Всероссийская конференция с международным участием «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов»,
г. Бердск, Новосибирская область, 2023

МОДЕЛИ РЕГРЕССИИ ДЛЯ ПРОГНОЗА УРОВНЯ ЗАГРЯЗНЕНИЯ АТМОСФЕРНОГО ВОЗДУХА ГОРОДА КРАСНОЯРСКА

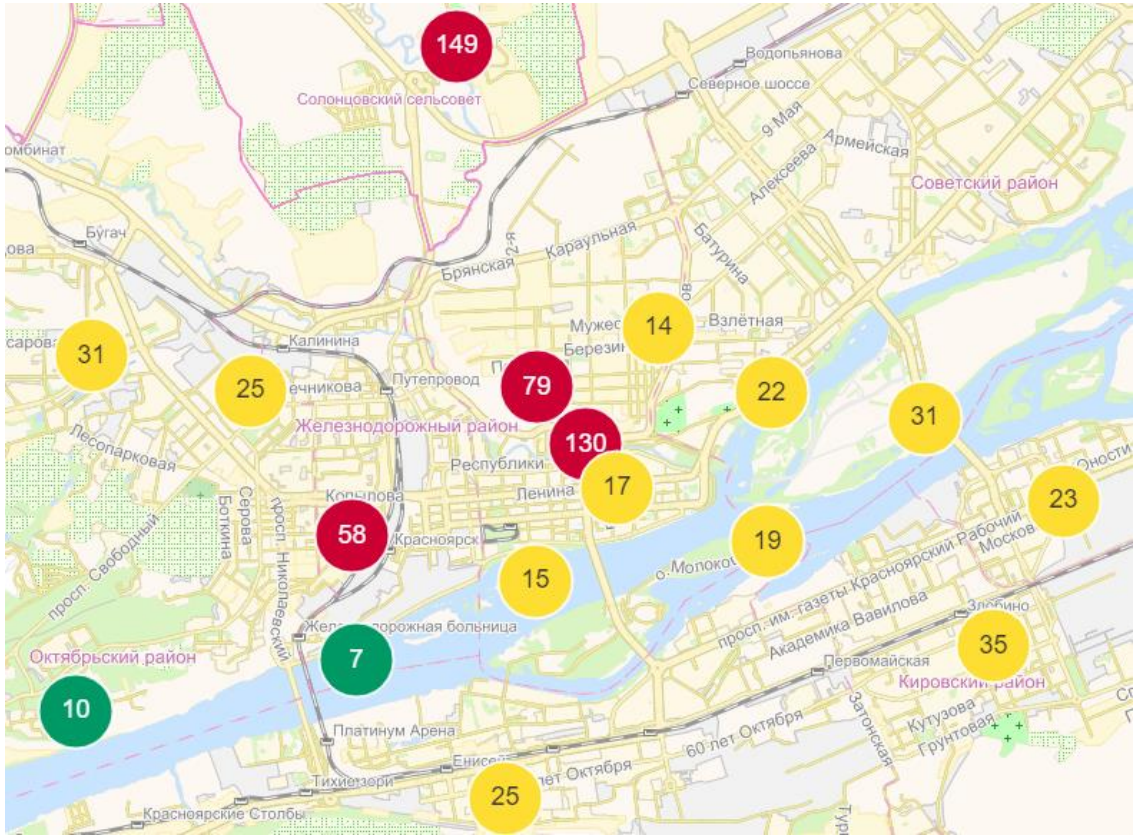
Володько Ольга Станиславовна[1], Буряк Никита Александрович[1],
Полянчикова Дарья Витальевна[2], Дергунов Александр Владимирович[3]

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

[2]Сибирский Федеральный Университет, г. Красноярск

[3]Федеральный исследовательский центр КНЦ СО РАН, г. Красноярск

Актуальность



Система мониторинга воздуха г. Красноярска:
http://air.krasn.ru/ Скриншот от 5 апреля 2023 19:00

[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 30.06.2023)
[2] WHO, 2022. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (дата обращения 30.06.2023)

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.

Атмосферные модели

GFS

- Kwok, L. et al. Developing a statistical based approach for predicting local air quality in complex terrain area // Atmospheric Pollution Research. 2017. – № 8(1). – p. 114–126.
- Peng, Z., et al. Impact of assimilating meteorological observations on source emissions estimate and chemical simulations // Geophysical Research Letters. 2020. – V. 47 (20). – p. e2020GL089030.
- Shin, U. et al. Predictability of PM2. 5 in Seoul based on atmospheric blocking forecasts using the NCEP global forecast system // Atmospheric Environment. 2021. – V. 246. – p. 118141.

WFR и WRF-Chem

- Zhong, M. et al. Sensitivity of projected PM2. 5-and O3-related health impacts to model inputs: A case study in mainland China // Environment international. 2019. – V. 123. – p. 256–264.
- Casallas, A., et al. Validation of PM 10 and PM 2.5 early alert in Bogotá, Colombia, through the modeling software WRF-CHEM // Environmental Science and Pollution Research. 2020. – V. 27 – p. 35930-35940.

Модели регрессии для прогноза уровня PM

Множественная регрессия

- Abdullah, S. et al. Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia // Atmosphere. 2020. V. 11, № 3. P. 289.
- Perez, P. et al. PM2. 5 forecasting in Coyhaique, the most polluted city in the Americas. // Urban Climate. 2020. V. 32. P. 100608.
- Kumar S., Mishra S., Singh S.K. A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere // Heliyon. 2020. V. 6, № 11. P. e05618.

Регрессия главных компонент

- Abdullah, S. et al. Evaluation for Long Term PM 10 Concentration Fore-casting using Multi Linear Regression (MLR) and Principal Component Regression (PCR) Models // Environment Asia. 2016. V. 9, № 2. P. 101–110.
- Alfiandy, S., Virgianto, R.H., Putri, A.S. Modeling of daily PM2. 5 concentration based on the princi-pal components regression in South and Central Jakarta // Journal of Physics: Conference Series. 2020. V. 1434, №. 1. P. 012012.

Район исследования

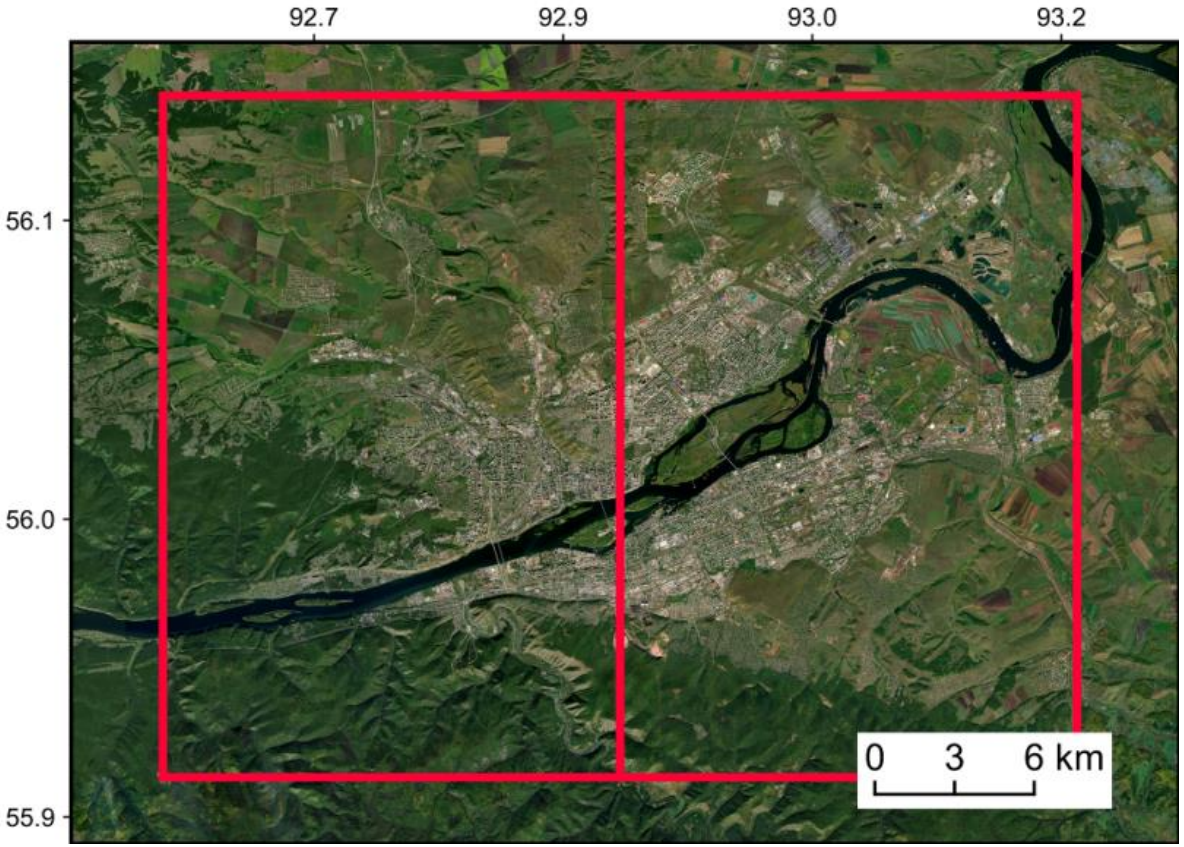


Рис. Район исследования – г. Красноярск. Красными рамками обозначены две ячейки регулярной сетки модели GFS

Данные получены с наземных станций мониторинга [4] и метеоинформации модели реанализа National Centers for Environmental Prediction Global Forecast System (NCEP GFS).

Данные по концентрации твердых взвешенных частиц PM2.5 получены с наземных станций мониторинга [5].

Для исследования были использованы данные с 2019 г. по 2022 г. с промежутком в 6 часов.

[3] Данные оперативного мониторинга // Геопортал ИВМ СО РАН. – 2021. – URL: http://sensor.krasn.ru/sc/
[4] The Global Forecast System (GFS) // National Centers for Environmental Prediction. 2019. – URL: https://www.emc.ncep.noaa.gov/emc/pages/numericalforecast_systems/gfs.ph

Модели регрессии главных компонент

Алгоритм:

1. Находим главные компоненты, которые независимы друг от друга, с помощью метода главных компонент.
2. Составляем уравнение множественной регрессии с использованием главных компонент.
3. Оцениваем параметры уравнения регрессии в соответствии с заранее определенными критериями.

Используемые модели:

- Линейная множественная регрессия;
- Линейная на экспоненцированных данных;
- Линейная с L1-регуляризацией (Lasso);
- Линейная (гребневая) регрессия с L2-регуляризацией (Ridge);
- Полиномиальные регрессии 2-ой и 3-ей степени.

[5] Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 2. М.: Финансы и статистика, 1986. 351 с.

Метрики качества модели

- MAE — средняя абсолютная ошибка

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

- MSE — средняя квадратичная ошибка

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

- R^2 — коэффициент детерминации

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

- $R^2_{\text{корр}}$ — скорректированный коэффициент детерминации

$$\boldsymbol{R^2_{\text{корр}} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)}}$$

y_i — истинное значение;
$\hat{y}_i$ — прогнозируемое значение;
$\bar{y}$ — среднее значение в выборке

n — количество наблюдений в выборке;
k — количество рассматриваемых признаков.

Данные для обучения и прогноза

- Исследования проводились на усреднённых по двум ячейкам сетки данным.

- Модели регрессии обучались на данных за 4 года с 2019-2022 гг. в разные периоды, выбранные на основании проведенного дисперсионного анализа [6] в зависимости от метеоусловий и среднего уровня концентрации $PM_{2.5}$.

[6] *Volodko O, Yakubailik O, Lapo T, Dergunov A.* Influences of meteorological conditions in PM2.5 levels in Krasnoyarsk city atmosphere // E3S Web of Conferences 2023. Vol. 392. P. 02022.

Оценка качества моделей на тестовой выборке

8 признаков (главных компонент) составляют 80 % дисперсии.
16 признаков (главных компонент) составляют 90 % дисперсии.

		Декабрь-февраль 2019-2022 гг.			Март, апрель, ноябрь 2019-2022 гг.		
Количество признаков	**Метрика**	**Линейная регрессия**	**Lasso**	**Ridge**	**Линейная регрессия**	**Lasso**	**Ridge**
8 признаков	***MSE***	526,18	528,69	526,18	22,76	23,76	**22,75**
	MAE	17,10	**16,93**	17,10	**3,51**	3,63	**3,51**
	R^2	0,46	0,45	0,46	**0,61**	0,59	**0,61**
	$R^2_{\text{корр}}$	**0,40**	0,39	**0,40**	**0.56**	0.54	**0.56**
16 признаков	***MSE***	495,28	510,20	**495,07**	23,61	23,57	23,60
	MAE	17,38	16,96	17,36	3,52	3,61	3,52
	R^2	**0,49**	0,47	**0,49**	0,60	0,60	0,60
	$R^2_{\text{корр}}$	0,35	0,33	0,35	0.48	0.48	0.48

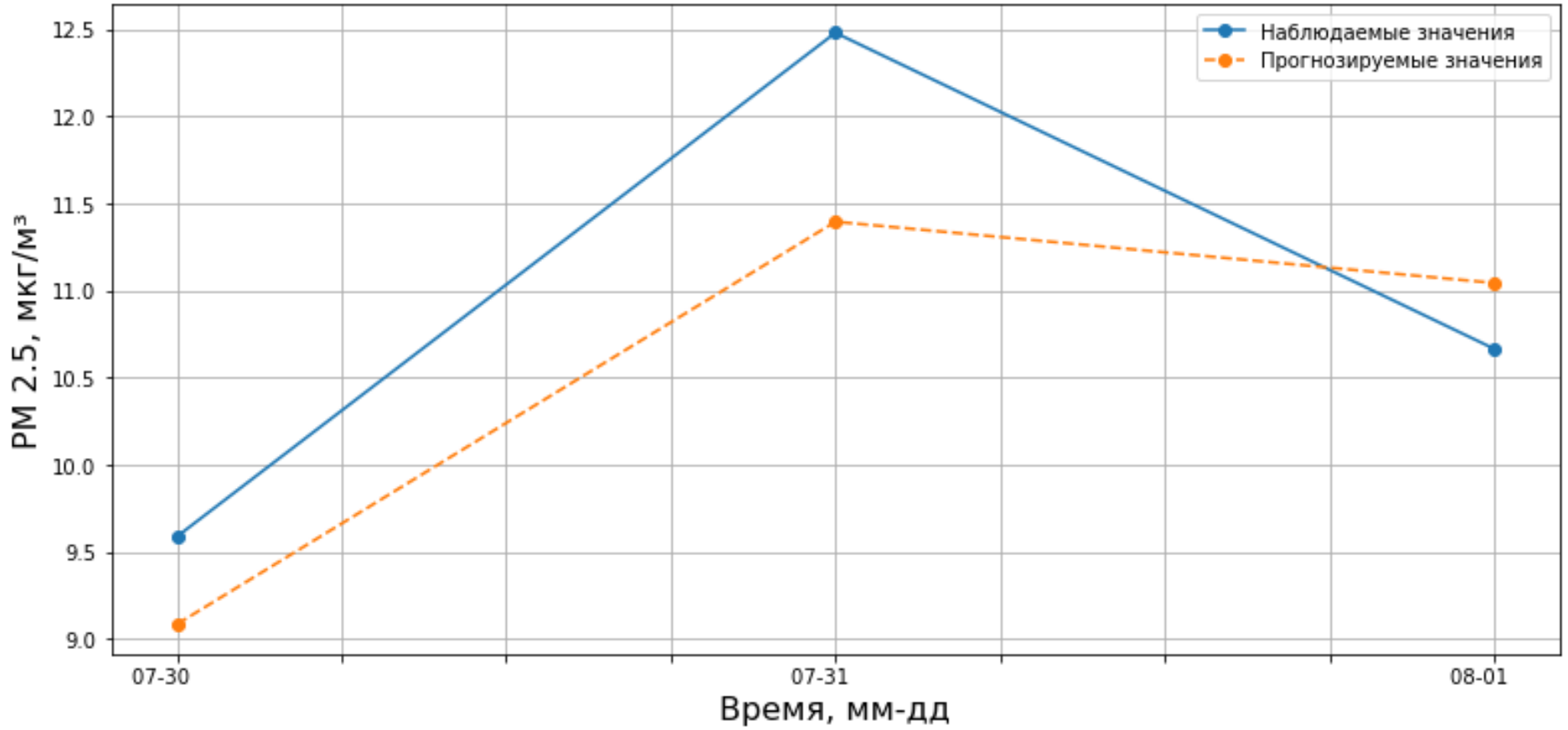
Оценка качества моделей на тестовой выборке

Количество признаков	Метрика	Май-июль 2019-2022 гг. Линейная регрессия	Lasso	Ridge	Август-октябрь 2019-2022 гг. Линейная регрессия	Lasso	Ridge
8 признаков	***MSE***	**17,09**	17,57	**17,09**	**14,23**	15,52	**14,23**
	MAE	**3,30**	3,31	**3,30**	**2,80**	2,80	**2,80**
	R^2	0,48	0,47	**0,49**	**0,43**	0,38	**0,43**
	$R^2_{\text{корр}}$	0,42	0,41	**0,43**	**0,35**	0,29	**0,35**
16 признаков	***MSE***	18,13	16,79	18,11	15,57	15,39	15,54
	MAE	3,25	3,13	3,24	2,73	2,77	2,73
	R^2	0,45	0,49	0,45	0,38	0,39	0,38
	$R^2_{\text{корр}}$	0,30	0,35	0,30	0,15	0,17	0,15

Прогноз моделью гребневой регрессии (Ridge), 8 главных компонент

Период прогноза 3 дня: с 30 июля по 1 августа 2023 г.

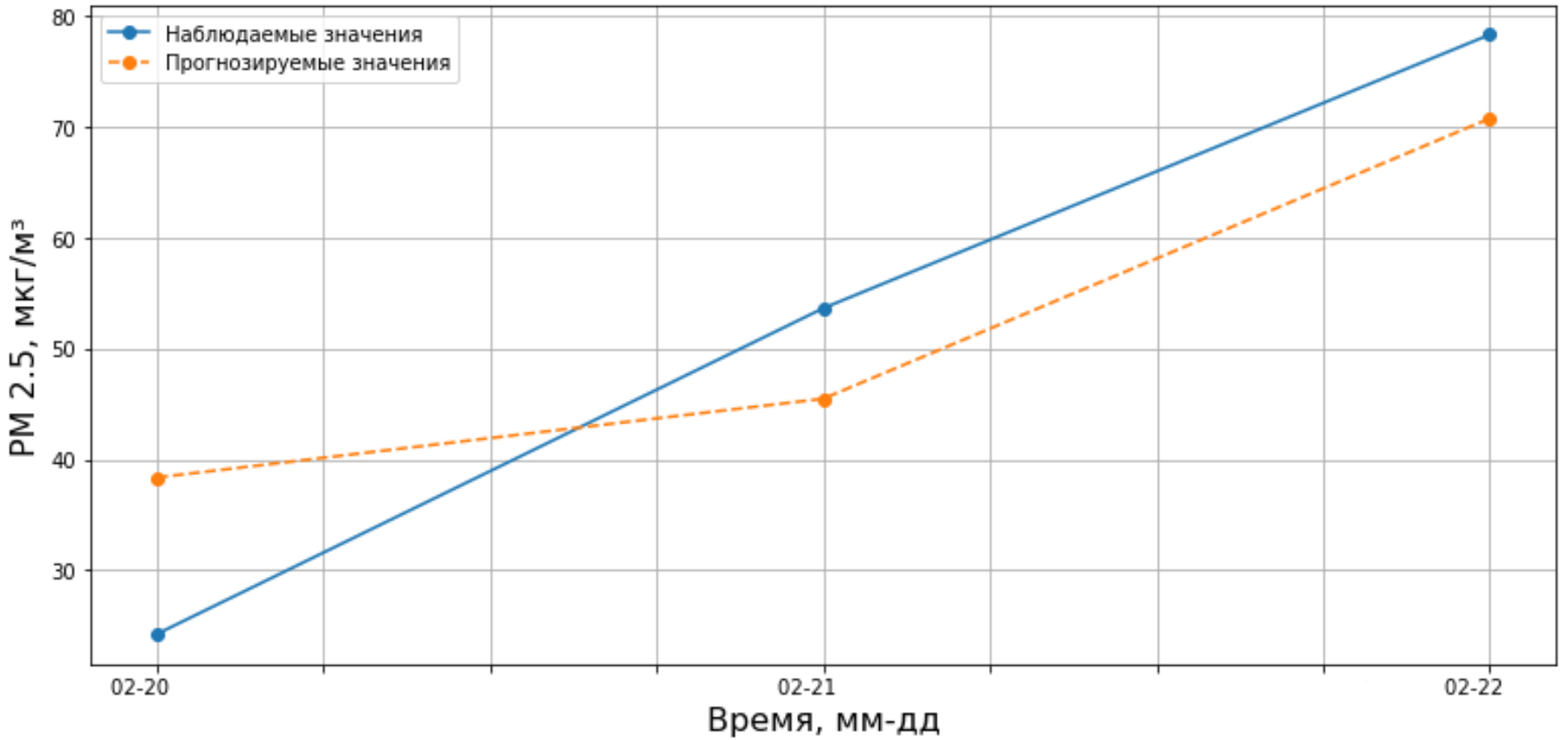
MSE = 0,65; MAE = 0,66; R^2 = 0,62



Прогноз моделью гребневой регрессии (Ridge), 8 главных компонент

Период прогноза 3 дня: с 20 по 22 февраля 2023 г.

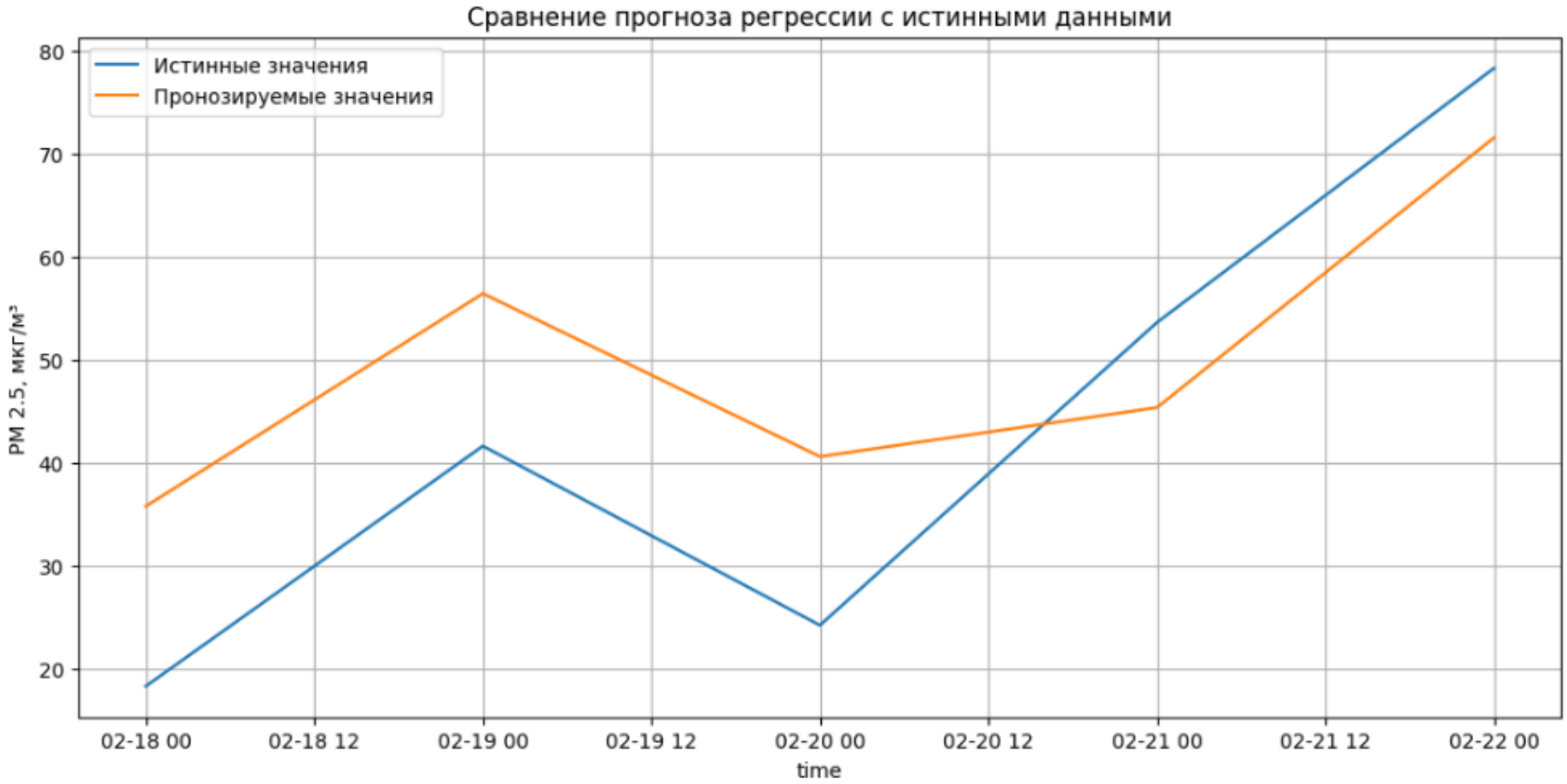
***MSE = 108,7; MAE* = 9,96; R^2 = 0,78**



Прогноз моделью гребневой регрессии (Ridge), 8 главных компонент

Период прогноза 5 дней: 18 по 22 февраля 2023 г.

***MSE = 181,71; MAE* = 12,75; R^2 = 0,61**



Заключение

Модели регрессии главных компонент были обучены на разных временных периодах с 2019-2022 гг. в зависимости от метеоусловий и уровня концентрации $PM_{2.5}$.

Модель регрессии главных компонент с **L2-регуляризацией (Ridge) для 8-ми признаков показала** лучшее качество и с достаточно хорошим уровнем точности может быть использована:

- для прогнозирования периодов повышенного уровня концентрации $PM_{2.5}$
- для построения гибридных моделей, которые в ряде исследований показывают наиболее высокую точность прогноза.

Исследование выполнено за счет средств гранта Российского научного фонда № 22-21-20117, Красноярского краевой фонда науки

Спасибо за внимание!