UDC 004.056:896
DOI: 10.15587/2312-8372.2020.202498

Lakhno V.,
Sagun A.,
Khaidurov V.,
Panasko E.

DEVELOPMENT OF AN INTELLIGENT SUBSYSTEM FOR OPERATING SYSTEM INCIDENTS FORECASTING

Об'єктом дослідження є підсистема прогнозування інцидентів роботи операційної системи (ОС) серверної платформи, яка функціонує на базі операційної системи сімейства Windows. Одним із найбільш проблемних місць при плануванні заходів запобігання шкідливим наслідкам мережевих атак типу dDOS, апаратним відмовам серверної системи тощо є отримання ефективної моделі передбачення інцидентів роботи операційної системи.

У ході досліджень використовувалися методи формування та дослідження часового ряду, експоненціального згладжування, елементи теорії машинного навчання на базі методу групового врахування аргументу (МГВА). Для отримання точних і надійних прогнозів роботи інтелектуальної підсистеми прогнозування інцидентів було застосовано елементи теорії евристичної самоорганізації та конкретна реалізація даної теорії – МГВА. Отримано алгоритм та розроблена програмна реалізація інтелектуальної системи прогнозування інцидентів роботи операційної системи та основні характеристики її роботи. Це стало можливо в результаті аналізу побудованої моделі порушника, системного журналу інцидентів безпеки та використанню МГВА. Запропоновано механізм формування вибірки подій інцидентів роботи ОС на основі системного журналу подій Windows. Проведене тестування запропонованої підсистеми прогнозування на базі тестових вибірок дозволяє стверджувати, що результати прогнозування, отримані при різних налаштуваннях системи машинного навчання та параметрів (степінь опорного поліному, кількість змінних в моделі характеристичного поліному, кількість рядів селекції) є задовільними. У результаті застосування створеного алгоритму прогнозування інцидентів роботи ОС було показано, що застосування великої кількості поліноміальних моделей у МГВА дає змогу отримати підсистему прогнозування, яка якісно переважає системи, засновані на класичних регресійних моделях та методах. Завдяки цьому можливо отримати значно точніший прогноз у порівнянні з класичними регресійними методами або методом експоненціального згладжування, що дає відсоткове відношення хибних обрахунків з використанням МГВА не більше 4 %.

Ключові слова: *часовий ряд, підсистема прогнозування, машинне навчання, поліноміальна модель, метод групового врахування аргументів.*

Received date: 28.01.2020
Accepted date: 26.02.2020
Published date: 30.04.2020

1. Introduction

Most authors do not raise the issue of classifying methods and models for predicting the operation of operating systems (OS). It concerns the forecasting of security events and algorithms, or forecasting models that should be used for this purpose, it is not possible to name specific algorithms or methods. As a review of the literature shows, currently the most popular are classical forecasting models (trending, regression), forecasting using neural networks and Markov models [1–3]. Scientists make a special contribution to the theory and practice of creating algorithms, methods and forecasting systems in [4, 5]. Therefore, it is relevant to analyze critical operating modes of operating systems using modern methods of forecasting time series, as well as developing new effective machine learning methods based on GMDH for use in incident forecasting subsystems. Thus, *the object of research* is the subsystem for forecasting incidents of the operating system of the server platform, which operates on the basis of the operating system of the Windows family. *The aim of research* is to create a software tool for the subsystem for predicting incidents of operating the server platform OS based on the Windows family of OS using time series forecasting using machine learning methods.

2. Methods of research

The subject of the model of the forecasting subsystem is the time series. As such a series, the system events log of the OS from the fixation system and the accounting of various security incidents in the Windows Server OS are used. In general, such systems react and register such violations of information properties as: violation of information confidentiality, violation of information integrity, violation of information availability, violation of system manageability and the like. The content of this journal, in general, correlates with the model of the intruder [6].

Among examples of the actions of the intruder and typical attacks on the operating system, information about which is contained in the logs of system events of the Windows OS, for example, there are following [7]:

- attempts to scan the file system and steal key information;
- password selection;
- collecting data from non-empty Windows recycle bin;
- excess of access authority;
- software bookmarks;
- greedy programs.

The time series for the subsystem for predicting incidents of OS operation is the sampling of critical events in the OS for 1, 2, 3, 4, 5, 6, 7, 8 days of the OS from the system log shown in Table 1.

Table 1

The appearance of the time series for the forecast

Day	Event	Number of samples	Day	Event	Number of samples
1	Error	234	5	Error	36
October 19th	Warning	30	October 23th	Warning	68
2	Error	39	6	Error	54
October 20th	Warning	18	October 24th	Warning	98
3	Error	39	7	Error	24
October 21th	Warning	53	October 25th	Warning	42
4	Error	16	8	Error	19
October 22th	Warning	19	October 26th	Warning	19

In MS Excel, there is quite a list of tools for statistical analysis. To test the forecast, select «Exponential Smoothing». In fact, the obtained time samples are elements of the time series (TS), which will be used later to obtain a forecast.

Let's obtain the following as a result of the analysis of the TS incidents of OS operation (Fig. 1).

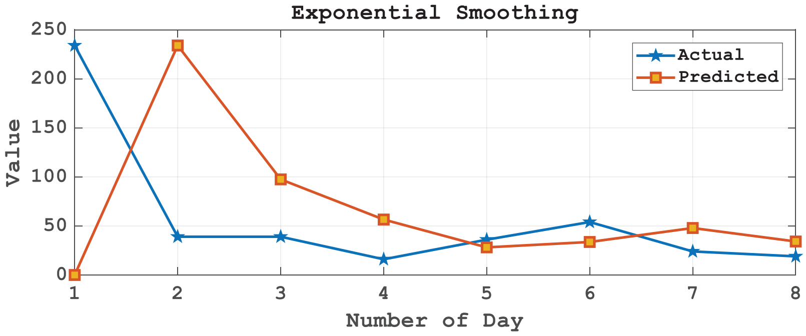


Fig. 1. The constructed forecast trend for graphically displaying the functional dependence of the number of events in time for the «Errors in the operating system» parameter for 8 days

For this TS, the average deviation is calculated, which is in the range from 18.67 to 119.85. The value of the smoothed levels for each of the 8 available values of the TS value indicator obtained in Fig. 1, allow to plan the expectations of such events for the next 8 days. The obtained polynomial forecasting model is not adequate for obtaining a forecast; therefore, the forecasting option using the method of group accounting of arguments (MGUA) is considered, and the resulting forecast is heuristic. The nature and absolute magnitude of the error do not allow to conclude about the reality of the trend.

To prove the correctness of this trend, it is possible to go in two ways: empirical; experimental. It is possible to use the predictive trend model using regression analysis [8]. But, to obtain accurate and reliable forecasts in the study of complex objects, for example, such as an incident registration system, the theory of heuristic self-organization and the concrete implementation of the theory – GMDH [9] are used. GMDH makes sense to use as a basic method for forecasting incidents, since the data sampling (Windows system event log) contains several elements [10–12]. Therefore, an inductive approach is used, according to which models of increasing complexity are successively generated until a minimum of some criterion of model quality is found. This quality criterion is called an external criterion, because when setting up models and evaluating the quality of models, various data are used. Achieving the global minimum of the external criterion when generating models means that a model that is able to find such a minimum is the desired one.

The algorithm for finding the optimal structure model for the incident forecasting subsystem can be represented in the form of the following steps [9]:

1. There is a sample in the form of the TS system log $D=\{(x_n,y_n)\}_{n=1}^{N}$, where $x\in\mathfrak{R}^m$. Due to the fact that for the GMDH operation it is necessary to conduct learning and testing, the sample is divided into educational and test ones. In the practical GMDH implementation, the percentage of these samples is manually selected.

Let l, C be sets from the range $\{1,..,N\}=W$. These sets satisfy the conditions for partitioning sets $l\cup C=$ $=\mathbb{W}_1 l\cap C=0$. The matrix X_l consists of row vectors x_n for which the index $n\in l$. The vector $\mathcal{Y}_l$ consists of those elements $\mathcal{Y}_n$ for which the index $n\in l$. The partition of the sample is written as follows:

$$X_{\mathbb{W}}=\left(\frac{X_l}{X_C}\right),\ \mathcal{Y}_{\mathbb{W}}=\left(\frac{\mathcal{Y}_l}{\mathcal{Y}_C}\right),\ \mathcal{Y}_{\mathbb{W}}\in\mathfrak{R}^{Nx1},$$

$$X_{\mathbb{W}}\in\mathfrak{R}^{Nxm},\ |l|+|C|=N.$$

2. Let's define the base model. This model describes the relationship between a dependent variable $\mathcal{Y}$ and free variables $\mathcal{X}$. For the forecasting algorithm being created, let's use the Voltaire functional series (the so-called Kolmogorov-Gabor polynomial):

$$\mathcal{Y}=\omega_0+\sum_{i=1}^{m}\omega_i x_i+\sum_{i=1}^{m}\sum_{j=1}^{m}\omega_{ij}x_i x_j+$$
$$+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\omega_{ijk}x_i x_j x_k+\ldots\ . \qquad (1)$$

In model (1), $x=\{x_i\,|\,i=1,..,m\}$ – set of free variables and ω – set of weights:

$$\omega=\langle\omega_i,\omega_{ij},\omega_{i,j,k},\cdots|\,i,j,k,\cdots=1,\cdots,m\rangle.$$

3. Based on the set objectives, the objective function are selected – an external criterion that describes the quality of the model. A few commonly used external criteria are described below.

4. Inductively generated candidate models. In this case, restrictions are introduced on the length of the polynomial of the base model. For example, the degree of polynomial of the base model should not exceed a specific natural value. Then the basic model is written as a linear combination of a given number $\mathbb{F}_0$ of products of free variables as follows:

$$\mathcal{Y} = f\left(x_1,x_2,..,x_1^2,x_1x_2,x_2^2,..,x_m^R\right), \tag{2}$$

where f – the linear combination function. Arguments (2) are redefined as follows:

$$x_1 \to a_1, x_2 \to a_2,..,x_1^2 \to a_\alpha, x_1x_2 \to a_\beta, x_2^2 \to a_\gamma,..,x_m^q \to a_{\mathbb{F}_0},$$

so,

$$\mathcal{Y} = f\left(a_1,a_2,..,a_{\mathbb{F}_0}\right).$$

For coefficients linearly included in the model, one-index numbering is specified in the following order: $\omega = \omega_1,..,\omega_{\mathbb{F}_0}$. In this case, the model can be represented as a linear combination of the form:

$$\mathcal{Y} = \omega_0 + \sum_{i=1}^{\mathbb{F}_0} \omega_i a_i. \tag{3}$$

Each model of the generated form (3) is defined by a linear combination of elements $\{(\omega_i,a_i)\}$ in which the set of indices $\{i\} = s$ is subset $\{1,..,\mathbb{F}_0\}$.

5. To configure these parameters, an internal criterion is used; it is calculated using the training sample. To each element of a vector x_n – a selection element D, a vector is mapped a_n. Next, a view matrix is constructed $A_{\mathcal{W}}$, which represents a set of column vectors a_i. The matrix $A_{\mathcal{W}}$ is divided into submatrices A_l and A_C. The smallest remainder of the form $\left|\mathcal{Y} - \hat{\mathcal{Y}}\right|$, where $\hat{\mathcal{Y}} = A\hat{\omega}$ returns the value of the parameter vector $\hat{\omega}$, which is calculated by the least squares method [10], respectively, of the expression:

$$\hat{\omega}_G = \left(A_G^T A_G\right)^{-1} A_G^T \mathcal{Y}_G,$$

where $G \in \{l, C, \mathbb{W}\}$. The internal criterion for the model applies the standard error of the form:

$$\varepsilon_G^2 = \left|\mathcal{Y}_G - A_G\hat{\omega}_G\right|^2.$$

In accordance with the criterion $\varepsilon_G^2 \to \min$, parameters ω are selected and errors are calculated on the test sample G, where $G = l$. When the model is complicated, the internal criterion does not give the minimum models of optimal complexity; therefore, it is not suitable for choosing a model.

6. To select the best models, let's calculate their quality. For this, a control sample and an external criterion are used. The error in the sample H is indicated as follows:

$$\Delta^2(H) = \Delta^2(H|G) = \left|\mathcal{Y}_H - A_H\hat{\omega}G\right|^2,$$

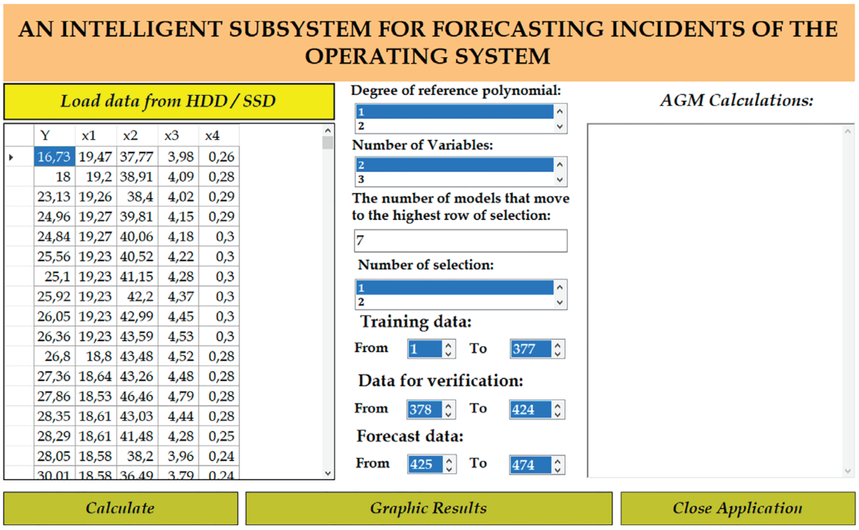
where $H \in \{l, C\}$, $H \cap G = 0$. This means that the error is calculated on the sample H with the model parameters obtained on the sample G.

7. A model that provides a minimum of external criteria is considered optimal.

3. Research results and discussion

An application in the C# forecasting language based on GMDH has been implemented. It contains an interface that allows to change the degree of the reference polynomial from 1 to 7, and the number of variables in the model of the characteristic polynomial from 2 to 7. It is assumed that a priori the number of models is unknown, go to the top row, therefore, this field on screen forms can be filled in manually. The number of selection rows for the model can be set from 1 to 10.

After downloading a sample of data from the security logs of Windows 10 OS of the investigated personal computer, the parameters of the generated GMDH models are set and the opportunity to enter the parameters of the separation of the sample for training/verification of the forecast is opened (Fig. 2).



AN INTELLIGENT SUBSYSTEM FOR FORECASTING INCIDENTS OF THE OPERATING SYSTEM

Y	x1	x2	x3	x4
16,73	19,47	37,77	3,98	0,26
18	19,2	38,91	4,09	0,28
23,13	19,26	38,4	4,02	0,29
24,96	19,27	39,81	4,15	0,29
24,84	19,27	40,06	4,18	0,3
25,56	19,23	40,52	4,22	0,3
25,1	19,23	41,15	4,28	0,3
25,92	19,23	42,2	4,37	0,3
26,05	19,23	42,99	4,45	0,3
26,36	19,23	43,59	4,53	0,3
26,8	18,8	43,48	4,52	0,28
27,36	18,64	43,26	4,48	0,28
27,86	18,53	46,46	4,79	0,28
28,35	18,61	43,03	4,44	0,28
28,29	18,61	41,48	4,28	0,25
28,05	18,58	38,2	3,96	0,24
30,01	18,58	36,49	3,79	0,24

Degree of reference polynomial:
1
2

Number of Variables:
2
3

The number of models that move to the highest row of selection:
7

Number of selection:
1
2

Training data:
From 1 To 377

Data for verification:
From 378 To 424

Forecast data:
From 425 To 474

AGM Calculations:

Fig. 2. Downloading a sample of a time series through a graphical interface of a software implementation of an algorithm for the intelligent prediction subsystem

ISSN 2664-9969

When calculating the parameters of the GMDH model, let's obtain the corresponding intermediate data of the stages of the calculation of models based on GMDH (Fig. 3):

- regularity criteria for S[1], S[2], S[3], S[4] models;
- global criterion at the selection level 1;
- module of the deviation of the estimation of the obtained forecasting models throughout the voters.

With an increase in the number of samples in the TS from the Windows security log, an improvement in the quality of the forecast (a decrease in the error criterion) may occur. But in this case, increasing the accuracy will require increasing the size of the data set for training and testing models from the supply TS for the model.

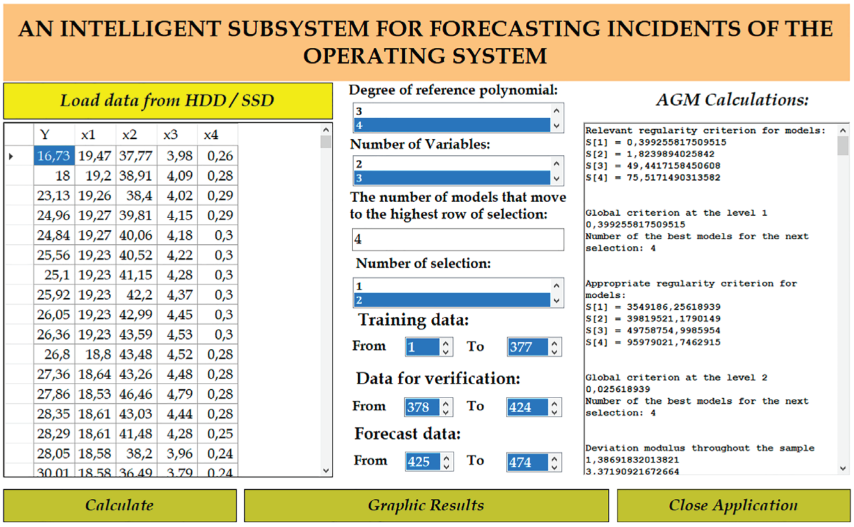
Depending on the goals and the expected duration of the forecast period for forecasting incidents, the parameters of the model can be changed on the dialogue form. But with an increase in the number of variables in the model and the degree of the reference polynomial, obtaining the best model for forecasting can increase significantly.

As a result of the program module, the mathematical model of training selects the best models, and as a result of the selection of the best models get the best model, which will be used to predict security incidents of the investigated OS (Fig. 4).

On the graph of the input and processed data of the forecasting model based on GMDH, it is possible to estimate the discrepancy between the data of the real sample and the data (shown in yellow) obtained on the basis of the best forecasting GMDH model (shown in red) (Fig. 5).

In order to predict the time series obtained from the incident log of the server OS based on GMDH, it is necessary to select the best models at each iteration of the method. This approach reduces the total number of calculations (processor time), and also reduces the amount of memory the work of the method itself. Comparing the forecasting results generated by the GMDH model and autoregression (Fig. 6), it is possible to conclude that it is possible to use the created software product in practice.

Prediction results are obtained with various system settings (Fig. 6) and various parameters (degree of the reference polynomial, number of variables of the characteristic polynomial model, number of selection series).


AN INTELLIGENT SUBSYSTEM FOR FORECASTING INCIDENTS OF THE OPERATING SYSTEM

Y	x1	x2	x3	x4
16,73	19,47	37,77	3,98	0,26
18	19,2	38,91	4,09	0,28
23,13	19,26	38,4	4,02	0,29
24,96	19,27	39,81	4,15	0,29
24,84	19,27	40,06	4,18	0,3
25,56	19,23	40,52	4,22	0,3
25,1	19,23	41,15	4,28	0,3
25,92	19,23	42,2	4,37	0,3
26,05	19,23	42,99	4,45	0,3
26,36	19,23	43,59	4,53	0,3
26,8	18,8	43,48	4,52	0,28
27,36	18,64	43,26	4,48	0,28
27,86	18,53	46,46	4,79	0,28
28,35	18,61	43,03	4,44	0,28
28,29	18,61	41,48	4,28	0,25
28,05	18,58	38,2	3,96	0,24
30,01	18,58	36,49	3,79	0,24

Degree of reference polynomial: 3 / 4

Number of Variables: 2 / 3

The number of models that move to the highest row of selection: 4

Number of selection: 1 / 2

Training data: From 1 To 377

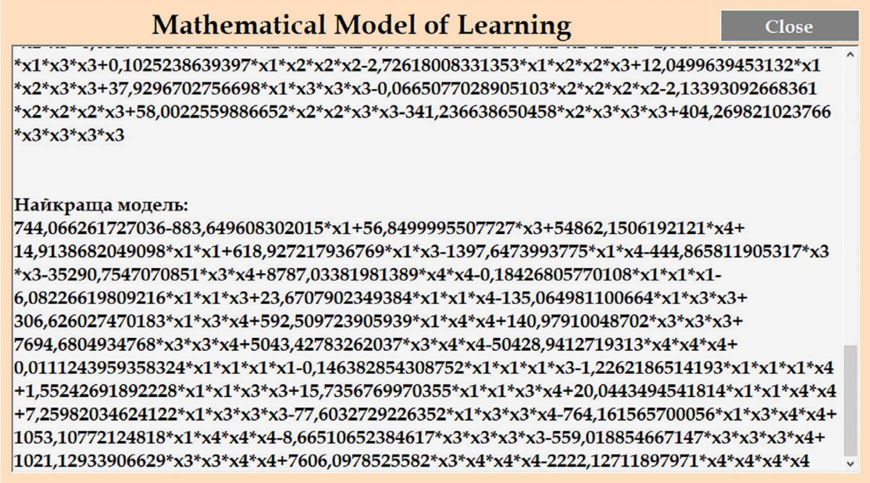
Data for verification: From 378 To 424

Forecast data: From 425 To 474

AGM Calculations:

```
Relevant regularity criterion for models:
S[1] = 0,399255817509515
S[2] = 1,8239894025842
S[3] = 49,4417158450608
S[4] = 75,5171490313582


Global criterion at the level 1
0,399255817509515
Number of the best models for the next
selection: 4


Appropriate regularity criterion for
models:
S[1] = 3549186,25618939
S[2] = 39819521,1790149
S[3] = 49758754,9985954
S[4] = 95979021,7462915


Global criterion at the level 2
0,025618939
Number of the best models for the next
selection: 4


Deviation modulus throughout the sample
1,38691832013821
3.37190921672664
```

Fig. 3. Downloaded input data, intermediate data on the stages of calculation of the forecasting model by the method of group accounting of arguments


Mathematical Model of Learning

```
*x1*x3*x3+0,1025238639397*x1*x2*x2*x2-2,72618008331353*x1*x2*x2*x3+12,0499639453132*x1
*x2*x3*x3+37,9296702756698*x1*x3*x3*x3-0,0665077028905103*x2*x2*x2*x2-2,13393092668361
*x2*x2*x2*x3+58,0022559886652*x2*x2*x3*x3-341,236638650458*x2*x3*x3*x3+404,269821023766
*x3*x3*x3*x3


Найкраща модель:
744,066261727036-883,649608302015*x1+56,8499995507727*x3+54862,1506192121*x4+
14,9138682049098*x1*x1+618,927217936769*x1*x3-1397,6473993775*x1*x4-444,865811905317*x3
*x3-35290,7547070851*x3*x4+8787,03381981389*x4*x4-0,18426805770108*x1*x1*x1-
6,08226619809216*x1*x1*x3+23,6707902349384*x1*x1*x4-135,064981100664*x1*x3*x3+
306,626027470183*x1*x3*x4+592,509723905939*x1*x4*x4+140,97910048702*x3*x3*x3+
7694,6804934768*x3*x3*x4+5043,42783262037*x3*x4*x4-50428,9412719313*x4*x4*x4+
0,0111243959358324*x1*x1*x1*x1-0,146382854308752*x1*x1*x1*x3-1,2262186514193*x1*x1*x1*x4
+1,55242691892228*x1*x1*x3*x3+15,7356769970355*x1*x1*x3*x4+20,0443494541814*x1*x1*x4*x4
+7,2598203462412*x1*x3*x3*x3-77,6032729226352*x1*x3*x3*x4-764,161565700056*x1*x3*x4*x4+
1053,1077212481*x1*x4*x4*x4-8,66510652384617*x3*x3*x3*x3-559,018854667147*x3*x3*x3*x4+
1021,12933906629*x3*x3*x4*x4+7606,0978525582*x3*x4*x4*x4-2222,12711897971*x4*x4*x4*x4
```

Fig. 4. The best and best forecasting models obtained by the method of group accounting of arguments obtained

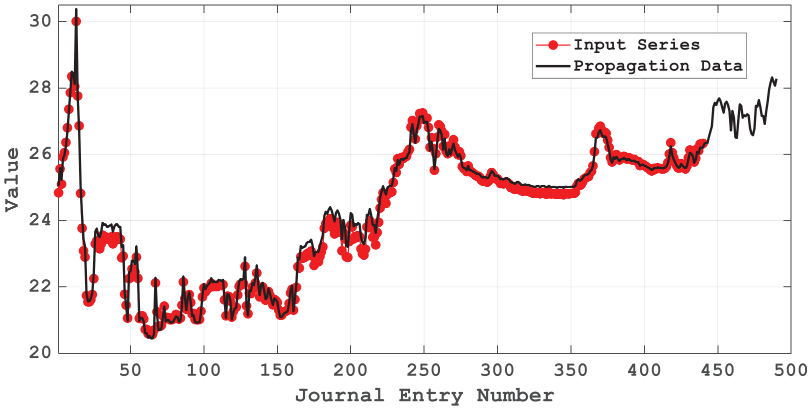


Fig. 5. Comparison of real sample data and data obtained on the basis of the best forecast model by the method of group accounting of arguments

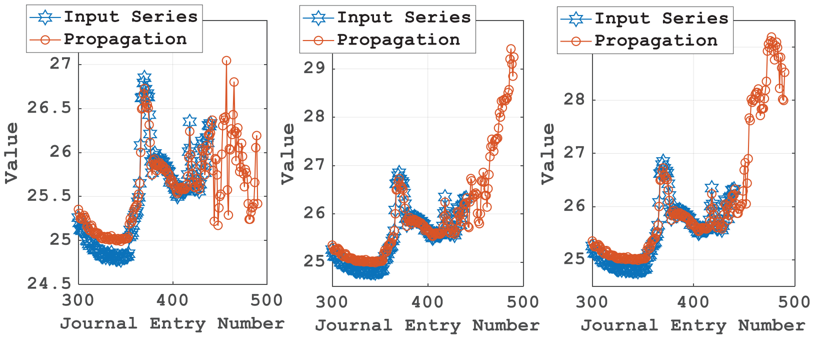


Fig. 6. Comparison of the forecasting results obtained by the model based on the method of group accounting of arguments and autoregression on 3 types of input parameters of machine learning

4. Conclusions

After testing the obtained forecasting subsystem and the generated test samples, it is found that the results of forecasting OS security incidents obtained with various system settings and parameters may differ slightly. The use of a large number of polynomial models, such as those used in GMDH, allows one to obtain a much more accurate forecast for the task of forecasting OS incidents than classical regression methods, as well as the method of exponential smoothing.

References

1. Zaichenko, Iu. P. (2008). *Nechetkie modeli i metody v intellektualnykh sistemakh*. Kyiv: Izd Dom «Slovo», 344.
2. Bidiuk, P., Romanenko, V., Tymoshchuk, O. (2010). *Analiz chasovykh riadiv*. Kyiv: Politekhnika, 317.
3. Krause, A. (2009). *Evaluating the performance of adapting trading strategies with different memory lengths*. Available at: https://arxiv.org/abs/0901.0447
4. Geisser, S. (1993). *Predictive inference: an introduction*. Chapman & Hall, 282.
5. Billings, S. A., Hong, X. (1998). Dual-orthogonal radial basis function networks for nonlinear time series prediction. *Neural Networks, 11 (3),* 479–493. doi: http://doi.org/10.1016/s0893-6080(97)00132-9
6. Hizun, A., Volianska, V., Ryndiuk, V., Hnatiuk, S. (2013). Main parameters for information security intruder identification. *Ukrainian Information Security Journal, 15 (1),* 66–74. doi: http://doi.org/10.18372/2410-7840.15.4221
7. Sidorov, V. V. (2019). *Windows 10: kak prosmotret zhurnaly sobytii Windows?* Available at: http://netler.ru/ikt/windows10-events.htm
8. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, 758.
9. MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 640.
10. *Metod gruppovogo ucheta argumentov* (2019). MachineLearning. Available at: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%93%D0%A3%D0%90
11. Armstrong, J. S. (1999). *Forecasting for Marketing. Quantitative Methods in Marketing*. London: International Thompson Business Press, 92–119.
12. Jingfei Yang, M. S. (2006). *Power System Short-term Load Forecasting*. Darmstadt: Elektrotechnik und Informationstechnik der Technischen Universitat, 139.

Lakhno Valeriy, *Doctor of Technical Science, Professor, Department of Computer Systems and Networks, National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine, e-mail: lva964@gmail.com, ORCID: http://orcid.org/0000-0001-9695-4543*

Sagun Andriy, *PhD, Associate Professor, Department of Informatics, Information Security and Documentation, Cherkasy State Technological University, Ukraine, e-mail: avd29@ukr.net, ORCID: http://orcid.org/0000-0002-5151-9203*

Khaidurov Vladyslav, *PhD, Senior Researcher, Institute of Engineering Thermophysics of the National Academy of Sciences of Ukraine, Kyiv, Ukraine, e-mail: allif0111@gmail.com, ORCID: http://orcid.org/0000-0002-4805-8880*

Panasko Elena, *PhD, Associate Professor, Department of Informatics, Information Security and Documentation, Cherkasy State Technological University, Ukraine, e-mail: lena.pa@ukr.net, ORCID: http://orcid.org/0000-0002-0510-7742*