Comparison of the SVR and ARIMA models for Prediction of daily imported new cases of COVID-19 in Shanghai, China

*Daren Zhao, Huiwu Zhang**
Medical administration Office, Sichuan Provincial Orthopedics Hospital
Chengdu,Sichuan,China
*Corresponding author: cdzhanghuiwu@163.com

2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML) | 978-1-6654-8290-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CACML55074.2022.00048

***Abstract*—The outbreak and spread of COVID-19 poses a tremendous threat to the health of people all over the world. We collected the new imported COVID-19 cases daily in Shanghai, China from September 1, 2021 to January 17, 2022 from the National Commission on Health of the People's Republic of China website. The SVR and ARIMA models were constructed and compared. On this base, it is provided for the early warning of the outbreak of COVID-19 and the targeted preventive measures proposed for this infectious disease.**

Keywords-COVID-19; SVR model; ARIMA model; prediction;

I. INTRODUCTION

COVID-19 was first reported in Wuhan, China, in December 2019(1), and rapidly spread worldwide, resulting in approximately 340,543,962 confirmed cases and 5,551,314 deaths on 19 January 2022(2). In particular, the emergence of the Delta and Omicron variant of coronavirus had accelerated the global epidemic and spread of COVID-19 due to the continued mutation of this infectious disease.

At present, the Omicron variant has become the predominant epidemic strain in the world(3). According to the National Commission on Health of the People's Republic of China, a total of 3,297 confirmed cases (including 16 severe cases) and 4,636 deaths were reported in 31 provinces on January 19, 2022, in mainland China(4). Recently, a few provinces breakout sporadic local cases in China, and the tested strains were Delta or Omicron variants.

Shanghai is one of the major cities for the new imported COVID-19 cases. According to the National Health Commission of the People's Republic of China, the new imported COVID-19 cases were reported almost every day in Shanghai during September 1, 2021 and January 17, 2022. Therefore, the prevention and control of COVID-19 are of great importance in China. Monitoring and early warning of COVID-19 have played an important role in the prevention and control of infectious disease outbreaks(5).

According to literature research, machine learning models have been used in fields such as healthcare, medicine, and smart homes. In the field of healthcare, Thakur et al(6).used 19 different machine learning methods fall detection systems to determine the optimal machine learning process for the development of such systems. In the medical field, Thakur et al(7). used machine learning models, human-computer interaction, the Internet of Things, pattern recognition, and ubiquitous computing to integrate and develop smart homes that have the potential to address the multiple needs of older people during ADL. And in the field of medicine, Hou et al(8).used the machine learning model XGboost to forecast the 30-days mortality for MIMIC-III patients with sepsis-3. KayvanJoo et al(9).used machine learning algorithms, Gini Index, Chi-Squared to predict the outcome of treatment with interferon/ribavirin of hepatitis C virus. Further, machine learning models have been widely used in the field of infectious disease surveillance. Dairi et al(10).used machine learning methods to predict the COVID-19 pandemic from the seven considered countries. Ayoobi et al(11).used LSTM, Convolutional LSTM, and GRU to predict the rate of new cases and new deaths of COVID-19 in Australia and Iran countries.

In this study, we collected daily imported new cases of COVID-19 from September 1, 2021, to January 17, 2022, in Shanghai, China. And all the cases have been laboratory confirmed. The SVR and ARIMA models were constructed and compared, respectively. On this basis, the best model is selected. The results will provide preventive measures and early warning of the outbreak of this infectious disease.

II. DATA SOURCE

In China, COVID-19 is classified as a Class B infectious disease. The National Health and Medical Commission of the People's Republic of China reports every day on COVID-19 in 31 provinces across China. In China, the month of September starts gradually in autumn. Autumn and winter are seasons of activity for infectious diseases. Therefore, we collected the daily imported new cases of COVID-19 in Shanghai, China between September 1, 2021 and January 17, 2022 from the National Health Commission of the People's Republic of China(http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml) for the database. In the study, the test set and validation set were the same, from September 1, 2021, and January 17, 2022, which were used to construct and evaluate the performance of the SVR and ARIMA models, respectively.

III. METHODS

A. SVR model

SVR is a machine learning algorithm, based on statistical theory(12). It has better regression performance for nonlinear, high-dimensional, and small-sample problems(13). The basic idea of the SVR model is to map the data to the high-level feature space through a nonlinear mapping (kernel function), and to perform linear regression in this space, introducing penalty parameters and converting the regression problem to

978-1-6654-8290-5/22/$31.00 ©2022 IEEE
DOI 10.1109/CACML55074.2022.00048

Authorized licensed use limited to: Universitas Brawijaya. Downloaded on November 07,2023 at 07:58:43 UTC from IEEE Xplore. Restrictions apply.

an optimization problem(14). The basic SVR function of the mathematical formula is expressed as(15):

$$f(x) = w^T \varphi(x) + b \tag{1}$$

Where f(x) is the prediction values, φ(x) is nonlinear mapping, and w and b represent modifiable coefficients. R(C) is the penalty function, ε is the insensitive loss factor, ξ_i and ξ_i^* are relaxation variables.

$$R(C) = \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{2}$$

$$\text{s.t.} \quad f(x_i) - y_i \le \varepsilon + \xi_i \tag{3}$$

$$y_i - f(x_i) \le \varepsilon + \xi_i^* \tag{4}$$

$$\xi_i \ge 0, \xi_i^* \ge 0, i = 1,2,...m \tag{5}$$

B. ARIMA model

The ARIMA model, a series forecasting method, was proposed by Box and Jenkins in the 1970s(16). The basic idea of the ARIMA model is to consider the observational data set over time as a random set and use a certain mathematical model to approximate this sequence(17).

The basic ARIMA model is expressed as ARIMA (p, d, q) (P, D, Q)s. p is the order of auto-regression, d is the trend difference, q is the order of moving average; P is the seasonal auto regression lag, D is the seasonal difference, Q is the seasonal moving average, s is the length of the cyclical pattern. If the time series is non-seasonal, the expression for the ARIMA model is ARIMA (p, d, q).

The modeling process of the ARIMA model mainly includes three steps(18). First, to determine whether the original sequence is stationary or not. The original sequence was plotted by SPSS software to judge it. If the original sequence is non-stationary, the difference should be carried out. Second, to determine parameters of p, q, P, Q orders and candidate ARIMA models by autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Third, to evaluate model adaptation by Ljung-Box (Q) test. If the ARIMA model passed the Ljung-Box (Q) test, its residuals were white noise time series. The optimal model was determined by the lowest value of the Bayesian information criterion of Schwarz(BIC) and its residuals were white noise time series.

C. Evaluation of Prediction Performance

Evaluation of prediction performance of SVR and ARIMA models were determined by MAE(Mean Absolute Error), MSE(Mean Square Error)and RMSE(Root Mean Square Error) indices, which were shown as(18, 19):

$$MAE = \frac{\sum_{t=1}^{n}\left|X_t - \hat{X}_t\right|}{n} \tag{6}$$

$$MSE = \frac{1}{n}\sqrt{\sum_{t=1}^{n}(X_t - \hat{X}_t)^2} \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(X_t - \hat{X}_t)^2}{n}} \tag{8}$$

D. Data Analysis

The R software version 4.1.1 was used to construct the SVR model, and the SPSS software version 21.0 (IBM Corp., Armonk, NY, USA) to construct the ARIMA model. The level of significance is 0.05.

IV. RESULTS

A. SVR model

We used the "e1071" "caret"and "tidyverse" packages of R software to construct the SVR model. The tune.svm ( ) function was applied to obtain the optimal SVR model, the parameters of C, γ , ε were 100, 0.1 and 0.1, respectively.

B. ARIMA model

We plotted the original time series by SPSS software. In our study, due to the time series of the daily imported new cases of COVID-19 from September 1, 2021 and January 17, 2022 showed non-seasonal characteristics, the expression for the ARIMA model is ARIMA(p, d, q).

Figure 1(a) showed that the original time series of the daily imported new cases of COVID-19 from September 1, 2021 and January 17, 2022 displayed a fluctuating trend, indicating that the time series was non-stationary. Accordingly, a transformation of the trend difference over the original time series was performed, so the parameter of d was 1(Figure 1(b)).

The autocorrelation function (ACF) and partial autocorrelation function (PACF) charts were plotted to determine parameters of p, q, P, Q orders and candidate ARIMA models. Figure 2 showed that the differenced time series had an obvious trailing with a slow decay at the first order, and began to decrease at the third orders. Therefore, the parameter q is 3, and the range of p is 0-3. The four candidate ARIMA models were determined, which were ARIMA(0,1,3), ARIMA(1,1,3), ARIMA(2,1,3), ARIMA(3,1,3)(TABLE Ⅰ).

Authorized licensed use limited to: Universitas Brawijaya. Downloaded on November 07,2023 at 07:58:43 UTC from IEEE Xplore. Restrictions apply.

TABLE I. THE FOUR CANDIDATE ARIMA MODELS

Candidate ARIMA models	Normalized BIC	Ljung-Box Q(18) Statistics	DF	*p-value*
ARIMA(0,1,3)	2.772	14.197	16	0.584
ARIMA(1,1,3)	2.862	14.995	14	0.381
ARIMA(2,1,3)	2.905	12.750	13	0.467
ARIMA(3,1,3)	2.914	9.864	12	0.628

TABLE II. THE PARAMETERS OF ARIMA(0,1,3) MODEL

ARIMA(0,1,3)		Estimate	SE	t	*p-value*
Difference		1	-	-	-
MA	Lag 1	0.440	0.078	5.607	0.000
	Lag 3	0.220	0.080	2.736	0.007

The four candidate ARIMA models all passed the Ljung-Box (Q) test, and their residuals were all white noise time series ($p > 0.05$). The optimal model was ARIMA(0,1,3), with the lowest value of the Bayesian information criterion of Schwarz(BIC), its residual was white noise time series ($p > 0.05$).(TABLE Ⅱ).

C. Comparison of the SVR and ARIMA models

Prediction performance of the SVR and ARIMA models was evaluated by calculating the MAE, MSE, and RMSE of the observed and predicted values. The results were as shown in (TABLE Ⅲ). Comparison of the observed and predicted values of SVR and ARIMA models were as shown in Figure 3.

V. DISCUSSION

With the emergence of the Delta and Omicron variant of coronavirus, the global spread of this infectious disease epidemic has been accelerated, resulting in a growing risk of preventing the new imported COVID-19 cases in mainland China. Recently, the flow of people across regions has increased in China due to the upcoming Spring Festival travel, and the situation of epidemic prevention and control is severe and complicated.

Shanghai is a major city for international trade and commerce. There is a frequent flow of people such as studying and working abroad, traveling, visiting relatives in Shanghai, China, causing huge pressure on the prevention and control of the new imported COVID-19. As a result, it is important to strengthen the monitoring and prediction of the new imported COVID-19 cases to effectively prevent and control the spread of this infectious disease in Shanghai, China.

In this study, research data is non-seasonal with a sample size of 139, which meets the modeling requirements of SVR and ARIMA models. Therefore, the SVR and ARIMA models were used to predict the number of new daily imported COVID-19 cases in Shanghai, China. The results

TABLE III. PREDICTION PERFORMANCE OF THE SVR AND ARIMA MODELS

Indices	SVR model	ARIMA model
MAE	2.4420	2.7681
MSE	20.1184	23.0217
RMSE	28.4517	32.5576

showed that the MAE, MSE, and RMSE values of the SVR model were all lower than the ARIMA model, indicating that the prediction performance of the SVR model was better than the ARIMA model.

There may be some potential reasons for this as follows. First, the ARIMA model may consider randomness, periodicity, and trend during the modeling, and can be constructed as long as there were at least 30 data samples(18). However, the this model cannot handle nonlinear problems. It is inferior to machine learning models in dealing with non-periodic and seasonal problems(14). Second, compared with the ARIMA model, the SVR model has some advantages for handling problems, which are nonlinear, high-dimensional, and small samples, therefore it is also increasingly applied to different sectors(13, 20, 21, 22). Moreover, the SVR model was one of the most effective machine learning methods and had been applied to predict the incidence of infectious diseases in recent years(12, 14, 23).

Therefore, in this study, the prediction performance of the SVR model was better than the ARIMA model. The prediction results of the SVR model well simulated the reality of epidemiological trends of the new imported COVID-19 cases in Shanghai, China.

VI. CONCLUSIONS

In the study, the new imported COVID-19 cases daily in Shanghai, China from September 1, 2021 to January 17, 2022 were taken from the National Commission on Health of the People's Republic of China website, and the SVR and ARIMA models were used to predict their epidemic trend. The study proved that the SVR had a better prediction performance than the ARIMA model. Prediction results can provide an early warning of the outbreak and take measures to prevent and control COVID-19 in Shanghai, China.

ACKNOWLEDGMENT

We thank the Projects of Sichuan Provincial Primary Health Service Development Research Center (grant no. SWFZ21-Q-59) and Projects of Sichuan Provincial Orthopedics Hospital (grant no. 2021GL01) for their funding to the publication.

REFERENCES

[1] Yüce M, Filiztekin E, Özkaya KG. "COVID-19 diagnosis -A review of current methods. Biosens Bioelectron," vol. 172, Jan. 2021, pp.112752, doi:10.1016/j.bios.2020.112752.

[2] WHO.WHO Coronavirus (COVID-19) Dashboard.[cited 2022 Jan 19]. Available from: URL: https://covid19.who.int/.

[3] Araf Y, Akter F, Tang YD, Fatemi R, Parvez SA, Zheng C, et al. Omicron variant of SARS-CoV-2: Genomics, transmissibility, and

Authorized licensed use limited to: Universitas Brawijaya. Downloaded on November 07,2023 at 07:58:43 UTC from IEEE Xplore. Restrictions apply.

responses to current COVID-19 vaccines. J Med Virol, 12 Jan. 2022, doi:10.1002/jmv.27588.

[4] The National Commission on Health of the People's Republic of China. Report of Epidemic Situation. [cited 2022 Jan 19]. Available from: URL: http://www.nhc.gov.cn/xcs/yqtb/202201/d9ec13c4b7c14ab39f81366653bbb382.shtml.

[5] Roy S, Bhunia GS, Shit PK. "Spatial prediction of COVID-19 epidemic using ARIMA techniques in India," Model Earth Syst Environ, Jul 16. 2020, doi: 10.1007/s40808-020-00890-y.

[6] Thakur N, Han CY. "A Study of Fall Detection in Assisted Living: Identifying and Improving the Optimal Machine Learning Method," Journal of Sensor and Actuator Networks, vol. 10, Jun. 2021, pp.39. https://doi.org/10.3390/jsan10030039.

[7] Thakur N, Han CY. "An Ambient Intelligence-Based Human Behavior Monitoring Framework for Ubiquitous Environments," Information, vol. 12, Feb. 2021, pp.81. https://doi.org/10.3390/info12020081.

[8] Hou, Nianzong et al. "Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost," Journal of translational medicine, vol. 18, Dec. 2020, pp.462. doi:10.1186/s12967-020-02620-5.

[9] KayvanJoo AH, Ebrahimi M, Haqshenas G. "Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms," BMC research notes, vol. 7, 2014, pp.565. doi:10.1186/1756-0500-7-565.

[10] Dairi A, Harrou F, Zeroual A, Hittawe MM, Sun Y. "Comparative study of machine learning methods for COVID-19 transmission forecasting," J Biomed Inform, vol. 118, 2021, pp.103791, doi:10.1016/j.jbi.2021.103791.

[11] Ayoobi N, Sharifrazi D, Alizadehsani R, Shoeibi A, Gorriz JM, Moosaei H, et al. "Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods," Results in physics, vol. 27, 2021, pp.104495. doi:10.1016/j.rinp.2021.104495.

[12] Dharani NP, Bojja P, Raja Kumari P. "Evaluation of Performance of an LR and SVR models to predict COVID-19 Pandemic," Mater Today Proc, 16 Feb. 2021, doi:10.1016/j.matpr.2021.02.166.

[13] Liu B, Jin Y, Li C. "Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR-SVR-ARMA combined model," Sci Rep, vol. 11, 2021, pp.348, doi:10.1038/s41598-020-79462-0.

[14] Norrulashikin MA, Yusof F, Hanafiah NHM, Norrulashikin SM. "Modelling monthly influenza cases in Malaysia," PLoS One, vol. 16, 2021, pp. e0254137, doi:10.1371/journal.pone.0254137.

[15] Xu D, Zhang Q, Ding Y, Zhang D. "Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting," Environ Sci Pollut Res Int, vol. 29, 2022, pp.4128-4144, doi: 10.1007/s11356-021-15325-z.

[16] Ilie OD, Cojocariu RO, Ciobica A, Timofte SI, Mavroudis I, Doroftei B. "Forecasting the Spreading of COVID-19 across Nine Countries from Europe, Asia, and the American Continents Using the ARIMA Models," Microorganisms, vol. 8, 2020, pp. 1158, doi:10.3390/microorganisms8081158.

[17] Wang L, Liang C, Wu W, Wu S, Yang J, Lu X, et al. "Epidemic Situation of Brucellosis in Jinzhou City of China and Prediction Using the ARIMA Model," Can J Infect Dis Med Microbiol, vol. 2019, Jun. 2019, pp.1429462, doi:10.1155/2019/1429462.

[18] Wang YW, Shen ZZ, Jiang Y. "Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China," PLoS One, vol. 13, 2018, pp.e0201987, doi:10.1371/journal.pone.0201987.

[19] Alim M, Ye GH, Guan P, Huang DS, Zhou BS, Wu W. "Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study," BMJ Open, vol. 10, 2020, pp. e039676, doi: 10.1136/bmjopen-2020-039676.

[20] Fan Y, Lu W, Miao T, An Y, Li J, Luo J. "Optimal design of groundwater pollution monitoring network based on the SVR surrogate model under uncertainty," Environ Sci Pollut Res Int, vol. 27, 2020, pp.24090-24102, doi:10.1007/s11356-020-08758-5.

[21] Ghazvinian H, Mousavi SF, Karami H, Farzin S, Ehteram M, Hossain MS, et al. "Integrated support vector regression and an improved particle swarm optimization-based model for solar radiation prediction," PLoS One, vol. 14, 2019, pp.e0217634, doi:10.1371/journal.pone.0217634.

[22] Leng T, Li F, Chen Y, Tang L, Xie J, Yu Q. "Fast quantification of total volatile basic nitrogen (TVB-N) content in beef and pork by near-infrared spectroscopy: Comparison of SVR and PLS model," Meat Sci, vol. 180, 2021, pp.108559, doi:10.1016/j.meatsci.2021.108559.

[23] Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. "Developing a dengue forecast model using machine learning: A case study in China," PLoS Negl Trop Dis, vol. 11, 2017, pp. e0005973, doi:10.1371/journal.pntd.0005973.

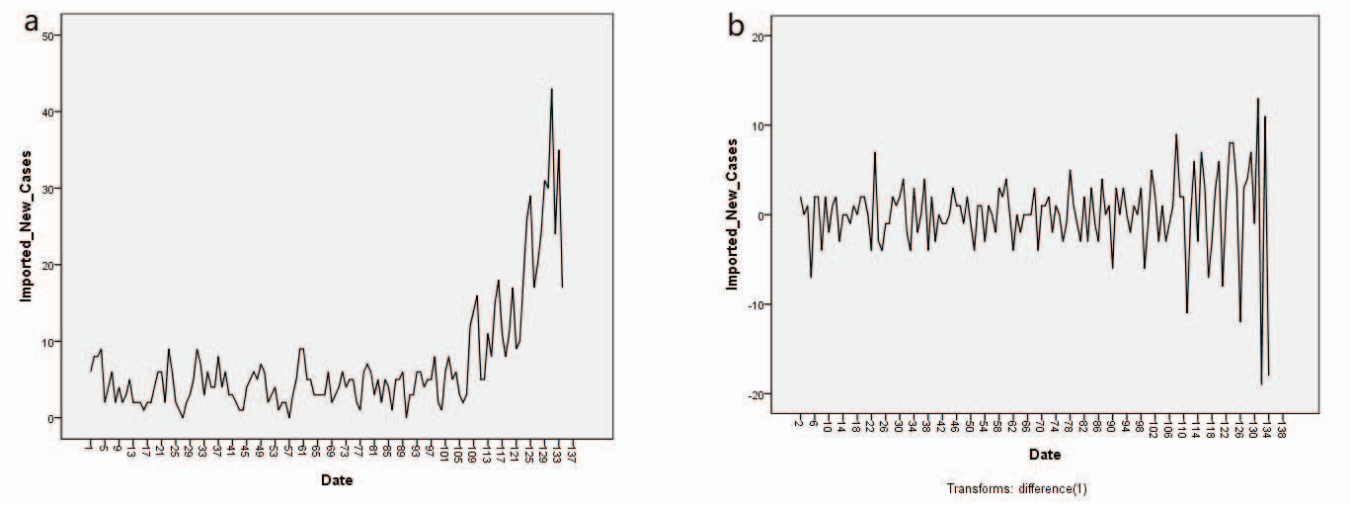

Figure 1. The original time series and after a transformation of the trend difference time series: (a) The original time series of the new imported COVID-19 cases daily in Shanghai, China. (b) After a transformation of the trend difference time series of the new imported COVID-19 cases daily in Shanghai, China

Authorized licensed use limited to: Universitas Brawijaya. Downloaded on November 07,2023 at 07:58:43 UTC from IEEE Xplore. Restrictions apply.

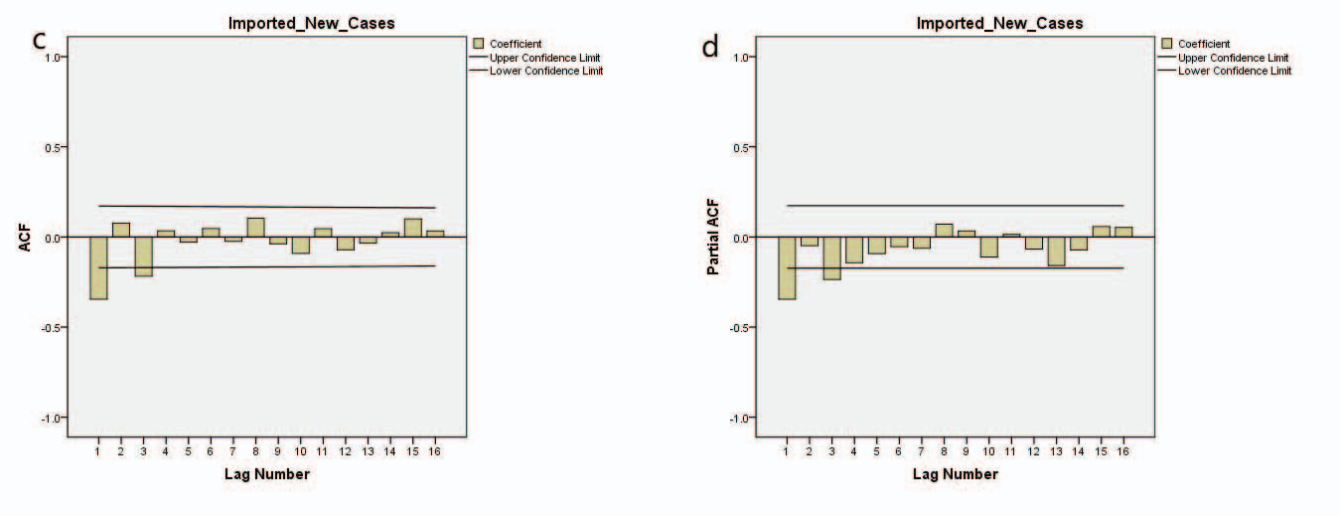

Figure 2. The plot of after a difference order of the new imported COVID-19 cases daily in Shanghai, China time series. (c) ACF plot (d) PACF plot.

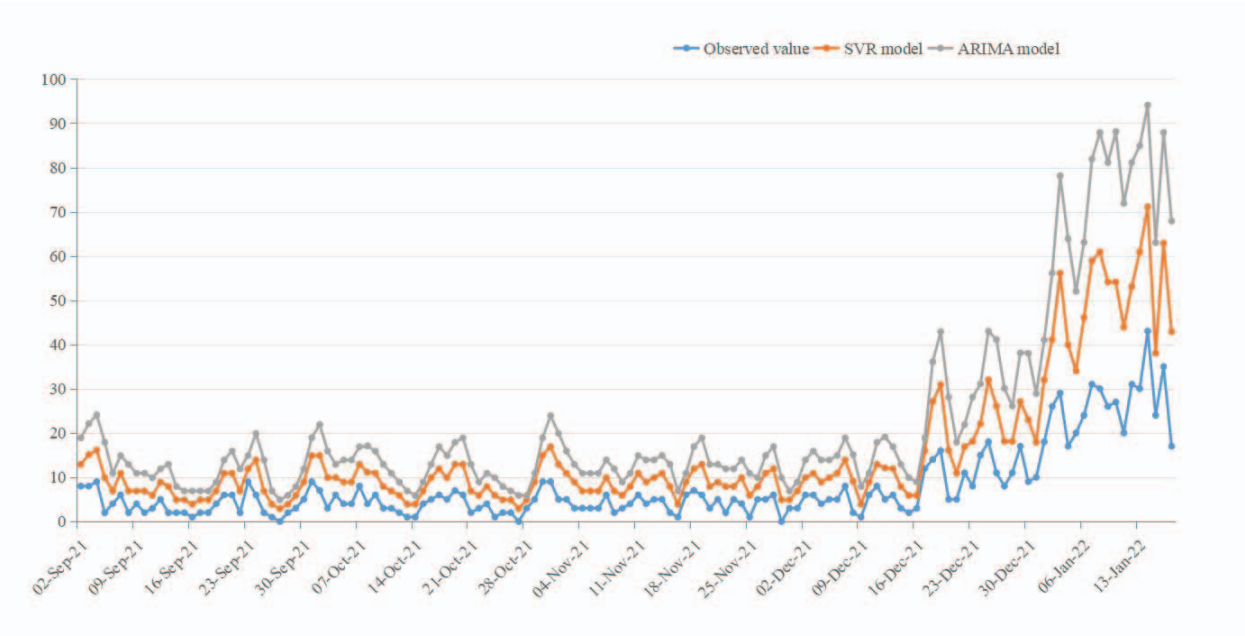

Figure 3. Comparison of the observed and predicted values of SVR and ARIMA models.

Authorized licensed use limited to: Universitas Brawijaya. Downloaded on November 07,2023 at 07:58:43 UTC from IEEE Xplore. Restrictions apply.