ARIMA Model of Chicago Bike Thefts with Transfer Function Using Gas Prices as Input Series

By

**Andrew J. Hillard**

Middlebush Hall, 900 University Ave

Columbia, MO 65211, USA

E-mail: [ajhkt6@mail.missouri.edu](mailto:ajhkt6@mail.missouri.edu)

**Abstract**

This paper proposes a model for monthly Chicago bike thefts using gas prices as an input series. The proposed model is an ARIMA (1, 0, 0) x (2, 1, 0)12 with a (3, 1, 0) transfer function. The proposed model is created using two datasets: a Chicago crime dataset that provides the number of bike thefts from January of 2001 to December of 2014, and a Chicago gas price dataset from October of 2000 to September of 2014. The ARIMA (1, 0, 0) x (2, 1, 0)12 models the trend of monthly Chicago bike thefts. The relationship between gas prices and bike theft is modeled by the transfer function with delay parameter of b=3 and an AR(1), or r=1. The paper concludes by comparing the model’s forecasts to actual observations as well as a discussion of model limitations.

**Introduction**

Biking is an environmentally friendly and healthy alternative to driving a car. Many cities throughout the U.S. have implemented policies that encourage biking for transportation. Even mid to small sized cities like Columbia, Missouri, are providing their residents with bike maps, bike parking, clearly marked bike lanes, and bike safety awareness programs (The District Downtown Columbia*)*.

Though more and more cities are providing resources to bikers, few cities are talking about bike security. In a recent study published in the International Journal of Sustainable Transportation, it was found that about half of all active cyclists have their bikes stolen (Van Lierop). The high risk of having your bike stolen makes it important to better understand bike security and the factors that influence bike theft.

In this paper, the relationship between bike theft and gas prices is explored and a time series model is created to quantify this relationship. If there is a relationship between bike theft and gas prices, the number of bike thefts will be more accurately predicted when gas prices are taken into account. With better prediction of bike theft, bike riders can be more aware when their bike is most at risk. The relationship between bike theft and gas prices was chosen because gas prices may have an impact on bike theft. When gas prices are higher, there is greater demand for cost effective means of transportation, i.e. bikes, and a greater incentive to steal and sell bikes to meet that demand.

The data in this study is from the City of Chicago. The number of monthly bike theft incidents is calculated using the City of Chicago’s database of crimes committed since 2001. The monthly gas prices in Chicago since 2000 are calculated using data from the U.S. Energy Information Administration. Though the model of bike thefts is specific to Chicago data, the study methods can be generalized to other cities.

**Data Cleaning**

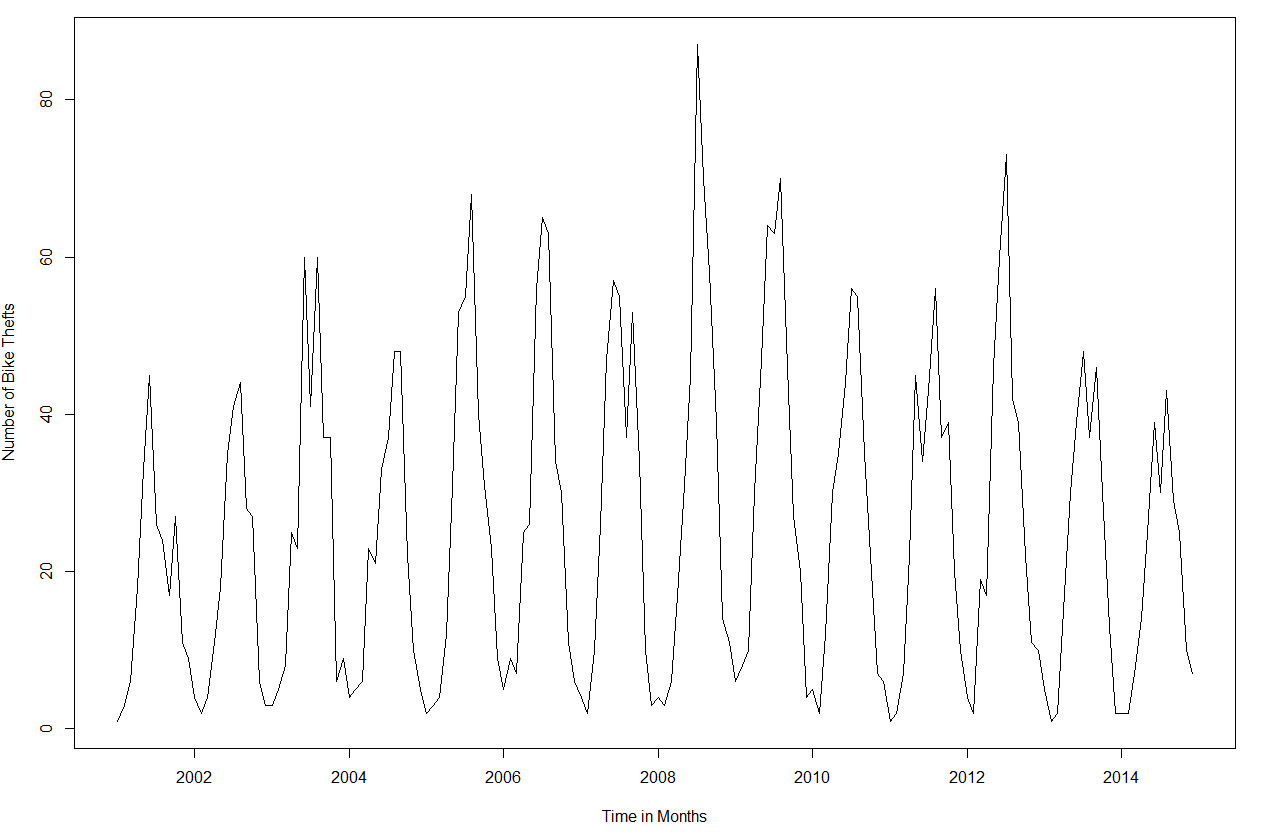
The City of Chicago database provides a record of crimes committed since January of 2001. Each row in the dataset corresponds to a reported crime. The crime data was filtered to only include incidents of cycle, scooter, and bike theft. The count of cycle, scooter, and bike theft by month was then calculated using R and Excel. The monthly count of cycle, scooter, and bike theft will be referred to as “theft.” The ARIMA transfer function model will be created using the theft data from January of 2001 to December of 2014. The remaining theft data, January 2015 to November of 2015, will be used to check the veracity of the model.

The U.S. Energy Information Administration provided a record of Chicago gas prices by week starting in June of 2000. The average monthly gas price was calculated using R and Excel. The cleaned average monthly gas price data will be referred to as “gas.”

**Model Specification**

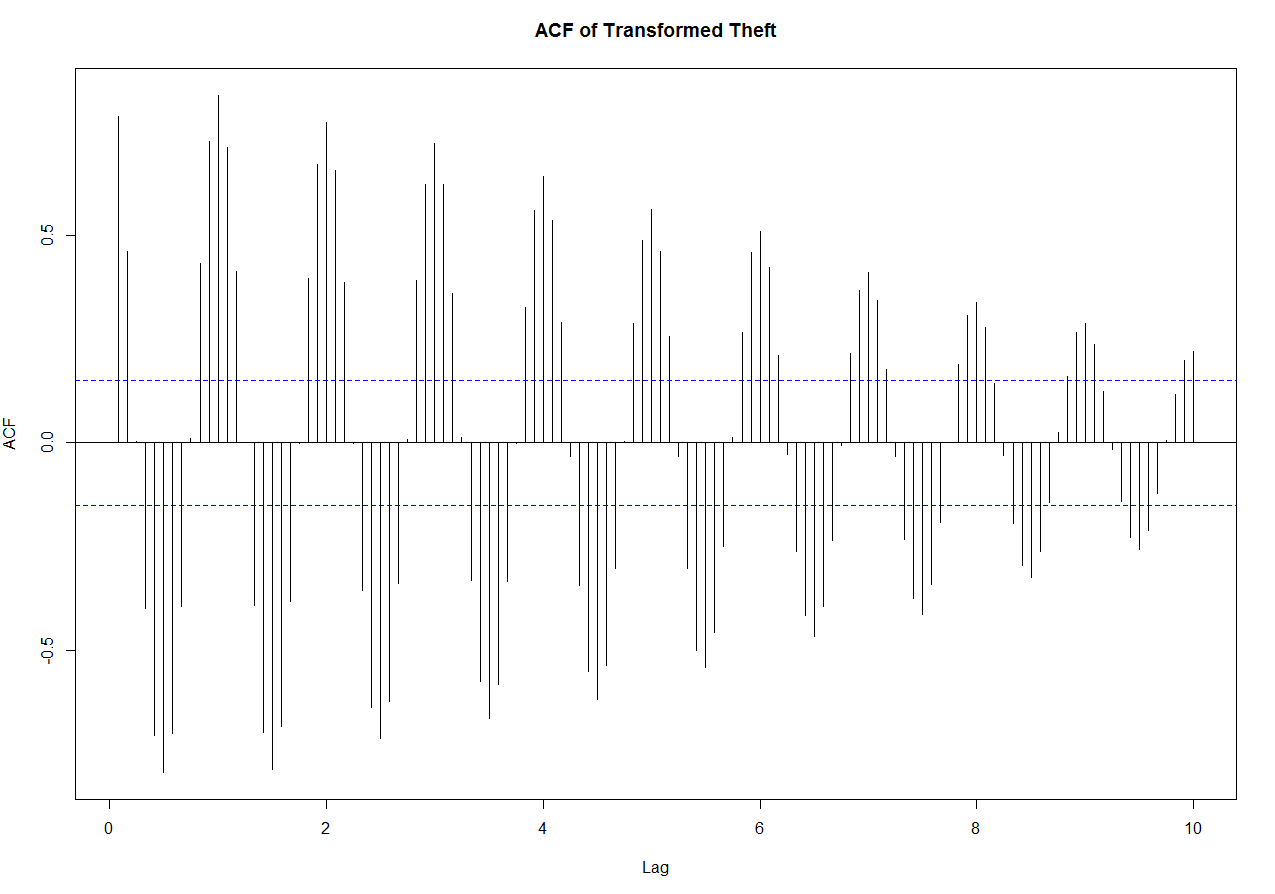
Exhibit 1 shows the time series plot for the theft data. The time series plot is graphed to check for stationarity of the mean and variance. The theft data appears to be stationary, which is confirmed by a Dickey Fuller test with p-value less than 0.01. The variance appears to be slightly unequal over time. The Box Cox test on theft data confirms that the variance is unequal over time and that a transformation with lambda = 0.4 would be appropriate. The time series for theft is transformed with lambda = 0.4. The plot of the transformed data looks very similar to Exhibit 1 and can be found in the Appendix A, Exhibit A.1.

**Exhibit 1: Time Series Plot for Theft Data**



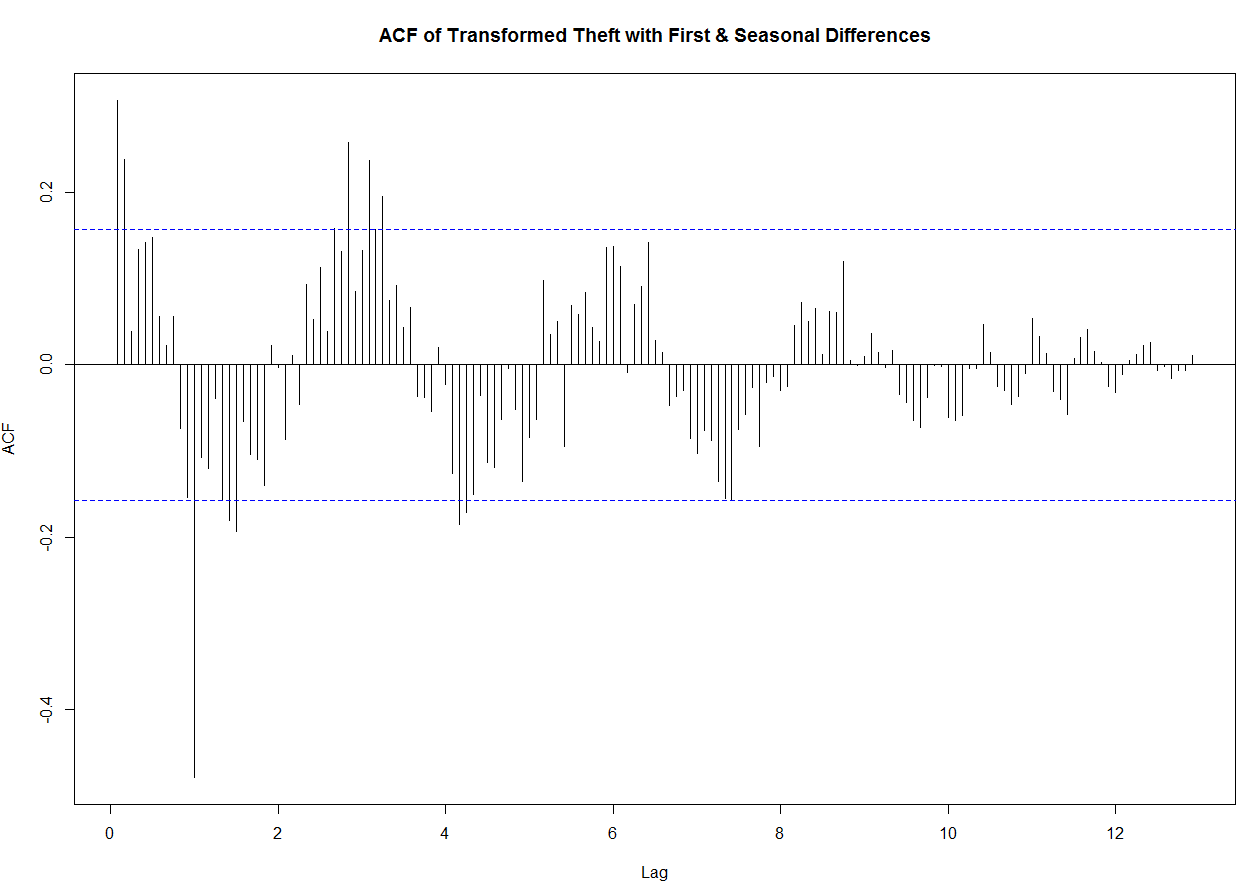
The time series plot for theft data shows 12 month seasonality. The ACF of the transformed theft data is plotted in Exhibit 2 to check if a seasonal difference operator is required. The ACF shows clear decay for spikes every 12 lags. Therefore, one seasonal difference is applied with the seasonal parameter s=12.

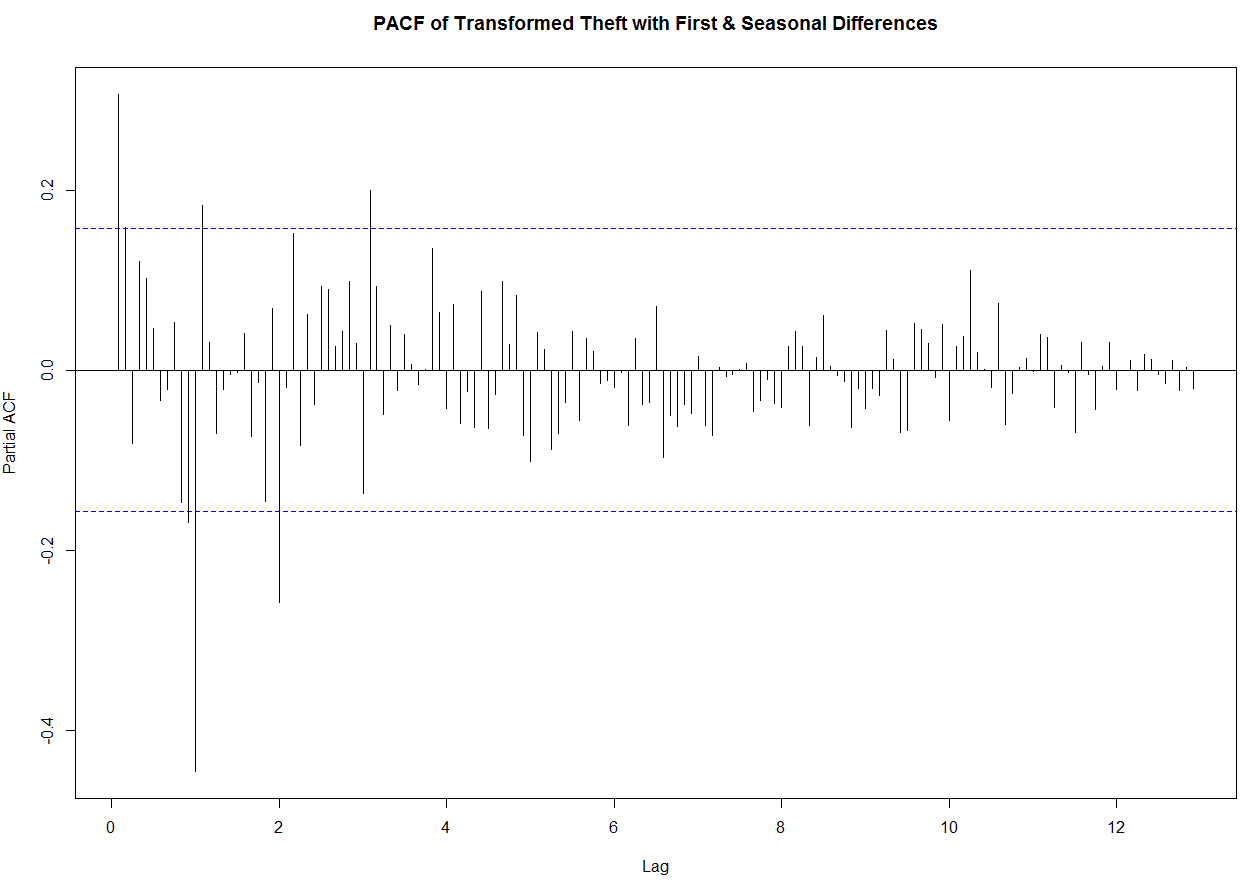
**Exhibit 2: ACF of Transformed Theft Data**



The times series plot of the theft data after the first difference and seasonal first difference is shown in Appendix A, Exhibit A.2. Though the plot in Exhibit A.2 no longer shows 12 month seasonality, there still appears to be some sort of seasonality with peaks and troughs appearing at equal distances over time. The ACF & PACF of the transformed theft data with first and seasonal difference is shown in Exhibit 3. The ACF shows a damped sine wave and the PACF cuts off after two seasonal spikes, indicating a seasonal AR(2) model. Therefore, the transformed theft data is fit with an ARIMA (0,0,0) x (2, 1, 0)12 model.

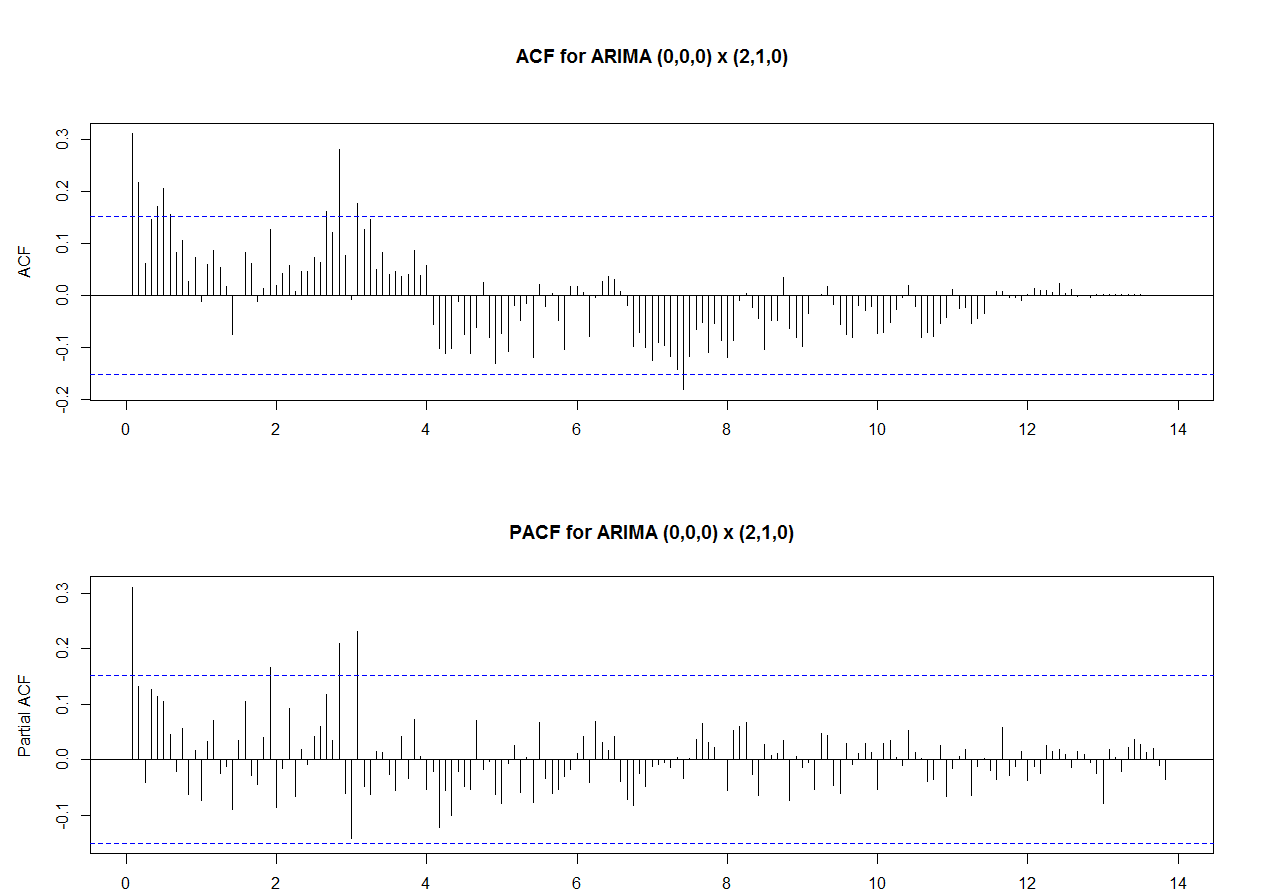
**Exhibit 3: ACF and PACF of Theft after First & Seasonal Differences**





The ACF and PACF of the residuals from the ARIMA (0,0,0) x (2, 1, 0)12 fitted model are shown in Exhibit 4 and analyzed to determine p and q in the non-seasonal part of the ARIMA model. The ACF in Exhibit 4 tails off over time. The PACF in Exhibit 4 shows cutting off behavior after non-seasonal lag 1, indicating that an ARMA(1,0) model would be a good fit.

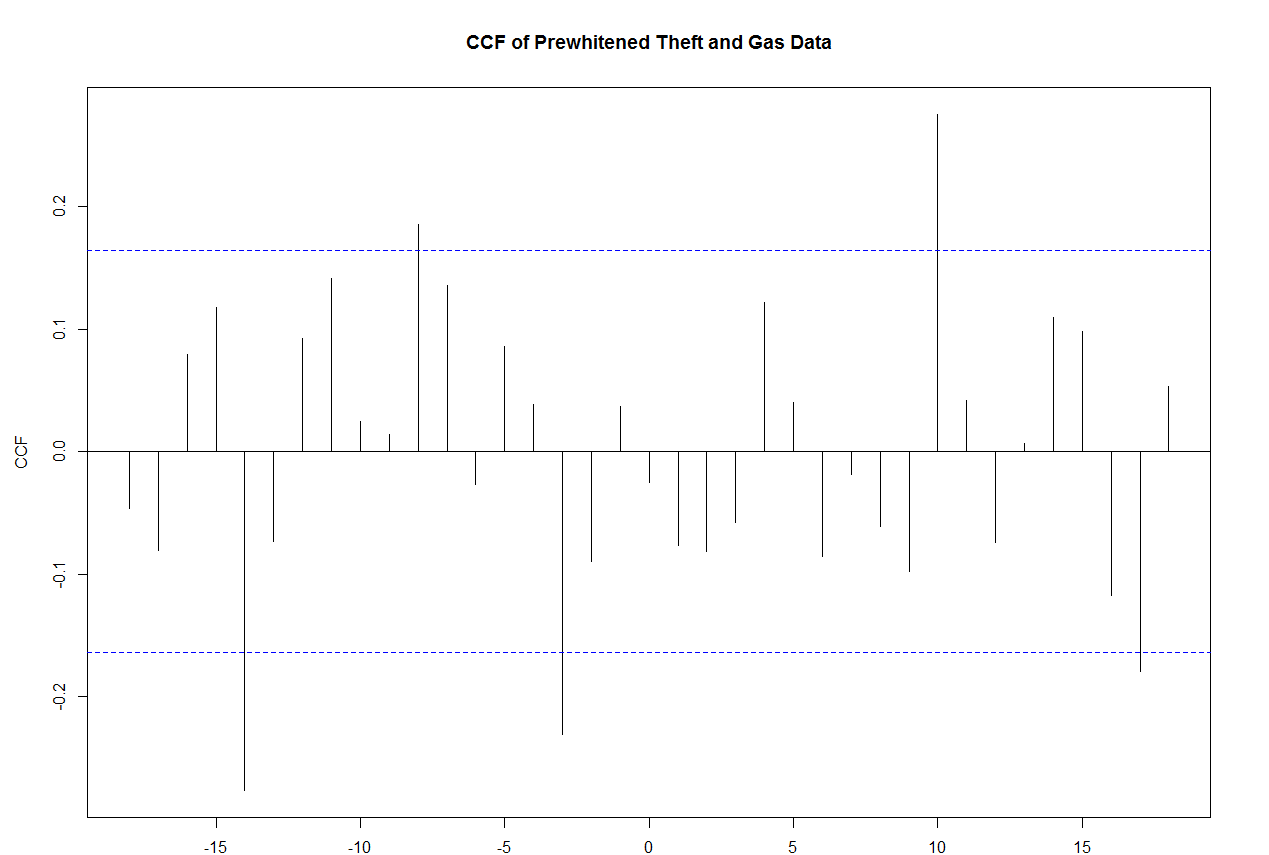
**Exhibit 4: ACF and PACF of ARIMA (0,0,0) x (2,1,0)12 fitted to Theft Data**



The transformed theft data is therefore fitted with an ARIMA (1, 0, 0) x (2, 1, 0)12 model The residuals for this fitted model are checked for normality and independence by looking at the residuals’ ACF, histogram, and normal probability plots as shown in Appendix A, Exhibit A.3. All graphs support the hypothesis that normality and independence of residuals can be assumed. The residuals are tested using the Shapiro-Wilk as well as the Runs test. At the 5% level, the two tests support normality and independence respectively. The ARIMA (1, 0, 0) x (2, 1, 0)12 model is therefore an appropriate fit for the theft data. In addition, the forecast of two seasons in Appendix A.4 using the ARIMA (1, 0, 0) x (2, 1, 0)12 closely matches what is expected and observed.

Since the ARIMA model has been selected for the theft data, the transfer function using gas data can now be determined. To determine the transfer function, the cross correlation between gas and theft is plotted in Appendix A, Exhibit A.5. The CCF shows spurious seasonal correlation. The spurious correlation is removed by prewhitening the data. The data is approximately prewhitened by taking the seasonal first difference of theft, and the seasonal and first difference of gas (Cryer 266). The CCF is plotted in Exhibit 5 with the prewhitened data. The CCF shows correlation between gas and theft data at a lag of -3, -11, and 17. The correlations at lag of -11 and 17 do not seem plausible as there seems to be little reason for there to be connection between theft and gas prices 11 months or 17 months in the past or future. The correlation at lag -3 is more plausible as gas prices 3 months ago could have an effect on the demand for bikes in the present, affecting the incentive to steal bikes. The transfer function (b,r,s) will therefore be modeled with time delay b=3. See Appendix B for transfer function representation in this paper

**Exhibit 5: CCF of Prewhitened Gas and Theft Data**



Once b is identified, the values of r and s are found using an ARMA model. The ARMA models considered were ARMA(1, 0), ARMA(0, 1), and ARMA(1, 1). These ARMA models were fitted using the ARIMAX R function with the underlying assumption that the theft data will be modeled with an ARIMA (1, 0, 0) x (2, 1, 0)12 as previously discussed. All the ARIMA transfer function models had similar AICs and log likelihood values. However, the only model that had significant estimates for the coefficients of the transfer function was when r and s were fitted using the ARMA(1, 0 ) model. The ARMA(1, 0) model also makes sense as the price of gas may have a lingering effect on consumer spending over time. The level of gas prices for one month may not affect the demand for bikes. However, sustained high or low gas prices over a number of months may affect the demand for bikes. The ARMA(1, 0) with lag b=3 was therefore selected to model the transfer function.

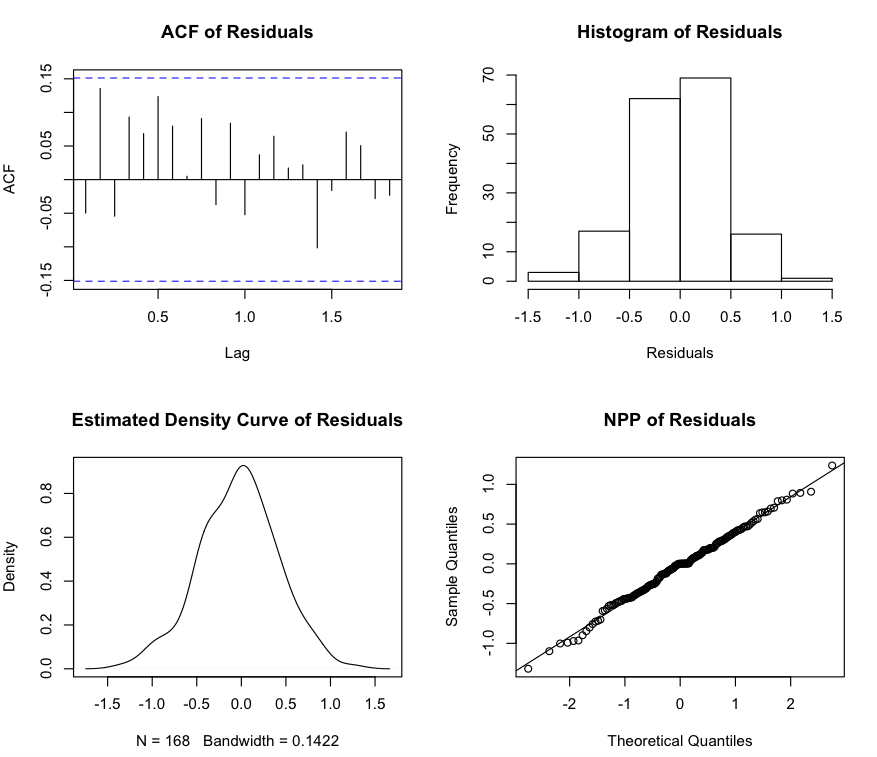
Below is the generic model for the ARIMA (1, 0, 0) x (2, 1, 0)12 with a transfer function (b=3, r=1, s=0). Y is the time series variable for theft, X is the time series variable for gas.

Below is the fitted model with coefficients estimates given by the ARIMAX R function.

**Fitting and Diagnostics**

The residuals for the ARIMA (1, 0, 0) x (2, 1, 0)12 with transfer function (b=3, r=1, s=0) are checked for independence and normality to assess the validity of the model. The residuals’ ACF, histogram, estimated density curve, and normal probability plot are graphed in Exhibit 6 and analyzed for independence and normality. All the graphs support the assumption that the residuals for the fitted model are independent and normally distributed. The residuals are also tested using the Shapiro-Wilk test, producing a p-value of 0.8828, supporting the assumption of normality. The residuals are finally tested with the Runs test, producing a p-value of 0.143, supporting the assumption of independence.

**Exhibit 6: ACF, Histogram, Estimated Density Curve, and Normal Probability Plot for Residuals of Fitted ARIMA Transfer Function Model**

****

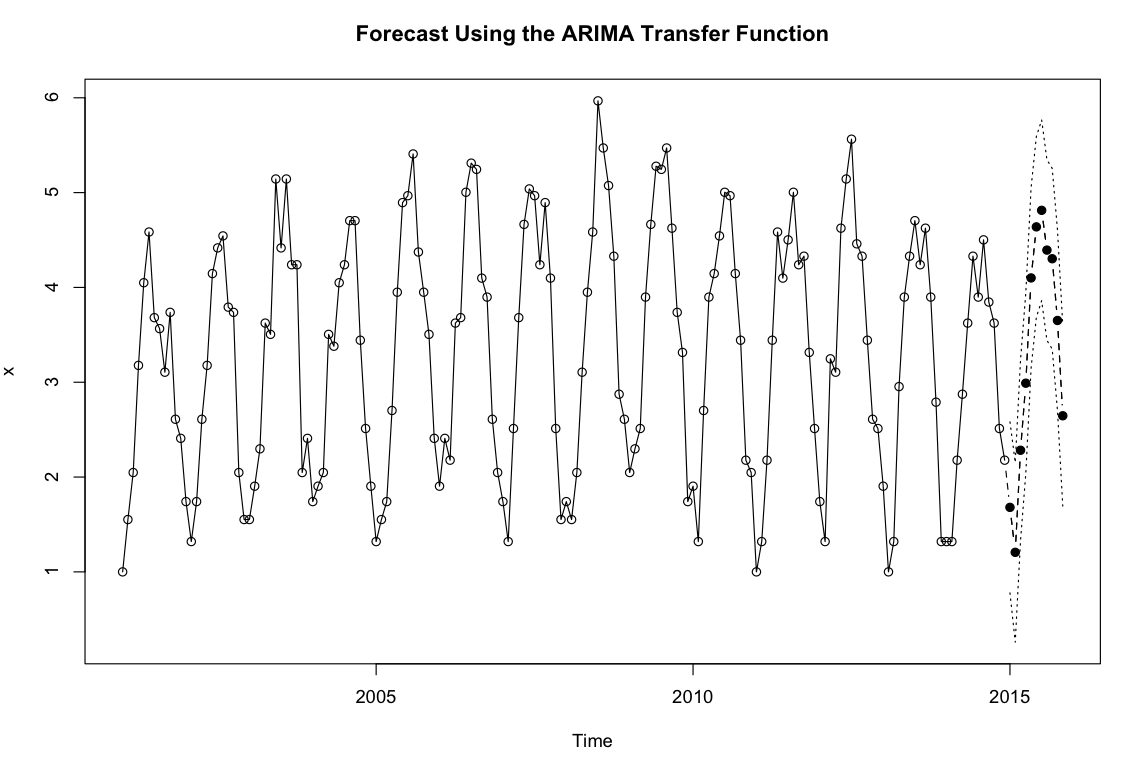
**Forecasting**

The ARIMA (1, 0, 0) x (2, 1, 0)12 with transfer function (b=3, r=1, s=0) model was used to forecast 11 months into the future, January 2015 to November 2015. The forecasts were compared with the actual observations of theft during these months, shown in Exhibit 6. The 11 forecasts differed from the actual observations by an average 12.8%. If the ARIMA (1, 0, 0) x (2, 1, 0)12 was used to model theft without the gas price transfer function, the 11 forecasts differ from the actual observations by an average of 15%, see Appendix Exhibit A.6. Since the % difference between actual and observed is less when the transfer function is included in the model, it appears that the transfer function helps to make better predictions. The forecasts are graphed in Exhibit 7 with 95% confidence intervals.

**Exhibit 6: Forecasts using ARIMA Transfer Function Compared with Actual Values of Theft for 2015**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Jan-15** | **Feb-15** | **Mar-15** | **Apr-15** | **May-15** | **Jun-15** |
| **Forecast** | 1.68 | 1.20 | 2.25 | 2.96 | 4.05 | 4.57 |
| **Actual** | 1.55 | 1.00 | 1.90 | 2.95 | 3.18 | 3.51 |
| **% Difference** | 0.09 | 0.20 | 0.18 | 0.00 | 0.27 | 0.30 |
|  |  |  |  |  |  |  |
|  | **Jul-15** | **Aug-15** | **Sep-15** | **Oct-15** | **Nov-15** |  |
| **Forecast** | 4.73 | 4.31 | 4.22 | 3.56 | 2.55 |  |
| **Actual** | 4.33 | 3.85 | 3.51 | 3.68 | 2.61 |  |
| **% Difference** | 0.09 | 0.12 | 0.20 | -0.03 | -0.02 |  |
|  |  |  |  |  |  |  |

**Exhibit 7: 95% Confidence Forecast Values using ARIMA Transfer Function**



**Discussion**

Despite the predictive improvement, the model still has serious limitations. There is a possibility that the approximate cross correlation at lag -3 was entirely spurious and that there is in fact little to no correlation between gas prices and bike theft. Or, gas prices and bike theft may be related through a third confounding variable. As a result, the model may predict well as long as gas prices behave in a way that is similar to the past. However, if gas price begin to fluctuate in ways that go beyond the status quo, the predictive capabilities of the model may suffer.

Another limitation is that the crime data was filtered for crimes of cycle, scooter, and bike theft. It was assumed for the purpose of this model that cycle and scooter theft could be taken as equivalent to bike theft since cycles and scooters are very similar to bikes, cheap alternative transportation. While cycle and scooter theft likely follow similar trends to bike theft, cycle and scooter theft may respond differently to gas prices than bike theft. As such, the model may be misrepresentative of bike theft predictions.

Gas prices three months in the past appear to have a negative correlation to present bike theft. Since it was expected that increases in gas price would increase bike theft, the negative correlation between gas price and bike theft is an additional reason to doubt the model.

In conclusion, despite its limitations, the ARIMA (1, 0, 0) x (2, 1, 0)12 was better at predicting the 2015 bike theft data when it incorporated gas price as an input series in a (3, 1, 0) transfer function. This type of model, an ARIMA transfer function with gas prices as the input series, has potential to be generalized to other cities. However, the model between gas and bike theft would need to be analyzed city to city as location may affect the relationship.

**Bibliography**

City of Chicago. 2015. Crimes – 2001 to Present. <https://data.cityofchicago.org/Public-Safety/Crimes-

2001-to-present/ijzp-q8t2>.

Cryer, Jonathan D., and Kung-Sik Chan. *Time Series Analysis With Applications in R*. Second ed. New York:

Springer, 2008. N. pag. Print.

Jaffe, Eric. "These 8 Depressing Bike Statistics Show Just How Bad the Problem Is." *CityLab*. The Atlantic,

16 Apr. 2014. Web. 13 Dec. 2015. <http://www.citylab.com/commute/2014/04/these-8

depressing-bike-theft-statistics-show-just-how-bad-problem/8890/>.

*The District Downtown Columbia*. N.p., n.d. Web. 13 Dec. 2015.

<http://discoverthedistrict.com/transit/biking/>.

U.S. Energy Information Administration. 2016. *Weekly Chicago, IL Regular Reformulated Retail Gasoline*

*Prices*. <https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=EMM\_EPMRR\_PTE\_YORD\_

DPG&f=W>

Van Lierop, D., Grimsrud, M., & El-Geneidy, A. (2015). Breaking into bicycle theft: Insights from

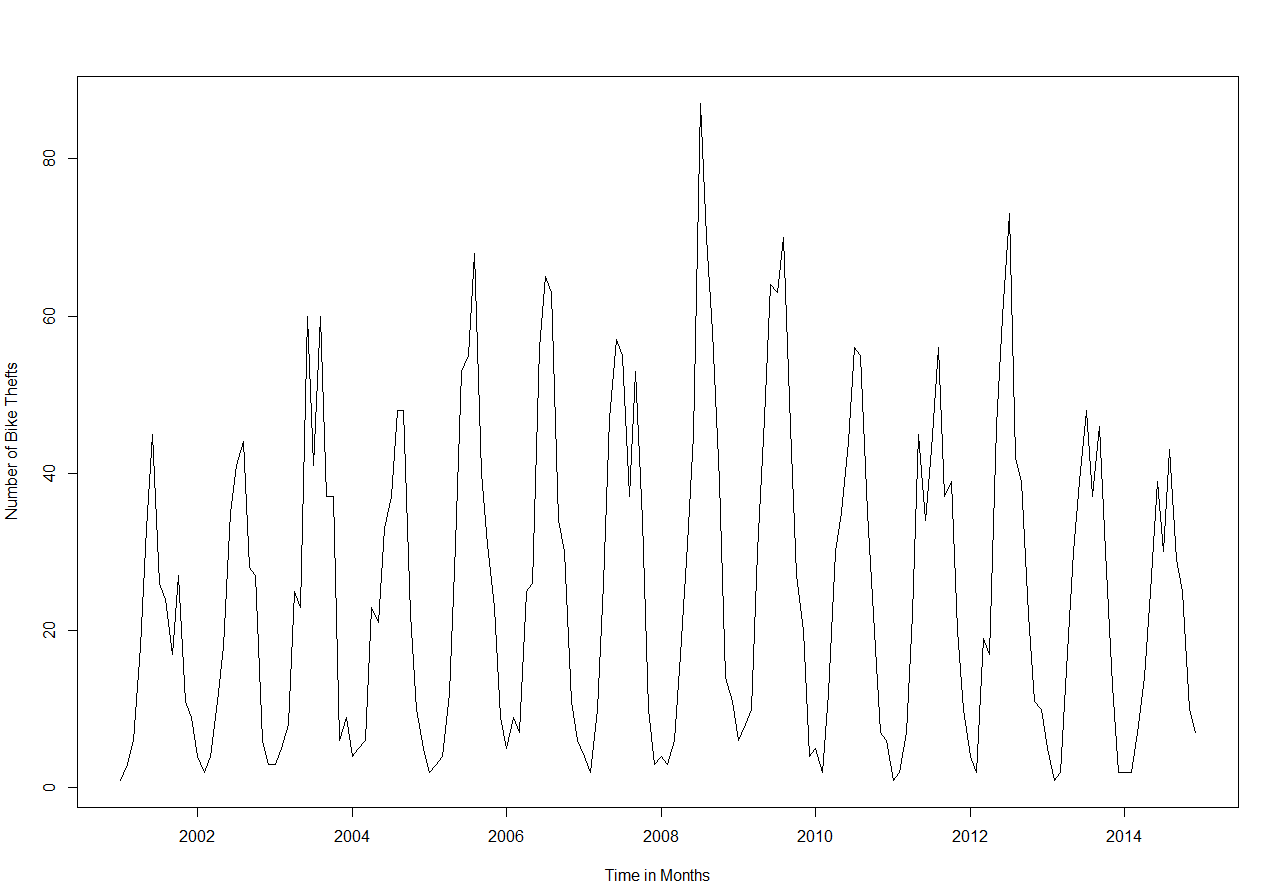
Montreal, Canada. International Journal of Sustainable Transportation, 9(7), 490-501.

Wei, William W. *Time Series Analysis Univariate and Multivariate Methods*. Second ed. Boston: Pearson

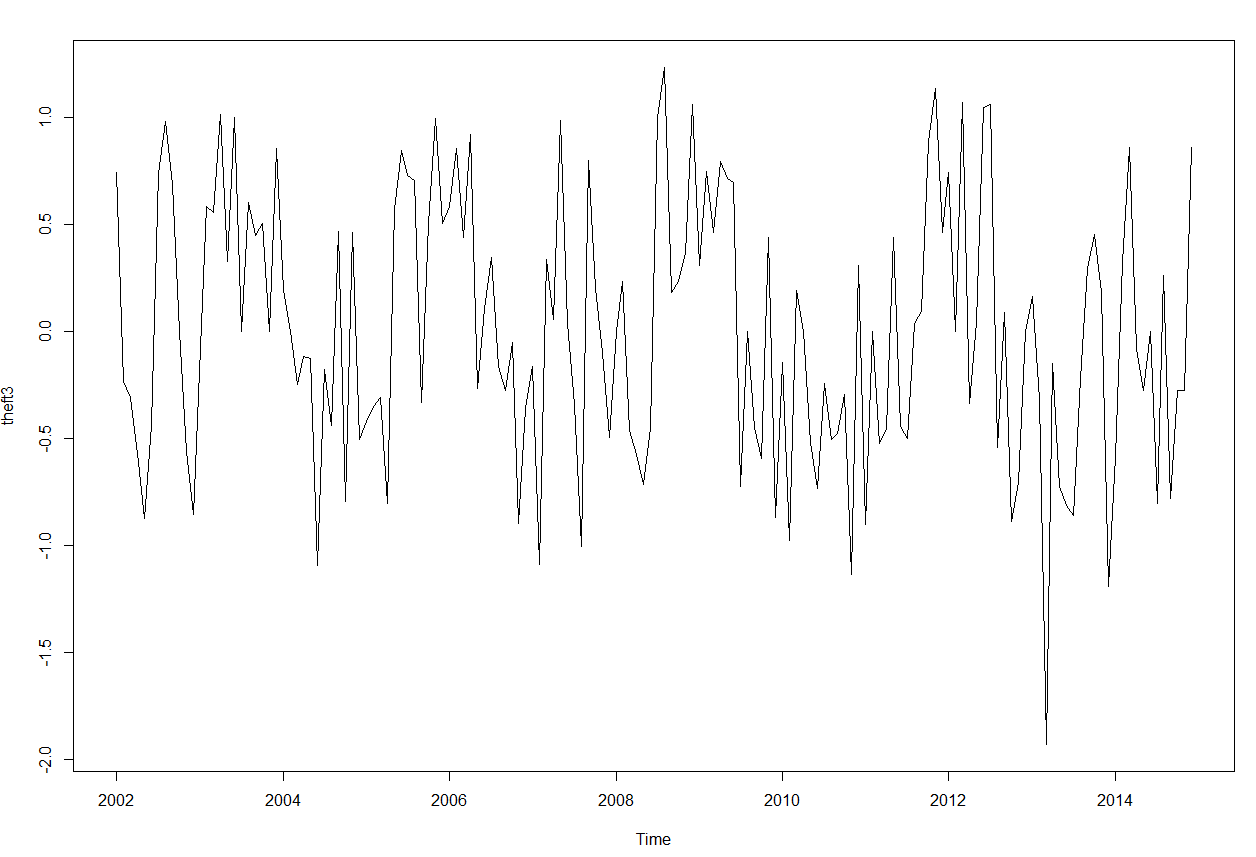
Education Inc., 2006. Print.

**Appendix A**

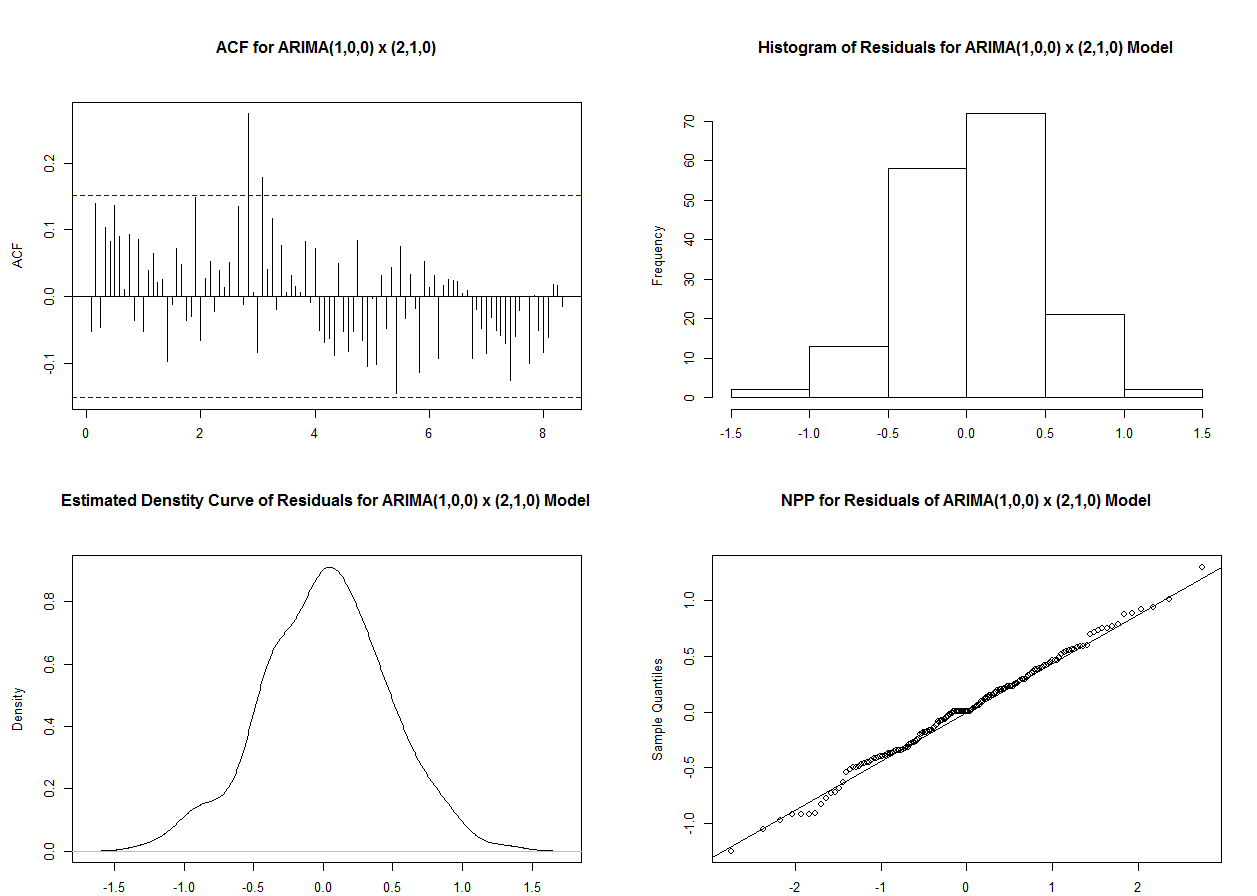
**Exhibit A.1: Time Series Plot of Transformed Theft Data**



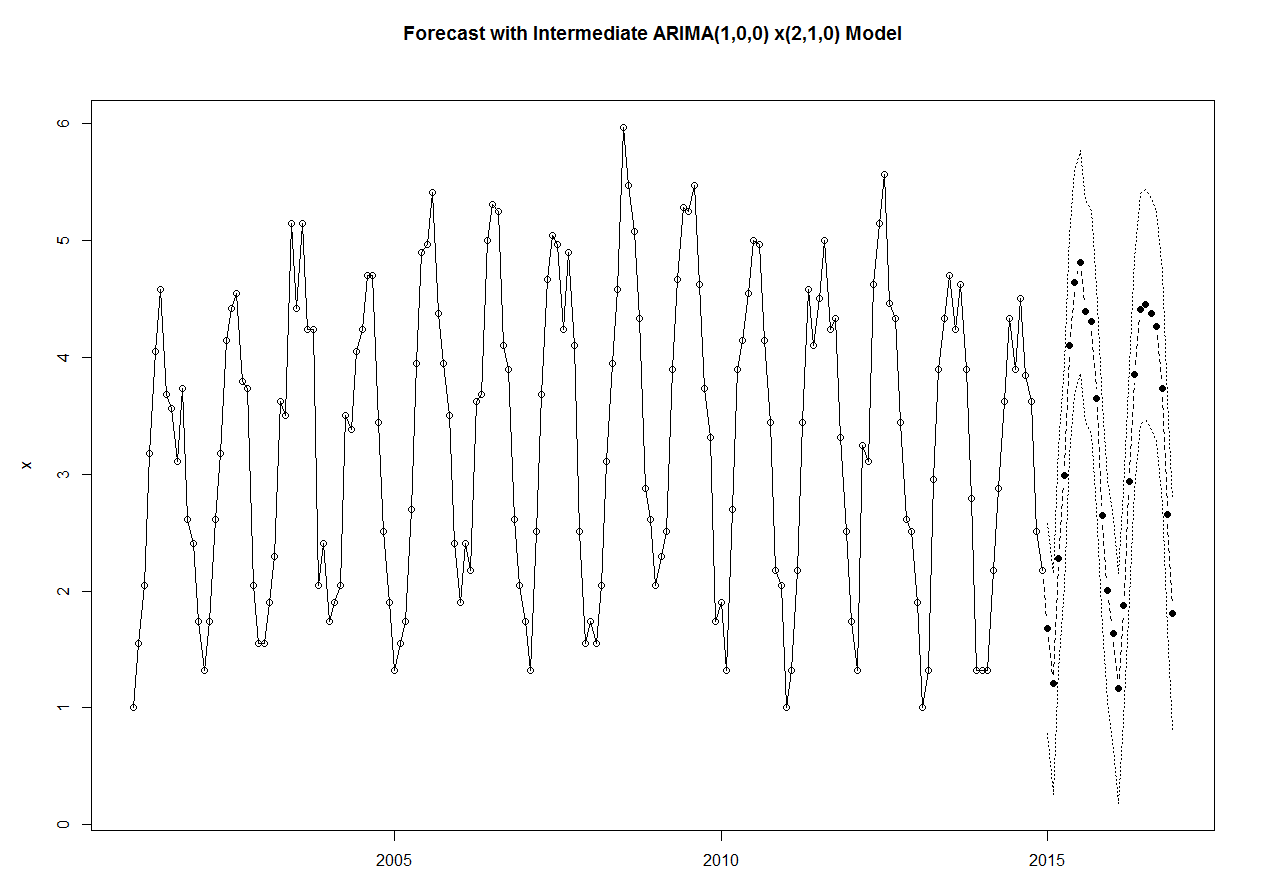
**Exhibit A.2: Times Series Plot of Theft after First Difference and Seasonal First Difference**



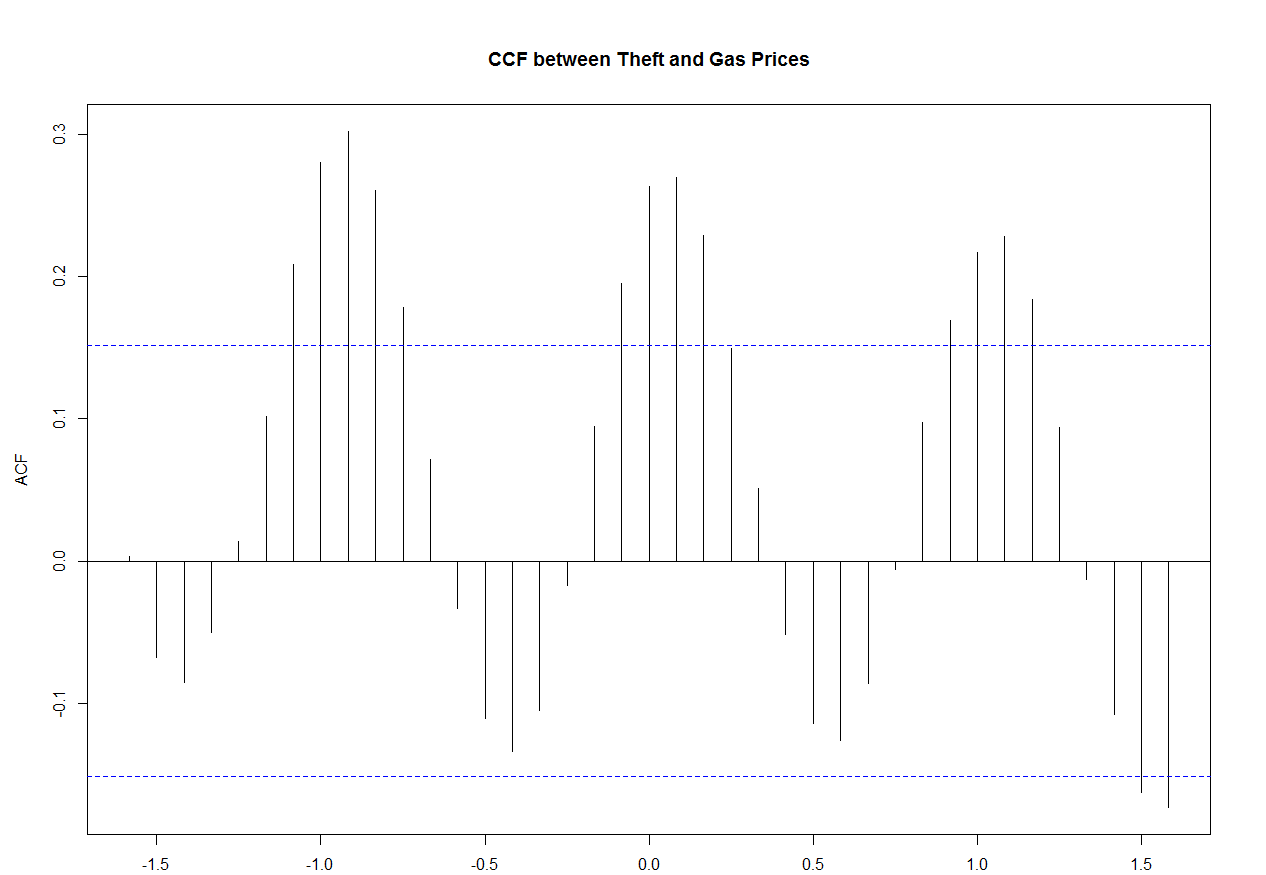
**Exhibit A.3: Diagnostic Checks for Intermediate ARIMA (1, 0, 0) x (2, 1, 0)12 Model**



**Exhibit A.4: Forecast with Intermediate ARIMA (1, 0, 0) x (2, 1, 0)12 Model**



**Exhibit A.5: CCF between Theft and Gas Prices – Spurious Correlation**



**Exhibit A.6: Forecasts Using ARIMA Model without Gas Price Transfer Function**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Jan-15** | **Feb-15** | **Mar-15** | **Apr-15** | **May-15** | **Jun-15** |
| **Forecast** | 1.68 | 1.21 | 2.28 | 2.99 | 4.10 | 4.64 |
| **Actual** | 1.55 | 1.00 | 1.90 | 2.95 | 3.18 | 3.51 |
| **% Difference** | 0.08 | 0.21 | 0.20 | 0.01 | 0.29 | 0.32 |
|  |  |  |  |  |  |  |
|  | **Jul-15** | **Aug-15** | **Sep-15** | **Oct-15** | **Nov-15** |  |
| **Forecast** | 4.81 | 4.39 | 4.30 | 3.65 | 2.65 |  |
| **Actual** | 4.33 | 3.85 | 3.51 | 3.68 | 2.61 |  |
| **% Difference** | 0.11 | 0.14 | 0.23 | -0.01 | 0.01 |  |

**Appendix B**

The following formulas are taken from Time Series Analysis: Univariate and Multivariate Methods, by William W.S. Wei.

, , and is a delay parameter

Transfer Function is therefore given as