УДК 004.75

ЗАЙЧЕНКО Ю. П.,
ЛАБЖИНСКИЙ В. А.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ДВУХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ ЗАГРЯЗНЕНИЯ ВОЗДУШНОГО БАССЕЙНА

В данной работе рассмотрена задача краткосрочного прогнозирования концентраций загрязняющих веществ в атмосфере с помощью двух статистических методов: Бокса-Дженкинса и метода группового учета аргументов. Приведено описание методик построения прогнозирующих моделей для обоих методов, а также сравнительная характеристика результатов прогноза.

In the given work short-term forecasting of pollutants' concentration in atmosphere have been considered with two statistical methods: Box-Jenkins and group method of data handling. The description of forecasting models building procedures and comparative characteristic of predictions' results has been adduced for both methods.

Введение

Резкий рост уровня антропогенного воздействия на окружающую среду в сочетании с низкой эффективностью и разобщенностью природоохранных мероприятий привел за последние десятилетия к значительному ухудшению экологической обстановки в различных регионах Украины. Наиболее уязвимыми оказались промышленно-развитые районы с высокой концентрацией промышленных производств топливно-энергетического, металлургического, химического и др. профилей, приведших к неконтролируемому превышению уровней допустимой экологической нагрузки на окружающую среду, возникновению потенциально опасных чрезвычайных ситуаций [1].

В связи с этим проблемы снижения эколого-экономического ущерба и управления экологической ситуацией приобретают в последнее время все возрастающее значение. Наряду с оценкой и контролем концентраций вредных примесей в воздухе в районе источников выбросов по данным наблюдений также необходимо осуществлять краткосрочные прогнозы загрязнения воздуха и использовать их для регулирования промышленных выбросов. Приоритетность регулирования выбросов вредных веществ в атмосферу обусловлена доминирующей ролью этой проблемы в качестве отправной точки при создании технологий, повышающих эффективность и действенность экологических нормативных ограничений загрязнения в целом.

Процесс загрязнения воздуха, как и многие другие атмосферные процессы, отличается большой сложностью. Методы исследования и прогноза, базирующиеся на использовании уравнений математической физики (теории диффузии и переноса примесей), учет метеофакторов (скорости и направления ветра, условий температурной стратификации), особенностей подстилающей поверхности, химического превращения примесей и т. п., обеспечивают получение результатов с высокой степенью точности [2–4]. Однако алгоритмы, реализующие эти методы, громоздки, требуют больших затрат машинного времени и ресурсов памяти ЭВМ. Указанные методы целесообразно применять при решении задач долгосрочного прогнозирования, исследования тенденций загрязнения воздушной среды за длительные промежутки времени (месяц, сезон и т. п.).

В свою очередь, для краткосрочного прогнозирования загрязнения воздуха в городах и промышленных районах, обусловленного действием многих источников, целесообразно использовать статистические методы, основанные на анализе материалов наблюдений. Здесь предполагается, что за период, к которому относится исследуемый материал, а также за срок прогноза, выбросы и расположение источников выбросов практически не изменяются. Поэтому рост среднего и суммарного загрязнения воздуха в городе связывается главным образом с изменением метеорологических условий или синоптической ситуации [5].

Разработка методов прогноза начинается в первую очередь с выявления периодов наибольшего загрязнения атмосферы. Затем устанавливаются корреляционные зависимости между наблюдавшимися в эти периоды степенью загрязнения воздуха и некоторыми метеорологическими величинами или их определенным сочетанием, рассматриваемыми в качестве предикторов. Таким путем вырабатывают различные прогностические правила. Также используют методы статистической экстраполяции во времени режима изменения загрязнения воздуха с учетом выявленных автокорреляционных зависимостей и инерционных факторов [2].

Постановка задачи

Рассмотрим один из вариантов постановки задачи моделирования загрязнения окружающей среды, а именно: моделирование загрязнения выполняется исключительно по данным измерений нескольких контрольно-измерительных станций (КИС) [6].

В этой задаче уравнение диффузии не задается априори и находится по экспериментальным данным по принципу самоорганизации при помощи перебора шаблонов и их нелинейностей. Оценки коэффициентов определяются с помощью метода наименьших квадратов (МНК). Решение находится при помощи пошагового интегрирования конечно-разностного уравнения. Знать это решение в аналитической форме не требуется.

Исходной информацией для прогнозирования концентрации загрязняющих веществ являются выборочные данные, представляющие собой замеры концентрации загрязняющих веществ через фиксированные интервалы времени в местах установки КИС. Таким образом, выборка представляет собой многомерный временной ряд.

Такая постановка задачи моделирования поля имеет целью построение поля загрязнения не только в области интерполяции (охваченной КИС), но и на значительном расстоянии за пределами этой области, т.е. построение экстраполяции поля и прогнозов его развития во времени. При этом предполагается, что сбросы загрязнения сравнительно мало изменяются во времени, и поэтому информация о них непосредственно не учитывается. Косвенно она содержится в данных измерений КИС. Поэтому в этой постановке множество аргументов включает только данные КИС.

Метод Бокса-Дженкинса. Модель ARIMA (AutoRegressive Integrated Moving Average – модель авторегрессии – проинтегрированного скользящего среднего)

Модель ARIMA относится к классу линейных моделей, которая может хорошо описывать поведение как стационарных, так и нестационарных временных рядов. Для прогнозирования модель ARIMA использует информацию, содержащуюся в исходном ряде. Модель ARIMA опирается, в основном, на автокорреляционную структуру данных временного ряда.

Внутренняя структура временного ряда, то есть зависимость его уровня y_t от предыдущих значений $y_{t-1}, y_{t-2}, ..., y_{t-p}$ может быть описана авторегрессионной функцией

$$y_t = \mu + a_1 \cdot y_{t-1} + a_2 \cdot y_{t-2} + ... + a_p \cdot y_{t-p} + \varepsilon_t,$$

где μ – константа, p – порядок авторегрессии, $a_1, ..., a_p$ – коэффициенты авторегрессии, ε_t – белый шум. Другими словами, каждое значение временного ряда есть сумма, содержащая в своем составе линейную комбинацию p его предыдущих значений и случайную ошибку ε_t.

Другим типом модели, описывающей поведение временного ряда, является модель скользящего среднего, в которой уровень ряда y_t линейно зависит от q предыдущих значений белого шума ε:

$$y_t = \mu + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - ... - \theta_q \cdot \varepsilon_{t-q},$$

где символы $-\theta_1, -\theta_2, ..., -\theta_q$ используются для обозначения конечного набора весовых параметров. Эта модель называется процессом скользящего среднего порядка q. Другими словами, каждое значение временного ряда есть сумма, содержащая в своем составе линейную комбинацию q предыдущих значений белого шума и случайную ошибку ε_t.

Для достижения большей гибкости при построении модели исследуемых процессов полезно включать в нее как выражение для скользящего среднего, так и выражение для авторегрессии. Это приводит к смешанной модели ARMA (p, q):

$$y_t = \mu + a_1 \cdot y_{t-1} + a_2 \cdot y_{t-2} + ... + a_p \cdot y_{t-p} + \\ + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - ... - \theta_q \cdot \varepsilon_{t-q}.$$

Если в модели необходимо учитывать наличие тренда, то это делается введением в ряд, описываемый моделью ARMA, разностей *d*-го порядка вместо значений y_t. В частности, $\nabla^1(y_t) = y_t - y_{t-1}$ (линейный тренд), $\nabla^2(y_t) = y_t - y_{t-2}$ (квадратичный тренд) и т. д. Разности ∇^d должны быть стационарными. В результате получаем модель ARIMA порядка (*p*, *d*, *q*) [7]:

$$\nabla^d(y_t) = \mu + a_1 \cdot \nabla^d(y_{t-1}) + a_2 \cdot \nabla^d(y_{t-2}) + ... \\ + a_p \cdot \nabla^d(y_{t-p}) + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - ... - \theta_q \cdot \varepsilon_{t-q}.$$

Вид модели ARIMA, ее адекватность реальному процессу и прогнозные свойства зависят от порядка авторегрессии *p* и порядка скользящего среднего *q*. Поэтому ключевым моментом моделирования является процедура идентификации – обоснования выбора модели того или иного вида. Выбор начальной модели ARIMA основывается на изучении графиков автокорреляций и частных автокорреляций между данными исходного временного ряда. Выбранная модель сопоставляется с исходными данными, чтобы проверить, насколько точно она описывает временной ряд. Модель считается приемлемой, если остатки малы и распределены случайно [8].

Метод группового учета аргументов (МГУА)

В основу МГУА [9, 10] положен принцип самоорганизации, и алгоритмы МГУА воспроизводят схему массовой селекции. В алгоритмах МГУА особым образом синтезируются и отбираются члены полинома, который называют обобщенным полиномом Колмогорова-Габора:

$$Y = a_0 + \sum_{i=1}^{N} a_i x_i + \sum_{j=1}^{N} \sum_{i \le j} a_{ij} x_i x_j + \sum_{i=1}^{N} \sum_{j \le i} \sum_{k \le j} a_{ijk} x_i x_j x_k + ...$$

где *N* – количество независимых переменных.

Этот синтез и отбор производится с нарастающим усложнением, и заранее нельзя предугадать, какой окончательный вид будет иметь обобщенный полином. Сначала обычно рассматривают простые попарные комбинации исходных переменных, из которых составляются уравнения решающих функций, как правило, не выше второго порядка. Каждое уравнение анализируется как самостоятельная решающая функция, и по обучающей выборке тем или иным способом находятся значения параметров составленных уравнений. Затем из полученного набора решающих функций отбирается часть в некотором смысле лучших. Проверка качества отдельных решающих функций осуществляется на контрольной (проверочной) выборке, что иногда называют принципом внешнего дополнения. Отобранные частные решающие функции рассматриваются далее как промежуточные переменные, служащие исходными аргументами для аналогичного синтеза новых решающих функций и т. д. Процесс такого иерархического синтеза продолжается до тех пор, пока не будет достигнут экстремум критерия качества решающей функции, что на практике проявляется в ухудшении этого критерия при попытках дальнейшего увеличения порядка членов полинома относительно исходных переменных [9].

Рассмотрим процесс синтеза модели оптимальной сложности более подробно. Представим функцию, аппроксимирующую набор исходных данных, в общем виде: $y = F(x_1, x_2, ..., x_N)$. Выше упоминалось, что такой функцией может быть полином Колмогорова-Габора, с помощью которого можно добиться весьма точной аппроксимации любой дифференцируемой функции. Заменим эту сложную зависимость множеством частных описаний, т.е. простых функций, аргументами которых являются все возможные пары исходных аргументов:

$y_1 = f(x_1, x_2), y_2 = f(x_1, x_3), ..., y_s = f(x_{N-1}, x_N)$,

где $s = C_N^2$, причем вид функции *f* одинаков для всех пар в течение всего процесса обучения. Очень часто в качестве функции *f* выбираются простые зависимости

$$y(x_i, x_j) = a_0 + a_1 \cdot x_i + a_2 \cdot x_j + a_3 \cdot x_i \cdot x_j$$

или

$$y(x_i, x_j) = a_0 + a_1 \cdot x_i + a_2 \cdot x_j + a_3 \cdot x_i \cdot x_j + \\ + a_4 \cdot x_i^2 + a_5 \cdot x_j^2.$$

Предварительно вся выборка разделяется на две части: обучающую и проверочную. Тем самым порождается внешнее дополнение (проверочная выборка), которая играет роль сита, отсеивающего худшие частные модели. Коэффициенты $a_0 - a_5$ частных опи-

саний определяются с помощью МНК по данным обучающей выборки.

В результате комбинирования всех возможных пар из N исходных аргументов получается множество решений, поскольку частное уравнение каждой пары рассматривается как некоторая упрощенная модель восстанавливаемой функции. Из полученного набора упрощенных моделей первого ряда отбирается часть, например, t в некотором смысле наилучших, показавших хорошие результаты на проверочной выборке, не участвовавшей в определении коэффициентов уравнений (т.е. на внешнем дополнении).

Далее вступает в действие принцип неокончательности решений: ни одна из полученных на первом этапе моделей не принимается за истину и наращивание сложности модели продолжается. Отобранные частные описания формируют множество новых переменных, которые являются исходными аргументами для частных описаний второго ряда селекции [10]:

$$z_1 = f\left(y_i, y_j\right), z_2 = f\left(y_i, y_k\right), ..., z_u = f\left(y_{t-1}, y_t\right),$$

где $u = C_t^2$.

Коэффициенты новых моделей находятся по МНК на точках той же обучающей последовательности. Новые модели проверяются на точках проверочной последовательности, и среди них выбирается v наилучших, которые используются в качестве аргументов следующего третьего ряда и т. д.

Сложность общей модели возрастает от ряда к ряду. Так, например, во втором ряду могут появиться нелинейные члены вида $\left(x_1 \cdot x_2 \cdot x_3\right), \left(x_1^2 \cdot x_3\right), \left(x_1^2 \cdot x_2 \cdot x_3\right)$ и т. д. Алгоритм останавливается сразу же по достижении минимума внешнего критерия (обычно среднеквадратичного отклонения или несмещенности), полученных на проверочной выборке. На практике усложнение модели прекращают, когда дальнейшее улучшение критерия селекции, например, средний квадрат ошибки прогноза, не будет превышать некоторого числа ε (параметр алгоритма). Тем самым выбирается модель оптимальной сложности, устанавливающая компромисс между сложностью и опасностью “переобучения” [10].

Существуют различные разновидности МГУА: однорядные, многорядные, гибридные. Для решения поставленной задачи будем использовать идею многорядного алгоритма МГУА.

Сравнение результатов прогнозирования с использованием двух методов

Выбор именно этих двух методов для сравнительного анализа точности получаемого прогноза объясняется следующими причинами. Метод Бокса-Дженкинса является классическим статистическим методом прогнозирования, широко применяемым, в частности, для получения краткосрочного прогноза состояния воздушного бассейна в автоматизированных системах контроля загрязнения воздуха [3]. Метод МГУА, в свою очередь, в этих целях почти не применяется [1]. Поэтому интересно сравнить результаты прогнозирования, полученные с помощью этих двух методов. Исходные данные по загрязнению воздуха оксидом углерода (CO), диоксидом азота (NO_2) и диоксидом серы (SO_2) одного из районов Лондона (Bexley) за май 2009 г. были взяты с сайта Национального архива загрязнения воздуха Великобритании [11]. Показания снимались в конце каждого часа. Объем выборки составил 744 наблюдения. Выборка была разделена на две части в соотношении 2:1 на обучающую и проверочную соответственно. Для каждого загрязнителя ниже представлены по две лучшие (каждая в своем классе) прогнозные модели.

Перед моделированием данные выборки были нормализованы с использованием формулы

$$\tilde{y}_t = 2 \cdot \frac{y_t - y_t^{(\min)}}{y_t^{(\max)} - y_t^{(\min)}} - 1$$

для приведения к диапазону изменения [-1; 1]. Оптимальный порядок моделей АРПСС выбирался на основе информационного критерия Акаике [12]. Оптимальный порядок моделей МГУА выбирался на основе критерия PSE (predicted squared error – квадрат ошибки прогноза) [13].

Лучшей моделью для описания концентрации CO (рис. 1), полученной с помощью метода Бокса-Дженкинса, оказалась модель АРПСС(1,0,0):

$$\tilde{y}_t = -0{,}8605 + 0{,}8041 \cdot \tilde{y}_{t-1} + \varepsilon_t .$$

Лучшей моделью для описания концентрации CO, полученной с помощью метода МГУА, оказалась следующая модель:

$$\tilde{y}_t = -0,15 + 1,4 \cdot \tilde{y}_{t-1} - 1,4 \cdot \tilde{y}_{t-3} + 0,65 \cdot \tilde{y}_{t-4} -$$
$$-0,37 \cdot \tilde{y}_{t-1}^3 + 0,4 \cdot \tilde{y}_{t-3}^3 - 0,96 \cdot \tilde{y}_{t-1} \cdot \tilde{y}_{t-3} + \quad .$$
$$+0,53 \cdot \tilde{y}_{t-1} \cdot \tilde{y}_{t-4} - 0,26 \cdot \tilde{y}_{t-1} \cdot \tilde{y}_{t-3} \cdot \tilde{y}_{t-4}$$

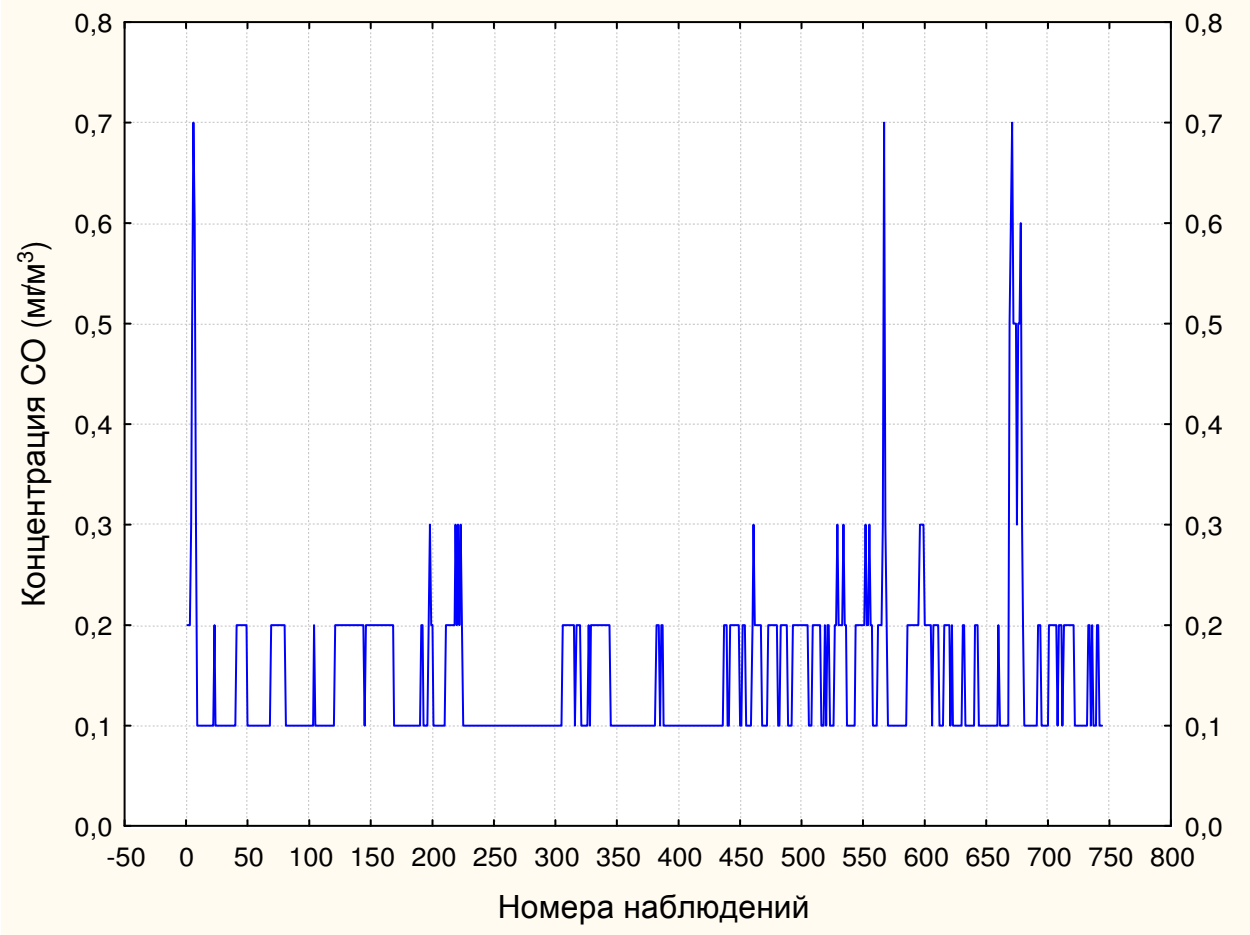

Рис. 1. Зависимость концентрации CO от времени

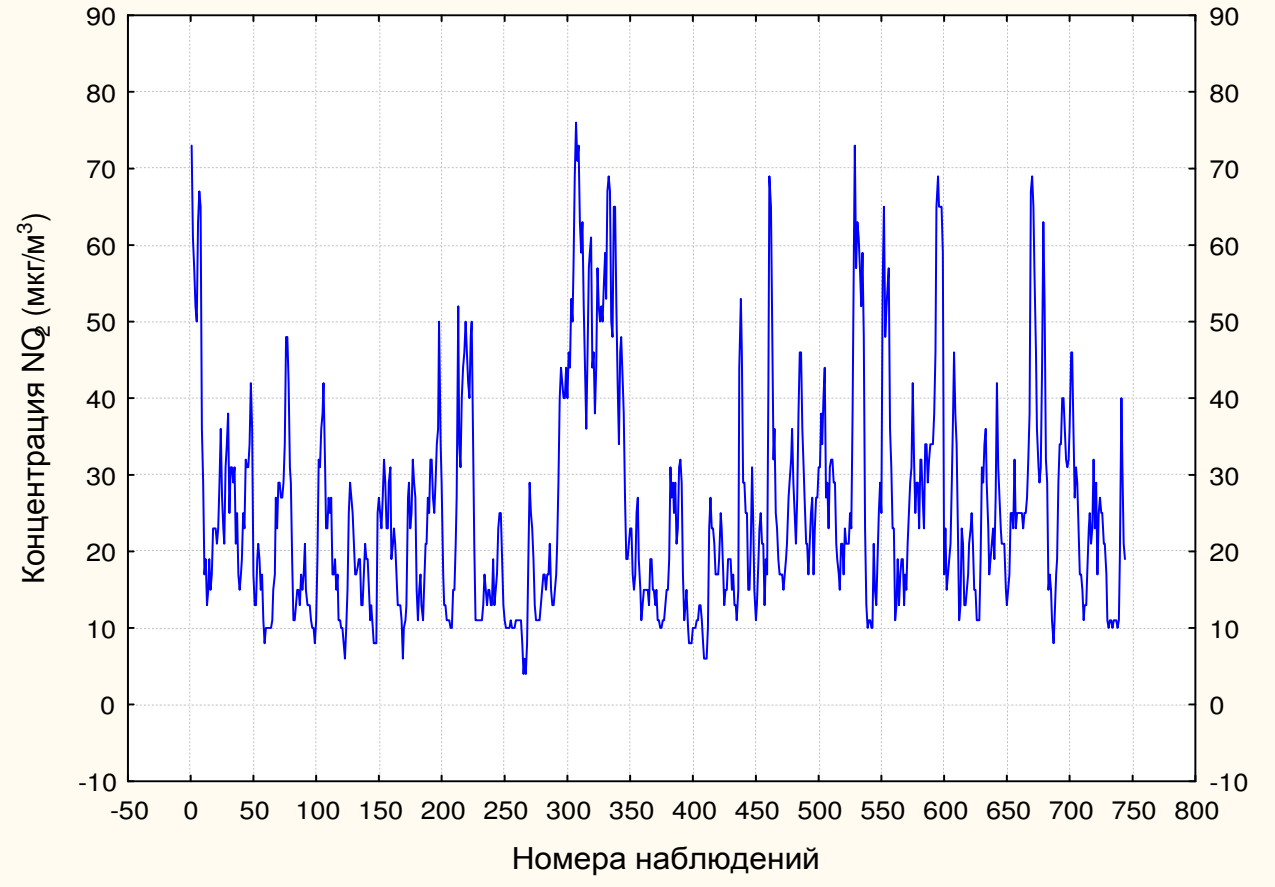

Рис. 2. Зависимость концентрации NO_2 от времени

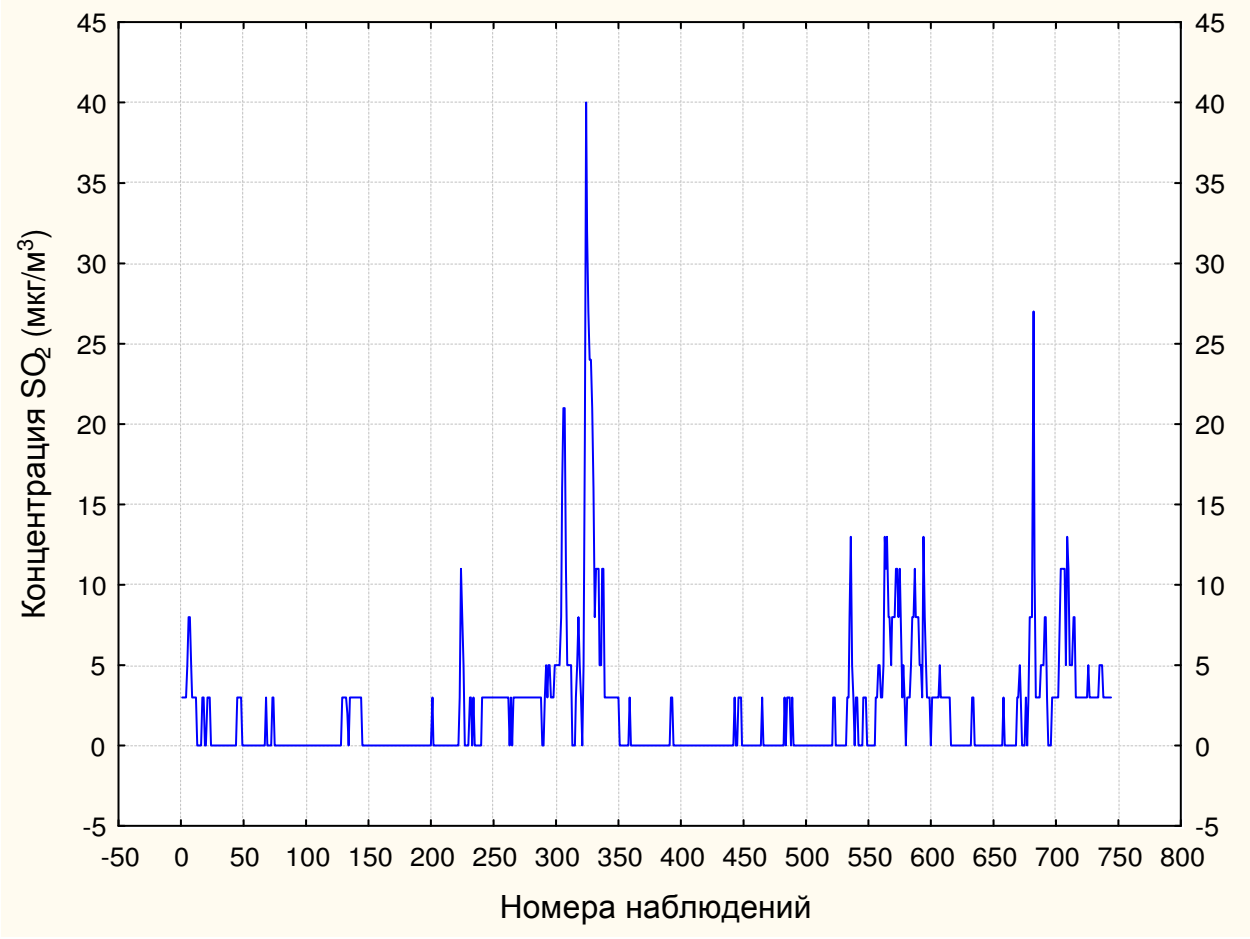

Рис. 3. Зависимость концентрации SO_2 от времени

Лучшей моделью для описания концентрации NO_2 (рис. 2), полученной с помощью метода Бокса-Дженкинса, оказалась модель АРПСС(1,0,1):

$$\tilde{y}_t = 0,9356 + 0,9915 \cdot \tilde{y}_{t-1} + \varepsilon_t + 0,1955 \cdot \varepsilon_{t-1}$$

Лучшей моделью для описания концентрации NO_2, полученной с помощью метода МГУА, оказалась следующая модель:

$$\tilde{y}_t = -0,045 + 1,1 \cdot \tilde{y}_{t-1} - 0,3 \cdot \tilde{y}_{t-2} + 0,13 \cdot \tilde{y}_{t-4} - -0,16 \cdot \tilde{y}_{t-1} \cdot \tilde{y}_{t-2} + 0,21 \cdot \tilde{y}_{t-2} \cdot \tilde{y}_{t-4}.$$

Лучшей моделью для описания концентрации SO_2 (рис. 3), полученной с помощью метода Бокса-Дженкинса, оказалась модель АРПСС(1,0,1):

$$\tilde{y}_t = 0,9928 \cdot \tilde{y}_{t-1} + \varepsilon_t + 0,1780 \cdot \varepsilon_{t-1}.$$

Лучшей моделью для описания концентрации SO_2, полученной с помощью метода МГУА, оказалась следующая модель:

$$\tilde{y}_t = -0,22 + 1,3 \cdot \tilde{y}_{t-1} - 0,77 \cdot \tilde{y}_{t-2} + 0,29 \cdot \tilde{y}_{t-4} - -0,59 \cdot \tilde{y}_{t-1}^2 + 1,9 \cdot \tilde{y}_{t-2}^2 - 0,15 \cdot \tilde{y}_{t-4}^2 - 0,76 \cdot \tilde{y}_{t-1}^3 + +0,88 \cdot \tilde{y}_{t-2}^3 - 2,7 \cdot \tilde{y}_{t-1} \cdot \tilde{y}_{t-2} + 0,44 \cdot \tilde{y}_{t-2} \cdot \tilde{y}_{t-4}$$

Результаты прогнозирования сведены в табл. 1.

Табл. 1. Средний квадрат ошибки прогнозирования загрязнения воздуха

Загрязняющее вещество	*Метод прогнозирования*			
	Бокса-Дженкинса		МГУА	
	Выборка			
	обучающая	проверочная	обучающая	проверочная
оксид углерода (CO)	0,0155	0,0509	0,0119	0,0709
диоксид азота (NO_2)	0,0281	0,0534	0,0249	0,0436
диоксид серы (SO_2)	0,0106	0,0229	0,0052	0,0393

Выводы

Подход Бокса-Дженкинса к анализу временных рядов является весьма мощным инструментом для построения точных краткосрочных прогнозов. Модели ARIMA достаточно гибкие и могут описывать многие временные ряды, встречающиеся на практике. Формальная процедура проверки модели на адекватность проста и доступна. Однако использование моделей ARIMA имеет несколько недостатков:

1. Необходимо относительно большое количество исходных данных. При использовании модели ARIMA для несезонных данных необходимо наличие значительного количества наблюдений (не менее 50), которые не всегда возможно осуществить.
2. Построение удовлетворительной модели ARIMA зачастую требует больших затрат временных и вычислительных ресурсов.

Кроме того, перед построением модели Бокса-Дженкинса необходимо провести анализ данных на однородность. В ряде случаев следует усилить однородность путем преобразования части исходных данных. Здесь к каждому ряду необходимо применять индивидуальный подход, перед построением модели проводить тщательный качественный анализ исходного ряда.

МГУА является лучшим методом для решения задач идентификации и краткосрочного прогноза. Математическая теория МГУА показала, что регрессионный анализ является частным случаем МГУА. Наиболее целесообразно функциональные предикторы использовать в краткосрочном прогнозировании. Надежность таких прогнозов может быть достаточно высокой.

МГУА по сравнению с другими статистическими методами имеет ряд преимуществ:

1. Основное преимущество – позволяет найти как структуру модели, так и ее коэффициенты.
2. Имеется возможность работы с короткими выборками.
3. Имеется возможность работы с зашумленными данными.
4. Имеется возможность решения задач большой размерности, какой является задача прогнозирования полей концентрации загрязняющих веществ.

К недостаткам МГУА можно отнести большое время вычислений, так как с увеличением числа аргументов N и/или степени обобщенного полинома p число возможных вариантов экспоненциально возрастает.

Точность прогнозов, полученных на основе двух рассмотренных выше статистических моделей, вполне удовлетворительна при установившемся состоянии атмосферных процессов и малом (до 8 часов) времени прогноза.

Список литературы

1. Ковальчук П. І. Моделювання і прогнозування стану навколишнього середовища: Навч. посібник. – К.: Либідь, 2003. – 208 с.
2. Берлянд М. Е. Прогноз и регулирование загрязнения атмосферы. – Л.: Гидрометеоиздат, 1985. – 272 с.
3. Примак А. В., Щербань А. Н., Сорока А. С. Автоматизированные системы защиты воздушного бассейна от загрязнения. – К.: Техніка, 1988. – 166 с.
4. Бретшнайдер Б., Курфюрст И. Охрана воздушного бассейна от загрязнений: технология и контроль: Пер. с англ. – Л.: Химия, 1989. – 288 с.
5. Каменева І. П., Яцишин А. В., Полішко Д. О., Попов О. О. Комплексний аналіз екологічної безпеки міста на основі сучасних ГІС-технологій // Екологія довкілля та безпека життєдіяльності. – 2008. – № 5. – С. 41–46.
6. Головін В. В. Інформаційно-логічна структура регіональної системи моніторингу довкілля // Екологія довкілля та безпека життєдіяльності. – 2004. – № 5. – С. 73–79.
7. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. Вып. 1: Пер. с англ. – М.: Мир, 1974. – 408 с.
8. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. – М.: Финансы и статистика, 2003. – 416 с.
9. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. - М.: Радио и связь, 1987. – 120 с.
10. Ивахненко А. Г., Степашко В. С. Помехоустойчивость моделирования. – К.: Наукова думка, 1985. – 216 с.
11. UK National Air Quality Archive. [Электронный ресурс]. (http://www.airquality.co.uk/data_selector.php). Проверено 12.10.2009.
12. Современные методы идентификации систем: Пер. с англ. / Под ред. П. Эйкхоффа. – М.: Мир, 1983. – 400 с.
13. Barron A. R. Predicted squared error: a criterion for automatic model selection // Self-organizing methods in modeling / Edited by S. J. Farlow. – New York: Marcel Dekker, 1984. – P. 87-103.

Поступила в редакцию 21.12.2009