Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 3, pp. 535–568
DOI : 10.47974/JSMS-964

Investigating trends in traffic count data utilizing exploratory data analysis and regression analysis

Edmund Baffoe-Twum *
Department of Engineering Management and Technology
University of Tennessee
Chattanooga
TN 37403
United States of America

Eric Asa †
Bright Awuku §
Adikie Essegbey @
Department of Construction Management and Engineering
North Dakota State University
Dept. # 2475
Box 6050
Fargo
ND 58108

Abstract

To decipher and understand data collected for any model or decision-making, a statistical investigation into a dataset is necessary to generate every unearthed information in the data through visual means or a summary of the dataset's patterns. Statistical analysis is utilized to test: normality, randomness, outliers, the relationship between the dependent and the independent variables, and possible temporal and spatial spread in the dataset. In this study, the annual average daily traffic (AADT) dataset obtained from Montana, Minnesota, and Washington were subjected to statistical investigations using descriptive statistics, regression analysis, and hypothesis testing to understand the data patterns. The results generated from exploring the various techniques indicated that the data acquired were generally random. However, the data's orderliness is not random. The data values had nearly no relationship with the collection locations, thus requiring a model that can generate

*E-mail: edmund.baffoetwum@mail.wvu.edu (Corresponding Author)
† E-mail: eric.asa@ndsu.edu
§ E-mail: bright.awuku@ndsu.edu
@ E-mail: adikie.essegbey@ndsu.edu

a precise fit for better decision-making. Data skewness varied from slight to high. Therefore, data ranged from approximating normality and nonnormality with high outliers.

Subject Classification: *62J05, 62C05.*

Keywords: *Annual average daily traffic, Exploratory data analysis, Regression analysis, Hypothesis testing.*

1. Introduction

The annual average daily traffic (AADT) is a primary element in transportation planning (Castro-Neto et al., 2009). However, the cost associated with the AADT data collection has brought about several estimation methods. Consequently, every technique developed and employed analyzes perspectives and intends to cut costs.

The funds invested in the design of transportation systems strongly correlate to the expected traffic volumes. Therefore, accurate AADT data or predicted data are the ultimate underlying condition for value for money(Castro-Neto et al., 2009). Therefore, AADT data is the primary input to drive any transportation predictive model. Unfortunately, optimization procedures to enable adequate coverage of all locations during AADT data collection by stakeholders are usually imaginary. Thus, it is sometimes difficult to justify and allocate the constrained budgets to various transportation agencies, especially those within the jurisdiction of low-volume roads.

As a result, data collected are subjected to statistical analysis to ensure the inadequate data collected bring out the needed information and trends for planning and modeling. Several studies have shown that the annual average daily traffic data collected can be extensive, and the processes of monitoring and gathering the data illustrate significant spatial variability. So, there is inherent temporal variability in traffic data. The variability can be introduced by alternative sampling procedures and estimation methods (Sharma et al., 1996; Albright, 1991; Staats, 2016; Cools et al., 2007 & 2009; Eom et al., 2006; Lui et al., 2019). The data and the spatial variability automatically translate into an influence on AADT data values obtained or estimated. Han and Song (2003) and Van Arem et al. (1997) have enumerated different techniques to generate excellent summaries for the variability in daily traffic counts. Also, the past few years have seen the development of several traffic-flow pattern predictive models with considerably enhanced prediction accuracy. It is noted that a proper understanding of data patterns and factors influencing the traffic functioning monitoring process would further improve traffic volume

prediction accuracy. The estimation of AADT data is essentially influenced by location optimization, the surveyed data, seasonal traffic pattern, and many other variables. An area's exceptional location condition and antiquity sometimes contribute to developing an unbalanced estimation of AADT data values (Chen et al., 2019). Thus, it generates disparities in traffic measurements at the targeted locations and, eventually, impacts the dataset's accuracy and stability collected for planning (Chen et al., 2019).

As a broad specialty, mathematics has a branch termed statistics that dispenses with development and procedure applications to collect, organize, analyze, summarize quantitative data, interpret, and present to make meaning and draw conclusions (UCI department of statistics). Statistics adopts the probability theory in estimating population parameters and makes predictions for the future. As a result, various data (univariate, bivariate, or multivariate) are dealt with by applying statistical techniques to show the intrarelationships or inter-relationship (UCI department of statistics).

Several information sets may be gathered during the collection of AADT data. However, this study's interest is the single variable AADT data. Therefore, the independent variable of interest is where AADT data were collected. In contrast, the AADT values obtained are the dependent variable. To know the impact levels of the AADT data collected at these locations, simple linear regression is explored to determine the dataset's trends.

Datasets from three states, Montana, Minnesota, and Washington, are utilized in this study. Each state has definite locations for AADT data collection, except for the state of Minnesota. The state of Minnesota varied data collection sites yearly. Montana and Washington's state AADT data collection locations remained almost the same; additional data collection sites were added yearly. This study concentrated on AADT datasets of 400, 500, 1000, and 2000 vehicles per day. This study intends to use statistical techniques to determine trends in the AADT datasets. Descriptive statistics, hypothesis testing, and regression analysis are exploited to examine the data set for inter and intra relationships trends. The process determined the technique that best explains the dataset for further investigations. In addition, hypothesis testing is used to test the statistical significance of data randomness.

2. Previous Studies

Traffic data collection has seasonal variabilities; therefore, seasonal patterns would need to be considered during the data analysis. According

to Stolwijk et al. (1999), these seasonal patterns can be studied with statistical tests. However, the tests necessitate retaining the information connected to the periods by describing seasonal patterns. With the ban on utilizing inferential statistics in concluding the analysis of datasets by their authors, the editors of basic and applied social psychology (BASP) have confirmed using descriptive statistics as the only means to communicate research results (Fricker Jr. et al., 2019). According to the publication, the descriptive statistics ensured that authors do not overstate conclusions beyond what the data analyzed will support if statistical significance is considered (Fricker Jr. et al., 2019).

Mishra et al. (2019) used descriptive statistics and normality tests to investigate their publication data. According to the authors, simple summaries of sampled data and measures are provided when data is exploited with descriptive statistics and normality tests. Nevertheless, quantitative data is best described by utilizing central tendency and dispersion measures. Mishra et al. (2019) confirmed that the central tendency and statistical methods for data analysis are essential to conducting a normality test. The parametric test is adopted when the data is normally distributed. However, nonparametric methods are employed when data is not normally distributed, making it possible to compare data (Mishra et al., 2019). Jato-Espino (2019) implemented descriptive statistics to model the relationships of the variables of elevation, land cover, and solar radiation, and inferential statistics explored spatiotemporal patterns of the urban heat island effect.

Rúa et al. (2019), to understand and determine the number of indicator variables necessary to designate the tendency to coal combustion, employed descriptive statistics, typical values detection, principal component determination, cluster analysis, and logistic regression. In a data analysis-related research, Weare (1979), using hypothesis testing, confirmed the general hypothesis of the relationship between ocean surface temperatures and the Indian monsoon statistically. Weare (1979) used rainfall and sea level pressure data from 1949 to 1972. Dushoff et al. (2018) discussed the misinterpretation and misuse of the null hypothesis's significance testing. Therefore, they suggested it is easier to interpret and explain statistical tests when the thought process is completed using 'statistical clarity.' Balakrishnan and Wasserman (2019) considered the goodness of fit testing and Hölder density with exponent $0<s\leq 10<s\leq 1$ as a statistical evaluation tool to analyze their data. The process ensured data from a quantified distribution against an alternative separated from the null in the total variation metric. The goodness of fit testing was adopted

for the discrete data with a possible growing or an unbounded number of null distribution categories. A Hölder density with exponent $0<s\leq 10<s\leq 1$ was adopted for the continuous case with possibly unbounded support.

Dul et al. (2020) presented a statistical significance test for their research's necessary conditions. The test was completed to ensure the assumed randomness test helped the research not make Type 1 errors and draw false-positive conclusions. Instead, appraise evidence against the null hypothesis of an effect due to chance. However, Fushiki and Maeda (2020) asserted that survey data should often be analyzed with regression analysis. In their publication, Ying et al. (2017) utilized linear regression methods to analyze eye data correlation. The process described and demonstrated the appropriateness of linear regression techniques for their data. In the publication, Wright (1995) reviewed the interpretation of model coefficients, test hypotheses, and data classification from actual research studies. The author, in the process, adopted a logistic regression analysis. Jin and Lee (2017) adopted logistic regression in a railway accident statistical study. Kleijnen (1995) investigated the system dynamics model in coal transportation using the statistical techniques of regression analysis and experiment design.

Nui et al. (2016) used statistical regression analysis to study predisposing factors for anorectal disease. According to Razza et al. (2019), "ordinary linear regression (OLR) is one of the most common statistical techniques used in determining the association between the outcome variable and its related factors." Cool et al. (2009) stated that reliable forecasts of travel performance, traffic functioning, and traffic safety evidence are essential to enacting governments' efficient traffic policies. Conclusion: Several methods determine variability in data collection and correction approaches to resolving the variabilities temporarily and spatially (Cool et al. 2009).

In their publication, Chen et al. (2010) asserted that the ever-increasing size of spatial data significantly demands the capability to unearth valuable but unreserved information from the dataset. The authors' interest was to generate a statistical approach capable of detecting spatial outliers from a dataset. The detection of spatial outliers makes it possible to uncover data values different from spatial neighbors. "In contrast to traditional outlier detection, spatial outlier detection must differentiate spatial and non-spatial attributes, and consider the spatial continuity and autocorrelation between nearby samples" (Chen et al. 2010). According to Tobler (1979), "Everything is related to everything else, but nearby things are more

related than distant things." Chen et al. (2010) classified the spatial outlier detection methods into local and global-based approaches.

Precise estimates of AADT data are essential for day-to-day operations at the various transportation department (Tsapakis et al., 2013). Therefore, several research studies have been completed to develop efficient techniques for estimating AADTs. As a result, Tsapakis et al. (2013) developed regression and Bayesian analysis-based models and used training datasets. Tsapakis et al. (2013) advocated that roadway functional class, population density, and spatial location were essential. In conclusion, the seasonal variations of the AADT dataset were best explained with a full Bayesian negative binomial model. Tsapakis et al.'s (2013) model accounted for the dataset variability between 87% and 92%. Based on the spatial regression model, Eom et al. (2006) presented a better-predicted model than ordinary regression.

3. Description of Dataset and Methodology

This section discusses AADT data acquisition, data processing, and the techniques used. The summary of the processes is shown in Figure 1. Through thorough random selection, Montana, Minnesota, and Washington were chosen for this study. The data acquired from the Minnesota, Montana, and Washington transportation departments were used to complete the statistical analysis. The statistical processes exploited were descriptive, linear regression, and hypothesis testing. Data downloaded were used because of the ease of locating, downloading, and completeness. The Dataset download was in a Microsoft Excel spreadsheet. Each state had a compiled dataset that contained several years of annual average daily traffic. In addition, 2009 to 2016 datasets were extracted and adopted for this study. A preliminary check on the dataset indicated data completeness, thus beingreliable for the studies, from each state's unzipped data file. The extracted information contained AADT data compiled for each state and was categorized into AADT volume of choice for the study. Data were categorized based on the Cornell Local Roads Program (CLRP) AADT for low volume roads. Theyincorporated the lower limit of the high AADT value. Classifications by Sharma et al., 2001; PennDOT, 2014; Apronti et al., 2016; and AASHTO 2019 were also considered. From the literature, most values limit low volume AADT values to 400 or fewer vehicles per day.

This study considered AADT values to range from greater than zero (0) to a maximum of 2000 vehicles per day. Therefore, AADT data

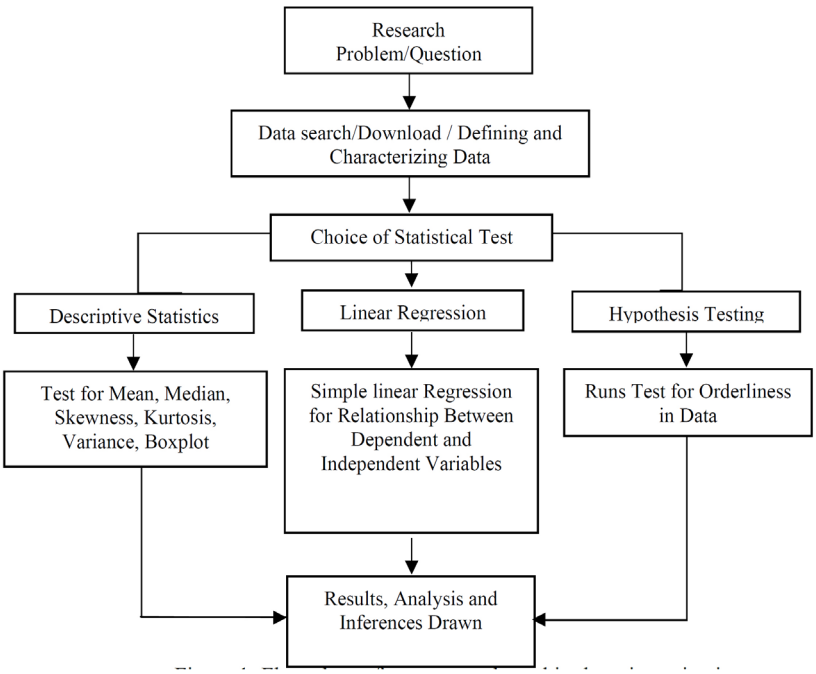

Figure 1
Flow chart of processes adopted in these investigations

considered were 400, 500, 1000, and 2000 vehicles per day from 2009 to 2016.

The cleaned data were subjected to descriptive statistics, regression analysis, and hypothesis testing analyses.

3.1 *Descriptive Statistics*

On completing the data categorization and cleaning, descriptive statistics describing and summarizing the dataset's features were explored using Minitab 18.0. The descriptive statistics were combined with appropriate graphs to explain the dataset and the trends. It helped simplify the large dataset, but some details were, at times, lost to the simplification.

Descriptive statistics differ from inferential statistics as the statistics describe what the dataset depicts. However, inferential statistics infers from the population's sample data or decides the probability and makes inferences from the dataset (McClave and Sincich, 2017). The critical

parameters were the sample mean, variance, mode, median, standard deviation, skewness, and kurtosis in these analyses. Equations 1 to 5 represent the population and sample means, variances, and standard deviations. Where n-1 represents the degrees of freedom. The mean is crucial for inferential statistics yet not resistant to outliers in a dataset. The median, which is the middle of the dataset as arranged in either descending or ascending order, is more resistant to outlier effects than inferential purposes. Finally, the central tendency (range, variance, and standard deviation) measures the dispersion or degree of clustering/ spread in the summarized dataset (McClave and Sincich, 2017).

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = u = \overline{x} \tag{1}$$

$$\text{Variance} = \sum_{i=1}^{n} \frac{(x_i - u)^2}{n} = \sigma^2 \tag{2}$$

$$\text{Standard deviation} = \sqrt{\sigma^2} = \sigma \tag{3}$$

$$\text{Variance} = \sum_{i=1}^{n} \frac{(x_i - u)^2}{n-1} = S^2 \tag{4}$$

$$\text{Standard deviation} = \sqrt{S^2} = S \tag{5}$$

Exploiting a dataset utilizing descriptive statistics primarily obtained the patterns and determined the dataset's normality. A dataset is normally distributed; if the data explicitly show no skewness (skewness equals zero), it must be symmetrical when the mean, mode, and median are equal. In contrast, kurtosis equals three (McClave and Sincich, 2017). The function for normal distribution and the standardized score (z-score) is shown in equations 6 and 7.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x-u}{2\sigma^2}} \tag{6}$$

$$z = \frac{x - \overline{x}}{\sigma_x} \tag{7}$$

3.2 *Regression Analysis*

A simple regression technique was adopted to determine the relationship between the AADT values and data collection locations. Although the regression analysis could be completed on more than two variables of interest to ascertain the relationships, in this study, the regression analysis adopted was conducted on the two variables of

interest to determine the relationship between the recording stations and the AADT values collected. In the literature, there are fifteen (15) types of regression, namely: linear regression, polynomial regression, logistic regression, quantile regression, ridge regression, lasso regression, elastic net regression, principal components regression (PCR), partial least squares (PLS) regression, support vector regression, ordinal regression, Poisson regression, negative binomial regression, quasi Poisson regression, Cox regression, and Tobit regression. When regression analysis is used as a predictive model, the relationship between the dependent (Target) and the independent variable(s) (Predictor) is explored and established. Thus, the regression analysis may predict time series models and determine the causal and effect relationship between variables (McClave and Sincich, 2017). *R*-squared (R^2) is the statistical measure of the variance's proportion for a dependent variable explained by an independent variable or variables in a regression model. The more complex the regression, the lesser the value of *R*-squared (R^2); however, the higher *R*-squared (R^2) indicates a better correlation (McClave and Sincich, 2017).

Regression analyses exist in many forms, but the basic understanding is that one dependent variable is examined to determine the impact of any number of independent variables (McClave and Sincich, 2017). As a result, it is a reliable way of categorizing which variables influence a matter of interest. Thus, it helps determine the factor of importance self-assuredly, what needs to be ignored, and the impact on each other. Therefore, it is essential to understand the dependent and independent variables (McClave and Sincich, 2017). Dependent variable: the main factor a researcher is trying to understand or predict (ADDT). Independent variables: factors hypothesized to impact the dependent variable (locations and population). In this study, the data collection locations assumed the independent variables. The linear regression equation is given as follows (McClave and Sincich, 2017):

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{8}$$

Y is the dependent or response variable, x is the independent or predictor variable, ε is the random error component. β_0 is the y-intercept of the line. When $x = 0$, the expected value of the response variable is β_0. β_1 is the slope of the line. The response (dependent) variable change is for every one-unit increase in the predictor (independent) variable.

$$E(\varepsilon) = 0 \tag{9}$$

$$E(y) = \beta_0 + \beta_1 x \tag{10}$$

3.3 *Hypothesis Testing*

Hypothesis testing is a statistical test used to determine if a hypothesis assumed for a sampled dataset from a population is correct and if there is enough statistical evidence to favor a particular belief or hypothesis about a parameter (McClave and Sincich, 2017). The inferred results showed whether the primary hypotheses are correct by comparing the null and alternative hypotheses. The null hypothesis is rejected if a predetermined significant level is not met. There are several types of hypotheses and are dependent on variables under testing.

A typical example is the chi-square test for testing univariate or multivariate datasets. The Minitab 18.0 software was used in this study to complete the hypothesis-testing. This study's test method was the nonparametric runs test purposed for continuous or discrete numeric data. The runs test determines the randomness in the order of a dataset (Minitab Express™ Support, 2019). Dataset to be selected randomly enables population representation. The runs test first determines differences in the observed and the expected number of runs. At the same time, data randomization is determined in the second step (Minitab Express™ Support, 2019). Data is sometimes not categorized as random when the observed runs are considerably greater or less than the expected run (Minitab Express™ Support, 2019). However, an inference is possible when the *p*-values are compared to the significant levels. A *p*-value less or equal to the assumed significance level of 0.05 (5%) allows for rejecting the null hypothesis, indicating that the data order is not random. The reverse holds if the *p*-value is more significant than the 0.05 (5%) significance level and fails to reject the null hypothesis, as there is insufficient evidence to conclude that the data order is not random (Minitab Express™ Support, 2019).

4. Results and Discussions

4.1 *Descriptive Statistics*

The descriptive statistics measured were meant to describe the dataset through analysis and representation. The results were on the moment (center – mean, spread – variance/standard deviation, skew-skewness, and peaked kurtosis) and non-mean-based measurements (center – mode, median, spread – range, interquartile range). In this write-up, the summary presented represents the descriptive statistics generated. Although descriptive statistics presented in either tables or graphs avoid inferences, they are good enough to infer in some cases. The two most

important concepts of concern to understand when generating descriptive statistics are the variables (AADT values in this study) and the distribution (normal and non-normal).

Each year of this study consists of varied data sizes; however, the maximum limits of the AADT data were set to counts of 400, 500, 1000, and 2000 vehicles per day. Thus, the data size was large enough to provide precise mean and standard deviation estimates. The mean and median were used to measure the central tendency of the data symmetry. The standard deviation determined how the data were spread out from the mean. Table 1 contains representative values from Montana's AADT dataset's descriptive statistics, spanning 2009 to 2016. The dataset's primary interest was mainly to determine how the data values were distributed about the data mean and serve as a guiding principle for future works. A non-normal distributed dataset predicts future datasets better when converted or approximated to normal conditions (McClave and Sincich, 2017).

One of the variables of interest is the measure of skewness in the dataset. A perfect symmetric (normal distribution) dataset has zero (0) skewness. However, the dataset rarely gets such perfection (McClave and Sincich, 2017). The dataset only approximates a symmetrical condition (normal distribution). Skewness is considered to be either positive or negative. However, a skewness value between 0.5 and -0.5 approximates a symmetric distribution in this study. In contrast, values between 0.5 and 1.0 or -0.5 and -0.1 are moderately skewed. Similarly, skewness values greater than 1.0 or less than – 1.0 are considered highly skewed.

Table 1 shows that Montana AADT of 400 and 500 vehicles per day from 2009 to 2015 is approximately symmetric. The skewness value for the year 2016 is considered moderately Skewed. Montana AADT of 1000 and 2000 vehicles per day have for all datasets under the years reviewed as moderately skewed. The datasets were positively (right) skewed. None of the values of skewness measured is considered highly skewed. Therefore, the data may not need to be transformed for further studies or predictive modeling. Besides the skewness, the interquartile range (IQR), a useful tool for determining outliers in a dataset, was explored. Outliers are values of individual observations outside of the overall pattern of the dataset (Taylor, 2020). In a dataset, the interquartile range (IQR) is defined by a five-number summary: The lowest or minimum value of the dataset, first quartile (Q1), median, third quartile (Q3), and the highest or maximum value of the dataset (Taylor, 2020). Identifying outliers using IQR demands determining any suspected outlier by adding to or subtracting 1.5X(IQR) from the third quartile (Q3) and first quartile(Q1), respectively.An outlier

Table 1
Descriptive statistics for Montana (2009-2016)-AADT 400, 500, 1000 and 2000

Statistics Montana _ AADT 400							
Variable	**Mean**	**SE Mean**	**StDev**	**Median**	**IQR**	**Skewness**	**Kurtosis**
AADT_09	176.11	3.41	113.50	160.0	200.0	0.36	-1.11
AADT_10	176.26	3.47	114.17	160.0	210.0	0.37	-1.11
AADT_11	178.03	3.42	113.49	160.0	210.0	0.33	-1.16
AADT_12	176.81	3.47	114.75	160.0	201.0	0.35	-1.17
AADT_13	174.09	3.41	113.61	160.0	200.0	0.36	-1.12
AADT_14	177.08	3.37	114.36	160.0	204.3	0.31	-1.18
AADT_15	175.51	3.37	116.40	150.0	210.0	0.34	-1.18
AADT_16	150.78	2.62	111.10	117.5	186.3	0.64	-0.84
Statistics Montana _ AADT 500							
Variable	**Mean**	**SE Mean**	**StDev**	**Median**	**IQR**	**Skewness**	**Kurtosis**
AADT_09	213.96	3.99	142.83	190.0	233.0	0.39	-1.07
AADT_10	215.77	4.06	144.28	190.0	240.0	0.38	-1.11
AADT_11	213.62	3.99	141.70	190.0	233.0	0.4	-1.04
AADT_12	216.03	4.03	143.87	193.5	247.5	0.35	-1.14
AADT_13	212.41	3.97	142.49	190.0	250.0	0.37	-1.11
AADT_14	211.35	3.89	140.86	190.0	250.0	0.37	-1.07
AADT_15	209.46	3.86	142.21	190.0	250.0	0.37	-1.11
AADT_16	182.32	3.12	140.17	139.0	228.0	0.67	-0.78
Statistics Montana _AADT 1000							
Variable	**Mean**	**SE Mean**	**StDev**	**Median**	**IQR**	**Skewness**	**Kurtosis**
AADT_09	380.15	6.55	283.64	310	480.0	0.56	-0.86
AADT_10	378.83	6.62	283.11	320	470.0	0.57	-0.83
AADT_11	383.09	6.54	282.50	320	470.0	0.52	-0.94
AADT_12	380.11	6.50	280.40	325	460.0	0.53	-0.88
AADT_13	376.84	6.50	280.48	320	460.0	0.54	-0.88
AADT_14	375.29	6.40	280.30	320	450.0	0.56	-0.86
AADT_15	372.16	6.40	283.43	310	470.0	0.58	-0.85
AADT_16	327.92	5.40	282.02	243	442.0	0.79	-0.57

Contd..

Statistics Montana _AADT 2000							
Variable	**Mean**	**SE Mean**	**StDev**	**Median**	**IQR**	**Skewness**	**Kurtosis**
AADT_09	688.50	11.10	567.30	520.0	918.8	0.68	-0.74
AADT_10	697.50	11.30	573.00	520.0	930.0	0.67	-0.77
AADT_11	697.80	11.10	569.60	540.0	912.5	0.68	-0.72
AADT_12	698.20	11.10	569.20	535.5	930.0	0.66	-0.77
AADT_13	701.70	11.10	574.40	540.0	937.5	0.65	-0.81
AADT_14	695.10	11.00	573.10	530.0	930.0	0.67	-0.79
AADT_15	681.90	10.90	570.50	502.0	920.0	0.69	-0.75
AADT_16	600.49	9.40	564.92	397.0	851.0	0.89	-0.39

is determined in data if the number is greater or less than the generated numbers.

From this background, no outliers were observed in the analyses of the datasets. Also, the approximation of the median to the mean shows the closeness of the data to normality. From Table 1, only data in the year 2016 category for all the AADT datasets explored showed much deviation from normality; thus, they were considered moderately skewed. However, the observed data indicated that the data spread was within one(1) standard deviation of the mean. The box plots in Figure 2 showed and affirmed no outliers in the AADT data from Montana for the years reviewed. Thus, for example, Figure 2 shows the highest skewed data in box plot number 4, whereas box plot number 1 showed the least and, with an approximation, the condition of normality or symmetry.

Similar processes were carried out on datasets from Minnesota and Washington states. The Minnesota state data descriptive statistics showed positively skewed values consistent with slightly, moderately, and highly skewed conditions. Skewness values for the years 2009 to 2013 for the AADT categories (400, 500, 1000, and 2000) were characterized as highly skewed. The datasets required to be transformed to about normality before using a model. Years 2014 to 2016 for AADT 400 and 500 vehicles per day have skewness approximating normality (slightly skewed), whereas AADT 1000 and 2000 for the same years were considered moderately skewed. The transformation of any datasets from 2014 to 2016 may cause the dataset to be left-skewed.

The interquartile range information generated from the Minnesota state AADT dataset indicated outliers within the dataset for all the

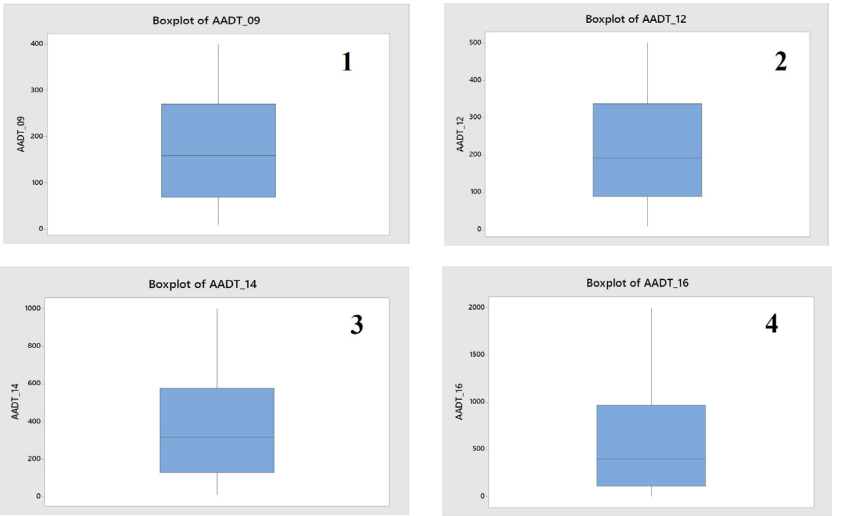

Figure 2
Representative boxplot for Montana datasets: 1. AADT- 400 for 2009, 2. AADT- 500 for 2012, 3. AADT- 1000 for 2014, and 4. AADT- 2000 for 2016

categories of AADT (400, 500, 1000, and 2000) from the year 2009 to 2013 and no outliers for all the AADT categories from 2014 to 2016. The dataset from all the categories (400, 500, 1000, and 2000 AADT) from 2009 to 2016 in Washington state approximates a normal distribution. The skewness values ranged between 0.04 and -0.15. The closeness of the skewness value to zero (0) indicated no need to transform the datasets before utilizing any further analyses. Furthermore, there were no outliers in the dataset for all the AADT data categories processed (Figure 3).

4.2 *Linear Regression*

Table 2 is the regression model summary of AADT data for Montana's 400 vehicles per day in 2014. The modeled *R* square (R^2) and *P* values explain the regression analysis. However, an attempt is made to explain other factors in the table. The degrees of freedom (DF) describes the amount of information contained in the data. The *R* square (R^2) value is the percentage of variation explained by the model. R^2 ranges between 0% and 100%. A higher percentage of R^2 indicates how well the model describes and fits the data.

In contrast, a low R^2 value indicates randomness in the data. An indication that the locations for ADDT data collection locations are

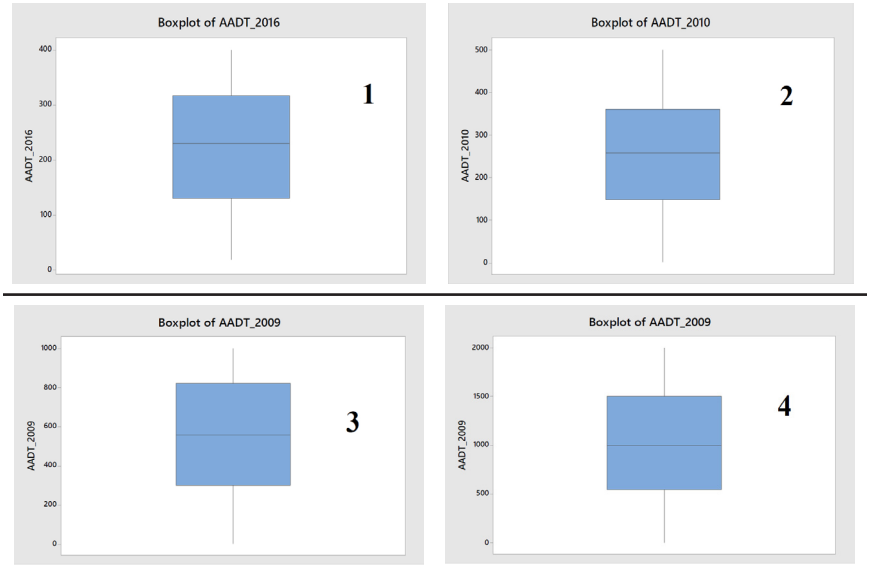

Figure 3
Boxplot for Washington state datasets: 1. AADT- 400 for 2016, 2. AADT- 500 for 2010, 3.AADT- 1000 for 2009, and 4. AADT- 2000 for 2009

randomly selected or sited. Thus, the model does not sufficiently describe or fit the data in any pattern. However, R^2 is only a single measure of model fitness. Accordingly, the model's residual plots are examined to verify if they meet the model assumptions. The model summary in Table 2 has R^2 as 1.34%, close to zero, thus, indicating a weak relationship between the data collection location and the obtained AADT data values.Also, the outcome exhibitedrandomness in the data used.

$$Df = n-k-1 \tag{11}$$

$$R^2 = 1-(1-R^2)*(n-1/n-k-1) \text{ or } 1-(SSE/SST)*(n-1/n-k-1) \tag{12}$$

The degrees of freedom relate directly to the R^2 values obtained. A very low degree of freedom of 1 relates to the low R^2. The adjusted mean squares (Adj MS), also known as the mean square error (MSE or S^2), measure how much variation a term or a model explains, assuming that all other terms are in the model, regardless of their entered order (Minitab Express™ Support, 2019; Cognos Analytics, 2021). The lower the value of the Adj MS, the better the prediction. However, the adjusted sum squared (Adj SS) is dependent on the degree of freedom, which quantifies

Table 2
A summary of the regression model for Montana (AADT- 400 for the year 2014)

Montana_ AADT_400_2014

Regression Analysis: AADT_14 versus OBJECTID Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	201337	201337	15.59	0.000
OBJECTID	1	201337	201337	15.59	0.000
Error	1148	14826070	12915		
Total	1149	15027407			

Model Summary

S	*R*-sq	*R*-sq(adj)	*R*-sq(pred)
113.643	1.34%	1.25%	0.99%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	209.78	8.94	23.48	0.000	
OBJECTID	-0.01067	0.00270	-3.95	0.000	1.00

Regression Equation

AADT_14	=	209.78 – 0.01067 OBJECTID

the amount of variation in the response data explained by the model (Minitab Express™ Support, 2019; Cognos Analytics, 2021). The adjusted sum squared error (Adj SS error) of 14826070 quantifies the variation in the data not explained by predictors; because the Adj SS is a measure of variation for different components of the model.

In contrast, the adjusted sum squared total (Adj SS total) of 15027407 is a total variation in the data. Nevertheless, the computed *P*-value and the R^2 are dependent on the adjusted sum squares. The term's coefficient represents the change in the mean response for a one-unit change in that term. An increase in the term implies a positive coefficient. Therefore, there is a decrease with a negative slope in the regression equation. At the same time, it approaches zero, thus less explained. The probability *P*-value measures the association between each term (recorder locations) and response (AADT) in the model and determines statistical significance.

A low *P*-value provides strong evidence against the null hypothesis. Thus, it helps determine the null hypothesis by comparing the *P*-value to a significant level (Minitab Express™ Support, 2019). In this regression analysis, if the null hypothesis coefficient is assumed to be zero, then there is an indication that there is no association between the term (recorder locations) and response (AADT). However, a *P*-value of 0.000, less than the significance level of 0.05 or 5%, suggests the model explains the variation in the response (AADT). Consequently, there is a statistically significant association between the response (AADT data values) variable and the term (recorder locations).

The variance inflation factor (VIF) explains how much the coefficients' variance is inflated because of the model predictors' correlations. If the VIF is equal to 1, there is no correlation. A VIF between 1 and 5 indicates a moderate correlation. A VIF value greater than 5 shows a high correlation among the model predictors (Minitab Express™ Support, 2019). In Table 2, a VIF of 1 implied no correlation between the predictors.

Figure 4 displays four different residual plots. The normal probability plot shows the residuals versus the expected values when the distribution is normal. The inverted *S*-curve pattern displayed by the normal probability plot indicates a distribution with short tails. It, therefore, violates the assumption that the residuals are normally distributed. Although there

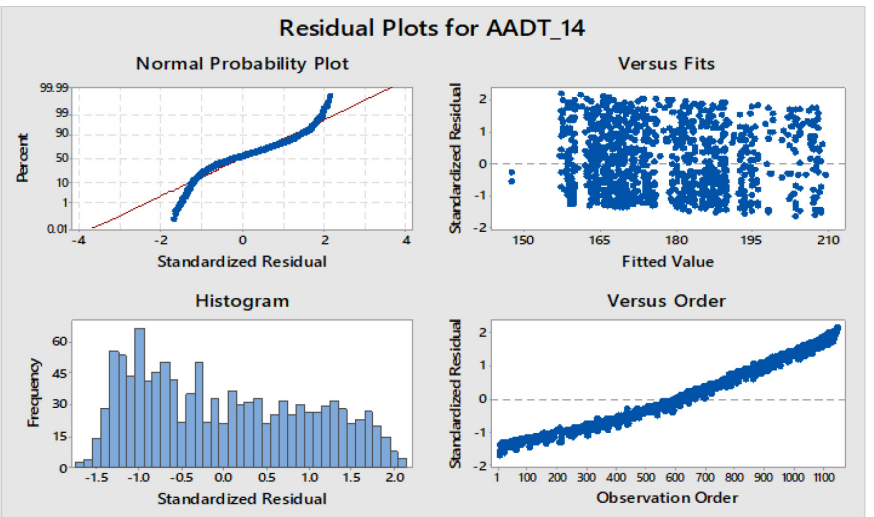

Figure 4
Residual plot representing AADT of 400 vehicles for the year 2014 (Montana)

are no outliers, skewness is considered at the lower end of the moderately skewed data. The histogram is a representation of the residuals for the observations. A visual inspection of the histogram does not indicate any significant skewness or possible outliers.

The standardized residuals versus the fits graph plot the residuals on the *y*-axis and the fitted values on the *x*-axis. The plot determines the assumption that the residuals are randomly distributed and have constant variance. It is ideally observed by the points falling randomly on both sides of the 0 line, without any recognizable patterns in the plotted points. The output pattern indicates that the model's assumptions are not met; thus, the data is not normally distributed. Finally, the residuals versus order plot explained the assumption that the residuals are independent of one another. The Independent residuals showed no trends or patterns, and points randomly fell around the centerline. The residuals versus order plot generated are dependent as no pattern or points do not randomly fall around the centerline. The result indicates no relationship and exhibits randomness in the data, thus the subsequent P and R^2 values.

According to Anderson (2016), there is no established relationship between *P*-values and *R*-squared; however, both are suitable data-dependent. The variations explained by a model are indicated with R^2; the *p*-value is about the *F* statistics hypothesis testing of the "fit of the intercept-only model and your model are equal" (Anderson, 2016). An R^2 of 0.1 implies that 10% of the data's variation is explained. If a *P*-value is less than the significance level, say 0.05, the model fits the data well. Therefore, the higher the R^2, the better the explained variation in the data.

Even so, Anderson (2016) outlines some situations:

- When a model results in a low *R*-square and low *p*-value (*p*-value < or = 0.05), the model is said to explain much less of the variation of the data; nonetheless, it is significant (it means having the model is better than not having a model)
- The worst consequence is experienced when a model has a low *R*-square and high *p*-value (*p*-value > 0.05); it indicates that the model does not explain much of the data's variation. However, simultaneously it is not significant.
- The best situation is when a high *R*-square and low *p*-value are obtained from a model. Then, the model explains many variations within the data, and it is again significant.

Table 3
Montana state regression summary (R^2 and P-Values)

State of Montana Regression Summary								
	400		500		1000		2000	
Year	R^2	P-Value	R^2	P-Value	R^2	P-Value	R^2	P-Value
2009	2.94%	0	2.87%	0	4.14%	0	4.16%	0
2010	2.72%	0	3.63%	0	3.98%	0	4.43%	0
2011	1.61%	0	2.50%	0	3.68%	0	3.96%	0
2012	1.26%	0	1.97%	0	3.41%	0	3.46%	0
2013	2.09%	0	3.03%	0	3.51%	0	4.09%	0
2014	1.34%	0	2.88%	0	2.92%	0	3.87%	0
2015	1.79%	0	3.02%	0	3.74%	0	3.92%	0
2016	5.42%	0	6.60%	0	7.99%	0	8.76%	0

- A high *R*-square and high *p*-value output from a model means that the model explains many variations within the data but is not significant. Thus, making the model worthless.

From the above situations, the Montana model summary in Table 3 indicates that much less variation in the data is explained, and it is significant for all the years and AADT levels.

For the state of Minnesota regression, the *P*-value at 0.028 (Table 4) is less than the significance level of 0.05 or 5%; the model explains the response (AADT of 2000). The R^2 at 0.86% indicates a less explanation of the relationship between the AADT and locations (Table 4). An indication of randomness in the data collection. The VIF is equal to 1; there is no correlation. The regression equation has a negative slope (- 0.001819),which indicates a decrease. The value also approached zero, thus less explaining the relationship between the AADT values and location.

The normal probability plot's downward curve (Figure 5) illustrates a long tail distribution. It violates the assumption that the residuals are normally distributed. Thus, skewness in the data is to the right or positively skewed. As indicated in the descriptive statistics, the skewness value for AADT year 2009 for AADT of 2000 vehicles per day is 4.23, indicating a highly skewed dataset. The residuals versus fit graph plots the residuals on the y-axis and the fitted values on the *x*-axis.The residual plots for the same year showed trends or patterns where some points are displayed away from others. The display is an expression of outliers, thus violating the assumption that the residuals are constant. Patterns in the points

Table 4
The state of Minnesota regression model summary (R^2 and P-Values)

State of Minnesota Regression Summary								
	400		500		1000		2000	
	R^2	P-Value	R^2	P-Value	R^2	P-Value	R^2	P-Value
2009	1.46%	0.006	1.56%	0.004	2.11%	0.001	0.86%	0.028
2010	0.47%	0.033	1.04%	0.001	0.86%	0.003	0.99%	0.001
2011	16.00%	0.000	13.96%	0.000	8.54%	0.000	4.76%	0.000
2012	0.61%	0.080	3.17%	0.000	3.36%	0.000	5.23%	0.000
2013	0.61%	0.036	0.44%	0.065	0.05%	0.503	0.71%	0.007
2014	0.00%	0.998	0.03%	0.457	0.14%	0.053	1.35%	0.000
2015	3.89%	0.000	2.12%	0.000	0.07%	0.164	0.57%	0.000
2016	1.00%	0.708	0.03%	0.468	0.02%	0.403	0.02%	0.300

indicate the dataset may need to be transformed though not independent. The residual versus order plot showed no pattern but evidence of the outliers.

The regression analyses for Washington state using data from 2009 and AADT 400 vehicles per day indicated a *P*-value of 0.00, which is less

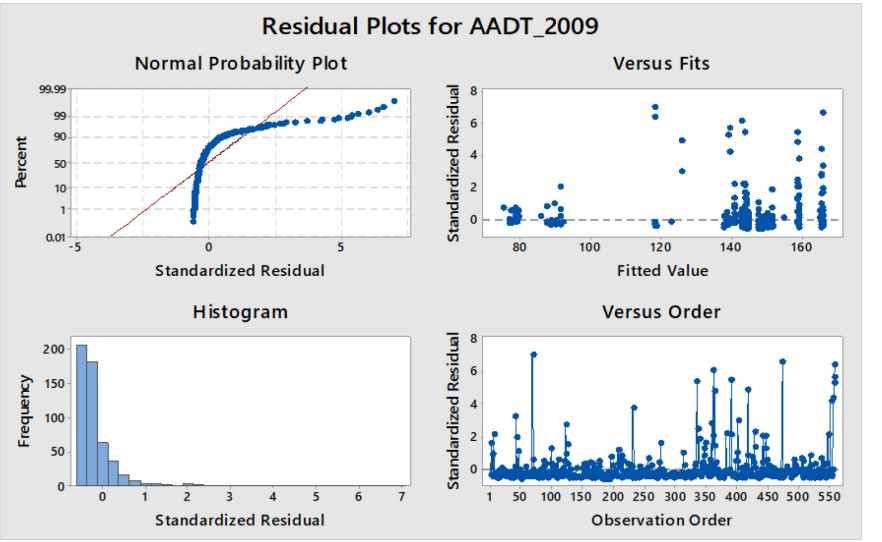

Figure 5
Residual Plot for Minnesota for AADT 2000 for the year 2009

than the significance level of 0.05 or 5%; the model explains the response (AADT) variation. The P-value relates to R^2; as the *P*-value increases, the R^2 decreases. The R^2 at 7.08% indicates a less explanation of the relationship between the AADT and locations. The output confirmed some randomness in the data collection. The VIF is equal to 1; there is no correlation. A negative slope with a coefficient of - 0.01404 for the regression equation indicated a decrease. It approached zero, thus explaining less of the AADT data values and location. The *S* value represents how far the data values fall from the fitted values. *S* is measured in the response; therefore, the smaller the *S* value, the better the model describes the response; however, it is not an indication of meeting the model assumptions. With an *S* value of 99.3746, the response does not meet the model assumptions. It describes the response as partial because the skewness approaches zero. The dataset showed slightly skewed, and maybe no need for transformation. The regression summary for Washington state is displayed in Table 5.

The model explained much less data variation for the Washington state dataset. It is not significant for all the years with low R^2 and *P*-values. For all AADT categories (400, 500, 1000, and 2000) for 2016, the *p*-values were more significant than the significance level of 0.05 (5%) for the model. Likewise, the model explained less in the data variation, which cannot be considered significant.

4.3 *Hypothesis testing*

Table 5
The Washington state regression model summary (R^2 and P-Values)

State of Washington Regression Summary								
	400		500		1000		2000	
	R^2	P-Value	R^2	P-Value	R^2	P-Value	R^2	P-Value
2009	7.08%	0.000	4.54%	0.000	4.76%	0.000	2.83%	0.000
2010	10.30%	0.000	10.54%	0.000	7.84%	0.000	5.45%	0.000
2011	8.95%	0.000	9.32%	0.000	8.07%	0.000	5.66%	0.000
2012	5.18%	0.000	8.19%	0.000	6.58%	0.000	4.00%	0.000
2013	4.71%	0.000	6.81%	0.000	6.10%	0.000	4.55%	0.000
2014	8.71%	0.000	10.32%	0.000	7.43%	0.000	4.02%	0.000
2015	9.81%	0.000	9.30%	0.000	7.36%	0.000	6.88%	0.000
2016	0.24%	0.419	0.41%	0.242	0.24%	0.221	0.22%	0.107

Table 6

Runs test summary for data order randomness for Montana AADT 400 (left) and AADT 2000 (right)

Runs Test: AADT_09, AADT_10, AADT_11, AADT_12, … ADT_15, AADT_16					Runs Test: AADT_09, AADT_10, AADT_11, AADT_12, … ADT_15, AADT_16				
Descriptive Statistics					**Descriptive Statistics**				
	Number of Observations					**Number of Observations**			
Variable	N	K	≤ K	> K	Variable	N	K	≤ K	> K
AADT_09	1110	176.108	607	503	AADT_09	2616	688.477	1526	1090
AADT_10	1083	176.259	595	488	AADT_10	2580	697.458	1508	1072
AADT_11	1101	178.027	594	507	AADT_11	2630	697.84	1521	1109
AADT_12	1091	176.808	588	503	AADT_12	2634	698.24	1533	1101
AADT_13	1107	174.091	603	504	AADT_13	2672	701.744	1558	1114
AADT_14	1150	177.078	615	535	AADT_14	2703	695.064	1579	1124
AADT_15	1193	175.509	658	535	AADT_15	2722	681.864	1601	1121
AADT_16	1802	150.782	1048	754	AADT_16	3583	600.485	2183	1400

Contd..

K = sample mean					*K = sample mean*				
Test					**Test**				
Null hypothesis	*H* : The order of the data is random				Null hypothesis	*H* : The order of the data is random			
Alternative hypothesis	*H* : The order of the data is not random				Alternative hypothesis	*H* : The order of the data is not random			
Number of Runs					**Number of Runs**				
Variable	Observed	Expected	P-Value		Variable	Observed	Expected	P-Value	
AADT_09	2	551.13	0		AADT_09	912	1272.67	0	
AADT_10	2	537.21	0		AADT_10	896	1254.16	0	
AADT_11	2	548.06	0		AADT_11	923	1283.73	0	
AADT_12	2	543.19	0		AADT_12	911	1282.57	0	
AADT_13	2	550.07	0		AADT_13	923	1300.11	0	
AADT_14	2	573.22	0		AADT_14	933	1314.2	0	
AADT_15	2	591.16	0		AADT_15	957	1319.68	0	
AADT_16	2	878.02	0		AADT_16	1138	1706.94	0	

Table 7
Runs test summary for data order randomness for Minnesota AADT 400 (left) and AADT 2000 (right)

Runs Test: AADT-2009, AADT_2010, AADT-2011, … DT_2015, AADT_2016					Runs Test: AADT_2009, AADT_2010, AADT_2011, … DT_2015, AADT_2016				
Descriptive Statistics					**Descriptive Statistics**				
			Number of					**Number of**	
			Observations					**Observations**	
Variable	N	K	≤ K	> K	Variable	N	K	≤ K	> K
AADT-2009	522	85.967	348	174	AADT_2009	560	141.054	440	120
AADT_2010	961	95.627	658	303	AADT_2010	1111	205.147	850	261
AADT-2011	1320	88.053	873	447	AADT_2011	1605	229.885	1155	450
AADT-2012	502	104.562	345	157	AADT_2012	686	352.405	483	203
AADT_2013	723	120.011	466	257	AADT_2013	1040	391.153	718	322
AADT_2014	1551	187.724	823	728	AADT_2014	3473	632.564	2120	1353
AADT_2015	1555	190.775	810	745	AADT_2015	3481	635.266	2096	1385
AADT_2016	1785	189.034	907	878	AADT_2016	4397	706.081	2562	1835

Contd..

K = sample mean					*K = sample mean*				
Test					**Test**				
Null hypothesis	*H* : The order of the data is random				Null hypothesis	*H* : The order of the data is random			
Alternative hypothesis	*H* : The order of the data is not random				Alternative hypothesis	*H* : The order of the data is not random			
Number of Runs					**Number of Runs**				
Variable	Observed	Expected	P-Value		Variable	Observed	Expected	P-Value	
AADT-2009	188	233	0.000		AADT_2009	146	189.57	0	
AADT_2010	329	415.93	0.000		AADT_2010	304	400.37	0	
AADT-2011	388	592.26	0.000		AADT_2011	416	648.66	0	
AADT-2012	169	216.8	0.000		AADT_2012	169	286.86	0	
AADT_2013	258	332.29	0.000		AADT_2013	324	445.61	0	
AADT_2014	677	773.59	0.000		AADT_2014	1294	1652.81	0	
AADT_2015	715	777.14	0.002		AADT_2015	1240	1668.89	0	
AADT_2016	781	893.26	0.000		AADT_2016	1554	2139.4	0	

Table 8

Nonparametric Runs Test for data order randomness for Washington AADT400 (Left) and 2000 (Right)

Runs Test: AADT_2009, AADT_2010, AADT_2011, ... DT_2015, AADT_2016					Runs Test: AADT_2009, AADT_2010, AADT_2011, ... DT_2015, AADT_2016				
Descriptive Statistics					**Descriptive Statistics**				
			Number of					**Number of**	
			Observations					**Observations**	
Variable	N	K	≤ K	> K	Variable	N	K	≤ K	> K
AADT_2009	234	233.004	115	119	AADT_2009	1277	1036.48	659	618
AADT_2010	316	212.668	157	159	AADT_2010	1401	993.34	720	681
AADT_2011	318	207.745	150	168	AADT_2011	1389	978.92	700	689
AADT_2012	322	205.134	155	167	AADT_2012	1412	984.57	717	695
AADT_2013	324	211.491	155	169	AADT_2013	1432	994.23	724	708
AADT_2014	316	215.484	154	162	AADT_2014	1383	998.97	694	689
AADT_2015	294	222.041	144	150	AADT_2015	1277	997.74	635	642
AADT_2016	276	225.036	134	142	AADT_2016	1194	990.26	593	601

Contd..

K = sample mean					*K = sample mean*				
Test					**Test**				
Null hypothesis	H : The order of the data is random				Null hypothesis	H : The order of the data is random			
Alternative hypothesis	H : The order of the data is not random				Alternative hypothesis	H : The order of the data is not random			
Number of Runs					**Number of Runs**				
Variable	Observed	Expected	P-Value		Variable	Observed	Expected	P-Value	
AADT_2009	110	117.97	0.296		AADT_2009	472	638.84	0.000	
AADT_2010	153	158.99	0.499		AADT_2010	499	700.96	0.000	
AADT_2011	142	159.49	0.049		AADT_2011	573	695.46	0.000	
AADT_2012	138	161.78	0.008		AADT_2012	649	706.83	0.002	
AADT_2013	140	162.70	0.011		AADT_2013	601	716.91	0.000	
AADT_2014	140	158.90	0.033		AADT_2014	646	692.49	0.012	
AADT_2015	124	147.94	0.005		AADT_2015	545	639.48	0.000	
AADT_2016	124	138.88	0.072		AADT_2016	587	597.97	0.525	

The nonparametric runs test for the data orderliness assumes the following conditions: When a P-value $\leq \alpha$: the data's order is not random (Reject H_0). "If the p-value is less than or equal to the significance level, the decision is to reject the null hypothesis and conclude that the order of the data is not random." When a P-value $> \alpha$: cannot conclude the order of the data is not random (Fail to reject H_0), similarly, "If the p-value is greater than the significance level, the decision is to fail to reject the null hypothesis (Minitab Support, 2019). There is not enough evidence to conclude that the order of the data is not random" (Minitab support, 2019). For example, From Table 6, the Montana AADT 400 (left column) and 2000 (right column) data have p-values of less than the significance level of 0.05 (5%). Therefore, the order of the data is not random; therefore, the null hypothesis is rejected. Similar results are obtained for Montana AADT 500 and 1000. Consequently, a similar interpretation of the results is given based on the conditions.

In the hypothesis testing, the observed and expected number of runs for the Minnesota AADT data of 400 vehicles per day in Table 7 (left), there is a considerable difference, thus a likely non-randomness in the data's order. Therefore, the null hypothesis is rejected based on a p-value less than the significance level of 0.05 (5%). Therefore, the order in the data acquired is not random. The results obtained for Minnesota AADT of 2000 vehicles per day are comparable. Therefore, an interpretation similar to the Minnesota AADT of 400 vehicles per day is based on the results. The result for the AADT of 2000 vehicles per day is presented in Table 7 (right). Based on the observed and expected number of runs in the Washington AADT 400 data in Table 7 (left), there exists a considerable difference, thus a likely non-randomness in the order of the data for AADT years 2012, 2013, 2014, and 2015. therefore, the null hypothesis is rejected based on the p-value of less than the significance level of 0.05 (5%). The same cannot be said about 2009, 2010, 2011, and 2016 as the p-values are more significant than 0.05 (5%). The results indicate randomness in the data and therefore fail to reject the null hypothesis.

There is not much evidence to conclude that the order of the data is not random. Therefore, the order of the data acquired is random. Similar results are not obtained for Washington AADT of 2000 vehicles per day. For AADT 2000 vehicles per day, all but 2016 have p-values less than the significance level of 0.05 (5%). The AADT for 2016 has a p-value of 0.525, which is greater than the significance level of 0.05 (5%). Therefore, there is no randomness in the AADT 2000 vehicles per day from 2009 to 2015.

The 2016 AADT of 2000 vehicles per day has a random order of data. The results for AADT 2000 are presented in Table 8 (right) and summarize the results.

5. Conclusions

This study adopted descriptive statistics and regression analysis and runs hypothesis tests to examine the AADT dataset acquired from Montana, Minnesota, and Washington. Generally, the following conclusions were drawn based on the dataset. The descriptive statistics gave summaries of the data from all three states. Essentially, most data are classified as slightly skewed and approximate normality. A few moderate and highly skewed datasets and the highly skewed datadeviated from normality. This category is associated with the data acquired from the state of Minnesota. Minnesota data exhibited outliers in the majority of the data. The spread of the dataset for Montana was one times the standard deviation, and Minnesota was less than one times the standard deviation. In contrast, the Washington data spread two times the standard deviation. The R-squared for all the states' datasets was very low and thus indicated randomness in the data. The data collection did not relate to the collection locations, emphasizing randomness and not optimized. The p-values less or equal to 0.05 or 5% indicated the model explains variation in association of the response (AADT) and the recording stations. It is also statistically significant.

There were exceptions to the p-values, especially with the Minnesota dataset and the 2016 AADT for Washington state. The VIF indicated no correlation between AADT values and data collection sites. The analyses of the residual outputs of the regression models showed no normality in the data from Minnesota and Montana. However, all the models confirmed no randomness in the order of the data. Finally, the data collection location in Minnesota was varied yearly, which may have impacted the models generated from the Minnesota datasets.

Data Availability Statements

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request. (List items)

Reference

[1] AASHTO. Guidelines for Geometric Design of Low-Volume Roads. *AASHTO Issues Second Edition of Low-Volume Roads Guidelines – AASHTO Journal* (2019).

[2] Albright, D., History of Estimating and Evaluating Annual Traffic Volume Statistics. Transportation Research Record 1305, pp 103-107 (1991).

[3] Apronti, D.,Ksaibati, K., Gerow, K. and Hepner, J.J., Estimating traffic volume on Wyoming low volume roads using linear and logistic regression methods. *Journal of Traffic Transportation*. Engineering. (English Edition). Vol. 3 (6), pp 493-506 (2016).

[4] Anderson, F., (2016). What is the relationship between R-squared and p-value in a regression? Communications on Research gate. Harrisburg University of Science and Technology (6/6/2016).

[5] Balakrishnan, S., and Wasserman, L., Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*. Vol. 47(4), pp. 1893-1927 (2019).

[6] Castro-Neto, M. M., Jeong, Y., Jeong, M. K., and L. Han, AADT prediction using support vector regression with data-dependent parameters. *Expert Systems with Applications.* Vol. 36(2), pp. 2979-2986 (2009). DOI: 10.1016/j.eswa.2008.01.073.

[7] Chen, P., Hu, S., Shen, Q., Lin, H., and Xie, C., Estimating Traffic Volume for Local Streets with Imbalanced Data. *Transportation Research Record*. Vol. 2673(3) pp 598–610 (2019). https://doi.org/10.1177/0361198119833347.

[8] Chen, Y., Yu, J., Khan, S., Spatial sensitivity analysis of multi-criteria weights in GIS-based land suitability evaluation. *Environmental Modelling & Software.* Vol. 25(12), pp. 1582-1591 (2010). https://doi.org/10.1016/j.envsoft.2010.06.001.

[9] Cognos Analytics (2021). https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=dashboards-statistical-tests.

[10] Cools, M., Moons, E., and Wets, G., Investigating Effect of Holidays on Daily Traffic Counts: Time Series Approach Transportation Research Record. *Journal of the Transportation Research Board*. 2019 (1), pp 22-31 (2007). https://doi.org/10.3141/2019-04.

[11] Cools, M., Moons, E., and Wets, G., Investigating the Variability in Daily Traffic Counts Through Use of ARIMAX and SARIMAX

Models Assessing the Effect of Holidays on Two Site Locations Transportation Research Record: *Journal of the Transportation Research Board*, No. 2136, Transportation Research Board of the National Academies, Washington, D.C., pp. 57–66 (2009). https://doi.org/10.3141/2136-07.

[12] Dul, J., Van der Laan, E., and Kuik, R., A Statistical Significance Test for Necessary Condition Analysis. *Organizational Research Methods.* Vol. 23(2) 385-395 (2020). https://doi.org/10.1177/1094428118795272.

[13] Dushoff, J., Kain, M. P., and Bolker, B. M., I can see clearly now: Reinterpreting statistical significance. Methods in Ecology and Evolution (2018). https://doi.org/10.1111/2041-210X.13159.

[14] Eom, J. K., Park, M. S., Heo, T-Y, and Huntsinger, L.F., Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method, Transportation Research Record: *Journal of the Transportation Research Board* (2006). https://doi.org/10.1177/0361198106196800103.

[15] Fricker Jr., R. D., Burke, K., Han, X., and Woodall, W.H., Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their *p*-Value Ban. The American Statistician, 73 : sup1, pp. 374-384 (2019), https://doi.org/10.1080/00031305.2018.153789.

[16] Fushiki, T., and Maeda, T., Nonresponse Bias Adjustment in Regression Analysis. *Journal of Statistical Theory and Practice*. 14 (20) (2020). https://doi.org/10.1007/s42519-020-0086-z.

[17] Han, C., and Song, S., A Review of Some Main Models for Traffic Flow Forecasting. Proc. 2003 IEEE International Conference on Intelligent Transportation Systems., Vol. 1., pp. 216–219 (2003). https://doi.org/10.1109/ITSC.2003.1251951.

[18] Jato-Espino, D., Spatiotemporal statistical analysis of the Urban Heat Island effect in a Mediterranean region, Sustainable Cities and Society, 46, 101427 (2019), https://doi.org/10.1016/j.scs.2019.101427.

[19] Jin, S-B., and Lee, J.-W., Study on Accident Prediction Models in Urban Railway Casualty Accidents Using Logistic Regression Analysis Model. *Journal of The Korean Society for Railway*. Vol.20 (4). pp. 482-490 (2017). http://dx.doi.org/10.7782/JKSR.2017.20.4.482.

[20] Kleijnen, J. P. C., Sensitivity analysis and optimization of system dynamics models: Regression analysis and statistical design of ex-

periments System Dynamics Review, Vol. 11(4), pp. 275-288 (1995). https://doi.org/10.1002/SDR.4260110403, Corpus ID: 62709384.

[21] Liu, H., Lee, M., and Khattak, A. J., Updating Annual Average Daily Traffic Estimates at Highway-Rail Grade Crossings with Geographically Weighted Poisson Regression, *Journal of the Transportation Research Board*, Vol. 2673(10) 105–117 (2019), https://doi.org/10.1177/0361198119844976.

[22] McClave, J., and Sincich, T., Statistics, 13th Edition. Pearson (2017).

[23] Minitab Express™ Support., 2019. https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/simple-regression/interpret-the-results/all-statistics-and-graphs/#p-value-regression.

[24] Minitab Express™ Support., (2019). https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/summary-statistics/descriptive-statistics/interpret-the-results/key-results/#step-2-describe-the-center-of-your-data.

[25] Minitab Express™ Support., (2019). https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/simple-regression/interpret-the-results/all-statistics-and-graphs/.

[26] Minitab Express™ Support., (2019). https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/inference/how-to/one-sample/runs-test/before-you-start/data-considerations/.

[27] Minitab Express™ Support., (2019). https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/inference/how-to/one-sample/runs-test/interpret-the-results/key-results/.

[28] Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C., and Keshri, A., Descriptive statistics and normality tests for statistical data. *Annals Cardiac Anesthesia,* Vol. 22(1), pp. 67-72 (2019). https://doi.org/10.4103/aca.ACA_157_18.

[29] Niu, M., Zhen, H., Wan, C., and Xi, Z., Statistical Regression Analysis Study on Anorectal Disease Predisposing Factors, 6th International Conference on Management, Education, Information and Control (MEICI 2016).

[30] PennDOT Traffic Count Policy, . Dirt, Gravel, and Low Volume Road Maintenance Program (DGLVRP) Traffic Count Policy SCC approved https://www.dirtandgravel.psu.edu/sites/default/files/PA%20Program%20Resources/Low%20Volume%20Roads/LVR_Traffic_Count_Policy.pdf.

[31] Rúa, M.O.B., Aragón, A.J.D., Baena, P.B., and Botero, J.D.O., Statistical analysis to establish an ignition scenario based on extrinsic and intrinsic variables of coal seams that affect spontaneous combustion, *International Journal of Mining Science and Technology*. 29, pp. 731–737 (2019).

[32] Raza, O., Mansournia, M. A., Foroushani, A. R., and Holakouie-Naieni, K., Geographically Weighted Regression Analysis: A Statistical Method to Account for Spatial Heterogeneity, *Archives of Iranian Medicine,* Vol. 22(3), pp: 155-160 (2019).

[33] Sharma, S. C., Gulati, B.M. and Rizak, S. N., Statewide traffic volume studies and precision of AADT estimates." *Journal of Transportation Engineering*, 122 (6), 430-439 (1996).

[34] Sharma, S.C., Lingras. P., Xu, F., and Kilburn, P., Application of Neural Networks to Estimate AADT On Low-Volume Roads. *Journal of Transportation Engineering*, Vol.127 (5), pp. 426-432 (2001).

[35] Staats, W. N., (2016). Estimation of Annual Average Daily Traffic On Local Roads In Kentucky. Theses And Dissertations--Civil Engineering, 36, HTTPS://UKNOWLEDGE.UKY.EDU/CE_ETDS/36, http://dx.doi.org/10.13023/ETD.2016.066.

[36] Stolwijk, A. M., Straatman, H., and Zielhuis, G. A., Studying seasonality by using sine and cosine functions in regression analysis, *Journal of Epidemiol Community Health*, 53, pp. 235–238 (1999).

[37] Taylor, C., (2020). "What Is the Interquartile Range Rule?" ThoughtCo, thoughtco.com/what-is-the-interquartile-range-rule-3126244.

[38] Tobler, W.R., Estimation of Attractivities from Interactions. Environment and Planning A: *Economy and Space*, 11(2), pp 121-127 (1979). https://doi.org/10.1068/a110121.

[39] Tsapakis, I., Schneider IV, W. H. Nichols, A. P., A Bayesian analysis of the effect of estimating annual average daily traffic for heavy-duty trucks using training and validation datasets. *Transportation Planning and Technology*, Vol. 36 (2), pp. 201-217 (2013), https://doi.org/10.1080/03081060.2013.770944.

[40] UCI Department of statistics, https://www.stat.uci.edu/what-is-statistics/.

[41] Van Arem, B., Kirby, H. R., Van Der Vlist, M. J. M., and Whittaker, J. C., Recent Advances and Applications in the Field of Short-Term Traffic Forecasting. *International Journal of Forecasting*. Vol. 13(1), pp. 1–12 (1997).

[42] Weare, B.C., A Statistical Study of the Relationship between Ocean Surface Temperatures and the Indian Monsoon. *Journal of the Atmospheric Sciences*. 36 (12), pp. 2279-2291 (1979), https://doi.org/10.1175/1520-0469(1979)036<2279:ASSOTR>2.0.CO;2.

[43] Wright, R. E., Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics. *American Psychological Association,* pp. 217–244 (1995).

[44] Ying, G-S.,Maguire, M. G., Glynn, R., and B.Rosner. Tutorial on Biostatistics: Linear Regression Analysis of Continuous Correlated Eye Data. *Ophthalmic Epidemiology,* Vol. 24(2) (2017), https://doi.org/10.1080/09286586.2016.1259636.

[45] Raza, O., Mansournia, M. A., Foroushani, A. R., and Holakouie-Naieni, K., Geographically Weighted Regression Analysis: A Statistical Method to Account for Spatial Heterogeneity, Arch Iran Med. 22(3), pp 155-160 (2019). PMID: 31029072, Scopus ID: 85065421618, http://www.aimjournal.ir/Article/aim-3740.

Received October 2021

Revised May 2022