ECON 509: Time Series Methods in Financial Econometrics

Paper Replication of:

Combining official and Google Trends data to forecast the Italian youth unemployment rate

Reginald Acquah (1623977)

University of Alberta

Department of Economics

Summary of the Paper

The paper's focus is to forecast the youth unemployment rate (YUR) over the short term by including external factors. The motivation behind the paper was from the lack of external factors inclusion into econometric models used for forecasting in all European countries. The department of statistics solely develops econometric models based on data collected from sample surveys. The lack of external data reduces the ability of the model to forecast variations expected in the short term.

The official estimates of the youth unemployment rate are based on a monthly survey, and the results are available 30 days after the end of the month. To get a fair idea of what youth unemployment rate, a reliable forecast can be made readily available as the youth unemployment rate is of importance to make preliminary estimates to use within macro-econometrics framework forecasting models to know how the economy is performing.

At the time the paper was published, the Italian National Institute of Statistics (ISTAT) uses the ARIMA methodology to provide short-term forecasts (at most 3 months) (Naccarato, et al., 2018). The paper aimed to provide an alternative model that will probably outperform the ARIMA model used. By so doing, the author included external data from Google Trends (GT) with the keyword "*offerte di lavoro*" (job offers). A VAR model using both official unemployment data and GT was developed to challenge the predictability of the ARIMA model that uses only unemployment data. The table below shows the VAR model used to predict the 3 months forecast.

Table 1: VAR model of the original paper for 3 months forecast

Parameter	Estimate	Std. Error	t-Value	p-Value
Estimation results for equation YUR:				
ΔYUR.l1	-0.4127	0.0798	-5.17	0.0000
ΔYUR.l2	-0.2927	0.0797	-3.67	0.0000
ΔYUR.l5	0.2580	0.0783	-3.29	0.0013

ΔGT.l5	1.0605	0.2386	4.44	0.0000
Estimation results for equation GT:				
ΔGT.l1	-0.26023	0.0851	-3.058	0.0027
ΔGT.l8	0.1826	0.0808	2.2593	0.0257

Source: (Naccarato, et al., 2018)

To assess the performance of each model, the rolling estimation was used to forecast for 1- and 3-months horizons and absolute mean (AE) values of the ARIMA and VAR model obtained. With the strong asymmetric nature of the AE distribution for each model, the author compared the median of the AE distribution for both models. The median of the VAR is lower than the ARIMA for both horizons. The Thiel index was also computed for both models and arrived at the same conclusion.

Table 2: Actual and forecast values of the youth unemployment rate

Month	Actual Value[1]	ARIMA Value	ARIMA AE	VAR Value	VAR AE
April 2015	127.3	133.2	5.9	124.0	3.3
June 2015	131.2	120.6	10.6	122.8	9.4

Source: (Naccarato, et al., 2018)

Table 3: Theil index for the ARIMA and VAR models

Forecast horizon	ARIMA Model	VAR Model
h = 1	0.789	0.775
h = 3	1.106	1.067

Source: (Naccarato, et al., 2018)

From the table above, the paper concluded that the VAR model had better out-of-sample performance over the ARIMA model used by ISTAT. Therefore, including GT as an auxiliary variable was recommended to be included in the model as it gives a more reliable forecast for the youth unemployment rate.

[1] The actual values as presented by the authors are greater than 100 which suggest that data has been standardized. However, the method for standardizing is not explained in the paper.

Literature Review

The vector autoregressive (VAR) model is an extension of the univariate autoregressive time series models. The VAR model is useful in describing the dynamic behaviour of economic and financial time series and forecasting it (Zivot & Wang, 2006). Forecasting the unemployment rate – an important macroeconomic indicator with the VAR model will produce much accurate results. Also, the ARIMAX model is an extension of the ARIMA model. The ARIMAX model adds a potential predictor of the target variable as an exogenous variable. Although the addition of the exogenous variable intricates the model, it enables the model to capture the effects of external factors (Andrews, et al., 2013).

Google trends provide information about the total volumes of a search for a keyword. The series is normalized with a scale between 0 to 100 such that the size of the geographical location will not affect the data. The value of 0 is assigned to insufficient data (Google, n.d.). Google trend series is problematic – the data changes in a vaguely erratic way. (Leinweber, 2013) noted that the plot for the keyword "debt" changed within a space of three days.

The choice of keyword to use for extracting the google trends data is vital. The risk of spurious correlations may arise – the keyword is primarily selected based on correlation with the target (Braaksma & Kees, 2017). The correlation may be completely random (i.e. Spurious). Hence, to ameliorate the risk, different keywords are considered, and the keyword that provides a statistically significant correlation with a verifiable intuition is selected. In the paper under review, the keyword "*offerte di lavoro*" proved popular among key job search-related keywords (Francesco & Juri, 2017).

Using the GT series comes with its limitations. In our study, we cannot only argue that not all workers go on the internet to search for jobs. Also, it does not carry information about the

employment status of the online job seekers – gainfully employed people may also be browsing for new jobs (Naccarato, et al., 2018). The age of the online job seeker is also not captured by the data. As we study the youth unemployment rate, it will be important to capture online job seekers between 15-24 years. The choice of the Google trend series is relevant because (Italian National Statistical Institute (ISTAT), 2016) reveals that online job seekers are primarily young people – 60% of internet search for a job search in December 2014 was among the young.

With the information of people's needs that they disclose on the internet (Ettredge, et al., 2005), it is possible to derive a wide range of information about social phenomena through online search data (Nikolaos & Klaus, 2015). Google Trends (GT) is a particular data source that provides information about online search data.

Performance metrics are necessary elements of model development. Various measures for accessing model performance are widely available. MAE is mean absolute error, MAPE is absolute percentage error, sMAPE is symmetric mean absolute percentage error, RMSE is the root mean squared error, and R2 is the coefficient of determination are some commonly used error metrics (Alessandrini & Sperati, 2017). Furthermore, the Diebold-Mariano test (Diebold & Mariano, 1995) is used as a measure to test the equal predictability of models. The Diebold-Mariano test statistic was modified by (Harvey, et al., 1998) to improve the finite sample properties of the test

Based on the literature, the study will include an ARIMAX model as a challenger to the VAR model as proposed by (Naccarato, et al., 2018). The equal predictability of the models will be assessed further to test if the VAR model is superior to the ARIMAX model.

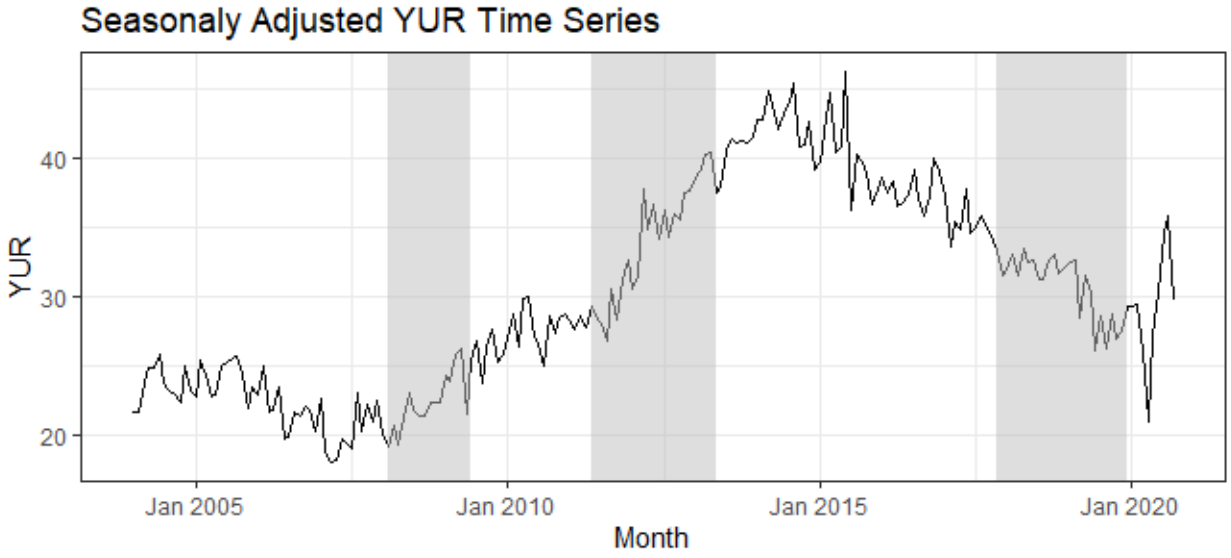
Data

The data used for the study is a monthly youth unemployment rate (YUR) in Italy. The time series source comes from the ISTAT labour force survey, the primary source of Italy's labour market information. The YUR definition used in the paper is the ratio of unemployed persons between 15-24 years against the corresponding labour force. The data window is between January 2004 to September 2020. For model replication, data will be limited to January 2004 to June 2015.

The Google Trends (GT) series will be extracted from Google. We will focus on Italy as the country and "*offerte di lavoro*" (job offers). Weekly GT data will be collected between January 2004 to September 2020 and aggregated into monthly averages.

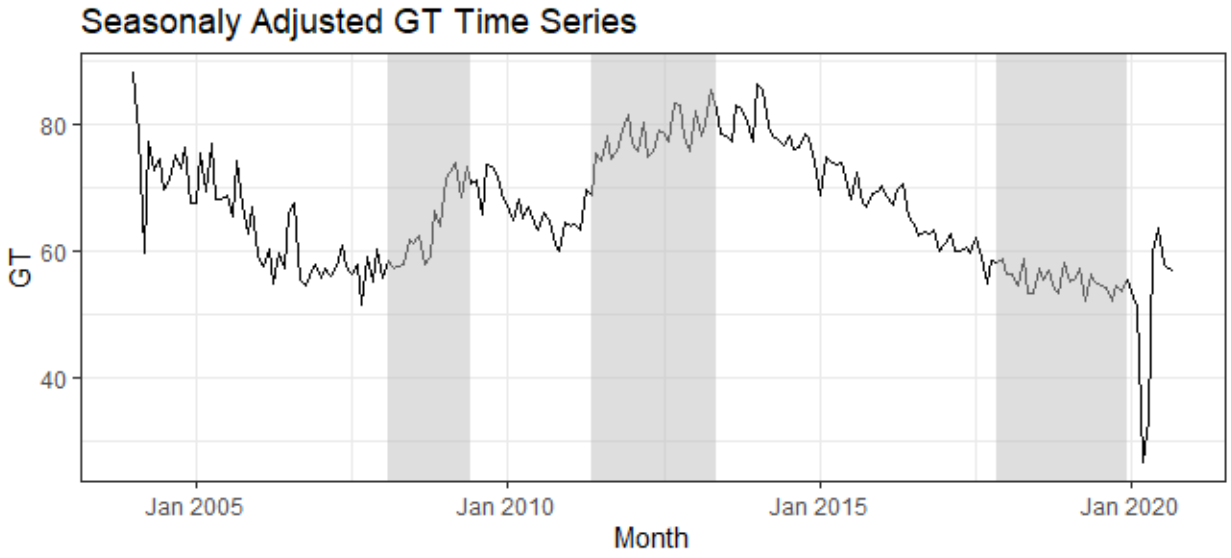
The series for YUR and GT were adjusted for seasonality to eliminate seasonal fluctuations. TRAMO-SEATS procedure (Hillmer & Tiao, 1982) is employed for seasonal adjustment.

Figure 1: Seasonally Adjusted Italy Youth Unemployment Rate (January 2004 to September 2020)



Source: ISTAT

*Figure 2: Seasonally Adjusted Google Trends ("*offerte di lavoro*") Series (January 2004 to September 2020)*



The figures above illustrate the seasonally adjusted time series of youth unemployment rate and google trends for "*offerte di lavoro*" keyword. We observe that the youth unemployment rate begins to rise during 2008 – the first recession depicted in the series. The increasing trend in the youth unemployment rate even after the recession until 2015 where the youth unemployment rate begins to diminish. The impact of the COVID-19 contracted industrial production by 30% in March 2020 and further decline of 19.1% in April 2020 due to lockdown measures (Colussi, 2020). The youth unemployment rate declined to an all time low after 2007 by April 2020; the sharp decline was due to the significant decrease in the labour force participation rate during the same period (Colussi, 2020). After April 2020, the youth unemployment rate momentarily increases. Similarly, the google trends data shows a similar relative to the youth unemployment rate. The correlation between the YUR and GT series is 44% and significant which corroborates the positive relationship between the two series.

Table 4: Summary Statistics of Seasonally Adjusted Time Series

	Youth Unemployment Rate	Google Trend
Minimum	18.03	26.78
1st Quintile	23.74	58
Median	29.27	65.72
Mean	30.29	66.29
3rd Quintile	36.65	74.42
Maximum	46.23	88.16

Methodology

In the original paper (Naccarato, et al., 2018), an ARIMA model and VAR model were estimated to verify GT data's relevance in forecasting the YUR. The VAR model combines the YUR and GT and uses cross-correlation of both series (Johansen, 1995). On the other hand, the ARIMA model was limited to only time-series of the youth unemployment rate as a benchmark model.

Our study will employ an ARIMAX model (Hamilton, 1994) by including GT as an exogenous variable, then compare the model performance to the VAR model and extend the model data to September 2020. The relevance of the internet search data in forecasting the unemployment rate is established (Naccarato, et al., 2018). Therefore, we can include it as an exogenous variable in the ARIMAX model.

First, we will test the stationarity of the two series using the Augmented Dickey-Fuller test (ADF). The Johansen test (Johansen, 1991) will be employed since the two series are to be studied together and hence it is necessary to test if they have common stochastic trends. The Johansen test involves two sequential tests. If the null of the first test is rejected, then we perform the second test. The conclusion from the test will determine if the series will be differenced.

Test one:

Null Hypothesis: Series are only integrated.

Alternate hypothesis: Series are integrated and linked by a cointegration relationship.

Test two:

Null Hypothesis: Series are cointegrated.

Alternate hypothesis: Series are stationary I(0).

The ARIMAX model firstly requires that the time series is stationary to avoid spurious regression. Transformation such as the first difference is required if the time series is non-stationary. Augmented Dickey-Fuller test (ADF) was performed to assess the stationarity of the series, including the exogenous variable. The autoregressive AR(p) orders and moving average *MA(q)* is determined by the ACF and PACF plots. Putting it together, we will get an *ARIMAX (p, d, q)* where d is the order of integration. The final ARIMAX model was selected based on the model with the lowest Akaike Information Criterion (AIC) (Akaike, 1974) and have stationary and white-noise residuals. The Ljung-Box test (Ljung & Box, 1978) was used to test for white-noise

ARIMAX model:

$$y_t = \beta X_t + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (1)$$

Where:

X_t is a covariate at time t, and β is its coefficient

The VAR model will have two equations. The first equation is an autoregressive equation for YUR and GT series added. The second equation will be an autoregressive equation for the GT component only.

VAR(p) model:

$$y_t = a + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t \qquad (2)$$

Where:

$Y_t = (y_{1t}, y_{2t})^T$: a (2×1) vector of YUR(1)and GT(2) variables

A: (2×1) vector of intercepts

$A_i = (i = 1, 2, \dots, P)$

E_t: (2×1) vector of the error term

The rolling estimation procedure was employed for the out-of-sample backtesting. With the rolling estimation, we divided our data into training and test set using an 80:20 ratio. The model is re-estimated with a fixed window to make out-of-sample forecasts. The rolling estimation is necessary because it eliminates past data that may result in biased parameter estimates.

Finally, to compare the estimated models' forecast ability, a rolling regression will be performed by setting the horizon to 1 month. The out-of-sample measures of error employed will be the MSE, RMSE, MAE, MAPE and SMAPE. The metrics are defined below.

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \; ; \qquad RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \; ; \qquad MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \; ;$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right| \; ; \qquad SMAPE = \frac{200}{n}\sum_{t=1}^{n} \frac{|e_t|}{|y_t| + |\hat{y}_t|}$$

$$Where; \; e_t = y_t - \hat{y}_t$$

A further study of the equal predictability of the models was performed. The Diebold-Mariano test (Diebold & Mariano, 1995) was used in this regard. The null hypothesis states that the models have equal predictive accuracy. The Diebold-Mariano test statistic (DM) is defined below.

Let $d_t = L(e^1_{t+h|t}) - L(e^2_{t+h|t})$ [2] be the sample loss difference.

$$DM = \frac{\bar{d}}{\sqrt{asymptotic\ variance\ of\ \bar{d}}}$$

[2] The sample loss difference can either be a squared error loss $\left(L(e_{t+h|t}) = (e_{t+h|t})^2\right)$ or absolute error loss $(L(e_{t+h|t}) = |e_{t+h|t}|)$. For this paper, we use the squared error loss

Where:

$$\bar{d} = \frac{1}{T}\sum_{t=t_0}^{T} d_t$$

Results and Discussion

Replication Results

It is necessary to study the nature of the time series we are working with. In our case, since we are studying the YUR and GT together, it is important to check whether they are cointegrated. The Johansen test (Johansen, 1991) is used to undertake this test because it allows us to know if the series are cointegrated and further determine the number of cointegration relationships present.

Before performing the cointegration test, we initially perform the unit root test to determine whether our series is non-stationary I(1). We perform the Augmented-Dickey Fuller test (Cheung & Lai, 1995) on both YUR and GT series (Table 5). The conclusion of the tests suggests that both the YUR and GT series are I(1) – we can therefore conduct the cointegration test.

Table 5: Unit root test of YUR and GT series

	Test statistics	Critical Value (95%)
YUR	0.398	-2.86
GT	-1.432	-2.86
ΔYUR	-16.793	-2.86
ΔGT	-2.74	-2.86

To perform the Johansen test (Johansen, 1991), we must first identify the lag length. (Naccarato, et al., 2018) uses the Akaike Information Criterion (AIC) (Akaike, 1974) to select the lag length. Using the same AIC recommendation, our data suggests a lag length of 4[3]. Although (Naccarato,

[3] The AIC for lag length 4 is 3.7524. for lag 8 the AIC is 3.7576 which is the second lowest after lag 4

et al., 2018) does not state the lag length used for the test, we noted that a lag length of 8 was used for the VAR modelling. We included the Johansen test of lags 4 and 8 in our study (Table 6).

Table 6: Johansen test for YUR and GT series

Null	Lag	Johansen Statistic (JS)	95% Confidence Interval
r = 0	4	15.32	19.96
r = 0	8	18.1	19.96

From table 3 above, since the Johansen statistics is smaller than the 95% critical value (19.96), we don't reject the null for both lag lengths (4 and 8). We, therefore, conclude that the YUR and GT series are non-stationary but not cointegrated. Since we reject the null hypothesis for r = 0, we do not proceed to r = 1. We can further conclude from the test that the relationship between YUR and GT cannot be assumed spurious. Hence, the choice of keyword "*offerte di lavoro*" is considered suitable.

To replicate the parameter estimates, we firstly fit the VAR(8) model then restrict the model to report only significant variables (Table 7). We noted that (Naccarato, et al., 2018) reported only significant coefficients in the paper. Using the significant variable restriction method, we realize that the variables reported in (Table 7) are different from the variables in the original paper (Table 1). We further fit the VAR(8) model to achieve results from the original paper using a manual restriction approach (Table 8). We restricted the variables according to the original paper.

Table 7: VAR model using significant[4] variable restriction

Parameter	Estimate	Std. Error	t-Value	p-Value
Estimation results for equation YUR:				
ΔYUR.l1	-0.53403	0.0855	-6.246	6.50E-09
ΔYUR.l2	-0.35653	0.0865	-4.122	6.92E-05
ΔGT.l5	0.16482	0.03963	4.159	6.00E-05
ΔGT.l6	0.11637	0.04388	2.652	0.00907
ΔGT.l7	0.09135	0.03745	2.439	0.01617

[4] A 5% significance level was used to restrict the variables. If the absolute value of t-statistics was less that 1.96, the parameter is rejected

Estimation results for equation GT:				
ΔGT.l1	-0.3802	0.0806	-4.717	6.34E-06
ΔYUR.l3	0.423	0.1743	2.427	0.0167

Table 8: VAR model using original paper restrictions

Parameter	Estimate	Std. Error	t-Value	p-Value
Estimation results for equation YUR:				
ΔYUR.l1	-0.45707	0.08348	-5.475	2.38E-07
ΔYUR.l2	-0.27063	0.08491	-3.187	0.001824
ΔYUR.l5	-0.14706	0.08076	-1.821	0.071064
ΔGT.l5	0.13312	0.03675	3.622	0.000427
Estimation results for equation GT:				
ΔGT.l1	-0.38695	0.08238	-4.697	6.88E-06
ΔGT.l8	0.03415	0.07125	0.479	0.633

From the results above, it is evident that we were unable to replicate the exact estimates from the original paper. The discrepancies can be attributed to the vaguely erratic nature of the Google trend (Leinweber, 2013). We further examined the VAR(4) model as suggested by AIC (Table 9). We impose the significant variable restriction on the VAR(4) model.

Table 9: VAR(4) Model

Parameter	Estimate	Std. Error	t-Value	p-Value
Estimation results for equation YUR:				
ΔYUR.l1	-0.47069	0.0861	-5.467	2.31E-07
ΔYUR.l2	-0.27263	0.08699	-3.134	0.00214
Estimation results for equation GT:				
ΔGT.l1	-0.46861	0.08633	-5.428	2.82E-07
ΔGT.l8	-0.22944	0.09196	-2.495	0.0139
ΔYUR.l3	0.41366	0.17168	2.409	0.0174
ΔGT.l3	-0.16266	0.07965	-2.042	0.0432

From table 8 above, we observe that using lag 4 eliminates the effects that the GT series will have on the YUR series. Since the purpose of the paper was to show the impact Google trends data can

have on forecasting the youth unemployment rate; it will not be prudent to use lag 4. We further observe that lag 8 had the second-lowest AIC, which justifies the selection of VAR(8) over VAR(4).

New Results

A new VAR model was developed as an extension of (Naccarato, et al., 2018) VAR model. We included data up to September 2020. An ARIMAX model was developed to challenge the predictability of the VAR model.

Similarly, we initially perform the Augmented-Dickey Fuller test (Cheung & Lai, 1995) on both YUR and GT series (Table 10). The conclusion of the tests suggests that both the YUR and GT series are I(1) – we can therefore conduct the cointegration test.

Table 10: Unit root test of YUR and GT complete series

	Test Statistics	Critical Value (95%)
YUR	-1.216	-2.86
GT	-1.451	-2.86
ΔYUR	-3.283	-2.86
ΔGT	-3.622	-2.86

Also, we perform the Johansen test (Johansen, 1991), with a lag length of 4 as suggested by the Akaike Information Criterion (AIC) (Akaike, 1974). From table 10 below, the Johansen statistics (10.73) is smaller than the critical value (19.96); we do not reject the null of r = 0. We, therefore, conclude that the YUR and GT series are non-stationary but not cointegrated. Since we reject the null hypothesis for r = 0, we do not proceed to r = 1.

Table 11: Johansen test for YUR and GT complete series

	Lag	Johansen Statistic (JS)	95% Confidence Interval
r <= 1	4	1.87	19.96
r = 0	4	10.73	19.96

Table 12: VAR(4) Model for complete series

Parameter	Estimate	Std. Error	t-Value	p-Value
Estimation results for equation YUR:				
ΔYUR.l1	-0.51736	0.07198	-7.187	1.43E-11
ΔGT.l1	0.09815	0.03299	2.976	0.0033
ΔYUR.l2	-0.27996	0.07218	-3.879	0.000144
ΔGT.l2	0.08728	0.03205	2.723	0.007063
Estimation results for equation GT:				
ΔGT.l1	-0.35022	0.07096	-4.935	1.73E-06
ΔGT.l2	-0.36216	0.07271	-4.981	1.41E-06
ΔGT.l3	-0.25587	0.06979	-3.666	0.000318
ΔGT.l4	-0.13697	0.06606	-2.073	0.039473

ARIMAX (0, 1, 2) was selected as the model to challenge the VAR model. The selection of the order follows the ACF and PACF plots as well as the AIC. We further tested if the residuals are white noise with the Ljung-Box test. The p-value (0.07) from the Ljung-Box test is greater than a 5% significant level. We, therefore, fail to reject the null hypothesis that the residuals are not correlated. Table 13 presents the estimates of the ARIMAX model.

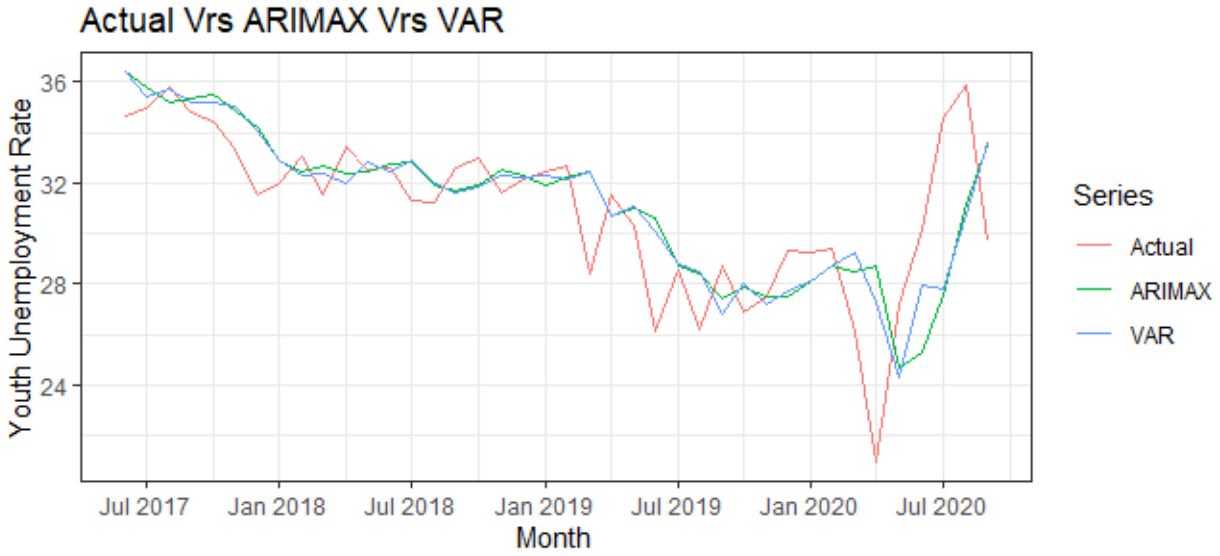
Table 13: ARIMAX Model for complete series

Parameter	Estimate	Std. Error	t-Value	p-Value
MA1	0.4782	0.0659	7.2564	<.00001
MA2	-0.5218	0.0646	-8.0774	<.00001
ΔGT	0.5005	0.0041	122.0732	<.00001

Rolling estimation and model performance

We performed an out-of-time sample backtesting for both models to assess the best performing model between the VAR and ARIMAX. The backtesting was done based on training and test set of 80% and 20%, respectively. Using the rolling estimation method, we forecast 1 horizon for each model. The figure below illustrates the performance of both models. From an initial look, we can see that the VAR and ARIMAX models have insignificantly different.

Figure 3: Forecast of the test set (Actual vs. ARIMAX vs. VAR)



Various error metrics were employed to assess the performance of each model. The errors from the out-of-time backtesting are computed in Table 14. We observe that the VAR model outperformed the ARIMAX model in all categories.

The Diebold-Mariano test (Diebold & Mariano, 1995) was further performed to evaluate the equal predictability of both models in forecasting the youth unemployment rate. From the test, the p-value (0.3238) is greater than 5%. We, therefore, fail to reject the null hypothesis that both models have equal predictive power. In conclusion, the VAR and ARIMAX models have an equal predictive ability to forecast the youth unemployment rate in Italy.

Table 14: Out-of-time backtesting

	MSE	RMSE	MAE	MAPE	SMAPE
ARIMAX	6.467	2.543	1.769	6.07%	5.95%
VAR	5.543	2.354	1.684	5.76%	5.66%

Conclusion

The paper's objective was to first replicate some results from (Naccarato, et al., 2018) and secondly challenge the VAR model with an ARIMAX model.

Replicating the results from the original paper was unsuccessful. A significant reason for the failure to get precisely the exact estimates is attributed to the vaguely erratic nature of the google trends data (Leinweber, 2013). We also observed that the authors used an unexplained scaling for the youth unemployment rate. From table 1, we noted that the actual values of the youth unemployment rate as used by the author were greater than 100. The actual youth unemployment rate values are expected to lie within 0% to 100%.

By including the new data, we challenged the VAR model with an ARIMAX model. The VAR model outperformed the ARIMAX model in all the error metrics. However, by testing for the equal predictive accuracy of the models, we can conclude that both have equal ability to predict the youth unemployment rate in Italy.

References

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6), pp. 716-723.

Alessandrini, S. & Sperati, S., 2017. Characterization of forecast errors and benchmarking of renewable energy forecasts. In: G. Kariniotakis, ed. *Renewable Energy Forecasting.* s.l.:Woodhead Publishing, pp. 235-256.

Andrews, B. H., Dean, M. D., Swain, R. & Cole, C., 2013. *Building ARIMA and ARIMAX Models for Predicting Long-Term Disability Benefit Application Rates in the Public/Private Sectors,* Maine: Society of Actuaries.

Braaksma, B. & Kees, Z., 2017. Big Data in Official Statistics. In: . T. Prodromou, ed. *Data Visualization and Statistical Literacy for Open and Big Data.* s.l.:IGI Global, pp. 274-296.

Cheung, Y. W. & Lai, K. S., 1995. Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics,* 13(3), pp. 277 - 280.

Colussi, T., 2020. *Crisis Response Monitoring.* [Online]
Available at: https://covid-19.iza.org/crisis-monitor/italy/
[Accessed 22 April 2021].

Diebold, F. X. & Mariano, R. S., 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics, American Statistical Association,* 13(3), pp. 253-263.

Ettredge, M., Gerdes, J. & Karuga, G., 2005. Using Web-based search data to predict. *Mag. Commun. ACM,* 48(11), pp. 87-92.

Francesco, D. & Juri, M., 2017. The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting,* 33(4), pp. 801-816.

Google, n.d. *Google Trends Help.* [Online]
Available at: https://support.google.com/trends/answer/4365533?hl=de&ref_topic=6248052
[Accessed 25 March 2021].

Hamilton, J. D., 1994. *Time Series Analysis.* Princeton: University Press.

Harvey, D. I., Leybourne, S. J. & Newbold, P., 1998. Tests for Forecast Encompassing. *Journal of Business & Economic Statistics,* 16(2), pp. 254-259.

Hillmer, S. C. & Tiao, G. C., 1982. An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association,* 77(377), pp. 63-70.

Italian National Statistical Institute (ISTAT), 2016. *Italian labour force survey.* [Online]
Available at: http://www.istat.it/it/archivio/8263

Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica,* 59(6), pp. 1551-1580.

Johansen, S., 1995. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Leinweber, D., 2013. *The Problem With The Google Trends Market Prediction Data.* [Online] Available at: https://www.forbes.com/sites/davidleinweber/2013/07/19/google-trends-market-prediction-still-in-the-shop-but-calls-market-flat/?sh=305d82de7665 [Accessed 25 March 2021].

Ljung, G. M. & Box, G. E. P., 1978. On a measure of lack of fit in time series models. *Biometrika,* 65(2), pp. 297-303.

Naccarato, A., Falorsi, S., Loriga, S. & Pierini, A., 2018. Combining official and Google Trends data to forecast the Italian youth. *Technological Forecasting & Social Change,* pp. 114-122.

Nikolaos, A. & Klaus, Z. F., 2015. The internet as a data source for advancement in social sciences. *International Journal of Manpower,* 36(1).

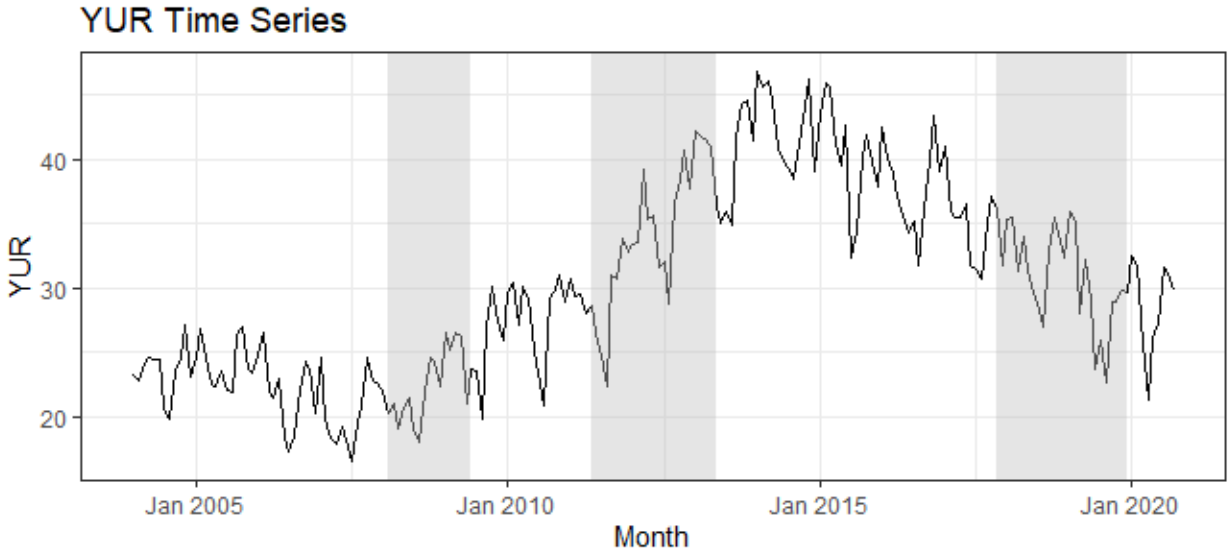
Zivot, E. & Wang, J. eds., 2006. Vector Autoregressive Models for Multivariate Time Series. In: *Modeling Financial Time Series with S-PLUS.* New York: Springer, pp. 383-429.

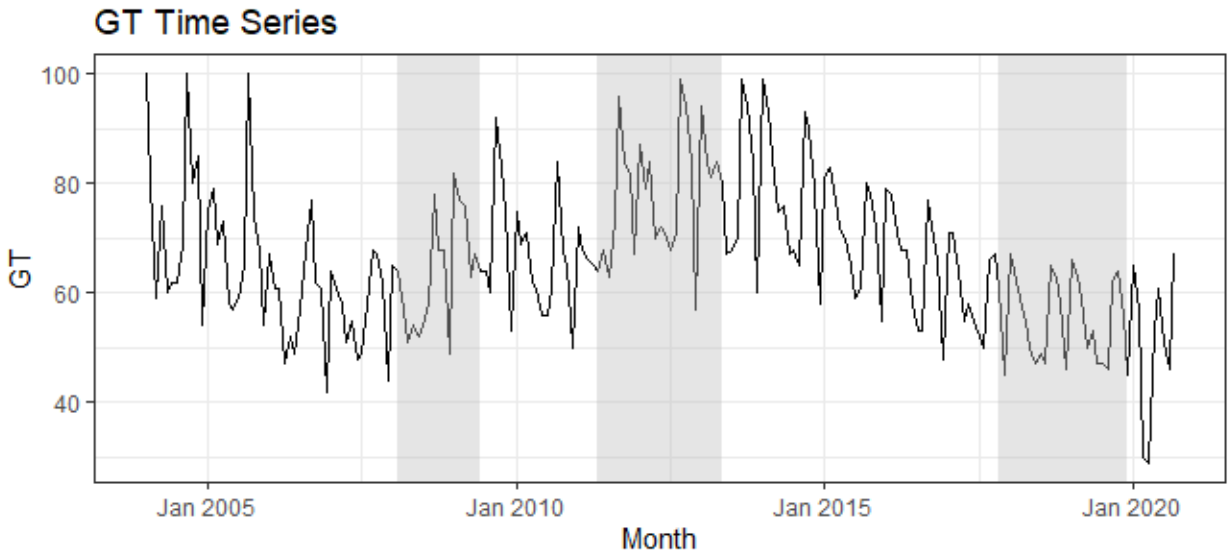
Appendix

Data Appendix

Variable	Abbreviation	Source
Italy youth unemployment rate	YUR	Italian National Statistical Institute (ISTAT)
Google trend series (job offer)	GT	Google

Plots for YUR and GT up to September 2020





Lag Length for data up to March 2015

Lag	1	2	3	4	5	6	7	8	9	10
AIC(n)	3.995	3.825	3.764	3.752	3.781	3.759	3.767	3.758	3.786	3.828
HQ(n)	4.050	3.917	3.892	3.918	3.984	3.998	4.043	4.070	4.135	4.214
SC(n)	4.131	4.052	4.081	4.160	4.279	4.347	4.446	4.527	4.645	4.778
FPE(n)	54.314	45.852	43.120	42.644	43.916	42.975	43.362	42.989	44.275	46.255

Lag Length for data up to September 2020

Lag	1	2	3	4	5	6	7	8	9	10
AIC(n)	4.484	4.329	4.222	4.199	4.206	4.228	4.246	4.259	4.285	4.306
HQ(n)	4.525	4.398	4.318	4.323	4.358	4.407	4.453	4.494	4.547	4.596
SC(n)	4.586	4.499	4.460	4.506	4.581	4.671	4.757	4.838	4.932	5.022
FPE(n)	88.552	75.867	68.145	66.631	67.120	68.613	69.867	70.821	72.700	74.308

R Code

```
## ----setup, include=FALSE----------------------------------------------------------------------

knitr::opts_chunk$set(echo=TRUE, cache=TRUE, warning=FALSE, message=FALSE)

library(tidyverse)

library(gridExtra)

library(fpp2)

library(sandwich)

library(lmtest)

library(xts)


##-----------------------------------------------------------------------------

#Set Working Directory

setwd("C:/Users/ReggyRyt/OneDrive - ualberta.ca/ECON 509 Time Series/Paper/Italy Unemployment")


##-----------------------------------------------------------------------------

#Import Data
```

```
YUR_Data <- read.csv("YUR Time Series.csv", sep=",",  header=TRUE)
#YUR_Data$Date <- seq(from = as.Date("2004-01-01"), to = as.Date("2020-09-30"), by = 'month')
GT_Data <- read.csv("multiTimeline (2).csv", sep=",",  header=TRUE, skip = 2)
names(GT_Data)[2] <- "GT"

##------------------------------------------------------------------------
# load Italy recession dates
Italy <-read.table("ITRES.csv", sep=",", header=TRUE)
ItRec <- ts(Italy[2], start=2004, frequency=12)

# find peaks and troughs
tmp0 <- diff(ItRec)
peak <- time(tmp0)[which(tmp0=="1")-1]
trough <- time(tmp0)[which(tmp0=="-1")]
# set as data frame
recessions.df <- data.frame(Peak=peak,Trough=trough)

##------------------------------------------------------------------------
##Preparing And Plotting Data
df.YUR <- ts(data.frame(YUR_Data),start=c(2004, 1), frequency=12)
df.GT <- ts(data.frame(GT_Data),start=c(2004, 1), end=c(2020, 9), frequency=12)

Data <- ts(cbind(df.YUR[,2], df.GT[,2]), start=c(2004, 1), frequency=12)
colnames(Data) <- c("YUR","GT")

p1 <- autoplot(Data[,2]) + xlab("Year") + ylab("GT") + theme_bw()
p2 <- autoplot(Data[,1]) + xlab("Year") + ylab("YUR") + theme_bw()
grid.arrange(p1,p2,nrow=2)
```

```
autoplot(as.xts(Data[,1])) + xlab("Month") + ylab("YUR") + theme_bw() +
  ggtitle("YUR Time Series") +
  geom_rect(data=recessions.df,aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
            fill='grey', alpha=0.4)

autoplot(as.xts(Data[,2])) + xlab("Month") + ylab("GT") + theme_bw() +
  ggtitle("GT Time Series") +
  geom_rect(data=recessions.df,aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
            fill='grey', alpha=0.4)

##-----------------------------------------------------------------------------
#TRAMO-SEATS Seasonal Adjustment
library(RJDemetra)
TS.YUR <- tramoseats(Data[,1])
TS.YUR.SA <- ts(data.frame(TS.YUR$final$series),start=c(2004, 1),frequency=12)

TS.GT <- tramoseats(Data[,2])
TS.GT.SA <- ts(data.frame(TS.GT$final$series),start=c(2004, 1), frequency=12)

df.TS <- ts(cbind(TS.YUR.SA[,"sa"],TS.GT.SA[,"sa"]),
            start=c(2004, 1), frequency=12)
#Date <- seq(from = as.Date("2004-01-01"), to = as.Date("2020-09-30"), by = 'month')
#df.TS1 <- cbind.data.frame(Date,df.TS)
colnames(df.TS) <- c("YUR","GT")

autoplot(df.TS) + xlab("Year") + ylab("") + theme_bw() +
  ggtitle("Seasonaly Adjusted Data")

p3 <- autoplot(df.TS[,1]) + xlab("Year") + ylab("YUR") + theme_bw()
```

```
p4 <- autoplot(df.TS[,2]) + xlab("Year") + ylab("GT") + theme_bw()
grid.arrange(p3,p4, nrow = 2)

autoplot(as.xts(df.TS[,1])) + xlab("Month") + ylab("YUR") + theme_bw() +
  ggtitle("Seasonaly Adjusted YUR Time Series") +
  geom_rect(data=recessions.df,aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
            fill='grey', alpha=0.5)

autoplot(as.xts(df.TS[,2])) + xlab("Month") + ylab("GT") + theme_bw() +
  ggtitle("Seasonaly Adjusted GT Time Series") +
  geom_rect(data=recessions.df,aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
            fill='grey', alpha=0.5)
##-----------------------------------------------------------------------------
library("ggpubr")
#Correlation
cor.test(Data[,1],Data[,2])
cor.test(df.TS[,1],df.TS[,2])

##-----------------------------------------------------------------------------
#Summary Statistics
summary(df.TS)
summary(Data)
##-----------------------------------------------------------------------------
## Replication
df_replication <- window(df.TS, end=c(2015, 3))
autoplot(df_replication) + xlab("Year") + ylab("") + theme_bw() +
  ggtitle("Plot of Replicated Time Series")

##Unit Root Test
```

```
source("urtests.r")
testYUR <- ur.test(diff(df_replication[,1]), trend="c", method="adf.ols", penalty="MAIC", kmax=10)
print.ur.test(testYUR)

testGT <- ur.test(diff(df_replication[,2]), trend="c", method="adf.ols", penalty="MAIC", kmax=10)
print.ur.test(testGT)

##Johansen Test
library(urca)
library(vars)
lag <- VARselect(df_replication, type="const", lag.max=10 )
laglenght <- data.frame(lag$criteria)
colnames(laglenght) <- c("1","2", "3", "4", "5", "6", "7", "8", "9", "10")
write.csv(laglenght,"laglenght.csv")

jotest=ca.jo(df_replication, type="trace", K=8, ecdet="const", spec="longrun")
summary(jotest)

fitvar1 <- VAR(diff(df_replication), p=8, type="none")
var <- restrict(fitvar1, method='ser', thresh = 2)
summary(var)

matrix <- matrix(c(1,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,
                   0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1),
                 nrow=2 ,ncol=16, byrow=TRUE)
var1 <- restrict(fitvar1, method="man", resmat = matrix )
summary(var1)

##-------------------------------------------------------------------------
```

```
##New Data

##Unit Root Test
testYUR_ <- ur.test(df.TS[,1], trend="c", method="adf.ols", penalty="MAIC", kmax=10)
print.ur.test(testYUR_)

testGT_ <- ur.test(df.TS[,2], trend="c", method="adf.ols", penalty="MAIC", kmax=10)
print.ur.test(testGT_)

##Johansen Test
lag_ <- VARselect(df.TS, type="const", lag.max=10 )
lag_
laglenght_ <- data.frame(lag_$criteria)
colnames(laglenght_) <- c("1","2", "3", "4", "5", "6", "7", "8", "9", "10")
write.csv(laglenght_,"laglenght_.csv")

jotest_=ca.jo(df.TS, type="trace", K=4, ecdet="const", spec="transitory")
summary(jotest_)

fitvar_ <- VAR(diff(df.TS), p=4, type="none")
var_ <- restrict(fitvar_, method='ser', thresh = 2)
summary(var_)

##------------------------------------------------------------------------------
##ARIMAX
p1 <- ggAcf(df.TS[,1]) + ggtitle("")
p2 <- ggPacf(df.TS[,1]) + ggtitle("")
grid.arrange(p1,p2,ncol=2)
```

```
p3 <- ggAcf(diff(df.TS[,1])) + ggtitle("") + theme_bw()
p4 <- ggPacf(diff(df.TS[,1])) + ggtitle("") + theme_bw()
grid.arrange(p3,p4,nrow=2)

ARIMAX <- Arima(df.TS[,1][-1], order=c(0,1,2), xreg = diff(df.TS[,1]))
summary(ARIMAX)
checkresiduals(ARIMAX, lag=10)
##-----------------------------------------------------------------------------
## Backtesting for ARIMAX and VAR
# train:test = 80:20
#end of training set 2017-05
n.end <- 2017+4/12

#test set: 2017-06 - 2020-09
#40 observations in test set
predVAR <- matrix(rep(NA,40*4),40,4)
colnames(predVAR) <- c("Actual","YUR.l1","Forecast","Level")

predARIMAX <- matrix(rep(NA,40*2),40,2)
colnames(predARIMAX) <- c("Actual","Forecast")

##Backtesting for VAR
#Parameter Restriction
mat <- matrix(c(1,1,1,1,0,0,0,0,
                1,0,1,0,1,0,1,0),
              nrow=2 ,ncol=8, byrow=TRUE)

#loop for rolling estimation
for(i in 1:40){
```

```
 start <- 2004+(i-1)*1/12
 end <- n.end+(i-1)*1/12
 train_set <- window(df.TS, start, end)
 predVAR[i,"Actual"] <- window(df.TS[,'YUR'], end+1/12, end+1/12)
 predVAR[i,"Forecast"] <-forecast(
   restrict(VAR(diff(train_set), p=4, type="none"),
          method="man", resmat = mat),h=1)$forecast$YUR$mean
 #Conditional forecast
 predVAR[i,"YUR.l1"] <- window(df.TS[,'YUR'], end, end)
 predVAR[i,"Level"] <- predVAR[i,"YUR.l1"] + predVAR[i,"Forecast"]
}

#Unconditional forecast
# for(i in 1:18){
#   for(j in 2:18){
#     predVAR[,"YUR.l1"][1] <- window(df.TS[,'YUR'], start=c(2019,2), end=c(2019,2))
#     predVAR[,"Level"][i] <- predVAR[,"YUR.l1"][i] + predVAR[,"Forecast"][i]
#     predVAR[,"YUR.l1"][j] <- predVAR[,"Level"][j-1]
#   }}


##Backtesting for ARIMAX
#loop for rolling estimation
for(i in 1:40){
 start <- 2004+(i-1)*1/12
 end <- n.end+(i-1)*1/12
 train_YUR <- window(df.TS[,'YUR'], start, end)
 train_GT <- window(df.TS[,'GT'], start, end)
 predARIMAX[i,"Actual"] <- window(df.TS[,'YUR'], end+1/12, end+1/12)
```

```
  predARIMAX[i,"Forecast"] <-forecast(Arima(train_YUR[-1], order=c(0,1,2),
                                    xreg = diff(train_GT)),
                h=1, xreg=diff(window(df.TS[,'GT'], end, end+1/12)))$mean
}

##Plot of Backtesting
fcast.all <- ts(cbind(predARIMAX,predVAR[,'Level']), start=c(2017,6), frequency = 12)
colnames(fcast.all) <- c("Actual","ARIMAX", "VAR")
#write.csv(fcast.all, "Forecast.csv")

autoplot(as.xts(fcast.all), facets=NULL) + xlab("Month") +
  ylab("Youth Unemployment Rate") + theme_bw() +
  ggtitle("Actual Vrs ARIMAX Vrs VAR")

##Compute OOT Error Metric
OOT <- matrix(rep(NA,2*5),2,5)
rownames(OOT) <- c("ARIMAX","VAR")
colnames(OOT) <- c("MSE","RMSE","MAE","MAPE","SMAPE")

library(Metrics)
library(scales)
for(i in 1:2){
  OOT[i,"MSE"] <- round(mse(fcast.all[,"Actual"],fcast.all[,1+i]),3)
  OOT[i,"RMSE"] <- round(rmse(fcast.all[,"Actual"],fcast.all[,1+i]),3)
  OOT[i,"MAE"] <- round(mae(fcast.all[,"Actual"],fcast.all[,1+i]),3)
  OOT[i,"MAPE"] <- percent(mape(fcast.all[,"Actual"],fcast.all[,1+i]), accuracy = 0.001)
  OOT[i,"SMAPE"] <- percent(smape(fcast.all[,"Actual"],fcast.all[,1+i]), accuracy = 0.001)
}
#write.csv(OOT, "OOT.csv")
```

```
# Compare Predictability of ARIMAX vs VAR
e1 <- fcast.all[,"Actual"]-fcast.all[,"ARIMAX"]
e2 <- fcast.all[,"Actual"]-fcast.all[,"VAR"]
# compute Diebold-Mariano statistic
dm.test(e1, e2, h=1, power=2)
```