David Oluwagbenga Agboola

SRM 637 Spring 2020

INTERVENTION ANALYSIS OF DAILY COUNT OF RENTAL BIKES IN WASHINGTON D.C.

Purpose of the Analysis

Bike sharing systems are getting popular and becoming a new generation of traditional bike rentals where the whole process from membership, rental and return has become automated which is easier than before. Through these systems, a user can easily rent a bike from a position (or station) and return the bike at another position (or station) without stress or paperwork. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. There are recent research interests in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike sharing systems, the features of data being generated by these systems make them attractive for research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position are explicitly recorded in these systems. These features turn bike sharing systems into a virtual sensor network that can be used for sensing mobility in a city or location. Hence, it is expected that most of the important events in a city or location could be detected via monitoring these data. Washington D.C. has hundreds of bike sharing stations operated by Capital Bikeshare. The analysis will involve using predictors - weather situation, holiday, feeling temperature, humidity, wind-speed, amongst others influence the average number of bikes rented daily.

The purpose of this study is to carry out an intervention analysis on the bike-sharing dataset.

Data Collection and Description

The data-set "Bike-Sharing-Dataset" was collected by the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto which contains the daily count of rental bikes for the years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA . The data-set has 731 observations. Data type of count is integer representing count of total rental bikes.

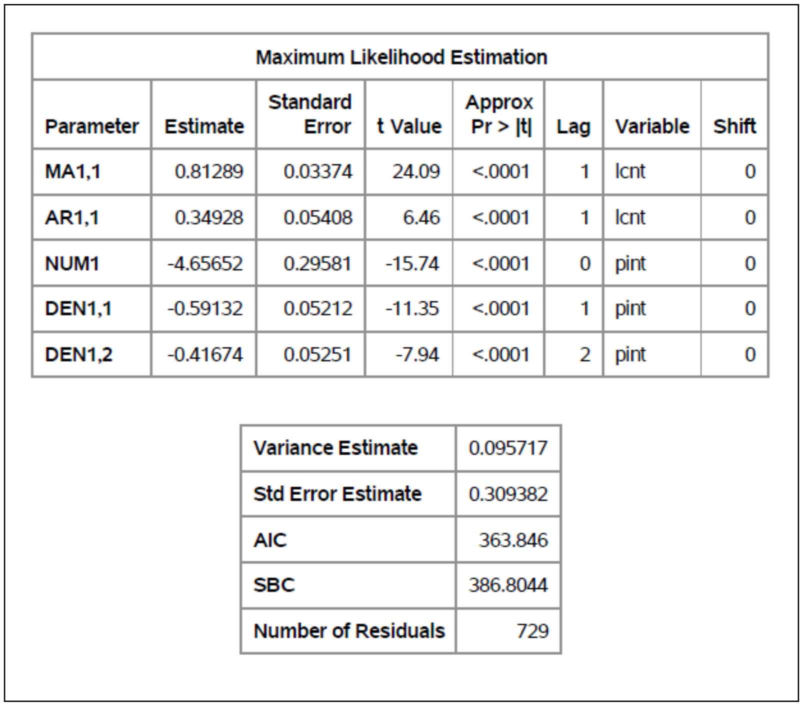
Analysis

The time plot, sample ACF, and PACF plots are shown in Figure 1.

Observe from the time plot in Figure 1(a) that there is a trend in the data and the variances are not stable. Though it seems like there is seasonality in the time plot, but the sample ACF and PACF do not show an evidence of seasonality. In fact, applying seasonal differencing will not make the process stationary and the model will not perform better according to the diagnostic checks. These were not shown in the report. It was explored while trying to find the most appropriate model.

The sample ACF shows that the data is not stationary (See Figure 1(b)). So, log transformation (using Box-Cox) and simple differencing are applied to stabilize the variance and remove the trend respectively. The outcome is shown in Figure 2.

Now, observe there is a pulse intervention on day 668. This is October 29, 2012 when D.C. declared a state of emergency to prepare for the Hurricane Sandy. It is no surprise that the count of bike for that day reached a low of 22 which is the lowest ever recorded in the data-set. The intervention effect died out gradually over time. So the pulse function input had a transfer function with denominator factor of order 2.



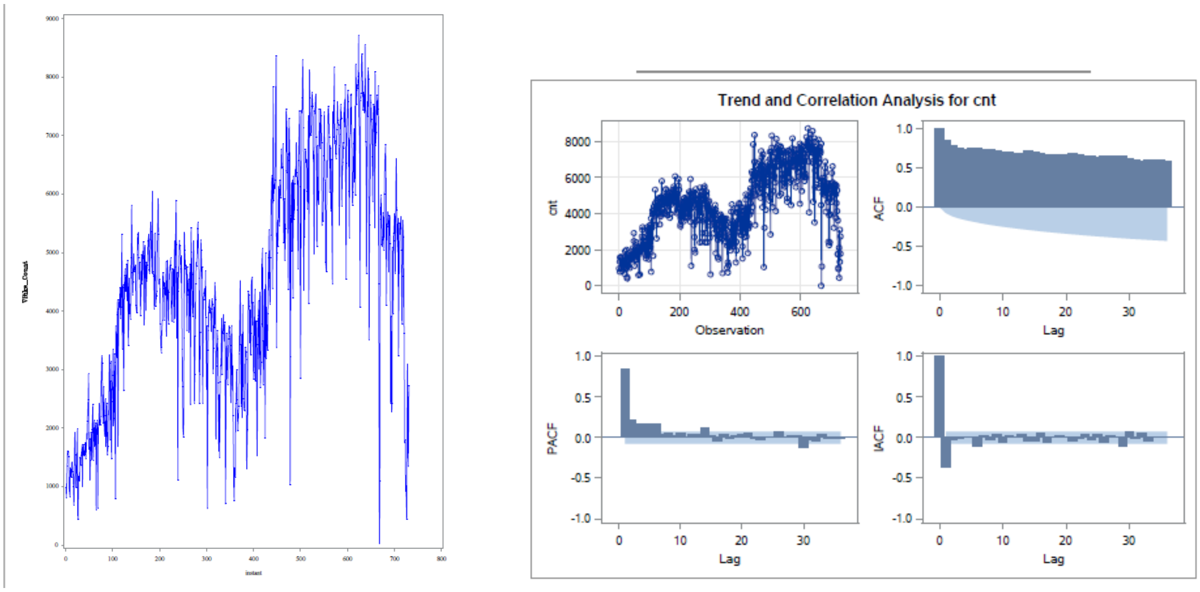
Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.81289	0.03374	24.09	<.0001	1	lcnt	0
AR1,1	0.34928	0.05408	6.46	<.0001	1	lcnt	0
NUM1	-4.65652	0.29581	-15.74	<.0001	0	pint	0
DEN1,1	-0.59132	0.05212	-11.35	<.0001	1	pint	0
DEN1,2	-0.41674	0.05251	-7.94	<.0001	2	pint	0

Variance Estimate	0.095717
Std Error Estimate	0.309382
AIC	363.846
SBC	386.8044
Number of Residuals	729

From above, the fitted model to the count of daily rented bikes in the Capital Bike-Share system in D.C. is:

$$\log(Y_t) = \frac{-4.657}{1-0.591\boldsymbol{B}-0.418\boldsymbol{B}^2} P_t^{668} + X_t$$

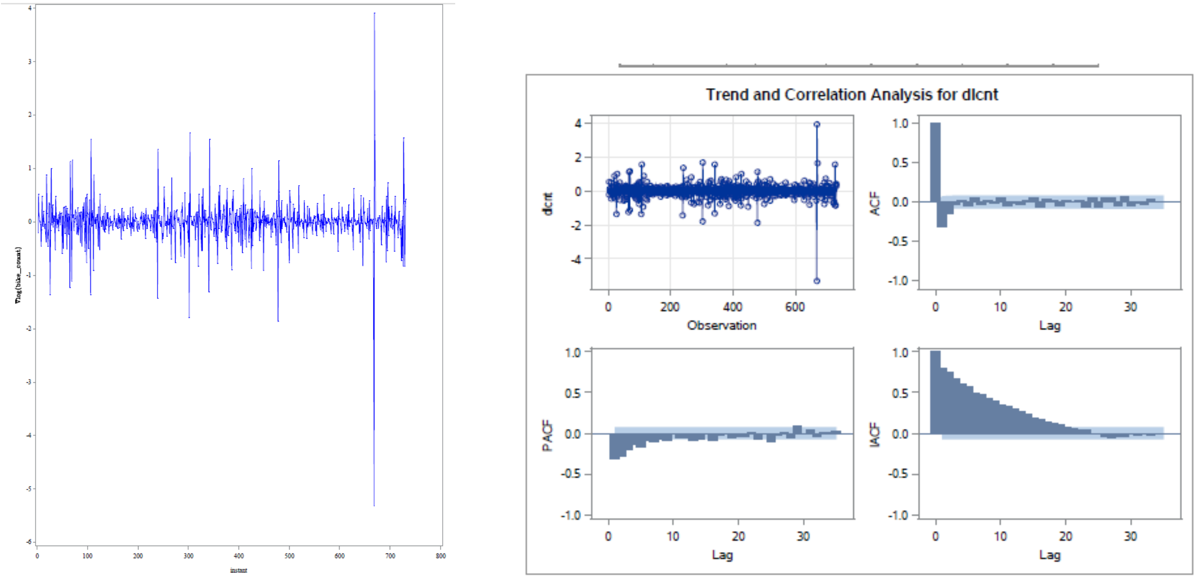
where $X_t = \frac{\Theta_q(\boldsymbol{B})}{\Phi_p(\boldsymbol{B})} Z_t = \frac{(1+0.813\boldsymbol{B})}{(1+0.349\boldsymbol{B})} Z_t, Z_t$ is a Gaussian white noise with mean 0 and variance 0.09572.



(a) Time Plot

(b) Trend Analysis

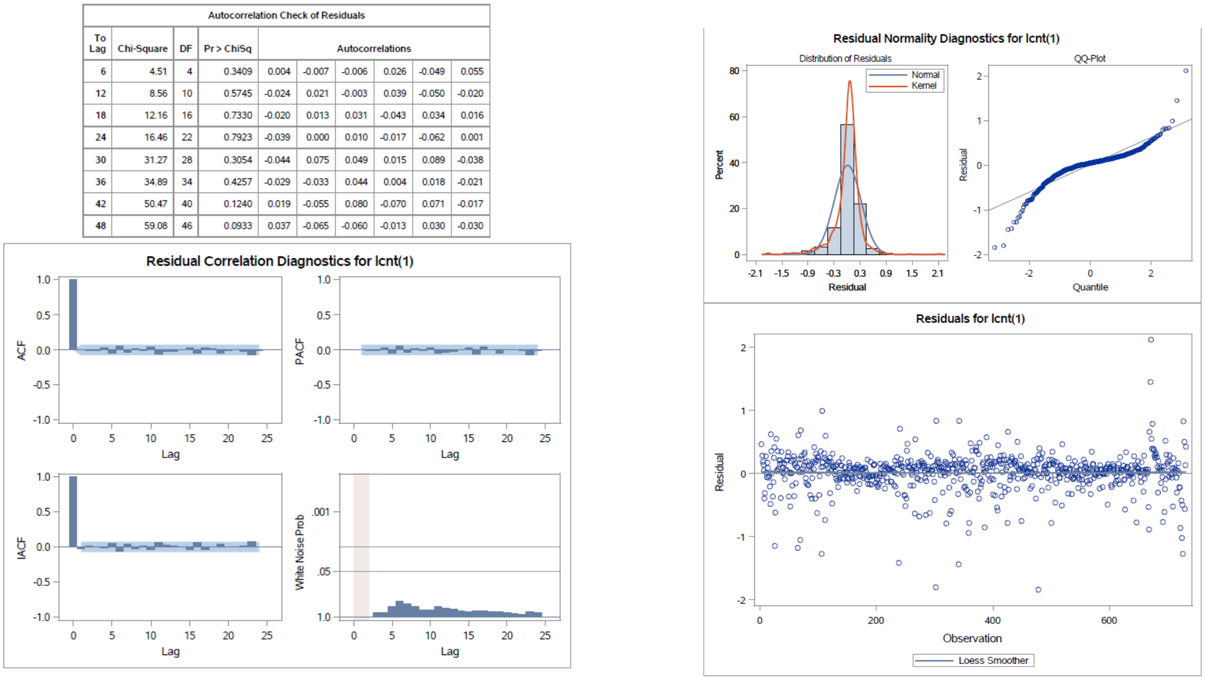
Figure 1: Trend and Correlation Analysis



(a) Time Plot

(b) Trend Analysis

Figure 2: Trend and Correlation Analysis after applying log transformation and simple differencing



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.51	4	0.3409	0.004	-0.007	-0.006	0.026	-0.049	0.055
12	8.56	10	0.5745	-0.024	0.021	-0.003	0.039	-0.050	-0.020
18	12.16	16	0.7330	-0.020	0.013	0.031	-0.043	0.034	0.016
24	16.46	22	0.7923	-0.039	0.000	0.010	-0.017	-0.062	0.001
30	31.27	28	0.3054	-0.044	0.075	0.049	0.015	0.089	-0.038
36	34.89	34	0.4257	-0.029	-0.033	0.044	0.004	0.018	-0.021
42	50.47	40	0.1240	0.019	-0.055	0.080	-0.070	0.071	-0.017
48	59.08	46	0.0933	0.037	-0.065	-0.060	-0.013	0.030	-0.030

(a) Residual Correlation Diagnostics

(b) Residual Normality Diagnostics

Figure 3: Residual Correlation and Normality Diagnostics

For residual diagnostic checking

Observe from the autocorrelation check of residuals plot in Figure 3(a) that that the p-values are large enough to fail to reject the randomness. Residuals were checked and they follow a normal distribution as shown in Figure 3(b).

Conclusion and Recommendation

In conclusion, carrying out an intervention analysis reveals the presence of an intervention which is not just an outlier. In this case, it was a pulse function resulting from a state of emergency declared in D.C. during the period of the data collection. Due to the nature of the intervention, it appeared to gradually die out over time. This informed the decision to apply a transfer function. The final pulse intervention model fitted appear to be appropriate for the data-set by checking the diagnostics.

It could also be of interest to forecast and check the forecast errors. For this, the pulse indicator function should be extended for the period of forecast before the forecast of the count can be estimated. This is peculiar to situations where independent variables are used in the model. In this case, the independent variable is the pulse function. I tried showing the forecast but I was getting inappropriate values. I would take time to check this later to figure out what is going wrong.

Finally, if I had more time, I would have compared the forecast with predicted values from classical methods like negative binomial regression or Bayesian time series and Bayesian regression methods. But since it is a correlated data, the regression methods would not be appropriate.

REFERENCES

Box, G.E.P. and Tiao, G.C. (1975). *Intervention analysis with applications to economic and environmental problems.* JASA, 70, 70-79.

Chatfield, C. (2005). *The analysis of time series: An introduction* Chappman & Hall/CRC.

Wei, W.W.S. (2006). *Time series analysis: Univariate and multivariate methods.* Pearson Education Inc.

Woodfield, T.J. (1987). *Time Series Intervention Analysis Using SAS Software.* Proceedings of the Twelfth Annual SAS Users Group International Conference, 331-339. Cary, NC: SAS Institute Inc.

APPENDIX

The output/code/data analysis for this paper was generated using SAS software, Version [9.4] of the SAS System for [Windows]. Copyright [2020] SAS Institute Inc.

```
proc import datafile="Z:\OneDrive - University of Northern Colorado\Spring
   2020\Time series\Data\day.csv"
    out= DData
    dbms=csv
    replace;
    getnames=yes;
run;

goptions ftext="Times New Roman" ctext=BLACK htext=1 cells;

axis1 width=1 label=( a=90 r=0 F=CGREEK ’V’ F=COMPLEX "Bike_Count" )
   minor=none;
axis3 width=1 label=( a=90 r=0 h=1 F=CGREEK ’V’ h=1.2 f=amigas ’12’ h=1
  F=CGREEK ’V’ F=COMPLEX "Bike_Count" ) minor=none;
axis2 width=1 minor=none;
/*order=(0 to 25 by 1);*/
symbol1 c=BLUE ci=BLUE v=dot height=.25 cells
       interpol=join l=1 w=1;
proc gplot data=ddata;
plot cnt* instant / vaxis = axis1 haxis = axis2 ;
run;
quit;

PROC ARIMA DATA=DData plots(only)=(series(all));
IDENTIFY VAR=cnt NLAG=36 ;
run;

/** Using Box-Cox Transformation to determine lambda **/
PROC TRANSREG DATA=DData;
  MODEL BOXCOX(cnt)=IDENTITY(instant);
QUIT;


/* Applying log transformation and taking simple differencing */
data ddata;
set ddata;
lcnt=log(cnt);
dlcnt= DIF(lcnt);
run;

goptions ftext="Times New Roman" ctext=BLACK htext=1 cells;

axis1 width=1 label=( a=90 r=0 F=CGREEK ’V’ F=COMPLEX "log(bike_count)" )
```

```
   minor=none;
axis3 width=1 label=( a=90 r=0 h=1 F=CGREEK 'V' h=1.2 f=amigas '12' h=1
  F=CGREEK 'V' F=COMPLEX "log(bike_count)" ) minor=none;
axis2 width=1 minor=none;
/*order=(0 to 25 by 1);*/
symbol1 c=BLUE ci=BLUE v=dot height=.25 cells
       interpol=join l=1 w=1;
proc gplot data=ddata;
plot dlcnt* instant / vaxis = axis1 haxis = axis2 ;
run;
quit;

PROC ARIMA DATA=DData plots(only)=(series(all));
IDENTIFY VAR=dlcnt NLAG=35 ;
run;


data pdata;
set ddata;
pint = instant=668;
run;

proc arima data=pdata plots(only)=(residual( all));
identify var=lcnt(1) crosscorr=pint;
estimate p=1 q=1 input=( / (1,2) pint) method=ml noconstant ;
run;
```