Алчевск

Организация и математическое планирование эксперимента

С. И. Кулакова
Л. Е. Подлипенская
Д. А. Мельничук

Учебное пособие

Алчевск

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ЛУГАНСКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ДОНБАССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ ИНСТИТУТ»

С. И. Кулакова, Л. Е. Подлипенская, Д. А. Мельничук

ОРГАНИЗАЦИЯ И МАТЕМАТИЧЕСКОЕ ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

Учебное пособие

Рекомендовано Ученым советом ГОУ ВО ЛНР «ДонГТИ»

Алчевск
2021

УДК 519.242 (075.8)
ББК В172я73
О 64

Кулакова Светлана Ивановна — старший преподаватель кафедры высшей математики ГОУ ВО ЛНР «ДонГТИ» (г. Алчевск);

Подлипенская Лидия Евгеньевна — кандидат технических наук, доцент кафедры экологии и безопасности жизнедеятельности ГОУ ВО ЛНР «ДонГТИ» (г. Алчевск);

Мельничук Дина Александровна — кандидат экономических наук, доцент кафедры высшей математики ГОУ ВО ЛНР «ДонГТИ» (г. Алчевск).

Рецензенты:

С. В. Капранов — доктор медицинских наук, и. о. гл. врача ГС «Алчевская городская МЗ ЛНР» (г. Алчевск);

Н. И. Русанова — кандидат физико-математических наук, заведующая кафедрой «Радиофизика» ГОУ ВО ЛНР «ДонГТИ» (г. Алчевск);

В. С. Федорова — кандидат фармацевтических наук, доцент, заведующая кафедрой экологии и безопасности жизнедеятельности ГОУ ВО ЛНР «ДонГТИ» (г. Алчевск).

Рекомендовано Ученым советом ГОУ ВО ЛНР «ДонГТИ»
(Протокол № 8 от 26.03.2021)

О 64 **Организация** и математическое планирование эксперимента : учебное пособие / С. И. Кулакова, Л. Е. Подлипенская, Д. А. Мельничук. — Алчевск : ГОУ ВО ЛНР «ДонГТИ», 2021. — 121 с.

В учебном пособии представлены материалы, которые полностью соответствуют рабочей программе учебной дисциплины «Организация и математическое планирование эксперимента» ГОУ ВО ЛНР «ДонГТИ». Рассмотрены вопросы организации, математического планирования и обработки результатов эксперимента, ориентированных на решение экологических проблем металлургического производства. Материал содержит наглядные примеры использования алгоритмов планирования эксперимента и обработки его результатов.

Предназначено для подготовки магистров по профилю «Экология металлургического производства», выполнения учебных научно-исследовательских и выпускных квалификационных работ.

УДК 519.242 (075.8)
ББК В172я73

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Предприятия металлургической отрасли несут высокую техногенную нагрузку на окружающую среду. Пылегазовые выбросы приводят к загрязнениям атмосферы, почвы, уничтожению растительности и образованию техногенных пустошей вокруг крупных заводов. Сточные воды металлургического производства содержат большое количество тяжелых металлов, которые имеют способность накапливаться в донных отложениях и аккумулироваться в трофических цепях. В результате деградируют экосистемы многих примыкающих к комбинатам природных объектов.

В связи с этим для обеспечения соблюдения действующего природоохранного законодательства на всех этапах производства требуется регулярно осуществлять контроль качества окружающей среды. В области охраны окружающей среды функционирует система экологического мониторинга, которая представляет собой комплексную систему наблюдений за состоянием окружающей среды, включающую также прогнозирование ее возможных изменений под воздействием антропогенных и природных факторов. Накапливаемая информационная база экологического мониторинга предусматривает непрерывную обработку и получение достоверных оценок необходимой точности и надёжности. Применяемые расчетные методики требуют от исполнителей высоких профессиональных компетенций, знаний и умений, в первую очередь в области математической статистики.

В данном учебном пособии изложены этапы организации и математического планирования эксперимента в природоохранных исследованиях, приведены примеры составления математических моделей в задачах экологической направленности, описаны методики первичной обработки статистических данных и особенности выполнения корреляционного, дисперсионного и регрессионного анализа. Большое внимание в пособии уделяется методам и практическим примерам построения статистических моделей путем активного эксперимента.

Целью дисциплины «Организация и математическое планирование эксперимента» является подготовка будущего специалиста к научно-технической и организационно-методической деятельности, связан-

ной с проведением экспериментальных исследований в области защиты окружающей среды от техногенного воздействия производственной деятельности металлургического производства: выбор и составление плана эксперимента; организация эксперимента и проведение измерений отклика объекта исследований; анализ результатов исследований, включая построение математических моделей объекта исследований, определение оптимальных условий, поиск экстремума функции (поверхности) отклика.

Задачи дисциплины: изучение математических методов, применяемых при планировании и оптимизации эксперимента, освоение современных методологических подходов к постановке и обработке результатов экспериментальных исследований, формирование практических навыков выполнения научных экспериментальных исследований и обработки их результатов.

1 ОБЩИЕ ПРИНЦИПЫ МОДЕЛИРОВАНИЯ В ЗАДАЧАХ ЭКОЛОГИИ

1.1 Математические модели в экологии

Математическое моделирование — важная часть современных исследований в экологии. Окружающая среда, которая представлена как природными, так и техногенными объектами, является сложным комплексом материальных объектов, процессов и явлений, их взаимосвязей и взаимовлияний. Антропогенная деятельность стала существенным фактором, ухудшающим состояние окружающей природной среды, поэтому в настоящее время любой вид человеческой деятельности, в особенности в таких техногенно опасных отраслях, как металлургия, теплоэнергетика, горнодобывающая и химическая отрасли, нуждается в экологической оценке степени негативного воздействия. Во многих случаях эти воздействия нельзя определить прямыми измерениями. Тогда прибегают к моделированию.

Особенное значение моделирования для экологии объясняется следующими обстоятельствами:

– природные объекты часто отличаются большими размерами, и это затрудняет их изучение. Моделирование позволяет наблюдать за их уменьшенными копиями;

– некоторые природные процессы протекают медленно. Моделирование оказывается необходимым для решения задач палеореконструкций и практически единственным методом решения прогнозных задач;

– природные и природно-техногенные объекты являются чрезвычайно сложными системами. Единственным методом, позволяющим учесть все действительно важные стороны такого объекта, является математическое моделирование;

– в геоэкологических исследованиях лабораторное воспроизведение процессов обычно невозможно, специалист вынужден делать заключение о них по неполным результатам. В этом случае моделирование становится важным инструментом анализа;

– в экологических науках самым важным направлением является поиск новых средств, способов и методов защиты окружающей среды от негативных изменений как природного, так и техногенного характера. Для того чтобы экологические мероприятия могли принести максимальный эффект, создаются модели, на которых отрабатываются разные варианты и определяются оптимальные природоохранные стратегии.

Модель — это упрощенная система, описывающая существенные для изучаемого объекта свойства. Различают физические и математические модели.

В математической модели реальный объект заменяется абстрактным, который описывается с помощью соответствующего математического аппарата. Математическое описание должно отражать целостность, структуру, динамику экологического объекта, его функционирование и взаимосвязи внешних и внутренних факторов воздействия. В математической модели учитываются не все свойства реального объекта, а наиболее существенные, связанные с решаемыми с ее помощью задачами. То есть подобие объекта-оригинала и его математической модели должно быть не полное, а частичное — в той мере, в которой это необходимо, чтобы модель адекватно представляла объект для поставленных целей исследования.

Наиболее известные направления моделирования в экологии и природопользовании связаны с геоэкологическими исследованиями следующих направлений:

– моделирование загрязнения атмосферы, гидросферы, литосферы и в целом биосферы в результате антропогенной и производственной деятельности предприятий;

– моделирование водных экосистем (трансформации компонент экосистемы, образования и превращения веществ, потребления, роста и гибели организмов);

– моделирование продукционного процесса развития растений;

– моделирование лесных сообществ (для описания лесных массивов на больших пространственных и временных масштабах и для моделирования популяций).

Особый статус имеют математические модели, в которых рассматриваются глобальные изменения биосферы, вызванные человече-

ской деятельностью или изменением климата в результате космических или геофизических причин. Классической является модель ядерной зимы, предсказавшая глобальное изменение климата на срок в несколько десятилетий в сторону понижения температур и гибель биосферы в случае широкомасштабной ядерной войны. Эта модель и ее последующее обсуждение имели несомненное политическое значение и в значительной мере послужили причиной приостановки гонки ядерных вооружений.

При моделировании глобальных экологических процессов необходимо учитывать огромное число факторов, пространственную неоднородность Земли, физические и химические процессы, антропогенные воздействия, связанные с развитием промышленности и ростом народонаселения. Сложность задачи требует применения системного подхода, впервые введенного в практику математического моделирования Дж. Форрестером (Principles of systems, 1968; World Dynamics, 1971).

Результатом работ, выполненных по заказу Римского клуба — международной группы выдающихся бизнесменов, государственных деятелей и ученых, стала построенная на основе идей Дж. Форрестера компьютерная модель «World 3». В 1972 году результаты этой работы были суммированы в книге D. Meadows «The limits to Growth», которая вызвала сенсацию. В модели Земля была рассмотрена как единая система, в которой происходят процессы, связанные с ростом населения, промышленного капитала, производства продуктов питания, потребления ресурсов и загрязнения окружающей среды. Результаты моделирования взаимодействия этих процессов привели к неутешительному выводу о том, что если существующие тенденции роста численности населения мира, индустриализации, загрязнения окружающей среды, потребления и истощения ресурсов останутся неизменным, пределы роста на нашей планете будут достигнуты в течение ближайших десятилетий.

На рисунке 1.1 представлен прогноз развития глобальной системы в случае сохранения существующих в настоящее время тенденций, выполненный на основании модели системной динамики (Д. Медоуз и др., 1994). Согласно расчетам человечеству грозит сценарий коллапса — падение темпов промышленного производства и производства продуктов питания, неуклонное снижение численности и продолжительности жизни населения.

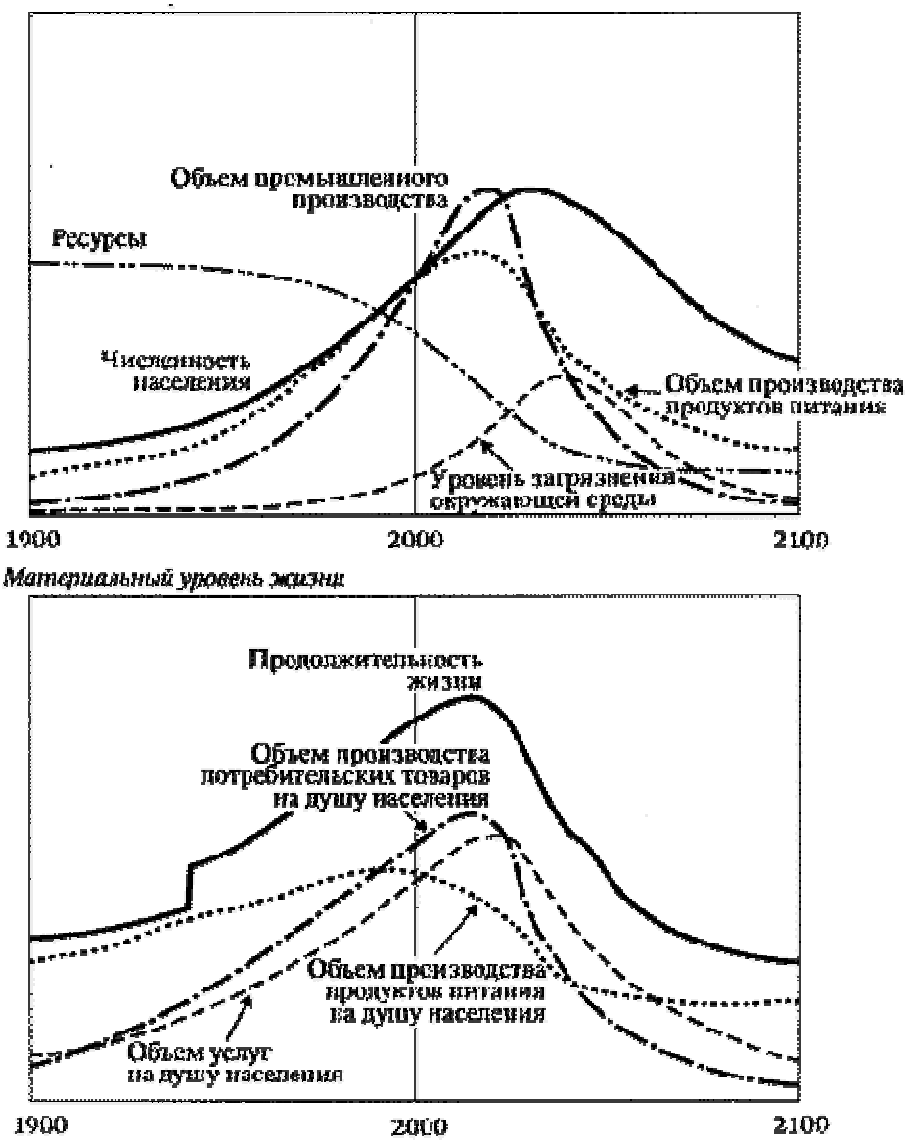


Рисунок 1.1 — Прогнозные модели развития глобальных показателей при сохранении существующих тенденций развития

Смысл таких глобальных моделей заключается в том, что они позволяют оценивать вклад отдельных процессов и регионов в общий баланс вещества и энергии на Земле и решать обратную задачу о влиянии на локальные процессы этих глобальных показателей. Такой всесторонний учет множества факторов и связей возможен только в рамках математических моделей, интегрирующих знания о тысячах взаимосвязей, содержащих сотни и даже тысячи параметров пространственно неоднородной системы, и возможен только с использованием современной вычислительной техники и геоинформационных технологий.

Чтобы избежать подобных сценариев, необходимо принятие программ стабилизации численности населения и объема промышленного производства, внедрения технологий, уменьшающих вредное воздействие человеческой деятельности на биосферу и повышающих эффективность использования природных ресурсов.

1.2 Основные термины и понятия моделирования

Объект — некоторая часть окружающего мира, рассматриваемая в исследовании как единое целое.

Модель представляет собой абстрактное описание объекта (системы, процесса, проблемы, понятия) в некоторой форме, отличной от формы его реального существования, но отражающей существенные свойства моделируемого объекта.

Объект — это оригинал, модель — его образ, прототип.

Моделирование — это процесс создания, анализа и использования модели.

Моделирование в самом общем виде преследует две цели:

– исследование свойств объекта-оригинала на модели его замещающей;

– создание нового объекта с улучшенными характеристиками.

Виды моделирования — физическое и математическое.

При **физическом моделировании** исследуемая модель имеет материальное воплощение, подобное по структуре, пропорциям, материалам, свойствам и пр. подобие с реальным объектом.

Математическое моделирование изучает свойства оригинала на модели, описанной с помощью математического языка.

Классификация моделей

Модели классифицируются по разным признакам:

по цели использования:

– для научного эксперимента;

– для производственного эксперимента;

– для решения оптимизационных задач;

по области применения:

– учебные;

– опытные;

– игровые;

– имитационные.

по фактору времени:

– статические;

– динамические;

по типу функций динамические модели разделяются на:

– дискретные;

– непрерывные;

по наличию случайного воздействия на систему:

– детерминированные (нет случайных изменений);

– стохастические (в модели есть воздействие случайного характера).

Процесс моделирования обычно проходит несколько взаимосвязанных стадий:

1. Рассмотрение структуры объекта-оригинала, его внешних и доступных внутренних свойств. Формулируется цель моделирования и определяются задачи, которые необходимо решить для реализации поставленной цели.

2. Проведение предварительных экспериментов для уточнения задач исследования, определения исходных данных. Обоснование и выбор параметров модели.

3. Формирование требований к математическому описанию модели исходя из предполагаемых задач исследования (поиск оптимальных технологических режимов на модели, разработка алгоритмов управления, построение системы оценки загрязнения окружающей среды в зависимости от влияющих факторов и др.). Выбор методов моделирования.

4. Выбор метода получения экспериментальной информации (в режиме нормальной эксплуатации или по заранее составленному плану, проведение длительных промышленных исследований или использование экспресс-методов) и определение методов обработки информации.

5. Составление математической модели.

6. Исследование и анализ полученной модели, проверка ее адекватности реальному объекту по отношению к выбранному критерию; выяснение необходимости определения последовательности дополни-

тельного уточнения модели; выбор требуемых технических средств для построения систем идентификации.

7. Следующим этапом является проведение расчетов на компьютере или вычислительный эксперимент. Анализ его результатов позволяет уточнить модельные представления авторов и компьютерную моделирующую систему. Часто это означает возврат к одному из первых шести пунктов. После того, как модель начинает вести себя в определенном смысле как реальный объект, ее можно использовать для различных имитаций и прогнозов.

Каждому из указанных этапов соответствует некоторый набор используемых методов, который определяется в процессе исследования. Для большинства реальных технологических процессов получение модели возможно только в режиме нормального функционирования.

Обычно основными взаимосвязанными задачами математического моделирования являются:

– уточнение представлений авторов об исследуемом объекте;

– прогноз поведения объекта при изменении его свойств или внешних воздействий.

Но иногда задача моделирования формулируется иначе: как при имеющихся исходных ресурсах и некоторых ограничениях получить нужный результат. Например, для предотвращения загрязнения поверхностных вод водоемов и водотоков сбросами металлургического производства необходимо разработать способы очистки вод и подобрать рациональное соотношение используемых при очистке реагентов. Решением задач выбора наилучших вариантов из множества имеющихся природоохранных способов и технологий занимается теория оптимального планирования. Одной из ее ветвей, очень важной для экологов, является планирование эксперимента. Даже лабораторные, а тем более натурные эксперименты требуют значительных затрат ресурсов различного рода. Поэтому вопрос об оптимальном числе опытов и условиях их проведения является чрезвычайно актуальным. Этими вопросами занимается такой раздел математики, как математическое планирование эксперимента.

Кратко охарактеризуем основные виды моделирования.

Детерминированное моделирование. Основное содержание — построение математических моделей природных объектов на основе генетических представлений об объекте с использованием известных математических зависимостей.

Вероятностное моделирование. Модель также строится на основе генетических представлений, но ситуация, в которой процесс проходит в природе, воспроизводится с помощью вероятностного, или стохастического подхода.

Статистическое моделирование. Очень часто природные генетические схемы недостаточно ясны для того, чтобы построить удовлетворительную модель. В этом случае используются статистические модели-отклики, известные также как модели черного ящика. Это функция или набор функций, описывающих какие-либо свойства природного объекта, известные из наблюдений.

Оптимизационное моделирование позволяет создать такие модели, которые обладают наилучшими характеристиками (с точки зрения исследователя) среди возможных. К такому моделированию прибегают, как правило, после создания математической модели процесса с помощью детерминированного, вероятностного или статистического моделирования. Оптимизационные методы позволяют создать модели, обладающие наилучшими характеристиками среди возможных. Например, эволюционно-оптимизационная модель определяет оптимальную жизненную стратегию организма, т.е. таких функций, которые обеспечивают максимальное ожидаемое значение приспособленности. На основе такой модели находят эволюционно оптимальные характеристики жизненного цикла, например, возраст репродуктивной зрелости, ожидаемую продолжительность жизни, массу тела и др., а также зависимость этих параметров от внешних условий, например, таких как обеспеченность пищей и агрессивность среды.

Имитационное моделирование используют в том случае, когда для исследуемого экологического объекта по различным причинам не разработана аналитическая модель, либо не найдены методы решения полученной модели. Имитационная модель — логико-математическое описание объекта, которое используется для экспериментирования на компьютере в целях проектирования, анализа и оценки функционирова-

ния объекта. К имитационному моделированию обращаются, когда экономически невыгодно или невозможно экспериментировать на реальном объекте.

Привлечение компьютеров существенно раздвинуло границы моделирования экологических процессов. С одной стороны, появилась возможность создания и исследования сложных математических моделей, не допускающих аналитического описания, с другой — возникли принципиально новые направления (например, имитационное моделирование). Применение математических методов в профессиональной деятельности уже не зависит от трудоёмкости решения и знаний последовательности вычисления конкретного параметра. Успешность исследований определяется в первую очередь корректностью применения математических методов, знаниями общих закономерностей изучаемых явлений и пониманием направленности влияния основных факторов на эти явления. При обосновании рациональных технических решений по нейтрализации негативного течения какого-либо процесса применение современного программного обеспечения играет важную роль, позволяет выявить ситуации, когда конкретные факторы имеют наибольшее влияние.

Таким образом, моделирование становится одним из универсальных методов познания, применяемых во всех современных науках, в том числе в экологии.

1.3 Эксперимент и его организация

Несмотря на высокую эффективность теоретических методов, при рассмотрении конкретных технологических проблем, особенно в условиях действующего производства, инженеру зачастую приходится сталкиваться с задачами, решение которых практически невозможно без организации и проведения экспериментального исследования.

В технической литературе термин эксперимент рассматривается как система операций, воздействий и (или) наблюдений, направленных на получение информации об объекте исследования.

Эксперимент предполагает наличие объекта исследования и цели. В экологии металлургического производства объектами исследования являются все те технологические процессы, явления, физические объек-

ты, которые могут оказать негативное воздействие на человека и компоненты окружающей среды.

Вредное воздействие металлургических предприятий связано с использованием на предприятии устаревших технологий, технических средств защиты от выбросов и сбросов загрязняющих веществ, а также с нерациональными в плане защиты окружающей среды архитектурно-планировочными решениями на территориях, прилегающих к промышленным зонам предприятий.

Все известные технологические процессы производства чугуна, стали, изготовления прокатной продукции сопровождаются образованием большого количества отходов в виде вредных газов и пыли, шлаков, шламов, сточных вод, содержащих различные химические компоненты, скрапа, окалины, боя огнеупоров, мусора и т.д., которые загрязняют атмосферу, воду и поверхность земли.

Экологические службы предприятия, государственных природоохранных ведомств обязаны контролировать ситуацию с загрязнением конкретными предприятиями компонент окружающей среды, для чего регулярно осуществляется экологический мониторинг, в задачи которого входит наблюдение, анализ и прогноз негативных изменений состояния окружающей среды. Большой сегмент в сфере техносферной безопасности занимают научные организации, занимающиеся разработкой перспективных ресурсосберегающих и экологически эффективных способов, средств и технологий.

Для успешного осуществления поставленных задач необходимо обнаружить и зафиксировать экологические нарушения, оценить степень нанесенного ущерба, спрогнозировать ситуацию на перспективу и разработать, в случае необходимости, систему природоохранных мероприятий. Во всех этих случаях необходимо проведение экспериментальных исследований и построение на их основе моделей, как правило, математических.

Первичное звено — экспериментальное, имеет чрезвычайно большое значение, как фундамент в основе большого здания. Использование точных математических моделей, современной вычислительной техники не поможет исправить ошибки, допущенные на этапе эксперимента. Поэтому организации эксперимента необходимо уделять должное внимание.

Охарактеризуем основные термины и понятия.

1. Эксперимент — это метод научного познания, при помощи которого исследуются явления реально-предметной действительности в определённых (заданных), воспроизводимых условиях путём их контролируемого изменения.

2. Опыт — это эксперимент в узком смысле слова, понимается как воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов.

В экспериментальных исследованиях используют различные виды проведения и формы представления результатов.

3. Качественный эксперимент дает результаты в виде качественной информации, зачастую только фиксирует наличие того или иного явления. Как правило, это первичный этап экспериментальных исследований, не требует значительных материальных и временных затрат.

Например, для пробы воды из водоема, в который осуществлялся сброс промышленных сточных вод металлургического производства, необходимо выполнить исследование на присутствие в ней железа. Использовался эксперимент, в котором по качественной реакции на вносимый реагент устанавливался факт наличия в воде железа (II и III).

Рассмотрим реакции двухвалентного и трехвалентного железа.

Качественные реакции на присутствие в растворе железа (II) — ионы Fe^{2} вступают в реакцию с красной кровяной солью $K_3[Fe(CN)_6]$ — гексацианоферратом калия. Если при добавлении соли образуется синеватый осадок, значит эти ионы присутствуют в растворе (рис. 1.2).

Рисунок 1.2 — Реакция пробы сточной воды при добавлении красной кровяной соли

Качественные реакции на присутствие железа (III) определяются при внесении в раствор щелочи. При наличии в растворе ионов трехвалентного железа образуется основание — гидроксид железа (III) $Fe(OH)_3$ (рис. 1.3). Бурый осадок указывает на присутствие в исходном растворе ионов железа (III).

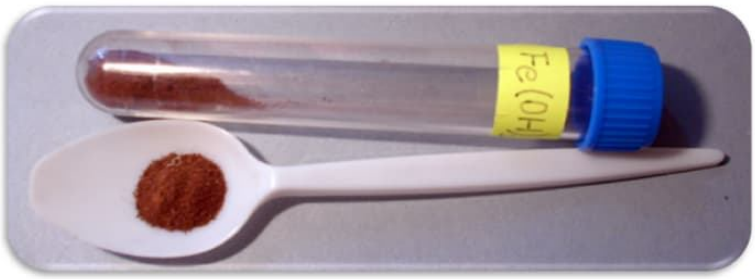


Рисунок 1.3 — Гидроксид железа (II) $Fe(OH)_3$

4. Количественный эксперимент не только фиксирует наличие того или иного явления, а позволяет оценить его количественной характеристикой. Например, для определения распространения в селитебной зоне окислов азота, которые образуются в доменных, мартеновских и нагревательных печах, в коксохимическом производстве, берут пробы воздуха и исследуют их в лаборатории фотометрическим методом с использованием сульфаниловой кислоты и і-нафтиламина. В результате получают значения концентрации этих веществ, выраженных количественно (ед. измерения — мг/м^3). Результаты таких исследований гораздо шире качественных измерений, поскольку они позволяют оценить степень опасности нахождения данных веществ в атмосфере для населения путем сравнения с установленными предельно допустимыми значениями.

5. Лабораторный эксперимент проводится в условиях специализированных лабораторий, позволяет уменьшить влияние случайных факторов на результаты исследований, дает возможность изучить объект эксперимента при различных комбинациях влияющих факторов. Как правило, воспроизводимость опытов в лабораторных условиях выше и достигается меньшим количеством опытов.

Так, в 2013–2014 годах аспиранткой кафедры экологии и безопасности жизнедеятельности ДонГТУ А. А. Карпо в лабораторных условиях была создана экспериментальная установка для изучения процессов

выноса угольной пыли при различных скоростях и направлениях воздушного потока. Объектом научных исследований являлись открытые склады угля коксохимического производства, которые были источниками значительного загрязнения атмосферного воздуха взвешенными веществами. Была создана физическая модель в виде соответствующего штабеля угля с уменьшенными размерами, изолированного от остального помещения лаборатории конструкцией, покрытой полиэтиленовой пленкой. Поток воздуха, действующий на модель штабеля, создавался воздуходувкой. Фиксировались параметры распространения и концентрация угольной пыли в зависимости от моделируемых условий. Эксперимент позволил обнаружить некоторые закономерности в интенсивности эмиссии угольной пыли от штабеля угля при различных ветровых условиях.

6. Промышленный эксперимент проводится в условиях реального производства, где возможности широкого экспериментирования значительно ниже, чем в лабораторном исследовании. В таких случаях особенно важно использовать статистический подход на этапе планирования эксперимента.

7. Пассивный эксперимент — исследование, при котором значения факторов в каждом опыте регистрируются исследователем, но не задаются.

8. Активный эксперимент — это исследование, в котором экспериментатор заранее определяет значения варьируемых факторов, для чего и использует планирование эксперимента.

9. План эксперимента — совокупность данных, определяющих количественные и качественные условия проведения опытов.

10. Планирование эксперимента — выбор плана эксперимента, удовлетворяющего заданным требованиям, совокупность действий, направленных на разработку стратегии экспериментирования (от получения априорной информации до получения работоспособной математической модели или определения оптимальных условий). Это целенаправленное управление экспериментом, реализуемое в условиях неполного знания механизма изучаемого явления.

Организация и планирование эксперимента являются важным этапом исследования, поскольку многие эксперименты требуют при

своем проведении больших ресурсов, как материальных, так и человеческих. Грамотно составленный план эксперимента и правильная организация его проведения обеспечивают достоверность результатов эксперимента и в последующем позволяют получить адекватные математические модели.

Исходя из вышесказанного, необходимость планирования эксперимента представляется очевидной. При этом обычно преследуется одна из двух целей:

– получение максимального количества информации при заданных ограничениях на затраты (включая затраты времени);

– минимизация затрат при получении необходимого количества информации.

Широкое применение экспериментальных методов, математической статистики привело к созданию теории эксперимента. Эта теория призвана дать ответы на следующие вопросы:

– как нужно организовать эксперимент, чтобы наилучшим образом решить поставленную задачу (в смысле затрат времени и средств или точности результатов);

– как следует обрабатывать результаты эксперимента, чтобы получить максимальное количество информации об исследуемом объекте (явлении);

– какие выводы можно сделать об объекте-оригинале на основании результатов эксперимента.

Контрольные вопросы

1. Определение математического моделирования.
2. Моделирование как метод познания.
3. Перечислите стадии процесса моделирования.
4. Назовите стадии процесса моделирования.
5. Понятие модели. Свойства моделей.
6. Классификация моделей.
7. Перечислите виды моделирования.
8. Сущность детерминированного моделирования.
9. Основной принцип математического моделирования.

10. Особенности вероятностного и статистического моделирования.

11. Основная идея оптимизационного моделирования.

12. В чем сущность имитационного моделирования?

13. Понятие математической модели.

14. Задачи математического моделирования.

15. Классификация математических моделей.

16. В чем состоит различие эксперимента и опыта?

17. Перечислите виды экспериментов.

18. Приведите примеры количественного и качественного эксперимента в экологии.

19. Перечислите формы проведения эксперимента.

20. Укажите различие лабораторного и промышленного эксперимента.

21. Укажите различия активного и пассивного эксперимента.

22. Перечислите методы планирования эксперимента.

23. Сущность планирования эксперимента.

24. Какие вопросы решаются в теории эксперимента?

25. Понятие плана эксперимента.

2 ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ПУТЕМ ПАССИВНОГО ЭКСПЕРИМЕНТА

2.1 Основные понятия пассивного эксперимента

Фактор — переменная величина, предположительно влияющая на результаты эксперимента. В отдельном опыте каждый фактор может принимать одно из возможных своих значений — уровень фактора.

Уровень фактора — фиксированное значение фактора относительно начала отсчета. Фиксированный набор уровней всех факторов в каждом опыте определяет одно из возможных состояний объекта исследований.

В зависимости от того, в какой степени исследователь вмешивается в процесс, различают такие виды факторов:

– контролируемые и управляемые — возможна регистрация уровня фактора и задание в каждом конкретном опыте любого возможного значения фактора;

– контролируемые и неуправляемые — возможна только регистрация уровня фактора в каждом конкретном опыте;

– неконтролируемые — это факторы, уровни которых не регистрируются экспериментатором.

Отклик — это свойство исследуемого явления, которое наблюдает исследователь, по предположению зависит от факторов.

Количественный эксперимент проводится с целью найти зависимость между откликом и факторами — функцию отклика.

По способу проведения исследования количественный эксперимент может быть пассивным или активным.

Пассивный эксперимент состоит в сборе и обработке данных, полученных в результате пассивного наблюдения за технологическим процессом в производственных условиях. Для анализа и обработки данных применяют корреляционный и регрессионный анализы, временные ряды.

С помощью корреляционного и регрессионного анализа исследуемого процесса в условиях производства изучается степень взаимосвязи факторов и возможность построения прогнозной модели. Временной ряд представляет собой совокупность измерений технологического показателя в течение некоторого периода времени. Основной чертой этого

анализа является существенность порядка, в котором производятся наблюдения. Природа ряда и структура порождающего его процесса определяют порядок образования последовательности.

Преимущество пассивного эксперимента состоит в том, что при его применении нет необходимости тратить время и средства на постановку опытов. Математические модели, полученные при пассивном эксперименте, можно использовать для управления процессом. При этом пассивный эксперимент имеет существенные недостатки, которые ограничивают его применение для оптимизации технологических процессов.

Эти недостатки вызваны тем, что в условиях производства нет возможности произвольно варьировать влияющими факторами, а изменение выходной исследуемой величины (функции отклика) в большей мере обусловлено воздействием неконтролируемых, случайных возмущений. Также при пассивном эксперименте часто нет возможности рассматривать все факторы, оказывающие существенное влияние на исследуемый процесс.

Если на объекте исследования по техническим, технологическим или экономическим причинам невозможно плановое варьирование входных факторов в необходимом диапазоне, то для накопления статистического материала применяют пассивный эксперимент, заключающийся в наблюдении и регистрации значений входных и выходных факторов в режиме нормального функционирования объекта. Необходимые условия проведения пассивного эксперимента: обоснованное время регистрации данных, обеспечение независимости соседних измерений и входных переменных друг от друга, достаточный объем экспериментальных данных.

Для осуществления пассивного эксперимента требуется знание математической статистики.

2.2 Первичная обработка результатов эксперимента

Основные понятия математической статистики

Генеральная совокупность — множество объектов, подлежащих изучению.

Выборка — это совокупность объектов, отобранных из генеральной совокупности с целью изучения определенного признака.

Объем выборки — количество наблюдений в выборке. Обозначается n.

Признак — это свойство объекта, которое необходимо изучить. Обозначается X, Y, Z.

Статистическое наблюдение — это сбор сведений, заключающийся в регистрации (учете) признаков и фактов, которые характеризуют каждую единицу исследуемой (изучаемой) совокупности.

Варианта — это значение признака, измеренное у определенного объекта выборки. Применяемые обозначения для вариант признака X: $x_1, x_2, \ldots, x_n$, где n — объем выборки.

Ряд распределения — ряд упорядоченных значений признака (вариант).

Задача математической статистики — найти обобщающие характеристики выборки, которые отражают свойства генеральной совокупности.

Основные статистические характеристики

Выборочное среднее:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \,. \tag{2.1}$$

Выборочное среднее характеризует среднее значение признака X по выборке.

Выборочная дисперсия:

$$D_в = \overline{x^2} - (\overline{x})^2 \text{ , где } \overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^{\,2} \,. \tag{2.2}$$

Выборочная дисперсия характеризует средний квадрат отклонения значений признака X от среднего значения по выборке.

Выборочное среднее квадратическое отклонение (СКО):

$$\sigma_в = \sqrt{D_в} \,. \tag{2.3}$$

Выборочное СКО характеризует среднее отклонение значений признака X от среднего значения по выборке, часто называется стандартным отклонением.

Поскольку $D_в$ является смещенной оценкой дисперсии (а именно заниженной), то обычно для малых выборок используют **исправленную выборочную дисперсию:**

$$S^2 = \frac{n}{n-1} \cdot D_в, \tag{2.4}$$

которая, как и $D_в$, характеризует средний квадрат отклонения значений признака X от среднего значения по выборке.

Исправленное выборочное среднее квадратическое отклонение:

$$S = \sqrt{S^2}\ , \tag{2.5}$$

S — характеризует то же, что и $\sigma_в$.

Коэффициент вариации:

$$V = \frac{\sigma_в}{\overline{x}} \cdot 100\%. \tag{2.6}$$

Коэффициент вариации характеризует степень вариации признака. При нормальном распределении коэффициент вариации обычно не превышает 45–50% и часто бывает гораздо ниже этого уровня.

Размах выборки:

$$R = x_{\max} - x_{\min}\ . \tag{2.7}$$

Размах выборки характеризует меру варьирования признака.

Выборочные характеристики, как правило, не совпадают по абсолютной величине с соответствующими генеральными параметрами. Величину отклонения выборочного показателя от его генерального параметра называют статистической ошибкой или ошибкой репрезентативности. В случае, когда распределение исходного признака не слишком отличается от нормального вида и объем выборки не слишком мал (на практике $n \geq 30$), **стандартная ошибка среднего** (ошибка репрезентативности среднего значения) находится по формуле:

$$S_{\overline{x}} = \frac{S}{\sqrt{n}}\ . \tag{2.8}$$

Для сравнения точности оценок, найденных по различным выборкам, используют относительный показатель точности оценки, который определяется по формуле:

$$C_s = \frac{S}{\overline{x}_в \cdot \sqrt{n}} \cdot 100\%\ . \tag{2.9}$$

Точность средних показателей, которые оценивают результаты наблюдений, считают вполне удовлетворительной, если коэффициент Cs не превышает 3–5 %.

Во многих случаях в качестве обобщающих характеристик выборки используются структурные средние — медиану, моду, квантили — конкретные значения, которые занимают особое место в упорядоченном ряду выборочных значений.

Медиана *Me* — средняя, относительно которой ряд распределения делится на две равные части: в обе стороны от медианы располагаются одинаковое число вариант признака. Для нахождения медианы в упорядоченном ряде значений выбирают центральную варианту. При четном числе членов ряда медиана определяется по полусумме двух соседних вариант, расположенных в центре упорядоченного ряда.

Мода *Mo* — это наиболее часто встречающееся значение признака. Также определяется по упорядоченному ряду.

Квантили отсекают в пределах ряда определенную часть вариант. К ним относят квартили, децили и перцентили (процентили). Квартили — это три значения признака ($x_{0,25}$, $x_{0,5}$, $x_{0,75}$), делящие упорядоченный ряд значений признака на четыре равные части. Между квартилями $x_{0,25}$ и $x_{0,75}$ находится 50 % всех членов ряда, а квартиль $x_{0,5}$ равен медиане. Аналогично, девять децилей делят ряд на десять равных частей, а 99 перцентилей — на 100 равных частей.

Проверка статистических данных на аномальность

В экологии при статистической обработке информации часто сталкиваются с наличием в исследуемой совокупности некоторого числа наблюдений, значения которых резко отличаются от основной массы наблюдений. Появление таких наблюдений может объясняться как естественной вариабельностью признака, так и неоднородностью статистической совокупности. Причин неоднородности может быть несколько, причем они могут быть связаны как с неоднородностью самой среды, так и с нарушением стандартных условий эксперимента (непредвиденных отступлений от методик исследований, неправильной работы приборов, ошибок при взятии точек отсчетов и т.д.) Поэтому прежде чем приступать к дальнейшей статистической обработке, следует проверить данные

на наличие аномальных наблюдений. Наиболее мощным является критерий, основанный на использовании нормированного отклонения:

$$\tau = \frac{\left| x - \bar{x} \right|}{S}. \qquad (2.10)$$

Проверка заключается в сравнении наблюдаемого значения критерия τ с табличным $\tau_{кр}$. Если $\tau > \tau_{кр}$ при выбранном уровне значимости, то это означает, что проверяемое значение аномально и должно быть отброшено.

Нормальный закон распределения

При статистическом анализе часто требуется определить вероятность, с которой значения признака в генеральной совокупности находятся в заданном интервале. Ответ на этот вопрос дает закон распределения, который характеризует закономерность появления значений признака.

Наиболее широко в природе распространен нормальный закон. Нормальное распределение означает, что в массе значений признака, большинство вариант оказывается близкими к среднему значению, и чем дальше варианты отстоят от среднего, тем реже встречаются. Это один из фундаментальных законов природы. Графическое представление этого закона называется кривой Гаусса, которая имеет «колоколообразный» вид и задается только двумя параметрами: средним значением признака и средним квадратическим отклонением. Для нормального закона характерно, что практически все варианты признака (99,7 %) находятся в пределах ± 3 стандартных отклонений от среднего. Это простое практическое правило проверки нормального закона называется правилом трех сигм.

Для нормального закона также характерно совпадение среднего значения, моды и медианы, т.е. теоретически нормальное распределение строго симметрично. В тех случаях, когда какие-либо причины благоприятствуют более частому появлению значений признака, которые больше или, наоборот, меньше среднего, образуется асимметричное распределение. Об отклонении от нормального распределения свидетельствуют коэффициенты асимметрии A_s и эксцесса E_k. Положительное значение коэффициента асимметрии свидетельствует о наличии

распределения с «длинным правым хвостом». Отрицательное значение асимметрии характеризует распределение с «длинным левым хвостом». Коэффициент эксцесса E_k показывает «остроту пика» распределения. В тех случаях, когда какие-либо причины способствуют преимущественному появлению средних или близких к средним значениям, образуется распределение с положительным эксцессом. Если в распределении преобладают крайние значения, то такое распределение характеризуется отрицательным эксцессом. Иногда в центре распределения может образоваться впадина, превращающая его в двухвершинное, что может свидетельствовать о неправомерном объединении двух разных выборок в одну.

Существует простое правило, используя которое, можно оценить, насколько распределение соответствует нормальному. Для нормального закона должны выполняться соотношения:

$$\left|A_S\right| \le 3 \cdot \sqrt{D(A)}, \quad \left|E_k\right| \le 3 \cdot \sqrt{D(E)}, \tag{2.11}$$

где $D(A)$ и $D(E)$ — дисперсии асимметрии и эксцесса, находятся по формуле:

$$D(A) = \frac{6(n-1)}{(n+1)(n+3)}, \quad D(E) = \frac{24(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \tag{2.12}$$

Знание закона распределения важно, поскольку в соответствии с типом распределения применяются разные принципы статистической обработки: параметрический и непараметрический. Параметрический принцип включает все методы анализа нормально распределенных количественных признаков. Непараметрический принцип используется во всех остальных случаях – для анализа количественных признаков независимо от вида их распределения и для анализа качественных признаков.

Статистические гипотезы

Статистическая гипотеза — это предположение о свойствах признака, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки. Основной принцип проверки: маловероятные события считаются невозможными, а события, имеющие большую вероятность — достоверными. Этот принцип реализуется следующим образом.

Этап 1: формулируется основная и альтернативная гипотезы. Основная гипотеза H_0 — это гипотеза об отсутствии различий (поэтому 0); альтернативная гипотеза H_1— гипотеза, которую принимают, когда основную отвергают (различия значимы).

Этап 2: для проверки гипотезы выбирают критерий. Статистический критерий — это решающее правило, обеспечивающее принятие истинной или отклонение ложной гипотезы с высокой вероятностью. Это специальная случайная величина с известным распределением, составлена и обоснована исследователем, например, Фишером, Стьюдентом и т. д. Согласно выбранному критерию по выборке рассчитывается наблюдаемое значение критерия.

Этап 3: находят критические и допустимые значения критерия. Допустимая область критерия — это область, где верна гипотеза H_0, критическая область — это область, где гипотеза H_0 не верна, а верна альтернативная гипотеза H_1. Границы областей — это критические значения критерия, они определяются по специальным таблицам в зависимости от выбранного критерия.

Выбор критических значений зависит от:

– уровня значимости α (0,05; 0,01; 0,001) — вероятности отвергнуть основную гипотезу H_0, в то время как на самом деле она верна (называется ошибкой первого рода);

– числа степеней свободы k, которое зависит от объема выборки и условий формирования самого критерия.

Этап 4: определяют, в какую область попадает наблюдаемое значение критерия (сравнивают наблюдаемое и критическое значения):

– если наблюдаемое значение принадлежит допустимой области, то нет оснований отвергнуть гипотезу H_0;

– если наблюдаемое значение принадлежит критической области, то H_0 отвергают и принимают гипотезу H_1.

Важно: гипотеза может быть отвергнута, но никогда не может быть окончательно принята. В итоге могут быть совершены ошибки:

– ошибка первого рода α — отвергнута основная гипотеза H_0, в то время как она верна (отклонить верную гипотезу), где α — вероятность того, что сочли различия существенными, когда на самом деле

они случайны. Например, $\alpha = 0,05$ означает, что в 5 случаях из 100 отвергают верную гипотезу.

– ошибка второго рода β — принята гипотеза H_0, в то время как она неверна. Это более опасная ошибка — принять ошибочную гипотезу.

В математической статистике изучено множество различных гипотез, каждая из которых проверяется своим способом. Рассмотрим некоторые из них, наиболее часто встречаемые в экологических статистических расчетах.

А. Сравнение двух дисперсий нормально распределенных генеральных совокупностей

Исходные параметры выборок:

по признаку X: n_1 — объем выборки, S_x^2 — исправленная выборочная дисперсия;

по признаку У: n_2 — объем выборки, S_y^2 — исправленная выборочная дисперсия.

Пусть для определенности $S_x^2 > S_y^2$. Требуется при заданном уровне значимости α сравнить дисперсии $D(X)$ и $D(Y)$ генеральных совокупностей.

Выдвинем основную и альтернативную гипотезы. Рассмотрим два случая:

а) $H_0 : D(X) = D(Y)$, $H_1 : D(X) > D(Y)$ б) $H_0 : D(X) = D(Y)$, $H_1 : D(X) \neq D(Y)$

Для проверки гипотез по результатам выборок вычисляем наблюдаемое значение критерия (отношение большей дисперсии к меньшей):

$$F_{набл} = \frac{S_x^2}{S_y^2} \qquad (2.13)$$

Этот критерий является случайной величиной, которая подчиняется закону распределения Фишера-Снедекора.

Критические области и точки зависят от выдвинутых альтернативных гипотез H_1.

Случай а) $H_1 : D(X) > D(Y)$.

Критическая область является правосторонней. Критическая точка находится по таблице критических точек распределения Фишера-Снедекора: $F_{кр} = F(\alpha; k_1; k_2)$, где α — заданный уровень значимости; $k_1 = n_1 - 1$ — число степеней свободы большей дисперсии ($S_x{}^2$); $k_2 = n_2 - 1$ — число степеней свободы меньшей дисперсии ($S_y{}^2$).

Если в результате сравнения окажется $F_{набл} < F_{кр}$, то нет оснований отвергнуть нулевую гипотезу H_0; если же $F_{набл} > F_{кр}$, то нулевая гипотеза H_0 отвергается; принимается гипотеза H_1.

Случай б) $H_1 : D(X) \neq D(Y)$.

Критическая область является двусторонней. Критическая точка находится по таблице критических точек распределения Фишера-Снедекора: $F_{кр} = F\left(\frac{\alpha}{2}; k_1; k_2\right)$, где α — заданный уровень значимости; $k_1 = n_1 - 1$ — число степеней свободы большей дисперсии ($S_x{}^2$); $k_2 = n_2 - 1$ — число степеней свободы меньшей дисперсии ($S_y{}^2$).

Если в результате сравнения окажется $F_{набл} < F_{кр}$, то нет оснований отвергнуть нулевую гипотезу H_0; если же $F_{набл} > F_{кр}$, то нулевая гипотеза H_0 отвергается; принимается гипотеза H_1.

Б. Сравнение двух математических ожиданий нормально распределенных генеральных совокупностей, дисперсии которых неизвестны и одинаковы

Исходные параметры выборок:

по признаку X: n_1 — объем выборки, $\overline{x}_в$ — выборочное среднее, $S_x{}^2$ — исправленная выборочная дисперсия;

по признаку У: n_2 — объем выборки, $\overline{y}_в$ — выборочная средняя; $S_y{}^2$ — исправленная выборочная дисперсия.

Требуется при заданном уровне значимости α сравнить математические ожидания $M(X)$ и $M(Y)$ генеральных совокупностей.

Перед тем, как решать поставленную задачу, нужно убедиться, что дисперсии сравниваемых совокупностей статистически не различимы (см. предыдущий случай А). Далее решение осуществляется следующим образом: выдвигается основная и альтернативная гипотезы. Рассмотрим три случая:

а) $\begin{matrix} H_0 : M(X) = M(Y) \\ H_1 : M(X) > M(Y) \end{matrix}$ б) $\begin{matrix} H_0 : M(X) = M(Y) \\ H_1 : M(X) < M(Y) \end{matrix}$ в) $\begin{matrix} H_0 : M(X) = M(Y) \\ H_1 : M(X) \neq M(Y) \end{matrix}$

Для проверки гипотез по результатам выборок вычисляем наблюдаемое значение критерия:

$$T_{набл} = \frac{\overline{x}_в - \overline{y}_в}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}. \qquad (2.14)$$

Этот критерий является случайной величиной, которая подчиняется закону распределения Стьюдента с $k = n_1 + n_2 - 2$ степенями свободы.

Критические области и точки зависят от выдвинутых альтернативных гипотез H_1.

Случай а) $H_1 : M(X) > M(Y)$.

Критическая область является правосторонней. Критическая точка находится по таблице критических точек распределения Стьюдента $t_{кр} = t\left(\frac{\alpha}{2}; k\right)$, где α — заданный уровень значимости.

Если в результате сравнения окажется $T_{набл} < t_{кр}$, то нет оснований отвергнуть нулевую гипотезу H_0; если же $T_{набл} > t_{кр}$, то нулевая гипотеза H_0 отвергается; принимается гипотеза H_1.

Случай б) $H_1 : M(X) < M(Y)$.

Критическая область является левосторонней. Критическая точка находится по таблице критических точек распределения Стьюдента, только с отрицательным знаком $t_{кр} = -t\left(\frac{\alpha}{2}; k\right)$, где α — заданный уровень значимости.

Если в результате сравнения окажется $T_{набл} > t_{кр}$, то нет оснований отвергнуть нулевую гипотезу H_0; если же $T_{набл} < t_{кр}$, то нулевая гипотеза H_0 отвергается; принимается гипотеза H_1.

Случай в) $H_1 : M(X) \neq M(Y)$.

Критическая область является двусторонней. Критическая точка находится по таблице критических точек распределения Стьюдента $t_{кр} = t(\alpha; k)$, где α — заданный уровень значимости.

Если в результате сравнения окажется $|T_{набл}| < t_{кр}$, то нет оснований отвергнуть нулевую гипотезу H_0; если же $|T_{набл}| > t_{кр}$, то нулевая гипотеза H_0 отвергается, принимается гипотеза H_1.

2.3 Корреляционный и регрессионный анализ

Корреляционный анализ — метод обработки статистических данных, заключающийся в изучении тесноты связи между признаками. Этот анализ позволяет выявить наиболее связанные между собой переменные. Основная цель корреляционного анализа — обеспечить получение некоторой информации об одном признаке с помощью другого признака. Наличие корреляции означает, что изменение одной переменной вызывает изменение другой. Такую зависимость называют статистической.

Нахождением формы зависимостей между признаками занимается регрессионный анализ. Статистическая зависимость может быть описана математической функцией. Общий вид статистической зависимости (статистической модели):

$$y = f(x) + \varepsilon, \tag{2.15}$$

где x — независимая, т.е. объясняющая переменная (фактор);

y — зависимая, объясняемая переменная;

ε — случайная составляющая.

При обозначении переменных (x или y) необходимо выбирать так, чтобы изменение x служило причиной изменения y.

Независимой переменной (x) называется переменная, которая варьируется исследователем, тогда как зависимая переменная (y) — это переменная, которая измеряется или регистрируется в результате изменения x. Иными словами, независимая переменная «независима» от реакций, свойств, намерений и т.д., присущих объектам исследования. Зависимость проявляется в ответной реакции исследуемого объекта на посланное воздействие.

Случайная составляющая (ε) может быть обусловлена множеством неучтённых факторов. Чем больше существенных для переменной y факторов будет добавлено в модель, тем меньше будет доля в сумме случайной составляющей. При выборе влияющих факторов важно понимать «физику» происходящих процессов, что позволит выделить среди множества факторов те, которые оказывают влияние, а среди них, в свою очередь, те, которые оказывают существенное (значимое) влияние.

Одномерная регрессия

Функция, которая описывает статистическую зависимость, называется регрессией. Уравнение регрессии имеет вид: $\overline{y} = f(x)$ — такая регрессия называется парной (одномерной) и характеризует, как объясняющая переменная влияет на зависимую результативную переменную «в среднем».

Уравнение регрессии строится с целью предсказания (прогнозирования) среднего значения $\overline{y}$ при фиксированном значении x.

Виды парной регрессии определяются видом функции $f(x)$:

– линейная $\overline{y} = a_0 + a_1 \cdot x$;

нелинейная:

– квадратичная $\overline{y} = a_0 + a_1 \cdot x + a_2 \cdot x^2$;

– логарифмическая $\overline{y} = a_0 + a_1 \cdot \ln x$;

– экспоненциальная: $\overline{y} = a_0 + a_1 \cdot e^x$ и др.

Определение коэффициентов a_0 и a_1 (параметров уравнения регрессии) осуществляется методом наименьших квадратов (МНК) по формулам:

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \qquad (2.16)$$

$$a_0 = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}. \qquad (2.17)$$

Важно не только получить уравнение регрессии, но и оценить его адекватность (качество и надежность), т. е. соответствие действительности. Для оценки качества модели рассчитывают:

1) Коэффициент детерминации:

$$R^2 = \frac{\sum (y_i^T - \overline{y})^2}{\sum (y_i - \overline{y})^2}, \qquad (2.18)$$

где $y_i^T = f(x_i)$ — расчетное значение y, найденное по уравнению регрессии.

R^2 характеризует долю вариации результативного признака y, которая объясняется вариацией признака x.

2) Коэффициент корреляции:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x^2}, \qquad (2.19)$$

где r характеризует тесноту линейной связи. Пределы изменения коэффициента корреляции $-1 \le r \le 1$. Если $r \approx 0$, то линейной связи между факторами нет. Если $r \approx \pm 1$ (точное значение ± 1 при статистических наблюдениях невозможно), то между факторами есть сильная почти функциональная линейная связь. Положительное значение r означает возрастающую зависимость, отрицательное — убывающую.

Коэффициент парной корреляции необходимо проверить на статистическую значимость по критерию Стьюдента. Выдвигают гипотезы:

основная $H_0 : r_z = 0$ (коэффициент корреляции не значим);

альтернативная $H_1 : r_z \neq 0$ (коэффициент корреляции значим).

Для проверки гипотезы H_0 вычисляется наблюдаемое значение критерия:

$$T_{набл} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (2.20)$$

Критерий $T_{набл}$ является случайной величиной, которая подчиняется закону распределения Стьюдента с $k = n-2$ степенями свободы. Критическая область является двусторонней. По таблице критических точек распределения Стьюдента (в Excel функция СТЬЮДРАСПОБР) определяется критическое значение критерия $t_{кр} = t(\alpha, k)$ при выбранном уровне значимости ошибки $\alpha = 0,05$ или $\alpha = 0,01$ и числе степеней свободы $k = n-2$.

Если $|T_{набл}| > t_{кр}$, то нулевая гипотеза отвергается. Это значит, что коэффициент корреляции статистически значим.

Если $|T_{набл}| < t_{кр}$, то нет оснований отвергнуть нулевую гипотезу.

Это значит, что коэффициент корреляции незначимо отличается от нуля. Делают вывод, что линейной связи между признаками нет.

Если коэффициент корреляции значим, то сила связи оценивается по шкале Чеддока (табл. 2.1).

Таблица 2.1 — Шкала Чеддока

$\|r\|$	<0,1	0,1–0,3	0,3–0,5	0,5–0,7	0,7–0,9	>0,99
Сила связи	несущественная	слабая	умеренная	заметная	высокая	весьма высокая

3) Средняя квадратическая ошибка уравнения (стандартная ошибка оценки) вычисляется по формуле:

$$S_{уравн} = \sqrt{\frac{\sum (y_i - y_i^{т})^2}{n-k}}, \qquad (2.21)$$

где y_i — наблюдаемые значения зависимой переменной;

$y_i^{т} = f(x_i)$ — теоретические значения зависимой переменной, найденные из уравнения регрессии;

n — объем выборки;

k — число параметров (коэффициентов) в уравнении регрессии.

Величина $S_{уравн}$ измеряется в тех же единицах, в которых задана зависимая переменная y и характеризует точность, с которой могут быть оценены значения зависимой переменной по найденному уравнению регрессии.

4) Средняя относительная ошибка аппроксимации (средняя относительная погрешность модели):

$$\varepsilon = \frac{1}{n} \cdot \sum \left| \frac{y_i - y_i^{\text{т}}}{y_i} \right| \cdot 100\%. \qquad (2.22)$$

Величина ε характеризует точность модели (табл. 2.2), т.е. среднее отклонение фактических значений от расчетных измеряется в %.

Таблица 2.2 — Шкала для оценки точности статистической модели

ε	$\varepsilon \le 5\%$	$5\% < \varepsilon \le 10\%$	$10\% < \varepsilon \le 20\%$	$\varepsilon > 20\%$
Точность модели	высокая	хорошая	удовлетворительная	неудовлетворительная

Прогноз по уравнению регрессии

Уравнение регрессии — это математическая модель, описывающая зависимость y от x. Уравнение, адекватно описывающее изучаемую зависимость, должно иметь максимальное R^2, максимальное r, минимальные ошибки ε и $S_{уравн}$. Адекватное уравнение зависимости используется для прогноза среднего значения результирующей (зависимой) переменной y для определенного (заданного либо выбранного самостоятельно) значения x. Для прогноза значение x_* подставляется в уравнение регрессии и рассчитывается значение y, то есть $y_{прогноз} = f(x_*)$.

Важно! Уравнение регрессии можно применять для прогноза в пределах тех значений независимой переменной x, для которых оно было получено.

Множественная регрессия

На практике во многих случаях изменение выходного показателя объясняется влиянием многих факторов, в таком случае для прогноза используются модели множественной регрессии.

Общий вид модели: $\overline{y} = f(x_1, x_2, ...x_k) + \varepsilon$, где $x_1, x_2, ...x_k$ — независимые значения признака, k — число независимых признаков, $\overline{y}$ — зависимый признак (результативный). Функция $f(x_1, x_2, ...x_k)$ может быть как линейной, так и нелинейной формы.

Очень важным является вопрос о том, сколько независимых факторов может быть в уравнении множественной регрессии при заданном объеме выборки n. Обычно используют такое правило: число наблюдений должно быть не менее чем в 8–10 раз больше числа факторов в уравнении регрессии.

Проверка данных на аномальность производится двумя подходами:

Первый подход — по τ-критерию с помощью формулы (2.10).

Второй подход — графически, который позволяет визуально определить наблюдения, резко отличающиеся от основной группы наблюдений. На рисунке 2.1 в качестве примеров показаны графики между зависимой переменной y и факторами x_1, x_2, x_3, x_4. Аномальные наблюдения обведены кружочком.

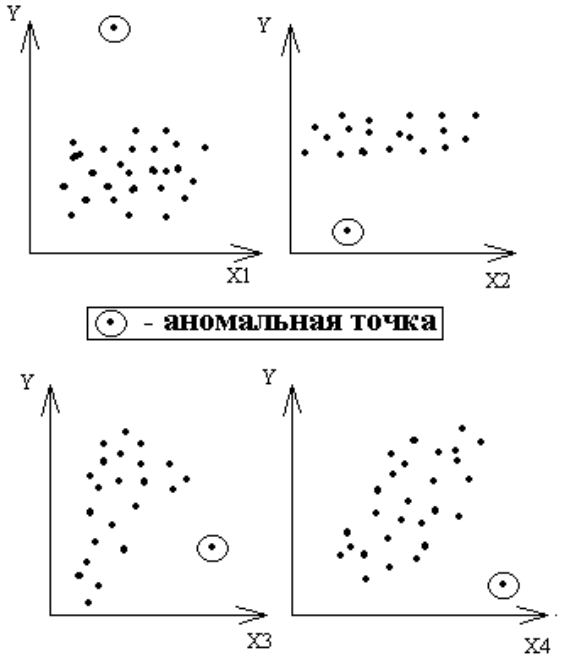


Рисунок 2.1 — Графики рассеяния с аномальными наблюдениями

Прежде чем приступить к определению типа уравнения множественной регрессии и его параметров, необходимо исследовать факторы (признаки $x_1, x_2, ...x_k$) на пригодность их корректного использования в модели. Факторы, входящие в уравнение, должны быть независимы между собой. Для этого определяют коэффициенты парной корреляции, которые характеризуют тесноту линейной связи между парой факторов. Изменяются в диапазоне –1 до 1. Вычисляют по формуле:

$$r_{x_i x_j} = \frac{\overline{x_i x_j} - \overline{x}_i \cdot \overline{x}_j}{\sigma_{x_i} \cdot \sigma_{x_j}}, \qquad (2.23)$$

где $\overline{x_i x_j}$ — среднее произведение признаков x_i и x_j, $\overline{x}_i$ и $\overline{x}_j$ — средние значения каждого признака, σ_{x_i} и σ_{x_j} — их средние квадратические отклонения.

Проверка коэффициентов парной корреляции на значимость по критерию Стьюдента проверяется так же, как для простой корреляции, изложенной ранее, по формуле (2.20).

После вычисления коэффициентов парной корреляции их записывают в корреляционную матрицу:

$$K = \begin{pmatrix} 1 & r_{x_1 x_2} & r_{x_1 x_3} \\ r_{x_2 x_1} & 1 & r_{x_2 x_3} \\ r_{x_3 x_1} & r_{x_3 x_2} & 1 \end{pmatrix}.$$

Анализ графа связей между факторами (рис. 2.2) является очень важным этапом предварительной подготовки данных при составлении уравнения регрессии. Для построения графа связей используют коэффициенты парной корреляции. Наиболее значимые связи показывают жирными стрелочками, слабые связи отмечают пунктирными линиями. Далее с помощью графа и с учетом значений корреляционной матрицы отбирают переменные для модели множественной регрессии. Если факторные переменные уравнения $x_1, x_2, ...x_k$ статистически взаимосвязаны, то они называются мультиколлинеарными. Модель с такими факторами имеет низкую надежность прогноза. Поэтому в уравнении регрессии мультиколлинеарность должна быть исключена.

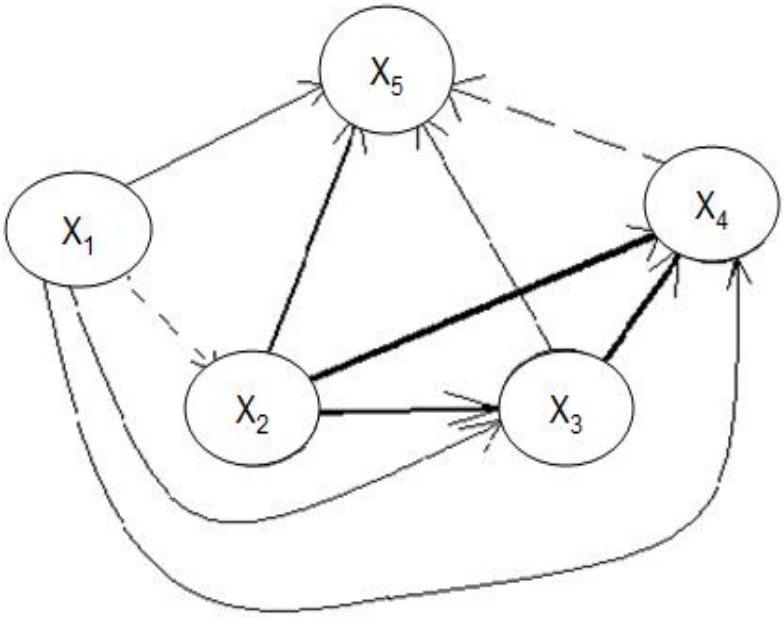


Рисунок 2.2 — Граф связей между факторами

После тщательного отбора факторов, входящих в уравнение множественной регрессии, приступают к определению его формы и параметров.

Линейное уравнение множественной регрессии имеет вид:

$$\overline{y}_x = a_0 + a_1 x_1 + a_2 x_2 + ... + a_k x_k . \qquad (2.24)$$

Пример нелинейного уравнения:

$$\overline{y} = a_o + a_1 \cdot x_1 + a_2 \cdot x_2^2 + ... + a_k \cdot x_k - \frac{b}{x_1} . \qquad (2.25)$$

Для нахождения параметров $a_o, a_1, a_2, ..., a_k$ линейной множественной регрессии (2.24), записываются матрицы:

$$A = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ ... \\ a_k \end{pmatrix} \qquad x = \begin{pmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ 1 & x_{31} & ... & x_{3k} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{nk} \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_n \end{pmatrix} .$$

Здесь x_{ij} обозначает наблюдаемое значение *i*-го признака для *j*-го наблюдения.

В матричной форме уравнение регрессии имеет вид: $X \cdot A = Y$. Умножим обе части уравнения слева на транспонированную матрицу X^T. Получим: $X^T \cdot X \cdot A = X^T \cdot Y$. Обозначим матрицу моментов $B = X^T \cdot X$. Тогда из матричного уравнения $B \cdot A = X^T \cdot Y$ можно найти матрицу оценок параметров уравнения множественной регрессии:

$$A = B^{-1} \cdot (X^T \cdot Y).$$

Дисперсионный анализ

Дисперсионный анализ — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. В отличие от t-критерия, позволяет сравнивать средние значения трёх и более групп. Разработан Р. Фишером для анализа результатов экспериментальных исследований. Во многих статистических программах обозначается как ANOVA (от англ. Analysis of variance).

Оценку адекватности модели в случае уравнения множественной регрессии выполняют на основании дисперсионного анализа после того, как уравнение получено и параметры оценены.

Основные этапы проверки:

1) Определение множественного коэффициента детерминации по формуле (2.18). Интервалы измерения: $0 \le R^2 \le 1$. Чем больше R^2, тем связь между результативным показателем y и факторами $x_1, x_2, ...x_k$ сильнее. Обозначает тесноту связей между признаками, объясненную данным уравнением регрессии.

При большом количестве факторов используют нормированный множественный коэффициент детерминации:

$$\tilde{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}. \qquad (2.26)$$

2) Определение множественного коэффициента корреляции, характеризующего тесноту линейной связи между y и $x_1, x_2, ...x_k$, по формуле:

$$R = \sqrt{R^2}. \qquad (2.27)$$

3) Проверка модели на адекватность по критерию Фишера.

Вычисляют наблюдаемое значение критерия по формуле:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}, \qquad (2.28)$$

где n — объем выборки, k — число переменных в уравнении регрессии.

Определяют критическое значение критерия по таблице критических точек распределения Фишера-Снедекора:

$$F_{кр} = F(\alpha, k_1, k_2),$$

где α — ошибка первого рода (обычно $\alpha = 0,05$ или $\alpha = 0,01$), $k_1 = k$ — число переменных в уравнении регрессии, $k_2 = n - k - 1$.

Если наблюдаемое значение критерия $F > F_{кр}$, то полученное уравнение множественной регрессии адекватно описывает фактические наблюдения; если наблюдаемое значение критерия окажется меньше критического $F < F_{кр}$, то построенная модель не адекватна реальной.

2.4 Временные ряды

Если значения признака расположены в хронологической последовательности, то выборку рассматривают как временной ряд. В отличие от элементов случайной выборки, члены временного ряда не являются статистически независимыми, т. е. фактор времени оказывает существенное влияние на формирование его значений. Как правило, временной фактор задает тенденцию (тренд) изменения признака. Тренд может быть определен в виде математической функции $x = f(t)$ или специальными непараметрическими методами, нацеленными на сглаживание наблюдаемых значений временного ряда. При определении вида функции $x = f(t)$, параметров уравнения тренда и оценки статистической надежности используются подходы и методы регрессионного анализа.

Временной ряд может иметь сезонную составляющую, которая формирует повторяющиеся в определенное время года колебания ана-

лизируемого признака. Сезонную составляющую также можно задать в виде периодической функции, с периодами, кратными сезонам. В аналитическом выражении этой функции участвуют гармоники (тригонометрические функции), периодичность которых обусловлена содержательной сущностью задачи.

Обозначение временного ряда $X_t : x_1, x_2, x_3 ... x_i ... x_N$, где N — количество наблюдений.

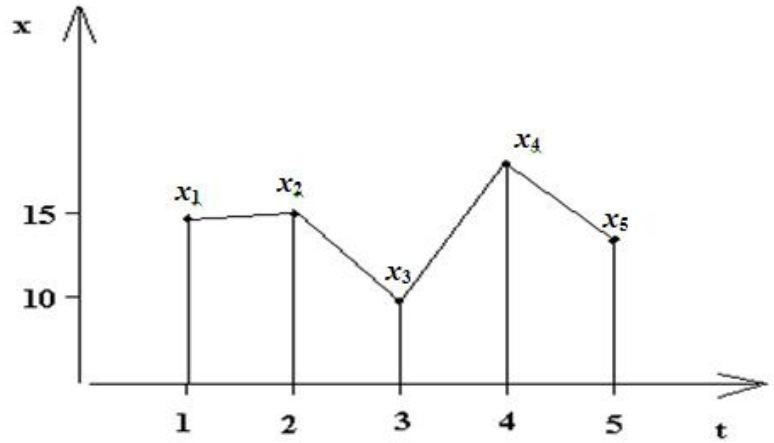


Рисунок 2.3 — Графическое изображение временного ряда

Различают числовые ряды:

– с одинаковыми интервалами между наблюдениями — эквидистантные ряды $\Delta t = const$ (рис. 2.4);

– с различными интервалами между наблюдениями — неэквидистантные ряды (рис. 2.5)

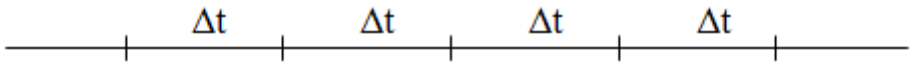


Рисунок 2.4 — Временная ось эквидистантного ряда

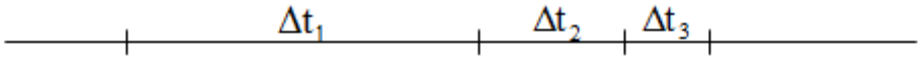


Рисунок 2.5 — Временная ось неэквидистантного ряда

Основные методы анализа временных рядов:

1. Графическая визуализация (визуальный экспресс-анализ).
2. Спектральный анализ.
3. Анализ автокорреляционной функции.
4. Сглаживание временного ряда.
5. Метод сезонной декомпозиции.
6. Нейросетевое прогнозирование.
7. ARIMA-модели.
8. Метод SSA.

Далее рассмотрим некоторые из этих методов.

Визуальный экспресс-анализ

На основе исходных данных строится график временного ряда (по оси OX отмечается время t, по оси OY — значения временного ряда) и делается предварительное заключение. Основные типы временных рядов приведены на рисунках (2.6–2.11).

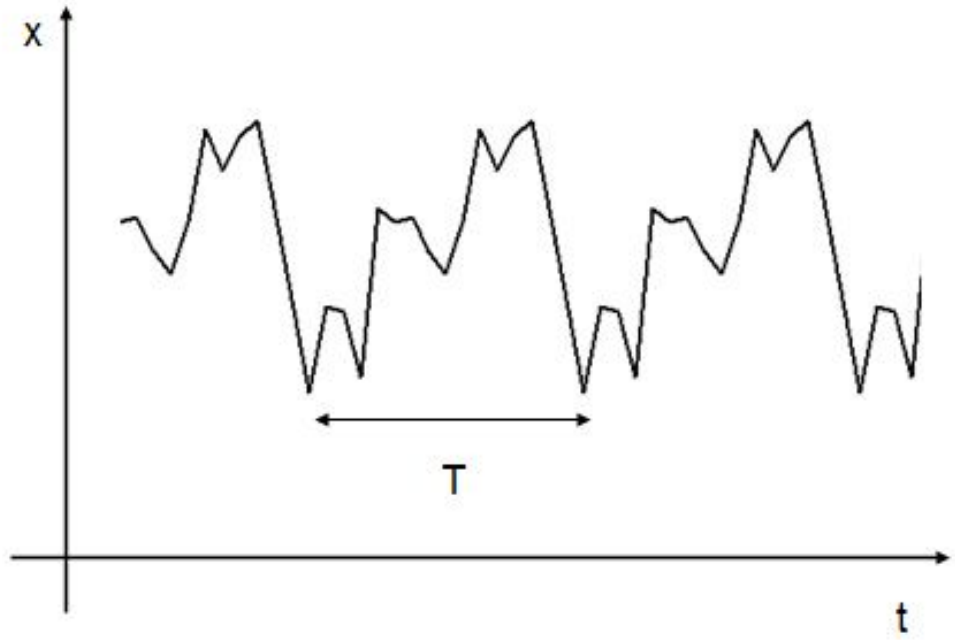


Рисунок 2.6 — Графическое изображение временного ряда циклического процесса

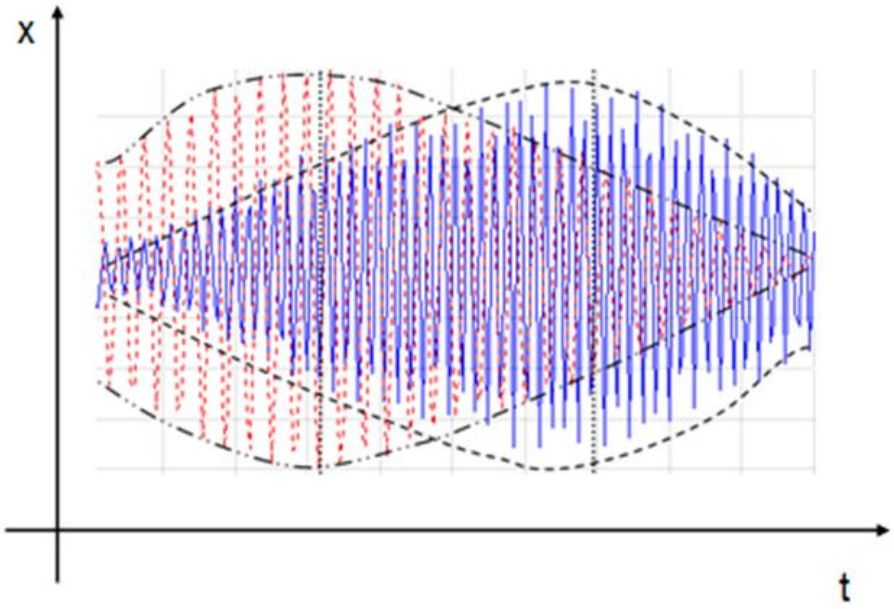


Рисунок 2.7 — Суперпозиция нескольких циклических компонент

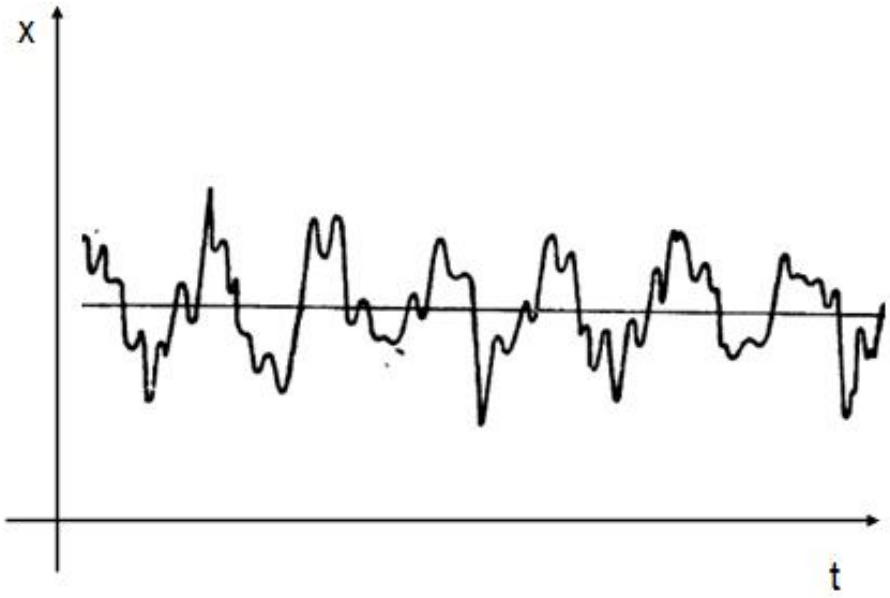


Рисунок 2.8 — Графическое изображение временного ряда стационарного процесса

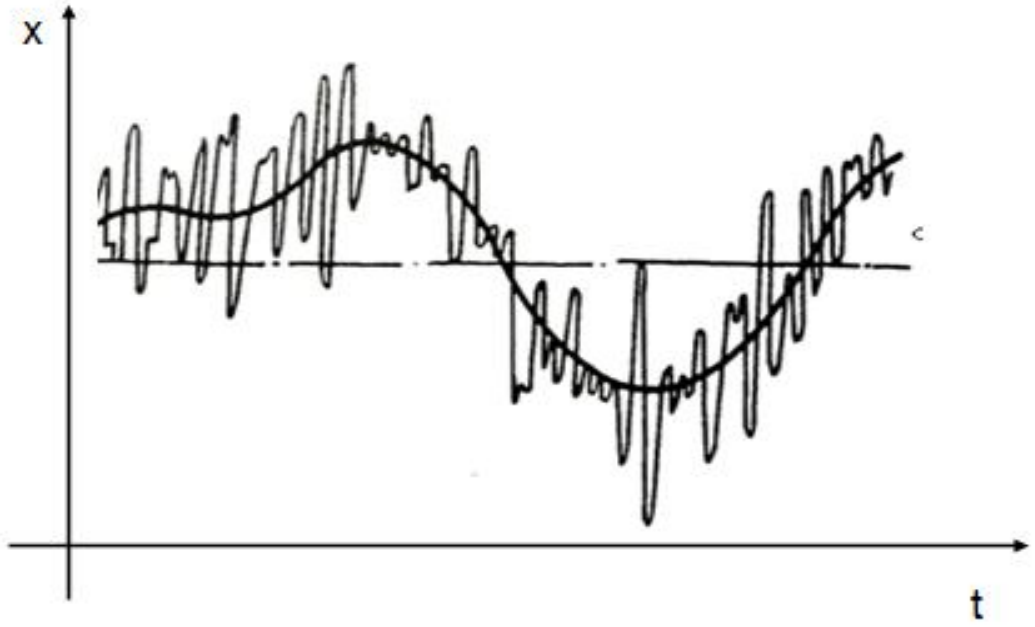


Рисунок 2.9 — Графическое изображение временного ряда нестационарного процесса по математическому ожиданию

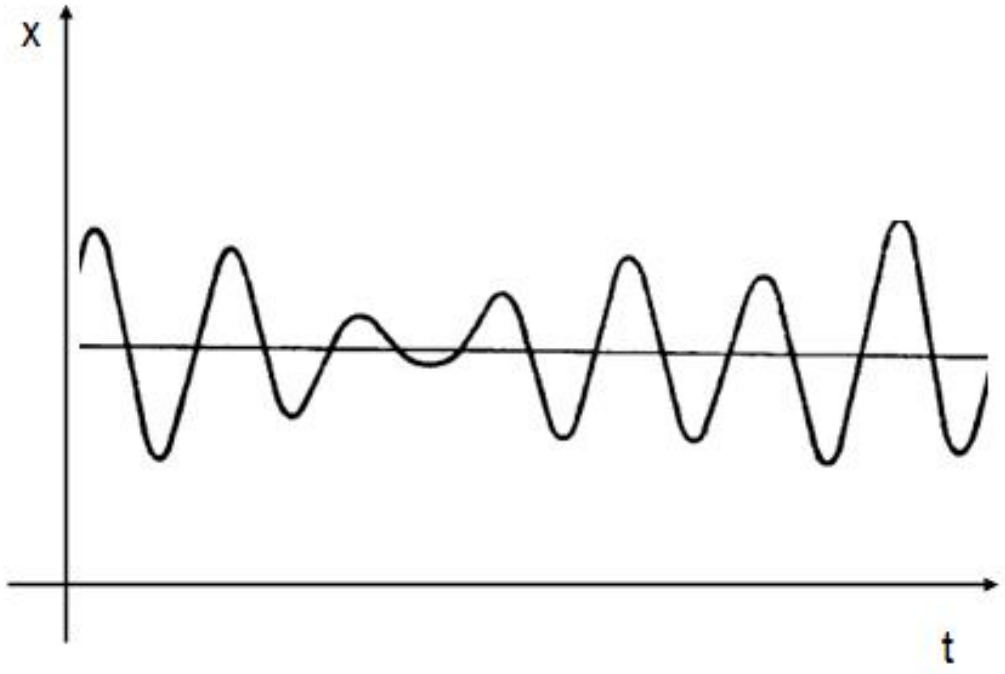


Рисунок 2.10 — Графическое изображение временного ряда нестационарного процесса по дисперсии

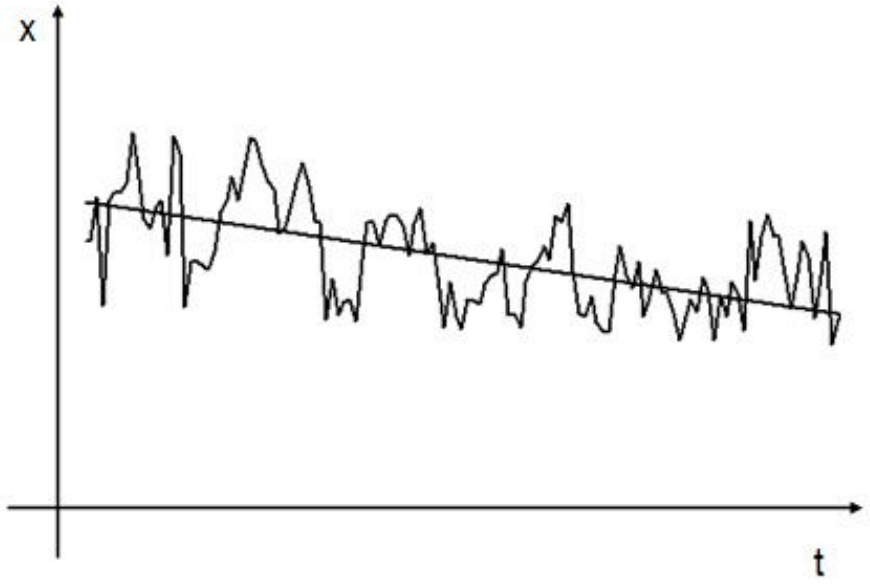


Рисунок 2.11 — Графическое изображение временного ряда с сезонной компонентой и трендом

Во многих случаях временной ряд можно представить в виде гармонической модели (рис. 2.12):

$$x = a_0 + a \cdot \sin(\omega t + \varphi), \tag{2.29}$$

где a_0 — средний уровень ряда, a — амплитуда колебаний, ω — циклическая частота, ϕ — фазовое смещение.

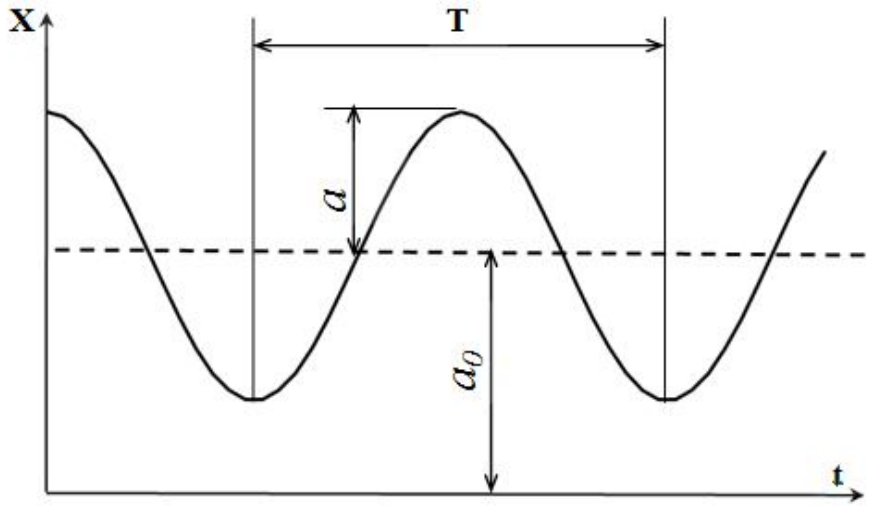


Рисунок 2.12 — Основные параметры модели гармонических колебаний на графике

В этом случае циклическая (круговая) частота колебаний определяется по формулам:

$$\omega = \frac{2\pi}{T}, \tag{2.30}$$

$$\omega = 2\pi \cdot f, \tag{2.31}$$

где Т — период колебаний, f — частота колебаний:

$$f = \frac{1}{T}. \tag{2.32}$$

Различают высокочастотные и низкочастотные колебания (рис. 2.13). Если период увеличивается, то частота уменьшается согласно формуле (2.32).

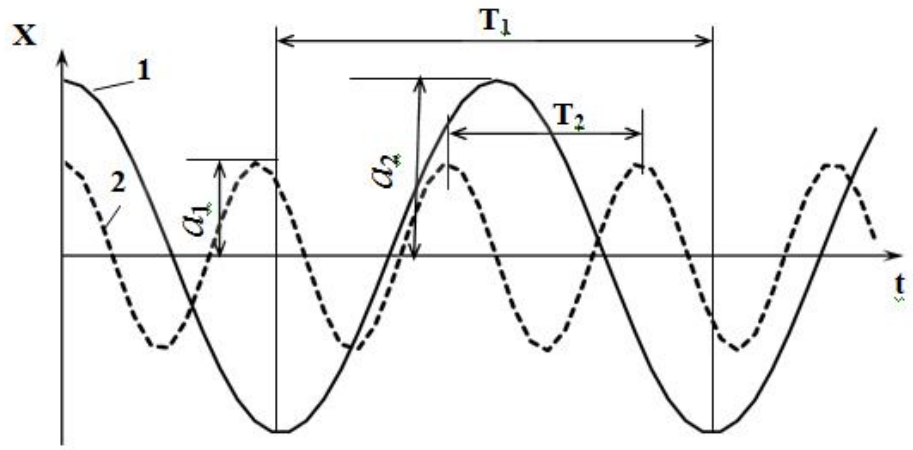


Рисунок 2.13 — Графики и параметры низкочастотных (1) и высокочастотных (2) колебаний

Этап визуального анализа динамики временных рядов очень важен, поскольку он помогает выявить аномальные наблюдения, оценить структуру ряда, сделать предположения о типе случайного процесса им представляемого и др.

Спектральный анализ

Метод состоит в разложении сложного процесса на простые гармонические составляющие, которые затем анализируются и интерпретируются. Спектральный анализ представляет исходный ряд данных в виде суммы гармонических функций ($\sin x$ и $\cos x$):

$$x(t_i) = a_0 + \sum_{k=1}^{\frac{N}{2}} \left[a_k \cos(2\pi f_k t_i) + b_k \sin(2\pi f_k t_i) \right] + e(t_i), \qquad (2.33)$$

где $i = \overline{1; N}$, N — количество наблюдений;

t_i — i-ый момент времени;

$x(t_i)$ — значение показателя x в момент времени t_i;

$a_0 = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x(t_i)$ — средний уровень ряда;

k — номер гармоники, $k = \overline{1; \frac{N}{2}}$;

$f_k = \frac{k}{T}$ — частота k-ой гармоники ряда;

T — длина реализации;

a_k, b_k — коэффициенты, которые находятся по формулам:

$$a_k = \begin{cases} \frac{2}{N} \sum_{i=1}^{N} x(t_i) \cos(2\pi f_k t_i), & k = \overline{1; \frac{N}{2} - 1} \\ \frac{1}{N} \sum_{i=1}^{N} x(t_i) \cos(2\pi f_k t_i), & k = \frac{N}{2} \end{cases},$$

$$b_k = \begin{cases} \frac{2}{N} \sum_{i=1}^{N} x(t_i) \sin(2\pi f_k t_i), & k = \overline{1; \frac{N}{2} - 1} \\ 0, & k = \frac{N}{2} \end{cases},$$

$e(t_i)$ — ошибка аппроксимации.

Амплитуда $c_k = \sqrt{{a_k}^2 + {b_k}^2}$ характеризует дисперсию k-ой гармоники. Величина ${c_k}^2 = {a_k}^2 + {b_k}^2$ характеризует интенсивность k-ой гармоники. Ордината периодограммы $p_k = \left({a_k}^2 + {b_k}^2 \right) \cdot \frac{N}{2}$ характеризует дисперсию (т.е. изменчивость показателя) за счет k-ой гармоники. Гармоники не коррелированы между собой, следовательно, никакие две гармоники не будут учитывать одну и ту же часть дисперсии. Значит, дисперсии, учитываемые различными гармониками, можно суммировать. Согласно теореме Парсеваля, суммарный вклад гармоник равен дисперсии ряда:

$$\sigma^2 = \frac{1}{2}\sum_{k=1}^{\frac{N}{2}-1}\left(a_k{}^2 + b_k{}^2\right) + a_{\frac{N}{2}}{}^2, \qquad (2.34)$$

где $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x(t_i) - \bar{x}\right)^2$ — оценка общей дисперсии ряда. Если N — нечетное, то в равенстве (2.34) коэффициент $a_{\frac{N}{2}} = 0$.

Выделение значимых гармоник возможно по двум критериям: по вкладу в общую дисперсию и по статистическому критерию сравнения дисперсий:

А) Вклад k-ой гармоники в общую дисперсию процесса определяется из теоремы Парсеваля:

$$\rho_k{}^2 = \frac{c_k{}^2}{2\sigma^2}\cdot 100\% = \frac{a_k{}^2 + b_k{}^2}{\sum_{k=1}^{\frac{N}{2}-1}\left(a_k{}^2 + b_k{}^2\right) + a_{\frac{N}{2}}{}^2}\cdot 100\%. \qquad (2.35)$$

Основными m гармониками принято считать те, для которых $\sum_{k=1}^{m}\rho_k{}^2 \geq 80\%$.

Б) Значимость k-ой гармоники оценивается по критерию Фишера.

Основная гипотеза $H_0: \frac{c_k{}^2}{\nu_1} = \frac{\sigma^2}{\nu_2}$.

Альтернативная гипотеза $H_1: \frac{c_k{}^2}{\nu_1} > \frac{\sigma^2}{\nu_2}$.

Величина $\frac{\frac{c_k{}^2}{\nu_1}}{\frac{\sigma^2}{\nu_2}}$ имеет распределение Фишера $F(\alpha,\nu_1,\nu_2)$, где α — уровень значимости ($\alpha = 0{,}05$), ν_1 — число степеней свободы, приходящееся на k-ю гармонику ряда ($\nu_1 = 2$), ν_2 — число степеней свободы исходного ряда ($\nu_2 = N - 1$). Критерий имеет правостороннюю критическую область.

Если сложить графики всех частот, то получим временной ряд:

$$x(t_i) = a_0 + \sum_{k=1}^{\frac{N}{2}}\left[a_k\cos(2\pi f_k t_i) + b_k\sin(2\pi f_k t_i)\right],$$

который отличается от исходного ряда (рис. 2.14) на случайную величину e(t_i).

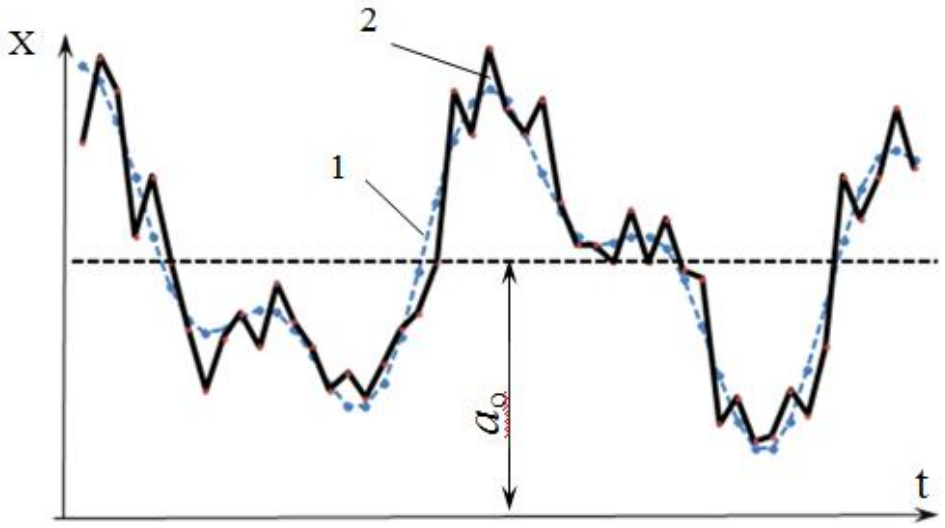


Рисунок 2.14 — Расчетные значения (1) и исходные значения (2) временного ряда

Структура ряда

Структура ряда — это математическая модель, объясняющая схему влияния факторов на формирование значений ряда. Временной ряд $x_1, x_2, x_3 ... x_i ... x_N$ может содержать: трендовую компоненту T (рис. 2.15); сезонную или циклическую компоненту S (рис. 2.13); случайную компоненту E.

Общая модель структуры ряда $X = f(T, S, E)$ имеет вид:

– аддитивная $X = T + S + E$ (рис. 2.16 а);

– мультипликативная $X = T \cdot S \cdot E$ (рис. 2.16 б).

Визуальные признаки аддитивной структуры ряда — отклонения значений ряда от линии тренда циклические и с относительно постоянной амплитудой. Визуальные признаки мультипликативной структуры ряда — отклонения значений ряда от линии тренда циклические, но с увеличивающейся или уменьшающейся амплитудой.

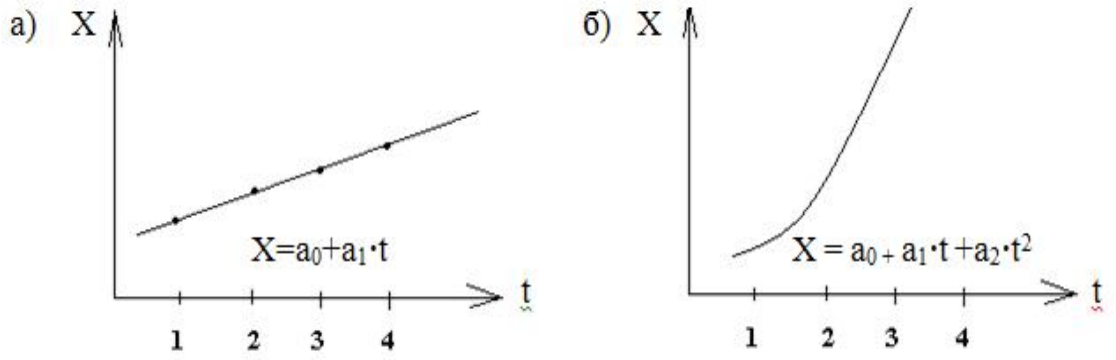


Рисунок 2.15 — Виды трендов: а) линейный; б) параболический

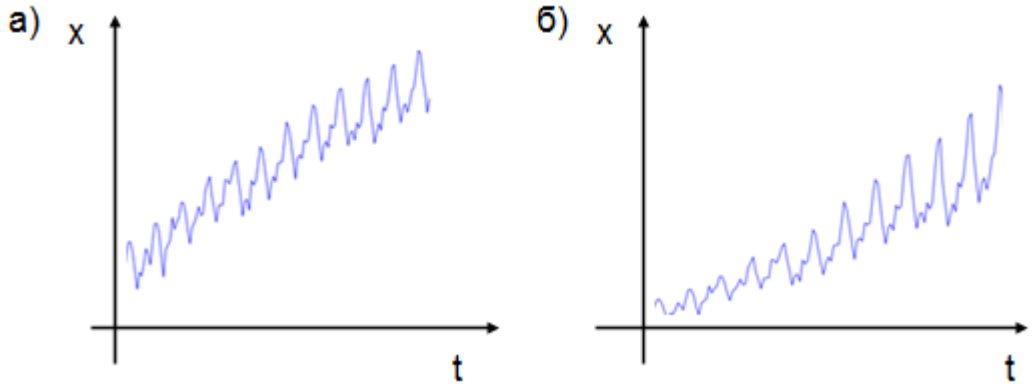


Рисунок 2.16 — Виды структуры ряда: а) аддитивная, б) мультипликативная

Автокорреляционная функция

Используется для выявления структуры ряда. В отличие от элементов случайной выборки, члены временного ряда не являются статистически независимыми, т. е. текущее значение ряда зависит от предыдущих. Для проверки степени зависимости между членами временного ряда используют ряды со смещением (лагом).

Для каждого смещенного ряда на лаг k и исходного ряда находят коэффициент автокорреляции:

$$r_k = \frac{\overline{x_t x_{t-k}} - \overline{x}_t \cdot \overline{x}_{t-k}}{\sigma_t \cdot \sigma_{t-k}},$$

где x_t — исходный ряд;

x_{t-k} — ряд, смещенный на k лагов.

Проверка коэффициентов автокорреляции на значимость по критерию Стьюдента проверяется так же, как для простой корреляции, изложенной ранее, по формуле (2.20).

Вычисленные значения коэффициентов автокорреляции r_k в зависимости от лага k наносят на график и делают следующие выводы:

– если значимый коэффициент автокорреляции наблюдается только при лаге $k=1$, то в исследуемом ряде присутствует тренд;

– если на некотором лаге k коэффициент автокорреляции значим и имеет наибольшее значение, то в ряде присутствует сезонная компонента периода k;

– если все коэффициенты автокорреляции незначимы, то ряд представляет собой случайную величину.

Сглаживание временного ряда

А) Метод скользящей средней

Любую функцию можно сгладить тремя, пятью и даже семью точками. Это позволяет уменьшить разброс значений выходных данных, выявить тенденцию временного ряда и предсказать будущие значения, т.е. сделать прогноз. Общая тенденция может быть положительной (рост параметра), отрицательной (убывание) или указывать на колебательный или экстремальный характер динамики.

Простейший алгоритм сглаживания скользящей средней по трем точкам осуществляется по формулам:

$$\overline{x_1}=\frac{5x_1+2x_2-x_3}{6};\ \overline{x_i}=\frac{x_{i-1}+x_i+x_{i+1}}{3};i=\overline{2N-1};$$

$$\overline{\overline{x_N}}=\frac{5x_N+2x_{N-1}-x_{N-2}}{6}.$$

Первые и последние средние вычисляются по фактическим значениям трех первых и трех последних значений исходного временного ряда, а промежуточные — как средние трех смежных значений.

Сглаженный с помощью скользящей средней временной ряд наносят на график исходного временного ряда (рис. 2.17). По характеру тренда оценивают тенденцию изменения выброса за весь период предыстории. Это может быть монотонное снижение или рост, колебания вокруг среднего значения или кривая с экстремумом.

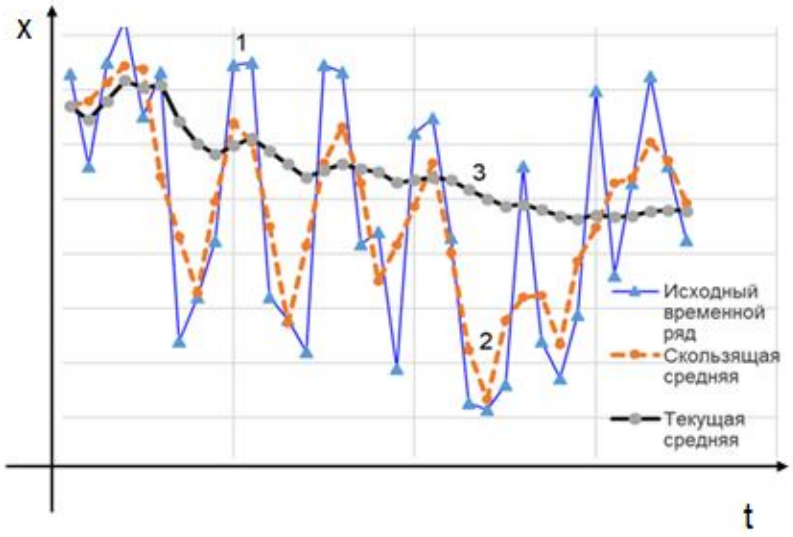


Рисунок 2.17 — Временные ряды: 1 — исходный, 2 — сглаженный скользящей средней по трем точкам, 3 — сглаженный по методу текущей средней

Б) Метод текущей средней

Часто для оценки прогнозного значения параметра в требуемый момент времени достаточно знать среднее значение по предыстории. В этом случае вычисляют среднее значение по формуле $\bar{x} = \frac{1}{N}\sum_{1}^{N} x_i$, что требует участия всех значений временного ряда. Поэтому иногда используют рекуррентный алгоритм накопления текущего значения средней, что требует запоминания только предыдущего значения средней $\bar{x}_{i-1}$, текущего значения временного ряда x_i и его порядкового номера i: $\bar{x}_i = \bar{x}_{i-1} + \frac{1}{i}(x_i - \bar{x}_{i-1})$, $i = 1,...,N$. На рисунке 2.17 представлен сглаженный ряд по методу текущей средней в виде кривой 3. Данная кривая отчетливо показывает убывающий по времени тренд.

В) Метод экспоненциального сглаживания

Метод экспоненциального сглаживания учитывает степень влияния каждой точки исходного временного ряда. Очевидно, что во многих случаях последние точки предыстории имеют значительно большее влияние на прогноз будущих значений, чем первые. Сущность метода заключается в сглаживании временного ряда с помощью взвешенной

скользящей средней, в которой вес наблюдений подчиняется экспоненциальному закону вероятности.

Рекуррентная формула Г. Брауна для определения экспоненциальной средней p-го порядка имеет вид:

$$S_t^{[p]} = \alpha \cdot S_t^{[p-1]} + \beta \cdot S_{t-1}^{[p]},$$

где α — безразмерный параметр сглаживания, выбирается в пределах $0 \le \alpha \le 1$, при этом он позволяет управлять влиянием данных временного ряда на прогнозируемую точку (при $\alpha = 0$ на прогноз будут одинаково влиять все точки временного ряда, а при $\alpha = 1$ — только последняя); $\beta = 1 - \alpha$.

Параметры выбирают двумя способами:

– с помощью интуиции;

– по формуле: $\alpha = \dfrac{2}{(t+1)}$, где t — количество точек стабильного ряда.

Параметр $p = 1,...,n$ равен порядку интерполирующего полинома. Характерно, что при $p = 1$, экспоненциальные средние $S_t^{p-1} = S_t^0$ представляют собой исходный временной ряд x_t. Количество решаемых уравнений зависит от p. При p=1 имеем модель нулевого порядка с одним рекуррентным уравнением вида:

$$S_t^{[1]} = \alpha \cdot x_t + \beta \cdot S_{t-1}^{[1]}.$$

Согласно этой модели получим тренд, начальная точка которого равна среднему значению предыстории, то есть

$$S_0^1 = \frac{\sum_{1}^{t} x_t}{t},$$

а прогнозное значение оценивается величиной: $\tilde{x}_{t+\tau} = S_t^{[1]}$.

Таким образом, последнее значение, вычисленное по рекуррентной модели нулевого порядка, численно равно прогнозу на следующий дискретный момент времени.

Точность прогноза оценивают по контрольной точке $x_{t+\tau}$ временного ряда, которая сравнивается с прогнозируемой:

$$\delta_{i+\tau} = \frac{\tilde{x}_{t+\tau} - x_{t+\tau}}{x_{t+\tau}} \cdot 100\%.$$

Контрольные вопросы

1. Укажите основные отличия активного и пассивного экспериментов, их преимущества и недостатки.
2. Приведите порядок проведения пассивного эксперимента в производственных условиях.
3. Что такое случайная величина? Приведите примеры в экологии.
4. Каков экологический смысл математического ожидания случайной величины, описывающей ущерб от определенного вида техногенного воздействия?
5. Перечислите основные статистические характеристики.
6. Сформулируйте общее правило проверки статистических гипотез.
7. В чем состоит сущность нормального закона распределения?
8. Как применяется правило трех сигм?
9. Основное назначение корреляционного анализа.
10. Смысл коэффициента корреляции.
11. Виды регрессии.
12. Каким методом проводится оценка коэффициентов регрессионного уравнения?
13. Что характеризует средняя квадратическая ошибка уравнения?
14. Что характеризует средняя относительная погрешность модели?
15. Основные термины временного ряда.
16. Что такое структура временного ряда?
17. Что характеризует автокорреляционная функция?
18. Какие задачи решает спектральный анализ?
19. Для чего сглаживают временные ряды?
20. Перечислите основные методы сглаживания временного ряда.

3. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ПУТЕМ АКТИВНОГО ЭКСПЕРИМЕНТА

3.1 Основные понятия активного эксперимента

При планировании эксперимента в экологии исследователь должен составить четкую и последовательную логическую схему исследования, максимально формализовать математическую модель и обеспечить высокую надежность результатов экспериментальных исследований. Этим требованиям удовлетворяют статистические методы планирования активного эксперимента. Наиболее распространенным статистическим методом планирования является полный факторный эксперимент.

Для изучения влияние ряда факторов x_1, x_2, … x_k на некоторую характеристику объекта исследования — функцию отклика y, проводят эксперименты по определенному плану, в котором реализуются все возможные комбинации факторов. Каждый фактор рассматривается на p фиксированных уровнях. Самый распространённый случай — два уровня: верхний и нижний. Величина интервала варьирования оказывает влияние на адекватность модели. Рекомендуется интервал варьирования выбирать в пределах 0,05–0,3 от возможного диапазона изменения исследуемого фактора. Факторы и интервалы их варьирования записываются в таблицу 3.1

Таблица 3.1 — Первичная таблица исходных факторов

Фактор	Единицы измерения	Обозначения	Диапазон варьирования	
			нижний уровень	верхний уровень
1		x_1	$x_1^{\min}$	$x_1^{\max}$
...				
k		x_k	$x_k^{\min}$	$x_k^{\max}$

Общее число экспериментов равно $n = p^k$, где p — число уровней, k — число факторов. Если в модели используется два фактора x_1, x_2, каждый из которых имеет два уровня, то общее число экспериментов

$n = 2^2 = 4$. Если три фактора x_1, x_2, x_3 на двух уровнях, то общее число экспериментов $n = 2^3 = 8$. Проведение экспериментов по такому плану называется полным факторным экспериментом типа 2^k (ПФЭ типа 2^k).

Первый этап исследования — составление плана эксперимента, т.е. определение условий для всех опытов, которые необходимо провести. План эксперимента задается в виде матрицы планирования, строки которой определяют условия опыта, а столбцы — значения контролируемых параметров процесса. В последнем столбце матрицы указываются средние значения функции отклика, полученные экспериментальным путем в серии опытов, проведенных в соответствии с условиями матрицы планирования.

При проведении эксперимента необходимо свести к минимуму влияние случайных параметров исследуемого процесса на функцию отклика. С этой целью проводят несколько параллельных опытов (серию опытов) при одинаковых условиях, предусмотренных соответствующей строкой матрицы планирования. Также при проведении эксперимента необходимо обеспечить взаимную компенсацию неконтролируемых параметров процесса. Для этого, перед непосредственной реализацией плана, последовательность проведения опытов на исследуемом объекте выбирают случайным образом, например, с помощью таблицы случайных чисел. Если дальнейшие расчеты покажут, что требования к модели не выполняются, то количество опытов увеличивают.

Для случая двух и трех факторов на двух уровнях матрица планирования эксперимента имеет вид (табл. 3.2 и 3.3):

Таблица 3.2 — Матрица планирования эксперимента типа 2^2

№ эксперимента	Значения факторов		Результаты m опытов		
	x_1	x_2	y_1	...	y_m
1	-1	-1	y_{11}	...	y_{1m}
2	+1	-1	y_{21}	...	y_{2m}
3	-1	+1	y_{31}	...	y_{3m}
4	+1	+1	y_{41}	...	y_{4m}

Таблица 3.3 — Матрица планирования эксперимента типа 2^3

№ эксперимента	Значения факторов			Результаты m опытов		
	x_1	x_2	x_3	y_1	…	y_m
1	-1	-1	-1	y_{11}	…	y_{1m}
2	+1	-1	-1	y_{21}	…	y_{2m}
3	-1	+1	-1	y_{31}	…	y_{3m}
4	+1	+1	-1	y_{41}	…	y_{4m}
5	-1	-1	+1	y_{51}	…	y_{5m}
6	+1	-1	+1	y_{61}	…	y_{6m}
7	-1	+1	+1	y_{71}	…	y_{7m}
8	+1	+1	+1	y_{81}	…	y_{8m}

Значение «–1» указывает, что во время опыта значение фактора устанавливается на нижнем уровне, а «+1» — значение фактора на верхнем уровне диапазона варьирования. Результаты m опытов в каждом i-ом эксперименте ($i = 1,...n$) записывают в столбцы матрицы планирования.

3.2 Обработка результатов эксперимента

Важным условием успешного планирования является контроль воспроизводимости результатов исследования (проверка однородности дисперсий функции отклика в результате проведения параллельных опытов). Если эксперименты признаны воспроизводимыми, то их результаты можно использовать для оценки параметров математической модели, если нет — необходимо увеличивать число параллельных опытов.

С этой целью сначала определяют среднее значение результатов всех опытов для каждого эксперимента по формуле:

$$\overline{y}_i = \frac{1}{m}\sum_{q=1}^{m} y_{iq},\ i = 1,...n, \qquad (3.1)$$

где y_{iq} — результат отдельного q-го опыта в i-ом эксперименте.

Затем рассчитывают дисперсию результатов каждого эксперимента по формуле:

$$S_i^2 = \frac{1}{m-1}\sum_{q=1}^{m}\left(y_{iq} - \overline{y}_i\right)^2, \qquad (3.2)$$

где m — число опытов в каждом эксперименте.

В последующем анализе адекватности модели используется дисперсия воспроизводимости $S^2_{\{y\}}$, которая характеризует ошибку всего эксперимента. При одинаковом числе опытов в каждом эксперименте $S^2_{\{y\}}$ рассчитывают по формуле:

$$S^2_{\{y\}} = \frac{1}{n}\sum_{i=1}^{n} S_i^2, \qquad (3.3)$$

где n — число экспериментов.

Для обработки результатов проведенных экспериментов факторы приводят к одному масштабу. Это достигается путем кодирования переменных. Обозначим нижний уровень фактора $x_j^{\min}$, а верхний уровень — $x_j^{\max}$ ($j = 1,...,k$). Тогда основной уровень (центр плана) определяем по формуле:

$$x_j^0 = \frac{x_j^{\max} + x_j^{\min}}{2},\ j = 1,...,k. \qquad (3.4)$$

Интервал варьирования вычисляется по формуле:

$$\Delta x_j = \frac{x_j^{\max} - x_j^{\min}}{2},\ j = 1,...,k. \qquad (3.5)$$

Безразмерные координаты (коды) рассчитываются по формуле:

$$\tilde{x}_j = \frac{x_j - x_j^0}{\Delta x_j},\ j = 1,...,k. \qquad (3.6)$$

При таком кодировании новые переменные будут принимать значения от –1 до +1, т.е. $\tilde{x}_j \in [-1;1]$, $j = 1,...,k$.

Следующий этап исследования состоит в разработке математической модели процесса. Под моделью понимается функция $y = f(x_1, x_2, ...x_k)$, которая связывает изучаемый показатель (функцию

отклика) со значениями факторов, лежащих в интервале между верхним и нижним уровнями. Модели разрабатываются по принципу «от простого — к более сложному». Согласно этому принципу сначала предполагают линейную модель исследуемого процесса, и если не достигается требуемая точность и надежность модели, то ее усложняют. Самыми простыми моделями считаются алгебраические полиномы. В случае ПФЭ типа 2^2 математическая модель имеет вид:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 . \qquad (3.7)$$

Для ПФЭ типа 2^3:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + b_{123} x_1 x_2 x_3 . \qquad (3.8)$$

Если необходимо учесть другие взаимодействия, то число слагаемых увеличивают.

Прежде чем определить коэффициенты модели, записывают матрицу планирования относительно новых переменных, в которую дополнительно включают столбцы взаимодействия факторов. Знаки этих столбцов получают с помощью исходной матрицы планирования путем перемножения соответствующих столбцов (табл. 3.4 и 3.5).

Таблица 3.4 — План-матрица для обработки результатов ПФЭ 2^2

№ опыта	Значения факторов				Совместное влияние	Вспомогательный столбец	Отклик (среднее результатов опытов)
	в натуральном масштабе		в безразмерном масштабе				
	x_1	x_2	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_1\tilde{x}_2$	$\tilde{x}_0$	$\overline{y}$
1	$x_1^{\min}$	$x_2^{\min}$	–1	–1	+1	1	$\overline{y}_1$
2	$x_1^{\max}$	$x_2^{\min}$	+1	–1	–1	1	$\overline{y}_2$
3	$x_1^{\min}$	$x_2^{\max}$	–1	+1	–1	1	$\overline{y}_3$
4	$x_1^{\max}$	$x_2^{\max}$	+1	+1	+1	1	$\overline{y}_4$

Таблица 3.5 — План-матрица для обработки результатов ПФЭ 2^3

№ эксперимента	Значения факторов						Совместное влияние				Вспомогательный столбец	Отклик (среднее результатов опытов)
	в натуральном масштабе			в безразмерном масштабе								
	x_1	x_2	x_3	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_1\tilde{x}_2$	$\tilde{x}_1\tilde{x}_3$	$\tilde{x}_2\tilde{x}_3$	$\tilde{x}_1\tilde{x}_2\tilde{x}_3$	$\tilde{x}_0$	$\overline{y}$
1	$x_1^{\min}$	$x_2^{\min}$	$x_3^{\min}$	–1	–1	–1	+1	+1	+1	–1	1	$\overline{y}_1$
2	$x_1^{\max}$	$x_2^{\min}$	$x_3^{\min}$	+1	–1	–1	–1	–1	+1	+1	1	$\overline{y}_2$
3	$x_1^{\min}$	$x_2^{\max}$	$x_3^{\min}$	–1	+1	–1	–1	+1	–1	+1	1	$\overline{y}_3$
4	$x_1^{\max}$	$x_2^{\max}$	$x_3^{\min}$	+1	+1	–1	+1	–1	–1	–1	1	$\overline{y}_4$
5	$x_1^{\min}$	$x_2^{\min}$	$x_3^{\max}$	–1	–1	+1	+1	–1	–1	+1	1	$\overline{y}_5$
6	$x_1^{\max}$	$x_2^{\min}$	$x_3^{\max}$	+1	–1	+1	–1	+1	–1	–1	1	$\overline{y}_6$
7	$x_1^{\min}$	$x_2^{\max}$	$x_3^{\max}$	–1	+1	+1	–1	–1	+1	–1	1	$\overline{y}_7$
8	$x_1^{\max}$	$x_2^{\max}$	$x_3^{\max}$	+1	+1	+1	+1	+1	+1	+1	1	$\overline{y}_8$

Коэффициенты модели (3.7) и (3.8) оцениваются независимо друг от друга по формулам:

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{i0}\,\overline{y}_i, \tag{3.9}$$

$$b_j = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{ij}\,\overline{y}_i,\ j = 1,...,k, \tag{3.10}$$

$$b_{ls} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{il}\tilde{x}_{is}\,\overline{y}_i,\ l < s;\ l = 1,...,k;\ \mathrm{s} = 1,...,k. \tag{3.11}$$

После вычисления коэффициентов оценивается их значимость для определения степени влияния факторов на функцию отклика. Проверка значимости осуществляется по критерию Стьюдента.

Для этого необходимо рассчитать дисперсии коэффициентов по формуле:

$$S^2_{\{b\}} = \frac{S^2_{\{y\}}}{n \cdot m}, \tag{3.12}$$

где $S^2_{\{y\}}$ — дисперсия воспроизводимости по всем проведенным экспериментам, найденная по формуле (3.3);

n — число экспериментов;

m — число опытов в каждом эксперименте.

Для линейной модели величина $S^2_{\{b\}}$ постоянна для всех коэффициентов модели b (b_0, b_j, b_{ls}).

Выдвигают гипотезы:

основная $H_0 : b = 0$ (коэффициент не значим);

альтернативная $H_1 : b \neq 0$ (коэффициент значим).

Для проверки гипотезы H_0 вычисляется наблюдаемое значение критерия:

$$T_{\,набл} = \frac{b}{\sqrt{S^2_{\{b\}}}}. \tag{3.13}$$

Критическая область является двусторонней. По таблице критических точек распределения Стьюдента (в Excel функция СТЬЮДРАСПОБР) определяется критическое значение критерия при выбранном уровне значимости ошибки $\alpha = 0,05$ или $\alpha = 0,01$ и числе степеней свободы $k = n(m-1)$: $t_{кр} = t_{кр}(\alpha, k)$.

Если $|T_{набл}| > t_{кр}$, то нулевая гипотеза отвергается. Это значит, что коэффициент статистически значим.

Если $|T_{набл}| < t_{кр}$, то нет оснований отвергнуть нулевую гипотезу. Это значит, что коэффициент незначимо отличается от нуля.

Если коэффициент модели статистически не значим, то влияние соответствующего фактора признается незначительным, и он исключается из модели. Статистическая незначимость коэффициента может быть обусловлена неправильным выбором интервала варьирования фактора, либо этот фактор не оказывает влияние на значение выходного показателя.

3.3 Проверка адекватности модели

Адекватность — это способность модели предсказывать результаты эксперимента в некоторой области с требуемой точностью. Адекватность математической модели подразумевает, что она достаточно верно качественно и количественно описывает свойства исследуемого явления.

Проверка на адекватность полученного уравнения регрессии (3.7) и (3.8) со значимыми коэффициентами осуществляется с помощью критерия Фишера. Расчетное значение F-критерия определяется по формуле:

$$F_{набл} = \frac{S^2_{ост}}{S^2_{\{y\}}}, \tag{3.14}$$

где $S^2_{\{y\}}$ — дисперсия воспроизводимости, найденная по формуле (3.3); $S^2_{ост}$ — остаточная дисперсия или дисперсия адекватности.

Остаточная дисперсия вычисляется по формуле:

$$S^2_{ост} = \frac{m}{n-r}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}_i\right)^2, \tag{3.15}$$

где n — число экспериментов;

m — число опытов в каждом эксперименте;

r — число значимых коэффициентов в уравнении регрессии;

$\hat{y}_i$ — значение функции отклика, расчитанное по уравнению модели со значимыми коэффициентами для i-го эксперимента;

$\overline{y}_i$ — среднее значение фукции отклика в i-м эксперименте.

Остаточная дисперсия представляет собой среднеквадратическую ошибку модели.

По таблице критических точек распределения Фишера-Снедекора (в Excel функция FРАСПОБР) определяется критическое значение критерия $F_{кр} = F(\alpha, k_1, k_2)$ при выбранном уровне значимости ошибки $\alpha = 0{,}05$ или $\alpha = 0{,}01$ и числе степеней свободы $k_1 = n - r$, $k_2 = n(m-1)$.

Если $F_{набл} < F_{кр}$, то уравнение регрессии признается адекватным, в противном случае — неадекватным.

Если модель нельзя считать адекватной, то необходимо либо выбирать более сложный вид уравнения связи входных факторов и функции отклика, либо, если это возможно, проводить эксперимент с меньшим интервалом варьирования влияющих факторов.

3.4 Анализ результатов моделирования

Анализ результатов предполагает интерпретацию полученной модели. Интерпретацию модели необходимо проводить в кодированных переменных. В этом случае величина каждого коэффициента уравнения функции отклика указывает на степень влияния соответствующего ему фактора на выходной показатель. Чем больше абсолютная величина коэффициента, тем больше фактор влияет на функцию отклика. Знак «плюс» у коэффициента означает, что с увеличением значения фактора растет величина отклика, знак «минус» — убывает.

Для получения математической модели в натуральных факторах в уравнение регрессии вместо $\tilde{x}_j$ необходимо подставить их выражение из формулы (3.6). При переходе к натуральным факторам коэффициенты уравнения изменяются и пропадает возможность оценивать величи-

ну влияния фактора на функцию отклика. Однако, если уравнение адекватно, то его можно использовать для расчета функции отклика по значениям факторов в натуральном масштабе.

3.5 Поиск экстремума функции отклика

Если модель признана адекватной, то ее можно использовать для управления процессом и его оптимизации путем движения по направлению к экстремуму функции отклика. Эта задача возникает при оптимизации производственных процессов, осуществляемых для улучшения свойств изделий, исследований предельных возможностей приборов и устройств и т. д.

Для поиска оптимальных значений функции отклика применяют следующие методы: градиентный, безградиентный и методы случайного поиска.

Наиболее широкое распространение получили градиентный метод крутого восхождения и безградиентный метод симплекс-планирования.

Рассмотрим подробно метод крутого восхождения (метод Бокса-Уилсона). Этот метод предусматривает стратегию последовательного проведения эксперимента. Известно, что изменение функции отклика будет наибольшим в направлении градиента функции, а оценки коэффициентов пропорциональны проекциям градиента на оси-факторы. Если факторы оценивать пропорционально значениям коэффициентов, то будет обеспечено движение вдоль градиента (линии крутого восхождения).

Алгоритм поиска экстремальных значений функции отклика методом крутого восхождения

1. Провести полный факторный эксперимент. Вычислить коэффициенты модели b_j.

2. Рассчитать произведение $b_j \Delta x_j$. Фактор, для которого это произведение наибольшее, принимается за базовый $b_б$.

3. Выбирают параметр шага движения к экстремуму: $\lambda = \dfrac{\mu}{|b_б|}$, μ принимает значение от 0 до 1.

4. Определяют шаг крутого восхождения для каждого фактора $\lambda \cdot (b_j \cdot \Delta x_j)$ и определяются новые точки плана.

5. Далее проводят «мысленные» опыты, которые заключаются в последовательном расчете значений функции отклика для значений факторов в новых точках плана. Для удобства расчетов используют уравнение регрессии в натуральных факторах.

6. Некоторые «мысленные» опыты реализуются для проверки соответствия исследуемому процессу.

Экстремум функции отклика достигается при выполнении условия: $y_{n-1} < y_n > y_{n+1}$ для максимума; $y_{n-1} > y_n < y_{n+1}$ для минимума; либо при выходе за диапазон варьирования факторов.

Контрольные вопросы

1. Что называется полным факторным экспериментом?
2. Как выбираются факторы планирования, их основные (базовые) уровни и интервалы варьирования?
3. Чем определяется величина интервала варьирования фактора?
4. Как выбрать центр плана эксперимента?
5. Указать порядок проведения эксперимента методом ПФЭ.
6. Как составляется матрица планирования ПФЭ?
7. Почему необходимо проведение параллельных опытов и их рандомизация?
8. В чем заключается смысл разработки математической модели по принципу «от простого – к сложному»?
9. Каков порядок статистической обработки и анализа результатов эксперимента?
10. При каких условиях не соблюдается требование воспроизводимости эксперимента и как следует поступить в этом случае?
11. Как проверить значимость оценок коэффициентов регрессии?
12. Как проверить адекватность математической модели?
13. При каких условиях не соблюдается требование адекватности математической модели и как следует поступить в этом случае?
14. Перечислите основные методы поиска оптимального плана.
15. Объясните сущность метода крутого восхождения.

4 СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРАКТИЧЕСКИХ ЗАДАЧ

Металлургические предприятия с точки зрения проектирования и управления техногенной безопасностью относятся к высшей категории сложности. Экологический мониторинг является составной и важной частью системы обеспечения экологической безопасности в зоне влияния металлургического производства. Для принятия правильных решений по охране окружающей среды необходим детальный анализ информации, получаемой из различных источников, в том числе сети государственного и производственного экологического мониторинга. Обработка и анализ регулярно получаемых данных наилучшим образом осуществляются методами математической статистики.

В данной главе рассмотрены типичные задачи, с решением которых сталкиваются при анализе экологических проблем, связанных в том числе с оценкой воздействия металлургического производства на окружающую среду.

4.1 Одномерный статистический анализ

Задача 1. Изучение влияния сезонного фактора на показатели качества воды в природных водоемах, в которые производится сброс сточных вод металлургического комбината.

Стоки предприятий металлургического комплекса выносят загрязняющие вещества в природные водные объекты. На концентрацию загрязняющих веществ в воде открытых водоемов может также влиять температурный режим и количество осадков, которые изменяются в течение года. Поэтому при определении степени антропогенного воздействия промышленных стоков конкретного предприятия необходимо учитывать влияние сезонного фактора.

Цель исследования. Используя математические методы, установить влияние сезонного фактора на концентрацию загрязняющего вещества в воде водоема.

Для осуществления поставленной цели необходимо решить следующие задачи:

1. По данным лаборатории контроля качества поверхностных вод сформировать две статистические выборки, содержащие концентрации загрязняющего вещества в воде в зимний и летний период.

2. Найти основные статистические характеристики выборок и оценить закон распределения показателя в течение каждого периода. Определить квантили уровня 0,25 и 0,75. Сделать выводы.

3. Сравнить средние значения показателя, используя соответствующие статистические гипотезы. Сделать вывод о влиянии сезонного фактора на средний уровень показателя.

Постановка задачи. Исследуется концентрация хлоридов в воде водоема, в который после некоторой очистки предприятие сбрасывает свои сточные воды.

Результаты контроля содержания хлоридов в воде в течение двух месяцев приведены ниже. Замеры велись ежедневно в течение января (признак X_1) и июля (признак X_2) месяца отчетного года.

X_1 (мг/дм^3): 93,2; 92,2; 89,7; 90,6; 88,3; 93,6; 91,7; 91,8; 91,5; 93,1; 89,7; 91,2; 88,4; 89,8; 90,1; 89,5; 87,8; 87,9; 91,5; 91,1; 93,1; 87,9; 89,7; 88,9; 90,1; 87,9; 90,7; 92,9; 92,4; 93,5; 92,3.

X_2 (мг/дм^3): 79,3; 79,5; 78,6; 78,6; 79,6; 78,6; 78,6; 73,6; 73,6; 78,6; 76,7; 76,7; 78,3; 78,2; 79,2; 78,3; 79,2; 81,2; 79,2; 83,2; 85,3; 81,2; 81,4; 81,4; 82,2; 79,4; 81,2; 80,2; 81,2; 81,3; 82,1.

Требуется проверить, подтверждают ли данные результаты предположение о влиянии сезонного фактора на концентрацию хлоридов в воде.

Решение поставленной задачи

Для выполнения статистических расчетов наиболее часто используют табличный процессор Excel. Это дает возможность намного упростить работу со статистическими характеристиками, которые рассчитываются по сложным формулам. В программе заложено множество групп формул, в том числе и статистических, также имеется возможность самостоятельно записать формулы.

Сначала необходимо создать файл Excel. Затем в столбцы электронной таблицы внести исходные данные по признакам X_1 и X_2 (каж-

дый признак в отдельном столбце). В названии столбцов указать названия признаков. Далее обозначить ячейки для результатов расчета. Расчет основных статистических характеристик осуществляется с помощью вставки функции f_x либо в пакете «Анализ данных/Описательная статистика».

В первом случае в меню «Формулы» выбирается «Вставить функцию» (f_x). При выборе этой опции появляется диалоговое окно (рис. 4.1). В диалоговом окне «Категория» указывается категория функций. В нашем случае обычно выбирается категория «Статистические».

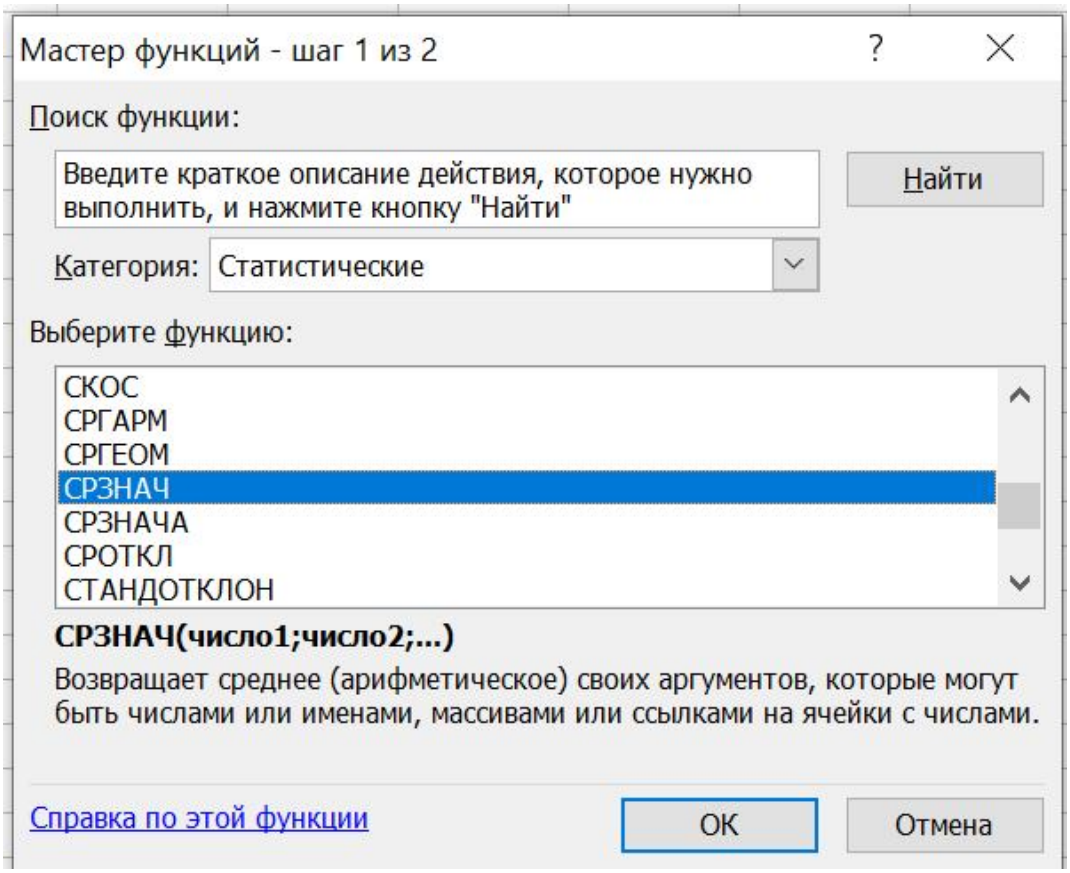


Рисунок 4.1 — Диалоговое окно для выбора функции

Ниже диалогового окна выводится список функций данной категории. Если отметить какую-то функцию в этом списке, то в нижней части окна появится форма обращения к указанной функции, а ниже приводится короткое объяснение функции на русском языке. Если выбрать нужную функцию (для этого надо после отметки функции щелкнуть мышкой по кнопке ОК), то откроется соответствующее диалоговое ок-

но. Далее необходимо ввести запрашиваемую информацию относительно массива (это значения признака X). Для этого достаточно указать мышкой на нужный массив (в массиве выделять все значения за один раз). После нажатия кнопки ОК результат вычисления записывается в ячейку, которая была выделена пользователем перед обращением к функции.

Некоторые наиболее известные статистические функции приведены в таблице 4.1. Более подробная справка по статистическим функциям, которые используются в табличном процессоре Excel, есть во встроенной справке.

Таблица 4.1 — Статистические функции в Excel

Название функции	Обозначение	Обращение в Excel
Выборочное среднее	$\overline{x}_в$	СРЗНАЧ
Выборочная дисперсия	$D_в$	ДИСПР
Исправленная выборочная дисперсия	S^2	ДИСП
Выборочное СКО	$\sigma_в$	СТАНДОТКЛОНП
Исправленное выборочное СКО	S^2	СТАНДОТКЛОН
Медиана	M_e	МЕДИАНА
Мода	M_o	МОДА
Коэффициент асимметрии	A_s	СКОС
Эксцесс	E_k	ЭКСЦЕСС

Для быстрого получения всех основных статистических характеристик используют специальный пакет «Анализ данных/Описательная статистика». Для установки этой надстройки необходимо в меню «Файл» выбрать: Параметры/Надстройки/Управление: надстройки Excel (перейти)/выбрать «Пакет анализа». После активации в меню «Данные» появляется вкладка «Анализ данных», в которой нужно выбрать категорию «Описательная статистика» (рис. 4.2). Появится окно (рис. 4.3), в котором нужно выставить необходимые позиции.

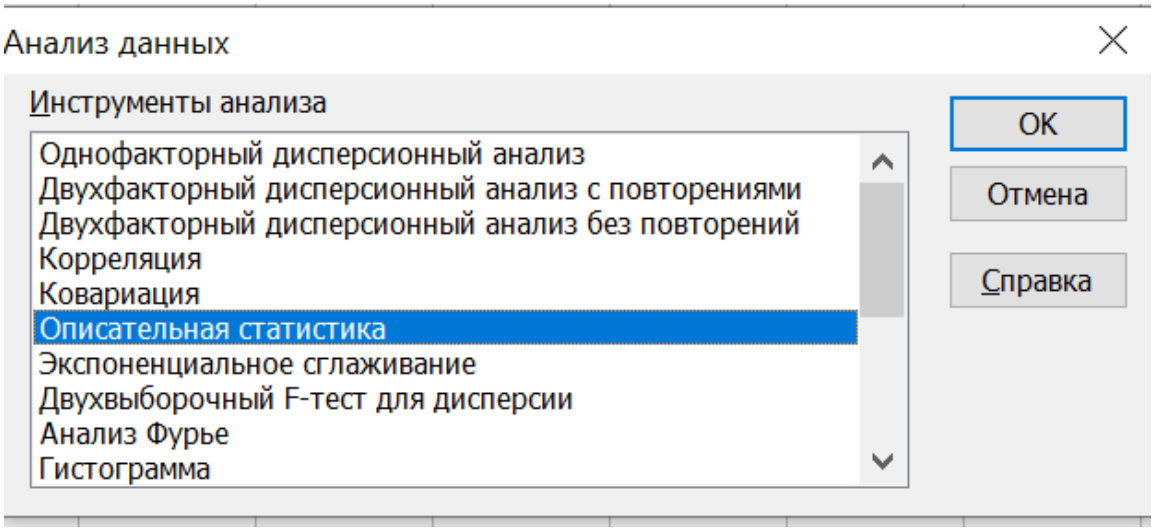


Рисунок 4.2 — Вкладка «Анализ данных»

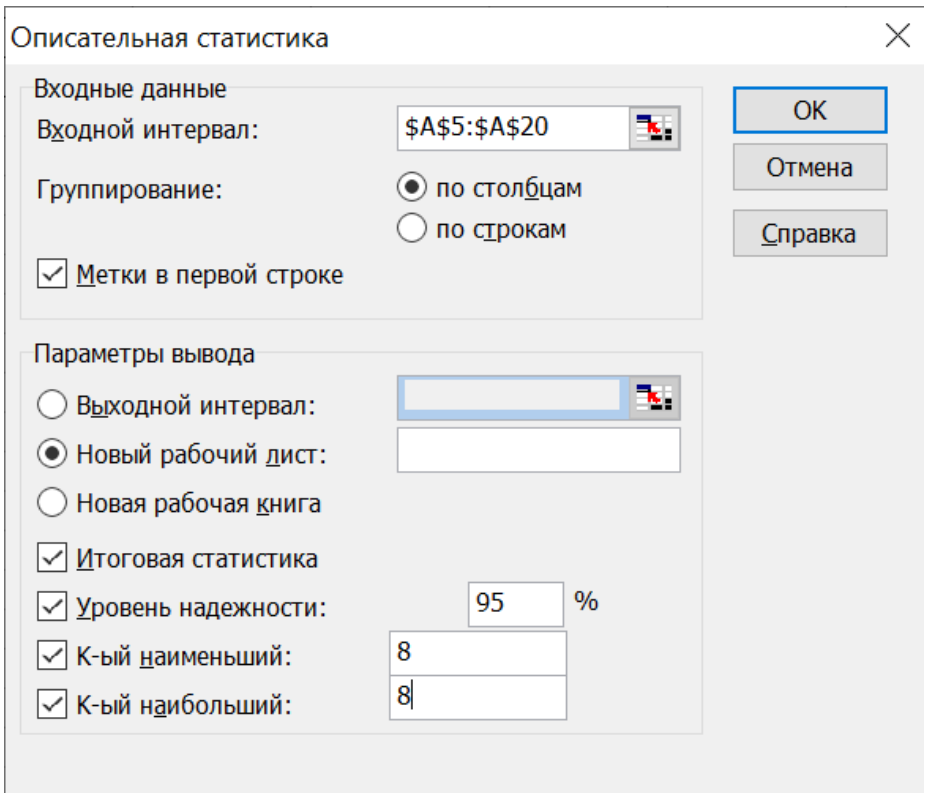


Рисунок 4.3 — Диалоговое окно «Описательная статистика»

Метка в первой строке выставляется, если исходный массив выбирается вместе с его обозначением в первой строке. После нажатия ОК на отдельном листе появляется таблица результатов, расшифровка которых приведена в таблице 4.2.

Таблица 4.2 — Расшифровка результатов работы надстройки «Анализ данных»

Обозначение в Excel	Обозначение математическое	Название статистической характеристики
Среднее	$\overline{x} = \overline{x}_в$	выборочное среднее
Стандартная ошибка	$\sigma_{\overline{x}} = \frac{S}{\sqrt{n}}$	стандартная ошибка среднего
Медиана	Me	медиана
Мода	Mo	мода
Стандартное отклонение	S	исправленное СКО
Дисперсия выборки	S^2	исправленная дисперсия
Эксцесс	Ек	эксцесс
Асимметричность	As	коэффициент асимметрии
Интервал	Δ= Xmax – Xmin	размах выборки
Минимум	Xmin	минимальное по выборке значение признака
Максимум	Xmax	максимальное по выборке значение признака
Сумма	Σx_i	сумма всех значений выборки
Счет	n	объем выборки
Наибольший (8)		квантиль 0,75
Наименьший (8)		квантиль 0,25
Уровень надежности (95.0%)	δ	точность определения генерального среднего с надежностью γ=0,95

Применяя пакет «Анализ данных» к решаемой задаче сразу к двум признакам X_1 и X_2, получим следующие результаты (табл. 4.3).

Таблица 4.3 — Основные статистические характеристики выборки

Статистическая характеристика	*Январь* *Признак* X_1	*Июль* *Признак* X_2
Среднее	90.71525	79.53871
Стандартная ошибка	0.331364	0.437605
Медиана	90.67678	79.3
Мода	#Н/Д*	78.6
Стандартное отклонение	1.844954	2.436483
Дисперсия выборки	3.403856	5.936452
Эксцесс	–1.18217	1.375957
Асимметричность	–0.09694	-0.39949
Интервал	5.760784	11.7
Минимум	87.80383	73.6
Максимум	93.56462	85.3
Сумма	2812.173	2465.7
Счет	31	31
Наибольший (8)	92.34	81.2
Наименьший (8)	89.53286	78.6
Уровень надежности (95.0%)	0.676735	0.893709

* — *распределение имеет несколько мод*

Анализ результатов. Среднее значение хлоридов в январе находилось в пределах $90,72 \pm 0,33$ мг/дм^3, а в июле — $79,54 \pm 0,44$ мг/дм^3. Но изменчивость показателя (стандартное отклонение) в июле на 25 % больше: 1,84 мг/дм^3 в январе и 2,44 мг/дм^3 в июле. Квантили в январе равны $x_{0,25}$=89,53 мг/дм^3 и $x_{0,75}$= 92,34 мг/дм^3, значит, в диапазоне (89,53–92,34) мг/дм^3 находится 50 % значений хлоридов. Следовательно, половина значений хлоридов в январе мало отличается от среднего, в отличие от июля. Таким образом, несмотря на то, что в июле содержание хлоридов в воде в среднем на 12 % меньше, этот показатель менее стабилен в теплое время года, т. е. может принимать и достаточно большие значения. Это косвенно подтверждается и коэффициентами

асимметрии. Асимметричность для двух выборок имеет незначительные значения, но отрицательные, что означает преобладание в выборках значений больше средних. Поэтому на этом этапе анализа нельзя сделать надежный вывод о различии в среднемесячной концентрации хлоридов в январе и июле отчетного года. Так как статистическая значимость расхождений средних в разные месяцы не доказана, то сравнение выборок должно быть продолжено — выполнено при помощи соответствующих статистических гипотез.

Для корректной проверки статистических гипотез требуется нормальный закон распределения признаков. Убедимся, что признаки X_1 и X_2 подчиняются нормальному закону распределения. Прежде всего, в пользу нормального закона свидетельствует, что медианные значения и мода незначительно отличаются от соответствующих средних значений.

Далее проверим выполнение условий (2.11). По формулам (2.12) при объеме выборки n=31:

$$D(A) = \frac{6(31-1)}{(31+1)(31+3)} = 0{,}165;$$

$$D(E) = \frac{24(31-2)(31-3)}{(31+1)^2(31+3)(31+5)} = 0{,}016.$$

Тогда для признака X_1: условие $|-0{,}09| \le 3 \cdot \sqrt{0{,}165}$, $0{,}09 \le 1{,}22$ выполняется; а условие $|-1{,}18| \le 3 \cdot \sqrt{0{,}016}$ не выполняется. Для признака X_2: условие $|-0{,}4| \le 3 \cdot \sqrt{0{,}165}$, $0{,}4 \le 1{,}22$ также выполняется; а условие $|1{,}38| \le 3 \cdot \sqrt{0{,}016}$ не выполняется. Таким образом, по этому критерию нельзя сделать однозначный вывод о том, что признак подчиняется нормальному закону.

Проверим выполнение правила трех сигм. Для признака X_1 в трехсигмовый интервал $(90{,}72 - 3 \cdot 1{,}84;\ 90{,}72 + 3 \cdot 1{,}84)$ или $(85{,}19;\ 96{,}23)$ попадают все значения признака, т.к. минимальное значение 87,80 мг/дм^3, а максимальное — 93,56 мг/дм^3. Для признака X_2 в трехсигмовый интервал $(72{,}23;\ 86{,}85)$ также попадает 100 % вариант. Следовательно, можно принять, что распределение концентрации

хлоридов в течение месяца подчиняется нормальному закону распределения.

Вопрос влияния сезона года на концентрацию хлоридов в воде сводится к проверке статистической гипотезы о равенстве двух средних (математических ожиданий) генеральных совокупностей. Для корректного решения необходимо убедиться в равенстве дисперсий двух генеральных совокупностей, из которых сделаны выборки X_1 и X_2. Для этого используют *F*-критерий Фишера.

Выдвинем основную и альтернативную гипотезы:

H_0: $D(X_2) = D(X_1)$;

H_1: $D(X_2) > D(X_1)$.

Для проверки гипотез по выборочным данным вычисляем наблюдаемое значение критерия (отношение большей дисперсии к меньшей):

$$F_{набл} = \frac{S_{x_2}^2}{S_{x_1}^2} = \frac{5,94}{3,40} = 1,75 .$$

Критическая область является правосторонней. Критическая точка находится по таблице критических точек распределения Фишера–Снедекора (в Excel функция FРАСПОБР(α; k_1; k_2)) при уровне значимости α=0,01, число степеней свободы большей дисперсии $k_1 = 31 - 1 = 30$, число степеней свободы меньшей дисперсии $k_2 = 31 - 1 = 30$: $F_{кр} = F(0,01;30;30) = 2,39$.

В результате сравнения получим $F_{набл} < F_{кр}$. Значит, нет оснований отвергнуть нулевую гипотезу H_0. Следовательно, принимаем гипотезу о равенстве дисперсий двух генеральных совокупностей.

Для выяснения влияния сезонного фактора на концентрацию хлоридов в воде проверим статистическую гипотезу о равенстве двух средних генеральных совокупностей. Используем критерий Стьюдента сравнения двух средних нормально распределенных генеральных совокупностей, дисперсии которых неизвестны и одинаковы.

Выдвинем основную и альтернативную гипотезы.

H_0: $M(X_1) = M(X_2)$;

H_1: $M(X_1) > M(X_2)$.

Принятие нулевой гипотезы H_0 дает основания считать, что сезонный фактор не приводит к изменению концентрации хлоридов. Принятие гипотезы H_1 будет означать, что в зимний период концентрация хлоридов в воде больше, чем в летний период.

Для проверки гипотез по результатам выборок вычисляем наблюдаемое значение критерия по формуле (2.14):

$$T_{набл} = \frac{90,72 - 79,54}{\sqrt{(31-1)\cdot 3,40 + (31-1)\cdot 5,94}} \cdot \sqrt{\frac{31\cdot 31\cdot (31+31-2)}{31+31}} = 20,36 \ .$$

Этот критерий является случайной величиной, которая подчиняется закону распределения Стьюдента с $k = 31+31-2 = 60$ степенями свободы.

Критическая область является правосторонней. Критическая точка находится по таблице критических точек распределения Стьюдента для односторонней критической области. В Excel выбираем функцию СТЬЮДРАСПОБР(2·0,01;60), получим: $t_{кр} = 2,39$.

В результате сравнения имеем: $T_{набл} > t_{кр}$. Значит, нулевую гипотезу H_0 отвергаем, принимаем H_1. Различия в средних значениях по двум выборкам статистически значимы. Следовательно, генеральные средние значения также различаются, т.е. сезонный фактор оказывает влияние на концентрацию хлоридов в воде: в зимний период концентрация выше.

4.2 Одномерная регрессия

Задача 2. Исследование изменения показателей качества воды в природных водоемах в зависимости от температурного фактора.

Качество воды определяется свойствами и концентрацией веществ, содержащихся в ней. Качество природной, питьевой и сточной воды зависит от конкретного содержания некоторых химических веществ, способных в концентрациях, превышающих предельно допустимые (ПДК), ухудшать органолептические и физико-химические свойства воды. Такими веществами, например, являются сульфаты, хлориды, железо, марганец и др. Концентрация этих веществ в воде подвер-

жена влиянию временных и климатических факторов, а также зависит от наличия и мощности в данном и соседних регионах промышленных, сельскохозяйственных и других предприятий. Температура воды влияет на протекание в ней физических, химических, биохимических и биологических процессов, на содержание растворенного кислорода и в конечном итоге на интенсивность процессов самоочищения. Температуру следует учитывать при проектировании и расчете многих очистных сооружений. Так, биологическая очистка сточных вод практически не идет при температуре ниже 10 °С. Для поверхностных и сточных вод характерны значительные изменения температуры, связанные с климатом и сезоном года.

Поэтому исследование влияния температуры на показатели качества воды является актуальной задачей при планировании экологических мероприятий.

Цель исследования. Построить математическую модель для прогноза влияния температурного фактора на концентрацию загрязняющего вещества в воде поверхностного водоема.

Для осуществления поставленной цели необходимо решить следующие задачи:

1. По данным лаборатории контроля качества воды сформировать две статистические выборки, содержащие концентрации загрязняющего вещества в воде и температуру воды в момент отбора проб.

2. Провести корреляционный анализ показателей. Сделать вывод о взаимосвязи показателей.

3. Построить линейную и нелинейные модели одномерной регрессии. Провести сравнительный анализ различных уравнений регрессии по коэффициенту детерминации, корреляционному отношению и средней относительной погрешности аппроксимации ε. Обосновать выбор лучшего уравнения регрессии и выполнить прогноз.

Постановка задачи. Получены результаты анализа проб воды, исследуемой на присутствие хлоридов в водоеме, при различных температурных режимах (табл. 4.4).

Таблица 4.4 — Результаты анализа проб воды

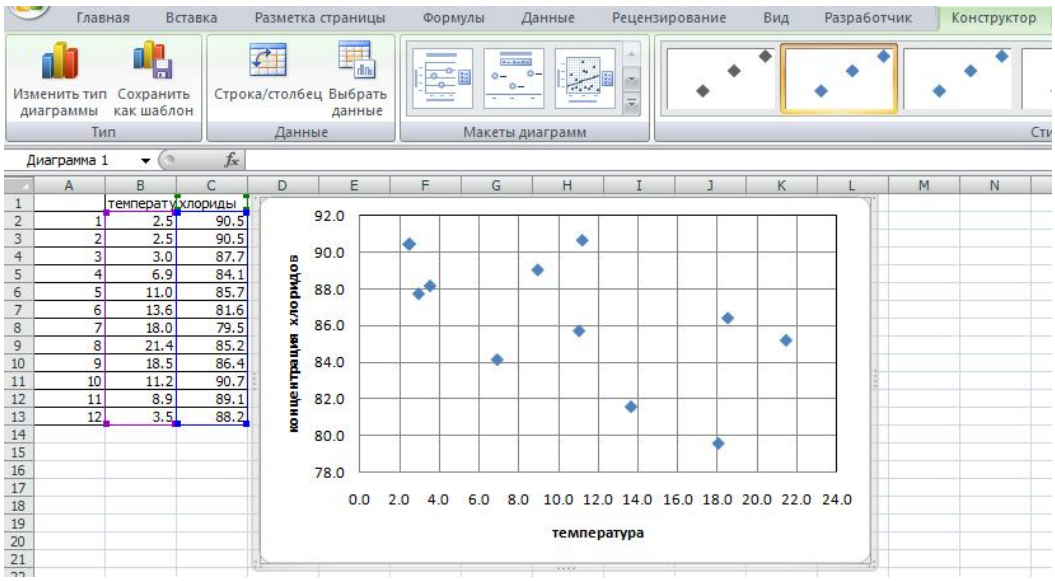
№ пробы	Температура воды, °С	Концентрация хлоридов, мг/дм^3
1	2,5	90,5
2	2,5	90,5
3	3,0	87,7
4	6,9	84,1
5	11,0	85,7
6	13,6	81,6
7	18,0	79,5
8	21,4	85,2
9	18,5	86,4
10	11,2	90,7
11	8,9	89,1
12	3,5	88,2

Требуется установить зависимость концентрации хлоридов от температуры воды водоема.

Решение поставленной задачи

Обозначим x — температура воды (независимый фактор); y — концентрация хлоридов (зависимый фактор).

Для визуального представления зависимости y от x используется корреляционное поле. Для этого выделяется массив исходных данных (x, y) и с помощью меню Вставка/Диаграмма/Точечная/Точечная с маркерами строится точечная диаграмма (рис. 4.4). Чтобы подписать названия осей необходимо при активной диаграмме (выделена пользователем) выбрать Конструктор/Макет диаграмм 1. Далее подписать оси на диаграмме.



	температура	хлориды
1	2.5	90.5
2	2.5	90.5
3	3.0	87.7
4	6.9	84.1
5	11.0	85.7
6	13.6	81.6
7	18.0	79.5
8	21.4	85.2
9	18.5	86.4
10	11.2	90.7
11	8.9	89.1
12	3.5	88.2

Рисунок 4.4 — Построение точечной диаграммы

Из вида корреляционного поля следует, что с ростом температуры концентрация хлоридов в среднем уменьшается. Для более обоснованного вывода найдем коэффициент корреляции, используя меню Вставка — f_x/КОРРЕЛ (рис. 4.5). В данном случае получим: $r = -0,61$, знак минус означает обратную зависимость между x и y. Этот результат был получен по выборочным данным. Чтобы распространить его на генеральную совокупность, необходимо доказать значимость найденного коэффициента корреляции.

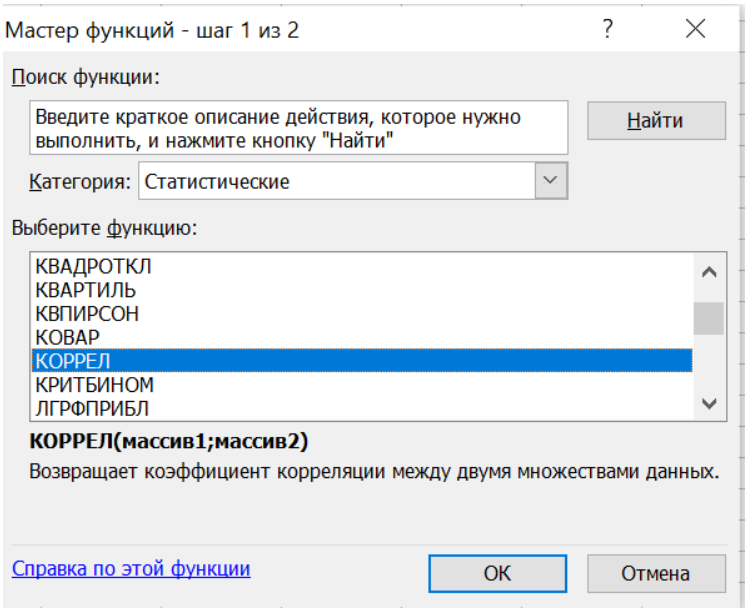


Рисунок 4.5 — Диалоговое окно для выбора функции

Для проверки статистической значимости коэффициента корреляции используют критерий Стьюдента. Выдвигают гипотезы:

основная $H_0 : r_г = 0$ (коэффициент корреляции не значим);

альтернативная $H_1 : r_г \neq 0$ (коэффициент корреляции значим).

Для проверки гипотезы H_0 вычисляется наблюдаемое значение критерия:

$$T_{набл} = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}} = \frac{-0,61\sqrt{12-2}}{\sqrt{1-(-0,61)^2}} = -2,40.$$

По таблице критических точек распределения Стьюдента (в Excel функция СТЬЮДРАСПОБР) определяется критическое значение критерия при выбранном уровне значимости ошибки $\alpha = 0,05$ и числе степеней свободы $k = 12-2 = 10$: $t_{кр} = 2,23$.

Так как $|T_{набл}| > t_{кр}$, то нулевая гипотеза отвергается. Это значит, что коэффициент корреляции статистически значим. Следовательно, между концентрацией хлоридов и температурой есть линейная связь. По шкале Чеддока (табл. 2.1) делаем вывод, что между концентрацией хлоридов в воде и температурой есть заметная связь.

Чтобы найти аналитическое выражение связи, используют уравнение регрессии. Для получения на диаграмме уравнения регрессии и коэффициента детерминации R^2, следует выделить полученную диаграмму; в строке меню выбрать Макет/Анализ/Линия тренда/Дополнительные параметры линии тренда. Появится диалоговое окно Формат линии тренда (рис. 4.6).

Далее нужно отметить параметры линии тренда. Например, когда надо построить параболическую зависимость, то выбирают модель «Полиномиальная» и выставляют степень 2. Затем ниже указывают опции:

- ✓ показывать уравнение на диаграмме;
- ✓ поместить на диаграмму величину достоверности аппроксимации (R^2) (коэффициент детерминации R^2).

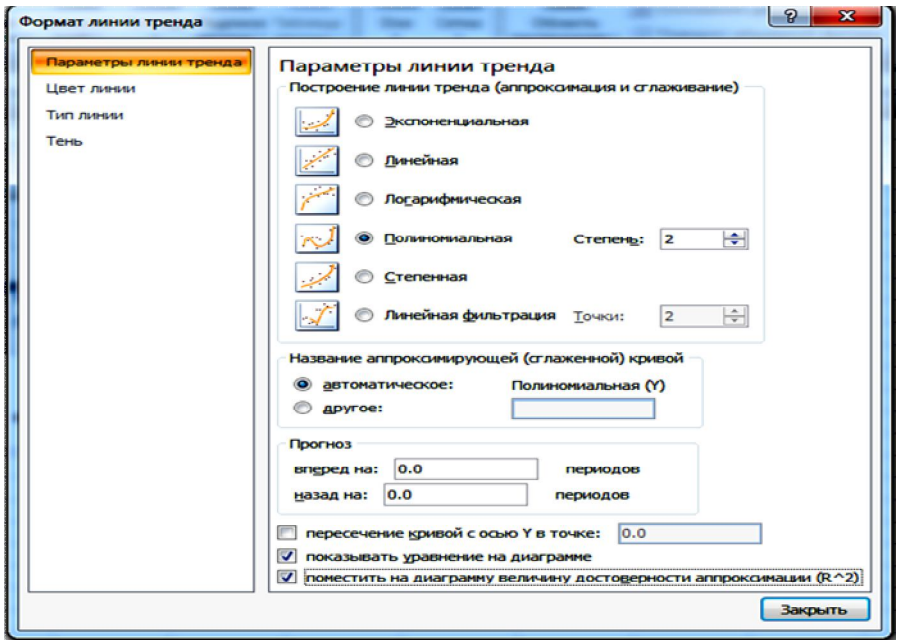


Рисунок 4.6 — Построение линии тренда

После выбора опции «Закрыть» появится необходимая информация и диаграмма примет требуемый вид. Необходимо построить все основные виды зависимостей: экспоненциальную, линейную, логарифмическую, полиномиальную, степенную. Затем выбрать одно уравнение, наиболее адекватно описывающее зависимость y от x. Выбирается уравнение с наибольшим значением коэффициента детерминации R^2. Если уравнения имеют близкие коэффициенты R^2, выбирают более простую. В данном случае, была выбрана линейная модель (рис. 4.7).

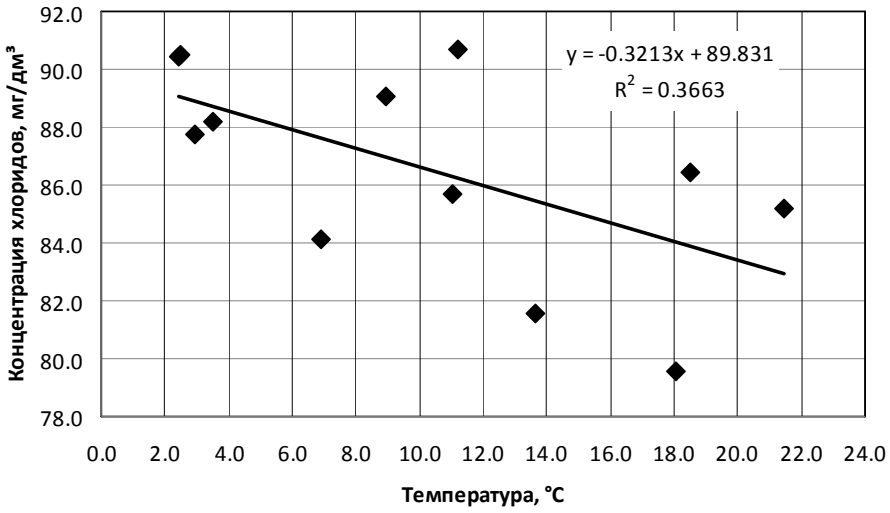


Рисунок 4.7 — Зависимость концентрации хлоридов от температуры воды

Сделаем прогноз концентрации хлоридов при температуре воды 15 °С:

$$\overline{y} = -0,32 \cdot 15 + 89,83 = 85,01 (\text{мг/дм}^3).$$

Точность прогноза по модели оценивается с помощью средней квадратической ошибки уравнения (стандартной ошибки оценки) $S_{уравн}$ и средней относительной ошибки аппроксимации (средней относительной погрешности модели), которые находятся по формулам (2.21) и (2.22). Для удобства расчета заполняют специальную таблицу (табл. 4.5).

Таблица 4.5 — Расчет ошибок уравнения линейной регрессии $\overline{y} = -0,32 \cdot x + 89,83$

i	x_i	y_i	$y_{теор}$	$y_i - y_{теор}$	ε	$(y_i - y_{теор})^2$
1	2,5	90,5	89,044	1,416	0,016	2,006
2	2,5	90,5	89,028	1,442	0,016	2,080
3	3,0	87,7	88,883	–1,143	0,013	1,307
4	6,9	84,1	87,614	–3,484	0,041	12,138
5	11,0	85,7	86,290	–0,580	0,007	0,337
6	13,6	81,6	85,452	–3,902	0,048	15,223
7	18,0	79,5	84,038	–4,498	0,057	20,232
8	21,4	85,2	82,946	2,244	0,026	5,038
9	18,5	86,4	83,887	2,523	0,029	6,366
10	11,2	90,7	86,239	4,431	0,049	19,635
11	8,9	89,1	86,959	2,091	0,023	4,374
12	3,5	88,2	88,700	–0,530	0,006	0,281
Сумма					0,331	89,016
Ошибка					2,76	2,98

Следовательно, при температуре 15 °С средняя концентрация хлоридов ожидается $85 \pm 2,98$ мг/дм^3. Отклонение от прогнозного значения может быть в среднем 2,76 %.

4.3 Множественная регрессия

Задача 3. Моделирование процессов загрязнения окружающей среды от автомобильного транспорта.

Современное состояние окружающей среды в городах зависит от развития не только промышленности, но и автотранспорта, эксплуатация которого приводит к значительному загрязнению атмосферного воздуха, в частности, таким веществом, как оксид углерода (CO). По данным исследований, социально-экологическая шкала оценки влияния оксида углерода на человека имеет следующий вид (табл. 4.6):

Таблица 4.6 — Шкала оценки влияния оксида углерода

Степень влияния CO	Содержание CO в 1 $м^3$ воздуха, $см^3$	Содержание CO, $мг/м^3$
Легкая	0–5	0,0–6,25
Слабая	5–10	6,25–12,5
Заметная	10–20	12,5–25,0
Значительная	20–30	25,0–37,5
Серьезная	30–40	37,5–50
Очень серьезная	40–50	50,0–62,5
Угрожающая	50–60	62,5–75,0
Опасная	Больше 60	Больше 75

Цель исследования. Получить количественные оценки и построить математическую модель, описывающую зависимость содержания CO в атмосферном воздухе от характеристик движения автотранспорта, параметров дороги и уличных строений вдоль автомобильной дороги.

Для осуществления поставленной цели необходимо решить следующие задачи:

1. Используя информацию экологической службы города собрать статистическую базу для оценки воздействия автотранспорта на состояние городской среды.

2. Провести корреляционный анализ показателей. Сделать вывод о взаимосвязи показателей.

3. Построить модель множественной регрессии. Оценить точность и качество полученной модели.

4. Сделать прогноз при некоторых данных, соответствующих конкретным условиям.

Постановка задачи. Для количественной оценки концентрации оксида углерода на примагистральной территории собраны статистические данные по следующим входным факторам и выходным показателям (табл. 4.7):

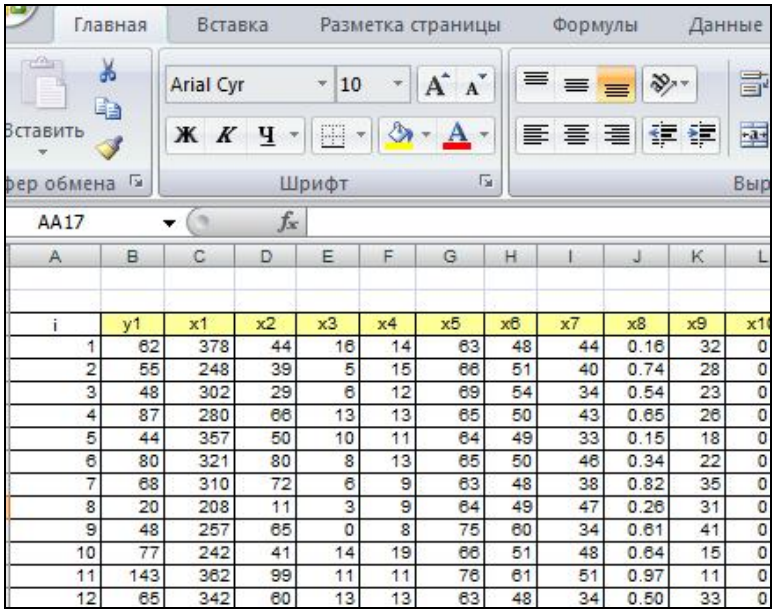
Таблица 4.7 — Исследуемые факторы

Фактор	Наименование	Единица измерения
$У_1$	Концентрация оксида углерода	мг/м^3
X_1	Общая интенсивность движения автотранспортных потоков	автомобиль/час
X_2	Доля грузовых автомобилей и автобусов в общем потоке	%
X_3	Продольный уклон проезжей части	градусы
X_4	Количество этажей уличного строения	
X_5	Ширина улицы возле строения	м
X_6	Ширина проезжей части	м
X_7	Средневзвешенная скорость движения автомобилей в потоке	км/час
X_8	Линейная плотность уличного строения	
X_9	Температурный показатель (T+26)	°C
X_{10}	Коэффициент неоднородности состава автотранспортных потоков	

Требуется установить зависимость концентрации оксида углерода от влияющих факторов.

Решение поставленной задачи

На первом этапе необходимо сформировать базу данных в Excel (рис. 4.8, показана часть базы).



i	y1	x1	x2	x3	x4	x5	x6	x7	x8	x9	x1
1	62	378	44	16	14	63	48	44	0.16	32	0
2	55	248	39	5	15	66	51	40	0.74	28	0
3	48	302	29	6	12	69	54	34	0.54	23	0
4	87	280	66	13	13	65	50	43	0.65	26	0
5	44	357	50	10	11	64	49	33	0.15	18	0
6	80	321	80	8	13	65	50	46	0.34	22	0
7	68	310	72	6	9	63	48	38	0.82	35	0
8	20	208	11	3	9	64	49	47	0.26	31	0
9	48	257	65	0	8	75	60	34	0.61	41	0
10	77	242	41	14	19	66	51	48	0.64	15	0
11	143	362	99	11	11	76	61	51	0.97	11	0
12	65	342	60	13	13	63	48	34	0.50	33	0

Рисунок 4.8 — База статистических данных

Для выяснения взаимосвязи между факторами необходимо получить корреляционную таблицу. Для этого используем меню Данные / Анализ данных (рис. 4.9).

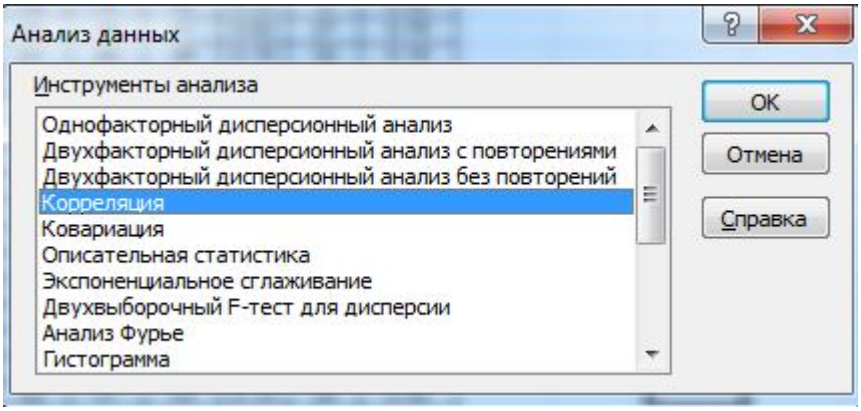


Рисунок 4.9 — Вкладка «Анализ данных»

В появившемся окне (рис. 4.9) выбираем «Корреляция». Появляется окно, изображенное ниже (рис. 4.10). В него необходимо внести:

– входной интервал — диапазон, который занимают все данные с заголовками;

– поставить галочку в «Метки в первой строке»;
– в параметрах ввода выбрать «Новый рабочий лист».

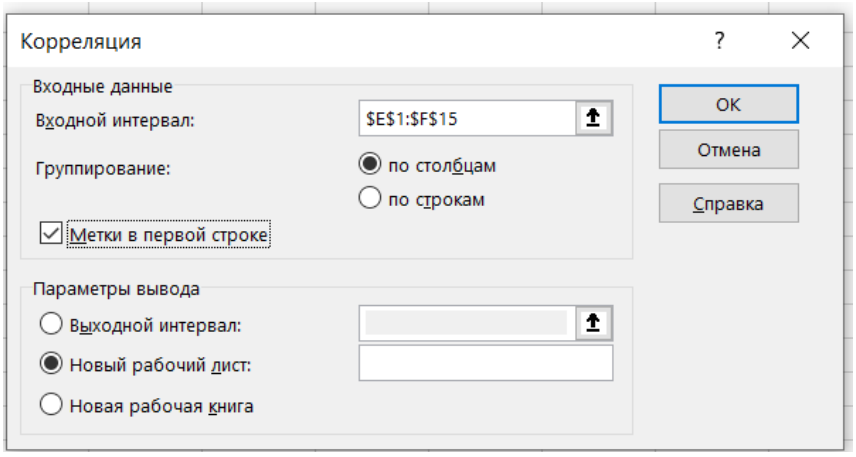


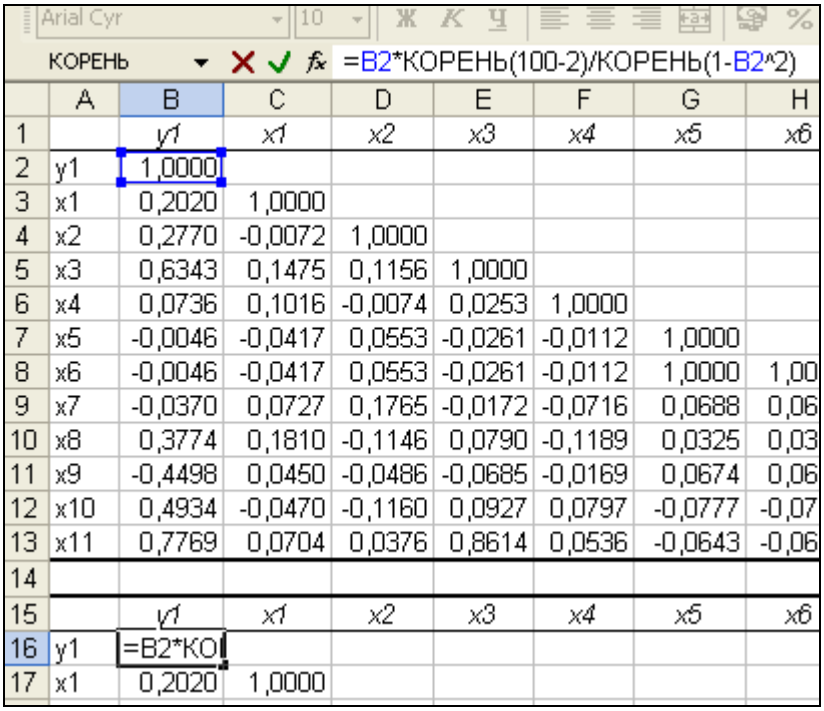
Рисунок 4.10 — Диалоговое окно «Корреляция»

На новом листе появится корреляционная матрица (табл. 4.8).

Таблица 4.8 — Корреляционная матрица

	y1	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*
y1	1,0000											
x1	0,2020	1,0000										
x2	0,2770	-0,0072	1,0000									
x3	0,6343	0,1475	0,1156	1,0000								
x4	0,0736	0,1016	-0,0074	0,0253	1,0000							
x5	-0,0046	-0,0417	0,0553	-0,0261	-0,0112	1,0000						
x6	-0,0046	-0,0417	0,0553	-0,0261	-0,0112	1,0000	1,0000					
x7	-0,0370	0,0727	0,1765	-0,0172	-0,0716	0,0688	0,0688	1,0000				
x8	0,3774	0,1810	-0,1146	0,0790	-0,1189	0,0325	0,0325	0,0016	1,0000			
x9	-0,4498	0,0450	-0,0486	-0,0685	-0,0169	0,0674	0,0674	-0,0825	0,1584	1,0000		
x10	0,4934	-0,0470	-0,1160	0,0927	0,0797	-0,0777	-0,0777	-0,1332	0,0812	-0,2390	1,0000	
x11	0,7769	0,0704	0,0376	0,8614	0,0536	-0,0643	-0,0643	-0,0805	0,1113	-0,1641	0,5652	1,0000

Каждый коэффициент таблицы необходимо проверить на значимость по критерию Стьюдента. Для этого надо найти $T_{набл}$, $t_{кр}$ и сравнить их. Если $|T_{набл}| > t_{кр}$, то коэффициент корреляции значимый. Для удобства выполнения этого задания делают копию корреляционной таблицы и заполняют ее следующим образом: в первой ячейке записывают формулу для определения $T_{набл}$ (2.20) и далее используют автозаполнение (рис. 4.11).



КОРЕНЬ | =B2*КОРЕНЬ(100-2)/КОРЕНЬ(1-B2^2)

	A	B	C	D	E	F	G	H
1		*y1*	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*
2	y1	1,0000						
3	x1	0,2020	1,0000					
4	x2	0,2770	-0,0072	1,0000				
5	x3	0,6343	0,1475	0,1156	1,0000			
6	x4	0,0736	0,1016	-0,0074	0,0253	1,0000		
7	x5	-0,0046	-0,0417	0,0553	-0,0261	-0,0112	1,0000	
8	x6	-0,0046	-0,0417	0,0553	-0,0261	-0,0112	1,0000	1,00
9	x7	-0,0370	0,0727	0,1765	-0,0172	-0,0716	0,0688	0,06
10	x8	0,3774	0,1810	-0,1146	0,0790	-0,1189	0,0325	0,03
11	x9	-0,4498	0,0450	-0,0486	-0,0685	-0,0169	0,0674	0,06
12	x10	0,4934	-0,0470	-0,1160	0,0927	0,0797	-0,0777	-0,07
13	x11	0,7769	0,0704	0,0376	0,8614	0,0536	-0,0643	-0,06
14								
15		*y1*	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*
16	y1	=B2*КО						
17	x1	0,2020	1,0000					

Рисунок 4.11 — Диалоговое окно «Корреляция»

Для определения $t_{кр}$ используют меню Вставка f_x / СТЬЮДРАСПОБР (рис. 4.12).

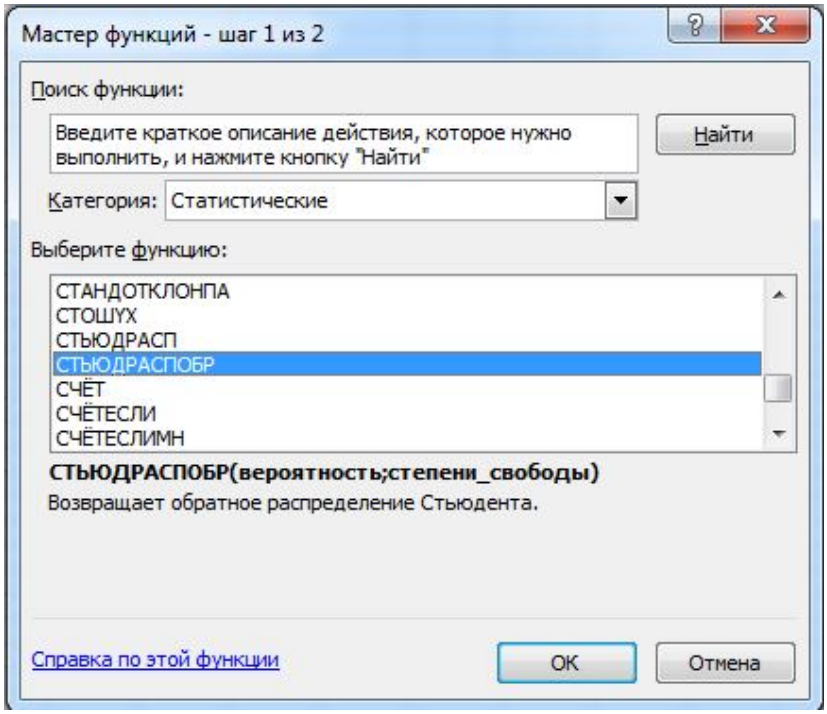


Рисунок 4.12 —Диалоговое окно для выбора функции

Заполняют параметры: вероятность (0,05 или 0,1 —допустимая ошибка), степень свободы: $n-2$, где n — объем выборки (рис. 4.13).

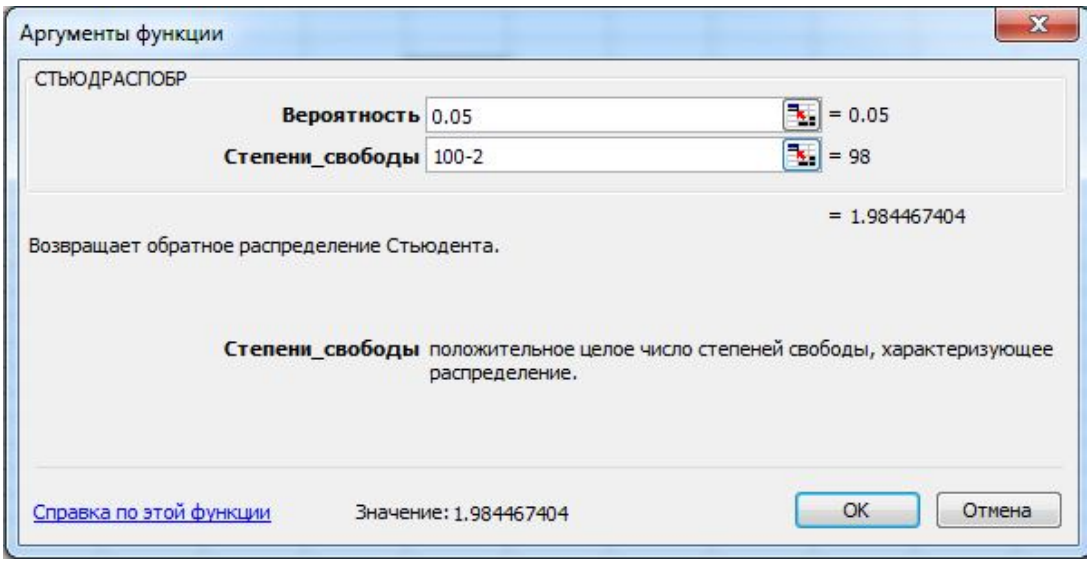


Рисунок 4.13 — Диалоговое окно «Аргумент функции»

Далее сравнивают рассчитанные в таблице (рис. 4.11) значения критерия Стьюдента с $t_{кр}$ и выделяют значимые связи.

На основе таблицы 4.8 строят граф связи, на котором отмечают значимые связи между признаками (рис. 4.14).

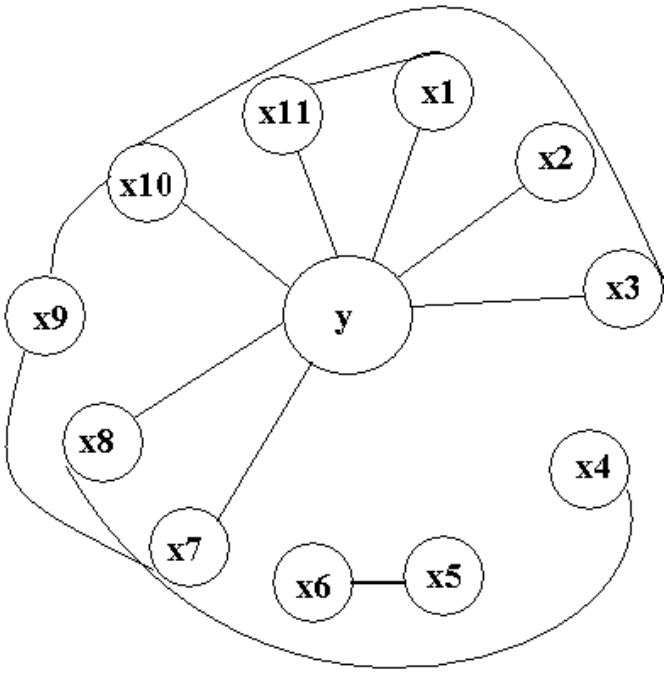


Рисунок 4.14 — Граф связей

Для составления модели отбирают те факторы, которые имеют большое влияние на y, но не связаны между собой. Возможные варианты моделей (исходя из вышеприведенного графа):

– модель 1: $y = f_1(x_1, x_2, x_3, x_7, x_8)$;

– модель 2: $y = f_2(x_2, x_7, x_8, x_{10}, x_{11})$;

– модель 3: $y = f_3(x_2, x_3, x_7, x_{11})$.

Для определения параметров модели необходимо на новый лист скопировать факторы, составляющие выбранную модель.

Используя меню Данные /Анализ данных, выбирают категорию «Регрессия» (рис. 4.15) и заполняют открывшееся окно (рис. 4.16).

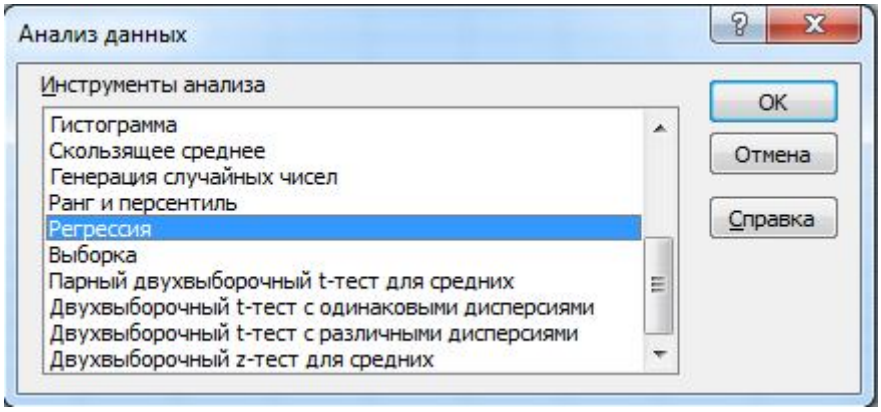


Рисунок 4.15 — Вкладка «Анализ данных»

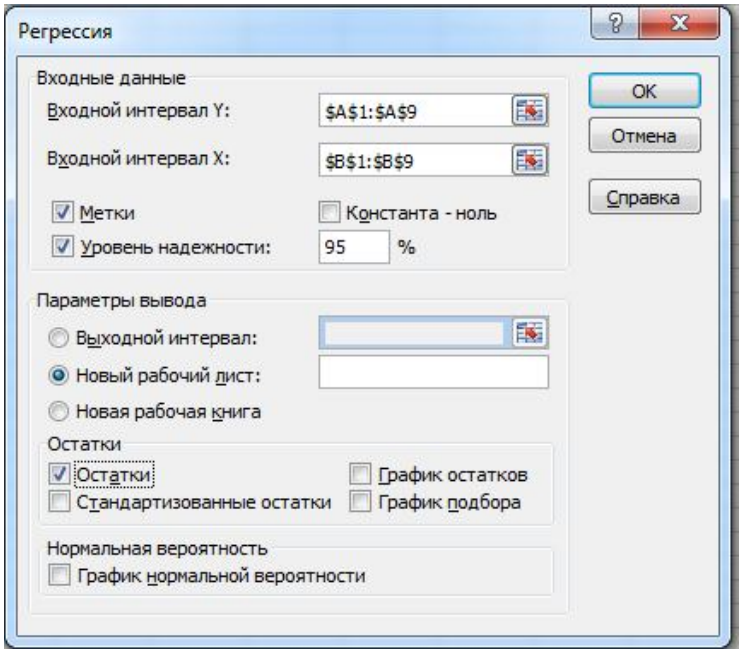


Рисунок 4.16 — Диалоговое окно «Регрессия»

В результате данных действий получится четыре таблицы (табл. 4.9–4.12).

Таблица 4.9 — Результат регрессии

Регрессионная статистика		
Множественный R	0,678187268	R
R-квадрат	0,45993797	R^2
Нормированный R-квадрат	0,431211266	нормированный R^2
Стандартная ошибка	14,47535722	S
Наблюдения	100	n

Таблица 4.10 — Результат регрессии

Дисперсионный анализ				
	df	*SS*	*MS*	*F*
Регрессия	5	16774,20912	3354,841824	16,01081607
Остаток	94	19696,38088	209,5359668	
Итого	99	36470,59		

Таблица 4.11 — Результат регрессии

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение (a0)	-7,5612	17,1602	-0,4406	0,6605
x1(a1)	0,0275	0,0190	1,4443	0,1520
x2(a2)	0,4372	0,1597	2,7388	0,0074
x3(a3)	2,2895	0,2983	7,6747	0,0000
x7(a7)	0,3176	0,4945	0,6422	0,5223
x8(a8)	0,0296	0,5079	0,0583	0,9536
	для уравнения	стандартное отклонение	для проверки на значимость	вероятность

Таблица 4.12 — Результат регрессии*

ВЫВОД ОСТАТКА				
Наблюдение	*Предсказанное y1*	*Остатки*	*y1*	*abs*
1	73,54547871	-9,545478714	64	14,91481049
2	54,49930565	4,500694345	59	7,628295501
3	61,7971047	-21,7971047	40	54,49276175
4	53,13046256	5,869537444	59	9,948368548
5	29,64355853	-1,643558535	28	5,86985191
6	38,68073428	-11,68073428	27	43,26197883
7	51,88133092	-3,881330918	48	8,08610608

* Представлено начало таблицы, которая выдается программой.

Таблица 4.12 используется для определения средней относительной ошибки аппроксимации (средней относительной погрешности модели) ε по формуле (2.22). Для этого к таблице добавляют 4-ый и 5-ый столбцы. В 4-ый копируют y, а в 5-ом выполняют расчет.

На основе (табл. 4.11) записывают уравнение модели 1:

$$\overline{y} = -7,5612 + 0,0275 \cdot x_1 + 0,4372 \cdot x_2 + 2,2895 \cdot x_3 + 0,3176 \cdot x_7 + 0,0296 \cdot x_8$$

Аналогично строятся модель 2 и модель 3. Далее модели сравниваются и по лучшей делается прогноз. Для удобства сравнения заполняют таблицу 4.13.

Таблица 4.13 — Сравнительный анализ моделей множественной регрессии

Параметры	Модель 1	Модель 2	Модель 3
Множественный коэффициент корреляции R	**0,678**	0,667	0,679
Коэффициент детерминации R^2	**0,460**	0,430	0,452
Нормированный R^2	0,431	0,402	**0,503**
Стандартная ошибка	**14,475**	15,342	15,322

В сравнении учитывается (в табл. 4.13 выделено жирным шрифтом):

- чем R^2 больше, тем модель лучше;
- чем R больше, тем модель лучше;
- чем нормированный R^2 больше, тем модель лучше;
- чем меньше стандартная ошибка, тем модель лучше;

– чем средняя относительная погрешность модели ε меньше, тем лучше.

По этим показателям лучшей является модель 1.

Далее осуществляют прогноз по лучшей модели при следующих значениях влияющих факторов: $x_1 = 35$, $x_2 = 62$, $x_3 = 35$, $x_7 = 13$, $x_8 = 0{,}8$:

$$\overline{y} = -7{,}5612 + 0{,}0275 \cdot 35 + 0{,}4372 \cdot 62 + 2{,}2895 \cdot 35 +$$

$$+0{,}3176 \cdot 13 + 0{,}0296 \cdot 0{,}8 = 72 \frac{мг}{м^3}.$$

Значит, при заданных условиях содержание CO в воздухе в среднем составит 72 мг/м^3. Такая концентрация оксида углерода в воздухе представляет большую угрозу для людей, находящихся длительное время в зоне воздействия автотранспорта. Поэтому необходимо разработать комплекс природоохранных мероприятий по снижению негативного воздействия автотранспорта в рассмотренной ситуации.

4.4 Анализ временных рядов

Задача 4. Исследование динамики показателей загрязнения окружающей среды в городе.

Цель исследования. Построить модель изменения во времени показателя, характеризующего загрязнение окружающей среды.

Для осуществления поставленной цели необходимо решить следующие задачи:

1. Построить график исходного временного ряда и сформулировать гипотезы о характере изменений уровней и тенденции ряда.

2. С помощью спектрального анализа выявить существование цикличности в изменении показателя.

4. Построить автокорреляционную функцию для выяснения структуры ряда.

5. Найти тренд ряда методами сглаживания (скользящего среднего, текущего среднего, экспоненциального сглаживания).

6. Методом сезонной декомпозиции построить математическую модель временного ряда.

Постановка задачи. Концентрация пыли в приземном слое атмосферного воздуха (мг/м^3), определяемая путем регулярного отбора проб

на наблюдательном посту с последующим лабораторным анализом, представлена в виде дискретного временного ряда за четыре года от начала наблюдений (номер месяца):

Номер месяца	Год				
	2017	2018	2019	2020	2021
1	0,13	0,26	0,32	0,44	0,38
2	0,18	0,22	0,29	0,43	0,43
3	0,27	0,32	0,34	0,38	0,39
4	0,36	0,34	0,35	0,45	
5	0,30	0,35	0,39	0,48	
6	0,34	0,37	0,32	0,40	
7	0,35	0,35	0,40	0,57	
8	0,33	0,32	0,38	0,4	
9	0,31	0,35	0,39	0,35	
10	0,32	0,30	0,36	0,37	
11	0,26	0,26	0,26	0,4	
12	0,27	0,31	0,31	0,38	

Требуется определить тенденцию ряда и дать прогноз концентрации пыли на два месяца вперед (апрель и май 2021 года).

Решение поставленной задачи

На отдельном листе Excel необходимо создать таблицу «Исходные данные» и заполнить ее данными, расположив их в один ряд по временной оси t. Для создания модели используем данные за 2017–2020 годы (48 значений за каждый месяц), которые записываем в таблицу 4.14. Данные за 2021 год оставляем для проверки модели на последнем этапе.

Таблица 4.14 — Исходные данные

t	x
1	0,13
2	0,18
…	…
48	0,38

По данным таблицы «Исходные данные» построим диаграмму и добавим линию линейного тренда. В результате получим фактический график концентрации пыли во времени (рис. 4.17).

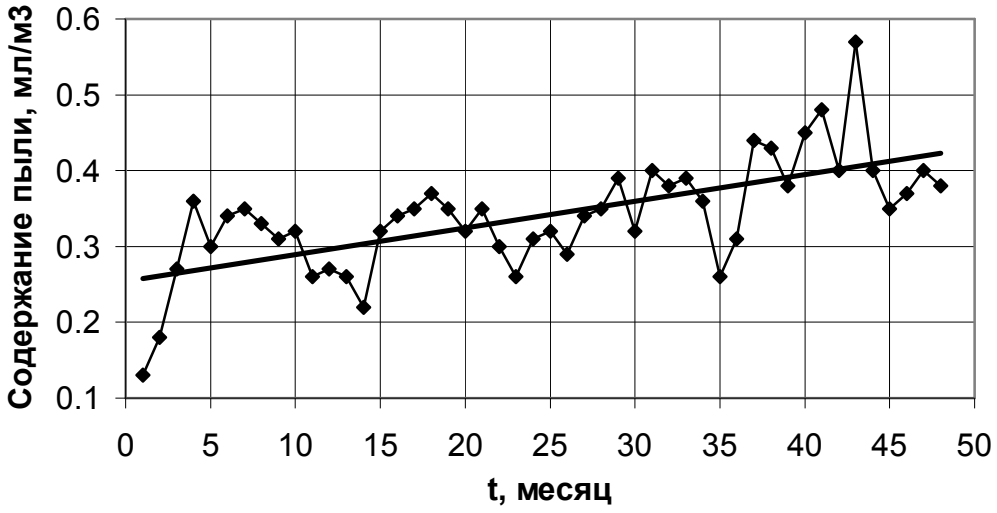


Рисунок 4.17 — График временного ряда

Используя типичные графики временных рядов (рис. 2.6–2.11) как эталоны, можно сделать вывод о структуре исследуемого ряда. Визуальный анализ графика содержания пыли в воздухе на рисунке 4.17 позволяет предположить наличие возрастающего тренда и сезонной циклической компоненты периода 10–12 месяцев.

Для проверки степени зависимости между членами временного ряда используют ряды со смещением (лагом) (табл. 4.15).

Таблица 4.15 — Исходный ряд и смещенный на k лагов*

t	X_t	X_{t-1}	X_{t-2}	X_{t-3}	X_{t-4}	X_{t-5}	X_{t-6}	X_{t-7}	X_{t-8}
1	0.13	-							
2	0.18	0.13	-						
3	0.27	0.18	0.13	-					
4	0.36	0.27	0.18	0.13	-				
5	0.30	0.36	0.27	0.18	0.13	-			
6	0.34	0.30	0.36	0.27	0.18	0.13	-		
7	0.35	0.34	0.3	0.36	0.27	0.18	0.13	-	
8	0.33	0.35	0.34	0.30	0.36	0.27	0.18	0.10	-

Продолжение таблицы 4.15

t	X_t	X_{t-1}	X_{t-2}	X_{t-3}	X_{t-4}	X_{t-5}	X_{t-6}	X_{t-7}	X_{t-8}
9	0.31	0.33	0.35	0.34	0.3	0.36	0.27	0.2	0.13
10	0.32	0.31	0.33	0.35	0.34	0.30	0.36	0.3	0.18
11	0.26	0.32	0.31	0.33	0.35	0.34	0.30	0.40	0.27
12	0.27	0.26	0.32	0.31	0.33	0.35	0.34	0.3	0.36
13	0.26	0.27	0.26	0.32	0.31	0.33	0.35	0.30	0.30
14	0.22	0.26	0.27	0.26	0.32	0.31	0.33	0.40	0.34
15	0.32	0.22	0.26	0.27	0.26	0.32	0.31	0.30	0.35

* Данные представлены в неполном объеме.

По данным таблицы 4.15 найдем коэффициенты корреляции между исходным рядом и смещенными рядами. Для нахождения коэффициента корреляции в Excel используется мастер функций $f(x)$ / Статистические / КОРРЕЛ (рис. 4.18).

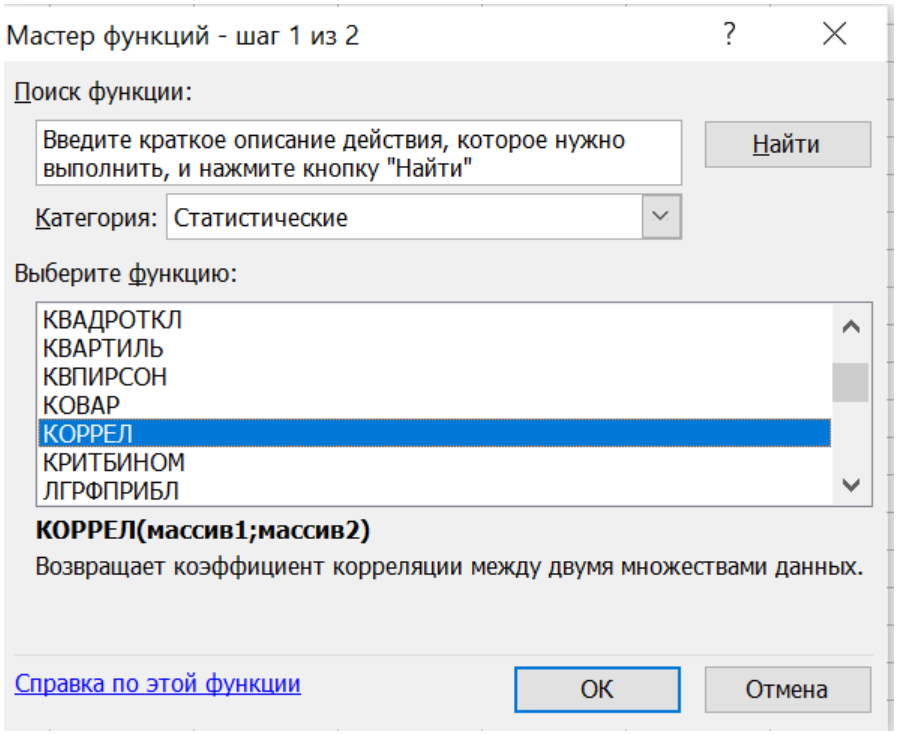


Рисунок 4.18 — Диалоговое окно для выбора функции

Для правильного вычисления коэффициента необходимо внимательно заполнить диапазоны массивов: выбирать одинаковое количество значений. Вот одна из вероятных ошибок (рис. 4.19):

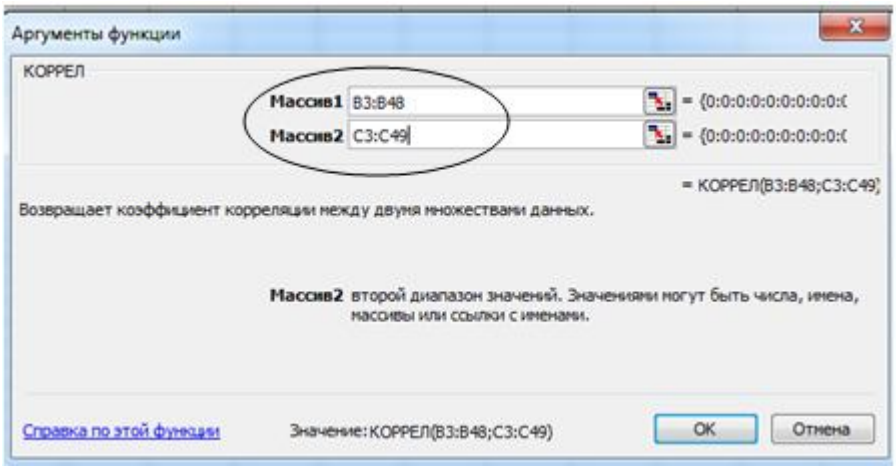


Рисунок 4.19 — Ошибка при задании массивов

После заполняем таблицу 4.16.

Таблица 4.16 — Коэффициенты корреляции и проверка их значимости

k	r	$T_{набл}$	$T_{крит}$
1	0,6377822	5,616146	2,012896
2	0,4406694	3,293084	2,014103
3	0,3896763	2,806684	2,015368
4	0,2954485	2,027914	2,016692
5	0,1814605	1,195852	2,018082
6	0,2220192	1,458005	2,019541
7	0,1342419	0,856775	2,021075
8	0,1688621	1,083538	2,021075
9	0,4445984	3,059726	2,024394
10	0,5379663	3,881912	2,026192
11	0,5300023	3,750033	2,028094
12	0,6632533	5,243033	2,030108
13	0,4615587	3,033815	2,032244

Полученные коэффициенты корреляции на каждом лаге отмечаются на графике, который называется коррелограммой (рис. 4.20).

По графику (рис. 4.20) видно, что наибольшие значения имеют коэффициенты корреляции при величине лага $k = 1$ и $k = 12$. Значит, модель имеет тренд и циклическую компоненту.

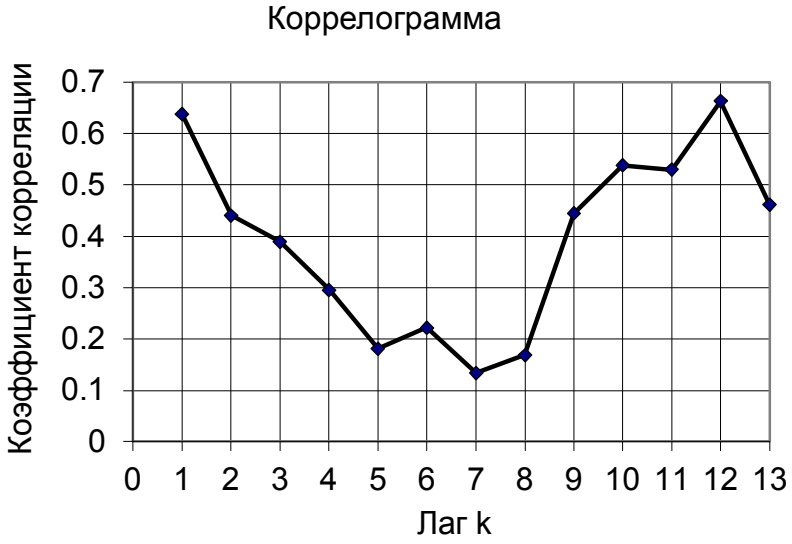


Рисунок 4.20 — График коррелограммы

Построим аддитивную модель исходного временного ряда методом сезонной декомпозиции: $X = T + S + E$.

На основе данных коррелограммы и учитывая естественную сезонность содержания пыли в приземном слое воздуха, принимаем период 12 месяцев.

Алгоритм расчета параметров модели:

1. Расчет сезонной компоненты проведем в таблице 4.17.

Таблица 4.17 — Расчет сезонной компоненты

месяц	t	Xt	скол.ср.	центр.ср.	Сезонная компонента
1	1	0,13	-		
2	2	0,18	-		
3	3	0,27	-		
4	4	0,36	-		
5	5	0,3	-		
6	6	0,34	0,285		
7	7	0,35	0,29583333	0,290417	0,059583333
8	8	0,33	0,29916667	0,2975	0,0325
• • •	•••	•••	•••	• • •	• • •
2	38	0,43	0,40583333	0,405	0,025
3	39	0,38	0,4025	0,404167	-0,024166667
4	40	0,45	0,40333333	0,402917	0,047083333
5	41	0,48	0,415	0,409167	0,070833333
• • •	•••	•••	•••	• • •	• • •
12	48	0,38	-		

1-й столбец — номер месяца (1...12, 1...12, 1...12...);

2-й столбец — дискретное время;

3-й столбец — концентрация пыли в приземном слое воздуха;

4-й столбец — скользящее среднее — значение 0,285 среднее арифметическое первых 12 значений, далее автозаполнение до 6 строки с конца;

5-й столбец — центрированное среднее — значение 0,290411 среднее арифметическое двух чисел 0,285 и 0,29583333, далее автозаполнение;

6-й столбец — сезонная компонента находится как разность 3-го и 5-го столбцов.

После этого необходимо распределить значения сезонной компоненты по годам (табл. 4.18).

Таблица 4.18 — Корректировка сезонной компоненты

Месяц	2017	2018	2019	2020	Среднее значение	Скорректированное значение
1		-0,048333	-0,0070833	0,042917	-0,004166667	-0,005648148
2		-0,087917	-0,0416667	0,025	-0,034861111	-0,036342593
3		0,010833	0,00416667	-0,02417	-0,003055556	-0,004537037
4		0,03	0,01	0,047083	0,029027778	0,027546296
5		0,040833	0,0475	0,070833	0,053055556	0,051574074
6		0,059167	-0,0225		0,018333333	0,016851852
7	0,06	0,035	0,0525		0,049027778	0,047546296
8	0,033	-0,000417	0,02166667		0,017916667	0,016435185
9	0,009	0,025833	0,02416667		0,019583333	0,018101852
10	0,018	-0,025417	-0,0116667		-0,006527778	-0,008009259
11	-0,04	-0,0675	-0,1195833		-0,076944444	-0,078425926
12	-0,04	-0,017083	-0,0766667		-0,043611111	-0,045092593
				сумма	0,017777778	0,000000
				среднее	0,001481481	

Далее в 6-ой столбец заносим среднее значений за каждый месяц и потом суммируем, должно быть 0. Если полученное число отлично от 0 (0,01777778), то необходимо поделить его на период 12, получится 0,0014814. Столбец 7 дает скорректированные средние сезонные значения, заполняется следующим образом: из 6 столбца вычитается поправ-

ка 0,0014814. Суммируем значения по месяцам, при правильном расчете сумма должна быть равна 0.

Таким образом, окончательно сезонная компонента представляется в виде столбца с 12-ю значениями и соответствующим графиком (рис. 4.21).

месяц	сезонная компонента
январь	-0,005648
февраль	-0,036342
март	-0,00453
апрель	0,027546
май	0,051574
июнь	0,016851
июль	0,047546
август	0,016435
сентябрь	0,018101
октябрь	-0,008009
ноябрь	-0,078425
декабрь	-0,045092

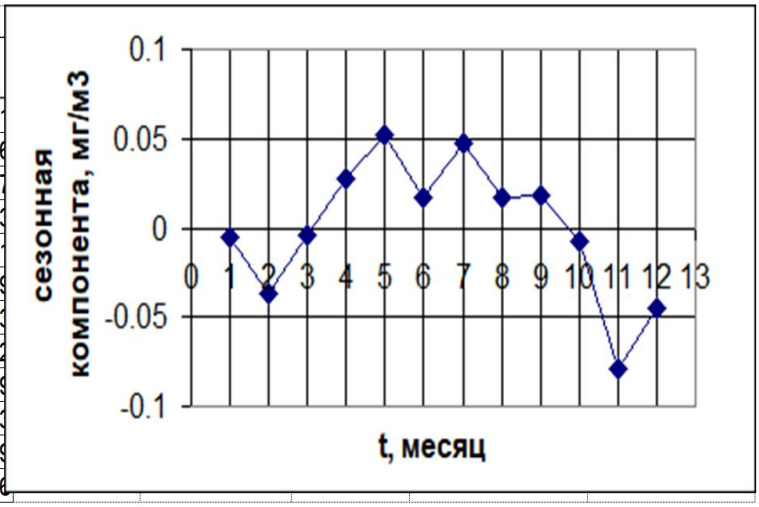


Рисунок 4.21 — Сезонная компонента

Для нахождения тренда заполняем таблицу 4.19.

Таблица 4.19 — Расчет тренда

	Нахождение тренда					
t	Xt	скоректир знач.	T=Xt-S	T=0,1878*t^0,1956	X^	остатки
1	0,13	-0,0056481	0,135648	0,1878	0,182151852	-0,052151852
2	0,18	-0,0363426	0,216343	0,215068623	0,178726031	0,001273969
3	0,27	-0,004537	0,274537	0,232820118	0,228283081	0,041716919
4	0,36	0,0275463	0,332454	0,24629666	0,273842956	0,086157044
5	0,3	0,05157407	0,248426	0,257284795	0,308858869	-0,008858869
6	0,34	0,01685185	0,323148	0,266625678	0,28347753	0,05652247
7	0,35	0,0475463	0,302454	0,274787368	0,322333664	0,027666336
8	0,33	0,01643519	0,313565	0,282059018	0,298494203	0,031505797
9	0,31	0,01810185	0,291898	0,288632627	0,306734479	0,003265521
...	...	...	...	...	...	...
46	0,37	-0,0080093	0,378009	0,397128467	0,389119208	-0,019119208
47	0,4	-0,0784259	0,478426	0,398802552	0,320376626	0,079623374
48	0,38	-0,0450926	0,425093	0,400448226	0,355355633	0,024644367

1-й столбец — дискретное время;

2-й столбец — концентрация пыли в приземном слое воздуха;

3-й столбец — сезонная компонента;

4-й столбец — значение тренда, определяется как разность фактических значений концентрации пыли и сезонной компоненты. После расчета этого столбца строят корреляционное поле по данным 1-го и 4-го столбцов. Затем на графике подбирают подходящее уравнение регрессии. В данном примере лучшее уравнение регрессии — степенная функция $T=0{,}1878t^{0.1956}$;

5-й столбец — расчетное значение тренда, найденное по уравнению регрессии $T=0{,}1878t^{0.1956}$;

6-й столбец — расчетное значение концентрации пыли, вычисляется как сумма сезонной компоненты и тренда (3-й и 5-й столбцы).

7-й столбец — остатки, разница между фактическими (2-й столбец) и расчетными (6-й столбец) значениями.

Сравним фактический и расчетный графики концентрации пыли (рис. 4.22).

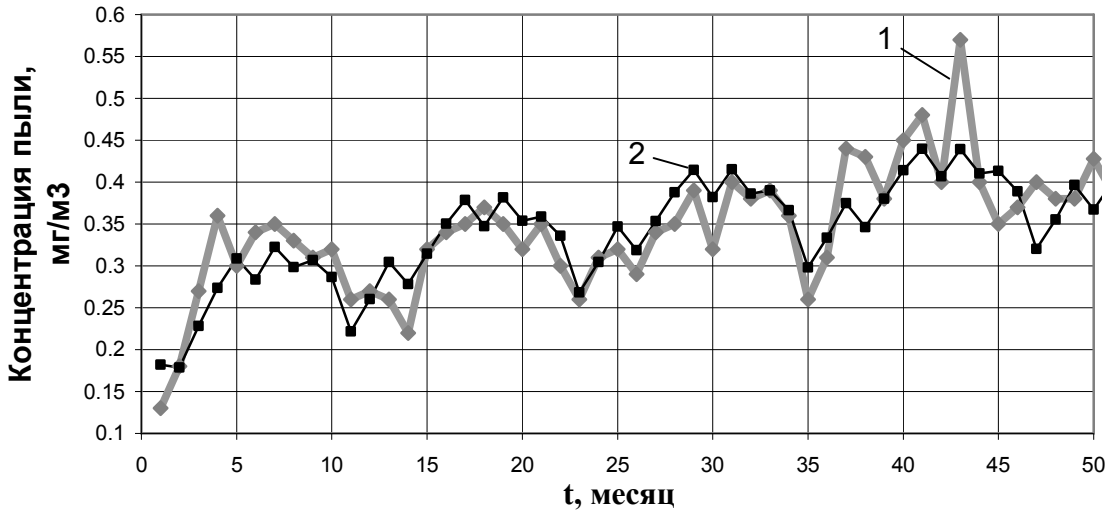


Рисунок 4.22 — Графики концентрации пыли (1 — фактический временной ряд, 2 — расчетные значения)

Графики рисунка 4.22 имеют некоторые отличия, хотя в целом характер поведения совпадает. Значит, исходный ряд содержит не только тренд (T) и сезонную компоненту (S), но и случайную компоненту (E).

Запишем в окончательном виде математическую модель содержания пыли в приземном слое воздуха:

$$X = 0{,}1878 \cdot t^{0{,}1956} + S + E,$$

где сезонная компонента задается соответствующими значениями (рис. 4.21).

Для прогноза случайную компоненту не учитываем. В таблице 4.20 приведены прогнозные значения на пять месяцев вперед, а также фактические значения показателя. В последнем столбце приведен расчет средней относительной погрешности по формуле (2.22).

Таблица 4.20 — Прогноз по модели временного ряда

Месяц	t	S	$\hat{T}$	$\hat{x}_t$	x_t	$\frac{\lvert x_t - \hat{x}_t \rvert}{x_t} \cdot 100$
1	49	-0,0056	0,4021	0,3964	0,380	4,3206
2	50	-0,0363	0,4037	0,3673	0,428	14,1785
3	51	-0,0045	0,4052	0,4007	0,387	3,5370
4	52	0,0275	0,4068	0,4343		
5	53	0,0516	0,4083	0,4599		
					сумма	22,0361
					ε	7,34

Из таблицы 4.20 следует, что прогнозные значения концентрации пыли в атмосферном воздухе в апреле и мае 2021 года составляют соответственно 0,434 и 0,460 мг/м^3. Средняя относительная погрешность модели равна 7,34 %, что свидетельствует согласно таблице 2.2 о хорошей точности модели.

4.5 Полный факторный эксперимент

Задача 5. Исследование эффективности процесса очистки воды от нитратов в зависимости от изменения факторов работы ионообменной колонки.

Загрязнение природных вод в результате антропогенной деятельности и техногенного воздействия предприятий наносит непоправимый ущерб экосистеме и делает воды малопригодными для водохозяйственного использования. Присутствие в питьевой воде азотсодержащих соединений (аммоний, нитраты, нитриты) приводит к заболеванию водо-

роднитратной метгемоглобинемией и способствует развитию различных степеней кислородного голодания организма.

Поэтому исследование методов физико-химической очистки природных вод является актуальной задачей.

Метод ионного обмена является перспективным методом очистки вод. Ионообменные фильтры обеспечивают максимальный уровень очистки. Они могут применяться не только для подготовки питьевой воды, но и для очистки промышленных стоков. Прочие методы не способны обеспечить достаточный уровень очистки.

Ионный обмен — это специфический случай сорбции заряженных частиц (ионов), когда поглощение одних ионов сопровождается выходом в раствор других ионов, входящих в состав сорбента. При этом ион, присутствие которого в воде нежелательно, фиксируется на сорбенте. Таким образом происходит «замещение» одних ионов на другие.

Сорбенты, работающие по такому механизму, называются ионообменными материалами или ионитами. Иониты способны извлекать из воды одни растворенные соли, замещая их другими солями (например, соли кальция и магния могут заменяться на соли натрия).

Чаще всего в процессе водоочистки ионный обмен используется для удаления из воды катионов тяжелых металлов (например, свинца), представляющих опасность для здоровья человека, а также для избавления от нитратов.

Водоподготовку путем ионного обмена выполняют при помощи специальных фильтрующих устройств (ионообменных колонок) — сначала их заполняют ионитами, а потом запускают воду.

Для улучшения характеристик фильтрующего устройства, работающего по принципу ионного обмена, необходимо экспериментальным путем изучить зависимость эффективности водоочистки от различных технологических и конструктивных факторов и затем подобрать оптимальные параметры. На рисунке 4.23 представлена экспериментальная установка для проведения серии опытов.

Рисунок 4.23 — Экспериментальная установка для исследования очистки воды от нитратов путем ионного обмена

При выполнении исследований рассматривались наиболее эффективные иониты следующих марок: «Purolite NRW600(OH)», «АВ-17-8чС», «LewatitMonoPlus®SR7».

Постановка задачи планирования эксперимента

Задача 5А. Трехфакторный эксперимент

Исследовалась эффективность процесса очистки воды от нитратов (Э, %) в зависимости от изменения следующих факторов работы ионообменной колонки:

– скорости фильтрации воды V, м/час;

– отношения высоты загрузки фильтрационной колонки к ее диаметру h/d;

– температуры очищаемой воды, t, °С.

Поскольку процесс определения оптимальных параметров ионообменной очистки является многофакторным, с целью сокращения затрат времени и материальных средств на выполнение исследований необходимо применить математический метод планирования эксперимента.

Факторы варьировались в пределах, описанных в таблице 4.21.

Таблица 4.21 — Параметры варьируемых факторов

Фактор	Единицы измерения	Нижняя граница	Верхняя граница
Скорость фильтрации воды	м/час	10	15
Отношение высоты загрузки фильтрационной колонки к ее диаметру		2	8
Температура очищаемой воды	°С	5	25

Цель исследования. Используя математический метод планирования эксперимента, разработать модель оценки эффективности процесса очистки воды от нитратов путем ионного обмена.

Для осуществления поставленной цели необходимо решить следующие задачи:

1. Определить уровни и интервалы варьирования факторов.
2. Составить исходную матрицу планирования и провести опыты согласно матрице планирования, результаты записать в таблицу.
3. Составить математическую модель.
4. Вычислить коэффициенты модели.
5. Определить дисперсию воспроизводимости и проверить значимость коэффициентов модели.
6. Проверить адекватность модели.
7. Записать уравнение оценки эффективности процесса очистки воды в натуральных факторах.
8. Провести интерпретацию полученной модели.

Решение поставленной задачи

1. Определим уровни и интервалы варьирования факторов. Для этого введем обозначения:

x_1 — скорость фильтрации воды (V, м/час), $x_1^{\min} = 10$, $x_1^{\max} = 15$;

x_2 — отношение высоты загрузки фильтрационной колонки к ее диаметру (h/d); $x_2^{\min} = 2$, $x_2^{\max} = 8$;

x_3 — температура очищаемой воды, (t, °C), $x_3^{\min} = 5$, $x_3^{\max} = 25$;

отклик y — эффективность процесса очистки воды от нитратов, (Э, %).

Основной уровень (центр плана) определяем по формуле (3.4), получим:

$$x_1^0 = \frac{15+10}{2} = 12,5;\ x_2^0 = \frac{8+2}{2} = 5,0;\ x_3^0 = \frac{25+5}{2} = 15,0.$$

Интервалы варьирования определяем по формуле (3.5):

$$\Delta x_1 = \frac{15-10}{2} = 2,5;\ \Delta x_2 = \frac{8-2}{2} = 3,0;\ \Delta x_3 = \frac{25-5}{2} = 10,0.$$

Безразмерные координаты по формуле (3.6):

$$\tilde{x}_1^{\,1} = \frac{10-12,5}{2,5} = -1;\ \tilde{x}_1^{\,2} = \frac{15-12,5}{2,5} = 1;$$

$$\tilde{x}_2^{\,1} = \frac{2-5,0}{3,0} = -1;\ \tilde{x}_2^{\,2} = \frac{8-5,0}{3,0} = 1;$$

$$\tilde{x}_3^{\,1} = \frac{5-15,0}{10,0} = -1;\ \tilde{x}_3^{\,2} = \frac{25-15,0}{10,0} = 1.$$

В данной задаче три фактора (k=3) на двух уровнях (p=2). Следовательно, общее число экспериментов $N = 2^3 = 8$. Коды и интервалы варьирования факторов запишем в таблицу 4.22.

Таблица 4.22 — Коды и интервалы варьирования факторов

Фактор	Единицы измерения	Обозначения	Диапазон варьирования	
			код = –1	код = 1
V	м/час	x_1	10	15
h/d	–	x_2	2	8
t	°C	x_3	5	25

2. Составим матрицы планирования трехфакторного эксперимента данной задачи (табл. 4.23 и 4.24).

Таблица 4.23 — Исходная матрица планирования для проведения опытов

№ эксперимента	Значения факторов						Результаты опытов		
	в натуральном масштабе			в безразмерном масштабе					
	x_1	x_2	x_3	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	y_1	y_2	y_3
1	10	2	5	-1	-1	-1	49,39	52,12	51,99
2	15	2	5	+1	-1	-1	56,71	54,06	55,82
3	10	8	5	-1	+1	-1	52,88	55,98	55,68
4	15	8	5	+1	+1	-1	55,53	59,61	58,4
5	10	2	25	-1	-1	+1	78,03	80,03	79,99
6	15	2	25	+1	-1	+1	79,52	78,61	77,68
7	10	8	25	-1	+1	+1	81,01	84,4	83,67
8	15	8	25	+1	+1	+1	85,69	84,48	82,69

Таблица 4.24 — План-матрица эксперимента 2^3

№ эксперимента	Значения факторов						Совместное влияние				Вспомогательный столбец	Отклик (среднее результатов опытов)
	в натуральном масштабе			в безразмерном масштабе								
	x_1	x_2	x_3	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_1\tilde{x}_2$	$\tilde{x}_1\tilde{x}_3$	$\tilde{x}_2\tilde{x}_3$	$\tilde{x}_1\tilde{x}_2\tilde{x}_3$	$\tilde{x}_0$	$\bar{y}$
1	10	2	5	-1	-1	-1	+1	+1	+1	-1	1	51,17
2	15	2	5	+1	-1	-1	-1	-1	+1	+1	1	55,53
3	10	8	5	-1	+1	-1	-1	+1	-1	+1	1	54,85
4	15	8	5	+1	+1	-1	+1	-1	-1	-1	1	57,85
5	10	2	25	-1	-1	+1	+1	-1	-1	+1	1	79,35
6	15	2	25	+1	-1	+1	-1	+1	-1	-1	1	78,60
7	10	8	25	-1	+1	+1	-1	-1	+1	-1	1	83,03
8	15	8	25	+1	+1	+1	+1	+1	+1	+1	1	84,29

3. Математическая модель, учитывающая взаимодействие факторов, согласно формуле (3.8) имеет вид:

$$y = b_0 + b_1\tilde{x}_1 + b_2\tilde{x}_2 + b_3\tilde{x}_3 + b_{12}\tilde{x}_1\tilde{x}_2 + b_{13}\tilde{x}_1\tilde{x}_3 + b_{23}\tilde{x}_2\tilde{x}_3 + b_{123}\tilde{x}_1\tilde{x}_2\tilde{x}_3 .$$

4. Коэффициенты модели найдем по формулам (3.9)–(3.11):

$$b_0 = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i0}\overline{y}_i = \frac{1}{8}\big(1\cdot 51{,}17 + 1\cdot 55{,}53 + 1\cdot 54{,}85 + 1\cdot 57{,}85 + 1\cdot 79{,}35 + 1\cdot 78{,}60 + 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 68{,}08$$

$$b_1 = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i1}\overline{y}_i = \frac{1}{8}\big(-1\cdot 51{,}17 + 1\cdot 55{,}53 - 1\cdot 54{,}85 + 1\cdot 57{,}85 - 1\cdot 79{,}35 + 1\cdot 78{,}60 - 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 0{,}98$$

$$b_2 = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i2}\overline{y}_i = \frac{1}{8}\big(-1\cdot 51{,}17 - 1\cdot 55{,}53 + 1\cdot 54{,}85 + 1\cdot 57{,}85 - 1\cdot 79{,}35 - 1\cdot 78{,}60 + 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 1{,}92$$

$$b_3 = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i3}\overline{y}_i = \frac{1}{8}\big(-1\cdot 51{,}17 - 1\cdot 55{,}53 - 1\cdot 54{,}85 - 1\cdot 57{,}85 + 1\cdot 79{,}35 + 1\cdot 78{,}60 + 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 13{,}23$$

$$b_{12} = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i1}\tilde{x}_{i2}\overline{y}_i = \frac{1}{8}\big(1\cdot 51{,}17 - 1\cdot 55{,}53 - 1\cdot 54{,}85 + 1\cdot 57{,}85 + 1\cdot 79{,}35 - 1\cdot 78{,}60 - 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 0{,}08$$

$$b_{13} = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i1}\tilde{x}_{i3}\overline{y}_i = \frac{1}{8}\big(1\cdot 51{,}17 - 1\cdot 55{,}53 + 1\cdot 54{,}85 - 1\cdot 57{,}85 - 1\cdot 79{,}35 + 1\cdot 78{,}60 - 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = -0{,}86$$

$$b_{23} = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i2}\tilde{x}_{i3}\overline{y}_i = \frac{1}{8}\big(1\cdot 51{,}17 + 1\cdot 55{,}53 - 1\cdot 54{,}85 - 1\cdot 57{,}85 - 1\cdot 79{,}35 - 1\cdot 78{,}60 + 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 0{,}42$$

$$b_{123} = \frac{1}{8}\sum_{i=1}^{8}\tilde{x}_{i1}\tilde{x}_{i2}\tilde{x}_{i3}\overline{y}_i = \frac{1}{8}\big(-1\cdot 51{,}17 + 1\cdot 55{,}53 + 1\cdot 54{,}85 - 1\cdot 57{,}85 + 1\cdot 79{,}35 - 1\cdot 78{,}60 - 1\cdot 83{,}03 + 1\cdot 84{,}29\big) = 0{,}42$$

Запишем полученные коэффициенты в таблицу 4.25.

Таблица 4.25 — Коэффициенты модели ПФЭ

b_0	b_1	b_2	b_3	b_{12}	b_{13}	b_{23}	b_{123}
68,08	0,98	1,92	13,23	0,08	-0,86	0,42	0,42

5. Оценка значимости коэффициентов модели.

Найдем выборочные дисперсии S_i^2 результатов опытов для i-го эксперимента ($i = 1,...,8$). Для удобства расчеты представим в таблице 4.26.

Таблица 4.26 — Расчет выборочных дисперсий

i	y_1	y_2	y_3	$\overline{y_i}$	$\left(y_{1i} - \overline{y}_i\right)^2$	$\left(y_{2i} - \overline{y}_i\right)^2$	$\left(y_{3i} - \overline{y}_i\right)^2$	S_i^2
1	49,39	52,12	51,99	51,17	3,16	0,91	0,68	2,37
2	56,71	54,06	55,82	55,53	1,39	2,16	0,08	1,82
3	52,88	55,98	55,68	54,85	3,87	1,28	0,69	2,92
4	55,53	59,61	58,4	57,85	5,37	3,11	0,31	4,39
5	78,03	80,03	79,99	79,35	1,74	0,46	0,41	1,31
6	79,52	78,61	77,68	78,60	0,84	0,00	0,85	0,85
7	81,01	84,4	83,67	83,03	4,07	1,89	0,41	3,18
8	85,69	84,48	82,69	84,29	1,97	0,04	2,55	2,28

Суммируя элементы последнего столбца таблицы 6, получим:

$$\sum_{i=1}^{8} S_i^2 = 19,12 .$$

Отсюда определим дисперсию воспроизводимости по формуле (3.3):

$$S_{\{y\}}^2 = \frac{1}{8}\sum_{i=1}^{8} S_i^2 = \frac{1}{8} \cdot 19,12 = 2,39 .$$

Определим дисперсию коэффициентов по формуле (3.12):

$$S_{\{b\}}^2 = \frac{S_{\{y\}}^2}{n \cdot m} = \frac{2,39}{8 \cdot 3} = 0,10 .$$

Среднее квадратическое отклонение коэффициентов:

$$S_{\{b\}} == \sqrt{0,10} = 0,32.$$

Используя распределение Стьюдента (в Excel функция СТЬЮДРАСПОБР) по числу степеней свободы $n(m-1) = 8 \cdot 2 = 16$ и уровню значимости $\alpha = 0,05$, находим $t_{кр} = 2,12$. Значит, $t_{кр} \cdot S_{\{b\}} = 2,12 \cdot 0,32 = 0,67$. Сравнивая коэффициенты модели (табл. 4.25) со значением $t_{кр} \cdot S_{\{b\}} = 0,67$, видим, что коэффициенты b_0, b_1, b_2, b_3, b_{13} больше по абсолютной величине 0,67, следовательно, являются статистически значимыми. Полагая незначимые коэффициенты равными нулю, запишем модель в кодированных переменных:

$$y = 68,08 + 0,98\tilde{x}_1 + 1,92\tilde{x}_2 + 13,23\tilde{x}_3 - 0,86\tilde{x}_1\tilde{x}_3. \qquad (4.1)$$

6. Проверим полученное уравнение (4.1) на адекватность по критерию Фишера. Для нахождения наблюдаемого значения критерия $F_{наб}$ необходимо найти остаточную дисперсию $S^2_{ост}$. Для этого найдем расчетные значения $\hat{y}_i$ по уравнению (4.1) для каждого эксперимента в таблице 4.23. Для удобства расчеты проведем в таблице 4.27.

Таблица 4.27 — Расчет дисперсии адекватности

i	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_1\tilde{x}_2$	$\tilde{x}_1\tilde{x}_3$	$\hat{y}_i$	$\overline{y}_i$	$\left(\hat{y}_i - \overline{y}_i\right)^2$
1	-1	-1	-1	+1	+1	51,09	51,17	0,0063
2	+1	-1	-1	-1	-1	54,77	55,53	0,5795
3	-1	+1	-1	-1	+1	54,93	54,85	0,0063
4	+1	+1	-1	+1	-1	58,61	57,85	0,5795
5	-1	-1	+1	+1	-1	79,27	79,35	0,0066
6	+1	-1	+1	-1	+1	79,53	78,60	0,8502
7	-1	+1	+1	-1	-1	83,11	83,03	0,0066
8	+1	+1	+1	+1	+1	83,36	84,29	0,8502

Суммируя элементы последнего столбца таблицы 4.27, получим:

$$\sum_{i=1}^{8}\left(\hat{y}_i - \overline{y_i}\right)^2 = 2,89.$$

Отсюда остаточная дисперсия (дисперсия адекватности) равна:

$$S^2_{ост} = \frac{m}{n-r}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y_i}\right)^2 = \frac{3}{8-5}\cdot 2,89 = 2,89.$$

Используем найденную ранее дисперсию воспроизводимости $S^2_{\{y\}} = 2,39$, найдем наблюдаемое значение критерия Фишера:

$$F_{наб} = \frac{S^2_{ост}}{S^2_{\{y\}}} = \frac{2,89}{2,39} = 1,21.$$

Используя распределение Фишера (в Excel функция FРАСПОБР) при уровне значимости $\alpha = 0,05$ по соответствующим степеням свободы $k_1 = n - r = 8 - 5 = 3$ и $k_2 = n(m-1) = 8\cdot 2 = 16$, находим $F_{кр} = 3,24$. Так как $F_{наб} = 1,21 < F_{кр} = 3,24$, то уравнение (4.1) адекватно.

7. Уравнение зависимости оценки эффективности процесса очистки воды от влияющих факторов в натуральном масштабе.

Запишем модель в натуральных факторах, для чего используем уравнение (4.1) и преобразование переменных по формуле (3.6):

$$\tilde{x}_1 = \frac{x_1 - 12,5}{2,5};\ \tilde{x}_2 = \frac{x_2 - 5,0}{3,0};\ \tilde{x}_3 = \frac{x_3 - 15,0}{10,0}.$$

Получим:

$$y = 68,08 + 0,98\frac{x_1 - 12,5}{2,5} + 1,92\frac{x_2 - 5,0}{3,0} +$$
$$+13,23\frac{x_3 - 15,0}{10,0} - 0,86\frac{x_1 - 12,5}{2,5}\cdot\frac{x_3 - 15,0}{10,0}.$$

Преобразовав уравнение, получим модель эффективности процесса очистки воды в натуральных влияющих факторах:

$$Э = 34,58 + 0,84\cdot V + 0,64\cdot\frac{h}{d} + 1,69\cdot t - 0,03\cdot V\cdot t.$$

8. Интерпретация коэффициентов полученной модели

Анализируя коэффициенты уравнения (4.1), выделяем факторы с наибольшими значениями коэффициентов. Это факторы X_3 – температура очищаемой воды (коэффициент 13,23) и X_2 – отношение высоты загрузки фильтрационной колонки к ее диаметру (коэффициент 1,92). Следовательно, наибольшее влияние на эффективность процесса очистки воды от нитратов путем ионного обмена оказывает температура очищаемой воды, на втором месте – отношение высоты загрузки фильтрационной колонки к ее диаметру. Поскольку связь положительная, то это значит, что с увеличением этих факторов эффективность очистки воды возрастает. Совместное воздействие скорости фильтрации и температуры очищаемой воды (коэффициент в уравнении (4.1) равен –0,86) оценивается отрицательной связью, приводящей к уменьшению эффективности очистки воды. Однако эта связь гораздо менее выражена, чем линейные связи по факторам X_2 и X_3.

Задача 5Б. Поиск оптимального плана двухфакторного эксперимента

На предыдущем этапе исследования было установлено, что на эффективность очистки наибольшее влияние оказывает температура очищаемой воды (t, °C) и скорость фильтрации (V, м/час), которые предположительно оказывают совместное влияние на изучаемый процесс. Поэтому с целью уточнения степени совместного влияния факторов необходимо провести полный факторный эксперимент для двух факторов (ПФЭ 2^2).

Для определения влияния на эффективность процесса очистки воды от нитратов (Э, %) требуется применить полный факторный эксперимент с последующим поиском оптимального плана.

Решение поставленной задачи.

1. На основе априорных данных были выбраны нижняя и верхняя граница варьирования факторов (табл. 4.28.)

Таблица 4.28 — Параметры варьируемых факторов

Фактор	Единицы измерения	Нижняя граница	Верхняя граница
Скорость фильтрации воды	м/час	10	15
Температура очищаемой воды	°С	5	25

Уровни и интервалы варьирования факторов

x_1 — скорость фильтрации воды (V, м/час), $x_1^{\min} = 10$, $x_1^{\max} = 15$;

x_2 — температура очищаемой воды (t, °С), $x_2^{\min} = 5$, $x_2^{\max} = 25$;

отклик Y — эффективность процесса очистки воды от нитратов, (Э, %).

Основной уровень (центр плана): $x_1^0 = \frac{15+10}{2} = 12{,}5$;

$x_2^0 = \frac{25+5}{2} = 15{,}0$.

Интервал варьирования: $\Delta x_1 = \frac{15-10}{2} = 2{,}5$; $\Delta x_2 = \frac{25-5}{2} = 10{,}0$.

Безразмерные координаты (коды):

$$\tilde{x}_1^{\,1} = \frac{10-12{,}5}{2{,}5} = -1;\ \tilde{x}_1^{\,2} = \frac{15-12{,}5}{2{,}5} = 1;$$

$$\tilde{x}_2^{\,1} = \frac{5-15{,}0}{10{,}0} = -1;\ \tilde{x}_2^{\,2} = \frac{25-15{,}0}{10{,}0} = 1.$$

В данной задаче два фактора (k=2) на двух уровнях (p=2). Следовательно, общее число экспериментов $N = 2^2 = 4$. Коды и интервалы варьирования факторов запишем в таблицу 4.29.

Таблица 4.29 — Коды и интервалы варьирования факторов

Фактор	Единицы измерения	Обозначения	Диапазон варьирования	
			код = –1	код = 1
V	м/час	x_1	10	15
t	°С	x_2	5	25

2. Составим матрицы планирования двухфакторного эксперимента данной задачи (табл. 4.30 и 4.31).

Таблица 4.30 — Исходная матрица планирования для проведения опытов

№ эксперимента	Значения факторов				Результаты опытов	
	в натуральном масштабе		в безразмерном масштабе			
	x_1	x_2	$\tilde{x}_1$	$\tilde{x}_2$	y_1	y_2
1	10	5	-1	-1	48,66	51,12
2	15	5	+1	-1	75,43	76,06
3	10	25	-1	+1	32,30	31,51
4	15	25	+1	+1	56,32	58,61

Реализация плана эксперимента (опыты) были проведены на экспериментальной установке (рис. 4.23) для ионита марки «Purolite NRW600(OH)». Для уменьшения систематических ошибок опыты по плану типа 2^2 проводили в случайном порядке. Для оценки ошибки эксперимента каждый опыт был осуществлен дважды.

Таблица 4.31 — План-матрица эксперимента 2^2

№ эксперимента	Значения факторов				Совместное влияние	Вспомогательный столбец	Отклик (среднее результатов опытов)
	в натуральном масштабе		в безразмерном масштабе				
	x_1	x_2	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_1\tilde{x}_2$	$\tilde{x}_0$	$\overline{y}$
1	10	5	-1	-1	+1	1	49,89
2	15	5	+1	-1	-1	1	75,75
3	10	25	-1	+1	-1	1	31,91
4	15	25	+1	+1	+1	1	57,47

3. Математическая модель, учитывающая взаимодействие двух факторов имеет вид (3.7):

$$y = b_0 + b_1\tilde{x}_1 + b_2\tilde{x}_2 + b_{12}\tilde{x}_1\tilde{x}_2$$

4. Найдем коэффициенты модели по формулам (3.9)–(3.11):

$$b_0 = \frac{1}{4}\sum_{i=1}^{4}\tilde{x}_{i0}\overline{y}_i = \frac{1}{4}\left(1\cdot 49{,}89 + 1\cdot 75{,}75 + 1\cdot 31{,}91 + 1\cdot 57{,}47\right) = 53{,}75;$$

$$b_1 = \frac{1}{4}\sum_{i=1}^{4}\tilde{x}_{i1}\overline{y}_i = \frac{1}{4}\left(-1\cdot 49{,}89 + 1\cdot 75{,}75 - 1\cdot 31{,}91 + 1\cdot 57{,}47\right) = 12{,}85;$$

$$b_2 = \frac{1}{4}\sum_{i=1}^{4}\tilde{x}_{i2}\overline{y}_i = \frac{1}{4}\left(-1\cdot 49{,}89 - 1\cdot 75{,}75 + 1\cdot 31{,}91 + 1\cdot 57{,}47\right) = -9{,}07;$$

$$b_{12} = \frac{1}{4}\sum_{i=1}^{4}\tilde{x}_{i2}\tilde{x}_{i3}\overline{y}_i = \frac{1}{4}\left(1\cdot 49{,}89 - 1\cdot 75{,}75 - 1\cdot 31{,}91 + 1\cdot 57{,}47\right) = -0{,}07.$$

Запишем полученные коэффициенты в таблицу 4.32.

Таблица 4.32 — Коэффициенты модели ПФЭ

b_0	b_1	b_2	b_{12}
53,75	12,85	-9,07	-0,07

5. Оценка значимости коэффициентов модели

Найдем выборочные дисперсии S_i^2 результатов опытов для i-го эксперимента ($i = 1,...,4$). Для удобства расчеты представим в таблице 4.33.

Величины S_i^2 рассчитываем по формуле (3.2). В данном случае:

$m = 2$, значит, $S_i^2 = \frac{1}{2-1}\sum_{j=1}^{2}(y_{ji} - \overline{y}_i)^2 = \sum_{j=1}^{2}(y_{ji} - \overline{y}_i)^2$, $i = 1,...,4$.

Таблица 4.33 — Расчет выборочных дисперсий

i	y_1	y_2	$\overline{y_i}$	$\left(y_{1i}-\overline{y}_i\right)^2$	$\left(y_{2i}-\overline{y}_i\right)^2$	S_i^2
1	48,66	51,12	49,89	1,51	1,51	3,03
2	75,43	76,06	75,75	0,10	0,10	0,20
3	32,30	31,51	31,91	0,16	0,16	0,31
4	56,32	58,61	57,47	1,31	1,31	2,62

Суммируя элементы последнего столбца таблицы 6, получим:

$$\sum_{i=1}^{4} S_i^2 = 6{,}16 .$$

Определим дисперсию воспроизводимости $S_{\{y\}}^2$:

$$S_{\{y\}}^2 = \frac{1}{n}\sum_{i=1}^{n} S_i^2 = \frac{1}{4}\sum_{i=1}^{4} S_i^2 = \frac{1}{4}\cdot 6{,}16 = 1{,}54 .$$

Определим среднее квадратическое отклонение коэффициентов:

$$S_{\{b\}} = \sqrt{\frac{S_{\{y\}}^2}{n\cdot m}} = \sqrt{\frac{1{,}54}{4\cdot 2}} = 0{,}44 .$$

Используя распределение Стьюдента (в Excel функция СТЬЮДРАСПОБР) по числу степеней свободы $n(m-1) = 4\cdot(2-1) = 4$ и уровню значимости $\alpha = 0{,}05$, находим $t_{кр} = 2{,}78$. Значит, $t_{кр}\cdot S_{\{b\}} = 2{,}78\cdot 0{,}44 = 1{,}22$. Сравнивая коэффициенты модели (табл. 4.32) со значением $t_{кр}\cdot S_{\{b\}} = 1{,}22$, видим, что коэффициенты b_0, b_1, b_2, больше по абсолютной величине 1,22, следовательно, являются статистически значимыми. Коэффициент b_{12} статистически незначим, следовательно, эффектом взаимодействия факторов можно пренебречь.

Полагая незначимые коэффициенты равными нулю, запишем модель в кодированных переменных:

$$y = 53{,}75 + 12{,}85\tilde{x}_1 - 9{,}07\tilde{x}_2 . \qquad (4.2)$$

6. Проверка модели на адекватность

Проверим полученное уравнение (4.2) на адекватность по критерию Фишера. Для нахождения наблюдаемого значения критерия $F_{набл}$ необходимо найти остаточную дисперсию $S^2_{ост}$. Для этого найдем расчетные значения $\hat{y}_i$ по уравнению (4.2) для каждого эксперимента в таблице 4.30. Для удобства расчеты проведем в таблице 4.34.

Таблица 4.34 — Расчет дисперсии адекватности

i	$\tilde{x}_1$	$\tilde{x}_2$	$\hat{y}_i$	$\overline{y}_i$	$\left(\hat{y}_i - \overline{y}_i\right)^2$
1	-1	-1	49,96	49,89	0,0054
2	+1	-1	75,67	75,75	0,0054
3	-1	+1	31,83	31,91	0,0054
4	+1	+1	57,54	57,47	0,0054

Суммируя элементы последнего столбца таблицы 4.34, получим:

$$\sum_{i=1}^{4}\left(\hat{y}_i - \overline{y}_i\right)^2 = 0,02.$$

Отсюда остаточная дисперсия (дисперсия адекватности) равна:

$$S^2_{ост} = \frac{m}{n-r}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}_i\right)^2 = \frac{2}{4-2}\cdot 0,02 = 0,02.$$

Наблюдаемое значение критерия Фишера:

$$F_{набл} = \frac{S^2_{ост}}{S^2_{\{y\}}} = \frac{0,02}{1,54} = 0,01.$$

Используя распределение Фишера (в Excel функция FРАСПОБР) при уровне значимости $\alpha = 0,05$ по соответствующим степеням свободы $k_1 = n - r = 4 - 2 = 2$ и $k_2 = n(m-1) = 4\cdot(2-1) = 4$, находим $F_{кр} = 6,94$. Так как $F_{набл} = 0,01 < F_{кр} = 6,94$, то уравнение (4.2) адекватно.

7. Запишем модель в натуральных факторах, для чего используем уравнение (4.2) и преобразование переменных по формуле (3.6):

$$\tilde{x}_1 = \frac{x_1 - 12,5}{2,5};\ \tilde{x}_2 = \frac{x_2 - 15,0}{10,0}.$$

Получим:

$$y = 53,75 + 12,85 \cdot \frac{x_1 - 12,2}{2,5} - 9,07 \cdot \frac{x_2 - 15,0}{10,0}.$$

После преобразований получим: $y = 3,105 + 5,140 \cdot x_1 - 0,907 \cdot x_2$.

Следовательно, модель эффективности Э, % процесса очистки воды в натуральных влияющих факторах имеет следующий вид:

$$Э = 3,105 + 5,140 \cdot V - 0,907 \cdot t, \qquad (4.3)$$

где V — скорость фильтрации воды (м/час),

t — температура очищаемой воды (°С).

8. Определим значения температуры очищаемой воды и скорость фильтрации, при которой эффективность процесса очистки воды от нитратов будет максимальной. Для этого решим задачу поиска оптимального плана методом крутого восхождения.

На первом этапе определяется базовый фактор. Вначале рассчитываются произведения $b_i \cdot \Delta x_i$ для каждого из факторов уравнения. Наибольшим оказалось произведение $b_2 \cdot x_2 = -90,7$, поэтому базовым фактором считаем x_2.

Далее рассчитываем параметр шага оптимизации $\lambda = \frac{\mu}{|b_2|}$, приняв $\mu = 0,4$. Тогда $\lambda = \frac{0,4}{|-9,07|} = 0,04$. После этого для каждого фактора определяем шаги при крутом восхождении по формуле: $\lambda(b_i \Delta x_i)$.

В результате получим:

– для фактора x_1 шаг $0,04 \cdot 32,13 = 1,42$, принимаем 1,5;

– для фактора x_2 шаг $0,04 \cdot 90,7 = 3,63$, принимаем 4,0.

Далее, изменяя основной уровень фактора ($x_1^0 = 12,5$ и $x_2^0 = 15,0$) на соответствующий шаг, рассчитываем значение функции отклика в натуральном масштабе по формуле (4.3), т.о. проводим «мысленный» опыт. С целью уточнения результата на следующем шаге проводим «реальный» опыт с заданными значениями факторов. Условия планирования и последовательность этапов оптимизации плана методом крутого восхождения представлены в таблице 4.35.

Таблица 4.35 — Оптимизации плана методом крутого восхождения

Последовательность этапов крутого восхождения	Факторы		Эффективность процесса очистки воды от нитратов Э, %
	скорость фильтрации воды V, м/час	температура очищаемой воды t, °С	
Условия планирования эксперимента			
Основной уровень	12,5	15,0	
Интервал варьирования Δx_i	2,5	10,0	
Верхний уровень ($x_i^{\max}$)	10,0	5,0	
Нижний уровень ($x_i^{\min}$)	15,0	25,0	
План ПФЭ 2^2	$\tilde{x}_1$	$\tilde{x}_2$	$\overline{y}$
	-1	-1	49,89
	+1	-1	75,75
	-1	+1	31,91
	+1	+1	57,47
Коэффициенты уравнения b_i	12,85	-9,07	
Произведение $b_i \cdot \Delta x_i$	32,13	-90,7	
Параметр λ	–	0,04	
Шаг $\lambda(b_i \Delta x_i)$	1,5	4,0	

Продолжение таблицы 4.35

Последовательность этапов крутого восхождения	Факторы		Эффективность процесса очистки воды от нитратов Э, %
	скорость фильтрации воды V, м/час	температура очищаемой воды t, °С	
Опыты на линии крутого восхождения			
1) мысленный	12,5+1,5=14,0	15,0-4,0=11,0	65,09
2) реализованный	14,0+1,5=15,5	11,0-4,0=7,0	75,50
3) мысленный	15,5+1,5=17,0	7,0-4,0=3,0	вне диапазона изменений факторов

Поиск оптимального решения был оставлен, когда значения факторов вышли за границы диапазона варьирования. По результатам эксперимента оптимальным был признан план: $x_1 = 15,5$, $x_2 = 7,0$, $y_{\max} = 75,5$. Следовательно, в условиях данного эксперимента максимальная эффективность очистки воды от нитратов может составлять 75,5 % при следующих факторах работы ионообменной колонки: скорость фильтрации воды должна быть 15,5 м/час, а температуры очищаемой воды — 7,0 °С.

ЛИТЕРАТУРА

1. Гмурман, В. Е. Теория вероятностей и математическая статистика [Текст] / В. Е. Гмурман. — М. : Высш. шк., 2001. — 479 с.

2. Подлипенская, Л. Е. Математическая статистика для горняков [Текст] : учеб. пособие / Л. Е. Подлипенская. — Алчевск : ДГМИ, 2004. — 171 с.

3. Юдин, Ю. В. Организация и математическое планирование эксперимента : учебное пособие [Рекомендовано методическим советом Уральского федерального университета для студентов вуза, обучающихся по направлению подготовки 22.03.01 — Материаловедение и технология материалов] / Ю. В. Юдин, М. В. Майсурадзе, Ф. В. Водолазский ; научный редактор А. А. Попов ; Министерство образования и науки Российской Федерации, Уральский федеральный университет имени первого Президента России Б.Н. Ельцина. — Екатеринбург : Издательство Уральского университета, 2018. — 124 с. — ISBN 978-5-7996-2486-6. Режим доступа: https://elar.urfu.ru/bitstream/10995/65224/1/978-5-7996-2486-6_2018.pdf.

4. Медведев, П. В. Математическое планирование эксперимента : учебное пособие / П. В. Медведев, В. А. Федотов ; Оренбургский государственный университет. – Оренбург : Оренбургский государственный университет, 2017. – 98 с. : табл., граф., схем., ил. – Режим доступа: https://biblioclub.ru/index.php?page=book&id=481785 (дата обращения: 14.02.2021). – Библиогр.: с. 72-74. – ISBN 978-5-7410-1759-3. – Текст : электронный.

5. Вершинин, В. И. Планирование и математическая обработка результатов химического эксперимента. [Электронный ресурс] / В. И. Вершинин, Н. В. Перцев. — Электрон. дан. — СПб. : Лань, 2017. — 236 с.

6. Математическое моделирование. Практикум [Электронный ресурс] : учебное пособие / Л. А. Коробова, Ю. В. Бугаев, С. Н. Черняева, Ю. А. Сафонова ; Министерство образования и науки РФ, Воронежский государственный университет инженерных технологий ; науч. ред. Л. А. Коробова. — Воронеж : Воронежский государственный универси-

тет инженерных технологий, 2017. — 113 с. (ЭБС «Университетская библиотека онлайн»).

7. Иванов, В. В. Математическое моделирование [Электронный ресурс] : учебно-методическое пособие / В. В. Иванов, О. В. Кузьмина ; Поволжский государственный технологический университет. — Йошкар-Ола : ПГТУ, 2016. — 88 с. (ЭБС «Университетская библиотека онлайн»).

8. Дерябин, В. А. Экология : учебное пособие / В. А. Дерябин, Е. П. Фарафонтова. — Екатеринбург : Изд-во Урал. ун-та, 2016. — 136 с.

9. Рогожников, Д. А. Экологические проблемы металлургического производства. Часть 1 : учебное пособие / Д. А. Рогожников, А. А. Шопперт, И. В. Логинова. — Екатеринбург : Издательство УМЦ УПИ, 2017. — 224 с.

10. Методы планирования и обработки результатов инженерного эксперимента : учебное пособие / Н. А. Спирин, В. В. Лавров, Л. А. Зайнуллин, А. Р. Бондин, А. А. Бурыкин; под общ. ред. Н. А. Спирина. — Екатеринбург : ООО «УИНЦ», 2015. — 290 с.

11. Большина, Е. П. Экология металлургического производства : Курс лекций. — Новотроицк : НФ НИТУ «МИСиС», 2012. – 155 с. — http://nf.misis.ru/download/mt/Ekology_metallurg_proizvodstva.pdf

12. Андреюк, С. В. Эффективность исселдований очистки воды от нитратов математическим планированием / С. В. Андреюк, Б. Н. Житенев // Аграрные ландшафты, их устойчивость и особенности развития : Сборник научных трудов по материалам Международной научной экологической конференции / составитель Л. С. Новопольцева; под редакцией И. С. Белюченко. — 2020. — С. 321–323.

13. Андреюк, С. В. Исследование методов физико-химической очистки природных вод от нитратов / С. В. Андреюк // Сборник научных статей Международной научно-практической конференции, Брест, 6–8 апреля 2016 г. : в 2-х ч. / УО «Брестский гос. технический ун-т.»; под ред. А.А. Волчека [и др.]. — Брест, 2016. — Ч. II. — С. 159–163.

14. Соколовская, И. Ю. Полный факторный эксперимент / И. Ю. Соколовская // Методические указания для самостоятельной работы студентов. — Новосибирск : НГАВТ, 2010. — 36 с.

УЧЕБНОЕ ИЗДАНИЕ

С. И. Кулакова, Л. Е. Подлипенская, Д. А. Мельничук

ОРГАНИЗАЦИЯ И МАТЕМАТИЧЕСКОЕ ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

Учебное пособие

В авторской редакции

Художественное оформление обложки Н. В. Чернышова

Заказ № 77. Формат 60x84 $^{1}/_{16}$.
Бумага офс. Печать RISO.
Усл. печат. л. 7,1 Уч.-изд. л. 6,1
Издательство не несет ответственность за содержание
материала, предоставленного автором к печати.
Издатель и изготовитель:
ГОУ ВО ЛНР «ДонГТИ»
пр. Ленина, 16, г. Алчевск, ЛНР, 94204
(ИЗДАТЕЛЬСКО-ПОЛИГРАФИЧЕСКИЙ ЦЕНТР, ауд. 2113, т/факс 2-58-59)
Свидетельство о государственной регистрации издателя, изготовителя
и распространителя средства массовой информации
МИ-СГР ИД 000055 от 05.02.2016