Transport and Telecommunication, 2011, Volume 12, No 2, 20–27
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

SHORT-TERM TRAFFIC FORECASTING WITH NEURAL NETWORKS

Irina Klevecka

Riga Technical University
Faculty of Electronics and Telecommunications
Riga, Latvia
Ph.: +371 26006970, e-mail: klevecka@inbox.lv

The paper is dedicated to an advanced algorithm aimed at short-term forecasting of electronic communications traffic by applying non-linear neural networks. The algorithm incorporates three procedures, which allow automating the process of determining an optimal solution with minimal involvement and influence of expert assessment and human factors.

Keywords: telecommunications, electronic communication networks, traffic, neural networks, multilayer perceptron, algorithm, forecasting

1. Introduction

Traffic forecasting plays an important role in design, management and optimization of modern telecommunications systems. Accurate and reliable forecasts allow planning the capacity of a network on time and sustaining the required level of quality of service. Besides, the properties of network traffic directly influence both capital costs of equipment and expected income of an operator.

The emergence of the packet-switched Internet networks as well as transformation of traditional telephone networks into multi-service systems provides new opportunities to a user in the sphere of his/her activities. It has changed not only the architecture of a network but also statistical nature of teletraffic, which is now characterized by the effects of self-similarity and strong long-range dependence [1, p. 150; 2]. Therefore, new approaches to the analysis and forecasting of states and parameters of the packet-switched networks are strongly required. A *non-linear artificial neural network* is one of these methods, which is rapidly gaining recognition in this field.

An *artificial neural network* is a massively parallel-distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making it available for use [3, p.2]. Neural networks have been developed as a generalization of the mathematical models of human cognition and neural biology. A special type of neural networks, *time-lagged feed-forward networks* (Fig. 1), is usually applied for the purposes of time series modelling and forecasting. A time-lagged feed-forward network is a powerful nonlinear filter consisting of a tapped delay memory of order Ω and a multilayer perceptron. The standard back propagation algorithm can be used to train this type of neural networks. More details on the operation of multilayer perceptrons and time-delayed neural networks can be found in [3, 4].

The main advantage of the neural networks in comparison with classic linear methods is their ability to model functions characterized by non-linear dynamics. High adaptive ability, tolerance to various outer noises as well as the influence of "heavy-tailed" distributions are the other frequently mentioned advantages of neural networks.

However, we can often hear that *neural networks are more art than science*. This is primarily due to the lack of a functional algorithm for applying the neural networks to time series forecasting. Because of that, the active expert assessment is still necessary at all the stages of implementation, which prevents the automation of a forecasting process. It is also important to understand that, in contrast to some classic linear models, the method of neural networks has not been initially aimed at time series modelling and forecasting. The attempts to predict traffic loads of both a conventional telephone network and a packet-switched Internet network by means of neural networks have been made many times in the past. However, most of these research papers solve a trivial task of time series approximation, with more or less success, without taking into account the general theory of neural networks and time series forecasting.

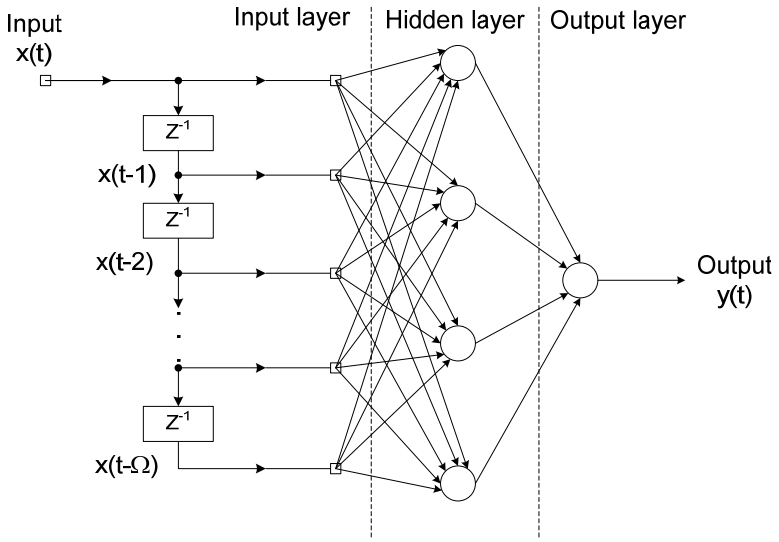


Figure 1. Time lagged feed-forward network [3, p.644; 4]

2. Brief Description of the Algorithm for Short-Term Traffic Forecasting

In order to facilitate and automate the process of time series modelling and forecasting, and compensate for the problems associated with instability of a produced solution, an advanced algorithm has been developed, the core diagram of which in shown in Fig. 2. The algorithm is especially aimed at producing the operative (24 hours ahead) and short-term (up to two weeks ahead) forecasts of network traffic represented as univariate / multivariate time series.

A forecasting task is solved under the assumption that optimal training parameters and an appropriate training algorithm are already selected according to some rules or procedures, and stay invariable during the whole training process.

Such parameters of network architecture as the number of input and output neurons are also constant during all the stages of training and forecasting. Moreover, the number of hidden layers of a regression network does not usually exceed one in accordance with the universal approximation theorem [5, 6]. Thus, the optimal value of only one variable parameter, the number of hidden neurons, has to be determined through test-and-trial routine.

Only few hidden neurons are often satisfactory for the purposes of traffic modeling and forecasting. Therefore, in order to identify the optimal number of hidden neurons, it is advised to apply a constructive method, which is implemented in the proposed algorithm. It involves a gradual increase of the number of hidden neurons with subsequent testing and verification of each network architecture.

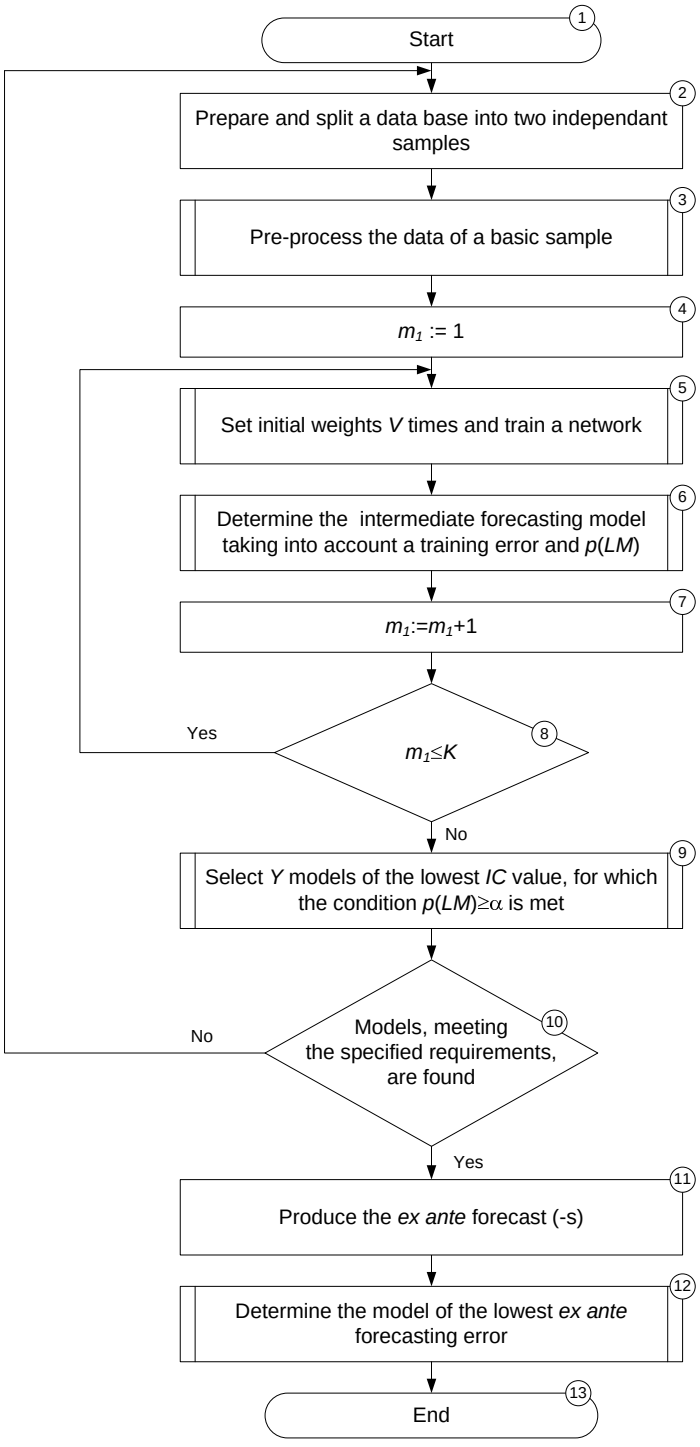
The important stage of the algorithm is pre-processing of input data. Some procedures of pre-processing are discussed in [7] and include reducing a time series to stationary behaviour, identifying seasonal and / or cyclic components, etc.

3. Novelty of the Algorithm

In contrast to some classic algorithms for accomplishing a pre-defined task (see, for example, [8; **Error! Reference source not found.**, p.84]), the proposed algorithm incorporates three procedures:

- implementation of multiple cycles of weight initialization of a neural network during a training process;
- method of selecting the intermediate forecasting models, taking into account the level of residual autocorrelation and the estimates of information criteria;
- method of selecting the final forecasting model, taking into account the accuracy of *ex ante* forecasts.

Let us consider the theoretical arguments in defence of the necessity of implementing these procedures to identify a relevant forecasting model.



Legend

m_1 – number of hidden neurons of a two-layer perceptron
K – maximum number of hidden neurons
V – total number of weight initialization cycles
IC – information criterion
LM – Lagrange multiplier type test
p(LM) – p-value (observed significance) of the LM-type test
α – significance level

Figure 2. Advanced algorithm for solving a traffic forecasting task by means of neural networks

3.1. Implementation of Multiple Cycles of Weight Initialization of a Neural Network

It is required to set some initial values to all the weights and biases of a neural network before a training process starts. The aim of initialization is, probably, to find the best approximation to an optimal solution and, in this way, to decrease training time and facilitate the convergence of a training algorithm.

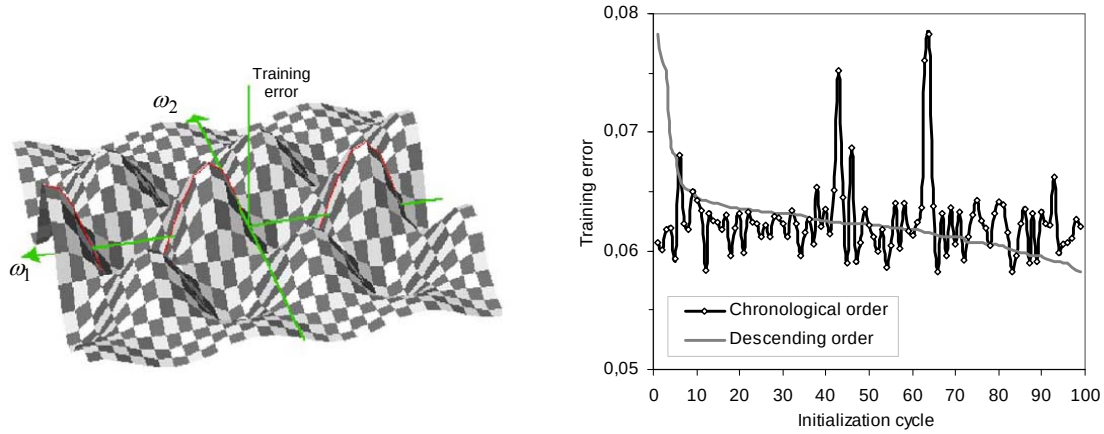


Figure 3. Illustration of the necessity of implementing multiple initialization cycles: (a) a simplified example of two-dimensional error surface, which demonstrates a key problem – several local minima can exist on the training error surface; (b) the training errors of a neural network (without applying cross-validation) produced as a result of a hundred cycles of weight initialization[1]

If the initial weights and biases are set to large values, then neurons usually approach the saturation level very quickly. In this case, the local gradients, calculated according to the back-propagation algorithm, take small values, which, in turn, would significantly increase training time. If the initial values are set to small values, then the algorithm works very slowly about the origin of the error surface. This is specifically true in the case of anti-symmetric activation functions such as hyperbolic tangent.

During two last decades many heuristic methods of weight initialization have been proposed (see [10] for more details). Despite this, the optimal solution of this issue has not still been found. Due to its simplicity, the most common method is the random initialization of weights and biases from a uniformly or normally distributed narrow range of small values.

Regardless of the initialization method, starting values of weight coefficients influence the final result of training. This is due to the properties of a training algorithm as well as the fact that several local minima do exist on the error surface (see Fig. 3-a). Therefore, in order to find an actual global minimum, it is necessary to train one and the same neural network multiple times under the same conditions, changing only the initial values of weights and biases.

In spite of these problems, most researchers still do not pay a proper attention to this aspect. It is possible to find references to 5 [11], 10 [12], 25 [13], 50 [14; 15] and 100 [13] cycles of initialization. However, the most practical studies restrict the number of initialization cycles to one, and this can mislead a researcher regarding the adequacy of a produced solution.

The example shown in Fig. 3-b illustrates the uncertainty in producing a final solution and its dependence on the starting values of weights and biases. The training errors shown here are produced as a result of training a neural network of the same architecture and applying the same training parameters but changing the initial values of weights and biases. It is easy to notice that the difference between the highest and lowest errors comprises more than 25 per cent, which can significantly influence the accuracy of final forecasts.

The number of initialization cycles is usually chosen mandatory, and depends on the complexity of a task as well as on time resources a researcher has on his / her disposal.

[1] The values of training errors are displayed in chronological order of their evaluation at the end of each training epoch as well as sorted in descending order to facilitate their comparison.

3.2. Selection of the Intermediate Forecasting Models Taking into Account the Residual Autocorrelation and Estimates of Information Criteria

According to the developed algorithm, the selection of the intermediate forecasting models among the trained networks is carried out taking into account the level of r*esidual autocorrelation* and the value of *information criterion*.

Some standard parameters, such as a correlation coefficient (R), mean squared error (MSE), mean absolute error (MAE) are traditionally used to evaluate a general forecasting ability of forecasting models. However, these parameters provide little information about the accuracy of a fitted model, and are also useless in identifying statically significant differences between the forecasts produced by various methods [16, 17, 18].

The most accurate indicator of the adequacy of a forecasting model can be the absence of *autocorrelation in residuals*. The residuals of a fitted model are defined as n differences given by $\varepsilon_t = x_t - \hat{x}_t, t = 1, 2, \ldots, n$, where x_t is the observed value and $\hat{x}_t$ is a corresponding predicted value produced by means of a fitted statistical model [**Error! Reference source not found.**, p. 94]. These differences cannot be explained by a produced forecasting model. Therefore, we can consider residuals ε_t to be observed errors.

The acceptance of the hypothesis of no autocorrelation in residuals at a pre-defined significance level means that the residuals are similar to *white noise* and further analysis will not discover any statistically significant dependencies. In classic regression analysis, the *Durbin-Watson* criterion [**Error! Reference source not found.**, p.245] is traditionally used for testing the autocorrelation of residuals. However, this test is not suitable for testing the autocorrelation if the regressor is a lagged explanatory variable [**Error! Reference source not found.**, p.256]. For the same reason, the *Box-Pierce* and *Ljung-Box* [21] criteria cannot be applied to neural networks, although the last one is widely used and, despite its theoretical inconsistence, is still included in most statistical and econometric software packages.

At present, the most relevant estimate of the residual autocorrelation of neural networks is a powerful *Lagrange Multiplier* (*LM*) *type test* [22]. The LM-type test belongs to classic asymptotic tests and is capable of identifying the autocorrelation of any order.

In turn, the use of *information criteria* is based on one of the main idea of time series forecasting – "choice a parsimonious model" (known as the *Ockham's razor principle*). It means that, all other things being equal, one should prefer the model with the fewest free parameters.

The mean squared residuals usually decrease once the model becomes more "complicated" with addition of new free parameters. Increasing the number of free parameters of a neural network (which is associated with the addition of new neurons / layers of neurons), one can fit a model to historical data with infinite accuracy. The *universal approximation theorem* [5, 6] explains this property of neural networks. However, such a neural network may have a poor ability to make generalizations due to *over-training* [3, p.206]. Besides, once a certain limit is reached, the gain in accuracy of fitting with addition of new parameters tends to be insignificant. On the other hand, time required for selecting the optimal values of free parameters can lead to a sharp decrease in the performance of a network. Therefore, it is very important to seek for the balance between the preciseness of approximation and the complexity of a statistical model.

Information criteria have been found to be quite useful in solving this problem. The estimate of the criterion consists from the penalty for poor fitting and the penalty for over-parameterization. The most popular criteria of this type are the *Akaike's information criterion* (AIC) and the *Bayesian information criterion* (BIC) given by [23, p.38; 24, p.373]:

$$AIC(l) = N_{ef} \ln \hat{\sigma}_\varepsilon^2 + 2l, \tag{1}$$

$$BIC(l) = N_{ef} \ln \hat{\sigma}_\varepsilon^2 + 2l + l \ln(N_{ef}), \tag{2}$$

where

N_{ef} – number of effective observations, to which the model is fitted;

l – number of adjusted parameters;

$\hat{\sigma}_\varepsilon^2$ – estimate of the residual variance, $\hat{\sigma}_\varepsilon^2 = {}^{1}\!/_{N_{ef}} \cdot \sum_{t=1}^{Nef} \varepsilon_t^2$.

Information criteria are evaluated separately for each analyzed specification (architecture) of neural networks. The models that possess the lowest value of the criterion should be selected for further analysis. It has been also noticed that, in practice, the BIC "selects" very parsimonious models with only few parameters. Therefore, this criterion is often used for non-linear models, where insignificant gain in fitting quality hinges upon the necessity of calculating many additional parameters.

3.3. Selection of the Final Forecasting Model Taking into Account the Accuracy of *Ex Ante F*orecasts

If the models, meeting the above specified conditions, are found, it is required to test their generalization ability (i.e., the ability to produce reliable forecasts) for an independent test set, which is not involved in training. The forecast developed for an independent test set we will call an *ex ante forecast*. The necessity of *ex ante* forecasting is related to the fact that even if a neural network provides a high accuracy of approximation and uncorrelated residuals, it would be still over-trained on historical data.

The approach of splitting a time series into two independent subsets has gained wide acceptance in the practical studies dedicated to time series forecasting (see, e.g., [25; 26]) but it is still rarely used in the case of neural networks. The first, largest data subset, called the *basic* or *retrospective sample*, is used to select and verify a statistical model. The second, *ex ante forecasting sample* is used to examine the quality of *ex ante* forecasts, comparing them against historical data. The latter provides the opportunity to evaluate independently the forecasting abilities of the model fitted to the basic sample.

The last historical values of an analyzed time series are traditionally used for developing the *ex ante* forecasting sample. However, it is necessary to keep in mind, that these observations influence the direction of the actual *real-life* forecast much more than the earlier ones. Therefore, the last historical data are the most valuable for the process of selecting an appropriate forecasting model, and using them as a testing sample is not always reasonable.

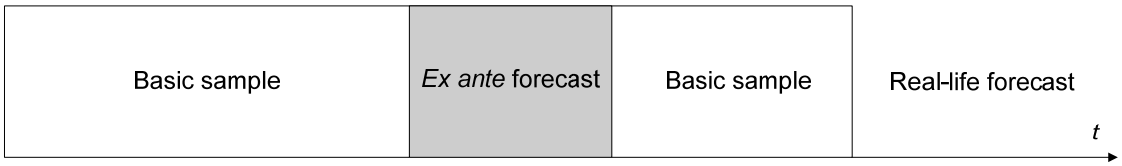


Figure 4. Chronological division of a time series into the basic and ex ante forecasting samples

According to the proposed algorithm, it is recommended to divide a time series into the basic and *ex ante* forecasting samples in the way shown in Fig. 4. This original approach allows increasing the quality of real-life forecasts, since the last historical observations are used for fitting a statistical model rather than testing.

The accuracy of *ex ante* forecasts is evaluated by one of the standard error parameters. In the practical studies of [10], the mean absolute percentage error (*MAPE*) was applied, given by [27, p. 347]:

$$MAPE = \frac{\sum_{t=1}^{L} \left| \frac{x_t - \hat{x}_t}{x_t} \right|}{L} \cdot 100\%, \tag{3}$$

where L – the size of an *ex ante forecasting* sample (i.e., forecasting horizon).

The *MAPE* is a relative, dimensionless measure of the accuracy of approximation or forecasting. It is helpful in comparing forecast performance across different data sets, or comparing the performance of different forecasting methods.

The lower the *MAPE*, the more accurate is a forecast. The model of the lowest *MAPE* is the final model assigned to further *real-life* forecasting. The interpretation of *MAPE* introduced in [28] allows judging about the accuracy of a forecast: less than 10 per cent is a highly accurate forecast, 11 to 20 per cent is a good forecast, 21 to 50 per cent is a reasonable forecast, and 51 per cent or more is an inaccurate forecast.

Thus, the choice of a final forecasting model is based on the results of multiple consecutive procedures and tests. The final model is characterized by the lowest value of the information criterion, uncorrelated residuals and the lowest error of an *ex ante* forecast.

4. Conclusion

The use of neural networks traditionally involves expert assessment at all stages of application. The proposed algorithm allows minimizing the influence of human factor and helps identifying the neural network solutions resulting in reliable forecasts with a minimum level of errors.

The effectiveness of the developed algorithm was examined on real data sets represented as ten time series of different lengths and averaging degrees, which characterized:

- the intensity of total carried traffic of a conventional telephone network;
- the transmission rate of outgoing international traffic measured at a transport level of a packet-switched Internet network.

The results of this practical study are summarized in [10]. It was shown that the MAPE estimates of examined *ex ante* forecasts varied from 10 per cent to 30 per cent. This confirms the possibility of applying these models in real-life conditions.

The criteria for selecting a final forecasting model proposed in the thesis – the lowest estimate of the information criterion and statistically insignificant residual autocorrelation – can be successfully applied to linear statistical methods as well.

References

1. Janevski, T. *Traffic Analysis and Design of Wireless IP Networks*. ARTECH HOUSE, INC., 2003.
2. Leland, W.E., M.S. Taqqu, W. Willinger, and D.V. Wilson. "On the Self-Similar Nature of Ethernet Traffic (Extended Version)." *IEEE/ACM Transactions on Networking* 2.1 (February 1994), pp. 1-15.
3. Haykin, S. *Neural Networks: A Comprehensive Foundation*. 2nd / Ed. Prentice Hall, 1999.
4. Mozer, M. C. "Neural Network Architectures for Temporal Pattern Processing." *Time Series Prediction: Forecasting the Future and Understanding the Past* Eds. A.S. Weigend and N. A. Gershenfeld. Perseus Book Publishing, 1993, pp. 243-264.
5. Hornik, K. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2 (1989), pp. 359-366.
6. Cybenko, G. Approximation by Superpositions of Sigmoidal Function. *Mathematics of Control, Signals, and Systems,* 2(1989), pp. 303-314.
7. Klevecka, I., and J. Lelis. Pre-Processing of Input Data of Neural Networks: The Case of Forecasting Telecommunication Network Traffic. *Telektronikk* 3/4 (2008), pp. 168-178.
8. Fildes, R. and K.-P. Liao. The Accuracy of Procedural Approach to Specifying Feedforward Neural Networks for Forecasting. *Computers & Operations Research* 32.8 (August 2005), pp. 2151-2169.
9. Komartseva, L., Maksimov, A. *Neurocomputers: Manual for higher schools*. Moscow: MGTU Publishing House named after N.E. Bauman, 2002. 320 p. (in Russian)
10. Klevecka, I. *Neural Networks for Short-Term Traffic Forecasting*. Promotion Thesis. Riga, Riga Technical University, 2011. [National Library of Latvia].
11. Kaastra, I., Boyd, M. Designing a neural network for forecasting financial and economic time series. *Neurocomputing* 10 (1996), pp. 215-236.
12. Tang, Z. and Fishwick, P.A. Feedforward Neural Nets as Models for Time Series Forecasting. *ORSA Journal on Computing* 5.4 (Fall 1993), pp. 374-385.
13. Kolen, P., and J.B. Pollack. Back Propagation in Sensitive to Initial Conditions. In: *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, Denver, Colorado, United States, 1990, pp. 860-867.
14. Faraway, J., Chatfield, C. Time Series Forecasting with Neural Networks: A Comparative Study Using the Airline Data. *Applied Statistics* 47.2 (1998), pp. 231-250.
15. Thimm, G., and E. Fiesler. *Optimal Setting of Weights, Learning Rate, and Gain*. IDIAP Research Report 97-04. Switzerland: IDIAP, 1997.
16. Clements, M.P., Hendry, D.F. "On the Limitations of Comparing Mean Square Forecast Errors". *Journal of Forecasting* 12 (1993), pp. 617-637.
17. Diebold, F.C. and Mariano, R.S. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics* 13.3 (July 1995), pp. 253-263.

18. Kunst, R.M. *Testing for Relative Predictive Accuracy: A Critical Viewpoint*. Economics Series 130/ Vienna: Institute for Advanced Studies (IHS), 2003.
19. Dreiper, N., Smith, G. *Applied Regression Analysis*, 3rd edition: Transl. from Enligsh. Moscow: Publishing House "Williams", 2007. 912 p. (in Russian)
20. Kantorovich, G.G. Time Series Analysis. *Economic Journal of HSE*. №1 2002, p.85-116. (in Russian)
21. Ljung, G.M., and G.E. Box. On a Measure of Lack of Fit in Time Series Models. *Biometrika* 65 (1978), pp. 297-303.
22. Medeiros, M.C., Teräsvirta, T. and Rech, G. Building Neural Network Models for Time Series: A Statistical Approach. *Journal of Forecasting* 25(2006), pp. 49-75.
23. Franses, H.P. and D. van Dijk. *Nonlinear Time Series Models in Empirical Finance*. Cambridge University Press, 2000.
24. Priestley, M.B. *Spectral Analysis and Time Series*. London: Academic Press, 1981.
25. Makridakis, S. and R.L. Winkler. Sampling Distributions of Post-sample Forecasting Errors. *Applied Statistics* 38.2 (1989), pp. 331-342.
26. Tashman, L.J. Out-of-sample Tests of Forecasting Accuracy: An Analysis and Review. *International Journal of Forecasting* 16.4 (October 2000), pp. 437-450.
27. Armstrong, J.S. *Long-Range Forecasting*. New York: John Willey & Sons, 1985.
28. Lewis, C.D. *Industrial and business forecasting methods: a practical guide to exponential smoothing and curve fitting*. London: Butterworth Scientific, 1982.