International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-9, Issue-11; Nov, 2022
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.911.48

Statistical estimation of traffic volume for the Minneapolis-St Paul Metropolitan area

Hitalo J.B. Nascimento

Federal University of Ceará, Itapajé - Brazil

Received: 28 Oct 2022,

Receive in revised form: 15 Nov 2022,

Accepted: 21 Nov 2022,

Available online: 27 Nov 2022

***Keywords*—Multiple Linear Regression, Queuing, Urban Traffic Management.**

***Abstract*—** *In this work we propose a multiple linear regression model to describe the relationship between traffic volume in the Minneapolis-St Paul metropolitan area as a function of several variables, which include weather conditions and holiday occurrences. Additionally, we present an analysis for a queue model with a general arrival process, denoted by (G/M/1) : (∞/FIFO). Our results indicate that the uniform distribution is more suitable to model the process of vehicle arrivals.*

I. INTRODUCTION

The development of efficient public policies aimed at traffic management is extremely important for any large city. Problems with congestion and long queues have been a problem that has been a concern for public management in different parts of the world for decades. It is an inevitable phenomenon, but it can be mitigated through good public policies, which can increase people's quality of life, in addition to reducing its impact on the environment. In this sense, a considerable amount of studies have been proposed to deal with this problem. In [1] two approaches are proposed based on graph theory to solve the problem of time information in public transportation systems, whereas in [2], techniques for route planning on public transportation networks are proposed. In [3], a study was conducted to verify the efficiency of public transport in smart cities, specifically in relation to vehicle fluidity.

In this work we propose a multiple linear regression model to describe the relationship between traffic volume in the Minneapolis-St Paul metropolitan area as a function of several variables, which include weather conditions and holiday occurrences. Additionally, we present an analysis for a queue model with a general arrival process, denoted by (G/M/1) : (∞/FIFO). This model deals with a queueing system where a single service channel, interarrival times and service times are independent and identically distributed random variables, given respectively by G, which represents a general probability distribution, and M modeled by a exponential distribution with a mean and PDF (Probability density function) and CDF (Cumulative distribution function) given by equations 1 and 2. Moreover, there is no limit on the system capacity while the customers are served on a first in, first out, basis.

$$\lambda e^{-\lambda t} I_{(0,\infty)}(t) \quad (1)$$

$$1 - e^{-\lambda t} I_{(0,\infty)}(t) \quad (2)$$

In this sense, our research contributions can be summarized as follows:

- We propose a multiple linear regression model to describe the relationship between traffic volume in the MinneapolisSt Paul metropolitan area as a function of several variables, which include weather conditions and holiday occurrences;

- We present an analysis for a model queue with a general arrival process, denoted by (G/M/1):(∞ / FIFO). We specifically investigate whether or not the arrivals follow a poisson distribution and which probability distribution is most suitable for the observed data;
- In addition, the analysis of variance (ANOVA) test is performed in order to verify the contributions of each variable in the regression model.

The rest of the paper is organized as follows: In section 2, we present the related work. The problem statement is described in section 3. Experimental results are described in section 4 and conclusions are presented in section 5.

II. RELATED WORK

There are several related works that address the topic in question, In [4] is proposed a Bayesian inference model for traffic prediction, capable of incorporating spatial and temporal components. Furthermore, according to the authors, the proposed solution works well with missing data points, taking advantage of previous information. In [5], a traffic forecast is proposed as a birth and death process to describe the behavior of vehicles on the road. [6] proposes to model the relationship between traffic congestion and weather. The authors used a multiple linear regression model to predict daily changes in congestion, based on eight weather forecast factors and six dummy variables to express the days of the week. In [7] is give an overview of some approaches for reducing and managing congestion so as to reduce this phenomenon, particularly, the effects of congestion on public transport. In [8] presents a novel integration of machine learning models into simulation to improve the realism of simulating a public transport system. The authors conclude that, with an efficient congestion prediction tool, it is possible to effectively predict the time delays in traffic.

III. PROBLEM STATEMENT

Estimating traffic behavior in large cities is important for the development of public policies that minimize congestion, which affect people's quality of life and impacts on the environment. In this sense, this study uses a multiple linear regression model, defined in the equation 3, to verify the relationship between the dependent variable traffic volume and a set of independent variables, such as the occurrence of holidays and weather conditions. In addition, it is statistically evaluated which probability distribution (*G*) is more suitable for the *G/M/*1 model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \qquad (3)$$

where Y represents the dependent variable; x1,x2 ⋯xp are the explanatory variables; βi, for i = 1,2,...,p, are slope coefficents for explanatory variables and ϵ is a random error with $\overset{i.i.d}{\sim} N\left(0, \sigma^2\right)$.

IV. RESULTS

In our experimental study, we considered the metro Interstate Traffic Volume Data Set, provided by the UCI data repository [9]. This dataset contains information about hourly from 20122018, Interstate 94 Westbound traffic volume for MN DoT ATR station 301, roughly midway between Minneapolis and St Paul, MN. Hourly weather features and holidays included for impacts on traffic volume. In this sense, we investigate the influence of variables like holidays and climate in relation the dependent variable variable traffic volume. The attributes informations are:

- Holiday Categorical US National holidays plus regional holiday, Minnesota State Fair;
- Temperature in kelvin;
- Amount in mm of rain that occurred in the hour;
- Amount in mm of snow that occurred in the hour;
- Numeric Percentage of cloud cover;
- Short textual description of the current weather;
- Longer textual description of the current weather;
- DateTime Hour of the data collected in local CST time;
- Numeric Hourly I-94 ATR 301 reported westbound traffic volume.

Figure 1 the box plot of vehicles arrivals between 2012 and 2018, indicating that there is no significant difference between the observations, whereas figure 2 shows the empirical cumulative distribution function (CDF) of the data.

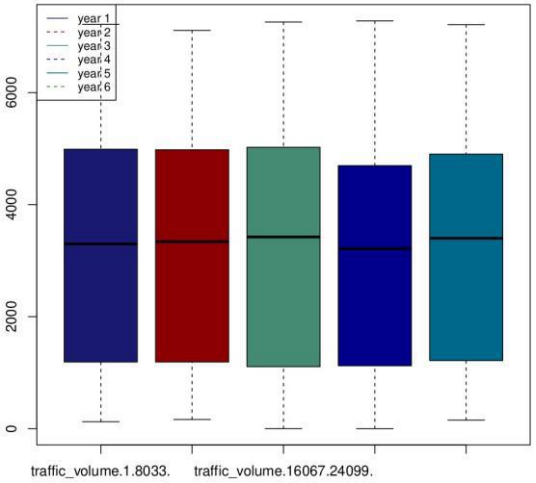

Fig. 1: boxplots of traffic volume per year.

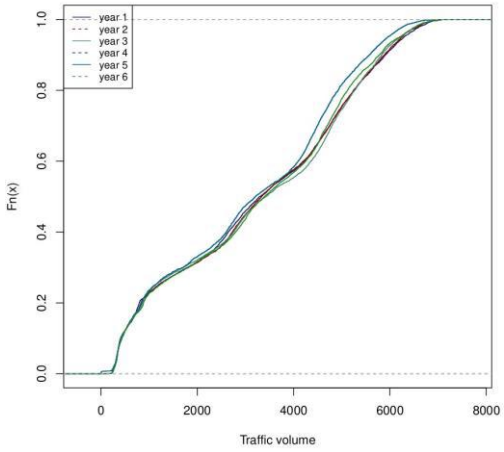

Fig. 2: CDF of vehicles arrivals between 2012 and 2018.

Regression results:

According to section 3, we formulate multiple linear regression analyses to explore the relationships between traffic volume for metropolitan Minneapolis and the independent variables defined at the beginning of this section. The source code in listing 1, implemented in R language, shows the results for the model in question. Note that the R-squared, which measures the strength of the relationship between the dependent and independent variables, was 0.933. That is, 93.3% of the variance in the data can be explained by the independent variables.

```
Call:
lm(formula = traffic_volume ~ holiday + rain + temperature +
    clouds + clear + mist + rain + snow + drizzle + haze + thunderstorm)

Residuals:
    Min      1Q  Median      3Q     Max
-1048.6  -452.6   -81.6   362.5  1350.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   915.286      7.047 129.883  < 2e-16 ***
holiday       -33.384     65.909  -0.507   0.6125
rain         4790.665      6.138 780.502  < 2e-16 ***
temperature  1471.889     17.053  86.311  < 2e-16 ***
clouds        -88.179      7.618 -11.575  < 2e-16 ***
clear         -35.825      7.762  -4.615 3.93e-06 ***
mist          -49.379      9.208  -5.362 8.25e-08 ***
snow         -114.147     11.526  -9.903  < 2e-16 ***
drizzle      -105.647     13.629  -7.752 9.25e-15 ***
haze          -31.747     15.332  -2.071   0.0384 *
thunderstorm   95.927     17.213   5.573 2.52e-08 ***

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 514.4 on 48193 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.933, Adjusted R-squared:  0.933
F-statistic: 6.709e+04 on 10 and 48193 DF,  p-value: < 2.2e-16
```

Listing 1: Results of multiple regression analysis.

In order to provide evidence for our approach, we now use ANOVA test to verify the null and alternative hypotheses, defined in equation 6. The source code in listing 2, shows the results for this test. ANOVA test is based on the sum of squares decomposition. In other words, the deviation of an observation from the mean can be decomposed as the deviation of the observation from the regression-fitted value plus the deviation of the fitted value from the mean, that is, we can write (Yi- $\bar{Y}$)

$$(Y_i - \bar{Y}) = (Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \quad (4)$$

[10].

Squaring both sides of equation 4, we obtain:

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k} \left(\hat{Y}_i - \bar{Y}\right)^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}_{\text{SSE}} \quad (5)$$

Where SST, is sum of squares; SSR is sum of squares due to regression and SSE is sum of of squares error/residuals.

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \cdots = \beta_n; \\ H_1 & : \beta_i \neq 0, i = 0, 1, 2, \cdots, n. \end{cases} \quad (6)$$

```
Response: traffic_volume
                Df  Sum Sq Mean Sq F value Pr(>F)
holiday          1 9.1830e+03 9.1830e+03 3.4700e-02 0.852219
rain             1 1.7540e+11 1.7540e+11 6.6289e+05 < 2.2e-16 ***
temperature      1 2.0402e+09 2.0402e+09 7.7103e+03 < 2.2e-16 ***
clouds           1 2.6270e+07 2.6270e+07 9.9280e+01 < 2.2e-16 ***
clear            1 7.3259e+04 7.3259e+04 2.7690e-01 0.598766
mist             1 9.6686e+05 9.6686e+05 3.6540e+00 0.055940 .
snow             1 2.3243e+07 2.3243e+07 8.7839e+01 < 2.2e-16 ***
drizzle          1 1.8947e+07 1.8947e+07 7.1606e+01 < 2.2e-16 ***
haze             1 2.3092e+06 2.3092e+06 8.7268e+00 0.003137 **
thunderstorm     1 8.2179e+06 8.2179e+06 3.1057e+01 2.519e-08 ***
Residuals    48193 1.2752e+10 2.6460e+05
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Listing 2: Variance analysis (ANOVA) results.

Since p-values associated with the F-statistic, for variables rain, temperature, clouds, mist, snow drizzle, haze and thunderstorm, respectively, are less than 0.05, we have enough evidence to reject the null hypothesis and conclude there is a significant amount of variation in the response that is explained by the proposed model. For variables holiday and clear, it is observed a low sum-of-squares value and a high p-value, which means there is not much variation that can be explained by the those variables.

Queue results:

Here our objective is to identify which probability distribution is best suited to model the arrivals process for the G/M/1 queue type. Typically, the arrival process is modeled as Poisson distribution. In this sense, we performed a statistical test, called a chi-square, to verify if the poisson distribution would be more suitable for this scenario. To develop this test, the following hypotheses were considered, with a significant level of $\alpha = 0.05$:

$$\begin{cases} H_0 & : \text{The data} \sim \text{Poison;} \\ H_1 & : \text{The data don't} \sim \text{Poison.} \end{cases} \tag{7}$$

To measure the degree of disagreement between observed and expected frequencies, we use the summation defined in equation 8.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \tag{8}$$

Where, foi represents the observed frequencies; fei the expected frequencies and k, the number of classes or intervals. The source code in listing 3, shows the result for the test in question.

```
# Chi-squared test in R.

# Output:

Chi-squared test for given probabilities


data:  traffic_volume

X-squared = 58373468, df = 48204, p-value < 2.2e-16
```

Listing 3: Results of Chi-Square test for traffic volume and poisson distribution.

Since P-value is less than 0.05, statistically we have enough evidence to reject the null-hipothesis, with a significance level of 0.05 and thus accept the alternative hypothesis, that the arrivals process does not follow a poisson distribution. This fact can also be observed in figure 3, in which a comparison between the theoretical distribution (poisson) and the empirical data is performed.

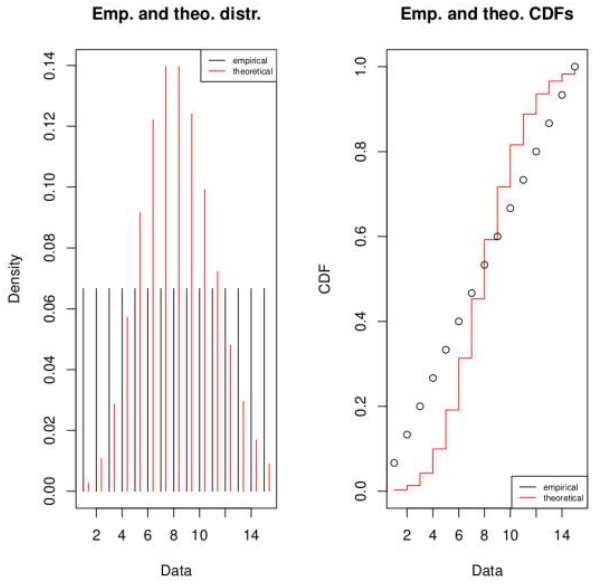

Fig. 3: Comparison between the poisson distribution and the empirical data.

Figure 4 presents Cullen and Frey graph [11], used to recognize a distribution among a set of parametric distributions on the basis of relations of skewness-kurtosis parameters. From a sample (Xi)i ~ (i.i.d) with observations (Xi)i , the skewness and kurtosis and their corresponding unbiased estimator are given by [12] [13].

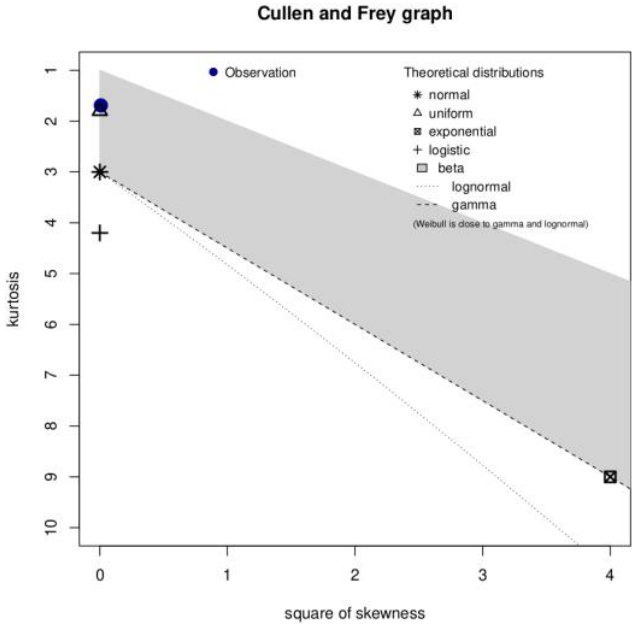

Fig. 4: Cullen and Frey graph of arrival.

$$sk(x) = \frac{\left[E\left(X - E\left(X\right)\right)^3\right]}{Var\left(X\right)^{\frac{3}{2}}} \tag{9}$$

$$\widehat{sk} = \frac{\sqrt{n\left(n-1\right)}}{n-2} \times \frac{m_3}{m_2^{\frac{3}{2}}} \tag{10}$$

$$kr(x) = \frac{\left[E\left(X - E\left(X\right)\right)^4\right]}{Var\left(X\right)^2} \tag{11}$$

$$\widehat{kr} = \frac{n-1}{(n-2)\left(n-3\right)}\left(\left(n+1\right) \times \frac{m_4}{m_2^2} - 3\left(n-1\right)\right) + 3 \tag{12}$$

where m2, m3, m4 denote empirical moments defined by mk = $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$, for k = 1, 2, 3, . . . and xi representing the n observations of variable x and x̄ their mean value. Skewness is the degree of asymmetry of a distribution. A normal distribution has a skewness value of zero [14]. Kurtosis is the degree of peakedness of a distribution. Usually taken relative to a normal distribution [15]. A distribution having a relatively high peak, is called leptokurtic and has a value less than 3, while a distribution of flat-topped is called platykurtic, with a value greater than 3. The normal distribution, which is not very peaked, is called mesokurtic and has a value equal to 3. As a result, note that the uniform distribution is closer to the real data (observation). Additionally, a comparison between empirical and theoretical data is presented in figure 5.

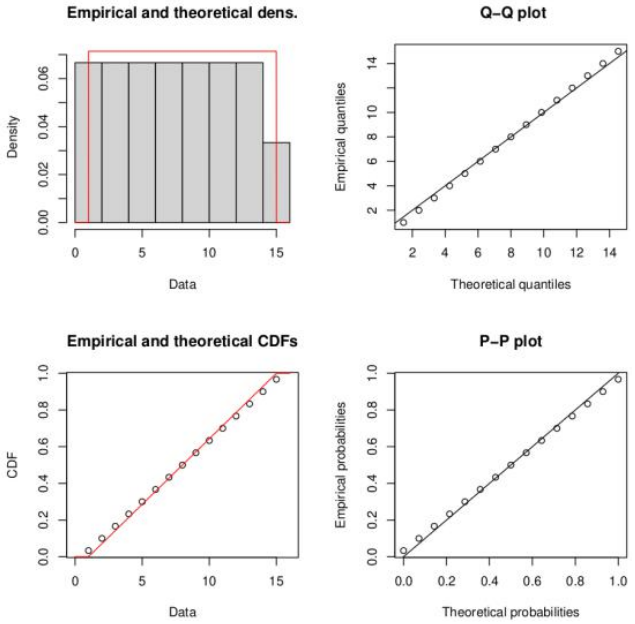

Fig. 5: Comparison between the uniform distribution and the empirical data.

The suitability of the uniform distribution is evident, both by the Q–Q plot (quantile-quantile plot) and by the P–P plot (probability–probability plot or percent–percent plot or P value plot) [16], in which the empirical data form approximately a straight line when compared to the theoretical data.

V. CONCLUSION

We propose a multiple linear regression model to describe the relationship between traffic volume in the Minneapolis-St Paul metropolitan area. The ANOVA test shows that we have enough evidence to conclude there is a significant amount of variation in

the response that is explained by the proposed model, by the variables rain, temperature, clouds, mist, snow drizzle, haze and thunderstorm. Additionally, we present an analysis for a model queue with a general arrival process, denoted by (G/M/1) : (∞/F IF O). Our results indicate that the uniform distribution is more suitable to model the process of vehicles arrivals.

The future research includes:

• Apply other regression models, in order to verify the adequacy of these models to the problem in question;

• Development a Bayesian approach to evaluate arrivals and queue service times;

REFERENCES

[1] Pyrga, E., Schulz, F., Wagner, D., and Zaroliagis, C. 2007. Efficient models for timetable information in public transportation systems. ACM J. Exp. Algor. 12, Article 2.4 (2007), 39 pages DOI 10.1145/1227161.1227166. http://doi.acm.org 10.1145/ 1227161.1227166.

[2] Wang, Sibo, et al. "Efficient route planning on public transportation networks: A labelling approach." Proceedings of the 2015 ACM SIGMOD. International Conference on Management of Data. 2015.

[3] Karoń, G., Żochowska, R. (2020). Problems of Quality of Public Transportation Systems in Smart Cities—Smoothness and Disruptions in Urban Traffic. In: Sładkowski, A. (eds) Modelling of the Interaction of the Different Vehicles and Various Transport Modes. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham. https://doi.org/10.1007/978-3-030-11512-8_9

[4] S. Mostafi, T. Alghamdi and K. Elgazzar, "A Bayesian Linear Regression Approach to Predict Traffic Congestion," 2021 IEEE 7th World Forum on Internet of Things (WF-IoT), 2021, pp. 716-722, doi: 10.1109/WF-IoT51360.2021.9595298.

[5] Y. Li, H. Chen and M. Feng, "A Novel Model for the Traffic of Urban Roads Based on Queuing Theory," 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS), 2020, pp. 190-194, doi: 10.1109/ICICAS51530.2020.00046.

[6] J. Lee, B. Hong, K. Lee and Y. -J. Jang, "A Prediction Model of Traffic Congestion Using Weather Data," 2015 IEEE International Conference on Data Science and Data Intensive Systems, Sydney, NSW, Australia, 2015, pp. 81-88. doi: 10.1109/DSDIS.2015.96.

[7] M. O. Ghali and M. J. Smith, "Managing traffic congestion by using traffic control," IEE Colloquium on Urban Congestion Management, London, UK, 1995, pp. 8/1-8/6. doi: 10.1049/ic:19951300.

[8] M. S. Bin Othman and G. Tan, "Machine Learning Aided Simulation of Public Transport Utilization," 2018 IEEE/ACM 22nd International Symposium on Distributed Simulation and Real Time Applications (DS-RT), Madrid, Spain, 2018, pp. 1-2. doi: 10.1109/DISTRA.2018.8601011.

[9] Hogue, J. MN Department of Transportation, Weather data from OpenWeatherMap. UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Metro%20Interstate%20Traffic%20Volume].

[10] Šmilauer, Petr., Lepš, Jan. Biostatistics with R: An Introductory Guide for Field Biologists. Cambridge University Press, 2020.

[11] Cullen, A. and Frey, H. (1999). Probabilistic Techniques in Exposure Assessment. Plenum Publishing Co., 1st edition.

[12] M. L. Delignette – Muller, C. Dutang, "fitdistrplus: An R Package for Fitting Distributions", in Journal of Statistical Software, vol. 64(4), 2015, pp. 1-34

[13] Casella, G. and Berger, R. (2002). Statistical Inference. Duxbury Thomson Learning, 2nd edition.

[14] Abbott, Dean. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Alemanha: Wiley, 2014.

[15] Murray R. Spiege. Theory and problems of statistics. Schaum's outline series. McGraw-Hill Publishing Company, 1972.

[16] Wilk, M.B.; Gnanadesikan, R. (1968), "Probability plotting methods for the analysis of data", Biometrika, Biometrika Trust, 55 (1): 1–17, doi:10.1093/biomet/55.1.1, JSTOR 2334448, PMID 5661047.