УДК 551.588 МРНТИ 28.23.27

ЭМУЛЯЦИЯ ПОКАЗАНИЙ ДАТЧИКОВ КАЧЕСТВА ВОЗДУХА В ГОРОДСКОЙ СРЕДЕ УМНОГО ГОРОДА

Р.И. Мухамедиев[1,2] *д.и.н.,* **А.Г. Терехов**[2] *к.т.н.,* **А.А. Оксененко**[1], **А.С. Еримбетова**[1,2*] *PhD., к.т.н.,* **Я.И. Кучин**[1,2], **А. Сымагулов**[1,2], **Д.Р. Құсайын**[1], **П. Рыстыгулов**[1]

[1]Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан
[2]Институт информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан
E-mail: aigerian8888@gmail.com

Загрязненность воздуха городской среды представляет собой серьезную угрозу здоровью людей. Для ее контроля используются как отдельные датчики, так и системы, позволяющие оценить концентрацию пылевых частиц PM1, PM2.5, PM10 и органических соединений. Однако, надежность системы датчиков не может быть 100 процентной. Время от времени те или иные датчики в распределенной системе выходят из строя. По этой причине весьма полезной является эмуляция их показаний на основании показаний оставшихся датчиков. В работе описан набор данных и предложена модель машинного обучения, которая на основе показаний работоспособных датчиков и погодных условий в местах сбора данных, моделирует показания датчика, вышедшего из строя. Оценена точность подобной эмуляции по отдельным видам загрязнений (коэффициент детерминации в пределах от 0.43 до 0.61).

Ключевые слова: качество воздуха, умный город, машинное обучение.

Поступила: 13.05.24
DOI: 10.54668/2789-6323-2024-114-3-87-99

ВВЕДЕНИЕ

Загрязнение окружающей среды явля-ются одной из серьезных проблем развития городов. Вследствие климатического особенностей, развития сельского хозяйства и промышленности, быстрого роста автомобильного транспорта, городов и недостаточного экологического контроля, ситуация в Казахстане одна из наиболее напряженных (Russell A. et al., 2018). Например, в 2022 году Казахстан занял 33-е место из 115 стран (чем выше место, тем выше загрязненность) по уровню загрязнения городов в мире (DKnews.kz, 2023). Устойчивый характер загрязненности связан как с географическими особенностями некоторых городов, находящихся в предгорных котловинах, как например, Алматы (рис.1), а также наличием промышленных производств, осуществляющих выбросы опасных веществ (например, Усть-Каменогорск).

Сказывается также недостаточное количество и низкое качество полигонов твердых бытовых отходов (ТБО) и возникающие вследствие этого стихийные свалки мусора.

Казахстан занимает второе место в мире по потреблению угля на душу населения в секторе домашних хозяйств (Kerimray A. et al., 2017). Парк автомобилей характеризуется высоким износом, а транспортные выбросы составляют почти треть всех выбросов в атмосферу (Әділет, 2023). Свой вклад вносит низкое качество топлива (Kazenergy, 2017). Основная часть производства электроэнергии и тепла (66 %) основана на сжигании угля (Kerimray A. et al., 2017, Karatayev M. et al.) при этом выделяется большое количество опасных загрязнителей воздуха. Совместно с выбросами автомобильного транспорта это делает воздушный бассейн Алматы (крупнейшего города Казахстана) наиболее загрязненным (Current Pollution Index, 2023, Kerimray A. et al., 2020). Как следствие наблюдается значительный рост легочных

заболеваний, почти вдвое превышающий средний уровень на постсоветском пространстве (Nugmanova D. et al., 2018). Снижение концентрации пылевых частиц является одним из путей решения этой серьезной социальной проблемы. Дополнительным средством является применение систем информирования населения. Для этого нужны системы контроля концентрации пыли в воздухе.

В Казахстане имеются системы, информирующие жителей городов о состоянии загрязненности воздушной среды, например, портал airkaz.kz. (рисунок 2).

Рис. 1. *Иллюстрация котловинного воздушного загрязнения*

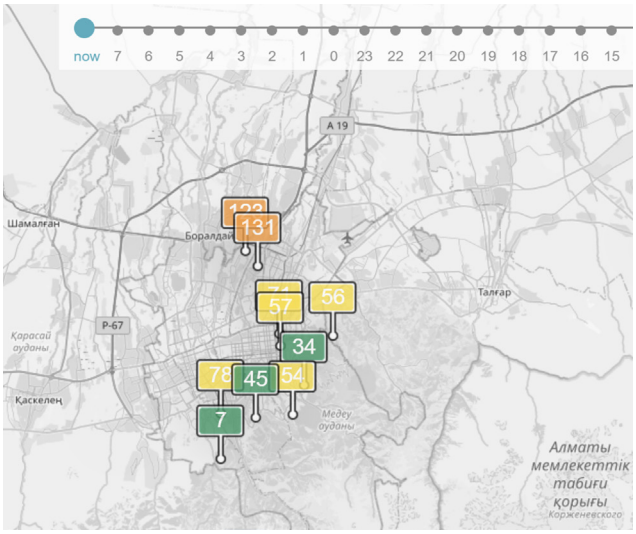

Рис. 2. *Портал airkaz.kz информирующий о показаниях датчиков PM2.5 в городах Казахстана*

Однако, как распределенная техническая система, портал обладает ограниченными возможностями предоставления данных вследствие естественного износа датчиков и возможных проблем связи. Для частичного решения данной проблемы в настоящей работе исследуется возможность эмуляции показаний датчиков качества воздушной среды на основе показаний ветровой модели потала Yandex и показаний датчиков, оставшихся работоспособными. Кроме этого, используя модель машинного обучения в работе исследуется влияние отдельных показателей погоды на результаты предсказаний.

Работа включает:

- Раздел Материалы и методы, который описывает данные, процесс их получения и применяемую модель машинного обучения.
- Раздел Результаты, где приводятся основные результаты многочисленных вычислительных экспериментов.
- Раздел Обсуждение результатов, где обобщаются и обсуждаются основные результаты.
- Заключение, в котором подводятся итоги, перечисляются ограничения текущего этапа исследования и перспективы дальнейших работ.

МАТЕРИАЛЫ И МЕТОДЫ

Исследование основано на применение регрессионных моделей машинного обучения. Для реализации метода был сформирован набор данных, который включает показатели качества воздуха и погодные данные собираемые несколько раз в день в трех точках города Алматы с января по март 2024 года. Упомянутые точки сбора данных нумеровались цифрами 0...1...6 (Рисунок 3). Расстояние между точками 1 и 6 составляет 2.7 км, тогда как расстояние между 0 и 1 около 7 км.

В качестве прибора, оценивающего воздушные загрязнения, применялся мобильный комплект датчиков загрязненности воздуха Atmotube Pro (рисунок 4). Прибор позволяет измерять концентрацию пылевых частиц (PM1, PM2.5, PM10) и органических соединений (VOCs), а также давление, влажность воздуха и температуру.

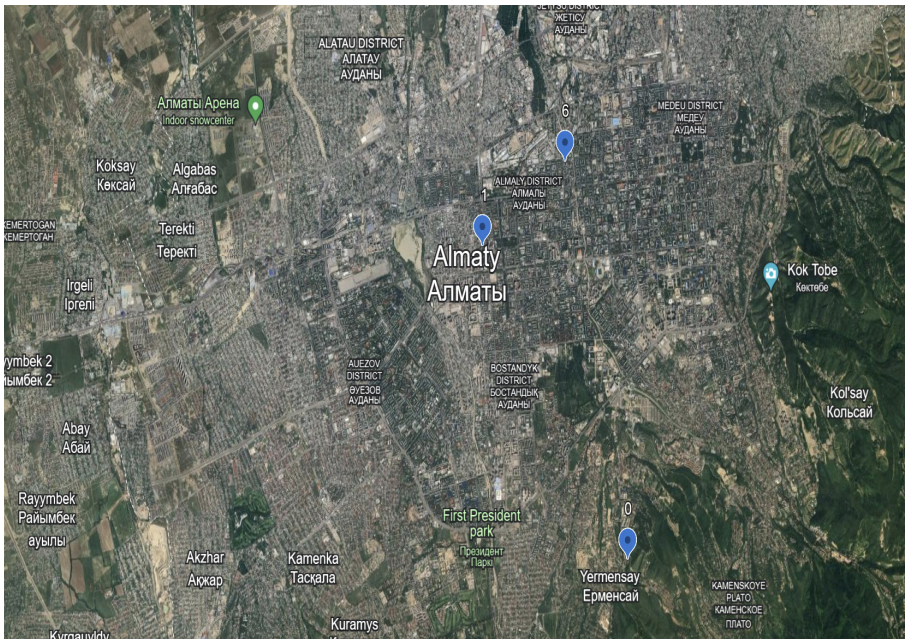

Рис. 3. *Места сбора данных на карте Алматы отмечены цифрами 0...1...6*

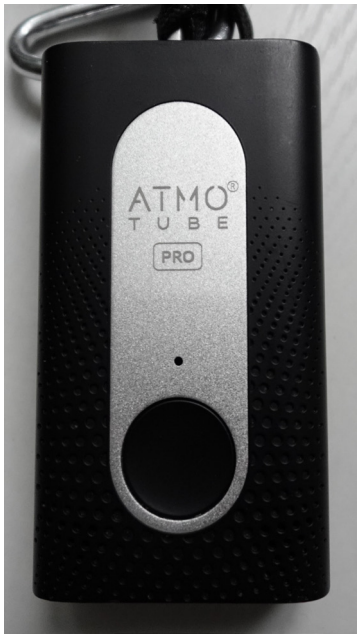

Рис. 4. *Мобильный прибор измерения качества воздуха*

В качестве источника погодных данных (направление и скорость ветра, влажность и температура) использовался портал *https://yandex.kz/pogoda*. Данные в каждой точке измерения собирались в таблицы, фрагмент которой показан в табл. 1.

Таблица 1

Фрагмент набора данных на одной из трех точек сбора данных

year	month	day	time	AQS	PM1	PM2.5	PM10	VOCs	bar	wind_dir	wind_speed (m/sec)	temp	humidity(%)
2024	1	7	21	0	206	232	242	0.64	694	4	1	2	81
2024	1	8	0	0	101	113	119	0.13	695	2	1	1	75
2024	1	8	9	78	12.3	14.4	15	0.36	698	7	0	4	88
2024	1	8	12	69	23.9	27.9	31.5	0.45	699	6	1.7	6	78
2024	1	8	15	72	21.7	25.3	27.8	0.35	698	2	1.9	5	83
2024	1	8	18	55	41.6	46.3	46.2	0.72	698	6	1.6	3	88
2024	1	8	21	0	140	158	169	1.14	698	6	1.6	3	88
2024	1	9	0	50	47.2	52	54.2	0.25	697	5	0	2	90
2024	1	9	9	80	12.9	15.1	16	0.33	697	4	0	1	73
2024	1	9	12	21	91.1	100	104	0.46	697	0	1.6	5	64
2024	1	9	15	23	86.6	99.8	104	0.32	696	2	1.7	6	63

Набор данных каждого датчика включает ежедневные показания, записанные в 9...12...15...18 и 21 час. PM1, PM2.5, PM10 – концентрация пылевых частиц в мкг на метр кубический размером 1...2.5 и 10 мкм, соответственно. VOCs (volatile organic compound) – концентрация огранических соединений, AQS (air quality score) – интегрированные показатель качества воздуха, bar – давление воздуха. Кроме этого, погодные данные (портал Яндекс):

- wind_dir – направление ветра в румбах (0 – север, 8 – юг, 12 – запад);
- wind_speed – скорость ветра в м/сек;

- temp – температура в точке измерения;

- humidity – влажность воздуха в процентах.

Всего исходный набор данных содержит либо 208, либо 384 строки данных и 78 колонок. Меньшее количество строк соответствует синхронизированному по времени файлу, содержащему данные всех трех точек измерения. Данный файл можно получить по ссылке *https://www.dropbox.com/scl/fi/r25jfow8k3uuwlyqa0n8x/air_pollution_0_1_6.xlsx?rlkey=wej2da7x0be5wr3zu6e9q23dq&dl=0*. В зависимости от цели регрессионной модели целевым параметром выбирались PM1, PM2.5, PM10, VOCs и AQS на одном из трех датчиков. Данные целевого датчика исключались из набора данных. Погодные данные в точке установки целевого датчика оставались в составе набора данных.

Модель машинного обучения и оценка ее качества

В качестве регрессионной модели использован ensemble learning method based the gradient boosted trees algorithm LightGBM (Ke G. et al., 2017, Bentéjac C. et al., 2021) достоинством которого является высокая скорость обучения и хорошее качество получаемых результатов моделирования даже при принятых по умолчанию настройках гиперпараметров модели. LightGBM является алгоритмом ансамбля деревьев решений, который использует технику усиления (boosting), когда следующие деревья ансамбля обучаются с учетом градиента ошибки предыдущих деревьев. То есть следующее дерево настраивается так, что целевым значением является не целевое значение регрессионной модели (target value $-y^{(i)})_{i=1}^{m}$ ($y^{(i)}$ – целевое значение для i-го примера из m обучающих примеров), а антиградиент функции ошибки предыдущего набора деревьев $-L'(y^{(i)}, h_\theta(x^{(i)}))_{i=1}^{m}$. То есть при обучении каждого следующего дерева вместо традиционных пар $(x^{(i)},y^{(i)})$ используются пары $(x^{(i)},-L'(y^{(i)},h_\theta(x^{(i)}))$, где $h_\theta(x)$ функция гипотезы предыдущего набора деревьев. Аналогичный метод усиления используется и алгоритмом Extremely gradient boosting (Chen T. et al., 2016), который также достигает высоких результатов моделирования в самых разных областях практики.

По окончании обучения модели ее необходимо оценить. Оценка качества работы модели машинного обучения чаще всего строится на общепринятом наборе метрик. Как известно, для регрессионных моделей машинного обучения основными метриками качества являются те, которые перечислены в таблице 2 (Mukhamediev R. et al., 2023, Mukhamediev R. et al., 2022).

При этом, поскольку собранный набор данных относительно не велик (390 записей) и достаточно вариативен, для оценки результатов моделирования разделение данных на тренировочные и тестовые проводилось многократно. Другими словами, качество работы модели оценивалось с применением cross-validation of random permutations ShuffleSplit, так же как это сделано в работах авторов исследований (Mukhamediev R. et al., 2023, Mukhamediev R. et al., 2023). В этом случае происходит многократное разделение данных на тренировочную и тестовую часть случайным образом, обучение и тестирование модели машинного обучения с последующим усреднением результата. Для получения статистически значимых оценок качества подобное разделение проводилось 200 раз. По окончании вычислений программа формирует итоговый результат, пример которого показан в таблице 3.

В таблице, кроме названия регрессионной модели (LGBM) и основных показателей MAE, MSE, R^2, R, приводится дисперсия основных показателей качества (Var) и продолжительность выполнения расчетов (Duration).

Вычислительные эксперименты проведены на компьютере, оснащенном процессором Intel(R) Core(TM) i7-10750H, с 32 GB оперативной памяти и дискретной видеокартой Nvidia Quatro T2000.

Таблица 2

Метрики качества моделей регрессии

Показатель точности	abbreviation	Equation	Пояснение
Коэффициент детерминации	R^2 /r2_score	$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$ $SS_{res} = \sum_{i=1}^{m_k} (y^{(i)} - h^{(i)})^2$ *Stot=i=1mk(y(i)−y_)2,y_=1mki=1mky(i),*	где $y^{(i)}$ – фактическое значение; $h^{(i)}$ – расчетное значение (значение функции гипотезы) для i-го примера; $m_k \in m$ – часть обучающей выборки (множества размеченных объектов).
Средняя абсолютная ошибка	MAE	$MAE = \frac{\sum_{i=1}^{n}(y^{(i)} - h^{(i)})}{n}$	при оценке работы модели на тестовом множестве n это размер тестового множества
Средняя квадратичная ошибка	MSE	$MSE = \frac{\sum_{i=1}^{n}(y^{(i)} - h^{(i)})^2}{n}$	
Коэффициент линейной корреляции	R	$R(y,h) = \frac{\sum_{i=1}^{n}(h^{(i)} - \bar{h})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(h^{(i)} - \bar{h})^2}$, $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h^{(i)}$	

Таблица 3

Пример оценки качества работы регрессионной модели

Regressor name	MAE	MSE	R^2	R	Var MAE	Var MSE	Var R^2	Var R	Duration
LGBM	20.225	876.11	0.35	0.61	3.592	46346.73	0.008	0.004	13.2117

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результат оценки загрязненности в точке 0 по данным значений погоды в этой же точке и приведен в таблице 4.

Корреляция между целевым параметром и входными параметрами модели показана на рисунке 5.

Таблица 4

Качество работы модели оценки уровня PM2.5 по данным собранным в точке 0

Regressor name	MAE	MSE	R^2	R	Var MAE	Var MSE	Var R^2	Var R	Duration
LGBM	20.638	879.303	0.367	0.613	3.507	37821.57	0.009	0.004	10.8266

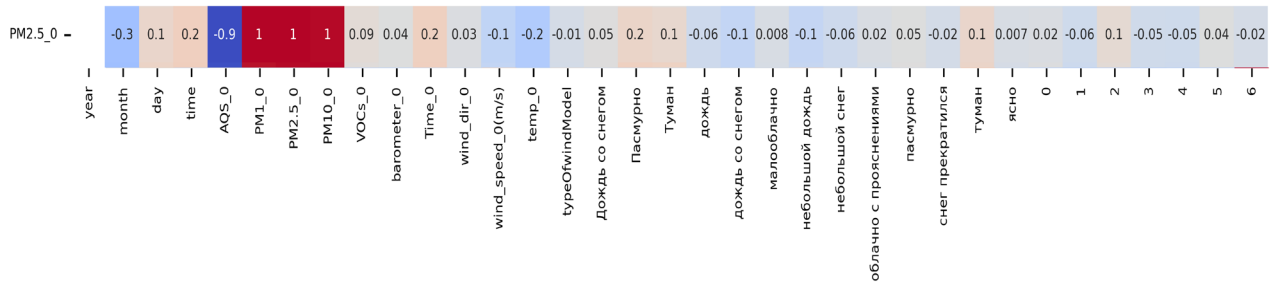

Рис. 5. *Корреляция между целевым параметром модели (PM2.5) и входными переменными*

Таблица корреляции показывает взаимосвязь между входными параметрами, однако она не показывает то, как эти входные параметры повлияли на выводы модели. Для этого можно воспользоваться моделью интерпретации SHapley Additive exPlanations (SHAP) (Lundberg S.M. et al., 2017).

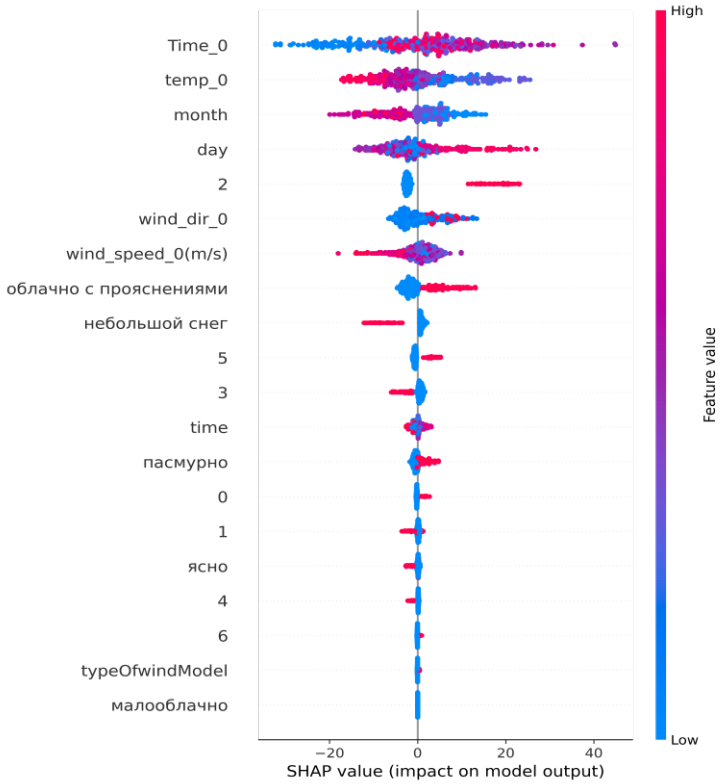

Рис. 6. *Влияние входных параметров на результаты работы регрессионной модели*

Дополнительным способом улучшения результатов моделирования является исключение некоторых малозначимых параметров, и параметров дающих противоречивые результаты. Для этого воспользуемся матрицей корреляции SHAP величин (рисунок 7). Попутно заметим, что параметр time на рисунке 6 во первых малозначим, во вторых противоречив, то есть его большие и малые значения влияют на выводы модели примерно одинаково.

Удалив незначимые параметры (пустые строки в матрице корреляции) и параметр time немного улучшим результаты до R^2=0.3705.

Результаты экспериментов по расчету показаний датчиков с применением данных другого датчика приведены в таблице 5. В ней показаны значения коэффициента детерминации для расчетных значений PM2.5 при использовании погодных данных в точке измерения, где датчик считается отказавшим, и полных данных (погода и загрязненность воздуха) другой точки.

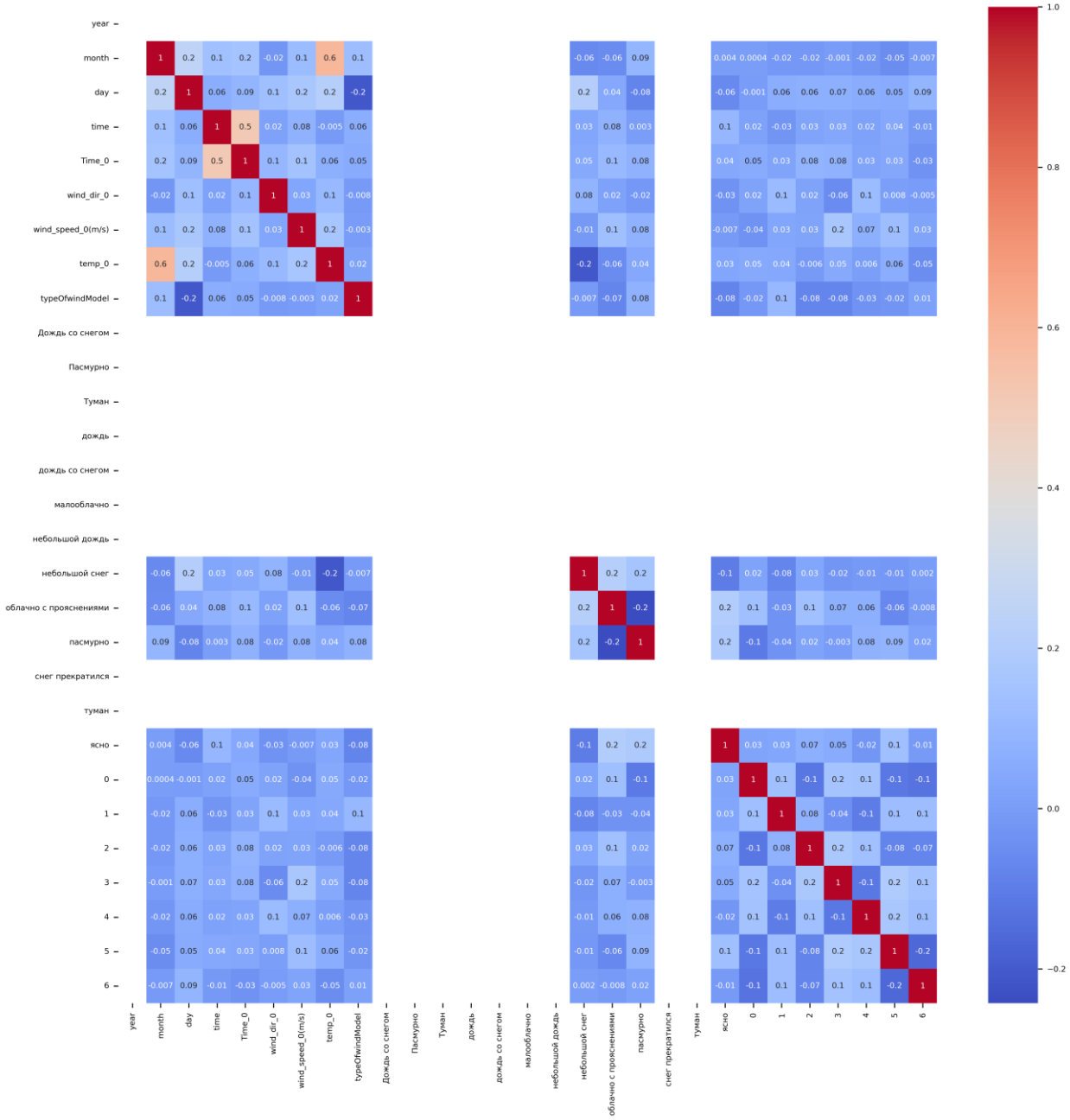

Рис. 7. *Матрица корреляции SHAP величин*

Таблица 5

Значения R^2 при расчете PM2.5 при различных сочетаниях входных данных

Расчет PM2.5 в указанной точке	Используемые данные		
	0	1	6
0	0.367	0.46	0.327
1	0.313	0.303	0.609
6	0.278	0.641	0.27

Значения в таблице следует интерпретировать следующим образом. Строки обозначенны номерами датчиков, в которых рассчитывается значение PM2.5 с помощью предобученной модели машинного обучения с дополнением полных данных (погода и загрязненность воздуха) в точке измерения, обозначенной номером столбца. Например, значение R^2 на пересечении строки 0 и столбца 0 означает результат работы

модели только при использовании данных точки 0. В свою очередь, значение на пересечении строки 0 и стобца 1 показывает результат, когда использовались погодные данные точки 0 и полные данные точки 1 для расчета значения PM2.5 в точке 0.

Полученные результаты подтверждают некоторые интуитивные ожидания, связанные с циркуляцией воздуха в предгорных районах. Например, интерпретируя результаты, показанные на рисунке 4, касающиеся точки 0, можно констатировать следующее:

1. Чем более раннее время получения данных (утро) тем ниже загрязненность PM2.5.

2. Чем выше температура воздуха, тем ниже загрязненность.

3. Чем больше номер месяца, тем ниже загрязненность.

4. Чем меньше направление ветра (восточнее) тем ниже загрязненность.

5. Когда погода «облачно с прояснениями» или «пасмурно» - загрязненность выше.

6. Чем ниже скорость ветра в точке измерения, тем выше загрязненность.

7. Небольшой снег снижает загрязненность.

Таблица корреляции (рисунок 3) показывает, что при отказе одного из датчиков измерения концентрации пыли его показания легко восстановить так как значения AQS, PM1, PM2.5, PM10 сильно коррелированы. Например, предсказание PM2.5 в точке 1 лишь по погодным данным дает значение R^2 =0.313. Однако, проведя вычислительный эксперимент при известных значениях PM10, видим, что точность расчета PM2.5 возрастает до R^2 =0.902. Другими словами, можно довольно точно восстанавливать значения запыленности одного размера частиц по данным другого размера пылевых частиц.

Как следует из таблицы 5 качество расчета загрязненности воздуха может значительно возрастать, особенно если в расчетах используются данные с близкорасположенного датчика. Например, если предсказывать показания в точке 0 с использованием данных 1, то значение R^2 увеличивается на 25 %. В тоже время для точки 6, с использованием данных 1, значение увеличивается более чем в 2 раза с 0.27 до 0.64. Можно сказать, что это ожидаемо, так как точка 6 примерно в два раза ближе к 1. Однако, если соответствующая пара подобрана не верно (например, 0 и 6) то качество расчета может даже уменьшаться.

ЗАКЛЮЧЕНИЕ

В настоящей работе мы рассматриваем распределенную систему сбора данных о качестве воздуха в городской среде. Мы обсудили задачу восстановления показателей отказавшего датчика по данным работоспособных датчиков в других точках измерения. Для решения задачи разработана модель машинного обучения с целью восстановления (расчета) показателей качества воздуха в случае отказа датчика. Модель обучена с использованием собственноручно собранного набора данных. Полученные результаты показывают, что возможно частичное восстановление показаний с ошибкой в пределах 20 мкм. При этом максимальные значения концентрации PM2.5 могут быть более 100, а минимальные менее 2. Расчеты подтверждают интуитивно ожидаемый результат – чем ближе работоспособный источник данных, тем точнее расчетные показатели загрязненности в точке, где датчик не работоспособен. Однако, в некоторых случаях добавление данных другой точки измерения может даже ухудшать результат (предсказание в точке 0 на основе 6 и 0). При этом если комплект датчиков отказал частично, то наличие значений хотя бы одного из показателей загрязненности позволяет весьма точно предсказывать остальные показатели.

Кроме этого, модель машинного обучения подтверждает наличие горнодолинной циркуляции воздуха в городе Алматы в период с января по март 2024 года и процессы очистки воздуха в утренние часы. Вместе с тем, анализируя полученные данные можно отметить, что в зимний период, во время морозов, данная циркуляция становится малоощутимой, что в значительной мере усугубляет проблему естественной вентиляции воздуха в городе Алматы в зимний период. В силу ограниченности временного цикла исследования и объема собранных данных можно отметить следующие огранничения проведенного исследования:

1. Ограниченность набора данных. Набор данных лишь за три месяца одного года и лишь в трех точках сбора данных.

2. Модель машинного обучения использовалась без оптимизации входных параметров и гиперпараметров.

Для исключения указанных недостатков в будущем полезно:

1. Расширить набор датчиков, используемых для измерения загрязненности воздуха.

2. Увеличить временной промежуток сбора данных.

3. Добавить погодные данные фиксируемые в различных частях города.

4. Проанализировать полезность методов, подбирающих параметры моделей машинного обучения (Scikit-learn, 2024, Scikit-optimize, 2014) и оптимизирующих список входных переменных модели (Raschka S., 2018).

5. Применить некоторые приемы генерации дополнительных входных параметров, например, плавающее окно данных.

Несмотря на отмеченные выше ограничения текущего этапа исследования, работа восполняет пробел, по количественной оценке, возможности восстановления показателей распределенной сети датчиков качества воздуха городской агломерации и текущих параметров горно-долинной циркуляции, оказывающей существеной влияние на перенос загрязненных воздушных масс.

Благодарности

Работа выполнена при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (грант №AP23488745 «Оперативная оценка засоленности почвы с применением маловысотных беспилотных летательных платформ», № BR21881908 «Комплекс экологического сопровождения городской агломерации» и № BR24992908 «Система поддержки агротехнических мероприятий в растениеводстве на базе комплекса средств мониторинга и методов искусственного интеллекта (Agroscope)».

Вспомогательные материалы

Набор данных можно скачать по ссылке https://www.dropbox.com/scl/fi/r25jfow8k3uuwlyqa0n8x/air_pollution_0_1_6.xlsx?rlkey=wej2da7x0be5wr3zu6e9q23dq&dl=0

СПИСОК ЛИТЕРАТУРЫ

1. Russell A., Ghalaieny M., Akhmetov K.K., Mukanov Y., McCann M., Vitolo C., Althonayan A. A spatial survey of environmental indicators for Kazakhstan: an examination of current conditions and future needs // International Journal of Environmental Research. – 2018. – Vol. 12. – P. 735-748. – DOI: https://doi.org/10.1007/s41742-018-0134-7

2. Международное информационное агентство «DKnews.kz». Казахстан в топ-позициях по уровню загрязнения.–URL:https://dknews.kz/ru/eksklyuziv-dk/221987-kazahstan-v-top-poziciyah-po-urovnyu-zagryazneniya (дата обращения 20.07.2023).

3. Kerimray A., Rojas-Solórzano L., Amouei Torkmahalleh M., Hopke P. K., Ó Gallachóir B.P. Coal use for residential heating: Patterns, health implications and lessons learned // Energy for Sustainable Development. – 2017. – Vol. 40. – P.19–30. – DOI:10.1016/j.esd.2017.05.005

4. Информационно-правовая системанормативных правовых актовРеспублики Казахстан «Әділет». О Стратегическом плане Министерства транспорта и коммуникаций Республики Казахстан на 2011 - 2015 годы. – URL: https://adilet.zan.kz/rus/docs/P1100000129 (дата обращения 21.07.2023).

5. KAZENERGY. Национальный энергетический доклад 2017 KAZENERGY. – URL: http://www.kazenergy.com/upload/document/energy-report/NationalReport17_ru.pdf (дата обращения 21.07.2023).

6. Kerimray A., Rocco M., Rojas-Solórzano L., Gallachoir B. Causes of energy poverty in a cold and resource-rich country: evidence from Kazakhstan // Local Environment. – 2017. – DOI: 10.1080/13549839.2017.1397613.

7. Karatayev M., Pedro R., Mourao Z.S., Konadu D.D., Nilay S., Michèle C. The water-energy-food nexus in Kazakhstan: challenges and opportunities // Energy Procedia. – 2017. – Vol.125. – P.63-70. – DOI: 10.1016/j.egypro.2017.08.064.

8. Current Pollution Index by City. – URL: https://www.numbeo.com/pollution/rankings_current.jsp (дата обращения 21.07.2023).

9. Kerimray A., Azbanbayev E., Kenessov B., Plotitsyn P., Alimbayeva D., Karaca, F. Spatiotemporal Variations and Contributing Factors of Air Pollutants in Almaty, Kazakhstan // Aerosol and Air Quality Research. – 2020. – Vol.20.–P.1340-1352.–DOI:10.4209/aaqr.2019.09.0464.

10. Nugmanova D., Feshchenko Yu., Iashyna L., Gyrina O., Malynovska K., Mammadbayov E., Akhundova I., Nurkina N., Tariq L., Makarova J., Vasylyev A. The prevalence, burden and risk factors associated with chronic obstructive pulmonary disease in Commonwealth of Independent States (Ukraine, Kazakhstan and Azerbaijan): Results of *the CORE study // BMC Pulmonary Medicine. – 2018. – 18. – DOI: 10.1186/s12890-018-0589-5.*

11 Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. Lightgbm. A highly efficient gradient boosting decision tree // Advances in neural information processing systems. – 2017. – Vol.30. – P.3149-3157.
12. Bentéjac C., Csörgő A., Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms // Artificial Intelligence Review. – 2021. – Vol. 54. – P.1937-1967.
13. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – P. 785-794.
14. Mukhamediev R., Amirgaliyev Y., Kuchin Y., Aubakirov M., Terekhov A., Merembayev T., Yelis M., Zaitseva E., Levashenko V., Popova Y., et al. Operational Mapping of Salinization Areas in Agricultural Fields Using Machine Learning Models Based on Low-Altitude Multispectral Images // Drones. – 2023. – Vol.7(357). https://doi.org/10.3390/drones7060357
15. Mukhamediev R.I., Kuchin Y., Amirgaliyev Y., Yunicheva N., Muhamedijeva E. Estimation of Filtration Properties of Host Rocks in Sandstone-Type Uranium Deposits Using Machine Learning Methods // IEEE Access. – 2022. – Vol.10. – P.18855–18872.
16. Mukhamediev R.I., Merembayev T., Kuchin Y., Malakhov D., Zaitseva E., Levashenko V., Popova Y., Symagulov A., Sagatdinova G., Amirgaliyev Y. Soil Salinity Estimation for South Kazakhstan Based on SAR Sentinel-1 and Landsat-8,9 OLI Data with Machine Learning Models // Remote Sens. – 2023. – Vol.15, 4269. – DOI:https://doi.org/10.3390/rs15174269
17. Mukhamediev R.I., Terekhov A., Sagatdinova G., Amirgaliyev Y., Gopejenko V., Abayev N., Kuchin Y., Popova Y., Symagulov A. Estimation of the Water Level in the Ili River from Sentinel-2 Optical Data Using Ensemble Machine Learning // Remote Sens. – 2023. – 15, 5544. – DOI: https://doi.org/10.3390/rs15235544
18. Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions // Adv. Neural Inf. Process. Syst. – 2017. – 30. – P.1–10.
19. Scikit-learn. Machine Learning in Python. Available online: https://scikit-learn.org/stable/ (accessed on 1 February 2024)
20. Scikit-optimize. Sequential model-based optimization in Python https://scikit-optimize.github.io/stable/ (accessed on 1 February 2024)
21. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack // Journal of Open Source Software. – 2018. – 3. – 638. – DOI:10.21105/joss.00638.

REFERENCES

1. Russell A., Ghalaieny M., Akhmetov K.K., Mukanov Y., McCann M., Vitolo C., Althonayan A. A spatial survey of environmental indicators for Kazakhstan: an examination of current conditions and future needs // International Journal of Environmental Research. – 2018. – Vol. 12. – P. 735-748. – DOI: https://doi.org/10.1007/s41742-018-0134-7
2. Mezhdunarodnoe informatsionnoe agentstvo «DKnews.kz». Kazakhstan v top-pozitsiyakh po urovnyu zagryazneniya. – URL: https://dknews.kz/ru/eksklyuziv-dk/221987-kazahstan-v-top-poziciyah-po-urovnyu-zagryazneniya (data obrashcheniya 20.07.2023).
3. Kerimray A., Rojas-Solórzano L., Amouei Torkmahalleh M., Hopke P. K., Ó Gallachóir B.P. Coal use for residential heating: Patterns, health implications and lessons learned // Energy for Sustainable Development. – 2017. – Vol. 40. – P.19–30. – DOI:10.1016/j.esd.2017.05.005
4. Informatsionno-pravovaya sistemanormativnykh pravovykh aktovRespubliki Kazakhstan «Ədilet». O Strategicheskom plane Ministerstva transporta i kommunikatsii Respubliki Kazakhstan na 2011 - 2015 gody. – URL: https://adilet.zan.kz/rus/docs/P1100000129 (data obrashcheniya 21.07.2023).
5. KAZENERGY. Natsional'nyi energeticheskii doklad 2017 KAZENERGY. – URL: http://www.kazenergy.com/upload/document/energy-report/NationalReport17_ru.pdf (data obrashcheniya 21.07.2023).
6. Kerimray A., Rocco M., Rojas-Solórzano L., Gallachoir B. Causes of energy poverty in a cold and resource-rich country: evidence from Kazakhstan // Local Environment. – 2017. – DOI: 10.1080/13549839.2017.1397613.
7. Karatayev M., Pedro R., Mourao Z.S., Konadu D.D., Nilay S., Michèle C. The water-energy-food nexus in Kazakhstan: challenges and opportunities // Energy Procedia. – 2017. – Vol.125. – P.63-70. – DOI: 10.1016/j.egypro.2017.08.064.
8. Current Pollution Index by City. – URL: https://www.numbeo.com/pollution/rankings_current.jsp (data obrashcheniya 21.07.2023).
9. Kerimray A., Azbanbayev E., Kenessov B., Plotitsyn P., Alimbayeva D., Karaca, F. Spatiotemporal Variations and Contributing Factors of Air Pollutants in Almaty, Kazakhstan // Aerosol and Air Quality Research. – 2020. – Vol. 20. – P.1340-1352. – DOI:10.4209/aaqr.2019.09.0464.
10. Nugmanova D., Feshchenko Yu., Iashyna L., Gyrina O., Malynovska K., Mammadbayov E., Akhundova I., Nurkina N., Tariq L., Makarova J., Vasylyev A. The prevalence, burden and risk factors associated with chronic obstructive pulmonary disease in Commonwealth of Independent States (Ukraine, Kazakhstan and Azerbaijan): Results of the CORE study // BMC Pulmonary Medicine. – 2018. – 18. – DOI: 10.1186/s12890-018-0589-5.

11. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. Lightgbm. A highly efficient gradient boosting decision tree // Advances in neural information processing systems. – 2017. – Vol.30. – P.3149-3157.
12. Bentéjac C., Csörgő A., Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms // Artificial Intelligence Review. – 2021. – Vol. 54. – P.1937-1967.
13. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – P. 785-794.
14. Mukhamediev R., Amirgaliyev Y., Kuchin Y., Aubakirov M., Terekhov A., Merembayev T., Yelis M., Zaitseva E., Levashenko V., Popova Y., et al. Operational Mapping of Salinization Areas in Agricultural Fields Using Machine Learning Models Based on Low-Altitude Multispectral Images // Drones. – 2023. – Vol.7(357). https://doi.org/10.3390/drones7060357
15. Mukhamediev R.I., Kuchin Y., Amirgaliyev Y., Yunicheva N., Muhamedijeva E. Estimation of Filtration Properties of Host Rocks in Sandstone-Type Uranium Deposits Using Machine Learning Methods // IEEE Access. – 2022. – Vol.10. – P.18855–18872.
16. Mukhamediev R.I., Merembayev T., Kuchin Y., Malakhov D., Zaitseva E., Levashenko V., Popova Y., Symagulov A., Sagatdinova G., Amirgaliyev Y. Soil Salinity Estimation for South Kazakhstan Based on SAR Sentinel-1 and Landsat-8,9 OLI Data with Machine Learning Models // Remote Sens. – 2023. – Vol.15, 4269. – DOI:https://doi.org/10.3390/rs15174269
17. Mukhamediev R.I., Terekhov A., Sagatdinova G., Amirgaliyev Y., Gopejenko V., Abayev N., Kuchin Y., Popova Y., Symagulov A. Estimation of the Water Level in the Ili River from Sentinel-2 Optical Data Using Ensemble Machine Learning // Remote Sens. – 2023. – 15, 5544. – DOI: https://doi.org/10.3390/rs15235544
18. Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions // Adv. Neural Inf. Process. Syst. – 2017. – 30. – P.1–10.
19. Scikit-learn. Machine Learning in Python. Available online: https://scikit-learn.org/stable/ (accessed on 1 February 2024)
20. Scikit-optimize. Sequential model-based optimization in Python https://scikit-optimize.github.io/stable/ (accessed on 1 February 2024)
21. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack // Journal of Open Source Software. – 2018. – 3. – 638. – DOI:10.21105/joss.00638.

ҚАЛАЛЫҚ АҚЫЛДЫ ҚАЛА ОРТАСЫНДАҒЫ АУА САПАСЫ ДАТЧИКТЕРІНІҢ ЭМУЛЯЦИЯСЫ

Р.И. Мухамедиев[1,2] *и.ғ.д.,* **А.Г. Терехов**[2] *т.ғ.к.,* **А.А. Оксененко**[1], **А.С. Еримбетова**[1,2*] *PhD., к.т.н.,* **Я.И. Кучин**[1,2], **А. Сымагулов**[1,2] , **Д.Р. Құсайын**[1], **П. Рыстыгулов**[1]

[1] Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан
[2]ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан
E-mail: aigerian8888@gmail.com

Қалалық ауаның ластануы адам денсаулығына үлкен қауіп төндіреді. Оны бақылау үшін PM1, PM2.5, PM10 шаң бөлшектерінің және органикалық қосылыстардың концентрациясын бағалау үшін жеке сенсорлар да, жүйелер де қолданылады. Бірақ сенсорлық жүйенің сенімділігі 100 пайыз бола алмайды. Кейде бөлінген жүйедегі белгілі бір сенсорлар істен шығады. Осы себепті, қалған сенсорлардың көрсеткіштеріне негізделген олардың көрсеткіштерін эмуляциялау өте пайдалы. Жұмыс деректер жиынтығын сипаттайды және деректерді жинау орындарындағы функционалды сенсорлардың көрсеткіштері мен ауа-райы жағдайларына негізделген, сәтсіз сенсордың көрсеткіштерін модельдейтін машиналық оқыту үлгісін ұсынады. Ластанудың жекелеген түрлері үшін мұндай эмуляцияның дәлдігі бағаланды (детерминация коэффициенті 0.43-тен 0.61-ге дейін).

Түйін сөздер: ауа сапасы, ақылды қала, машиналық оқыту.

EMULATION OF AIR QUALITY SENSORS IN AN URBAN SMART CITY ENVIRONMENT

R. Mukhamediev[1,2] *doctor of engineering science,* **A. Terekhov**[2] *candidate of technical science,* **A. Oksenenko**[1] **A. Yerimbetova**[1,2*] *PhD., candidate of technical science,* **Ya. Kuchin**[1,2], **A. Symagulov**[1,2], **D. Kusayin**[1], **P. Rystygulov**[1]

[1]*Kazakh National Research Technical University named after K.I. Satbayev, Almaty, Kazakhstan*
[2]*Institute of Information and Computing Technologies CS MSHE RK, Almaty, Kazakhstan*
E-mail: aigerian8888@gmail.com

Urban air pollution poses a serious threat to human health. To monitor it, both individual sensors and systems are used to assess the concentration of dust particles PM1, PM2.5, PM10 and organic compounds. However, the reliability of the sensor system cannot be 100 percent. From time to time, certain sensors in a distributed system fail. For this reason, it is very useful to emulate their readings based on the readings of the remaining sensors. The work describes a data set and proposes a machine learning model, which, based on the readings of functional sensors and weather conditions at the data collection sites, simulates the readings of a failed sensor. The accuracy of such emulation for certain types of pollution has been assessed (the coefficient of determination ranges from 0.43 to 0.61).

Keywords: air quality, smart city, machine learning.

Сведения об авторах/Авторлар туралы мәліметтер/Information about authors:

Мухамедиев Равиль Ильгизович – д.и.н., профессор, Заведующий лабораторией прикладного машинного обучения Satbayev University, г. Алматы, Сатпаева, 22, Главный научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, г. Алматы, Шевченко, 28, *ravil.muhamedyev@gmail.com*
Терехов Алексей Геннадьевич – к.т.н., Главный научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, г. Алматы, Шевченко, 28, *aterekhov1@yandex.ru*
Оксененко Алексей Алексеевич – Инженер, заведующий лабораторией беспилотных летательных аппаратов Satbayev University, эксперт по дронам Международной авиационной федерации, г. Алматы, ул. Сатпаева, 22, *alex-ok@bk.ru*
Еримбетова Айгерим Сембековна – PhD, к.т.н., профессор Satbayev University, Казахстан, Алматы, Сатпаева 22, Ведущий научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, Казахстан, г. Алматы, Шевченко, 28, *aigerian8888@gmail.com*
Кучин Ян Игоревич – магистр инженерных наук, старший научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, г. Алматы, Шевченко, 28, *ykuchin@mail.ru*
Сымагулов Адилхан – магистр технических наук, инженер-программист Института информационных и вычислительных технологий КН МНВО РК, г. Алматы, Шевченко, 28, *asmogulove00@gmail.com*
Құсайын Диас – бакалавр студент Satbayev University, г.Алматы, ул. Сатпаева, 22, *diac.kusain@gmail.com*
Рыстыгулов Панабек Абашович – магистр технических наук, PhD студент Satbayev University, г. Алматы, ул. Сатпаева, 22, *panabek1993@gmail.com*

Мухамедиев Равиль Ильгизович – инж. ғылым. докторы, профессор, Сәтбаев университетінің Қолданбалы машиналық оқыту зертханасының меңгерушісі, Алматы қ., Сәтбаев, 22; ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының бас ғылыми қызметкері, Алматы, Шевченко, 28, *ravil.muhamedyev@gmail.com*
Терехов Алексей Геннадьевич – техн. ғылым. кандидаты, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының бас ғылыми қызметкері, Алматы қ., Шевченко, 28, *aterekhov1@yandex.ru*
Оксененко Алексей Алексеевич – инженер, Сәтбаев университетінің ұшқышсыз ұшатын аппараттар зертханасының меңгерушісі, Халықаралық авиация федерациясының ұшқышсыз аппаратының сарапшысы, Алматы қ., Сәтбаев, 22, *alex-ok@bk.ru*
Еримбетова Айгерим Сембековна – PhD, техн. ғылым. кандидаты, Сәтбаев университетінің профес-

соры, Алматы қ., Сәтбаев, 22, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының жетекші ғылыми қызметкері, Алматы қ., Шевченко, 28, *aigerian8888@gmail.com*
Кучин Ян Игоревич – инж. ғылым. магистрі, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының аға ғылым қызметкері, Алматы қ., Шевченко, 28, *ykuchin@mail.ru*
Сымагулов Адилхан – техн. ғылым. магистрі, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының инженер-программисті, Алматы қ., Шевченко, 288, *asmogulove00@gmail.com*
Құсайын Диас – Сәтбаев университетінің бакалавр студенті, Алматы қ., Сәтбаев, 22, *diac.kusain@gmail.com*
Рыстыгулов Панабек Абашович – техн. ғылым. магистрі, Сәтбаев университетінің PhD студенті, Алматы қ., Сәтбаев, 22, *panabek1993@gmail.com*

Mukhamediev R. – Doctor of Science (Engineering), professor, Head of The Applied Machine Learning Laboratory of Satbayev University, Almaty, Satbayev, 22, Chief Researcher of Institute of Information and Computational Technologies CS MSHE RK, Almaty, Shevchenko, 28, *ravil.muhamedyev@gmail.com*
Terekhov A. – Candidate of Technical Science, Chief Researcher of Institute of Information and Computational Technologies CS MSHE RK, Almaty, Shevchenko, 28, *aterekhov1@yandex.ru*
Oxenenko A. – Engineer, Head of Unmanned Aerial Vehicle Laboratory of Satbayev University, drone expert in Fédération Aéronautique Internationale, Almaty, Satbayev, 22, *alex-ok@bk.ru*
Yerimbetova A. – PhD, Candidate of Technical Science, Associate Professor, professor of Satbayev University, Leading Researcher of Institute of Information and Computational Technologies CS MSHE RK, Almaty, *aigerian8888@gmail.com*
Kuchin Yan – Master of Engineering Science in Computer Systems, Senior Researcher of Institute of Information and Computational Technologies CS MSHE RK,Almaty, Shevchenko, 28, *ykucnin80@gmail.com*
Symagulov A. – Master of Technical Science, software engineer Institute of Information and Computational Technologies CS MSHE RK, Kazakhstan, Almaty, Shevchenko, 28, *asmogulove00@gmail.com*
Kussaiyn D. – bachelor of Satbayev University, Almaty, Satbayev, 22, diac.kusain@gmail.com
Rystygulov P. – Master of Technical Science, PhD student of Satbayev University, Almaty, Satbayev, 22, *panabek1993@gmail.com*

Вклад авторов/ Авторлардың қосқан үлесі/ Authors contribution:

Мухамедиев Р.И. – разработка методологии
Терехов А.Г. – разработка концепции
Оксененко А.А. – проведение исследования
Еримбетова А.С. – ресурсы, подготовка и редактирование текста, визуализация
Кучин Я.И. – создание программного обеспечения
Сымагулов А. – создание программного обеспечения, проведение исследования
Құсайын Д. – создание программного обеспечения
Рыстыгулов П.А. – проведение исследования

Мухамедиев Р.И. – әдістемені әзірлеу
Терехов А.Г. – тұжырымдаманы әзірлеу
Оксененко А.А. – зерттеу жүргізу
Еримбетова А.С. – ресурстар, мәтінді дайындау және өңдеу, көрнекілік
Кучин Я.И. – бағдарламалық жасақтама жасау
Сымагулов А. – бағдарламалық жасақтама жасау, зерттеу жүргізу
Құсайын Д. – бағдарламалық жасақтама жасау
Рыстыгулов П.А. – зерттеу жүргізу

Mukhamediev R. – methodology development
Terekhov A. – concept development
Oxenenko A. – conducting a research
Yerimbetova A. – resources, preparing and editing the text, visualization
Kuchin Yan – creating software
Symagulov A. – creating software, conducting a research
Kussaiyn D. – creating software
Rystygulov P. – conducting a research