Time series forecasting by global optimization method

V Balyk¹, E Balyk² and V Vakulchuk¹*

¹Department of design of aerohydrocosmic systems, Moscow Aviation Institute (National Research University), 4 Volokolamsk highway, Moscow 125993, Russia
²Federal Institute of Industrial Property, 30 Berezhkovskaya nab., building 1, Moscow 125993, Russia
*E-mail: vakulvi@yandex.ru

Abstract. The actual problem of time series forecasting associated with the choice of the type of forecasting functions is considered. The aim of the work is to develop a new forecasting method, in which, along with the construction of the predictive function, the values of the predicted characteristics are directly optimized at the entire depth of the forecast. The developed method belongs to the class of probabilistic statistical forecast methods and partially to the class of feedback methods. Typical representatives of the methods of these classes are correlation and regression analysis, factor and variance analysis, statistical modeling. In the new method the predictive function is constructed in the basis of trigonometric polynomials, which leads to multi-extremality of the forecasting problem. In this regard, to solve the problem of forecasting, the method of global optimization, hereinafter called the criterion shifts method, is used. The developed forecasting method allows solving a wide range of practical forecasting problems. Since the method of global solution search is used, it guarantees the exact finding of the predicted characteristics obtained as a result of their optimization by the criterion of regularity. The results obtained in this article showed high accuracy.

1. Introduction

The problem of forecasting time series, the solution of which is based on the theoretical provisions presented in [1, 2], is considered. The predictive function is constructed in the basis of trigonometric polynomials, and it, along with high approximating qualities, is multi-extreme. A minimum of regularity corresponds to an accurate prediction, and we consider the method of searching the global minimum based on the construction of process of extremum search in the extended space of variables, the dimension of which is determined by the dimensionality of the coefficients of the prediction function and the number of predictable characteristics. The peculiarity of the method is that the ordinates of the predicted process for the entire depth of the forecast for the entire sampling interval are taken as variables.

The relevance of the problem is due to numerous practical applications. So, in [3], an algorithm is considered based on innovative models of the state space that underlie the methods of exponential smoothing. It also provides a step-by-step forecasting algorithm using ARIMA models. In [4], the ecological problem of developing a model for forecasting based on time series is considered. In [5], approaches related to forecast aggregation, the so-called 'coherence forecast' methods, are presented. Clustering of time series data, used in various scientific fields to identify patterns, is given in [6, 7]. A similar problem of clustering, but for granular time series, is discussed in [8].

Forecasting of time series is shown on the problem of fuel cost forecasting according to statistics taken in the period from 2002 to 2008. The predictive function is constructed in the class of...
trigonometric polynomials, which describe quite well the various processes occurring in time, but the problem of choosing a prediction function remains. Optimization of forecast points, as is done in this method, allows you to implement a parallel forecast, which gives additional justification when choosing a forecast function.

The novelty in this work is that here, in addition to the coefficients of the predictive trigonometric polynomial, the actual predicted parameters vary, the optimal values of which are taken as the final forecast. This is the fundamental difference between this work and similar works [3-9].

The aim of the work is to develop a new method for forecasting time series and compare the quality of the forecast from [9] with the forecast obtained by this method. All comparisons are made with the training of 20, 40 and 60 points respectively.

2. Time series forecasting by the criterion of regularity

The initial data for prediction is the implementation of the investigated random process specified with a certain sampling interval:

\[ X(t_k), \quad k = 0, n, \quad t_{k+1} - t_k = \Delta, \]

in this case, the sampling interval should be less than the expected characteristic interval. The statistical sample size of \( N \) must be greater than the decay period of its autocorrelation function. It should be noted that the choice of too small sampling interval significantly increases the volume of initial data and complicates their processing-estimation of statistical characteristics and forecasting.

This method assumes the formalization of the problem of forecasting a random process by the criterion of regularity (minimum of the mean square of the error of the predictive function), which is the most common in statistical optimization problems. In General, the optimal operator is such an operator that provides an extreme value of the criterion for the quality of the functioning of the system, the input of which is a known random process \( X(t) \) [1, 2].

In the statement of the problem, it is assumed that at the system output it is required to obtain a given output signal \( Y_h(t) \), where \( h \) is the observation interval, when the system generates an output signal \( Y(t) \). In the process prediction problem \( X(t) \) for the interval \( \gamma \), the required output signal is determined by the expression:

\[ Y_h(t) = X(t + \gamma). \]

The forecast error is defined as:

\[ E(t) = Y(t) - Y_h(t), \]

which in the case where \( Y(t) \) and \( Y_h(t) \) are random processes is also a random process.

The initial statistical sample in the forecasting problem has the form presented in table 1.

| Table 1. Statistical sampling in the forecasting problem. |
|---------------------------------------------------------|
| \( t \)     | \( X(t); Y_h(t) \) |
|-------------|-------------------|
| \( t_1 \)   | \( X(t_1) \)     |
| \( t_2 \)   | \( X(t_2) \)     |
| ...         | ...              |
| \( t_{N_a} \)| \( X(t_{N_a}) \) |
| \( t_{N_a+1}\) | \( Y_h(t_{N_a+1}) \) |
| \( t_{N_a+2}\) | \( Y_h(t_{N_a+2}) \) |
| ...         | ...              |
| \( t_N \)   | \( Y_h(t_N) \)   |
First part of the statistical sample \((N_A)\) is the training part. Second part is the predicted part of the statistical sample. A predictive function is constructed on a segment \([t_i, t_{N_A}]\) of long \(h\). The forecast \(Y_h(t_{N_A+1}), Y_h(t_{N_A+2}), \ldots, Y_h(t_N)\) is based on the segment \([t_{N_A+1}, t_N]\).

The predicted values \(Y_h(t_{N_A+1}), \ldots, Y_h(t_N)\) are also variable parameters \(x_1, x_2, \ldots, x_{N-N_A}\). Construction of the predictive function is carried out on a statistical sample from table 2.

**Table 2.** Statistical sampling of predictive function construction.

| \(t\) | \(X(t)\) |
|---|---|
| \(t_i\) | \(X(t_i)\) |
| \(t_2\) | \(X(t_2)\) |
| \(\ldots\) | \(\ldots\) |
| \(t_{N_A}\) | \(X(t_{N_A})\) |
| \(t_{N_A+1}\) | \(x_1\) |
| \(t_{N_A+2}\) | \(x_2\) |
| \(\ldots\) | \(\ldots\) |
| \(t_N\) | \(x_{N-N_A}\) |

A trigonometric polynomial was adopted as a predictive function:

\[
Y_h(t) = a_0 + \sum_{m=1}^{m} \left[ a_m \cos(\omega t) + b_m \sin(\omega t) \right]
\]

where \(a_0\) - the average value of the predicted process, \(m\) - the Harmonic number, \(a_m, b_m\) - Fourier coefficients, \(\omega_t\) - the frequency of the process.

Parameters \(a, b, \omega_t\), \(i = 1, m\) as well as parameters \(x_i, i = 1, N-N_A\) are determined from the minimum condition of the regularity criterion:

\[
J^{(1)} = \min_{\{a_i, b_i, \omega_i, i = 1, m\}} \sum_{i=1}^{N-N_A} \left[ Y_h(t_i) - x_i \right]^2
\]

\[
J^{(2)} = \min_{\{a_i, b_i, \omega_i, i = 1, m\}} \frac{\sum_{i=1}^{N} [X(t_i) - Y_h(t_i)]^2 + \sum_{i=1+N_A}^{N} [Y_h(t_i) - x_i] \cdot \sum_{i=1}^{N-N_A} x_i^2}{\sum_{i=1}^{N-N_A} x_i^2}
\]

For the purpose of generality, a criterion is introduced \(J\). Under the criterion \(J\), we will understand either a criterion \(J^{(1)}\) or a criterion \(J^{(2)}\) depending on which variables are optimized. If the criterion \(J\) is optimized for forecast points \(x_i, i = 1, N-N_A\), then \(J = J^{(1)}\). If the criterion \(J\) is optimized by Fourier coefficients and frequencies of the trigonometric polynomial \(a, b, \omega_t, i = 1, m\), then \(J = J^{(2)}\).

The minimization tasks (1), (2) are multi-extreme and in practical applications - of large dimension. To solve such problems, a method of criterion shifts has been developed, which is given below.
3. Research methods

According to the considered approach, solution of the time series forecasting problem is reduced to the problem of minimizing the regularity criterion of types (1) and (2) by choosing variable parameters \( x_1, x_2, \ldots, x_n \), where \( x_1, x_2, \ldots, x_n \) - Fourier coefficients and the frequencies of the trigonometric polynomial, and \( x_{n+1}, x_{n+2}, \ldots, x_n \) - variable predicted parameters. The global extremum search method from [9] was chosen as the minimization method.

Let \( J_0 \) be the initial value of the criterion and let \( J(x_1, x_2, \ldots, x_n) \) be the criterion function. Since the criterion function must be minimized, a shift is made:

\[ J_1 = J_0 - \Delta J \]

and the condition is checked:

\[ J(x_1, x_2, \ldots, x_n) \leq J_1. \]

This check is carried out by solving the following problem:

\[ \inf h(x), \quad (3) \]

where \( h(x) = |J(x_1, x_2, \ldots, x_n) - J_1| \).

Let's replace the variables:

\[
\begin{align*}
x_1 &= a_1 \cos(\omega_1 J) + b_1 \sin(\omega_1 J); \\
x_2 &= a_2 \cos(\omega_2 J) + b_2 \sin(\omega_2 J); \\
&\vdots \\
x_n &= a_n \cos(\omega_n J) + b_n \sin(\omega_n J).
\end{align*}
\]

Then task (3) will be rewritten as:

\[ \inf_{\{a_i, b_i, \omega_i, J\} \in \mathbb{R}^n} h(x). \quad (4) \]

If in the process of solving problem (4) it turns out that \( J(x_1, x_2, \ldots, x_n) \leq J_1 \),

is done shift:

\[ J_2 = J_1 - \Delta J \]

and the described process continues.

If in the process of solving problem (4) it turns out that \( J(x_1, x_2, \ldots, x_n) > J_1 \),

the shift is made in the opposite direction, i.e.

\[ J_2 = J_1 + \Delta J \]

and further similar checks are carried out.

The search step of the method is based on the rule:

\[
J_{\text{lim}}^{(k+1)} = \begin{cases} 
J_{\text{lim}}^{(k)} - \delta, & \delta = \frac{J_{\text{lim}}^{(1)}}{2k} \quad \text{if} \ H[f, J_{\text{lim}}^{(k)}] < \varepsilon; \\
J_{\text{lim}}^{(k)} + \delta, & \delta = \frac{J_{\text{lim}}^{(1)}}{2k} \quad \text{if} \ H[f, J_{\text{lim}}^{(k)}] > \varepsilon, \\
k = 1, 2, \ldots 
\end{cases} \quad (5)
\]

Note that the process of minimizing the incompatibility measure of equation (4) stops if the inequality is established at some step \( s(k) : f(x_1, x_2, \ldots, x_n) \leq J_{\text{lim}}^{(k)} \).

As a result of the iterative process (5), a sequence of criteria is formed:

\( J_{\text{lim}}^{(1)}, J_{\text{lim}}^{(2)}, \ldots, J_{\text{lim}}^{(k)} \),

from which a decreasing subsequence can be distinguished:
and increasing subsequence:
\[ J_{\text{lim}}^{(2)} < J_{\text{lim}}^{(3)} < J_{\text{lim}}^{(6)} < \ldots < J_{\text{lim}}^{(2k)} , \]
the limits of which at \( k \to \infty \) are equal to each other, and this limit corresponds to the global extremum.

The iterative process is terminated if
\[ \left| J_{\text{lim}}^{(2k+1)} - J_{\text{lim}}^{(2k)} \right| \leq \varepsilon . \]

The method is described. We prove the convergence of the method by value.

Theorem:
A bounded, continuous function is considered \( J = J(x) \).
\[ J(x) : X \to \mathbb{R}^n , \text{where } x \in X = \mathbb{R}^n , \]
where \( X \) is the compact set.

Let the global minimum be defined as:
\[ J_* = \text{globmin} J(x) = J(x_*), \]
where \( x_* \) is any point of the global minimum equal to \( J_* \).

Let the sequence of records \( \{ J_k \} \) be constructed by the method (1) – (4). Then
\[ \lim_{k \to \infty} J_k = J_* . \]

Evidence:
Let's replace the variables:
\[
\begin{aligned}
x_1 &= a_1 \cos(\omega_1 J) + b_1 \sin(\omega_1 J) ; \\
x_2 &= a_2 \cos(\omega_2 J) + b_2 \sin(\omega_2 J) ; \\
&\vdots \\
x_n &= a_n \cos(\omega_n J) + b_n \sin(\omega_n J) .
\end{aligned}
\]

\( a_1, b_1, \omega_1, \ldots, a_n, b_n, \omega_n \) - selectable parameters that can be interpreted as Fourier coefficients and frequency, \( J \) - a new argument, it is also the criterion of optimality.

Suppose that the point \( J^{(k)} \) is already known at any \( k \geq 0 \). Calculate the value \( J^{(k+1)} \):
\[
\begin{aligned}
J^{(k+1)} &= J^{(k)} - \frac{J^{(1)}}{2k} \quad \text{if } f \left( x_1, x_2, \ldots, x_n \right) \leq J^{(k+1)} ; \\
J^{(k+1)} &= J^{(k)} + \frac{J^{(1)}}{2k} \quad \text{if } f \left( x_1, x_2, \ldots, x_n \right) > J^{(k+1)} .
\end{aligned}
\] (6)

Rule (6) forms an iterative process whereby the optimality criterion decreases as long as the criterion function is less than or equal to the value of the sequentially decreasing argument \( J^{(k)} \). If there is no such solution for some \( k \), then \( k = k+1 \) the criterion \( J^{(k)} \), according to rule (6), increases sequentially. The stepwise increase of the criterion stops when the inequality is fulfilled \( f \left( x_1, x_2, \ldots, x_n \right) \leq J^{(k+1)} \), is accepted \( k = k+1 \), and the points obtained by rule (6) will be called 'search points', as it is done in [2]. Each search point has a double designation: if the point appeared for the first time and it happened on the step, it receives the number \( s \) and is denoted by \( J_s \). At each subsequent step with a number \( k \geq s \), the same point \( J_s \), depending on the location on the segment \( [J_i, J_s] \), receives one of the numbers \( i \) \((0 \leq i \leq k) \) and is denoted by \( J_i \). Here the number \( i = i(k) \) as well as the number \( s = s(k) \) depend on the step number \( k \).

We show that the point \( J_{s+1} \) satisfies inequalities \( J_{s+1} < J_{s+2} < J_* \).
The search for the minimum of the function \( f(x) \) starts from any valid point \( x_i \in X \) at which it is calculated \( J_x = f(x_i) \).

Suppose \( k \) iterations are made. The point at which the inequality \( f(x_1, x_2, \ldots, x_n) \leq J^{(k+1)} \) is changed to inequality \( f(x_1, x_2, \ldots, x_n) > J^{(k+1)} \) is taken as the left boundary \( J^{(k)} \) of the segment \([J_{k-1}, J_k]\). The next iteration \( k = k + 1 \) starts from a point \( J_{k-1} \), but with a step twice smaller than in the previous step. This ensures that the new point \( J_x \) of change of inequality \( f(x_1, x_2, \ldots, x_n) \leq J^{(k+1)} \) to inequality \( f(x_1, x_2, \ldots, x_n) > J^{(k+1)} \) does not exceed the point \( J^{(k+1)} \), i.e. \( J_x \leq J^{(k+1)} \). Similarly, the point \( J_{k-1} \) obtained by changing inequality \( f(x_1, x_2, \ldots, x_n) \leq J^{(k+1)} \) to inequality \( f(x_1, x_2, \ldots, x_n) > J^{(k+1)} \) is not less than the point \( J^{(k)} \), i.e. \( J_x > J^{(k)} \). It follows that

\[
J_{k-1} < J^{(k+1)} < J_x.
\]

Thus, following the new search point \( J_{k+1} \) on \( k + 1 \) the step, two new segments \([J_{k-1}, J_{k+1}]\) and \([J_{k+1}, J_x]\) are obtained. As a result of iterative process (6) the system of nested segments is formed:

\[
[J_{2}, J_1],
[J_4, J_3],
\ldots
[J_{2k}, J_{2k-1}],
\]

(7)

Let \( J_* \) be the limit point of the sequence \( \{J_k\} \), then in the system of segments (7) each segment contains a point \( J_* \) and finite set of search points other than \( J_* \).

We introduce a set consisting \( N(J_*) \) of all those numbers \( k \geq 1 \) for which either \( J_{m(k)-1} < J^{(k+1)} \), or \( u_{m(k+1)} < u_{m(k)} \). This means that at \( k \in N(J_*) \), the search point \( J_{k+1} \) falls inside the segment, \([J_{m(k)-1}, J_{m(k)}]\) forming one of the ends of the next segment \([J_{m(k)-1}, J^{(k+1)}]\).

Since each of the segments \([J_{m(k)-1}, J_{m(k)}]\) contains infinitely many search points, the set \( N(J_*) \) also contains infinitely many numbers.

The length of the segment \([J_{m(k)} - J_{m(k)-1}]\) is estimated by analogy with the method of dividing the segment in half. The formation of points in the considered method and in the method of dividing the segment in half is organized by a single mechanism: if the sign of inequality in system (5), at the new point, is reversed, then \( k = k + 1 \). In the method of dividing a segment in half, if at the \( k\)-th point \( J(x_{2k-1}) \leq J(x_{2k}) \), then the left border of the segment \([a_k, b_k]\) does not change, if \( J(x_{2k-1}) > J(x_{2k}) \), then the right border of the segment \([a_k, b_k]\) does not change. The length of the \( k\)-th segment in the method of dividing the segment in half is equal to

\[
b_k - a_k = \frac{b - a - \delta}{2^k} + \delta.
\]

Here \( \delta \) is a method parameter, \( 0 < \delta < b - a \), and it determines the accuracy of the calculations.

When estimating the length of a segment \([J_{m(k)} - J_{m(k)-1}]\), the value \( \delta \) is neglected, then

\[
J_{m(k)} - J_{m(k)-1} = \frac{J_0 - J_{\min}}{2^k},
\]

where \( J_{\min} \) is the lower estimate of the investigated function, \( J_0 \) is the initial value.
Hence, and from the countable set, $N(J_*)$ it follows that

$$\lim_{k \to \infty} (J_m(k) - J_m(k-1)) = 0$$

and

$$\lim_{k \to \infty} J_m(k) = J_*.$$

4. Results and discussion

All of the following results were obtained on the basis of retrospective prediction, that is, the test part of the sample $[Y_s(t)]$ is considered unknown and its elements are considered as variable parameters.

Studies of the dependence of the forecast quality on the volume of the training sample were modeled by the volume of the testing part of the sample. Here the depth of the forecast was consistently taken equal $\gamma = 12, 32, 52$.

The initial data were taken from state statistics on fuel prices [10] during the period from 2002 to 2008. The forecast was obtained for the initial sample. Then it was compared with the data from the international report [9], the results of which were obtained on the basis of the same initial data.

Figure 1 shows the initial process of changing costs over time $c(t)$. Dates are deferred on the abscissa axis. The sample size of 72 points.

![Figure 1. Initial statistical dependence.](image)

The following three graphs (figures 2-4) show the results of forecasting the cost of fuel taken from [9]. Figure 2 shows the forecast made for the training period of 20 points, obtained on the basis of data from [10]. As can be seen from figure 2, the forecast and the tabular curve diverge significantly from each other.
Figure 2. Forecast, based on the algorithm from [9], made by the period of training 20 points.

The same large divergence is observed in the forecast for 40 points (figure 3) and 60 points (figure 4).

Figure 3. Forecast, based on the algorithm from [9], made by the period of training 40 points.

As can be seen from figures 2-4, the forecast obtained by the algorithm from [9] differs significantly from the forecast made by the new method, figures 5-7. The calculation of the mean absolute percentage error (MAPE) was made for each forecast. Among all the forecasts obtained by the algorithm from [9], for most accurate the MAPE value is 62.0397%. This forecast was made on the basis of the training sample containing 40 values for fuel cost (figure 3). Whereas the most accurate of the forecasts obtained by the new method corresponds to the MAPE value of 2.6917%. This forecast was obtained on the basis of the training sample of 60 points (figure 5). Thus, the new method of time series forecasting using
global optimization demonstrates the most accurate results from the reviewed. Figure 6 shows the prediction obtained by the new method when training on 40 points. Figure 7 shows the prediction obtained by the new method when training on 20 points.

Figure 4. Forecast, based on the algorithm from [9], made by the period of training 60 points.

Figure 5. Forecast, based on the new method, made by the period of training 60 points.
5. Conclusion

As a result of numerical experiments the following results were obtained:

1. Figures 2-4 show a significant divergence of the predicted process from the real one, which is due to the unsuccessful choice of the predictive function in the form of a trigonometric polynomial.
2. Application of the new method to the problem solution shows its efficiency. As can be seen from figures 5-7, the difference between the actual process and its forecast is not more than 7.5%.

3. The accuracy of the forecast is related to the size of the training part of the sample. The more elements are considered in the process of training the model, the higher the accuracy of the forecast.

References
[1] Box G E P, Jenkins G M, Reinsel G C, Ljung G M 2015 *Time Series Analysis: Forecasting and Control* (New York: John Wiley & Sons. Inc)
[2] Anderson T W 1971 *The Statistical Analysis of Time Series* (New York: John Wiley & Sons. Inc.)
[3] Hyndman R J, Khandakar Y 2008 Automatic time series forecasting: The forecast package for R. *Jour. of Statistical Software* 27(1) doi: 10.18637/jss.v027.i03
[4] Leigh C, Kandanaarachchi S, McGree J M, Hyndman R J, Alsibai O, Mengersen K, Peterson E E 2019 Predicting sediment and nutrient concentrations in rivers using high-frequency water quality surrogates, available at: https://arxiv.org/abs/1810.12499
[5] Wickramasuriya S L, Athanasopoulos G, Hyndman R J 2019 Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J. American Statistical Association* 114(526) 804 doi: 10.1080/01621459.2018.1448825
[6] Aghabozorgi S, Shirkhorshidi A S, Wah T Y 2015 Time-series clustering – A decade review. *Information Systems* 53 16 doi: 10.1016/j.is.2015.04.007
[7] Liao T W 2005 Clustering of time series data – a survey. *Pattern Recognition* 38(11) 1857 doi: 10.1016/j.patcog.2005.01.025
[8] Duan L, Fusheng Yu, Pedrycz W, Xiao W, Xiyang Y 2018 Time-series clustering based on linear fuzzy information granules. *Applied Soft Computing* 73 1053 doi: 10.1016/j.asoc.2018.09.032
[9] Balyk V M, Balyk E V, Thuong N Q, Osadchuk M A 2019 Inverse Forecasting Problems of Time Series in the Harmonic Basis. *Int. Jour. of Applied Mathematics and Statistics* 58(4) 86
[10] *Central Statistical Database under Federal Service of State Statistics*, available at: https://www.fedstat.ru/indicator/31448.