STAT 6363: Course Project

Modeling Route and Mode Choice Using Time Series

Duleepa Liyanage

1. Introduction

Travel demand forecasting is quintessential for any country that provides for its transport supply infrastructure. The overall travel demand reflects how much transport supply is required in terms of road capacity, number of lanes, vehicles, and other supportive infrastructure such as bridges. A fictitious forecasting of travel demand would result in over or underestimate of travel supply. For example, if road capacity is determined based on its average daily demand – an easily accessible metric – the corridor reaches its peak volume for half the day. Therefore, we need robust and more accurate ways to predict travel demand rather than relying on descriptive measures.

Transport experts have developed various models to estimate the demand for specific road sections, among which is the Gravity Model [1]. This model relies on factors like population size, distances between locations, and socio-economic data of cities of interest. However, the Gravity Model presents challenges in measuring these variables accurately and lacks a standardized form, leading to less reproducible estimates. Additionally, as these factors often exhibit minimal variance over time, their predictive power is limited. To address these limitations, we opted for time series analysis to forecast trip demand more precisely, without the need for additional covariates typically required by the Gravity Model.

The dataset utilized in this project comprised monthly time series data sourced from the U.S. Department of Transportation [2] from 1996 to 2022. It includes 88 pairs of ports between the United States and Canada, from which we selected eight adjacent port combinations for analysis: Sault Sainte Marie, Detroit, Port Huron, Buffalo Niagara Falls, Alexandria Bay, Ogdensburg, Massena, and Champlain Rouses Point. For convenience, we refer to these ports by their U.S. names only. Each port has two types of time series: Y_t represents the overall trip volume between the U.S. and Canada for each route, while X_t decomposes Y_t into its modal components. Therefore, for each vehicle type such as buses, trains, personal vehicles, trucks, and loaded rail containers, there exists an X_t time series.

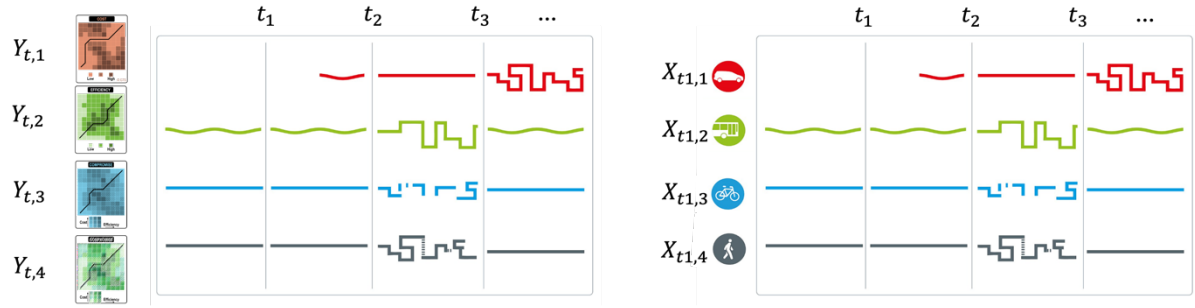

Figure 1 Route Choice and Modal Choice Time Series

Figure 1 illustrates the relationship between the two types of time series. $Y_{t,1}$ denotes the time series of route 1. $X_{t1,1}$ denotes the modal specific (e.g. car) time series of route 1. This decomposition of Y_t into its modal components allows us to understand how different vehicle sizes contribute to traffic, with larger vehicles typically occupying more road space and contributing to higher traffic volumes. Given that we have chosen adjacent ports for analysis, our focus is on examining the relationship between these ports, as users are likely to consider proximity when making port selection decisions. We anticipate that nearby ports will exhibit stronger correlations compared to more distant ports. We are also interested in observing trip patterns of commuters other than their route and mode selections.

Figure 2 illustrates the variation in traffic volumes across different ports (Y_t). We demonstrated only three ports to preserve the clarity of the plot.

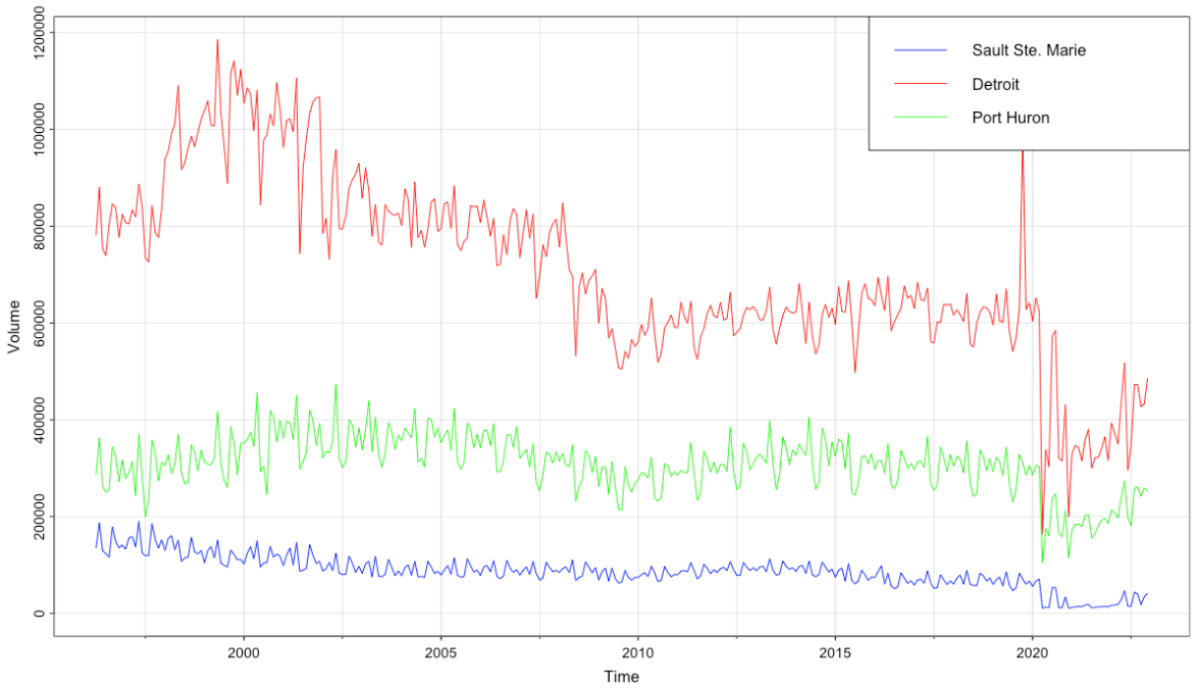

Figure 2 Time series of ports

Figure 3 highlights the contributions of decomposed modal time series (X_t) to the overall time series of Detroit. By breaking down the total trip volume into modal components such as buses, trains, and personal vehicles, we gain insight into how each type of vehicle contributes to the overall traffic pattern.

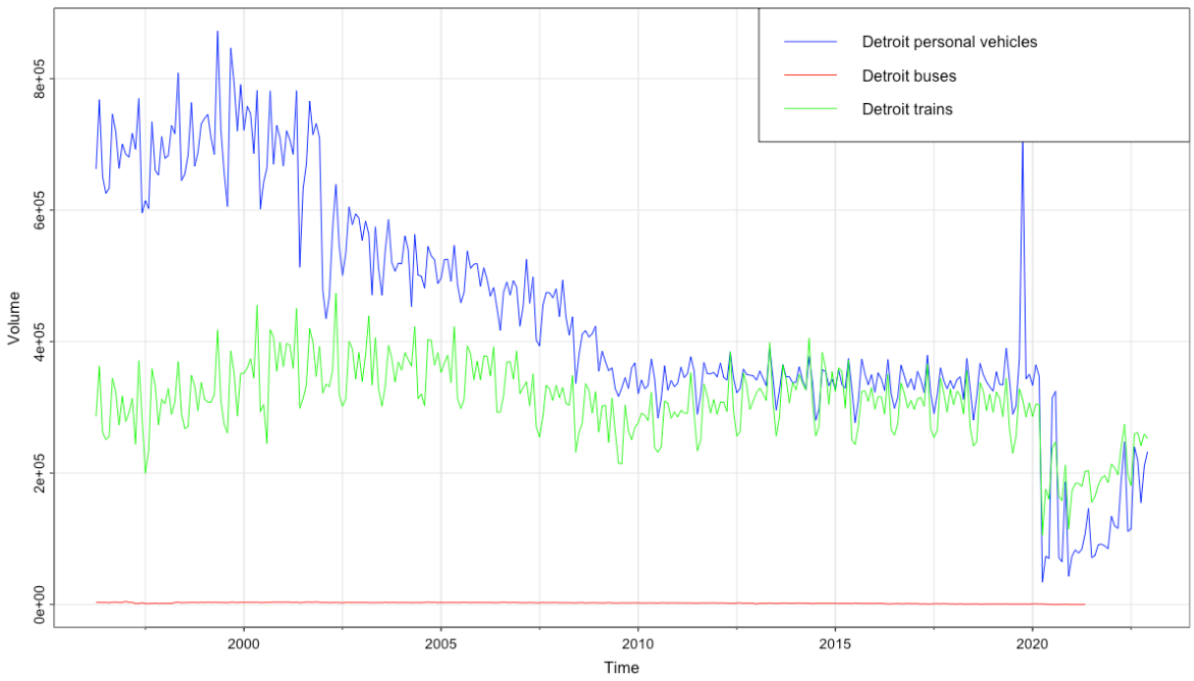

Figure 3 Time series of modes

2. Statistical Analysis

2.1 Characteristics

We employ multivariate time series analysis to explore the correlation among adjacent ports. Initially, we examine the characteristics exhibited by these different port-based time series.

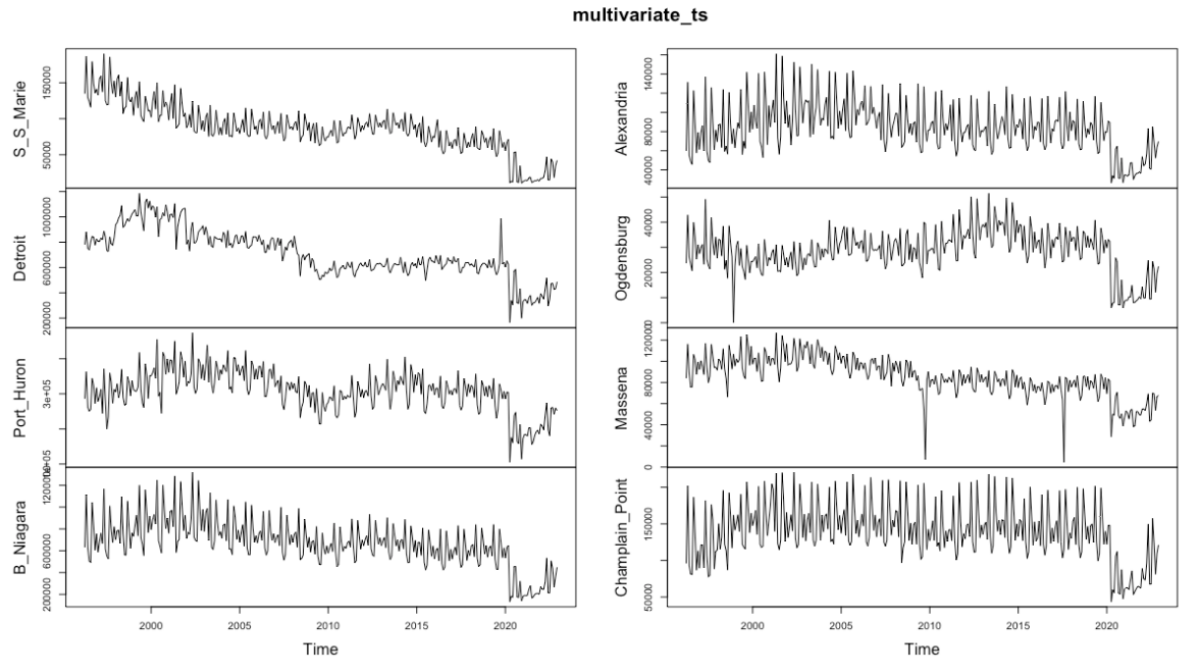

Figure 4 Port Time series (Individual)

Figure 5 Port Time Series (Group)

2.2 Stationarity

We first checked the stationarity of the time series using the Augmented Dickey Fuller (ADF) test. Table 1 displays the p-values of each time series.

Port	P value
S_S_Marie	0.64557331
Detroit	0.09955494
Port_Huron	0.31577861
B_Niagara	0.34594490
Alexandria	0.27964280
Ogdensburg	0.82640610
Massena	0.02020696
Champlain_Point	0.40620513

Table 1 Stationarity based on ADF test

Based on the table 1 we can see except Massena, all ports have non-stationary time series (large p-values). Therefore, we get the first differencing of the multivariate time series and recalculated the ADF test.

Port	P value
S_S_Marie	0.01
Detroit	0.01
Port_Huron	0.01
B_Niagara	0.01
Alexandria	0.01
Ogdensburg	0.01
Massena	0.01
Champlain_Point	0.01

Table 2 ADF test after differencing

Table 2 demonstrates that the p-values based on ADF test after differencing. We can see all eight-time series are now stationary (small p-values).

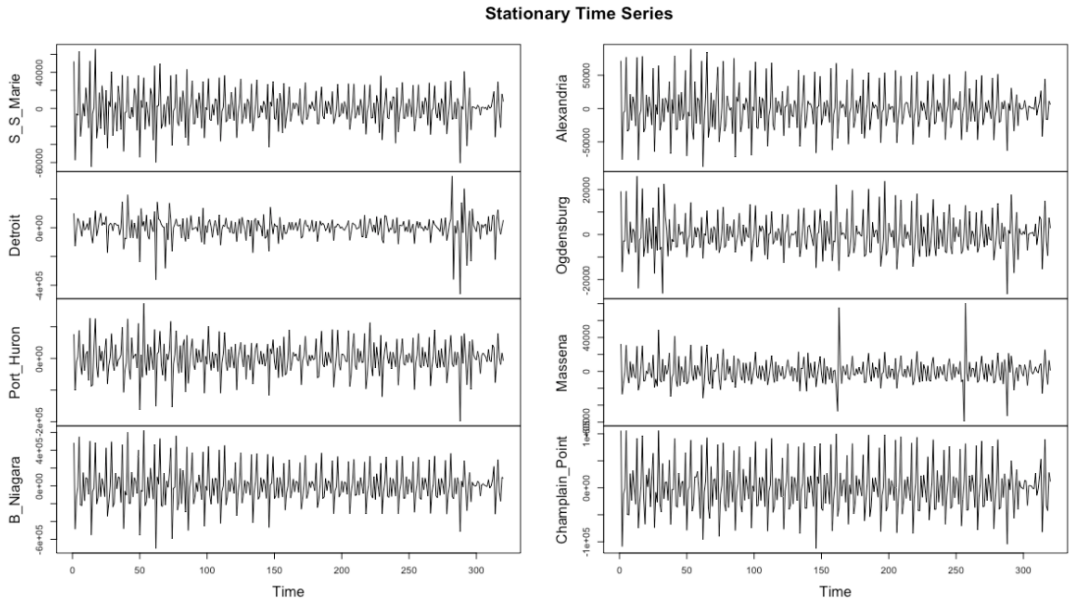

Figure 6 Stationary multivariate time series

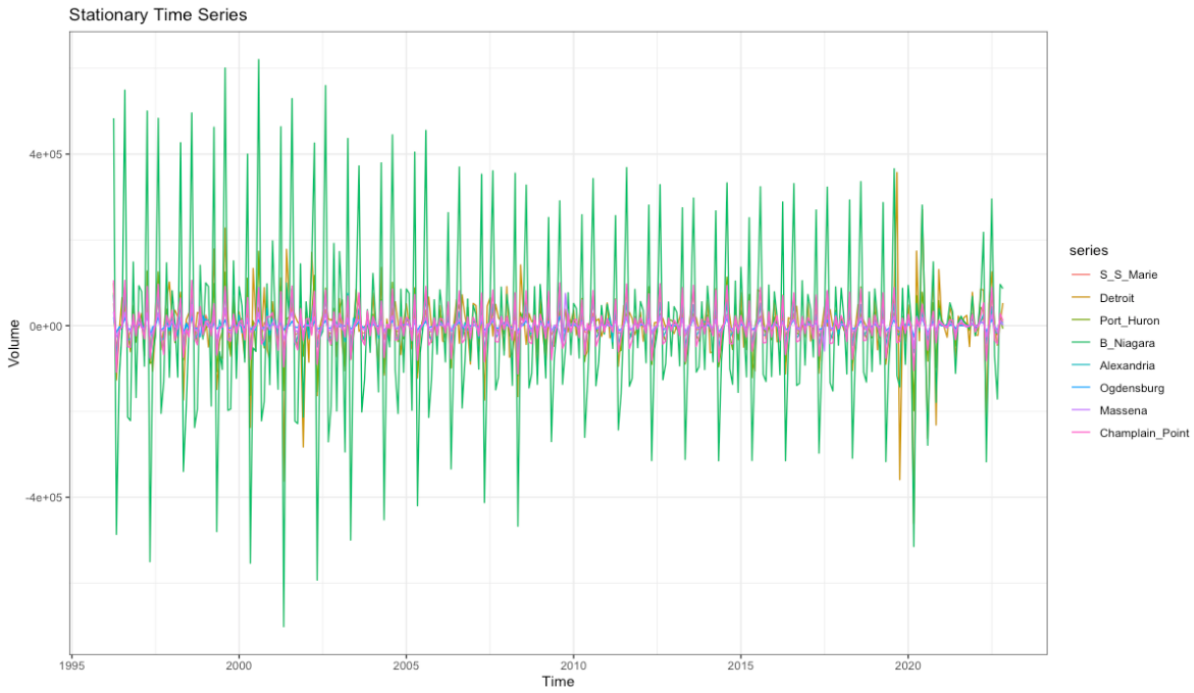

Figure 7 Stationary time series

Figure 6 and figure 7 show the stationary time series.

2.3 Model

Since we have dealt with stationarity, now we can build the model. The model we use is VAR(2) with seasonality (12) based on the ACF plots that we will show below. $Y_{t,1}$, $Y_{t,2}, \dots, Y_{t,88}$ follow multivariate vector autoregressive model of order 2 as follows.

$$\begin{bmatrix} Y_{t,1} \\ Y_{t,2} \\ \vdots \\ Y_{t,88} \end{bmatrix} = \begin{bmatrix} \beta_{t,1} \\ \beta_{t,2} \\ \vdots \\ \beta_{t,88} \end{bmatrix} + \sum_{j=1}^{2} \begin{bmatrix} \phi_{j,11} & \cdots & \phi_{j,1q} \\ \vdots & \ddots & \vdots \\ \phi_{j,q1} & \cdots & \phi_{j,qq} \end{bmatrix} \begin{bmatrix} Y_{t-j,1} \\ Y_{t-j,2} \\ \vdots \\ Y_{t-j,88} \end{bmatrix} + \begin{bmatrix} a_{t,1} \\ a_{t,2} \\ \vdots \\ a_{t,88} \end{bmatrix}$$

$a_{t,1}, a_{t,2,\dots,}a_{t,88}$ are 88 independent white noise time series. Note that we only considered $Y_{t,1}, Y_{t,2}, \dots, Y_{t,8}$ subset from total of 88 time series for the illustration.

2.4 Model diagnostics and residuals

Figure 8 displays the ACF plots of the multivariate time series for the last four ports. The leading diagonal plots represent the ACF plots of the ports themselves, while the off-diagonal plots depict the cross-correlation. Some seasonality is evident in these plots, underscoring the rationale behind incorporating a seasonality component in our VAR model. Without accounting for seasonality, the ACF plots would have exhibited very high lags (we have checked this beforehand).

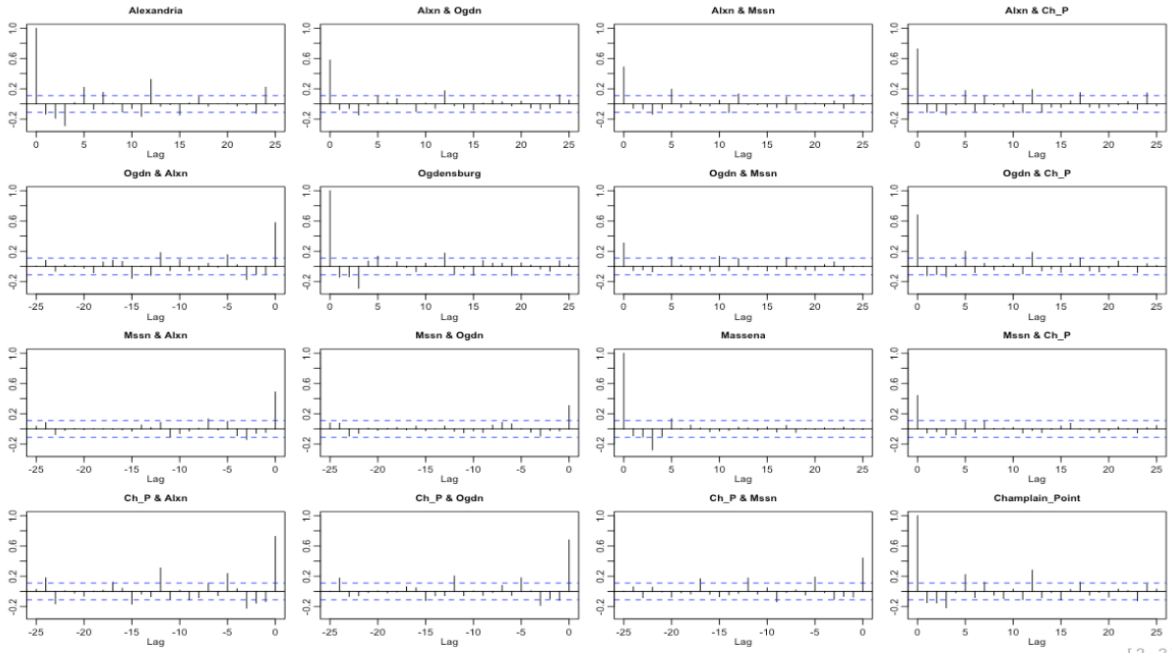

Figure 8 ACF plot of Multivariate time series

2.5 Causality

Since we have now dealt with modeling of time series, we can now investigate the causality. We use ***"Granger causality"*** test [3] and *"**instantaneous causality**"* to identify the causality between different time series above and beyond their own past lags. We use each port at a time and tested for its two types of causalities as in table 3.

Port	Granger-P value	Instant-P value
S_S_Marie	0.3999	0.0000
Detroit	0.0026	0.0000
Port_Huron	0.0106	0.0000
B_Niagara	0.0591	0.0000
Alexandria	0.3177	0.0000
Ogdensburg	0.0336	0.0000
Massena	0.9549	0.0000
Champlain_Point	0.1361	0.0000

Table 3 Causality

As we can see that all instantaneous p-values are significant meaning that all ports significantly contribute towards predicting the current lags of other port traffics. However, based on Granger causality we can infer that only Detroit, Port Huron, and Ogdensburg contribute to predict future traffic at other ports.

2.6 Forecasting

Figure 9 shows the forecasting of eight time series based on the model we picked (i.e. VAR(2) with seasonality (12)).

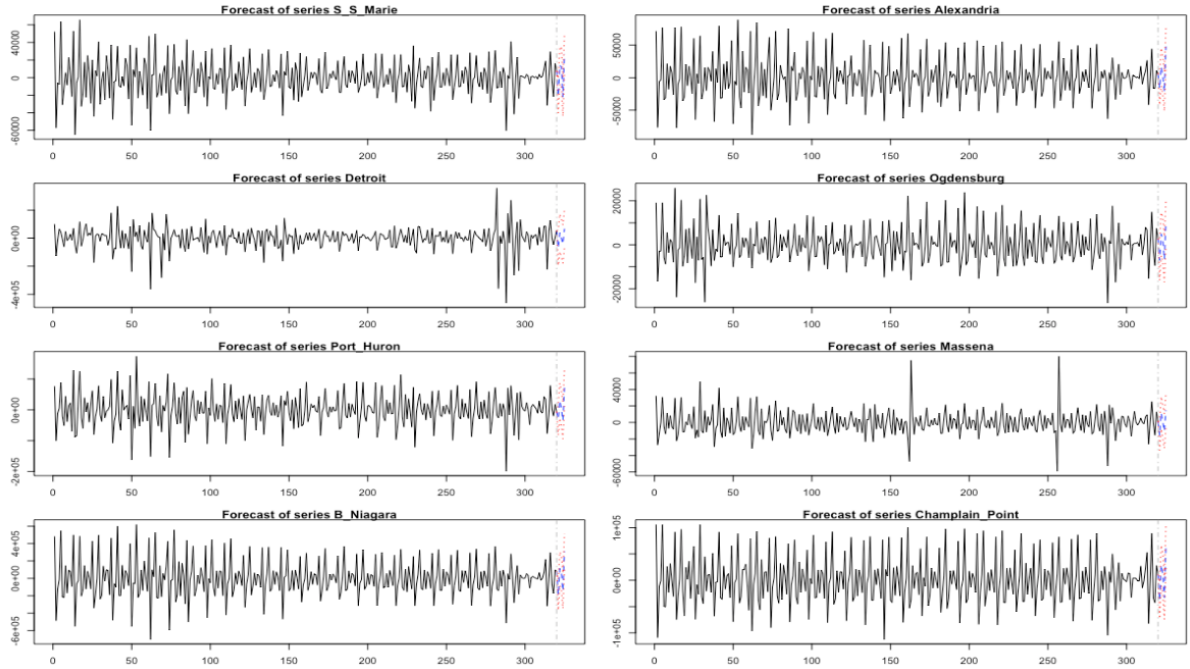

Figure 9 Forecasting

We compared the forecasting error with ARIMA models based on orders set by `auto.arima()` function and VAR models as follows. Table 4 provides the forecasting error for first two ports.

Model	Series	Forecast Error
ARIMA Univariate	Yt_1	262452430
ARIMA Univariate	Yt_2	7777405896
ARIMA Univariate	Yt_1 + Yt_2	8039858326
VAR Multivariate	Yt_1	345681233
VAR Multivariate	Yt_2	6057116931
VAR Multivariate	Yt_1 + Yt_2	6402798165

Interestingly, we have a lower forecast error in ARIMA compared VAR in $Y_{t,1}$. However, VAR outperforms ARIMA in forecasting $Y_{t,2}$. Thus, overall error is less in VAR than ARIMA.

3. Conclusion

In this project we mainly considered the Y_t multivariate time series. However, X_t is also a multivariate time series that one can make all the same analysis. We first made the Y_t stationary using the first order differencing. It was challenging to identify the (P) order in VAR(P) model, so we had to use few iterations to figure out the model order using residual diagnostics such as ACF plots. Despite these efforts, residual correlations persisted, potentially affecting our selection of the model order. We used `serial.test()` function to show that still there is serial autocorrelations in residuals. We found that all ports exhibit causal relationships in the current time period, while some ports exhibit causality in future time periods, as determined by instantaneous and Granger causality tests, respectively. Finally, we conducted forecasting using both ARIMA and VAR(2) models to compare their accuracy. Our analysis revealed that the VAR(2) model outperformed ARIMA in forecasting the multivariate time series.

References

[1] https://www.princeton.edu/~alaink/Orf467F12/The%20Gravity%20Model.pdf

[2] https://www.bts.gov/

[3] http://www.scholarpedia.org/article/Granger_causality