Mathematics and Statistics 11(2): 294-299, 2023 http://www.hrpub.org
DOI: 10.13189/ms.2023.110207

Comparing The Forecasting Accuracy Metrics Of Support Vector Regression and ARIMA Algorithms For Non-Stationary Time Process

Youness Jouilil*, **Driss Mentagui**

Department of Mathematics, Faculty of Sciences, Ibn Tofail University of Kenitra, Morocco

Received August 17, 2022; Revised January 20, 2023; Accepted February 15, 2023

Cite This Paper in the following Citation Styles

***(a)**: [1] Youness Jouilil, Driss Mentagui , "Comparing The Forecasting Accuracy Metrics Of Support Vector Regression and ARIMA Algorithms For Non-Stationary Time Process," Mathematics and Statistics, Vol.11, No.2, pp. 294-299, 2023. DOI: 10.13189/ms.2023.110207*

***(b)**: Youness Jouilil, Driss Mentagui, (2023). Comparing The Forecasting Accuracy Metrics Of Support Vector Regression and ARIMA Algorithms For Non-Stationary Time Process. Mathematics and Statistics, 11(2), 294-299. DOI: 10.13189/ms.2023.110207*

Abstract Univariate time series forecasting is a crucial machine learning issue across many fields notably sentiment analysis, economy, medicine, agriculture, and finance. In this working paper, we tackled comparing the Support Vector Regression (SVR) to the traditional Autoregressive Integrated Moving Average (ARIMA) algorithms in terms of forecasting through a real case study. In fact, the data set used in this investigation has been extracted from the World Bank. The target time series is the American Foreign direct investment, net outflows (% of GDP) which includes the data for 50 years from 1972 to 2021. For analytical and comparison purposes, all the compilations have been done using the R programming language for Windows 10. The statistical findings revealed that, in short-term prediction, the forecast accuracy of both algorithms reduces in terms of error accuracy, significantly. Comparatively, the analysis conducted in this investigation demonstrates that the machine learning algorithms, especially the SVM one perform better than the ARIMA in short-term forecasting since its accuracy functions are the lowest. Thus, we highly recommend future research to compare the advanced machine learning algorithms especially the recurrent neural network algorithms with the classical algorithms, especially with the ARIMA approach in order to choose the best algorithm in terms of results and predictive performance.

Keywords SVR, ARIMA, Time Series, Forecasting

1 Introduction

Nowadays, time series forecasting is widely employed in many fields especially finance, economy, and medicine. In this working paper, we try to compare the Support Vector Regression (SVR) to the traditional Autoregressive Integrated Moving Average Model (ARIMA) models in terms of forecasting.

The remainder of this manuscript will be structured as follows. In the first part, we present briefly the ARIMA model. In the second part, we expose the SVR algorithm. In the third part, we compare the SVR to the traditional ARIMA models in terms of forecasting through a real case study. In the last one, we discuss and conclude.

2 Methodology

2.1 ARIMA model

As known, the Autoregressive Integrated Moving Average (ARIMA) Model is a generalized algorithm of Autoregressive Moving Average one.
A given process $X = \{x_t \ , \ t \in \mathbf{Z}\}$ is an ARIMA(p,d,q) if it can be expressed as follows :

$$\Phi_p(B)(1-B)^d x_t = \Theta_q(B)\eta_t$$

$$\nabla^d x_t = (1-B)^d x_t \quad ; \quad \forall t \in \mathbf{Z}^{*+}$$

$$\eta_t \sim WN(0, \sigma^2)$$

where WN indicates the traditional white noise, B the backward shift operator, Φ_p represents the p-coefficients of the moving average model, and Θ_q represents the coefficients of the autoregressive model.

2.2 SVR algorithm

SVR is a famous learning technique which has been widely used especially in regression and classification. It has been invented by the researchers Vapnik and Boser in 1992. Nowadays, it is used in finance, economics and medicine filed.

The idea of SVR is to find a function that can predict future values accurately. The generic SVR can be expressed as follows :

$$f(x_t) = w\phi(x_t) + b$$

where t $\in \{1, 2 \ldots, n\}$and

$$Min \ \frac{1}{2}\,||w||^2 + C\sum_{t=1}^{n}(\zeta_t + \zeta_t^*)$$

Subject to :

$$y_t - f(x_t) \leq \epsilon_t + \zeta_t$$
$$f(x_t) - y_t \leq \epsilon_t + \zeta_t^*$$
$$C, \zeta_t, \zeta_t^* \geq 0\,;\ t \in \{1, 2 \ldots, n\}$$

2.3 Accuracy Metrics

To compare the performance of ARIMA and SVR models, we use the following indicators :

$$RMSE = \sqrt{\frac{\sum(x_t - \hat{x_t})^2}{p}}$$

where :

- x_t indicates actual observed values;
- $\hat{x_t}$ indicates the forecast values;
- p indicates the number of target output data.

$$MAPE = \frac{1}{p}\frac{\sum|x_t - \hat{x_t}|}{|x_t|}$$

where MAPE indicates the Mean Absolute Percentage Error. We have also :

$$MAE = \frac{\sum|x_t - \hat{x_t}|}{p}$$

where MAE indicates the Mean Absolute Error.

3 Experimental Results

3.1 Data source and software

Our data set has been extracted from the World Bank. The target time series is the American Foreign direct investment, net outflows (% of GDP) which includes the data for 50 years from 1972 to 2021. For analytical and comparison purposes, all our compilations have been done using the R programming language for Windows 10.

3.2 Estimation

Description of data

In this paragraph, we expose the main characteristic of our data set. See Figure 1 and Table 1 for more details.

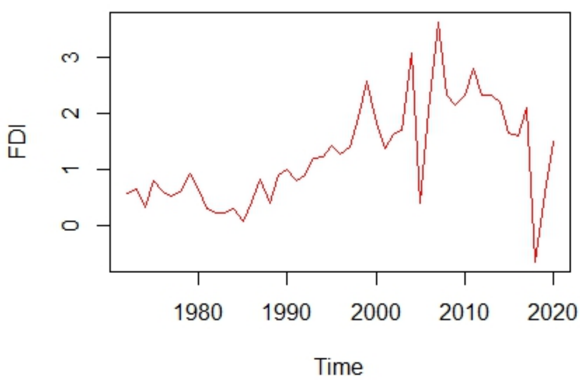

Figure 1. Time Plot of American FDI (x)

Table 1. Descriptive statistics

	Min	1. Q	Median	Mean	3.Q	Max
FDI	-0.633	0.573	1.207	1.269	1.928	3.619
Year	1972	1984	1996	1996	2008	2021

Data preprocessing

First of all, we have split our target time series into two species train and test regions using the window function. This step is very crucial, especially in time series since it allows us to assess the performance of our eventual machine learning algorithms. The figure below (Figure 2) exposes the Training and Testing Sets.

Code 1: R code for splitting the data set.

```
trainData = window(economics$FDI, start=min(
    economics$Year), end=2010)
testData = window(economics$FDI, start=2011, end
    =max(economics$Year))
# Visualize data and training/testing regions
economics %>%
  ggplot(aes(x = economics$Year, y =
    economics$FDI)) +
  geom_rect(xmin = 2011,
            xmax = 2021,
            ymin = -1, ymax = 10,
            fill = "yellow", alpha = 0.01) +
  annotate("text", x = 1990, y = 4,
           color = "black", label = "Train
    Region") +
  annotate("text", x = 2016, y = 7,
           color = "black", label = "Test Region
    ") +
  geom_point(alpha = 0.5, color = "red") +
  labs(title = "Visualize the TS and training/
    testing regions", x = "economics$Year")
```

Source: Authors' manipulations under R programming language.

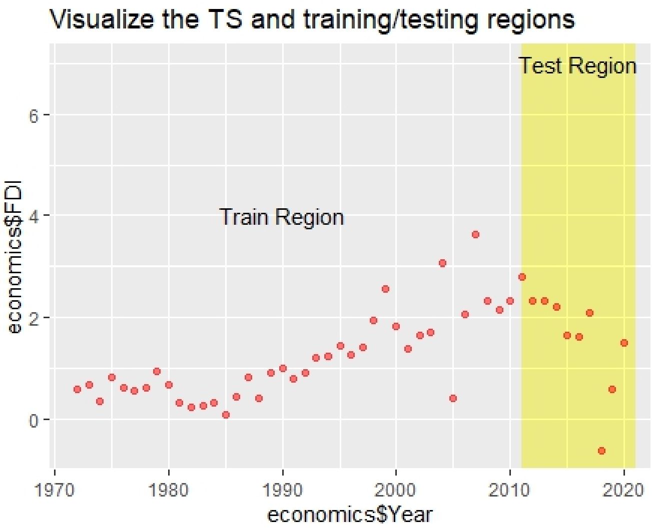

Figure 2. Data splitting into a training set and test set.

Table 2. Augmented Dickey-Fuller Test

Dickey-Fuller		Lag order	p-value
d=0	-2.1228	3	0.5253
d=1	-5.6062	3	0.001

According to the visualization graphs (Figure 1, Figure 3 and Figure 4), it appears that our series is not stationary. This result could be approved statistically by computing the Augmented Dickey-Fuller Test (ADF).

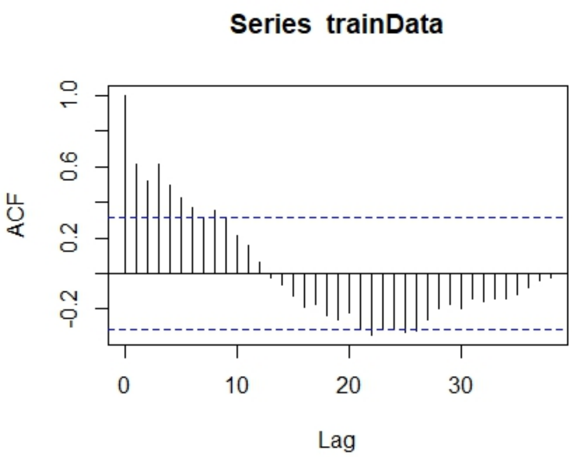

Figure 3. ACF of FDI

Table 2 shows that after the first difference (d=1), we confirm that it is now indeed stationary. Hence, we argue statistically that result by computing the ADF test. In fact, the associated p-value is less than 0.05, therefore we approve that

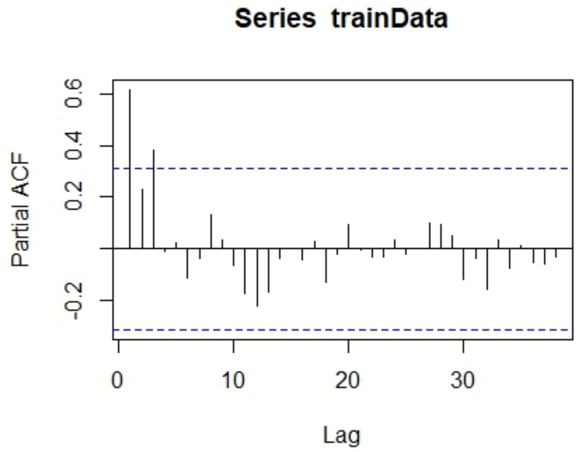

Figure 4. PACF of FDI

the data set is stationary at the first level (Figure 5).

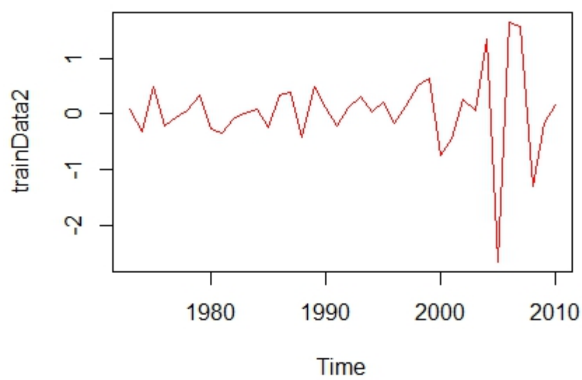

Figure 5. Time Plot of ∇x

After making sure that our TS is stationary, we move to the process of modeling using the Box-Jenkins approach. Indeed, to determine the orders p and q, we can be based on ACF and PACF plots. Graphically, it appears that p=2 and q = 1 (See Figure 8 and Figure 9). We can confirm those results by using auto.arima in R, which is based on BIC criterion to choose the best-fitted model. Outputs show that ARIMA(0,1,1) is the best model for our target process [1].

After understanding our univariate time series, we can move to modeling and forecasting.

Table 3. Estimation of ARIMA model

	term	estimate	std.error
1	ma1	-0.8037	0.1098
2	drift	0.0528	0.0196
log likelihood		-31.02	
AIC		68.04	
AICc		68.74	
BIC		72.95	

Hence, this is an ARIMA(0,1,1) model (Table 3) which can be written as :

$$\nabla y_t = \underset{(0.0196)}{0.0529} - \underset{(0.1098)}{0.8037}\eta_{t-1} + \eta_t \ ; \ \eta_t \sim WN(0, \sigma^2)$$

where ∇ is white noise with a standard deviation of $\sqrt{0.3077} = 0.5547071$ and brackets refer to standard errors. Thus, our model can be re-written as follows :

$$y_t = y_{t-1} + \underset{(0.0196)}{0.0529} - \underset{(0.1098)}{0.8037}\eta_{t-1} + \eta_t \ ; \ \eta_t \sim WN(0, \sigma^2)$$

Afterwards, we carry out the Ljung-Box in order to see whether or not the residuals are white noise and whether the data is independently distributed. The associated X-squared and P-value are equal to 10.335 and 0.9617 respectively. This value is quite high, with a non-zero correlation at lag 1 to 20. Hence, we can use our fitting ARIMA(0,1,1) model for forecasting [2]. Using historical data information, we forecast the future values of our target variable (Figure 6).

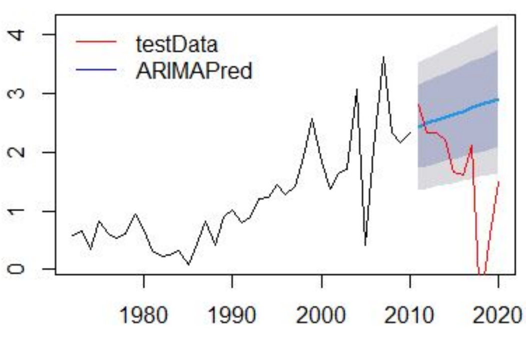

Figure 6. ARIMA forecast using training set

4 Comparison of SVR and ARIMA architecture

In this section, we will try to evaluate the performance of our models using the measures of accuracy, especially RMSE, MAE, and MAPE. See section 2 for more theoretical details [3].

Code 2: R code for assessment of ARIMA and SVR algorithms using accuracy criterion.

```
acf(trainData, lag.max=100)
pacf(trainData, lag.max=100)

adf.test(trainData, alternative = "stationary" )
#Differencing
trainData2 = diff(trainData)
adf.test(trainData2)

acf(trainData2, lag.max=100)
pacf(trainData2, lag.max=100)

arimaMod <- auto.arima(trainData, stepwise=FALSE
    , approximation=FALSE)
arimaMod.Fr <-forecast(arimaMod,h=10)

plot(arimaMod.Fr)
lines(testData, col="red")
legend("topleft",lty=1,bty = "n",col=c("red","
    blue"),c("testData","ARIMAPred"))
accuracy(arimaMod.Fr,testData)
tsdiag(arimaMod)

func <- economics$FDI ~ economics$Year #
    position x dependent on time
svr_model <- svm(func, trainData , type = "eps-
    regression",
                kernel="radial", gamma=13, cost
    =10, epsilon = 0.01)
k_hat <- predict(svr_model, economics)

plot(as.ts(trainData), xlab="Time", ylab="FDI",
    main = "")
lines(fitted(arimaMod), col="red", type = "l",
    lty=2) # ARIMA predicted position
points(x = economics$Year, y = k_hat, col = "
    blue", type = 'o', lty=3) # SVR predicted
    position
legend(x = "topleft",         title = "Series",
     # Title
       title.col = "blue",      # Color of the
    title
       # Position
       legend = c("FDI", "ARIMA", "SVR"),
       bg = rgb(1, 0, 0, alpha = 0.15),# Legend
    texts
       cex = 0.7, # Change legend size
       box.lty = 2, # Line type of the box
       box.lwd = 2,
       lty = c(1, 2, 3),           # Line types
       col = c("black", "red", "blue"), lwd = 2)
```

Source: Authors' manipulations under R programming language.

Table 4. Accuracy metrics and R squared of our models

	RMSE	MAE	MPE	MAPE	R^2
ARIMA	0.532	0.344	-50.566	67.450	0.587
SVR	0.483	0.249	-9.74	34.25	0.706

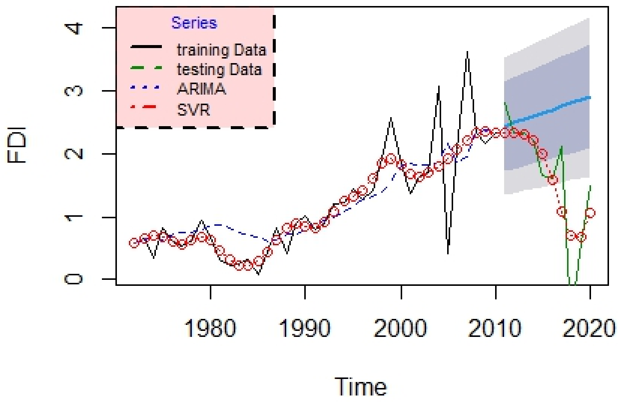

Figure 7. Performance of ARIMA and SVR

Regarding the RMSE criterion (Table 4), the lowest value is recorded at the SVR model with 0.4832 against more than 0.53 for ARIMA model. Also, based on the MAE, SVR has a very small value of 0.2491 compared to ARIMA (about 0.3441). Those results indicate clearly that the SVR algorithm performs better in terms of forecasting compared to ARIMA one. The visualization plot shows clearly that the SVR algorithm fits perfectly our target time series than the ARIMA model (for more details, see Figures 6, 7, 8, 9, 10 & 11).

4.1 Results and discussion

As demonstrated, the best ARIMA model of annual FDI is ARIMA(0,1,1). Also, the fitting of SVR in our target time series has been done by using the radial basis function (RBF). To assess the performance of our models in terms of error reduction and error variance reduction, we employed the following criterion namely MAPE, MAE, and RMSE.

Thus, the experimental findings comparing the performance of ARIMA and SVR models, show that the second model is much better than the traditional ARIMA one. Indeed, the MAPE for the ARIMA is 67.450 against only 34.25 for SVR. Same result has been given by RMSE, and MAE. Furthermore, the findings obtained from this research can be approved by computing the coefficients of determination of the two models (R-squared coefficient and adjusted R-squared coefficient).

5 Conclusion

Time series forecasting using the ARIMA technique has been widely investigated whereas the SVR one has not [4, 5, 6]. This paper proposes to assess the performance of ARIMA and SVR models in terms of error reduction and error variance reduction.

Regarding our technical manipulations, it can be inferred that the SVR technique generates the smallest errors and the biggest R squared. This technique can be robust and effective for time series prediction, especially for financial and economic series [7, 8].

We highly recommend future research to compare the advanced machine learning algorithms especially the recurrent neural network algorithms with the classical algorithms, especially with the ARIMA approach in order to choose the best program in terms of results and predictive performance.

Appendices

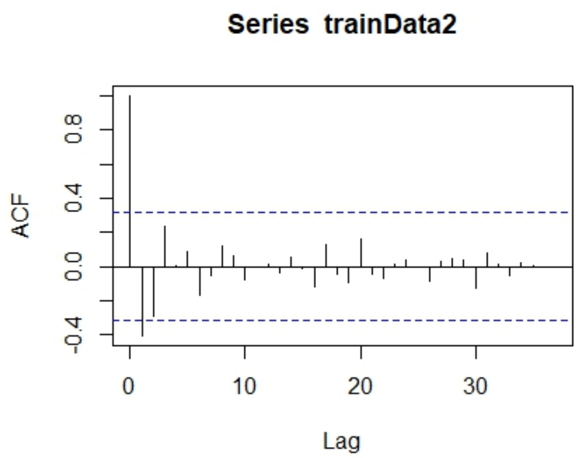

Figure 8. ACF of ∇x

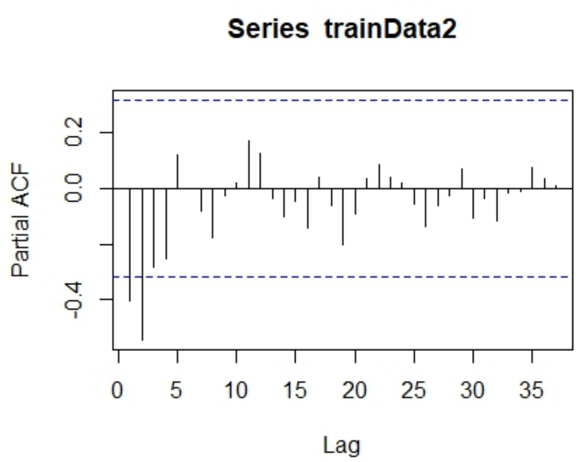

Figure 9. PACF of ∇x

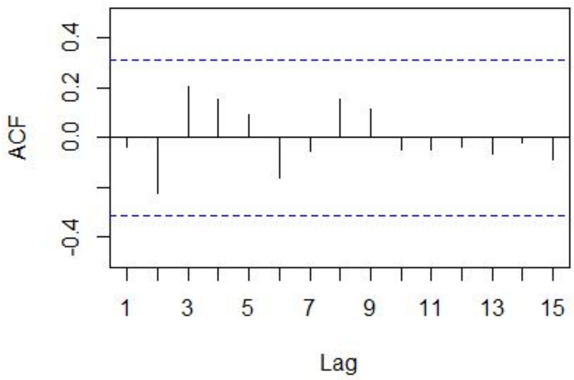

Figure 10. ACF of residuals

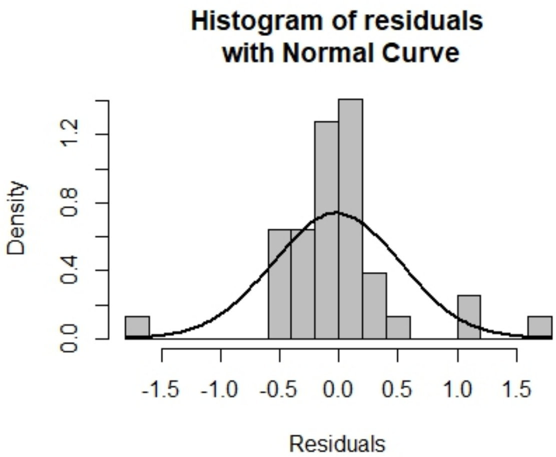

Figure 11. Histogram of residuals with Normal Curve

Acknowledgements

We're very thankful for the editor's and anonymous reviewers' helpful feedback and suggestions, which made this work much better.

REFERENCES

[1] Bourbonnais R., Terraza M., Braams J., Analyse des series temporelle en économie, 1998.

[2] Box, E.,P. Time Series Analysis: Forecasting and Control ed Second, 1976.

[3] Woodwardw A., On the relationship between the S-array and the Box-Jenkins method of ARMA model identification, Journal of the American Statistical Association Vol. 76, n 375, 1981.

[4] Youness J., and Driss M. An ARIMA Model for Modeling and Forecasting the Dynamic of Univariate Time Series: The case of Moroccan Inflation Rate 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), 2022. DOI: 10.1109/ISCV54655.2022.9806073.

[5] Guo, Y., Li, X., Bai, G., Ma, J. Time Series Prediction Method Based on LS-SVR with Modified Gaussian RBF. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds) Neural Information Processing. ICONIP 2012. Lecture Notes in Computer Science, vol 7664. Springer, Berlin, Heidelberg, 2012.

[6] Jouilil Y. and Mentagui D. "LSTM Deep Learning vs ARIMA Algorithms for Univariate Time Series Forecasting: A case study". In 2022. 8th International Conference on Optimization and Applications (ICOA). 2022.

[7] Müller, K.-R., Smola, A.J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., Vapnik, V.: Predicting Time Series with Support Vector Machines. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 999–1004. Springer, Heidelberg, 1997.

[8] Hao, W., Yu, S. Support Vector Regression for Financial Time Series Forecasting. In: Wang, K., Kovacs, G.L., Wozny, M., Fang, M. (eds) Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management. PROLAMAT 2006. IFIP International Federation for Information Processing, vol 207. Springer, Boston, MA, 2006.