EXPLORING THE CONSISTENCY OF BICYCLIST COUNT BETWEEN TYPICAL PERMANENT COUNTER DATA AND EMERGING CROWDSOURCING DATA

Seyedbamdad Sharifiilierdy, College of Engineering, California State Polytechnic University, Pomona, CA 91768, 949-519-9846, seyedbamdads@cpp.edu
Wen Cheng, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-2957, wcheng@cpp.edu
Yasser Salem, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-4312, ysalem@cpp.edu

ABSTRACT

There are typical methods for collecting bicyclist counts like permanent counter count. However, some emerging method utilizing crowdsourced data, (e.g. StreeLight) also benefits from some advantages. This paper aims to conduct a consistency checking between StreetLight count data and permanent counter data. The differences and similarities between the two types of data was evaluated through a number of statistical methods such as T-Test, Wilcoxon Test, Linear Regression R-Squared, Pearson's Correlation Coefficient, Spearman Correlation, and Kendall's Tau Correlation. The results of these tests and measures reveal that StreetLight can be a suitable alternative for permanent counter in terms of consistency.

Keywords: Bicyclists; Crowdsourced data; StreetLight; Permanent Counter Count; Association.

INTRODUCTION

Active transportation is that one can travel relying on human energy, which basically includes walking and biking. Active transportation can bring a wide range of advantages, namely reduce fatness as well as different chronic conditions such as diabetes and reduce the possibility of crimes in neighborhoods because of increased number of people in the neighborhoods (DoT, 2015; Schlossberg et al., 2012). Additionally, by emphasizing more on active transportation, transportation costs and required maintenance of roads will be decreased contributing to economic improvement (Litman, 2013; Litman, 2015). Another worth mentioning merit of active transportation is the reduction in fuel consumption (e.g., gasoline or diesel) and electricity as well (Hong, 2018). Concerning the above-mentioned benefits, it is necessary to ensure that bike paths and sidewalks will operate properly under different conditions and volumes, in order to keep them convenient for people to use. Therefore, finding the appropriate method of collecting data, or in other words, the number of people who use this mode of transportation, needs to be addressed. Contributions of this data can lead us to a better planning of transportation in many different aspects such as predicting the volume of bicyclists and pedestrians, modifying diverse transportation models, development of multimodal safety performance functions, to name but a few [Cheng et al., 2018a; Cheng et al., 2018b).

Counting methods for pedestrians and bicyclists have been experiencing many changes throughout the past decades. Generally, they can be classified into traditional and emerging methods. Manual data collection, one of the traditional methods of collecting data, is based on the field observations as well as recorded videos (Somasundaram et al., 2009). In addition, it includes active data which can be achieved from different surveys such as American Community Survey or global positioning system-oriented surveys and passive data like manual or automated counting. Although manually collected data has some privileges like being more precise and more available compared to automated data collection, it suffers from such problems as limitation of the locations which data need to be collected, more labor effort (data collectors), more safety risks as well as human error, and fewer sample collection (John and Johnson, 2000; Toth et al., 2013). Automated data collection mostly utilizes loop detectors and cameras linked with computer vision technology (Anderson, 1970; Uke and Thool, 2013). However, loop detectors need to be installed under the road surface and be maintained properly, and in addition the capability of counting pedestrians or bicyclists is not sufficiently precise (Han et al., 2009). Similarly, cameras include some drawbacks like being expensive and susceptible to different weather circumstances (Fries et al., 2007). While both manual and automated data collection have their own merits and demerits, by the development of newer methods, many cities have started counting pedestrians and bicyclists relying on crowdsourced data.

Crowdsourced data is generally classified to two subgroups, active and passive. Active data is collected from different devices like smartphones, tablets, and smart watch devices, which run specific apps actively such as various fitness-related apps, bike-share-related apps, to name a few. Active data can be collected based on whether an individual want to provide data or not. On the other hand, passive data is provided from programs that are passively running in the background. Location-based services (LBS) and mobile phone positioning (MPP) are among the different types of passive data collection. Considering these points, a more thorough data collection can be achieved by utilizing passive data.

Even though emerging methodologies have shown some benefits like efficiency and accuracy, to the authors' best knowledge, there is no consistency checking between typical data collected from loop detectors or permanent counters and the data collected from StreetLight (SL). Data that is collected from StreetLight contains the data of its location-based service, and in addition it identifies the location, direction, and speed of the device that is being collected by utilizing global positioning systems (GPS) features (StreetLight Data, 2021). This paper aims to fill this research gap by conducting a comprehensive consistency checking between crowdsourced data which is collected by StreetLight and permanent counter data available by Portland State University's Active Transportation Database and the City of San Jose. First, paired t-test and paired Wilcoxon signed-rank test was conducted to determine statistical difference between the two types of count data. Second, to identify linear association of the two types of data, R-squared and Pearson's Correlation Coefficient are computed. Finally, by assuming a non-linear relationship between the two types of data, Spearman's Correlation Coefficient and Kendall's Tau tests were performed for association checking.

DATA DESCRIPTION

The data used for this paper gathered from different sources. The first dataset obtained from the Location Based Services (LBS) crowdsourced data or in other words, StreetLight. StreetLight dataset includes bicyclists counts and are collected using smartphones' location-based services as well as different references. The authors collected this data based on the data availability of Portland State University's Active Transportation Database and the City of San Jose for each particular zone to be comparable with the mentioned data sources. Following this, the data organized into year, month, weekday, and hourly and then represented in the format of average hourly volume for bicycle counts in a specific month or year. The second dataset collected from the national archive for bicycle count data which is gathered using human resources and temporary or permanent counters like loop detectors (Bike-Ped Archive, 2021). The data captured from different cities in California, the location of which includes but not limited to San Diego, San Jose, and Imperial Beach. Additionally, the original data was structured in the form of 15-minutes intervals for each counter, which was converted to 1 hour duration by the authors, to be able to compare it with StreetLight data set. The last part is obtained from a permanent bike counter which was purchased during the research period. The permanent bike counter is located on the three creeks trail between Coe and Broadway in the Willow Glen neighborhood of San Jose and is presented by the City of San Jose. This dateset can be utilized to check the quality of the SL data, as it is relatively newer than the second data set. Finally, all three data sets were gathered, and in addition some data outliers of the SL data were removed to conduct a comprehensive consistency checking between the SL data and the permanent counter data. There are a total 6403 observations for bicyclists. The detailed information of the compiled data can be found in Table 1 (Cheng, 2022).

Table 1: Detailed Statistics for the Data used for Consistency Checking between StreetLight Counts and Permanent Counter Counts

Numerical Variables					
Variables	**Description**	**Minimum**	**Maximum**	**Mean**	**S.D.**
StreetLight Calibrated Counts	Average Hourly Volume for Bicyclist of Calibrated Streetlight Data of specific month and year	0	340	19.24	25.87
Permanent Counter Counts	Average Hourly Volume Bicyclist of Permanent Counter Data of specific month and year	0	222	13.79	18.88
Categorical Variables					
Variables	**Description**	**Details of categories (frequency, percentage)**			
Year	Year in which data were collected	2018 (2181, 34.06%); 2019 (2489, 38.87%); 2020 (1703, 26.59%); 2021 (30, 0.46%)			
Month	Month in which data were collected	January (432, 6.74%); February (406, 6.34%); March (509, 7.94%); April (452, 7.05%); May (616, 9.62%); June (672, 10.49%); July (727, 11.35%); August (698, 10.90%); September (567, 8.85%); October (453, 7.07%); November (460, 7.18%); December (411, 6.41%)			
Day	Day of the week when the data were collected	Monday (725, 11.32%); Tuesday (827, 12.91%); Wednesday (815, 12.72%); Thursday (835, 13.04%); Friday (912, 14.24%); Saturday (1226, 19.14%); Sunday (1063, 16.60%)			
Hour	Hour of the day when the data were collected	12 AM (18, 0.28%); 1 AM (12, 0.18%); 2 AM (8, 0.12%); 3 AM (3, 0.04%); 4 AM (15, 0.23%); 5 AM (24, 0.37%); 6 AM (121, 1.88%); 7 AM (253, 3.95%); 8 AM (331, 5.16%); 9 AM (417, 6.51%); 10 AM (460, 7.18%); 11 AM (475, 7.41%); 12 PM (523, 8.16%); 1 PM (504, 7.87%); 2 PM (540, 8.43%); 3 PM (580, 9.05%); 4 PM (499, 7.79%); 5 PM (489, 7.63%); 6 PM (435, 6.79%); 7 PM (283, 4.41%); 8 PM (196, 3.06%); 9 PM (131, 2.04%); 10 PM (60, 0.93%); 11 PM (26, 0.40%)			

METHODOLOGY

This paper aims to conduct a comprehensive consistency checking between the data collected from StreetLight (crowdsourced data) and the data provided by the City of San Jose and the national archive. To do so, a series of statistical techniques were conducted.

T-Test

First, due to the difference in variances between the two datasets, Welch's T-Test was conducted (Welch, 1947). By doing so, it is possible to identify the statistical difference between the datasets. Equations (1) and (2) are used to perform Welch's T-Test.

$$t = \frac{m_b - m_a}{\sqrt{\frac{S_b^2}{n_b} + \frac{S_a^2}{n_a}}} \tag{1}$$

$$df = \left(\frac{S_b^2}{n_b} + \frac{S_a^2}{n_a}\right) \Bigg/ \left(\frac{S_b^2}{n_b^2(n_b - 1)} + \frac{S_a^2}{n_a^2(n_a - 1)}\right) \tag{2}$$

In the equations (1) and (2), t and df represent T-value and Degrees of Freedom respectively, m_a and m_b are the sample means, S_a and S_b are sample standard deviations, and n_a and n_b are sample sizes. In this study, sample sizes are equal, as for comparison purposes.

Wilcoxon Test

In addition to the Welch's T-Test, Wilcoxon Signed Rank Test was conducted to identify the statistical differences between the datasets (Hayes, 2021) This can be achieved through equation (3).

$$V = \sum_{i=1}^{N} \left[sgn\left(x_{2,i} - x_{1,i}\right) * R_i\right] \tag{3}$$

In equation (3), R_i denotes the rank number, x represents the count, the subscripts "1" and "2" represent SL and Permanent counter, respectively. sgn() is the sign function that is used to determine the sign of a real number, and the subscript "i" represent the observation id. N is the sample size.

Linear Regression R-Squared

To identify the linear association between two variables, R-squared method was conducted. The association is calculated based on the linear regression from the two datasets (Frost, 2021). First, the sum of squares of residuals and the total sum of squares are calculated with equations (4) and (5). Then, R-squared is calculated by deducting the ratio between the mentioned sum of squares from 1, which is shown in equation (6).

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{4}$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{5}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6}$$

In these equations, SS_{res} and SS_{tot} are the sum of squares of residuals and the total sum of squares, respectively. The subscript "i" represents the observation number, y is the value of i^{th} observation, f_i is the result of the equation of the line of best fit for a given x value, y^bar denotes the mean of y values, and R^2 is R-squared value (Steel, 1960).

Pearson's Correlation Coefficient

Pearson's Correlation Coefficient can be used to measure the correlation between two datasets. Pearson's Correlation Coefficient is calculated with equation (7), resulting in a value between -1 and +1, which +1 indicates a perfect positive correlation between two variables and -1 show perfect negative correlation (Sedgwick, 2012).

$$r = \frac{\sum(x - m_x)\left(y - m_y\right)}{\sqrt{\sum(x - m_x)^2 \sum\left(y - m_y\right)^2}} \tag{7}$$

In equation (7), m shows the mean for x and y variables.

Spearman Correlation

Spearman Correlation assumes ordinal relationship between the datasets, in contrast to Pearson's Correlation which identifies linear relationship between the variables (Lehman, 2005). Unlike Linear Regression R-Squared and Pearson's Correlation Coefficient, the Ordinal association within the variables was evaluated with Spearman's Rank Correlation Coefficient in this study.

$$\rho = \frac{\sum(x' - m_{x'})\left(y_i' - m_{y'}\right)}{\sqrt{\sum(x' - m_{x'})^2 \sum\left(y' - m_{y'}\right)^2}} \tag{8}$$

In equation (8), m denotes the mean, and x' and y' are the ranks of x and y, respectively.

Kendall's Tau

Similar to spearman's Rank Correlation Coefficient, Kendall's Rank Correlation Coefficient is a popular method for identifying the ordinal association between the two datasets. The result ranges from +1 to -1, where +1 shows positive correlation and -1 indicates negative correlation between the two variables. To calculate Kendall's Tau, the equations (9), (10), and (11) can be used (Muñoz-Pichardo et al., 2021).

$$n_c = num(y_j > y_i) \tag{9}$$

$$n_d = num(y_j < y_i) \tag{10}$$

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{11}$$

In the equations, num() is the function that counts the satisfied observations for each criterion, and n is the number of counts.

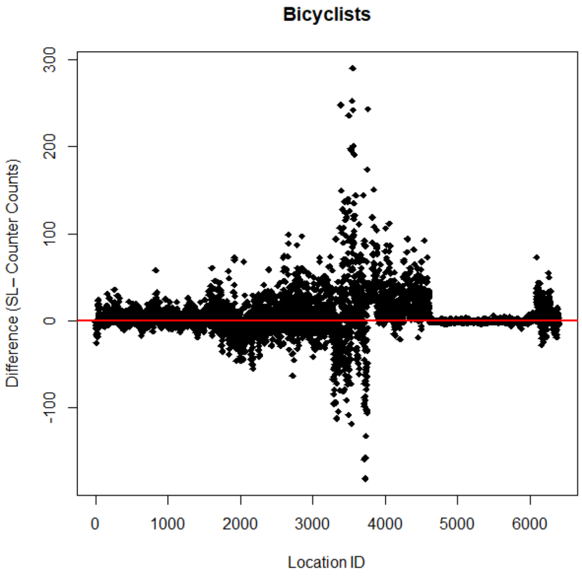
RESULTS

The main purpose of this paper is to monitor the consistency between the counts from permanent counters and SL. To achieve this goal, a number of statistical tests with different assumptions (linear or non-linear) was conducted. The results of these tests are illustrated in the following tables.

To assess the difference between the counts, both paired t-test and Wilcoxon test was utilized. In Table 2, all measures of t-value, degree of freedom, p-value, 95% confidence interval, and mean of difference are shown. It is recognized that SL and permanent counter counts are statistically significantly different, considering the p-value, or 2.2e-16. Nevertheless, the mean of difference is not relatively large. Figure 1 visualizes the difference between the counts (SL – Counter). The difference ranges from -200 to 300, and the larger values are between the location IDs of 3000 and 4000. Considering the facts that the mean of difference is not relatively large, and there are a few proportionally large differences in some locations, it can be identified that LBS-based (or SL) active transportation counts can perform efficiently in the absence of permanent counter. In addition to paired t-test, Wilcoxon signed-rank test results indicate that the two types of the counts are statistically significantly different. The V value and p-value are shown in Table 3.

Table 2: Paired T-Test Results between Streetlight and Counter Counts for Bicyclists

	t	**df**	**p-value**	**95% CL**	**Mean of Difference**
Bicyclists	19.296	6,402	< 2.2e-16	[4.903, 6.012]	5.457



Note: Points below the red line indicate that Streetlight values are greater than counter counts, where (SL -Counter counts) is negative.

Figure 1: Plot of Difference between Streetlight and Counter Counts for Bicyclists

Table 3: Paired Wilcoxon Signed-Rank Test Results between Streetlight and Counter Counts for Bicyclists

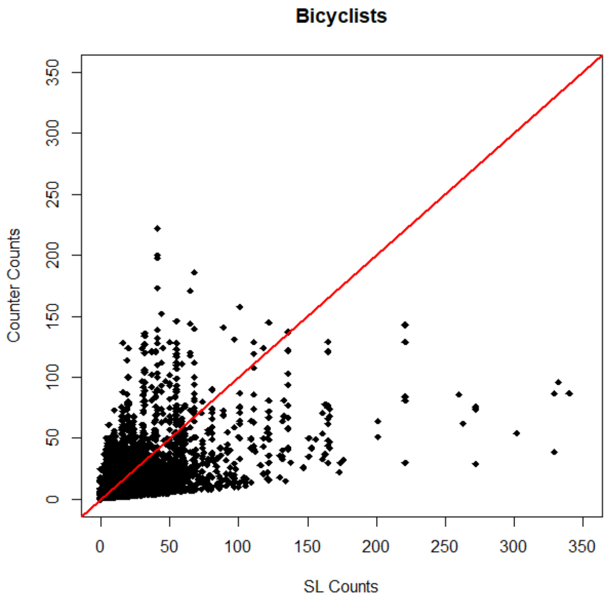
	V	**p-value**
Bicyclists	10,649,226	< 2.2e-16

Apart from checking the difference between the count types, R-squared method and Pearson's Correlation Coefficient was used to assess the linear association of the variables (SL and permanent counter counts). For R-squared method, the linear model between the count types includes one independent variable and one dependent variable, which are permanent counter count and SL count, respectively. The value of R-squared is 0.2765 as shown in Table 4. R^2 value illustrates how well the dependent variable can be explained by the independent variable. Accordingly, usually, more predictors in the linear model can generate larger values of R^2. In this study, given the fact that the linear model involves one predictor, it can be identified that the two count types have a remarkable linear association. Additionally, Pearson's Correlation Coefficient measure was used to evaluate the linear association. As shown in Table 4, it is demonstrated that the SL counts, and the permanent counter counts are positively correlated with relatively larger coefficient value of 0.5259 compared to the R-squared method. In Figure 2, the positive correlation between the two count types is visualized.

Table 4: Results of Linear Association between Streetlight and Counter Counts for Bicyclists

	R-square of simple linear test	**Pearson's correlation coefficient**
Bicyclists	0.2765	**0.5259 [0.5079, 0.5472]**

Notes: 1. The numbers in the square bracket represent the 95% confidence level for the correlation coefficient. 2. The bolded font indicates the statistical significance at the level of 0.05.



Note: The red line indicates the reference line that bisects streetlight counts and counter counts

Figure 2: Plot of StreetLight Counts vs. Counter Counts for Bicyclists

The former results show that the two count types are positively correlated, however, assuming a non-linear association in between, SL and permanent counter counts can be correlated. Hence, this paper utilized two other methods, Spearman Correlation and Kendall's Tau, to adequately evaluate the association between the datasets from a different perspective. As shown in Table 5, both values of Spearman correlation coefficient and Kendall correlation coefficient reveals the statistically significantly positive correlation between the SL counts and permanent counter counts, and in addition expresses the strong consistency and association between the two types of data.

Table 5: Results of Ordinal Association between Streetlight and Counter Counts for Bicyclists

	Spearman correlation coefficient (rho)	**Kendall correlation coefficient (tau)**

Bicyclists	**1.3498e+10** (0.691)	**59.57** (0.517)

Notes: 1. The numbers in the parenthesis represent the p-values for the correlation coefficients. 2. The bolded font indicates the statistical significance at the level of 0.05.

CONCLUSION

Considering the growing popularity of active transportation among people in the recent years and its invaluable contributions to the environment and economy, it is of vital importance to keep this mode of transportation convenient for people who wish to use it. In order to develop a comprehensive planning and designing (e.g., designing and constructing adequate bike lanes), there is a need to understand the demand. Common methods of collecting data for bicyclists are based on permanent counters such as personnel who count manually, loop detectors that count automatically, or other equipment. However, crowdsourced data has become more useful recently, in the light of its simplicity of collecting data in larger scales compared to the traditional methods. To assess the reliability of crowdsourced data, a comprehensive consistency checking between two types of data was conducted in this paper. The data collected from StreetLight for crowdsourced data, and national archive maintained by Portland State University and the City of San Jose for permanent counter count data. To explore the association and correlation between these two types of datasets, several statistical methods such as T-Test, Wilcoxon Test, Linear Regression R-Squared, Pearson's Correlation Coefficient, Spearman Correlation, and Kendall's Tau Correlation was utilized.

The findings show that the StreetLight (SL) can be an adequate substitute for permanent counter counts. Assuming different assumptions between datasets, the permanent counter count shows a significant consistency with the SL data. These assumptions consist of statistical difference and association in between through linear and non-linear relationship. However, it is worth mentioning that the above results were obtained by removing some of the data outliers to meet the criteria that defined by the authors. In addition, many data outliers were removed which constituted a big proportion of the data from City of San Jose. Overall, such findings place a great emphasis on the potential power of crowdsourced data in collecting data (e.g., SL data for bicyclists in this study). This is a brilliant method to use, considering its advantages over the common methods. Nonetheless, the data for this study was collected from different cities in California and some data outliers were removed during the project. Hence, it must be considered that more data from different cities and resources may be needed to evaluate the SL data accuracy more precisely.

REFERENCES

Anderson, R. L. (1970). Electromagnetic loop vehicle detectors. IEEE Transactions on Vehicular Technology, 19(1), 23-30.

Bike-Ped Archive. http://bikeped.trec.pdx.edu/bp/. (March 2021)

Cheng W., Zhang Y., and Clay E. "Comprehensive Performance Assessment of Passive Crowdsourcing for Counting Pedestrians and Bikes" Mineta Transportation Institute Publications (2022). https://doi.org/10.31979/mti.2022.2025

Cheng, W., Gill, G. S., Ensch, J. L., Kwong, J., & Jia, X. (2018b). Multimodal crash frequency modeling: Multivariate space-time models with alternate spatiotemporal interactions. Accident Analysis & Prevention, 113, 159-170.

Cheng, W., Gill, G. S., Vo, T., Zhou, J., & Sakrani, T. (2018a). Use of bivariate dirichlet process mixture spatial model to estimate active transportation-related crash counts. Transportation research record, 2672(38), 105-115.

DoT. (2015, August 24). Active transportation. U.S. Department of Transportation. https://www.transportation.gov/mission/health/active-transportation#:~:text=Benefits%20of%20active%20transportation,as%20diabetes%20and%20cardiovascular%20disease.

Fries, R., Chowdhury, M., & Ma, Y. (2007). Accelerated incident detection and verification: A benefit to cost analysis of traffic cameras. Journal of Intelligent Transportation Systems, 11(4), 191-203.

Frost, J. (2021, August 25). How to interpret R-squared in regression analysis. Statistics By Jim. Retrieved September 22, 2021, from https://statisticsbyjim.com/regression/interpret-r-squared-regression/.

Han, B., Yu, X., & Kwon, E. (2009). A self-sensing carbon nanotube/cement composite for traffic monitoring. Nanotechnology, 20(44), 445501.

Hayes, A. (2021, May 19). How the Wilcoxon test is used. Investopedia. Retrieved September 22, 2021, from https://www.investopedia.com/terms/w/wilcoxon-test.asp.

Hong, A. (2018). Environmental Benefits of Active Transportation. In Children's Active Transportation (pp. 21-38). Elsevier.

John, W. S., & Johnson, P. (2000). The pros and cons of data analysis software for qualitative research. Journal of nursing scholarship, 32(4), 393-397.

Lehman, A. (2005). JMP for basic univariate and multivariate statistics: a step-by-step guide. SAS Institute.

Litman, T. (2013). Evaluating active transport benefits and costs: guide to valuing walking and cycling improvements and encouragement programs.

Litman, T. (2015). Evaluating active transport benefits and costs (pp. 134-140). Victoria Transport Policy Institute.

Muñoz-Pichardo, J. M., Lozano-Aguilera, E. D., Pascual-Acosta, A., & Muñoz-Reyes, A. M. (2021). Multiple Ordinal Correlation Based on Kendall's Tau Measure: A Proposal. Mathematics, 9(14), 1616.

Schlossberg, M., Evers, C., Kato, K., & Brehm, C. (2012). Active Transportation, Citizen Engagement and Livability: Coupling Citizens and Smartphones to Make the Change. Journal of the Urban & Regional Information Systems Association, 24(2).

Sedgwick P. Pearson's correlation coefficient. Bmj. 2012 Jul 4;345.

Somasundaram, G., Morellas, V., & Papanikolopoulos, N. (2009, October). Counting pedestrians and bicycles in traffic scenes. In 2009 12th International IEEE Conference on Intelligent Transportation Systems (pp. 1-6). IEEE.

Steel, R. G. (1960). Principles and procedures of statistics: with special reference to the biological sciences (No. 04; QA276, S82.).

StreetLight Data, https://www.streetlightdata.com/bike-pedestrian-traffic-analytics/. 2021

Toth, C., Suh, W., Elango, V., Sadana, R., Guin, A., Hunter, M., & Guensler, R. (2013). Tablet-based traffic counting application designed to minimize human error. Transportation research record, 2339(1), 39-46.

Uke, N., & Thool, R. (2013). Moving vehicle detection for measuring traffic count using opencv. Journal of Automation and Control Engineering, 1(4).

Welch, B. (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. Biometrika, 34(1/2), 28-35. doi:10.2307/2332510