Forecasting the server status using the triple exponential smoothing model

M G Dubrovin¹, I N Gluhii¹, I Y Karyakin¹

¹Tyumen State University, 6, Volodarskogo Ave., Tyumen, 625003, Russia

E-mail: mikle1203@yandex.ru

Abstract. This article examines the issue of forecasting the time series of server systems and suggests applying the triple exponential smoothing model to solve this problem. It presents a mathematical formulation of the problem and describes the specifics of forecasting the time series of server systems. After this the article gives a comparative analysis of the autoregressive, neural network and exponential smoothing models in terms of their application to this problem. It argues that the triple exponential smoothing model (Holt-Winters method) offers a number of advantages when modelling the time series of server systems. It then provides experimental research to evaluate the accuracy of the Holt-Winters method with respect to the indicated time series. The research shows that the triple exponential smoothing model exhibits high results and can be applied to the solution of practical problems.

1. Introduction

The server is one of the most important elements of the corporate information system of an enterprise. A server is understood to mean a hardware-software complex that provides clients with access to certain resources or services [1]. The high cost of servers, as well as incidents of significant losses due to downtime caused by their malfunction, determine the relevance of the problem of ensuring the efficient functioning of corporate servers. Various automated monitoring tools are utilized to analyze the status and identify events in the functioning of servers.

Monitoring systems that receive information on the functioning of a server in real time are commonly used. This approach allows for the detection of many critical situations progressing in the system. A drawback of this approach is that the diagnosis and localization of errors is carried out after the detection of malfunctions, and as a result only problems that are already present in the system are identified. A more efficient approach is the introduction of proactive monitoring systems [2], which allow forecasting the occurrence of incidents in the functioning of server systems and generate appropriate warnings, making it possible to eliminate potential problems at the stage of their inception.

In this context an important task is to determine the forecasting model that would adequately describe the status of a server, detecting anomalies in its functioning with sufficient accuracy [3]. To forecast the status of a server system it is necessary to establish a set of information parameters and to collect statistical data on the parameters of the server, which is then presented in the form of time series.

In this paper we have examined the problem of forecasting the server status on the basis of time series, analysed commonly used models, and conducted experimental research on the applicability of the Holt-Winters’ triple exponential smoothing method (TES model) for solving this problem.
2. Materials and methods
Let \( S \) be the corporate server of an enterprise. Each server \( S \) can be described in terms of a list of parameters characterising its status: \( P = \{ x_1, x_2, ..., x_n \} \). Accordingly, the status of the server \( Z_t \) at a point in time \( t \) is described by the vector of the values of its parameters.

\[
Z_t = \begin{bmatrix} x_{t,1} \\ \vdots \\ x_{t,n} \end{bmatrix}
\]

The values of each parameter \( x_i \in P, i = 1, n \), known at discrete points in time \( t = 1, T \), represent the time series \( X_t = (x_{i,1}, x_{i,2}, ..., x_{i,T}) \). To solve the problem of forecasting the server’s status, for each parameter \( x_i \in P \) at the point in time \( T \) it is necessary to find the functional relationship between the values of the time series \( X_t \) and the forecasted future value \( \hat{x}_{i,T+m} \), where \( m \) is the interval of the forecast and \( p \) is the depth of immersion of the time series.

\[
\hat{x}_{i,T+m} = F(x_{1,T}, ..., x_{n,T-p})
\]

The forecasting model determines the functional relationship. Currently, there are dozens of various models and corresponding methods of forecasting time series [4,5]. When choosing a model it is necessary to take into account the specifics of the task at hand and the characteristics of the series being analyzed [6].

The following server system characteristics can be distinguished:

- The status of each server is described by a number of parameters. To obtain complete information on the status of a server, it may be necessary to record anywhere from a dozen to several hundred parameters.
- For the prompt detection of an inoperative status, short-term forecasts of the status of a server (minutes or hours sequence) are of interest. Long-term forecasts (weeks or months sequence) are of less interest, as it can be expected that the time series will develop in a dynamic environment.
- The status of each server changes over time due to shifts in the functional load of the server, scheduled activities or other external circumstances, which allows us to conclude that the processes under consideration are non-stationary.

Let us consider several models that have found practical application in forecasting the time series of servers. The use of a modification of the autoregressive model, which takes into account the seasonal component of the series (SARIMA), has shown good results in forecasting the loads of server systems [7]. According to the Box-Jenkins methodology [8], if the original series is non-stationary, it must be brought to a stationary form for this model to be used. If an enterprise has a distributed server infrastructure, each server of which is described by a number of non-stationary series, bringing them all to a stationary form can be a labor-intensive process.

Various neural network models are quite commonly used for forecasting computer systems [9] and allow analyzing non-linear patterns in data sets. On the other hand, the use of neural networks also requires additional data pre-processing, which includes the normalization of the values of the series and bringing them to stationary form. Moreover, neural networks operate on the black box principle, which complicates the interpretation of the model.

The TES model belongs to the class of exponential smoothing models. It is used in various fields: in IT-systems in particular there are examples of its use to forecast and detect anomalies in network traffic [10,11]. The multiplicative model is described by the following system of equations [12]:

\[
R_t = a * \frac{x_t}{S_{t-L}} + 1 - a * R_{t-1} + T_{t-1}
\]

\[
T_t = \beta * (R_t - R_{t-1}) + 1 - \beta * T_{t-1}
\]
\[ S_t = \gamma \frac{x_t}{S_t} + 1 - \gamma \times S_{t-L} \]

\[ \bar{x}_{t+m} = S_t + m \times T_t \times S_{t-L+m} \]

where \( R \) is the exponentially smoothed series; \( T_t \) is the value of the trend; \( S_t \) is the seasonal component of the series; \( \bar{x}_{t+m} \) is the forecast for \( m \) steps; \( \alpha, \beta, \gamma \) are the coefficients of the smoothing of the series, trend and seasonal component; \( L \) is the seasonal period.

The TES model takes into account the trend – the seasonal dynamic of the series, by which it arguably allows for creating accurate forecasts for the non-stationary time series of a server. Working with the model does not require the labor-intensive procedures of preliminary data processing, which is an advantage when analysing many server parameters. The model belongs to the class of statistical models [12], and so the forecasting results are not difficult to interpret, which can be an advantage within the framework of the tasks of our research.

The TES model belongs to the class of parametric models, for its implementation it is necessary to determine the seasonal period \( L \) and to train the model by selecting the optimal values of the smoothing parameters \( \alpha, \beta, \gamma \). The training of a model consists of going through possible combinations of parameters with the aim of minimizing the loss function. The mean squared error (MSE) [13] can be used as the loss function.

\[ \text{MSE} = \frac{1}{N} \sum_{t=1}^{N} (x_t - \bar{x}_t)^2 \]

where \( x_t \) is the actual value of the series, \( \bar{x}_t \) is the forecasted value, \( N \) is the number of observations.

Optimization is performed by evaluating the loss function for given parameters during cross-validation. The standard cross-validation process is not applicable to time series, as during the random mixing of values the time structure of the series is lost [14]. Therefore, sliding window cross-validation (Figure1) is used for time series.

![Figure 1. Cross-validation of a time series.](image)

Taking into account the above-mentioned advantages and characteristics of the TES model, we have chosen it for further experimental research with the aim of determining its applicability to solving the problem of forecasting the time series of server systems.

3. Results and Discussion
The data necessary for the conduction of experimental research was prepared. It was presented in the form of time series. Over the duration of a week the values of 4 parameters were recorded from two functioning servers of a manufacturing enterprise with an increment of 10 minutes:

1. CPU load.
2. Memory usage.
3. Input network traffic.
4. Outgoing network traffic.

All of the operations with the time series and their graphs were carried out in the specialized software called «STATISTICA» [15]. The accuracy of the triple exponential smoothing model was evaluated with the help of the MAPE [13] metric.

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|x_i - \bar{x}|}{x_i} \times 100
\]

Let us consider the graph of the time series for the first server’s central processor load (Figure 2). A visual analysis shows that the graph is quite full of noise, and so it is difficult to visually assess the presence of the seasonal component. The graph of the autocorrelation function for the first two days demonstrates the presence of a season of 24-hour duration, comprising of 144 observations (Figure 3).

![Figure 2. Server 1 – CPU load graph.](image1)

![Figure 3. Server 1 – Autocorrelation function graph.](image2)

Let us now consider the graph of the time series for changes in memory usage (Figure 4). With a linear approximation of the series, a weeklong downwards trend can be observed. Figure 5 shows the day-ahead forecast of the first server’s CPU load, created using the triple exponential smoothing model. The dark line on the graph denotes the time series of statistical data. The light line denotes the
time series of forecast values. The graph shows that the structure of the time series is taken into account when the forecast values are generated.

For each of the initial series, a multiplicative Holt-Winters model with a linear trend was built. The lag of the seasonal component \( L \) is selected as 144 observations. The selection of the smoothing coefficients was carried out with the help of the inbuilt automatic search algorithm of STATISTICA.

The results of the experiment are presented in Table 1. The average MAPE score is 8.7%. It is generally agreed that a model shows high accuracy when the MAPE score is less than 10% and good accuracy when the MAPE score is from 10% to 20% [16].

| No | Server   | Parameter        | \( L \) | \( \alpha \) | \( \beta \) | \( \gamma \) | MAPE  |
|----|----------|------------------|--------|-----------|---------|--------|------|
| 1  | CPU load | 144              | 1      | 0         | 0       | 0      | 14.8 |
| 2  | Memory usage | 144 | 0.9       | 0         | 0       | 2.3    |
| 3  | Network traffic in | 144 | 0.5       | 0.3      | 0       | 3.7    |
| 4  | Network traffic out | 144 | 0.4       | 0.1      | 0       | 11.3   |
| 5  | CPU load | 144              | 0.5    | 0         | 0       | 13.7   |
| 6  | Memory usage | 144 | 0.9       | 0.1      | 0       | 6.4    |
| 7  | Network traffic in | 144 | 0.2       | 0.1      | 0       | 4.4    |
| 8  | Network traffic out | 144 | 0.3       | 0.1      | 0       | 12.7   |

Thus it can be concluded that the TES model has exhibited a sufficiently high accuracy in forecasting server parameters. For comparison, an experimental analysis of the SARIMAX model’s ability to forecast the hourly load of two servers has shown accuracy scores of 11.46% and 11.67% respectively [7]. In the course of the experiment it had also been found that the model exhibits less accurate results for series with a high level of noise.

4. Conclusion

The article explores the problem of forecasting the time series of server systems and the possibility of utilizing the triple exponential smoothing model to solve this problem. A mathematical formulation of the problem was presented and its characteristics were described. A comparative analysis of certain forecasting models in terms of their application to this problem was carried out. It was argued that the triple exponential smoothing model possesses several advantages when modelling server time series. Experimental research to evaluate the efficiency of using the model to forecast server system parameters was conducted. The MAPE metric scores that were obtained range from 2.3% to 14.8%, the average MAPE score is 8.7%. The conclusion was made that the TES model shows sufficiently high results and can be used to solve such practical problems as detecting anomalies and forecasting incidents in the functioning of servers.

References
[1] Oluwatosin H 2014 Client-server model IOSRJ Comput. Eng. 16(1) 2278-8727
[2] Simonova S and Sprync O 2011 Proactive IT/IS monitoring for business continuity planning E+ M Ekonomie a Management 14(3) 57-65
[3] Jinka P and Schwartz B 2015 Anomaly Detection for Monitoring (Sebastopol: O'Reilly Media, Inc.) pp 15-31
[4] Mahalakshmi G, Sridevi S and Rajaram S 2016 A survey on forecasting of time series data International Conference on Computing Technologies and Intelligent Data Engineering – IEEE pp 1-8
[5] Ahmed N et al 2010 An empirical comparison of machine learning models for time series forecasting Econometric Reviews 29(5-6) 594-621
[6] Chambers J, Mullick S and Smith D 1971 How to choose the right forecasting technique (Brighton: Harvard Business Review) p 2

[7] Debusschere V et al 2012 Hourly server workload forecasting up to 168 hours ahead using seasonal ARIMA model IEEE international conference on industrial technology pp 1127-1131.

[8] Makridakis S and Hibon M 1997 ARMA models and the Box–Jenkins methodology Journal of Forecasting. 16(3) 147-163

[9] Debusschere V et al 2013 Neural networks for web server workload forecasting IEEE International Conference on Industrial Technology pp 1152-1156

[10] Brutlag J 2000 Aberrant Behavior Detection in Time Series for Network Monitoring LISA 14 139-146.

[11] Jasek R Szmit A and Szmit M 2013 Usage of modern exponential-smoothing models in network traffic modelling Nostradamus 2013: Prediction, Modeling and Analysis of Complex Systems. – Springer, Heidelberg 210 435-444

[12] Kalekar P 2004 Time series forecasting using holt-winters exponential smoothing Kanwal Rekhi School of Information Technology. 4329008(13)

[13] Shcherbakov Met al 2013 A survey of forecast error measures World Applied Sciences Journal 24 171-176

[14] Hyndman R and Athanasopoulos G 2018 Forecasting: principles and practice (Melbourne, Australia: OTexts) pp 68-69

[15] Hilbe J 2007 STATISTICA 7: an overview The American Statistician 61(1) 91-94

[16] Moreno J et al 2013 Using the R-MAPE index as a resistant measure of forecast accuracy Psicothema 25(4) 500-506