Modeling of forecasts variance reduction at multiple time series prediction averaging with ARMA (1, q) functions

Danila Musatov [1] and Denis Petrusevich [1]

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

Abstract

Combination of time series forecasts is usually considered as good technique in practice. But it has got weak theoretical explanation. In this research variance of time series forecasts and variance of combined models are considered. One is interested in the view of variance of forecasts function over one, two and three periods. Conditions which can lead to improvement of averaged time series predictions are in scope of this research. In this paper a few examples of the most popular time series models are observed: the moving average models MA(q), the autoregressions AR(p) models and their combination in the form of ARIMA(p, d, q) or ARMA(p, q) model. In particular, AR(1) and ARMA(1, q) are investigated. Nowadays there are researches about time series averaging. Approaches based on bagging and boosting are implemented very often in classification and regressions. It's very appealing to use such strategy in tine series modeling. At the same time it's easy to construct learning set and test set in classification tasks. But it's a complex task in case of time series processing. There's a need of two sets: to train time series models and to construct their combinations. Thus nowadays combination of time series models, combination of their forecasts or of their prediction intervals are in scope of view of a few complex researches. In this paper we investigate behaviour of time series predictions' variance in order to have another useful approach in time series prediction averaging. Russian macroeconomical time series statistics is used as experimental time series.

Keywords

Time series forecasting, prediction averaging, ARIMA, forecast variance, information criteria

1. Introduction

Mathematical models used to predict certain value are often used in combinations. The most simple combination function in case of time series processing is just averaging of all models' predictions or selection of the best one [1]. It's often widely believed that if averaged models are "good" enough and reflect some part of described process' behaviour their averaging also leads to a0 "good" model. Though, mathematical statement of this problem isn't researched well. There are attempts to choose the best model [1], to construct mean model [2], to implement bagging strategy to processed models of certain time series [3, 4]. Selection of the best model is traditional way but in practice there's always a lot of models and there's no mathematically strict way to choose the best one. Then researchers thought that multiple models can describe various sorts of processed time series behaviour from various points of view and thus their combination is better than selection of the best one [5-7]. Bagging [8] of time series is very appealing but complex task. Usually time series models are built in two stages: there's a training set used to construct them and a test set to evaluate their quality. But in case of bagging one needs three parts or training set should be subdivided in some parts to use them in construction of the models and in evaluation of their combinations [9-11]. In this research combination of forecasts is

Proceedings of MIP Computing-V 2022: V International Scientific Workshop on Modeling, Information Processing and Computing, January 25, 2022, Krasnoyarsk, Russia
EMAIL: danilamusatov20@mail.ru (Danila Musatov), petrdenis@mail.ru (Denis Petrusevich)
ORCID: 0000-0003-0673-5393 (Danila Musatov); ORCID: 0000-0001-5325-6198 (Denis Petrusevich)

investigated from the prediction intervals point of view. Behaviour of prediction intervals is often out of scope of research in a lot of ensemble models but this question is especially important in case of time series forecasting [11-12]. In the paper ARIMA models are considered. Though some steps considering combination of GARCH models (in case if time series is heteroscedastic) have already been done in [7], combination of heterogeneous models are still in scope of future work. These models are the most popular but sometimes frequential analysis is also a good tool [13, 14], so further analysis should consider combinations of such models. Here we get variance of time series forecasts at 1, 2, 3 periods of time and investigate its behaviour in case of model averaging.

2. Time series prediction intervals

In order to evaluate prediction interval one has to transform time series into MA(∞) form. According to Wald theorem [15] this transformation (also called psi-representation) can be implemented to any stationary time series. If one treats ARIMA models, stationary process can be achieved with time series differentiation [16]. Each process in such view can be expressed via infinite (in common case) summation of moving averages MA(q). Coefficients ψ_j in this series are usually called psi-weights. .
Thus, the simplest case is, of course, handling moving average MA(q) models:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}.$$

Here today value of the X time series is expressed via white noises ε_{t-q} of order less than or equal to q [16]. These models are already considered in the form when all coefficients are equal to psi-weights. Thus, variance of these models' predictions [16] is (1):

$$Var(\hat{x}_n - x_n) = \sigma^2 \sum_{j=0}^{n-1} \psi_j^{\,2}, \tag{1}$$

here $\hat{x}_n$ is a predicted value at time n, x_n is a real value of time series, σ is a standard error value (got at learning procedure) and $\psi_j = \theta_j$ for MA(q) series. If the assumption of normally distributed errors is hold, a 95% prediction interval of $\hat{x}_n$ is $[\hat{x}_n - 1.96\sqrt{Var(\hat{x}_n - x_n)}, \hat{x}_n + 1.96\sqrt{Var(\hat{x}_n - x_n)}]$. Here variance can be found via psi-weights be means of (1).

Another significant part of time series models consists of the autoregression AR(p) models. In this case today value of the X time series depends on its own past values of order less than or equal to p:

$$X_t = c + \varphi_1 X_{t-1} + \ldots + \varphi_q X_{t-p}.$$

Consideration of Wald's theorem to autoregression processes AR(p) leads to (2):

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 &= \varphi_1, \\ \psi_2 &= \varphi_1 \psi_1 + \varphi_2, \\ \psi_3 &= \varphi_1 \psi_2 + \varphi_2 \psi_1 + \varphi_3. \end{aligned} \tag{2}$$

These expressions are recurrent and can be reinterpreted in term of summation (3):

$$\psi_k = \sum_{i=1}^{k} \varphi_i \psi_{k-i}. \tag{3}$$

And in common case handling ARIMA (p, d, q) models (or ARMA(p, q)) consideration of psi-weights leads to (4):

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 &= \varphi_1 + \theta_1, \\ \psi_2 &= \varphi_1 \psi_1 + \varphi_2 + \theta_2, \\ \psi_3 &= \varphi_1 \psi_2 + \varphi_2 \psi_1 + \varphi_3 + \theta_3. \end{aligned} \tag{4}$$

Thorough analysis of these expressions can be found in [16, 17].

In this research we consider AR(p) and ARMA(p, q) models with p<3. The main goal is to express variance via terms of the models and to explain conditions at which averaging technique is going to lead its improvement. Here only averaging of the models is taken into account. But the same approach can be used in case of bagging and non-linear combinations of models. In further calculations and at averaging stage of experiments all models are supposed to have the same model of seasonality because this part is non-linear and its summation will lead to models with another complex seasonality.

2.1. Variance of the prediction made with AR(1) model

First of all, here AR(p) models are analyzed. Considering the simplest case of AR(1) model psi-weights (2) are elements of geometric progression:

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 &= \varphi_1, \\ \psi_2 &= \varphi_1\psi_1 = \varphi_1^2, \\ \psi_k &= \varphi_1\psi_{k-1} = \varphi_1^n. \end{aligned} \tag{5}$$

Variance of its prediction at time n is proportional to sum of geometric progression with ratio equal to φ_1^2:

$$Var_{AR(1)}(\hat{x}_n - x_n) = \sigma^2 \sum_{j=0}^{n-1} \psi_j^2 = \sigma^2 (1 + \sum_{j=1}^{n-1} \varphi_1^{2j}) = \sigma^2 \frac{1-\varphi_1^{2n}}{1-\varphi_1}. \tag{6}$$

This sum (6) gets finite if $|\varphi_1| < 1$ even if $n \to \infty$. Thus, in case of AR(1) process variance of the forecast over infinite period can be finite number. Logically it's close to the case of fluctuations with descending amplitude (7).

$$\lim_{n\to\infty} Var_{AR(1)}(\hat{x}_n - x_n) = \sigma^2 \frac{1}{1-\varphi_1}. \tag{7}$$

Variance of prediction over 1, 2 and 3 timesteps is presented in expressions (8):

$$\begin{aligned} Var_{AR(1)}(\hat{x}_1 - x_1) &= \sigma^2, \\ Var_{AR(1)}(\hat{x}_2 - x_2) &= \sigma^2 (1 + \varphi_1^2), \\ Var_{AR(1)}(\hat{x}_3 - x_3) &= \sigma^2 (1 + \varphi_1^2 + \varphi_1^4). \end{aligned} \tag{8}$$

Variance of prediction over 1 timestep ahead depends only on the standard error of model under investigation. Plots of these functions for predictions over 2 and 3 timesteps are presented in Figure 2, Figure 3. They are bulging downward (all degrees are even and there are only plus signs in the expression (8)). There's only one minimum (which is global one) in case when $\varphi_1 = 0$ and its prediction is also zero. If there are a few models their prediction variances can be various points at plots 1, 2. If predictions of a few models AR(1) are averaged, one can describe it as averaging of the very models (because the ARMA(p, q) model is linear). Thus, averaged model can be considered as a new ARMA(p, q) model with the same orders p, q but different coefficients.

Variance of combined model is lower than variance of the source ones if one moves towards minimum at plots 1, 2. If there are two models with variances situated at the same side from zero, variance of their combination will be situated between them. So, in the case of two models one of them is going to have higher variance and one of them is going to have lower variance than variance of the combined model. At the same time the combined model is going to have lower level of variance if variances of the source models are situated from different sides of zero.

Basically, this discourse is true in a case of a lot of models. The variance function (8) is bulging downward and it has got only one minimum. So, if all models are marked as points at plots 1,2, variance of averaged model is the same function of averaged parameter. Because variance function is bulging downward, point marking variance of averaged model is situated under line connecting models with extreme values of parameter φ_1 (minimum and maximum).

$$\min VarM_i \le Var\frac{\sum_i^N M_i}{N} \le \max VarM_i, \tag{9}$$

here M_i denotes enumeration of N models of AR(1) type.

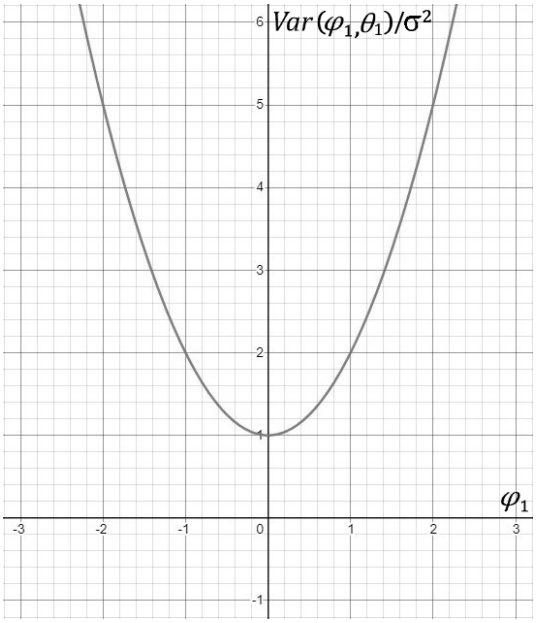

Figure 1: Forecast variance of AR(1) model at 2 timesteps ahead

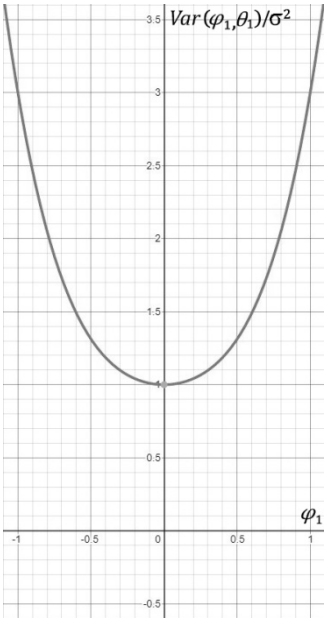

Figure 2: Forecast variance of AR(1) model at 3 timesteps ahead

But this combination gets the best model (with lower variance) only if equal numbers of models are situated to the left of zero and to the right of it. This situation can be seen at Figure 3. Thus, averaging leads to the best model (usually to a "good" model because there's too much of them) only if all models are divided into two equal parts: with negative value of φ_1 and positive ones. Basically, it means that half of models predict that value of time series is going to be lower and the other ones predict tendency to grow. Such situation can take place if investigated time series has got complex behaviuor of researchers haven't got enough data to make prediction. In case of "usual" situation there's always some model better than averaged one (with lower variance). If there's no tool to choose the best model, the

averaged one is better than worse models. So, it can be used as another tool for prediction and it has got "good" quality.

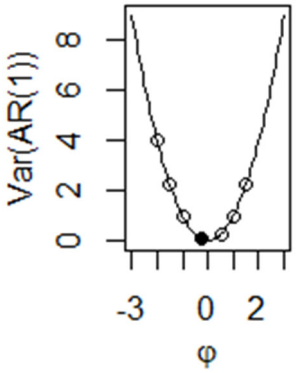

Figure 3: Variance of handled AR(1) models (shown with transparent dots) and variance of averaged model (black dot)

2.2. Variance of the prediction made with ARMA(1, q) model

Psi-weights of ARMA(1, q) models form geometric progression starting from time q:

$$\begin{aligned} &\psi_0 = 1, \\ &\psi_1 = \varphi_1 + \theta_1, \\ &\psi_2 = \varphi_1\psi_1 + \theta_2 = \varphi_1(\varphi_1 + \theta_1) + \theta_2, \\ &\psi_{q+1} = \varphi_1\psi_q. \end{aligned} \tag{10}$$

Thus, in case if order of moving average part is less than time of prediction ($q \le n$) there's no geometric progression. For example, considering variance of the ARMA(1, 2) model for 3 timesteps, one has got:

$$Var_{ARMA(1,2)}(\hat{x}_3 - x_3) = \sigma^2 \sum_{j=0}^{2} \psi_j{}^2 = \sigma^2 \left[1 + (\varphi_1 + \theta_1)^2 + (\varphi_1(\varphi_1 + \theta_1) + \theta_2)^2 \right]. \tag{11}$$

For q>n time series variance includes sum of geometric progression starting from time q:

$$\begin{aligned} &Var_{ARMA(1,q)}(\hat{x}_n - x_n) = \sigma^2 \sum_{j=0}^{n-1} \psi_j{}^2 = \\ &\sigma^2 \left[1 + (\varphi_1 + \theta_1)^2 + (\varphi_1(\varphi_1 + \theta_1) + \theta_2)^2 + \ldots + \psi_{q-1}{}^2 + \psi_q{}^2 \frac{1 - \varphi_1^{2n}}{1 - \varphi_1} \right]. \end{aligned} \tag{12}$$

Here, the last term denotes sum of the progression starting at q-th psi-function.

2.3. Variance of the prediction made with ARMA(1, 1) model

Variances of predictions over 1, 2 and 3 timesteps of ARMA(1, 1) model are shown at (13):

$$\begin{aligned} &Var_{ARMA(1,1)}(\hat{x}_1 - x_1) = \sigma^2, \\ &Var_{ARMA(1,1)}(\hat{x}_2 - x_2) = \sigma^2 \left[1 + (\varphi_1 + \theta_1)^2 \right], \\ &Var_{ARMA(1,1)}(\hat{x}_3 - x_3) = \sigma^2 \left[1 + (\varphi_1 + \theta_1)^2 + \varphi_1^2(\varphi_1 + \theta_1)^2 \right]. \end{aligned} \tag{13}$$

First of all, one considers variance of prediction over 2 timesteps ahead. Extreme value of this function is situated at the point where first derivatives are zeros:

$$(Var_{ARMA(1,1)}(\hat{x}_2 - x_2))'|_{\varphi_1} = (Var_{ARMA(1,1)}(\hat{x}_2 - x_2))'|_{\theta_1} = 2\sigma^2(\varphi_1 + \theta_1) = 0. \quad (14)$$

It happens if $\varphi_1 = -\theta_1$, predictions of this model ($X_t = c + \varphi_1 X_{t-1} - \varphi_1 \varepsilon_{t-1}$) and its variance are minimal (it can be seen at plot 4). To check whether this is minimum, one uses the second derivatives:

$$(Var_{ARMA(1,1)}(\hat{x}_2 - x_2))''|_{\varphi_1^2} = (Var_{ARMA(1,1)}(\hat{x}_2 - x_2))''|_{\theta_1^2} = (Var_{ARMA(1,1)}(\hat{x}_2 - x_2))''|_{\varphi_1\theta_1} = 2\sigma^2. \quad (15)$$

Hessian of variance is zero:

$$H_{ARMA(1,1)}(\hat{x}_2 - x_2) = \begin{vmatrix} 2\sigma^2 & 2\sigma^2 \\ 2\sigma^2 & 2\sigma^2 \end{vmatrix} = 0. \quad (16)$$

$\Delta_1 = 2\sigma^2, \Delta_2 = 0$. One can't make definite conclusions by means of Sylvester's theorem. It means that the function hasn't got one minimum. But taking a glance at Figure 4 one can see that minimum is achieved at models with $\varphi_1 = -\theta_1$. At the same time this function is also bulging downward and results for the AR(1) are also true in this case.

Using the same approach we've considered variance of prediction in three timesteps. Extreme values of variance are obtained equalizing first derivatives to zero:

$$(Var_{ARMA(1,1)}(\hat{x}_3 - x_3))'|_{\varphi_1} = 2\sigma^2(\varphi_1 + \theta_1)(2\varphi_1^2 + \varphi_1\theta_1 + 1) = 0, \quad (17)$$

$$(Var_{ARMA(1,1)}(\hat{x}_3 - x_3))'|_{\theta_1} = 2\sigma^2(\varphi_1 + \theta_1)(\varphi_1^2 + 1) = 0.$$

Here one can see the same result as in the case of prediction over two timesteps. The second equation in (17) has got solution $\varphi_1 = -\theta_1$.

The second derivatives for prediction of ARMA(1, 1) model over three timesteps are:

$$(Var_{ARMA(1,1)}(\hat{x}_3 - x_3))''|_{\varphi_1^2} = 2\sigma^2(6(\varphi_1 + \theta_1) + \theta_1^2 + 1), \quad (18)$$

$$(Var_{ARMA(1,1)}(\hat{x}_3 - x_3))''|_{\theta_1^2} = 2\sigma^2(\varphi_1^2 + 1),$$

$$(Var_{ARMA(1,1)}(\hat{x}_3 - x_3))''|_{\varphi_1\theta_1} = 2\sigma^2(3\varphi_1^2 + 2\varphi_1\theta_1 + 1).$$

In case if $\varphi_1 = -\theta_1$, all derivatives in (18) are equal to $2\sigma^2(\varphi_1^2 + 1)$.

Hessian has got the form:

$$H_{ARMA(1,1)}(\hat{x}_3 - x_3) = \begin{vmatrix} 2\sigma^2(6(\varphi_1 + \theta_1) + \theta_1^2 + 1) & 2\sigma^2(3\varphi_1^2 + 2\varphi_1\theta_1 + 1) \\ 2\sigma^2(3\varphi_1^2 + 2\varphi_1\theta_1 + 1) & 2\sigma^2(\varphi_1^2 + 1) \end{vmatrix}. \quad (19)$$

Hessian Δ_2 and minor Δ_1 have got complicated form. But in the case of extreme $\varphi_1 = -\theta_1 : \Delta_1 = 2\sigma^2(\theta_1^2 + 1) \geq 0, \Delta_2 = 0$. One can't confirm that case of $\varphi_1 = -\theta_1$ leads to the model with minimal variance. But it's clearly seen at the Figure 5.

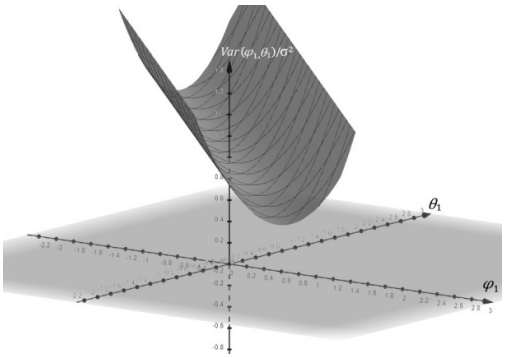

Figure 4: Forecast variance of ARMA(1, 1) model at 2 timesteps ahead

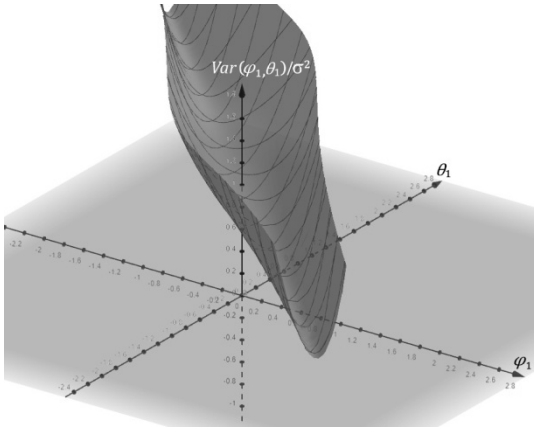

Figure 5: Forecast variance of ARMA (1, 1) model at 3 timesteps ahead (13)

2.4. Variance of the prediction made with ARMA(1, 2) model

Variances of predictions over 1, 2 and 3 timesteps of ARMA(1, 1) model are shown at (20):

$$\begin{aligned} &Var_{ARMA(1,2)}(\hat{x}_1 - x_1) = \sigma^2, \\ &Var_{ARMA(1,2)}(\hat{x}_2 - x_2) = \sigma^2\left[1+(\varphi_1+\theta_1)^2\right], \\ &Var_{ARMA(1,2)}(\hat{x}_3 - x_3) = \sigma^2\left[1+(\varphi_1+\theta_1)^2+(\varphi_1(\varphi_1+\theta_1)^2+\theta_2)^2\right]. \end{aligned} \tag{20}$$

Variance of forecast over 1 and 2 timesteps are the same as in the case of ARMA(1,1) model (13). So, all calculations and conclusions made there remain true. The same result is repeated in case of ARMA(1, q) models, q>2. For predictions over 1, 2 and 3 timesteps variance functions will be the same as in (20). Differences are expected only for forecasts over more timesteps. In case of prediction over 3 timesteps extreme point is situated in the place where:

$$\begin{aligned} &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))'|_{\varphi_1} = 2\sigma^2\left[\varphi_1+\theta_1+(\varphi_1(\varphi_1+\theta_1)+\theta_2)(2\varphi_1+\theta_1)\right] = 0, \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))'|_{\theta_1} = 2\sigma^2\left[\varphi_1+\theta_1+2\varphi_1(\varphi_1(\varphi_1+\theta_1)+\theta_2)\right] = 0, \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))'|_{\theta_2} = 2\sigma^2\left[\varphi_1(\varphi_1+\theta_1)+\theta_2\right] = 0 \end{aligned} \tag{21}$$

Substituting the last equation into the first one in the system (21), one gets $\varphi_1 = -\theta_1$ and after that from the last equation one has got $\theta_2 = 0$. So, structure of the model that has got minimal variance of prediction is the same as in cases of ARMA(1, 1) and of ARMA(1, 2) but for prediction at 2 timesteps. Determinant of hessian gets very large and here only values of the second derivatives are shown in expression (22):

$$\begin{aligned} &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))''|_{\varphi_1^2} = 2\sigma^2\left[1+2\left(\varphi_1(\varphi_1+\theta_1)+\theta_2\right)+\left(2\varphi_1+\theta_1\right)^2\right], \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))''|_{\theta_1^2} = 2\sigma^2(\varphi_1^2+1), \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))''|_{\theta_2^2} = 2\sigma^2, \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))''|_{\varphi_1\theta_1} = 2\sigma^2\left[1+\varphi_1\left(2\varphi_1+\theta_1\right)+\varphi_1(\varphi_1+\theta_1)+\theta_2\right], \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))''|_{\varphi_1\theta_2} = 2\sigma^2\left[2\varphi_1+\theta_1\right], \\ &(Var_{ARMA(1,2)}(\hat{x}_3 - x_3))''|_{\varphi_1\theta_2} = 2\sigma^2\varphi_1. \end{aligned} \tag{22}$$

Structure of hessian after simplification with summation of rows and columns with appropriate multipliers is presented in (23):

$$H_{ARMA(1,2)}(\hat{x}_3 - x_3) = \sigma^6 \begin{vmatrix} \left(\varphi_1(\varphi_1+\theta_1)+\theta_2\right)^2 & & \\ & 2 & \\ & & 2 \end{vmatrix}. \quad (23)$$

It's positive except solutions of the equations (21). At extreme points it's zero. This result is close to the previous cases. There's a hyperplane at which variation is minimal: $\varphi_1 = -\theta_1, \theta_2 = 0$. For models ARMA(1, q), q>2, variance of predictions over 1, 2, 3 timesteps has got the same form as in cases presented above. Differences are going to appear only in predictions over more timesteps.

3. Experiments

The Dynamic series of macroeconomic statistics of the Russian Federation (monthly wage index and monthly income index) [18] have been handled in the experimental part. Last 12 values of the time series data have been used as test while the first 300 values were used to train models. In previous part only ARIMA(1, d, q) models have been handled. So, here they are used as parts of combined model. At the same time the best model by value of information criteria [16, 17] is also highlighted. In order to compare forecasts RMSE and MAE metrics of errors are used:

$$RMSE = \sqrt{\frac{\sum_t (\tau(t) - ts(t))^2}{N}}, \quad (24)$$

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

Here N is length of test period, *τ(t)* denotes predicted values of the processed time series, *ts(t)* is a part of the investigated time series at the test period (real data).

Two experiments have been performed. In the first one 200 values were used to train models and 12 values to test them. Order of moving average part was limited with 3. Order of the autoregression part was set to 1. One can see its results in the Table 1. The best model (by value of information criteria) is ARIMA(0, 1, 1) one. It's marked with bold font in the table.

Table 1
The ARIMA (1, d, q), q<4, models of the wage index

ARIMA(p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(1, 1, 0)	1518.43	18.68	12.41
ARIMA(1, 1, 1)	1496.72	17.94	11.62
ARIMA(1, 1, 2)	1498.41	17.91	11.60
ARIMA(1, 1, 3)	1495.42	17.75	11.46
ARIMA(0, 1, 1)	**1494.6**	**15.14**	**9.21**
Combined model	1495.73	17.75	11.46

Among four tested models there are signs of terms: $\varphi_1 > 0$ one time and $\varphi_1 < 0$ 3 times, $\varepsilon_1 > 0$ one time and $\varepsilon_1 < 0$ 2 times (for ARIMA(1, 1, 0) it's zero), there's one positive ε_2 term and one negative. So, theoretically averaging is expected to achieve "good" results. One can see that combined model looks very close to the ARIMA(1, 1, 3) model and it's better than other ones (of 1st order) by value of Akaike information criterion and by quality of forecast at tested period. Especially it exceeds results of the worst model ARIMA(1, 1, 0) . The best model has got different orders and wasn't used in combination but its results are close to the combined model and to the ARIMA(1, 1, 3) one.

In the second experiment models ARIMA(1, d, q), q<6, were used. Length of the training set was 300 values. Volume of the training set was also 12. Its results are shown in the Table 2.

Table 2
The ARIMA (1, d, q), q<6, models of the wage index

ARIMA(p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(1, 1, 0)	942.44	37.59	36.50
ARIMA(1, 1, 1)	911.05	25.79	14.54
ARIMA(1, 1, 2)	**901.17**	**22.11**	**14.20**
ARIMA(1, 1, 3)	901.51	22.24	14.16
ARIMA(1, 1, 4)	903.51	22.23	14.14
ARIMA(1, 1, 5)	905.48	22.47	14.13
Combined model	906.7	22.47	14.13

Here one can see that the combined model is better than the worst ones but is worse than the best models. In this case almost all terms in various models have got the same signs. That's why averaging doesn't give results close to the best ones. But it's far better than the worst models used in combination.

The same experiment as presented in the Table 1 has been done in case the income index [18].

Among four tested models with orders q<4 there are signs of terms: $\varphi_1 > 0$ 2 times and $\varphi_1 < 0$ also 2 times, $\varepsilon_1 < 0$ 3 times (for ARIMA(1, 1, 0) it's zero), $\varepsilon_2 > 0$ for the both models with non-zero MA(2) term. Here, we expect that averaging is going to be good practice because signs of AR(1) part vary. Results are shown in the Table 3.

Table 3
The ARIMA (1, d, q), q<4, models of the income index

ARIMA(p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(1, 1, 0)	1742.29	26.08	15.64
ARIMA(1, 1, 1)	1685.3	27.02	16.61
ARIMA(1, 1, 2)	1676.16	28.13	18.21
ARIMA(1, 1, 3)	1678.15	28.13	18.21
ARIMA(2, 1, 2)	**1678.15**	**28.13**	**18.21**
Combined model	1678.15	28.13	18.21

Here four models (including the best one by information criteria values and the combined model) make almost the same predictions by quality. At the same time models with q<2 make better predictions but they've got worse information criteria values. It can be explained with overfitting to the learn data. Thus, researcher can have one more model that's close by quality to some number of the best ones. It can be used to analyze signs of terms where they vary in implemented models.

In the last experiment models ARIMA(1, d, q), q<6, were used. Length of the training set (income index) was 300 values. Volume of the training set was also 12. Its results are shown in the Table 4.

Table 4
The ARIMA (1, d, q), q<6, models of the income index

ARIMA(p, d, q) models	Akaike information criterion	RMSE of forecast	MAE of forecast
ARIMA(1, 1, 0)	1012.57	29.16	15.28

ARIMA(1, 1, 1)	976.78	28.91	16.49
ARIMA(1, 1, 2)	956.53	29.60	14.66
ARIMA(1, 1, 3)	958.36	29.89	14.58
ARIMA(1, 1, 4)	903.51	29.34	14.74
ARIMA(1, 1, 5)	959.62	29.44	14.74
ARIMA(0, 0, 1)	**973.22**	**29.60**	**20.84**
Combined model	962.75	29.44	14.74

There are a few models better than the best one (chosen by values of combination of information criteria). Again in case of a lot of models the combined model shows results which are better than some models (the worst ones) and worse than some of them (the best ones by quality of prediction).

Also, again one can see that parameters of combined models are close to traits of the model with the highest order of moving average part. Various approaches to averaging technique have also been tested in [2].

4. Conclusion

In this research averaging of predictions of ARIMA(p, d, q) time series models is investigated. To describe variance of various time series models it has been expressed in form of psi-weights for AR(1) and ARMA(1, q) models. It should also be mentioned that MA(q) series have already got appropriate form and their coefficients are equal to psi-weights. Having got explicit form of prediction variance over 1, 2, 3 periods of time it's possible to find models with the lowest variance and to evaluate whether averaging leads to improvement of prediction. Models AR(1), ARMA(1, 1) and ARMA(1, 2) have been considered. Models ARMA(1, q), q>2, have got the same variance of prediction as in considered cases. Differences in their structure are going to appear only for predictions over more timesteps.

If there's a few models averaged one can be the best. It happens when models predict various behaviour of investigated time series. In common case averaged model has got variance better than the worst model by variance and worse than the best one (expression (9)). These results are based on the form of variance of prediction function. It's shown that for models ARMA(1, q) it's bulging downward. Thus, combined models are going to give the best results or close to the best ones in case when there are equal number of terms with various signs (terms with large modules are the most important because they've got larger weights in averaging). This situation may take place if investigated time series is difficult for analysis and forecasting and models at hand predict various behaviour (in case of models with low orders; if orders increase, some terms in models have got various signs).

Time series from the Russian macroeconomical statistics [18] have been used as test data for computational experiments. Combined models are always better than the worst models. So, they can be used as another tool of analysis. In some cases results of combined models are close to results of the best ones.

Nowadays this theme is very important and there are papers on averaging and bagging of time series predictions, bagging with use of non-linear functions of time series terms [3, 4, 9-11,19-21]. In future work bagging of time series and averaging of ARMA(p, q) models with higher orders of autoregression parts (p>1) are going to be investigated.

5. References

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

[2] D. Petrusevich, Improvement of time series forecasting quality by means of multiple models prediction averaging, in: E. Semenkin, I. Kovalev (eds.), Proceedings of the III International Workshop on Modeling, Information Processing and Computing (MIP: Computing-2021),Volume 2899, Krasnoyarsk, Russia, 2021, pp. 109-117. doi: 10.47813/dnit-mip3/2021-2899-109-117.

[3] 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.

[4] 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.
[5] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, H. L. Shang, Optimal combination forecasts for hierarchical time series, Computational Statistics & Data Analysis 55.9 (2011) 2579-2589. doi: 10.1016/j.csda.2011.03.006.
[6] N. Shafik, G. Tutz, Boosting nonlinear additive autoregressive time series, Computational Statistics & Data Analysis 53.7 (2009) 2453-2464 doi: 10.1016/j.csda.2008.12.006.
[7] J. M. Matías, M. Febrero-Bande, W. González-Manteiga, J.C. Reboredo, Boosting GARCH and neural networks for the prediction of heteroskedastic time series, Mathematical and Computer Modelling 51.3-4 (2010) 256-271. doi: 10.1016/j.mcm.2009.08.013.
[8] L. Breiman, Random forests, Machine Learning 45 (2021) 5-32. doi: 10.1023/A:1010933404324
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] K. A. Boikov, M. S. Kostin, G. V. Kulikov, Radiosensory diagnostics of signal integrity in-circuit and peripheral architecture of microprocessor devices, Russian Technological Journal (In Russ.) 9.4 (2021) 20-27. doi: 10.32362/2500-316X-2021-9-4-20-27.
[14] N. M. Legkiy, N. V. Mikheev, Selection of location of radiators in a non-equivident antenna array, Russian Technological Journal (In Russ.) 8(6) (2020) 54-62. (In Russ.) doi: 10.32362/2500-316X-2020-8-6-54-62.
[15] H. Wald, A study in the analysis of stationary time series, 2nd edition, Almqvist and Wiksell Book Co., Uppsala, 1954.
[16] R. J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, 2nd edition, OTexts, Melbourne, Australia, 2018.
[17] J. H. Stock, M.W. Watson, Introduction to Econometrics, Pearson, 2019.
[18] Dynamic series of macroeconomic statistics of the Russian Federation. Wage index, income index, 2021. URL: http://sophist.hse.ru/hse/nindex.shtml
[19] S. F. Stefenon, M. H. Dal Molin Ribeiro, A. Nied, K.-C. Yow, V .C. Mariani, L. dos Santos Coelho, L.O. Seman, Time series forecasting using ensemble learning methods for emergency prevention in hydroelectric power plants with dam, Electric Power Systems Research 202 (2022) 107584. doi: 1016/j.epsr.2021.107584.
[20] M. Larrea, A. Porto, E. Irigoyen, A.J. Barragán, J.M. Andújar, Extreme learning machine ensemble model for time series forecasting boosted by PSO: Application to an electric consumption problem, Neurocomputing, 452 (2021) 465-472. doi: 10.1016/j.neucom.2019.12.140.
[21] R. Godahewa, K. Bandara, G. I. Webb, S. Smyl, C. Bergmeir, Ensembles of localised models for time series forecasting, Knowledge-Based Systems 233 (2021) 107518. doi: 10.1016/j.knosys.2021.107518.