FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 21, Nº 2, 2022, pp. 59 - 75
https://doi.org/10.22190/FUACR220330006R

Regular Paper

COMBINED TECHNIQUES FOR FORECASTING THE VOLUME OF PACKAGES IN INTERNAL POSTAL TRAFFIC OF SERBIA

UDC ((621.391+004.032.26):656.8)

Ivana Rogan, Olivera Pronić-Rančić

University of Niš, Faculty of Electronic Engineering,
Department of Telecommunications, Niš, Republic of Serbia

Abstract. *The main goal of time series analysis is to explain the main features of the data in a chronological order and in the general case to predict future processes, products, service requirements, etc., using appropriate statistical models. In this paper, time series prediction was performed using a seasonal autoregressive integrated moving average model (SARIMA) in the XLSTAT add-in for Excel environment, as well as two artificial neural network (ANN) models - long short-term memory (LSTM) network and relatively new machine learning technique - extreme learning machines (ELM). The proposed approaches were used for forecasting the volume of packages in the internal postal traffic of Serbia for the period 2014-2020. A comparison of the obtained modeling results with the original data was made and it was shown that the best modelling results were achieved by using ELM.*

Key words: *Time series analysis, forecasting, ANN, SARIMA, LSTM, ELM*

1. INTRODUCTION

A time series is a sequence of observations of a random, here, real, non-negative, variables. Time series analysis provides tools, mostly of the essential, statistical and/or analytical-approximate type, to select a model that can be used to anticipate future events. In this context, predicting future values is an appropriate mathematical method for extrapolating future data, depending on external influences and chronologically arranged numerical information, [1-7].

In cases of time series data prediction, several different techniques have been applied, [1-22]. Theoretical, general, and special techniques and softwares for numerical prediction

Received March 30, 2022 / Accepted June 16, 2022
Corresponding author: Ivana Rogan
University of Niš, Faculty of Electronic Engineering, Department of Telecommunications,
Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
E-mail: ivana84p@gmail.com

were presented in [1-7]. For example, [4] provides a modern overview of a broad range of methods, principles, and theoretical approaches to prepare, produce, organize, and evaluate forecasts. Concrete hybrid models, deterministic, (S)ARIMA with Artificial Neural Networks (ANN) were presented in [8] and, in [9], integration of EWT (Empirical Wavelet Transform), ARIMA with the improved ABC Optimized ELM was proposed for financial time series forecasting. In [11], [12] MATLAB machine learning (ML) and deep learning models (among others, LSTM (Long Short-Term Memory) and ELM (Extreme Learning Machines)) were considered. The directions of development of LSTM methods were observed in [13-17]. The properties and variations of ELM technique were shown in [18-24].

Forecasting the revenue and volume of some postal services is presented in [27-29]. Savitzky-Golay filter modification for forecasting the volume of postal services was presented in [28]. In [29] time series analysis techniques based on the SARIMA model, as well as the LSTM model, for predicting the volume of received express mail services in international traffic in the Republic of Serbia were developed. A bias correction to the minimum Akaike information criterion, AIC, is derived for regression and autoregressive time series models in [30].

In this paper, we will consider different techniques for forecasting the volume of packages in the internal postal traffic in Serbia. Parcel Services of the Post of Serbia are used to transfer goods and other items. A parcel is a closed postal item containing goods and other items, with or without indicated value, with registered reception number, [31]. The aim of the paper is to introduce new ANN methods, due to expected better results, into the methodology of forecasting the volume of packages in the Serbian internal postal traffic and thus contribute to more adequate or automated decision-making in relation to them. Time series prediction will be performed using a seasonal autoregressive integrated moving average model (SARIMA) in the XLSTAT add-in for Excel environment, as well as two ANN models - long short-term memory (LSTM) network and extreme learning machines (ELM).

The paper is organized as follows. After Introduction, a brief description of used forecasting models is given in Section 2. The most illustrative numerical results are presented in Section 3, and finally conclusion remarks are given in Section 4.

2. FORECASTING MODELS

In this paper, we considered several techniques for time series data forecasting: SARIMA, LSTM and ELM.

a) SARIMA

ARIMA [5] is basically a linear model assuming that time series data is stationary. Therefore, there is a limited ability to capture nonlinearities and non-stationarities in the data. ARIMA models effectively consider the serial linear correlation among observations, whereas Seasonal AutoRegressive Integrated Moving Average (SARIMA) models can satisfactorily describe time series that exhibits simple periodic non-stationarity both within and across seasons, [5]. The SARIMA approach to modeling was introduced as a statistical method of choice in the case of data derived from observations collected over a sufficiently long period of time.

ARIMA(p,d,q)(P,D,Q)$_s$ or SARIMA models [5], are usually used when the time series has short-term correlations, trend or seasonality. The time series function $y_t=y(t)$ is defined as:

$$y_t : \{1,2,...,n\} \to R^+ \cup \{0\}. \tag{1}$$

An observable time series y_t in which successive values are highly dependent can frequently be regarded as generated from a series of independent variations a_t - random deviations from normal distribution, having mean equals zero and variance σ_a^2.

For time shift operator, B, ($B^s y_t = y_{t-s}$, $s = 0,1,2,...$), a differential linear operator, ∇, is defined as follows: $\nabla y_t = (1-B) y_t$.

The basic relation describing the SARIMA model is [5]:

$$\varphi_p(B)\delta_P(B^s)\nabla^d(1-B^s)^D y_t = \mu + \chi_Q(B^s)\theta_q(B)a_t., \tag{2}$$

where

$$\varphi_p(B) = 1-\varphi_1 B-...-\varphi_p B^p, \tag{3}$$

$$\theta_q(B) = 1-\theta_1 B-...-\theta_q B^q, \tag{4}$$

$$\delta_P(B^s) = 1-\delta_1 B^s-...-\delta_P B^{sP}, \tag{5}$$

$$\chi_Q(B) = 1-\chi_1 B^s-...-\chi_Q B^{sQ}. \tag{6}$$

In equations (2) – (6), constants d and D indicate the degrees of non-seasonal and seasonal differences, s represents the seasonal time shift, and μ is the trend component. In previous equations, $\delta_1,...,\delta_P$ and $\phi_1,...,\phi_p$ are seasonal and non-seasonal autoregressive constant parameters (p and P are adequate constant autoregressive parameters); $\chi_1,...,\chi_Q$ and $\theta_1,...,\theta q$ - represent seasonal and non-seasonal constant parameters of the moving averages (q and Q are adequate constant moving averages parameters), respectively. AR(p, P) - autoregressive part (p-seasonal, P-non-seasonal indexes) refers to relationship between the data variable y_t with its own lagged values. Parameter values p and P are derived from PACF (partial autocorrelation function) plots. Integrated part I(d, D) refers to order of differencing and it is essential when the series is non-stationary. In ARIMA model, Moving Average order MA(q, Q) indicates the dependence of present value of the time series variable on the lagged error terms. The order of MA part can be inferred from the Auto-Correlation Function (ACF) plot.

The (S)ARIMA model is usually treated by a three-stage iterative procedure based on identification, assessment-estimation, and diagnostic verification [4-6]. Identification generally means using all information about how batch-data is generated; assessment means the efficient use of data in order to determine the conclusions about the model parameters by the adequacy of the chosen model; and diagnostic verification means the verification of this model against data with the intent to detect any inadequacies and to make improvements to the model (**Fig 1**.). ACF and PACF are often used in the identification.

The autocorrelation function (ACF) of series y_t at lag k (r_k) is, [5]

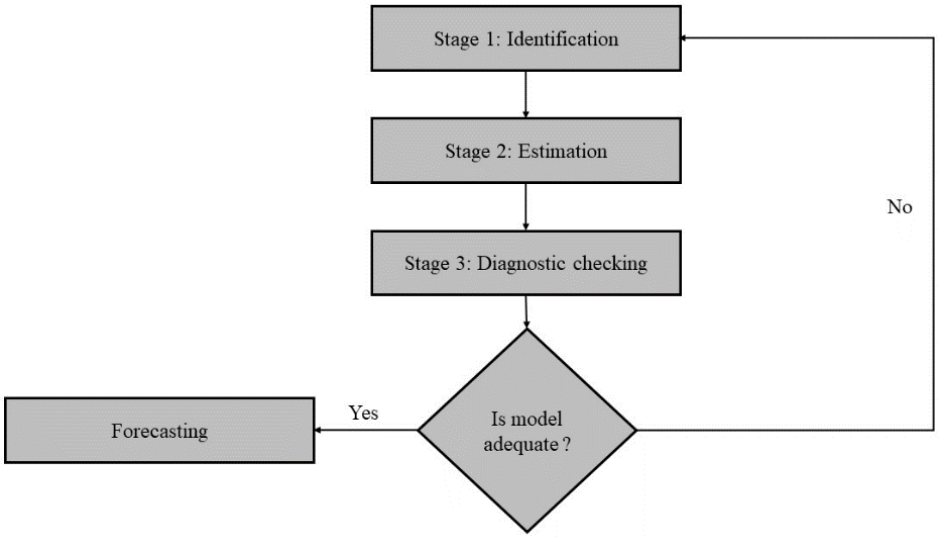

Fig. 1 Three stage of the Box-Jenkins methodology [2].

$$r_k := \frac{E\left[\left(y_t - \mu\right)\left(y_{t+k} - \mu\right)\right]}{\sigma_y^2}, \tag{7}$$

where μ is the mean value of the series, and the variance σ^2_y of the stochastic process can be estimated by $\sigma_y^2 = \sum_{t=1}^{n}\left(y_t - \mu\right)^2 / n.$

The partial autocorrelation function (PACF) of series y_t at lag k (f_{kk}) is, [5]

$$f_{kk} := \begin{cases} r_{1,\ldots,} & k = 1; \\ \dfrac{r_k - \sum_{j=1}^{k-1} f_{k-1j} r_{k-1j}}{1 - \sum_{j=1}^{k-1} f_{k-1j} r_{k-1j}}, \ldots, & k = 2,3,\ldots, \end{cases} \tag{8}$$

where $f_{kj} = f_{k\text{-}1j} - f_{kk}\, f_{k\text{-}1k\text{-}j}$.

For a complete understanding of the estimation situation, it is necessary to make a thorough analytical study of the likelihood probability function. In addition, as the ACF and the PACF determine more than one model, the Akaike Information Criteria (AIC) is used to identify the best fitted model among them [5]

$$AIC = -2\ln(L(\beta)) + 2\omega, \tag{9}$$

where ω is the number of estimated parameters, and $\hat{\beta}$ is the maximum likelihood function values. AIC consists of two parts. The first item reflects the model precision and the second marks the number of model parameters, which presents a positive relation with the order number. Akaike's Small Sample Correction Information Criterion (AICC) is one of the best criteria for selecting SARIMA models, AICC is the smallest and has a

negative value. In practice, the AICC gives the best model when it has the lowest per module, a negative value.

The *Q*-statistics test is applied to verify the tentative adequateness of the model

$$Q = N\sum_{k=1}^{m}\hat{r}_k^2 \approx \chi^2(m) \tag{10}$$

where m is the specified delay lags, and N is the length of the residuals. If the calculated value of Q exceeds the critical value of $\chi^2(m)$ (m - degrees of freedom obtained from the chi-square tables), the tentative model is tuned as inadequate; otherwise, the model is adequate.

b) LSTM

Deep (structured) learning [11], [12], [17], is a subfield of machine learning methods, which is essentially a neural network with three or more layers. There are supervised, semi-supervised or unsupervised deep learning networks. ANN with a single layer can still make approximate predictions, and then additional hidden layers can aid to optimize accuracy. Deep-learning architectures in general speaking, exists as include Restricted Boltzmann Machine (RBM) based deep belief network (DBN), Convolutional Neural Network (CNN), deep Auto-encoders, and deep Recurrent Neural Network (RNN) [4], [11], [17]. In contrast to Feedforward Neural Networks, which only pass data forward, RNN have returning connections which enable using the old cell state in addition to new cell input. Hence, it can be said that the neural network has some form of memory and output at any given time is based on new and past input. As time steps pass, the influence of older inputs fades, which is why more recent inputs affect output more than older ones. However, RNNs suffer from the Vanishing Gradient problem, i. e. due to the multiplying of a large number of small values, the gradient can become a very small value [4], [12], [17]. An RNN using LSTM units can be trained in a supervised shape, on a set of training series, using an optimization algorithm, like gradient descent, combined with backpropagation through time to compute the gradients needed during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with respect to corresponding weight. A problem with using gradient descent for standard RNNs is that error gradients vanish exponentially quickly with the size of the time lag between important events. This is due to $\lim n \to \infty$ then $W_n = 0$ if the spectral radius of W -weighted matrices, is smaller than 1. However, with LSTM units, when error values are back-propagated from the output layer, the error remains in the LSTM unit's cell. This error propagating process continuously feeds error back to each of the LSTM unit's gates, until they learn to cut off the value. The main advantage of LSTM neural networks is the ability to store long-term dependencies in data [14].

LSTM used in the field of deep learning, is an artificial recurrent neural network (RNN) architecture [4], [12], [14], [17]. A frequent LSTM unit is composed of an input gate, an output gate, a cell and a forget gate. Every cell remembers values over arbitrary time intervals, the three gates regulate the flow of information into and out of the neuron. The LSTM neuron provides nonlinear mechanism for controlling the information flow into and out of the LSTM cell. The LSTM architecture is depicted in Fig. 2.

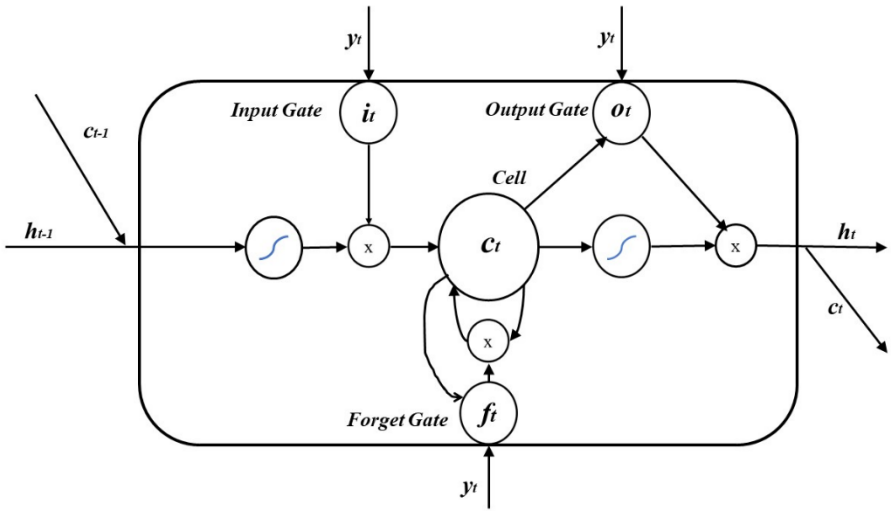

Fig. 2 The architecture of LSTM unit-neuron

As shown in the figure, the LSTM neuron provides a nonlinear mechanism for controlling information flow into and out of the LSTM cell. The forget gate determines the information that need to be discarded or forgotten from the previous cell states. The input gate determines what information will be allowed to enter into the neuron state. Finally, the output gate decides the information to be passed out of neuron state. Mathematically, the representation for the forward pass of an LSTM unit with a forget gate for $y_t : \{1, 2, ..., n\} \to R^d$ (input vector to the LSTM unit; for (1), d=1), is as follows, [14]

$$f_t = \sigma_g \left(W_f y_t + U_f h_{t-1} + b_f \right) \tag{11}$$

$$i_t = \sigma_g \left(W_i y_t + U_i h_{t-1} + b_i \right) \tag{12}$$

$$o_t = \sigma_g \left(W_o y_t + U_o h_{t-1} + b_o \right) \tag{13}$$

$$c'_t = \sigma_c \left(W_c y_t + U_c h_{t-1} + b_c \right) \tag{14}$$

$$C_t = f_t {}^o c_{t-1} + i_t {}^o c'_t \tag{15}$$

$$h_t = o_t {}^o \sigma_h (c_t) \tag{16}$$

The initial values are $c_0 = 0$ and $h_0 = 0$ and the operator ^{o} denotes the Hadamard product (element-wise product). Variables are: $f_t \in (0,1)^h$ - forget gate's activation vector; $i_t \in (0,1)^h$ - input/update gate's activation vector; $o_t \in (0,1)^h$ - output gate's activation vector; $h_t \in (-1,1)^h$ - hidden state vector or output vector of the LSTM unit; $c'_t \in (-1,1)^h$ - cell input activation vector; $c_t \in R^h$ - cell state vector; $W \in R^{h \times d}$, $U \in R^{h \times h}$ and $b \in R^h$ weight matrices and bias vector parameters which need to be learned during training where the superscripts, f, i and h refer to the number of forget, input features and number of hidden units, respectively. Activation functions σ - $\sigma_{g,c,h}$ are, respectively: sigmoid, hyperbolic, or, identic function. SGDM is a stochastic optimization method that adds the expression of the corresponding impulse to the known stochastic descent gradient in the parameter

space. This achieves a faster convergence of the local minimum. Also, in the case of a shallow local minimum, this moment may be sufficient for the gradient to eject the local solution, which is a great advantage over the standard method.

c) ELM

ELM [4], [18-23] was first introduced to improve the efficiency and speed of a single-hidden-layer feedforward network. The ELM algorithm, as contrasting to the conventional belief of ANN theory, linear theory, and control theory, does not require hidden nodes/neurons and is a training algorithm for the single hidden layer feedforward neural network (SLFN). Unlike standard ANN, that periodically assigns hidden nodes, ELM randomly assigns hidden nodes, constructs biases and input weights of hidden layers, and determines the output weights using least squares methods. This significantly justifies the low computational time of ELM. Different than gradient based methods, ELM assigns random values to the weights between input and hidden layer and the biases in the hidden layer, and these parameters are frozen during training. The nonlinear activation functions in hidden layer provide nonlinearity for the system. Then, it can be regarded as a linear system. The only parameter that network needs to learn is the weight between a hidden layer or the threshold of the hidden layer and output layer. Hence, ELM converges much faster than traditional algorithms because it learns without iteration. Random hidden nodes promise the universal approximation ability. Theoretical analysis showed that ELM is more likely to reach global optimal solution with random parameters than traditional networks with all the parameters to be trained [18-22]. Compared with the support vector machine (SVM) [13]. ELM tends to yield better classification performance with less optimization constrains. Due to its superior training speed and good generalization capability, ELM is widely applied in a variety of learning problems, such as classification, regression, clustering, and feature mapping, [4], [18-23].

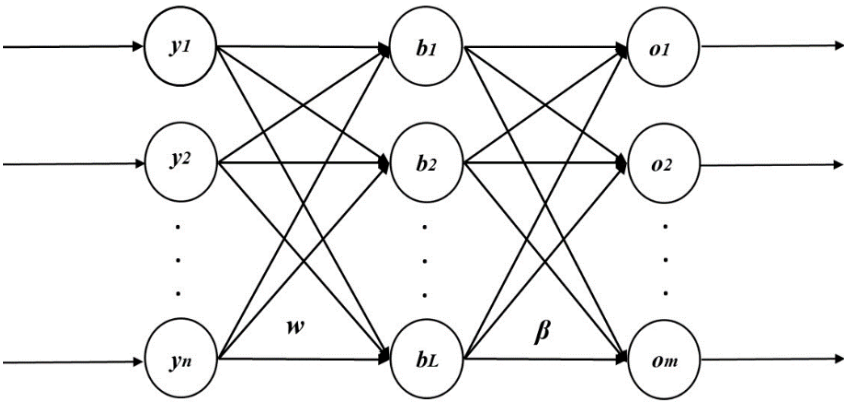

Fig. 3 Structure of SLFN.

The training problem for ELM is given in [22]. The schematic diagram of ELM is presented in Fig.3, [22]. A training set is $S = \{(\boldsymbol{y}_i, \boldsymbol{t}_i) \mid \boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})^T \in R^n, \boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{im})^T \in R^m\}$, where $\boldsymbol{y}_i$ denotes the input value and $\boldsymbol{t}_i$ represents the target,

where m is the number of output layer nodes. The output $\boldsymbol{o}_j$ of an ELM with L hidden neurons, N- number of training samples, can be expressed as

$$\mathbf{o}_j = \sum_{i=1}^{L} \beta_i \sigma\left(\mathbf{w}_i \mathbf{y}_j + b_i\right), j = 1,...,N. \tag{17}$$

In ELM, activation functions σ are nonlinear ones to provide nonlinear mapping for the system. The set of vectors $\mathbf{w}_i$ is the weight vector for the input layer in the i-th hidden node, b_i is the value of the bias in the i-th hidden node, $\boldsymbol{\beta}_i$ is the weight vector for the output layer in the i-th hidden node. The goal of training is to minimize the error between the target and the output of ELM. The most commonly used object function is mean squared error (MSE), defined as

$$MSE = E\left(\sum_{j=1}^{N}\left(\mathbf{y}_j - \mathbf{o}_j\right)\right) \tag{18}$$

where N is the number of training samples, and i and j are the indexes for the training sample and output layer node.

The basic training of ELM can be regarded as involving two steps: random initialization and linear parameter solution. Firstly, ELM uses random parameters wi and bi in its hidden layer, and they are frozen during the whole training process. The input vector is mapped into a random feature space with random settings and nonlinear activation functions which is more efficient than those of trained parameters. In the second step, βi can be obtained by Moore-Penrose inverse [22].

Besides MSE, for assessing the quality of prediction, RMSE (root mean square error)

$$RMSE = \sqrt{MSE} \tag{19}$$

and determination coefficient R^2

$$R^2 = \frac{E\left(\sum_{j=1}^{N} \mathbf{o}_j \mathbf{y}_j\right) - E\left(\sum_{j=1}^{N} \mathbf{o}_j\right) E\left(\sum_{j=1}^{N} \mathbf{y}_j\right)}{E\left(\sum_{j=1}^{N} \mathbf{o}_j\right) E\left(\sum_{j=1}^{N} \mathbf{y}_j\right)}. \tag{20}$$

are also used.

Other criteria for assessing the quality of time series models known in the literature are: Schwarz–Bayesian information criteria (SBC), sum of squares error (SSE), mean absolute percentage error (MAPE), Mean Average Error (MAE), and final prediction error (FPE), [1-6], [28-30].

Basic ELM training includes two steps: random initialization and solving a linear parameter problem. First, ELM uses the generated random weights and bias parameters $\boldsymbol{w}_i$ and b_i in its hidden layer, whose values do not change throughout the training process. The input vector is mapped to a random state space determined by random properties and nonlinear activation functions, which turns out to be a more efficient way than that with trained parameters. In the second step, $\boldsymbol{\beta}_i$ is calculated via the Moore-Penrose inverse of $\mathbf{H}$, acting on $\boldsymbol{t}_j$ [22]. They are further used to calculate the vector of output values $\boldsymbol{o}_j$

Possible areas of application of specific ELM algorithms are: recognition of objects in images, various classification problems, analysis of large amounts of data, hybrid online

learning, Self-Organizing Extreme Learning Machine (SOELM), processing unbalanced data, extremely fast ML, etc. However, ELM is faster than many ANN methods. With simple implementation, ELM performance is good in terms of accuracy as well.

3. NUMERICAL RESULTS

The most illustrative numerical results related to the application of the SARIMA model and ANN models in forecasting the volume of packages in the internal postal traffic of Serbia are presented here. We considered the monthly data for the period 2014-2020.

The logarithm of the total volume of internal parcel postal traffic in the Republic of Serbia for the entire considered period is shown in Fig. 4. This procedure of logarithmic scaling reduces the scope of the data set and thus simplifies the calculation.

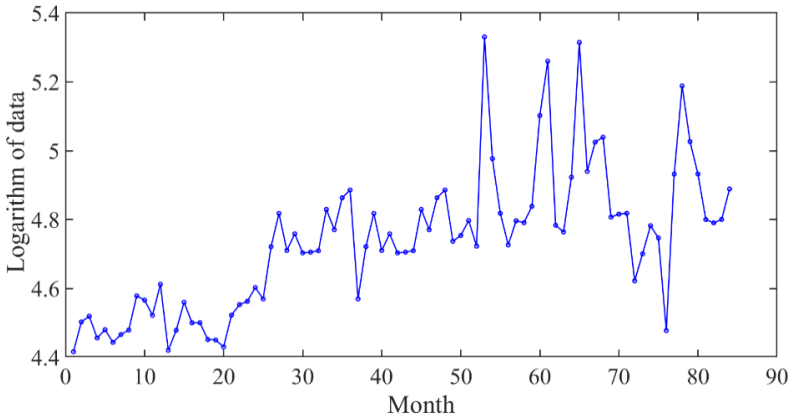

Fig. 4 Logarithm of the volume of internal parcel postal traffic of Serbia - original data, 84 monthly observations - period from January 2014 to December 2020.

XLSTAT software environment was used to find the best (S)ARIMA basic model properties and parameters, on the basis of which the analysis was continued in the MATLAB environment. MATLAB was also used for forecasting time series data using the LSTM network and ELM.

The best values of (S)ARIMA parameters (p, P, d, D, q, Q, s) can be determined in XLSTAT. For thus obtained values the model evolves by the aid of the process presented in Fig. 1. The procedure for obtaining the numerical values of parameters, criteria and graphic solutions is fast, and the work on establishing the group of statistical models is efficient. Some parameters and criteria of models are described in [5], [26-29] and supplied in this paper. The convergence values (10^{-5}) and maximal number of iterations (5000) were specifically selected, s=0, 12, and confidence interval is always 95%. In an XLSTAT environment the maximum values for d, D are set to be 1, and for p, P, q, Q are 4. The best obtained value for AICC is -77,8888 within the work on several hundreds of cases of various parameters and model parameters for this model are selected. The convergence towards this solution has been made in 40 iterations over the measurement subset for these 72 observations.

All the optimal affirmative statistics is shown in the Table 1. The model parameters, without s=0, (S)ARIMA (Eq.(2)) are provided in the Tables 2 and 3. The value of the trend component is provided in the Table 2. Based on hundreds of analyzed (S) ARIMA approaches in the XLSTAT environment of a given time series, in addition to noticing small values of basic parameters, it is recognized that the best model is not seasonal. The values of the parameters AR(1), MA(1) and MA(2) are presented in the Table 3. The value of RMSE=0.131536.

Table 1 Goodnes of fit statistics

Observations	72
DF	68
SSE	1.24572
MSE	0.017302
RMSE	0.131536
WN Variance	0.017302
MAPE(Diff)	1.779189
MAPE	1.779189
-2Log(Like.)	-86.4858
FPE	0.017789
AIC	-78.4858
AICC	-77.8888
SBC	-69.3791
Iterations	40

Table 2 SARIMA trend component

Parameter	Value	Hessian standard error
Trend component	4.680	0.127

Table 3 SARIMA model parameters

Parameter	Value	Hessian standard error
AR(1)	0.975	0.030
MA(1)	-0.493	0.130
MA(2)	-0.215	0.120

The value of the basic probability parameter is -2Log (Likelihood) =-86.4858. The SBC parameter has a value of -69.3791, which is, in theory, correct, that is, the smallest. This value serves as the criterion for selecting the model within the final set of models and is related to the Akaike information criterion.

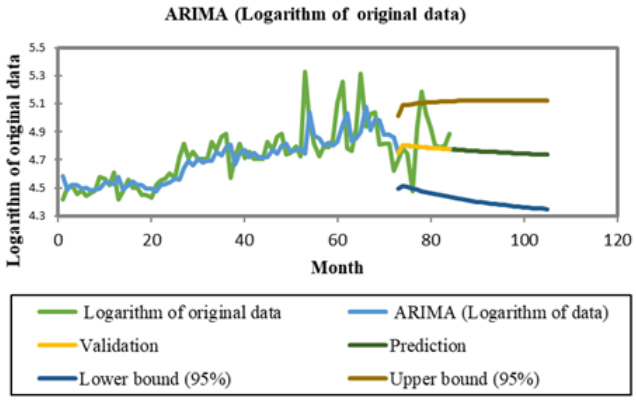

Fig. 5 Results of XLSTAT simulation for the logarithm of internal parcel postal traffic of Serbia, 84 monthly observations - period from January 2014 to December 2020.

The total results of the XLSTAT simulation are summarized in Fig. 5. In addition to the logarithms of the original data, their ARIMA values, validation, and prediction data, as well as the lower and upper 95% bounds of the prediction are given. Simulation results for 84 elements are given, including validation (last 12 members of the time series) and prediction (21 new members, from month 85 to 105). Confidence interval bounds (95%) are set for validation and prediction. For last 21 members of the time series, the values of characteristic validation parameter values are RMSE=0.1804 and R^2 = 0.0673.
The graph of the dependence of residual values on the number of months (84 in total) is shown in Fig. 6. The maximum of residuals is for month 54.

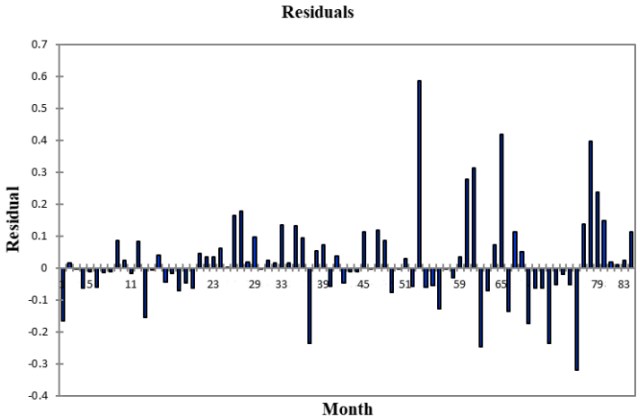

Fig. 6 XLSTAT residual graph for the logarithm of the volume of internal parcel postal traffic of Serbia.

XLSTAT descriptive analysis – determining the values of p and q (ACF and PACF) for a given time series is shown in Fig. 7 and Fig. 8. The obtained values are PACF – p = 1, ACF – q = 1 (cutoff after lag =1, according to the previously stated procedure after Eqs. (7) and (8)).

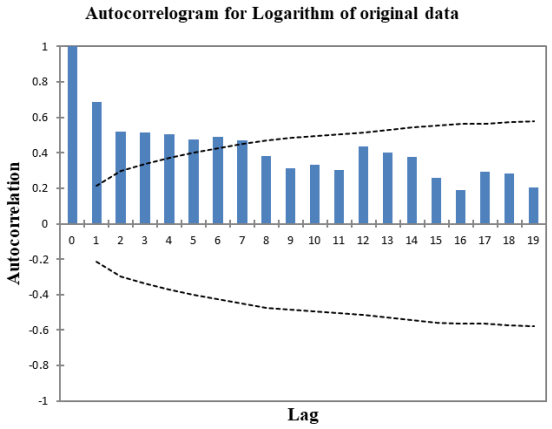

Fig. 7 XLSTAT ACF graph for the logarithm of the volume of internal parcel postal traffic of Serbia.

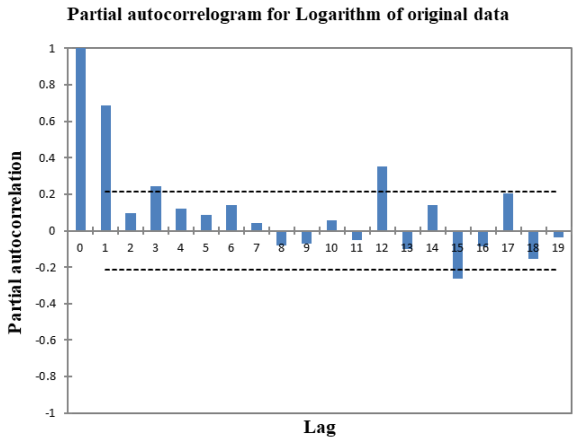

Fig. 8 XLSTAT PACF graph for the logarithm of the volume of internal parcel postal traffic of Serbia.

XLSTAT descriptive analysis (ACF and PACF) for residuals is shown in Fig. 9 and Fig. 10. By definition, PACF provides a partial correlation of a stationary time series with its own lagged values. Here, the results are P=0, Q=0 (cutoff after lag=0, due to small autocorrelations, according to the previously stated procedure). According to the results presented in Fig. 5 – Fig. 10, integrated process ARIMA (1,0,1) $(0,0,0)_0$ is proposed.

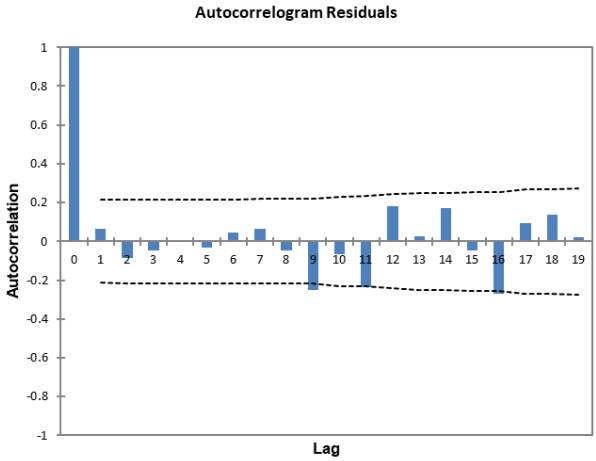

Fig. 9 XLSTAT ACF graph for residuals of the logarithm of the volume of internal parcel postal traffic of Serbia.

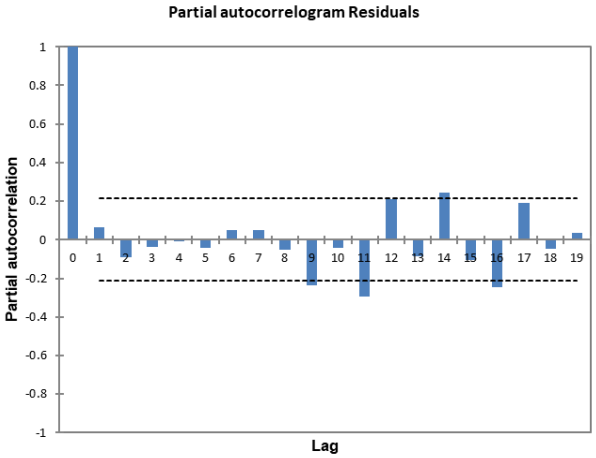

Fig. 10 XLSTAT PACF graph for residuals of the logarithm of the volume of internal parcel postal traffic of Serbia.

Prediction of data of the same time series was also done using the LSTM network in the MATLAB Deep Learning Toolbox.

For forecasting the last 21 time series values, LSTM is trained by sequence-to-sequence regression, using previous 63 members, and the training sequence responses are compared to actual time series values, shifted one step forward. LSTM network contains a sequence input layer, an LSTM layer having 200 hidden units, a fully connected layer, and a regression output layer. For the LSTM regression network training, an SGDM optimizer is used, and it was trained for 400 epochs, with the use of the state of forecasting and improvement. Sub-commands used, apart from the standard ones: *Momentum* (moment

value from interval (0, 1) which corrects the current value of stochastic gradient) and *L2Regularization* (multiplicator of network layers' parameters, introduced for their regularization, field of value is up to 0.1). The values of characteristic parameters are RMSE = 0.24585 and $R^2 = 0.011866$. The comparison of the LSTM model forecast the data of the original time series, as well as their difference, is shown in Fig. 11.

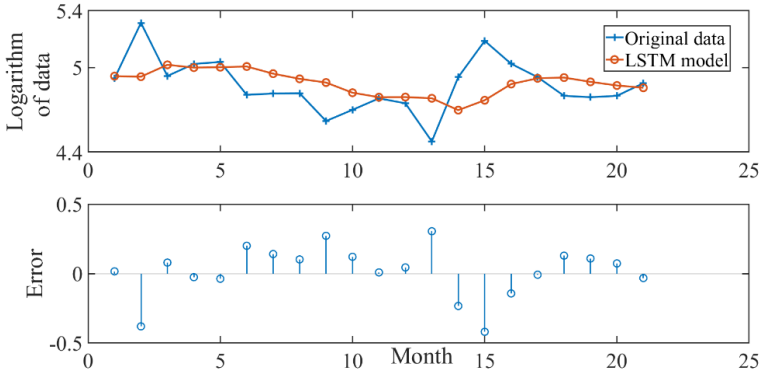

Fig. 11 LSTM forecast and corresponding errors - the differences between validation and observed datasets, values related to the logarithm of the volume of internal parcel postal traffic of Serbia.

ELM programs within MATLAB were used for the improvement of ARIMA and LSTM forecast data. As for both, the first 63 members of the original sequence are used for the training, and the remaining 21 data (test group) are from the original ARIMA and LSTM series. The initial values of the weight vector, **w**, and bias, **b**, (equation (17)) are generated randomly. Various transfer functions are tested, such as sigmoid, sinusoidal, and so-called unit hardlim function and the first one proved to be the best. In order to reduce the value of RMSE and increase the value of R^2, in the simplest case, the procedure was repeated several times, with the appropriate program conditions. It has been observed that the optimal number of iterations is 200,000 per 1 million such iterations for different initial machine-generated random conditions. Regarding the improvement of the ARIMA forecasting, the following values RMSE=0.12985 and R^2=0.59921 are obtained. The ELM method was applied to the results of 63 LSTM training data. The obtained results (21 of them) were compared with the original data and the following prediction quality is achieved: RMSE = 0.14561 and $R^2 = 0.59905$.

In order to reach 200000 iterations in ELM, the simulations of up to 10^6 iterations for each problem were made, with the increases up to 100000. There was no improvement in RMSE and R^2 values above 200000, which means that in this way, the relatively fast convergation of these parameters value cannot be implemented.

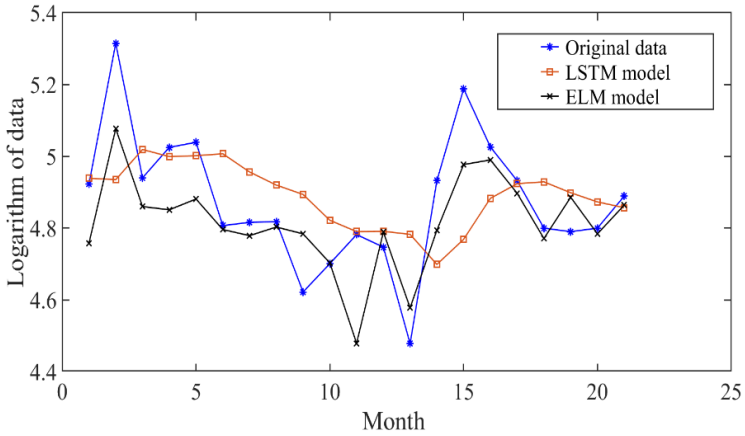

Fig. 12 Comparison of the test set results (21 data) of LSTM method and ELM method with the original data

The comparison of the results obtained by applying the two proposed methods with the original data is shown in Fig.12. While LSTM proved to be better for tracking the original data at the ends, ELM better represented the initial group, as a whole. The ELM method has the best performances in terms of RMSE and R^2 values compared to ARIMA and LSTM. The practical division into 63 training elements and 21 test group elements within all possible modeling was selected after studying the partition containing 72 training elements and 12 test group elements. It was estimated that the possible improvements of all methods in the two cases regarding RMSE according to the number of test data are approximately 10^{-5}.

5. Conclusion

The combined data analysis techniques based on the ARIMA statistical model, as well as LSTM and ELM neural network models are presented in the paper. The analysis was performed on a series of consecutive monthly data representing the volume of packages in the internal postal traffic of Serbia for the period 2014-2020.

The easy-to-use XLSTAT EXCEL software environment was first used to find the basic parameters of the (S)ARIMA model. The results were supplemented by the analysis of source and residual ACF and PACF time series data. A similar way of processing data, of the same time series, using the LSTM network trained in the MATLAB Deep Learning Toolbox was also considered. The modeling results obtained by the LSTM method were compared with the original data. The results of the ARIMA and LSTM methods were then used to improve the predicted ELM values. The ELM method was applied to the results of the ARIMA and LSTM methods in order to improve and obtain a more accurate prognosis. The obtained results of ELM methods, in both cases (for the LSTM model and for the ARIMA model), were compared with the original data. The obtained RMSE value for the ELM model was found to be about 28% lower than the corresponding one for the ARIMA model.

This way of combining and comparing the above forecasting methods is a significant novelty in relation to the application of previous individual techniques, both in terms of quality of results and methodology.

Acknowledgement: *This work was supported by the Ministry of Education, Science and Technological Development of Republic of Serbia (Grant No. 451-03-9/2021-14/200102).*

REFERENCES

[1] Taeho Jo, *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Springer Nature Switzerland AG, 2021.
[2] Manjusha Pandey; Siddharth Swarup Rautaray, *Machine Learning: Theoretical Foundations and Practical Applications*. Springer Nature Singapore Pte Ltd. 2021.
[3] https://ai.stanford.edu/people/nilsson/mlbook.html
[4] arXiv:2012.03854v2 [stat.AP] 3 Jun 2021. Available: https://arxiv.org/pdf/2012.03854v1.pdf
[5] G.E. Box, G.M. Jenkins, G.C. Reinsel, G. M. Ljung, *Time Series Analysis, Forecasting and Control*. New Jersey, John Wiley and Sons, 2016.
[6] R. J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice*. 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 26 September 2021.
[7] D. C. S. Bisht (eds.), M. Ram (eds.), *Recent Advances in Time Series Forecasting*. Boca Raton, CRC Press, 2021.
[8] A. Ticherahine, A. Boudhaouia, P. Wira and A. Makhlouf, "Time series forecasting of hourly water consumption with combinations of deterministic and learning models in the context of a tertiary building," Int.Conf. on Decision Aid Sciences and Application, 2020.
[9] H. Yu, L. Jingming, R. Sumei, Z. Shuping, "Hybrid Model for Financial Time Series Forecasting – Integration of EWT, ARIMA with The Improved ABC Optimized ELM," IEEE Access, 1–1. doi:10.1109/access.2020.2987547.
[10] B. Predić, N. Radosavljević, A. Stojčić, "Time Series Analysis: Forecasting Sales Periods in Wholesale Systems," Facta universitatis, Series: Automatic Control and Robotics, vol. 18, no 3, pp. 177–188, 2019.
[11] M. Paluszek, S. Thomas, *MATLAB Machine Learning Recipes: A Problem-Solution Approach*. Second Edition, New Jersey, Apress, 2019.
[12] M. Paluszek, S. Thomas, *Practical MATLAB Deep Learning: A Project-Based Approach*. New Jersey, Apress, 2020.
[13] S. Hochreiter, J. Schmidhuber, "Long short-term memory", Neural Computation. 9 (8): 1735–1780, 1997.
[14] arXiv:1303.5778v1 [cs.NE] 22 Mar 2013.
[15] https://www.mathworks.com/help/deeplearning/ref/sgdmupdate.html .
[16] A. Voelker, I. Kajić, C. Eliasmith, "Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks," 33rd Conference on Neural Information Processing Systems NeurIPS Vancouver, Canada, 2019.
[17] P. Poonia, V. K. Jain, "Short-Term Traffic Flow Prediction: Using LSTM", IEEE, International Conference on Emerging Trends in Communication, Control and Computing, 2020.
[18] Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew, "Extreme learning machine: theory and applications," Neurocomputing. 70 (1): 489–501, 2006.
[19] Guang-Bin Huang, Qin-Yu Zhu, K. Z. Mao, Chee-Kheong Siew, P. Saratchandran and N. Sundararajan, "Can threshold networks be trained directly?" in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 53, no. 3, pp. 187-191, March 2006, doi: 10.1109/TCSII.2005.857540.
[20] G. -B. Huang, H. Zhou, X. Ding and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 2, pp. 513-529, April 2012, doi: 10.1109/TSMCB.2011.2168604.
[21] Guang-Bin Huang, "What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle," Cognitive Computation, volume 7, issue 3 2015.
[22] J. Wang, S. Lu, Shui-Hua Wang, Yu-Dong Zhang, "A review on extreme learning machine," Multimed Tools Appl, 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-11007-7 .
[23] https://elmorigin.wixsite.com/originofelm .
[24] https://www.programmersought.com/article/231014377/

[25] https://www.xlstat.com/en/.
[26] https://www.mathworks.com/help/econ/estimate-multiplicative-arimamodel-using-econometric-modeler.html .
[27] N. Knežević, N. Glišović, M. Milenković, N. Bojović, "Neural Networks Based on Metaheuristics for Forecasting the Revenue of Postal Services," (in Serbian), PosTel 2018, pp.33-42, Belgrade, Serbia, 2018.
[28] I. D. Rogan, O.R. Pronić-Rančić, "Forecasting the volume of postal services using Savitzky-Golay filter modification," Sozopol, Bulgaria, 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), 2021. DOI: 10.1109/ICEST52640.2021.9483459.
[29] I. D. Rogan, O.R. Pronić-Rančić, "SARIMA and ANN Approaches in Forecasting the Volume of Postal Services," Niš, 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications, TELSIKS, 2021.
[30] C. M. Hurvich, Chih-Ling Tsai, "Regression and time series model selection in small samples", Biometrika, vol. 76, no. 2, pp. 297-307, 1989.
[31] https://www.posta.rs/eng/stanovnistvo/usluga.aspx?usluga=postal-services/parcel-services-serbia/sending-parcels-within-serbia