Использование Lasso-регрессии для отбора значимых экзогенных признаков в построении прогноза временных рядов

А. И. Балюк*
Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра математического моделирования и информатики
Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2
(Поступила в редакцию 16.06.2025; подписана в печать 27.10.2025)

Методом наименьших квадратов с регуляризацией предложено проводить предварительную обработку многомерных временных рядов и оптимизировать выбор экзогенных переменных с целью дальнейшего использования полученных результатов в модели SARIMAX. Проведенный анализ погодных и атмосферных данных Тверской области показал, что отбор значимых экзогенных признаков с использованием Lasso-регрессии позволяет минимизировать ошибку прогноза и предотвратить переобучение модели. Полученные результаты подтверждают, что правильный выбор экзогенных переменных улучшает качество предсказательной модели.

PACS: 20.50.Sk УДК: 519.2
Ключевые слова: анализ временных рядов, прогнозирование временных рядов, SARIMAX модель, Лассо-регрессия, оптимальный выбор экзогенных переменных.

ВВЕДЕНИЕ

Прогнозирование временных рядов и заполнение пропусков данных являются важными задачами в прикладных науках. В работе исследуется один из основных инструментов прогнозирования — статистическая модель SARIMAX, являющаяся расширением модели ARMA, а именно: включающая интегрирующую, сезонную компоненты и экзогенные переменные [1–6]. Однако, избыточное число экзогенных переменных может привести к переобучению модели, и, как следствие, возрастанию ошибки прогноза. В настоящей работе предложено использовать Lasso-регрессию для выбора наилучшей комбинации экзогенных признаков.

Для проверки предложенного метода обработки данных использовались погодные и атмосферные данные Тверской области. В качестве целевой переменной, пропуски которой необходимо заполнить, была взята концентрация углекислого газа в атмосфере (*ppm*). В качестве дополнительных признаков — экзогенных переменных — использовались следующие параметры: температура на уровне 30 м над уровнем моря (°C) (*Температура*), относительная влажность воздуха на уровне 50 м над уровнем моря (%) (*Относительная влажность*) и солнечная радиация (Вт/м^2) (*Солнечная радиация*).

1. СТАТИСТИЧЕСКИЕ МОДЕЛИ ВРЕМЕННЫХ РЯДОВ

1.1. MA-модель

Модель скользящего среднего (MA, Moving Average) — это статистическая модель временных рядов, основанная на предположении, что каждое значение временного ряда может быть представлено как сумма взвешенных случайных ошибок и математического ожидания [1–6].

Процесс y_t называется процессом скользящего среднего порядка q (MA(q)) относительно белого шума, если:

$$y_t = \mu + \varepsilon_t + \alpha_1\varepsilon_{t-1} + \ldots + \alpha_q\varepsilon_{t-q}, \qquad (1)$$

где: $\alpha_q \neq 0$, ε_t — белый шум.

A. 1.2. AR-модель

Модель авторегрессии (AR, Autoregressive) — это статистическая модель временных рядов, основанная на предположении, что каждое значение ряда может быть выражено как линейная комбинация его предыдущих значений [1–6].

Процесс y_t называется процессом AR(p) относительно белого шума u_t, если выполняются следующие условия:

1. y_t является процессом MA(∞) относительно белого шума;
2. $(y_t \ - - - \mu)$ выражается в виде:

$$(y_t - \mu) = \beta_1(y_{t-1} - \mu) + \beta_2(y_{t-2} - \mu) + \ldots \\ \ldots + \beta_p(y_{t-p} - \mu) + \varepsilon_t. \qquad (2)$$

1.3. ARMA(p,q)-модель

Перед рассмотрением модели ARMA необходимо ввести понятие оператора лага [3].
Оператором лага называется такой оператор, что:

$$L[\{y_t\}] = \{y_{t-1}\}. \qquad (3)$$

* baliuk.ai20@physics.msu.ru

Модель ARMA объединяет авторегрессию (AR) и скользящее среднее (MA), описывая временной ряд как линейную комбинацию его предыдущих значений и белого шума [1–6].

Процесс y_t называется процессом ARMA(p,q) относительно белого шума ε_t, если выполняются следующие условия:

1. y_t является процессом MA(∞) относительно ε_t.
2. y_t удовлетворяет соотношению:

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

где: $P_{AR}(L)$ — полином AR-оператора порядка p относительно оператора лага L, $P_{MA}(L)$ — полином MA-оператора порядка q относительно оператора лага L, $P_{AR}(0)=1$ и $P_{MA}(0)=1$, $P_{AR}(L)$ и $P_{MA}(L)$ не имеют общих множителей (не сокращаются).

1.4. ARIMA-модель

Модель ARIMA расширяет ARMA, добавляя компоненту интегрирования (I) для приведения нестационарных временных рядов к стационарному виду [1–6].
Процесс y_t называется процессом ARIMA(p,d,q), если выполнены следующие условия:

1. y_t становится стационарным после d-кратного применения оператора разности:

$$\Delta^d y_t = (1-L)^d y_t. \tag{5}$$

2. После приведения к стационарному виду, процесс $\Delta^d y_t$ может быть описан как ARMA(p,q)-модель.

1.5. SARIMAX-модель

Модель SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) является расширением ARIMA с учётом экзогенных переменных и сезонной составляющей [1–6].

Процесс y_t называется процессом SARIMAX$(p,d,q)(P,D,Q)[T]$ относительно белого шума ε_t, если он удовлетворяет соотношению:

$$P_{AR}(L)\cdot P_{SAR}(L^T)\cdot \Delta^d \Delta_s^D y_t = \\ = P_{MA}(L)\cdot P_{SMA}(L^T)\cdot \varepsilon_t + \sum_{i=1}^{n}\theta_i x_t^i, \tag{6}$$

где: $P_{AR}(L)$ — полином авторегрессии порядка p, $P_{SAR}(L^T)$ — полином сезонной авторегрессии порядка P, $P_{MA}(L)$ — полином скользящего среднего порядка q, $P_{SMA}(L^T)$ — полином сезонного скользящего среднего порядка Q, L — оператор лага, L^T — оператор сезонного лага с периодом T, $\Delta = 1-L$ — оператор разности, $\Delta_s = 1-L^T$ — оператор сезонной разности, $P_{AR}(0)=P_{SAR}(0)=P_{MA}(0)=P_{SMA}(0)=1$, Полиномы $P_{AR}(L)$ и $P_{SAR}(L^T)$ не имеют общих множителей с $P_{MA}(L)$ и $P_{SMA}(L^T)$, d и D выбраны как наименьшие возможные для достижения стационарности, x_t^i — экзогенные переменные.

2. LASSO-РЕГРЕССИЯ

Рассмотрим уравнение линейной модели:

$$Y = X\beta + \varepsilon, \tag{7}$$

где: Y — вектор наблюдений, X — матрица признаков (регрессоров), β — вектор коэффициентов, ε — вектор ошибок.

Задача МНК:

$$S(\beta) = \|Y - X\beta\|^2 \to \min_{\beta}. \tag{8}$$

Решение задачи:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{9}$$

Метод наименьших квадратов с регуляризацией — это модификация МНК, включающая штрафной член в целевую функцию. Он помогает отобрать наиболее значимые признаки и предотвращает переобучение модели. Рассмотрим два метода МНК с регуляризацией, а именно МНК с L2-регуляризацией и МНК с L1-регуляризацией [7–10]:

- Ridge-регрессия (L2-регуляризация):

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

Решение задачи:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y, \tag{11}$$

где: $\lambda > 0$ — параметр регуляризации.

- Lasso-регрессия (L1-регуляризация):

$$S(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \to \min_{\beta}. \tag{12}$$

Решение задачи:

$$\hat{\beta}_{lasso} = \arg\min_{\beta}\Big(\|Y - X\beta\|^2 + \lambda\|\beta\|_1\Big), \tag{13}$$

где $\lambda > 0$ — параметр регуляризации. Задача решается численно, так как L1-норма не дифференцируема.

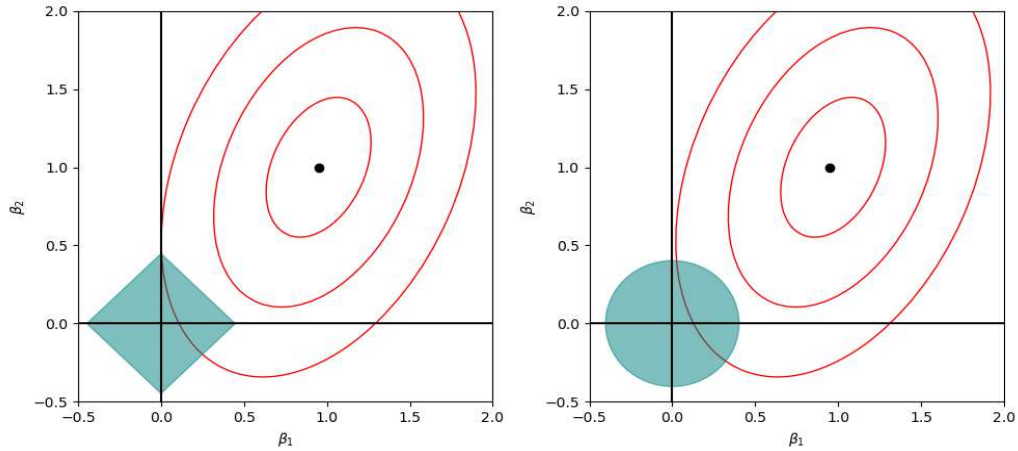


Рис. 1. Геометрическая интерпретация регуляризации: Lasso (L1), Ridge (L2)

Таблица 1. Значения коэффициентов значимости признаков при различных параметрах λ

Параметр	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$
Относительная влажность	8.905	8.899	8.794	7.741	2.536
Солнечная радиация	3.110	3.090	2.795	0.000	0.000
Температура	−10.676	−10.665	−10.458	−8.390	−3.198

Таблица 2. Значения ошибки RMSE для различных наборов экзогенных переменных в модели SARIMAX

Параметры: экзогенные переменные	Ошибка*, %
Без экзогенных переменных (SARIMA-модель)	4.2
Температура	4.2
Относительная влажность	3.9
Солнечная радиация	5.3
Температура + *Относительная влажность*	4.1
Температура + *Солнечная радиация*	4.9
Относительная влажность + *Солнечная радиация*	4.8
Все экзогенные переменные	4.7

*Ошибка: RMSE$mean$, где $mean(\mathrm{CO_2}\ \text{concentration}) = 414.01$ ppm за весь временной период (RMSE = Root Mean Square Error)

.

Разница в регуляризационных членах Ridge-регрессии и Lasso-регрессии влияет на форму оптимизационного пространства:

- L1-регуляризация (Lasso) создает ромбовидные ограничения на коэффициенты

$$\|\beta\|_1 = |\beta_1| + |\beta_2| \le \alpha \rightarrow \text{ ромб},$$

где α — радиус (масштаб) ограничения.

- L2-регуляризация (Ridge) создает круглые ограничения

$$\|\beta\|^2 = \beta_1^2 + \beta_2^2 \le \alpha \rightarrow \text{ круг},$$

где α — радиус (масштаб) ограничения.

При оптимизации решения пересечение ромбовидных ограничений Lasso-регресии с уровнями ошибки часто происходит на осях, приводя коэффициенты к нулю (рис. 1). В Ridge-регрессии пересечение происходит в любой точке круга, что не даёт коэффициентам стать строго нулевыми.

Способность занулять коэффициенты незначимых признаков делает Lasso-регрессию одним из методов отбора значимых признаков.

3. ПРИМЕНЕНИЕ LASSO-РЕГРЕССИИ ДЛЯ ОТБОРА ЭКЗОГЕННЫХ ПЕРЕМЕННЫХ В МОДЕЛИ SARIMAX

Если экзогенные переменные x_t^i сильно коррелируют с целевой переменной y_t или, наоборот, не оказывают влияния на итоговый прогноз, их вклад в значение $\hat{y}_t$ может оказаться чрезмерным, что приведёт к переобучению модели. Таким образом, важно отбирать те переменные x_t^i, которые действительно дополняют информацию к значению проноза $\hat{y}_t$.

Как отмечено выше, Lasso-регрессия позволяет отбирать признаки, которые способны внести вклад в предсказание целевой переменной.

4. ОТБОР ЗНАЧИМЫХ ЭКЗОГЕННЫХ ПЕРЕМЕННЫХ ПРИ ПОМОЩИ LASSO-РЕГРЕССИИ

Результаты, представленные в табл. 1, получены с использованием Lasso-регрессии на всех имеющихся данных. Экзогенные переменные использовались в качестве регрессоров.

Согласно результатам, представленным в табл. 1, значимыми признаками будут *Относительная влажность* и *Температура*, поскольку они имеют ненулевые коэффициенты при всех параметрах λ. *Солнечную радиацию* необходимо исключить как незначимый признак, т.к. переменная имеет нулевой коэффициент при некоторых параметрах λ. Стоит отметить, что в модель SARIMAX экзогенные переменные включены *линейно*, следовательно, в первую очередь необходимо обратить внимание на экзогенные переменные с положительной корреляцией с целевой переменной (экзогенные переменные с положительными коэффициентами значимости).

5. РЕЗУЛЬТАТЫ

Результаты исследования показали (табл. 2), что включение *Относительной влажности* уменьшает ошибку прогноза. Следовательно, значимый признак с положительной корреляцией (с целевой переменной) демонстрирует результат лучше, чем SARIMAX-модель без экзогенных переменных (SARIMA-модель).

Включение *Солнечной радиации* в модель повысило ошибку прогноза (табл. 2). Отсюда можно сделать вывод, что незначимый признак ухудшает качество прогноза и показывает результат хуже, чем SARIMAX-модель без экзогенных переменных (SARIMA-модель).

Несмотря на то, что *Температура* и является значимым признаком, включение этой переменной в модель не повлияло на ошибку прогноза. Согласно данным из табл. 2, в сравнении с моделью SARIMA ошибка не изменилась. Как отмечалось выше, в модель SARIMAX экзогенные переменные включены *линейно*, следовательно, для улучшения качества прогноза необходимо включать признаки с положительной корреляцией с целевой переменной.

В качестве примера представлены графики (рис. 2, 3, 4, 5) заполненных пропусков моделью SARIMAX с экзогенными переменными: *Относительная влажность*, *Солнечная радиация*. На рис. 2, 3, 4, 5 показано, что модель SARIMAX, в которой учитывается *Относительная влажность* дает результат лучше, чем модель SARIMAX, в которой учитывается *Солнечная радиация*. Таким образом, включение только значимых экзогенных переменных, имеющих положительную корреляцию с целевой переменной, в построение прогноза способно уменьшить ошибку.

ЗАКЛЮЧЕНИЕ

Показано, что Lasso-регрессия может быть использована для отбора экзогенных переменных. Это позволяет минимизировать ошибку прогноза и предотвратить переобучение модели SARIMAX для прогнозирования временных рядов.

На примере погодных и атмосферных данных Тверской области сделан вывод, что *Относительная влажность* и *Температура* показали положительный эффект, уменьшая или не ухудшая ошибку прогноза. Напротив, включение *Солнечной радиации* увеличило ошибку и привело к возможному переобучению модели. Это подтверждает целесообразность предварительной обработки данных с использованием Lasso-регрессии для включения конечных результатов в модель SARIMAX.

[1] *Brockwell P.J., Davis R.A.* Introduction to Time Series and Forecasting. Springer Texts in Statistics. 2016
[2] *Chatfield C.* The Analysis of Time Series: An Introduction. CHAPMAN & HALL/CRC. 1995.
[3] *Hyndman R. J., Athanasopoulos G.* Forecasting: Principles and Practice. Monash University. 2018.
[4] *Montgomery D. C., Jennings C. L., Kulahci M.* Time Series Analysis and Forecasting. Wiley. 2015.
[5] *Nielsen A.* Practical Time Series Analysis. O'REILLY Media. 2020.
[6] *Cowpertwait P.S.P., Metcalfe A.V.* Introductory Time Series with R. Springer. 2009.
[7] *Drapper N.R., Smith H.* Applied Regression Analysis. Moscow «Finance and Statistics». 1986.
[8] *Harrel F. E.* Regression Modeling Strategies. Springer Series in Statistics. 2015.
[9] *Пытьев Ю. П.* Методы анализа и интерпретации эксперимента. Московский Государственный Университет им.

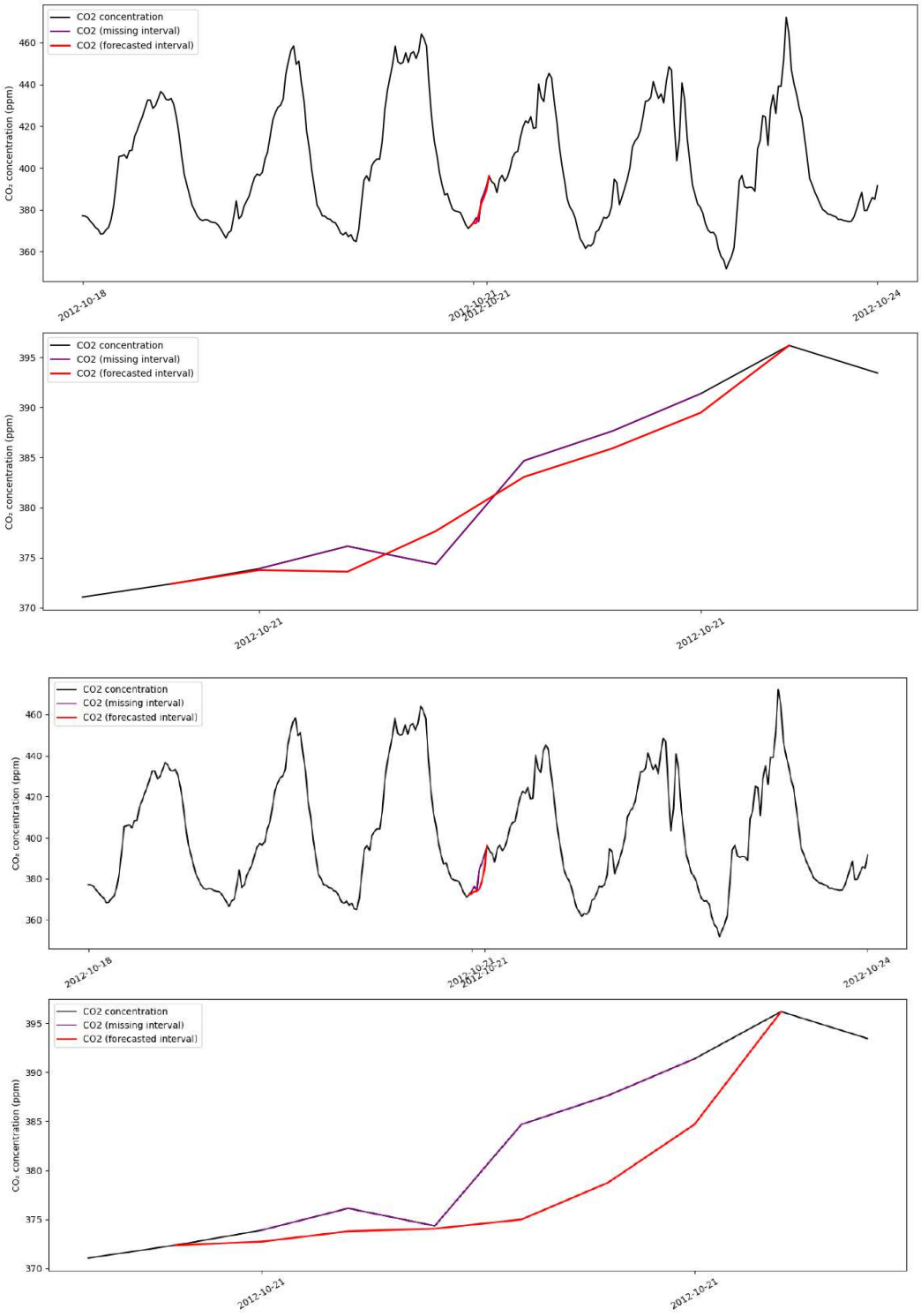
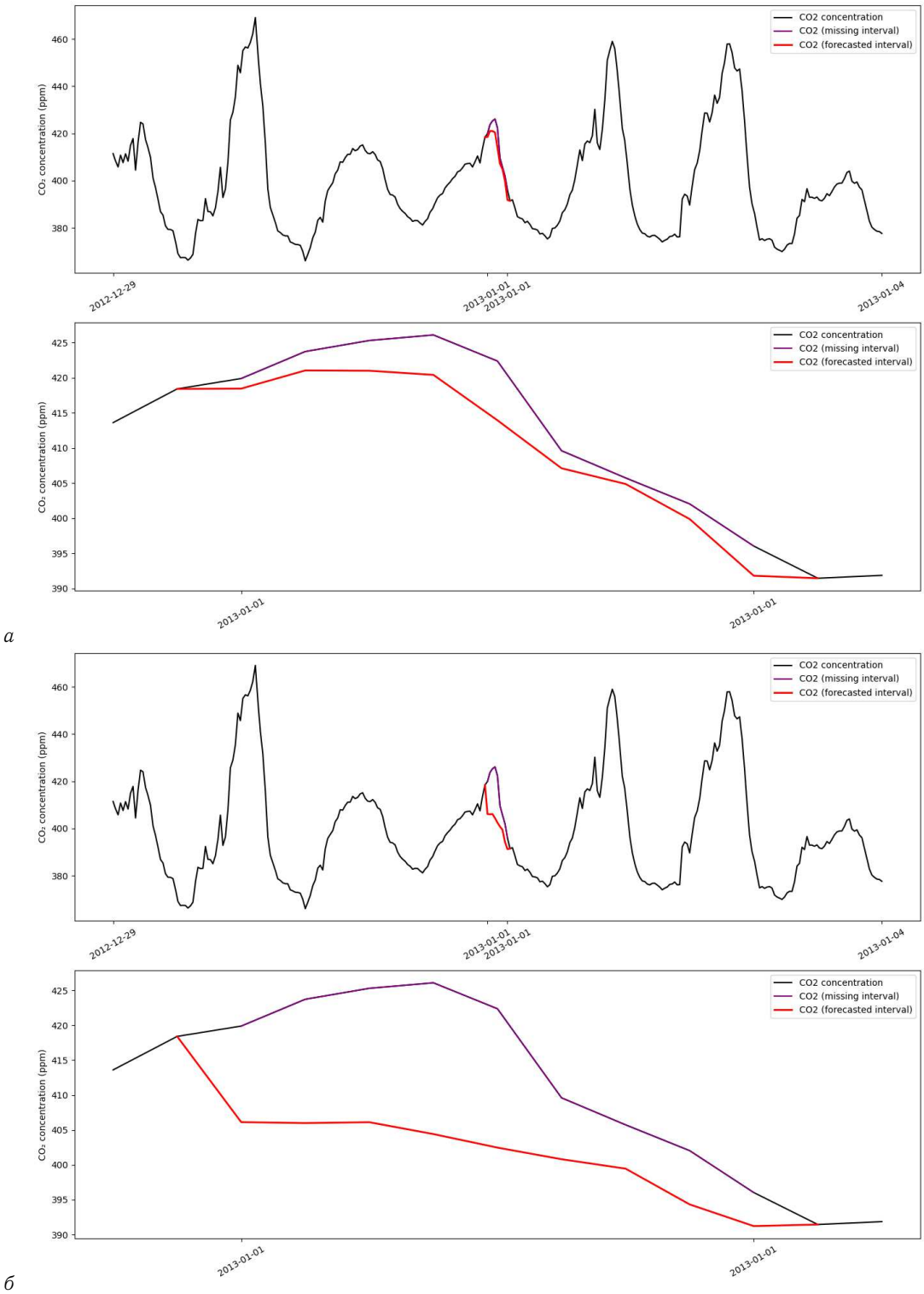


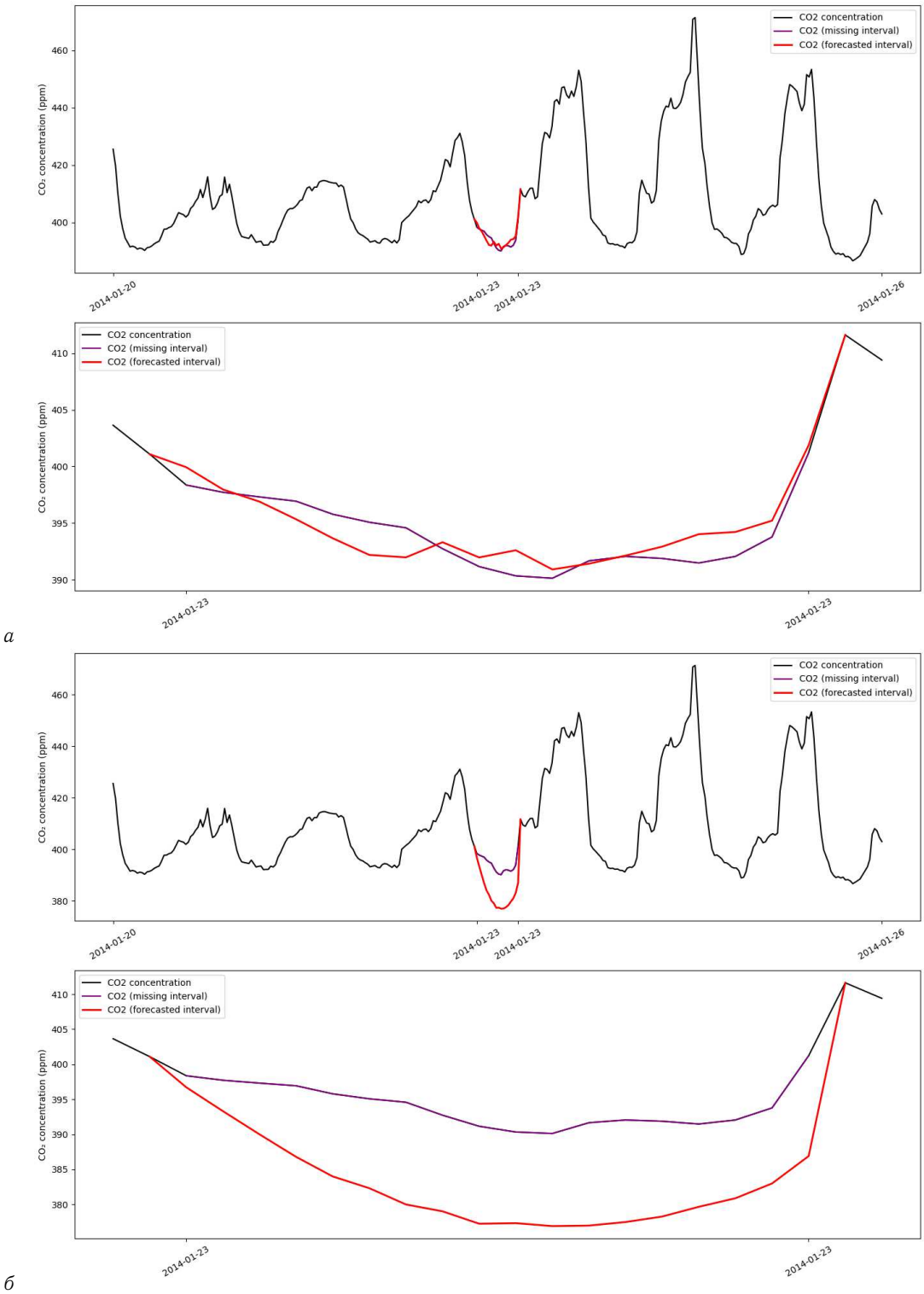
Рис. 2. Графики заполненных пропусков моделью SARIMAX с экзогенными переменными: *Относительная влажность*, *Солнечная радиация*. Начало пропуска: 2012-10-21 16:30:00. Конец пропуска: 2012-10-21 19:00:00. а — Модель SARIMAX с *Относительной влажностью*, б — Модель SARIMAX с *Солнечной радиацией*



а

б

Рис. 3. Графики заполненных пропусков моделью SARIMAX с экзогенными переменными: *Относительная влажность*, *Солнечная радиация*. Начало пропуска: 2013-01-01 05:30:00. Конец пропуска: 2013-01-01 09:30:00. а — Модель SARIMAX с *Относительной влажностью*, б — Модель SARIMAX с *Солнечной радиацией*



а

б

Рис. 4. Графики заполненных пропусков моделью SARIMAX с экзогенными переменными: *Относительная влажность*, *Солнечная радиация*. Начало пропуска: 2014-01-23 11:00:00. Конец пропуска: 2014-01-23 19:30:00. а — Модель SARIMAX с *Относительной влажностью* пропуск №3, б — Модель SARIMAX с *Солнечной радиацией* пропуск №3

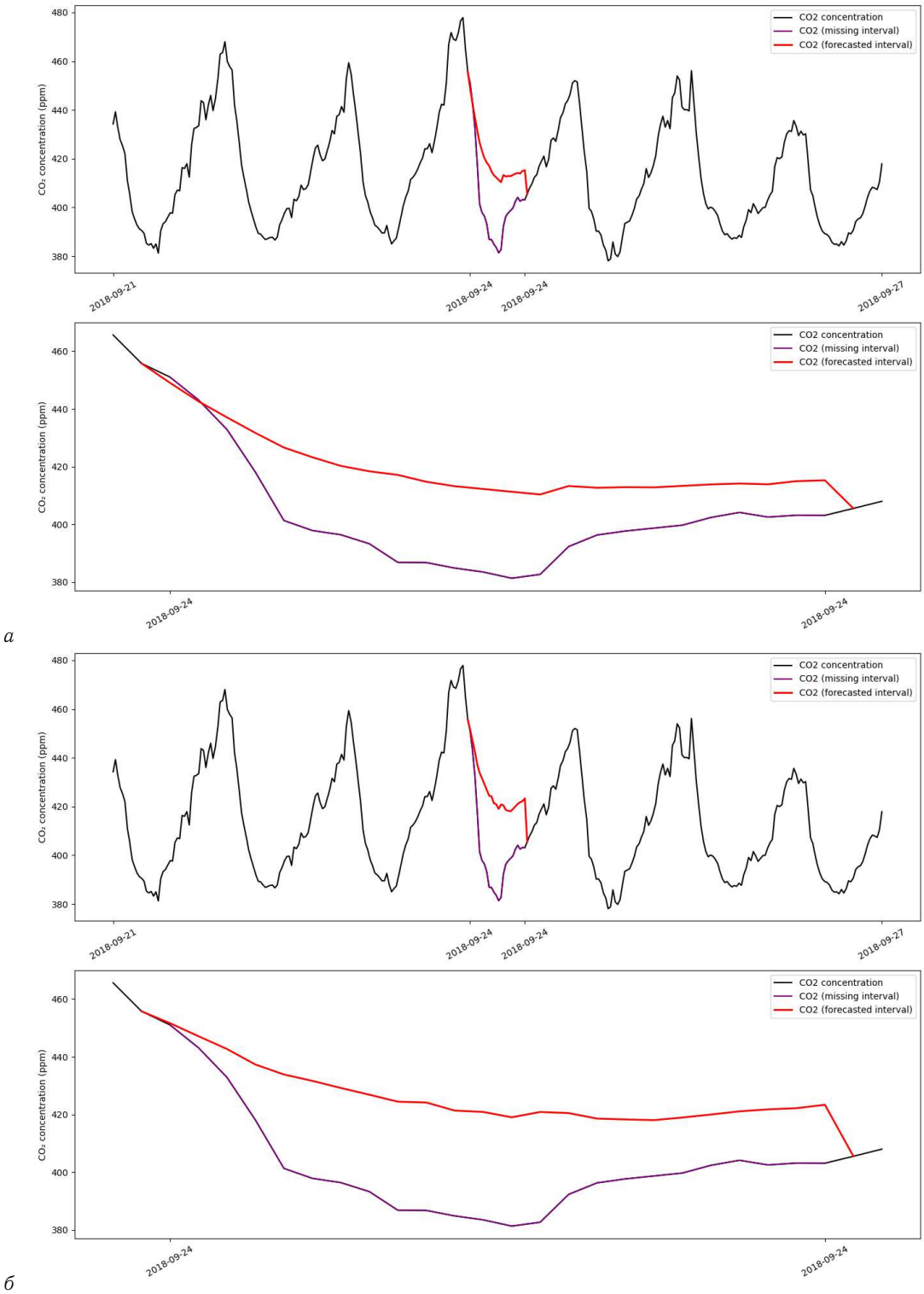


Рис. 5. Графики заполненных пропусков моделью SARIMAX с экзогенными переменными: *Относительная влажность*, *Солнечная радиация*. Начало пропуска: 2018-09-24 08:30:00. Конец пропуска: 2018-09-24 20:00:00. *а* — Модель SARIMAX с *Относительной влажностью*, *б* — Модель SARIMAX с *Солнечной радиацией*

М.В. Ломоносова. Физический Факультет. 1990.
[10] *Сердобольская М.Л.* Методы функционального анализа в задачах редукции. Московский Государственный Университет им. М.В. Ломоносова. Физический Факультет. 2014. https://cmp.phys.msu.ru/ru/study/special/mfazr

Using Lasso-regression to select significant exogenous features in time series forecasting

A. I. Baliuk

Department of Mathematical Modeling and Informatics, Faculty of Physics, Lomonosov Moscow State University. Moscow 119991, Russia
E-mail: baliuk.ai20@physics.msu.ru

It is proposed to perform preprocessing of multivariate time series and optimize the selection of exogenous variables using the ordinary least squares approach (OLS) with regularization, with the aim of further using the obtained results in the SARIMAX model. The analysis of weather and atmospheric data from the Tver region showed that selecting significant exogenous features using Lasso-regression helps minimize the forecast error and prevent model overfitting. The obtained results confirm that the correct selection of exogenous variables improves the quality of the predictive model.

PACS: 20.50.Sk
Keywords: time series analysis, time series forecasting, SARIMAX model, Lasso-regression, optimal choice of exogenous variables.
Received 16 June 2025.

Сведения об авторе Балюк Анфиса Игоревна — студент; e-mail: baliuk.ai20@physics.msu.ru.