Shahid Bahonar University of Kerman
Journal of Mahani Mathematical Research
Print ISSN: 2251-7952 Online ISSN: 2645-4505

FORECASTING EDUCATED UNEMPLOYED PEOPLE IN INDONESIA USING THE BOOTSTRAP TECHNIQUE

U. Mahmudah, S. Surono ✉, P.W. Prasetyo, and A.E. Haryati

Article type: Research Article
(Received: 21 April 2022, Received in revised form: 02 June 2022)
(Accepted: 06 July 2022, Published Online: 07 July 2022)

Abstract. Forecasting is an essential analytical tool used to make future predictions based on preliminary data. However, the use of small sample sizes during analysis provides inaccurate results, known as asymptotic forecasting. Therefore, this study aims to analyze the unemployment rate of educated people in Indonesia using the bias-corrected forecasting bootstrap technique. Data were collected from a total of 30 time series of educated unemployed from 2015 to 2019 using the bias-corrected bootstrap technique and determined using the interval prediction method. The bootstrap replication used is at intervals of 100, 250, 500, 1000, and 10000. The results obtained using the R program showed that the bootstrap technique provides consistent forecasting results, better accuracy, and unbiased estimation. Moreover, the results also show that for the next 10 periods, the number of educated unemployed people in Indonesia is projected to decline. The bootstrap coefficient also tends to decrease with an increase in the number of replications, at an average of 0.958. The interval prediction is also known to be smooth, along with a large number of bootstrap replications.

Keywords: AR model, Bias-corrected, Bootstrap, Forecasting.
2020 MSC: Primary 62P99, 62R07, 62R01.

1. Introduction

Forecasting future data is an important process used to determine human resource requirements in order to make effective and efficient decisions while planning design. The Autoregressive (AR) model in time series is a tool used to make future predictions by analyzing previous values. It is sensitive to small sample sizes, thereby affecting the accuracy of forecasting results [10]. According to [19] during forecasting, accuracy is highly dependent on the length of the previous data series. However, sometimes it is difficult to get long data series, especially in the annual form. This study combined a nonparametric bootstrap approach and an AR model to determine the relatively small sample size. The principle of the bootstrap approach is resampling without return based on the generated data, and it is an effective alternative used to produce

✉ sugiyarto@math.uad.ac.id, ORCID: 0000-0001-6210-7258
DOI: 10.22103/jmmr.2022.19368.1239

Publisher: Shahid Bahonar University of Kerman
How to cite: U. Mahmudah, S. Surono, P.W. Prasetyo, A.E. Haryati, *Forecasting Educated Unemployed People in Indonesia Using the Bootstrap Technique*, J. Mahani Math. Res. 2023; 12(1): 171-182.

better forecasting values. Moreover, the bootstrap approach is distribution-free [7] and is believed to provide relatively accurate prediction results using large and small samples [4]. Point forecasts under the bootstrap approach as well as bootstrap prediction interval are studied. Furthermore, this study follows the bootstrap procedure on the prediction interval model time series AR model [9, 10].

Numerous studies have been applied to the forecasting analysis using time series data [2, 4, 12]. Clements & Kim applied 3 methods to deal with small samples in forecasting analysis based on the bootstrap approach, namely the bias-corrected OLS, the Roy–Fuller, and the Andrews–Chen estimator. According to them, the Roy-Fuller estimator is generally superior to the other 2 [13]. Moreover, [14] used the bootstrap technique, which allowed for uncertainty in parameter estimation to reduce small sample bias. Meanwhile, a study carried out by [8] applied sieve bootstrap to autoregressive.

Furthermore, the bootstrap techniques application has also been found in forecasting future values in various fields. For instance, it is used to determine probabilistic forecasts from the past single-valued analysis offered by the Numerical Weather Prediction model in Eastern Canada [18]. This method is also used to predict pax in the air transportation industry [15]. Forecasting in the airline industry is also carried out, especially related to accurate demand [17]. Moreover, it is also used to get more accurate predictions on wind power integrated in a sustainable manner [5].

The bootstrap method is a non-parametric method that employs sample replication. Since Efron's introduction, it has become widely used [3]. The bootstrap method is a sampling procedure that repeats as many as N new samples from the original data of size n, where each new sample is taken one by one from the original data up to n times with returns. In other words, the bootstrap method estimates a sampling distribution by generating a large number of new samples from the original sample. The benefit of this technique is that it can analyze a sample of data with a small sample size. Furthermore, carrying out distribution assumptions and initial assumptions to estimate the shape of the distribution and statistical tests is not required [6, 16]. Efron demonstrated through simulation comparisons that using bootstrap bias correction could provide better estimates of classification error rate than the widely used cross-validation approach [16].

This study relies on using time-series data to determine the number of unemployment in Indonesia using the bootstrap technique. Data were collected from 30 time-series from 2005 to 2019, published by Statistics Indonesia twice a year and sourced from the national labor force survey (Sakernas).

2. Materials and Methods

2.1. **Autoregresive (AR)Model.** The autoregressive model is defined by $AR(p)$, which contains an arbitrary deterministic component, $D_(j,t)$, and includes intercepts, time trends, and dummy variables. Furthermore, the AR model is written as follows [10]:

$$Y_t = \Sigma_{i=1}^{p}\gamma_i Y_{t-1} + \Sigma_{j=1}^{m}\beta_j D_{j,t} + u_t \tag{1}$$

Where u_t indicates the error term. W is a size nxp matrix which is lagged dependent variable, while D is defined as a size nxm matrix with deterministic components. Then $Z = [W : D]$ is a size nxk matrix, where $k = p + m$. If $Y = (Y_1, Y_2, ..., Y_n)', u = (u_1, u_2, ..., u_n)'$ then equation 1 is written as follows:

$$Y = Z\alpha + u \tag{2}$$

If the vector of the unknown coefficients is defined by $\alpha = (\gamma : \beta)'$, then the least-square estimator of α and σ is defined as follows:

$$\hat{\alpha} = (\hat{\gamma}, \hat{\beta}) = \frac{Z'Y}{Z'Z} \tag{3}$$

$$s^2 = \frac{e'e}{(n-k)}$$

where $e = (e_1, e_2, ..., e_n)'$ shows residuals, assuming the optimal forecast is defined as follows:

$$Y_{n+h} = Y_n(h) + u_{n+h} \tag{4}$$

where

$$Y_n(h) = \Sigma_{i=1}^{p}\gamma_i Y_n(h-i) + \Sigma_{j=1}^{m}\beta_j D_{j,n+h} \tag{5}$$

if $h \leq 0$ then $y_n(h)$ is estimated as follows:

$$\hat{Y}_n(h) = \Sigma_{i=1}^{p}\hat{\gamma}_i Y_n(h-i) + \Sigma_{j=1}^{m}\hat{\beta}_j D_{j,n+h} \tag{6}$$

Equation 5 $\hat{Y}_n(h)$ is consistent and asymptotically normal [10]. Therefore, by using normal approximation, an asymptotic prediction interval is constructed. However, when the sample size used is a small category, the asymptotic predictions give poor results (deficiently). Therefore, alternative ways are needed to deal with small sample sizes and provide better predictive results [9, 20]. This is followed by using a bias-corrected method as an alternative to the bootstrap approach to provide good predictive results based on small sample sizes.

2.2. **Bias-Corrected Estimators for AR models.** Kim et al. [10] applied a nonparametric bootstrap approach using residual resampling to estimate the bias of $\hat{\alpha}$ on $O(n^{-1})$ based on equation 1. The sample bootstrap is defined with $\{Y_t^*\}_{t=1}^n$, which is degenerated using the starting point $\{Y_t\}_{t=1}^p$ as:

$$Y_t^* = \Sigma_{i=1}^p \hat{\gamma}_i Y_{t-i}^* + \Sigma_{j=1}^m \hat{\beta}_j D_{j,t} + e_t^* \tag{7}$$

where e_t^* is a random sample taken with returns based on $\{e_t\}_{t=1}^n$. Furthermore, a bootstrap estimator for α is generated and symbolized by $\hat{\alpha}^* = \frac{Z^* Y^*}{Z^{*\prime} Z^*}$. The estimator of the bias-corrected bootstrap is obtained from:

$$\hat{\alpha}_B^C = \hat{\alpha} - bias(\hat{\alpha}) \tag{8}$$

$$\hat{\alpha}_B^C = [\hat{\gamma}_B^C : \hat{\beta}_B^C]$$

Using equation 8, the following are the steps of the bias-correction bootstrap estimation method:

Step 1:

Calculate the $\hat{\alpha}$ and s^2 estimators and the bias-correction estimator using equations 1 and 8. The residual from the calculation of $\hat{\alpha}$ is defined by $\{e_t^C\}_{t=1}^n$.

Step 2:

Use starting point $\{Y_t\}_{t=1}^p$ to generate sample bootstrap $\{Y_t^*\}_{t=1}^p$ as $Y_t^* = \Sigma_{i=1}^p \hat{\gamma}_i^C Y_{t-i}^* + \Sigma_{j=1}^m \hat{\beta}_j^C D_{j,t} + e_t^*$, where e_t^* sample is randomly determined using the bias estimation in step 1 defined as $\hat{\alpha}^{C*} = \hat{\alpha}^* - bias(\alpha)$.

Step 3:

Repeat step 2 in B times to generate a bootstrap distribution for the forecast, $\{Y_n^{C*}(h;j)\}_{j=1}^B$.

The nominal coverage rate of $100(1-\theta)\%$ for bias-corrected bootstrap on the prediction interval is given as follows:

$$[Y_n^*(h,\tau), Y_n^*(h,1-\tau)], \tag{9}$$

where $Y_n^*(h,\tau)$ is the $100^{\text{th}}\tau$ percentile of the bootstrap distribution

$$\{Y_n^{C*}(h;i)\}_{i=1}^B$$

and $\tau = 0.5\theta$.

3. Results and Discussion

This study was carried out using 30 educated unemployment data series from 2005-2019 published by BPS in 2020. Indonesia's open unemployment data is tabulated based on the National Labor Force Survey (SAKERNAS) and released twice a year. The time series consists of 2 categories, namely diplomas and universities, which represent the total number of educated unemployed in Indonesia. Although data for the 2020 period was available at the time of this

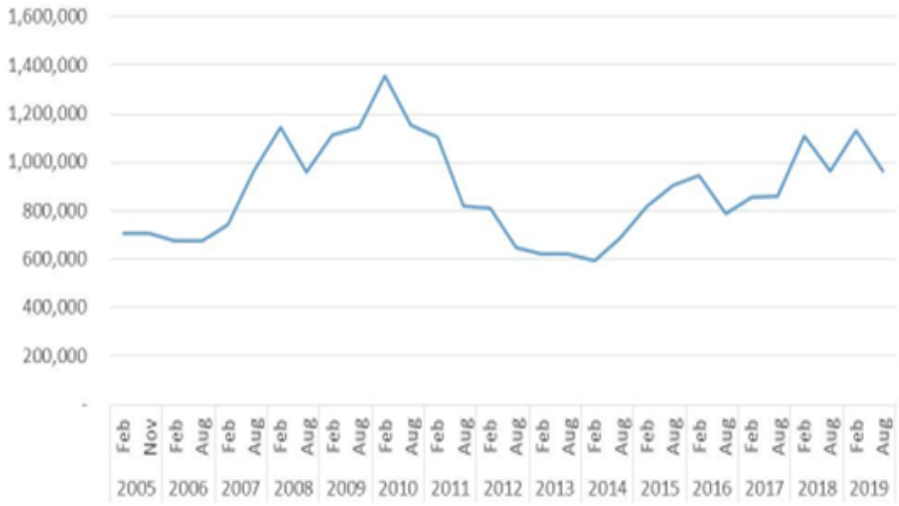

FIGURE 1. Plot of graduate unemployment series

study, it was not included during the model estimation and forecasting process due to the emergence of the COVID-19 pandemic. [10] stated that the spread of the SARS virus is unpredictable because it had a negative impact due to extreme observations. Therefore, the last data series used was in August 2019, and it is hoped that the prediction results still reflect the real situation on the ground. Moreover, the results also represent normal conditions that do not pay attention to special events capable of effecting changes in data. In other words, when there is no Covid-19, the forecasting results tend to reflect on the future state of the number of educated unemployed in Indonesia.

The time-series data for the educated unemployment in Indonesia from 2005 to 2015 had an average number of 886,180. Meanwhile, the smallest and the highest numbers were 593,556 and 1,358,206, and they both occurred in February of 2014 and 2010. Figure 1 shows the time series data on educated unemployment from 2005-2019, collected only in February and August.

Figure 1 shows that the data series used are inconsistent due to the increase and decrease in the number of educated unemployed in Indonesia from 2005 to 2019. Deeper analysis found a significant increase from February to August 2011 at 284,419. Meanwhile, the largest decline occurred from August 2017 to February 2018, at a rate of 246,673. The lowest decline was experienced from 2005 to 2019, at a rate of 8,864.

However, before performing predictive analysis with the AR model, a stationarity test on the required time series data. This is significant because the AR model cannot be applied to non-stationary data. The differencing process

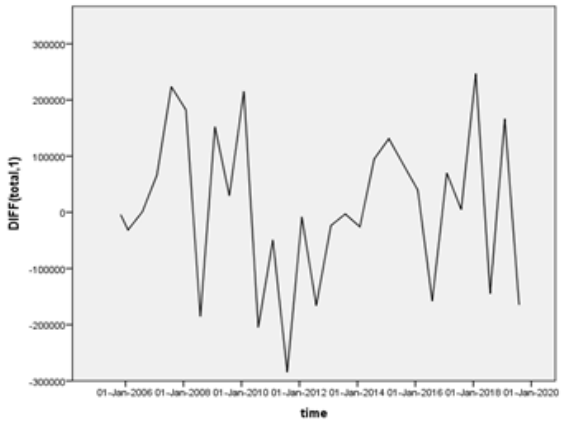

FIGURE 2. Plot of Differencing Process

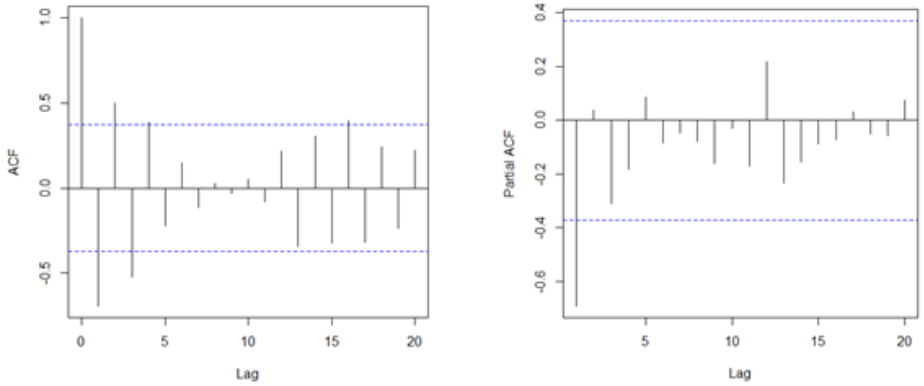

FIGURE 3. ACF and PACF of educated unemployment data

is commonly used to obtain stationary data. For this purpose, the Dickey-Fuller test is carried out. The Augmented Dickey-Fuller test indicated that it required two differencing processes to obtain stationary data (Dickey-Fuller = -4.238, Lag order = 3, and p-value 0.014). Figure 2 shows the plot of the time series data after the differencing process.

Figure 2 illustrates the stationary data on the educated unemployment data series. Further, Figure 3 below illustrates the results of the autocorrelation function (ACF) and Partial autocorrelation function (PACF) analysis on the educated unemployment data series from 2005-2019.

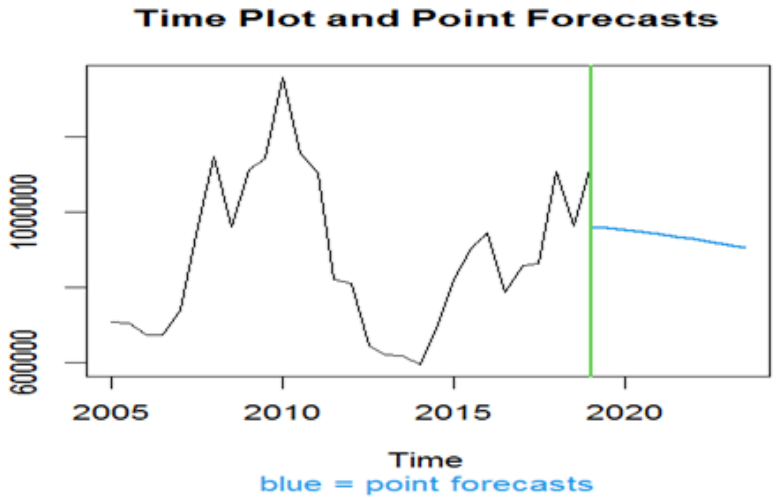

FIGURE 4. Point Forecasts of graduate unemployment series

From Figure 3, because the PACF cutoff is on lag 1, the AR (1) model could be used. By using the AR (1) model, the prediction analysis is carried out. Based on the results of these statistical tests, this study applies the bootstrap technique to the graduate unemployment series data to provide better accuracy. To determine the point forecasts and prediction intervals, this study applied a bias-correction method to generate point forecasts and prediction intervals. Point forecasts are inseparable from the prediction interval because the predictive analysis only uses point estimates. The prediction results are also difficult to determine, therefore, without interval prediction values, point forecasts cannot be ascertained [19]. This makes it difficult to ascertain the value of the prediction interval in expressing the uncertainties that need to arise in the forecast analysis. Some important things that need to be highlighted in prediction intervals include uncertainty in the model and parameter estimates and individual randomness associated with the prediction of certain points.

Figure 4 shows the prediction points for 10 data using the bootstrap approach from February 2020 to August 2024. The predictive analysis under the bootstrap approach was carried out using the bootstrap replications number of 100, 250, 500, 1000, and 10000.

Figure 4 shows that the blue line comprises a total of 10 predictive data. A point forecast is interpreted as a function that contains a points summary of the predictive distribution. Meanwhile, the green line is the deadline for data on the number of open unemployment used in August 2019. In general, Figure 4 shows that for the next 10 periods, the number of educated unemployed in Indonesia declined. The results also found that the bootstrap coefficient

decreases with increase in the number of replications at an average coefficient of 0.958.

A good forecasting system, which produces residuals with an average of zero, is needed to make unbiased findings [1,19]. The predictive results indicate that the average residual is 0.000 for all bootstrap (B) while using the replication counts. In other words, the bootstrap technique model carries out a decent job of projecting the educated unemployment rate in Indonesia. Furthermore, the results also indicate that point forecasts and bias-corrected-parameter estimates also provide consistent values, increasing the number of bootstrap replications due to a decrease in average value. Table 1 shows point forecasts based on bootstrap bias-corrected-parameter estimates for 10 future prediction data.

TABLE 1. Point Forecasts under bootstrap bias-corrected estimation

Point Forecasts	B=100	B=250	B=500	B=1000	B=10000
h1	959580	960960	961479	961016	960877
h2	953695	956175	957128	956277	956024
h3	947665	950974	952274	951111	950769
h4	941499	945374	946932	945538	945131
h5	935209	939395	941122	939575	939127
h6	928802	933052	934858	933240	932775
h7	922288	926364	928158	926548	926092
h8	915675	919345	921035	919517	919092
h9	908969	912010	913505	912160	911790
h10	902177	904373	905582	904492	904202

Table 1 shows the predicted value in the future indicated by the point forecasts denoted by h. The results indicate that the average numbers of educated unemployed in Indonesia from February 2020 to August 2024 are 931556, 934802, 936207, 934947, and 934588 for B values of 100, 250, 500, 1000, and 10,000. The figures are consistent and do not make a significant difference. Therefore, all point forecasts generated from the number of bootstrap replications indicate consistent results [19]. Figure 5 shows the point forecast for the five selected bootstrap replications.

Figure 5 clearly shows that the bootstrap approach in predictive analysis produces consistent results, with the point forecast decreasing for all bootstrap replications. Figure 5 also shows that when the number of replications is increased, the results tend to be more consistent. Figure 5 and Table 1 indicate that the point forecasts for the next 10 data periods decline due to the lack of fluctuations in the prediction data. This indicates that for the next 5 years, the number of educated unemployed will continuously decline. Therefore, the government's strategy to control or reduce the number of educated

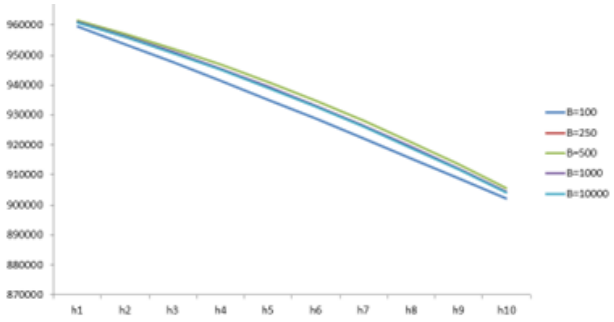

FIGURE 5. Bootstrapped point forecast

unemployed is appropriate. However, this research also focuses on the analysis of prediction intervals because these values cannot be separated from the point forecast. Without prediction intervals, a prediction analysis cannot provide a precision that accurately describes the real situation in the field, which is prone to uncertainty. Figure 5 shows that the plotting prediction intervals and point forecast that use bootstrap replications are B=100, 250, 500, 1000, and 10000. This study uses quantile estimates which are from the approximate distribution. Moreover, interval prediction is used to determine the range of coverage probabilities defined under the distribution. This study used 95% prediction intervals determined by the forecast distribution's 2.5% and 97.5% quantiles. The 95% prediction interval is commonly used in forecasting analysis besides the 80% [11, 19].

Figure 6 shows the green line of the data series in the forecasting analysis, which denotes the boundary. Meanwhile, the blue line shows point forecasts for predictive future data, using the next 10 periods. It is important to note that point forecast predictive analysis cannot provide the accuracy of the projected value without accompanying the prediction interval value [11, 19]. Moreover, when predictive analysis provides a greater degree of uncertainty, the prediction interval tends to widen.

The analysis results in Figure 6 indicate that Indonesia's actual educated unemployment rate needs to be within the prediction interval of 85% at a probability of 0.95. Furthermore, the prediction interval for 95% is used when forecasting over a longer period continuously.

The educated unemployed people in Indonesia are predicted to experience a continuous decline for each number of bootstrap replications used, such as 100, 250, 500, 1000, and 10000. Figure 6 shows that the prediction interval of the number of bootstrap replications at 100 and 250 are less smooth, while the number of replications for 500 and 1000 was smoother. Furthermore, Figure 5 shows that the prediction interval with 10000 replications appears to be

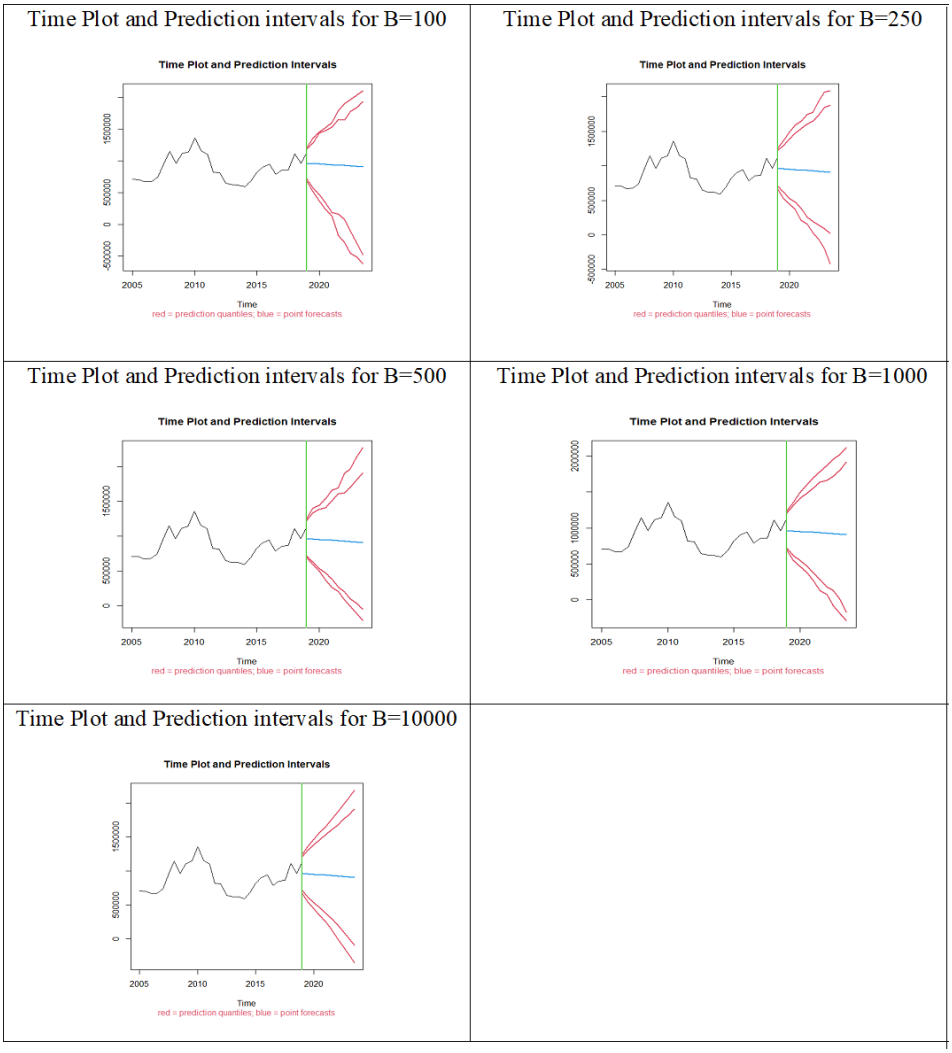

FIGURE 6. Time Plot and Prediction Intervals under Bootstrap Estimation

the smoothest. Therefore, the bootstrap technique has the ability to provide forecasting results with a better level of accuracy than traditional methods [7].

4. Conclusion

This study successfully used the bootstrap technique to forecast models dealing the problems arise due to small data series. It is also free from data distribution assumptions because it belongs to the category of the nonparametric approach. The time-series data on the number of educated unemployed people in Indonesia from 2005 to 2019 was used to determine the approach used. The results indicate that the number is projected to decline in the next 10 periods. Moreover, the number of bootstrap replications also impacts the forecasting accuracy, where the higher the number of replications used, the smoother and more accurate the resulting prediction intervals. Furthermore, the bootstrap technique is suitable for future value forecasting models compared to traditional techniques.

Aknowledgement

The authors thanks to head of center of study data science Department of Mathematics, Universitas Ahmad Dahlan.

Conflict of interest

The authors declare no conflict of interest.

References

[1] A. M. De Livera, R. J. Hyndman, and R. D. Snyder *Forecasting time series with complex seasonal patterns using exponential smoothing," J. Am. Stat. Assoc.* vol. 106, no. 496 (2011), 1513–1527.

[2] A. Staszewska-Bystrova, A. Staszewska-Bystrova, and A. Staszewska-Bystrova, *Bootstrap prediction bands for forecast paths from vector autoregressive models* Journal of Forecasting vol. 30, no. 8 (2011), 721-–735.

[3] B. Efron*Bootstrap Methods: Another Look at the Jackknife* Ann. Stat. vol. 7, no. 1 (1979), 1-–26.

[4] E. Paparoditis, *Sieve bootstrap for functional time series* the Annals of Statistics vol. 46, no. 6B (2018) 3510—3538.

[5] F. Harrou, A. Saidi, and Y. Sun *Wind power prediction using bootstrap aggregating trees approach to enabling sustainable wind power integration in a smart grid* Energy Convers. Manag. vol. 201 (2019), p.112077.

[6] G. Dikta and M. Scheer *Bootstrap Methods: with Application in R*, Springer, New York, 2021.

[7] G. Masarotto, *Bootstrap prediction intervals for autoregressions* International Journal of Forecasting vol. 6, no. 2 (1990) 229-–239.

[8] J. P. Kreiss, E. Paparoditis, and D. N. Politis, *On the range of validity of the autoregressive sieve bootstrap* Ann. Stat. vol. 39, no. 4 pp. 2103–2130, 2011.

[9] J. H. Kim, *Bootstrap-after-bootstrap prediction intervals for autoregressive models* Journal of Business & Economic Statistics vol. 19, no. 1 (2001), 117-–128.

[10] J. H. Kim, H. Song, and K. K. F. Wong, *Bias-corrected bootstrap prediction intervals for autoregressive model: new alternatives with applications to tourism forecasting* Journal of Forecasting vol. 29, no. 7 (2010) 655-–672.

[11] M. Chamdani, U. Mahmudah, and S. Fatimah *Prediction of Illiteracy Rates in Indonesia Using Time Series* Int. J. Educ. vol. 12, no. 1 (2019), 34-–41.

[12] M. La Rocca, F. Giordano, and C. Perna, *Clustering nonlinear time series with neural network bootstrap forecast distributions* International Journal of Approximate Reasoning vol. 137 (2021), 1-–15.
[13] M. P. Clements and J. H. Kim, *Bootstrap prediction intervals for autoregressive time series* Comput. Stat. Data Anal. vol. 51, no. 7 (2007), 3580-–3594.
[14] M. P. Clements and N. Taylor, *Bootstrapping prediction intervals for autoregressive models* Int. J. Forecast. vol. 17, no. 2 (2001), 247—267.
[15] M. R. M. R. Nieto, R. B. Carmona-BenÃtez, and R. B. Carmona-Benítez ARIMA+GARCH+Bootstrap forecasting method applied to the airline industry J. Air Transp. Manag. vol. 71, no. C (2018), 1-–8.
[16] M. R. M. R. Chernick and R. A. R. A. R. A. LaBudde, *An introduction to bootstrap methods with applications to R*, John Wiley & Sons, New York, 2014.
[17] N. A. Mobarakeh, M. K. Shahzad, A. Baboli, and R. Tonadre Improved forecasts for uncertain and unpredictable spare parts demand in business aircraft's with bootstrap method IFAC-PapersOnLine vol. 50, no. 1 (2017), 15241—15246. 2017.
[18] R. Errouissi, J. Cardenas-Barrera, J. Meng, E. Castillo-Guerra, X. Gong, and L. Chang *Bootstrap prediction interval estimation for wind speed forecasting* Proceeding of the 2015 IEEE Energy Conversion Congress and Exposition (ECCE),(Montreal,2015), 1919-–1924.
[19] R. J. R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.
[20] Y. S. Lee and S. Scholtes *Empirical prediction intervals revisited* Int. J. Forecast. vol. 30, no. 2 (2014), 217-–234.

Umi Mahmudah
Orcid number: 0000-0002-8793-322X
Department of Mathematics Education
IAIN Pekalongan
Pekalongan, Central Java, Indonesia
Email address: umi.mahmudah@iainpekalongan.ac.id

Sugiyarto Surono
Orcid number: 0000-0001-6210-7258
Department of Mathematics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
Email address: sugiyarto@math.uad.ac.id

Puguh Wahyu Prasetyo
Orcid number: 0000-0002-9188-2728
Department of Mathematics Education
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
Email address: puguh.prasetyo@pmat.uad.ac.id

Annisa E. Haryati
Orcid number: 0000-0002-0995-0665
Department of Mathematics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
Email address: annisa2007050015@webmail.uad.ac.id