

УДК 681.3

Лыченко Н.М.*Кыргызско-Российский славянский университет, Бишкек, Кыргызстан**E-mail: nlychenko@mail.ru*

МОДЕЛИ ПРОГНОЗА КАЧЕСТВА ВОЗДУХА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ С УЧЕТОМ ВРЕМЕННЫХ И ПРОСТРАНСТВЕННЫХ ЗАВИСИМОСТЕЙ

Представлен аналитический обзор наиболее используемых моделей прогноза процессов загрязнения атмосферного воздуха, характеризующихся сложными и нестационарными временными зависимостями. Пространственное положение, в дополнение к динамическим изменениям с течением времени, также оказывает влияние на качество воздуха. Относительно новый инструмент для решения задач прогноза качества воздуха с учетом не только временных, но и пространственных зависимостей - графовые нейронные сети. Основная характеристика этого типа сетей – использование информации, полученной в результате динамического взаимодействия между соседними географическими точками (городами, районами, улицами), которые взвешиваются в зависимости от расстояния между ними.

Ключевые слова: модели прогноза качества воздуха, временные ряды, пространственно-временные зависимости, машинное обучение, графовые нейронные сети.

Введение. Загрязнение воздуха является актуальной проблемой, которая представляет значительную угрозу здоровью человека и окружающей среде, приводит к дисбалансу в экосистеме. Термин «качество воздуха» относится к количеству и составу различных загрязняющих веществ, присутствующих в атмосфере, включая угарный газ, диоксид серы, диоксид азота, твердые частицы PM_{2.5}, PM₁₀ и другие. Индекс качества воздуха (AQI) рассчитывается на основе уровня концентрации загрязняющих веществ и степени их воздействия на здоровье. Для прогнозирования качества воздуха используется профессиональное оборудование и технологии мониторинга для сбора и анализа атмосферных параметров (таких как температура, влажность, скорость ветра, атмосферное давление, осадки и др.), географического положения, индекса качества воздуха и других данных, чтобы предсказывать и предупреждать о качестве воздуха в будущем. Точный прогноз качества воздуха имеет большое значение для принятия мер по минимизации воздействия опасных загрязняющих веществ на население и поддержания баланса экосистемы. Хотя для прогнозирования качества воздуха было разработано множество методов, прогнозирование качества воздуха остается непростой задачей, поскольку процессы, протекающие в атмосферном воздухе, характеризуются сложными и нестационарными временными и пространственными зависимостями [1].

Временную зависимость характеризуют периодичность и тренд. Периодичность — это появление схожих паттернов или регулярных изменений в течение определенного времени. Для индекса качества воздуха г. Бишкек, например, характерна явно выраженная периодичность сезонная (определяемая холодным и теплым временем года), явно выраженная суточная и – слабо выраженная недельная периодичность [2].

Пространственное положение, в дополнение к динамическим изменениям с течением времени, также оказывает влияние на качество воздуха. Качество воздуха в одной географической точке (разные города, разные районы, разные улицы) будет зависеть от атмосферных условий в другой географической точке. В сочетании с динамическими процессами в атмосфере, учет пространственных зависимостей в процессах загрязнения представляет собой очень сложную задачу.

Существует множество методов оценки качества воздуха, которые учитывают корреляцию со временем, включая авторегрессионную интегрированную модель скользящего среднего, модель рекуррентных нейронных сетей и модель долгосрочной и краткосрочной памяти (LSTM). Как правило, эти подходы учитывают динамические изменения качества воздуха, но не учитывают пространственные отношения между качеством воздуха и географическим положением.

В последние годы в прогнозировании качества воздуха стали использоваться графовые нейронные сети (Graph Neural Networks, GNN) [3-5]. Основная характеристика этого типа сетей – это использование информации, полученной в результате динамического взаимодействия между соседними географическими точками (городами, районами, улицами), которые взвешиваются в зависимости от расстояния между ними.

В настоящей работе приведен аналитический обзор наиболее используемых моделей прогноза процессов загрязнения атмосферного воздуха, характеризующихся сложными и нестационарными временными и пространственными зависимостями. Особое внимание уделено описанию графовых нейронных сетей как инструменту для решения задач прогноза качества воздуха с учетом не только временных, но и пространственных зависимостей.

1. Машинное обучение и классические регрессионные алгоритмы. Методы машинного обучения являются наиболее распространенными методами прогнозирования качества воздуха. С начала XXI века в литературе можно найти сотни произведений, которые предлагают реализации различных моделей для достижения максимальной точности прогнозирования индекса AQI или прогноза концентрации загрязняющих веществ. В [1] было проведено подробное исследование по применению различных алгоритмов машинного обучения как для прогнозирования AQI так и уровня концентрации конкретных загрязняющих веществ, связанных с качеством воздуха..

1.1 Классические статистические алгоритмы. Регрессионный анализ используется для установления отношения между зависимой переменной и набором независимых переменных. На основе этого отношения и с использованием значений независимых переменных оценивается значение зависимой переменной. Далее рассмотрим классические алгоритмы для построения регрессии.

Множественная линейная регрессия (Multiple Linear Regression, MLR). Пусть y и $x_1, \dots, x_p$ будут зависимой переменной и независимыми переменными соответственно. Цель линейной регрессии — определить линейную функцию $f(x_1, \dots, x_p)$, которая минимизирует среднюю квадратичную ошибку, то есть:

$$\min\{(y - f(x_1, \dots, x_p))^2\}.$$

Примеры применения множественной линейной регрессии в задачах прогноза качества воздуха можно найти в литературе, например [6, 7].

Авто-регрессионное интегрированное скользящее среднее (Auto-Regressive Integrated Moving Average, ARIMA). Эта модель использует временные ряды для прогнозирования, то есть она делает предсказания на основе прошлых и настоящих данных. Модель ARIMA включает три компонента:

1) авто-регрессионный компонент - характеризует количество задержек, используемых в модели. Например, авто-регрессионное значение 3 устанавливает, что только три предыдущих значения будут использоваться для объяснения текущего значения;

2) интегрированный компонент - представляет собой степень дифференцирования, необходимую для преобразования временного ряда в стационарный ряд;

3) скользящее среднее - относится к количеству прошлых ошибок, необходимых для объяснения текущей ошибки. Снова, если значение скользящей средней равно 3, то только три предыдущих значения ошибок могут быть использованы для объяснения текущей ошибки. Когда ВР содержит сезонность, используется вариант ARIMA, называемый SARIMA.

Модель ARIMA является одной из репрезентативных статистических моделей для прогнозирования качества воздуха. Например, ARIMA использовалась для прогнозирования загрязняющих веществ и показала хорошие результаты [8, 9, 10]. Однако определение параметров в статистических методах сильно зависит от теоретических предположений и предварительных знаний о данных, что делает затруднительным соответствие нелинейным и нестационарным данным о качестве воздуха. Таким образом, статистические методы могут привести к смещению прогноза AQI, особенно для точного долгосрочного прогноза AQI [4].

1.2 Алгоритмы машинного обучения на основе регрессионного анализа

Метод опорных векторов для регрессии (Support vector regression, SVR). Методы опорных векторов в основном применяются в задачах классификации. Однако их также можно использовать для регрессии. В этом случае подход называется регрессией на основе опорных векторов. Пусть y и $x_1, \dots, x_p$ будут зависимой переменной и независимыми переменными соответственно. В основном, SVR работает следующим образом. Сначала нужно определить линейную регрессионную функцию, то есть гиперплоскость $h(x) = w_1x_1 + \dots + w_px_p + b$. Затем предполагается, что все данные будут находиться на расстоянии не более чем ε от гиперплоскости. Если отклонение некоторых точек превышает это значение, можно ввести свободные переменные $\xi, \xi' \geq 0$ для работы с ними. Конечная цель — найти минимум функции:

$$\min \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi + \xi') \right)$$

при ограничениях:

$$y - wx_i - b \leq \varepsilon + \xi$$

$$wx_i + b - y_i \leq \varepsilon + \xi'.$$

В случае нелинейных функций SVR использует ядерные функции для преобразования данных в пространство более высокой размерности, чтобы разработать линейную регрессионную трансформацию

Деревья решений (Decision trees, DT). Цель этого алгоритма — разработать модель для прогнозирования количественной переменной на основе набора независимых переменных. Алгоритм основан на рекурсивном делении. Деревья состоят из узлов решений и листьев. Регрессия DT обычно строится с учетом уменьшения стандартного отклонения для определения того, как разделить узел на две или более ветки. Корневой узел — это первый узел решения, который делится на основе наиболее релевантной независимой переменной. Узлы снова делятся, принимая во внимание переменную с наименьшей суммой квадратов оценки ошибок (SSE) в качестве узла решения. Набор данных делится на основе значений выбранной переменной. Процесс завершается, когда выполнено заранее установленное условие остановки. Последние узлы известны как

листовые узлы и предоставляют предсказание зависимой переменной. Это значение соответствует среднему значениям, связанным с листьями. Рисунок 3а [11] показывает графическое представление общей структуры стандартного DT.

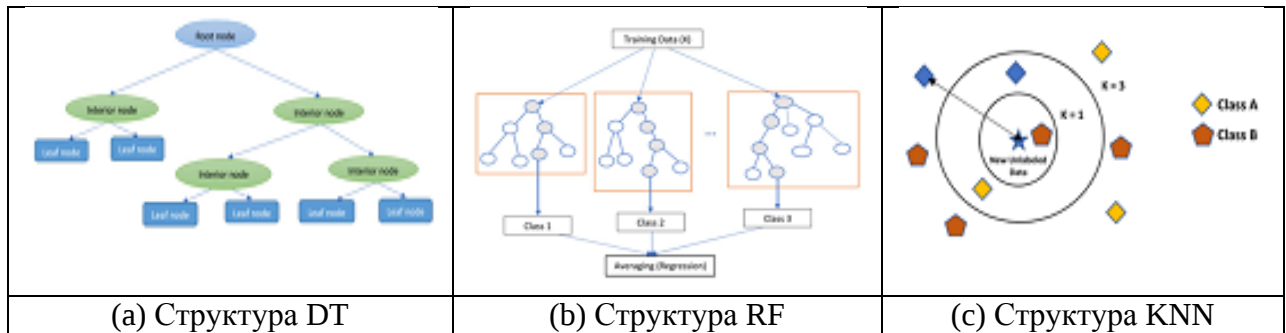


Рисунок 3- Графическое представление алгоритмов на основе регрессионного анализа.

Случайный Лес (Random Forest, RF). Случайный Лес основан на генерации нескольких деревьев решений. Прогноз будет средним значением прогнозов, предоставленных различными деревьями. Для построения каждого дерева решений выбирается выборка данных из обучающего набора данных. Остальные данные будут использованы для оценки ошибки дерева решений. Подмножество независимых переменных, которое может быть использовано для разделения каждого узла, выбирается случайным образом. Экстремально рандомизированные деревья (Extremely randomised trees, ERT) — это слегка модифицированный алгоритм случайного леса. Рисунок 3b [11] показывает графическое представление общей структуры регрессора RF.

Регрессия по k-ближайшим соседям (K-nearest neighbours regression, KNN). Алгоритм k -ближайших соседей часто применяется к задачам классификации, хотя его также можно применить и к задачам регрессии. Идея этого алгоритма проста. Учитывая расстояние (евклидово расстояние, например) и значение k , алгоритм вычисляет расстояние между точкой данных и точками обучающего набора данных для выбора k ближайших и устанавливает их среднее значение в качестве прогноза. Улучшением этого алгоритма является алгоритм, известный как взвешенные k -ближайшие соседи (WKNN). В этом случае прогноз учитывает взвешенное арифметическое среднее для расчета прогноза. Рисунок 3с [12] показывает графическое представление общей структуры модели KNN.

Традиционные методы машинного обучения совместно со статистическими моделями достаточно широко применяются к решению задачи прогноза качества воздуха. Так, например, методы K -ближайших соседей KNN, опорных векторов SVR, деревьев решений DT, RandomForest RF были реализованы для создания системы прогнозирования статуса индекса качества воздуха [13], а также были исследованы в [14] (KNN), [15] (SVM), [16] (DT, RF), и многих других. Основываясь на ограниченных наборах данных, эти методы машинного обучения могут лишь частично уловить ограниченные нелинейные временные и пространственные корреляции, влияющие на качество воздуха. Они больше подходит для поверхностных скрытых признаков, как показывают результаты. Учет очень сложных динамических нелинейных пространственно- временных связей им не поддается. Для эффективного преодоления этой проблемы создаются многочисленные комбинированные методы, использующие преимущества статистических методов и методов машинного обучения (см., например [17]).

В отличие от статистических методов, традиционные методы машинного обучения и методы глубокого обучения не зависят от предварительных физических знаний о данных и теоретических предположений. Они автоматически обучают свои параметры модели на

основе исторических данных, чтобы уловить нелинейные взаимосвязи и знания о данных качества воздуха. По сравнению со статистическими методами, традиционные методы машинного обучения, как правило, обеспечивают лучшую производительность в прогнозировании загрязняющих веществ в воздухе. Однако их эффективность, как правило, ограничена в практическом применении. Это связано с тем, что поверхностные структуры традиционных моделей машинного обучения часто не могут уловить сложные и нестационарные колебания данных о загрязнителях воздуха, что приводит к их плохой производительности в долгосрочном прогнозировании загрязняющих веществ в воздухе, а это имеет решающее значение для превентивных мер по предотвращению загрязнений воздуха и управления ими [4].

2. Алгоритмы глубокого обучения. Алгоритмы глубокого обучения используют искусственные нейронные сети (ИНС). ИНС — это алгоритм, основанный на связях, состоящих из нейронов или узлов. Эти связи организованы в три типа слоев (рисунок 4). Входной слой получает на вход исходные предикторные переменные. Выходной слой производит предсказанные значения для заданных входов. Эти два слоя соединены скрытыми слоями. Скрытые слои (более одного в случае глубокого обучения) содержат ненаблюдаемые нейроны, отвечающие за вычисления. Каждый узел в слое соединен с узлами в следующем слое. Каждое соединение имеет присвоенный вес, который используется для комбинации входных данных. Каждый узел или нейрон в следующем слое получает взвешенное значение и преобразует его с помощью активационной функции. Наиболее популярная функция - сигмоида. Полученный результат — это значение, которое передается в качестве входа узлам следующего слоя. Этот процесс продолжается, пока не будет достигнут выходной слой. На этом этапе производится выходное предсказание.

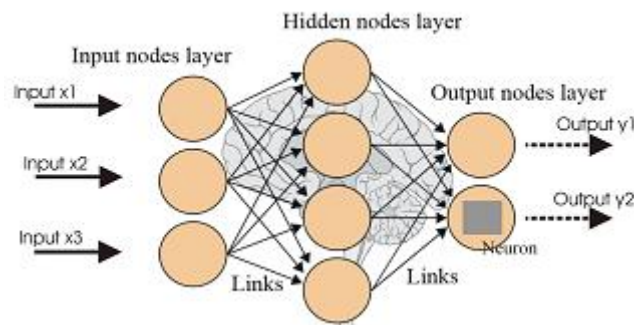


Рисунок 4- Структура ИНС [17]

Конечная цель ИНС — настроить веса для минимизации функции ошибки, обычно квадратичной функции. Для этого ИНС использует алгоритм обратного распространения ошибки. Этот алгоритм применяет метод градиентного спуска, используя частные производные слоев, чтобы найти оптимальный вес для каждого узла.

Типы ИНС, наиболее часто используемые при прогнозе качества воздуха с учетом временных зависимостей, — следующие.

Многослойные перцептронные нейронные сети (Multi-layer perceptron neural networks, MLP). MLP — это классические нейронные сети с одним или несколькими скрытыми слоями. Широко используемыми моделями в задачах прогноза качества воздуха являются нейронные сети с обратным распространением (BPNN) [18, 19] и обобщенно-регрессионные нейросетевые модели (GRNN) [20, 7]. Эти методы машинного обучения обладают четкой математической логикой, в которой связь между входными и выходными данными относительно определенная. Кроме того, у них относительно

простая структура сети, что приводит к ограниченной способности моделирования зависимостей во временных рядах данных PM2.5 и AQI.

Сверточные нейронные сети (Convolutional neural networks, CNN). CNN обычно применяются к изображениям, однако также используются в задачах прогноза качества воздуха. Это расширение MLP, в котором чередуются два типа слоев: сверточные и слой пуллинга (подвыборки). Цель сверточных слоев — извлечение признаков из входного изображения. Сверточная операция, которая выводит свёрнутые признаки, выполняется с помощью матрицы, называемой ядром или фильтром. Эти признаки становятся входом для подвыборочных слоев. Задача подвыборочного слоя — уменьшить размер свёрнутых признаков с целью снижения вычислительных ресурсов, необходимых для обработки изображения. В частности, в [21] представлен метод прогнозирования AQI с помощью одномерной сверточной нейронной сети и оптимизации экспоненциальных адаптивных градиентов.

Рекуррентные нейронные сети (Recurrent neural networks, RNN). RNN работают с временными рядами или последовательными данными, то есть с информацией, упорядоченной и взаимосвязанной. RNN имеют внутреннюю память в том смысле, что нейрон может обратным образом получать на вход выходные данные, которые он ранее произвел. Это позволяет модели приобретать краткосрочную память, которая необходима для прогнозирования временных рядов. Модели на основе RNN обладают отличной способностью улавливать временные зависимости во входных временных рядах данных PM2.5 и AQI [22].

Нейронные сети с долгосрочной и краткосрочной памятью (Long-short term memory neural networks, LSTM). LSTM — это расширение RNN. Они имеют расширенную память, которая позволяет работать с долгосрочными зависимостями. LSTM могут запоминать информацию на произвольных временных интервалах. Основным компонентом является состояние ячейки, которое переносит информацию на протяжении обработки данных. Информация обновляется на основе трех слоев (рисунок 5). Каждое из них контролирует информацию, которая должна находиться в состоянии ячейки, с использованием сигмоидной функции активации. Слой забывания (forget gate) определяет, какая часть информации из предыдущего состояния должна быть забыта. Слой ввода (input gate) решает, какая новая информация будет использована для обновления памяти. Используя гиперболическую тангенс-функцию, он создает вектор-кандидат для добавления к состоянию ячейки. Последний слой, известный как выходной слой (output gate), использует функцию активации tanh для определения, какая часть обновленного состояния ячейки будет использоваться в качестве выхода.

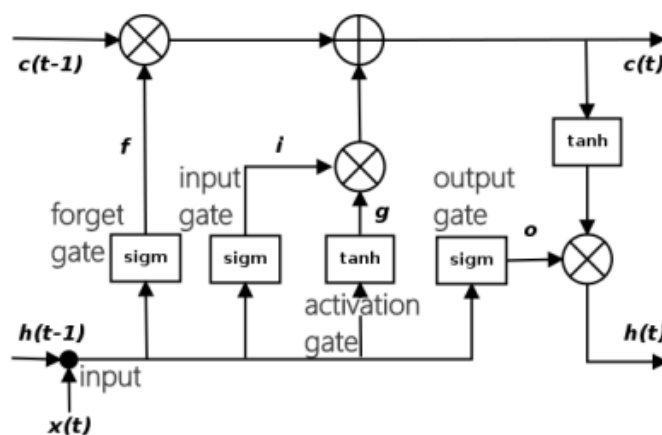


Рисунок 5 - LSTM - ячейка: [22]

Модели LSTM успешно применяются для предсказания концентраций загрязняющих веществ в воздухе, улавливая долговременные зависимости и паттерны во временных данных (см., например, [23, 24]) и являются наиболее часто используемыми среди методов глубокого обучения (в 36.13% статей [1]).

Упрощенная рекуррентная единица (Gated recurrent unit GRU). GRU — это упрощенная версия LSTM, которая объединяет слой забывания и слой ввода. Они дают результаты, похожие на те, что получаются с помощью LSTM.

Нейронные сети типа "кодировщик-декодировщик" (Encoder-Decoder neural networks, EDNN). Модель кодировщик-декодировщик — это рекуррентная нейронная сеть, используемая для задач предсказания последовательностей. Ее архитектура состоит из трех компонентов: кодировщика, промежуточного вектора и декодировщика. Кодировщик и декодировщик состоят из набора рекуррентных единиц (обычно LSTM или GRU). В случае кодировщика каждая единица обрабатывает элемент входной последовательности и пытается запечатлеть связанную информацию в промежуточном векторе, чтобы повысить точность предсказания декодировщика. Декодировщик использует этот вектор для формирования предсказания.

3. Графовые нейронные сети (Graph Neural Network, GNN). В последнее время все более популярными в прогнозировании качества воздуха становятся графовые нейронные сети [3 – 5]. Основная характеристика этого типа сетей заключается в информации, полученной от динамического взаимодействия между соседями (например, разные города, разные районы, разные улицы), которая взвешивается в зависимости от расстояния. Например, в [5] для прогноза концентрации PM2.5 на основе метеорологических данных и географического положения трех станций мониторинга применена графовая нейронная сеть на реальных наборах данных, которая показала лучшие результаты по сравнению с базовыми моделями [1].

Подобно представленным в [3] определениям определим некоторые понятия для описания проблемы прогноза качества воздуха для случая, когда информация о загрязнении воздуха приходит не с одного, а с некоторого множества датчиков.

Определение 1: датчики и группы датчиков. Для представления некоторого набора датчиков используется множество $S = \{s_i\}_{i=1}^{N_s}$, где N_s представляет число (количество) датчиков, а местоположение датчика задаётся его долготой и широтой, и матрица локации расположения датчиков выражается как $L \in R^{N_s \times 2}$. Кроме того, множество $U = \{u_i\}_{i=1}^{N_u}$ используется для представления множества групп датчиков, где N_u представляет количество групп датчиков.

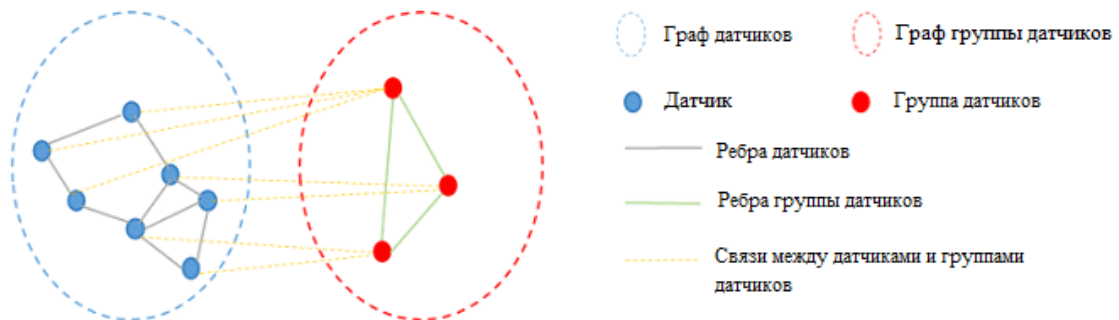


Рисунок 6 – Графы датчиков и групп датчиков.

Определение 2: граф группы датчиков. Будем рассматривать группы датчиков как датчики с сильными зависимостями и строить граф групп датчиков как полносвязный граф. Для представления топологической структуры графа групп датчиков используется неориентированный граф $g(v, \varepsilon)$, как показано на рисунке 6, где $v = \{v_1, v_2, \dots, v_n\}$ представляет собой множество всех узлов групп датчиков, при этом для любого узла v_i значения его атрибутов обозначаются как R_i , значения атрибутов всех узлов в v могут быть выражены как матрица R . $\varepsilon = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\}$ представляет собой множество рёбер всех групп датчиков. Для любого ребра ε_i значения его атрибута - M_i , значения атрибутов всех рёбер в ε могут быть выражены как матрица M .

Определение 3: граф датчиков. Для представления топологии графа датчиков используется неориентированный граф $G(V, E)$, как показано на рисунке 6, где $V = \{v_1, v_2, \dots, v_n\}$ - множество всех узлов датчиков, и для любого узла V_i определены его атрибуты X_i , значения атрибутов всех узлов в V могут быть представлены в виде матрицы X . $E = \{e_1, e_2, \dots, e_n\}$ представляет собой множество всех рёбер датчиков. Для любого ребра E_i используется Y_i для обозначения значений его атрибутов, а значения атрибутов всех рёбер в E могут быть представлены в виде матрицы Y . Граф датчиков строится следующим образом:

$$Y_{i,j} = \frac{1}{\sqrt{(a_i - a_j)^2 + (b_i - b_j)^2}},$$

$$d_{i,j} = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2}, \quad 0 < d_{i,j} < r_n,$$

где $Y_{i,j}$ - атрибут ребра между датчиком S_i и датчиком S_j , $[a_i, b_i]$ представляет локацию (местоположение) датчика S_i , а $[a_j, b_j]$ представляет локацию датчика S_j . $d_{i,j}$ - евклидово расстояние между датчиком S_i и датчиком S_j , а r_n представляет собой порог расстояния. Когда расстояние между двумя датчиками очень велико, датчики почти не влияют друг на друга. Если расстояние больше r_n , можно считать, что эти датчики принадлежат разным населённым пунктам.

Определение 4: задача прогноза индекса качества воздуха. Необходимо предсказать будущий индекс качества воздуха AQI датчика, исходя из исторической последовательности наблюдений за местоположением датчика, временем, AQI и данными о погоде. То есть, для датчика S_i , учитывая местоположение датчика L , вектор времени T_t и историческую последовательность наблюдений $H = (h_i^{t_1}, h_i^{t_2}, \dots, h_i^{t_m})$ за m временных шагов в момент времени t , можно предсказать последовательность наблюдений AQI датчика $\hat{H} = (\hat{h}_i^{t_{m+1}}, \hat{h}_i^{t_{m+2}}, \dots, \hat{h}_i^{t_{m+n}})$ на n временных шагов.

Итак, рассматривая данные о качестве воздуха, зарегистрированные на мониторинговой площадке, как сигнал соответствующего узла, GNN могут автоматически извлекать пространственную зависимость из данных о качестве воздуха без каких-либо предварительных знаний. Эта особенность позволила активно использовать GNN в моделировании прогнозирования качества воздуха. Например, в [25] предложено объединить графовую свертку (GCN) с рекуррентной нейронной сетью (RNN), структурированной как Gated Recurrent Unit (GRU), чтобы захватить временные и пространственные корреляции для улучшения прогнозирования PM2.5. В [26] авторы разработали три распределенные компонента на основе GCN для извлечения недавних, ежедневных периодических и еженедельных периодических временных пространственных зависимостей данных о качестве воздуха и объединили выходы трех компонентов с компонентом извлечения внешней информации для генерации прогнозов качества воздуха.

Отличная способность GNN к извлечению пространственных особенностей обеспечила этим методам хорошую предсказательную производительность. Однако способность этих методов извлекать временную зависимость все еще нуждается в улучшении. Например, точность модели с использованием RNN [25] значительно снижается с увеличением длины времени предсказания из-за ее ограничения на извлечение глобальной информации.

В последующих исследованиях необходимо дополнительно изучить, как графовые нейронные сети могут быть интегрированы с другими моделями временных рядов, чтобы повысить предсказуемость и универсальность этих моделей.

Литература

1. Manuel Méndez, · Mercedes G. Merayo, Manuel Núñez. Machine learning algorithms to forecast air quality: a survey// Artificial Intelligence Review (2023) 56:10031–10066 <https://doi.org/10.1007/s10462-023-10424-4>
2. Лыченко Н.М. Корреляционный анализ метеопараметров и концентраций твердых частиц PM_{2.5} в атмосферном воздухе г. Бишкек и его применение в моделях машинного обучения // Проблемы автоматизации и управления. – 2024. – № 2(50). – С. 88-96. – EDN ZIWQVG
3. A Hybrid Model for Spatiotemporal Air Quality Prediction Based on Interpretable Neural Networks and a Graph Neural Network
4. Pengfei Li, Tong Zhang and Yantao Jin. A Spatio-Temporal Graph Convolutional Network for Air Quality Prediction, Sustainability 2023, 15, 7624. <https://doi.org/10.3390/su15097624>
5. Li D, Yu H, Geng Y-a, Li X, Li Q (2021) DDGNet: A dual-stage dynamic spatio-temporal graph net-work for PM_{2.5} forecasting. In: 2021 IEEE International Conference on Big Data (Big Data), pp 1679–1685. <https://doi.org/10.1109/BigData52589.2021.9671941>
6. He, Z.; Liu, P.; Zhao, X.; He, X.; Liu, J.; Mu, Y. Responses of Surface O₃ and PM_{2.5} Trends to Changes of Anthropogenic Emissions in Summer over Beijing during 2014–2019: A Study Based on Multiple Linear Regression and WRF-Chem. Sci. Total Environ.2022, 807, 150792.
7. Великанова, Л. И. Мультирегрессионные и обобщенно-регрессионные нейросетевые модели краткосрочного прогноза загрязнения PM_{2.5} в г. Бишкек с учетом метеорологических параметров / Л. И. Великанова, Н. М. Лыченко // Проблемы автоматизации и управления. – 2019. – № 2(37). – С. 16-25. – DOI 10.5281/zenodo.3594777. – EDN MNNDYU.
8. Kulkarni, G.E.; Muley, A.A.; Deshmukh, N.K.; Bhalchandra, P.U. Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India. Model. Earth Syst. Environ. 2018, 4, 1435–1444.
9. Mirche Arsov, Eftim Zdravevski, Petre Lameski, Roberto Corizzo, Nikola Koteli, Sasho Gramatikov, Kosta Mitreski, and Vladimir Trajkovik. Multi-Horizon Air Pollution Forecasting with Deep Neural Networks // Sensors (Basel). 2021 Feb; 21(4): 1235. doi: 10.3390/s21041235
10. Верзунов С.Н., Лыченко Н.М. Анализ и ARIMA- модели динамики изменения концентрации PM_{2.5} в атмосферном воздухе г. Бишкек // Проблемы автоматизации и управления.- N1. Бишкек: Илим, 2019. –С. 21-30.
11. Balogun A-L, Tella A (2022) Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression. Chemosphere 299:134250. <https://doi.org/10.1016/j.chemosphere.2022.134250>

12. Tella A, Balogun A-L, Adebisi N, Abdullah S (2021) Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. *Atmos Pollut Res* 12(10):101202. <https://doi.org/10.1016/j.apr.2021.101202>
13. Morapedi TD and Obagbuwa IC (2023) Air pollution particulate matter (PM2.5) prediction in South African cities using machine learning techniques. *Front. Artif. Intell.* 6:1230087. doi: 10.3389/frai.2023.1230087
14. Keramat-Jahromi, M., Mohtasebi, S. S., Mousazadeh, H., Ghasemi-Varnamkhasti, M. & Rahimi-Movassagh, M. Real-time moisture ratio study of drying date fruit chips based on on-line image attributes using kNN and random forest regression methods. *Measurement* 172, 108899 (2021).
15. Pisner, D. A. & Schnyer, D. M. Support Vector Machine. *Machine Learning: Methods and Applications to Brain Disorders* (Elsevier Inc, 2019).
16. Sarkhosh, M. et al. Indoor Air Quality associations with sick building syndrome: An application of decision tree technology. *Build. Environ.* 188, 107446 (2021).
17. Mr. V. Devasekhar, Dr. P. Natarajan: Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques. In: *International Journal of Advanced Computer Science and Applications*. Vol. 14, No. 4 (2023)
18. Wang, W., Zhao, S., Jiao, L., Taylor, M., Zhang, B., Xu, G., et al. (2019). Estimation of PM2.5 Concentrations in China Using a Spatial Back Propagation Neural Network. *Sci. Rep.* 9 (1), 13788–13810. doi:10.1038/s41598-019-50177-1
19. Великанова, Л. И. Моделирование краткосрочного прогноза загрязненности воздуха твердыми частицами на основе искусственных нейронных сетей с учетом фактора загрязнения / Л. И. Великанова, Н. М. Лыченко // *Проблемы автоматизации и управления*. – 2022. – № 3(45). – С. 110-118. – EDN TGNBLC.
20. Zhou, Q., Jiang, H., Wang, J., and Zhou, J. A Hybrid Model for PM 2.5 Forecasting Based on Ensemble Empirical Mode Decomposition and a General Regression Neural Network. *Sci. Total Environ.* 496, 264–274, 2014. doi:10.1016/j.scitotenv.2014.07.051
21. Ragab, M. G., Abdulkadir, S. J., Aziz, N., Al-Tashi, Q., Alyousifi, Y., Alhussian, H., et al. A Novel One-Dimensional CNN with Exponential Adaptive Gradients for Air Pollution Index Prediction. *Sustainability* 12 (23), 2023. doi:10.3390/su122310090
22. Stojov, V., Koteli, N., Lameski, P., Zdravevski, E.: Application of machine learning and time-series analysis for air pollution prediction. *CIIT 2018* (2018).
23. Loy-Benitez, J.; Vilela, P.; Li, Q.; Yoo, C. Sequential Prediction of Quantitative Health Risk Assessment for the Fine Particulate Matter in an Underground Facility Using Deep Recurrent Neural Networks. *Ecotoxicol. Environ. Saf.* 2019, 169, 316–324.
24. Лыченко, Н. М. Применение LSTM-нейронных сетей для классификации индекса качества воздуха г. Бишкек / Н. М. Лыченко, А. В. Сорокова // *Проблемы автоматизации и управления*. – 2020. – № 1(38). – С. 70-80. – DOI 10.5281/zenodo.3904130. – EDN QQZNFI.
25. Wang, S.; Li, Y.; Zhang, J.; Meng, Q.; Meng, L.; Gao, F. PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2.5 Forecasting. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, Seattle, WA, USA, 3–6 November 2020; pp. 163–166.
26. Chang, F.; Ge, L.; Li, S.Y.; Wu, K.Y.; Wang, Y.Q. Self-adaptive spatial-temporal network based on heterogeneous data for air quality prediction. *Connect. Sci.* 2021, 33, 427–446.