Ensemble forecasting method

V A Ivanyuk\textsuperscript{1,2} and A D Tsvirkun\textsuperscript{3}

\textsuperscript{1}Department of Data Analysis and Machine Learning, Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, 125993, Moscow, Russia
\textsuperscript{2}Department of Higher Mathematics, Bauman Moscow State Technical University, Moscow, Russia
\textsuperscript{3}V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia

E-mail: VAIvanyuk@fa.ru

Abstract. The purpose of this article is to analyze the time series based on aggregate forecasting methods. Forecasting time series comprises an important scientific and technical task which is relevant in various sectors of economy and production. The main objective of any forecasting process is to get the ability to assess trends in changes to a particular factor. To predict the behaviour of future processes, a qualitative analysis of data is required, which provides the basis for the forecast. It is also important to choose the appropriate forecasting method. The forecast obtained may not be always correct, so it is crucial to determine how accurately it is built. In the present paper, characteristics of time series are provided. The main mathematical methods of time series forecasting are analyzed and their classification is presented. Criteria for the accuracy and reliability of forecasting models are defined. The accuracy of the forecast is an important criterion for evaluating the forecasting method. The aggregate ensemble forecast has proven to be an effective method for enhancing the accuracy of forecasting. The article provides an example of constructing an aggregate ensemble forecast. The types of nonlinear models are described and several methods are selected for constructing the aggregate ensemble forecast. The practical application of exponential smoothing is considered, the MA (moving average) model and the AR (autoregressive) model are constructed.

1. Introduction
Forecasting financial time series holds a special place in modern economics. A forecast is important for making many decisions, such as buying or selling assets in the financial market, investing capital, and identifying cost-effective options for long-term, medium-term, or short-term planning. To improve the accuracy of the forecast, various forecasting methods are developed. For example, the method of combining forecasting techniques, also called ensemble forecasting, provides an effective way of improving the accuracy of forecasting.

2. Main types of models
There are the following types of dynamic models:

- distributed lag models, i.e. models in which only independent (explanatory) variables can be lagged. Models with lags have the following form (1):
\[ y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + K + \beta_k x_{t-k} + \epsilon_t \quad (1) \]

- autoregressive models, i.e. models in which dependent variables are lagged. Autoregressive models have the following form (2):
\[ y_t = \alpha + \beta y_{t-1} + \epsilon_t \quad (2) \]

An important step in modelling is the model performance evaluation [1, 2]. Mathematical statistics offers a number of estimation methods:
1. The proportion of variance explained (coefficient of determination).
2. Pearson’s chi-squared test.
3. Graph of observed and predicted values.
4. Normal and half-normal residual plot.
5. The covariance matrix of parameter estimates.
Nonlinear models have the following advantages: flexibility and consistency of analysis. However, these models are not free from shortcomings such as the ambiguity of the estimated coefficients and the complexity of their estimation [3, 4, 5, 6].

The most popular nonlinear models are those of the ARMA (autoregressive–moving-average model) class (3), which generalizes two simpler models – AR (autoregressive model) and MA (moving average model).
\[ X_t = c + \theta \epsilon_{t-1} + \sum_{i=1}^{p} \phi_i X_{t-i}, \epsilon_t \sim N(0, \sigma^2) \quad (3) \]

The parameter \( p \) is the order of the regression part, and the parameter \( q \) is the order of the moving average.

Keeping in mind the problem of time series stationarity, integration should be added to the model to solve it. As a result, we have ARIMA (4) (autoregressive integrated moving average model). The model is an extended version of ARMA for nonstationary series, which can be turned into stationary by differencing the time series. In this model, the parameter \( d \) is added, which indicates the degree of differencing.
\[ \Delta^d X_t = c + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{p} \phi_i \Delta^d X_{t-i}, \epsilon_t \sim N(0, \sigma^2) \quad (4) \]

Financial time series are characterized by seasonality. Seasonality describes the periodic fluctuations observed in a time series. The seasonally adjusted model is referred to as SARIMA or ARIMA (P,D,Q)s. The parameter \( P \) is the order of the seasonal component SAR(P), \( D \) is the integration order of the seasonal component, \( Q \) is the order of the seasonal component SMA(Q), and \( s \) represents the dimension of seasonality (week, month, etc.).

In ARIMA modelling, \( d \) cannot be fractional, but there are fractionally integrated models (5) in which the coefficient \( d \) can take fractional values. The fractional value is explained by expanding the operator in a power series. This model is named ARFIMA (autoregressive fractionally integrated moving average model).
\[ \Delta^d = (1 - L)^d = \sum_{k=0}^{\infty} \sum_{j=0}^{k-1} \frac{(-1)^k}{k!} L^k \quad (5) \]
The process of building an ARMA class model involves the following steps:

1. Postulating. At this stage, the model class is selected.
2. Defining the subclass of models. Choosing a subclass assumes the selection of specific parameters for the model.
3. Estimation. The values of the model parameters are estimated using the maximum likelihood method.
4. Diagnostics. The final stage is to check the model for inconsistency, investigate the properties of the model and recognize its adequacy.

3. Forecast quality evaluation
The accuracy indices of the forecasting model determine the value of the resulting error \[7, 8, 9\]. To confirm the quality and suitability of the model used, one needs to analyze the system of indicators that reflect both the adequacy of the model and its accuracy. The accuracy of the forecast is determined by the amount of error in the forecast. Forecast error \(\Delta t\) is a value defined as the difference between the actual and predicted values of the indicator. As the error value decreases, the accuracy of the forecast increases accordingly.

The absolute error \(\Delta t\) of the time series forecast is an analytical indicator, the use of which makes it possible to quantify and estimate the value of the forecast error. The absolute error \(\Delta t\) is calculated using the following formula (6):

\[
\Delta t = y_t - \hat{y}_t
\]

where \(\hat{y}_t\) – predicted value of the indicator;

\(y_t\) – actual value.

The average squared error is calculated using the following formula (7):

\[
\Delta t = \frac{(y_t - \hat{y}_t)^2}{y^2}
\]

The forecast error is identical in dimension to the predicted indicator and directly depends on the scale of measurement of time series levels. Thus, the forecast accuracy increases as this value decreases. For processes that are defined by a large number of indicators, one needs to compute a summarizing indicator of accuracy.

To assess the accuracy of the forecast, the mean squared error MSE is used, which highlights major errors by squaring each error (8).

\[
MSE = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}
\]

4. Ensemble forecast
To build an ensemble forecast, we consider combining forecasts that were developed using three forecasting methods \[10, 11, 12\].

1. The method of exponential smoothing (9).

\[
F_t = aX_{t-1} + (1 - a)F_{t-1},
\]

where \(t\) – the period preceding the forecast period, \(F_t\) – the predicted indicator, \(X_{t-1}\) – the actual value of the studied indicator for the period preceding the forecast period, \(a\) – the smoothing coefficient. A formally correct procedure for selecting \(a\) does not exist, so the coefficient will be determined based on the length of the smoothing interval. Then \(a\) is computed using the formula:

\[
a = \frac{2}{n + 1},
\]
where \( n \) is the number of observations within the smoothing interval. One of the main advantages of this method is the ability to solve long-term forecasting problems [13, 14].

2. Moving average (MA) model (10):

\[
F_t = \frac{1}{n} \sum_{i=0}^{n-1} (F_{t-i} - F_{t-i-1}),
\]

where \( F_t \) – the predicted indicator, \( t \) – the period preceding the forecast period, \( m_{t-2} \) – the moving average of two periods prior to the forecast period, \( n \) – the number of equations within the smoothing interval, \( F_{t-1} \) – the actual value of the examined indicator over the period preceding the forecast period, \( F_{t-2} \) – the actual value of the examined indicator over two periods preceding the forecast period. The main advantages of this method are simplicity and visibility.

3. Autoregressive model AR(p). In this model, the values of the time series depend linearly on past values. The AR(p) model is well suited for making a forecast for 1 period ahead. The order of the model is equal to 1, \( p=1 \). The predicted value depends on the previous values (11):

\[
F_t = c + aX_{t-1} + \varepsilon_t,
\]

where \( a \) – autoregressive coefficients (model parameters), \( c \) – constant, and \( \varepsilon_t \) – white noise (random error). The main advantages of this model include simplicity and transparency of modelling, it is used in solving problems of forecasting time series [15, 16].

To make a forecast, let us take the data of the EUR/RUB currency pair rate over the period 06.09.2019–06.04.2020. The time interval is 1 week. As a training sample for generating a weighted moving average, the period 07.09.2019 – 14.10.2019 was selected. As a test sample, the period 14.10.2019 – 06.04.2020 was selected.

All calculations are performed using MS Excel [17, 18]. First, we make a forecast for the period of the test sample 07.10.2019 – 06.04.2020 using three methods that were selected and described earlier. These methods will provide the basis for constructing the ensemble forecast. The actual data is denoted as \( a_j \).

To build the first forecast «F1» we use exponential smoothing. The graph with the original data and the F1 values is shown in figure 1.

Let us denote the forecast made using the moving average method by «F2». Using the MS Excel tools for statistical analysis, we make a forecast for one period ahead based on the moving average method.

The graph with the original data and the F2 forecast values is shown in figure 2.

![Figure 1. Exponential smoothing. Forecast F1.](image1)

![Figure 2. Moving average. Forecast F2.](image2)

The forecast generated using the method of autoregression is denoted as «F3». Using the MS Excel tools for statistical data analysis, we calculate regression coefficients: \( a = 2.998, b = 0.965 \), hence, we get the equation describing the dynamics of the rate changes:

\[
f_j = 2.998 + 0.965 a_{j-1}
\]
The graph with the original data and the F3 forecast values is shown in figure 3. The ensemble forecast is built based on the arithmetic mean of original forecasts using a formula of the form (12):

$$ f_{mj} = \frac{1}{n} \sum_{i=1}^{n} f_{ij} $$  \hspace{1cm} (12)

where $i$ is the number of methods combined.

Accordingly, using equation (15), we calculate aggregate forecasts consisting of individual forecasts F1-F2, F1-F3, F2-F3, F1-F2-F3, or F12, F13, F23, F123.

Figure 4 shows the graph of forecasts F1, F2, F3, F12, F13, F23, F123.

The graph with the original data and the F3 forecast values is shown in figure 3. The ensemble forecast is built based on the arithmetic mean of original forecasts using a formula of the form (12):

$$ f_{mj} = \frac{1}{n} \sum_{i=1}^{n} f_{ij} $$  \hspace{1cm} (12)

where $i$ is the number of methods combined.

Accordingly, using equation (15), we calculate aggregate forecasts consisting of individual forecasts F1-F2, F1-F3, F2-F3, F1-F2-F3, or F12, F13, F23, F123.

Figure 4 shows the graph of forecasts F1, F2, F3, F12, F13, F23, F123.

**Table 1.** The results of the MSE forecast error.

| Forecast | F1   | F2   | F3   | F12  | F13  | F23  | F123 |
|----------|------|------|------|------|------|------|------|
| MSE      | 5.96 | 7.22 | 4.71 | 6.46 | 5.13 | 5.57 | 5.64 |

5. Conclusion
Thus, we considered methods for analyzing financial time series. Optimal forecasting models were selected and their constructive aspects were studied. Methods and algorithms for estimating the parameters of all the models under consideration were defined. An aggregate ensemble forecast was constructed. Combining forecasting methods for the purpose of improving the forecast accuracy showed a good result.

References
[1] Zhou M and Jin M 2017 Holographic ensemble forecasting method for short-term power load *IEEE Transactions on Smart Grid* **10**(1) 425-434
[2] Zhao Y, Li J and Yu L 2017 A deep learning ensemble approach for crude oil price forecasting. *Energy Economics* **66** 9-16
[3] Avci E, Ketter W and van Heck E 2018 Managing electricity price modeling risk via ensemble forecasting: The case of Turkey *Energy policy* **123** 390-403
[4] Buizza R 2019 Introduction to the special issue on “25 years of ensemble forecasting” *Quarterly Journal of the Royal Meteorological Society* **145** 1-11
[5] Dai Z and Zhu H 2020 Forecasting stock market returns by combining sum-of-the-parts and ensemble empirical mode decomposition *Applied Economics* **52**(21) 2309-2323
[6] Sunchalin A M, Kochkarov R A, Levchenko K G, Kochkarov A A and Ivanyuk V A 2019 Methods of risk management in portfolio theory Revista ESPACIOS 40(16)
[7] Elizarov M, Ivanyuk V, Soloviev V and Tsvirkun A 2017 Identification of high-frequency traders using fuzzy logic methods In 2017 Tenth International Conference Management of Large-Scale System Development (MLSD) IEEE pp. 1-4
[8] Koroteev M V, Terelyanskii P V and Ivanyuk V A 2016 Fuzzy Inference as a Generalization of the Bayesian Inference Journal of mathematical sciences 216(5) 685-691
[9] Koroteev M V, Terelyanskii P V and Ivanyuk V A 2016 Arithmetic of fuzzy numbers in generalized trapezoidal form Journal of Mathematical Sciences 216(5) 696-701
[10] Tang L, Dai W, Yu L and Wang S 2015 A novel CEEMD-based EELM ensemble learning paradigm for crude oil price forecasting International Journal of Information Technology & Decision Making 14(01) 141-169
[11] Wei Y, Sun S, Ma J, Wang S and Lai K K 2019 A decomposition clustering ensemble learning approach for forecasting foreign exchange rates Journal of Management Science and Engineering 4(1) 45-54
[12] Yu L, Zhao Y and Tang L 2017 Ensemble forecasting for complex time series using sparse representation and neural networks Journal of Forecasting 36(2) 122-138
[13] Xiao J, Li Y, Xie L, Liu D and Huang J 2018 A hybrid model based on selective ensemble for energy consumption forecasting in China Energy 159 534-546
[14] Ekinci A and Erdal H I 2017 Forecasting bank failure: Base learners, ensembles and hybrid ensembles Computational Economics 49(4) 677-686
[15] Simonov B, Ivanyuk V and Simonova I 2016 Existence of Best Approximation Elements in the spaces L Journal of Mathematical Sciences 217(5)
[16] Kim D and Hur J 2018 Short-term probabilistic forecasting of wind energy resources using the enhanced ensemble method Energy 157 211-226
[17] Ivanyuk V 2018 Econometric Forecasting Models Based on Forecast Combination Methods In 2018 Eleventh International Conference Management of large-scale system development"MLSD IEEE pp. 1-4
[18] Ivanyuk V and Soloviev V 2019 Efficiency of Neural Networks in Forecasting Problems In 2019 Twelfth International Conference Management of large-scale system development"(MLSD) IEEE pp. 1-4