UDC 004.852

N. D. Uvarov[1], M. P. Kuznetsov[2], A. S. Malkova[3], K. V. Rudakov[4], V. V. Strijov[5]

Selection of superposition of models for railway freight forecasting*

Our aim is to construct an optimal superposition of models for the short-term railway traffic forecasting. The historical data constitutes daily railway traffic volume between pairs of stations for different cargo types. The given time series are highly volatile, noisy, and non-stationary. We propose a system that finds an optimal superposition of forecasting models with respect to historical data features. Among the candidate models the system considers: moving average model, exponential and kernel smoothing models, ARIMA model, Croston's method and LSTM neural networks.

Keywords: time series, forecasting, superposition, forecasting models, asymmetric distribution.

1. Introduction. The Russian Railways are highly important for the Russian economy. Railroad transportations rank second in cargo turnover, outclassed by pipeline transportations only. Around 80% of Russian Railways profits are generated due to cargo transportation [1]. In order to compete with other modes of transport available for clients, the company needs to optimize economical efficiency of the traffic. Therefore the transportation volume forecasting problem naturally arises in the business. The extensive development does not lead to the desired economical outcome. Hence, intensive approach consisting of thorough optimization of existing capacities with the use of modern data analysis is suggested to improve company efficiency and increase profits.

Not all time series are subject to normal distribution of noise component [2]. A separate model for residual forecast may improve the forecast quality for such series. In [3] the Markov chain is used to forecast residuals; the resulting accuracy of the wind farm electricity output forecast is improved by 10%. In the tourism industry the Fourier spectre correction is common approach [4].

The aim of our work is to construct a model capable of forecasting time series with asymmetric noise distribution. We consider superposition models, which are described in [5]. Two models are used in the superposition: one of them predicts the main, usually smooth, component of a time series, whereas the other corrects the forecast in order to minimize error introduced by the noise present in the series. Errors of the first model constitute training

[1]CAM department of MIPT, student, E–mail: nikita.uvarov@phystech.edu
[2]Yahoo! Research, researcher, E–mail: mikhail.kuznecov@phystech.edu
[3]CAM department of MIPT, student, E–mail: malkova@phystech.edu
[4]CS department of MSU, prof., Dr. sc. math, corr. member of RAS, E–mail: rudakov@ccas.ru
[5]FRC CSC RAS, prof., Dr. sc. math, E–mail: strijov@ccas.ru
*The research was supported by the Russian Foundation for Basic Research, project 17-20-01212.

data for the second model. Whenever more than one forecast point is requested, either the forecast of the first model or the corrected superposition, forecast can be used to calculate the errors. In the computation experiment we compare the obtained forecast quality with the quality of base algorithms. We also investigate the forecast quality dependency on the magnitude of noise distribution asymmetry.

We use the following base algorithms as superposition components in the experiment: the autoregressive moving average model, ARIMA [6], the singular spectrum analysis model, SSA [7], the LSTM neural network model [8]. Among the less complex algorithms we consider the exponential [9] and kernel [10] smoothing models. An optimal combination of the mentioned algorithms varies depending on the time series. We consider all possible pair variants and build the superposition forecast quality matrix to find the optimal pair.

The historical data in our research is the Russian Railways transportation journal. The data includes the following information: shipment date, departure and destination stations, number of wagons, cargo type code, wagons type, total weight, and routing sign. The series are highly volatile, non-uniform, and noisy; some of them are non-stationary. Additionally, there are stations pairs with infrequent transportations that produce series with a lot of zero values. We use the Dickey–Fuller procedure to test whether the series is stationary.

The daily railway load data is represented in [11] and contains information about shipment date, station pair, cargo code, and volume. The length of the series is 1 year; the data contains shipments history for 78 regions, 4000 stations and 43 cargo types such as oil, ore, peat, automobiles, cotton, sugar and grain.

We also test our approach on the airline passenger data [12], German electricity prices [13], and a synthetic series generated as a sine wave with additive noise sampled from the Wald distribution.

2. Optimal superposition construction problem.

A parametric function $f \circ g$, which evaluates an unknown time series value $\hat{x}_{t+1}$ at the time moment $t+1$ given prior series observations $\mathbf{x} = [x_1, ..., x_t]^\intercal$ is called the *forecasting model*:

$$\hat{x}_{t+1} = f \circ g(\hat{\mathbf{w}}, x_t, x_{t-1}, \ldots, x_1). \tag{1}$$

The parameters $\hat{\mathbf{w}}$ are obtained by minimizing the error function.

Forecasting models are constructed from candidate models $f, g \in \mathfrak{F}$. The family $\mathfrak{F}$ is described in section 4. A forecast error (regression residual) is a difference $\varepsilon_e = x_t - \hat{x}_t$ between the historical time series value and its forecast $x_{t+1} = f \circ g(x_t, x_{t-1}, \ldots, x_1) + \varepsilon_{t+1}$. We assume that residuals expectation is not zero, $\mathsf{E}(\varepsilon) \neq 0$, and that their variance is constant, $\mathsf{D}(\varepsilon) = \sigma^2$.

Function f forecasts trend and seasonal components of time series whereas g forecasts residuals of f. The final forecast $\hat{x}_{t+1}$ accounts for both components. The idea is that g increases the accuracy of the forecast given by f:

$$f : x_t \to \hat{x}^f_{t+1},$$

$$g : x^f_{t+1} \to \hat{x}^{f,g}_{t+1}.$$

The problem is to construct a forecast (1) at r points in the future, that is, to determine $[\hat{x}_{t+1}, \ldots, \hat{x}_{t+r}]^\mathsf{T}$. The *forecast request* is the number of points at which the forecast must be calculated.

The *forecast horizon* is the number of steps h, $h \leq r$, after which the difference between the forecast value $\hat{x}$ and the historical value x is significant:

$$||\hat{x}_{t+h-1} - x_{t+h-1}|| \leq \mu, \text{ whereas } ||\hat{x}_{t+h} - x_{t+h}|| > \mu. \tag{2}$$

In order to obtain the model parameters $\hat{\mathbf{w}}$ we minimize the loss function $S(\mathbf{w})$:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{\tau=1}^{t} S(\mathbf{w}, x_\tau - \hat{x}_\tau).$$

The forecast error S is estimated using MAPE (5).

3. Retrospective forecast by superposition. We consider two variants of a superposition forecast algorithm and an ordinary sequential forecast algorithm. Both superposition forecast algorithms aim to increase the accuracy by using an additional model to forecast the residuals. Therefore, two functions are used to build the forecast: the first function builds the base forecast and the second function corrects it by forecasting the errors of the first function. The difference between the algorithms lies in the inputs to the first function. When building forecasts for requests greater than 1, we can use outputs of the model as inputs for future points. However, when building a superposition forecast we can also use the corrected superposition value as input.

3.1. Parallel superposition forecast algorithm. The inputs are the time series $\mathbf{x}$ and the functions f and g. The algorithm (3) consists of the following steps.

1. The function f is used to calculate n forecasts of the history end using the points $\hat{x}^f_t, ..., \hat{x}^f_{t-n+1}$ as inputs.

2. n residuals $\hat{\varepsilon}_t, ..., \hat{\varepsilon}_{t-n+1}$ are calculated as differences

$$\hat{\varepsilon}_{t-k} = x_{t-k} - \hat{x}^f_{t-k},$$

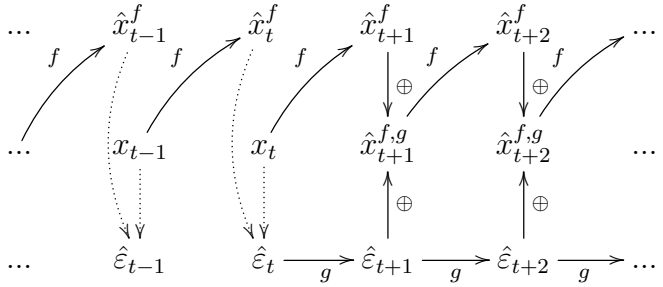
where t is the length of historical data.

3. The function g is used to calculate the residual forecast $\hat{\varepsilon}_{t+i}$ for h points ahead.

4. The final forecast is calculated as

$$\hat{x}^{f,g}_{t+i} = \hat{x}^{f}_{t+i} + \hat{\varepsilon}_{t+i}$$

using the sequential forecast $\hat{x}^{f}_{t+i}$ obtained from the function f.



(3)

3.2. Accumulating superposition forecast algorithm. The inputs are time series $\mathbf{x}$ and functions f and g. A *forecast depth* i is the number of time series elements that participate in the forecast of a point. The algorithm (4) consists in the following steps.

1. The function f is used to calculate the retrospective forecast $\hat{x}^{f,(1)}_{t+1}, ..., \hat{x}^{f,(i)}_{t+i}$; the forecast of the point $t+i$ is on the horizon i and depth i.

2. The function f is used to calculate r sets of the history end forecasts, $\hat{x}^{f,(i)}_{t}, ..., \hat{x}^{f,(i)}_{t-n+1}$, each set is a forecast — on depth i, $i = 1, ..., r$.

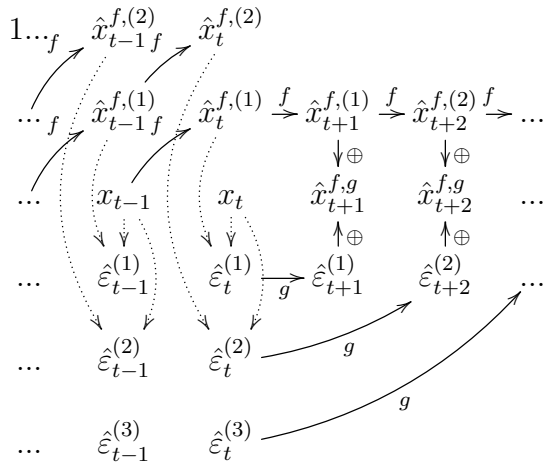
3. r sets of residues $\hat{\varepsilon}^{(i)}_{t}, ..., \hat{\varepsilon}^{(i)}_{t-n+1}$ are calculated as

$$\hat{\varepsilon}^{(i)}_{t-k} = x_{t-k} - \hat{x}^{f,(i)}_{t-k}, \quad i = 1, ..., r.$$

4. The function g is used to forecast the residues $\hat{\varepsilon}^{(i)}_{t+i}$; each forecast is one-point and is based on the calculated series $\hat{\varepsilon}^{(i)}_{t}, ..., \hat{\varepsilon}^{(i)}_{t-n(g)+1}$.

5. The final forecast is calculated as

$$\hat{x}^{f,g}_{t+i} = \hat{x}^{f,(i)}_{t+i} + \hat{\varepsilon}^{(i)}_{t+i}.$$



(4)

3.3. Sequential forecast algorithm. The inputs are time t_0 and time series $\mathbf{x}_{1\div t_0} = [x_1, x_2, \cdots, x_{t_0}]$. A forecast for request r is $\mathbf{x}_{t_0+r}$, $t_0+r < T$, where T is the length of historical data. The algorithm consists of the following steps.

1. The function f is used to forecast the next series element $\hat{x}_{t+1}$.
2. The new element $\hat{x}_{t+1}$ and the whole series history now serves as an input to f in order to obtain $\hat{x}_{t+2}$.
3. First two steps are repeated until the forecast for the request r is obtained.

3.4. Block forecast algorithm. To evaluate the quality of the forecasts on requests larger than 1 we introduce the *block forecast*.

The time series is split into training part $\mathbf{x}_{1\div t_0}$ and validation part $\mathbf{x}_{t_0+1\div t_1}$. To estimate the quality we build $t_1 - t_0$ forecasts, i-th forecast ($0 \le i < t_1 - t_0$) is built using values $\mathbf{x}_{1\div t_0+i}$ as a one-point forecast.

A block length r is a maximal forecast request that does not exceed the forecast horizon of the model used. The forecast to be compared with the validation part of the series is then built as follows.

1. Using the training data $\mathbf{x}_{1\div t_0}$ and the sequential forecast algorithm build an r-points ahead forecast.
2. Now treat the first r values of the validation part as known. Use series $\mathbf{x}_{1\div t_0+r}$ to build a new r-step ahead forecast.
3. Repeat step 2 until there are values in the validation part.

The resulting forecast contains equal high number of values obtained from the $1, 2, \ldots, r$-step ahead forecasts. Therefore, its quality is an average quality of the algorithm on forecast requests from 1 to r.

4. Computational experiment. In our experiment we analyse the quality and stability of the forecast with the use of superposition forecasting models under the condition of asymmetric distribution of the residuals. We use the following time series: railroad transportation series [14], electricity consumption series [15], electricity price series [13], airline passenger series [12], and a sine wave with additive noise sampled from Wald distribution.

We investigate the forecasts built by the base models and according to their superpositions. We provide visualizations of the forecasts and numerical values of the error. It is estimated using MAPE (5) and MSE (6):

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\varepsilon_i}{x_i}\right|, \tag{5}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2. \tag{6}$$

In all experiments the quality of a block forecast (3.4) is used.

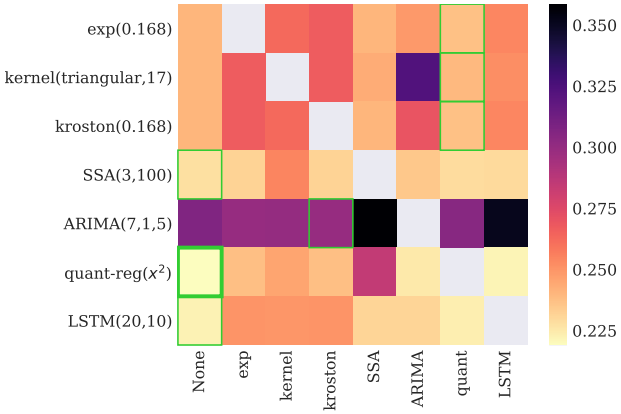


Fig. 1: Superposition forecast quality matrix (MAPE)

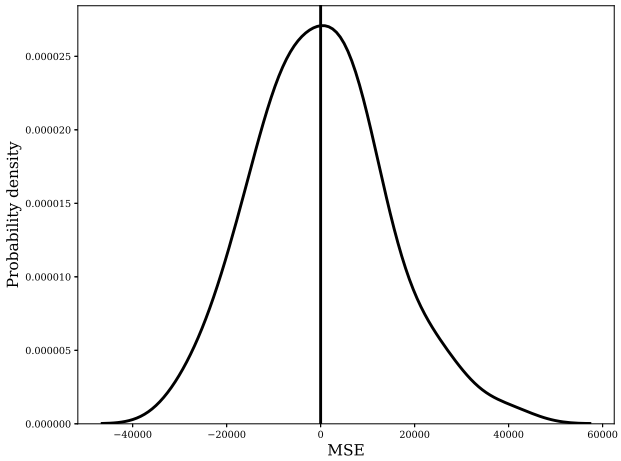


Fig. 2: Empirical residual distribution density for the best base model (LSTM)

For each time series (airline, synthetic, german, railroads) we provide:

1. A plot of series with a split into training and validation part visualized.
2. Plots of the base models forecasts with the optimal hyperparameters. All forecasts are built using (3.4). The block length in all experiments is 10.
3. The forecast of the best superposition and separately of this superposition main model.
4. The superposition forecast quality matrix in logarithmic scale. Its rows correspond to main forecasting models, its columns correspond to residual forecasting models. The

first column corresponds to no residual forecast model and thus contains base algorithm results. Cells with a border coresspond to the best forecast in a row.

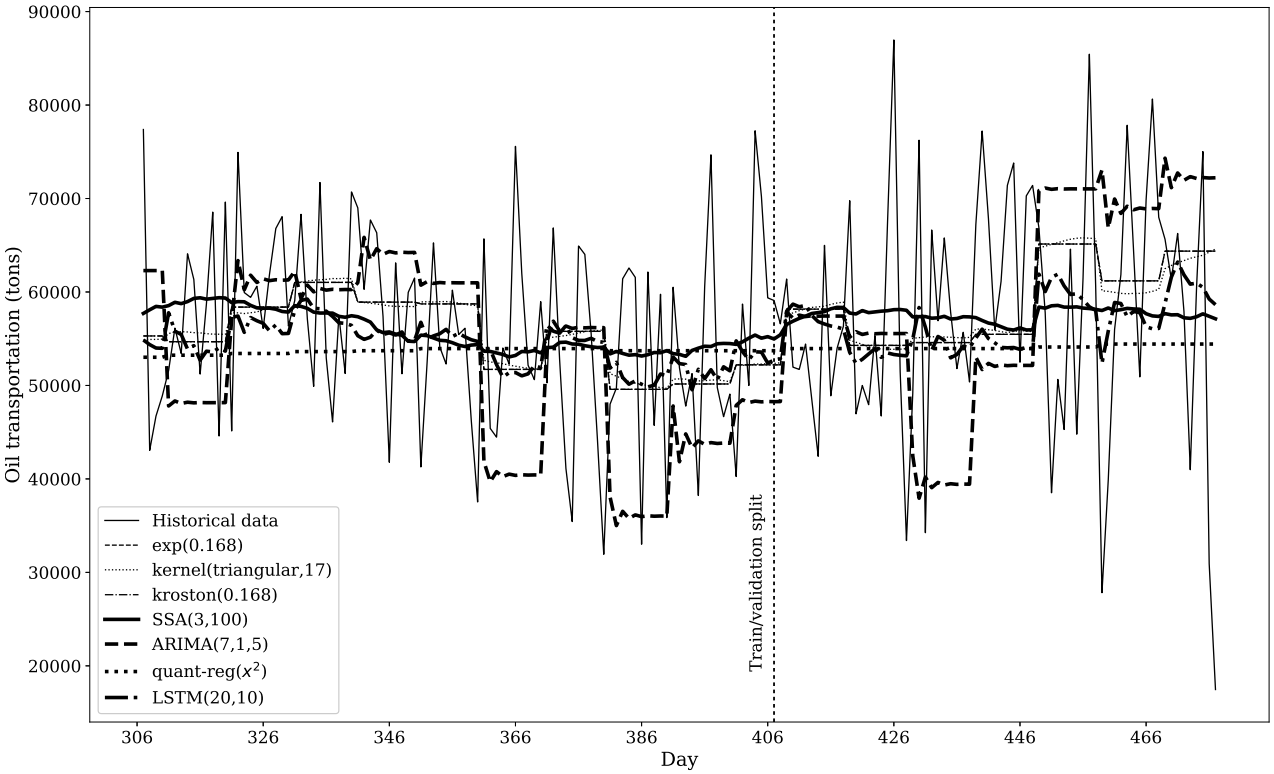


Fig. 3: Base models' forecasts

5. Experimental results: railroads.

Figure 2 shows that the distribution of the best model residuals is close to normal. Therefore, using a superposition does not improve the forecast quality.

6. Forecast horizon in the experiments. In order to determine the forecast horizon (2) we plot the MAPE (5) and its sample deviation (represented on the figure through the upper confidence bound) dependency on the forecast request in the experiments. The forecast horizon is then determined graphically by the "broken cane rule" (fig. 4).

The forecast horizon is equal to 7, whereas the used block length is 10, which is slightly greater.

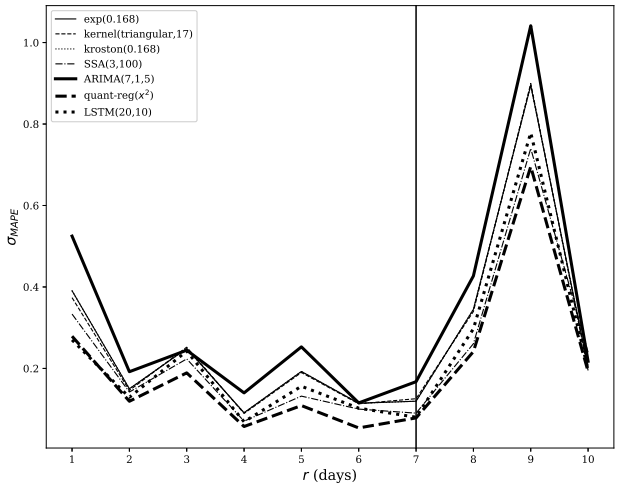


Fig. 4: Broken cane rule: a sharp rise in sample deviation of error indicates that the forecast horizon is reached

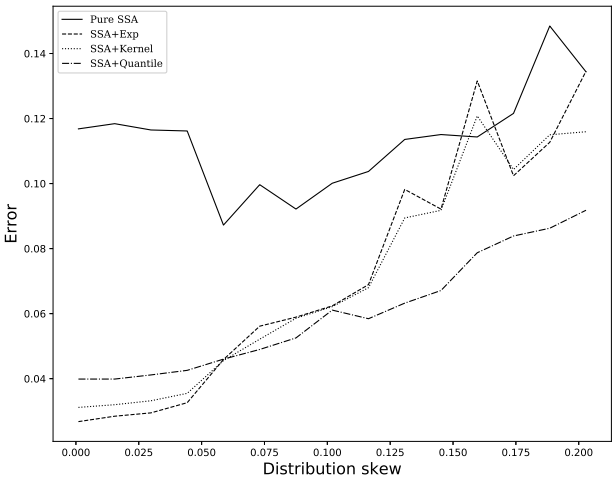


Fig. 5: Quality of the forecast versus the "skew" of a distribution

7. Experiment: dependency of the quality on the noise distribution asymmetry magnitude.

The noise distribution is obtained from a one-parametric subset of the Wald distribution. A single parameter l defines the distribution parameters as $\mu = 1 + \frac{l}{30}, \lambda = \frac{l^2}{4}$. The measure of the distribution asymmetry is

$$S = \left|\mathbb{P}\left(x > x_{\text{m.p.}}\right) - \mathbb{P}\left(x < x_{\text{m.p.}}\right)\right|,$$

where $x_{\text{m.p.}}$ is the most probable value of the random noise variable, corresponding to the probability density maximum and the peak on fig. 5. Clearly $S \in [0, \frac{1}{2})$; a value of 0 corresponds to the distribution, which probability mass is symmetric relative to the peak, e.g. normal distribution; values close to $\frac{1}{2}$ correspond to highly skewed distributions. The loss function used is an asymmetric piecewise-linear function

$$l(x) = \begin{cases} x, & x > 0 \\ -\frac{1}{5}x, & x < 0 \end{cases}.$$

8. Conclusion. We consider the problem of building a superposition model for time series forecasting. We include exponential smoothing, kernel smoothing, Croston's method, SSA, ARIMA, quantile regression, and LSTM in the family of base algorithms. We build the superposition forecast quality matrix, the residual distribution plot, and determine the forecast horizon for several datasets: railroad transportation, airline passenger data, electricity prices and synthetic data. The computational experiment results show that the superposition model quality can be greater than base model quality when the noise distribution

is asymmetric. Additionally we investigate the dependency of the error on the magnitude of noise distribution asymmetry and show that the superposition becomes better as the noise becomes more skewed. The experiment source code is available at [16].

REFERENCES

[1] Буряк Е.В., Кульпина В.П., Голяшев А.В., Лобанова А.А. Динамика грузоперевозок в России // Бюллетень социально-экономического кризиса в России. 2015. **1**. № 4. С. 12-15.

[2] Lijuan L., Liu H., Wu J. et al. A novel model for wind power forecasting based on Markov residual correction // Renewable Energy Congress. 2015. N 6. P. 1-5.

[3] Cohen D. A., Lys T. Z. A note on analysts' earnings forecast errors distribution // Journal of Accounting and Economics. 2003. **36**. N 1-3. P. 147-164.

[4] Kan M. L., Lee Y. B., Chen W. C. Apply grey prediction in the number of Tourist // Genetic and Evolutionary Computing. 2010. N 4. P. 481-484.

[5] Rudakov K.V., Kuznetsov M.P., Motrenko A.P., Stenina M.M., Kashirin D.O., Strijov V.V. Optimal model selection for rail freight forecasting // Automation and Remote Control. 2017. **78**. P. 75-87.

[6] Guha B., Bandyopadhyay G. Gold price forecasting using ARIMA model // Journal of Advanced Management Science Vol. 2016. **4**. N 2. P. 117-121.

[7] Golyandina N. E., Stepanov D.V. SSA-based approaches to analysis and forecast of multidimensional time series // Proceedings of the 5th St. Petersburg workshop on simulation. 2005. **293**. P. 298.

[8] Laptev N., Yosinski J., Li L. E. et al. Time-series extreme event forecasting with neural networks at Uber // International Conference on Machine Learning. 2017. N 34. P. 1-5.

[9] Kalekar P. S. Time series forecasting using holt-winters exponential smoothing // Kanwal Rekhi School of Information Technology. 2004. P. 1-13.

[10] Ralaivola L., D'Alch-Buc F. Time series filtering, smoothing and learning using the kernel Kalman filter // Neural Networks, proceedings of IEEE International Joint Conference. 2005. **3**. P. 1449-1454.

[11] Uvarov N.D., Kuznetsov M.P., Malkova A.S., Rudakov K.V., Strijov V.V. Addition to "Selection of superposition of models for railway freight forecasting". 2018. URL: `http://svn.code.sf.net/p/mlalgorithms/code/Group474/Uvarov2017SuperpositionForecasting/doc/addition_english.pdf` (date of access: 27.03.2018).

[12] Box G. E., Jenkins G. M., Reinsel G. C. et al. Time series analysis, forecasting and control. New York: John Wiley & Sons, 1976.

[13] Electricity price (German). URL: http://svn.code.sf.net/p/mlalgorithms/code/Group474/Uvarov2017SuperpositionForecasting/data/electricity_price_german/GermanSpotPrice.csv (date of access: 27.03.2018).

[14] Railroad transportation time series. URL: http://svn.code.sf.net/p/mlalgorithms/code/Group474/Uvarov2017SuperpositionForecasting/data (date of acccess: 27.03.2018).

[15] Electricity consumption (Poland). URL: http://gdudek.el.pcz.pl/files/PL.xls (date of access: 27.03.2018).

[16] Experiment source code for "Selection of superposition of models for railway freight forecasting" (Python module and Jupyter notebook). URL: http://svn.code.sf.net/p/mlalgorithms/code/Group474/Uvarov2017SuperpositionForecasting/code/ (date of access: 27.03.2018).