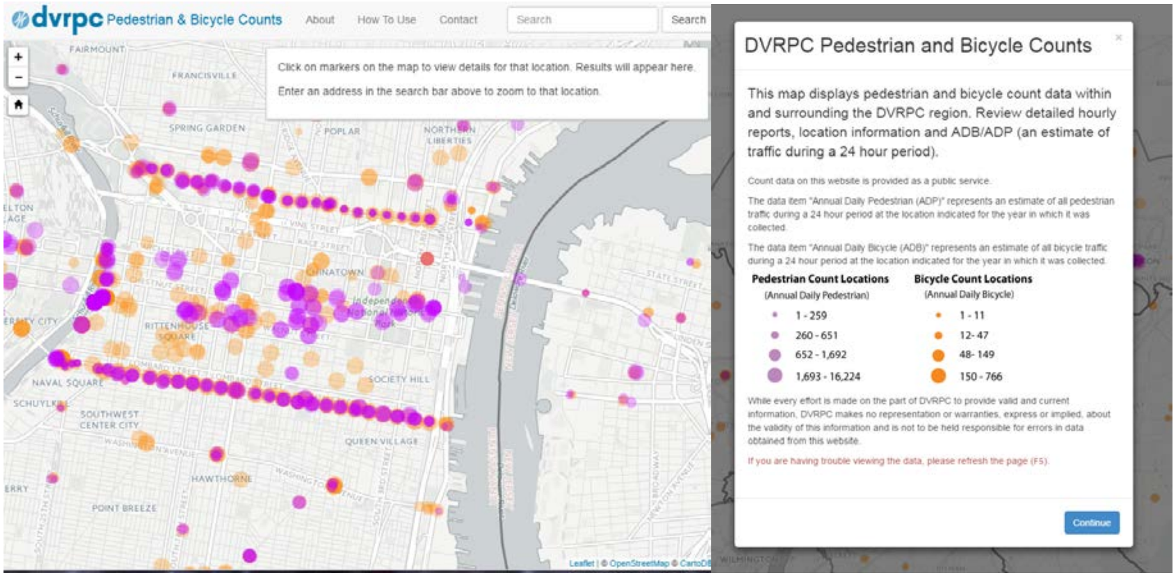

Figure 1-4 Delaware Valley Regional Planning Commission Pedestrian and Bicycle Counts Webpage

The National Bicycle and Pedestrian Documentation Project (NBPDP - http://bikepeddocumentation.org/index.php/downloads) is a joint effort of Alta Planning & Design and the Institute of Transportation Engineers (ITE) Pedestrian and Bicycle Council, and has made major strides in developing consistent data collection protocol and data entry forms. The project accepts data deposits, though it does not have a public portal or interface to the count data.

1.3 BIKEPED PORTAL QUALITY CHECKS

Counters, whether automated or manual, can have errors. In addition, the variety of sources of data and conversions to the BikePed Portal format may introduce errors. Currently, BikePed Portal has basic quality control measures such as rejecting null counts and duplicate count records. This work, when implemented, will add additional quality control measures to BikePed Portal.

With the quantities of data in BikePed Portal come both the need to ensure that the data is high quality, and the opportunity to use this large data set to develop tests and checks. This project set out to improve BikePed Portal's ability to identify potentially erroneous data and provide opportunities for data owners to identify the data as valid, erroneous, or something in between (e.g. valid data but caused by unusual circumstances, such as unusually high counts due to special events or unusually low counts due to closures and/or construction). In addition, due to the goal to implement the identified checks in BikePed Portal, focus was given only to tests that are structured in such a manner as to be automatable.

1.4 LITERATURE REVIEW

Quality bicycle and pedestrian data are critical for the study of bicycle and pedestrian safety. The lack of such data mean bicycle and pedestrian safety is not well understood. These data are also important for regional bicycle model validation, signal timing and economic impact studies. Jurisdictions around the country have started bicycle and pedestrian count programs, but little is known about how to automatically quality check these data.

Turner and Lasley's 2013 paper on establishing quality assurance procedures for non-motorized traffic count data began by putting forth several key principles, including that: quality assurance starts before data are collected; acceptable quality is determined by the data's use; and measures can quantify data quality dimensions.

There are established data quality checks for archived motor vehicle traffic data (e.g. Turner 2007 and Turner 2002), especially for freeway data. However, because non-motorized data has vastly different characteristics, including being much more variable and much less recorded and studied, the checks for motorized traffic data cannot be applied.

1.4.1 Expected error

This project did not undertake the task of examining expected error including systematic undercounts of pedestrians or bikes due to inherent detection challenges, such as counters being unable to identify individuals in a group, the potential for bikes and pedestrians to deviate from lanes where counters are installed, etc. Other research has explored that topic. For example, a study of different counting technologies (NCHRP 07-19) found that most automated counters undercounted pedestrian and bicycle traffic. It is important to note that although the checks and tests discussed in this report can identify many erroneous data points, agencies and individuals overseeing counts need to understand how a specific counter will operate in the specific locations and conditions in which they plan to install it, and to choose the counter, installation, and calibration appropriately – this was a key finding from NCHRP Project 07-19 (Ryus et al 2015b). Figure 1-5 lists correction factors for undercounting as calculated in the NCHRP project. We do not calculate calibration and adjustment factors in the current report.

Table 4-2. Simple counter correction factors developed by NCHRP Project 07-19.

Sensor Technology	Adjustment Factor	Hours of Data
Passive infrared	1.137	298
Product A	1.037	176
Product B	1.412	122
Active infrared*	1.139	30
Radio beam	1.130	95
Product A (bicycles)	1.470	28
Product A (pedestrians)	1.323	27
Product B	1.117	40
Bicycle-specific pneumatic tubes	1.135	160
Product A	1.127	132
Product B	1.520	28
Surface inductive loops	1.041	29
Embedded inductive loops	1.054	79
Piezoelectric strips*	1.059	58
Combination (pedestrians)	1.256	47

Notes: *Factor is based on a single sensor at one site; use caution when applying.

.

Figure 1-5 Counter Correction Factors for Undercounting by Sensor Technology From NCHRP 07-19 (Ryus et el., 2015b).

1.4.2 Sources of non-standard or unexpected error

NCHRP 797 provides valuable guidance on pedestrian and bicycle volume data collection, but stops short of providing the detailed guidance on automated quality checks [Ryus, 2015a]. The report does provide a discussion of a number of potential sources of error for automated bicycle and pedestrian counters. Among the sources of error outlined in the report are:

- Occlusion. When multiple people cross a screen line counter simultaneously, the counter may undercount. This is more likely to happen at higher volumes.
- Environmental conditions like extreme heat (thermal counters in particularly may not catch human if the ambient temp is near the temp of humans), extreme cold (minor error if subjects are wearing very heavy thermally protective coats, and pneumatic tubes can undercount due to hardening of the tube rubber – though not well documented), rain (can interfere with optical counters resulting in high over counts during heavy rain or snow events – particularly active infrared), and low lighting (may cause problems for optical counters).
- Counter bypassing. Loops or tubes can be bypassed, and most sensors can have blind spots.
- The effects of motor vehicles or other road users interacting with bicycle or pedestrian counters.

- Mechanical malfunction.
- User error or setup problems.

1.4.3 Checks for non-standard or unexpected error

Due to the relative lack of robust bicycle and pedestrian count data, there are few investigations into the data quality issues associated non-motorized count data. Among the investigations that do exist, Turner and Lasley (2013) examined the data quality issues associated with infrared count data. NCHRP 797 and associated web-only Document 205 discuss how to clean data and include an extensive discussion of the sources of error for automated counting devices (as discussed above), but does not provide specific tests that would be needed to automate data checking (Ryus, 2015a; Ryus, 2015b).

In terms of identifying potentially erroneous counts (notwithstanding expected error), existing work on error checking are summarized in Table 1-1. The work in this table has focused almost entirely on bicycle counts. Pedestrian counts may need different checks. Consensus on which tests and values to use has not been reached. Some of the checks are very simple and easy to implement, such as a single hourly or daily cap on count volume or set frequencies of repeating zero or non-zero values. Other checks require somewhat more existing data or computation, such as calculating the interquartile range or examining the standard deviation from surrounding days. TMAS uses an adjustable calculation based on the count volume itself.

Table 1-1. Summary of Quality Control Checks for Non-motorized Traffic Counts (Adapted from Nordback et al. 2016, Table 5-2)

Source	Upper bound [lower bound]	Identical non-zero values	Consecutive zeros	Directional Split
Turner & Lasley	Interquartile range (IQR) = 2.5 (Q_3-Q_1) + Q_3	-	-	-
Seattle DOT	3 standard deviations above surrounding days	-	-	-
Univ. of Minn.	2 to 3 standard deviation above average	-	-	-
Colorado DOT	Weekly check: daily count 3 times higher previous year's average daily traffic; Quarterly check: IQR = 2.5 (Q_3-Q_1) + Q_3	-	Over 2 days of zero counts	splits > than 70 percent/ 30 percent
North Carolina State Univ.	3 standard deviations above [or below] predicted daily count based on model from previous 6 months of cleaned data (model includes weather and day of week)	-	Over 3 days of zero counts	Splits > than 3 stand. dev. of average
BikePed Portal / Portland State University *initial	1,500 per hour, 5,000 per day	Over 6 identical non-zeros	Over 15 hours of zero counts	-
BikePed Portal / Portland State University (Report WA-RD 875.2)	1,000 per hour			
FHWA TMAS V2.7	For hourly counts <100: flag if 100% over/ under the previous interval count For hourly counts >100: Flag if 100 higher/ lower than previous interval count Over 50,000 daily count; over 4,000 hourly count For daily counts under 1,000: Flag if 100% > [or <] than average of past 6 previous. If daily count over 1,000: flag if 1000 over [or under] the average of past 6 previous.	Over 3 identical non-zero values	>7 hours with consecutive zeros	-

Another research team using BikePed Portal data conducted manual and semi-automated checks for data accuracy. Summarized on Page 68 of that report (Nordback et al., 2017) was an analysis of bicycle and pedestrian capacity and saturation flow rate:

> "The 1,000 per hour threshold was determined after a review of bicycle and pedestrian capacity and saturation flow rate studies (Tables 17-19). This threshold was considered to be flexible for the person conducting the quality checks to use discretion in determining if the data appeared to be real or a malfunction. This

discretion was necessary in order to not over clean the data at higher volume sites. The threshold was also loosely based off the peak hourly volumes found for one bicycle flow direction on the Hawthorne Bridge, a site with some of the largest bicycle volumes in the U.S. (e.g., roughly 5,000 riders daily) (Figure 20). Since many of the sites checked had much lower daily volumes than this, it would be an anomaly for 1,000 users to pass within one hour given demonstrated travel patterns and volumes."

1.4.4 Implementation of checks for archived databases

One notable pattern emerges from most examinations of quality assurance and quality control for non-motorized traffic count data: established checks are primarily manual in nature, and often assume that the check is being done by the agency collecting the data or an experienced professional. According to NCHRP Report 797, most practitioners use a spreadsheet to compile and analyze data, though some agencies develop their own databases or use a vendor's software.

Further, the checks developed to date and discussed above have primarily been administered on an ad-hoc basis or for testing purposes. Non-motorized count archives that conduct quality checks on archived data include FHWA's TMAS, the vendor-specific database managed by Eco-Counter which has the ability to send warnings to clients on potential data problems on upload and MS2's Non-Motorized Database System which checks data on upload (includes checks for runs of zeros for a given day).

Turner and Lasley (2013) also explore the transferability of checks developed for motorized transportation databases (including those outlines in Turner 2002), and noted that "Several motorized traffic database applications already have automated validity criteria built into their data import process. Therefore, it is possible to use existing software applications to perform validity reviews of pedestrian and bicyclist count data. However, many of these existing validity criteria use thresholds and parameter values that were developed and refined for typical motorized traffic patterns." Turner and Lasley also recommend, and other resources concur (e.g. Minge et al., 2017) that visual checks are an essential element of quality assurance for non-motorized count data.

2.0 METHODOLOGY

Recognizing the challenges of developing automated quality control and assurance checks, particularly for non-motorized data that often displays considerable variability, the checks developed for this project had the goals of:

1) being implementable in the BikePed Portal non-motorized count archive. To achieve this, the checks would need to be high level and flag primarily extreme values.
2) allowing for borderline or questionable data to be flagged as suspicious or valid by the data owner or uploader. In order to do this, there needs to be a middle ground between data flagged as likely bad or suspicious data, and data that appears to be good data.
3) being flexible, in terms of allowing for future checks and improvement of the current checks. Ensuring that the implementation allowed for updating of checks and inclusion of new checks required that the project consider the database design and operation.

We also sought to include opportunities for the data owner or uploader to visually inspect the data, specifically flagged or suspicious data, and update the flags according to their judgment.

The research team sought to employ the quantity of data contained within the BikePed Portal to scan for patterns in count volumes and count characteristics such as runs of repeated values. The goal was to identify trends around expected and potentially aberrant counts that could be deployed within the BikePed Portal to carry out simple data quality checks quickly.

Key to the understanding of the project goals was that data quality checks can identify suspect or potentially bad data, but that in most cases there would need to be a user who was familiar with the data to have a final say on flagged data.

2.1 CONSIDERED CHECKS

The checks that this project sought to test and consider for implementation into the BikePed Portal include:

Repetitive zero counts: Test a set of counts with a variety of daily volumes. Test hourly counts of 5, 10, 12, 15, 20, 24, and 48 hours of consecutive 0 counts for 15-minute or one hour counts.

Repetitive non-zero counts: Test a set of counts for consecutive non zeros counts. Runs with flags for 3, 4, 5, 6, 8, 10 consecutive non-zeros for 15-minute counts. Are runs of consecutive non-zeros more likely with lower volume counts than with higher volume counts?

Excessive 15-minute or daily hard cap count: Test count thresholds of 100, 200, 500, 1000, 1500. Review flagged counts and days for whether or not flag captured bad data, special event, good data (i.e. false positive), or unknown. Which thresholds appear to be most effective based on expected volume?

Several other tests were considered but not implemented during the study project phase. These will be considered for future implementation. Examples include:

- Inverted AM/PM: Test to identify if counts between 6pm and 6am are higher than those between 6am and 6pm. For each location, test for months, weeks, and days that exhibit this potential inversion, and flag. Review flags.
- Unusual data at night (12am-6am) - hourly counts between midnight and 6am with counts above 25, 50, 100, 250.

2.2 DATA INCLUDED IN TESTS

We sought to include sites with continuous counters and at least 30 days of data in 2015 to 2016. These restrictions and years were chosen to establish a constrained analysis period, while also using a data set with large count sets to understand what counts would be expected. These sites contained 12,627,239 total count entries.

Table 2-1 Count Sites included in Tests, by State, County and Count Type

State / City	Bicycle	Pedestrian	Total
CA	64	30	94
San Diego County	64	30	94
CO	10		10
Boulder County	10		10
OR	12		12
Multnomah County	12		12
VA	39	27	66
Arlington County	39	27	66
WA	19	17	36
Chelan County	2	2	4
King County	11	9	20
Spokane County	4	4	8
Thurston County	2	2	4
Grand Total	144	74	218

2.3 FLAGGING AND REVIEWING COUNTS

Within this report and in the proposed data quality checks, we apply several terms that should be defined based on their use herein. As noted above, the proposed checks would be implemented through a flagging system that would run automatically upon (or shortly after) data upload.

Definitions related to the flagging approach include:

- Suspicious – Unusual data that may be erroneous – should be reviewed.
- Possibly Suspicious – In some cases, data may not be suspicious (e.g. for higher volume sites), but in other cases may be suspicious. Handling of these scenarios requires further decision-making.
- Not Suspicious – Data is not unusual based on the criteria under consideration, and does not merit further review.
- Flag – Mark a data point for further review, such as validation by data uploader or project staff. Without validation, data users may wish to exclude this data.

In the tables in this report, we color code these groupings as follows: suspicious data in red, possibly suspicious data in yellow, not suspicious data in green.

It is not uncommon to utilize the standard deviation to identify potential bad data. NCHRP 797 suggests using two standard deviations from above or below the average value for a comparable same time of week counts (for an 8-week period before and after the test date) as a means of identifying probably incorrect data. However, this approach would in practice likely flag around 5% of data, which could prove unwieldly in a data archive scenario with millions of counts. Instead, we have opted to push to about three standard deviations for the tests deployed in this report. This results in about half a percent of data being flagged. It is worth noting, however, that the number of standard deviations used could be changed up or down as needed in the future.

3.0 FINDINGS

3.1 ZERO RUNS

Runs of counts of zero are just about as common as non-zero runs in one regard. We identified 689,334 runs of zero, compared to 636,177 runs of non-zeros. However, runs of zeros averaged about 10 counts per run, as opposed to just about 2.3 count per run for non-zero runs. Table 3-17 and Table 3-18 show the number of zero runs and percent of zero runs, respectively, by run length and average daily volume. Runs of up to 49 consecutive zero counts are not uncommon regardless of average daily volume. Suggested thresholds occur at drop-off points of 50 or more zero counts or just over 12 hours of 15-minute counts (possibly suspicious), or 100 or more zero counts, just over 24 hours of 15-minute counts (suspicious).

As before for non-zero runs, each “count” refers to the smallest count duration in the raw data originally provided by the counting device for that site. This analysis uses sites with 15-minute counts. The count includes traffic in one direction, unless the original count aggregated counts from both directions for a given mode on a given road or path segment.

Table 3-1 Count of Zero Runs by run length, by Average Daily Volume

Run Length	100	100 to 500	500+	Grand Total
2	97,080	88,843	25,316	211,239
3	55,945	44,591	13,266	113,802
4 to 5	59,510	43,985	12,995	116,490
6 to 9	47,887	35,465	10,488	93,840
10 to 15	26,202	23,038	6,749	55,989
16 to 25	18,753	23,137	5,838	47,728
26 to 49	21,980	16,574	3,626	42,180
50 to 99	5,731	974	599	7,304
100 to 149	168	41	34	243
150 to 249	142	55	32	229
250 to 999	94	59	31	184
1000+	42	48	16	106
Grand Total	**333,534**	**276,810**	**78,990**	**689,334**
All counts	4,607,615	5,811,197	2,208,427	12,627,239

Table 3-2 Percent of Zero Runs by Run Length, by Average Daily Volume

Run Length	100	100 to 500	500+	Grand Total
2	29.106%	32.095%	32.050%	30.644%
3	16.773%	16.109%	16.795%	16.509%
4 to 5	17.842%	15.890%	16.451%	16.899%
6 to 9	14.357%	12.812%	13.278%	13.613%
10 to 15	7.856%	8.323%	8.544%	8.122%
16 to 25	5.623%	8.358%	7.391%	6.924%
26 to 49	6.590%	5.988%	4.590%	6.119%
50 to 99	1.718%	0.352%	0.758%	1.060%
100 to 149	0.050%	0.015%	0.043%	0.035%
150 to 249	0.043%	0.020%	0.041%	0.033%
250 to 999	0.028%	0.021%	0.039%	0.027%
1000+	0.013%	0.017%	0.020%	0.015%
Grand Total	**333,534**	**276,810**	**78,990**	**689,334**
All counts	4,607,615	5,811,197	2,208,427	12,627,239

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Examining zero counts by percentage of all counts contained in runs of zero (Table 3-19) shows that very long runs of 1000 or more zero counts account for a significant portion of all data – about 9%, even though only accounting for 0.015% of runs (as shown in Table 3-18). It's also worth noting that, for sites with average daily volume of less than 100, an average of 75% of counts are contained within runs of zero – compared to 45% for medium volume sites and 31% for high volume sites.

Table 3-3 Percent of all Counts in Zero Runs, by Run Length and Average Daily Volume

Run Length	100	100 to 500	500+	Grand Total
2	4.214%	3.058%	2.293%	3.346%
3	3.643%	2.302%	1.802%	2.704%
4 to 5	5.688%	3.329%	2.590%	4.061%
6 to 9	7.455%	4.394%	3.422%	5.341%
10 to 15	6.841%	4.824%	3.738%	5.370%
16 to 25	8.132%	8.036%	5.244%	7.582%
26 to 49	16.916%	9.296%	5.413%	11.397%
50 to 99	7.656%	1.024%	1.818%	3.582%
100 to 149	0.426%	0.083%	0.174%	0.224%
150 to 249	0.554%	0.171%	0.274%	0.329%
250 to 999	0.954%	0.472%	0.608%	0.671%
1000+	12.908%	8.033%	3.291%	8.983%
Grand Total	75%	45%	31%	54%
All counts	4,607,615	5,811,197	2,208,427	12,627,239

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Compared to the non-zero run counts, the suggested flag thresholds for zero run counts result in a significant percentage of overall count data being flagged. As shown in Table 3-20, the suggested thresholds result in 10.2% of counts being flagged as suspicious and 3.6% flagged as possibly suspicious. However, the 10.2% of counts flagged as suspicious make up only 0.1% of runs. Using the 2015-2016 data, the average flagged zero run would be 1691 consecutive zero counts.

Table 3-4 Percent of Counts and Runs Flagged in Zero-Run Counts

Percent of all counts flagged		
Red	1,288,849	10.207%
Yellow	452,366	3.582%
Total	1,741,215	13.789%
Percent of runs flagged		
Red	762	0.111%
Yellow	7,304	1.060%
Total	8,066	1.170%

3.2 NON-ZERO RUNS

Non-zero runs is defined as anytime two consecutive counts reflect the same (non-zero) volume. It makes intuitive sense that repeated counts will occur at times, and that they are more likely to do so at lower volumes (due to the higher frequency of specific lower volume counts), and that counters are increasingly less likely to repeat specific count volumes for higher volumes.

At the same time, higher volume locations may be more likely than lower volume sites to repeat higher volume counts, primarily due to the simple fact that they are more likely to produce higher counts at all.

If the same count volume occurs repeatedly for a given site, there is a possibility of an error or glitch caused by the counting equipment. Prior to this study no such glitch had been documented for non-motorized counting equipment; however, FHWA proposed including this check in its TMAS quality checking procedure. As discussed in the following section, using a check developed in this project we did identify at least one near-certain glitch of this type.

This section details the frequency of non-zero runs by run length (how many times the same count repeats) and count volume. Here "count" refers to the smallest count duration in the raw data originally provided by the counting device for that site. For some equipment this is a 15-minute count and for other equipment this is the count per hour. The count includes traffic in both directions for a given mode on a given road or path segment. Then, we break the data down by average daily count volume for the site to explore the differences by expected volumes.

3.2.1 Non-zero runs by run length and volume - overall

First, we examine non-zero runs (regardless of count volume) to understand how often they occur. Within the 2015-2016 count data, we identified 636,177 non-zero runs, consisting of 1,434,792 counts and 11.36% of all count data. The vast majority of the runs were runs of two, accounting for over 80% of all runs, and just over 8% of all data. Runs of three consecutive non-zero values accounted for 14.86% of runs and 2.25% of all data. In most cases, the unexpectedness of a non-zero run (and thus the likelihood that it could represent bad data and should be flagged), will depend on the count volume, and factors such as the expected site volume. However, non-zero runs of seven or more are very rare regardless of count volume, at less than 0.1% of all runs. This suggests that if checking for non-zeros, runs of seven or longer should be flagged regardless of traffic volume at the site.

Table 3-5 2015-2016 sites with at least 30 days of Data: Non-Zero Runs Frequency

Run Length	Number of Runs	Counts in Runs	Percentage of Runs	Percentage of all data
2	513,108	1,026,216	80.65%	8.127%
3	94,518	283,554	14.86%	2.246%
4	21,034	84,136	3.31%	0.666%
5	5,356	26,780	0.84%	0.212%
6	1,507	9,042	0.24%	0.072%
7	414	2,898	0.065%	0.023%
8	143	1,144	0.022%	0.0091%
9	53	477	0.008%	0.0038%
10	26	260	0.00409%	0.0021%
11	5	55	0.00079%	0.00044%
12	2	24	0.00031%	0.00019%
13	2	26	0.00031%	0.00021%
16	1	16	0.00016%	0.00013%
19	1	19	0.00016%	0.00015%
20	3	60	0.00047%	0.00048%
21	3	63	0.00047%	0.00050%
22	1	22	0.00016%	0.00017%
Grand Total	636,177	1,434,792	100%	11.36%

A first look at run length by count volume is shown in Table 3-2 and Table 3-3, presenting the number of runs and the percent of runs in each count volume category, respectively. Although these tables do not take into account expected site volumes, they still provide insight into the frequency of runs by count volume that can inform quality checks. Generalized information about the relatively unusualness of non-zero runs can inform flagging of potentially suspicious data even absent information about the expected daily volume of the site. Runs of 6 or more of volumes greater than two are also very rare and should likely always be flagged. Similarly, runs of five or more with volumes above 5, of four or more with volumes above 16, and of three or more with volumes above 100 should be flagged. Possibly suspicious non-zero runs (highlighted in yellow in the tables below), may be flagged depending on the desired sensitivity threshold.

The runs of ten consecutive identical counts in the 6 to 9 category and 26 to 99 category were both at the same location (the Curtis Trail in Rosslyn, Virginia), and for the same period. The lesser count was 10 consecutive counts of 9 pedestrians, while the greater count was 10 consecutive counts of 51. We hypothesize that these values have been manual or automated filler for a time period when the counter was down.

Table 3-6 Number of Non-Zero Runs by Count Volume – all 2015-2016 data

	Count of runs for each length by volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**
2	297,815	119,350	52,118	25,010	12,378	5,931	506	513,108
3	70,433	16,668	4,861	1,715	638	194	9	94,518
4	17,819	2,497	551	127	29	11	0	21,034
5	4,933	363	50	9	1	0	0	5,356
6	1,443	60	3	0	1	0	0	1,507
7	402	10	1	1	0	0	0	414
8	142	1	0	0	0	0	0	143
9	53	0	0	0	0	0	0	53
10	24	0	1	0	0	1	0	26
11	5	0	0	0	0	0	0	5
12	2	0	0	0	0	0	0	2
13	2	0	0	0	0	0	0	2
16	1	0	0	0	0	0	0	1
19	1	0	0	0	0	0	0	1
20	3	0	0	0	0	0	0	3
21	3	0	0	0	0	0	0	3
22	1	0	0	0	0	0	0	1
Grand Total	393,082	138,949	57,585	26,862	13,047	6,137	515	636,177

Table 3-7 Percent of Non-Zero Runs for each Run Length, by Count Volume – all 2015-2016 data

	Percent of Runs for each length by count volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**
2	58.04%	23.26%	10.16%	4.87%	2.41%	1.16%	0.10%	513,108
3	74.52%	17.63%	5.14%	1.81%	0.68%	0.21%	0.01%	94,518
4	84.72%	11.87%	2.62%	0.60%	0.14%	0.05%	0%	21,034
5	92.10%	6.78%	0.93%	0.17%	0.02%	0%	0%	5,356
6	95.75%	3.98%	0.20%	0%	0.07%	0%	0%	1,507
7	97.10%	2.42%	0.24%	0.24%	0%	0%	0%	414
8	99.30%	0.70%	0%	0%	0%	0%	0%	143
9	100%	0%	0%	0%	0%	0%	0%	53
10	92.31%	0%	3.85%	0%	0%	3.85%	0%	26
11	100%	0%	0%	0%	0%	0%	0%	5
12	100%	0%	0%	0%	0%	0%	0%	2
13	100%	0%	0%	0%	0%	0%	0%	2
16	100%	0%	0%	0%	0%	0%	0%	1
19	100%	0%	0%	0%	0%	0%	0%	1
20	100%	0%	0%	0%	0%	0%	0%	3
21	100%	0%	0%	0%	0%	0%	0%	3
22	100%	0%	0%	0%	0%	0%	0%	1
Grand Total	393,082	138,949	57,585	26,862	13,047	6,137	515	636,177

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Applying the thresholds above, 0.13% of runs, and 0.05% of counts would be flagged as suspicious, while 0.13% of counts and 0.6% of runs could be flagged as possibly suspicious (see Table 3-4)

Table 3-8 Percent of Counts and Runs Flagged by Suggested Non-Zero Count Run Flag Thresholds

Percent of all counts flagged		
Sensitivity	# flagged	Percent flagged
Red	5,935	0.05%
Yellow	16,693	0.13%
Total	22,628	0.18%
Percent of runs flagged		
Red	827	0.13%
Yellow	3,822	0.60%
Total	4,649	0.73%

3.2.2 Non-zero runs frequency, Sites less than 100 per day

Next, we examine how the frequency of runs for sites with daily volumes of 100 or less per day (low-volume sites). For low-volume sites in 2015-2016, we identified 157,569 runs accounting for 367,037 counts, or just under 8% of all counts for these sites.

The tables below detail the thresholds for sites with expected daily volumes of less than 100, which could be self-identified expected volumes by the data inputter or owner, or could be calculated based on existing data.

Table 3-9 2015-2016 Low-Volume Sites with at least 30 days of Data: Non-Zero Runs Frequency

Run Length	Number of Runs	Counts in Runs	Percentage of Runs	Percentage of all data
2	120,104	240,208	76.2231%	5.213%
3	27,477	82,431	17.4381%	1.789%
4	7,058	28,232	4.4793%	0.613%
5	2,001	10,005	1.2699%	0.217%
6	625	3,750	0.3967%	0.081%
7	175	1,225	0.1111%	0.027%
8	75	600	0.0476%	0.013%
9	32	288	0.0203%	0.006%
10	11	110	0.0070%	0.002%
11	2	22	0.0013%	0.000%
12	2	24	0.0013%	0.001%
19	1	19	0.0006%	0.000%
20	3	60	0.0019%	0.001%
21	3	63	0.0019%	0.001%
Grand Total	157,569	367,037	100%	7.966%

Table 3-6 breaks down the run length by count volumes for low-volume sites. Most runs were runs of 2 with volumes less than 10, runs of 3 with volumes of 5 or less, or runs of 4 to 5 and volumes of 1 to 2. Runs beyond those thresholds were relatively unusual. No runs with volumes of 100 or more were observed, and runs with count volumes of 10 or more were very rare for these sites. Table 3-7 shows the percentage of runs by each count volume for low volume sites.

Table 3-10 Number of Non-Zero Runs by Count Volume – Low-Volume 2015-2016 Sites

	Count of runs for each length by volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**
2	106,197	12,432	1,231	192	33	19	0	120,104
3	25,881	1,478	109	7	2	0	0	27,477
4	6,831	217	10	0	0	0	0	7,058
5	1,964	36	1	0	0	0	0	2,001
6	623	2	0	0	0	0	0	625
7	175	0	0	0	0	0	0	175
8	75	0	0	0	0	0	0	75
9	32	0	0	0	0	0	0	32
10	11	0	0	0	0	0	0	11
11	2	0	0	0	0	0	0	2
12	2	0	0	0	0	0	0	2
19	1	0	0	0	0	0	0	1
20	3	0	0	0	0	0	0	3
21	3	0	0	0	0	0	0	3
Grand Total	141,800	14,165	1,351	199	35	19	0	157,569

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Table 3-11 Percent of Non-Zero Runs for each Run Length, by Count Volume – Low-Volume 2015-2016 sites

	Percent of Runs for each length by count volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**
2	88.4%	10.35%	1.02%	0.16%	0.03%	0.02%	0%	120,104
3	94.2%	5.38%	0.40%	0.03%	0.01%	0%	0%	27,477
4	96.8%	3.07%	0.14%	0%	0%	0%	0%	7,058
5	98.2%	1.80%	0.05%	0%	0%	0%	0%	2,001
6	99.7%	0.32%	0%	0%	0%	0%	0%	625
7	100%	0%	0%	0%	0%	0%	0%	175
8	100%	0%	0%	0%	0%	0%	0%	75
9	100%	0%	0%	0%	0%	0%	0%	32
10	100%	0%	0%	0%	0%	0%	0%	11
11	100%	0%	0%	0%	0%	0%	0%	2
12	100%	0%	0%	0%	0%	0%	0%	2
19	100%	0%	0%	0%	0%	0%	0%	1
20	100%	0%	0%	0%	0%	0%	0%	3
21	100%	0%	0%	0%	0%	0%	0%	3
Grand Total	141,800	14,165	1,351	199	35	19	0	157,569

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Applying the thresholds above, 0.06% of counts, and 0.23% of runs would be flagged as suspicious, while 0.12% of counts and 0.76% of runs could be flagged as possibly suspicious (see Table 3-8).

Table 3-12 Percent of Counts and Runs Flagged by Suggested Non-Zero Count Run Flag Thresholds – Low-Volume Sites

Percent of all counts flagged		
Red	2,675	0.06%
Yellow	5,421	0.12%
Total	8,096	0.18%
Percent of runs flagged		
Red	362	0.23%
Yellow	1,193	0.76%
Total	1,555	0.99%

3.2.3 Non-zero runs frequency, Sites 100 to 499 per day

Next, we examine how the frequency of runs for sites with daily volumes of 100 to 499 per day (medium-volume sites) – see Table 3-9. For medium-volume sites in 2015-

2016, we identified 359,232 runs accounting for 806,282 counts, or just under 14% of all counts for these sites. Over 80% of runs, accounting for just over 10% of data for these sites, were in runs of 2.

Table 3-13 2015-2016 Medium-Volume Sites with at least 30 days of Data: Non-Zero Runs Frequency

Run Length	Number of Runs	Counts in Runs	Percentage of Runs	Percentage of all data
2	291,551	583,102	81.1595%	10.03%
3	52,752	158,256	14.6847%	2.72%
4	11,192	44,768	3.1155%	0.77%
5	2,727	13,635	0.7591%	0.23%
6	726	4356	0.2021%	0.07%
7	189	1323	0.0526%	0.02%
8	59	472	0.0164%	0.01%
9	17	153	0.0047%	0%
10	12	120	0.0033%	0%
11	3	33	0.0008%	0%
13	2	26	0.0006%	0%
16	1	16	0.0003%	0%
22	1	22	0.0003%	0%
Grand Total	359,232	806,282	100%	13.87%

Table 3-10 breaks down the run length by count volumes for medium-volume sites. Most runs were runs of 2 with volumes less than 26, runs of 3 with volumes of 10 or less, runs of 4 with volumes of 5 or less, and runs of 5 with volumes of 1 to 2. Runs beyond those thresholds were relatively unusual. Table 3-11 shows the percentage of runs by each count volume for medium-volume sites.

Table 3-14 Number of Non-Zero Runs by Count Volume – Medium-Volume 2015-2016 Sites

	Count of runs for each length by volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**
2	153,559	84,990	35,433	13,198	3,741	622	8	291,551
3	35,987	12,255	3,425	883	182	20	0	52,752
4	8,830	1,881	396	76	9	0	0	11,192
5	2,418	271	35	2	1	0	0	2,727
6	670	53	2	0	1	0	0	726
7	180	8	1	0	0	0	0	189
8	58	1	0	0	0	0	0	59
9	17	0	0	0	0	0	0	17
10	11	0	1	0	0	0	0	12
11	3	0	0	0	0	0	0	3
13	2	0	0	0	0	0	0	2
16	1	0	0	0	0	0	0	1
22	1	0	0	0	0	0	0	1
Grand Total	201,737	99,459	39,293	14,159	3,934	642	8	359,232

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Table 3-15 Percent of Non-Zero Runs for each run length, by Count Volume – Medium-Volume 2015-2016 Sites

	Percent of runs for each length by volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**

2	53%	29%	12%	5%	1%	0.213%	0.003%	291551
3	68%	23%	6%	2%	0.35%	0.04%	0%	52752
4	79%	17%	4%	1%	0.08%	0%	0%	11192
5	89%	10%	1%	0%	0.04%	0%	0%	2727
6	92%	7%	0.3%	0%	0%	0%	0%	726
7	95%	4%	1%	0%	0%	0%	0%	189
8	98%	2%	0%	0%	0%	0%	0%	59
9	100%	0%	0%	0%	0%	0%	0%	17
10	92%	0%	8%	0%	0%	0%	0%	12
11	100%	0%	0%	0%	0%	0%	0%	3
13	100%	0%	0%	0%	0%	0%	0%	2
16	100%	0%	0%	0%	0%	0%	0%	1
22	100%	0%	0%	0%	0%	0%	0%	1
Grand Total	201,737	99,459	39,293	14,159	3,934	642	8	359,232

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Applying the thresholds above, 0.05% of counts, and 0.14% of runs would be flagged as suspicious, while 0.2% of counts and 0.84% of runs could be flagged as possibly suspicious (see Table 3-12).

Table 3-16 Percent of Counts and Runs Flagged by Suggested Non-Zero Count Run Flag Thresholds – Medium-Volume Sites

Percent of all counts flagged		
Red	3,107	0.05%
Yellow	11,398	0.20%
Total	14,505	0.25%
Percent of runs flagged		
Red	491	0.14%
Yellow	3,024	0.84%
Total	3,515	0.98%

3.2.4 Non-zero runs frequency, Sites 500 + per day

Next, we examine how the frequency of runs for sites with daily volumes of 500 or more per day (high-volume sites) – see Table 3-13. For high -volume sites in 2015-2016, we identified 119,376 runs accounting for 261,473 counts, or just under 12% of all counts for these sites. Over 84% of runs, accounting for just over 9% of data for these sites, were in runs of 2.

Table 3-17 2015-2016 High-Volume Sites with at least 30 days of Data: Non-Zero Runs Frequency

Run Length	Number of Runs	Counts in Runs	Percentage of Runs	Percentage of all data
2	101,453	202,906	84.986%	9.188%
3	14,289	42,867	11.970%	1.941%
4	2,784	11,136	2.332%	0.504%
5	628	3140	0.526%	0.142%
6	156	936	0.131%	0.042%
7	50	350	0.042%	0.016%
8	9	72	0.008%	0.003%
9	4	36	0.003%	0.002%
10	3	30	0.003%	0.001%
Grand Total	119,376	261,473	100%	11.840%

Table 3-14 breaks down the run length by count volumes for high-volume sites. Compared to low- and medium-volume sites, there were far more runs with volumes about 100, as well as between 26 and 99, and 16 to 25. Such higher volume runs were almost always runs of 2, or occasionally runs of three. Table 3-15 shows the percentage of runs by each count volume for high-volume sites.

Table 3-18 Number of Non-Zero Runs by Count Volume – High-Volume 2015-2016 Sites

	Count of runs for each length by volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**
2	38,059	21,928	15,454	11,620	8,604	5,290	498	101,453
3	8,565	2,935	1,327	825	454	174	9	14,289
4	2,158	399	145	51	20	11	0	2,784
5	551	56	14	7	0	0	0	628
6	150	5	1	0	0	0	0	156
7	47	2	0	1	0	0	0	50
8	9	0	0	0	0	0	0	9
9	4	0	0	0	0	0	0	4
10	2	0	0	0	0	1	0	3
Grand Total	49,545	25,325	16,941	12,504	9,078	5,476	507	119,376

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Table 3-19 Percent of Non-Zero Runs for each run length, by Count Volume – High-Volume 2015-2016 Sites

	Percent of Runs for each length by count volume							
Run Length	**1 to 2**	**3 to 5**	**6 to 9**	**10 to 15**	**16 to 25**	**26 to 99**	**100 +**	**Total**

2	38%	22%	15%	11%	8%	5%	0%	101,453
3	60%	20.54%	9.29%	5.77%	3.18%	1.22%	0.06%	14,289
4	78%	14.33%	5.21%	1.83%	0.72%	0.40%	0%	2,784
5	88%	8.92%	2.23%	1.11%	0%	0%	0%	628
6	96%	3.21%	0.64%	0%	0%	0%	0%	156
7	94%	4%	0%	2%	0%	0%	0%	50
8	100%	0%	0%	0%	0%	0%	0%	9
9	100%	0%	0%	0%	0%	0%	0%	4
10	67%	0%	0%	0%	0%	33.33%	0%	3
Grand Total	41.5%	21.2%	14.2%	10.5%	7.6%	4.6%	0.4%	119,376

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Applying the thresholds above, 0.04% of counts, and 0.11% of runs would be flagged as suspicious, while 0.22% of counts and 1.28% of runs could be flagged as possibly suspicious (see Table 3-12).

Table 3-20 Percent of counts and runs flagged by suggested non-zero count run flag thresholds – high-volume sites

Percent of all counts flagged		
Red	780	0.04%
Yellow	4,844	0.22%
Total	5,624	0.25%
Percent of runs flagged		
Red	133	0.11%
Yellow	1,528	1.28%
Total	1,661	1.39%

3.3 HARD CAP COUNT

We sought to identify extreme counts beyond which counts might be suspicious. To start examining this topic, we identified the five highest 15-minute counts for each site for one direction for a given mode, again using the sites from 2015-2016 that had at least 30 days of continuous data, and 15-minute counts. We also calculated the 99th percentile count volume for each site.

Table 3-21 shows the mean, median, standard deviation, and minimum/maximum of these data point by average daily volume category.

Table 3-21 Maximum 15-minute Count Volume Statistics by Daily Volume Categories

Daily Average Volume Category	n		99th	Top1	Top2	Top3	Top4	Top5
less than 100	77	Mean	5	200	89	67	51	47
		Median	4	36	23	21	19	18
		Std. Deviation	3	706	239	134	73	70
		Minimum	1	4	3	3	3	3
		Maximum	19	4,740	2,032	1,024	432	415
100 to 499	98	Mean	19	231	157	124	114	105
		Median	18	99	89	77	69	64
		Std. Deviation	10	522	247	146	141	131
		Minimum	7	17	13	13	13	13
		Maximum	78	4,094	2,048	921	909	874
500 or more	40	Mean	204	652	614	582	559	538
		Median	78	356	327	273	219	200
		Std. Deviation	519	985	968	949	941	939
		Minimum	29	77	76	71	70	70
		Maximum	3,306	4,737	4,642	4,611	4,611	4,610
Total	215	Mean	48	298	217	189	174	165
		Median	13	98	81	73	64	60
		Std. Deviation	234	712	506	465	455	450
		Minimum	1	4	3	3	3	3
		Maximum	3,306	4,740	4,642	4,611	4,611	4,610

We extracted the high daily 15-minute count for each day (2015-2016) for each site, and examined various volume thresholds for how frequently they flagged the daily high value. Values were calculated for a total of 131,544 days.

Tests were run for hard caps at volume of 75, 100, 250, 500, 1000, 1500, and 2000. As shown in Table 3-22, sites with volumes of 100 or less only surpassed the count of 75 thresholds on about 0.7% of days (the sum of days flagged at cut-offs of 75 or higher), and only surpassed 100 on about 0.27% of days. For sites with average daily volumes of 100 to 499, counts above 250 (possibly suspicious) to 500 (suspicious) were unusual. For sites with average daily volumes of 500 or more, appropriate thresholds for possibly suspicious values would be over 1000 to over 2000.

These checks would be possible either with a calculated average daily counts, or a user/uploader provided value.

Table 3-22 15-minute Volume Caps – Hard Cut-Off

	Average Daily Volume group	100 or less	100 to 499	500 +	Total
	Days of Data	47,998	60,539	23,007	131,544
Days flagged with hard cut-off . . . . per 15-minutes	75	202	442	4,601	5,245
		0.42%	0.73%	20.00%	3.99%
	100	113	283	3,384	3,780
		0.24%	0.47%	14.71%	2.87%
	250	10	63	1,396	1,469
		0.02%	0.10%	6.07%	1.12%
	500	3	23	489	515
		0.01%	0.04%	2.13%	0.39%
	1000	2	2	251	255
		0.00%	0.00%	1.09%	0.19%
	1500	2	2	218	222
		0.00%	0.00%	0.95%	0.17%
	2000	2	2	206	210
		0.00%	0.00%	0.90%	0.16%

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

Table 3-23 uses each sites' average daily volume and checks thresholds calculated as a percentage of the daily average. In general, 15-minute counts above 1.25 times the average daily count for that site would be suspicious, while counts above 50% of the daily average would be possibly suspicious. As an exploration into the effect of various percentage of daily average thresholds, Table 3-23 also presents the percentage of days that would be flagged for each threshold. On one extreme, using a threshold of flagging a 15-minute count that exceeded 10% of the daily average for that site would result in flags for 24% of all days of data. On the far end, only flagging 15-minute counts that exceeded three times the daily average for that site would result in flagging 0.035% of all days.

Table 3-23 15-minute Volume Caps – Percent of Daily Average Cut

	Average Daily Volume group	100 or less	100 to 500	above 500	Total
	Days of Data	47,998	60,539	23,007	131,544
Days flagged with calculated cut at	10% daily average	21,025	8,213	2,426	31,664
		43.80%	13.57%	10.54%	24.07%
	25% daily average	3,325	749	367	4,441
		6.93%	1.24%	1.60%	3.38%
	50% daily average	892	228	97	1,217
		1.86%	0.38%	0.42%	0.93%
	75% daily average	484	135	68	687
		1.01%	0.22%	0.30%	0.52%
	1x daily average	282	83	56	421
		0.59%	0.14%	0.24%	0.32%
	1.25x daily average	199	60	49	308
		0.41%	0.10%	0.21%	0.23%
	1.5x daily average	138	46	41	225
		0.29%	0.08%	0.18%	0.17%
	2x daily average	86	23	41	150
		0.18%	0.04%	0.18%	0.11%
	3x daily average	39	7	0	46
		0.08%	0.01%	0.00%	0.03%

Color coding: Green = not suspicious; Yellow = possibly suspicious; Red = suspicious/flag.

As an exploration into the effect of various percentage of daily average thresholds, the total column in Table 3-23 presents the percentage of days that would be flagged for each threshold. On one extreme, using a threshold of flagging a 15-minute count that exceeded 10% of the daily average for that site would result in flags for 24% of all days of data. On the far end, only flagging 15-minute counts that exceeded three times the daily average for that site would result in flagging 0.035% of all days.

3.4 ADAPTIVE RUNNING THRESHOLDS

Two checks have been identified and strategized, but current funding and time constraints do not allow us to implement them as part of this project. However, brief descriptions of the approaches conceptualized during this project are included below.

Both approaches employ analysis that calculates interquartile median based thresholds for overall daily volume or time of day expected volume. The benefit of these tests is that they present individualized thresholds to the specific count location (and time of day or day of week in the latter test). This test is expected to be better at identifying potentially suspicious data, but also would likely be more prone to flagging unusual but valid events, which will place more dependence on the data owner/uploader to be engaged and review and confirm if flagged data is valid or not.

3.4.1 Daily maximum value

This test uses the daily volume from the previous 27 days of data for the same counter, and calculates a moving maximum threshold for the expected daily volume. By examining the past 27 days of data, an estimate of the expected daily count and normal variation for a relatively comparable time period can be established. Thus, if a given day is outside the expected range, the data would be flagged. We used a calculation based on the interquartile range (IQR, or the difference between the third and first quartile). The details of the calculation and the cutoff thresholds for flagging data are shown below:

A) Calculate the daily volume for the past 27 days, along with the interquartile range for that time period.
B) Calculate third quartile plus 2x IQR (slightly greater than standard interquartile ratio outlier) of these values.
C) Establish thresholds as shown in Table 3-24:

Table 3-24 IQR Cutoff Thresholds

If Q3+2*IQR is…	The threshold cutoff should be…
0 to 10	10
10,1 to 25	25
25.1 to 50	50
50.1 to 100	100
100.1 to 150	150
150.1 to 250	250
250.1 to 500	500
500.1 to 750	750
750.1 to 1000	1000
1000.1 to 1500	1500
1500.1 to 2000	2000
2000.1 to 5000	5000
5000.1 to 7500	7500
7500.1 to 10,000	10,000
10000.1 to 20,000	20,000
20,000.1+	No cutoff

3.4.2 Time of day maximum and minimum value

This test uses counts from specific hours of the day (separating out weekend versus weekday) and is completed by carrying out an analysis similar to the daily maximum value test described above. This test provides an even more specific check on expected counts for a given hour of the day, and allows for identifying counts that are either above (Q3+2*IQR) or below (Q1-2*IQR) a specific range. An example of this calculation is shown in Table 3-25.

Table 3-25 Time of Day Max. and Min. Value overview

Hour	**Weekday**	**Weekend**
0	IQR for any weekday counts in the 0 hour (midnight to 1am), for past 27 days	IQR for any weekend counts in the 0 hour (midnight to 1am), for past 27 days
1	IQR for any weekday counts in the 1am hour (1am to 2am), for past 27 days	IQR for any weekend counts in the 1am hour (1am to 2am), for past 27 days
2	IQR for any weekday counts in the 2am hour (2am to 3am), for past 27 days	IQR for any weekend counts in the 2am hour (2am to 3am), for past 27 days
…	…	…

4.0 RECOMMENDED CHECKS AND THRESHOLDS

Section 3 described the findings of the analysis; this section describes the checks and thresholds which will be used to flag data. The following section, Section 5, provides an implementation framework for these checks and thresholds

4.1 RECOMMENDED CHECKS

The following checks are recommended to be used to flag data:

> *Zero-Count Thresholds:* Thresholds for each category (Suspicious; Possibly Suspicious; Not Suspicious) for runs of zero count volume values based on run length.
>
> *Non-Zero Count Thresholds:* Thresholds for each category (Suspicious; Possibly Suspicious; Not Suspicious) for runs of non-zero count volume values based on run length and count volume.
>
> *Volume-Specific Non-Zero Count Thresholds:* Thresholds as above, but specified for sites with expected daily volumes of less than 100; between 100 and 500; and greater than 500.
>
> *Volume-Specific Hard Cap Counts:* Hard cap counts for sites with expected daily volumes of less than 100, between 100 and 500, and greater than 500, as well as non-volume-specific thresholds.

Expected Daily Volume: The implementation assumes that sites with expected daily volumes of less than 100, between 100 and 500, and greater than 500 have been identified and that this identification is done through a process separate from the flagging process. The expected daily volume for a site may be identified by the data inputter or site owner or through a periodic calculation of average daily volume. The identification of the expected daily volume of the site is not addressed in this section.

Flag Categories: As described in Section 2 and employed in Section 3, data will be flagged into three categories: Suspicious; Possibly Suspicious; Not Suspicious. As listed above, runs of non-zero counts, runs of zero counts and hard caps are all used to flag data.

4.2 RECOMMENDED THRESHOLDS

For each of the recommended checks, thresholds differentiated by expected daily volume are provided. These thresholds will be used to flag data; data flagged as Suspicious or Possibly Suspicious will be provided to the data inputter for review. The definition of these thresholds was based on the goal of capturing the extreme ends of the range of data values and of providing a frequency of flagging that is acceptable to the user. These thresholds are intended to provide a lower rate of flagging that is intended to be used as a starting point. The thresholds can be updated (e.g. sensitivity adjusted) based on a combination of user experience (is the user receiving “too many” flags) and effectiveness – that is, is the system producing too many false positives or no false positives (which could be a sign that the system should be flagging slightly more). The thresholds in this section have generally been established based on flagging approximately 0.1% or less of counts as suspicious and 0.25% or less of counts as possibly suspicious, which will flag data that falls outside of approximately three standard deviations of the mean. The run length and count value thresholds will be customizable in the implementation to allow data inputters to regulate the amount of suspicious and possibly suspicious data flagged for review.

4.2.1 Zero-Count Thresholds

Runs of zero-value counts are flagged for review by the data inputter based on the following thresholds on run length (Table 4-5). Since these are runs of count value zero, the count value column is not needed in this table. Although we examined zero-value runs by expected daily volume categories, the suggested thresholds did not vary by expected volume.

Table 4-1 Zero-Count Thresholds – Not Volume Specific

Flag	***Run Length***
Suspicious	>=100
Possibly Suspicious	50 to 99

4.2.2 Non-Zero Count Thresholds

Although our initial findings sought to identify non-zero runs beyond two to three standards deviations from the mean, we recommend looser thresholds since these errors appear to be rare, with few known examples.

If checking for non-zeros, runs of nine or longer should be flagged regardless of traffic volume at the site.

Table 4-1 lists the data flags for non-zero count thresholds by run length and count value. These thresholds may be used absent an expected daily volume. If an expected daily volume is known, then the volume-specific thresholds are preferred.

Table 4-2 Non-Zero Count Thresholds – not Volume Specific

Flag	***Run Length***	***Count Value***
Suspicious	>=9	Any
Suspicious	>=8	> 2 (>=3)
Suspicious	>=7	> 5 (>=6)
Suspicious	>=6	> 9 (>=10)
Suspicious	>=5	> 15 (>=16)
Suspicious	>=3	> 100 (>= 100)
Possibly Suspicious	8	1 to 2

Possibly Suspicious	7	3 to 5
Possibly Suspicious	6	6 to 9
Possibly Suspicious	5	10 to 15
Possibly Suspicious	4	16 to 99
Possibly Suspicious	2	>= 100

4.2.3 Volume-Specific Non-Zero Count Thresholds

Table 4-2 to Table 4-4 lists the flag thresholds for non-zero counts by run length and count value for sites with expected daily volumes less than 100, between 100 and 500, and greater than 500, respectively.

Table 4-3 Non-Zero Count Thresholds: Expected Daily Volume < 100

Flag	***Run Length***	***Count Value***
Suspicious	>=9	Any
Suspicious	>=6	> 2 (>=3)
Suspicious	>=5	> 5 (>= 6)
Suspicious	>=4	>= 10
Suspicious	>=2	> =100
Possibly Suspicious	8	1 to 2
Possibly Suspicious	5	3 to 5
Possibly Suspicious	4	6 to 9
Possibly Suspicious	3	10 to 99

Table 4-4 Non-Zero Count Thresholds: Expected Daily Volume 100 - 500

Flag	***Run Length***	***Count Value***
Suspicious	>=9	Any
Suspicious	>=8	> 2 (>=3)
Suspicious	>=7	> 5 (>= 6)
Suspicious	>=5	>= 10
Suspicious	>=4	>= 26
Suspicious	>=3	>= 100
Possibly Suspicious	8	1 to 2
Possibly Suspicious	7	3 to 5
Possibly Suspicious	6	6 to 9
Possibly Suspicious	5	10 to 25
Possibly Suspicious	3	26 to 99

Table 4-5 Non-Zero Count Thresholds: Expected Daily Volume > 500

Flag	***Run Length***	***Count Value***
Suspicious	>=9	Any
Suspicious	>=7	> 2 (>=3)
Suspicious	>=6	> 5 (>= 6)
Suspicious	>=5	>= 16
Suspicious	>=4	>= 100
Possibly Suspicious	8	1 to 2
Possibly Suspicious	6	3 to 5
Possibly Suspicious	5	6 to 15
Possibly Suspicious	4	16 to 99
Possibly Suspicious	3	>= 100

4.2.4 Volume-Specific Hard Cap Counts

Count volumes exceeding a certain hard cap become increasingly unusual the higher the volumes are. Suggested thresholds for situations in which the expected daily volume is unknown, or is known, are presented in Table 4-6.

Table 4-6 Volume-Specific Hard Cap Thresholds

Flag	Suspicious – flag 15-minute counts above:	Possibly Suspicious– flag 15-minute counts above:
Expected Volume unknown	1000	500
Expected Daily Volume < 100	250	100
Expected Daily Volume 100 - 500	500	250
Expected Daily Volume > 500	2000	1000

4.2.5 Long Term Trends

Significant changes in data values over time, e.g. a gradual, but significant, decrease in counts that could be indicative of obstruction of a detector by the growth of a shrub or fouling of the detector are important to detect. Checks recommended are to compare a month of data versus the prior three months of data and same month over the prior three years. Data values that are more than three standard deviations above the mean for the prior three months or same month over the prior three years will be flagged as Suspicious. If there are less than three months of data for either case (most recent three months or prior three years, same month), the threshold will be adjusted to 2.5 standard deviations and the data will be flagged as Potentially Suspicious. We expect that more testing will be needed to validate that these thresholds produce the appropriate amount of flagged data.

4.3 APPLYING CHECKS

4.3.1 Using data from a subsequent time period

In order to conduct a measure of validation of the checks, we applied our checks to a set of sites for count data in 2017 (the checks and thresholds were developed using data from 2015 and 2016). Sites included location in Colorado, California, Oregon and Virginia, as shown in Table 4-7.

Table 4-7 2017 count data site locations

Segment Name	State	Location	Detector ID
W&OD Trail (Bon Air Park-West)	VA	Arlington	344
Hawthorne Bridge	OR	Portland	462
Torrey Pines Rd E of Almafi St	CA	San Diego	1042
13th St N of Walnut St	CO	Boulder	1129

Figure 4-1 demonstrates the hard cap threshold, assuming the expected daily volume is not specified. Because this check is designed to capture extreme outliers, it is not surprising that no counts are identified as exceeding the hard cap threshold. The adaptive running thresholds identified in Section 3.3.1 will be better able to identify unexpected fluctuations from day to day (both above and below expected count ranges).

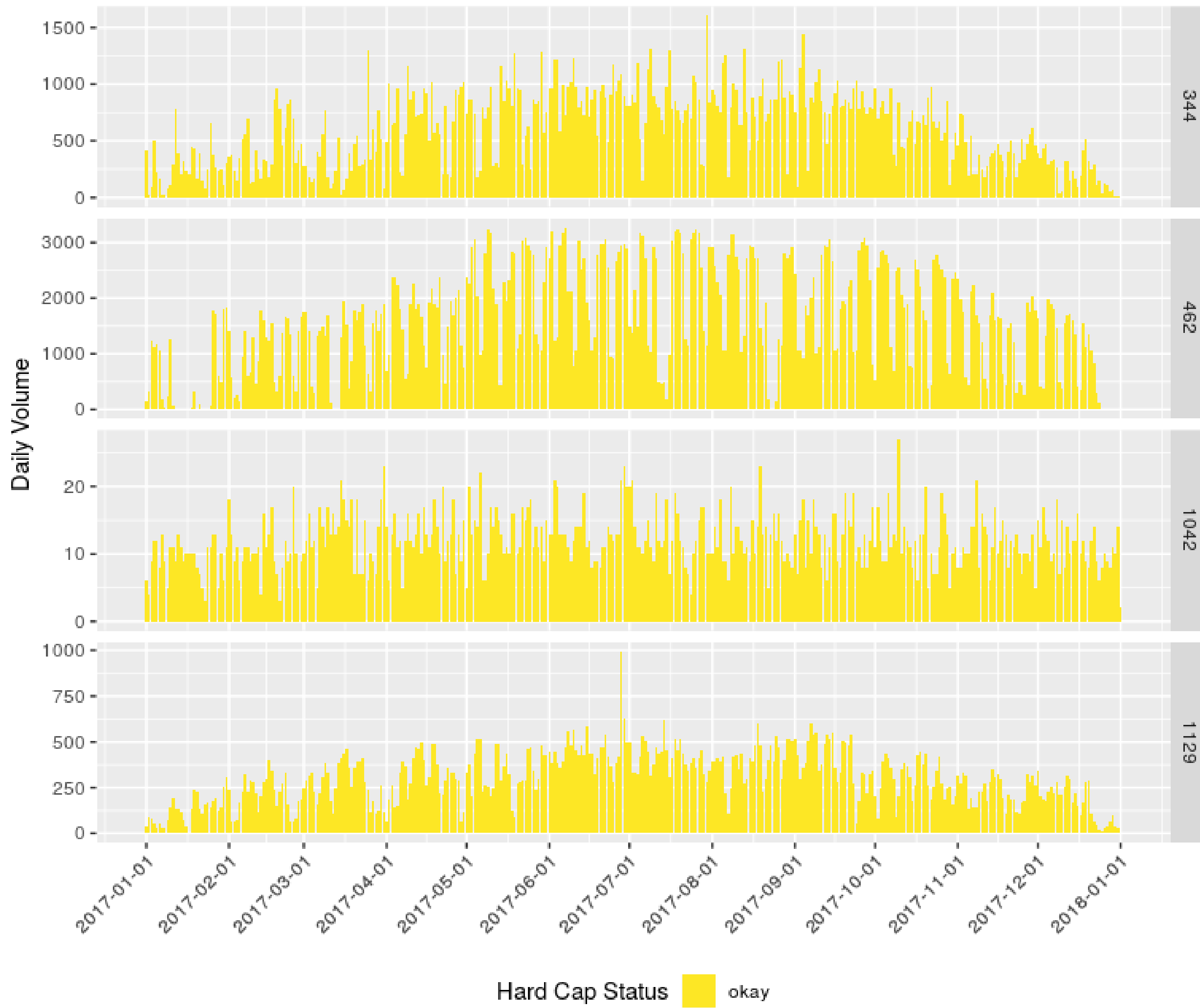

Figure 4-1 Application: Hard cap volume

Figure 4-2 demonstrates the check of non-zero runs. Note that longer runs, or runs with higher volumes are noted as possibly suspicious or suspicious. Site 344 had 13 possibly suspicious non-zero runs (3 count runs of 21, 23 and 26, 4 count runs of 7, 10, 12, and 15, 5 count runs of 3 two times, 6 count runs of 1 four times), along with a suspicious 7 count run of 1. Site 462 had three possibly suspicious runs – a two count run of 115, and three count runs of 17 and 29. Site 1042 had no suspicious or possibly suspicious non-zero runs. Site 1129 had 12 possibly suspicious runs, including a three count run of 16 two times, four count runs of 6, 7, 8, 10, 10 and 12, five count runs of 3 and 5, and six count runs of 1 and 2. The site also had two suspicious runs – a six count run of 4 and a seven count run of 1. As these flagged data show,

The values flagged in Figure 4-2 seem less likely to be erroneous data than just natural variation. For this reason, the non-zero runs checks may not be identifying errors in the data in most cases, but rather natural stochastic variation in the data.

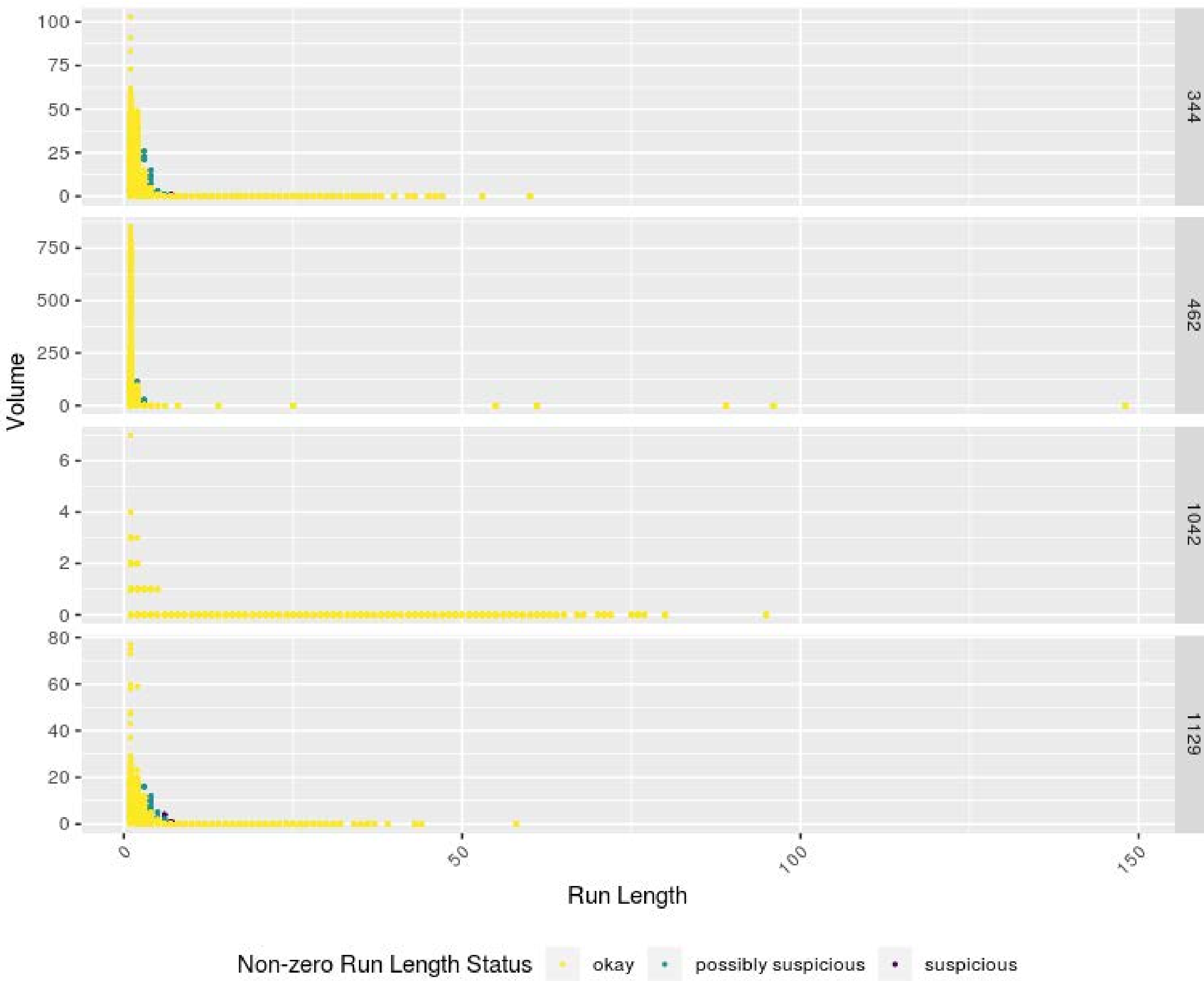

Figure 4-2 Application: Non-zero runs

Figure 4-3 shows the check of the runs of zeros, with runs between 50 and 99 marked as possibly suspicious and 100 or more counts of zero as suspicious. Site 344 had possibly suspicious runs of 53 and 60. Site 462 had possibly suspicious runs of 55, 61, 89, and 96. Site 1042 had a total of 93 runs of zeros that were identified as possibly suspicious. Site 1042 was a lower volume site, but this number of possibly suspicious counts could put a greater burden on the data owner during the flagging and checking process than the other sites would. An implementation next step would involve identifying if, for such low volume sites, the threshold needed to be adjusted along with if the burden is justified and bearable for the user. Finally, site 1129 had one possibly suspicious run of 58 consecutive zero counts.

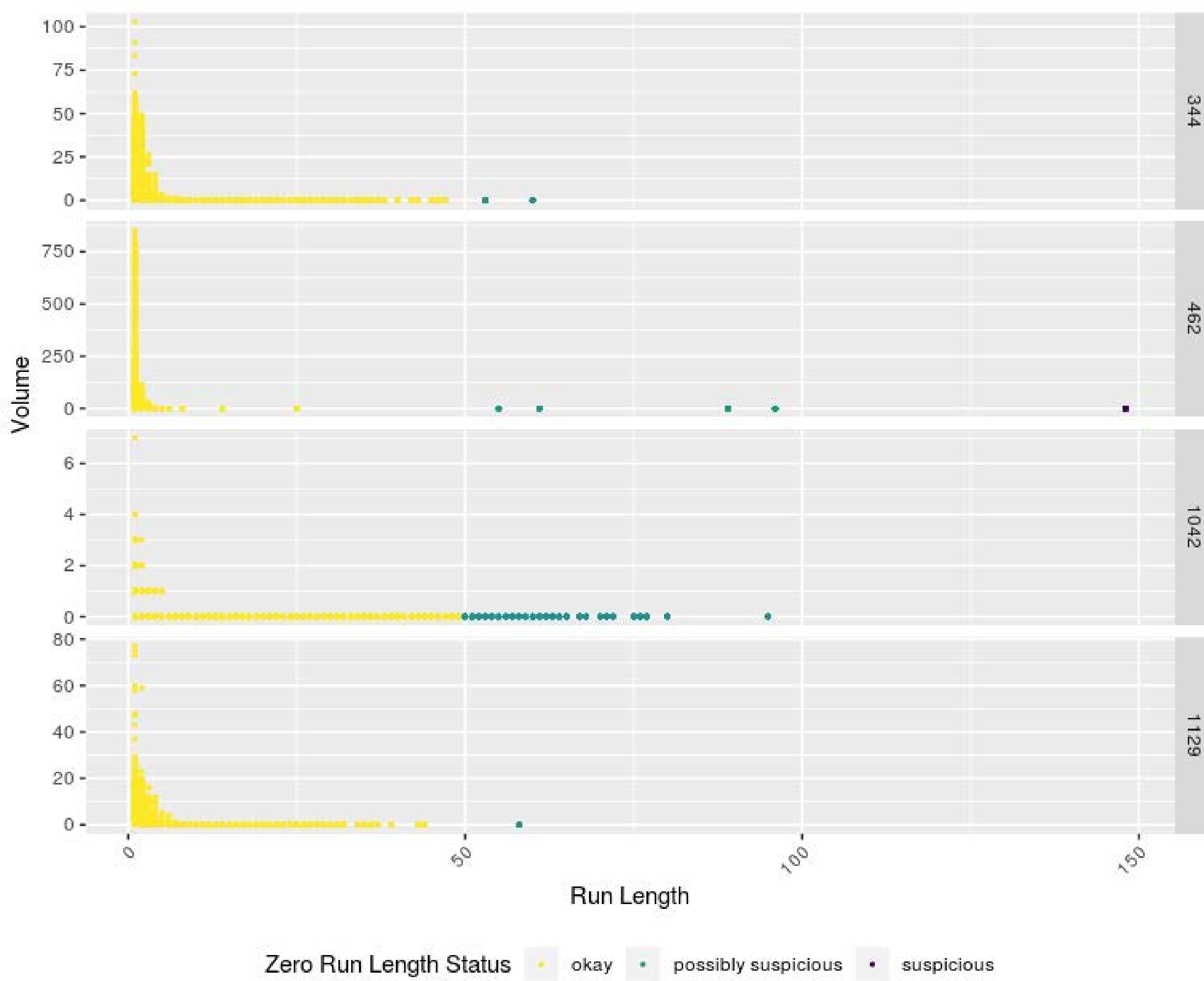

Figure 4-3 Application: Runs of Zero

Figure 4-4 uses a separate set of sites to illustrate the daily maximum value check based on the trailing 27-day interquartile ratio (IQR) calculation described in Section 3.4. Although not finalized, the check demonstrates how outliers are flagged, though the sensitivity may need adjusting to reduce the flagging of unusual but valid data. In the example shown, the four counters that flagged the same day (July 4th, 2016) are at the same location (Imperial Beach Bayshore Bikeway in San Diego) counting pedestrians eastbound, bicycles eastbound, pedestrians westbound and bicycle westbound. High traffic might be expected here on a summer holiday, so the data would likely be valid.

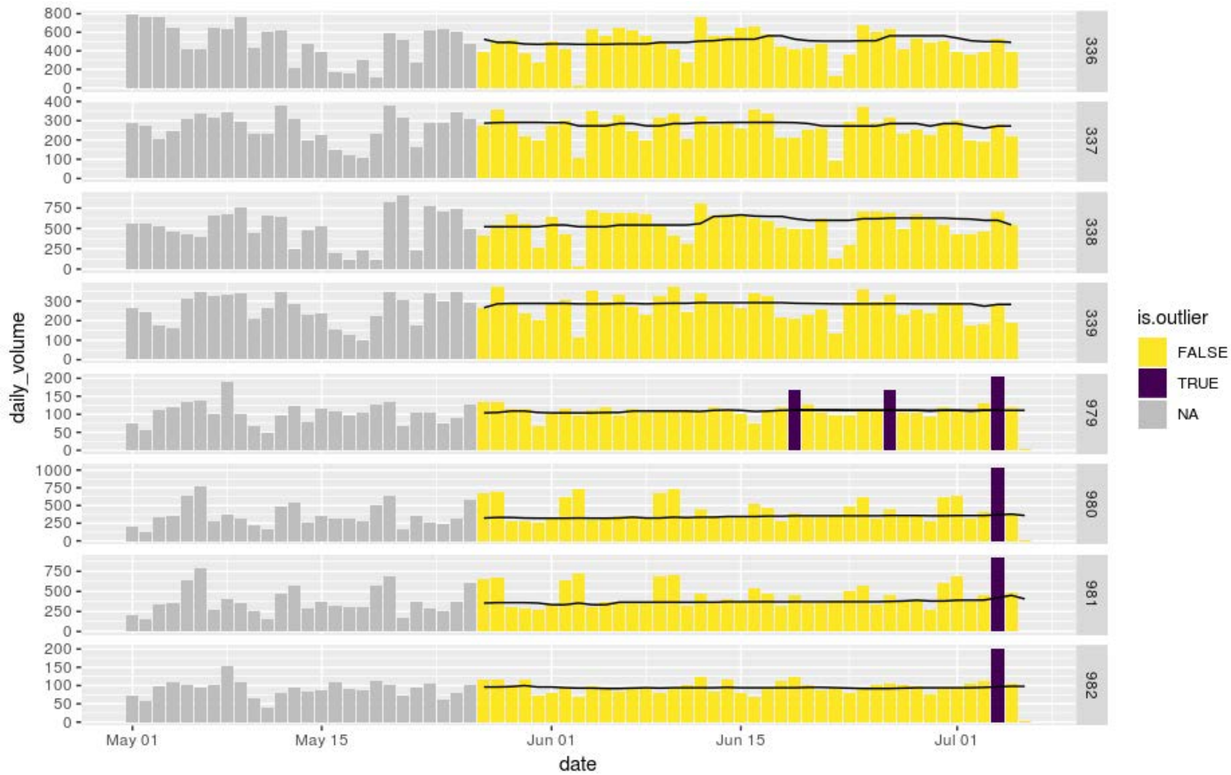

Figure 4-4 Application: Daily maximum value

4.3.2 Using known suspicious or bad data

In addition to applying the checks on datasets and years that were not used in the development of the checks, we sought to run several of the checks on a set of data that had known data quality issues. The sites were in the City of Portland, and are shown in Table 4-8. The visual outputs from the checks are shown in Figure 4-5 (zero run check), Figure 4-6 (non-zero run check), and Figure 4-7 (daily maximum value check).

Table 4-8 Sites with known data issues

Segment Name	State	Location	Detector ID	Manually identified data issue	Automated check outcome
PDX-Broadway-Bridge – Bicycle traffic	OR	Broadway Bridge	333, 334	Zero run from March 27 to April 2, 2015	Flagged as suspicious data, with 631 and 626 consecutive 0s (see Figure 4-5)
PDX-Rose-Garden-Ped-Trail	OR	Washington Park	1378	Zero run from 1/22-1/24/15 and from 3/27 to 3/29/15	Flagged as possibly suspicious (55 and 53 consecutive 0s)
PDX-Burlington-Creek Pedestrian Traffic	OR	Burlington Creek	1391	Zero run from 7/28/15 to 8/26/15 Zero run from 1/9/16 to 1/22/16	Flagged as suspicious data, with 694 and 311 consecutive 0s (see Figure 4-5)
PDX-Upper-Madison-Trail Pedestrian Traffic	OR	Washington Park	1399	Three Zero runs from 7/4/15 to 8/1/15	Flagged as suspicious data, with 256 and 188, and 187 consecutive 0s (see Figure 4-5)
PDX-Wildwood-at-MAC Pedestrian Traffic	OR	Washington Park	1422	Jump from average daily counts in the 100s to 1000s or 10,000s February 5 to 17, 2016	Flagged as suspicious in daily maximum value check (see Figure 4-7)

As outlined in Table 4-8, most of the known bad or suspicious data came from counters with excessive runs of zero counts. In most cases (i.e. counters 333, 334, 1391, and 1399) these excessive runs were caught by the check and flagged as suspicious. The zero runs that were previously identified as suspicious for site 1378 were only flagged as possibly suspicious. This is due to the fact that the dataset used to develop the check consisted of 15 minute counts, while this site provides hour long counts. Updating the check for hour counts is a known update which will be addressed before final implementation.

Site 1422 had a different know issue – namely a period of excessively high counts. These counts were appropriately flagged by the daily maximum value check. As can be seen in Figure 4-7, other sites also had some values flagged, which may have been good data, or previously unknown data issues. Under the check system, these values would be presented to the data owner to provide a decision on the validity of those counts.

Finally, the non-zero run check (Figure 4-6) identified five instances of possibly suspicious runs (counts of 15 five times in a row, 116 three times in a row, 1 eight times in a row, and 20 four times in a row).

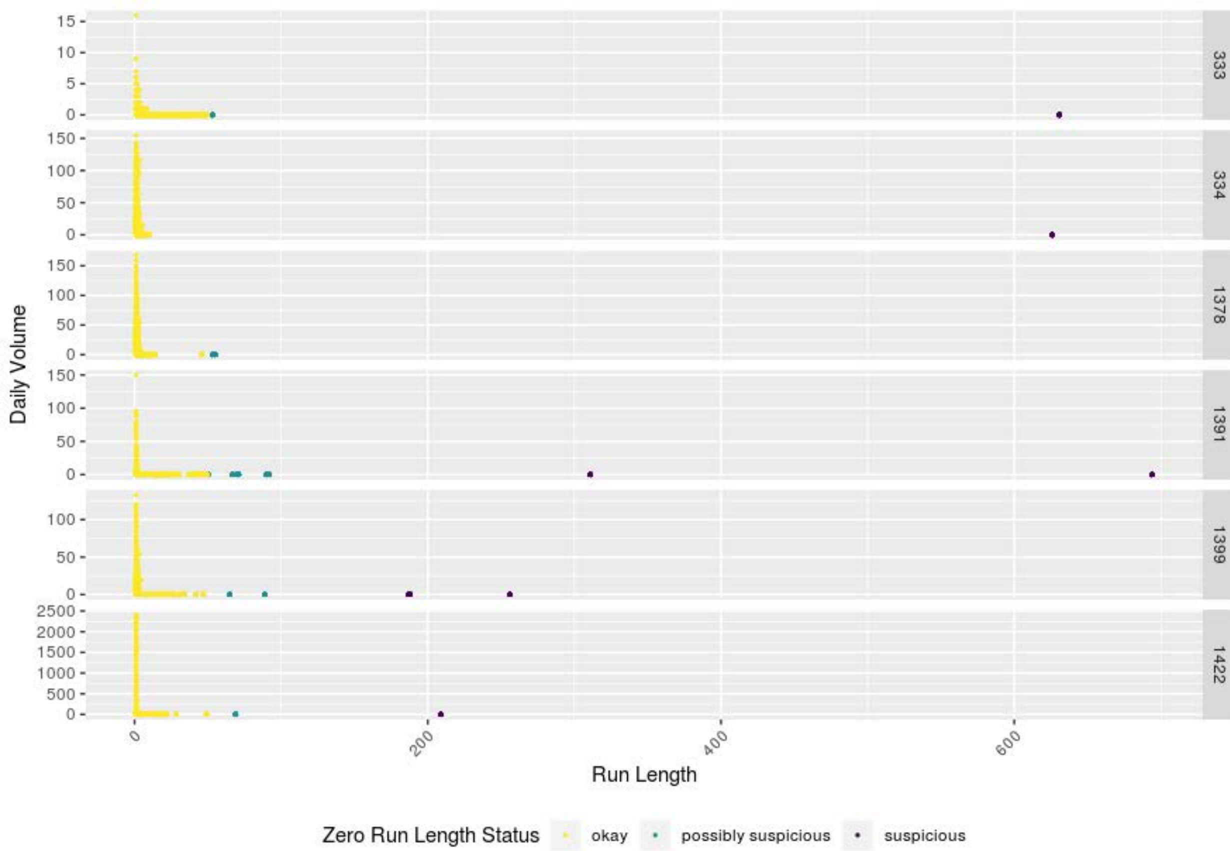

Figure 4-5 Zero run check

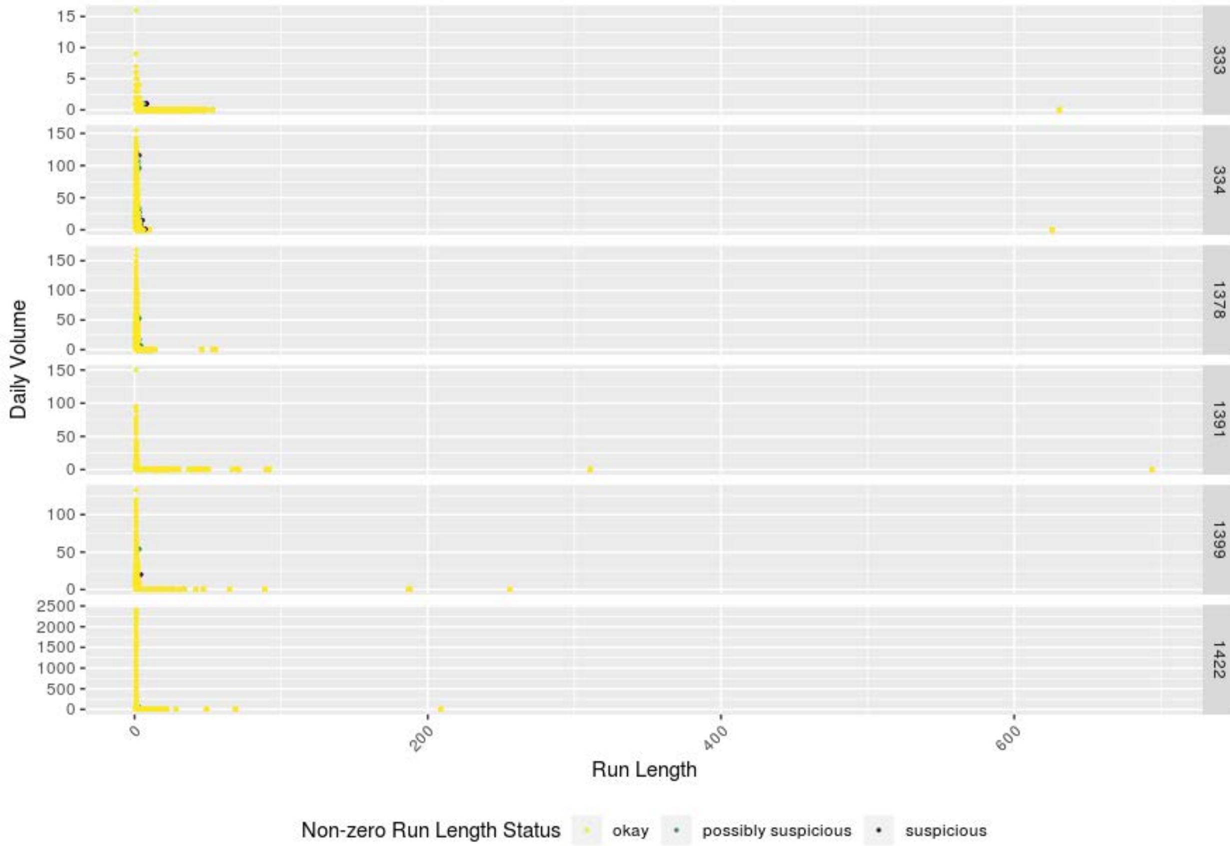

Figure 4-6 Non-zero run check

Figure 4-7 Daily Maximum Value check

5.0 IMPLEMENTATION CONSIDERATIONS

The end goal of this project is to implement data quality tests in BikePed Portal to provide the bike-ped data user with easy-to-access information about bike-ped data quality. In order to implement the recommended checks and thresholds identified in this report, a number of practical considerations for implementation need to be taken into account. This section describes the implementation framework for the checks and thresholds described in prior sections.

5.1 AUDIENCE

The end goal of the data checks and thresholds is to provide data users with data quality information. The data user will be provided with data quality information both on data download and as part of the BikePed Portal user interface. Towards this end goal, the data inputter will be provided with data flags for review. The implementation is designed to provide the data inputter with the ability to receive information about flagged data, but without receiving an overload of messages. In order to find the balance between too few and too many data flags to review, customizations are built into the implementation framework so that the data inputter can receive an appropriate level of flags for review.

5.2 STORAGE OF DATA FLAGS AND THRESHOLDS

Data flags will be stored along with each data value in the data table in the database. For each row (data value) in the database, additional columns will be added – one column will be added for each check (non-zero, zero, hard cap) to indicate if the data value was flagged as Suspicious, Partially Suspicious, or Not Suspicious for that check. An alternative method of storage considered was to store date ranges of flagged data. Storing flags for individual data values does increase storage requirements; however, storing flags for individual values is simpler and makes provision of data quality information to users significantly more straightforward.

In addition to storing data flags, additional information will be stored for each flow. For each flow, an expected daily value will be stored. The expected daily value will be stored in the Flows table in the database. Thresholds for the various checks will also be stored in the Flows table to allow for customizations to the thresholds.

5.3 IMPLEMENTATION OF DATA FLAGGING

Data flagging can be done at two different time points: on data input and periodically. On data input, the inputted data is checked and flagged and feedback is given to the data inputter at the time of input. The checks done at the time of data input look at only the data that is being currently input. In addition, data flagging is done periodically for the purpose of detecting deviations from historical trends. Periodic data flagging looks at historical data in addition to the data being flagged. Note that data input is done in a variety of ways: via manual upload, automatic upload via automatic scripts or semi-automatic upload via manually run scripts. Flagging will be implemented for all types of input.

5.3.1 Flagging on Data Input

Each time a data file is input into the database, a set of scripts will be run to flag the data based on the criteria set out in the prior section. These scripts will flag data based on the non-zero, zero and hard cap thresholds specified above. Each data item will be flagged individually and the flags for each threshold will be stored with the data item in the database. The scripts that execute flagging on data input will use only the data currently being input, these scripts will not use any (historical) data previously stored in the database. By having these scripts use only the data currently being input, we ensure efficiency of data flagging and simplicity of process. Periodic flagging, as described below, will use historical data stored in the database and will be used to evaluate changes over time.

5.3.2 Periodic Flagging

Periodic flagging will be done once each month to check for changes in long term trends and will be implemented using use scripts that execute once a month. The periodic flagging scripts will flag data per flow one month at a time. That is, on execution, for a given flow, the periodic flagging scripts will flag all months of data for that flow which are complete, but which have not been previously flagged. Thus, the scripts will first identify month-flow pairs of data that are to be flagged and for each month-flow pair combination, the scripts will compare the data for that month-flow pair to the three prior months of data for that flow and to data from the same flow in the same month over the prior three years. If the full three months or three years of data is not available, the data that is available – e.g. one or two months/years of data – will be used with an appropriately adjusted threshold. Data items that fall above or below the specified thresholds will be flagged. As with the non-zero, zero and hard cap flags, these flags (3-month and 3-year) will be stored with the data values in the database and will use the categories suspicious, possibly suspicious and not suspicious.

5.4 PROVIDING DATA QUALITY RESULTS TO INPUTTER

The end goal of data flagging is to provide data quality results to the data user. The automatic flagging process described above provides a first-pass at identifying suspicious and potentially suspicious data. These first-pass results are not intended to be the final data quality results, but are intended to be reviewed by the data inputter. That is, flags are designed to identify when human attention is needed to view suspect data. Data that is likely to be “bad” should be flagged; however, if flags are too frequent or common, reviewing those flags may become overly burdensome for the data inputter. The goal would be to have a tolerable number of flags be presented to the data inputter for review, and to have most flagged data confirmed as bad data by the user (i.e. minimize false positive flags).

Data that has been flagged must be presented to the data inputter for review. The review process must take into account the variety of ways data is input into the database and the various time schedules for data input. Data files can be manually uploaded to the database through the web interface; data files can be uploaded to the database by using a script run by the data inputter; and finally, data can be loaded into the database with fully automatic scripts that fetch data once a day from data vendors and load that data into the database. For each of these various data input methods, a mechanism for providing the data flags to the inputter for review is required. There are two pieces to this mechanism: the data flag report (for review) and how that report is provided to the inputter.

The data flag report will be web-based and will give the inputter/reviewer a graphical, high-level report of data quality as well as a detailed data item-by-data item report of data quality. The inputter/reviewer will be able to indicate which data values that have been flagged as Suspicious or Possibly Suspicious are actually valid and should be used in data analysis. The graphical report will display one month of daily count values on a line chart backed by a color-coded bar chart displaying the amount of suspicious and possibly suspicious data for each day. The data inputter/reviewer is intended to use this chart to identify time periods of concern (high levels of suspicious or possibly suspicious data) for further review. The detailed data item-by-data item report will be a table showing dates, data values and data flags. The inputter/reviewer will be able to specify that ranges of data are valid or invalid and hence should or should not be used in data analysis. The inputter/reviewer will not be able to change the original data flags, those data flags will be kept for reference and for use by researchers and other data users.

The inputter/reviewer will specify ranges of valid/invalid data by inputting a list of date ranges for valid/invalid data into a web interface. As with the data flags, the valid/invalid classification will be stored as a separate column in the data table in the database.

If the inputter/reviewer does not provide a valid/invalid classification for a data item, the following defaults will be used for BikePed Portal data analysis: data that has no flags is by default valid and will be included in analysis, data having only possibly suspicious flags is by default valid and will be included in analysis, data with one or more suspicious flags is by default invalid and will be excluded from analysis.

To avoid overwhelming the data inputter/reviewer, the implementation has been designed to enable tracking the number of flags that are generated, the percent of data that are deemed to be bad data, and the percent that are marked as good data. By storing data flags and the invalid / valid classification in the data table, reports can be generated that communicate what percent of data is being flagged by each check. In addition, by comparing flags to valid / invalid classifications, reports can be generated to identify which flags are associated with the most valid/invalid data. These reports are a practical method for determining if flags are generating high or low levels of false positives and false negatives. For the initial implementation, these reports will be manually generated using database queries. The results can be used to adjust the thresholds for each flow.

5.5 USER INTERFACE

The data users will be provided with data quality information in the BikePed Portal user interface as well as when they download data. The information provided to users will be similar to the information provided to the inputter/reviewer.

When downloading data, the data flags and valid/invalid classification information will be included for each data value. That is for each data item, the user will receive a date, a count value and a set of flags and the valid/invalid classification. Upon download, users will be provided the ability to select data meeting certain criteria (i.e. only valid data). The data flag storage described in Section 5.2 will be the basis for providing data flags to users for downloaded data.

For the user interface, graphics will be used to communicate data quality as well as quantitative percentages. As described in the section above, plots of data values will include color-coded bar charts that indicate the level of data quality for each point in the plot. In addition, each plot will be accompanied by text indicating the percentage of valid data in the plot.

5.6 LABELING SITE DAILY VOLUME CATEGORIES

Daily volume categories – low (<100), medium (between 100 and 500) and high (> 500) will be stored in the database. The expected data volume category will be stored in the flows table in the database. The data volume category will be based on a calculation of average daily count. The inputter will be asked to review the data volume category and update as necessary. Each month a new average daily count value will be calculated and will be compared with the existing daily volume category and if a significant discrepancy exists, the user will be asked to verify the category.

5.7 ADJUSTING THESHOLDS

The thresholds described in this report are based on a review of data from continuous counts in the BikePed Portal for the years 2015 to 2016. It is possible that the threshold will need to be adjusted up or down, either for all sites or for specific subsets of sites.

These adjustments might be needed if the flags were capturing so many data points as to make the checking process overly cumbersome to the inputter. However, at the same time, some false positive flags (e.g. counts flagged as suspicious or potentially suspicious that turn out to be good data) are desirable in that they are an indicator that the threshold is set at a level that is not allowing too many bad data points to slip through the filter. The implementation framework has been designed to allow adjustments of thresholds and to support analysis to inform threshold adjustment.

5.8 ADDITIONAL CHECKS

The BikePed Portal implementation framework has been specifically designed to allow customization and the addition of new checks. Section 3.3.1 describes two additional, valuable data quality checks which are expected to be implemented in the future in BikePed Portal. The database structure laid out in the sections above, especially the storage of the flags in the data table in the database and the implementation of flagging both periodically and on input, will support the addition of new checks such as those proposed in Section 3.3.1 with limited modifications to the existing system.

5.9 FUTURE MEASURES

As with all research studies, this exploration into potential quality checks for nonmotorized traffic counts opens the door to more research questions. This include investigation into if there are differences between pedestrian and bicycle counts in terms of how to check quality.

Future work will incorporate weather data, events and street closures into expected counts and thresholds. Weather, event and street closure data will be stored in the database.
In the near future, we plan to incorporate the median / interquartile range based adaptive running thresholds to identify site / location specific high and low count thresholds based on day and hourly count.

6.0 REFERENCES

Minge, Erik, Cortney Falero, Greg Lindsey, Michael Petesch, Tohr Vorvick. Bicycle and Pedestrian Data Collection Manual. Minnesota Department of Transportation Report Number MN/RC 2017-03. January 2017.

Nordback, Krista, Dylan Johnston and Sirisha Kothuri. Optimizing Short Duration Bicycle and Pedestrian Counting in Washington State. Report WA-RD 875.2. December 2017.

Nordback, K., Kothuri, S., Petritsch, T., McLeod, P., Rose, E., & Twaddell, H. (2016). Exploring Pedestrian Counting Procedures. Washington, D.C.: Federal Highway Administration.

Nordback, K., Tufte, K., Harvey, M., McNeil, N., Stolz, E., and J. Liu. “Creating a National Nonmotorized Traffic Count Archive. Process and Progress”. Transportation Research Record: Journal of the Transportation Research Board, Jan 2015, Vol. 2527, pp. 90-98. https://doi.org/10.3141/2527-10

Nordback, K., Tufte, K., McNeil, N., Harvey, M., and M. Watkins. 2016. “Bike-Ped Portal: Development of an Online Nonmotorized Traffic Count Archive” Final Report. National Institute of Transportation and Communities. http://nitc.trec.pdx.edu/research/project/817

Oddgeir Andersen, Vegard Gundersen, Line C. Wold & Erik Stange (2014) Monitoring visitors to natural areas in wintertime: issues in counter accuracy, Journal of Sustainable Tourism, 22:4, 550-560, DOI: 10.1080/09669582.2013.839693

Ryus, Paul; Ferguson, Erin; Laustsen, Kelly M.; Schneider, Robert J; Proulx, Frank R; Hull, Tony; Luis Miranda-Moreno. (2015) “Guidebook on Pedestrian and Bicycle Volume Data Collection”. NCHRP Report 797. ISBN 978-0-309-30826-7 | DOI 10.17226/22223

Ryus, Paul; Ferguson, Erin; Laustsen, Kelly M.; Schneider, Robert J; Proulx, Frank R; Hull, Tony; Luis Miranda-Moreno. (2015b). Methods and Technologies for Pedestrian and Bicycle Volume Data Collection.

Turner, Shawn. 2002. Defining and Measuring Traffic Data Quality. White Paper on Recommended Approaches. Transportation Research Record: Journal of the Transportation Research Board, No. 1870, TRB, National Research Council, Washington, D.C., 2004, pp. 62–69.

Turner, Shawn. 2007. Quality Control Procedures for Archived Operations Traffic Data: Synthesis of Practive and Recommendations. Texas Transportation Institute.

Turner, Shawn; Lasley, Philip. (2013). Quality Counts for Pedestrians and Bicyclists: Quality Assurance Procedures for Non-Motorized Traffic Count Data. Submitted for presentation and publication for the Transportation Research Board’s 92nd Annual Meeting.