УДК 004.89

МОДЕЛИ ПРОГНОЗА УРОВНЯ ЗАГРЯЗНЕНИЯ АТМОСФЕРНОГО ВОЗДУХА Г. БИШКЕК

Н.М. Лыченко, Л.И. Великанова, С.Н. Верзунов, А.В. Сороковая

Исследована динамика изменения концентрации вредных веществ (в частности, твердых частиц PM2.5) в приземном слое атмосферы г. Бишкек, а также индекс качества воздуха AQI. Построена модель для прогноза их уровней на основе актуальных для города данных. Представлены четыре варианта моделей для прогнозирования уровня загрязнения атмосферного воздуха г. Бишкек: ARIMA-модели (интегрированные модели авторегрессии – скользящего среднего) для краткосрочного прогноза концентраций PM2.5 и AQI; линейные мультирегрессионные модели процессов загрязнения на основе регрессионного анализа метеорологических факторов и концентраций частиц PM2.5; модель краткосрочного прогноза концентраций PM2.5 с учетом метеорологических факторов на основе обобщенно-регрессионной нейронной сети GRNN; модель среднесрочного прогноза классов AQI на основе классификатора индекса качества воздуха на базе LSTM-нейронных сетей. Показаны некоторые особенности и возможности каждой модели, а также приведены результаты прогнозирования. Информационная база исследования формировалась на основе данных загрязнения атмосферного воздуха г. Бишкек за период с 09.02.2019 по 31.03.2020, опубликованных на сайте «AirNow» и архивных данных метеослужбы.

Ключевые слова: прогнозирование; концентрации PM2.5; индекс качества воздуха AQI; метеорологические параметры; ARIMA-модель; мультирегрессионная модель; обобщенно-регрессионная нейронная сеть GRNN; LSTM-нейронная сеть; ошибка прогноза.

БИШКЕК ШААРЫНЫН АТМОСФЕРАЛЫК АБАСЫНЫН БУЛГАНУУ ДЕҢГЭЭЛИН БОЛЖОЛДОО МОДЕЛИ

Н.М. Лыченко, Л.И. Великанова, С.Н. Верзунов, А.В. Сороковая

Бул макалада Бишкек шаарынын атмосферасынын жерге жакын катмарында зыяндуу заттардын концентрациясынын (тактап айтканда PM2.5 катуу бөлүкчөлөрүнүн) өзгөрүү динамикасы, ошондой эле AQI абанын сапатынын индекси изилдөөгө алынды. Шаар үчүн актуалдуу болгон маалыматтардын негизинде алардын деңгээлин болжолдоо үчүн модель түзүлдү. Бишкек шаарынын атмосфералык абасынын булгануу деңгээлин болжолдоо үчүн моделдердин төрт варианты сунушталды: PM2.5 жана AQI концентрацияларын кыска мөөнөткө болжолдоо үчүн ARIMA-моделдери (авторегрессиянын интеграцияланган моделдери – жылып туруучу орточо); метеорологиялык факторлорго жана PM2.5 бөлүкчөлөрүнүн концентрациясына регрессивдүү талдоо жүргүзүүнүн негизинде булгануу процесстеринин тилкелик мультирегрессивдүү моделдери; GRNN жалпыланган-регрессивдүү нейрондук тармактын негизинде метеорологиялык факторлорду эске алуу менен PM2.5 концентрациясын кыска мөөнөткө болжолдоо модели; LSTM-нейрондук тармактардын базасында абанын сапатынын индексинин классификаторунун негизинде AQI класстарын орто мөөнөттүк болжолдоо модели. Ар бир моделдин айрым өзгөчөлүктөрү жана мүмкүнчүлүктөрү көрсөтүлдү, ошондой эле болжолдоонун жыйынтыктары берилди. Изилдөөнүн маалымат базасы «AirNow» сайтында жарыяланган 2019-жылдын 9-февралынан 2020-жылдын 31-мартына чейинки мезгилдеги Бишкек шаарынын атмосфералык абасынын булгануусу тууралуу маалыматтардын жана метео кызматтын архивдик маалыматтарынын негизинде түзүлдү.

Түйүндүү сөздөр: болжолдоо; PM2.5 концентрациясы; индекс качества воздуха AQ I абанын сапатынын индекси; метеорологиялык параметрлер; ARIMA-модели; мультирегрессивдүү модель; GRNN жалпыланган-регрессивдүү нейрондук тармагы; LSTM-нейрондук тармагы; болжолдоо катас.

FORECASTING MODEL THE ATMOSPHERIC AIR POLLUTION OF BISHKEK

N.M. Lychenko, L.I. Velikanova, S.N. Verzunov, A.V. Sorokovaja

The article ivestigates the dynamics of changes in the concentrations of harmful substances (in particular, PM2.5 particulate matter) in the surface layer of the atmosphere in Bishkek as well as the air quality index AQI. It is built the model to predict their levels based on relevant for the city data. The paper presents four versions of models for predicting the level of air pollution in Bishkek: the most popular and effective statistical ARIMA models (integrated autoregressive moving average models) for short-term forecasting of PM2.5 and AQI concentrations; linear multi-regression models of pollution processes based on regression analysis of meteorological factors and concentrations of PM2.5 particles, model of short-term forecast of PM2.5 concentrations taking into account meteorological factors based on generalized regression neural network GRNN, model of medium-term forecast AQI based on air quality index classifier based on LSTM neural networks. Some features and capabilities of each model are discussed, as well as the forecasting results. The information base of the study was formed on the basis of data on atmospheric air pollution in Bishkek for the period 09.02.2019-31.03.2020, published on the AirNow website and archived data from the meteorological service.

Keywords: forecasting of time series; PM2.5 concentrations; the air quality index AQI; meteorological parameters; ARIMA-model; multi-regression model; generalized regression neural network GRNN; LSTM-neural network; forecasting error.

Введение. В последнее время в литературе большое внимание уделяется исследованиям процессов загрязнения атмосферного воздуха городов в связи с ухудшающейся экологической ситуацией. Отмечено, что «физико-географические и климатические условия г. Бишкек, а также относительная замкнутость Чуйской долины, способствуют возникновению интенсивных приземных и приподнятых инверсий, что ведет к формированию высокого потенциала загрязнений» [1]. В последние годы из-за быстрого роста плотности населения, плотности хаотичной застройки, возросшего потребления энергии на отопление, увеличения количества автотранспорта качество атмосферного воздуха в городе значительно ухудшилось. Поэтому решение задачи анализа динамики изменения концентраций вредных веществ, в частности, твердых частиц PM2.5 и индекса качества воздуха (Air Quality Index, AQI) города Бишкек, а также построения моделей для прогноза их уровней, представляет особый интерес.

Универсальных моделей для прогноза концентрации PM2.5 и AQI быть не может, поскольку в них необходимо учитывать региональные природные, экономические, антропогенные и климатические особенности территории. В литературе опубликовано немало работ, связанных с построением моделей прогноза качества воздуха для различных регионов и городов. Для г. Бишкек таких работ практически нет, главным образом, ввиду того, что открытая информация об уровне загрязненности атмосферного воздуха в городе стала публиковаться на сайте «AirNow» [2] лишь с февраля 2019 г. Основываясь на этой информации, авторами выполнен ряд работ, связанных с оценкой временной изменчивости концентраций в атмосферном воздухе твердых частиц PM2.5 и индекса качества воздуха AQI. На основе данных наблюдений для различных периодов 2019–2020 гг. разработаны: интегрированные модели авторегрессии – скользящего среднего (AutoRegressive Integrated Moving Average model, ARIMA-модели) для краткосрочного прогноза концентраций в атмосферном воздухе твердых частиц PM2.5 и индекса качества воздуха AQI [3, 4], линейные мультирегрессионные модели процессов загрязнения на основе регрессионного анализа метеорологических факторов и концентраций частиц PM2.5 [5], модель краткосрочного прогноза концентраций PM2.5 с учетом метеорологических факторов на основе обобщенно-регрессионной нейронной сети GRNN [5], модель среднесрочного прогноза класса AQI на основе классификатора индекса качества воздуха на базе LSTM-нейронных сетей [6]. В настоящей работе обсуждены некоторые особенности и возможности каждой из перечисленных моделей, а также приведены результаты прогнозирования PM2.5 и AQI на их основе.

Модели и методы. ARIMA-модели – наиболее популярные, и эффективные статистические модели для прогноза временных рядов (ВР). В основу этих моделей положена идея, что «будущие значения временного ряда можно представить как некоторую линейную функцию прошлых наблюдений, дополненную случайной ошибкой (белым шумом). В силу того, что инерционность во ВР загрязнений велика, ARIMA-модели широко используют для прогноза загрязнений атмосферного воздуха.

В общем виде ARIMA(*p,d,q*)-модель имеет вид:

$$\phi(B)(1-B)^d y_{ft} = \theta(B)\varepsilon_t ,$$

где y_{ft} – значения модельного ряда в момент времени t; ε_t – случайная ошибка с нулевым средним и постоянной дисперсией; $\phi(B)$, $\theta(B)$ – полиномы степени p и q; B – лаговый оператор, с учетом которого $By_{ft} = y_{ft-1}$, $B^2y_{ft} = y_{ft-2}$, … , $B^dy_{ft} = y_{ft-d}$, $B^j\varepsilon_{ft} = \varepsilon_{ft-j}$, $j=0,1,…$; d – порядок взятия последовательной разности $\Delta y_t = y_{ft-1} – y_{ft}$, которая вычисляется для того чтобы сделать ряд стационарным» [7]. Для того, чтобы определить порядок (*p, d, q*) ARIMA-модели, обычно используются автокорреляционная (ACF) и частичная автокорреляционная (PACF) функции ВР наблюдений. После выбора порядка модели, продолжая следовать методологии Бокса–Дженкинса, «оценивают параметры модели, исходя из минимума «разногласий» между модельным ВР и ВР наблюдений. Затем адекватность модели проверяется на основе статистической обработки прогнозных значений ряда и анализа остатков, т. е. разницы между значением ряда y_t и его предсказанным на модели значением y_{ft}: $e_t = y_t – y_{ft}$. Если модель не адекватна, должна быть определена новая предварительная модель. Этот трехэтапный процесс построения модели обычно повторяется несколько раз, пока удовлетворительная модель будет окончательно выбрана» [7]. Как показано в [4], выбранная таким образом модель может быть использована для прогнозирования степени загрязненности атмосферного воздуха.

Мультирегрессионная модель прогноза. ARIMA-модели позволяют учитывать динамику процесса загрязнения, но не учитывают метеорологические факторы. Поэтому представляется интересным использовать модели множественной регрессии, которые позволяют учитывать большое количество факторов, характеризующих причинно-следственные связи, имеющие место в объекте исследования (в процессе загрязнений воздуха в нашем случае).

Как показано в [8], «общее уравнение линейной множественной регрессии может быть записано как:

$$y_{ft}=b+k_1x_{1t}+ k_2x_{2t}+ \ldots + k_mx_{mt}+e_t,$$

где b – константа регрессии; $k_1, k_2, \ldots, k_m$ – коэффициенты регрессии; m – количество независимых переменных (факторов, предикатов). Значения константы и коэффициентов определяются с использованием метода наименьших квадратов, который минимизирует ошибку $e_t = (y_t – y_{ft})$ или остатки модели. Для построения регрессионной модели обычно используют метод ступенчатой (пошаговой) регрессии, суть которого заключается в отборе из большого количества предикатов x небольшой подгруппы переменных, которые вносят наибольший вклад в вариацию зависимой переменной» [8]. При построении мультирегрессионной модели в работе [5] концентрации PM2.5 рассматривались как зависимые переменные, а метеопараметры (температура воздуха, температура точки росы, влажность воздуха, скорость ветра, давление, интенсивность осадков) рассматривались как независимые переменные.

Модель прогноза на основе обобщенно-регрессионной нейронной сети. Модели на основе искусственных нейронных сетей (ИНС), активно используются в прогнозировании, поскольку они обладают способностью находить зависимости между данными путем анализа большого количества подобных примеров или паттернов. Выбор обобщенно-регрессионной нейронной сети (Generalized Regression Neural Network, GRNN) [9], как варианта ИНС для прогнозирования степени загрязненности атмосферного воздуха, обусловлен следующими ее преимуществами: архитектура сети фиксирована и не нуждается в определении, и эта сеть характеризуется достаточно высокой скоростью обучения. В литературе достаточно подробно описана архитектура этой сети: «GRNN-сеть содержит два слоя: радиально-базисный слой с числом нейронов, равным числу элементов обучающего множества и линейный слой, а окончательная выходная оценка сети получается как взвешенное среднее выходов по всем обучающим наблюдениям, где величины весов отражают расстояние от этих наблюдений до той точки, в которой производится оценивание (и, таким образом, более близкие точки вносят больший вклад в оценку)» [9].

Модель прогноза на основе LSTM-нейронной сети. На степень загрязнения воздуха влияют метеорологические условия не только на текущий момент, но и история изменения этих условий. Это говорит об инерционности процессов загрязнения. Очевидно предположить, что сети, учитывающие историю наблюдений, будут работать лучше, чем сети, основанные на оценке исключительно текущих данных. Для анализа исторических данных, как правило, используются рекуррентные нейронные сети. Однако степень загрязнения воздуха может реагировать на различные факторы с разной скоростью. Поэтому история каждого фактора должна быть оценена по-разному – для каких-то факторов важны только последние данные, влияние других может сказываться продолжительное время. Учитывая этот факт, для прогнозирования степени загрязненности воздуха интересным представляется использование LSTM-сетей (Long Short-Term Memory, LSTM – долгая краткосрочная память) [10, 11]. Их архитектура содержит так называемые «фильтры, которые в процессе обучения настраиваются сохранять/забывать информацию выборочно о различных факторах и, таким образом, могут обнаруживать как длинные, так и короткие шаблоны в данных» [11].

Структура используемой LSTM-сети представлена в [6]. «Сеть содержит: 1) входной слой длиной 6 (6 признаков входного вектора BB), который принимает последовательность длиной S векторов признаков; 2) первый скрытый слой – слой прямого распространения с числом нейронов 100 и тангенциальной функцией активации, отображает входные векторы в векторы большей длины, для дальнейшего внесения шума в данные; 3) второй скрытый слой – слой регуляризации, который меняет некоторый процент значений выхода предыдущего слоя для предотвращения переобучения (вносит шумы); 4) третий скрытый слой – LSTM-сеть с 50 нейронными модулями и тангенциальной функцией активации как основной классифицирующий слой; 5) четвертый скрытый слой – слой прямого распространения с числом нейронов 10 и тангенциальной функцией активации; 6) выходной слой – слой с 2 нейронами и функцией активации SOFTMAX. Функция взвешивает входы и предсказывает вероятности активации каждого нейрона. При этом сумма выходов нейронов всегда равна 1. Все слои сети полносвязные, т. е. каждый нейрон имеет связь с каждым предыдущим нейроном, а для рекуррентных слоев (LSTM-сеть) каждый вход слоя также связан с каждым выходом слоя» [6].

Некоторые решения задач прогнозирования

Данные для исследования. Данные, представленные в настоящей работе и использованные в работах [3–6] – данные о концентрациях PM2.5 в mkg/m3 и индексе качества воздуха AQI г. Бишкек, размещенные на сайте [1]. Этот сайт публикует данные, начиная с 06.02.2019 по настоящее время (период измерения – 1 час) в CSV- формате.

На рисунке 1 показано изменение содержания PM2.5 в атмосферном воздухе и индекс качества воздуха г. Бишкек за период с 09.02.2019 по 31.03.2020 г. с интервалом в 3 часа. На графиках отчетливо отмечаются два скачка в значениях PM2.5 и AQI: 23.03.2019 г. значения резко падают и 01.11. 2019 г. – значения резко возрастают. Указанные даты – конец и начало отопительного сезона в городе. Среднее (математическое ожидание) *mean* и стандартное (среднеквадратическое) отклонение σ за весь период наблюдений, а также за периоды: 09.02.2019–23.03.2019, 24.03.2019–31.10.2019, 1.11.2019–31.03.2020 годы представлены в таблице 1.

Временные ряды (ВР) на рисунке 1 явно нестационарные. Это подтверждают и проведенные с использованием функции MatLab *adftest* расширенные тесты Дики-Фуллера [12]. Для анализа исходных рядов построены автокорреляционные и частичные автокорреляционные функции (ACF и PACF). Анализ ACF PM2.5 (рисунок 2) показывает, что они не затухают быстро, что также характерно для нестационарных ВР, и содержит периодичность в 8 лагов, что соответствует 24 часам. Последний значимый лаг на PACF PM2.5 равен 80, что соответствует 10 суткам и не согласуется с данными [13], согласно которым время оседания частиц PM2.5 составляет 5 суток. ACF AQI показывает, что ВР также содержит периодичность в 8 лагов.

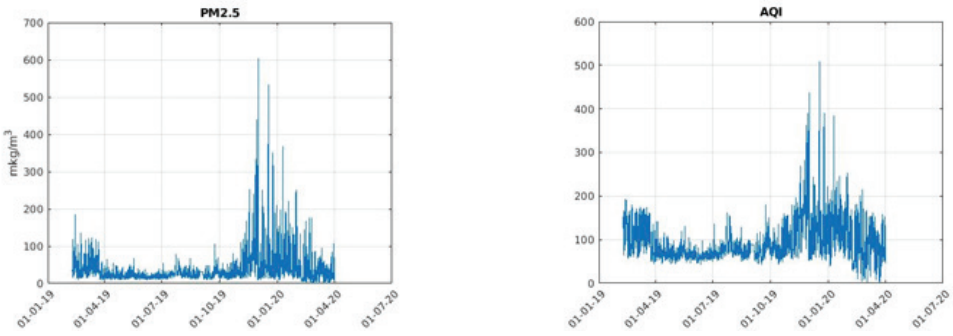

Рисунок 1 – Значения концентраций PM2.5 и индекса качества воздуха
в г. Бишкек за период 09.02.2019–31.03.2020 гг.

Таблица 1 – Средние значения и стандартные отклонения концентраций PM2.5 и AQI

Период наблюдений, год	PM2.5		AQI	
	mean, mkg/m^3	σ, mkg/m^3	mean	σ
09.02.2019–25.11.2019	29.6322	21.9179	85.2881	32.1227
09.02.2019–23.03.2019	43.9942	23.8089	115.1541	32.9580
24.03.2019–31.10.2019	22.2688	10.4384	69.9941	21.1958
01.11.2019–31.03.2020	40.7214	122.0221	102.2294	128.8958

Уровень загрязненности атмосферного воздуха в немалой степени зависит от метеорологических факторов. Исследования, широко представленные в литературе, показывают, что «высокие уровни содержания загрязняющих веществ наблюдаются в период длительных неблагоприятных метеорологических условий, которым свойственны температурные инверсии, слабые ветры, туманы» [4].

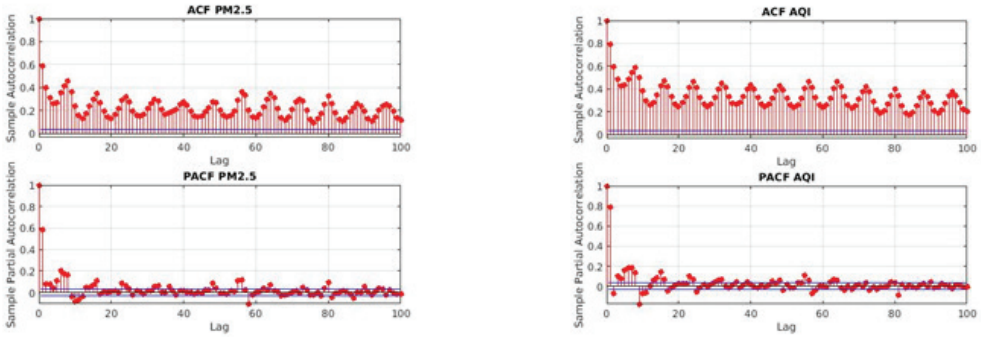

Рисунок 2 – Автокорреляционные (ACF) и частичные автокорреляционные функции исходных ВР

Для учета влияния метеорологических факторов на уровень загрязнения атмосферного воздуха, авторами использованы данные с сервера международного обмена NOAA (США) в формате SYNOP. На этом сервере хранятся измерения метеопараметров, произведенные более чем 90000 наземными станциями мира, в том числе, метеостанцией Бишкек (WMO_ID=38353). Данные получены через сайт [14] в виде Excel-файла за период с 09.02.2019 по 31.03.2020 г. В [15] «выполнена оценка влияния метеорологических факторов (таких, как скорость ветра, температура, относительная влажность воздуха, температура точки росы, интенсивность осадков и атмосферное давление) на процесс загрязнения воздуха г. Бишкек частицами PM2.5 в период с февраля по ноябрь 2019 г., и выявлены умеренные корреляции (как положительные, так и отрицательные) между концентрациями

PM2.5 и метеорологическими параметрами, измеренными в текущий и прошлые сроки измерений» [15]. Эти оценки учтены при разработке моделей прогноза уровня загрязненности воздуха с учетом влияния метеофакторов.

Прогнозирование на ARIMA-модели. Для иллюстрации прогностических возможностей ARIMA-моделей в [3] представлены результаты прогнозирования индекса качества воздуха AQI на основе ARIMA-модели, построенной на данных за период с 24.03.2019 по 29.08.2019 г. Поскольку исходный ВР нестационарный, была произведена декомпозиция исходного ряда на трендовую, сезонную и недетерминированную составляющую. Недетерминированная составляющая ВР была разделена на две части: обучающую выборку и тестовую. В результате анализа частичной автокорреляционной функции PACF и после проведения вычислений с различными значениями коэффициентов *p, d* и *q*, в ARIMA-моделях получены различные варианты прогнозов. Наименьшая среднеквадратическая ошибка $MSE = 17.14$ и наилучший коэффициент детерминации $R^2 = 0.88$ соответствуют варианту модели: $p = 4, d = 1, q = 0$ [3], результат прогноза с использованием которой также представлен на рисунке 3.

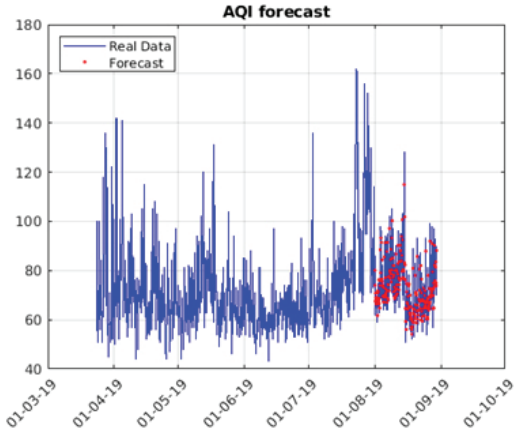

Рисунок 3 – Динамика реальных и прогнозных AQI

Прогнозирование на линейной мультирегрессионной модели. Возможности прогнозирования загрязнений воздуха на мультирегрессионных моделях показаны в [5]. Архивные данные были разделены на две части: обучающее множество – данные измерения Pm2.5 и метеорологических параметров по г. Бишкек за период с 24.03.2019 по 31.07.2019 г.; тестовое множество – данные измерения за период с 01.08.2019 по 18.08.2019 г. В [15] показано, что для летнего периода времени (24.03.2019–31.10.2019 гг.) корреляция между концентрациями PM2.5 и давлением несущественная, а «наилучшие взаимосвязи обнаружены между концентрациями PM2.5 и измерениями температуры воздуха T в предшествующий срок измерений *i-1* (т. е. тремя часами ранее), температуры точки росы Td в срок измерений *i-2* (шестью часами ранее), интенсивностью осадков PR в срок измерений *i-2*, влажностью RH в срок измерений *i-1*, скоростью ветра Ws в срок измерений *i-5*» [15]. Используя эту информацию, на обучающей выборке 24.03.2019–31.07.2019 гг. была построена линейная мультирегрессионная модель (см. таблицу 2), учитывающая все перечисленные относительно существенные факторы.

В таблице 2 также представлены коэффициенты множественной корреляции R между концентрациями PM2.5 и метеорологическими параметрами, средняя квадратическая ошибка модели MSE, средняя абсолютная ошибка MAE, средняя абсолютная процентная ошибка MAPE.

Затем, используя метод пошаговой множественной линейной регрессии, получена итоговая регрессионная модель, также представленная в таблице 2:

$$PM2.5(i) = b + k1*Td(i–2) + k_2*T(i–1) + k_3*PR(i–2) + k_4*PM2.5(i–1) + k_5*PM2.5(i–2).$$

Таблица 2 – Линейные мультирегрессионные модели

Teaching set: 24.03.2019–31.07.2019 гг.									
*PM2.5(i)= b+ k1*Td(i-2)+ k2*T(i-1)+ k3*PR(i-2)+ k4*Ws(i-5)+ k5*RH(i-1)*									
b	*k1*	*k2*	*k3*	*k4*	*k5*	*R*	*MAPE*	*MAE*	*MSE*
13.1054	-0.4092	0.4569	-1.9980	-0.0671	0.0550	0.3183	0.284	7.440	134.4
*PM2.5(i)= b+ k1*Td(i-2)+ k2*T(i-1)+ k3*PR(i-2)+ k4*PM2.5(i-1)+ k5*PM2.5(i-2)*									
b	*k1*	*k2*	*k3*	*k4*	*k5*	*R*	*MAPE*	*MAE*	*MSE*
10.4621	-0.4235	0.4930	-2.9192	0.0537	0.0432	0.3514	0.2772	5.867	70.11
Testing set: 01/08/2019-18/08/2019 гг.									
*PM2.5(i)= b+ k1*Td(i-2)+ k2*T(i-1)+ k3*PR(i-2)+ k4*PM2.5(i-1)+ k5*PM2.5(i-2)*									
b	*k1*	*k2*	*k3*	*k4*	*k5*	*R*	*MAPE*	*MAE*	*MSE*
10.4621	-0.4235	0.4930	-2.9192	0.0537	0.0432	0.3514	0.2087	4.395	43.57

В этой модели исключены факторы: скорость ветра *Ws* и относительная влажность *RH*, как имеющие наименьшие регрессионные коэффициенты. Для учета инерционности процесса загрязнения в модель включены измерения PM2.5 в (*i-1*)-ый и (*i-2*)-ой сроки измерения.

Полученная модель применена для восстановления концентрации PM2.5 на тестовой выборке 05.09.2019–15.09.2019 гг. На рисунке 4 представлены вычисленные на основе этой модели и измеренные (наблюденные) значения концентраций.

Прогнозирование на нейросетевой модели GRNN. Возможности прогнозирования загрязнений воздуха на нейросетевой модели GRNN показаны в [5]. На первом этапе были исследованы различные варианты обучения и моделирования прогнозных значений PM2.5 с использованием вышеприведенных параметров (исключая давление, как несущественный фактор) с различной историей и прогнозом метеослужбы на всем диапазоне сроков измерения *i* (полная выборка, *i* = 3, 6, 9, 12, 15, 18, 21, 24 час). Данные были разделены на две части: обучающее множество – данные измерения Pm2.5 и метеорологических параметров по г. Бишкек за период 24.03.2019–31.07.2019 гг.; тестовое множество – данные измерения за период 01.08.2019–18.08.2019 гг. С целью повышения точности обучения была проведена декомпозиция измеренных значений PM2.5 по срокам измерения *i*.

Сравнение полученных мультирегрессионной и нейросетевой моделей прогноза производилось на выборке 5.09.2019–15.09.2019 гг. и показало, что средняя абсолютная ошибка MAE и среднеквадратическая ошибка MSE, рассчитанные для нейросетевой модели меньше, чем соответствующие ошибки мультирегрессионной модели [5]. На графиках (рисунок 4) представлены наблюденные (измеренные) значения концентраций PM2.5 и их прогнозные значения на 3 часа вперед (что соответствует одному сроку наблюдений).

В [5] также показано, что идея декомпозировать данные концентраций PM2.5 и метеопараметры по срокам измерения оправдала себя, поскольку позволила снизить ошибку прогноза.

Прогнозирование класса AQI на основе LSTM-классификатора. Анализ распределения AQI по классам [16] в период 06.02.2019–31.03.2020 гг. показал, что в 65 % наблюдений принадлежит классу «Умеренный» (см. рисунок 5, *а*). Числа наблюдений AQI, принадлежащих классам «Хороший», «Нездоровый для чувствительных групп», «Нездоровый», «Очень нездоровый», «Опасный» недостаточно для решения задачи классификации. Поэтому было решено объединить наблюдения, принадлежащие классам «Хороший» и «Умеренный», а также принадлежащие классам: «Нездоровый для чувствительных групп», «Нездоровый», «Очень нездоровый» и «Опасный» (рисунок 5, *б*). Это позволило рассмотреть задачу классификации AQI в зависимости от метеофакторов по двум интегрированным классам, условно названных «Хороший» и «Нездоровый». При этом если всегда предсказывать «Хороший» класс, то точность прогноза составит 70 %, поэтому будем считать классификатор приемлемым, если точность прогноза класса AQI на его основе превысит точность 70 %.

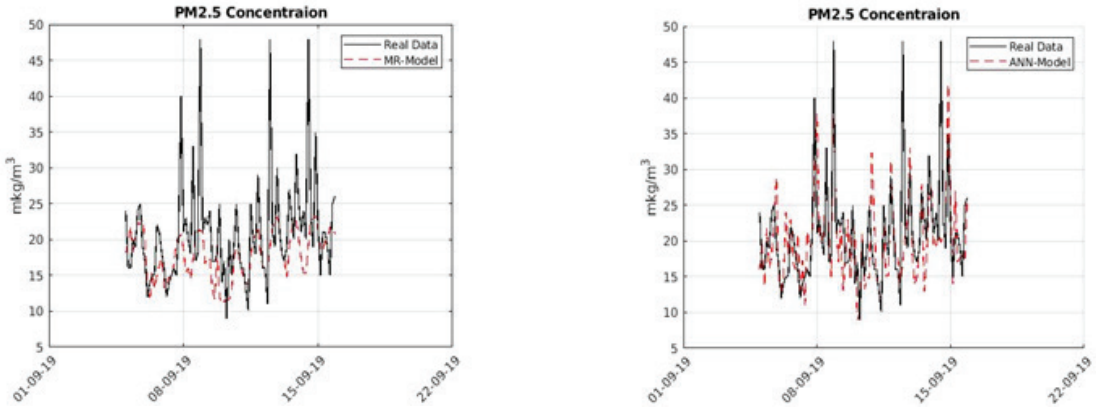

Рисунок 4 – Сравнение мультирегрессионной и нейросетевой GRNN моделей

Рисунок 5 – Распределение наблюдений по классам AQI (а)
и распределение наблюдений по объединенным классам AQI (б)

Применение различных нейросетевых классификаторов (многослойный перцептрон, обычная рекуррентная сеть и LSTM-сеть) показало, что наилучшую точность в прогнозировании класса AQI показала LSTM-сеть, поскольку она учитывает исторические данные и оценивает их в зависимости от временного удаления вектора входа до прогнозируемого выхода [11].

Как показано в [6], «входной вектор LSTM-классификатора определяется параметрами: температура воздуха, атмосферное давление, относительная влажность, скорость ветра, температура точки росы, показатель интенсивности осадков. При этом данные о температуре воздуха, атмосферном давлении, температуре точки росы нормализуются с помощью Z-нормы. Выходной вектор, к которому должен приближаться выход классификатора, определяется 2-мя параметрами – вероятность отнесения выхода классификатора к классу «Хороший», равная 1 для AQI <= 100 и вероятность отнесения выхода классификатора к классу «Нездоровый», равная 0 для AQI > 100. Проводя эксперименты по прогнозированию класса AQI, было решено варьировать следующие параметры: S – длину последовательности векторов исторических данных – входных векторов (ВВ) классификатора; P – глубину прогноза (на сколько шагов вперед прогнозируется AQI). Шаг прогноза – 3 часа. Эксперименты с LSTM-сетью показали [6], что лучшую точность прогнозирования дали классификаторы, учитывающие историю данных глубиной 12–16 шагов (1,5–2 дня), при этом прогноз AQI возможен до 4-х дней вперед с точностью 88–90 %» [6].

Заключение. Прогнозирование значений концентраций частиц PM2.5 в атмосферном воздухе и индекса качества воздуха AQI – непростая задача, поскольку на уровень загрязненности воздуха влияют многие факторы. Известны различные модели для прогнозирования – от классических статистических моделей до моделей глубокого обучения. В настоящей работе для прогнозирования уровня загрязненности воздуха представлены четыре варианта моделей: ARIMA-модели (интегрированные модели авторегрессии – скользящего среднего) для краткосрочного прогноза PM2.5 и AQI с использованием суточной истории наблюдений; линейные мультирегрессионные модели на основе

регрессионного анализа метеорологических факторов и концентраций частиц PM2.5 в предшествующие сроки наблюдений, модель краткосрочного прогноза концентраций PM2.5 с учетом метеорологических факторов на основе обобщенно-регрессионной нейронной сети GRNN, модель среднесрочного прогноза двух интегрированных классов AQI («Хороший»/«Нездоровый») в зависимости от метеопараметров на базе LSTM-нейронных сетей.

Указанные модели использованы для построения прогноза уровня загрязненности воздуха г. Бишкек в различные периоды 2019–2020 гг. При разработке нейросетевой модели предложено декомпозировать ряды наблюдений в соответствии со сроками наблюдений, что позволило повысить точность прогноза. Задача прогнозирования AQI в зависимости от метеопараметров рассмотрена как задача нейросетевой классификации. Последующее накопление данных позволит проводить классификацию AQI на большее число классов. Представленные в работе модели позволяют заключить, что затруднительно однозначно рекомендовать определенную модель для прогноза уровня загрязненности атмосферного воздуха, поскольку выбор должен определяться спецификой временного ряда наблюдений, размером выборки и целью прогнозирования.

Литература

1. Качество воздуха в Центральной Азии: пора бить тревогу. URL: https://livingasia.online/2020/09/01/kachestvo-vozduha-v-czentralnoj-azii-pora-bit-trevogu/ (дата обращения: 30.10.2020).
2. AirNow Depatment of State // https://airnow.gov/index.cfm?action=airnow. global_summary #U.S._ Department_ of_State $Bishkek (дата обращения: 02.11.2020).
3. *Верзунов С.Н.* Краткосрочное прогнозирование индекса качества воздуха на основе ARIMA-моделей / С.Н. Верзунов, Н.М. Лыченко // Математическое и компьютерное моделирование: сборник матер. VII межд. научн. конф. (22 ноября 2019 г.). Омск: Изд-во Омск. гос. ун-та, 2019.
4. *Верзунов С.Н.* Анализ и ARIMA-модели динамики изменения концентрации PM2.5 в атмосферном воздухе г. Бишкек / С.Н. Верзунов, Н.М. Лыченко // Проблемы автоматики и управления. 2019. № 1. С. 21–30.
5. *Великанова Л.И.* Мультирегрессионные и обобщенно-регрессионные нейросетевые модели краткосрочного прогноза загрязнения PM2.5 в г. Бишкек с учетом метеорологических параметров / Л.И. Великанова, Н.М. Лыченко // Проблемы автоматики и управления. 2019. № 2. С. 42–51.
6. *Лыченко Н.М.* Применение LSTM-нейронных сетей для классификации индекса качества воздуха г. Бишкек / Н.М. Лыченко, А.В. Сороковая // Проблемы автоматики и управления. 2020. № 1 (38). С. 70–80. DOI: 10.5281/zenodo.3904130.
7. *Бокс Д.* Анализ временных рядов: прогноз и управление / Д. Бокс, Т. Дженкинс. М.: Мир, 1974. 242 с.
8. Современное прогнозирование. URL: https://forecasting.svetunkov.ru/ etextbook/ (дата обращения: 30.09.2020).
9. *Филатова Т.В.* Применение нейронных сетей для аппроксимации данных / Т.В. Филатова // Вестник Томского госуд. ун-та. 2004. № 284. С. 121–125.
10. *Zhao X.* A Deep Recurrent Neural Network for Air Quality Classification / X. Zhao, R. Zhang, J.-L. Wu, P.-C. Chang // Journal of Information Hiding ultimedia Signal Processing. 2018. V. 9. N. 2, March.
11. Understanding LSTM Networks. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (дата обращения: 14.09.2020).
12. *MacKinnon J.G.* Critical Values for Cointegration Tests / J.G. MacKinnon // Queen's University, Dept. of Economics, Working Papers, 2010. http://ideas.repec.org/p/qed/wpaper/1227.html (дата обращения: 19.05.2019).
13. *Голохвост К.С.* Атмосферные взвеси и экология человека / К.С. Голохвост, П.Ф. Кику, Н.К. Христофорова // Экология человека. 2012. № 10. С. 5–10.
14. Сайт «Расписание погоды rp5.ru» Архив погоды в Бишкеке. URL: https://rp5.ru/%D0%90%D1%80%D1%85%D0%B8%D0%B2_%D0%BF%D0%BE%D0%B3%D0%BE%D0%B4%D1%8B_%D0%B2_%D0%91%D0%B8%D1%88%D0%BA%D0%B5%D0%BA%D0%B5 (дата обращения: 30.04.2020).
15. *Лыченко Н.М.* Регрессионный анализ метеорологических факторов и концентраций частиц PM2.5 в атмосферном воздухе г. Бишкек / Н.М. Лыченко // Проблемы автоматики и управления. 2019. № 2. С. 5–15.
16. *Air Quality* Index (AQI) – A Guide to Air Quality and Your Health. US EPA. 9 December, 2011.