Математичні моделі та методи

УДК 62.5 **doi: 10.26906/SUNZ.2018.2.034**

V.M. Galai

Poltava National Technical Yuri Kondratyuk University, Poltava

THE RESEARCH INTO THE PROBLEM OF STATISTICALLY INDETERMINATE TIME SERIES PREDICTION

In the article under consideration the appropriateness of prediction task optimizing according to related external prediction quality requirements based on the multitude of elements expanded by identification methods is proved by 15 mathematical models of time series and 4 methods of their identification.

Keywords: *prediction, identification, measurement, signals, obstacles, models, optimization.*

Introduction

The response rate is one of the most fundamental qualities of all the processes that occur in any real-time objects. Any physical, economic, biological, social or other value cannot change momentary. As possessing the quality of the response rate, it keeps its current value at the finite, though possibly infinitely small, interval of time after the impact of finite disturbance power. Accordingly, this value is smooth time function, i.e. it has one or more finite time derivatives. Then, according to the 1st and 2nd Weierstrass theorem, it can be approximated at the finite amount of time by Taylor or Fourier series, differential or difference equations as their discrete analog. The latter may be equally spaced or unequally distant. Naturally, the measurement of value of any origin contains, apart from its precise value, the measure of inaccuracy which, as a rule, is of random nature or a consequence of many indeterminable factors.

Different time series models, their identification methods and the criteria for optimality related to the prediction task are applied depending on the length of data selection, insights of the process (trend) which is being predicted, level and aprior information concerning the measure of inaccuracy. For example, there can be autoregressive model (AR), autoregressive moving-average model (ARMA), autoregressive integrated moving average model (ARIMA), Kalman filters, neural network (NN), etc [1, 4, 6].

Different criteria are used for the choice of mathematical models of time series:

– visual estimation of the inaccuracy graph; – autocorrelation functions of approximation inaccuracy series by the model with its importance factor based on the following criteria: Durbin-Watson model (DW), Q-statistic, Student statistic (t-statistic), Fisher statistic (f-statistic), Akaike informative criterion (AIC), Bayes-Schwarz information criterion (BSC) and other statistical values of model adequacy to time series [2, 3, 5].

Valid usage of the criteria given above is possible on the condition that time series measurements meet the statistical representation requirements. As a rule, these are long series (radio-technical, seismic and other systems and signals). In economics and other time (trend, obstacle) varying control systems the series are typically short (tens of measurements in time) with the reasonable prior uncertainty in characteristics. In this case the complex of DW, Q, T, F, AIC, BSC, etc. criteria is used to boost the identification process reliability. However, alongside with that, one of the parameter estimation method for model series parameter is applied.

Taking into account the high computerization and algorithm development level of time series prediction process, it is required to make a research into the problem of prediction optimizing appropriateness as exemplified in actual time series, to expand the multitude "model criteria" by subset "the methods of model parameter estimation" and to evaluate the results of such expansion in prediction system components which are being optimized.

The task setting

The criteria multitude. While selecting different structure options, one can chose the best structure which meets the I criterion as for the accuracy of prediction. Power models are more suitable for short series, while autoregressive ones are better for long series. Index I of the prediction accuracy, that is to be realized physically, is presented as the sum of quotient values I_i (i=1,2,3), which account for the quality of model series separate properties. Index I_1 :

$$I_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\beta_i^{\Pi} - \hat{\beta}_i^{n}\right|}{\hat{\beta}_i}, \qquad (1)$$

where $\hat{\beta}_i^{\Pi}$, $\hat{\beta}_i^{n}$ i $\hat{\beta}_i$ is the estimation of i model parameter, obtained at the selection of paired and unpaired dis-

crete samples k of time t_k; this is the so-called [7] unbiasedness of estimate index. β. Index:

$$I_2 = \left(\varepsilon^T \varepsilon\right) \cdot \left(x^T x\right)^{-1}, \quad (2)$$

where $\varepsilon^E = [\varepsilon(1),...,\varepsilon(M)]$, $\hat{x}^T = [\hat{x}(1),...,\hat{x}(M)]$, $\varepsilon(k)$ is inaccuracy of signal approximation x(k) by the corresponding $\hat{x}(k)$ model in k point within the series; this is so-called [8] unbiasedness or modeling accuracy by model series index. Index I_3 indicates the prediction accuracy at L - last points by the model based on the selection of M-L points:

$$I_3 = |1 - K|,$$

$$K = \frac{\sum_{i=1}^{L} \eta_i \left| x(M-i) \cdot \hat{x}(M-i) \right|}{\sum_{i=1}^{L} \eta_i \left| x(M-i) \cdot \sum_{i=1}^{L} \eta_i \left| \hat{x}(M-i) \right| \right|}. \quad (3)$$

Here η_i is distribution coefficient of desired prediction accuracy according to L - last points of selection X(k), $k = \overline{1, M}$; $\sum_{i=1}^{L} \eta_i = 1$; $\hat{x}(M-i)$ are predicted values x(M-i), obtained from the model, based on the selection reduced at L last points $k = \overline{1, M-L}$. It is generally accepted that the predicted series x(k) is made of the insight determined component, smooth in time, and the component close to White Gaussian Noise. That is why in the set of variate values models, arranged by their difficulty (the measurability of β vector of unknown parameter), indexes I_1 and I_3 limit the measurability of n vector β, whilst index I_2 at increasing n decreases.

Weight indexes g_i as weighted sum of the following three components, are given depending on the identification objective:

$$\hat{I} = \sum_{i=1}^{3} g_i I_i, \ \sum_{i=1}^{3} g_i = 1, \ g_i \geq 0. \quad (4)$$

To control the parameters β_i of the model with known structure, the maximum weight is g_1; for the problem of series exact approximation x(k) by model $\hat{x}(k)$ is g_2; for the prediction problem is g_3. Indexes I_1,I_2, I_3 in the aggregate provide the trading-off for the model estimation stability, accuracy approximation and prediction.

The aim of the experiment: to demonstrate the appropriateness of applying the group of external estimation criteria not only for predicting quality by different time series models, but also the advisability of applying groups of methods for these models' identification.

The research content

Through the example of the time series containing 43 discrete samples x(k) with uniform lead $\Delta t = 4$ months (one of the power economy parameter in Ukraine, Fig. 1), the solution for the problem of prediction x(k), $k = \overline{1.37}$, for 6 last points, regarded as unknown, shall be considered.

Such task setting for the research allows implementing physically infeasible for implementation in terms of the future prediction objective index I of the relative prediction accuracy for these 6 points, i.e. to calculate relative standard deviation

$$\varepsilon(k) = \hat{x}(k) - x(k), \ k = \overline{38.43}$$

of the predicted values $\hat{x}(k)$ from the known x(k), which is being optimized in the multitude of 15 models and 4 methods of their identification.

$$I = \frac{[\varepsilon(38),...,\varepsilon 43] \cdot [\varepsilon(38),...,\varepsilon 43]^T}{[x(38),...,x43] \cdot [x(38),...,x43]^T}. \quad (5)$$

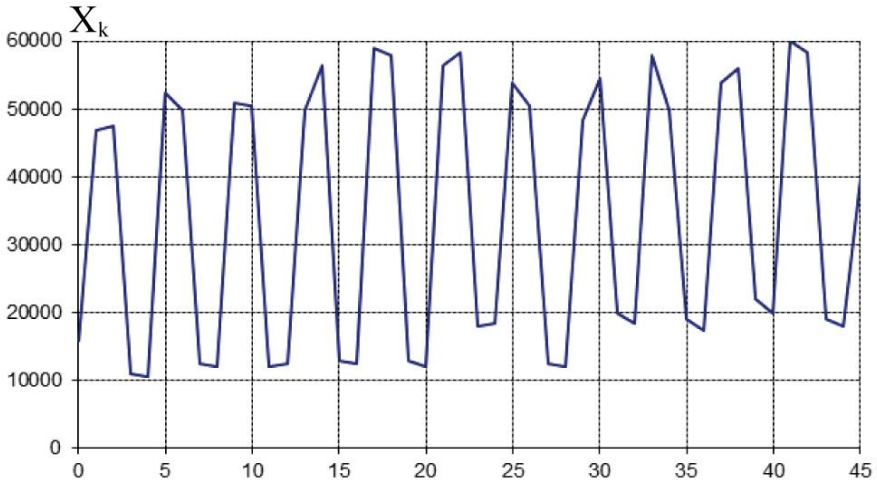

Fig. 1. Time series

In criterion (3)

$$\eta = \frac{1}{L} = \frac{1}{6}$$

is taken, in generalized criterion (4):

$$g_1=0.9;\ g_2=g_3=0.4.$$

Then it should be determined, in what way index (4), which is being implemented, corresponds to the ideal (5), that cannot be implemented physically.

Mathematical models multitude.

1. Model in the form of power polynomial for discrete samples of time k:

$$\hat{x}(k) = \beta_0 + \beta_1 k, \quad (6)$$

$$\hat{x}(k) = \beta_0 + \beta_1 k + \beta_2 k, \quad (7)$$

$$\hat{x}(k) = \beta_0 + \beta_1 k + \beta_2 k^2 + \beta_3 k^3, \quad (8)$$

$$\hat{x}(k) = \beta_0 + \beta_1 k^{\frac{1}{4}} + \beta_2 k^{\frac{1}{3}} + \beta_3 k^{\frac{1}{2}} + \beta_4 k^{\frac{3}{2}}, \quad (9)$$

$$\hat{x}(k) = \beta_0 + \beta_1 k + \beta_2 k^{-1} + \beta_3 k^{-3}. \quad (10)$$

2. Autoregressive model for fixed and variable lead k:

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-1); \quad (11)$$

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-1) + \beta_2 x(k-2), \quad (12)$$

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-1) + \beta_2 x(k-2) + \beta_3 x(k-3); \quad (13)$$

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-4); \quad (14)$$

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-1) + \beta_2 x(k-2) + \beta_3 x(k-3) + \beta_4 x(k-4); \quad (15)$$

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-1) + \beta_2 x(k-4); \quad (16)$$

$$\hat{x}(k) = \beta_0 + \beta_1 x(k-1) + \beta_2 x(k-4) + \beta_3 x(k-8). \quad (17)$$

3. Combined polynomial-time and autoregressive models:

$$\hat{x}(k) = \beta_0 + \beta_1 k + \beta_2 x(k-1), \quad (18)$$

$$\hat{x}(k) = \beta_0 + \beta_1 k + \beta_2 x(k-4), \quad (19)$$

$$\hat{x}(k) = \beta_0 + \beta_1 k + \beta_2 x(k-1) + \beta_3 x(k-4). \quad (20)$$

The multitude of model identification methods (6 ÷ 20)

1. Least square method (LSM) The estimation $\hat{\beta}$ of models (6...20) parameter vector β is calculated under the condition that:

$$\hat{\beta} = \arg\min_{\beta} \left[\hat{\varepsilon}(1),...,\hat{\varepsilon}(37)\right]\cdot\left[\hat{\varepsilon}(1),...,\hat{\varepsilon}(37)\right]^{T} \frac{\partial^2 \Omega}{\partial u \partial v}, \quad (21)$$

where $\varepsilon(k) = x(k) - \hat{x}(k)$, $k = \overline{1.37}$.

2. Generalized least squares method (GLSM) The estimation $\hat{\beta}$ of models $(6 \div 20)$ parameter vector β is calculated under the condition that:

$$\hat{\beta} = \arg\min_{\beta} [\tilde{\varepsilon}(1),...,\tilde{\varepsilon}(37)]\cdot[\tilde{\varepsilon}(1),...,\tilde{\varepsilon}(37)], \quad (22)$$

where

$$\tilde{\varepsilon}(k) = \tilde{x}(k) - \hat{x}(k),\ k = \overline{1.37};\ \tilde{x}(k)$$

is moving average x(k):

$$\tilde{x}(k) = \frac{1}{5}\sum_{i=k-2}^{k+2} x(k+i).$$

3. *Correlation method (CM) [9].* The estimation $\hat{\beta}$ of models (6...20) parameter vector β is calculated under the condition that:

$$\hat{\beta} = \arg\min_{\beta} \sum_{p=1}^{5} \left[\varepsilon(1),...,\varepsilon(37-p)\right]\cdot\left[\varepsilon(p),...,\varepsilon(37)\right]^{T}, \quad (23)$$

So, under the condition of sum minimum displaced at p discrete Δt running time $\varepsilon(k)$ for $\varepsilon(k+p)$.

4. Intermediate variable method (IVM) The estimation $\hat{\beta}$ of models (6...20) parameter vector β is calculated in the same way as in LSM– estimation (21), but, instead of sensitivity function $\frac{\partial \varepsilon}{\partial \hat{\beta}}$ some auxiliary function U with components U_i. is taken. In our example U_i is equal to signum function of $\frac{\partial \varepsilon}{\partial \beta_i}$.

Numeral experiment

The effectiveness of the criteria applying that is to be realized physically upon the object of its proximity till the required criteria that cannot be implemented physically has been checked on the multitude of 15 models and 4 methods (CM). The results of number modeling for models (6)...(20) are given in 15 lines of table 1. In columns 1...10 the following data are given:

1 – model types (power (6) - (10), autoregressive (11) – (17), combined (18) – (20));

2 – relative mean-square error of the series modeling.

By the corresponding model for $k = \overline{1.37}$ at its identification according to LSM; 3 – ideal criterion (5), that cannot be implemented physically for the model obtained by LSM;

4 – the criterion (4), that is to be realized physically at model identification according to LSM;

5 – the best, according to the criterion (4), identification method for the corresponding model line;

6 – the ideal criterion value (5), selected according to the method for the real criterion (4) and the corresponding model line;

7 – the criterion value (4) for the selected best identification model for the corresponding model line;

8 – the best, according to the ideal (5), identification method or the corresponding model line;

9 – the ideal criterion value (5) for the criterion (4) for the best identification method for the corresponding model line according to the criterion (5).

The analyses of the experiment results:

1. The autoregressive model (17) with variable delay at k-1, k-4 i k-8 steps (IVM method) has proved to be the best one according to the ideal criteria (5) basing on the multitude of 15 models and 4 identification methods The same results have been obtained according to the real criterion (4). Generally, at 8 cases out of 15 under consideration the optimal identification method according to real criteria has been selected correctly (lines 2, 3, 6, 7, 10, 11, 12, 15 in the table), so, it coincides with the method, selected according to the ideal criterion (5). In other 7 cases (lines 1, 4, 5, 8, 9, 13, 14) the ideal value (5) for the method, selected according to the real value (4), is just a little worse than the same value for the optimal, according to the ideal value, method (columns 6 and 9).

2. For power series (6), (7), (8), value (2) (II column of the table) relative mean-square approximation error of the series modeling (6) – (8) is decreasing, with is the natural consequence of Weierstrass theorem about the approximation by Taylor series. At the same time the ideal criterion of prediction accuracy at the model complication worsens (lines 1,3 of the third column in table). This proves the biased nature of the inner approximation criterion (2) and the inexactness of its application for the prediction problem-solving.

Table 1

Modeling results

№	1	2	3	4	5	6	7	8	9	10	11
1	6	0.49	0.47	0.25	**GLSM**	0.41	0.24	**IVM**	0.36	0.26	1.3
2	7	0.48	0.59	0.26	**ILSM**	0.41	0.22	**CM**	0.41	0.22	1.4
3	8	0.47	0.88	0.41	**ILSM**	0.38	0.2	**CM**	0.38	0.2	2.32
4	9	0.485	0.593	0.27	**GLSM**	0.43	0.226	**CM**	0.365	0.235	1.62
5	10	0.488	0.49	0.25	**ILSM**	0.45	0.23	**GLSM**	0.425	0.237	1.15
6	11	0.49	0.435	0.24	**GLSM**	0.42	0.235	**GLSM**	0.415	0.235	1.05
7	12	0.62	0.58	0.28	**GLSM**	0.56	0.262	**GLSM**	0.558	0.262	1.04
8	13	0.123	0.143	0.04	**LSM**	0.143	0.04	**IVM**	0.096	0.048	1.49
9	14	0.133	0.1	0.03	**LSM**	0.1	0.03	**IVM**	0.088	0.126	1.13
10	15	0.113	0.122	0.037	**IVM**	0.092	0.03	**IVM**	0.092	0.03	1.33
11	16	0.131	0.103	0.034	**IVM**	0.091	0.031	**IVM**	0.091	0.031	1.13
12	17	0.087	0.092	0.015	**IVM**	0.063	0.011	**IVM**	0.063	0.011	1.46
13	18	0.488	0.47	0.245	**IVM**	0.489	0.225	**GLSM**	0.411	0.237	1.14
14	19	0.132	0.108	0.035	**LSM**	0.108	0.035	**IVM**	0.081	0.038	1.33
15	20	0.131	0.111	0.036	**LSM**	0.111	0.037	**LSM**	0.111	0.037	1

3. The situation is slightly different for the autoregressive and combined polynomial-time and autoregressive models (11)-(20). In this case, the inner criteria (2) of mean-square proximity measurement at the approximation section and both the ideal (5) and the ($k = \overline{1.37}$) real (4) criteria become substantially correlated as a result of LSM regulation property, when variables are noise-contaminated.

Consequently, for this model class the application of the approximation criterion (2) for the prediction problem at points (38...43) for the noise-contaminated data at points (1...37) is less crucial. This is the case of self-regulation.

The more complex auto regression is, the worse the conditioning of information LSM matrix for the exact data becomes. But for more noise-contaminated and obstacle non-correlated data the diagonal elements of this matrix expand and, as a result, they cause the decrease according to LSM module - the estimation of model indexes, alongside simplifying (or, according to A. M. Tihonov [3], regulating) the model.

4. Let us compare the value of the ideal criterion (5) for the models, obtained according to LSM (column 3) and for one of the suggested methods (column 6) with real criterion (4) optimization.

The index (5) was unessentially smaller only for model (18) out of 15 models. So, only in this case according to criterion (4) LSM was mistakenly selected instead of IVM. In other 14 cases the method, obtained under the condition of minimal criterion (4) of prediction accuracy, provides better or practically the same results as LSM, if according to (4) it has been chosen as better than LSM (columns 6 and 3 in table 1).

5. In terms of one method identification, e.g. IVM (column 6, lines 10...13) ideal criterion dispersion (5) depending on model structure makes up from 0.063 to 0.489, which proves the effectiveness of model structure selection.

Within the framework of one model, e.g. model (17) optimal according to the criteria (5), the optimal solution in terms of the multitude of 4 methods (LSM, IVM, GLSM, CM) gives advantage of 1.5 times (0.092 for LSM and 0,063 for IVM as an optimal method). This proves the effectiveness of identification method selection.

6. Generally, the optimization of identification models and methods provides substantial profit in the prediction accuracy.

The advantage can be defined as the ratio of criterion (5) for the index model obtained according to LSM (column 3 of table 1) to the same criterion (5) value for the index model obtained according to the optimal method (5) (column 9 of table 1).

In column 11 of table 1 this ratio is given as to be calculated in the range from 1 to 2.32.

Conclusion

In general, the optimization of identification models and methods provides substantial profit in the prediction accuracy.

References

1. Bidyuk P.I., Polovtsev O.V. Analiz i modelyuvannya ekonomichnykh protsesiv perekhidnoho periodu. – Kyiv: NTU KPI, 1999. – 210 p.

2. Ferster E., Rents B. Metodyi korrelyatsionnogo i regressionnogo analiza. – M.: Finansyi i statistika. 1983. – 302 p.

3. Enders W. Applied Econometric Time Series. – New York: Wiley & Sons, Inc., 1995. – 433 p.

4. Minaev Yu.N., Filimonova O.Yu., Benameur Lies. Metodyi i algoritmyi identifikatsii i prognozirovaniya v usloviyah neopredelennosti v neyrosetevom logicheskom bazise. – M.: Goryachaya liniya-Telekom, 2003. – 205 p.

5. Ivahnenko A.G. Samoobuchayuschiesya sistemyi raspoznavaniya i avtomaticheskogo upravleniya. – K.: Tekhnika, 1969. – 392 p.

6. Yakovlev V.L., Yakovlev G.L., Lisitskiy L.A. Sozdanie matematicheskih modeley prognozirovaniya pri pomoschi neyrosetevyih algoritmov // Informatsionnyie tehnologii.

7. Ivahnenko A.G. Dolgostrochnoe prognozirovanie i upravlenie slozhnyimi sistemami. – K.: Tekhnika, 1975. – 312 p.

8. Silvestrov A.N, Chinaev P.I. Identifikatsiya i optimizatsiya avtomatichiskih sistem – M.: Energoatomizdat, 1983. – 280 p.

9. Tihonov A.N., Arsenin V.Ya. Metodyi resheniya nekorrektnyih zadach – M.: Nauka, 1979 – 286 p.

Надійшла до редакції 03.03.2018

Рецензент: д.т.н., проф. А.М. Сільвестров, Національний технічний університет України “КПІ”, Київ.

ИССЛЕДОВАНИЕ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ СТАТИСТИЧЕСКИ НЕОПРЕДЕЛЕННЫХ ВРЕМЕННЫХ РЯДОВ

В.Н. Галай

В статье доказано на 15 математических моделях временных рядов и 4 методах их идентификации целесообразность оптимизации задачи прогнозирования с соответствующим внешним критерием качества прогноза на расширенном методами идентификации множестве элементов.

Ключевые слова: *прогноз, идентификация, измерения, сигналы, помехи, модели, оптимизация.*

ДОСЛІДЖЕННЯ ЗАДАЧІ ПРОГНОЗУВАННЯ СТАТИСТИЧНО НЕВИЗНАЧЕНИХ ЧАСОВИХ РЯДІВ

В.М. Галай

В статті доведено на 15 математичних моделях часових рядів та 4 методах їх ідентифікації доцільність оптимізації задачі прогнозування за відповідним зовнішнім критерієм якості прогнозу на розширеній методами ідентифікації множині елементів.

Ключові слова: *прогноз, ідентифікація, вимірювання, сигнали, перешкоди, моделі, оптимізація.*