Improvement of time series forecasting quality by means of multiple models prediction averaging

Denis Petrusevich[1]

[1] *MIREA – Russian Technological University, Moscow, 119454, Prospekt Vernadskogo, 78, Russia*

Abstract

Construction of time series models is usually based on Akaike and Bayes information criteria. Requirement of model simplicity is built inside the information criteria structure and the best fitted models aren't always best in terms of criteria values. Often there are a few best models that fit investigated time series well and there's problem of choice between them. Usually information criteria values allow to choose among them. But if one needs the best model by forecasts quality, best fitted models, or there are other thoughts a research has to choose and test models manually. At the same time when a few models are chosen it's possible to construct their combination. The simplest way is to count mean value of their forecasts and to use it as a combined prediction. Practical researches confirm that forecast error gets lower in this approach. Also, more complex construction than averaging of forecasts can be used (for example, weighted voting that is widely used in bagging technique in solution of classification problems). But this approach hasn't got enough theoretical base. From theoretical point of view confidence intervals of time series forecasts (usually here they're considered as prediction intervals) is also very complex task. Intervals tend to be very wide even for models with good prediction quality in terms of mean forecast errors. Thus, prediction intervals are rare in use. In this paper a few time series models are constructed for wage and income indices of Russian macroeconomic time series. Their predictions are combined into one forecast and it's quality is compared to individual ones. Transformation of forecast variance and prediction intervals in case of simple (moving averages MA(q) and autoregressions AR(p) of low order) models is also considered but is a part of further work.

Keywords

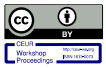
Time series forecasting, prediction averaging, ARIMA, information criteria.

1. Introduction

Time series modeling is usually based on information criteria evaluation (Bayes or Akaike criteria) in order to choose the best model for investigated time series [1-3]. In practice often a few models have got good criteria values and it's difficult to choose the best one. At the same time it's possible to make averaging of forecasts made with each model. Such technique is investigated in [4-7] and it has got good results. But this approach isn't confirmed well theoretically. Prediction quality is usually evaluated with confidence intervals (for example, in case of linear regression model). But in case of time series investigation usual 95% prediction intervals are too wide and are difficult for evaluation [1, 4, 8, 9]. Thus, there's no theoretical connection between forecast averaging and narrowing of prediction intervals. If used models are of the same type their prediction intervals remain the same and prediction interval of combined model isn't better. Also, it should be mentioned that prediction intervals behaviour in case of averaging of different models (for example, ARIMA and GARCH models) still isn't investigated. In this research practical experiments on time series models averaging are presented.

III International Workshop on Modeling, Information Processing and Computing (MIP: Computing-2021), May 28, 2021, Krasnoyarsk, Russia

EMAIL: petrdenis@mail.ru (Denis Petrusevich)
ORCID: 0000-0001-5325-6198 (Denis Petrusevich)



Prediction intervals behaviour for simple models (autoregression AR(p) models of low order, p<3, and moving average MA(q) models) and for their combination are in the scope of current research program and is going to be under further investigation.

2. Time series forecast errors and prediction intervals

One of the most widely used time series models is ARIMA that's constructed out of two parts: the p-order autoregression AR(p) model and the q-order moving average MA(q) model [1, 9]. They are used in combination with time series integration technique in order to get stationary time series. These models have got wide prediction intervals. They usually have got complex form (for example, in case of autoregression models AR) and are difficult to evaluate. In order to evaluate forecasts often mean square errors are used (formula (1):

$$MSE = \frac{\sum_t (\tau(t) - ts(t))^2}{N}, \quad (1)$$

or root mean square errors are implemented (formula (2)):

$$RMSE = \sqrt{\frac{\sum_t (\tau(t) - ts(t))^2}{N}}. \quad (2)$$

Here $\tau(t)$ denotes predicted values of the processed time series, $ts(t)$ is real values of the investigated time series, t enumerates all predicted time points and N is their quantity.

In practice also mean absolute error of forecast is used (formula (3):

$$MAE = \frac{\sum_t |\tau(t) - ts(t)|}{N}. \quad (3)$$

Thus, quality of forecasts at certain period is used instead of prediction intervals evaluation. At the same time there's idea of averaging forecasts made by means of a few models. If all the models make good forecasts it's obvious that combination of their forecasts should behave well also [4-7, 10]. But this result needs more thorough check and strict confirmation. Also, idea of connection between forecast quality improvement and prediction intervals narrowing should be tested. If forecasts become better, it should have some influence at prediction interval. Evaluation of prediction intervals of time series forecasters has been in scope of scientific research for a long time [8].

The idea to select the best models in some set their by predictions evaluating is considered in [11]. But there the best model is chosen. In this research predictions of a set of models are combined into one forecast. Prediction intervals should also be evaluated as well as prediction quality. Implementation of non-linear combination forecasters [12] is an idea of next stage of research in this area comparing to linear combination technique used in [4-7]. Competitive selection of best models by their prediction intervals quality is under investigation of [13]. So, this problem is in scope of modern research.

The tested method is close to bagging technique. In the classification problem at which bagging is used very often there can be constructed a lot of weak classifiers. Their accuracy must be higher than 50% but isn't supposed to be very high. This technique constructs strong classifier (with high accuracy) out of a lot of weak ones. Each weak classifier is constructed using its own training set and test set, its own set of variables (features). One of main examples of bagging principle implementation is random forest classifier. This approach also takes place in combining linear regression models into a new one with higher accuracy [14].

In this research quality of combination constructed out of ARIMA models [1, 9] is under investigation: their RMSE and MAE values (formulae (2) and (3)). First of all, mean value of all predictors' forecasts is counted:

$$\tau(t+1) = \frac{\sum_N f(t)}{N}. \quad (4)$$

Here $\tau(t+1)$ denotes predicted value of the processed time series, $f(t)$ are forecasters' predictions and N is their quantity.

At the same time voting technique used in random forests classifiers [14] is tested. At each experiment time series forecaster makes some error (difference between forecast at time point and its real value). Level of confidence to each model is constructed into the model (4). Weight *w* of model's vote is increased if its prediction is close to real value and it's decreased if other models make better predictions:

$$\tau(t+1) = \sum_N w_i f_i(t), \quad \sum_N w_i = 1. \tag{5}$$

The same approach can be used in a slightly different way. The models themselves can be combined into a new one. At the first stage standard models are tested at training set. After that the best models *m(t)* by their forecasts at test set are combined into a new model *M(t)* with use of averaging:

$$M(t) = \frac{\sum_K m(t)}{K}. \tag{6}$$

Here K is number of best models used in a combination. This approach can also be transformed with voting technique implementation:

$$M(t) = \sum_K w_i m_i(t), \quad \sum_K w_i = 1. \tag{7}$$

Here *w* is weight of each vote. Weights change at training stage. Model's weight is increased when its forecast is close to reality and is decreased otherwise.

Classifiers used in [14] should be trained at different sets of data. In time series models this requirement can be reinterpreted in use of different training periods for each time series model in combination and also in implementation of different by structure time series predictors. If ARIMA (p, d, q) models are considered, models with varying p and q parameters should be used. Thus, models with close orders p and q are going to make close predictions and they can be considered as dependent. If models have got different structure (varying p and q in large intervals) these models' predictions tend to independence. Also, predictions of different models by type (for example, GARCH and ARIMA) should have good results and are supposed to be implemented. But it's task for further investigation because prediction intervals in such combinations can't be constructed with usual statistics. There's some research that can be used as basis for GARCH models testing in this approach [15]. Is one takes into account that information criteria values depend not only on model quality (as difference between model's prediction and real value) but requirement of model simplicity is also used [1], other metrics of time series quality can be used: for example, only likelihood function value of each time series. At this stage even game-based techniques [16, 17] can be used.

3. Experiments

The Dynamic series of macroeconomic statistics of the Russian Federation (monthly wage index and income index) [18] have been handled in the experiment section. Last 12 values were used as test period predicted by models. Good ARIMA (p, d, q) models by RMSE and MSE of their forecasts were chosen to construct combined models. These models aren't trained with regard to information criteria values or likelihood function level. Their coefficients are mean values of appropriate coefficients of the chosen set of models (in accordance with expression (6)). Also mean value of all chosen models' predictions is obtained and compared to real values.

3.1. ARIMA (p, d, q) models of wage and income indices of Russian macroeconomic statistics and combination of models

The ARIMA (p, d, q) models of orders p<4 and q<4 have been tested for the monthly income index [18] of Russian macroeconomic statistics. According to automatic fitting function that's based on time differentiation and on choice of the best model by means of information criteria values [1, 9], the ARIMA (2, 1, 2) is best fitted model. In this experiment other "good models" are defined as models that have got good forecasts. At the same time it's necessary to select models with varying structure. So, if there are two models with orders differing only by 1 and they've got close values of coefficients, only one of these models is selected for further experiment. These models are presented in the Table 1. Columns contain Akaike information criterion values, RMSE and MAE of forecasts at period of 12 months.

Table 1

The ARIMA (p, d, q) models of the income index

ARIMA(p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(0, 1, 1)	2775.14	30.12	17.13
ARIMA(0, 1, 3)	2754.3	30.59	19.10
ARIMA(1, 1, 2)	2741.58	30.21	18.10
ARIMA(2, 1, 2)	**2742.66**	**30.29**	**18.71**
ARIMA(3, 1, 1)	2753.03	30.47	19.14
Mean forecast	-	30.22	18.18
Combined model	-	26.36	15.84

It's clearly seen that result of automatic fitting procedure (ARIMA (2, 1, 2)) isn't the best model according to errors of forecasts. The best model according to automatic fitting procedure is marked with bold font. Mean forecast of all these models' predictions is presented in the end of the table. It's better than forecasts of some other models and it can be used in practice. Quality of the forecast made by the combined model is shown in the last row. It's much closer to the real values than all other models shown above.

Appropriate models of wage index [18] are shown in the Table 2.

Table 2

The ARIMA (p, d, q) models of the wage index

ARIMA (p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(0, 1, 3)	2509.22	21.60	12.66
ARIMA(1, 1, 1)	2520.11	22.63	13.51
ARIMA(1, 1, 2)	2522.29	22.90	13.72
ARIMA(2, 1, 1)	2515.67	22.25	12.70
ARIMA(2, 1, 2)	2508.03	21.43	12.88
ARIMA(3, 1, 2)	**2472.72**	**21.54**	**17.61**
Mean forecast	-	20.93	12.29
Combined model	-	20.64	14.11

Here the combined model and mean forecast are closer to real values than other ARIMA models. Mean forecast has got the best forecast according to MSE value and the combined model's forecast is the best according to RMSE value.

Plots of predictions made by the best fitted model ARIMA(3, 1, 2), the combined model, mean forecast of five models shown in the Table 2 and real values are presented as Figure 1:, Figure 2:, Figure 3:, Figure 4:.

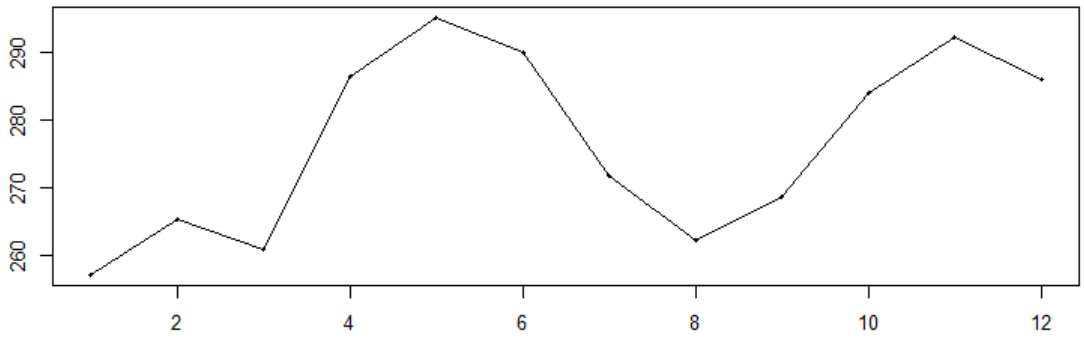


Figure 1: Prediction of the wage index by the best fitted model ARIMA(3, 1, 2) (12 months)

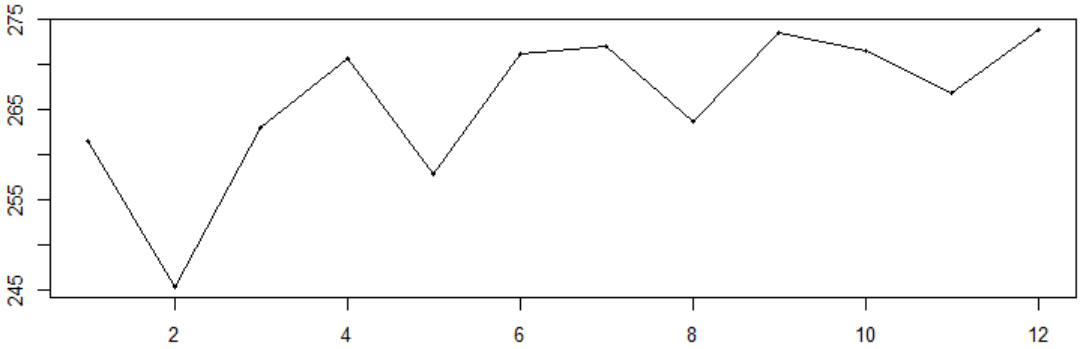


Figure 2: Prediction of the wage index by the combined model (12 months)

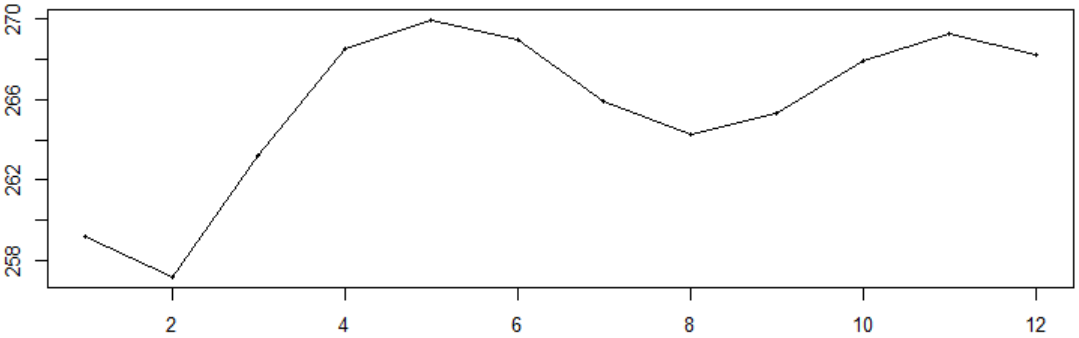


Figure 3: Prediction of the wage index by the combined model (12 months)

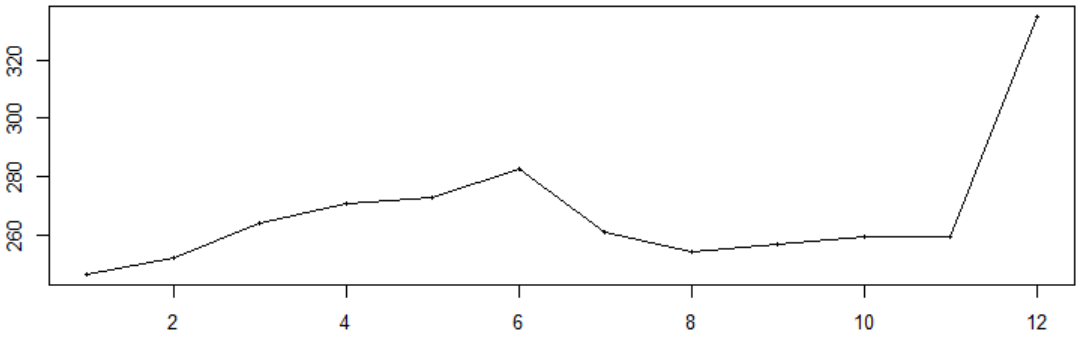


Figure 4: Real values of the wage index at the predicted period

3.2. Combination of ARIMA (p, d, q) models trained with voting technique

The voting technique based on expression (7) is also used analyzing the monthly wage index [18] of Russian macroeconomic statistics. 2/3 of training set was used to construct ARIMA models. Five "good" models were chosen to use them at the next stage. Then equal weights for their votes were added into the model. Last 1/3 part of the training set was used to implement voting technique. All chosen models made prediction for one month ahead. Their predictions were compared to the real data. Weight of the model with the best prediction was increased. Other weights were decreased. Their sum must equal to 1. At the same time it's necessary to control that weights shouldn't become less than zero.

In case of the income index the ARIMA (1, 1, 0), ARIMA (2, 1, 0), ARIMA (3, 1, 0) models were chosen. Weights of the combination: 0.15 for the ARIMA (1, 1, 0) model, 0.03 for the ARIMA (2, 1, 0) model and 0.82 for the ARIMA (3, 1, 0) one. This result itself can be used for evaluation of models and their predictions.

Unfortunately, variance of long-term prediction is very high and it's impossible to compare constructed model with models of previous section trained at 100% of data. But one can compare it with the models used in this combination. Though in the «voting» model data of the predicted period have been implemented as training set. So, evaluation method for such models is still to be discussed and constructed. Forecasts have been made for 12 months ahead. Their comparison is shown in the Table 3.

Table 3

The ARIMA (p, d, q) models of the income index and the model trained with voting technique

ARIMA (p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA (1, 1, 0)	2588.28	26.08	15.64
ARIMA (2, 1, 0)	2547.15	26.09	15.70
ARIMA (3, 1, 0)	2536.94	25.93	15.52
ARIMA (2, 1, 2)	1678.15	28.12	18.20
Voting model	-	25.93	15.52

In the voting model weight of the ARIMA (3, 1, 0) model is maximal and they give close predictions. Their forecasts are even better than the automatically fitted function prediction.

In case of the wage index the ARIMA (0, 1, 3), ARIMA (1, 1, 3), ARIMA (2, 1, 0), ARIMA (2, 1, 2) and ARIMA (3, 1, 0) models were chosen. Because of "winner takes it all" principle usage at the end of the voting training stage there are only two models with non-zero weights: weight of the ARIMA (0, 1, 3) is about 0.2 and weight of the ARIMA (2, 1, 0) model is about 0.8. Comparison of the voting model for the wage index with the chosen models are shown in the Table 4 (the best fitted model hasn't been shown because that's ARIMA (0, 1, 1) and it can't make stable predictions).

Table 4

The ARIMA (p, d, q) models of the wage index and the model trained with voting technique

ARIMA (p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(0, 1, 3)	1493.54	17.69	11.33
ARIMA(1, 1, 3)	1495.42	17.75	11.46
ARIMA(2, 1, 0)	1493.33	17.54	11.11
ARIMA(2, 1, 2)	1495.77	17.68	11.28
ARIMA(3, 1, 0)	1494.76	17.61	11.16
Voting model	-	14.79	10.03

One can conclude that this voting model has shown the best result. Usually conclusions made by researcher are based on consideration of a lot of models and of their forecasts. So, such approach can give one more model in such set and here this model has become the best one by prediction quality.

Predictions of the ARIMA (2, 1, 0) model, voting model at period of 12 months and real values of the wage index at the predicted period are shown at Figure 5:Figure 6:,Figure 7:.

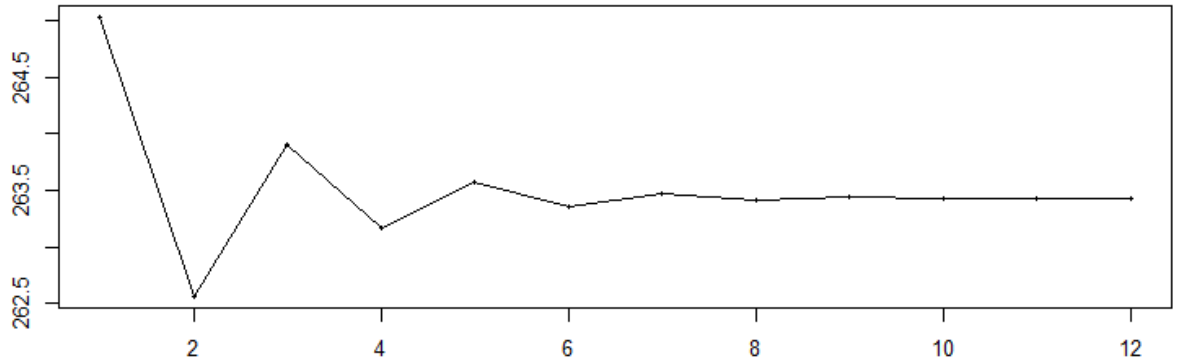


Figure 5: Prediction of the wage index by the model ARIMA (2, 1, 0) (12 months)

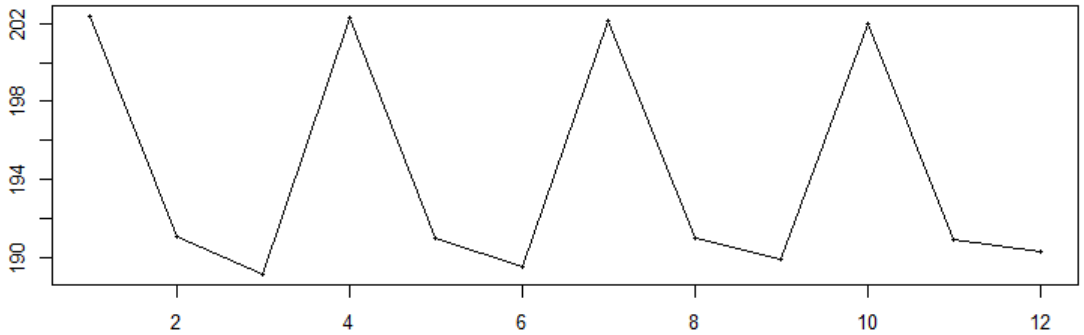


Figure 6: Prediction of the voting model (12 months)

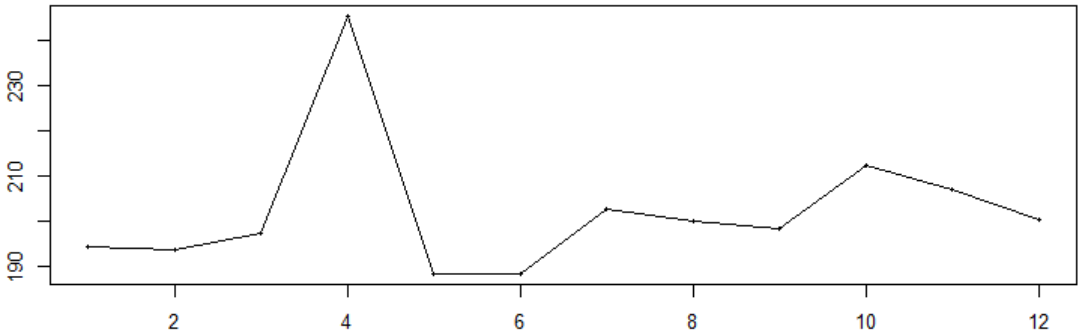


Figure 7: Real values of the wage index at the predicted period

4. Conclusion

In this research three ways of combining ARIMA models into a new one is investigated. The main goal is to construct a new model that's going to make better forecasts or predictions of the same quality. Russian macroeconomical statistics was used in the experimental part of the research.

The first method of averaging is just constructing mean forecast out of predictions of a few good models. Their quality is confirmed with prediction at test period. This approach has been tested in the first experiment and it has got good results that are close to the best models.

The second method is averaging of the best models. The model with maximal orders p and q inside of set of the best ones (ARIMA (p, d, q) models) is constructed. Its coefficients are mean values of

appropriate coefficients of this set. Combined model makes the best prediction in case of income index and good one in case of the wage index.

The third type of the models is based on implementation of voting technique. First of all, good models are chosen while comparing quality of their predictions. At the first stage of votes evaluation all of them are equal and this model is the same like in the second case. But then if a model has got forecast closest to reality its weight is increased, weights of other models are decreased. Here, only winners' weight is increased. But in further research "soft" methods implementation (like softmax method) are going to be investigated. Combined models are the best ones by prediction quality in the both experiments.

Also, prediction intervals of combined models are in scope of further research. One can conclude that combination of models has got forecast of the same quality or even better. This conclusion is confirmed in [4-7] and in this paper. But narrowing of prediction intervals for combination of models is still investigated [12, 13] and is in the scope of further research.

Forecasting with combination of time series models is close to bagging technique [14] used in classification and regression tasks. But there are requirements to weak classifiers. Analogical requirements should be formulated and tested for time series models combined into a set. This problem is going to be under investigation in further work.

5. References

[1] R. J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia, 2018.

[2] E. G Andrianova, S. A. Golovin, S. V. Zykov, S. A. Lesko, E. R. Chukalina, Review of modern models and methods of analysis of time series of dynamics of processes in social, economic and socio-technical systems, Russian Technological Journal (In Russ) 8.4 (2020) 7-45. doi: 10.32362/2500-316X-2020-8-4-7-45.

[3] D. Petrusevich, Time series forecasting using high order ARIMA functions, in: Proceedings of the International Multidisciplinary Scientific GeoConference: SGEM 2019, Volume 19(2.1), Sofia, Bulgaria, 2019, pp.673-679. doi: 10.5593/sgem2019/2.1/S07.088.

[4] F. Petropoulos, R. J. Hyndman, C. Bergmeir, Exploring the sources of uncertainty: Why does bagging for time series forecasting work? European Journal of Operational Research 268.2 (2018) 545-554. doi: 10.1016/j.ejor.2018.01.045.

[5] M. H. Dal Molin Ribeiro, L. dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Applied Soft Computing 86 (2020) 105837. doi: 10.1016/j.asoc.2019.105837.

[6] K. Chen, Y. Peng, S. Lu, B. Lin, X. Li, Bagging based ensemble learning approaches for modeling the emission of PCDD/Fs from municipal solid waste incinerators, Chemosphere 274 (2021) 129802. doi: 10.1016/j.chemosphere.2021.129802.

[7] M. Yang, J. Zhang, H. Lu, J. Jin, Regularized ELM bagging model for Tropical Cyclone Tracks prediction in South China Sea, Cognitive Systems Research 65 (2021) 50-59. doi: 10.1016/j.cogsys.2020.09.005.

[8] C. Chatfield, Prediction Intervals for Time Series Forecasting, Principles of Forecasting (2001) 475-494. doi: 10.1007/978-0-306-47630-3_21.

[9] J. H. Stock, M. W. Watson, Introduction to Econometrics, Pearson, 2019.

[10] S. S. Sarkisov, N. V. Lomonosova, A. V. Zolkina, T. S. Sarkisov, Integration of digital technology in mining and metallurgy industries, Tsvetnye Metally (in Russ) 3 (2020) 7-14. doi: 10.17580/tsm.2020.03.01.

[11] P. Hansen, A. Lunde, J. Nason, Model confidence sets for forecasting models, Econometrica 79.2 (2005): 453-497.

[12] W. Chen, H. Xu, Z. Chen, M. Jiang, A novel method for time series prediction based on error decomposition and nonlinear combination of forecasters, Neurocomputing 426 (2021) 85-103. doi: 10.1016/j.neucom.2020.10.048.

[13] E. Meira, F. L. C. Oliveira, J. Jeon, Treating and Pruning: New approaches to forecasting model selection and combination using prediction intervals, International Journal of Forecasting 37.2 (2021) 547-568. doi: 10.1016/j.ijforecast.2020.07.005.
[14] L. Breiman, Random forests, Machine Learning 45 (2021) 5-32. doi: 10.1023/A:1010933404324.
[15] S. Pellegrini, E. Ruiz, A. Espasa, Prediction intervals in conditionally heteroscedastic time series with stochastic components, International Journal of Forecasting 27.2 (2011) 308-319. doi: 10.1016/j.ijforecast.2010.05.007.
[16] A. V. Zolkina, N. V. Lomonosova, D. A. Petrusevich, Gamification as a tool of enhancing teaching and learning effectiveness in higher education: needs analysis, Science for Education Today 3 (2020) 127-143. doi: 10.15293/2658-6762.2003.07.
[17] O. Osipova, N. Lomonosova, Application of online courses in higher education system, in: Proceedings of the International Multidisciplinary Scientific GeoConference: SGEM 2019, Volume 19(5.4), Albena, Bulgaria, pp. 49-54, 2019.
[18] Dynamic series of macroeconomic statistics of the Russian Federation. Wage index, income index. Retrieved from (the 21nd of March 2021). URL: http://sophist.hse.ru/hse/nindex.shtml