**стохастическое моделирование и прогноз загрязнения атмосферы с использованием нелинейной регрессии**

**А. Пашкова**

# Содержание

ВВЕДЕНИЕ

[1. ОБЗОР И ПОСТАНОВКА ЗАДАЧИ](#_Toc230853229)

2. МЕТОДОЛОГИЯ РЕШЕНИЯ

[2.1 Модель](#_Toc230853231)

2.2 Алгоритм

[3.3 Описание методов](#_Toc230853233)

2.4 Программа

[3. ОПИСАНИЕ РЕЗУЛЬТАТОВ](#_Toc230853235)

3.1 Исходные данные

[3.2 Подготовительный этап](#_Toc230853237)

3.3 Преобразование предикторов

[3.4 Построение регрессионного уравнения](#_Toc230853239)

3.5 Оценка эффективности прогностической схемы

[3.6 Сравнительный анализ результатов](#_Toc230853241)

3.7 Результаты для других городов

[ЗАКЛЮЧЕНИЕ](#_Toc230853243)

ЛИТЕРАТУРА

# ВВЕДЕНИЕ

Разработка методов прогнозирования загрязнения воздуха является одной из важнейших задач, возникающих в рамках проблематики охраны воздушного бассейна. Прогнозы и предупреждения о высоком уровне загрязнения воздуха служат основанием для проведения мероприятий по регулированию выбросов и уменьшению антропогенной нагрузки на окружающую среду в периоды неблагоприятных метеорологических условий.

Актуальность указанной тематики обусловлена тем, что в последние годы, несмотря на применяемые природоохранные меры, проблема чистоты атмосферы городов Российской Федерации не только не решена, но даже обострилась. Как следует из анализа результатов наблюдений, за последнее десятилетие в крупнейших (численностью более 500 тыс. жителей) городах России, высокий уровень загрязнения воздуха сохранился и, согласно прогнозу, такая тенденция будет иметь место в течение ряда лет. Сравнение средних за год концентраций примесей с национальными стандартами качества воздуха показало, что они превышают установленные нормативы предельно допустимых концентраций (ПДК). Максимальные концентрации, превышающие ПДК в десятки раз, регулярно регистрировались в большинстве (55-80%) крупнейших городов страны.

Проводимая в нашей стране природоохранная политика предусматривает необходимость регулирования (т.е. сокращения) выбросов в периоды неблагоприятных метеорологических условий (НМУ). Чтобы обеспечить практическую возможность такого сокращения, соответствующие органы управления, предприятия и др. должны быть оповещены заранее о необходимости своевременного сокращения выбросов вредных примесей.

В этой связи практический интерес представляют краткосрочные прогнозы - большей частью с заблаговременностью от 1 до 3 суток. Их внедрение при обеспечении регулирования выбросов позволяет в ближайшее время без значительных затрат улучшить состояние воздушного бассейна и предотвратить появление высоких уровней загрязнения воздуха.

Целью дипломной работы является разработка стохастической модели, которая позволяет производить краткосрочный прогноз абсолютных уровней загрязнения воздуха на территории городов методами регрессионного анализа. В соответствии с указанной целью в работе были поставлены следующие задачи:

* выбор метода и разработка статистической модели прогноза загрязнения;
* выбор вычислительных алгоритмов;
* программная реализация математической модели;
* построение соответствующие этой модели прогностических схем загрязнения воздуха по данным измерений и оценка их эффективности на основе данных, полученных в нескольких городах;
* проведение анализа результатов, полученных на основе прогностических схем.

Оперативные методы прогноза загрязнения воздуха изложены в действующем «Руководстве по прогнозу загрязнения воздуха» РД.52.04.306-92 [5] и внедрены во всех управлениях Росгидромета. Работы по прогнозированию загрязнения воздуха проводятся в 250 городах Российской Федерации, предупреждения его возможного роста передаются более чем на 5000 предприятий, на которых принимаются конкретные меры по снижению выбросов в неблагоприятные периоды.

Однако, методы прогноза, используемые в оперативной практике Росгидромета, не позволяют предсказывать наибольшие концентрации примесей, формирующиеся в воздухе отдельных районов промышленных городов. А ведь предотвращение появления высоких концентраций имеет существенное значение для решения проблемы защиты атмосферы от загрязнения в период НМУ.

Модель, разработанная в данной работе, не имеет подобных недостатков, так как позволяет прогнозировать не относительные характеристики загрязнения воздуха на территории городов, а их абсолютные уровни.

# 1. ОБЗОР И ПОСТАНОВКА ЗАДАЧИ

Основная задача данной работы ⎯ краткосрочный прогноз максимальных концентраций загрязняющих примесей. Рабочая гипотеза состоит в том, что формирование опасного загрязнения воздуха на достаточно большой территории города обусловлено процессами рассеивания антропогенных выбросов, то есть, процессами, регулярно встречающихся при обычных технологических режимах работы промышленных источников или в суточном ходе изменения в интенсивности движения автотранспорта. Задача прогноза последствий аварийных выбросов вредных веществ в атмосферу вследствие нарушения технологических процессов на предприятиях в данной работе не рассматривается.

Количественной оценкой наибольшего загрязнения может служить максимальная (CМАХ) из измеренных за день концентраций на посту. Измеренные на каком-то конкретном посту концентрации значительно варьируются в течение суток за счет изменчивости:

* выбросов,
* «внешних» метеорологических условий,
* локальных флуктуаций концентраций, связанных с флуктуациями локальных метеопараметров, определяющих перенос и рассеивание примесей.

При переходе к дневным максимумам влияние локальных флуктуаций сильно уменьшается, уменьшается также роль систематических изменений выбросов в течение суток, так что связь с внешними метеорологическими условиями становится более устойчивой.

Решение основной задачи работы сводится к разработке метода прогноза максимальной за день концентрации в точке наблюдения. При такой постановки задачи необходимо выполнение следующего требования: разрабатываемый метод должен быть эффективным как для области часто наблюдаемых значений, так и для высоких квантилей (экстремумов) функции распределения CМ.

При наличии многолетних рядов наблюдения за загрязнением атмосферы для этой цели можно применить хорошо разработанный математический аппарат регрессионного анализа. Он позволяет на основе статистического анализа зависимостей между переменными прогнозировать одну из них, если известны значения других.

Воспользуемся для решения поставленной задачи методом регрессионного анализа с определенными преобразованиями - достаточно простым и удобным математическим аппаратом.

В рамках регрессионного анализа модель представляется в виде:

, (1)

где СМАХ – предиктант (в нашем случае максимальная концентрация рассматриваемой примеси за сутки), xi – предикторы (в качестве предикторов используют различные метеорологические характеристики), а bi – коэффициенты регрессии, которые требуется оценить.

ε ⎯ вектор ошибки (остатка). Предполагается, что ε ⎯ независимая случайная величина, имеющая нормальное распределение N(0, σ2).

Уравнение решается методом наименьших квадратов, т.е. из условия минимума среднего квадрата ошибок. После того как определены параметры bi, получаем прогностическое уравнение, которое отличается от исходного модельного уравнения тем, что не содержат случайной ошибки.

Успешное применение выбранного математического аппарата требует обеспечить выполнение следующих условий: двумерное (совместное) распределение плотности вероятности переменных (предиктанта с каждым из предикторов) подчиняется нормальному закону; форма связи между переменными должна быть близкой к линейной. При этом, однако, заранее известно, что экстремумы случайной величины обычно не распределены нормально, и что связи между предикторами и предиктантом могут оказаться нелинейными. Поэтому перед тем, как строить уравнение регрессии, необходимо преобразовать переменные таким образом, чтобы линеаризовать и нормализовать соответствующие связи.

После соответствующих преобразований переменных, включающих нормализацию предиктанта и предварительное исключение нелинейности связей, получаем рабочее уравнение:

, (2)

Задача с преобразованными предикторами решается методом пошагового регрессионного анализа. Данный вид анализа позволяет включать в схему только те факторы, которые имеют значимую корреляцию с показателями загрязнения. Применение такого аппарата обусловлено тем, что нет никакой гарантии, что между используемыми предикторами отсутствует тесная корреляционная связь. Если же такая связь существует, то соответствующая система уравнений метода наименьших квадратов, используемая для определения коэффициентов в уравнении регрессии, оказывается плохо обусловленной, а ее решение может привести к накоплению вычислительных ошибок.

После того как определены параметры bi, получаем прогностическое уравнение. По этому уравнению рассчитываются прогностические значения максимальной концентрации загрязняющей примеси. Применимость этого уравнения проверяется его испытанием на независимой выборке.

Из значений СМАХ и СМАХПРОГ, полученных с использованием прогностических уравнений по зависимому и независимому рядам, формируется таблица результатов прогноза и рассчитываются статистические характеристики эффективности прогноза максимальной концентрации примеси.

Исходные данные для разработки стохастических моделей были предоставлены ГУ «ГГО» по таким городам, как Санкт-Петербург, Обнинск, Милан, Мадрид, Новосибирск и др.

# 2. МЕТОДОЛОГИЯ РЕШЕНИЯ

## 

## 2.1 Модель

Регрессионный анализ — это эффективный метод, который позволяет анализировать значительные объемы информации с целью исследования вероятной взаимосвязи двух или больше переменных.

В регрессионном анализе рассматривается связь между одной, зависимой, переменной и несколькими другими независимыми переменными. Эта связь представляется с помощью математической модели, то есть уравнением, которое связывает зависимую переменную с независимыми. В рамках регрессионного анализа модель представляется в виде:

, (3)

где СМАХ – предиктант (в нашем случае максимальная концентрация рассматриваемой примеси за сутки), Xi – предикторы (в качестве предикторов используют различные метеорологические характеристики и концентрации других загрязняющих примесей), а bi – коэффициенты регрессии, которые требуется оценить.

Регрессионный анализ используется по двум причинам.

1. Описание зависимости между предикторами и предиктантом помогает установить наличие возможной причинной связи.
2. Получение аналитической зависимости между переменными дает возможность предсказывать будущие значения СМАХ по значениям предикторов.

Успешное применение этого математического аппарата требует выполнение двух условий:

1. Функции распределения переменных (предиктанта и каждого из предикторов) подчиняются нормальному случайному закону.
2. Форма связи между переменными должна быть близкой к линейной.

## 2.2 Алгоритм

Предварительный этап разработки прогностической схемы состоит в подготовке исходного ряда данных:

1. Ряд разбивается на «обучающую» и «независимую» выборки. В данной работе прогностическая модель загрязнения атмосферы разрабатывается с использованием длительного ряда данных наблюдений. Две трети ряда рассматриваются, как «обучающая» выборка для построения прогностической схемы, а оставшаяся одна треть применяется для проверки её эффективности на независимом материале (т.е. как «независимая» выборка). К «независимой выборке» относятся данные наблюдений, соответствующие неделям года с номерами, кратными трём (т. е. третья, шестая, девятая и т.д. недели). Остальные данные относятся к «обучающей» выборке.
2. По «обучающей» выборке строится функция распределения суточных максимумов концентраций и определяется её 60-ый процентиль С60.
3. Устанавливается граничное значение СГР для прогноза суточных максимумов, которое принимается равным С60.

Прогноз CMAX осуществляется по следующим правилам:

1. Если максимальная за предыдущие сутки C’MAX концентрация была ниже СГР, то прогнозируемая максимальная концентрация на очередные сутки CMAXПРОГ принимается равной C’MAX («инерционный прогноз»).
2. Если максимальная за предыдущие сутки C’MAX концентрация была выше или равна СГР, то прогноз осуществляется с использованием прогностических схем.

Применение метода линейной регрессии требует, чтобы корреляционные связи между предиктантом с каждым из предикторов были близки к линейным, однако это условие не всегда выполняется. Для исключения нелинейности связей предикторы нужно преобразовать с помощью кривых зависимости показателя загрязнения воздуха от отдельных метеопараметров, построенных по использованному для разработок материалу наблюдений. При этом каждое значение предиктора меняется на соответствующее ему среднее значение характеристики загрязнения.

* Для каждой градации предиктора (их должно быть не менее 5) рассчитать среднее значение CMAX. При недостаточном количестве случаев в одной из градаций, она объединяется с соседних. Таким образом, получаем набор точек с абсциссами M(CMAX) и ординатами, соответствующими серединам отрезков осреднения.
* Построить график кусочно-линейной функции, у которой полученные точки являются угловыми.
* Каждому значению преобразованного предиктора сопоставляется значение кусочно-линейной функции в соответствующей точке.

Связь преобразованных таким образом предикторов с предиктантом в значительной степени линеаризуется. Этот прием позволяет учесть реальный вид связи в каждом конкретном случае. Он близок к так называемому «кусочно-линейному» преобразованию, применяемому при построении моделей для прогноза погоды.

При возникновении трудностей, связанных с тем, что данные, подчиняющиеся какому-нибудь несимметричному распределению, должны быть подвергнуты анализу, теория которого разработана в основном для нормального распределения, можно преобразовать эмпирическое распределение в нормальное («нормализовать переменные») и затем продолжить анализ на базе известной теории.

Для нормализации переменных используется стандартное преобразование выборочной функции распределения в нормальную (гауссову) со средним, равным 0, и стандартным отклонением, равным 1. Это преобразование осуществляется по формуле

, (4)

где Ф-1(t) – обратная функция к функции распределения нормальной случайной величины со средним значением ноль и стандартным отклонением единица, а F(x) - выборочная функция распределения рассматриваемой случайной величины X.

Задача с преобразованными предикторами решается методом многомерной пошаговой регрессии. На каждой итерации этого метода ищется предиктор, имеющий наибольшую связь с предиктантом. Таким образом определяются наиболее значимые предикторы, которые следует включить в уравнение регрессии.

Если значимыми оказались два предиктора, соответствующие двум срокам измерения одного и того же метеорологического параметра, то в уравнение регрессии включается тот, который больше связан с предиктантом. В итоге должны остаться 4 – 7 наиболее информативных предикторов, связь которых с предиктантом наиболее значима.

Данный вид анализа позволяет включать в схему только те факторы, которые имеют значимую корреляцию с показателями загрязнения. Применение такого аппарата также обусловлено тем, что нет никакой гарантии, что между используемыми предикторами отсутствует тесная корреляционная связь. Если же такая связь существует, то соответствующая система уравнений метода наименьших квадратов, используемая для определения коэффициентов в уравнении регрессии, оказывается плохо обусловленной, а ее решение может привести к накоплению вычислительных ошибок. После того как определены параметры bi, получаем стохастическую модель процесса, которая может быть для краткости представлена в виде:

, (5)

Здесь [Xi] –преобразованные предикторы, I – количество использованных предикторов, b0 и bi – коэффициенты регрессии. Значения b0 и bi определяется с помощью метода наименьших квадратов. По этому уравнению рассчитываются прогностические значения максимальной концентрации загрязняющей примеси.

Из значений Смах и Смахпрог, полученных с использованием прогностических уравнений по зависимому и независимому рядам, формируется таблица результатов прогноза. Рассчитываются статистические характеристики эффективности прогноза максимальной концентрации примеси.

Эффективность разработанных прогностических схем проверяется по зависимым (использованным для построения уравнения регрессии) и независимым (не использовавшимся для построения уравнения регрессии) данным наблюдений.

Оправдываемость индивидуального прогноза максимальной концентрации примеси Смахпрог за конкретные сутки оценивается путем сопоставления этой прогностической концентрации с определенной по данным наблюдений фактической максимальной за сутки концентрацией Смах. Прогноз считается оправдавшимся, если при Смах > ПДК выполняется условие:

, (6)

или если при Смах  ПДК выполняется условие

, (7)

где ПДК - установленная Минздравом РФ максимальная разовая предельно допустимая концентрация примеси в атмосферном воздухе населенных мест.

## 

## 2.3 Описание методов

**1. Нахождение СГР.**

n-й процентиль - это такое значение, ниже которого расположено n процентов наблюдений рассматриваемой переменной. График функции распределения случайной величины X имеет ступенчатый вид. Значение функции F(X) равно:

, k = 0…M-1, (8)

где M – объём выборки, а k – порядковый номер события в упорядоченном по возрастанию массиве. Как известно, то α-квантиль однозначно задаётся [уравнением](http://ru.wikipedia.org/wiki/%D0%A3%D1%80%D0%B0%D0%B2%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5): F(xα) = α. Значит за 60 процентиль можно принять элемент с порядковым номером k = 0.6M (округление производим в большую сторону).

**2. Нормализации.**

Нормализация осуществляется по формуле:

, (9)

График функции распределения случайной величины X имеет ступенчатый вид. Значение функции F(X) равно:

, k = 0…M-1, (10)

Так как при k = 0 F(Xk) обращается в ноль, то [Xk] становится равным минус бесконечности, что является нежелательным, заменим формулу (10) на:

, k = 0…M-1. (11)

При достаточно больших M погрешность в значениях F(Xk), вычисляемых по формуле(11) становится мала. При этом F(X) нигде не обращается в ноль или M, а значит [Xk] принимают только конечные значения.

Вместо функции, обратной к функции распределения нормальной случайной величины, Ф-1 можно использовать её аппроксимацию (погрешность e-16).

**3. Пошаговая регрессия.**

Имеется набор независимых переменных X1…Xn, которые являются кандидатами на роль предикторов СМАХ, и случайная выборка объема М. Рассмотрим стандартную пошаговую процедуру (F-метод), которая состоит из правила включения переменных и правила исключения. Включение и удаление переменных осуществляются с помощью критерия, который имеет F-распределение, и называется либо F-включения, либо F-удаления.

Более точно, предположим, что в набор *с* уже включено k переменных, k = 0, 1… M-1. Тогда значение F-включения для переменной X (не входящей в *с*) вычисляется по формуле:

, (12)

где rСмахX\*с – множественный коэффициент корреляции. Величина F служит статистикой критерия для проверки гипотезы о том, что предсказание СМАХ значимо не улучшается при включении X в набор *с*. Аналогично, величина F-удаления для какой-либо переменной X из *с* служит статистикой критерия для проверки гипотезы о том, что набор *с'* получающийся из *с* при удалении X и содержащий k'=k-1 переменных, предсказывает СМАХ так же хорошо как и набор *с*.

, (13)

Правило остановки, обычно используемое в стандартной процедуре, основано на задании допустимого минимума F-включения. По умолча­нию предполагается, что минимум F-включения равен 4. Для удаляемых переменных также выбирается допустимый минимум. F-удаления величина должна быть меньше минимума F-включения. По умолчанию принимается, что ми­нимум F-удаления равен 3,9. Рассмотрим теперь подробно шаги стандартной процедуры.

*Шаг 0*. Вычисляются величины F-включения для i = 1…n. Статистика критерия дается выражением

, (14)

которое получается из формулы (?) подстановкой k=0.

*Шаг 1*. Переменная Xi1, которой отвечает наибольшее значение F-включения (или, что эквивалентно, наибольшая величина квадрата коэффициента корреляции с СМАХ), выбирается как наилуч­ший предиктор. Величина F-удаления для Xi1 в этом случае совпадает с величиной F-включе­ния. Далее вычисляются значение F-включения и F-удаления.

Если все вычисленные значения F-включения меньше установленного минимума, то далее выполняется шаг S.В противном случае происходит переход на шаг 2.

*Шаг 2*. Переменная Хi2, имеющая наибольшее значение F-вклю­чения (или, что эквивалентно, наибольший квадрат частного коэффициента корреляции с СМАХ при фиксированном значении X*i1*), выбирается как наилучший предиктор для СМАХ при условии, что уже выбрана переменная Xi1. Если все значения F-включения меньше установленного минимума, то далее выполняется шаг S. В противном случае происходит переход на шаг 3.

*Шаг 3*. *а)* Пусть L обозначает набор из *l* независимых перемен­ных, которые включены в уравнение регрессии. Если какое-либо из значений F-удаления для переменных из L меньше, чем соответ­ствующий минимум, то переменная, которой соответствует наи­меньшее значение F-удаления, удаляется из набора и выполняется шаг Зb с заменой *l* на *l*-1. Если для всех переменных, не входя­щих в L, значение F-включения меньше установленного минимума, то выполняется шаг S. В противном случае в набор L добавляется переменная, которой соответствует максимальное значение F-включения, и l заменяется на l + 1. *b)* Вычисляются значения F-удаления между СМАХ и переменной Хi из L при заданных остальных *l*-1 переменных из L и значение F-включения между СМАХ и каждой переменной Xi, не входящей в L, при данных переменных из L.

*Шаги 4, 5...* Рекуррентно повторяется шаг 3. шаг S выполняется а) если F-включения для всех переменных, не входящих в L, меньше установленного минимума, b) если для всех переменных из L значение F-удаления больше установленного минимума или с) число включенных переменных равно р.

*Шаг S*. Вывод результатов.

**4. Метод наименьших квадратов.** Метод наименьших квадратов - простой и быстрый способ получить неизвестные параметры в функциональной зависимости () и оценить их погрешности. Минимизируется сумма квадратов отклонений реально наблюдаемых CMAX от их оценок CMAXПРОГ (имеются в виду оценки с помощью прямой линии, претендующей на то, чтобы представлять искомую регрессионную зависимость):

, (15)

Для решения задачи регрессионного анализа методом наименьших квадратов вводится понятие функции невязки:

, (16)

Условие минимума функции невязки – равенство нулю всех частых производных, то есть:

, (17)

Полученная система является системой I+1 линейных уравнений с I+1 неизвестными b0...bI. Так как матрица системы уравнений является симметричной, то система решается методом LU-разложения.

**5. Сортировка.** Сортировка массивов осуществляется методом разделения (Quicksort). Шаги алгоритма таковы:

* 1. Выбираем в массиве некоторый элемент, который будем называть опорным элементом. С точки зрения корректности алгоритма выбор опорного элемента безразличен. С точки зрения повышения эффективности алгоритма выбираться должна медиана, но без дополнительных сведений о сортируемых данных её обычно невозможно получить.
  2. Операция разделения массива: реорганизуем массив таким образом, чтобы все элементы, меньшие или равные опорному элементу, оказались слева от него, а все элементы, большие опорного — справа от него. Обычный алгоритм операции:
     + два индекса — *l* и r, приравниваются к минимальному и максимальному индексу разделяемого массива соответственно;
     + вычисляется опорный элемент m;
     + индекс *l* последовательно увеличивается до m или до тех пор, пока *l*-й элемент не превысит опорный;
     + индекс r последовательно уменьшается до m или до тех пор, пока r-й элемент не окажется меньше опорного;
     + если r = *l* — найдена середина массива — операция разделения закончена, оба индекса указывают на опорный элемент;
     + если *l* < r — найденную пару элементов нужно обменять местами и продолжить операцию разделения с тех значений *l* и r, которые были достигнуты. Следует учесть, что если какая-либо граница (*l* или r) дошла до опорного элемента, то при обмене значение m изменяется на r или *l* соответственно.
  3. [Рекурсивно](http://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BA%D1%83%D1%80%D1%81%D0%B8%D1%8F) упорядочиваем подмассивы, лежащие слева и справа от опорного элемента.
  4. Базой рекурсии являются наборы, состоящие из одного или двух элементов. Первый возвращается в исходном виде, во втором, при необходимости, сортировка сводится к перестановке двух элементов. Все такие отрезки уже упорядочены в процессе разделения.

Алгоритм называется быстрой сортировкой, поскольку для него оценкой числа сравнений и обменов является O(n\*lg n).

## 2.4 Программа

Для автоматизации расчёта по описанной модели была составлена прикладная программа. Она была написана на языке C++ в среде разработки Microsoft Visual Studio 2005, и её объём составляет несколько тысяч операторов. Программа состоит из трёх основных модулей: управляющая часть, блок с функциями, реализующими логику всех необходимых методов, и модулем, который отвечает за взаимодействие программы с пользователем. Таким образом, программа составлена в соответствии с шаблоном проектирования модель-вид-контроллер (model-view-controller).

Управляющая часть – контроллер – позволяет разграничить модель и представление модели. Она включает в себя последовательность команд, которые вызывают требуемые функции.

Основные функции реализованы во втором модуле программы. В нём реализованы все этапы, необходимые для построения прогностической схемы, начиная от разбиения всех данных на зависимую и независимую выборки, и заканчивая прогнозированием концентрации загрязняющего вещества для введённых параметров.

Основные функции, реализованные во втором модуле программы:

* Считывание данных из файла. Для удобства пользователя исходные данные читаются из файла в формате xls.
* Деление выборки на зависимую и независимую части. В «обучающую» выборку попадают первые две недели из каждых трёх, а в независимую – каждая третья неделя. Отсчёт начинается от первого понедельника года.
* Определение граничного значения и отбрасывание из выборки неподходящих данных. В качестве аргумента данной функции передаётся процентиль, по которому определяется граничное значение. Если значение концентрации меньше граничного, то соответствующая строка отбрасывается из выборки.
* Линеаризация. Функция преобразует предикторы так, чтобы исключить нелинейность связей. Для этого все значения предиктора делятся на несколько интервалов и для каждого интервала находится среднее значение концентрации. Таким образом, строится кусочно-линейная функция, которая проходит через точки с координатами: седина интервала, соответствующее среднее значение. Координаты этих точек запоминаются в специальной структуре для дальнейшего использования.
* Нормализация. Данная функция для каждой преобразуемой величины вычисляет значение выборочной функции распределения вероятности. Затем она сопоставляет этому значению значение функции, обратной к функции распределения нормальной случайной величины. Таким образом, происходит преобразование всех случайных величин к нормальным.
* Определение предикторов, которые будут включены в модель. Необходимо выявить наиболее значимые предикторы. Для этого используется метод пошагового регрессионного анализа. При пошаговом алгоритме для критерия Фишера назначаются значения включения и исключения переменных. За счёт этого можно регулировать количество включаемых в модель предикторов.
* Функция составления системы для метода наименьших квадратов и её решения - определение коэффициентов регрессии.
* Проверка полученной схемы на «обучающей» и «независимой» выборках. Функция с помощью сохранённых параметров для преобразования линеаризует, нормализует необходимые величины и вычисляет прогностическое значение концентрации загрязняющего вещества.
* Сохранение. Функция записывает все промежуточные данные и результаты вычислений в выходной файл out.xls. Это позволяет пользователю легко анализировать результаты, строить все необходимые графики и оценки для модели средствами MS Excel, OpenOffice Calc, Statistica и другие.

Также в этом модуле программы реализованы вспомогательные функции такие как:

* Вычисление функции вероятности нормального распределения со средним 0 и ско 1.
* Вычисление обратной функции вероятности нормального распределения.
* Вычисление определителя матрицы.
* Решение системы с помощью LU-разложения.

Благодаря третьему модулю, который отвечает за визуализацию, пользователь имеет возможность, получать некоторые промежуточные результаты, в зависимости от них вводить различные параметры и корректировать работу программы.

Весь алгоритм программы можно представить в виде блок-схемы (рис.1).



Рис. 1. Блок-схема алгоритма.

# 3. ОПИСАНИЕ РЕЗУЛЬТАТОВ

## 

## 3.1 Исходные данные

Исходные данные (рис. 2) для разработки стохастической модели были предоставлены ГУ "ГГО" по станции, расположенной на ул. Пестеля (г. Санкт-Петербург). Эти данные характеризуют загрязнение атмосферного воздуха озоном за 2002 год.

В разрабатываемой стохастической модели связь межу предиктантом и предикторами описывается в виде , где

* - предиктант, максимальная за сутки концентрация озона (мкг/м3);
*  – линейная функция от n предикторов;
* в качестве предикторов  используются следующие величины:

 - максимальная концентрация озона (мкг/м3) за предыдущие сутки;

-7- концентрация оксида азота (мкг/м), измеренная в 7 часов;

-7 - концентрация диоксида азота (мкг/м3), измеренная в 7 часов;

 - концентрация озона (мкг/м3), измеренная в 7 часов;

 - скорость ветра (м/с) в 6 и 12 часов;

 - направление ветра (дес.град) в 6 и 12 часов;

 - атмосферное давление (мб) в 6 и 12 часов;

 - температура воздуха (°С) в 6 и 12 часов;

 - относительная влажность воздуха (%) в 6 и 12 часов;

 - атмосферные явления (шифр), наблюдаемые в 6 и 12 часов.

Длина массива данных составляет 273.

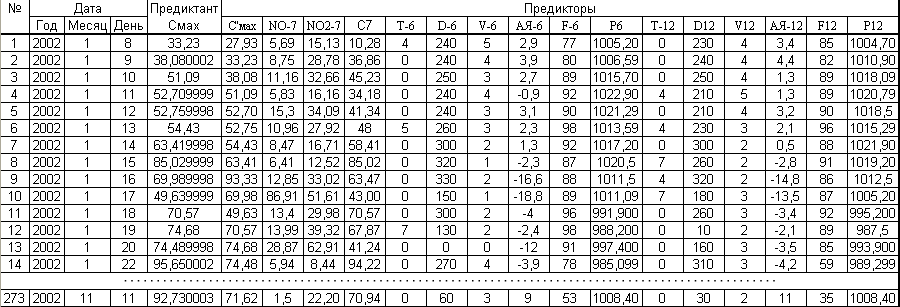


Рис.2. Исходные данные

## 3.2 Подготовительный этап

Чтобы вычислить коэффициенты функции  при каждом из предикторов необходимо подготовить исходные данные. С этой целью:

* 1. Ряд значений  максимальных за сутки концентраций озона разбивается на обучающую и независимую выборки. Далее все преобразования производятся только с обучающей выборкой.
  2. Осуществляется цензурирование выборки: сортируется массив данных в соответствие с ростом переменной , где  - значение  за сутки до срока, на который дается прогноз. Находим значение 60%-ного квантиля функции распределения . - В рассматриваемом примере значению 60%-ного квантиля функции распределения  соответствует концентрация  = 63,42 мкг/м3. Делим исходные данные на две группы. В первую группу, объем которой составляет 60% от общей выборки, войдут значения , сопутствующие метеоусловия и расчетные параметры, которые наблюдались при  < 63,42 мкг/м3. Для этой группы прогноз  осуществляется по уравнению . Во вторую группу, объем которой составляет 40% от общей выборки, - значения ,  , , , сопутствующие метеоусловия и расчетные параметры, которые наблюдались при  >= 63,42 мкг/м3. Для этой группы прогноз  осуществляется с использованием прогностических схем.

## 3.3 Преобразование предикторов

1. Линеаризация (рис.3). Исключение нелинейности связей между предиктантом  и предикторами.



Рис. 3. Линеаризация предикторов

1. Для повышения эффективности модели в области высоких значений необходимо нормализовать все переменные, входящие в модель (рис. 4). Для этого каждое значение переменной сопоставляем значение обратной функции к функции распределения нормальной случайной величины со средним значением ноль и стандартным отклонением единица. Ее аргументом является значение выборочной функции распределения рассматриваемой случайной величины.



Рис. 4. Графики функции распределения предиктанта (до и после преобразования)

## 3.4 Построение регрессионного уравнения

Подготовленные исходные данные применяются для получения коэффициентов уравнения. Применяем метод пошаговой многомерной регрессии. Устанавливаем параметры: F-включения = 2, F-удаления = 1,9. После девяти итераций получаем предикторы, которые необходимо включить в систему. Так как два предиктора, F-6 и F-12, соответствуют одному и тому же метеорологическому параметру, в систему включаем только тот, который больше коррелирует с предиктантом, то есть F-6.

Таблица 1. Оценка предикторов

|  |  |  |  |
| --- | --- | --- | --- |
| F-включения | F-удаления | In/Out | Предиктор |
| 89,0213 |  | In |  |
| 31,3170 |  | In | С7 |
| 3,4080 |  | In | F-6 |
| 6,1487 |  | In | D-6 |
| 4,7916 |  | In | NO2-7 |
| 2,2418 |  | In | F12 |
| 3,5270 |  | In | V12 |
| 3,0150 |  | In | Т-6 |
| 2,3492 |  | In | NO-7 |
|  | 0,4037 | Out | Р6 |
|  | 0,0001 | Out | Т-12 |
|  | 0,0001 | Out | D12 |
|  | 0,4391 | Out | V-6 |
|  | 0,0083 | Out | АЯ-12 |
|  | 0,9863 | Out | АЯ-6 |
|  | 0,6304 | Out | Р12 |

После того как определены наиболее значимые предикторы, ищем коэффициенты b0 и bi. Для этого используем метод наименьших квадратов. В результате получаем прогностическое уравнение:

, (18)

Для того чтобы получить окончательные значения концентрации , необходимо выполнить преобразование, обратное нормализации.

## 

## 3.5 Оценка эффективности прогностической схемы

По полученному уравнению определяется ожидаемое значение .

Убедимся в эффективности построенной прогностической схемы с помощью корреляционного графика между спрогнозированными и измеренными концентрациями озона для «обучающей выборки» (рис. 5).



Рис.5. Корреляционный график измеренных и прогностических суточных максимумов концентрации озона

Квадрат коэффициента корреляции между данными измерения и прогноза составил 0,83, а угловой коэффициент регрессии оказался равным 1,06.

Проверим применимость прогностической схемы на «независимой выборке». Корреляционный график, показывающий связь между прогностическими и фактическими концентрациями озона, приведен на рисунке (рис. 6).



Рис.6. Корреляционный график измеренных и прогностических суточных максимумов концентрации озона

Эффективность построенной прогностической схемы подтверждают выполненные оценки: квадрат коэффициента корреляции между данными измерения и прогноза составил 0,78, а угловой коэффициент регрессии оказался равным 1,09.

Оправдываемость индивидуального прогноза максимальной концентрации примеси Смахпрог за конкретные сутки оценивается с помощью формул 6-7. Прогноз оказался оправдавшимся для «обучающей выборки» в 89% случаев, а для «независимой выборки» - в 87%.

Таким образом, построенную схему прогноза концентраций озона следует считать эффективной.

## 3.6 Сравнительный анализ результатов

Сравним результаты, построенные с помощью приведённой выше прогностической схемы и с помощью метода инерционного прогноза. Полученные оценки сведём в таблицу.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Прогностическая схема | | Инерционный прогноз | |
| "Обучающая выборка" | "Независимая выборка" | "Обучающая выборка" | "Независимая выборка" |
| Уравнение линии тренда |  |  |  |  |
| Квадрат коэффициента регрессии R2 | 0,8256 | 0,7808 | 0,7305 | 0,6585 |
| Оправдываемость | 89% | 87% | 75% | 79% |

## 

## 3.7 Результаты для других городов



Рис.7. Корреляционный график измеренных и прогностических суточных максимумов концентрации озона (Милан)

прогноз концентрация примесь программа



Рис.8. Корреляционный график измеренных и прогностических суточных максимумов концентрации озона (Winfield)



Рис.9. Корреляционный график измеренных и прогностических суточных максимумов концентрации озона (Обнинск)



Рис.9. Корреляционный график измеренных и прогностических суточных максимумов концентрации озона (Новосибирск)

# ЗАКЛЮЧЕНИЕ

При выполнении данной работы получены следующие основные результаты:

1. Разработан статистический метод прогноза максимальных за день концентраций примесей в отдельных точках города и построены соответствующие прогностические схемы (для Санкт-Петербурга, Новосибирска, Милана). Проведенный анализ результатов и графические материалы показали, что применение выбранной стохастической модели позволяет прогнозировать максимальные концентрации в период повышенного загрязнения воздуха достаточно эффективно. Оценка эффективности схем по использованному и независимому материалу показала, что оправдываемость прогнозов наибольших концентраций составила в среднем 80%, а их предсказуемость — 85%.
2. Разработана и отлажена компьютерная программа на алгоритмическом языке C++, реализующая указанный метод и позволяющая прогнозировать суточные максимумы концентрации вредных примесей в городах. С учетом универсальности использованного алгоритма, модно надеяться, что данная программа может быть также использована для решения задачи прогнозирования в других областях.
3. Показано, что применение метода множественной линейной регрессии с предварительным исключением нелинейности связей и нормализацией предиктанта позволяет успешно прогнозировать максимальные концентрации и их экстремумы в период повышенного загрязнения воздуха в городе. Необходимость преобразований переменных вызвана нелинейной формой зависимости предиктанта от предикторов и асимметрией распределения функции плотности вероятности.

Выполненная работа направлена на повышение качества оценки уровня загрязнения на основе использования новой методики прогноза в целях повышения эффективности охраны чистоты воздушного бассейна. Полученная модель может быть рекомендована для оперативного использования в промышленных городах, в том числе для составления предупреждений об опасных уровнях загрязнения воздуха.

# ЛИТЕРАТУРА

1. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. Пер. с английского. - М.: Мир, 1982. - 488с.
2. Берлянд М.Е. Современные проблемы атмосферной диффузии и загрязнения атмосферы.// Л. :Гидрометеоиздат, 1975.—448с.
3. Кириллова В.И. …. Автореферат диссертации ….. кандидата геогр. наук
4. Норман Дрейпер, Гарри Смит. Прикладной регрессионный анализ. Множественная регрессия = Applied Regression Analysis. — 3-е изд. — М.: «Диалектика», 2007. — С. 912.
5. «Руководство по прогнозу загрязнения воздуха» РД 52.04.306-92. // СПб.: Гидрометеоиздат, 1993. - 104 с.
6. Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980, с. 456.