Time Series Analysis on US Border Crossing Entry
Written Report

George Washington University
Department of Decision Science

Tian Tian G25258451
Xiang Fan G29708057
Hsin-Yu Chen G34247940
Huang-Chin Yen G33998025
Qi Gui G47997189

DNSC 6219 | Spring 2020
Apr 26th, 2020

Outline

1. Introduction and Overview

   1.1 Expectation

   1.2 Data Cleaning

2. Univariate Time-series models.

   2.1 Deterministic Time Series Models (Seasonal Dummies and Trend, Cyclical Trend) and Error model.

   2.2 ARIMA models (with seasonal ARIMA components if relevant)

   2.3 Comparison of models (in terms of fit and validation)

3. Multivariate Time Series Models

4. Conclusion

1. Introduction and Overview

The United States is located between Canada and Mexico, this geographic location generates tons of border crossing entry data on a daily basis. Especially Mexico, each year, we can see various news about the United States-Mexico border frictions. Hence, we would like to apply the knowledge of times series models to this project in order to:

1.1 Expectation

- See if there is any trend and seasonality of the United States-Mexico border entries.
- Apply deterministic and stochastic models to the data to see how each model's performance is when predicting the hold-out samples.
- Find possible relationships between the United States-Canada and the United States-Mexico border entries by using the Transfer Function model.

1.2 Data Cleaning

The dataset, Border Crossing Entry Data, is found from Kaggle. The dataset was firstly recorded in September 1996 by ports of entry by U.S. Customs and Border Protection (CBP). Currently, it contains 349,000 rows with 7 Columns (Port Name, State, Port Code, Border, Date, Measure, Value, Location). To better estimate the time series, we took the following steps:

- Split United States-Canada Border and United States-Mexico Border
- The variable combination for grouping is Border, Year, Month, and Value. As for the variable Value, it is the sum of values in a month

After grouping, our datasets are monthly basis and there are 279 rows for each border. We decided our hold-out sample size to be 30 observations, starting from October 2016.

2. Univariate Time-series models

2.1 Deterministic Time Series Models

Seasonal Dummies and Trend model.

Seasonal Dummies and Trend Model would help us identify if there is seasonal trend in our data. Firstly, we apply boxplot to observe the seasonal trend.

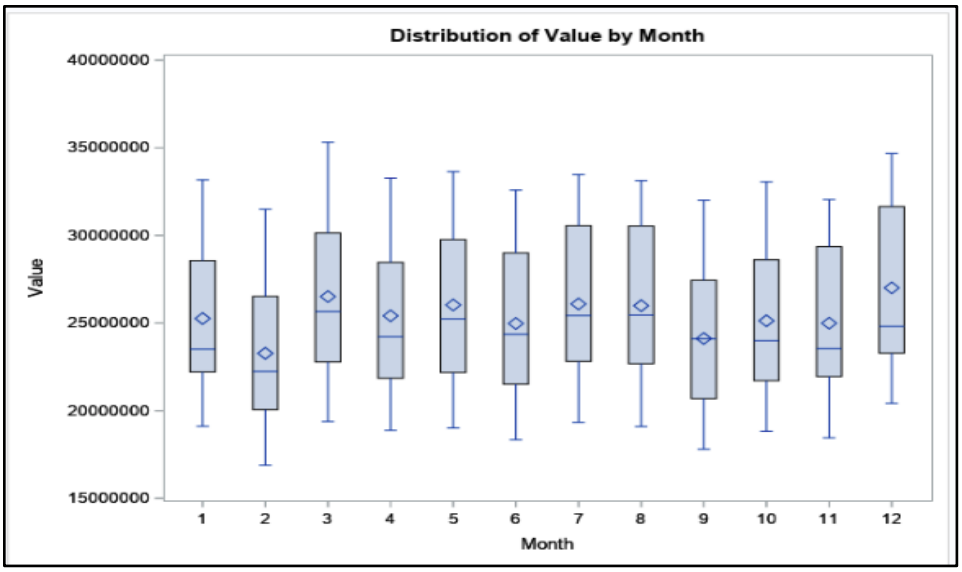


Figure2.1.1

From Figure 2.1.1, we can see since the means are different from month to month and the variances are also different, we can find that there is evidence of seasonality in this series.

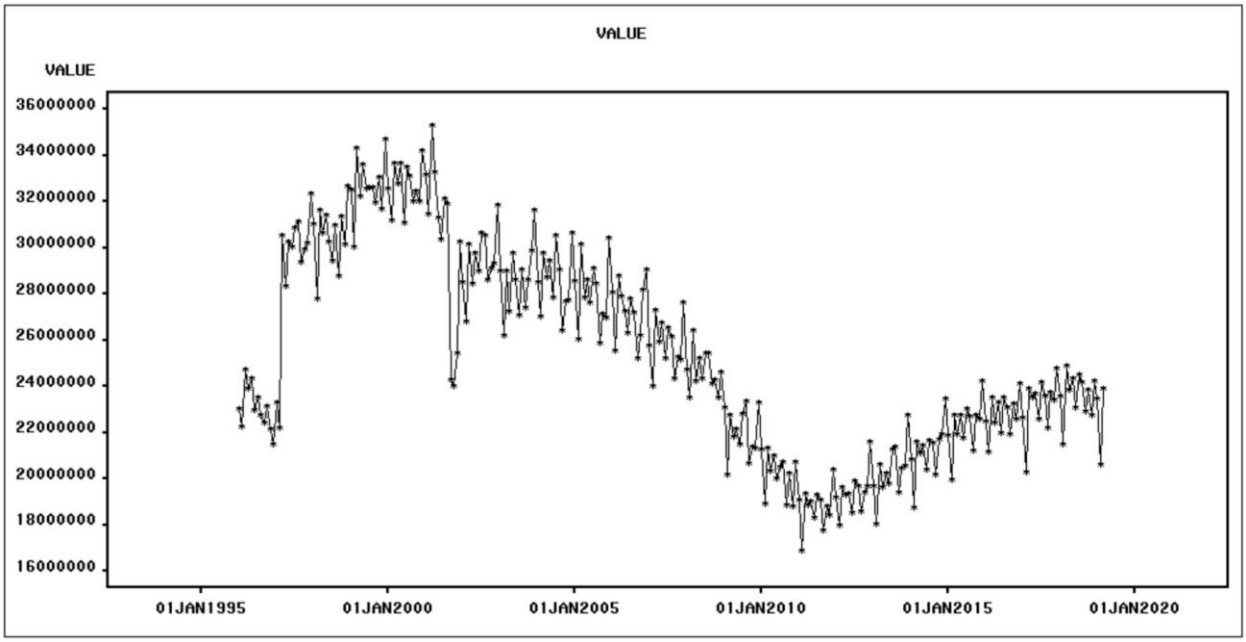


Figure 2.1.2

From Figure 2.1.2, the overall plot of US-Mexico Border, we can also observe changing means, again proving that there is a seasonality in this series.

Then we obtain the parameter estimate table to see if there are significant seasonal dummies.

Parameter Estimates

VALUE

Seasonal Dummies + Linear Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	10758536	207786	51.7771	<.0001
Seasonal Dummy 1	-811864	258682	-3.1385	0.0019
Seasonal Dummy 2	-977202	258674	-3.7777	0.0002
Seasonal Dummy 3	432675	258668	1.6727	0.0956
Seasonal Dummy 4	407397	261451	1.5582	0.1204
Seasonal Dummy 5	1647026	261439	6.2999	<.0001
Seasonal Dummy 6	2237356	261428	8.5582	<.0001
Seasonal Dummy 7	4457605	261419	17.0516	<.0001
Seasonal Dummy 8	4916183	261411	18.8063	<.0001
Seasonal Dummy 9	1842992	261406	7.0503	<.0001
Seasonal Dummy 10	1220022	261401	4.6672	<.0001
Seasonal Dummy 11	172169	261399	0.6586	0.5107
Linear Trend	-17149	659.1887	-26.0150	<.0001
Model Variance (sigma squared)	7.85783E11	.	.	.

Figure 2.1.3

We can see from figure 2.1.3, seasonal Dummy1,2,5,6,7,8,9,11 are significant, while dummy 3,4,11 are not significant, which means they are not different from December.

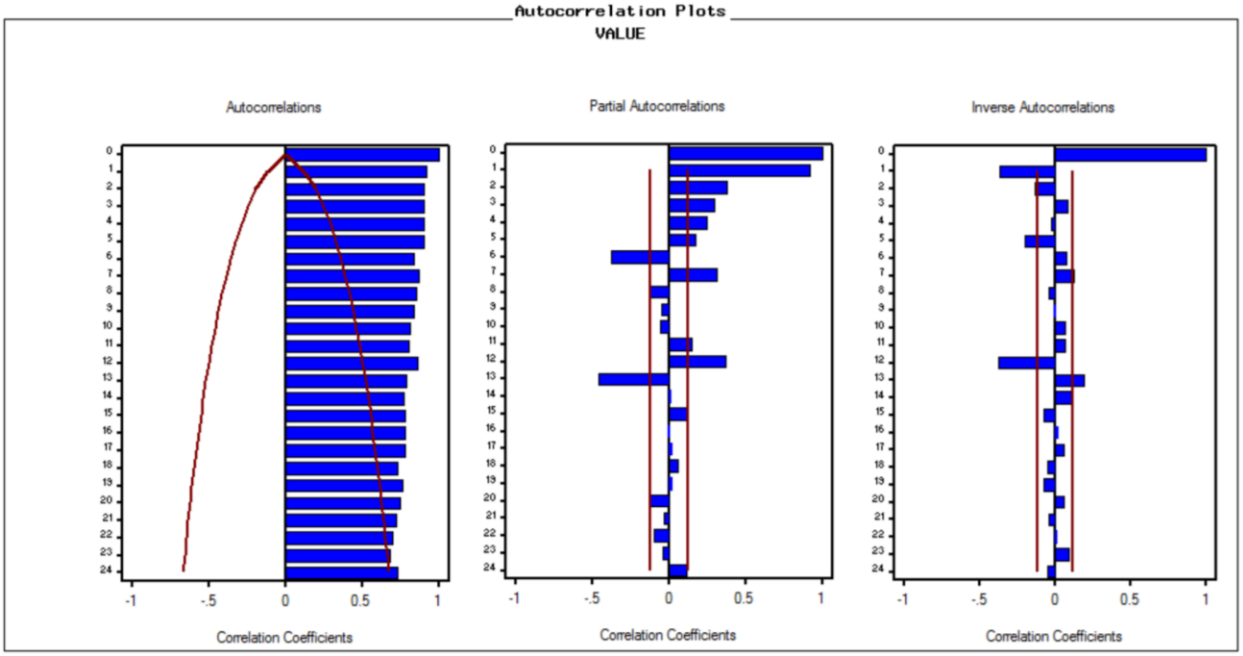


Figure2.1.4

From Figure2.1.4, The ACF decays very slowly. The PACF does not entirely chop off after 10 lags. Both graphs suggest that the time series is not stationary and not white noise. More transformations should be done for further analysis.

VALUE
Seasonal Dummies + Linear Trend

Statistic of Fit	Value
Mean Square Error	1.82252E13
Root Mean Square Error	4269098.0
Mean Absolute Percent Error	18.06757
Mean Absolute Error	4236727.8

Figure2.1.5

From the Figure2.1.5, the model with seasonal dummies and linear trend fitting Mexico data generates very high RMSE (4269098, higher than the Canada data generated model) and relatively low MAPE (18.06757, higher than the Canada data generated model).

The ACF plot and the test accuracy table both suggest the seasonal dummies and trend table is not appropriate, so we further explore the error model.

Seasonal Dummies with Error Model

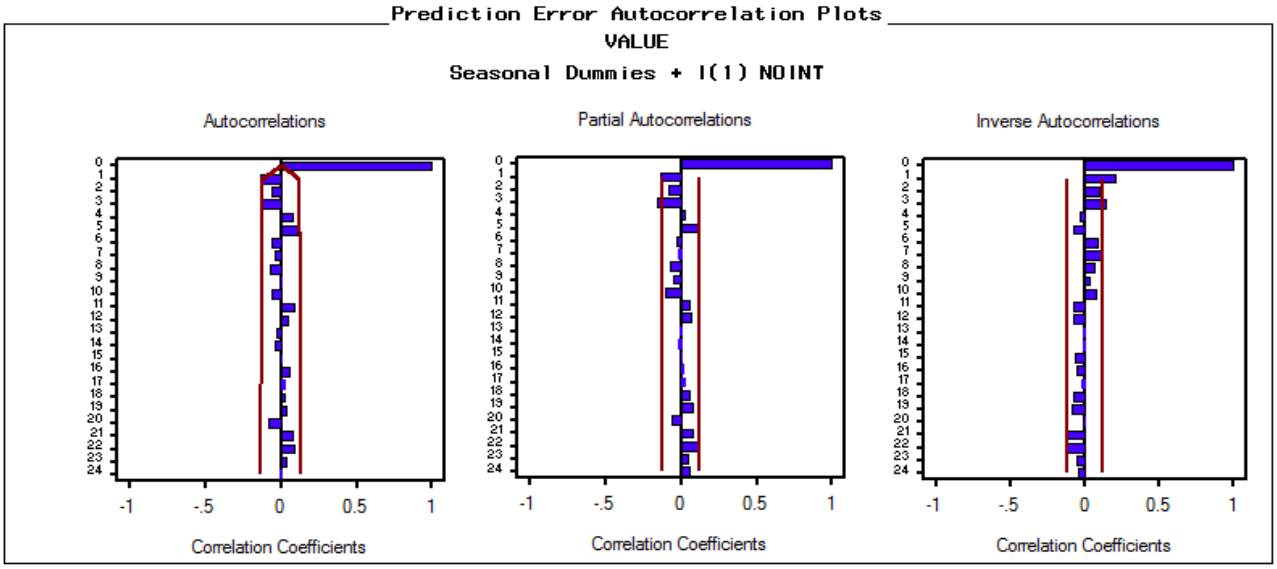


Figure2.1.6

After taking the first difference as well as adding seasonal dummies to the model, we can see that the ACF decays quickly, which we can say the difference of the series is stationary. And also since the ACF chopped off after lag1 and the IACF decays exponentially, so we apply the first order Moving Average process, MA(1) process.

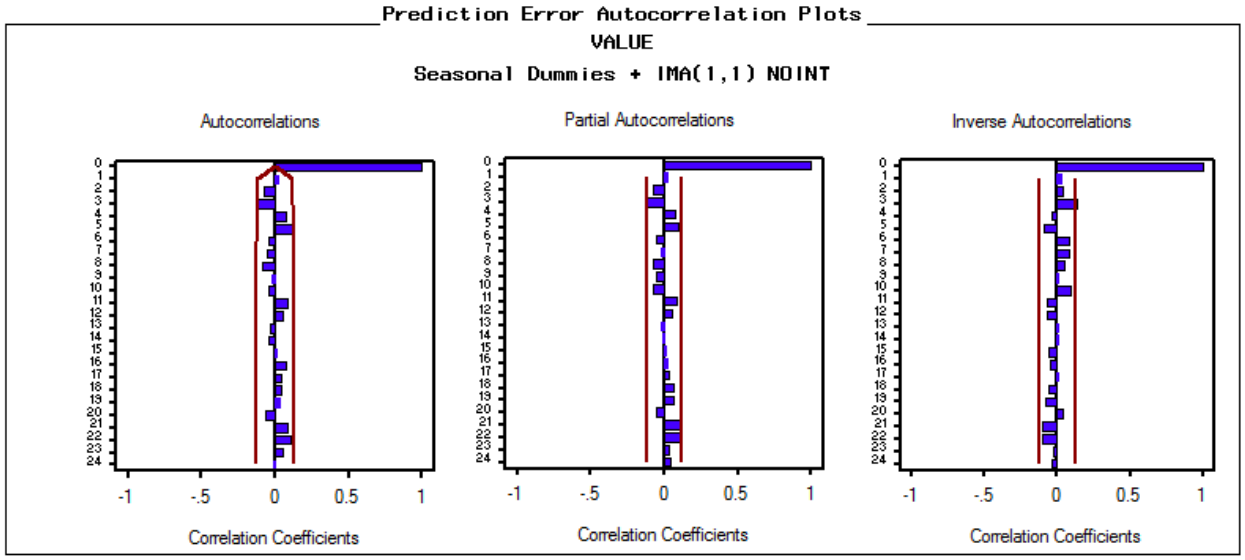


Figure2.1.7

After applying the first order Moving Average process(Figure2.1.7), we can see that the ACF at all non-zero lags are now inside the two standard error bounds.

Parameter Estimates

VALUE

Seasonal Dummies + IMA(1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.17048	0.0605	2.8156	0.0052
Seasonal Dummy 1	-1664374	189578	-8.7794	<.0001
Seasonal Dummy 2	-1984537	185696	-10.6870	<.0001
Seasonal Dummy 3	3239519	185696	17.4453	<.0001
Seasonal Dummy 4	-1213544	189579	-6.4013	<.0001
Seasonal Dummy 5	620230	189690	3.2697	0.0012
Seasonal Dummy 6	-1056064	189690	-5.5673	<.0001
Seasonal Dummy 7	1112921	189690	5.8671	<.0001
Seasonal Dummy 8	-97240	189690	-0.5126	0.6086
Seasonal Dummy 9	-1872423	189690	-9.8710	<.0001
Seasonal Dummy 10	1010578	189690	5.3275	<.0001
Seasonal Dummy 11	-138924	189690	-0.7324	0.4646
Seasonal Dummy 12	2018371	189690	10.6404	<.0001
Model Variance (sigma squared)	8.04301E11	.	.	.

Figure2.1.8

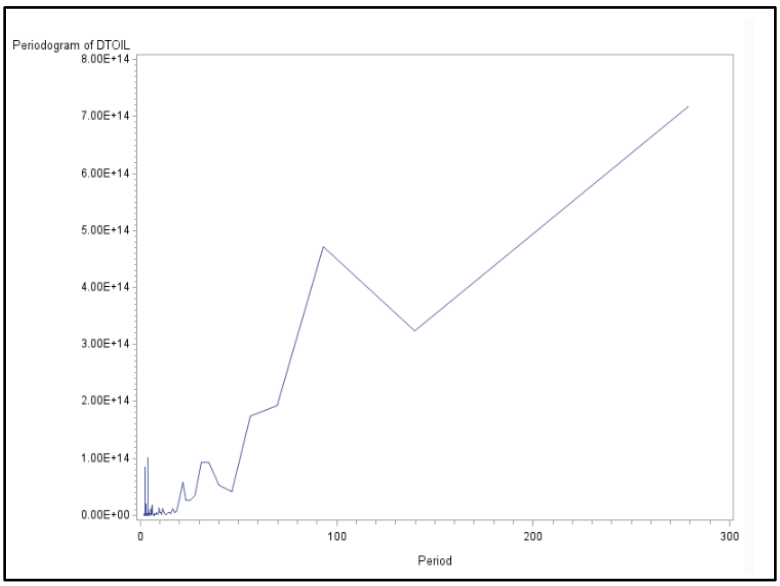
Statistic of Fit	Value
Mean Square Error	7.6683E11
Root Mean Square Error	875688.0
Mean Absolute Percent Error	2.17498
Mean Absolute Error	556454.6

Figure2.1.9

The MA(1) model with seasonal dummies has a Mean Absolute Percent Error of 2.17 and Model Variance is 804301E11.

Cyclical Model

We use the Cyclical model to find cyclical patterns in the Mexico border. Firstly, we obtain the periodogram to discover hidden dominant periods in data(FIgure 2.3.1). The table (Figure 2.3.2)shows the top 20 harmonics. We can see the dominant periods are not related to the monthly trend.



Obs	FREQ	PERIOD	P_01/1,000,000,000
2	0.02252	279	717629
4	0.06756	93	471813
3	0.04504	139.5	323413
5	0.09008	69.75	191897
6	0.1126	55.8	174878
71	1.57643	3.986	100997
10	0.20268	31	93477
9	0.18016	34.875	91945
117	2.61236	2.405	84380
14	0.29276	21.462	58383
8	0.15764	39.857	52318
7	0.13512	46.5	41101
11	0.2252	27.9	34996
15	0.31529	19.929	26583
13	0.27024	23.25	26334
12	0.24772	25.364	26184
94	2.0944	3	19776
47	1.03594	6.065	18335
70	1.55391	4.043	14754
30	0.65309	9.621	13169

Figure2.3.1

The top 10 harmonics with the highest amplitudes to include in a cyclical trend model:

Harmonic 1 with period 279, Harmonic 2 with period 139.5, Harmonic 3 with period 93, Harmonic 4 with period 69.75, Harmonic 5 with period 55.8,Harmonic 70 with period 3.986.Harmonic 9 with period 34.875. Harmonic 116 with period 2.405. Harmonic 13 with period 21.452.

Then we Created the corresponding sine and cosine pairs we have identified from the periodogram.

Specifying 30 observations as hold-out samples and we estimate cyclical models by using “log Linear Trend” and the sine and cosine pairs we have created. Then we obtain the plot of the model and the parameter estimate.

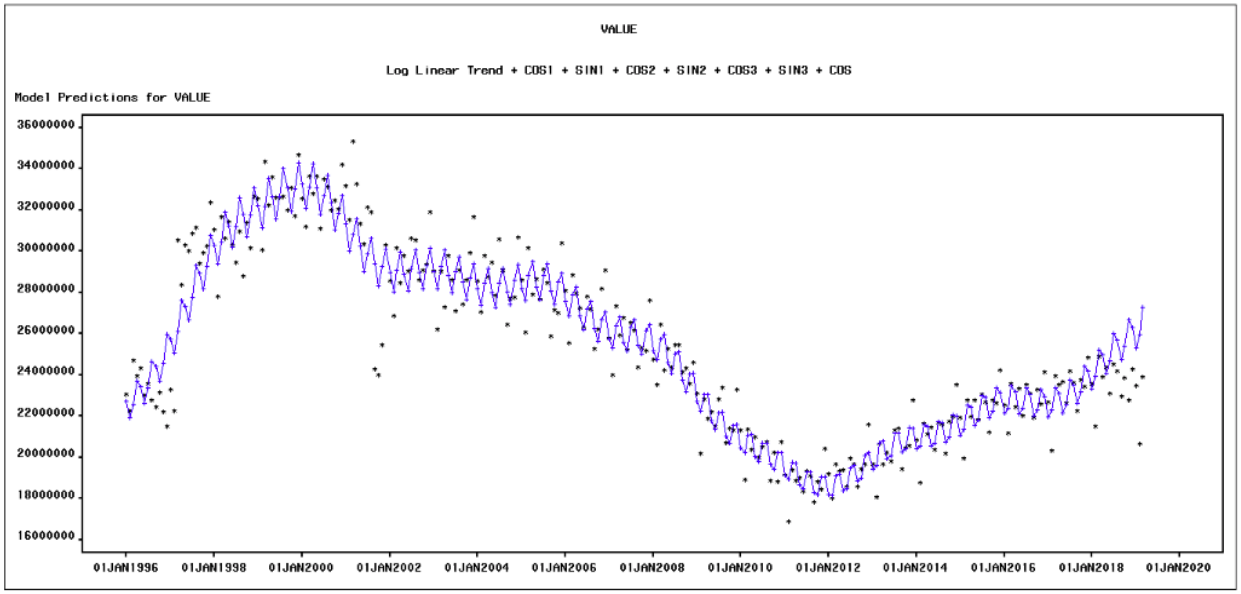


Figure 2.3.2

The Figure 2.3.2 shows how the model with top8 sine and cosine pairs and log linear trend capture the data.We can see the trend can be captured by the Cyclical model but some outliers as well as hold-out samples can not be captured very well.

Parameter Estimates

VALUE

Log COS1 + SIN1 + COS2 + SIN2 + COS3 + SIN3 + COS4 + SIN4 + COS5 + SIN5 + COS70 + SIN70 + Linear Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	16.91131	0.0807	209.6849	<.0001
COS1	0.03446	0.0191	1.8053	0.0899
SIN1	0.30215	0.0531	5.6930	<.0001
COS2	0.01900	0.0168	1.1320	0.2743
SIN2	0.03276	0.0212	1.5440	0.1421
COS3	-0.05483	0.0135	-4.0474	0.0009
SIN3	0.02983	0.0102	2.9159	0.0101
COS4	-0.00969	0.0101	-0.9582	0.3522
SIN4	0.01010	0.0064	1.5876	0.1319
COS5	0.00336	0.0073	0.4622	0.6502
SIN5	-0.00990	0.0056	-1.7728	0.0953
COS70	0.03184	0.0047	6.8398	<.0001
SIN70	0.01106	0.0047	2.3767	0.0303
Linear Trend	0.0009443	0.000644	1.4654	0.1622
Model Variance (sigma squared)	0.00270	.	.	.

Figure 2.3.3

Statistic of Fit	Value
Mean Square Error	3.17798E12
Root Mean Square Error	1782687.7
Mean Absolute Percent Error	5.85815
Mean Absolute Error	1337140.8

Figure2.3.4

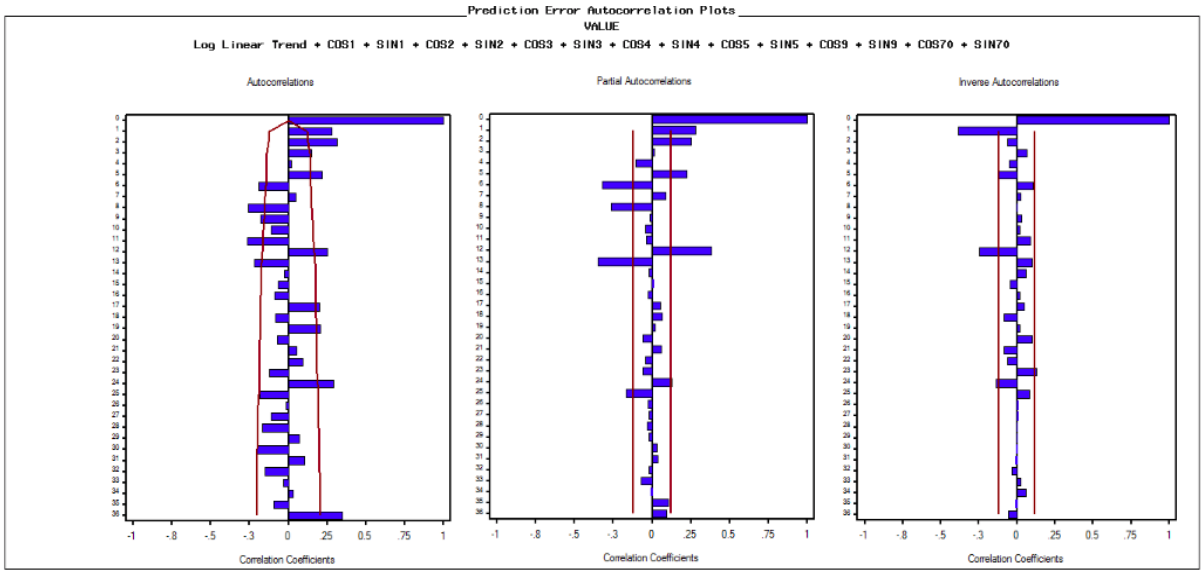


Figure 2.3.5

From Figure 2.3.5, we find the MAPE is 5.86. From Figure 2.3.3 and Figure 2.3.4, we find that most parameters are not significant, and the residuals are not stationary nor white noise, there exists some seasonal patterns.

2.2 ARIMA models (with seasonal ARIMA components if relevant)

From Figure 2.1.4, We can see the ACF of MEXICO data decays slowly ,which is not stationary.So constructing an ARIMA model seems necessary.

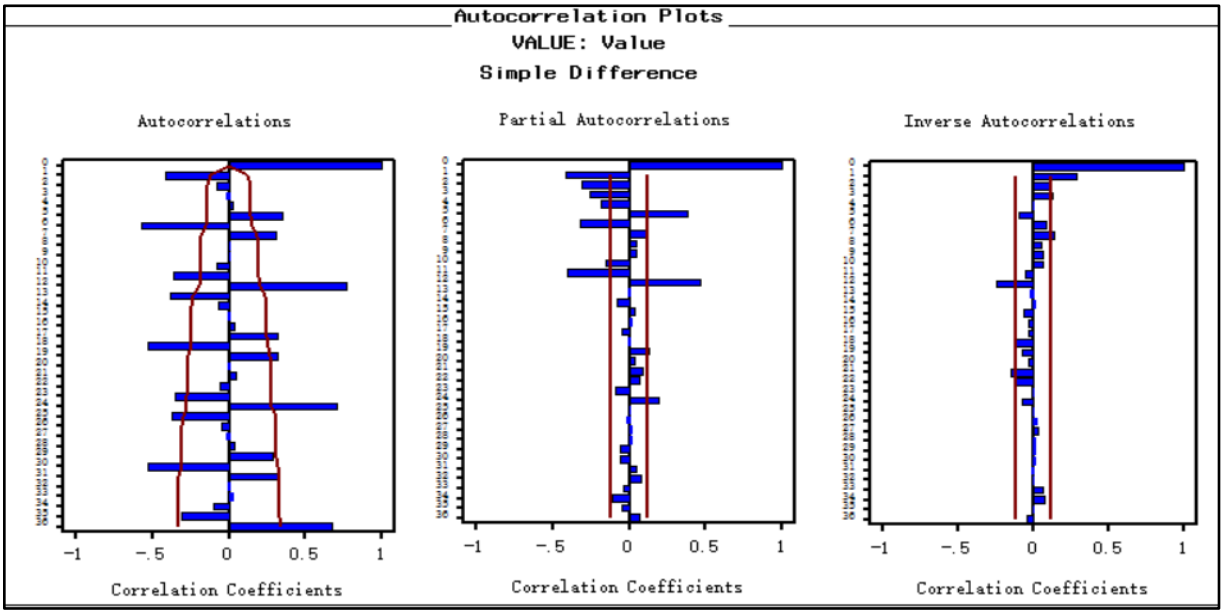


Figure 2.2.1

After taking the nonseasonal difference of the series, from figure 2.2.1, we observe

lag12,24 are significant, and ACF shows seasonality, thus the series has seasonal behavior of s=12.So we try a simple and seasonal difference.

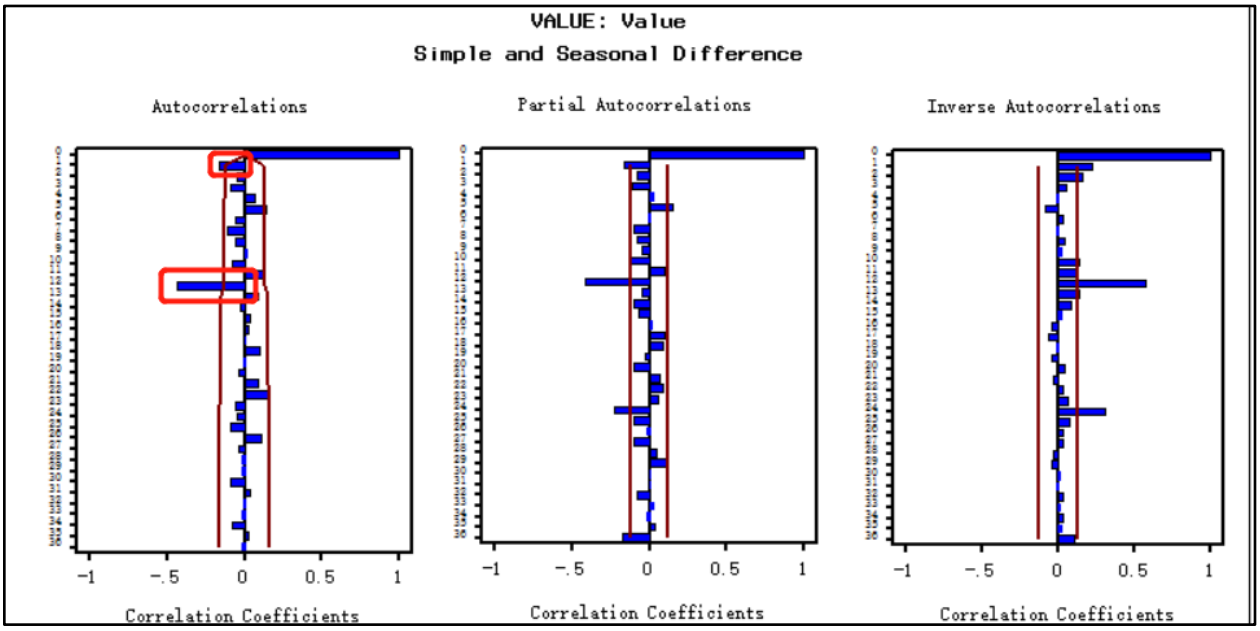


Figure 2.2.2

After taking the seasonal difference, from Figure2.2.2, the autocorrelation chopped off after seasonal lags1 and lags12. And Inverse autocorrelation plot decay quickly. so, the series is stationary.

The significant spike at lag 1 in the ACF suggests some additional non-seasonal terms need to be included in the model, that means we need a non-seasonal MA(1,1) component. And ACF comes up again at lag12 but then drops to 0.suggests we also need a seasonal MA(1,1) component.,

Then we try the seasonal multiplicative model: ARIMA(0,1,1)(0,1,1)seasonal=12 model. The series is stationary and the residual is white noise as we can see from Figure2.2.3 and Figure 2.2.4.

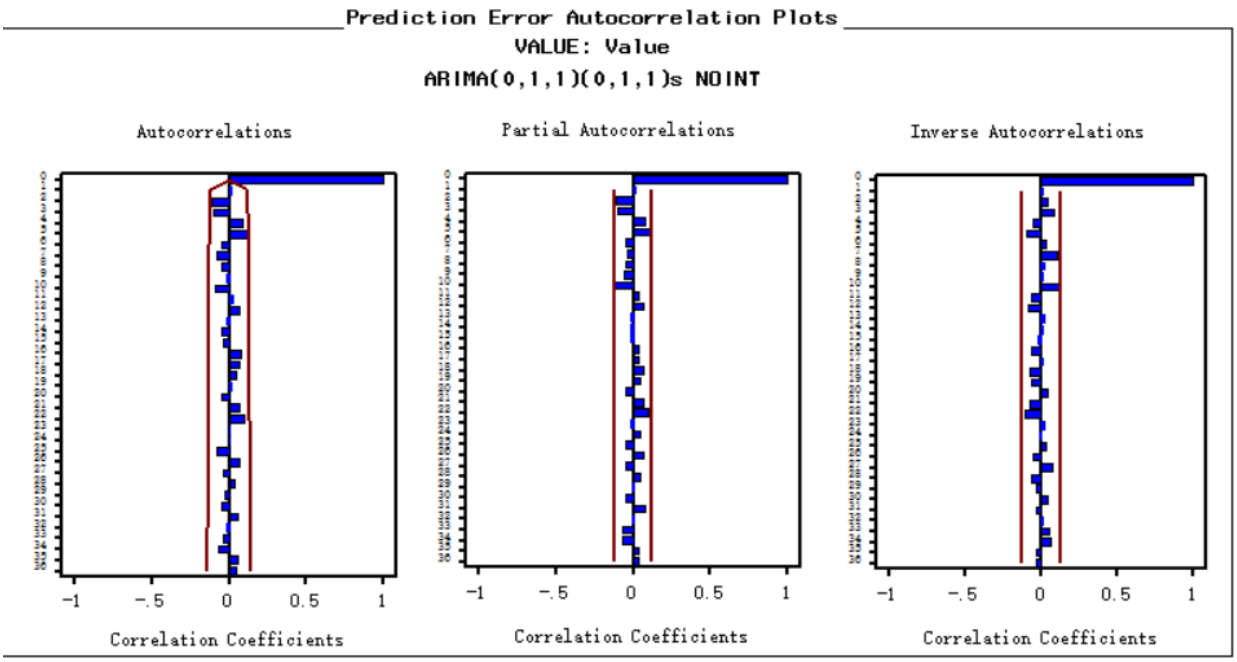


Figure2.2.3

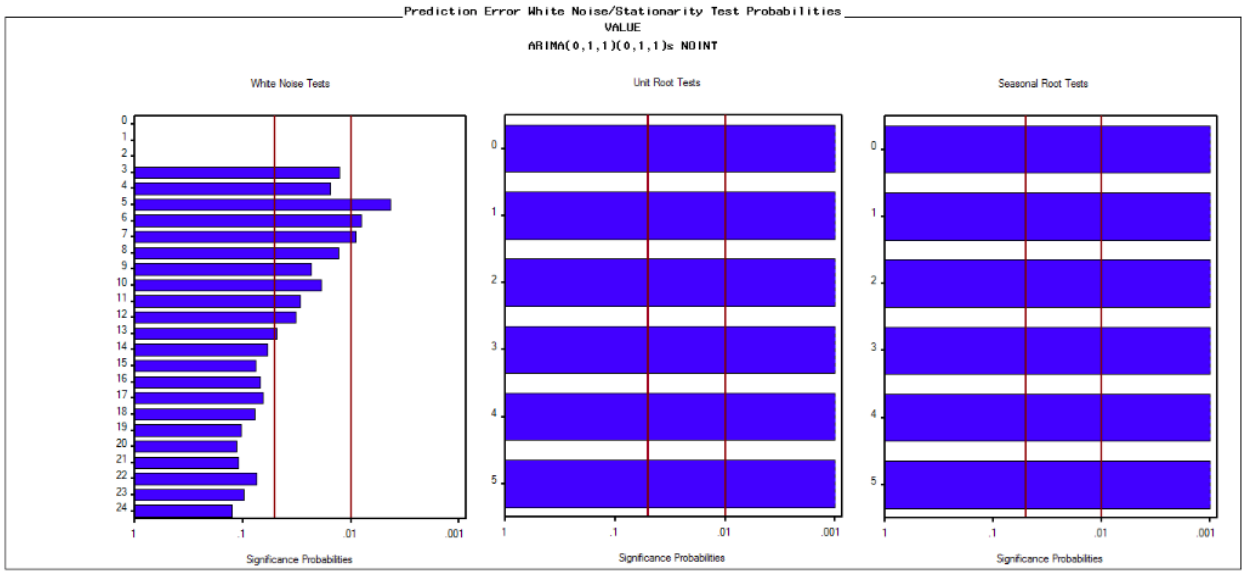


Figure 2.2.4

VALUE: Value
ARIMA(0,1,1)(0,1,1)s NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.16770	0.0633	2.6477	0.0132
Seasonal Moving Average, Lag 12	0.84290	0.0524	16.0946	<.0001
Model Variance (sigma squared)	9.50715E11	.	.	.

Figure 2.2.5

VALUE: Value
ARIMA(0,1,1)(0,1,1)s NOINT

Statistic of Fit	Value
Mean Square Error	2.0066E11
Root Mean Square Error	447951.0
Mean Absolute Percent Error	1.62211
Mean Absolute Error	376214.2

Figure 2.2.6

We can see from Figure2.2.5, the T test shows the p-value of theta1 and theta 12 is lower than 0.05, which means they are significant.
From Figure2.2.3, the autocorrelations are all inside the 2 standard error bounds, so it's stationary, and from Figure2.2.4, the p-value of white noise test is larger than 0.05, so the residuals are white noise.
The above shows the ARIMA(0,1,1)(0,1,1)S=12 is appropriate, and the prediction error is quite low which is 1.622(MAPE).

2.3 Comparison of models (in terms of fit and validation)

After analyzing the four models, we obtain a comparison table of their predicted errors and variance.

Models	RMSE	MAPE	MAE	Variance
Seasonal dummy and trend	4269098.0	18.06757	4236727.8	7.85783E11
Error Model	875688	2.17498	556454.6	8.04301E11
Cyclical Model	1782687.7	5.858	1337140.8	0.00259
ARIMA Model	447951.0	1.62211	376214.2	9.50715E11

Figure 2.3

The Root Mean Squared error (RMSE) is the square root of the MSE, Mean Squared error (MSE) is a measure of how close a fitted line is to data points. We compare the RMSE of models to observe variation in measurements of a typical point.

As for Mean Absolute Percent error (MAPE) and Mean Absolute Error, they are used to measure the forecast accuracy of models. We did not consider R-square in this project because R-square is unappropriate in SAS.

We will compare the Mean Absolute Percent error of four models. We can see the Log

ARIMA model has the best performance fitting the United States-Mexico border entry data, which generates the lowest Mean Absolute Percent error, that is 1.62211. The second best model is the 1st order Moving Average Error model with taking first difference and seasonal dummies. It has 2.12796 Mean Absolute Percent error. Cyclical model has Mean Absolute Percent error 5.858, while the model with seasonal dummies and trend has the highest Mean Absolute Percent error, which is about 18.

3. Multivariate Time series model

Transfer Function Model

Firstly ,we want to use Mexico as the predictor to find the relationship between US-Canada border value and US-Mexico border value .

We take the Border value of Mexico as X and the Border value of Canada as Y to develop the cross correlation function analysis to see whether they are correlated in time series.

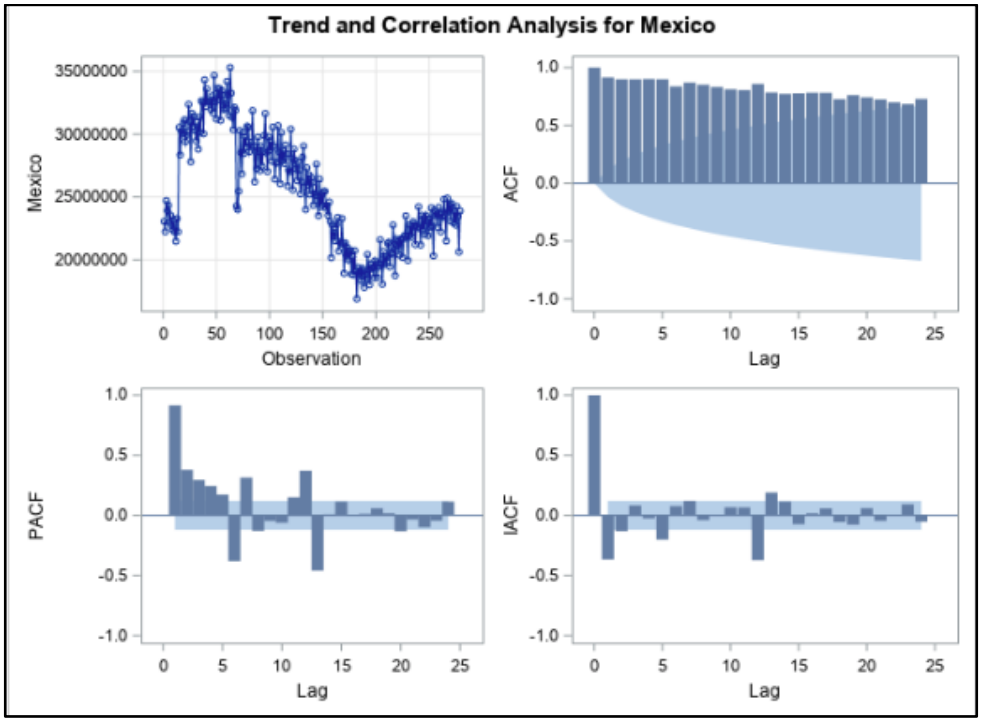


Figure 3.1

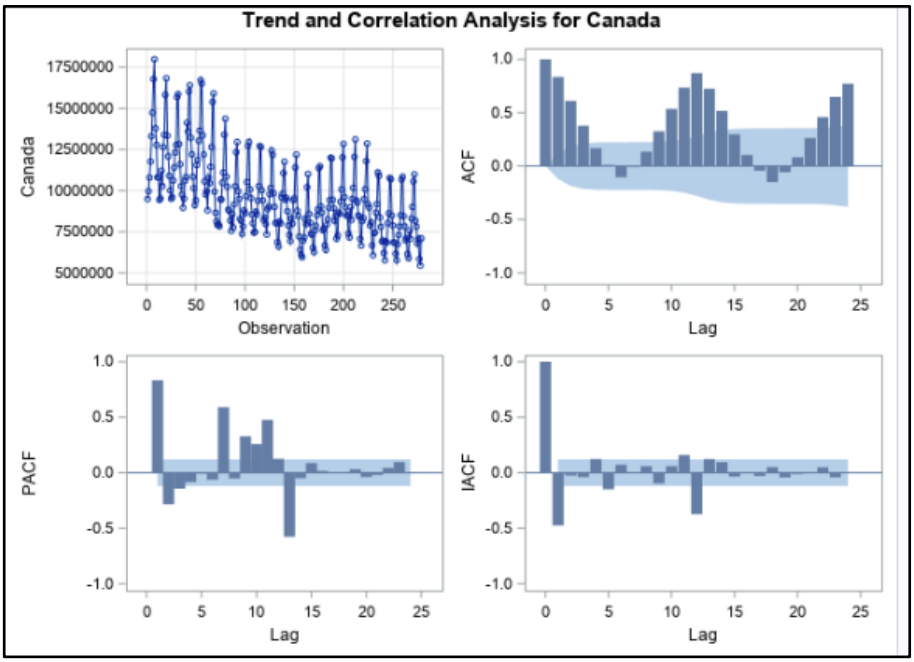


Figure 3.2

Based on figure 3.1 and figure 3.2, we can see two series are not stationary, so we take the seasonal difference and simple difference.

Name of Variable = Mexico	
Period(s) of Differencing	1,12
Mean of Working Series	-4632.06
Standard Deviation	1208865
Number of Observations	266
Observation(s) eliminated by differencing	13

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	16.65	6	0.0107	-0.159	-0.052	-0.085	0.063	0.138	-0.056
12	79.09	12	<.0001	-0.112	-0.055	0.018	-0.079	0.115	-0.433
18	84.80	18	<.0001	0.089	-0.031	0.031	0.028	0.000	0.097
24	97.28	24	<.0001	-0.006	-0.035	0.090	0.165	-0.062	-0.048

Figure 3.3

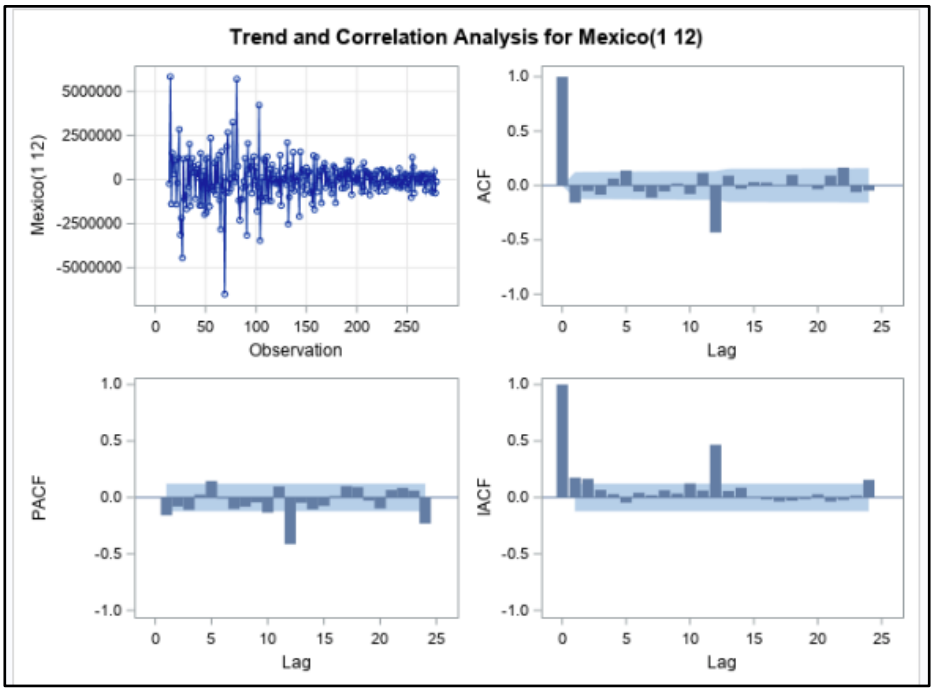


Figure 3.4

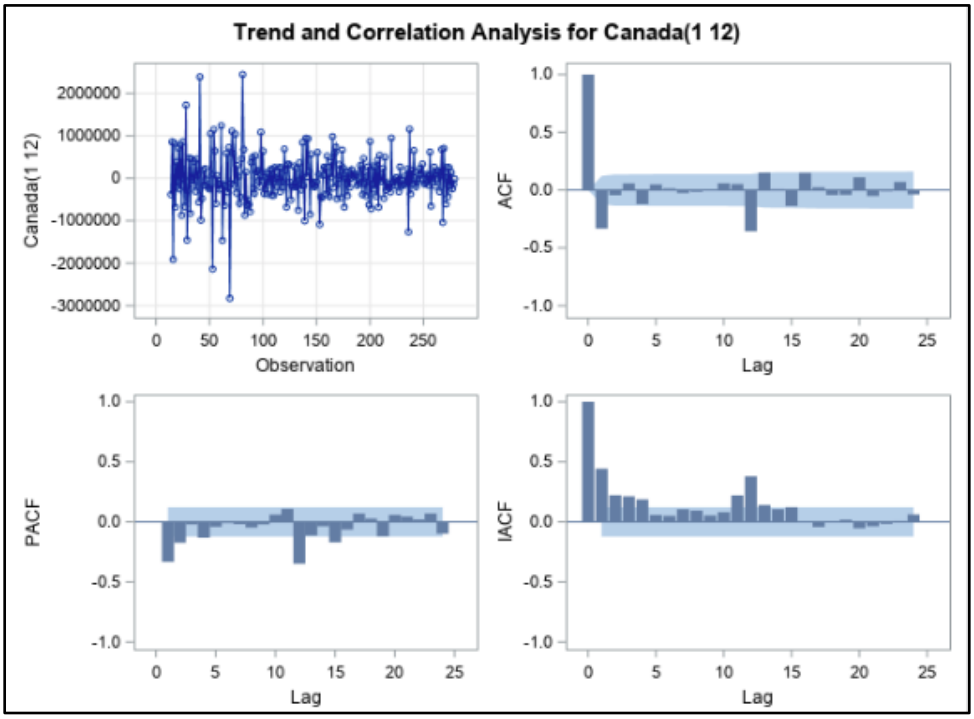


Figure 3.5

After taking the seasonal difference, shown in the figure3.4 and figure3.5, we find the two series are stationary. However, Mexico as the independent variable is not White noise(Figure3.3). So we apply the ARIMA (0,1,1)(0,1,1)s=12 model on Mexico.

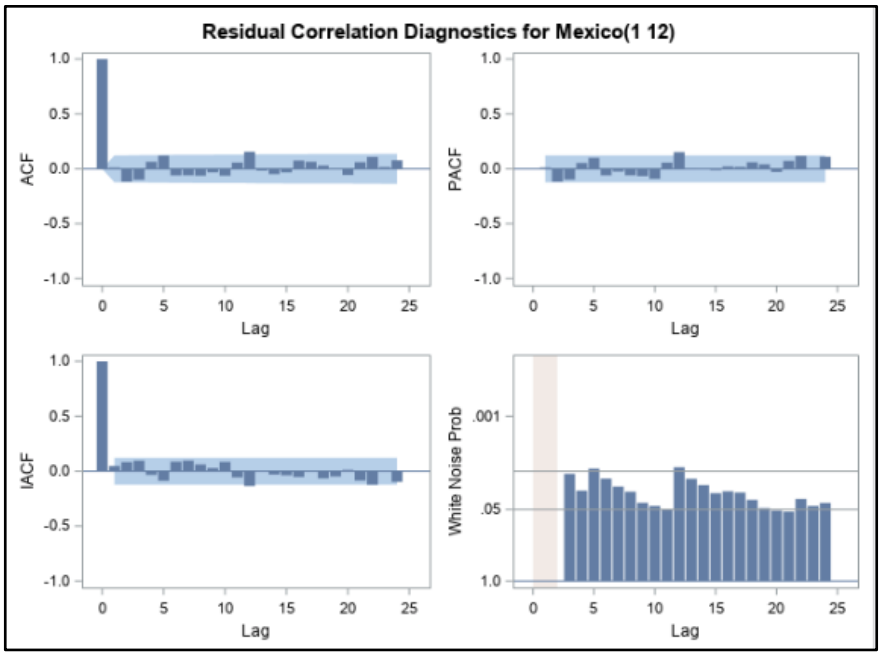


Figure 3.6

After the prewhitening process to Mexico,both ACF plot and White noise test shows the residual is white noise(Figure3.6). Since Mexico is white noise after taking the difference and Canada is stationary after differencing, then we apply the cross correlation function on Mexico and Canada.

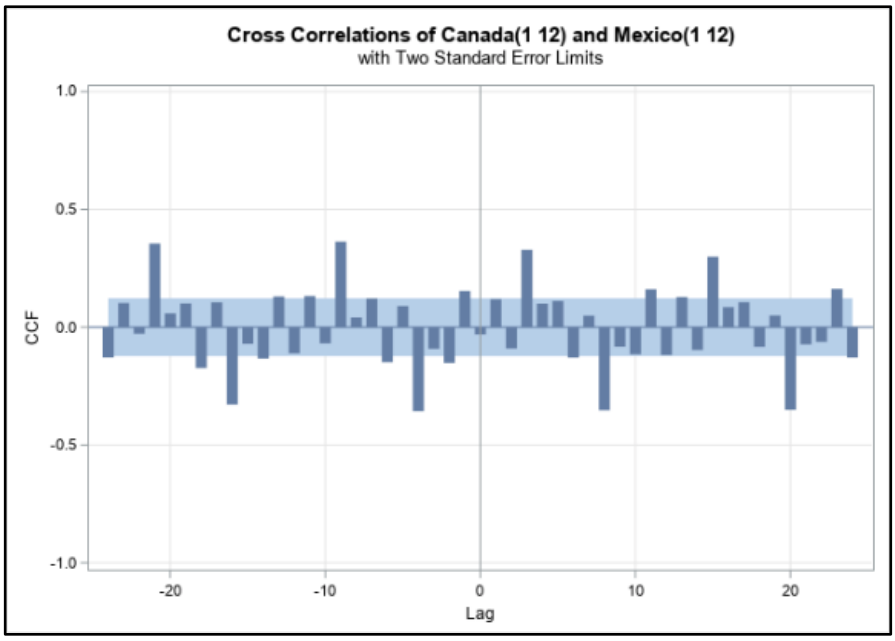


Figure 3.7

We can obviously find that the CCF is significant both at negative and positive lags(Figure3.7),

which means the border values of US-Mexico and US-Canada are related to each other. However, TF function is not appropriate since CCF is significant at negative lags.

4. Conclusion

From this project, we know that the ARIMA model predicts the dataset the best due to the lowest Mean Absolute Percent error rate.
Secondly, although we cleaned our data by grouping month, it does not show a monthly trend. We discovered hidden dominant periods by using a cyclical model.
On top of that, the Cross Correlation Function (CCF) of multivariate time series analysis indicates that there is a relationship between United States-Mexico border entries and United States-Canada border entries. However, it seems there are patterns for both positive and negative lags. We will not be able to reveal the relationship by using the Transfer Function model.
However, different presidents and ruling parties seem to have an effect on border entry. From the below seasonal dummies and trend plot, we see that when the ruling party is republican, the number of entries is decreasing, and when the ruling party is democratic party, the border entry seems to be more flexible and has an upward trend. We will need more data and try more time series models to figure the relationships between two borders out.