Traffic Flow Study

VEE - Time Series Student Project
Winter 2012
Lina Wang

Introduction

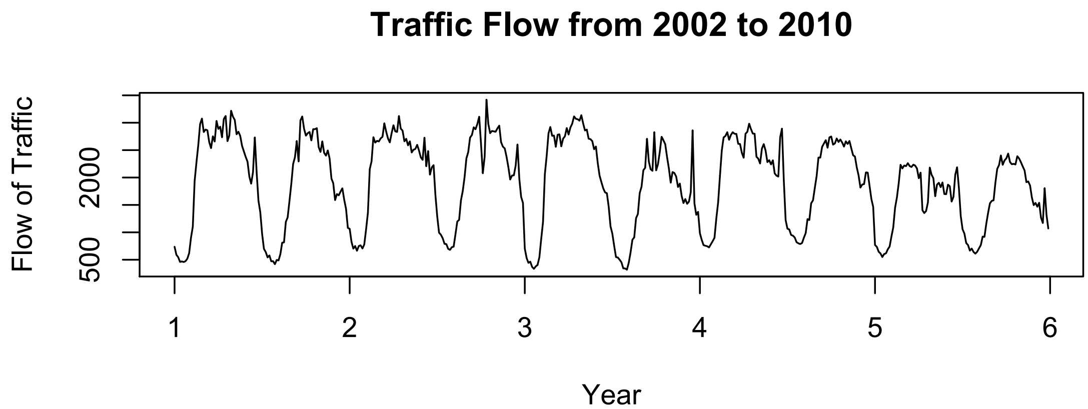
The purpose of this study is to test the time series of the traffic flow during the university game season on highway. This paper will focus on an important event to Bay-Area college students – the Big Game. The Big Game is a football game held between and University of California – Berkeley and Stanford University each year, the home filed alternates each year. I will use time series model to fit the flow of traffic on I880 North, the direction from Palo Alto (Stanford) to Berkeley on the Game day. We will use R for all the data analysis.

Analysis

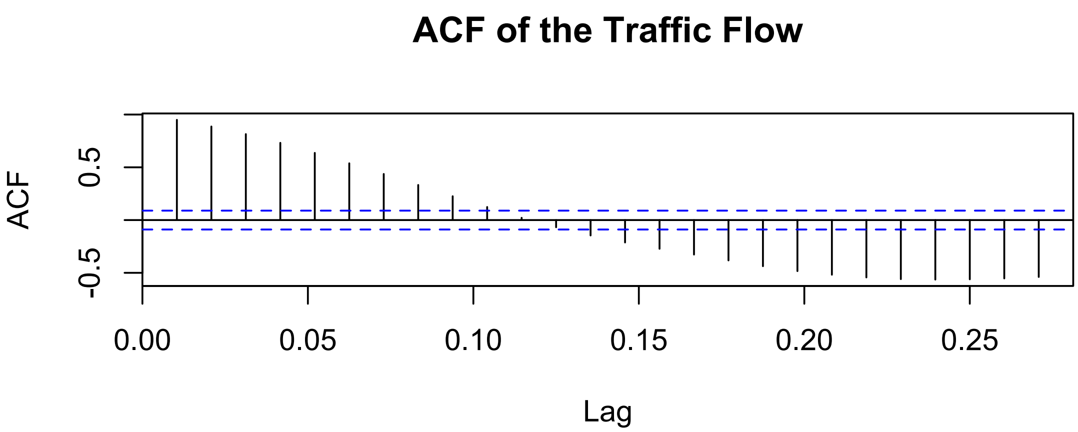
The traffic data were found on the PeMS website which summarized historical traffic data of California. The system collects, filters, processes, aggregates and examines continuous traffic data that were recorded by detectors and tag readers. As we will research the traffic flow on Big Game day, we took the time period from the day before the game day 12:00 am to 12:00 pm on the game day on 2002, 2004, 2006, 2008 and 2010. The flow of traffic is showed as number of cars passing by the detectors every five minutes period.

To shorten the number of points for analysis, we converted the five minutes period into 30 minutes period. The final data file adjunct all five year data with chronologic order (from 2002 to 2010). Then we transformed it into a time series with frequency of 96, which indicates the data of two days in one year.

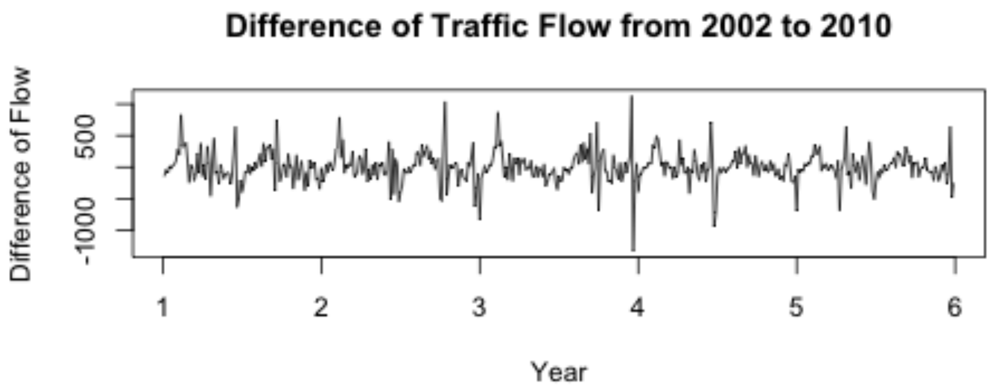
Below is the plot of the traffic flow time series.



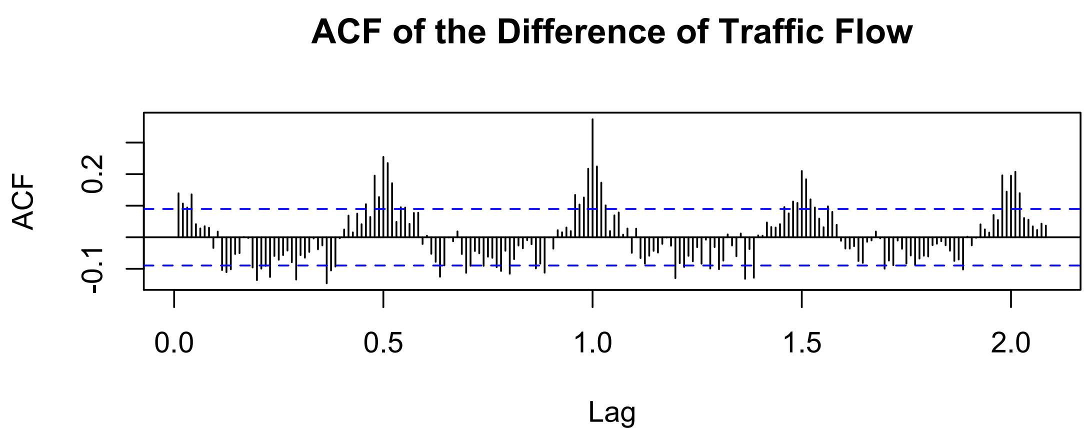
As we can see from the raw data, it suggested a strong seasonality, which can be interpreted as that traffic is normally slower on prime daytime and faster at night. The ACF suggested the same.



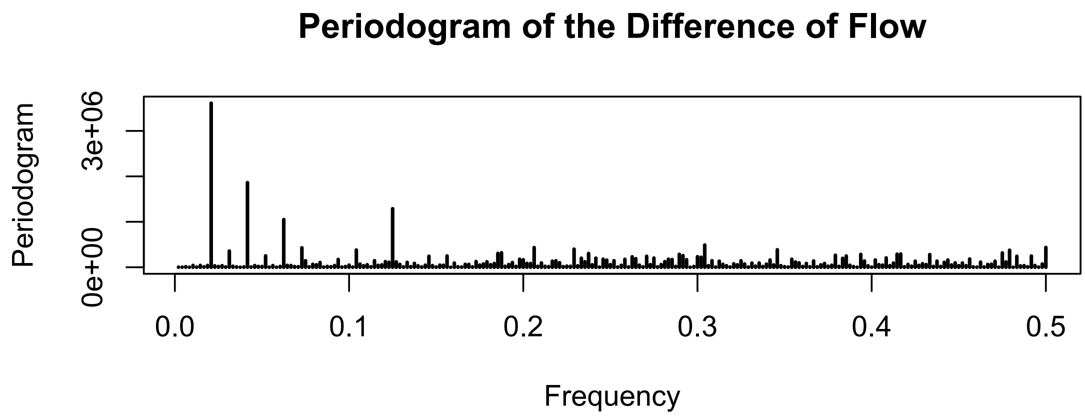
To reduce the seasonality, we will take the difference. Transformed data are plotted as follows.



Let’s look at the ACF and PACF of the difference, which improves a lot comparing to the original data.



By observing the ACF, we can assume the pattern follows a sine and cosine wave. A periodogram below suggested 4 spikes.



The textbook introduced spectral analysis. It suggested that a signal plus noise model can be used to fit the difference of the traffic flow as follows.

$$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_2\cos(2\pi f_2 t) + B_2\sin(2\pi f_2 t) + \ldots + W_t$$

where ∇Y_t is the transformed (difference) time series. (From the textbook chapter 13)

The periodogram clearly shows that the series contains four cosine-sine pairs. The frequency ordered from high to low is f1=10/240*0.5 = 0.0208, f2= 20/240*.5=0.0417, f3= 60/240*0.5 = 0.125, f4=30/240*0.5 =0.0625

Note f1 is the higher-frequency component that is much stronger. There are some other very small spikes in the periodogram, apparently caused by the additive white noise component. Let's test the model with f1. The summary table is as follows

Signal plus noise model	Multi R^2	Adjusted R^2	F-Stat	P-Value
$\nabla Y_t = A_1\cos(2\pi f_1 t) + W_t$	0.02931	0.02728	14.41	0.0001664
$\nabla Y_t = B_1\sin(2\pi f_1 t) + W_t$	0.093	0.0911	48.91	9.097e-12
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + W_t$	0.1224	0.1187	33.18	3.235e-14

Based on the table above, it is found that the coefficient of both $\cos(2\pi f_1 t)$ and $\sin(2\pi f_1 t)$ are significant at 5% level.

Next step we will repeat testing model with addition of f2, f3 and f4's coefficient and assess the corresponding significant level. Table result as follows:

Signal Plus Noise Model	Multi R^2	Adjusted R^2	F-Stat	P-Value
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_2\cos(2\pi f_2 t) + W_t$	0.1847	0.1796	35.87	< 2.2e-16
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_2\sin(2\pi f_2 t) + W_t$	0.123	0.1174	22.2	1.817e-13
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_3\cos(2\pi f_3 t) + W_t$	0.1489	0.1436	27.71	< 2.2e-16
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_3\sin(2\pi f_3 t) + W_t$	0.1396	0.1342	25.7	2.017e-15
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_4\cos(2\pi f_4 t) + W_t$	0.1339	0.1285	24.49	9.499e-15
$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_4\sin(2\pi f_4 t) + W_t$	0.1465	0.1411	27.18	3.087e-16

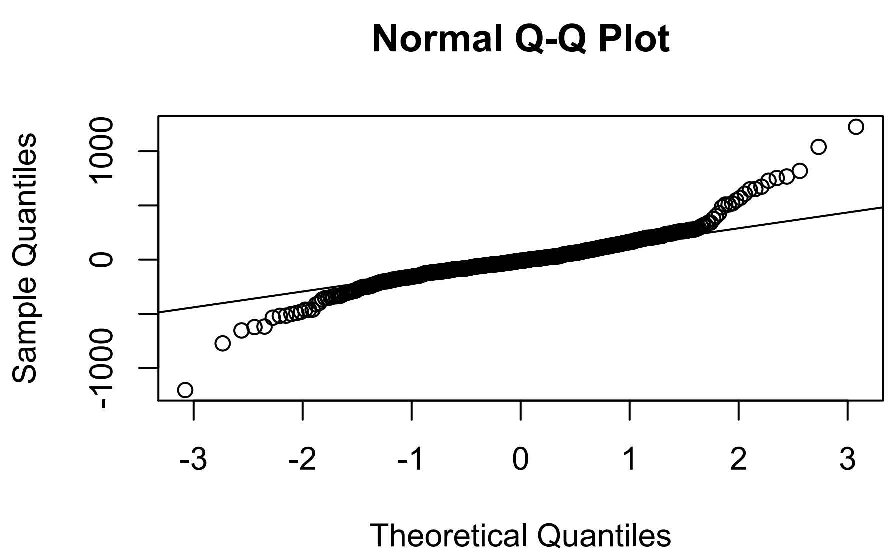
The testing results indicate that $\cos(2\pi f_1 t)$, $\sin(2\pi f_1 t)$, $\cos(2\pi f_2 t)$, $\cos(2\pi f_3 t)$ and $\sin(2\pi f_4 t)$ can potentially be part of the model as they give the lowest p-value.

After we tried different combination of the model in R. It is found that a model of

$$\nabla Y_t = A_1\cos(2\pi f_1 t) + B_1\sin(2\pi f_1 t) + A_2\cos(2\pi f_2 t) + A_3\cos(2\pi f_3 t) + W_t$$

has the lowest p-value. Please refer to the appendix (R-Code) below for details.

Lastly, we perform the normal Q-Q plot to verify the goodness of fit.



Overall we can see the model fits well into a normal distribution except a few residual outliers on the tail. These outliers may be explained by inconsistent driving conditions (i.e. road maintenance, car model evolutes over years, etc)

Conclusion

The following model has been considered a potential good model that will fits the time series data.

$$\nabla Y_t = A_1 \cos(2\pi f_1 t) + B_1 \sin(2\pi f_1 t) + A_2 \cos(2\pi f_2 t) + A_3 \cos(2\pi f_3 t) + W_t$$

Please refer to the appendix (R-Code) below for details. However, please note that there could be other time series model might also fit well, it is not restricted to this model only.

Reference

1. Cryer, Jonathan cand Chan, Kung-Sik, 2008. Time series Analysis With Applications in R
2. PEMS website: <http://pems.dot.ca.gov>

Appendix: R-Code with process explanation

```
library(TSA)

# Reading the 2002 to 2010 traffic flow data into R
flow10 <- read.csv('~/Documents/VEE TS/Data/flow_10.csv')
flow08 <- read.csv('~/Documents/VEE TS/Data/flow_08.csv')
flow06 <- read.csv('~/Documents/VEE TS/Data/flow_06.csv')
flow04 <- read.csv('~/Documents/VEE TS/Data/flow_04.csv')
flow02 <- read.csv('~/Documents/VEE TS/Data/flow_02.csv')

# Mulnipliate the data from 5 min into 30 mins period (shorten the
points)
flow10 <- as.numeric(flow10[1:576,2])
flow08 <- as.numeric(flow08[1:576,2])
flow06 <- as.numeric(flow06[1:576,2])
flow04 <- as.numeric(flow04[1:576,2])
flow02 <- as.numeric(flow02[1:576,2])

fl10 <- rep(0,96)
fl.10 <- matrix(flow10, ncol=6, byrow=TRUE)
for(i in 1:96){
     fl10[i] <- sum(fl.10[i, 1:6])}

fl08 <- rep(0,96)
fl.08 <- matrix(flow08, ncol=6, byrow=TRUE)
for(i in 1:96){
     fl08[i] <- sum(fl.08[i, 1:6])}


fl06 <- rep(0,96)
fl.06 <- matrix(flow06, ncol=6, byrow=TRUE)
for(i in 1:96){
     fl06[i] <- sum(fl.06[i, 1:6])}


fl04 <- rep(0,96)
fl.04 <- matrix(flow04, ncol=6, byrow=TRUE)
for(i in 1:96){
     fl04[i] <- sum(fl.04[i, 1:6])}


fl02 <- rep(0,96)
fl.02 <- matrix(flow02, ncol=6, byrow=TRUE)
for(i in 1:96){
     fl02[i] <- sum(fl.02[i, 1:6])}

flow_all <- c(fl02, fl04, fl06, fl08, fl10)
```

```r
# convert the data into a time series
flow.all <- ts(flow_all, start=1, frequency=96)
plot(flow.all, main="Traffic Flow from 2002 to 2010", xlab="Year",
ylab="Flow of Traffic")

# observe the ACF & PACF
acf(flow.all, lag.xax=200, main="ACF of the Traffic Flow" )

# Transform data. Takeing difference to reduce the reduce/remove the
signial of seasonality
diff <- diff(flow.all)
plot(diff, main="Difference of Traffic Flow from 2002 to 2010",
xlab="Year", ylab="Difference of Flow")

# test the ACF & PACF of the diff
acf(diff, lag.max=200, main="ACF of the Difference of Traffic Flow")
pacf(diff, lag.max=200, main= "PACF of the Difference of Traffic Flow")

##check the periodogram to see the period of the series
spec.pgram(diff, k=kernel("modified.daniell", c(4,4)), taper=0,
detrend=FALSE, demean=TRUE, log="no", main="Smoothed Periodogram of the
Difference of Flow")
periodogram(diff, main="Periodogram of the Difference of Flow")

# find that f1=10/240*0.5 = 0.0208, f2= 20/240*.5=0.0417, f3=
60/240*0.5 = 0.125, f4=30/240*0.5 =0.0625, try sin & cos model with
frequence = 1
# sin(2*w*pi*f1)+cos(2*w*pi*f1), w=2*pi*t, t=1:length(flow.t)
t=1:length(diff)
w=2*pi*t
f1 <- 0.0208
model.l1 <- lm(diff~cos(w*f1))
summary(model.l1)
model.l2 <- lm(diff~sin(w*f1))
summary(model.l2)
model.l3 <- lm(diff~sin(w*f1)+cos(w*f1))
summary(model.l3)

# Test with frequency = 2, 3 & 4

f2 <- 0.0417
f3 <- 0.125
f4 <- 0.0625
model.4 <- lm(diff~cos(w*f1)+sin(w*f1)+cos(w*f2))
summary(model.4)
model.5 <- lm(diff~cos(w*f1)+sin(w*f1)+sin(w*f2))
summary(model.5)
```

```
model.6 <- lm(diff~cos(w*f1)+sin(w*f1)+cos(w*f3))
summary(model.6)
model.7 <- lm(diff~cos(w*f1)+sin(w*f1)+sin(w*f3))
summary(model.7)
model.8 <- lm(diff~cos(w*f1)+sin(w*f1)+cos(w*f4))
summary(model.8)
model.9 <- lm(diff~cos(w*f1)+sin(w*f1)+sin(w*f4))
summary(model.9)

# In order to improve the goodness of fit, let's try couple more
models.
model.l4 <- lm(diff~sin(w*f1)+cos(w*f1)+cos(w*f2)+cos(w*f3)+sin(w*f4))
summary(model.l4)
model.l5 <- lm(diff~sin(w*f1)+cos(w*f1)+cos(w*f3)+sin(w*f4))
summary(model.l5)
model.l6 <- lm(diff~sin(w*f1)+cos(w*f1)+cos(w*f2)+cos(w*f3))
summary(model.l6)
model.l7 <- lm(diff~sin(w*f1)+cos(w*f2))
summary(model.l7)
model.l8 <- lm(diff~sin(w*f1)+cos(w*f3)+sin(w*f4))
summary(model.l8)
model.l81 <- lm(diff~cos(w*f1)+cos(w*f3))
summary(model.l81)
model.l9 <- lm(diff~cos(w*f1)+cos(w*f2)+cos(w*f3))
summary(model.l9)
model.l91 <- lm(diff~cos(w*f1)+cos(w*f3)+sin(w*f4))
summary(model.l91)

# After we tried different combination of the model, model l6 fitts the
best among all. See below test statistics

Call:
lm(formula = diff ~ sin(w * f1) + cos(w * f1) + cos(w * f2) +
    cos(w * f3))

Residuals:
     Min       1Q   Median       3Q      Max
-1203.95  -100.90   -10.71    95.90  1226.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2185    10.1264   0.022    0.983
sin(w * f1) 106.9159    14.2945   7.480 3.65e-13 ***
cos(w * f1) -60.5966    14.3471  -4.224 2.88e-05 ***
cos(w * f2) -88.0002    14.3300  -6.141 1.74e-09 ***
cos(w * f3) -57.6861    14.3359  -4.024 6.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 221.6 on 474 degrees of freedom
Multiple R-squared: 0.2116, Adjusted R-squared: 0.205
F-statistic: 31.81 on 4 and 474 DF,  p-value: < 2.2e-16

# Goodness of Fit, Q-Q plot on residuals
res.l6 <- diff - fitted(model.l6)
qqnorm(res.l6)
qqline(res.l6)
```