**Восстановление пропусков данных в многомерных временных рядах с помощью методов машинного обучения**

***Алешновский В.С.*1**

1*аспирант*

*Московский государственный университет имени М.В.Ломоносова,*  
*физический факультет, Москва, Россия*  
*E–mail*: [*aleshnovskii.vs17@physics.msu.ru*](mailto:aleshnovskii.vs17@physics.msu.ru)

В настоящее время актуальным является вопрос эффективного контроля за процессами изменения климата, который порождает множество дополнительных задач, таких как сбор данных, их валидация и последующая обработка. Изучение динамики различных геофизических показателей, таких как концентрация СО2 и температура воздуха на различных высотах в разные моменты времени, представляется важным в этом контексте. Анализ таких многомерных временных рядов позволяет отслеживать изменения параметров в течение времени и выявлять как типичное, так и нетипичное поведение этих параметров [1] [2].

Однако, из-за недостатков в методах сбора, хранения и передачи данных, возникают проблемы с неполными данными по различным компонентам временных рядов. Это может привести к искажению результатов последующей обработки информации. Решение этой проблемы заключается в использовании методов математической статистики и машинного обучения для восстановления потерянных данных. Это позволяет улучшить качество анализа и повысить достоверность выводов, основанных на временных рядах геофизических показателей [3].

Исследованы способы восстановления многомерных временных рядов в сравнении с одномерным случаем с использованием моделей из семейства ARIMA [4]. Представлено описание и сравнительный анализ точности восстановления пропущенных данных алгоритмами ARIMA (3, 0, 2), ARIMAX (3, 0, 2) (рис. 1), с учетом тех или иных дополнительных компонент ряда, а также с использованием методов машинного обучения LSTM и GRU[5] (рис. 2).

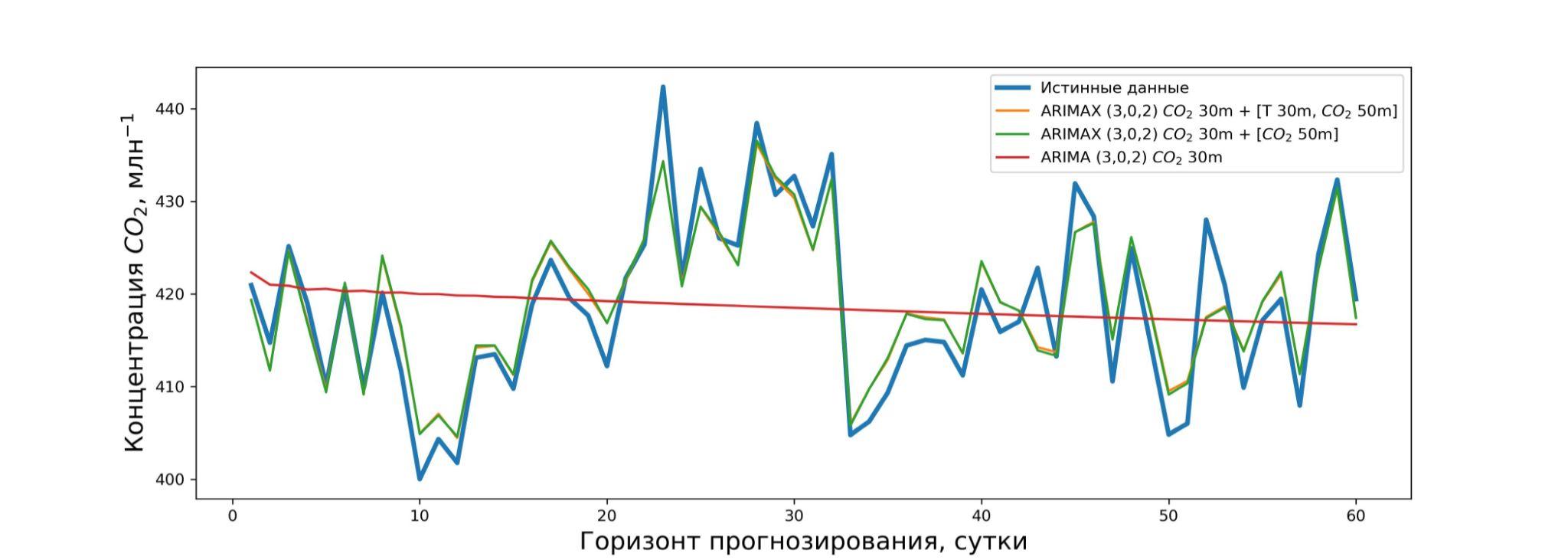


Рис. 1. График истинного ряда CO2 на высоте 30 метров, а также ряды, рассчитанные с использованием моделей из семейства ARIMA.

Результаты восстановления показателя концентрации углекислого газа на высоте 30 метров над уровнем земли с помощью различных моделей представлены на рис. 1 и 2.

При сравнении моделей (табл. 1) были использованы следующие критерии: среднеквадратическая ошибка (MSE), коэффициент детерминации (R²) и время восстановления (T).

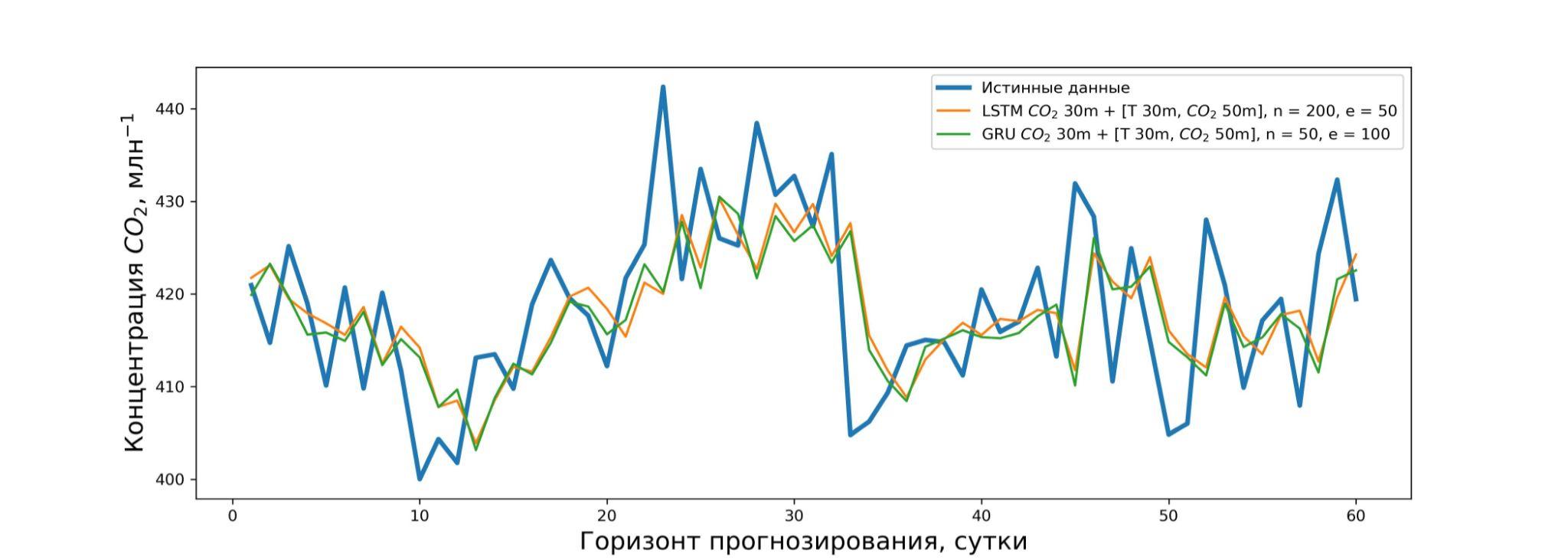


Рис. 2. График истинного ряда CO2 на высоте 30 метров, а также ряды, рассчитанные с использованием моделей машинного обучения.

| Модель | MSE | R2 | T, сек |
| --- | --- | --- | --- |
| ARIMAX (3,0,2) CO2 30m + [T 30m, CO2 50m] | **11.21** | **0.87** | 2.3 |
| ARIMAX (3,0,2) CO2 30m + [CO2 50m] | 11.41 | **0.87** | **1.13** |
| GRU CO2 30m + [T 30m, CO2 50m], n = 50, e = 100 | 70.63 | 0.18 | 372 |
| LSTM CO2 30m + [T 30m, CO2 50m], n = 200, e = 50 | 72.20 | 0.16 | 388 |
| ARIMA (3,0,2) CO2 30m | 90.03 | -0.05 | 2.9 |

Табл. 1. Анализ метрик различных моделей для восстановления ряда CO2 на высоте 30 метров.

Комплексный анализ всех моделей позволяет сделать вывод о том, что при наличии сильной корреляции между несколькими компонентами многомерного временного ряда наиболее эффективной моделью для восстановления данных является ARIMAX, при условии использования в качестве дополнительного параметра компоненты с высокой корреляцией. В случае отсутствия таких связей модель GRU, основанная на рекуррентных нейронных сетях, представляется хорошей альтернативой.

**Литература**

1. Kurbatova J., Tatarinov F., Molchanov A. et al. Partitioning of ecosystem respiration in a paludified shallow-peat spruce forest in the southern taiga of European Russia. // Environ. Res. Lett., 2013. T. 8. №. 4. C. 045028.
2. Box George E.P., Jenkins Gwilym M. Time series analysis, forecasting and control. Holden-day, Inc., 1976. С.300
3. Hocke K., Kämpfer N. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram // Atmospheric Chemistry and Physics., 2009. Т. 9. №. 12. С. 4197-4206.
4. Pirani M., Thakkar, P., Jivrani, P. et al. A comparative analysis of ARIMA, GRU, LSTM and BiLSTM on financial time series forecasting // 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2022. С. 1-6.
5. Atabay F. V., Pagkalinawan R. M., Pajarillo S. D. et al. Multivariate Time Series Forecasting using ARIMAX, SARIMAX, and RNN-based Deep Learning Models on Electricity Consumption // 2022 3rd International Informatics and Software Engineering Conference (IISEC). IEEE, 2022. С. 1-6.