U.S. – North American Trade and Freight Import Transportation by Land mode Highlights

Abstract. This research is a study model of analyzing 1995-2011 U.S. – North American Trade and Freight Import Transportation by Land mode. The objective of the research is to find an appropriate SARIMA model for analyzing and forecasting U.S. – North American Trade and Freight Import by land mode. SARIMA (2,1,0,1,1,1,12) was chosen to fit data year 1995-2011 and been used for one year ahead forecasting.

Key words and phase: SARIMA Model, Box-Ljung test, U.S. – North American Trade and Freight (NAFTA), AIC, BIC, forecasting, model estimating, model specification, model diagnostics.

1. Introduction

Nowadays, the economic activities of US and whole world are becoming more integrated and globalized, more goods produced by U.S. factories and farms are bound for export, and imports from more than 200 countries. Moving large quantities of goods across the county and around the world, US depend on the Nation's freight transportation system- a vast network of roads, rail tracks, airport, bridges, seaports, pipelines, and equipment. People are really enjoying buying fresh fruits and vegetables in mid-winter, fast and reliable next-day deliveries of Internet purchases, and buying electronic appliances manufactured thousands of miles away. According to research of U.S. Department of Transportation Bureau of Transportation statistics, US-North American trade amounted to nearly one-third of total U.S. – international merchandise trade, which reached an historic high of $2.3 trillion in 2004. Moreover, land modes (truck, rail, and pipeline) moved freight shipments worth $634 billion with Canada and Mexico which comprising 89 percent of our trade with these two counties. Canada has been number one trading partner for several decades, and Mexico became number two when it surpassed Japan in 1999.

Meanwhile, although US department of Transportation Bureau of Transportation statistics has published papers on the all details of analysis of NAFTA data which include modal perspective, modal roles by value and weight, land freight by state and port and ect by basic comparison and trend analysis, seldom research paper was created to analysis the relationship of the historic data and fit a time series ARMA model to descript the NAFTA freight transportation data, and furthermore provide forecasting data to provide better service of department of transportation and end user or client of the goods.

This paper includes three sections which are theoretical proven of SARIMA model (section2), a brief introduction of the data used in this research (section3), analysis of the model and results of this specific research data (section4), and discussion(section5). Also fours steps of the research also have been adapted which including Model specification, Parameter estimation, Model diagnostics, and Forecasting.

2. Theoretical proven
   1) The first step of any data before adapting any ARMA model is to plot ACF and PACF of the data. ACF is a function to descript the autocorrelation of time series data with lags information, PACF is a function to descript the partial autocorrelation between Xs and Xt with the linear effect of everything “in the middle” removed. Most of important is to use ACF/PACF to identify p, q and P, Q in the SARIMA model. Two tables from text book can be a good explanation of the function of ACF/PACF.

Table 3.1. Behavior of the ACF and PACF for ARMA Models

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Table 3.3. Behavior of the ACF and PACF for Pure SARMA Models

	$AR(P)_s$	$MA(Q)_s$	$ARMA(P,Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots$, are zero.

2) In order choosing better model, several criteria have to been used in this research.
   a) Box-Ljung test

   Box-Ljung test is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero.

   The Ljung–Box test test can be defined as follows.

   H_0: The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).
   H_a: The data are not independently distributed.

   The test statistic is:

$$Q = n\,(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

   where *n* is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag *k*, and *h* is the number of lags being tested. For significance level α, the critical region for rejection of the hypothesis of randomness is

$$Q > \chi^2_{1-\alpha,h}$$

   where $\chi^2_{1-\alpha,h}$ is the α-quantile of the chi-squared distribution with *h* degrees of freedom.

   b) AIC and BIC

$$AIC = 2k - 2\ln(L)$$

where *k* is the number of parameters in the statistical model, and *L* is the maximized value of the likelihood function for the estimated model.
Given a set of candidate models for the data, *the preferred model is the one with the minimum AIC value.* Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting (increasing the number of free parameters in the model improves the goodness of the fit, regardless of the number of free parameters in the data-generating process).

$$\mathrm{BIC} = n \cdot \ln(\hat{\sigma_e^2}) + k \cdot \ln(n)$$

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC.

c) Forecasting

Fitted SARIMA model will be applied for forecasting purpose.

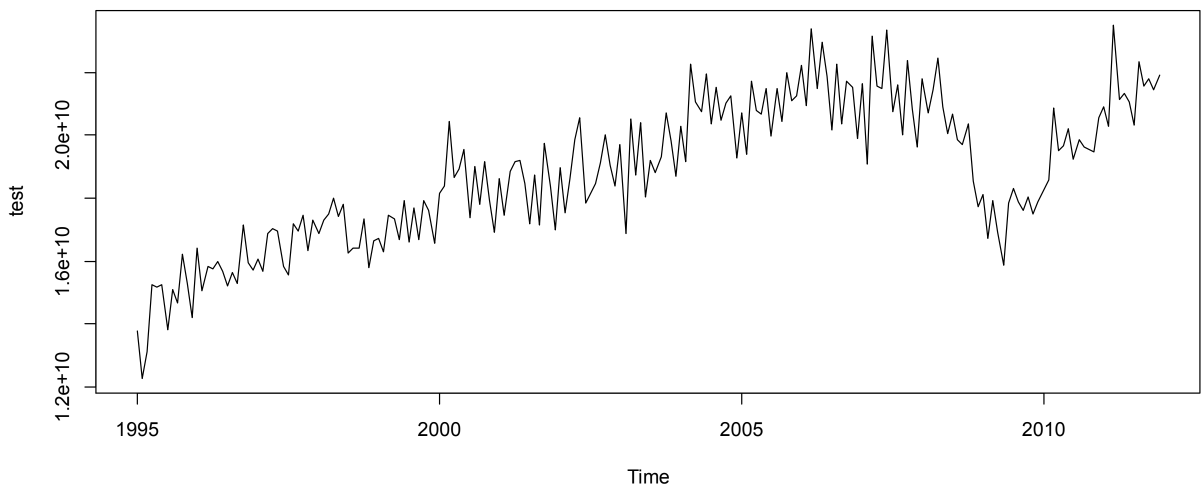
3. Introduction of sample data set

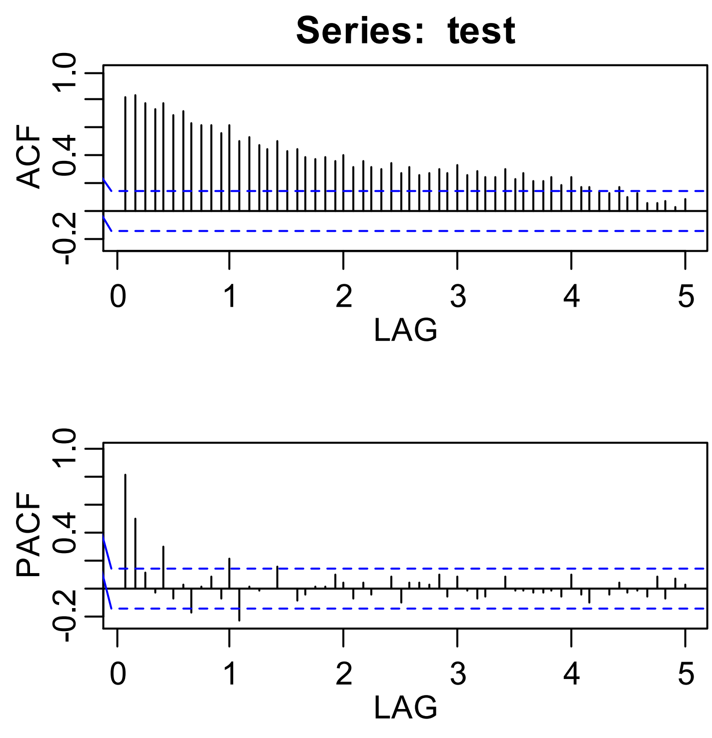
In order to analyze SARIMA model base on sufficient data, only land mode import data by weight were used in this research which including 1995-2011 with all 12 months detail data. Export data were abandoned due to the small size of data which only including data from 1995-2004. The reason only weight unit was chosen in this research is because that more complicate current and inflation-adjusted economic model have to been adapted if value unit will be chosen for the research. In the other hand, most physical goods are valued by weights in the nowadays technology level except computer instrument. However, more weight of computer instrument base on less weight technology also proved the increased amount of demand in US market. In other words, the sample data was chosen base on thorough design for time series model fitting purpose.

4. Analysis of the model and results

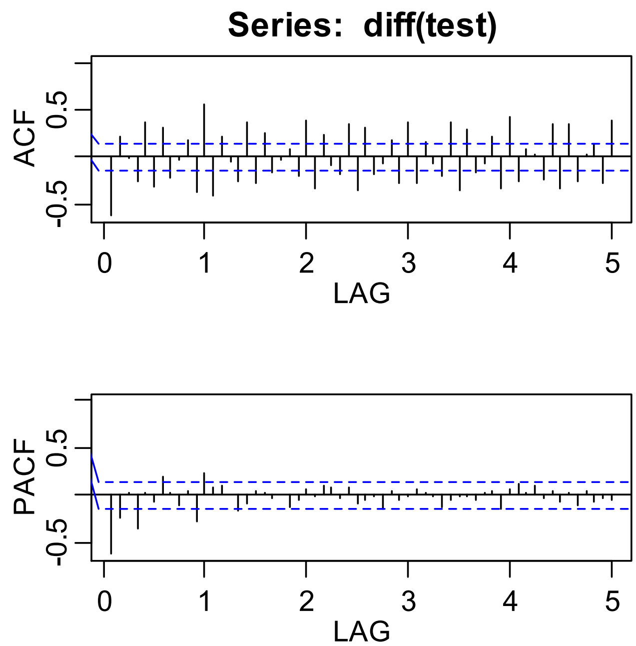
a) Model specification

A time series plot of US-NAFATA import monthly data from 1995-2011 by land mode.

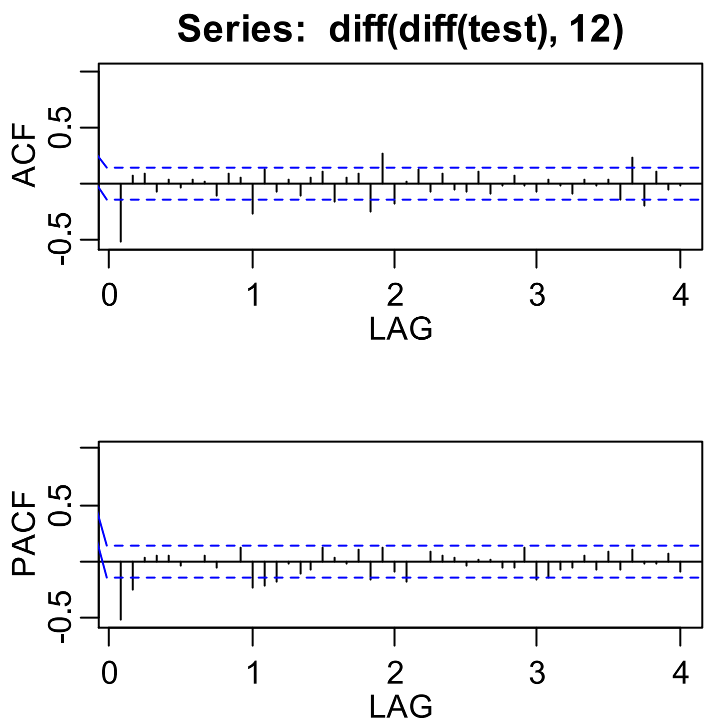




From the table, we can easily find out the trend in the time series data. Also from ACF and PACF, we can notice the slow decay in ACF and the fact that PACF at the first lag is nearly 1 which indicates nonstationary behavior. So next step, we adapted the first difference and plot ACF and PACF of it.



We can note the peaks at seasonal lags at 1s, 2s, 3s, 4s, and 5s which corresponding five 12 months. This indicates a seasonal difference. Base on the data structure of US NAFATA, this shows us 12 month seasonal lags. Plot out ACF and PACF of seasonal difference.



From the plot of ACF and PACF, we can clearly see both ACF and PACF are both tailing off in the seasonal lags with no spikes in both plot. This clearly indicate an SARMA of orders P=1 and Q=1 base on table 3.3. The next step is to assess ACF and PACF at the within lags from 1 to 11, it appears that either 1) both of ACF and PACF are tailing off; 2) PACF cuts off at lag 2. We can easily get two candidate models base on table 3.1 1) p=1, q=1 or 2) p=2, q=0. Until now, we successfully dig out two models combination from ACF and PACF analysis.

b) Model fitting and Model diagnostics

After we go through the ACF and PACF plot analysis, we can list the candidate models out:

1) testfit1 = SARIMA(test, 1,1,1,1,1,1,12, no.constant=TRUE)
2) testfit2 = SARIMA(test, 2,1,0,1,1,1,12, no.constant=TRUE)

The first thing we need to check is the Box-Ljung test which can give us a very important p value of the hypothesis at lag 48.

	Model testfit1	Model testfit2
Box-Ljung test pvalue	0.478	0.797

From the output of the Box-Ljung test and its p value, we can easily find out that the second model has better fitting properties. From the AIC and BIC output value, we also can prove our decision assumption of choosing model 2. So we can write out the model

The second step we need to diagnostic is the AIC and BIC of these two models.

	Model testfit1	Model testfit2
AIC	42.03513	42.0289
BIC	41.10019	41.09396

testfit2

Coefficients:

```
         ar1      ar2     sar1      sma1
     -0.6694  -0.2694  0.2008  -0.8994
s.e.  0.0708   0.0699  0.0928   0.0877
```

sigma^2 estimated as 6.333e+17: log likelihood = -4193.41, aic = 8396.81

$AIC

[1] 42.0289

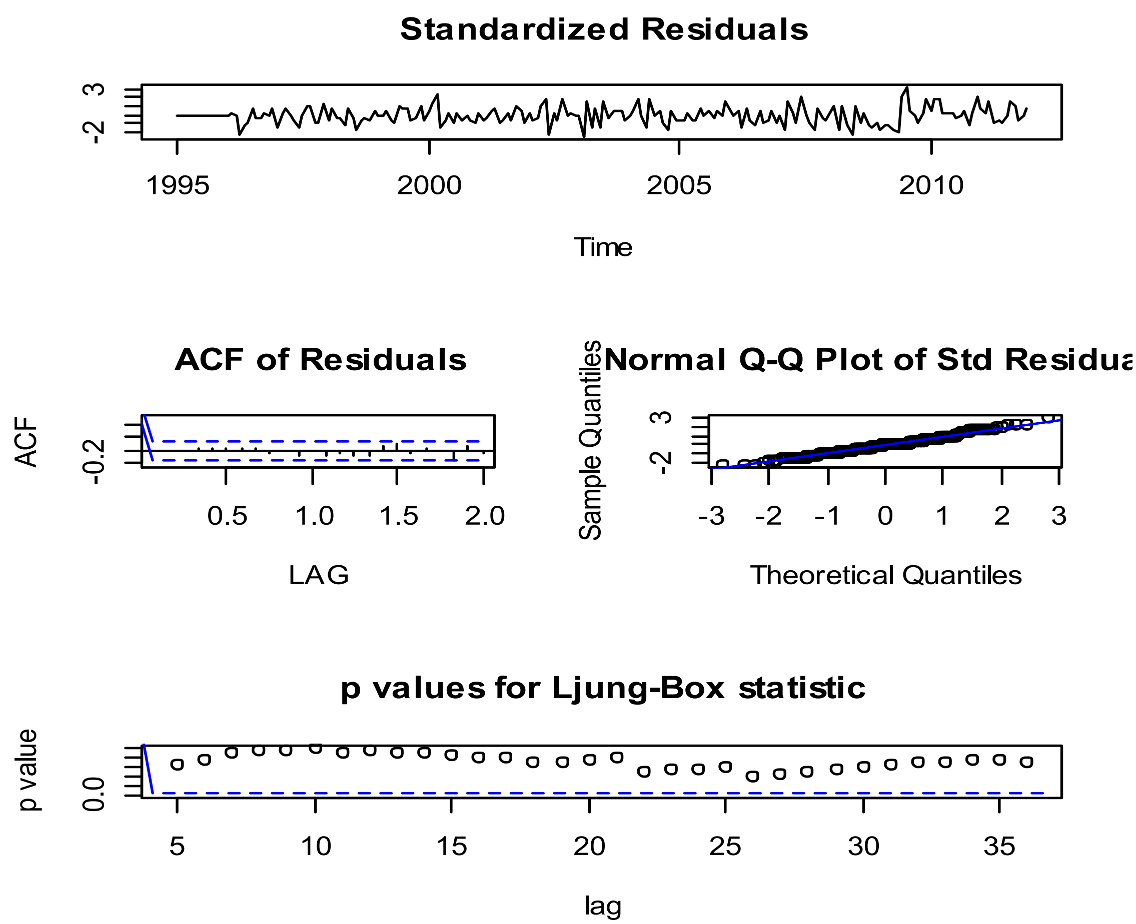
$AICc

[1] 42.04019

$BIC

[1] 41.09396

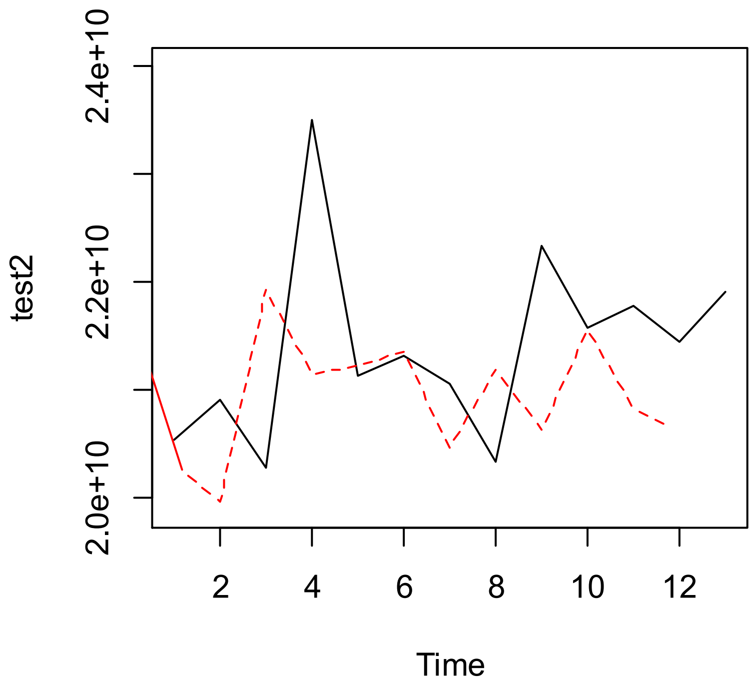
After we choose the second model to fit our data, we can print out the diagnostics information out at below plot. It shows few outliers in the series in the first plot of standardized residuals and also few outliers in Q-Q plot which is acceptable and reasonable for a fitted model. And only one ACF of residuals close to -0.2 and others reminds inside -0.2 to 0.2 which indicates good fitting of the source data.



c) Forecasting

The last step of this research is model forecasting. Forecasting function of SARIMA model will be used here. In the data analysis stage, last 12 months data were withheld for the purpose of comparison with forecasted last 12 months data. A plot of the real data and forecasted data is listed below. The black solid line is real data, red dashed line is the forecasted data.

```
test.pr=sarima.for(test, n.ahead=12, 2,1,0,1,1,1,12)
```



As shown in the plot is a year forecasting base on fitted model. This can help US department of transpiration has a better idea of the future trend of cross border freight status.

5. Discussion

In this research paper, SARIMA model was introduced and explored in very details. After model diagnostics, SARIMA(2,1,0,1,1,1,12) was used for the US-NAFTA 1995-2011 land mode import data. As discussed in the introduction of data section, only import data was used in this research paper due lack of export data after before 2004. However, we still could explore the export data and combined import and export data in future. This will provide more comprehensive statistics data analysis of US-NAFTA transportation data and information. As result of this research paper, the data shows a pretty normal seasonal pattern and increasing trend from 1995 to 2011. However, there is a obvious decline in year 2008 to 2009 due to US economic. This also proves that US-NAFTA transportation data precisely reflect US economic trend. This information can be useful in two ways. First, this is very helpful and can be a directory to US department of transportation to manage all the limit resource to facilitate US economic climbing. Also this can be useful for other US government departments as a decision making factor for their purpose.

Reference

[1] U.S. Department of Transportation, Research and Innovative Technology Administration, Bureau of Transportation Statistics, Freight Transportation: Global Highlights, 2010 (Washington, DC: 2010)

[2] U.S. Department of Transportation, Research and Innovative Technology Administration, Bureau of Transportation Statistics, Freight Shipments in America, 2004 (Washington, DC: 2004)

[3] U.S. Department of Transportation (USDOT), Research and Innovative Technology Administration (RITA), Bureau of Transportation Statistics (BTS). 1996. *Transportation Statistics Annual Report*. Washington, DC.

Appendices

Code used in this research paper:

```
test = ts(scan("mydata.txt"), start=1995, frequency=12)

plot(test)

acf2(test, 60)

acf2(diff(test), 60)

acf2((diff(test), 12), 60)

testfit1<-sarima(test,1,1,1,1,1,1,12,no.constant=TRUE)

testfit1

testfit1$resid=innov

Box.test (testfit1$resid, lag = 48,type="Ljung")

testfit2<-sarima(test, 2,1,0,1,1,1,12, no.constant=TRUE)

testfit2

testfit2$resid=innov

Box.test(testfit2$resid, lag=48, type="Ljung")

test.pr=sarima.for(test, n.ahead=12, 2,1,0,1,1,1,12)

test.pr

test2=test[192:204]

test2

test3=test1.pr$pred[1:12]

test3

ts.plot(test2, ylim=c(1.99e+10, 2.4e+10))

lines(test3, col="red", lty="dashed")
```