УДК 519.25 **DOI** 10.15622/ia.23.2.2

Е.Д. Карепова, В.С. Петракова

СТАТИСТИЧЕСКИ ОБОСНОВАННАЯ КОРРЕКТИРОВКА ПОКАЗАНИЙ ДАТЧИКОВ СТАНЦИЙ CITYAIR УРОВНЯ КОНЦЕНТРАЦИИ ВЗВЕШЕННЫХ ЧАСТИЦ PM2.5 В ПРИЗЕМНОМ СЛОЕ АТМОСФЕРЫ ГОРОДА

Карепова Е.Д., Петракова В.С. **Статистически обоснованная корректировка показаний датчиков станций CityAir уровня концентрации взвешенных частиц PM2.5 в приземном слое атмосферы города.**

Аннотация. В качестве маркера, характеризующего загрязнение воздуха в приземном слое атмосферы современных городов, часто используется уровень концентрации твердых частиц диаметром 2.5 микрона и меньше (Particulate Matter, PM2.5). В работе обсуждается практика применения для измерения концентрации PM2.5 в условиях городской среды относительно дешевого оптического датчика, входящего в состав станции CityAir. В статье предложена статистически обоснованная корректировка получаемых станциями CityAir первичных данных о значениях концентрации взвешенных частиц PM2.5 в приземном слое атмосферы г. Красноярска. Для построения регрессионных моделей эталонными считались измерения, получаемые от анализаторов Е-BAM, расположенных на тех же постах наблюдения, что и корректируемые датчики. Для анализа использовались первичные данные 1) с 9 автоматизированных постов наблюдения краевой ведомственной информационно-аналитической системы данных о состоянии окружающей среды Красноярского края (КВИАС); 2) с 21-й станции CityAir системы мониторинга Красноярского научного центра СО РАН. В работе продемонстрировано, что при корректировке показаний датчиков необходимо учитывать метеорологические показатели. Кроме того, показано, что коэффициенты регрессии существенно зависят от сезона. Проведено сравнение методов обучения с учителем для решения задачи корректировки показаний недорогих датчиков. Дополнительная информация по результатам анализа данных, не вошедшая в текст статьи, размещена на электронном ресурсе https://asm.krasn.ru/.

Ключевые слова: уровень концентрации PM2.5, обучение с учителем, регрессионные модели, корректировка системы датчиков.

1. Введение. По данным Министерства природных ресурсов и экологии РФ город Красноярск является одним из нескольких городов России с самым грязным воздухом, концентрация вредных веществ в атмосфере города часто превышает допустимые нормы. Только за февраль 2023 года Красноярск дважды (5 и 13 февраля) попал на первое место в рейтинге крупных городов мира с высоким уровнем загрязнения атмосферы по версии сервиса IQAir, отслеживающего качество воздуха в реальном времени (https://www.iqair.com/ru/world-air-quality-ranking).

Общепринятым маркером и одновременно одним из самых вредных загрязнителей воздуха в приземном слое атмосферы современных городов являются твердые частицы диаметром 2.5 микрона и меньше (Particulate Matter, PM2.5). Взвешенные частицы PM2.5 имеют как

естественное происхождение (частицы почвы, пыль, сажа, споры растений, цветочная пыльца, а также дым от лесных пожаров) так и антропогенное (выхлопные газы двигателей автомобилей, выбросы промышленных предприятий, продукты сгорания угля или дров при отоплении). Концентрация взвешенных частиц PM2.5 является базовым показателем загрязнения городов и широко обсуждается в научной литературе [1 – 6]. Модели многофакторной линейной регрессии широко описаны в силу легкости их получения и возможности применения в практических задачах оперативного прогноза загрязнения [7 – 10]. Представляет также интерес развивающийся подход к моделированию согласованности измерений [11, 12].

В последнее время количество и качество собираемых данных, а также их детализация имеют тенденцию к росту. Поэтому помимо непосредственной работы с данными о загрязнениях и моделями их распространения уделяется большое внимание проблемам сбора и накопления информации о загрязнениях. В связи с этим представляет особый интерес оценка эффективности использования недорогих сенсоров [13 – 18].

Красноярск является одним из городов России в котором ведется мониторинг качества атмосферного воздуха на стационарных постах наблюдения. Во-первых, Министерство экологии и рационального природопользования Красноярского края поддерживает краевую ведомственную информационно-аналитическую систему данных о состоянии окружающей среды Красноярского края (КВИАС). Девять автоматизированных постов наблюдений (АПН) КВИАС расположены в г. Красноярске. Раз в 20 минут выполняется автоматическое измерение метеорологических параметров и концентрации загрязнений в приземном слое атмосферы. В КВИАС для мониторинга концентрации PM2.5 используются анализаторы пыли модели E-BAM (Met One Instruments Inc., США) [19], принцип действия которых основан на измерении поглощения β-излучения частицами пыли, осажденными на фильтрующую ленту. Эта методтка сертифицирована U.S. EPA (United States Environmental Protection Agency) [20]. Анализаторы этого класса рекомендованы для измерения содержания фракций PM10 и PM2.5 в атмосфере, сертифицированы и аккредитованы во многих странах мира, в том числе и в России (№ 57884-14 в Госреестре средств измерений).

Во-вторых, в г. Красноярск действует система мониторинга качества воздуха Красноярского научного центра СО РАН (КНЦ СО РАН) [21]. Каждый пост оснащен станцией мониторинга воздуха CityAir [22], разработанной группой компаний из новосибирского

технопарка и инновационного центра Сколково. Станция раз в 20 минут выдает основные метеорологические параметры и концентрацию аэрозольных частиц PM2.5 и PM10. Для мониторинга концентрации PM2.5 используются оптические датчики (№ 75984-19 в Госреестре средств измерений), в которых проходящий через поток загрязненного воздуха фокусированный лазерный луч рассеивается на твердых частицах, что регистрируется фотодиодом и позволяет количественно оценить загрязнение. Практика применения таких датчиков показала, что они уступают по точности анализаторам Е-BAM, однако пригодны для оценки уровня концентрации PM2.5 [18]. Система мониторинга КНЦ СО РАН имеет около 30 постов, расположенных в разных районах г. Красноярск, что обеспечивает хорошую детализацию информации о загрязнениях.

В статье анализируется согласованность показаний датчиков, принадлежащих разным системам мониторинга и предлагается статистически обоснованная корректировка первичных данных о концентрации PM2.5, получаемых с постов системы мониторинга КНЦ СО РАН.

Статистический анализ выполнен на языке Python с использованием библиотек `numpy`, `pandas`, `sklearn`, `statsmodels`. Отметим, что дополнительная информация по результатам анализа данных, не вошедшая в текст статьи, размещена на электронном ресурсе [23].

2. Используемые для анализа данные и их обозначения. Для анализа использовались первичные данные 1) с 9-ти автоматизированных постов наблюдения сети КВИАС; 2) с 21-й станции CityAir системы мониторинга КНЦ СО РАН. Далее в зависимости от принадлежности поста АПН КВИАС или системе мониторинга КНЦ СО РАН показатели будут иметь префикс “m_” или “s_”, соответственно. Кроме того, для краткости датчики концентрации PM2.5 АПН КВИАС и станции CityAir будем упоминать как “анализаторы Е-BAM” и “оптические датчики”, соответственно.

По каждому посту данные представлены временными рядами измерений в приземном слое атмосферы температуры (t) в °C, давления (p) в мм рт. ст., относительной влажности воздуха (h) в % и концентрации взвешенных частиц PM2.5 (PM) в мкг/м^3. В скобках указаны используемые далее обозначения факторов. Каждый ряд содержит до 105192 измерений (с 01.01.2019 00:00 по 31.12.2022 23:40, 3 измерения в час).

Поскольку целью работы является построение регрессионной модели для корректировки данных о концентрациях PM2.5 станций CityAir, мы будем рассматривать данные не как временные ряды,

а как связанные выборки случайных величин. Описательная статистика показывает, что распределение температуры имеет ярко выраженную трехмодальность (зимний, летний и демисезонный периоды), причем зимний сезон имеет «тяжелый хвост» в сторону низких отрицательных температур. Распределение давления близко к нормальному, распределение влажности имеет большую асимметрию с «тяжелым хвостом» влево. Гистограммы распределения для концентрации PM2.5, температуры, давления и влажности для 4-х дублирующих датчиков (карусель изображений), а также описательная статистика данных представлены в разделе «Описательная статистика» ресурса [23].

Распределение концентрации PM2.5 близко к логнормальному (таблица 1), однако имеет две моды, первая из которых соответствует фоновым значениям концентрации, а вторая, менее выраженная, – периодам высоких концентраций. Отметим, что значения, обычно определяемые в описательной статистике как выбросы (т.е. превышающие в полтора межквартильного расстояния значение третьего квартиля [24]), для нас таковыми не являются, поскольку отражают ситуацию значительного превышения предельно допустимых концентраций (ПДК) PM2.5 в атмосфере.

Таблица 1. Уровни концентрации твердых взвешенных частиц PM2.5. Описательная статистика. Данные поста «Ветлужанка»

Статистика	Первичные данные (мкг/м^3)		Первичные данные после очистки (мкг/м^3)		Логарифм от первичных данных		Логарифм от первичных данных после очистки	
	s	m	s	m	s	m	s	m
Максимум	797.00	403.00	797.00	403.00	6.68	6.00	6.68	6.00
Среднее	45.90	24.78	48.19	25.99	2.85	2.63	2.94	2.70
Ошибка среднего	0.34	0.14	0.37	0.16	0.01	0.00	0.01	0.00
Среднеквадратичное отклонение	79.12	33.04	81.33	34.13	1.37	1.09	1.33	1.04
25% (Q_1)	6.50	7.00	7.17	8.00	1.87	1.95	1.97	2.08
Медиана (Q_2)	14.50	14.00	15.50	14.00	2.67	2.64	2.74	2.64
75% (Q_3)	42.50	27.00	45.26	28.00	3.75	3.30	3.81	3.33
$IQR = Q_3 - Q_1$	36.00	20.00	38.00	20.00	1.88	1.35	1.84	1.25
Асимметрия	3.16	3.33	3.06	3.24	0.35	-0.03	0.42	0.11
Эксцесс	14.64	17.71	13.83	16.72	2.56	3.08	2.54	3.02

В данных присутствуют пропуски, соответствующие периодам поломки аппаратуры. Для нашего исследования заполнение пропусков нецелесообразно, такие данные были просто удалены из выборок. Кроме того, данные о концентрации PM2.5 содержали небольшое количество случаев, которые были расценены как сбои аппаратуры. Например, если в период фоновых концентраций PM2.5 в трех идущих подряд измерениях между двумя измерениями, которые соответствуют небольшим значениям концентраций PM2.5, происходит резкий скачок показаний датчика, мы считали его сбоем аппаратуры. Такие ситуации редки (< 10 случаев), хаотично разбросаны по 30 выборкам и, следовательно, не отражают какую-либо тенденцию. Эти данные тоже были удалены из выборок. Следует иметь в виду, что не все ошибки измерений и сбои аппаратуры мы могли выявить, и они остались в выборке.

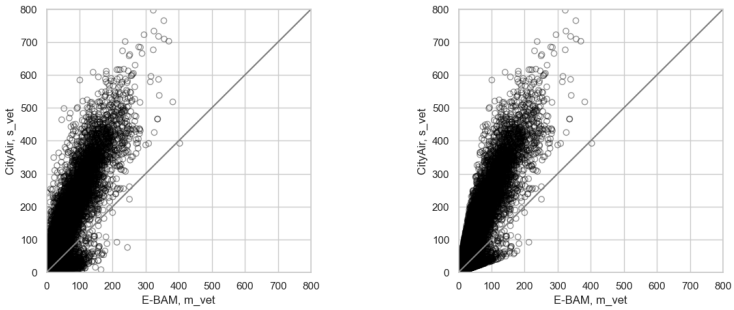
Для сравнения показаний датчиков, принадлежащих разным системам мониторинга существует 4 поста, на которых установлена измерительная аппаратура обоих типов. Это посты “Ветлужанка”, “Покровка”, “Свердовский”, “Кировский”[1].

Уже на этапе описательной статистики становится ясно, что имеются существенные различия в данных, полученных с помощью измерительной аппаратуры разного типа. Например, среднее значение концентрации PM2.5 по выборке измерений оптического датчика станции CityAir, почти в 2 раза превышают среднее значение в выборке анализатора Е-BAM.

Поскольку методика измерения концентрации PM2.5, используемая анализаторами Е-BAM, тщательно верифицирована [20], а оптические датчики принято использовать, как сравнительно дешевую альтернативу [13 – 18], то актуально построение статистически обоснованного правила корректировки первичных данных о концентрации PM2.5 оптических датчиков по данным, полученным от Е-BAM.

Диаграмма рассеяния рисунка 1(а) показывает систематическое завышение концентрации оптических датчиков. Мы приняли решение для каждой пары датчиков исключить из анализа $\sim$5 % пар значений, которые соответствуют максимальным расхождениям в показаниях уровней концентраций PM2.5 в паре. В результате из последующего анализа были исключены пары значений, в которых показания оптического датчика более чем в 6 раз отличались от показаний анализатора Е-BAM. Диаграмма рассеяния после корректировки отображена на рисунке 1(б). В таблице 1 приведена описательная статистика для скорректированных выборок концентраций PM2.5.

[1] раздел «Карты» ресурса [23]



а) б)

Рис. 1. Диаграмма рассеяния показаний о концентрации взвешенных частиц PM2.5 (мкг/м^3) на анализаторе E-BAM (`m_vet`) и оптическом датчике станции CityAir (`s_vet`), расположенных в Ветлужанке а) до; б) после очистки данных

Далее в тексте будут приведены результаты анализа для пары дублирующих датчиков с поста “Ветлужанка”. Для краткости будем помечать “s_vet” и “m_vet” измерения станции CityAir и датчиков АПН КВИАС, соответственно. Анализ для оставшихся трех пар датчиков, расположенных в Покровке, Свердловском и Кировском районах г. Красноярска, проводился аналогично, необходимые ссылки на электронный ресурс [23], содержащий результаты анализа по этим постам, будут даны в тексте статьи.

3. Описание используемых для анализа методов. Основными хорошо формализованными средствами статистики, которые используются при поиске взаимосвязей между выборками, являются 1) корреляционный анализ, выявляющий наличие линейных связей между двумя выборками, и 2) аппарат парной или множественной регрессии, устанавливающий более общую взаимосвязь между наборами данных [25 – 27]. В обоих случаях, выборки должны быть достаточного объема и хорошо описывать генеральную совокупность.

Пусть известно множество $X_{\mathcal{T}} \in R^{D\times n}$, состоящее из n *связанных* выборок (факторов) мощности D. Каждая выборка представляется вектор-столбцом $x^j = (x_1^j, \ldots, x_D^j)^T$, $j = 1, \ldots, n$. Отметим, что $X_{\mathcal{T}}$ также можно рассматривать как множество строк-наблюдений $x_i = (x_i^1, \ldots, x_i^n)$, $i = 1, \ldots, D$. Пусть каждому наблюдению x_i соответствует скалярный отклик $y_i^{\mathcal{T}}$, $i = 1, \ldots, D$. На основе этой информации в регрессионном анализе необходимо построить алгоритм оценки значения отклика $\hat{y}$ по входному набору значений факторов $\hat{x} = (\hat{x}^1, \ldots, \hat{x}^n) \notin X_{\mathcal{T}}$. Алгоритм называют *регрессионной моделью*, а множество пар $(y_i^{\mathcal{T}}, x_i)$ – *обучающим*.

В настоящее время в регрессионном анализе наряду с классическими статистическими методами используются методы машинного обучения с учителем. Кратко опишем методы, используемые в данной работе.

В большинстве случаев, мы рассматривали линейные регрессионные модели, для которых прогнозное значение отклика $\hat{y} \in R$ ищется в виде:

$$\hat{y} = f(\hat{x}) := \omega_0 + \omega_1 \hat{x}^1 + ... + \omega_n \hat{x}^n, \tag{1}$$

линейной комбинации факторов $\hat{x} = \left(\hat{x}^1, ..., \hat{x}^n\right) \in R^n$. В этом случае построение регрессионной модели сводится к определению набора параметров $\omega = (\omega_0, ..., \omega_n)$ на основе обучающего множества пар $(y_i^{\mathcal{T}}, x_i)$, где $x_i \in X_{\mathcal{T}}$, $i = 1, \ldots, D$.

Классический метод определения параметров ω, разработанный еще К.Ф. Гауссом и А.А. Марковым – это метод наименьших квадратов (МНК), в котором минимизируется сумма квадратов невязок:

$$\min_{\omega} \sum_{i=1}^{D} \left(y_i^{\mathcal{T}} - f(x_i)\right)^2. \tag{2}$$

Если обучающее множество содержит выборки почти коллинеарных факторов (нечеткая мультиколлинеарность набора факторов), то МНК становится неустойчив, в том числе вычислительно. В таких случаях используют некоторую регуляризацию (2), не позволяющую параметрам ω сильно возрастать. В методе *гребневой регрессии* (Ridge) используется регуляризация Тихонова в норме пространства L_2 [28], в методе *LASSO* (Least Absolute Shrinkage and Selection Operator) – в норме пространства L_1 [29], а модель *эластичной сети* (Elastic Net) использует обе эти регуляризации [30]. При этом в моделях появляются дополнительные параметры, которые необходимо подбирать на основе специального исследования набора данных, например, методом кросс-валидации. Метод LASSO часто приводит к разреженным регрессионным моделям, поскольку некоторые компоненты в ω могут стать нулевыми и, следовательно, будут исключены из модели. Метод гребневой регрессии может делать часть компонент в ω малыми, но редко зануляет их. Эластичная сеть дает более гладкую зависимость ω от параметра регуляризации, чем метод LASSO. Таким образом, эти методы также

можно использовать для обоснования уменьшения размерности пространства факторов.

Метод опорных векторов (SVM, Support Vector Method) был предложен в [31], как метод классификации (метод обобщенных портретов) полезный для нечетко разделенных классов. В [32, 33] метод был расширен на решение задач регрессии, аппроксимации и оценивания функций. В методе опорных векторов наряду с минимизацией функционала (2) и его регуляризацией в пространстве L_2 оптимизируется и множество наблюдений (опорных векторов), по которым вычисляется сумма в (2). Метод игнорирует ошибки, меньшие заданного ε.

Предсказание отклика по линейной модели регрессии, обученной по обучающему множеству без выбросов, довольно устойчиво. Однако, поскольку накладывается сильное ограничение на структуру модели (линейный вид функции $f(x)$ в (1)), предсказание отклика может быть неточным [27]. Существует широкий набор методов, в которых априорные допущения о структуре модели очень слабые. В результате модель, обычно, перестает быть линейной и гибко подстраивается под обучающее множество $X_{\mathcal{T}}$. При отсутствии переобучения (о чем необходимо заботиться специально) предсказания таких методов могут быть весьма точны. Однако, обратной стороной адаптивности модели является ее неустойчивость. Структура модели сильно зависит от обучающего множества; и предсказания откликов могут сильно отличаться, если они получены по моделям, обученным даже на мало отличающихся обучающих множествах.

В работе мы сравнили результаты, полученные по описанным выше линейным моделям и двум методам, в которых структура модели заранее не фиксируется. Оба метода разрабатывались для задач классификации, но были адаптированы к задачам регрессии. В этих методах пространство объектов рассматривается, как n-мерное метрическое пространство с определенным расстоянием $\rho(\cdot,\cdot)$, а наблюдения x – как точки в нем.

Метод k *ближайших соседей* предсказывает значение отклика $\hat{y}$ по входному набору значений факторов $\hat{x} = (\hat{x}^1, \ldots, \hat{x}^n)$ на основе осреднения откликов от k наблюдений из обучающего множества $X_{\mathcal{T}}$, которые являются ближайшими к значению входной переменной $\hat{x}$ по метрике ρ.

В регрессионной модели *дерева решений* [34, 35] пространство объектов представляется дизъюнктным объединением непересекающихся областей R_m, $m = 1, \ldots, M$, в каждом из которых настраивается простая

регрессионная модель (например, константа) для предсказания отклика:

$$\hat{y} = f(\hat{x}) := \sum_{m=1}^{M} c_m(\hat{x}) I(\hat{x} \in R_m).$$

Здесь $I(x \in R)$ – идентификационная функция области R, а $c_m(x)$ – соответствующая R_m регрессионная модель. Для идентификации регрессионной модели необходимо определить правило разбиения пространства объектов на области и соответствующую каждой области c_m. В большинстве пакетов в настоящее время для обучения деревьев решений используется итерационный алгоритм CART [35], в котором разбиение на подобласти происходит гиперплоскостями, параллельными одной из координатных осей, а $c_m = \mathtt{avg}(y_i^{\mathcal{T}} | x_i \in R_m)$ усредняет отклики из обучающей выборки, соответствующие всем $x_i \in R_m$.

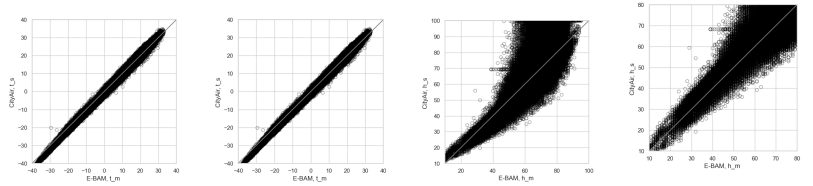
Поскольку модели, полученные по алгоритму CART являются неустойчивыми относительно обучающего множества и имеют тенденцию к переобучению, то обычно используют специальную процедуру баггинга [36], повышающую устойчивость прогноза. Мы использовали алгоритм *случайного леса* (Random Forest) [37], который генерирует ансамбль решающих деревьев и усредняет по нему предсказание.

4. Корреляционный анализ. Сезонность. Построенные на рисунке 2 диаграммы рассеяния для а) температуры; б) давления; в) влажности, полученных с АПН КВИАС (`m_vet`) и станции CityAir (`s_vet`) иллюстрируют[2] согласованность показаний температуры и давления (коэффициент детерминации показаний одного типа датчика относительно другого в этом случае $R^2 > 0.98$). Однако измерения влажности разными приборами существенно отличаются (рисунок 2(в)). Регрессия вида:

$$h'_{\mathtt{m\_vet}} = 0.000028 h^3_{\mathtt{s\_vet}} - 0.01305 h^2_{\mathtt{s\_vet}} + 1.9667 h_{\mathtt{s\_vet}} - 14.62964,$$

с коэффициентом детерминации $R^2 = 0.78$ улучшает согласованность показаний разных типов датчиков (рисунок 2(г)), но все равно при повышенной влажности разброс показаний станции CityAir относительно АПН КВИАС остается более 40 %.

[2]Аналогичный анализ для оставшихся пар дублируемых датчиков доступен в разделе Диаграммы рассеяния ресурса [23]



а) температура б) давление в) влажность г) скорректированная влажность

Рис. 2. Диаграммы рассеяния метеорологических параметров, полученных со станции CityAir и АПН КВИАС на посту Ветлужанка

Климат г. Красноярск континентальный с относительно морозной малоснежной зимой и жарким летом с малым количеством осадков. Более того, в холодный сезон большой вклад в концентрацию загрязнителя в атмосфере вносят работающие на полную мощность ТЭЦ и печное отопление. Поэтому естественна гипотеза о разном проявлении связи значений концентрации загрязнителя в атмосфере и метеорологических параметров в разные сезоны, что неизбежно будет влиять на модель корректировки датчиков. Для подсети дублирующих датчиков сгруппированные по месяцам коэффициенты корреляции между концентрацией PM2.5 и температурой, давлением и влажностью приведены в таблицах 2 (а–в), соответственно[3]. Заметим, что взятые по месяцам значения метеорологических параметров распределены нормально.

Корреляционный анализ позволяет быстро оценить возможность линейной связи между значениями концентраций PM2.5 и метеопараметрами. На его основе можно сделать следующие предварительные выводы. Во-первых, в холодное время года существенны отрицательная корреляция концентрации PM2.5 с температурой и положительная с давлением и влажностью, а поздней весной и летом эти корреляции практически отсутствуют. Это легко объясняется следующими причинами. Низкие температуры, отсутствие ветра и высокая влажность являются причиной температурных инверсий в нижних слоях атмосферы, что затрудняет рассеяние загрязняющих веществ [38] и приводит к периодам устойчивой повышенной концентрации их маркера PM2.5. Более того, с понижением температуры повышается интенсивность печного отопления и работы ТЭЦ, что тоже увеличивает загрязнение нижних слоев атмосферы в холодный сезон.

[3] Данные для всех 30 датчиков доступны в разделе «Корреляционный анализ» ресурса [23].

Таблица 2. Коэффициент корреляции между показаниями подсети дублирующих датчиков о концентрации частиц PM2.5 и измерениями а) температуры; б) влажности; в) давления

а)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
m_vet	-0.48	-0.58	-0.49	-0.05	-0.13	-0.04	0.03	0.08	-0.06	-0.36	-0.10	-0.64
m_pok	-0.45	-0.47	-0.36	0.08	0.10	0.12	0.12	0.21	0.02	-0.12	-0.28	-0.42
m_svr	-0.35	-0.42	-0.32	0.21	0.13	0.27	0.12	0.21	0.20	-0.05	-0.22	-0.52
m_kir	-0.32	-0.44	-0.37	0.08	0.10	0.17	-0.05	0.14	-0.01	-0.11	-0.22	-0.38
s_vet	-0.55	-0.57	-0.52	-0.15	-0.22	-0.07	-0.16	0.07	-0.25	-0.33	-0.33	-0.67
s_pok	-0.48	-0.55	-0.40	-0.09	-0.09	-0.01	-0.12	0.09	-0.13	-0.17	-0.26	-0.55
s_svr	-0.46	-0.36	-0.38	0.16	0.02	0.07	0.12	0.22	0.04	-0.13	-0.22	-0.45
s_kir	-0.45	-0.38	-0.42	-0.02	-0.06	-0.05	-0.06	0.08	-0.15	-0.24	-0.24	-0.41

б)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
m_vet	0.49	0.46	0.40	0.12	0.04	0.10	0.18	0.21	0.28	0.27	0.37	0.49
m_pok	0.39	0.38	0.22	-0.01	-0.15	-0.04	-0.00	0.08	0.06	0.04	0.36	0.32
m_svr	0.48	0.38	0.28	-0.05	-0.07	-0.18	-0.04	0.02	0.01	0.04	0.36	0.52
m_kir	0.54	0.47	0.35	0.03	-0.04	0.01	0.11	0.11	0.23	0.18	0.40	0.41
s_vet	0.35	0.19	0.34	0.18	0.14	0.18	0.03	0.12	0.30	0.22	0.25	0.05
s_pok	0.31	0.12	0.18	0.07	0.01	0.11	0.05	0.12	0.14	0.14	0.20	0.19
s_svr	0.35	0.21	0.18	0.02	0.04	0.21	0.13	0.01	0.16	0.12	0.15	0.15
s_kir	0.35	0.34	0.23	0.15	0.05	0.24	0.27	0.13	0.27	0.14	0.34	0.32

в)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
m_vet	0.13	0.21	0.07	0.05	0.15	-0.12	-0.28	-0.21	0.02	0.19	0.03	0.26
m_pok	0.31	0.34	0.05	0.00	0.04	-0.02	-0.02	-0.11	0.06	0.16	0.31	0.14
m_svr	0.14	0.26	0.02	-0.05	-0.03	-0.04	-0.08	-0.14	-0.07	0.09	0.27	0.40
m_kir	0.20	0.34	0.02	-0.00	-0.01	-0.15	-0.06	-0.09	-0.05	0.16	0.20	0.31
s_vet	0.35	-0.07	0.08	0.09	0.19	-0.02	-0.07	-0.09	0.09	0.23	0.29	-0.32
s_pok	0.33	0.43	0.07	0.03	0.02	0.04	0.03	0.00	0.10	0.10	0.32	0.37
s_svr	0.30	0.26	0.08	-0.05	0.02	-0.23	-0.14	-0.12	-0.06	0.17	0.31	0.34
s_kir	0.28	0.33	0.09	0.05	0.06	-0.16	-0.10	-0.10	0.05	0.17	0.27	0.30

Есть и более общее замечание. Летом в целом концентрации PM2.5 ниже, чем в холодный период, а точность измерений анализатора E-BAM [19] и оптического датчика при низких концентрациях взвешенных частиц ниже, чем при высоких. Таким образом, дисперсия обоих измерений растет, что плохо отражается на их согласованности. Во-вторых, в холодный период положительная корреляция между показаниями концентрации PM2.5 и влажностью для датчиков эталонной подсети КВИАС выше, чем для датчиков станций CityAir, что, по всей видимости, объясняется невысокой точностью измерений станции CityAir.

Таким образом, из результатов корреляционного анализа следует, что параметры регрессионной модели согласования показаний парных

датчиков должны зависеть от сезона. Для последующего анализа нами были условно выделены три периода: 1) зимний (октябрь – март) с сильной корреляцией между температурой и концентрацией PM2.5); 2) летний (июнь – август); 3) демисезонный (апрель, май, сентябрь). Следует отметить, что в августе наблюдается слабая положительная корреляция загрязнений с температурой, тогда как в другие летние месяцы эта связь отсутствует. Такое поведение косвенно подтверждает, что измерения при низких концентрациях плохо согласованы, поскольку в августе в Красноярске часто возникают периоды повышения концентрации PM2.5, связанные с лесными пожарами.

5. Модели регрессии для корректировки значений концентрации PM2.5 от датчиков станций CityAir. Сначала мы выполним регрессионный анализ для выявления зависимости значений концентрации взвешенных частиц PM2.5, полученных с помощью оптического датчика станции CityAir, от эталонных значений концентраций PM2.5, полученных от анализатора Е-BAM, с учетом значений метеорологических параметров. Кроме того, мы сравним методы обучения регрессионной модели, описанные в разделе 3.

Имеющиеся данные были разбиты на обучающую (80% объема) и тестовую (20% объема) выборки. Качество модели оценивалось коэффициентом детерминации R^2, который показывает, какую долю дисперсии тестовой выборки концентраций PM2.5 объясняет модель. Поскольку значение коэффициента детерминации зависит от разбиения данных на обучающую и тестовую выборки, то процедура повторялась для 100 случайных разбиений. Далее в таблицах приведен средний коэффициент детерминации по всем попыткам, при этом среднеквадратичное отклонение не превышает процента. Ниже приведены результаты для дублирующих датчиков с поста Ветлужанка[4].

В регрессионном анализе рассматривались различные комбинации следующих факторов: концентрации PM2.5, полученные анализатором Е-BAM (PM_m); температура (t_s), давление (p_s) и влажность (h_s), полученные с помощью датчиков станции CityAir. В качестве отклика рассматривались значения концентрации PM2.5, полученные с оптического датчика CityAir (PM_s). Сделаем несколько замечаний относительно всех построенных в статье линейных регрессионных моделей. Во-первых, все остатки имеют нулевое среднее, медиану в районе 0,8 мкг/м^3, слабую отрицательную симметрию ($\sim$-0,2) и умеренный эксцесс ($\sim$ 15). По критерию Дарбина-Уотсона автокорреляции первого

[4]Результаты регрессионного анализа по остальным парам датчиков представлены в разделе «Регрессионный анализ» ресурса [23].

порядка у остатков отсутствуют. Тем не менее, статистическую проверку на нормальность остатки не проходят. Во-вторых, t-статистика с 5% уровнем значимости показывает, что во всех случаях коэффициенты значимо отличны от нуля. F-статистики с 5% уровнем значимости показывает, что во всех случаях существуют коэффициенты отличные от нуля, т.е. в этом смысле линейная модель приемлема. В-третьих, с помощью информационного критерия Акаике AIC и байесовского критерия BIC [40, 41] проведено сравнение качества линейных регрессий с различным набором факторов, построенных на одном и том же обучающем множестве. Анализ показал, что наименьшие AIC и BIC имеет регрессия, учитывающая PM_m, t_s и p_s.

В таблице 3 представлены коэффициенты детерминации, вычисленные для каждой из рассмотренных регрессионных моделей, обученных на всем объеме обучающей выборки. R^2 оценивался на основе полного объема данных тестовой выборки с учетом множественности факторов.

Таблица 3. Коэффициент детерминации R^2 регрессионных моделей, обученных на полном объеме данных обучающей выборки

Модель	Факторы				
	PM_m	PM_m, t_s	PM_m, t_s, h_s	PM_m, t_s, p_s	PM_m, t_s, p_s, h_s
Линейная регрессия (МНК)	0.844	0.856	0.857	0.858	0.859
LASSO	0.844	0.856	0.856	0.858	0.859
Эластичная сеть	0.844	0.856	0.856	0.858	0.859
Метод опорных векторов	0.831	0.850	0.853	0.760	0.748
k ближайших соседей	0.830	0.870	0.882	0.883	0.889
Дерево решений	0.847	0.789	0.803	0.815	0.825
Случайный лес	0.848	0.864	0.883	0.893	*0.902*

На основе результатов регрессионного анализа можно сделать следующие выводы. Во-первых, множественная линейная регрессия с помощью наименьших квадратов (МНК) даёт хорошее приближение, сравнимое по точности с более сложными и вычислительноемкими методами машинного обучения. При этом линейная регрессия позволяет в явном виде получать коэффициенты, отражающие зависимость значения отклика от значений факторов. Во-вторых, для всех пар дублирующих датчиков лучшую точность предсказания отклика дают непараметрические методы “случайный лес” и “k ближайших соседей”. В-третьих, добавление в анализ зависимости факторов влажности и давления не дает значительного улучшения точности моделей.

Это можно объяснить двумя причинами: 1) всесезонность выборки гасит разнонаправленное влияние этих факторов в разные сезоны; 2) обсуждаемая ранее некорректность измерений влажности.

Множественная линейная регрессия, учитывающая максимальное количество факторов, для парных датчиков поста Ветлужанка имеет следующий вид:

$$PM_s = a_0 + a_1 \cdot PM_m + a_2 \cdot t_s + a_3 \cdot p_s + a_4 \cdot h_s = \\ = 88.068 + 2.100\, PM_m - 0.781\, t_s - 0.127\, p_s + 0.054\, h_s, \quad (3)$$

где коэффициенты определены со следующими доверительными интервалами: $a_0 \in [86.584; 89.209]$, $a_1 \in [2.097; 2.103]$, $a_2 \in [-0.785; -0.777]$, $a_3 \in [-0.129; -0.125]$, $a_4 \in [0.054; 0.056]$. Отметим, что учет логнормальности распределений концентраций PM2.5 не дает существенного улучшения в прогнозе[5].

Построенные по полной обучающей выборке регрессионные модели мы оценили для отдельных групп значений тестовой выборки (таблица 4). Во-первых, коэффициент детерминации R^2 был вычислен по группам значений тестовой выборки, относящимся к одному сезону (строки Зима, Лето и Демисезон). Во-вторых, R^2 был вычислен для наблюдений из тестовой выборки, соответствующих моментам, когда скользящее среднее за сутки значение концентрации PM2.5 не превышало принятого в России [39] среднесуточного значения ПДК PM2.5 (35 мкг/м^3) и наблюдениям, в которых среднесуточная концентрация превышала ПДК. В таблице 4 соответствующие строки помечены “Не превышает ПДК” и “Превышает ПДК”, соответственно. В этом случае для того, чтобы избежать запаздывания периодов роста и спада концентрации PM2.5 текущее среднее значение вычислялось по наблюдениям, взятым за период 12 часов до текущего значения и 12 часов, начиная с текущего значения. Наконец, R^2 был вычислен для усредненных скользящим средним данных с окном 1 час, 6 часов, сутки.

Из данных таблицы 4 следует, что линейная регрессия хорошо приближает скользящее среднее, и тем лучше, чем больше окно. В то же время ожидаемо предсказания отклика в период высоких концентраций точнее, чем в период низких. Это подтверждается и очень низкой точностью модели для демисезонного, и, особенно, летнего периодов. Кроме того, учет в модели температуры повышает ее точность

[5]раздел «Учет логнормальности» ресурса [23]

(R^2 увеличивается на проценты), добавление в модель давления еще незначительно улучшает точность прогноза отклика (R^2 увеличивается на десятые доли процента). Введение в модель влажности нецелесообразно.

Таблица 4. Коэффициент детерминации R^2, рассчитанный для групп значений уровня концентрации PM2.5 датчика "`s_vet`" тестовой выборки, для линейной регрессии (МНК), построенной на всём объеме данных обучающей выборки

Обучающая выборка	Факторы				
	PM_m	PM_m, t_s	PM_m, t_s, h_s	PM_m, t_s, p_s	PM_m, t_s, p_s, h_s
Скользящее среднее за сутки	0.939	0.944	0.952	*0.954*	0.953
Скользящее среднее за 6 ч.	0.936	0.941	0.947	0.947	0.947
Скользящее среднее за час	0.895	0.901	0.907	0.907	0.907
Не превышает ПДК	0.324	0.304	0.357	0.356	0.357
Превышает ПДК	0.739	0.770	0.764	0.771	0.771
Зима	0.862	0.863	0.866	0.867	0.868
Лето	0.220	0.314	0.338	0.367	0.363
Демисезон	0.582	0.600	0.603	0.603	0.602

Из проведенного исследования следует, что для повышения точности корректировки датчика необходимо обучать модели не на всей совокупности обучающей выборки, а предварительно выделять из всего множества данных целевую группу. В таблице 5 представлены значения коэффициента детерминации R^2 после обучения трех моделей: линейной регрессии на основе метода наименьших квадратов (LR), случайного леса (RF) и «k ближайших соседей» (k-N). В названии строки указана группа данных, на которой проходило обучение. В частности, из таблицы 5 следует, что обучение на осредненных данных с учетом всех факторов дает R^2 близкий к единице. Кроме того, мы видим значительное улучшение моделей на данных с небольшими значениями концентрации PM2.5. Таким образом, результаты анализа, представленные в таблицах 4, 5, показывают, что корректировка концентраций PM2.5 для оптического датчика CityAir относительно анализатора E-BAM должна выполнятся, по крайней мере, по сезонным данным.

Для оперативной корректировки PM_s удобно использовать параметрическую модель множественной линейной регрессии методом наименьших квадратов, учитывающую показания температуры и давления. В этом случае фактором является концентрация PM2.5, измеренная оптическим датчиком CityAir, а для обучения модели в качестве отклика используются концентрации PM2.5, измеренные анализатором E-BAM.

Таблица 5. Коэффициент детерминации R^2, рассчитанный для значений уровня концентрации PM2.5 оптического датчика "s_vet" тестовой выборки для нескольких моделей регрессии, построенных на указанной в строке группе данных обучающей выборки

Обучающая выборка	PM_m			PM_m, t_s			PM_m, t_s, p_s		
	LR	RF	k-N	LR	RF	k-N	LR	RF	k-N
Скользящее среднее									
за сутки	0.945	0.945	0.949	0.959	0.991	0.988	0.959	0.997	0.997
за 6 часов	0.944	0.944	0.945	0.955	0.978	0.977	0.955	0.985	0.984
за час	0.904	0.904	0.898	0.914	0.930	0.931	0.915	0.943	0.938
Не превышает ПДК	0.616	0.621	0.577	0.631	0.654	0.683	0.631	0.725	0.712
Превышает ПДК	0.772	0.769	0.752	0.795	0.796	0.799	0.807	0.840	0.820
Зима	0.889	0.893	0.883	0.890	0.879	0.889	0.890	0.898	0.893
Лето	0.795	0.782	0.788	0.797	0.787	0.799	0.799	0.874	0.823
Демисезон	0.693	0.709	0.667	0.715	0.712	0.723	0.716	0.773	0.737

Ниже приведены формулы пересчета показаний оптического датчика для поста "Ветлужанка"[6]:

$$PM^{corr}_{\texttt{s\_vet}} = 36.109 + 0.367\, PM_{\texttt{s\_vet}} - 0.143\, t_{\texttt{s\_vet}} - 0.041\, p_{\texttt{s\_vet}}, \quad (4)$$

$$PM^{corr}_{\texttt{s\_vet}} = 286.693 + 0.535\, PM_{\texttt{s\_vet}} + 0.300\, t_{\texttt{s\_vet}} - 0.385\, p_{\texttt{s\_vet}}, \quad (5)$$

$$PM^{corr}_{\texttt{s\_vet}} = 28.500 + 0.361\, PM_{\texttt{s\_vet}} + 0.166\, t_{\texttt{s\_vet}} - 0.0310\, p_{\texttt{s\_vet}}, \quad (6)$$

для зимнего ($R^2 \sim 0.88$), летнего ($R^2 \sim 0.81$) и демисезонного ($R^2 \sim 0.67$) периодов, соответственно.

Более того, однопараметрическая регрессия $PM^{corr}_s = a \cdot PM_s$, построенная по тем же данным дает коэффициенты пересчета a равные 0.43, 0.7 и 0.54 для зимнего, летнего и демисезонного периодов, соответственно. В этом случае, коэффициенты детерминации равны 0.85, 0.57 и 0.40.

[6]Аналогичные формулы для трех других постов можно найти в разделе «Формулы для корректировки» ресурса [23].

6. Корректировка значений концентрации PM2.5 по постам системы мониторинга КНЦ СО РАН. В результате описанного статистического анализа мы имеем подсистему из четырех оптических датчиков (`s_vet`, `s_pok`, `s_svr`, `s_kir`), откалиброванных по показаниям эталонных анализаторов Е-BAM.

Будем использовать эти датчики для корректировки оптических датчиков CityAir всех других постов системы мониторинга КНЦ СО РАН.

На рисунке 3 представлены коэффициенты корреляции между показаниями о концентрации PM2.5 каждого откалиброванного оптического датчика CityAir и всеми другими оптическими датчиками CityAir постов системы мониторинга КНЦ СО РАН. Для каждого датчика найден откалиброванный, у которого с ним максимальный коэффициент корреляции. На рисунке 4 все датчики нанесены на карту и одним цветом закрашены топографические области, объединяющие датчики, коэффициенты корреляции которых максимальны с одним из четырех откалиброванных датчиков.

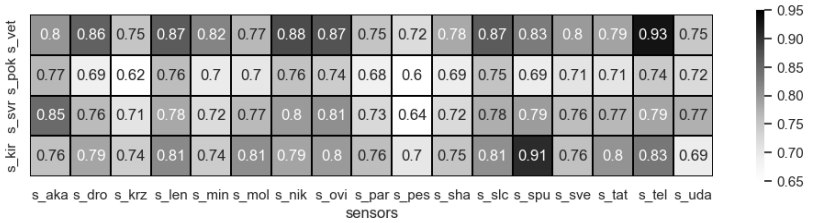


Рис. 3. Корреляция между показаниями об уровнях концентрации частиц PM2.5 подсети дублирующих датчиков и датчиков, оставшейся подсети КНЦ СО РАН

Полученные топографические области хорошо объясняются следующими географическими фактами. Во-первых, датчик `s_pok` расположен на хорошо проветриваемом высоком холме вдали от транспортных развязок. Ожидается, что этот датчик будет наименее коррелирован с другими. Поэтому в зону его действия не попал ни один датчик. Во-вторых, самая большая зона влияния (Зона 1) оказалась у датчика `s_vet`. Это можно объяснить преобладанием северо-западных ветров и расположением датчика `s_vet` на северо-западе в жилом районе Красноярска. Все датчики Зоны 1 относятся к левому берегу Енисея. Эти датчики будут аналогичным образом реагировать на суточные колебания уровней концентрации PM2.5, связанные, например, с пробками на дорогах, и длительные периоды высокой концентрации PM2.5, связанные с неблагоприятными метеорологическими условиями.

Рис. 4. Области максимальной корреляции показаний датчиков сети КНЦ СО РАН с показаниями скорректированных датчиков постов "vet "pok "svr "kir"

Может показаться удивительным, что в зону влияния (Зона 2) датчика `s_svr`, расположенного на правом берегу Енисея, попали датчики с левого берега. Однако все эти датчики расположены в хорошо проветриваемой относительно чистой части города, расположенной в долине реки Енисей. Поэтому понятна и их хорошая корреляция с датчиком `s_svr`, находящимся в аналогичных условиях. Зона влияния (Зона 3) датчика `s_kir` объединяет небольшую группу датчиков, расположенных в наиболее загрязненной части долины реки Енисей, а также на островах, через которые проходит мост с плотным транспортным потоком.

7. Благодарности. Авторы благодарят Алексея Токорева за предоставленные данные.

8. Заключение. В статье предложена статистически обоснованная корректировка первичных данных об уровне концентрации взвешенных частиц PM2.5 в приземном слое атмосферы г. Красноярска, получаемых оптическими датчиками станции CityAir. Для построения регрессионных моделей эталонными считались измерения, получаемые от анализаторов E-BAM, расположенных на тех же постах наблюдения, что и корректируемые датчики.

На основе проведенного анализа можно сделать следующие выводы.

1. При корректировке показаний датчиков необходимо учитывать метеорологические параметры. В нашем случае в линейной регрессионной модели лучше учитывать зависимость от температуры и давления. Точность измерения влажности на станциях CityAir не позволяет

учитывать этот показатель при корректировке, однако не исключает его влияния на показания датчика.

2. Параметры корректировки уровня концентрации PM2.5 с помощью регрессионной модели существенно зависят от сезона.

3. Показания оптических датчиков станций CityAir даже после корректировки не могут быть использованы в качестве эталонного оборудования для определения уровня концентрации PM2.5. В то же время их показания полностью отражают тренды показателей загрязнения, поэтому эти датчики могут использоваться для описания сценариев и прогнозов периодов повышенных концентраций взвешенных частиц в атмосфере городов.

4. Для оперативной корректировки значений концентрации PM2.5 датчиков станций CityAir достаточно использовать линейную многофакторную регрессионную модель на основе метода наименьших квадратов. Для научных ретроспективных исследований временных рядов концентраций PM2.5 рекомендуется использовать многофакторную регрессию, основанную на методе случайного леса.

Литература

1. Chae S., Shin J., Kwon S., Lee S., Kang S., Lee D. PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network // Science Report. 2021. vol. 11(1). no. 11952.
2. Kim B., Lim Y., Wan Cha J. Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based maching learning algorithms // Atmospheric Pollution Research. 2022. vol. 13(10). no. 101547.
3. Perrino C., Catrambone M., Pietrodangelo A. Influence of atmospheric stability on the mass concentration and chemical composition of atmospheric particles: A case study in Rome, Italy // Environment International. 2008. vol. 34. pp. 621–628.
4. Perez P., Menares C., Ramirez C. PM2.5 forecasting in Coyhaique, the most polluted city in the Americas // Urban Climate. 2020. vol. 32. no. 100608.
5. Zhang Zh., Wu L., Chen Y. Forecasting PM2.5 and PM10 concentrations using GMCN(1,N) model with the similar meteorological condition: Case of Shijiazhuang in China // Ecological Indicators. 2020. vol. 119. no. 106871.
6. Yang J., Yan R., Nong M., Liao J., Li F., Sun W. PM2.5 concentrations forecasting in Beijing through deep learning with different inputs model structures and forecast time // Atmospheric Pollution Research. 2021. vol. 12(9). no. 101168.
7. Лыченко Н.М., Великанова Л.И., Верзунов С.Н., Сороковая А.В. Модели прогноза уровня загрязнения атмосферного воздуха г. Бишкек // Вестник Кыргызско-Российского Славянского университета. 2021. Т. 21. № 4. С. 87–95.
8. Vlachogianni A., Kassomenos P., Karppinen A., Karakitsios S., Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki // Science of the total environment. 2011. vol. 409. pp. 1559–1571.
9. Iglesias-Gonzalez S., Huertas-Bolanos M.E., Hernandez-Paniagua I.Y., Mendoza A. Explicit Modeling of Meteorological Explanatory Variables in Short-Term Forecasting

of Maximum Ozone Concentrations via a Multiple Regression Time Series Framework // Atmosphere. 2020. vol. 11(12). no. 1304.

10. Zhou Q., Jiang H., Wang J., Zhou J. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network // Science of the Total Environment. 2014. vol. 496. pp. 264–274.
11. Аронов П.М. Оценка согласованного значения результатов межлабораторных измерений с минимальным увеличением их неопределённости // Эталоны. Стандартные образцы. 2019. Т. 15. № 4. С. 49–52.
12. Носков С.И. Метод максимальной согласованности в регрессионном анализе // Известия ТулГУ. Технические науки. 2021. № 10. С. 380–385.
13. Badura M., Batog P., Drzeniecka-Osiadacz A., Modzel P. Evaluation of Low-Cost Sensors for Ambient PM 2.5 Monitoring // Journal of Sensors. 2018. vol. 1. no. 5096540.
14. Shen H., Hou W., Zhu Y., Zheng S., Ainiwaer S., Shen G., Chen Y., Cheng H., Hu J., Wan Y., Tao S. Temporal and spatial variation of PM2.5 in indoor air monitored by low-cost sensors // Science of The Total Environment. 2021. vol. 770. no. 145304.
15. Jayaratne R., Liu X., Ahn K.H., Asumadu-Sakyi A., Fisher G., Gao J., Mabon A., Mazaheri M., Mullins B., Nyaku M., Ristovski Z., Scorgie Y., Thai P., Dunbabin M., Morawska L. Low-cost PM2.5 sensors: An assessment of their suitability for variousapplications // Aerosol and Air Quality Research. 2020. vol. 20. no. 3. pp. 520–532.
16. Gao M., Cao J., Seto E. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China // Environmental Pollution. 2015. vol. 199. pp. 56–65.
17. Wang W., Lung S., Liu Ch. Application of Machine Learning for the in-Field Correction of a PM2.5 Low-Cost Sensor Network // Sensors. 2020. vol. 20(17). no. 5002.
18. Bi J., Stowell J., et al. Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA // Environmental research. 2020. vol. 180. no. 108810.
19. E-BAM particulate monitor operation manual. Available at: https://metone.com/wp-content/uploads/2022/06/E-BAM-9805-Manual-Rev-G.pdf (accessed: 08.05.2023).
20. Environmental Technology Verification Report. Available at: https://archive.epa.gov/nrmrl/archive-etv/web/pdf/01_vr_metone_bam1020.pdf (accessed: 18.08.2023).
21. Заворуев В.В., Якубайлик О.Э., Кадочников А.А., Токарев А.В. Система мониторинга воздуха Красноярского научного центра СО РАН // Региональные проблемы дистанционного зондирования Земли: Материалы VII Международной научной конференции (г. Красноярск, 29 сентября – 2 октября 2020 г.). Красноярск: СФУ, 2020. С. 70–73.
22. Станция мониторинга воздуха CityAir. Электронный ресурс. URL: https://cityair.ru/ru/equipment/ (дата обращения: 08.05.2023).
23. Мониторинг состояния воздуха. Электронный ресурс. URL: https://asm.krasn.ru/ (дата обращения: 08.05.2023).
24. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. М. 1981. 696 с.
25. Себер Дж. Линейный регрессионный анализ. М. 1980. 456 с.
26. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика. 1981. 302 с.
27. Хасти Т., Тибришани Р., Фридман Д. Основы статистического обучения. Интеллектуальный анализ данных, логический вывод и прогнозирование. М: Вильямс, 2020. 768 с.

28. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems // Technometrics. 1970. vol. 12. no. 1. pp. 55–67.
29. Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society, Series B: Statistical Methodology. 1996. vol. 58. pp. 267–288.
30. Zou H., Hastie T. Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2005. vol. 67. no. 2. pp. 301–320.
31. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М. 1974. 416 с.
32. Vapnik V. The Nature of Statistical Learning Theory. New- York: Springer Verlag N.Y. 1995. 188 p. DOI: 10.1007/978-1-4757-2440-0.
33. Vapnik V., Golowich S., Smola A. Support Vector Method for Function Approximation Regression Estimation and Signal // Advances in Neural Information Processing Systems. 1996. p. 281–287.
34. Morgan J.N. Sonquist J.A. Problems in the analysis of survey data, and a proposal // Journal of the American Statistical Association. 1963. vol. 58. pp. 415–434.
35. Breiman L., Friedman J., Olshen R., Stone C. Classification and regression trees. CA: Wadsworth and Brooks/Cole Advanced Books and Software. 1984. 368 p.
36. Breiman L. Bagging Predictors // Machine Learning. 1996. vol. 24. pp. 123–140.
37. Breiman L. Random Forests // Machine Learning. 2001. vol. 45. pp. 5–32.
38. Wallace J., Kanaroglou P. The effect of temperature inversions on ground-level nitrogen dioxide (NO2) and fine particulate matter (PM2.5) using temperature profiles from the Atmospheric Infrared Sounder (AIRS) // Science of The Total Environment. 2009. vol. 407. no. 18. pp. 5085–5095.
39. Санитарные правила и нормы СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания». 2021 г. URL: https://docs.cntd.ru/document/573500115 (дата обращения: 26.01.2024).
40. Akaike H. A new look at statistical model identification // IEEE Transactions on Automatic Control. 1974. vol. 19. pp. 716–723.
41. Stoica P., Selen Y. Model-order selection: a review of information criterion rules // IEEE Signal Processing Magazine. 2004. vol. 21. no. 4. pp. 36–47.

Карепова Евгения Дмитриевна — канд. физ.-мат. наук, доцент, ведущий научный сотрудник, Отдел "Вычислительная математика", Институт вычислительного моделирования СО РАН. Область научных интересов: вычислительная математика, математическое моделирование природных и технических процессов, анализ данных, высокопроизводительное программирование. Число научных публикаций — 175. e.d.karepova@icm.krasn.ru; Академгородок, 50/44, 660036, Красноярск, Россия; р.т.: +7(391)290-7476.

Петракова Виктория Сергеевна — канд. физ.-мат. наук, научный сотрудник, Отдел "Вычислительная математика", Институт вычислительного моделирования СО РАН. Область научных интересов: математическое моделирование, вычислительная математика, анализ данных. Число научных публикаций — 22. rikka@icm.krasn.ru; Академгородок, 50/44, 660036, Красноярск, Россия; р.т.: +7(923)267-3748.

Поддержка исследований. Исследование выполнено при финансовой поддержке «Красноярского краевого фонда поддержки научной и научно-технической деятельности» в рамках реализации научного проекта № 2022110809055 «Оценка эффективности использования сети недорогих сенсорных датчиков для сбора данных о загрязнениях в пограничных слоях атмосферы на основе анализа наблюдений за динамикой концентрации взвешенных частиц PM2.5».

DOI 10.15622/ia.23.2.2

E. KAREPOVA, V. PETRAKOVA

STATISTICAL SUBSTANTIATION OF THE REVISING OF READINGS BY THE CITYAIR STATION OF PM2.5 CONCENTRATION LEVELS IN THE ATMOSPHERIC BOUNDARY LAYER OF THE CITY

Karepova E., Petrakova V. **Statistical Substantiation of the Revising of Readings by the CityAir Station of PM2.5 Concentration Levels in the Atmospheric Boundary Layer of the City.**

Abstract. As a marker characterizing air pollution in the surface layer of the atmosphere of modern cities, the concentration level of particulate matter with a diameter of 2.5 microns or less (Particulate Matter, PM2.5) is often used. The paper discusses the practice of using a relatively cheap optical sensor, which is part of the CityAir station, to measure the concentration of PM2.5 in an urban environment. The article proposes a statistically justified correction of the primary data obtained by CityAir stations on the values of the concentration of suspended particles PM2.5 in the surface layer of the atmosphere of Krasnoyarsk. For the construction of regression models, measurements obtained from E-BAM analyzers located at the same observation posts as the corrected sensors were considered as a reference. For the analysis, primary data was used 1) from 9 automated observation posts of the regional departmental information and analytical system of data on the state of the environment of the Krasnoyarsk Territory (KVIAS); 2) from the 21st CityAir station of the monitoring system of the Krasnoyarsk Scientific Center of the Siberian Branch of the Russian Academy of Sciences. The paper demonstrates that when correcting sensor readings, it is necessary to take into account meteorological indicators. In addition, it is shown that the regression coefficients significantly depend on the season. Supervised learning methods are compared for solving the problem of correcting the readings of inexpensive sensors. Additional information on the results of data analysis, which was not included in the text of the article, is available on the electronic resource https://asm.krasn.ru/.

Keywords: particulate Matter, PM2.5 concentration level, supervised learning, regression models, sensor system revising.

References

1. Chae S., Shin J., Kwon S., Lee S., Kang S., Lee D. PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network. Science Report. 2021. vol. 11(1). no. 11952.
2. Kim B., Lim Y., Wan Cha J. Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based maching learning algorithms. Atmospheric Pollution Research. 2022. vol. 13(10). no. 101547.
3. Perrino C., Catrambone M., Pietrodangelo A. Influence of atmospheric stability on the mass concentration and chemical composition of atmospheric particles: A case study in Rome, Italy. Environment International. 2008. vol. 34. pp. 621–628.
4. Perez P., Menares C., Ramírez C. Perez P., Menares C., Ramirez C. PM2.5 forecasting in Coyhaique, the most polluted city in the Americas. Urban Climate. 2020. vol. 32. no. 100608.
5. Zhang Zh., Wu L., Chen Y. Forecasting PM2.5 and PM10 concentrations using GMCN(1,N) model with the similar meteorological condition: Case of Shijiazhuang in China. Ecological Indicators. 2020. vol. 119. no. 106871.

6. Yang J., Yan R., Nong M., Liao J., Li F., Sun W. PM2.5 concentrations forecasting in Beijing through deep learning with different inputs model structures and forecast time. Atmospheric Pollution Research. 2021. vol. 12(9). no. 101168.
7. Lychenko N.M., Velikanova L.I., Verzunov S.N., Sorokovaya A.V. [Models for predicting the level of atmospheric air pollution in Bishkek]. Vestnik Kyrgyzsko-Rossiyskogo Slavyanskogo universiteta – Bulletin of the Kyrgyz-Russian Slavic University. 2021. vol. 21. no. 4. pp. 87–95. (In Russ.).
8. Vlachogianni A., Kassomenos P., Karppinen A., Karakitsios S., Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Science of the total environment. 2011. vol. 409. pp. 1559–1571.
9. Iglesias-Gonzalez S., Huertas-Bolanos M.E., Hernandez-Paniagua I.Y., Mendoza A. Explicit Modeling of Meteorological Explanatory Variables in Short-Term Forecasting of Maximum Ozone Concentrations via a Multiple Regression Time Series Framework. Atmosphere. 2020. vol. 11(12). no. 1304.
10. Zhou Q., Jiang H., Wang J., Zhou J. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Science of the Total Environment. 2014. vol. 496. pp. 264–274.
11. Aronov P.M. [Estimation of consensus value of interlaboratory measurement results accompanied by a minimum increase in associated uncertainty]. Etalony. Standartnye obrazcy – Standards. Reference materials. 2019. vol. 15. no. 4. pp. 49–52. (In Russ.).
12. Noskov S.I. [Maximum Consistency Method in Regression Analysis]. Izvestija TulGU. Tehnicheskie nauki – Proceedings of TulGU. Technical science. 2021. № 10. С. 380–385. (In Russ.).
13. Badura M., Batog P., Drzeniecka-Osiadacz A., Modzel P. Evaluation of Low-Cost Sensors for Ambient PM 2.5 Monitoring. Journal of Sensors. 2018. vol. 1. no. 5096540.
14. Shen H., Hou W., Zhu Y., Zheng S., Ainiwaer S., Shen G., Chen Y., Cheng H., Hu J., Wan Y., Tao S. Temporal and spatial variation of PM2.5 in indoor air monitored by low-cost sensors. Science of The Total Environment. 2021. vol. 770. no. 145304.
15. Jayaratne R., Liu X., Ahn K.H., Asumadu-Sakyi A., Fisher G., Gao J., Mabon A., Mazaheri M., Mullins B., Nyaku M., Ristovski Z., Scorgie Y., Thai P., Dunbabin M., Morawska L. Low-cost PM2.5 sensors: An assessment of their suitability for variousapplications. Aerosol and Air Quality Research. 2020. vol. 20. no. 3. pp. 520–532.
16. Gao M., Cao J., Seto E. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. Environmental Pollution. 2015. vol. 199. pp. 56–65.
17. Wang W., Lung S., Liu Ch. Application of Machine Learning for the in-Field Correction of a PM2.5 Low-Cost Sensor Network. Sensors. 2020. vol. 20(17). no. 5002.
18. Bi J., Stowell J., et al. Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. Environmental research. 2020. vol. 180. no. 108810.
19. E-BAM particulate monitor operation manual. Available at: https://metone.com/wp-content/uploads/2022/06/E-BAM-9805-Manual-Rev-G.pdf (accessed: 08.05.2023).
20. Environmental Technology Verification Report. Available at: https://archive.epa.gov/nrmrl/archive-etv/web/pdf/01_vr_metone_bam1020.pdf (accessed: 18.08.2023).
21. Zavoruev V.V., Yakubailik O.E., Kadochnikov A.A., Tokarev A.V. [Air monitoring system of the Krasnoyarsk Scientific Center of the SB RAS]. Regionalnyye problem distantsionnogo zondirovaniya Zemli: Materialy VII Mezhdunarodnoy nauchnoy konferentsii [Regional problems of remote sensing of the Earth: Proceedings of the VII International Scientific Conference]. Krasnoyarsk: SFU. 2020. pp. 70-73. (In Russ.).

22. Stancija monitoringa vozduha CityAir. Jelektronnyj resurs [Air Monitoring Station CityAir. Electronic resource]. Available at: https://cityair.ru/ru/equipment/ (accessed: 08.05.2023). (In Russ.).
23. Monitoring sostojanija vozduha. Jelektronnyj resurs [Air state monitoring. Electronic resource.] Available at: https://asm.krasn.ru/ (accessed: 08.05.2023). (In Russ.).
24. Tyuki J. Analiz rezul'tatov nabljudenij. Razvedochnyj analiz [Analysis of observation results. Exploratory analysis]. Moscow. 1981. 696 p. (In Russ.).
25. Seber G. Linejnyj regressionnyj analiz [Linear regression analysis]. Moscow. 1980. 456 p. (In Russ.).
26. Demidenko E.Z. Linejnaja i nelinejnaja regressii. [Linear and nonlinear regression]. Moscow: Finansy i Statistika. 1981. 302 p. (In Russ.).
27. Hastie Tr., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer. 2017. 768 p.
28. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970. vol. 12. no. 1. pp. 55–67.
29. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 1996. vol. 58. pp. 267–288.
30. Zou H., Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2005. vol. 67. no. 2. pp. 301–320.
31. Vapnik V., Chervonenkis A. Teorija raspoznavanija obrazov (statisticheskie problem obuchenija) [Pattern recognition theory (statistical learning problems)]. M.: Nauka. 1974. 416 p. (In Russ.).
32. Vapnik V. The Nature of Statistical Learning Theory. New- York: Springer Verlag N.Y. 1995. 188 p. DOI: 10.1007/978-1-4757-2440-0.
33. Vapnik V., Golowich S., Smola A. Support Vector Method for Function Approximation Regression Estimation and Signal. Advances in Neural Information Processing Systems. 1996. pp. 281–287.
34. Morgan J.N. Sonquist J.A. Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association. 1963. vol. 58. pp. 415–434.
35. Breiman L., Friedman J., Olshen R., Stone C. Classification and regression trees. CA: Wadsworth and Brooks/Cole Advanced Books and Software. 1984. 368 p.
36. Breiman L. Bagging Predictors. Machine Learning. 1996. vol. 24. pp. 123–140.
37. Breiman L. Random Forests. Machine Learning. 2001. vol. 45. pp. 5–32.
38. Wallace J., Kanaroglou P. The effect of temperature inversions on ground-level nitrogen dioxide (NO2) and fine particulate matter (PM2.5) using temperature profiles from the Atmospheric Infrared Sounder (AIRS). Science of The Total Environment. 2009. vol. 407. no. 18. pp. 5085–5095.
39. anitary rules and norms SanPiN 1.2.3685-21 «Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans». 2021. Available at: https://docs.cntd.ru/document/573500115 (accessed: 26.01.2024). (In Russ.).
40. Akaike H. A new look at statistical model identification. IEEE Transactions on Automatic Control. 1974. vol. 19. pp. 716–723.
41. Stoica P., Selen Y. Model-order selection: a review of information criterion rules. IEEE Signal Processing Magazine. 2004. vol. 21. no. 4. pp. 36–47.

Karepova Eugeniya — Ph.D., Associate Professor, Leading researcher, Department "Computational mathematics", Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences. Research interests: computational mathematics, mathematical modeling of natural and technical processes, data analysis, high-performance programming. The number of publications — 175. e.d.karepova@icm.krasn.ru; 50/44, Akademgorodok, 660036, Krasnoyarsk, Russia; office phone: +7(391)290-7476.

Petrakova Viktoriya — Ph.D., Researcher, Department "Computational mathematics", Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences. Research interests: mathematical modeling, computational mathematics, data analysis. The number of publications — 22. rikka@icm.krasn.ru; 50/44, Akademgorodok, 660036, Krasnoyarsk, Russia; office phone: +7(923)267-3748.

Acknowledgements. The study was financially supported by the Krasnoyarsk Regional Fund of Science and Technology Support, project No. 2022110809055 «Evaluation of the effectiveness of using a network of the low-cost sensors to collect data on pollution in the atmospheric boundary layers based on an analysis of observations of the dynamics of the concentration of suspended particles PM2.5».