Monte Carlo Simulation in Sampling Technique of Traffic Data Collection: An Extension

By Stacy Lee

Background

A great deal of transportation research has been dedicated to modeling vehicular traffic behaviors and volumes. A common application in traffic studies is using Monte Carlo simulation of statistical distributions for modeling traffic flow to achieve a more efficient and accurate estimate of traffic flow than the estimates produced by regional transportation models. The approach Williamson used implements a Monte Carlo simulation of a two-parameter Weibull distribution using parameters estimated from traffic volume count data.[1] The two-parameter Weibull distribution was selected based on past research. Previously, Stathopoulos and Karlaftis proved that traffic flows are best modeled by the Weibull distribution, but the specific values of the Weibull parameters for scale and shape are different and must be determined separately for different traffic samples.[2]

The traffic data utilized in Williamson's paper contains the traffic volume counts based on video traffic data collected from downtown Birmingham, Alabama.[1] The volume counts were collected and manually counted from a video tape of traffic for each one-minute interval in one hour. Of all the traffic data that was collected, the traffic volume counts selected for presentation in the study were collected between 2:00pm and 3:00pm on June 20, 2000. The traffic for all vehicle classes was tested for the possibility of identifying a shorter time interval that would be representative of the one-hour traffic data. Subintervals of one hour were specified as 5-minute, 10-minute, 20-minute, and 30-minute intervals.

A computer program was developed to process the video data, compute the Weibull parameters for each specified subinterval using the L-moment method, and generate samples based on the specified Weibull probability function with the estimated parameters. Then, 90% quantile intervals, which are referred to as confidence intervals in the paper, were calculated based on the samples in each specified subinterval within the hour. If the mean of the one-hour traffic data fell within the 90% quantile interval, the conclusion that the subinterval of traffic data is representative of one-hour traffic data. Williamson et al. concluded that 20-minute and 30-minute intervals are representative of the one-hour traffic data in their case study.

Objective

The data utilized in this study are the traffic volume counts in New York City (NYC) from 2012 to 2013 provided by the NY Department of Transportation and downloaded from the NYC Open Data portal.[3] Traffic vehicle counts for each hour on different dates and road segments in both directions. Compared to the data in Williamson's paper, the NYC traffic data set provides traffic volume counts for one-hour intervals rather than one-minute intervals. Therefore, Williamson's approach cannot be reproduced using the NYC dataset. However, a similar approach to Williamson's method will be applied to this dataset.

Instead of generating samples of traffic volume counts for subintervals of an hour, samples of traffic volume counts will be generated for each hour in the morning, between 6am and 11am, and evening, between 5pm and 10pm, where each hour would be considered a subinterval of the five-hour time period. This can be accomplished since the dataset has the traffic volume counts for each hour of the day with respect to a road segment, driving direction, and day. The same quantile interval method applied in Williamson's paper is used to calculate the quantile intervals for finding the representative hours of the morning and evening traffic periods. Unlike Williamson's method, this study will utilize conventional moments, which is reasonable since L-moments are analogous to conventional moments.[4]

Assumptions

Since the NYC traffic volume data is limited in the number of samples for each road segment in a single direction, where the maximum sample size is 18, the fitted Weibull distribution in this study will not represent a single road as in Williamson's paper. In order to achieve a clear distribution shape to estimate the parameters of the Weibull distribution, the fitted Weibull distribution in this study will represent all the traffic in NYC at each hour where the sample size of this data for each hour is 5945. To reduce the effect of outliers, the data for each hour is truncated by removing the observations above the 95% quantile. The resulting traffic volume data for each hour is assumed to follow a Weibull distribution.

Method/Procedure

1. Calculate the mean traffic volume for the morning and evening period, which will represent the population means.
2. For each hour:
    a. Remove the observations above the 95% quantile.
    b. Calculate the sample mean and variance of the observations.
    c. Estimate the shape and scale parameters of the two-parameter Weibull distribution using method of moments estimation.
    d. Use the Naive Monte Carlo method to generate 1000 independent and identically distributed samples from the Weibull distribution using the estimated shape and scale parameters.
    e. Calculate the 90% quantile interval of the generated samples.
    f. If the population mean falls into the 90% quantile interval of the generated samples, then the hour is concluded to be representative of the morning, or evening, period.
    g. Visually compare the probability density function of the fitted Weibull curve to the five-hour morning period.

Results

In Figure 1, the blue curved line represents the fitted Weibull probability density function of the samples, the green curve represents the probability density function for the five-hour morning period, the red vertical lines represent the upper or lower bounds of the 90% quantile interval, and the green vertical line represents the population mean, or mean traffic volume of the five-hour morning period. For each hour, the shape of the fitted Weibull probability density curve (in blue) is noticeably similar to the shape of the probability density curve for the morning (in green). A closer look at the plots reveal that the two probability density curves are most similar in the hour between 7 and 8 am as shown in Figure 2.

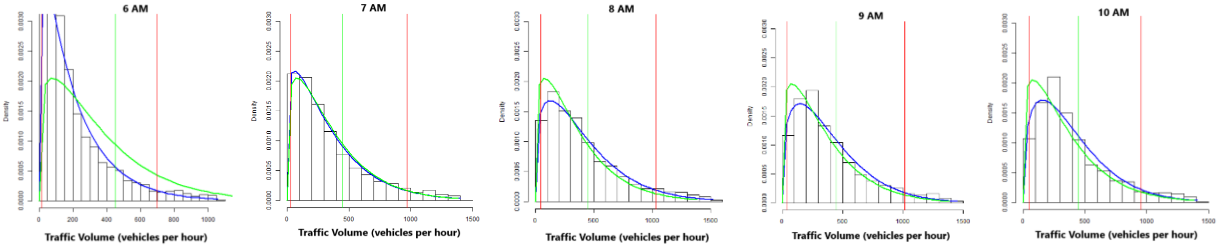

Figure 1. Probability Density Plots of Traffic Volume Per Hour in Order from 6 to 10 AM

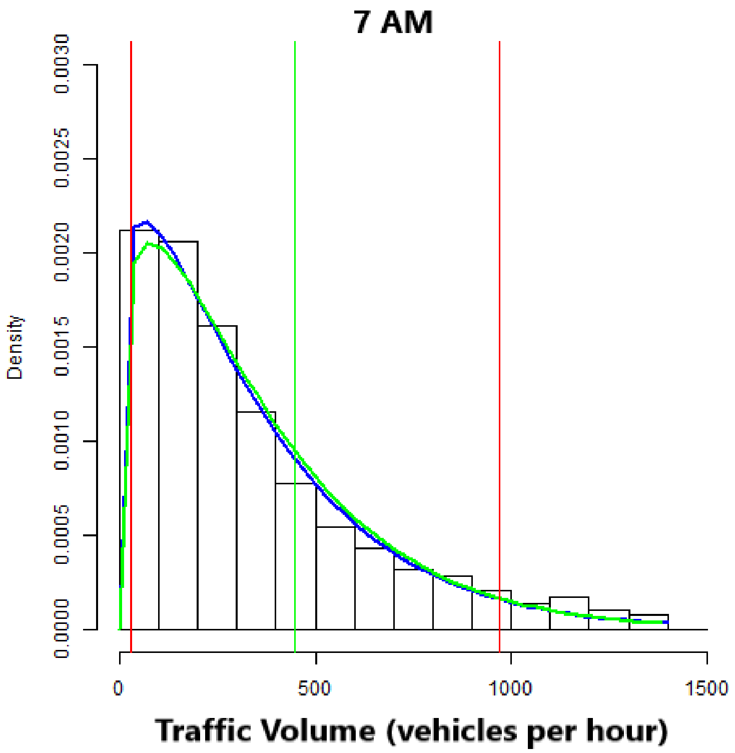

Figure 2. Probability Density Plot of Traffic Volume for the Hour Starting at 7 AM

Table 1. Statistical Analysis for the Morning Traffic Data of Total Vehicles

Time	Sample Mean	Sample Variance	Fitted Weibull Distribution 90% Range	Population Mean	Shape	Scale	Rep
6 – 7 am	210.44	43007.96	(11.18, 617.94)	449.64	1.01	211.73	1
7 – 8 am	341.12	89546.97	(24.46, 891.92)	449.64	1.14	357.76	1
8 – 9 am	408.90	106974.10	(40.40, 1049.61)	449.64	1.26	439.71	1
9 – 10 am	395.69	92261.92	(44.23, 1011.73)	449.64	1.31	429.39	1
10 – 11 am	392.44	86282.08	(42.98, 954.65)	449.64	1.35	428.00	1

In Figure 3, the blue curved line represents the fitted Weibull probability density function of the samples, the purple curve represents the probability density function for the five-hour evening period, the red vertical lines represent the upper or lower bounds of the 90% quantile interval, and the purple vertical line represents the population mean, or mean traffic volume of the five-hour evening period. For each hour, the shape of the fitted Weibull probability density curve (in blue) is noticeably similar to the shape of the probability density curve for the evening (in green). A closer look at the plots reveal that the two probability density curves are most similar in the hour between 7 and 8 pm as shown in Figure 4.

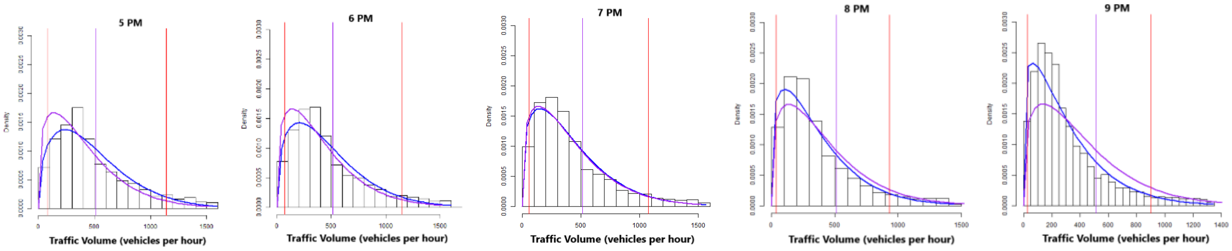

Figure 3. Probability Density Plots of Traffic Volume Per Hour in Order from 5 to 9 PM

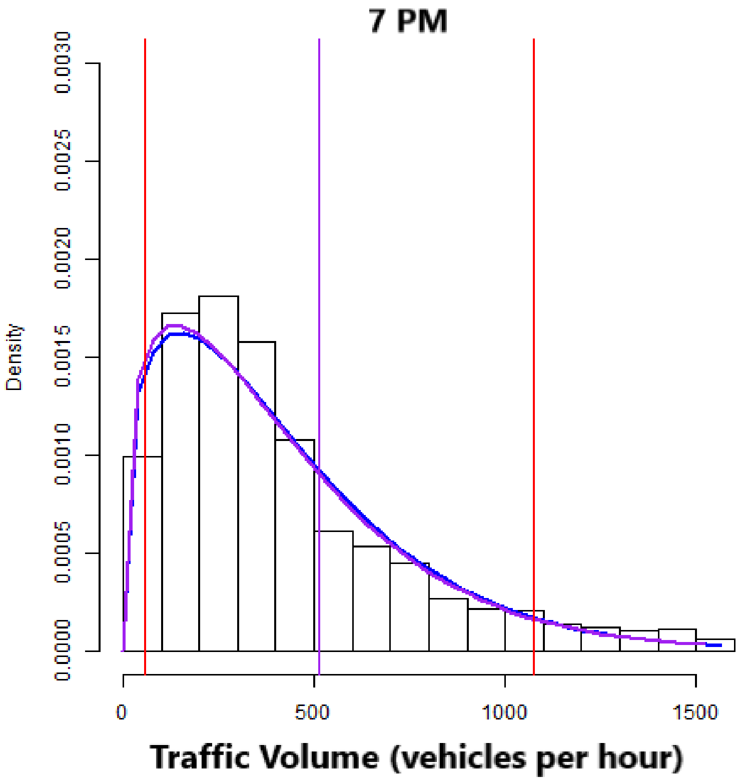

Figure 4. Probability Density Plot of Traffic Volume for the Hour Starting at 7 PM

Table 2. Statistical Analysis for the Evening Traffic Data of Total Vehicles

Time	Sample Mean	Sample Variance	Fitted Weibull Distribution 90% Range	Population Mean	Shape	Scale	Rep
5 – 6 pm	490.44	120102.94	(69.79, 1121.22)	512.77	1.44	540.20	1
6 – 7 pm	469.19	117631.03	(69.43, 1151.44)	512.77	1.39	513.92	1
7 – 8 pm	418.11	104252.18	(45.21, 1035.26)	512.77	1.31	453.15	1
8 – 9 pm	363.29	86689.42	(49.91, 996.11)	512.77	1.24	389.43	1
9 – 10 pm	312.79	72886.40	(25.55, 852.27)	512.77	1.16	329.66	1

In Table 1 and Table 2, the population means for morning and evening periods were within the 90% quantile interval based on the fitted Weibull distribution. Following Williamson's verification method, this implies that each of the one-hour samples are representative of their corresponding morning, or evening, period.

Summary/Conclusion

Based on the results, it can be concluded that any hour of the morning or evening traffic is representative of their respective time period for NYC traffic in 2012 to 2013. Furthermore, the results of this study suggest that the Weibull distribution can potentially represent the traffic volumes throughout NYC at each hour of the morning and evening periods, regardless of the location of the given roads. By visual inspection, the hour that appears most representative is 7 to 8 am for the morning period and 7 to 8 pm for the evening period.

A concerning factor of this study is that the assumptions made may oversimplify the conclusion. For example, the sample variance for each hour is very high, which results in a very wide 90% interval and is most likely due to treating the different roads of multiple directions in NYC as independent and identically distributed. A suggestion for future steps to verify the conclusion of this study would be to obtain traffic volume counts per minute rather than traffic volume counts per hour.

References

1. Williamson, Derek G., Maosheng Yao, and John McFadden. "Monte Carlo Simulation in Sampling Techniques of Traffic Data Collection." Transportation Research Record 1804, no. 1 (January 2002): 91–97. doi:10.3141/1804-13.
2. Stathopoulos, A., and M. Karlaftis. "Temporal and Spatial Variations of Real-Time Traffic Data in Urban Areas." Transportation Research Record 1768, no. 1 (January 2001): 135–40. doi:10.3141/1768-16.
3. The City of New York. "Traffic Volume Counts (2012-2013) | NYC Open Data." Data.cityofnewyork.us. https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2012-2013-/p424-amsu (Accessed November 13, 2018)
4. Hosking, J. R. M. "Moments or L Moments? An Example Comparing Two Measures of Distributional Shape." *The American Statistician* 46, no. 3 (1992): 186-89. doi:10.2307/2685210.

Appendix

Data: Please refer to the “Traffic_Volume_Counts__2012-2013.csv” file

Code:

```
install.packages("EnvStats")
install.packages("glue")
library(EnvStats)
library(glue)
set.seed(525525)

traffic = read.csv("Traffic_Volume_Counts__2012-2013.csv")

M1 = mean(as.numeric(as.matrix(traffic[,14:18]))) #  06:00-11:00
M2 = mean(as.numeric(as.matrix(traffic[,25:29]))) #  17:00-22:00

X1= as.numeric(as.matrix(traffic[,14:18])) #  06:00-11:00
X1 = X1[X1<=quantile(X1, 0.95)]
Est1 = eweibull(X1, method = "mme")$parameters
h = hist(X1, breaks = 20, freq = FALSE ,main = "06:00-11:00")
Xfit1 <- seq(min(X1),max(X1),length=40)
Yfit1 <- dweibull(Xfit1,Est1[1],Est1[2])
lines(Xfit1, Yfit1, col="green", lwd=2)
abline(v = M1, col = "green")
dev.copy(png, "06.00_11.00.png")
dev.off()

X2= as.numeric(as.matrix(traffic[,25:29])) #  17:00-22:00
X2 = X2[X2<=quantile(X2, 0.95)]
Est2 = eweibull(X2, method = "mme")$parameters
h = hist(X2, breaks = 20, freq = FALSE, main = "17:00-22:00")
Xfit2 <- seq(min(X2),max(X2),length=40)
Yfit2 <- dweibull(Xfit2,Est2[1],Est2[2])
lines(Xfit2, Yfit2, col="purple", lwd=2)
abline(v = M2, col = "purple")
dev.copy(png, "17.00_22.00.png")
dev.off()

road = 1:nrow(traffic)

Hour = colnames(traffic)[8:ncol(traffic)]
SampleMean = c()
SampleVar = c()
Lower = c()
Upper = c()
PopulationMean = c()
```

```
Shape = c()
Scale = c()
Rep = c()

for (i in 8:ncol(traffic)){
  # extract the column to be analied
  x = traffic[road,i]

  # remove observations above 95 quantile
  cap = quantile(x, 0.95)
  x = x[x<=cap]

  # calculate sample mean and sample variance
  m = mean(x)
  SampleMean = c(SampleMean, m)
  v = var(x)
  SampleVar = c(SampleVar, v)

  # estimate parameters for weibull distribution
  est = eweibull(x, method = "mme")$parameters
  Shape = c(Shape, est[1])
  Scale = c(Scale, est[2])

  # generate 1000 random sample for the weibull distribution with estimated parameters
  y = rweibull(1000, est[1], est[2])
  # calculate empirical confidence interval
  y1 = quantile(y, 0.05)
  y2 = quantile(y, 0.95)
  Lower = c(Lower, y1)
  Upper = c(Upper, y2)

  # create plot
  h = hist(x, breaks = 20, freq = FALSE, main = glue("{colnames(traffic)[i]}"))
  xfit<-seq(min(x),max(x),length=40)
  yfit<-dweibull(xfit,est[1],est[2])
  lines(xfit, yfit, col="blue", lwd=2)

  # test if the population mean falls in the interval
  r = NaN
  M = NaN

  if (i>=14 && i<=18){
    r = (y1<=M1 & M1<=y2)
    M = M1
  abline(v = M1, col = "green")
  lines(Xfit1, Yfit1, col="green", lwd=2)
 }
 if (i>=25 && i<=29){
  r = (y1<=M2 && M2<=y2)
  M = M2
  abline(v = M2, col = "purple")
  lines(Xfit2, Yfit2, col="purple", lwd=2)
 }
 Rep = c(Rep, r)
 PopulationMean = c(PopulationMean, M)

 abline(v = y1, col = "red")
 abline(v = y2, col = "red")

 dev.copy(png, glue("{colnames(traffic)[i]}.png"))
 dev.off()
}

result.one = data.frame(SampleMean, SampleVar, Lower, Upper, PopulationMean, Shape, Scale,
Rep, row.names = Hour)
```