The Usage of Lasso-Regression for Optimal Selection of Exogenous Variables for SARIMAX Models

Anfisa I. Baliuk, Alexey I. Chulichkov

Lomonosov Moscow State University, Faculty of Physics,
Department of Mathematical Modeling and Informatics, Moscow, Russia - baliuk.ai20@physics.msu.ru

Keywords: Time series analysis, Time series forecast, SARIMAX model, Lasso-regression, Exogenous variables selection.

Abstract

Time series forecasting and data gap filling are significant tasks in applied science. Nowadays there are a lot of different useful methods for forecasting missing gaps in the data. However, it should be taken into account that the inclusion of a large number of features may lead to overfitting of the model and a decrease in the forecast quality. This paper examines the problem of forecast error minimization in time series with exogenous variables, in which the missing gaps are forecasted by the SARIMAX model. The analysis of Tver region data showed that the selection of significant exogenous variables using Lasso-regression allows for minimization of the forecast error and prevents model overfitting. The obtained results confirm that the correct choice of exogenous variables significantly improves the forecast quality.

1. Introduction

A time series is a *sequence* of random variables so that each value of the series corresponds to a certain point in time. The time interval over which the values of a time series are recorded is constant and does not change for a particular series. Examples of time series can be data from a variety of fields, including economics (e.g., data on the daily flow of customers), finance (e.g., data on stock price fluctuations), and engineering (e.g., weather data). Time series are divided into one-dimensional, in which the target variable is analyzed and predicted only on the basis of its previous values, and multidimensional, in which additional (exogenous) variables are used in addition to the target variable (for example, the target variable can be the value of electricity consumption in the city and the additional variable can be the value of electricity consumption in the suburbs) (Brockwell and Davis, 2016, Chatfield, 1995, Hyndman and Athanasopoulos, 2018, Montgomery et al., 2015).

There is a wide range of forecasting tools, including statistical and neural networks. One of the most widely used statistical models is SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors). SARIMAX is an extension of the ARIMA model that accounts for seasonal patterns and incorporates external (exogenous) variables that may influence the target variable. This makes SARIMAX particularly useful for making forecasts in time series in which the predicted values depend not only on past values, but also on external factors such as economic indicators, weather conditions or market trends.

However, working with multidimensional time series may require a method of selecting significant features to increase the quality of the forecast and avoid overfitting the model. In this paper, Lasso-regression is used to select the significant exogenous features. Lasso-regression (Least Absolute Shrinkage and Selection Operator) is a linear regression method that applies regularization to feature selection. Further results can be taken into account for choosing the right combination of exogenous variables. This can be particularly useful in multidimensional time series analysis where exogenous factors are involved.

2. Statistical models

2.1 MA model

MA (Moving Average) is a statistical model used to analyze time series, based on the assumption that each value of a time series can be expressed as the sum of weighted random errors and the initial value of the time series (Brockwell and Davis, 2016, Chatfield, 1995, Hyndman and Athanasopoulos, 2018, Montgomery et al., 2015).

Definition: y_t is a Moving Average (MA) process of order q in relation to white noise u_t if:

$$\begin{aligned} y_t = \mu + u_t + \alpha_1 \cdot u_{t-1} + \dots \\ + \alpha_q \cdot u_{t-q}, \end{aligned} \tag{1}$$

where $\alpha_q \neq 0$
$\rho_k = cov(y_t, y_{t-k}) \Rightarrow$
$\rho_q \neq 0, \rho_{q+1} = \rho_{q+2} = ... = 0$

The MA process can also be written with the usage of the lag operator:

$$y_t = P_{ma}(L) \cdot u_t, \tag{2}$$

where $P_{ma}(L) = 1 + \alpha_1 \cdot L + \alpha_2 \cdot L^2 + ...$

Publisher's note: Copernicus Publications has not received any payments from Russian or Belarusian institutions for this paper.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-39-2025 | © Author(s) 2025. CC BY 4.0 License.

2.2 AR model

AR (AutoRegressive) is a statistical model used to analyze time series, based on the assumption that each value of a time series can be expressed as a linear combination of previous values of the time series (Brockwell and Davis, 2016, Chatfield, 1995, Hyndman and Athanasopoulos, 2018, Montgomery et al., 2015).

Definition: y_t is an AR(p) model in relation to white noise u_t if:

$$\begin{aligned}(y_t - \mu) = \beta_1 \cdot (y_{t-1} - \mu) + \beta_2 \cdot (y_{t-2} - \mu) + \dots \\ + \beta_p \cdot (y_{t-p} - \mu) + u_t,\end{aligned} \tag{3}$$

where y_t is a MA(∞) process
in relation to white noise u_t

2.3 ARMA model

ARMA (AutoRegressive Moving Average) is a statistical model for analyzing time series, which is a combination of AR model and MA model. The ARMA model describes a time series as a linear combination of its past values and white noise. The AR part of the model uses previous values of the series, and the MA part uses white noise (Brockwell and Davis, 2016, Chatfield, 1995, Hyndman and Athanasopoulos, 2018, Montgomery et al., 2015).

Definition: y_t is an ARMA(p,q) process in relation to white noise u_t if it fits the equation:

$$P_{AR}(L) \cdot y_t = P_{MA}(L) \cdot u_t, \tag{4}$$

where degree $(P_{AR}) = p$
degree $(P_{MA}) = q$
$P_{AR}(0) = 1, P_{MA}(0) = 1$
P_{AR}, P_{MA} are irreducible
y_t is a MA(∞) process
in relation to white noise u_t

2.4 ARIMA model

ARIMA (Autoregressive Integrated Moving Average) is an extension of the ARMA model that includes an integrating component (I). Integration is used to transform non-stationary time series into a stationary ARMA model (Brockwell and Davis, 2016, Chatfield, 1995, Hyndman and Athanasopoulos, 2018, Montgomery et al., 2015).

Definition: y_t is an ARIMA(p,q) process if:

Δy_t - non-stationary

$\Delta^2 y_t$ - non-stationary

$\Delta^3 y_t$ - non-stationary

$\dots$

$\Delta^{d-1} y_t$ - non-stationary

$$\Delta^d y_t - ARMA(p,q), \tag{5}$$

where $\Delta^d y_t = (1-L)^d \cdot y_t$

The ARIMA model equation can be rewritten in the following form:

$$\phi(L)(1-L)^d y_t = \mu + \theta(L)\varepsilon_t, \tag{6}$$

where $\phi(L) = (1 - \phi_1 L - \dots - \phi_p L^p)$
$\theta(L) = (1 + \theta_1 L + \dots + \theta_q L^q)$

2.5 SARIMAX model

SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) is an extension of the classical ARIMA model by including exogenous variables (Brockwell and Davis, 2016, Chatfield, 1995, Hyndman and Athanasopoulos, 2018, Montgomery et al., 2015).

Definition: y_t is a SARIMAX(p,d,q)(P,D,Q)[T] is a process in relation to white noise u_t if:

$$\begin{aligned}P_{AR}(L) \cdot P_{SAR}(L^T) \cdot \Delta^d y_t = P_{MA}(L) \cdot P_{SMA}(L^T) \cdot u_t + \\ + \sum_{i=1}^{n} \theta_i \cdot x_t^i,\end{aligned} \tag{7}$$

where $P_{AR}(L), P_{SAR}(L^T)$ are not
conjugate with $P_{MA}(L), P_{SMA}(L^T)$
degree $P_{AR}(L) = p$
degree $P_{SAR}(L^T) = P$
degree $P_{MA}(L) = q$
degree $P_{SMA}(L^T) = Q$
$P_{AR}(0) = P_{SAR}(0) =$
$= \mathrm{P}_{MA}(0) = P_{SMA}(0) = 1$
d, D are the smallest possible values
$\Delta = 1 - L, \Delta_s = 1 - L^T$
L is the lag operator
x_t^i - exogenous variables

In this paper, the SARIMAX model is used as a key method to fill in data gaps:

1. The SARIMAX model combines additional components to provide a more accurate prediction of real data.

2. Weather data has a strong seasonality, so it is necessary to use models that take this into account.

3. Forecasting process using SARIMAX models

3.1 Dickey-Fuller test

The Dickey-Fuller test is used to determine whether a time series is stationary or non-stationary. The Dickey-Fuller test is a statistical test that checks the presence of unit roots in a time series (Hyndman and Athanasopoulos, 2018).

Let the model of a series be given by the formula:

$$\Delta y_t = \mu + \beta \cdot t + \gamma \cdot y_{t-1} + \sum_{i=1}^{p} \alpha_i \cdot y_{t-i} + \sum_{i=1}^{q} \theta_i \cdot u_{t-i}, \tag{8}$$

where $\Delta y_t = y_t - y_{t-1}$ — first difference of series
$\beta \cdot t$ — trend of series
γ - stationarity test parameter
y_{t-i} - lag values of the target variable
u_{t-i} - lag values of the white noise

The Dickey-Fuller test is a hypothesis testing task:

Hypothesis: $\gamma = 0$.

Even with trend subtraction, the process remains non-stationary.

Alternative: $\gamma < 0$.

Non-stationarity is removed by subtracting the deterministic trend.

If the p-value (significance level) is less than the specified level (usually: 0.05), the hypothesis is rejected and it is concluded that the time series is stationary.

3.2 Information criteria

Information criteria are used to determine model orders. In this paper, the AIC criterion was used:

$$AIC = -2 \log L + 2 \cdot (p + q + k + 1), \tag{9}$$

where L is the likelihood function
$k = 1$ if the series has non-zero mathematical expectation
$k = 0$ if the series has zero mathematical expectation

The best model is selected by finding the smallest possible value of the AIC criterion among all possible values. This follows from maximizing the likelihood function (the logarithm of the likelihood function enters the formula with a negative sign, so the criterion itself is minimized), as well as the penalty for choosing too many parameters (Hyndman and Athanasopoulos, 2018).

It is important to note that information criteria are good optimization methods to find the best orders of p, q. However, they are not capable of picking the order of differentiation - for this purpose, stationarity tests must be used. A series is differentiated until the $p - value$ in the Dickey-Fuller test is less than a given significance level. The inability to use information criteria to determine the order d is due to the fact that differentiation changes the data from which the likelihood function is computed. This makes the AIC values for models with different orders of differentiation incomparable.

3.3 Forecasting procedure

For making a forecast with a usage of the SARIMAX model, the following procedure should be followed (Hyndman and Athanasopoulos, 2018).

1. Determine whether a series is stationary or non-stationary by using the stationarity test (Dickey-Fuller test). If the data are non-stationary, differentiate the series and use the test. If necessary, repeat the procedure until the differentiated series becomes stationary. The number of differentiations of the series will determine the order d in the SARIMAX-model,

2. Put the required seasonality T in the model parameters

3. Find the best model. Two options are possible:
   (a) Use automatic algorithms, which are able to find right p and q orders by minimizing AIC value,
   (b) Manual search by modeling and comparing several models with different p and q orders,

4. Analyze the residuals of the selected models. If the distribution of residuals is similar to the distribution of white noise, then this model is suitable for further forecasting

3.4 Lasso-regression

The ordinary least squares approach (OLS) with regularization is an improved version of the OLS that adds a penalty term to the target function. The penalty parameter helps not only find the optimal coefficients, but also select the necessary features to avoid overfitting the model (Draper and Smith, 1986, Harrell, 2015, Pyt’ev, 1990, Serdobolskaya, 2014).

Lasso-regression is a least squares method with L1-regularization. The L1-norm of the coefficient vector is added as a penalty term.

Lasso-regression can be written as **a regularization problem**

$$S(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \sim \min_{\beta}, \tag{10}$$

Decision:

$$\hat{\beta}_{lasso} = \arg\min_{\beta}(\|Y - X\beta\|^2 + \lambda\|\beta\|_1), \tag{11}$$

where $\lambda > 0$ – regularization parameter

Lasso-regression compares insignificant features with the zero value of their corresponding coefficients. This happens because L1-regularization creates diamond-shaped constraints on the coefficients (the contours of the penalty look like a rhombus). That is why in solution optimization, the intersection of Lasso diamond-shaped constraints with error levels often occurs on the axes, driving the coefficients to zero.

4. Research question

4.1 The selection of exogenous variables using lasso regression for SARIMAX models

If the exogenous variables x_t^i are highly correlated with the target variable y_t or, on the contrary, do not affect the final forecast value, their influence on $\hat{y}_t$ may become excessive, leading to overfitting and, consequently, to an increase in the error. Thus, it is important to select those x_t^i that complement the forecast information $\hat{y}_t$ rather than duplicate this information or make no contribution to the forecast value $\hat{y}_t$.

As mentioned in the previous chapter, Lasso-regression allows selection of only significant features that contribute to the information about the target variable. It can help identify those variables among the x_t^i variables that have predictive power and can reduce the error in predicting the target variable.

An optimization problem can be written based on the formula 10:

$$\min_{\chi i=1,\ldots,k} \sum_t \left(y_t - \sum_{i=1}^{k} \chi_i \cdot x_t^i \right)^2 + \lambda \sum_{i=1}^{k} |\chi_i|, \qquad (12)$$

where x_t^i — exogenous variables
χ_i — coefficients of exogenous variables, which are subject to regularization

If there is a need for the addition of the lag values of the target variable, they should be added to the part with exogenous variables. In this case, the regularization has to be done for both lag values and exogenous variables. However, it has to be taken into account that *the variables which are highly correlated with the target variable may be detected by Lasso-regression as unnecessary features*. Hence, both variants of regularization have to be done to reach more precise results.

4.2 Data description

In this paper, real weather and atmospheric data from Tver region were taken as an example of working with time series. The values data were measured every half hour, the beginning of data collection was November 12, 2011 at 11:00, the end of data collection was December 23, 2020 at 14:00. Data collection included the measurement of several weather parameters.

In this paper the problem of the gap filling in the carbon dioxide concentration variables was studied. Hence, the carbon dioxide concentration (million^{-1}) was taken as the target variable. The following data were taken as additional parameters (exogenous variables): Temperature at 30 meters ($°C$) (further *Temperature*), Relative humidity at 50 meters (%) (further *Relative humidity*), and Solar radiation (W/m^2) (further *Solar radiation*).

The following procedure was done to find the best combination of exogenous variables:

1. 70 artificial gaps containing 3 to 47 missing values were randomly cut out in the carbon dioxide concentration data for subsequent filling and estimation of the prediction error.

2. The gaps were cut in such a way that there was a distance of at least 500 points between them.

3. 240 points were used for model training to fill in the gaps.

4. The exogenous variables did not contain gaps in either the training sample or the test sample (where gaps in the carbon dioxide concentration data were filled in).

5. The obtained gaps were predicted using the SARIMAX model with different sets (combinations) of exogenous variables with a preliminary check for stationarity using the Dickey-Fuller test to assess the adequacy of the selected model parameters.

6. The model parameters were selected automatically using the auto_ARIMA (Python, sktime library).

7. The predicted residuals were analyzed for the lack of correlation (the residuals should be white noise).

8. RSS (Residual Sum of Squares) was calculated for each gap, then the total SSE (Sum of Squared Error) value was found for the entire sample of gaps, which was then normalized by the total size of all gaps and reduced to the RMSE form.

9. RMSE (Root Mean Squared Error) values were compared for different combinations of exogenous variables to identify the best model implementation.

4.3 Evaluation

This subsection contains tables that were obtained in the study.

Firstly, the table of significance coefficients (β) is presented. The results were obtained by the usage of Lasso-regression on all available data, regarding the target variable with exogenous variables as regressors (features).

It can be seen from the Table 1 that the use of Lasso-regression allowed us to identify the most significant variables – *Temperature* and *Relative humidity* with no zero coefficients, excluding *Solar radiation*, which has zero coefficients for some of the penalty parameters.

Secondly, the error table for the SARIMAX models with different sets of exogenous variables is presented (error values are obtained using the procedure described in the previous subsection).

Overall, it can be seen from the Table 2 that the forecast error can be decreased by the usage of excluding redundant features.

As presented in the Table 2, the SARIMAX model with the *Relative humidity* as the exogenous variable has the forecast error less than the SARIMAX model without any exogenous variables. Hence, including *Relative humidity* as a significant variable is bound to reduce the error.

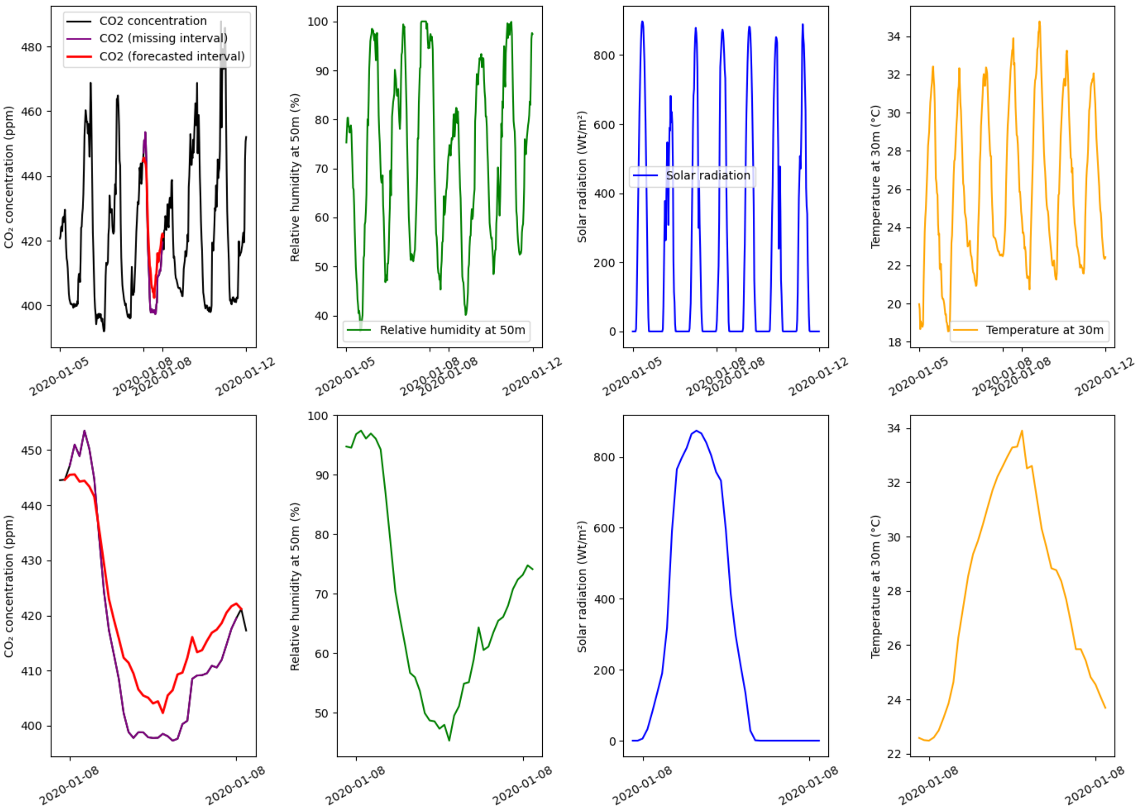

Figure 1. The result of filling the gap with the SARIMAX model with exogenous variables: Temperature, Relative humidity, Solar radiation together with graphs of exogenous variables

Exogenous variable	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$
Relative humidity	8.905	8.899	8.794	7.741	2.536
Solar radiation	3.110	3.090	2.795	0.000	0.000
Temperature	-10.676	-10.665	-10.458	-8.390	-3.198

Table 1. Significance coefficients of exogenous variables obtained by Lasso-regression for different values of the penalty parameter

Including *Temperature* into the model can reduce the forecast error, however, the difference in the errors (comparing the SARIMAX model with *Temperature* as the exogenous variable and the SARIMAX model without any exogenous variables) is not as big as it might be expected. Nevertheless, including *Temperature* into the model at least is not likely to spoil the forecast error. Also, it has to be taken into account that *Temperature* has a negative correlation with the target variable (Table 1). Consequently, the forecast error could be higher than in the model with exogenous variables which have a positive correlation with the target variable since the exogenous variables are included in the model linearly.

As it was expected from the Table 1, the inclusion of *Solar radiation* as an insignificant variable worsened the forecast error compared to the SARIMAX model without exogenous variables. Hence, insignificant variables should be excluded from the construction of the forecast, as they are more likely to increase the error and affect the accuracy.

4.4 Results

This subsection contains some examples of the final graphs that were obtained in the study.

Firstly, Figure 1 shows the results of gap filling by the SARIMAX model with the following set of exogenous variables: *Relative humidity*, *Temperature* and *Solar radiation*, as one of the examples of gap filling.

It is shown that

- *Relative humidity* has the same pattern as the target variable
- *Solar radiation* and *Temperature* have a negative correlation with the target variable
- *Temperature* and *Relative humidity* have the similar pattern, with the difference in the correlation with a target variable

Thus, the usage of Lasso-regression allows us to exclude the exogenous variable - *Solar radiation*, to avoid overfitting.

Secondly, Figures 2 and 3 show the results of gap filling by the SARIMAX model with *Relative humidity* and *Solar radiation* respectively. It can be seen from the graphs that the SARIMAX model with *Relative humidity* (Figure 2) has the better forecast (with the less RMSE) than the the SARIMAX model with *Solar radiation* (Figure 3).

Parameters: exogenous variables	**RMSE**	**Error*, %**
Without exogenous variables (SARIMA model)	17.50	4.23
Temperature	17.40	4.20
Relative humidity	15.96	3.85
Solar radiation	21.98	5.31
Temperature + Relative humidity	17.02	4.11
Temperature + Solar radiation	20.17	4.87
Relative humidity + Solar radiation	20.05	4.84
Temperature + Relative humidity + Solar radiation	19.42	4.69

Table 2. RMSE values for different sets of exogenous variables in the SARIMAX model

* Error is RMSE$/mean$, where $mean$ of CO2 concentration $= 414.01\ mln^{-1}$ in the given period

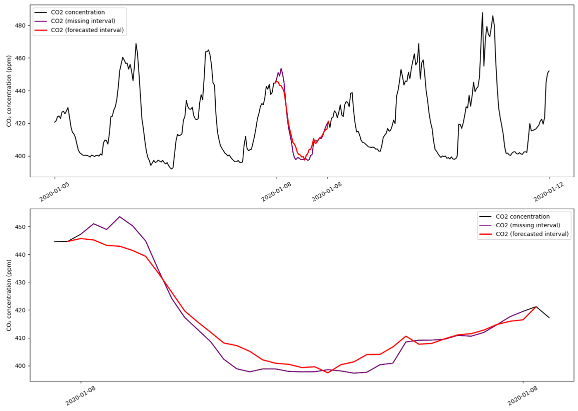

Figure 2. The result of filling the gap with the SARIMAX model with exogenous variables: Relative humidity

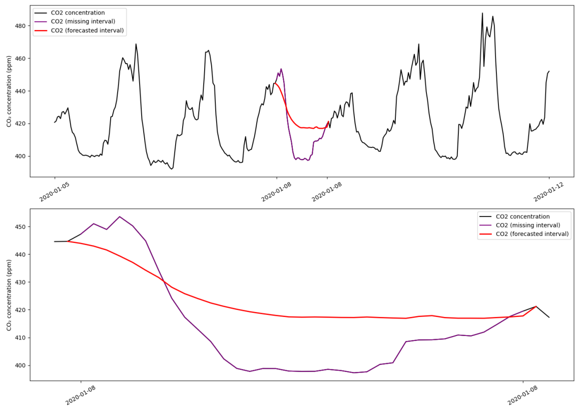

Figure 3. The result of filling the gap with the SARIMAX model with exogenous variables: Solar radiation

5. Conclusion

Lasso-regression can be used for exogenous variables' selection. This allows to minimize the forecast error and prevent overfitting of the SARIMAX model for forecasting time series.

Thus, in the following weather data of Tver region it has been concluded that *Temperature* and *Relative humidity* showed a positive effect, reducing the forecast error. At the same time, the inclusion of *Solar radiation* in the model led to an increase in the error and possible overfitting. This confirms the advisability of pre-processing the data using Lasso-regression to incorporate the final results into the SARIMAX model.

The obtained results can be useful in various fields of science where time series analysis is required.

Acknowledgements

The study was conducted under the state assignment of Lomonosov Moscow State University.

References

Brockwell, P. J., Davis, R. A., 2016. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics, Springer.

Chatfield, C., 1995. *The Analysis of Time Series: An Introduction*. 5th edn, Chapman and Hall/CRC.

Draper, N. R., Smith, H., 1986. *Applied Regression Analysis*. Finance and Statistics, Moscow. Russian edition.

Harrell, F. E., 2015. *Regression Modeling Strategies*. Springer Series in Statistics, Springer.

Hyndman, R. J., Athanasopoulos, G., 2018. *Forecasting: Principles and Practice*. 2nd edn, Monash University.

Montgomery, D. C., Jennings, C. L., Kulahci, M., 2015. *Time Series Analysis and Forecasting*. Wiley.

Pyt'ev, Y. P., 1990. *Methods of Analyzing and Interpreting the Experiment*. Lomonosov Moscow State University, Faculty of Physics.

Serdobolskaya, M. L., 2014. Methods of functional analysis in reduction problems. Lecture notes.