Всероссийская конференция с международным участием «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов», посвященная памяти академика Ю.И. Шокина
г. Белокуриха, Алтайский край, Россия, 2025

НЕЙРОСЕТЕВАЯ МОДЕЛЬ ВРЕМЕННЫХ РЯДОВ ДЛЯ ПРОГНОЗИРОВАНИЯ КОНЦЕНТРАЦИИ ЗАГРЯЗНЯЮЩИХ ВЕЩЕСТВ В АТМОСФЕРЕ Г. КРАСНОЯРСКА

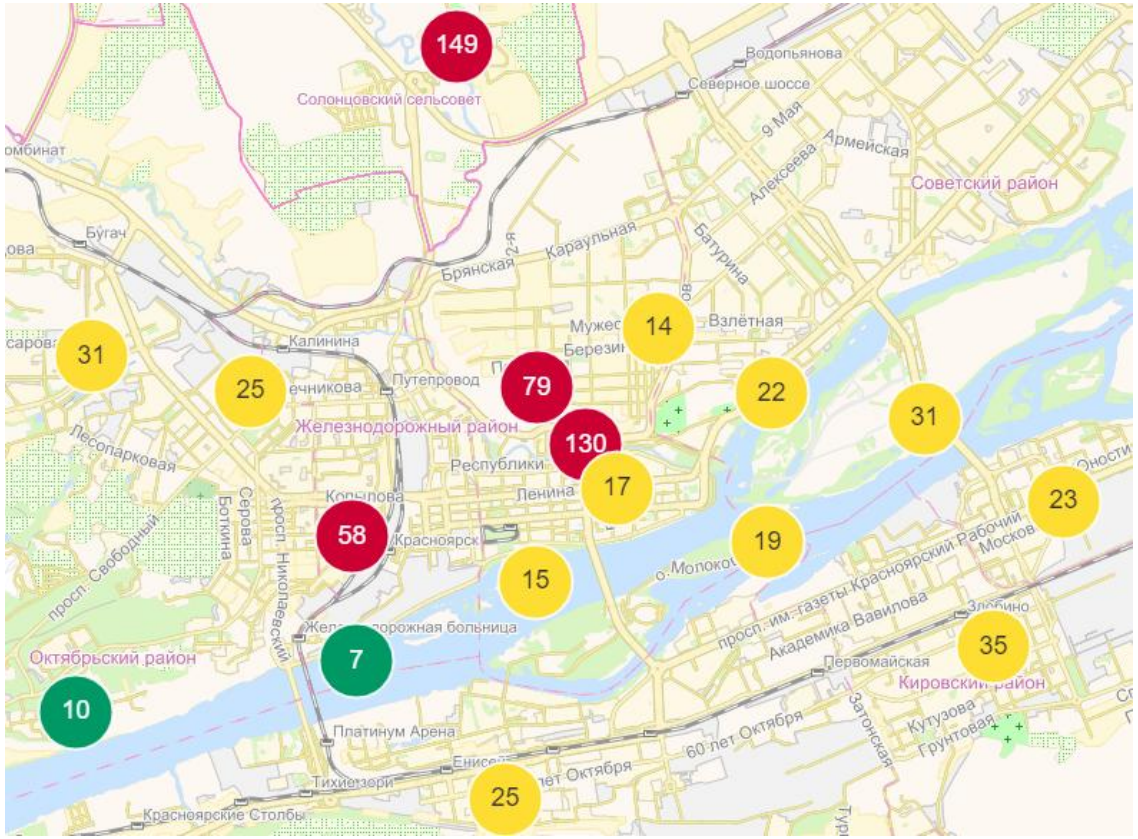
Володько Ольга Станиславовна, к.ф.-м.н.,
Лев Никита Андреевич[2], Полянчикова Дарья Витальевна[2]

[1]Институт Вычислительного Моделирования СО РАН, г. Красноярск

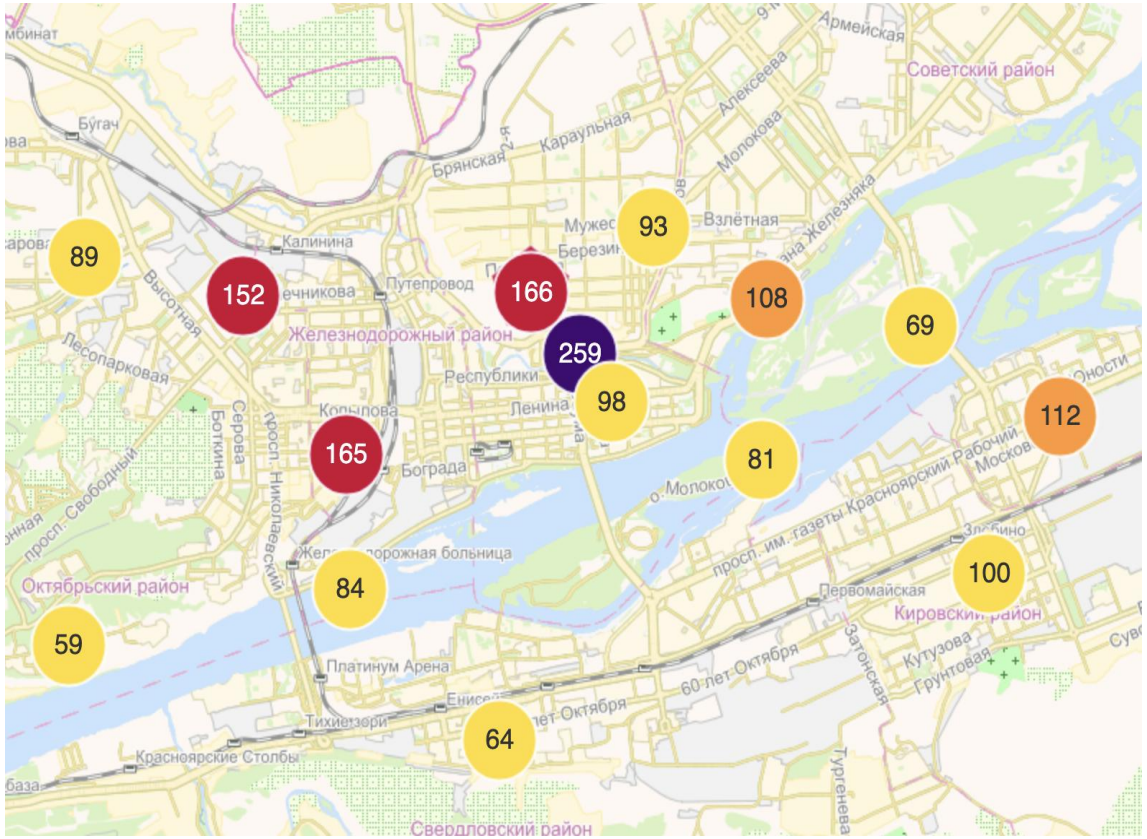
Федеральный исследовательский центр КНЦ СО РАН, г. Красноярск

[2]Сибирский Федеральный Университет, г. Красноярск

Актуальность



Система мониторинга воздуха г. Красноярска:
http://air.krasn.ru/ Скриншот от 5 апреля 2024 19:00



Индекс качества воздуха AQI в г. Красноярске
16 марта 2025 13:00

[1] Санитарные правила и нормы СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания". https://docs.cntd.ru/document/573500115 (дата обращения 30.06.2024)

Метеоусловия и PM

- Stafoggia, M. et al. Estimation of daily PM10 and PM2.5 concentrations in Italy 2013–2015, using a spatiotemporal land-use random-forest model // Environment international. – 2019. – № 124. – p. 170–179.

- Liu, X. et al. Air pollution in Germany: Spatio-temporal variations and their driving factors based on continuous data from 2008 to 2018 // Environmental Pollution. 2021. – № 276. – p. 116–732.

- Toro, R. et al. Exploring atmospheric stagnation during a severe particulate matter air pollution episode over complex terrain in Santiago // Environmental pollution. 2019. – № 244. – p. 705–714.

- Du, H. et al. Assessment of the effect of meteorological and emission variations on winter PM2.5 over the North China Plain in the three-year action plan against air pollution in 2018–2020 // Atmospheric Research. 2022. – № 280. – p. 106–395.

Модели машинного обучения для прогноза уровня PM

Модели регрессии

- Abdullah S., et al. Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia / S. Abdullah // Atmosphere. – 2020. – Vol. 11. – P. 289.

Модели прогнозирования временных рядов

ARIMA и ARIMAX

- Agarwal, P. A Study on Time Series Forecasting using Hybridization of Time Series Models and Neural Networks / P. Agarwal // Recent Advances in Computer Science and Communications. – 2020. – Vol. 13. – P. 827-832.

Ансамблевые модели

Случайный лес

- Hui Liu. Intelligent modeling strategies for forecasting air quality time series: A review / Liu Hui // Applied Soft Computing Journal – 2021. – Vol. 102. – P. 106957.

Градиентный бустинг

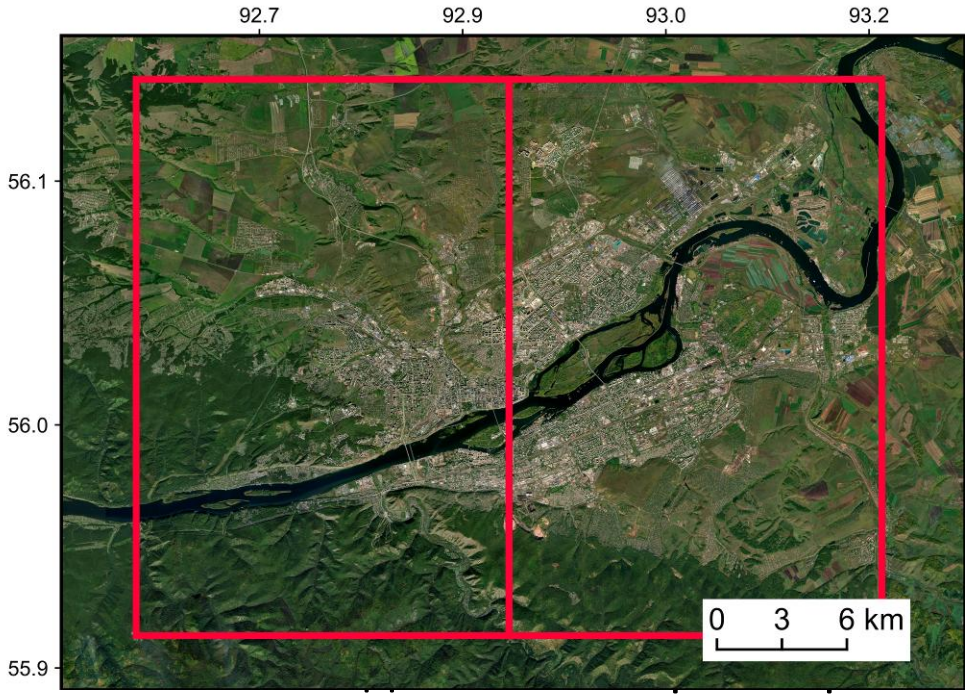
- Das R., Middya A. I., Roy S. High granular and short-term time series forecasting of PM2.5 air pollutant – a comparative review / R. Das, A. Middya, S. Roy // Artif. Intell. Rev. – 2022. – Vol. 55. – P. 1253-1287.

Нейросетевые модели прогнозирования временных рядов

LSTM

- Lin, M.D., Liu, P.Y., Huang, C.W. and Lin, Y.H. The application of strategy based on LSTM for the short-term prediction of PM2.5 in city. Science of The Total Environment, V.906, P. 167892 (2024).
- Das R., Middya A. I., Roy S., High granular and short-term time series forecasting of PM2.5 air pollutant – a comparative review. Artif. Intell. Rev, V.55, P. 1253-1287 (2022).
- Agarwal, P. A Study on Time Series Forecasting using Hybridization of Time Series Models and Neural Networks. Recent Advances in Computer Science and Communications, V.13, P. 827-832 (2020).

Район исследования



Красными рамками обозначены две ячейки регулярной сетки модели GFS

Данные метеоусловий получены:

1) Наземные станции мониторинга [2]

2) Метеоинформации модели реанализа National Centers for Environmental Prediction Global Forecast System (NCEP GFS) [3]

Период: 2019 - 2024 гг.

[2] Геопортал – данные оперативного мониторинга: сайт. – URL: http://sensor.krasn.ru/sc/ (дата обращения 14.04.2024)
[3] The Global Forecast System (GFS): сайт. – URL: https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php (дата обращения: 21.06.2023)

Данные для анализа

Пример данных NCEP GFS

Название	Расшифровка
DT1	Разность температур на изобарических слоях 1000 мбар – 925 мбар
DT2	Разность температур на изобарических слоях 925 мбар – 850 мбар
DT3	Разность температур на изобарических слоях 1000 мбар – 850 мбар
APTMP	Ощущаемая температура (K)
GUST	Скорость ветра (м/с)
POT	Потенциальная температура (K)
RH_m_2	Относительная влажность при 2м над землёй (%)
RH_mb_70	Относительная влажность на 700mb (%)
SOILW_g_0	Влажность почвы на 0-0.1 м над землёй (доля)
TMP_m_2	Температура на 2м над землёй (K)
TMP_m_80	Температура на 80м над землёй (K)
TMP_mb_750	Температура на 750mb (K)
TOZNE	Общий озон (DU)
UGRD_m_10	U-компонента ветра на высоте 10 м (м/с)
UGRD_m_80	U-компонента ветра на высоте 80 м (м/с)

Данные для анализа

Наземные станции мониторинга	Модель реанализа NCEP GFS
Температура воздуха	Разность температур на изобарических слоях 1000 мбар – 925 мбар
Влажность воздуха	Разность температур на изобарических слоях 925 мбар – 850 мбар
Атмосферное давление	Разность температур на изобарических слоях 1000 мбар – 850 мбар
Скорость ветра	
Направление ветра	
Концентрация частиц $PM_{2.5}$	

ФИЦ КНЦ: Николаевка, Овинный, Песчанка, Телевизорный и Ветлужанка.
Министерство: Кировский, Северный, Покровский, Черёмушки, Солнечный, Свердловский, Ветлужанка.

Модели регрессии

- Линейная регрессия

$$\tilde{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i$$

- Функция ошибки

$$Q(y, x) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}(x_i) - y_i)^2 \to min$$

- Lasso-регрессия (L1-регуляризация)

$$Q_{lasso}(y, x) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}(x_i) - y_i)^2 + \lambda |w_i|$$

- Ridge-регрессия (L2-регуляризация)

$$Q_{ridge}(y, x) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}(x_i) - y_i)^2 + \lambda {w_i}^2$$

- Полиномиальная регрессия степени P:

$$\tilde{y}(x) = w_0 + \sum_{k=1}^{P} \sum_{i=1}^{n} w_i x_i^k,$$

x_i – признаки,

w_i – параметры модели,

$\tilde{y}(x_i)$– прогнозируемые значения целевой переменной для признака x_i,

y_i – истинные значения целевой переменной,

n – количество используемых признаков,

λ – коэффициент регуляризации,

P – степень регрессии.

Модель ARIMAX (p, d, q)

- Состоит из трёх частей: авторегрессионной (AR) части, скользящего среднего (MA) и экзогенных переменных (X):

$$y_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{r} \beta_k x_{k_t} + \varepsilon_t,$$

y_t – значение временного ряда в момент времени t,
ε_t – шумовая компонента в момент времени t,
p – количество предыдущих значений временного ряда, используемых в модели,
q – количество предыдущих ошибок прогноза авторегрессии,
r – количество экзогенных признаков,
x_{k_t} – значение экзогенного признака k в момент времени t,
$a_i, \beta_j, \theta_j, c$ – параметры модели, которые необходимо оценить.

[4] Володько О. С., Лев Н. А. Методы прогнозирования временных рядов в задаче анализа уровня концентрации загрязняющих веществ в атмосфере г. Красноярска / О. С. Володько, Н. А. Лев // Безопасность и мониторинг природных и техногенных систем. – 2023. – Р. 205-208.

Модель случайного леса

$$\tilde{y}(X) = \frac{1}{k} * \left(b_1(X) + \cdots + b_n(X)\right),$$

$\tilde{y}(X)$ – прогноз модели на выборке X,

$b(X)$ – модель решающего дерева,

k – количество решающих деревьев в лесу.

Критерий ошибки в каждом узле дерева:

$$Q(X_m) = \frac{|X_l|}{|X_m|} H(X_l) + \frac{|X_r|}{|X_m|} H(X_r) \to min$$

$$H(X) = \frac{1}{|X|} \sum_{i \in X} \left(y_i - \bar{y}(X)\right)^2,$$

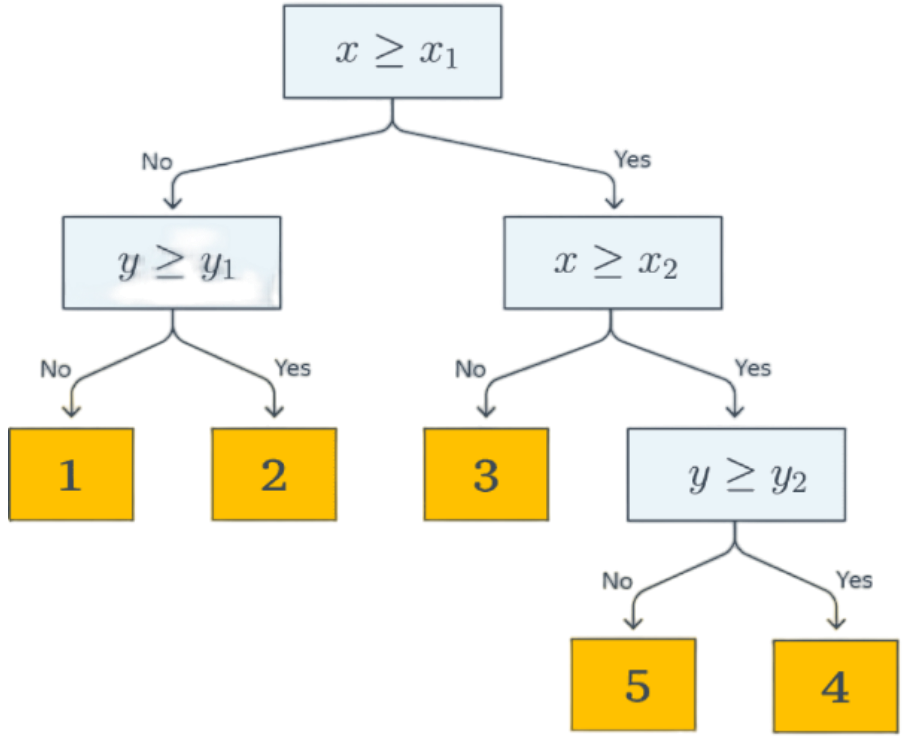
X – обучающая выборка,

$X_{m,\,l,\,r}$ – обучающая выборка в текущей (m), левой (l), правой (r) вершине,

y_i – значение целевой переменной,

$\bar{y}$ – среднее значение целевой переменной.

Решающее дерево



Решающие деревья — Учебник по машинному обучению: сайт. – URL: https://education.yandex.ru/handbook/ml/article/reshayushchiye-derevya (дата обращения: 08.07.2024).

Модель градиентного бустинга

$$a_N(X) = \sum_{n=1}^{N} b_n(X),$$

$b_n(X)$ – базовый алгоритм, X – выборка, N – количество базовых алгоритмов,

- Ошибка композиции моделей:

$$F(b) = \sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + b(x_i)) \to min_b,$$

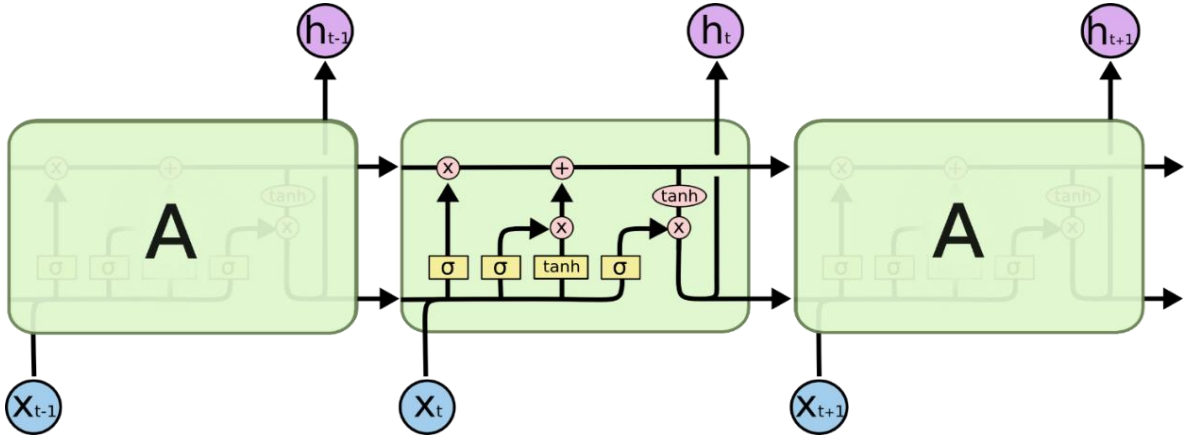
L – функция потерь (MSE), l – количество объектов обучающей выборки, b – произвольное дерево решений, y_i – значение целевой переменной.

b представим как вектор сдвигов $s = (s_1 \dots s_n)$, тогда

$$s = -\nabla F = \begin{pmatrix} -L'_{a_{N-1}(x_1)}(y_1, a_{N-1}(x_1)) \\ \dots \\ -L'_{a_{N-1}(x_l)}(y_l, a_{N-1}(x_l)) \end{pmatrix}$$

Оптимальный алгоритм $b_N(x) = argmin_b L(s, b(x))$.

Модель рекуррентной нейронной сети LSTM



LSTM из трёх блоков

- Вентиль забывания: $f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$
- Вентиль входного состояния: $i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$
- Вентиль выходного состояния: $o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$
- Внутреннее состояние: $c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c)$

σ – сигмоида, x_t – вход блока t, h_{t-1} – выход блока $t-1$, c_t – внутреннее состояние блока t, $\odot$ – поэлементное произведение, $W_f, W_i, W_o, W_c\ U_f, U_i, U_o, U_c, b_f, b_i, b_o, b_c$ – параметры модели, которые необходимо оценить.

Understanding LSTM Networks [Электронный ресурс]. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (дата обращения: 14.08.2025)

Метрики качества моделей

- *MAE* – средняя абсолютная ошибка

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- *MAPE* – средняя абсолютная процентная ошибка

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- *MSE* – средняя квадратичная ошибка

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- R^2 – коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

y_i – истинное значение,
$\hat{y}_i$ – прогнозируемое значение,
$\bar{y}$ – среднее значение в выборке,
n – количество объектов выборки.

Построение регрессионных моделей

Обучающая выборка разделялась на временные периоды в зависимости от значений концентрации $PM_{2.5}$:

- Зимний период: ноябрь – февраль.
- Весенний период: март, апрель.
- Летний период: май – июль.
- Осенний период: август – октябрь.

Вариации признаков обучающей выборки:

A: 8 главных компонент;

B: 8 главных компонент с отбором по корреляции с $PM_{2.5}$;

C: признаки, входящие в главные компоненты с наибольшими весами;

D: исходные данные метеоусловий NCEP GFS + данные станций мониторинга;

E: исходные данные метеоусловий NCEP GFS + данные станций мониторинга, коррелирующие с $PM_{2.5}$.

Volodko, O. Influences of meteorological conditions in PM 2.5 levels in Krasnoyarsk city atmosphere / O. Volodko, O. Yakubailik, T. Lapo, A. Dergunov // E3S Web of Conferences. – 2023 – Vol. 392. – P. 02022.

Сравнение регрессионных моделей на тестовой выборке

MSE на тестовой выборке

Период	A	B	C	D	E
Зимний	582.64	496,1	572.02	349,67	**310.24**
Весенний	35,01	26.9	37,7	27.25	**26,18**
Летний	35.86	37,83	38.23	35,1	**33.4**
Осенний	53,41	54.01	54,01	**45.02**	48.96

Лучшие модели для каждого временного периода:

- Зимний период – полиномиальная регрессия 2 степени.
- Весенний период – полиномиальная регрессия 2 степени.
- Летний период – Lasso регрессия.
- Осенний период – градиентный бустинг.

Сравнение качества моделей ARIMA и ARIMAX

Значения MAE на тестовой выборке

Временной период	ARIMA	ARIMAX
2019-03-12 — 2019-05-10	**4.89**	5.63
2019-05-10 — 2019-07-08	4.17	**3.89**
2019-08-28 — 2019-11-10	10.74	**8.16**
2019-11-25 — 2020-02-25	**5.52**	10.99
2020-03-15 — 2020-05-15	5.37	**3.46**
2020-05-01 — 2020-08-01	0.85	**0.73**
2020-08-01 — 2020-10-28	2.66	**2.56**
2020-12-15 — 2021-02-25	48.82	**13.60**
2021-04-01 — 2021-07-15	3.20	**2.49**
2021-08-15 — 2021-10-15	8.11	**7.22**
2022-03-01 — 2022-05-12	8.22	**7.22**
2022-04-01 — 2022-06-15	4.33	**1.97**
2022-07-16 — 2022-10-01	3.53	**2.79**
2022-12-01 — 2023-02-25	38.34	**29.67**
2023-02-25 — 2023-04-06	5.90	**5.12**

Построение моделей случайного леса

Гиперпараметры:

- Количество деревьев в лесу (n_estimators).
- Доля признаков в каждом узле (max_features).

Обучающая выборка разделялась двумя способами:

- По годам: 2019-2024, 2020-2024, 2021-2024, 2022-2024, 2019, 2020, 2021, 2022, 2023.
- По величине концентрации PM2.5 [5]:
  - зима (ноябрь - февраль);
  - весна (март, апрель);
  - лето (май - июль);
  - осень (август - октябрь).

[5] Volodko O., Yakubailik O., Lapo T., Dergunov A. Influences of meteorological conditions in PM 2.5 levels in Krasnoyarsk city atmosphere / O. Volodko, O. Yakubailik, T. Lapo, A. Dergunov // E3S Web of Conf.– 2023 – Vol. 392. – P. 02022.

Построение моделей градиентного бустинга

Гиперпараметры подбирались при помощи поиска по сетке:

- Шаг спуска (learning_rate).
- Количество деревьев (n_estimators).
- Глубина дерева (max_depth).
- Доля объектов в узле (subsample).

Оценка качества модели проводилась по кросс-валидации (5-fold).

Сравнение моделей градиентного бустинга и случайного леса

Значения MSE

Временной период	Бустинг	Случайный лес
Зима	**444.17**	450.19
Весна	**51.32**	63.01
Лето	**71.43**	74.19
Осень	**27.29**	27.94

Значения R^2

Временной период	Бустинг	Случайный лес
Зима	**0.79**	**0.77**
Весна	**0.70**	0.63
Лето	**0.16**	0.14
Осень	**0.64**	0.63

Сравнение моделей бустинга регрессий с ARIMAX

MSE на тестовой выборке

Сезон	Градиентный бустинг регрессий	ARIMAX
Зима	**251.46**	302.67
Весна	15.74	**14.66**
Лето	6.75	**1.24**
Осень	1.9	**0.87**

[5] Володько О. С., Лев Н. А. Методы прогнозирования временных рядов в задаче анализа уровня концентрации загрязняющих веществ в атмосфере г. Красноярска. Безопасность и мониторинг природных и техногенных систем, P. 205-208 (2023)

Построение моделей LSTM

Гиперпараметры подбирались при помощи поиска по сетке:

- Количество нейронов (units).
- Количество эпох (epochs).
- Доля прореживания нейронов (dropout).
- Количество шагов во времени в прошлое (n_steps).

Сравнение моделей LSTM по способу разделения выборки

Значения MSE

Временной период	Срез 1-2 месяца	Сезонная модель
Зима	366.82	**247.87**
Весна	103.12	**64.27**
Лето	35.35	**35.07**
Осень	29.89	**29.64**

Значения MAE

Временной период	Срез 1-2 месяца	Сезонная модель
Зима	14.04	**10.30**
Весна	6.81	**5.09**
Лето	4.01	**2.62**
Осень	3.46	**3.07**

Сравнение моделей LSTM и ARIMAX

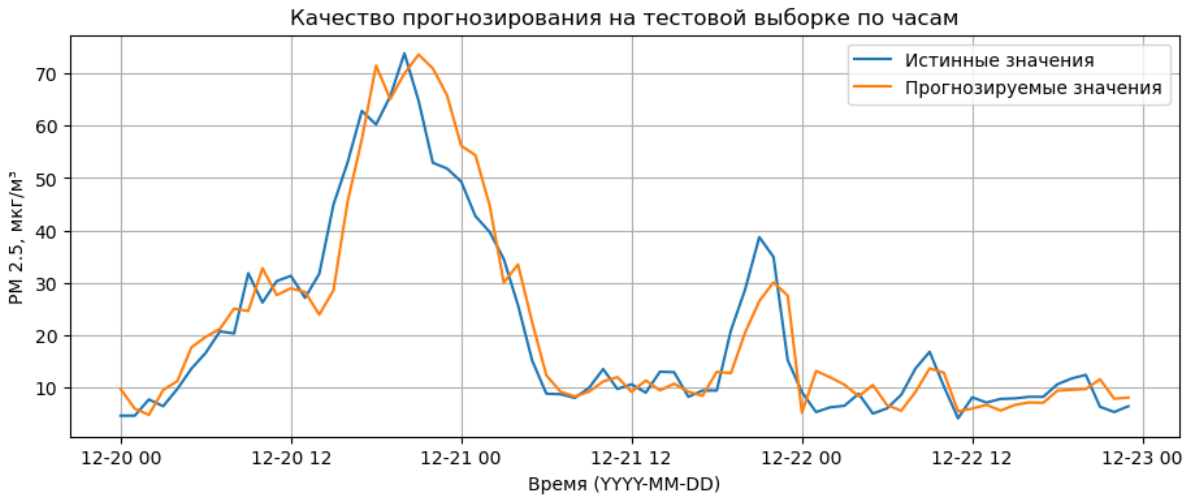
Значения MAE на тестовой выборке

Период прогнозирования		ARIMAX	LSTM
20 – 22 декабря	2022	8.34	**1.63**
20 – 22 июля	2022	2.53	**0.35**
12 – 14 августа	2022	3.09	**1.61**
10 – 12 марта	2023	5.43	**4.05**
10 – 12 апреля	2023	3.81	**0.72**
10 – 12 мая	2023	5.07	**0.77**
10 – 12 июля	2023	**0.29**	0.87
23 – 25 сентября	2023	4.77	**0.84**
23 – 25 октября	2023	5.26	**0.82**
10 – 12 января	2024	11.83	**3.11**
10 – 12 февраля	2024	13.21	**0.93**
17 – 19 марта	2024	7.01	**0.48**

Прогноз модель LSTM
Срез зимнего периода

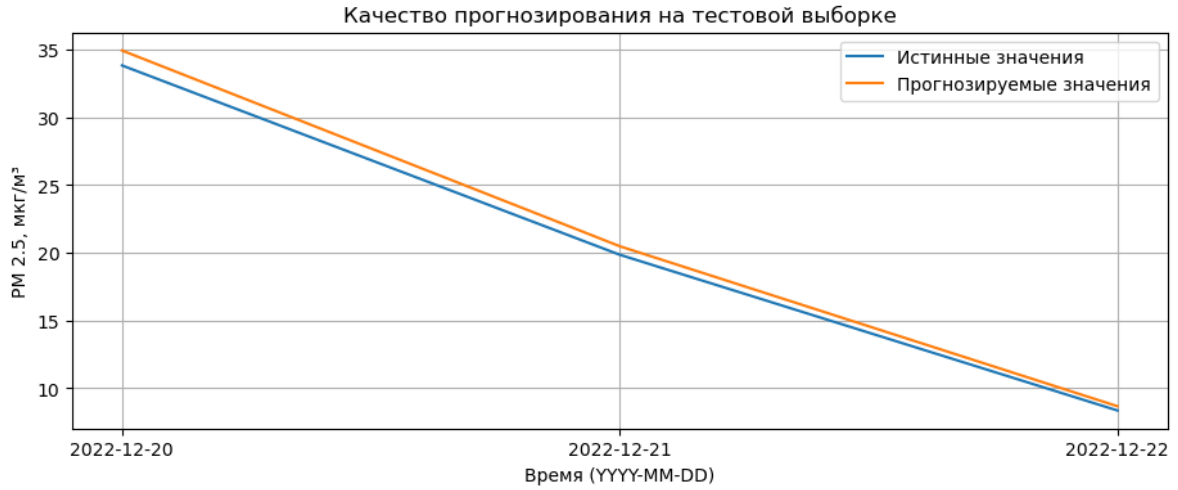
Период прогноза 3 дня: с 20 по 22 декабря 2022 г.

Почасовой прогноз



MAE = 4.44; R^2 = 0.89

Среднесуточный

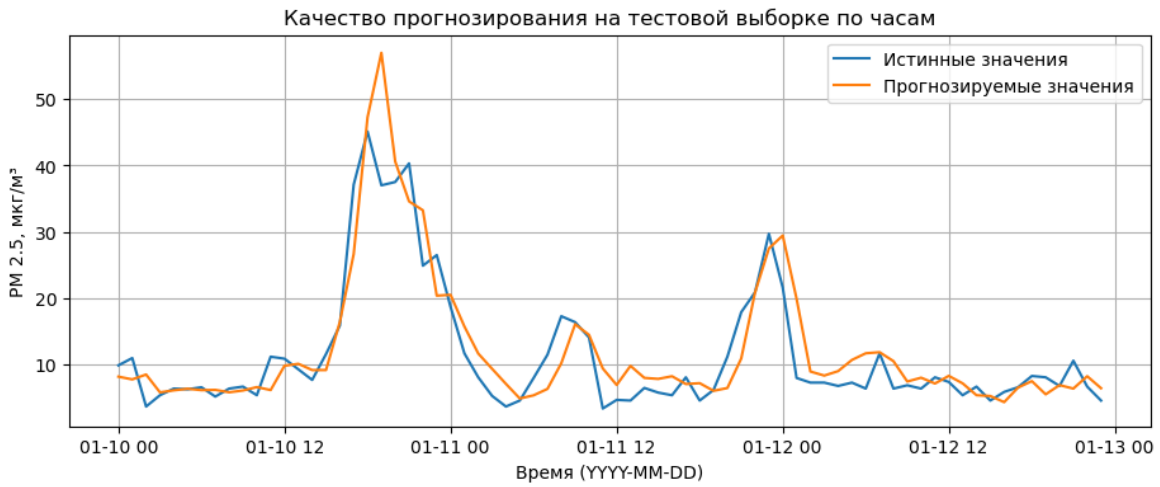


MAE = 0.68; R^2 = 0.99

Прогноз модель LSTM
Срез зимнего периода

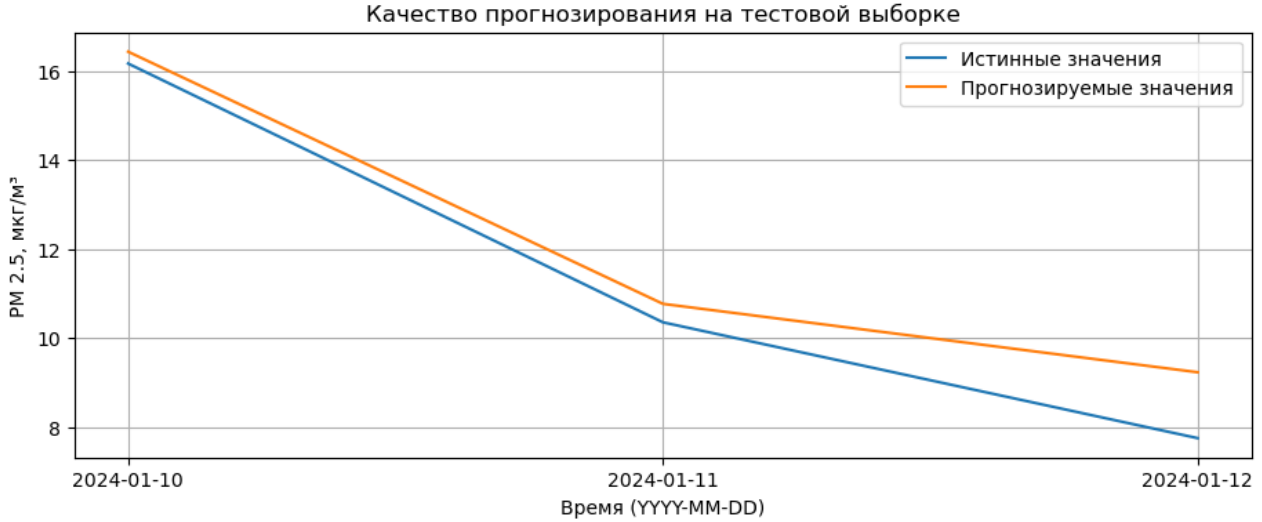
Период прогноза 3 дня: с 10 по 11 января 2024 г.

Почасовой прогноз



MAE = 2.92; R^2 = 0.79

Среднесуточный

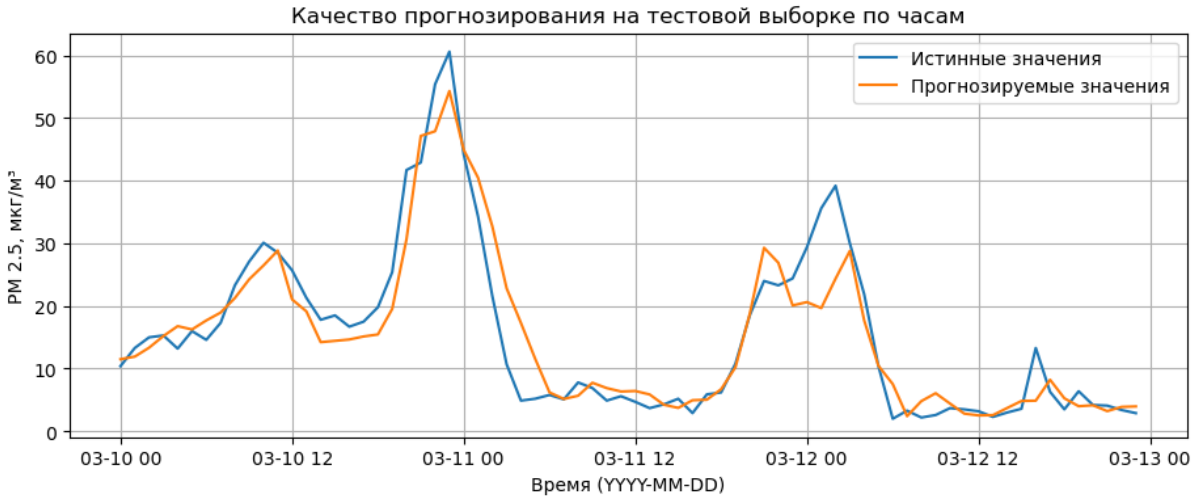


MAE = 0.72; R^2 = 0.93

Прогноз модель LSTM
Срез весеннего периода

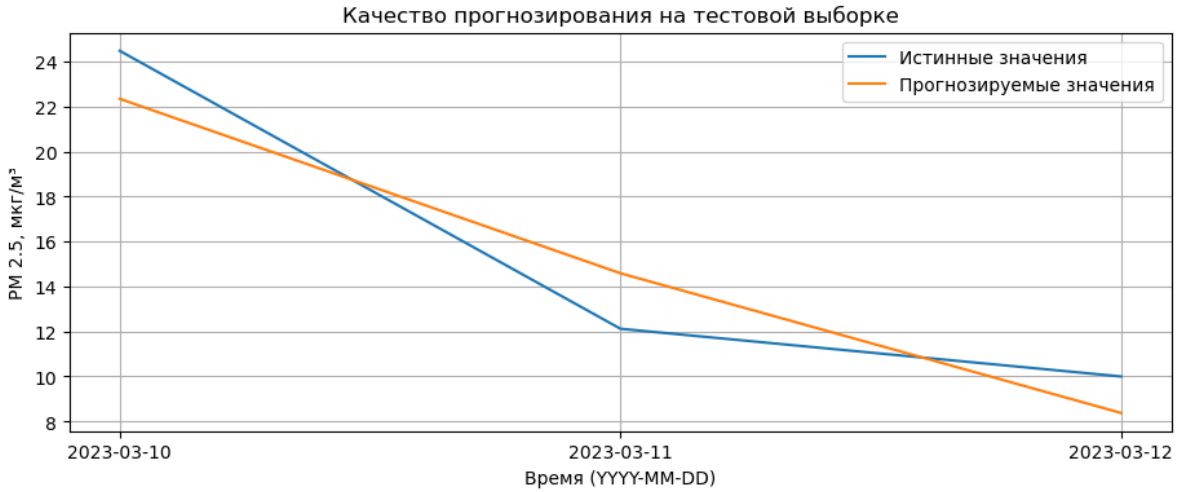
Период прогноза 3 дня: с 10 по 12 марта 2023 г.

Почасовой прогноз



MAE = 3.29; R^2 = 0.87

Среднесуточный



MAE = 2.07; R^2 = 0.89

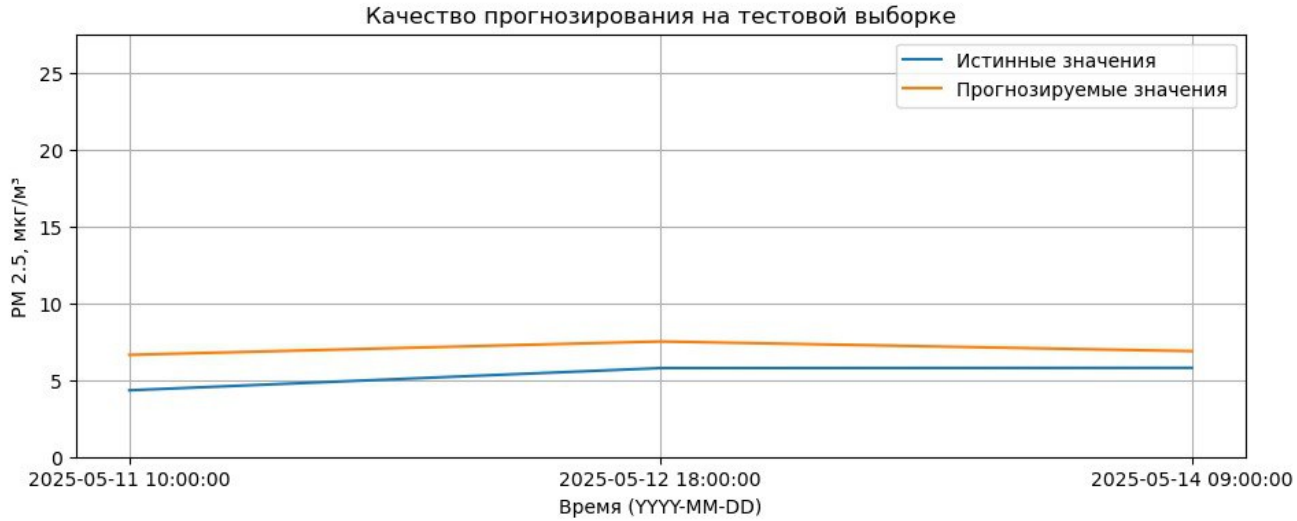
Прогноз модель LSTM
Срез весеннего периода

Период прогноза 3 дня: с 11 по 13 мая 2025 г.

Почасовой прогноз

MAE = 2.60; R^2 = 0.82

Среднесуточный



MAE = 1.48; R^2 = 0.90

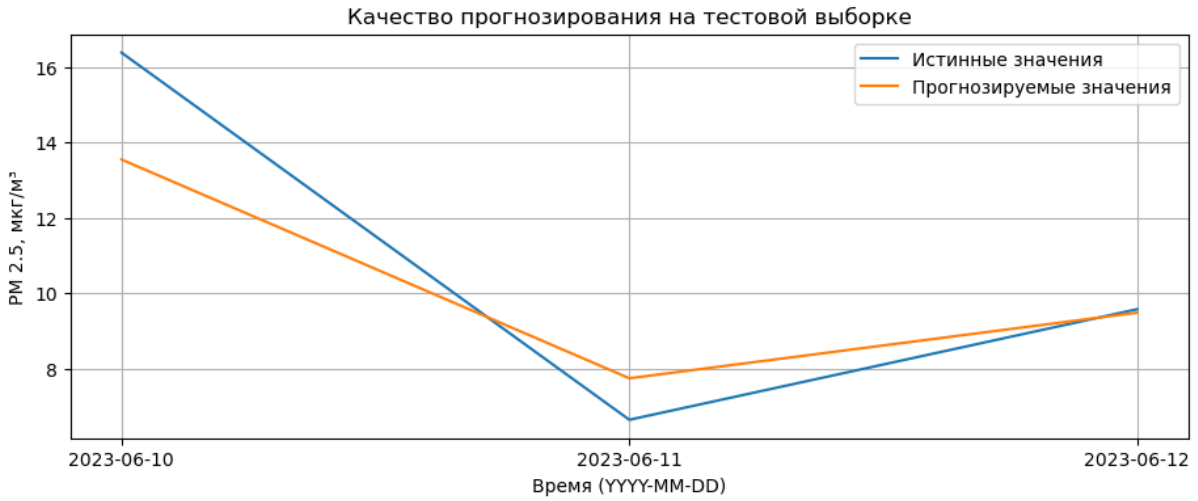
Прогноз модель LSTM
Срез летнего периода

Период прогноза 3 дня: с 10 по 12 июня 2023 г.

Почасовой прогноз

MAE = 2.26; R^2 = 0.74

Среднесуточный

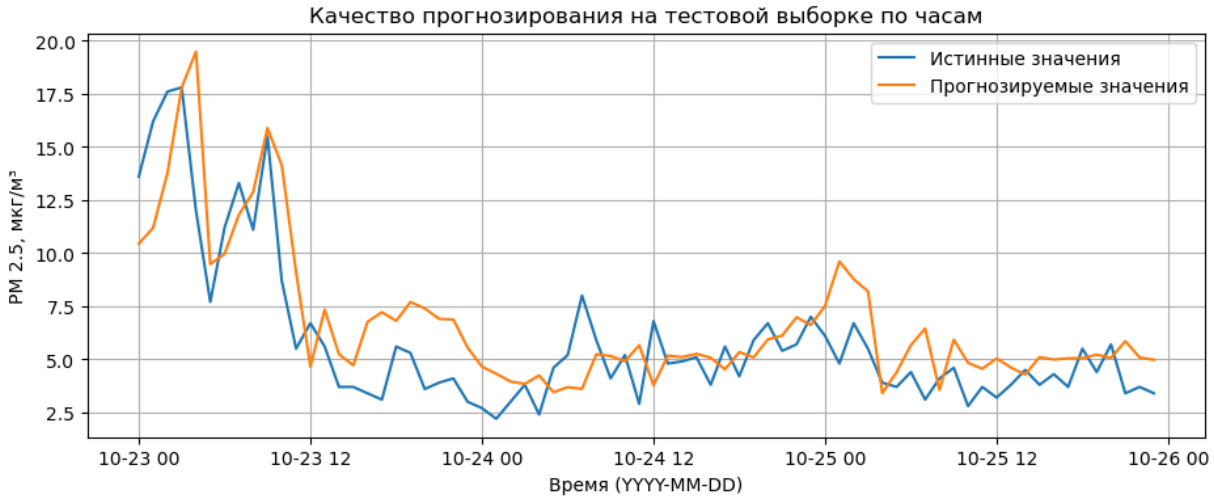


MAE = 1.34; R^2 = 0.81

Прогноз модель LSTM
Срез осеннего периода

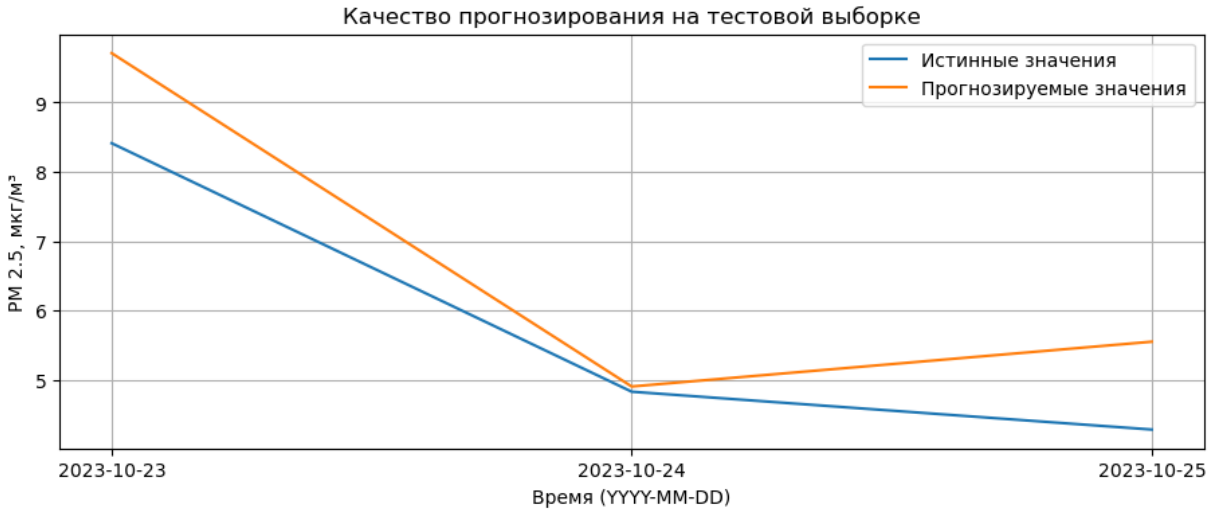
Период прогноза 3 дня: с 23 по 25 октября 2023 г.

Почасовой прогноз



MAE = 1.80; R^2 = 0.76

Среднесуточный



MAE = 0.88; R^2 = 0.83

Заключение

- Сравнение качества прогноза нейросетевой моделью LSTM с построенными ранее моделями машинного обучения, показало преимущество нейросетевых моделей LSTM по сравнению с лучшими моделями машинного обучения ARIMAX.
- Нейросетевую модель LSTM планируется использовать для прогноза величины концентрации PM 2.5 в г. Красноярске в рамках задач экологического мониторинга качества атмосферного воздуха.

Спасибо за внимание!

Володько Ольга Станиславовна
к.ф.-м.н., н.с. ИВМ СО РАН
E-mail: olga.pitalskaya@gmail.com