УДК 911.2:51(075.8)
ББК 26.82В.я73
Ч-50

Р е ц е н з е н т ы:
кафедра физической географии Белорусского государственного педагогического университета имени Максима Танка;
доктор географических наук, доцент Института проблем использования природных ресурсов и экологии НАН Беларуси *В. С. Хомич*

Чертко, Н. К.

Ч-50 Математические методы в географии : учеб.-метод. пособие / Н. К. Чертко, А. А. Карпиченко. – Минск : БГУ, 2009. – 199 с.
ISBN 978-985-518-128-7.

Рассматриваются математические методы (дисперсионный, информационный, кластерный, корреляционный, регрессионный, линейного программирования, теории графов, моделирования, тренд-анализа), применяемые в географических исследованиях для группировки, классификации объектов и выявления пространственных закономерностей на базе картографирования.

Для студентов, обучающихся по географическим специальностям в вузах.

УДК 911.2:51(075.8)
ББК 26.82В.я73

ISBN 978-985-518-128-7

ВВЕДЕНИЕ

Географические исследования и практические задачи базируются на большом объеме количественной информации, которую необходимо сгруппировать, классифицировать, а также объективно оценить. Эти вопросы успешно решаются с помощью математических методов и соответствующих программ, разработанных для ПЭВМ. Исследователь или практик должен лишь четко сформулировать задачу, выбрать наиболее подходящий для конкретных условий математический метод анализа и дать объективную интерпретацию результатов.

Математика позволяет нам решать задачи частные и общие. Например, расход воды в реке рассчитывается на основе специальной частной формулы, а загрязнение воды в реке под воздействием предприятия оценивается с применением факторного анализа – общего для решения многих специальных географических задач.

В данном пособии рассматриваются те математические методы анализа, которые можно применять исполнителю независимо от географической специализации. Во избежание ошибок много внимания уделяется систематизации экспериментальных данных, формулировке задач, обоснованию применения метода анализа, решению конкретных примеров, интерпретации результатов. В приложении приведены алгоритмы выполнения задания на ПЭВМ по важнейшим методам анализа.

Механический подход при использовании математических методов недопустим. В конкретной ситуации надо выбрать надежный математический прием, так как каждый из методов анализа имеет свои возможности и ограниченную область применения. Большинство методов статистического анализа универсальны и могут применяться в разнообразных отраслях деятельности человека. Поэтому все программные средства, которые можно использовать для статистической обработки на персональных компьютерах, разделяют на специализированные пакеты, статистические пакеты общего назначения, табличные процессоры и электронные таблицы. Сопроводительные описания рассчитываются для пользователей со специальной подготовкой в области математики.

Значительное влияние на развитие математических методов оказал закон больших чисел, открытый Яковом Бернулли (1654–1705), и теория вероятности, основы которой разработал французский математик и астроном Пьер Симон Лаплас (1749–1827). На основе теории вероятности, позволяющей выявить определенные тенденции в кажущемся хаосе случайных событий, появилась математическая статистика, с помощью которой можно дать количественную оценку вероятностей различных явлений, не имеющих постоянных, всегда одних и тех же исходов, поскольку большинству из них свойственна вариабельность (изменение в определенных пределах). Например, температура воздуха меняется ежечасно, ежедневно, ежемесячно, непостоянна прибыль предприятия. Однако многие хаотические явления имеют упорядоченную структуру и поэтому могут иметь конкретную оценку. Главное условие для этого – их статистическая устойчивость, которую можно описать математическими методами статистики.

По виду учетные признаки могут быть качественными или количественными. Качественные (описательные, атрибутивные) признаки характеризуют качество отдельных единиц совокупности (пол мужской и женский; образование начальное, среднее, высшее). Количественные признаки характеризуют числовые выражения (масса – кг, скорость – км/час). Аналитическая оценка взаимосвязи качественных и количественных признаков проводится только после разбиения количественных признаков на качественные группы.

Следует иметь в виду, что чрезмерное увеличение объема исходной информации ведет к увеличению «информационного шума» (роста числа помех). Достигнув известного предела, «шум» подавляет исходную информацию. Чем сложнее система, тем больше рассматриваемых взаимосвязанных переменных, тем труднее установить множество отношений. Количественные методы анализа помогают выбрать ведущие факторы, причины, признаки. В таких случаях важно понимание смысла математических методов, чтобы не допустить ошибочных выводов. Начинать изучение системы необходимо с усвоения методологических основ организации самих исследований и важнейших элементов системологии, которые определяют последовательность дальнейших действий.

Современные географические методы исследования – сравнительно-географический, системный и другие – необходимо использовать в сочетании с математическим обоснованием результатов. Математические методы позволяют широко использовать системный анализ как наиболее совершенный. Любой географический объект исследования может быть представлен как *система* – определенный объект, состоящий из множе-

ства частей, которые взаимосвязаны не только между собой, но и с соседними объектами-системами. Установить целостность и структуру, иерархичность, величину и направленность связей в системе, их характер позволяют математические методы создания формализованных систем. *Системный подход* основан на исследовании объектов как систем, создает единую теоретическую модель. *Системный анализ* представляет собой совокупность методологических средств. Успешное использование системного анализа возможно при реализации следующих важнейших принципов, опирающихся на математические методы: определяется конечная цель исследования; система-объект рассматривается как единое целое, в ней выявляются все взаимосвязи и их результаты; строится обобщенная комбинированная модель (модели), где отображаются структура, иерархия и взаимосвязи.

Выделяются две группы систем: *материальные и абстрактные*. Традиционные методы географии изучают материальные системы (статичные, динамичные, закрытые, открытые). Социальные системы через техногенез могут оказывать воздействие на природные. По развитию выделяют системы статичные (предприятия) и динамичные (ландшафт). По характеру взаимодействия системы делятся на закрытые (в них не поступает и из них не выводится вещество, происходит лишь обмен энергией) и открытые (постоянно происходит ввод и вывод вещества и обмен энергией). В открытой системе, например ландшафте, постоянно протекающие процессы и явления создают подвижное равновесие, т. е. некоторую стабильность в определенных условиях среды и общества.

Среди абстрактных систем на основе различных систематизирующих отношений можно выделить: функциональные (математическая модель), структурные (глобус), временные (прогноз погоды), геометрические (линия регрессии на графике). В научную литературу введено понятие «*управляющая система*», рассматриваемая как схематическое отображение реальных объектов. Она задается элементами, схемой и координатами. Элементы определяются через их свойства. Схема отображает характер соединений между элементами. Координаты показывают относительное положение выделенных элементов управляющей системы. Любая управляющая система не мыслится без понятия функции отображения одного множества в другом как действия с реальными предметами или как вещественного процесса (например, функция растительности – создание органического вещества из неорганического с использованием солнечной энергии в процессе фотосинтеза).

Впервые математические методы в географии предложено было использовать в 20-е годы XX в. российскими географами П. П. Семеновым-Тян-Шанским и М. М. Протодьяконовым. Положительно отозвался о воз-

можности применения математики в географии академик А. А. Григорьев в 1934 г. Пионером внедрения математики в географию является Д. Л. Арманд (1949). Первая работа, посвященная использованию математической статистики в географии, была опубликована В. А. Червяковым (1966).

Успехи применения математических методов в географии позволили в 1968 г. на базе Московского государственного университета провести первое Всесоюзное совещание по данной проблеме. В решении совещания обращалось внимание на необходимость фундаментальной подготовки молодых специалистов-географов в области математики.

Дальнейшее развитие всех областей географической науки дает возможность использовать в экспериментах многие разделы математики (теория вероятности, теория информации, линейное программирование, теория графов, теория игр, временные ряды).

Основоположником практического применения линейного программирования является академик Л. В. Канторович. Им разработаны методы решения транспортных задач и сетевой постановки с применением теории графов.

С 1978 г. в издательстве Московского государственного университета выходят учебные пособия В. С. Михеевой, в которых рассматривается использование математических методов в экономической географии (методы линейного программирования и теория графов).

В настоящее время основные математические методы анализа обеспечены программными продуктами для ПЭВМ. Простейшие статистические расчеты можно выполнять с помощью Microsoft Excel, входящего в состав Microsoft Office. Однако лучшие результаты дает специализированное программное обеспечение. Наиболее распространенные и универсальные статистические программные пакеты – это Statistica, Systat, NCSS, SPSS, различающиеся в деталях, версиях, а также полнотой представления материала. Наиболее полно типичные задачи представлены в пакете статистических программ Statistica.

Глава 1

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Источником материала для статистической обработки могут быть собственные экспериментальные исследования, статистическая информация, аналитические данные других исследователей, фондовые материалы, литературные источники, географические карты, аэрофотоснимки. При изучении территориальных комплексов низших рангов (фаций, урочищ), промышленных предприятий, объектов сельскохозяйственного назначения наиболее ценными для статистической обработки являются материалы собственных исследований. При изучении объектов среднего ранга возрастает роль отраслевых и специальных карт вместе с авторскими данными и литературными источниками. Для исследования объектов высоких рангов (области, провинции, регионы) используются карты, литературные источники, обобщающие материалы по объектам более низких рангов.

1.1. Генеральная совокупность и выборка

Первичным элементом в статистике является *единица наблюдения* (варианта, дата): 3 4 3 4 3 3 3 3. Их ряд образует *статистическую совокупность*, которая характеризует *объект* исследования. Большинство единиц наблюдения имеет *вероятностный, случайный* характер. По виду исследуемые *признаки* могут быть *качественными* и *количественными.* Количественные признаки имеют числовое выражение, качественные – словесное (образование начальное, среднее, высшее). Качественным признакам при статистической обработке присваивают балл или ранг соответственно их смыслу (начальное образование – 1 балл, среднее – 2, высшее – 3). Исследуемые признаки можно подразделить на *факторные* (факториальные) и *результативные* (результирующие); вторые изменяются под влиянием первых. Все единицы наблюдения, входящие в статистическую совокупность, объединены *единством места и времени исследования.*

Чрезмерное увеличение объема любой исходной информации ведет к увеличению «информационного шума» (погрешностей), который подавляет искомую исследователем информацию. Это отражается на *вариабельности* (изменчивости, случайности) процессов и явлений.

По *времени* наблюдение может быть текущим (непрерывным) и единовременным (в один и тот же момент времени в разных точках – метеонаблюдения на постах). По *охвату* исследование может быть *сплошное* и *несплошное* (*выборочное*). Эта особенность определяет ход и методику статистического анализа.

Сплошное статистическое исследование (перепись всего населения республики) образует *генеральную совокупность*. Общее число членов генеральной совокупности называют *объемом генеральной совокупности*. Из-за больших размеров генеральной совокупности или из-за отсутствия определенных границ этой совокупности (Белорусская гряда) оно проводится редко. На исследование генеральной совокупности затрачивается много средств и времени, поэтому ограничиваются методом *выборочного исследования* из *генеральной совокупности.* Выборка образует совокупность наблюдений, полученных с целью объективной характеристики и получения информации о генеральной совокупности. Число ее членов называют *объемом выборочной совокупности.*

Выборочное исследование можно проводить такими методами, как *монографический, основного массива и выборочный.* Монографический метод используется для описания объекта с какими-либо особенностями (зонирование города с развитой машиностроительной промышленностью). Выводы могут быть распространены только на группу аналогичных объектов. Метод основного массива дает представление о конкретном объекте, поэтому переносить полученные закономерности на другие объекты нельзя (бассейн р. Неман). Наиболее распространен метод выборочного исследования из генеральной совокупности.

Выборка может быть представлена следующими основными типами отбора: *случайным, направленным* (типическим), *смешанным*.

При случайном отборе все объекты имеют одинаковую возможность попасть в выборку. В его основе лежит перемешивание. Для этого можно использовать таблицу случайных чисел (прил. 2). Начав с любой четырехзначной колонки и двигаясь по столбцу сверху вниз или снизу вверх, выписывают первые или последние однозначные цифры для объема выборки до 9, двухзначные – для объема выборки от 10 до 99, трехзначные – для объема выборки от 100 до 999 и т. д. По второму варианту из объектов в списке (алфавитном или ином) в выборку включают каждый третий или пятый, или десятый (механическая выборка) и т. д.

Случайная выборка может не отвечать условиям исследования из-за неоднородности. Тогда производят целенаправленный (когортный) отбор, выбирая для исследования типичные объекты. Правила отбора остаются те же, что и при случайном отборе.

Смешанный отбор производят в тех случаях, когда необходимо дать характеристику неоднородного объекта. Например, холмисто-моренный ландшафт делят фации с однородными условиями, в каждой из которых производят случайный отбор. Полученные результаты объединяют в одну выборку.

Соблюдения правил составления выборки дают возможность *наиболее полно и точно, т. е. репрезентативно,* характеризовать генеральную совокупность. Величина ошибки репрезентативности зависит от изменчивости изучаемого признака. Чем больше разброс значений изучаемого признака, тем больше статистическая ошибка. Отбор для выборки должен быть также *научно обоснованным* с учетом принятых методических правил, т. е. *рендомизированным.* В таких случаях при меньшем числе наблюдений уменьшается вероятность систематических ошибок наблюдений.

На втором этапе статистического исследования проводят *сводку* и *группировку* данных. Варианты группировок следующие: разделение анализируемой статистической совокупности на группы по тем или иным признакам; объединение мелких однородных групп в более крупные; комплексная группировка на основе многих учетных признаков, даже если они разнородные.

Типологическая группировка выделяет в совокупности качественно однородные в существенном отношении группы. Группировка по своей сути представляет собой процесс классификации. В государственной статистике используют *классификаторы* – специальные справочники, инструкции, указания.

Самым сложным является определение объема наблюдений в исследованиях, который необходим для получения надежного представления о характере изменчивости признака в генеральной совокупности. Если объект исследуется впервые, то определить объем выборки трудно. В большинстве случаев достаточно точные результаты получают при объеме выборки около 100. Оптимальный объем выборки обычно пропорционален степени изменчивости признака. Если признак сильно изменяется, то количество измерений следует увеличить. Предложены также другие способы определения величины выборочной совокупности при исследованиях: *по таблице достаточно больших чисел* (прил. 1), а также *расчетным способом.* В обоих случаях количество наблюдений определяется исходя из величины допускаемой вероятности, с какой

предполагается делать заключения, и величины точности опыта. Например, при допускаемом уровне вероятности P = 0,95 (95 %) и точности опыта p = 5 % число наблюдений по таблице достаточно больших чисел составит 384. Если точность опыта увеличить до 1 %, то число наблюдений следует увеличить до 9603.

Чаще всего ориентировочный объем (N) выборочной совокупности рассчитывают по формулам, в которых вероятность заменяют степенью варьирования:

$$N = \sigma^2 / m^2_M,$$

где σ – среднее квадратическое отклонение; m_M – ошибка среднего арифметического.

Допустим, варьирование признака (колебание температуры) составляет 7 °С, тогда число наблюдений выборочной совокупности с ошибкой среднего арифметического $m = \pm 0{,}5$ °С составит: $N = \sigma^2 / m^2_M = 7^2 / 0{,}5^2 = 196$.

Объем выборочной совокупности можно также определить по ожидаемому коэффициенту вариации (V) и точности опыта (p) с учетом поправочного коэффициента (1,96) для уровня вероятности 0,95 и 0,99:

$$N = (1{,}96 \cdot V)^2 / p^2.$$

Пример. Для расчета коэффициента увлажнения в зависимости от количества выпадающих осадков и испарения с ожидаемой точностью опыта 3 % и коэффициента вариации 30 % потребуется следующий объем выборочной совокупности $N = (1{,}96 \times 30)^2 / 3^2 = 384$.

Определение объема выборочной совокупности методом расчета минимального, но объективного количества наблюдений необходимо для получения достоверной информации о генеральной совокупности. Полученные параметры по выборке могут служить приблизительными оценками аналогичных параметров генеральной совокупности, т. е. указывать пределы, в которых они заключены ($M \pm m_M$; $\sigma \pm m_\sigma$).

1.2. Обработка вариационного ряда

Варианты в статистической совокупности подвергаются обработке. Для этого составляется *вариационный ряд*, т. е. варианты располагают по возрастающим или убывающим величинам. Варианты в выборке, относящиеся к одному и тому же признаку, практически не совпадают между собой, или *варьируют*. Те варианты, которые резко отличаются от вариантов статистической совокупности и вызывают сомнение у исследователя, определяются как *артефакт*. Они располагаются в начале или в конце вариационного ряда. Артефакт исключается из статистической

совокупности и не подлежит обработке. Например, в приведенных вариационных рядах 2, 9, 11, 12, 13, 15 и 25, 27, 29, 32, 55 почти все соседние показатели весьма близки по значению. Вызывают сомнение варианты 2 в первом ряду и 55 во втором. Их можно принять за артефакт и исключить (выбраковать) из обработки. Выбраковка должна быть статистически доказана.

Существующие критерии выбраковки основываются, как правило, на допущении, что выборка распределяется по нормальному или близкому к нему закону. В качестве критерия выбраковки может быть использован критерий τ (прил. 3). Если критерий τ вычисленный (фактический) больше или равен критерию τ табличному ($\tau_ф \geq \tau_т$) при объеме выборки N и уровне значимости α (0,05 или 0,01), то соответствующие значения вариантов выборки (x) допустимо отбросить как артефакт. Значения τ для вызывающей сомнение величины вычисляются по следующим формулам:

$$\tau_1 = (x_2 - x_1) / (x_{n-1} - x_1) \qquad (1.1)$$

для наименьшего значения переменной величины в вариационном ряду (x_1);

$$\tau_n = (x_n - x_{n-1}) / (x_n - x_2) \qquad (1.2)$$

для максимального значения переменной в вариационном ряду.

П р и м е р. При составлении вариационного ряда по урожайности сельскохозяйственных культур в разрезе хозяйств одного из районов получен следующий ряд значений: 10,8; 12,5; 12,9; 13,2; 20,2 (ц/га). Вызывает сомнение максимальное значение в выборке варианты 20,2. Следует доказать ее принадлежность к артефакту. Подставляем необходимые данные в формулу 1.2:

$$\tau_5 = (x_5 - x_4) / (x_5 - x_2) = (20{,}2 - 13{,}2) / (20{,}2 - 12{,}5) = 0{,}958.$$

Вычисленное значение критерия ($\tau_5 = 0{,}958$) сравнивают с табличным значением ($\tau_т$), учитывая объем выборки ($N = 5$). В прил. 3 критические значения критерия артефакта для $N = 5$ и уровня значимости $\alpha = 0{,}05$ и 0,01 соответственно будут равны 0,807 и 0,916, что меньше расчетного значения ($\tau_5 = 0{,}958$). Поэтому варианту 20,2 признают артефактом и исключают из статистической обработки как сомнительную. Затем приступают к вычислению показателей описательной статистики при условии, что *тип распределения* вариант соответствует *нормальному или логнормальному закону распределения*. В иных случаях с выборкой работают как с *непараметрической*, на которую теория вероятности не распространяется.

При установлении типа распределения принимается следующий порядок действий. Сначала определяется величина классового интервала i, которая зависит от принятого числа классов k и объема выборки N:

$$i = (x_{max} - x_{min}) / k. \qquad (1.3)$$

Число классов в зависимости от объема выборки определяется по формуле:

$$k = 1 + 3{,}3 \lg N. \qquad (1.4)$$

Исходя из формулы (1.4), можно рекомендовать следующее число классов в зависимости от объема выборки:

N	30–50	51–10	101–400	401–1000	1001–2000
k	4–5	6–7	8–9	9–10	11–12

Величина классового интервала должна быть одинаковой на протяжении всего вариационного ряда. Границы классов выбираются такими, чтобы каждый вариант мог быть отнесен только к одному классу. Примеры правильной границы классов: 5–9, 10–14, 15–19 или 5,1–9,1, 9,2–13,2, 13,3–17,3, первый и последний классы могут быть неполными. Границы классов желательно выбирать так, чтобы крайние варианты ряда по возможности оказались ближе к середине интервала своего класса.

П р и м е р. Пусть в выборке объемом $N = 64$ по количеству осадков за время наблюдения $x_{max} = 179$ мм, $x_{min} = 103$ мм. Согласно формуле (1.4), вариационный ряд разбиваем на 8 классов. Затем находим классовый интервал:

$$i = (179 - 103) / 8 = 9{,}5, \text{ или округленно } 10.$$

Исходя из величины классового интервала и минимального значения в выборке, за начало левой границы первого класса удобно принять величину 100. Прибавляя к 100 классовый интервал 10, получаем левые границы последующих классов: 110, 120, 130, 140, 150, 160, 170 мм. Правые границы классов должны отличаться на единицу точности наблюдения от левой границы следующего класса, чтобы граничные значения вариант были отнесены к определенному классу. В нашем примере точность измерения составляет 1,0 мм, поэтому правые границы классов будут следующими: 109, 119, 129, 139, 149, 159, 169, 179 (табл. 1.1).

Срединное значение класса (x) вычисляем сложением границ классов и делением суммы на два. Для первого класса срединное значение равно: (100 + 109) / 2 = 104,5. Срединное значение последующих классов определяется путем последовательного прибавления классового интервала к срединному значению предыдущего класса: 104,5 + 10= 114,5.

Затем производим разноску вариант по классам (подсчитываем количество вариант, вошедших в тот или иной класс в зависимости от их абсолютных величин). Получаем частоту (f) класса (см. табл. 1.1). Сумма частот должна соответствовать объему выборки (64), сумма частостей $f'_ч$ (частота, выраженная в процентах) должна равняться 100 %.

Таблица 1.1

Группировка вариант в классы при дискретной изменчивости признака

Границы класса	Середина класса, x	Частота, f	Частость, $f_ч$, %
100–109	104,5	6	9,37
110–119	114,5	10	15,62
120–129	124,5	12	18,75
130–139	134,5	14	21,87
140–149	144,5	10	15,62
150–159	155,5	6	9,37
160–169	165,5	4	6,25
170–179	175,5	2	3,12
$i = 10$	$k = 8$	$N = 64$	$\sum = 100{,}00$

По частоте и середине класса представим вариационный ряд графически в виде полигона и кривой распределения частот (рис. 1.1)

При построении вариационной кривой по оси абсцисс откладываются значения середины класса, по оси ординат – частоты. При построении гистограммы по оси абсцисс откладываются границы классов, а число вариант каждого класса обозначается высотой или площадью соответствующего прямоугольника. При сравнении изменчивости одинаковых условий или признаков полученные вариационные кривые распределения частот наносятся на один график. Группировка вариант в классы для срав-

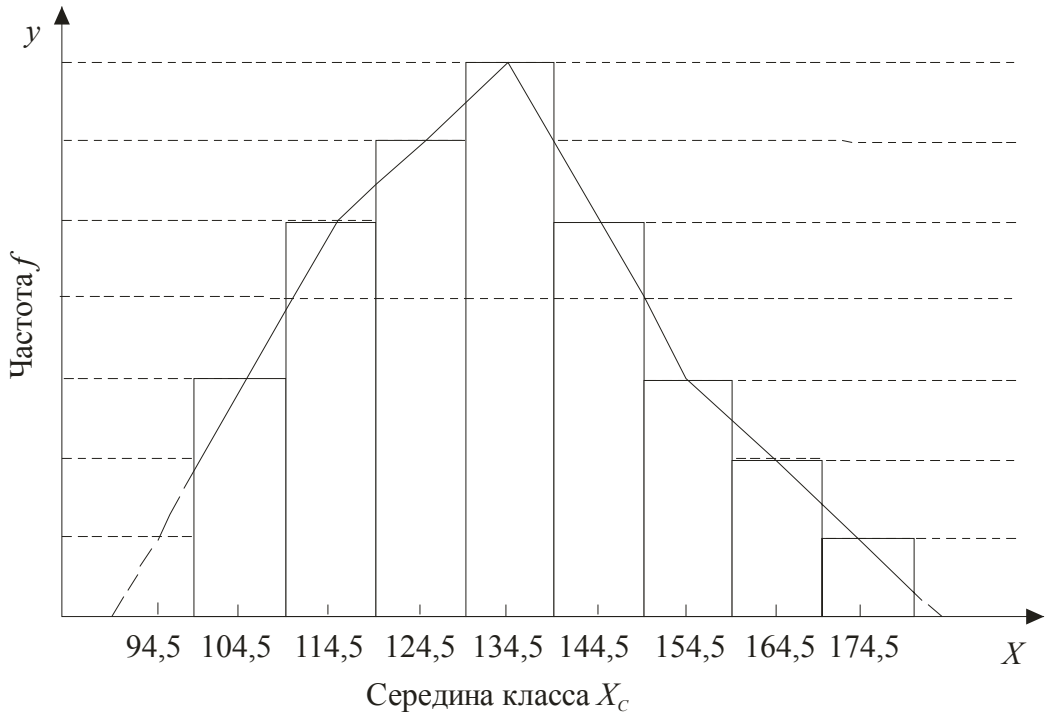


Рис. 1.1. Способы графического представления вариационного ряда: кривая распределения и гистограмма

ниваемых выборок должна быть одинаковой. Если объем выборок не одинаков, все частоты должны быть выражены в процентах от объема выборки по каждой совокупности.

Показатели асимметрии и эксцесса. Распределение частот в изучаемом объекте не всегда подчиняется закону нормального распределения. Это особенно четко проявляется при выражении вариационного ряда в виде графика. Распределение частот может быть представлено асимметричной, островершинной или туповершинной кривой.

Асимметрия кривой распределения обусловлена неравномерным размещением вариант по обе стороны от модального значения признака. Если число вариант больше справа от моды, распределение имеет положительную асимметрию, если слева – отрицательную (рис. 1.2).

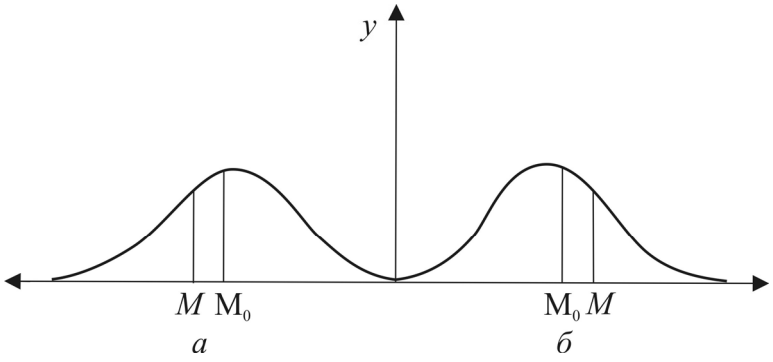


Рис. 1.2. Асимметричное распределение:
а – отрицательная асимметрия, *б* – положительная асимметрия

При получении асимметричной кривой следует проверить асимметричность распределения. Если асимметричность не будет доказана по критерию Стьюдента, то рассматриваемое распределение относят к симметричному. Для проверки асимметричности распределения вычисляют коэффициент асимметрии, его ошибку, затем на основании показателя достоверности устанавливают вид кривой распределения. Коэффициент асимметрии равен:

$$K_{as} = (M - \text{Mo}) / \sigma, \text{ или } K_{as} = (M - \text{Me}) / \sigma.$$

П р и м е р. При изучении содержания подвижного бора в дерново-подзолистых почвах были получены следующие показатели: M = 0,25 мг/кг, Mo = 0,28, σ = 0,02, N = 20. Для получения представления о форме кривой распределения бора предварительно вычисляем коэффициент асимметрии:

$$K_{as} = (0{,}25 - 0{,}28) / 0{,}02 = -1{,}5.$$

Полученная величина указывает на наличие отрицательной асимметрии в распределении вариант содержания подвижного бора в дерново-подзолистых почвах. Затем находим ошибку коэффициента асимметрии:

$$m_{as} = \sqrt{6/(N+3)} = \sqrt{6/(20+3)} = 0{,}51.$$

Достоверность коэффициента асимметрии определяется по критерию Стьюдента: $t = K_{as} / m_{as} = -1{,}5 / 0{,}51 = -2{,}94$.

Величина критерия Стьюдента (см. прил. 4) для $P_{0,99}$ при $v \to \infty$ составляет 2,58 (число степеней свободы принимается равным бесконечности). Рассчитанный критерий Стьюдента (2,94) больше табличного для $P_{0,99}$ (2,58), что указывает на асимметричность распределения подвижного бора. Если бы расчетная величина критерия Стьюдента была меньше табличной, то распределение отнесли бы к симметричному даже при наличии незначительной асимметрии.

Эксцесс кривой распределения (E) имеет место в тех случаях, когда большинство вариантов совокупности сосредоточено около среднего арифметического. Тогда эмпирическая кривая распределения отклоняется от нормальной теоретической кривой у ее вершины и количественно выражается показателем эксцесса (рис. 1.3).

Положительный эксцесс представлен островершинной кривой (эксцессивной, или лептокуртичной) (см. рис. 1.3, *а*), отрицательный – плосковершинной (депрессивной, или платикуртичной) (см. рис. 1.3, *б*). При сильном отрицательном эксцессе кривая может приобрести вид двухвершинной.

Показатель эксцесса определяется по формуле:

$$E = \left[\sum (x - M)^4 / N \cdot \sigma^4 \right] - 3.$$

Вычисляют ошибку коэффициента эксцесса: $m_E = 2\sqrt{6/(N+5)}$. Оценка достоверности показателя эксцесса производится аналогично оценке показателя асимметрии по критерию Стьюдента: $t = E / m_E$.

Оценить достоверность показателей эксцесса и асимметрии можно более простым способом. Отклонение эмпирического ряда по асимметрии и эксцессу от нормального распределения считают существенным,

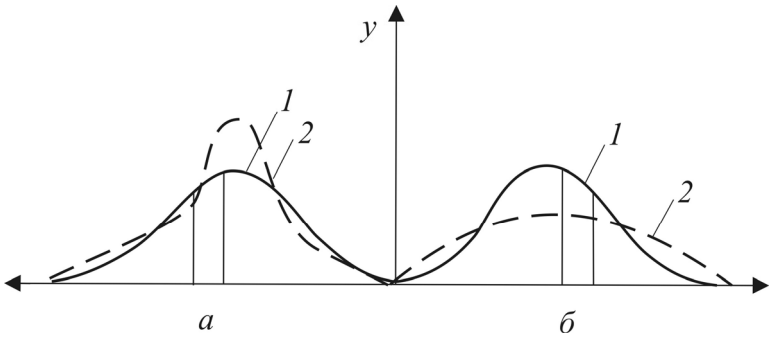


Рис. 1.3. Эксцесс кривой распределения положительный (*а*) и отрицательный (*б*):
1 – теоретическая линия распределения,
2 – эмпирическая линия распределения

если K_{as} и E более, чем в 3 раза, превышают свои ошибки (m_{as}, m_E). Если показатель эксцесса меньше –2, это указывает на наличие в выборке вариант, относящихся к разным совокупностям. Эксцесс считается незначительным, если $|E| < 0{,}4$. Чем меньше показатель эксцесса, тем ближе распределение к нормальному.

Асимметрия и эксцесс эмпирических кривых указывают иногда на важные особенности объекта исследования, например, на изменение признака в ходе усовершенствования технологии на предприятии при выпуске той же продукции. В таких случаях изучение степени и характера асимметрии и эксцесса вариационных кривых может быть самостоятельной задачей при проведении исследовательских работ.

1.3. Показатели описательной статистики

Одна из основных задач статистической обработки – нахождение параметров, представляющих в обобщенном виде распределение данной статистической совокупности. Для решения этих задач используются методы описательной статистики (табл. 1.2.).

Таблица 1.2

Статистические показатели распределения

Показатели	Назначение показателей	Показатели распределения
Центра распределения (средние величины)	Описывают положение середины распределения	*Структурные (непараметрические) средние*: мода (Мо) и медиана (Ме). *Степенные средние*: среднее арифметическое (M), среднее гармоническое ($M_{гар}$), среднее геометрическое ($M_г$), среднее квадратическое ($M_{кв}$), среднее кубическое ($M_{куб}$), среднее взвешенное ($M_{взв}$)
Рассеивания вариант	Описывают степень разброса (вариабельности, изменчивости) вариант	Лимит (lim) Размах варьирования (*ampl*) Среднеквадратическое отклонение (σ) Дисперсия (σ^2) Коэффициент вариации (V) Квантили ($V_{0,25;\ 0,5;\ 0,75}$)
Формы распределения	Описывают симметрию и островершинность распределения вариант около центра	Коэффициент ассиметрии (*as*) Эксцесс (E) Гистограмма Полигон распределения

Показатели центра распределения. Для обоснования представления о генеральной совокупности на основании выборки необходимо использовать наиболее характерные параметры признаков. К ним относятся показатели центра распределения, или среднего положения: мода, медиана, среднее арифметическое, гармоническое, геометрическое, квадратическое, кубическое, взвешенное. Средняя величина выражает характерную, типичную для данного ряда величину признака и является равнодействующей всех факторов, влияющих на признак. В ней погашаются индивидуальные различия вариант в ряду, обусловленные случайными обстоятельствами.

Мода (Мо) представляет собой наиболее часто встречающуюся варианту в вариационном ряду. На графике она соответствует максимальной ординате и находится на вершине вариационной кривой. Если вариационный ряд разбит на классы, то мода соответствует максимальной частоте класса, который называется *модальным*. При полимодальном (многовершинном) распределении вариационный ряд имеет несколько значений моды.

Медиана (Ме) представляет собой среднюю варианту в ранжированном вариационном ряду, которая делит его на две равные части. При нечетном числе вариант середину ряда будет составлять одна варианта (медиана). При четном числе вариант середину ряда образуют две варианты, среднее арифметическое которых будет характеризовать медиану.

При наличии в вариационном ряду сильно отличающихся вариант медиана будет характеризовать середину ряда более точно, чем среднее арифметическое. Мода и медиана используются в тех случаях, когда о выборочных параметрах необходимо иметь ориентировочное представление.

Среднее арифметическое (M, $\bar{x}$) представляет собой величину, сумма положительных и отрицательных отклонений от которой равна нулю. Оно является основной характеристикой статистической совокупности и вычисляется по формуле: $M = \sum x_i / N$, где $\sum x_i$ – сумма всех вариант совокупности. Среднее арифметическое рассчитывается в тех случаях, когда противопоказано вычислять другие средние.

П р и м е р. Определено следующее количество осадков, выпавших в трех пунктах наблюдений: 10, 15 и 20 мм (N = 3). Среднее арифметическое равно: M = (10+15+20) / 3 = 15 мм.

Среднее гармоническое ($M_{гар}$) вычисляется при усреднении меняющихся скоростей процессов (скорость течения воды), показателей обратно пропорциональной зависимости между процессами и явлениями, сложных абсолютных величинах измерений (тонна/километр). Оно рас-

считывается по формуле: $M_{\text{гар}} = N / \sum(1/x_i)$. Его величина меньше средней арифметической. Для вычисления сохраним те же количественные варианты, что и для определения среднего арифметического.

П р и м е р. При измерении скорости воды в реке на трех створах русла получены следующие результаты: 10, 15 и 20 м/с:

$$M_{\text{гар}} = 3 / \sum[(1:10)+(1:15)+(1:20)] = 13,8 \text{ м/с}.$$

Среднее геометрическое ($M_{\text{г}}$) вычисляется в тех случаях, когда в вариационном ряду отдельные значения распределяются в геометрической прогрессии (резко различаются между собой, например 4 и 16). В данном случае среднее геометрическое равно 8. Оно в два раза меньше 16 и в два раза больше 4. Среднее арифметическое из этих вариант 10, т. е. больше среднего геометрического. При наличии нулевой варианты рассчитывается приближенное среднее арифметическое. Если варианты представлены логарифмами чисел (pH и др.), то вычисляют среднее логарифмическое.

П р и м е р. Строение стоит 100 тыс. у. е. Одним лицом оно оценивается в 10 млн, другим – в 1000 млн. С арифметической точки зрения в первом случае получаем ошибку в 90 млн у. е., во втором – в 900 млн у. е. Если оценивать, во сколько раз ошиблись покупатели, то получаем один ответ в обоих случаях – в 10 раз.

Среднее квадратическое ($M_{\text{кв}}$) используется, когда необходима проверка результатов эксперимента на единство суммарного действия (средний радиус или диаметр объекта, площадь земельных участков, функциональных зон и т. д.).

П р и м е р. Имеются данные по величине радиусов трех спилов дуба: 10, 15 и 20 см. Среднее квадратическое будет равно:

$$M_{\text{кв}} = \sqrt{\sum x_i^2 / N} = \sqrt{\sum(10^2 + 15^2 + 20^2)/3} = 15,56 \text{ см}.$$

Среднее кубическое ($M_{\text{куб}}$) применяется при проверке на единство суммарного действия, например при нахождении средней величины объема.

П р и м е р. Кубатура древесины по трем ключевым участкам составляет 10, 15, и 20 м^3. Определяем среднее кубическое по формуле:

$$M_{\text{куб}} = \sqrt[3]{\sum x_i^3 / N} = \sqrt[3]{\sum(10^3 + 15^3 + 20^3)/3} = 16,03 \text{ м}^3.$$

Величина средней кубической максимальна по сравнению с другими средними и находится в ряду справа всех средних: $M_{\text{гар}} < M_{\text{г}} < M < M_{\text{кв}} < M_{\text{куб}}$.

Средневзвешенная ($M_{взв}$). Сгруппированный вариационный ряд по классам иногда называют взвешенным из-за той роли, которую выполняют частоты. Чем больше частота вариант в классе, тем *больший вес* она имеет в характере распределения числового ряда. Среднее арифметическое, рассчитанное в этом ряду, называют *взвешенным средним*:

$$M_{взв} = \sum[(x_1 \cdot f_1) + (x_2 \cdot f_2) + ... + (x_n \cdot f_n)] / \sum f_i,$$

где x_n – варианты; f_i – частоты по классам.

Если совокупность вариант разбита на несколько неравных по численности групп, то среднюю арифметическую вычисляют для каждой группы. Затем их объединяют, определяя *общее среднее* ($M_{общ}$):

$$M_{общ} = \sum M_j \cdot n_j / \sum n_j,$$

где M_j – среднее по группам; n_j – число вариант в группе.

Вычисление ошибки среднего приведено в п. 1.4.

Показатели рассеивания вариант. Для характеристики распределения в вариационном ряду недостаточно лишь средней арифметической. В совершенно разных по величине вариантах двух выборок средняя может быть одной и той же:

$$-100; -20; 100; 20; \ M = 0,$$
$$0{,}1; -0{,}2; 0{,}1; \ M = 0.$$

Для получения более полного представления о выборочных совокупностях используют показатели рассеяния вариант, или разнообразия признаков: лимит, размах варьирования (амплитуда), среднеквадратическое (стандартное) отклонение, средний квадрат отклонений (дисперсия), коэффициент вариации, квантили. Эти показатели признаков характеризуют различную степень и особенности разброса.

Лимит указывает границы вариационного ряда: $\text{Lim} = x_{max} \div x_{min}$.

Амплитуда (вариационный размах, размах варьирования) – разность между максимальным и минимальным значениями вариант: $\text{Ampl} = x_{max} \div x_{min}$.

Чем ближе минимальные и максимальные варианты к среднему и чем меньше амплитуда, тем меньше степень разнообразия между переменными в вариационном ряду, тем надежнее характеризуют статистические показатели искомую закономерность.

Более точно степень разнообразия признака следует характеризовать другими показателями. Среднеквадратическое отклонение и дисперсию используют как составляющие параметры нормального распределения при вычислении ряда параметрических статистических критериев.

Среднеквадратическое отклонение, или сигма (σ), показывает степень рассеяния значений статистической совокупности около среднего

значения, а точнее интервал (M±σ), в который входит до 75 % вариант выборочной совокупности. Считается, если 75 % вариант выборки находится в пределах M±σ, то это соответствует норме (стандартному отклонению); если в пределах M±2σ, то имеется незначительное отклонение от нормы; если выходит за пределы M±3σ, то можно утверждать о наличии анормального явления, процесса. Величина сигмы прямо пропорционально зависит от разброса вариант в вариационном ряду. Чем больший разброс, тем больше значение сигмы. Однако пределы колебаний не дают оценки разброса, как и дисперсия, независимо от его величины.

Среднеквадратическое отклонение используется:

- для оценки данных одноименных вариационных рядов при близких средних: чем больше сигма, тем больший разброс вариант в совокупности, соответственно среднее арифметическое менее типично для данного ряда;
- оценки типичности среднего арифметического в ряду, используя правило трех сигм (M±3σ);
- определения доверительных интервалов статистических коэффициентов и репрезентативности выборочных исследований.

Недостаток сигмы, как и дисперсии, заключается в том, что критерий представляет собой абсолютную именованную величину, поэтому его нельзя использовать при сравнении разнородных рядов, выраженных в различных единицах измерения. Для этой цели подходит коэффициент вариации.

Среднеквадратическое отклонение можно определить двумя путями:

$$\sigma = \sqrt{\sum (x_i - M_x)^2 / (N-1)}, \qquad (1.5)$$

$$\sigma = (x_{\max} - x_{\min}) / 6, \qquad (1.6)$$

где (x_i – M_x) – отклонение от среднего индивидуальных вариант; N – объем выборочной совокупности.

Формулу (1.6) можно использовать для приближенного расчета сигмы. Алгебраически сигма представляет собой корень квадратный из дисперсии.

П р и м е р. Получены следующие данные по относительной высоте холмов в пределах моренно-эрозионного ландшафта в метрах: 20, 20, 22, 23, 24, 25, 25, 26, 27, 28, 30.

Для расчета сигмы составляем табл. 1.3 исходных данных. Подставив в формулу (1.5) данные, определяем сигму: $\sigma = \sqrt{100{,}85/10} = 3{,}17$ м. Среднее арифметическое равно 24,54 м. Если значение сигмы 3,17 прибавить к среднему арифметическому и вычесть ее из него, то определим граничные значения, в которых будет находиться определенная часть вариант (до 75 %) исследуе-

мой статистической выборки (24,54 ± 3,17). В этот интервал (от 21,37 до 27,71) вошли варианты 22, 23, 24, 25, 25, 26, 27. Это означает, что 68 % вариант в выборке находится в пределах от 21,33 (24,54 – 3,17) до 27,71 м (24,54 + + 3,17 = 27,71). Лишь 32% вариант выходит за указанные пределы.

Таблица 1.3

Форма записи и расчета среднеквадратического отклонения

x_i	$x_i - M_x$	$(x_i - M_x)^2$	x_i	$x_i - M_x$	$(x_i - M_x)^2$
20	–4,54	20,61	26	1,46	2,27
20	–4,54	20,61	27	2,46	6,05
22	–2,54	6,45	28	3,46	11,97
23	–1,54	2,37	30	5,46	29,81
24	–0,54	0,29			
25	0,46	0,21	$\sum x_i = 270$	$\sum = -0,06$	$\sum (x_i - M_x)^2 = 100,85$
25	0,46	0,21	$M = 24,54$		

Вычисление ошибки сигмы приведено в п. 1.4.

Средний квадрат отклонений, или дисперсия, указывает колебание значений признака внутри выборочной совокупности через отклонение всех вариант от среднего значения, т. е. показывает интервал, в который входят все варианты выборки (100 %). Однако сумма всех отрицательных и положительных отклонений от среднего равна нулю. Поэтому все отклонения от среднего возводятся в квадрат и суммируются: $\sum (x_i - M_x)^2$. При усреднении всех отклонений числового ряда путем деления на $(N - 1)$ получаем средний квадрат отклонений, или дисперсию (D, σ^2).

Если вычислена сигма (σ), то дисперсию получаем путем возведения ее в квадрат: σ^2.

При упрощенном способе расчета дисперсии не вычисляют отклонений вариант от среднего $(x_i - M_x)$, используя следующий расчет:

$$\sigma^2 = \sum x_i^2 / N - M^2,$$

где $\sum x_i^2$ – сумма квадратов всех вариант выборки; M^2 – квадрат среднего арифметического; N – число вариант в выборке.

Более точно значение дисперсии вычисляется с использованием данных в табл. 1.3 по формуле:

$$\sigma^2 = \sum (x_i - M_x)^2 / (N - 1). \qquad (1.7)$$

Средний квадрат отклонений выражается в тех же единицах, что и варианты. Форма записи исходных данных для вычисления дисперсии

такая же, как и для сигмы (см. табл. 1.3). Подставив значения из таблицы в формулу, получим значение дисперсии: $\sigma^2 = 100{,}85 / 10 = 10{,}08$ м.

Исходя из величины дисперсии, можно определить интервал, в пределы которого входят все варианты выборки: $M \pm \sigma^2$, от 14,5 м (24,5 – 10,0) до 34,5 м (24,5 + 10,0). В этот интервал вошли 100 % вариант выборочной совокупности.

При объединении нескольких аналогичных выборок в общую выборку можно рассчитать общий средний квадрат отклонений, если имеются сведения о дисперсии по каждой выборке в отдельности:

$$\sigma^2_{\text{общ}} = \sum (N_i - 1) \cdot \sigma_i^2 / \left(\sum N_i - k \right), \qquad (1.8)$$

где σ_i^2 – дисперсия индивидуальной выборки; N_i – объем частных выборок; k – число частных выборок.

П р и м е р. Вычислим общий средний квадрат отклонений для четырех выборок, отражающих содержание кальция в озерных водах Беларуси: $\sigma_1^2 = 2$; $N_1 = 8$; $\sigma_2^2 = 2{,}5$; $N_2 = 6$; $\sigma_3^2 = 3{,}0$; $N_3 = 7$; $\sigma_4^2 = 3{,}5$; $N_4 = 8$. По формуле (1.8) имеем:

$$\sigma^2_{\text{общ}} = \frac{(8-1) \cdot 2 + (6-1) \cdot 2{,}5 + (7-1) \cdot 3 + (8-1) \cdot 3{,}5}{(8+6+7+8) - 4} = 2{,}76.$$

Если извлечь корень квадратный из полученной величины, получим общее среднеквадратическое отклонение, или сигму ($\sigma_{\text{общ}} = 1{,}66$).

Практическое применение дисперсии следующее:

- оценка вариабельности рядов распределения;
- факторный и дисперсионный анализ;
- статистическая оценка двух совокупностей по критерию Фишера.

Дисперсия выражается в тех же единицах, что и показатели выборки.

Коэффициент вариации представляет собой относительный показатель разнообразия признаков, выражается в процентах. Он показывает отношение среднеквадратического отклонения к средней арифметической:

$$V = (\sigma / M) \cdot 100. \qquad (1.9)$$

В случаях, когда значение среднеквадратического отклонения не рассчитывается, величина коэффициента вариации может быть определена следующим образом:

$$V = 100 \sqrt{\frac{\sum x_i^2 / (M^2 - N)}{N - 1}}, \qquad (1.10)$$

где $\sum x_i^2$ – сумма квадратов индивидуальных вариант в совокупности.

Чем меньший по размаху варьирования будет признак, тем меньший будет коэффициент вариации для данной совокупности. Соответственно меньшими будут сигма и дисперсия.

Коэффициент вариации позволяет оценить вариабельность (разброс) признака в нормированных границах. Если его значение меньше 10 %, то разброс вариант относительно средней арифметической считается слабым, при 10–30 % – средним, 30–60 % – высоким, 60–100 % – очень высоким, более 100 % – аномальным.

О преимуществе использования коэффициента вариации при оценке разнородных признаков можно судить по табл. 1.4.

Таблица 1.4

Сравнительная оценка состава работников предприятия

Учетный признак	Среднее арифметическое, M	Среднеквадратическое отклонение, σ	Коэффициент вариации, V
Стаж работы (лет)	8,7	2,8	32,1
Возраст (лет)	37,2	4,1	11,0
Образование (класс)	9,2	1,1	11,9

В табл. 1.4 абсолютные величины средних и сигмы близки по стажу работы и образованию. Однако по коэффициенту вариации сходны по возрасту и образованию. В данном случае сравнение по сигме проводить некорректно, так как все три признака разнородны и не сравнимы между собой. Выручает неименованный коэффициент вариации, который позволяет оценить разброс признака в нормированных границах.

Коэффициент вариации нельзя рассчитывать при наличии вариант признака с отрицательным числом (отрицательные температуры, отметка поверхности ниже уровня воды в океане и др.). В таких случаях коэффициент вариации рекомендуется вычислять по формуле с учетом модуля:

$$V = 100\sigma / |x_i| + M, \qquad (1.11)$$

где $|x_i|$ – модуль наименьшей отрицательной величины без учета знака.

В данном случае имеется в виду, что при вычислении коэффициента вариации среднее арифметическое и среднеквадратическое отклонения должны быть представлены отрезками на числовой оси. Приведем алгоритм вычисления коэффициента вариации для величин с разными знаками.

Пример. Температура воздуха в течение суток в октябре составила (в градусах Цельсия): –4, –3, –1, +1, +3. Среднее арифметическое равно –0,6, среднеквадратическое отклонение – 1,95. Если не учитывать наличия интер-

вальной шкалы и определять коэффициент вариации по формуле (1.9), то получим следующую величину: $V = (1{,}95 \cdot 100) / (-0{,}6) = -325$ %. Результаты противоречат исходным данным, которые фактически характеризуются небольшим размахом варьирования температур в течение суток. Если среднее арифметическое представить как отрезок от точки -4 до $-0{,}6$, то оно будет равно: $|-4| + (-0{,}6) = 3{,}4$. Используя формулу (1.11), получаем коэффициент вариации, соответствующий условиям задачи: $V = (100 \cdot 1{,}95) / (|-4| + (-0{,}6)) = 54{,}16$ %.

Квантили. В открытых вариационных рядах и рядах распределения качественных признаков для сжатого описания распределений используется другой параметр разброса – *квантиль* (синонимы: перцетиль, персентиль). Этот параметр может использоваться для перевода количественных признаков в качественные. В практике статистического анализа наиболее часто используются следующие квантили:

$V_{0,5}$ – медиана;

$V_{0,25}$, $V_{0,75}$ – квантили четверти, соответственно нижняя и верхняя квантиль;

$V_{0,1}$, $V_{0,2}$, …, $V_{0,9}$ – децили (десятые);

$V_{0,01}$, $V_{0,02}$, …, $V_{0,99}$ – процентили, или центили (сотые).

Квантили делят область возможных изменений вариант в выборке на определенные интервалы. Статистическая суть квантилей лучше раскрывается при построении графика.

1.4. Оценка статистических параметров по выборочным данным

Оценка в статистике – это правило вычисления оцениваемого параметра. Она указывает приближенное значение показателей выборки относительно этих параметров генеральной совокупности. По мере увеличения числа наблюдений выборочные средние и другие параметры все больше приближаются к этим значениям генеральной совокупности. Степень соответствия показателей оценивается *ошибкой* (m). Ее запись производится вместе с оцениваемым параметром, например, $M \pm m_M$, $\sigma \pm m_\sigma$, $V \pm m_V$. Ошибка указывает интервал, в пределах которого находится этот показатель в генеральной совокупности. Чем меньше ошибка, тем ближе значение выборочного показателя к этому показателю генеральной совокупности. Чем больше число наблюдений и чем однороднее выборка, тем меньшая ошибка среднего и других показателей. Расчеты ошибок параметров в дальнейшем будут приводиться после характеристик самих параметров. Здесь покажем расчеты ошибок важнейших статистических параметров.

Представление средней арифметической выборки приводится обязательно с ее ошибкой. Стандартная ошибка средней рассчитывается:

$$m_M = \sqrt{\frac{\sum (x_i - M_x)^2}{N(N-1)}}, \text{ или } m_M = \sqrt{\frac{\sigma^2}{N}}, \text{ или } m_M = \frac{\sigma^2}{\sqrt{N}}. \quad (1.12)$$

Ошибка среднеквадратического отклонения определяется по формуле:

$$m_\sigma = \sigma / \sqrt{2(N-1)}. \quad (1.13)$$

Ошибка дисперсии вычисляется путем возведения в квадрат ошибки среднеквадратической.

Ошибка коэффициента вариации рассчитывается следующим образом:

$$m_V = \frac{V}{\sqrt{N}} \cdot \sqrt{\frac{1}{2} + (V / 100)^2}. \quad (1.14)$$

Поскольку параметр *m* характеризует ошибку утверждения (прогноза) о том, что выборочное среднее равно генеральному среднему, то чем выше требование к вероятности этого вывода, тем шире должен быть обеспечивающий точность такого прогноза интервал, называемый *доверительным интервалом*. Его величина задается вероятностью безошибочного прогноза, которую принято *называть доверительной вероятностью* (*уровень вероятности,* надежность опыта, вероятность безошибочного прогноза). В исследованиях допускается доверительная вероятность (*P*) не менее 95 % (0,95 частей от 1). В этих случаях *P* для средних арифметических при достаточно большом числе наблюдений ($N > 30$) равен $\pm 2m$. Предельная ошибка выборки $\Delta = M \pm 2m$. При доверительной вероятности 99 % (0,99) доверительный интервал составит $\pm 3m$, $\Delta = M \pm 3m$. По иному в отношении доверительного интервала можно сказать так: *он показывает, какой процент вариант выборки* (*выборок*) *подтверждает искомую статистическую закономерность*.

Каждому значению доверительной вероятности соответствует свой *уровень значимости* (α). Он выражает вероятность нулевой гипотезы: вероятность того, что выборочная и генеральная средние не отличаются друг от друга. Иначе говоря, чем выше уровень значимости, тем меньше можно доверять утверждению, что различия существуют, т. е. *он показывает, какой процент вариант совокупности* (*выборок*) *отвергают искомую статистическую закономерность*. Уровень значимости 5 % (0,05) дополняет доверительную вероятность 95 % (0,95). В сумме они составляют 100 % (1). Если доказано подобие между выборками при α = 5 % (0,05), то из этого следует, что до 5 % вариант выборки подобие не подтверждают. В таблицах приложения приводятся численные значения для

P или α соответственно: 0,95 и 0,99; 0,05 и 0,01. В этих случаях при интерпретации мы можем утверждать нулевую гипотезу (H_0). При более высоких уровне вероятности 0,99 и уровне значимости 0,01 мы получаем сильный довод для утверждения нулевой гипотезы.

Проверка статистических гипотез. Методологической основой любого исследования является формулировка рабочей гипотезы. В ходе исследования рабочая гипотеза либо принимается, либо отвергается. Статистической называют гипотезу о виде неизвестного распределения или о параметре распределения. Примеры гипотез:

- генеральная совокупность распределяется по закону Пуассона;
- средние арифметические двух совокупностей не равны между собой;
- дисперсии двух совокупностей равны между собой.

Выдвинутую гипотезу называют *основной или нулевой* (H_0). Гипотезу, которая противоречит нулевой, называют *конкурирующей или альтернативной* (H_1). Если нулевая гипотеза предполагает, что $M = 20$, то логическим отрицанием будет $M \neq 15$. Простая гипотеза содержит одно предположение, сложная состоит из конечного или бесконечного множества простых гипотез. Выдвинутую гипотезу проверяют на правильность ее статистическими методами, т. е. проводят статистическую проверку. При проверке могут быть допущены ошибки двух родов.

Ошибка первого рода – отвергается правильная гипотеза. Вероятность совершить ошибку первого рода называют *уровнем значимости* (α). Это значит, что в 5 случаях из 100 мы рискуем допустить ошибку первого рода.

Ошибка второго рода – принимается неправильная гипотеза, значимость ошибки которой допускается 0,95 и выражается символом P. Это значит, что в 95 случаях из 100 мы рискуем допустить ошибку второго рода.

Для проверки нулевых гипотез применяют статистические критерии. При сравнении дисперсий используют критерий Фишера. В большинстве исследований для статистической проверки гипотез существенности различий средних арифметических используют параметрический критерий Стьюдента. Если нулевая гипотеза принимается, это не означает ее доказательство. Доказать на основании однократной или косвенной проверки гипотезу нельзя, а опровергнуть можно. Для повышения точности статистических данных необходимо уменьшить вероятности ошибок первого и второго рода, увеличить объем выборок. Область применения того или иного критерия задается законом его распределения.

Оценка точности опыта. При исследованиях *методического* характера необходимо приводить их оценку по показателю *точность опыта*

(p). Его смысл состоит в установлении величины ошибки среднего арифметического (m_M) в процентах от величины среднего арифметического (M). Показатель точности опыта можно определить по одной из двух формул:

$$p = (m_M / M) \cdot 100; \quad p = V / \sqrt{N}, \tag{1.15}$$

где V – коэффициент вариации.

Опыт считается достаточно точным, если $p < 3$ %, удовлетворительным – при его величине 3–5 %. Если величина точности опыта более 5 %, к полученным выводам следует относиться осторожно и увеличить число повторностей в опыте. Эти градации обязательны для полевых опытов с растениями. Некоторые приборы для анализа могут давать значительно большую погрешность (p до 15 %).

Ошибка показателя точности опыта вычисляется следующим образом:

$$m_p = \pm\ p\ \sqrt{(1/2N) + (p/100)^2}. \tag{1.16}$$

П р и м е р. Среднее арифметическое общей биомассы многолетних трав в луговом ландшафте прирусловой поймы $M = 235$ ц/г, ошибка среднего арифметического $m_M = \pm 4$ ц/га, $N = 20$. Используя формулу (1.15), выполним расчет показателей:

$$p = (4 / 235) \cdot 100 = 1{,}7\ \%.$$

Полученная величина достаточно точная.

1.5. Теоретические функции распределения

В ходе работы с выборочной совокупностью иногда возникает необходимость описать вариационную кривую с помощью математической функции. Для характеристики вариационной кривой можно подобрать ряд математических зависимостей. Выбирают ту, которая наиболее реально отражает сущность объекта исследования. Выбор математической зависимости, описывающей распределение, проводится путем подбора подходящей математической модели, которая определяет вид функции распределения. Затем находят параметры функции и проверяют ее соответствие эмпирическому распределению.

В географии большинство закономерно повторяющихся явлений, процессов можно представить в виде *нормального и логнормального распределения*. Реже встречается *биномиальное распределение, распределение Пуассона и другие.*

Биномиальное распределение (распределение Бернулли) возникает, когда оценивается сколько раз происходит событие в серии определенного числа независимых, выполняемых в одинаковых условиях на-

блюдений. Разброс вариант – следствие влияния ряда независимых и случайно сочетающихся факторов (есть событие или его нет). Характерно для альтернативного типа изменчивости признака.

Распределение Пуассона рассматривается как предельный случай биномиального распределения и используется для характеристики редких событий. Отличительная особенность распределения Пуассона – величина дисперсии близка к величине среднего арифметического, например длительное наводнение. Это проявляется в ситуациях, когда в определенный отрезок времени или на определенном пространстве происходит случайное число каких-либо событий, например длительно повторяющиеся ураганы в течение одного летнего периода. На графике это распределение представляется в виде резко выраженной асимметрии.

Рассмотрим более детально наиболее характерные типы теоретических распределений в природе и обществе: нормальное и логнормальное.

Нормальное распределение. Нормальное (распределение Гаусса) используется для приближенного описания явлений, которые носят вероятностный, случайный характер. Приоритет в открытии этого закона принадлежит Де Муавру (1733), но его связывают с именем Гаусса, исследовавшего его в начале XIX в.

Распределение Гаусса имеет место среди природных и экономических явлений. В системе признак варьирует под влиянием большого количества взаимно независимых факторов, каждый из которых мало влияет на его общую вариабельность. Причем одни факторы приводят к возрастанию величины признака, другие – к уменьшению. Встречаемость вариантов, занимающих середину совокупности, максимальна. Такое распределение считается нормой для случайных величин, поэтому оно получило название нормального. Графически нормальное распределение выражается плавной симметричной куполообразной кривой с приближающимися к оси абсцисс ветвями (кривая плотности нормального распределения) (рис. 1.4).

Кривая показывает, что большие отклонения от средней встречаются реже, чем малые. С уменьшением среднего квадратического отклонения (σ) кривая нормального распределения становится все более островершинной. Площадь, заключенная под кривой нормального распределения, всегда принимается равной единице.

При нормальном распределении среднее, мода и медиана совпадают. Кривая плотности не пересекает оси абсцисс, что подтверждает вероятность существования неограниченно больших отклонений. Уравнение нор-

мального распределения можно записать в нескольких модификациях. Наиболее часто используется следующая формула:

$$f' = \frac{N_i}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x_i - M)^2}{2\sigma^2}}, \qquad (1.17)$$

где f' – искомая ордината кривой (теоретическая частота); в степень числа e входит величина $(x_i - M_x)/\sigma$, получившая название нормированного отклонения.

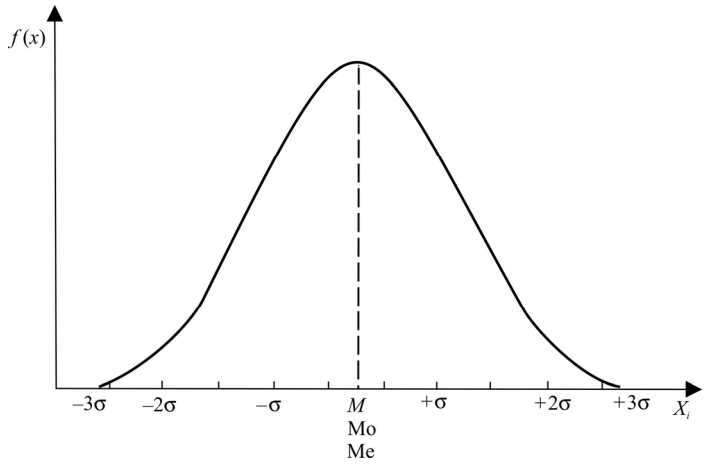


Рис. 1.4. Кривая нормального распределения

Подставив необходимые значения по исследуемой статистической совокупности в формулу (1.17), рассчитаем теоретические частоты нормального распределения f' для каждого класса совокупности. Получим ряды теоретических (f') и эмпирических (f) данных:

f	8	17	34	70	141	165	253	187	145	85	51	25	19
f'	7	16	39	79	131	180	206	196	154	101	55	24	13

Были приняты следующие исходные данные для расчета: N = 1200, M = 10,22, σ = 2,31, i = 13.

Произведем проверку соответствия эмпирических частот вычисленным частотам нормального распределения. Для этого, используя критерий χ^2, составляем таблицу по форме:

f	f'	$f - f'$	$(f - f')^2$	$(f - f')^2 / f'$

Сумма показателей в последнем столбце будет составлять величину χ^2, равную 21,184. Полученная величина сравнивается со стандартной

величиной χ^2 (см. прил. 6) при числе свободы: $\nu = i - 3 = 13 - 3 = 10$ (см. выше ряды по частотам).

Табличные значения χ^2 следующие: для $P = 0{,}95$ и $0{,}99$ $\chi^2 = 18{,}307$ и $23{,}209$ соответственно. Рассчитанное значение $\chi^2 = 21{,}184$ находится между указанными табличными значениями. Поскольку расчетное значение χ^2 не превышает табличной величины при $P = 0{,}99$, можно считать, что эмпирическое распределение признака удовлетворительно подчиняется нормальному закону распределения.

При нормальном распределении около 68,3 % всех вариант отклоняется от среднего значения не более, чем на величину среднего квадратического отклонения ($\pm\sigma$). Соответственно в пределах от -2σ до $+2\sigma$ находится 95,5 % вариант, в пределах от -3σ до $+3\sigma$ – 99,7 %.

Отклонение вариант от нормального закона распределения указывает на влияние какого-либо другого фактора на статистическую совокупность.

Логнормальное распределение. Некоторые распределения при изучении географических объектов имеют выраженную асимметрию, поэтому представляет практический интерес преобразование асимметричного распределения в симметричное (нормальное). Иногда это возможно, если каждую варианту выборки выразить в виде логарифма ($\lg x_i$). В тех случаях, когда логарифм случайной величины (x_i) подчиняется нормальному распределению, а сами значения случайных величин распределены асимметрично, распределение случайной величины принято называть логарифмически нормальным, или логнормальным. Уравнение логнормального распределения имеет вид:

$$f' = \frac{1}{\sigma_{x_i}} \cdot \frac{e^{\frac{-(\lg x_i - M)^2}{2\sigma^2}}}{\sqrt{2\pi}}.$$

Например, к логнормальному распределению можно отнести распределение микроэлементов в почвах, породах.

1.6. Статистические критерии различия

Проведение географических исследований предполагает не только изучение строения, развития, закономерностей распространения исследуемых объектов, явлений, но и установление сходства или различия между одноименными генеральными совокупностями изучаемых систем. Это зависит от условий, в которых протекает один и тот же процесс. Сопряженный анализ одноименных признаков в выборках используется для классификации и районирования по одному или нескольким параметрам.

При этом возникает необходимость применения объективного метода выделения классификационных групп или районов на основе методов математической статистики с использованием критериев достоверности. Если достоверность различия между выборочными совокупностями доказана, то генеральные совокупности, сравниваемые по какому-либо признаку, выделяют как самостоятельные. В случае отсутствия достоверных различий их объединяют в одну группу.

Различие между двумя выборками устанавливается с помощью ряда критериев: t – распределение Стьюдента, наименьшего существенного различия (НСР), F – распределения Фишера, критерия соответствия (χ^2).

Каждый из критериев применяется при определенных условиях, которые задаются целью исследования. Несоблюдение указанных условий может привести к ошибочным выводам.

Прежде чем приступать к статистической обработке и расчету критериев различия, следует убедиться в отсутствии артефакта в сравниваемых выборках. Если в малых совокупностях распределение нормально, то для установления артефакта достаточно использовать правило трех сигм. Согласно этому правилу в пределах M±3σ находится 99,7 % всех вариант выборки. Если крайние варианты попадают в этот интервал, то они включаются в статистическую выборку, так как не являются артефактом. Наличие артефакта можно проверить по формулам (1.1, 1.2).

Критерий Стьюдента. Используется для оценки сходства или различия между выборочными совокупностями по разности величин их средних арифметических ($d = M_{\text{большая}} - M_{\text{меньшая}}$) и ее отношения к ошибке этой разности (m_d) при условии распределения вариант в группах по закону нормального распределения и подтверждает равенство разброса вариант в выборке (близкие дисперсии сравниваемых выборок). Не допускается применения критерия в случае балльного характера сравниваемых числовых признаков.

Выбор конкретной методики оценки различий по критерию Стьюдента зависит от учета следующих особенностей выборочных совокупностей: сравниваются средние арифметические в *независимых* (*несвязанных*) выборках; устанавливаются различия в *сопряженных* (*парных*) *выборках*, а также между выборочными и генеральными средними (теоретическими стандартами).

Независимые статистические совокупности могут быть получены на одном или нескольких объектах, но *при одинаковых условиях* проведения эксперимента: например, измерение температуры воздуха в январе в г. Бресте на протяжении нескольких лет и установление достоверных различий между этими показателями по годам исследований; сравнение экономиче-

ского показателя в хозяйстве или на предприятии по пятилеткам между собой; сравнение чистого дохода в хозяйствах с одинаковым экономическим развитием, но расположенных на значительном расстоянии. При сравнении независимых выборочных совокупностей объемы выборок могут быть одинаковые ($N_1 = N_2$) или разные ($N_1 \neq N_2$). В двух сравниваемых независимых выборках с одинаковым или разным объемом наблюдений *степень свободы* определяется по формуле: $\nu = (N_1 - 1) + (N_2 - 1) = N_1 + N_2 - 2$.

При *малых объемах независимых совокупностей, если дисперсии сравниваемых выборок нельзя считать одинаковыми*, число степеней свободы определяется сложнее:

$$\nu = \frac{1}{u^2/(N_{x_1}-1) + (1-u)^2/(N_{x_2}-1)},$$

где $u = m_{x_1}^2 / (m_{x_1}^2 + m_{x_2}^2)$; m_{x_1} и m_{x_2} – ошибки среднего арифметического первой и второй выборок соответственно.

Сопряженные статистические совокупности получают на одном или на разных объектах, но в разных условиях. Например, сравнение температуры воздуха в июле и январе г. Могилева; сравнение прибыли фермерских и подсобных хозяйств в любом районе или фермерских хозяйств Витебской и Гомельской области. Объем сравниваемых выборок должен быть одинаков ($N_1 = N_2$). Определение *степени свободы* для *сопряженных* выборок определяется как: $\nu = N_{\text{пар}} - 1$.

Ошибка разности между средними выборок (m_d) в зависимости от вида наблюдений (независимые, сопряженные) и объема наблюдений рассчитывается по разным формулам. Рассмотрим их ниже.

Вариант первый. Сравниваемые выборки имеют одинаковый объем наблюдений ($N_1 = N_2$) и *независимы*:

$$m_d = \sqrt{m_{x_1}^2 + m_{x_2}^2}, \qquad (1.18)$$

где m_{x_1} и m_{x_2} – ошибка средней арифметической первой и второй выборки.

Критерий Стьюдента определяют по формуле:

$$t = d / m_d = (M_{\text{большая}} - M_{\text{меньшая}}) / m_d. \qquad (1.19)$$

Сопоставляя вычисленный критерий Стьюдента с табличным, устанавливают или отвергают с некоторой долей уверенности различия между средними арифметическими выборок.

П р и м е р. При исследовании глубины расчленения рельефа в северной (x_1) и центральной (x_2) провинциях Беларуси необходимо установить, объединять их в один геоморфологический район по степени расчленения рельефа или различать их как самостоятельные. Исходные данные и их обработка при-

водятся в табл. 1.5. Из полученной информации по средним арифметическим (M_{x_1} = 16,6 и M_{x_2} = 15,2 м) различие по глубине расчленения рельефа можно признать как существенным, так и несущественным. Для объективных выводов используем критерий Стьюдента.

Таблица 1.5

Форма обработки вариант в независимых совокупностях

X_{i_1}	$X_{i_1} - M_{x_1}$	$(X_{i_1} - M_{x_1})^2$	X_{i_2}	$X_{i_2} - M_{x_2}$	$(X_{i_2} - M_{x_2})^2$
20	3,4	11,56	17	1,8	3,24
17	0,4	0,16	16	0,8	0,64
16	–0,6	0,36	15	–0,2	0,04
15	–1,6	2,56	14	–1,2	1,44
15	–1,6	2,56	14	–1,2	1,44
$\sum = 83$	0	17,20	76	0	$\sum = 6,80$
$M_{x_2} = 16,6$			$M_{x_2} = 15,2$		

Определяем разницу между средними: d = 16,6 – 15,2 = 1,4. Ошибки средних по каждой выборке равны:

$$m_{x_1} = \sqrt{\sum (x_{i_1} - M_{x_1})^2 / N_{x_1}(N_{x_1} - 1)} = \sqrt{\sum (17,2)/5(5-1)} = 0,93;$$

$$m_{x_2} = \sqrt{\sum (x_{i_2} - M_{x_2})^2 / N_{x_2}(N_{x_2} - 1)} = \sqrt{6,8/20} = 0,58.$$

Ошибка разности средних составляет:

$$m_d = \sqrt{m_{x_1}^2 + m_{x_2}^2} = \sqrt{0,93^2 + 0,58^2} = 1,20.$$

Полученные данные подставляем в формулу (1.19) и вычисляем $t_ф$ = 1,4 / 1,2 = = 1,17. Число степеней свободы $\nu = N_{x_1} + N_{x_2} - 2$ = 5 + 5 – 2 = 8.

Сопоставляем табличные значения критерия Стьюдента 2,31 и 3,36 (см. прил. 4) при P = 0,95 и 0,99 для степени свободы ν = 8 с фактическим (расчетным) $t_ф$ = 1,17. Поскольку $t_т$ (2,31 и 3,36) > $t_ф$ (1,17) при обоих уровнях значимости, то разность между средними признается недостоверной (несущественной). При выделении геоморфологических районов по глубине расчленения рельефа их объединяют.

Вариант второй. Сравниваемые *независимые* совокупности имеют различие по объему ($N_1 \neq N_2$). Порядок вычисления критерия Стьюдента такой же, как и при установлении достоверности в независимых выборках с одинаковым числом наблюдений. Различие состоит в вычислении по другой формуле ошибки разности средних:

$$m_d = \sqrt{\frac{\left(\sum (x_{i_1} - M_{x_1})^2 + \sum (x_{i_2} - M_{x_2})^2\right) \cdot (N_{x_1} + N_{x_2})}{(N_{x_1} + N_{x_2} - 2) \cdot (N_{x_1} \cdot N_{x_2})}}. \qquad (1.20)$$

Вариант третий. Сравниваемые *сопряженные* совокупности имеют одинаковый объем выборки ($N_1 = N_2$). Ошибка разности средних определяется по формуле:

$$m_d = \sqrt{\frac{\sum (d_i - d)^2}{N_{\text{пар}}(N_{\text{пар}} - 1)}}. \qquad (1.21)$$

Обозначения для формул (1.20) и (1.21): x_{i_1} и x_{i_1} – индивидуальные значения вариант первой и второй выборок соответственно; M_{x_1} и M_{x_2} – средние первой и второй выборочной совокупности соответственно; N_{x_1} и N_{x_2} – объем выборки первой и второй соответственно; d_i – разность между индивидуальными сопряженными вариантами в выборках; d – разность между средними сопряженных выборок.

П р и м е р для сопряженных наблюдений. Сравним глубину расчленения рельефа в пределах конечно-моренного (x_1) и донно-моренного (x_2) ландшафта. Для обработки данных составляем исходную табл. 1.6.

Таблица 1.6

Форма обработки данных сопряженных наблюдений

X_{i_1}	X_{i_2}	d_i	d_i^2	$d_i - d$	$(d_i - d)^2$
20	17	3	9	+1,6	2,56
17	16	1	1	–0,4	0,16
16	15	1	1	–0,4	0,16
15	14	1	1	–0,4	0,16
15	14	1	1	0,4	0,16
$\sum = 83$	$\sum = 76$	$\sum = 7$	$\sum = 13$	$\sum = 0$	$\sum = 3,20$
$M_{x_1} = 16,6$	$M_{x_2} = 15,2$				
$d = 1,4$					

Число пар в выборках $N_{\text{п}} = 5$. Разность между средними арифметическими сопряженных выборок $d = 16,6 - 15,2 = 1,4$. Ошибку разницы средних рассчитываем по одной из формул:

$$m_d = \sqrt{\sum (d_i - d)^2 / N_n (N_n - 1)} = \sqrt{3,2 / 5(5-1)} = 0,40, \qquad (1.22)$$

$$m_d = \sqrt{\frac{\sum d_i^2 - \left(\sum d_i\right)^2 / N_n}{N_n (N_n - 1)}} = \sqrt{\frac{13 - 7^2 / 5}{5(5-1)}} = 0,40. \qquad (1.23)$$

Результаты расчетов по приведенным формулам не выявили расхождений. Критерий Стьюдента получим следующий: $t = 1,4 / 0,40 = 3,5$. Число степеней

свободы $\nu = N_n - 2 = 5 - 2 = 3$. Для $\nu = 3$ при $P = 0{,}95$ и 0,99 табличное значение критерия Стьюдента равно 3,18 и 5,84 соответственно (см. прил. 4). Поскольку $t_ф > t_т$ при $P_{0,95}$, то различие по глубине расчленения рельефа в сравниваемых ландшафтах признается существенным. Такие ландшафты образуют самостоятельные группы.

Если при проведении эксперимента не учитывать сопряженность и независимость выборок, то можно получить противоположный вывод.

При сравнении средних, полученных на основе большого объема наблюдений при соблюдении нормального распределения, определение достоверности и различий средних можно выполнить упрощенно:

$$(M_1 - M_2)^2 / (m_1^2 + m_2^2) \geq 9.$$

Различия средних арифметических можно считать статистически достоверными, если получена величина 9 и более, если меньше – недостоверными. Пример нахождения сходства и отличия выборок с помощью критерия Стьюдента в MS Excel приведен в прил. 10.

Наименьшая существенная разность (НСР). Используется в дисперсионном анализе. Она показывает то минимальное различие между средними, начиная с которого при выбранном уровне вероятности средние сравниваемые показатели существенно отличаются друг от друга. Величина критерия выражается в тех же единицах, что и сравниваемые средние выборочных совокупностей, и определяется по формуле:

$$НСР = t_{табл} \cdot m_d, \qquad (1.24)$$

где m_d – ошибка разницы средних; $t_{табл}$ – табличное значение критерия Стьюдента при уровне вероятности 0,95 или 0,99 и степени свободы, определяемой экспериментом.

Если разность между сравниваемыми средними в условиях эксперимента больше или равна величине НСР при $P = 0{,}95$ или 0,99, то различие существенно. Используя предыдущий пример по глубине расчленения рельефа, проверим достоверность разницы между средними арифметическими с использованием критерия НСР для случаев независимого и сопряженного наблюдений по формуле (1.24):

$НСР_{0,95} = 2{,}31 \cdot 1{,}40 = 3{,}23$ м; $НСР_{0,99} = 3{,}36 \cdot 1{,}40 = 4{,}70$ м (для независимых наблюдений);

$НСР_{0,95} = 3{,}18 \cdot 0{,}40 = 1{,}27$ м; $НСР_{0,99} = 5{,}84 \cdot 0{,}40 = 2{,}33$ м (для сопряженных наблюдений).

Разница между средними арифметическими глубины расчленения рельефа при независимых и сопряженных наблюдениях одна и та же (1,4 м). Сравнивая ее с величиной НСР, приходим к тем же выводам, что и при использовании критерия Стьюдента.

Критерий Фишера. В выборочных совокупностях дисперсии могут существенно отличаться друг от друга. В таких случаях установление различий между выборочными совокупностями проводится по критерию Фишера (F – положительное асимметричное распределение). Расчет производится по формуле:

$$F = \sigma^2_{\text{большая}} / \sigma^2_{\text{меньшая}}. \quad (1.25)$$

Если величина расчетного критерия Фишера ($F_ф$) не превышает величины приведенного в таблице ($F_т$) (прил. 5), то различие между сравниваемыми дисперсиями считается недостоверным. При $F_ф > F_т$ эти дисперсии достоверно различны, как и сравниваемые по ним генеральные совокупности. Степень свободы рассчитывается для сравниваемых выборок отдельно по формуле $\nu = N - 1$.

П р и м е р. Необходимо установить достоверность различия в содержании гумуса в дерново-подзолистой заболоченной суглинистой почве для северной (x_1) и центральной (x_2) провинций Беларуси. Объем выборочных совокупностей одинаков (N_1, N_2). В результате обработки данных получены следующие средние и дисперсии: $M_{x_1} = 3{,}53$ %, $\sigma_{x_1} = 0{,}0024$ %; $M_{x_2} = 3{,}32$ %, $\sigma_{x_2} = 0{,}00032$ %. Сравниваемые совокупности весьма сходны, и можно констатировать отсутствие различия между ними. Однако пределы колебаний вариант в совокупностях существенно различны (более чем в 2 раза). В данном случае для сравнения следует использовать критерий Фишера. В результате вычислительных операций получены следующие результаты: $F_ф = \sigma_{x_1} / \sigma_{x_2} = 0{,}0024 / / 0{,}00032 = 7{,}5$. Степень свободы одинакова для первой и второй совокупностей (5–1 = 4). Для $P = 0{,}95$ и 0,99 табличное значение критерия Фишера равно 6,39 и 15,98 соответственно. Поскольку $F_ф > F_т$, то различие в содержании гумуса по провинциям признается существенным при $P = 0{,}95$.

Критерий Пирсона (хи-квадрат, χ^2). Для оценки соответствия или расхождения полученных эмпирических данных и теоретических (расчетных, прогнозных) распределений применяются статистические *критерии согласия.* Среди них наибольшее распространение получил непараметрический критерий К. Пирсона – хи-квадрат. Его можно использовать с различными формами распределения совокупностей. Как и любой другой статистический критерий, он не доказывает справедливость нулевой гипотезы, а лишь устанавливает с определенной вероятностью ее согласие или несогласие с экспериментальными данными. Критерий применяется при условии наличия не менее 5 наблюдений или частот в каждой группе, классе или совокупности. Малые частоты объединяют. Вычисление проводят по формуле:

$$\chi^2 = \sum \left[(\varphi - \varphi')^2 / \sum \varphi' \right], \quad (1.26)$$

где φ, φ′– наблюдения или частоты в опыте, ожидаемые эмпирически или теоретически соответственно.

Значения χ^2 могут быть только положительными и возрастать от нуля до бесконечности. Если вычисленный критерий χ^2 больше табличного (теоретического) значения, нулевая гипотеза, которая предполагает соответствие эмпирического и теоретического распределений, отвергается, при $\chi^2_{выч} < \chi^2_{табл}$ нулевая гипотеза принимается.

Достоверность различий можно определить по правилу Романовского: нулевая гипотеза отвергается, если соблюдается неравенство:

$$D = (\chi^2 - \nu) / \sqrt{2\nu} > 3. \qquad (1.27)$$

Степень свободы при проверке гипотезы о нормальном распределении вычисляется по формуле $\nu = k - 3$, где k – число классов. Различия между экспериментальными и теоретическими вариантами считаются достоверными, если $D > 3$.

Критерий Пирсона тем меньше, чем меньше различаются эмпирические и теоретические частоты. Он не позволяет обнаружить различия, которые скрадывает группировка (объединение малых частот в одну группу). Его удобно использовать, так как не требуется вычислений средних дисперсий.

П р и м е р. Следует определить число сельских жителей с бронхолегочными заболеваниями, обострение болезни у которых связано с природными условиями местожительства. Для обработки выборочных вариантов составляем табл. 1.7.

Таблица 1.7

Сравнение эмпирических и теоретических частот с использованием критерия Пирсона

Число обследованных жителей (классы)	Число фактически больных, φ	Число теоретически больных, φ′	φ – φ′	(φ – φ′)²	(φ – φ′)² / φ′
1–71	1	2			
72–142	3 **11**	4 **15**	–4	16	1,06
143–213	7	9			
214–284	**10**	**13**	–3	9	0,69
285–355	**15**	**14**	1	1	0,07
356–426	**12**	**10**	2	4	0,40
427–497	**10**	**11**	–1	1	0,09
498–568	8 **13**	6 **8**	5	25	3,12
569–639	5	2			
$I = 9$	$N_1 = 71$	$N_2 = 71$			$\chi^2_{выч} = \sum 5{,}43$

Всего выявлен 71 больной житель из 639 обследованных одного возраста и пола – по 9 человек в каждом населенном пункте. Для обработки данных количество обследованных сгруппировано в 9 классов. Поскольку частота в каждом классе φ, φ′ должна быть не менее 5, объединяем первые три и последние два класса в столбцах 2 и 3. Получаем новые классы с частотами 11 и 13 (всего по 6 классов распределения). Частоты в новых классах выделены жирным шрифтом в табл. 1.7. Затем производим расчеты, которые позволяют получить критерий χ^2 (см. табл. 1.7).

Сравниваем $\chi^2_{\text{выч}}$ с $\chi^2_{\text{табл}}$ при степени свободы $\nu = k - 3 = 6 - 3 = 3$, $P_{0,95}$. Поскольку $\chi^2_{\text{выч}} = 5,43 < \chi^2_{\text{табл}} = 7,815$, теоретическое распределение частот несущественно отличается от эмпирического, а гипотеза признается состоятельной.

Определим достоверность χ^2 по формуле (1.27):

$$D = (\chi^2 - \nu) / \sqrt{2\nu} > 3 = (5,43 - 3) / \sqrt{2 \cdot 3} = 0,99.$$

Полученная величина $D = 0,99 < 3$, следовательно, нулевая гипотеза признается состоятельной, т. е. влияние природных условий на распространение бронхолегочных заболеваний достоверно.

Глава 2
ДИСПЕРСИОННЫЙ АНАЛИЗ

При планировании эксперимента бывают ситуации, когда исследуемую систему необходимо разбить на группы, отличающиеся между собой в количественном отношении, и установить сходство или различие между ними по влиянию различных факторных величин на признак. Например, определить степень влияния географических условий на ход тех или иных процессов, явлений. Таким условиям лучше всего отвечает дисперсионный анализ, который нашел применение в физической географии.

Дисперсионный анализ позволяет утверждать с определенной долей уверенности наличие влияния на изучаемый объект каждого из условий в отдельности или в их сочетаниях. *Обязательным условием применения дисперсионного анализа является разбивка каждого учитываемого фактора не менее чем на две группы.* Они могут быть представлены как качественными, так и количественными показателями. Качественные показатели приводятся в виде баллов. Анализу подвергаются лишь определяющие поведение объекта факторы, которые установлены исследователем. По количеству определяющих факторов дается название виду дисперсионного анализа (одно-, двух-, трехфакторный и т. д.).

Обработка данных дисперсионного анализа – весьма трудоемкий процесс; облегчает вычисления правильная организация опыта. Порядок расчета в разных видах дисперсионного анализа будет различным, но логическая схема остается единой. Факторы в дисперсионном анализе должны быть независимыми друг от друга; каждый фактор следует разделить на группы, количество которых зависит от поставленной задачи.

Дисперсионный анализ применяется в случаях нормального или близкого к нему распределения выборочных совокупностей. Выборки должны иметь близкие по значению показатели дисперсии σ^2. Количество повторностей в каждой выделенной группе принимается одинаковым.

Основная трудность при использовании дисперсионного анализа – составление комбинационной таблицы для обработки данных (*дисперсионный комплекс*). Если число наблюдений над результативным признаком по отдельным группам изучаемого фактора одинаково, то дисперсионный комплекс называется *равномерным*, если разное, то *неравномерным.* Об-

щее число наблюдений над результативным признаком принято называть *объемом дисперсионного комплекса.*

Порядок действия по каждому виду дисперсионного анализа определяется его основной задачей, которая состоит в делении суммарного или общего варьирования изучаемого признака на доли: варьирование, вызываемое действием отдельных факторов; варьирование, вызываемое взаимодействием факторов между собой; остаточное варьирование объекта, которое определяется неучитываемыми факторами.

2.1. Однофакторный дисперсионный анализ

Среди различных видов дисперсионного анализа наиболее часто используется однофакторный. Для выполнения однофакторного анализа в опыте должно быть предусмотрено две повторности и более. Исследуемый фактор разбивается на группы с целью выявления его оптимальной величины, влияющей на результативный признак. Для облегчения расчета можно уменьшить все показатели в пределах дисперсионного комплекса на определенную величину, а затем увеличить конечные результаты на ту же величину.

Географы исследуют не только природные, но и сельскохозяйственные ландшафты (агроландшафты), претерпевающие существенные изменения под воздействием агротехногенеза. Использование дисперсионного анализа позволяет не только констатировать изменения в агроландшафте, но и активно включаться в его преобразование.

Известно, что оптимальным условиям питания растений соответствует дерновая легкосуглинистая гумусированная нейтральная почва. Ее можно создать путем внесения в пахотный горизонт добавок минерального грунта определенного механического состава и торфа. Формирование искусственной антропогенной почвы требует полевых экспериментов. В связи с этим поставлена следующая задача: определить влияние на урожай зерна ячменя разных доз торфа (200, 300, 400 т абсолютно сухого вещества на гектар) при внесении его на фоне минеральных, органических удобрений и доломитовой муки. Исходная почва – дерново-подзолистая глееватая связносупесчаная осушенная. После получения сведений об урожайности ячменя в названных условиях составляется таблица дисперсионного комплекса (табл. 2.1), куда заносится исходная информация по группам влияющего фактора (вариантам опыта) и некоторые результаты расчетов (для удобства сделано округление по урожайности до целых чисел). Вначале производим расчет данных по вариантам опыта (строкам).

Результаты разносим по столбцам. Суммарный урожай ячменя по повторностям $\sum x_i$ и по каждому варианту опыта вносим в столбец 6 в чис-

лителе. Аналогично поступаем с квадратами этих показателей $\sum x_i^2$. Затем в столбце 7 приводим квадраты суммарного урожая ячменя по повторностям $\left(\sum x_i\right)^2$. И, наконец, вычисляем среднее арифметическое M_i по каждому варианту опыта, заносим в столбец 8; вычисляем общее среднее $M_{\text{общ}}$.

После получения данных по вариантам опыта производим расчет необходимых показателей по повторностям (x_k). Сначала суммируем данные урожайности ячменя и приводим в строке под чертой $\sum x_k$. Суммы сумм урожайности ячменя по вариантам опыта и повторностям должны совпасть и дать сумму всех вариант $\left(\sum\sum x_{i,k} = 495\right)$. Аналогично суммируем квадраты этих показателей по повторностям $\left(\sum x_k^2\right)$. Суммы сумм квадратов по вариантам и повторностям опыта должны совпасть и дать сумму квадратов всех вариант $\left(\sum x_i^2 = \sum x_k^2 = 15\,935\right)$. Ниже вписываем результаты возведения в квадрат сумм вариант по каждой повторности $\left(\sum x_k\right)^2$ и суммируем их: $\sum\left(\sum x_k\right)^2 = 61\,269$. Вычисляем средние арифметические по каждой повторности опыта M_k. Общее среднее арифметическое всех вариант опыта составляет $M_{\text{общ}} = \left(\sum x_{i,k}\right) / N = 495 : 16 = 30{,}93$.

Таблица 2.1

Однофакторный дисперсионный анализ

<table>
<tr><td colspan="2" rowspan="2">Варианты опыта (фактор)</td><td colspan="4" rowspan="2">Урожай ячменя по повторностям, ц/га*</td><td colspan="3">По повторностям (признакам) (i)</td></tr>
<tr><td>$\frac{\sum x_i}{\sum(x_i^2)}$</td><td>$\left(\sum x_i\right)^2$</td><td>$M_i$</td></tr>
<tr><td colspan="2">Контроль (фон)</td><td>$\frac{20}{400}$</td><td>$\frac{21}{441}$</td><td>$\frac{22}{484}$</td><td>$\frac{20}{400}$</td><td>$\frac{83}{1725}$</td><td>6889</td><td>20,75</td></tr>
<tr><td colspan="2">Фон+200 т/га торфа</td><td>$\frac{30}{900}$</td><td>$\frac{32}{1024}$</td><td>$\frac{32}{1024}$</td><td>$\frac{31}{961}$</td><td>$\frac{125}{3909}$</td><td>15625</td><td>31,25</td></tr>
<tr><td colspan="2">Фон+300 т/га торфа</td><td>$\frac{35}{1225}$</td><td>$\frac{36}{1296}$</td><td>$\frac{35}{1225}$</td><td>$\frac{36}{1296}$</td><td>$\frac{142}{5042}$</td><td>20164</td><td>35,50</td></tr>
<tr><td colspan="2">Фон+400 т/га торфа</td><td>$\frac{36}{1296}$</td><td>$\frac{35}{1225}$</td><td>$\frac{37}{1369}$</td><td>$\frac{37}{1396}$</td><td>$\frac{145}{5259}$</td><td>21025</td><td>36,25</td></tr>
<tr><td rowspan="4">По факторам</td><td>$\sum x_k$</td><td>121</td><td>124</td><td>126</td><td>124</td><td rowspan="3">$\sum\sum x_{i,k} = 495$
$\sum\sum x_{i,k}^2 = 15935$</td><td rowspan="3">$\sum\left(\sum x_i\right)^2 = 63703$</td><td rowspan="4"></td></tr>
<tr><td>$\sum(x_k^2)$</td><td>3821</td><td>3986</td><td>4102</td><td>4026</td></tr>
<tr><td>$\left(\sum x_k\right)^2$</td><td>14 641</td><td>15 376</td><td>15 876</td><td>15 376</td></tr>
<tr><td>M_k</td><td>30,25</td><td>31,00</td><td>31,50</td><td>31,00</td><td colspan="2">$\sum\left(\sum x_k\right)^2 = 61\,269$</td></tr>
</table>

Примечание: * В числителе – опытные данные, в знаменателе – квадраты этих показателей.

Следующий этап работы – нахождение сумм квадратов отклонений, т. е. расчленение общего варьирования признака на составные части исходя из равенства:

$$\Theta = \Theta_1 + \Theta_2 + \Theta_3,$$

где Θ – сумма квадратов отклонений по общему варьированию данных, Θ_1 – по группам фактора (варианты опыта), Θ_2 – по повторностям опыта, Θ_3 – по остаточному варьированию, вызванному неучтенными факторами.

Общая сумма квадратов отклонений вычисляется следующим образом:

$$\Theta = \sum\left(\sum x_{i,k}^2\right) - \left(\sum\sum x_{i,k}\right)^2 / N.$$

Подставив данные из табл. 2.1, получим: Θ =15 935 – 495^2: 16 = 621. Затем находим сумму квадратов отклонений по группам фактора (варианты опыта) по формуле:

$$\Theta_1 = \left[\sum\left(\sum x_i\right)^2 - \left(\sum\sum x_{i,k}\right)^2 / k\right] / i, \qquad (2.1)$$

где k – число групп фактора, т. е. 4; i – число повторностей, т. е. 4. В данном случае должно выдержаться равенство $N = ki = 4 \cdot 4 = 16$. По формуле (2.1) вычислим:

$$\Theta_1 = [63\ 703 - 495^2 : 4] : 4 = 611{,}75.$$

Сумму квадратов отклонений по повторностям опыта находим по формуле

$$\Theta_2 = \left[\sum\left(\sum x_k\right)^2 - \left(\sum\sum x_{i,k}\right)^2 / i\right] / k, \qquad (2.2)$$

где i – число повторностей, т. е. 4; k – число слагаемых в каждой сумме $\sum x_k$, т. е. 4.

Вычисляем Θ_2 по формуле (2.2):

$$\Theta_2 = [61269 - 495^2 : 4] : 4 = 3{,}25.$$

Таблица 2.2

Результаты однофакторного дисперсионного анализа

Варьирование данных	Сумма квадратов отклонений, Θ	Степень свободы, ν	Дисперсия, $\sigma^2 = \Theta/\nu$	Критерий Фишера	
				$F_ф$	$F_т$
Общее по опыту	621,00	15	41,40	–	–
По вариантам опыта	611,75	3	203,91	304,31	8,81
По повторностям	3,25	3	1,08	1,61	8,81
Случайное (остаточное)	6,00	9	0,67	–	–

Сумма квадратов отклонений по остаточному варьированию определяется из равенства

$$\Theta_3 = \Theta - \Theta_1 - \Theta_2. \qquad (2.3)$$

Подставив значение вычисленных сумм соответствующих квадратов отклонений в формулу (2.3), получим

$$\Theta_3 = 621 - 611{,}75 - 3{,}25 = 6{,}00.$$

Проводим дисперсионный анализ данных урожая ячменя (табл. 2.2). Вносим в таблицу рассчитанные суммы квадратов отклонений (Θ, Θ_1, Θ_2, Θ_3). Число степеней свободы получаем следующим образом: по общей сумме квадратов отклонений $\nu = N - 1 = 16 - 1 = 15$; по вариантам опыта $\nu_1 = n_1 - 1 = 4 - 1 = 3$; по повторностям $\nu_2 = n_2 - 1 = 4 - 1 = 3$; по остаточной сумме $\nu_3 = \nu - \nu_1 - \nu_2 = 15 - 3 - 3 = 9$.

Дисперсия определяется путем деления сумм квадратов отклонений (Θ, Θ_1, Θ_2, Θ_3) на соответствующие им числа степеней свободы (ν, ν_1, ν_2, ν_3), что можно выразить в общем виде формулой $\sigma^2 = \Theta/\nu$, получим $\sigma^2 = 621 : 15 = 41{,}40$.

Оценку сходства или различия между вариантами опыта можно проводить по критерию Фишера, критерию Стьюдента или НСР.

Поскольку (по вариантам опыта) $F_ф > F_т$ (см. табл. 2.2 и прил. 5), то это позволяет сделать вывод, что внесение больших доз торфа положительно влияет на величину урожая ячменя в агроландшафте.

Наиболее распространен в дисперсионном анализе для оценки результатов опыта критерий НСР, алгоритм которого приводим ниже. Вначале определяем среднее квадратическое отклонение из дисперсии, полученной в результате случайного варьирования (см. табл. 2.2): $\sigma = \sqrt{\sigma_3^2}$, затем вычисляем обобщенную ошибку среднего: $m_M = \sigma / \sqrt{N_{повт}}$. Поскольку ошибка среднего для всех сравниваемых вариант одна и та же, формула для расчета ошибки разности может быть преобразована: $m_d = \sqrt{2m^2}$. Наименьшую существенную разность рассчитываем по формуле (1.24). Используя исходные данные, вычислим НСР по указанному выше алгоритму:

$$\sigma = \sqrt{0{,}67} = 0{,}82; \quad m_M = 0{,}82 / \sqrt{4} = 0{,}41;$$

$$m_d = \sqrt{2 \cdot 0{,}41^2} = 0{,}58; \quad НСР_{0{,}95} = 0{,}58 \cdot 2{,}26 = 1{,}31;$$

$$НСР_{0{,}99} = 0{,}58 \cdot 3{,}25 = 1{,}88.$$

Из полученных результатов дисперсионного анализа вытекает следующий вывод (табл. 2.3). Величина $НСР_{0{,}95}$ и $НСР_{0{,}99}$ меньше величины

Таблица 2.3

Влияние высоких доз торфа на урожай ячменя

Вариант опыта	Урожай ячменя по повторностям				Среднее	Прибавка
Контроль (фон)	20	21	22	20	20,75	–
Фон+200 т/га	30	32	32	31	31,25	10,50
Фон+300 т/га	35	36	35	36	35,50	14,75
Фон+400 т/га	36	35	37	37	36,25	15,50
$НСР_{0,95}$, ц/га $НСР_{0,99}$, ц/га p	1,31 1,88 1,32 %					

прибавки урожая зерна ячменя, поэтому внесение высоких доз торфа положительно влияет на урожай. Лучший результат получен в варианте с дозой внесения торфа 400 т/га, где прибавка зерна ячменя составила 15,5 ц/га.

В случае необходимости можно рассчитать ошибки частных средних арифметических: по повторностям:

$$m_{\text{п}} = \sqrt{\sigma^2_{\text{ост}} / N_{\text{п}}} = \sqrt{0{,}67 / 4} = 0{,}41;$$

по вариантам опыта:

$$m_{\text{в}} = \sqrt{\sigma^2_{\text{ост}} / N_{\text{в}}} = \sqrt{0{,}67 / 4} = 0{,}41.$$

Ошибку общего среднего арифметического используют для вычисления точности опыта. Показатель точности опыта для общего среднего арифметического вычисляется следующим образом:

$$p_{M\text{общ}} = (m_{\text{общ}}/M_{\text{общ}}) \cdot 100 = (0{,}41 : 30{,}90) \cdot 100 = 1{,}32\ \%.$$

Поскольку $p = 1{,}32\%$, т. е. $< 3\ \%$, то опыт признается достаточно точным.

Аналогичным образом вычисляется точность опыта для частных средних арифметических по вариантам опыта и по повторностям:

$$p_{\text{в}} = (m_{\text{в}}/M_{\text{в}}) \cdot 100; \quad p_{\text{п}} = (m_{\text{п}}/M_{\text{п}}) \cdot 100.$$

2.2. Двухфакторный дисперсионный анализ

Если в дисперсионный анализ включают несколько факторов, влияющих на результативный признак, то они должны быть независимыми друг от друга. Рассмотрим обработку данных с двумя факторами, каждый из которых делится на две группы. Для этого составляем комбинационный дисперсионный комплекс (табл. 2.4). Каждый фактор характеризуется тремя наблюдениями (повторностями). Аналогичную схему можно использовать для двухфакторного анализа с большим числом групп и повторностей в каждом факторе.

Таблица 2.4

Двухфакторный дисперсионный комплекс*

Повторность опыта по фактору II	Биомасса, кг/м^3		$\frac{\sum y_i}{\sum y_i^2}$	$\left(\sum y_i\right)^2$	M_y
	Группы по фактору I				
	1982 г. (сухой)	1984 г. (влажный)			
Группа по фактору II (неосушенный агроландшафт)					
Первая	$\frac{5}{25}$	$\frac{4}{16}$	$\frac{9}{41}$		
Вторая	$\frac{6}{36}$	$\frac{5}{25}$	$\frac{11}{61}$		
Третья	$\frac{5}{25}$	$\frac{6}{36}$	$\frac{11}{61}$		
$\frac{\sum}{\sum}$	$\frac{16}{86}$	$\frac{15}{77}$	$\frac{31}{163}$	961	5,16
Группа по фактору II (осушенный агроландшафт)					
Первая	$\frac{3}{9}$	$\frac{5}{25}$	$\frac{8}{34}$		
Вторая	$\frac{4}{16}$	$\frac{6}{36}$	$\frac{10}{52}$		
Третья	$\frac{4}{16}$	$\frac{6}{36}$	$\frac{10}{52}$		
$\frac{\sum}{\sum}$	$\frac{11}{41}$	$\frac{17}{97}$	$\frac{28}{138}$	784	4,66
$\frac{\sum x_i}{\sum x_i^2}$	$\frac{27}{127}$	$\frac{32}{174}$	$\frac{59}{301}$	$\sum\left(\sum y_i\right)^2 = 1745$	$M_{общ} = 4{,}90$
$\left(\sum x_i\right)^2$	729	1024	$\sum\left(\sum x_i\right)^2 = 1753$		
M_x	4,50	5,33			

Примечание: * фактор I – погода 1982 и 1984 гг., фактор II – осушенный и неосушенный агроландшафт.

Двухфакторный дисперсионный анализ можно представить в виде равенства:

$$\Theta = \Theta_1 + \Theta_2 + \Theta_3 + \Theta_4 + \Theta_5, \qquad (2.4)$$

где Θ – общая сумма квадратов; Θ_1, Θ_2 – сумма квадратов отклонений для факторов I и II соответственно; Θ_3 – сумма квадратов отклонений, возникающих при взаимодействии факторов I и II; Θ_4 – сумма квадратов отклонений по повторностям; Θ_5 – остаточная сумма квадратов отклонений неучтенных факторов.

Следует определить влияние метеорологических условий (фактор I) и мелиорации (фактор II) на урожай биомассы трав в агроландшафте.

При обработке данных исходной информации (см. табл. 2.4) порядок расчета не отличается от описанного выше алгоритма однофакторного дисперсионного комплекса (см. п. 2.1). Дальнейшие расчеты проводятся в следующем порядке.

Общую сумму квадратов отклонений находим по формуле

$$\Theta = \sum\sum x_i^2, y_i^2 - \left[\left(\sum\sum x_i, y_i \right)^2 / N \right] = 301 - \left[(59)^2 : 12 \right] = 10{,}92,$$

где N – общий объем выборки.

Сумма квадратов отклонений по фактору I вычисляется по формуле

$$\Theta_1 = \left[\sum \left(\sum x_i \right)^2 - \left(\sum\sum x_i, y_i \right)^2 / n_x \right] / k_x = \left[1753 - (59)^2 : 6 \right] = 2{,}08,$$

где n_x – число групп фактора I (n_x = 2); k_x – число вариант в каждой отдельной сумме (k_x = 6).

Сумма квадратов отклонений по фактору II вычисляется аналогично определению суммы квадратов отклонений по фактору I:

$$\Theta_2 = \left[\sum \left(\sum y_i \right)^2 - \left(\sum\sum x_i, y_i \right)^2 / n_y \right] / k_y = \left[1745 - (59)^2 : 2 \right] : 6 = 0{,}75.$$

Сумма квадратов отклонений, вызываемых взаимодействием факторов I и II, определяется следующим образом:

$$\Theta_3 = \left[\sum \left(\sum z_i \right)^2 - \left(\sum\sum x_i, y_i \right)^2 / n_z \right] / k - \Theta_1 - \Theta_2, \qquad (2.5)$$

где $\sum \left(\sum z_i^2 \right)$ – сумма квадратов сумм значений вариант по группам выборки комбинационной таблицы ($16^2 + 15^2 + 11^2 + 17^2 = 891$); n_z – число сумм вариант по группам; k_z – число слагаемых вариант в каждой группе выборки.

Подставляем данные в формулу (2.5):

$$\Theta_3 = [891 - (59)^2 : 4] : 3 - 2{,}08 - 0{,}75 = 4{,}08.$$

Сумма квадратов отклонений по повторностям Θ_4 определяется по формуле (2.6) путем подстановки конкретных данных задачи:

$$\Theta_4 = \left[\sum \left(\sum x_i \right)^2 - \left(\sum\sum x_i, y_i \right)^2 / n_{x,y} \right] / k_{x,y}, \qquad (2.6)$$

где $n_{x,y}$ – число сумм по повторностям (по 3); $k_{x,y}$ – число слагаемых в каждой сумме (равное 4); $\sum \left(\sum x_i \right)^2$ – сумма квадратов сумм исходных данных по повторностям фактора I сверху вниз: $[(5+4) + (3+5)]^2 + [(6+5) + (4+6)]^2 + [(5+6) + (4+6)]^2 = 1171$. Подставив данные в исходную формулу (2.6), получим

$$\Theta_4 = [1171 - (59)^2 : 3] : 4 = 2{,}67.$$

Сумму квадратов отклонений по остаточному варьированию определяем из равенства (2.4):

$$\Theta_4 = 10{,}92 - 2{,}08 - 0{,}75 - 4{,}08 - 2{,}67 = 1{,}14.$$

Затем вычисляем число степеней свободы: для Θ $\nu = N - 1 = 12 - 1 = 11$; для Θ_1 и Θ_2 число степеней свободы равно числу градаций фактора минус единица: $\nu_1 = n_1 - 1 = 2 - 1 = 1$; $\nu_2 = n_2 - 1 = 2 - 1 = 1$; для Θ_3 $\nu_3 = \nu_1 \cdot \nu_2 = 1 \cdot 1 = 1$; для Θ_4 число степеней свободы равно числу повторностей минус единица: $\nu_4 = 3 - 1 = 2$; для Θ_5 этот показатель определяется следующим образом: $\nu_5 = \nu - \nu_1 - \nu_2 - \nu_3 - \nu_4 = 11 - 1 - 1 - 1 - 2 = 6$.

Показатели дисперсии (табл. 2.5) вычисляются путем деления значений сумм квадратов отклонений на соответствующие значения степеней свободы (например 10,92 : 11 = 0,99).

Фактический критерий Фишера определяется путем деления каждой из величин дисперсий на значение остаточной. Критическое значение критерия Фишера находим в прил. 5 на пересечении значений большей и меньшей степеней свободы, которые устанавливаем по величине сравниваемых дисперсий (см. табл. 2.5). Например, по фактору II отношение дисперсий равно $F_ф = 3{,}94$. В данном случае большей будет дисперсия по фактору II σ^2=0,75 с числом степеней свободы $\nu = 1$, для меньшей величины остаточной дисперсии $\sigma^2 = 0{,}19$ и $\nu = 6$. Пересечение ν =1 и $\nu = 6$ дает величину $F_т = 5{,}99$ для $P = 0{,}95$. Если $F_ф > F_т$, то действие данного фактора признается существенным, при $F_ф < F_т$ – несущественным.

Таблица 2.5

Результаты двухфакторного дисперсионного анализа

Варьирование данных	Сумма квадратов отклонений, Θ	Степень свободы, ν	Дисперсия, σ^2	Критерий Фишера	
				$F_ф$	$F_т$
Общее по опыту	10,92	11	0,99	5,21	4,31
По фактору I	2,08	1	2,08	10,94	5,99
По фактору II	0,75	1	0,75	3,94	5,99
По взаимодействию факторов I и II	4,08	1	4,08	21,47	5,99
По повторностям	2,67	2	1,33	7,00	5,14
Остаточное	1,14	6	0,19	1,00	–

Исходя из анализа критерия Фишера, можно заключить, что влияние исследуемых параметров на биомассу признается существенным в целом по опыту, по фактору I, по взаимодействию факторов и по повторностям, т. е. во всех случаях $F_ф > F_т$. Действие фактора II на объект не доказано ($F_ф < F_т$).

Оценку результатов эксперимента можно сделать по критериям НСР и Стьюдента. Для вычисления НСР и t находим ошибку среднего арифметического m_M всего опыта и ошибку разности средних m_d по следующим формулам:

$$m_M = \sqrt{\sigma^2_{\text{ост}} / N}; \quad m_d = \sqrt{2\sigma^2_{\text{ост}} / n},$$

где n – численность меньшей из сравниваемых частных групп (в нашем примере обе группы одинаковы и равны шести). Произведем расчет необходимых показателей:

$$m_M = \sqrt{0{,}19 : 12} = 0{,}1258; \quad m_d = \sqrt{(2 \cdot 0{,}19) : 6} = 0{,}25;$$

$$\text{НСР} = m_d \cdot t_{\text{т}} = 0{,}25 \cdot 2{,}45 = 0{,}61;\ \nu_{\text{ост}} = 6.$$

По критерию Стьюдента сравниваем средние арифметические данные по осушенному и неосушенному агроландшафту:

$$t = (M_{y,1} - M_{y,2}) / m_{dy} = (5{,}16 - 4{,}66) : 0{,}25 = 2{,}0.$$

Сравниваем также средние арифметические по метеорологическим условиям:

$$t = (M_{x,1} - M_{x,2}) / m_{dy} = (5{,}33 - 4{,}50) : 0{,}25 = 3{,}32\,.$$

По прил. 4 критерия Стьюдента $t_{\text{т}} = 2{,}45$ при $P = 0{,}95$ для $\nu = 6$.

Таким образом, на биомассу трав в агроландшафтах не влияет мелиорация (т. е. фактор II), так как $t_{\text{ф}} = 2{,}0 < t_{\text{т}} = 2{,}45$ при $P = 0{,}95$; метеорологические условия (фактор I) достоверно влияют на биомассу трав при P = 0,95. Выводы, сделанные при использовании критериев Фишера и Стьюдента, совпадают.

В заключение обычно определяют точность опыта, которая равна:

$$p = (m_M / M_{\text{общ}}) \cdot 100 = (0{,}1258 : 4{,}9) \cdot 100 = 2{,}56\ \%.$$

Точность опыта признается достаточно высокой, поскольку $p < 3$ %.

Если имеется необходимость, вычисляется коэффициент варьирования опытных данных:

$$V = \left(\sqrt{\sigma^2_{\text{общ}}} / M_{\text{общ}} \cdot 100\right), \quad V = \left(\sqrt{0{,}99} : 4{,}9\right) \cdot 100 = 20{,}0\ \%.$$

Коэффициент варьирования опытных данных незначителен, что также удовлетворяет требованиям опыта.

Глава 3

КЛАСТЕРНЫЙ АНАЛИЗ

При проведении географических исследований, как правило, возникает проблема *объединения по сходству* (*кластеризация*) объектов, которые характеризуются множеством признаков, выраженных в разных единицах измерения. Для этой цели используется *кластерный анализ*. Поскольку кластерный анализ занимается классификацией объектов, а факторный исследует связи между ними, то оба метода дополняют друг друга и между ними иногда трудно провести четкие границы.

Методологические особенности кластерного анализа сводятся к выявлению единой меры, охватывающей ряд исследуемых признаков. Эти признаки объединяются с помощью метрики (расстояния) в один кластер сходства группируемых объектов.

Состояние любого объекта может быть описано с использованием *многомерного признака*, или *многомерной случайной величины* (x_1, x_2,..., x_n). Примером количественных признаков при зонировании территории города может служить площадь строений (x_1), количество исторических памятников (x_2), количество промышленных предприятий (x_3) и т. д. Их можно объединить в один качественный признак – инфраструктурные условия города. Таким способом состояние любого объекта может быть описано с помощью многомерного признака.

Исследование нескольких аналогичных объектов (городов) обязывает проводить разбиение совокупности объектов на однородные группы, т. е. провести их классификацию по сходству признаков (x_1, x_2, ...). Содержательная постановка задачи при кластерном анализе заключается в следующем. Имеется некоторая совокупность объектов, которые характеризуются рядом признаков. Объекты необходимо разбить на несколько кластеров (классов) таким образом, чтобы объекты из одного класса были сходными по характеризующих их признакам, например, сравнение ландшафтов, выявление сходных тенденций в развитии экономических субъектов.

В зависимости от специальности и природы используемых методов исследователи называют классификацию многомерных наблюдений как

распознавание образов с учителем (численной таксономией), *кластер-анализом без учителя, дискриминантным* анализом.

Таксономические методы классификации объектов основываются на выделении групп объектов, наиболее близких в многомерном пространстве. Для определения степени сходства объектов вычисляются таксономические расстояния между ними. Если исследователь имеет перед собой образы будущих групп – обучающие выборки, то группировка выполняется методом дискриминантного анализа. При отсутствии обучающих выборок используется кластерный анализ (В. В. Глинский, В. Г. Ионин, 1998). В отличие от дискриминантного анализа (С. А. Айвазян и др., 1984), отсутствие классифицированных обучающих выборок в кластерном анализе значительно усложняет решение задачи классификации.

При относительной формализации методов кластерного анализа они носят эвристический (теоретический) характер, реализуют принцип здравого смысла. Для оценки сходства объектов по ряду признаков используют три типа мер:

- *коэффициент подобия* – для группировки объектов и признаков, если уровни показателей являются действительно целыми числами;
- *коэффициенты связи* – чаще применяются для группировки признаков с использованием коэффициента корреляции;
- *показатели расстояния* – характеризуют степень взаимной удаленности признаков и применяются в основном для кластеризации объектов; признаки объектов должны быть независимыми, что предварительно можно уточнить с помощью корреляционного анализа.

Многомерное наблюдение может быть интерпретировано геометрически в виде точки в многомерном пространстве. Геометрическая близость точек в пространстве означает близость физических состояний объектов, их однородность. Решающим в интерпретации остается выбор масштаба метрики, т. е. задание расстояния между объектами, которые объединяют или разъединяют объекты. В результате разбиения объектов на группы по сходству признаков образуются *кластеры* (*таксоны, образы*). Необходимость разбиений совокупности объектов на однородные группы возникает при проведении социально-экономических, землеустроительных, географических исследований и т. д.

Выбор метрики (меры близости) является важнейшим моментом исследования, который определяет окончательный вариант разбиения объектов на группы. Это зависит от цели исследования, физической и статистической природы вектора наблюдений (x), полноты априорных сведений о характере вероятностного распределения x.

В задачах кластер-анализа широко используются следующие метрики: Эвклида, Махаланобиса, Хемминга, меры близости, задаваемые потенциальной функцией. Эвклидова метрика наиболее употребительна.

Обычно среднее Эвклидовое расстояние рассчитывается по формулам:

$$d_{kl} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(Z_{k_j} - Z_{l_j}\right)^2}, \qquad (3.1)$$

где m – число признаков x; Z_{k_j}, Z_{l_j} – стандартизированные значения признака j для k и l объектов соответственно, или:

$$d_{kl} = \sqrt{\frac{(Z_{k_{j1}} - Z_{l_{j1}})^2 + (Z_{k_{j2}} - Z_{l_{j2}})^2 + \ldots + (Z_{k_{jn}} - Z_{l_{jn}})^2}{m}}.$$

Если не учитывать число признаков x – m, формула примет вид:

$$d_{kl} = \sum_{j=1}^{m}\left(Z_{k_j} - Z_{l_j}\right)^2. \qquad (3.2)$$

Формула (3.2) менее объективна, так как не учитывает число признаков, количество которых может изменяться от трех и более.

Расчет упрощается, если в качестве метрики использовать l_1-*норму*:

$$d_{kl} = \sum_{j=1}^{m}\left(Z_{k_j} - Z_{l_j}\right). \qquad (3.3)$$

Эти метрики применяются в следующих случаях:

- наблюдения x извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей (совместное изменение двух признаков), где компоненты x взаимно независимы и имеют одинаковую дисперсию;
- компоненты $x_1, x_2, \ldots, x_p$ вектора наблюдений x однородны по своему физическому смыслу и все важны;
- факторное пространство совпадает с геометрическим; понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве.

«Взвешенное» эвклидово расстояние определяется:

$$d_{kl} = \sqrt{\omega_1 (Z_{k_{j1}} - Z_{l_{j1}})^2 + \omega_2 (Z_{k_{j2}} - Z_{l_{j2}})^2 + \ldots + \omega_n (Z_{k_{jn}} - Z_{l_{jn}})^2}, \qquad (3.4)$$

или

$$d_{kl} = \sqrt{(Z_k - Z_l)' \wedge' \sum^{-1} \wedge (Z_k - Z_l)},$$

где $\sum$ – ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения Z; $\wedge$ – некоторая симметричная неотрица-

тельно-определенная матрица «весовых» коэффициентов λ_{mq}, которая чаще всего выбирается диагональной; ′ – штрих-символ операции транспонирования вектора; –1 – обращение матрицы $\sum$.

Расстояние Хемминга используется как мера различия объектов, задаваемых дихотомическими признаками (деление объекта на две составляющие):

$$d_{kl} = \sum_{j=1}^{m} \left| Z_{k_j} - Z_{l_j} \right|. \tag{3.5}$$

Обычно признаки заданы в виде набора нулей и единиц: 0, если $a_i = b_i$; 1, если $a_i \neq b_i$.

Средним линейным отклонением оценивается расстояние между объектами по Хеммингу:

$$d_{kl} = \frac{1}{m} \sum_{j=1}^{m} \left| Z_{k_j} - Z_{l_j} \right|. \tag{3.6}$$

Таким образом, при решении задач классификации могут быть использованы разные меры сходства между объектами. Выбор метрики зависит от вида информации, характеризующей объекты в пространстве признаков, и требует тщательного критического анализа.

Покажем на общих примерах основные приемы кластерного анализа. На основании данных, содержащихся во множестве *x*, необходимо разбить множество объектов *I* на *m* кластеров (подмножеств) так, чтобы каждый объект I_i принадлежал лишь одному подмножеству разбиения, а объекты, принадлежащие одному кластеру, были сходными. Объекты, принадлежащие разным кластерам, должны быть разнородными (несходными). В качестве объекта *I* рассмотрим *n* стран, каждая из которых характеризуется валовым национальным продуктом на душу населения (C_1), природными условиями (C_2), природными ресурсами (C_3) и т. д. Тогда x_1 (вектор измерений) представляет собой набор указанных характеристик (показателей) для первой страны, x_2 – для второй, x_3 – для третьей и т. д. Задача заключается в том, чтобы сгруппировать *n* стран по уровню развития с учетом природных факторов. Для выполнения поставленной задачи лучше подходит кластерный анализ, чем другие методы с использованием группировки.

При субъективном разбиении множества показателей на группы остается неизвестным, действительно ли такое разбиение оптимально. Еще не разработан удовлетворительный статистический критерий, который позволил бы оценить проведенное разбиение и принадлежность данного

показателя к определенной группе. В практической работе исследователя это может привести к ошибке в таких сложных вопросах, как группировка ландшафтов, их классификация и районирование. Лишь проведение кластерного анализа на моделях с четкой структурой является наиболее объективным.

Число кластеров определяется в ходе разбиения имеющегося объема совокупности. При большом числе вариант в совокупности пользуются методом случайного отбора. Оптимальное число разбиений является функцией заданной доли оптимальных разбиений во множестве всех возможных. Общее рассеяние множества кластеров будет тем больше, чем выше доля допустимых разбиений. Находим необходимое число разбиений S в зависимости от значений вероятности P и заданной доли допустимых разбиений во множестве всех возможных β (табл. 3.1).

Таблица 3.1

Число разбиений в зависимости от их заданной доли и вероятности

β	*P*					
	0,80	0,90	0,95	0,99	0,999	0,9999
0,20	8	11	14	21	31	42
0,10	16	22	29	44	66	88
0,05	32	45	59	90	135	180
0,01	161	230	299	459	689	918
0,001	1626	2326	3026	4652	6977	9303
0,0001	17 475	25 000	32 526	55 000	75 000	100 000

В качестве меры разнородности рассматривается мера принадлежности. При решении задач кластерного анализа принимаются следующие условия: а) выбранные характеристики допускают желательное разбиение на кластеры; б) единицы измерения (масштаб) выбраны правильно (это обусловлено тем, что разбиение на кластеры зависит от выбора масштаба). Наиболее прямой способ решения задачи заключается в полном переборе всех возможных разбиений на кластеры и отыскании такого, которое ведет к оптимальному (минимальному) значению целевой функции. Целевая функция как критерий оптимальности представляет собой некоторый функционал, выражающий уровни возможности различных разбиений и группировок. Например, в качестве целевой функции может быть использована внутригрупповая сумма квадратов отклонений $\sigma_1^2 + \sigma_2^2$. Приведем пример кластеризации с помощью полного перебора (все возможные варианты сочетаний). Если число объектов $n = 8$, кластеров $m = 4$, то число возможных разбиений составляет 1701, т. е. существует 1701 способ разбить 8 объектов на 4 кластера (табл. 3.2). Число разбиений можно определить также по формуле $S(n,m) \approx m^n / m \approx m^{n-1}$.

Таблица 3.2

Число разбиений в зависимости от сочетаний числа кластеров и объектов

n	*m*							
	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	15	1		
7	1	63	301	350	140	21	1	
8	1	127	966	1701	1050	266	28	1

Разбиение в конечном итоге должно удовлетворять критерию оптимальности, т. е. целевому функционалу (целевой функции).

Метод дендритов. Исследуемые объекты, разделенные на кластеры, можно изобразить в виде дендрограммы, которая представляет собой графическое изображение матрицы расстояний или сходства. Такой анализ объектов исследования носит название метода дендритов. Имея *n* объектов, можно построить большое количество дендрограмм, которые соответствуют избранной процедуре кластеризации. Для конкретной матрицы расстояний или сходства существует только одна дендрограмма.

Представим дендрограмму с шестью объектами ($n = 6$) (рис. 3.1). Объекты 1 и 3 наиболее близки, т. е. наименее удалены друг от друга, поэтому объединяются в один кластер на уровне сходства, равном 0,9 (образуют 1-й шаг). Объекты 4 и 5 объединяются при уровне сходства 0,8 (2-й шаг). На 3-м и 4-м шагах процесса образуются кластеры 1, 3, 6 и 5, 4, 2, соответствующие уровню сходства 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне сходства 0,5.

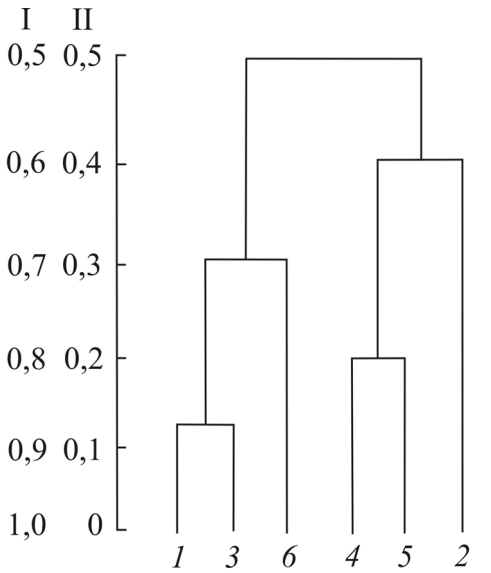


Рис. 3.1. Общий вид дендрограммы: I – сходство, II – расстояние

Вид дендрограммы зависит от выбора меры сходства или расстояния и метода кластеризации. Например, разработаны алгоритмы кластерного анализа, позволяющие проводить классификацию (группировку) многомерных наблюдений (строк и столбцов матрицы *x*) с помощью следующих мер сходства: выборочного

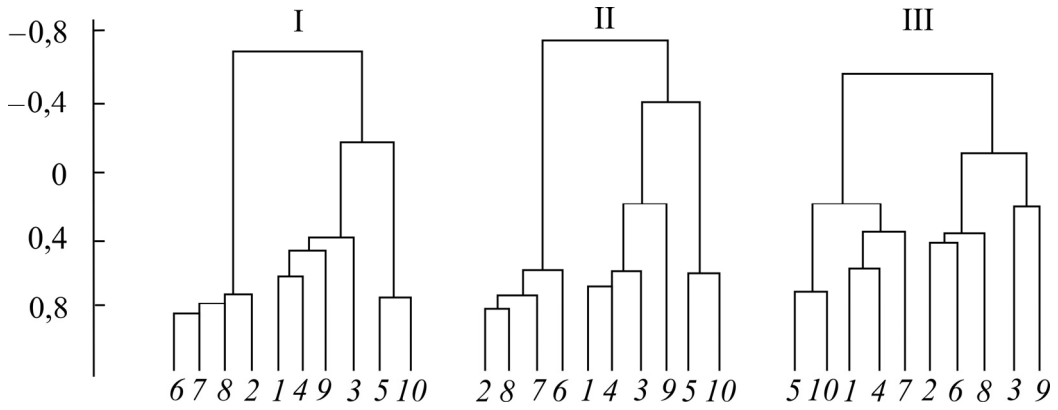


Рис. 3.2. Дендрограммы корреляционных связей почвенно-биогеохимических показателей темнохвойной (I), мелколиственной (II) фаций и залежи-пашни (III): *1* – влажность, *2* – температура; органическое вещество: *3* – водорастворимое, *4* – кислоторастворимое; ионы водной вытяжки: *5* – Ca^{2+}, Mg^{2+}, *6* – HCO_3^-; подвижные формы железа: *7* – FeO, *8* – Fe_2O_3; *9* – анаэробные бактерии; *10* – оксид углерода почвенного воздуха

коэффициента корреляции, модуля выборочного коэффициента корреляции, косинуса угла между векторами, модуля косинуса угла между векторами, эвклидова расстояния и т. д.

Выделяются группы взаимосвязанных признаков (см. рис. 3.2). Достоверно положительно связаны температура и содержание оксидов железа и гидрокарбонат-иона. На среднем уровне положительно связаны влага, подвижные формы органического вещества и анаэробные бактерии. Еще одну группу образуют концентрация щелочноземельных элементов и углекислоты почвенного воздуха. Сравнение дендрограмм показывает, что изучаемые признаки хвойной и мелколиственной фации однотипны. Это свидетельствует о внутренней однородности протекающих в них процессов и подтверждает их генетическое единство. На залежи как производной от природных ландшафтов наблюдаются менее тесные связи между показателями внутри фации.

3.1. Этапы работ в кластерном анализе

Решение задач классификации объектов с использованием кластерного анализа проводится в определенной последовательности. Многомерный анализ делится на три этапа:

- составляется таблица исходной информации с указанием объектов и их признаков;
- проводится нормализация исходной информации с использованием среднего квадратического отклонения;

• по нормализованным данным рассчитывается метрика, сроится дендрограмма и проводится содержательная интерпретация полученных результатов.

На первом этапе при формировании таблицы выбор объекта зависит от места и масштаба исследования. Каждый объект должен быть пространственно локализован и одного ранга (уровня). Показатели должны отражать существенные черты или свойства исследуемых объектов и характеризовать их всесторонне.

На втором этапе нормализация значений исходных показателей по объектам проводится потому, что исходные данные выражены обычно в разных единицах измерения и проводить между ними арифметические действия невозможно без перевода их в безразмерные единицы.

Наиболее распространенный способ нормализации показателей проводится с использованием среднего квадратического отклонения по формуле:

$$\hat{Z}_{ij} = (Z_{ij} - \bar{Z}_{ij}) / \sigma_j; \tag{3.7}$$

$$\sigma_j = \sqrt{\frac{\sum (Z_{ij} - \bar{Z}_{ij})^2}{N_j}}, \tag{3.8}$$

где $\hat{Z}_{ij}$ – нормализованная безразмерная величина; Z_{ij} – индивидуальные значения по столбцам матрицы; $\bar{Z}_{ij}$ – среднее значение по столбцам матрицы; σ_j – среднее квадратическое отклонение по столбцам; N_j – объем выборки по столбцам.

Составляется матрица нормализованных показателей.

На третьем этапе по нормализованным показателям рассчитывается метрика по одному из предложенных выше способов, учитывая условия задачи. Классификацию объектов производят приемами таксономического или факторного анализа.

При количестве координат (показателей) в многомерном пространстве более трех графически интерпретировать таксономические расстояния невозможно. Поэтому таксономические расстояния определяют на основе функции расстояний. Чаще всего используется эвклидова метрика.

На основе матрицы таксономических расстояний производится группировка объектов с использованием разных приемов, из них наиболее распространенные – вроцлавская таксономия, дендро-дерево Берри, метод дендритов.

3.2. Вроцлавская таксономия

По матрице таксономических метрик (табл. 3.3) строится граф-дерево, вершинами которого будут объекты группировки.

Таблица 3.3

Матрица таксономических метрик

Объекты	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
A	0	**1,15**	5,05	4,22	3,54	3,30	**2,56**	3,62	3,10	**1,67**
B	**1,15**	0	6,41	4,53	3,81	3,84	2,99	4,53	3,88	2,63
C	5,05	6,41	0	4,04	4,82	4,06	4,83	**3,07**	4,34	4,14
D	4,22	4,53	4,04	0	**1,66**	1,68	2,34	2,80	2,99	4,02
E	3,54	3,81	4,82	**1,66**	0	**0,96**	1,34	2,76	2,26	3,72
F	3,30	3,84	4,06	1,68	**0,96**	0	**1,11**	1,80	1,51	3,22
G	**2,56**	2,99	4,83	2,34	1,34	**1,11**	0	2,24	**1,38**	3,01
H	3,63	4,53	**3,07**	2,80	2,76	1,80	2,24	0	**1,33**	3,09
I	3,10	3,88	4,34	2,99	2,26	1,54	**1,38**	**1,33**	0	3,18
J	**1,67**	2,63	4,14	4,02	3,76	3,22	3,01	3,09	3,18	0

Порядок построения графа следующий (рис. 3.3). В каждом столбце или ряде зеркальной матрицы (по диагонали нули) находится минимальная величина метрики. Вначале откладывается в выбранном масштабе наименьшая среди метрик матрицы между объектами (*EF* = 0,96). Затем последовательно к отложенным объектам откладываем минимальные

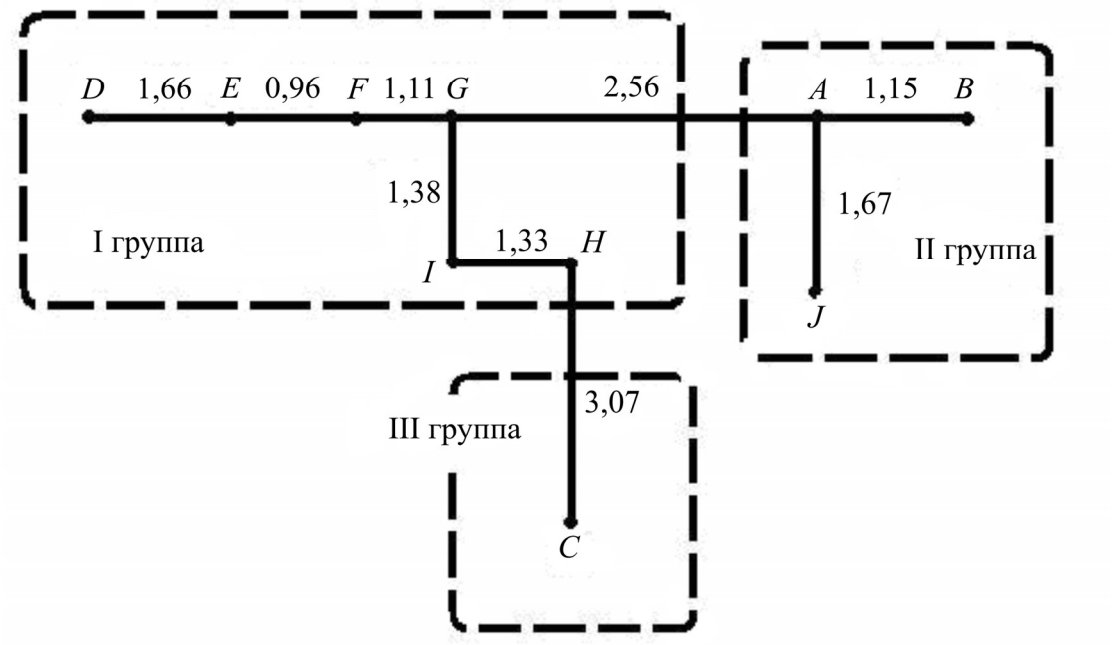


Рис. 3.3. Вроцлавский дендрит

метрики других столбцов-объектов: FG = 1,11, ED = 1,66, GI = 1,38, IH= = 1,36, HC = 3,07, GA = 2,56, AB = 1,15, AJ = 1,67. Метрика используется только один раз. Если при построении графа на нем образуется замкнутый цикл, то замыкающее ребро цикла во внимание не принимается и вместо него откладывается ребро, которое отвечает другой минимальной метрике в данном столбце матрицы.

После построения графа с нанесением всех объектов проводится группировка (классификация) объектов. Задается определенная величина таксономической метрики, которая является основой классификации. Таким образом граф разбивается на подграфы, в пределах которых объекты должны располагаться компактно (близко друг к другу) (см. рис. 3.3). В конце дается интерпретация полученных результатов с учетом исходной таблицы первоначальных данных. Чем меньшая метрика объединяет объекты на графе, тем более близкие по своим значениям исходные показатели в этих объектах.

3.3. Метод дендро-дерева Б. Берри

В матрице таксономических метрик выбирается наименьший элемент, который связывает два объекта (см. табл. 3.3): EF = 0,96. Метрика свидетельствует, что объекты E и F находятся на минимальном и одинаковом расстоянии по отношению к другим объектам. Поэтому их можно заменить одним, присвоив символ M (рис. 3.4).

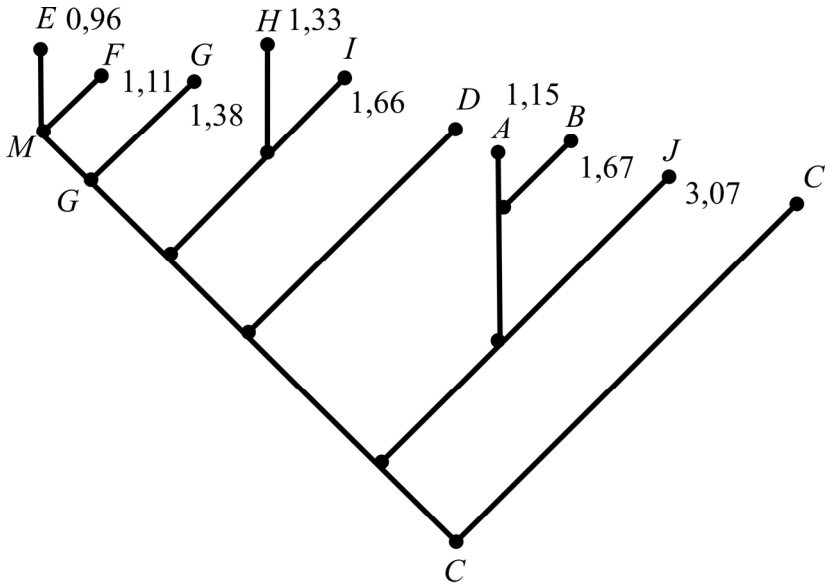


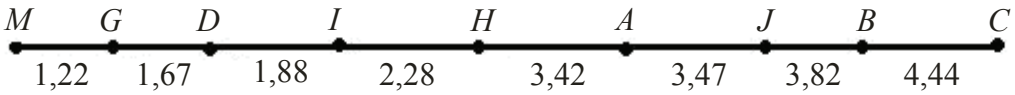
Рис. 3.4. Дендро-дерево Берри

В дальнейшем на горизонтальной линии размещаем объекты последовательно по мере увеличения их метрик с учетом связи с первыми объектами *EF*. Объект *G* связывается с *F* метрикой 1,11, объект *I* с *G* – 1,38, *I* с *H* – 1,33, *E* с *D* – 1,66. Далее связь неотложенных объектов (*A, B, J, C*) с отложенными прерывается. В таких случаях внутри этих объектов и ищем наименьшие метрики между ними: *A* и *B* связывает минимальная метрика 1,15; *A* и *J* – 1,67. Объект *C* связан наименьшей метрикой 3,07 с ранее отложенной *H*, поэтому он выделяется самостоятельно в конце по прямой линии (см. рис. 3.4).

Отложенные объекты на горизонтальной линии с минимальными метриками связываются между собой (*H* и *I* : *A* и *B*) или выделяются самостоятельно с общей наклонной линией *M* – *C*, на которой откладываются вычисленные метрики от объекта *M* (*E* – *F*) путем вычисления усредненных величин, используя данные матрицы (см. рис. 3.3) по строкам *E* – *F*:

$A = (3{,}54+3{,}30)/2 = 3{,}42$; $B = (3{,}81+3{,}84)/2 = 3{,}82$;
$C = (4{,}82+4{,}06)/2 = 4{,}44$; $D = (1{,}66+1{,}68)/2 = 1{,}67$;
$G = (1{,}34+1{,}11)/2 = 1{,}22$; $H = (2{,}76+1{,}80)/2 = 2{,}28$;
$I = (2{,}26+1{,}51)/2 = 1{,}88$; $J = (3{,}72+3{,}22)/2 = 3{,}47$.

Располагаем объекты относительно *M* по возрастающей величине на линии и производим группировку:



В нашем примере объекты можно объединить в 4 класса (*EFG*; *HID*; *ABJ*; *C*) по минимальным метрикам между объектами и по усредненным относительно объекта *M* (*E*, *F*).

Расчленение графа на подгруппы для определения количества групп объектов может производиться в процессе его построения (см. рис. 3.4): *EF*; *HI*; *ABJ*.

При делении объектов на классы важным критерием является минимизация внутригрупповой и максимизация межгрупповой дисперсии. Практически количество классов определяется априорно, т. е. по внешнему виду дендро-дерева. В выделенном классе объекты по анализируемым признакам являются сходными (однородными). Если они соседние в пространстве, то образуют однородный регион.

Пример кластерного анализа по способу «Вроцлавский дендрит»

Задача: провести зонирование территории города по предложенным признакам.

Таблица 3.4

Количественные показатели для зонирования города

Минск	Площадь застройки, га		Количество истори-ческих памятников	Количество архитек-турных памятников	Количество промыш-ленных предприятий	Площадь лесной зоны, га	Шумовое загрязнение, балл
	дерев.	бетон					
Объект № 1	0,1	25	5	10	2	2	80
Объект № 2	0,5	10	7	12	3	3	40
Объект № 3	1,5	15	3	16	5	0,5	30
Объект № 4	2,0	17	4	5	4	0,7	50
Объект № 5	3,0	18	1	4	7	5	20
Объект № 6	3,5	30	1	1	1	4	35

Этапы работы:

1. Подсчитываем сумму, среднее и сигму по столбцам:

	$\sum$	среднее	σ
1 столбец	10,6	1,8	1,2
2 столбец	115	19,2	6,6
3 столбец	21	3,5	2,14
4 столбец	48	8	5,1
5 столбец	22	3,7	1,97 и т. д.

2. Трансформируем количественные показатели в числа без измерений (табл. 3.4) с использованием формулы (3.7, 3.8).

Таблица 3.5

Нормализованные безразмерные данные

1	–1,42	0,88	0,7	0,39	–0,86	–0,31	1,96
2	–1,08	–1,4	1,63	0,78	–0,36	0,31	–0,13
3	–0,25	–0,64	–0,23	1,56	0,66	–1,25	–0,65
4	0,17	–0,33	0,23	–0,58	0,15	–1,12	0,4
5	1,00	–0,18	–1,16	–0,78	1,68	1,56	–1,18
6	1,42	1,64	–1,16	–1,37	–1,37	0,93	–0,39

3. Рассчитываем расстояния (метрику) между объектами по формуле (3.2) и проставляем в матрицу ниже:

	1	2	3	4	5	6
1	0	2,21	6,26	3,11	5,62	4,10
2	2,21	0	3,01	3,01	4,52	5,40
3	6,26	3,01	0	2,32	5,21	5,67
4	3,11	3,01	2,32	0	3,83	3,90
5	5,62	4,52	5,21	3,83	0	3,76
6	4,10	5,40	5,67	3,90	3,76	0

4. По полученным расстояниям (метрикам) по столбцам или строкам выбираем наименьшие расстояния и откладываем их в масштабе до тех пор, пока не нанесем все объекты. Первое расстояние выбирается наименьшим.

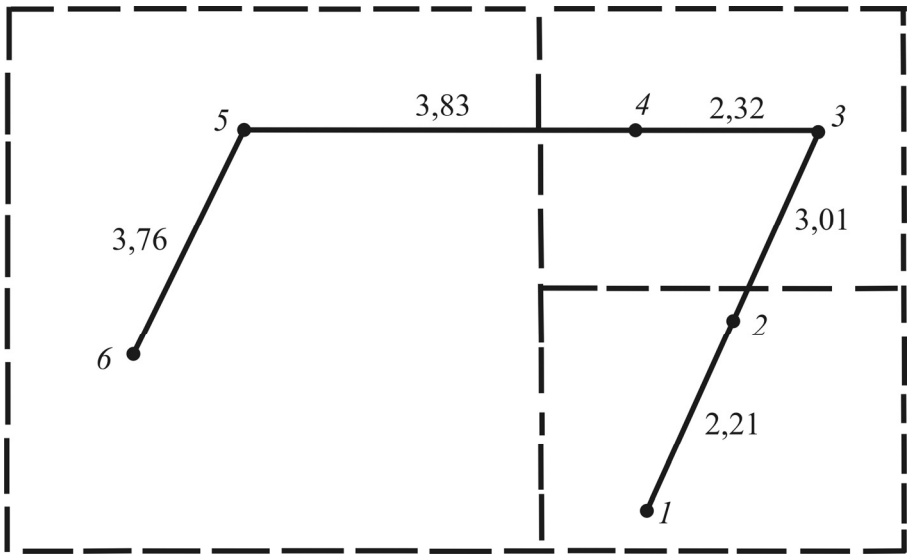


Рис. 3.5 Кластеризация объектов города («Вроцлавский дендрит»)

По рис. 3.5 выделяем 2 группы объектов, близких по первоначальным показателям: *3–4* с расстоянием от 2 до 3 и *1*, *2*, *5*, *6* с расстоянием от 3,01 до 4. Анализ первоначальной таблицы показывает, что объекты *3* и *4* характеризуются средними параметрами, а остальные объекты характеризуются как большими, так и меньшими параметрами. Вторая группа характеризуется контрастными значениями, поэтому внутри этой группы выделяются 2 подгруппы – в первую выделяем объекты *1* и *2*, а во вторую – *5* и *6*.

Глава 4

ИНФОРМАЦИОННЫЙ АНАЛИЗ

Научно-техническая революция привела к ускоренному росту объема информации в различных областях науки, включая географию. Математическая теория информации возникла, когда появилась потребность в оценке количества передаваемых сведений. Первоначально она опиралась на отдельные положения теории вероятности; постепенно вырабатывалась собственная методика, определялся свой круг задач. На современном этапе развития теория информации ставит своей целью оценку объема информации, выявление разнообразия в природе, установление различия и сходства в этом разнообразии.

По теории вероятности информацию содержат лишь такие данные, которые устраняют существующую до их получения неопределенность. Однако не всегда приходится использовать информацию вероятностного характера, например в картографии, где обычно имеют дело с определенными данными. Это привело к разработке иных подходов в теории информации: комбинаторного и алгоритмического. *Комбинаторный подход* рассматривает количество информации как функцию числа элементов в конечной совокупности. Он широко применяется, например, при измерении объема картографической информации. *Алгоритмический подход* определяет количество информации как минимальную длину программы, которая позволяет однозначно преобразовать один объект в другой.

Существует также представление об информации как о *мере разнообразия.* В целом разнообразие связано с различием, т. е. с отрицанием неразличимости. Простейшей единицей измерения информации является *элементарное различие* – различие двух объектов. Чем больше в совокупности попарно различных элементов, тем больше она содержит информации. Если рассматриваемые объекты отождествляются, то информация исчезает.

Информационный анализ применяется в некоторых областях географии при соответствующих условиях. В настоящее время разработан

способ определения количества информации, содержащейся в рельефе, подсчитан объем информации субаквального биоценоза; ведутся поиски критерия связи на примерах зависимости между физическими свойствами горных пород, климатом и растительностью, компонентами и структурными частями биогеоценозов. Теория информации помогла разработать критерий пространственной дифференциации и однородности. Информационный анализ предпочтительнее использовать для выявления закономерностей в общих, а не частных явлениях.

Весь процесс информационного анализа изучаемого явления можно разбить на следующие этапы.

Предварительный этап. При сборе материалов необходимо, чтобы сопоставляемые факторы и явления территориально и во времени соответствовали друг другу во избежание неслучайных ошибок, которые могут привести к возникновению «шума». Факторы и явления должны быть представлены возможно большим числом своих состояний. Они объединяются в более широкие классы в процессе анализа.

Анализ информации. После подготовки материала к обработке оценивается связь изучаемого явления с каждым из возможных факторов, из них отбираются наиболее информативные. Рассчитываются попарные каналы связи. Оценивается общая информативность всей совокупности выбранных факторов. Определяется величина «новой информации» и размеры косвенной связи.

Процесс моделирования и его оценка. На основе анализа частных каналов связи в сопоставлении с общими строится логическая функция зависимости явлений от совокупности факторов. Оценивается ошибка распознаваний явления по величине «шума» и для составленной логической функции. Проверку достоверности анализа целесообразно проводить и после построения частных каналов связей. Если логическая функция недостаточно полно описывает изменения состояний явления (по распределению ошибок), пытаются найти дополнительные факторы, которые смогли бы улучшить распознающую систему.

Прогноз. Если в анализ вошли материалы с достаточным разнообразием состояний и собранные на значительной территории, то прогноз можно осуществить для любой точки, характеристики которой соответствуют состояниям факторов, включенных в анализ.

Преимущество информационных методов заключается в том, что они, в отличие от статистического, не требуют применения закона нормального распределения, линейности связей, независимости признаков, метричности и упорядоченности.

С практической точки зрения важно уметь численно оценивать степень неопределенности проводимых исследований (энтропия), чтобы их

сравнить между собой. Степень неопределенности каждого опыта выражается числом K, поэтому искомая численная характеристика степени неопределенности должна являться функцией числа K. Для K =1 (неопределенность полностью отсутствует) функция должна обращаться в нуль и возрастать при увеличении числа K.

За меру неопределенности опыта (показатель энтропии), имеющего K равновероятных исходов, принято число lg K. Чаще всего пользуются логарифмами при основании два ($f(K) = \log_2 K$). В данном случае за единицу измерения степени неопределенности принимается неопределенность опыта, имеющая два равновероятных исхода (например при подбрасывании монеты равная вероятность появления орла или решки). Такая единица измерения неопределенности называется двоичной единицей (*бит*). Если пользоваться десятичными логарифмами, то за единицу степени неопределенности принимается неопределенность опыта, имеющего 10 равновероятных исходов. Такая десятичная единица примерно в 3,32 раза крупнее двоичной единицы ($\log_2 K \approx 3{,}32$).

Для перевода десятичных единиц в биты полученную величину делят на log 2 = 0,30103.

При применении натуральных логарифмов энтропия выражается в *нитах.* Если величина энтропии получена с применением натуральных логарифмов, а ее требуется перевести в биты, т. е. в двоичную систему, то этот расчет осуществляется путем деления величины в нитах на ln 2 = = 0,69315.

Чтобы перевести логарифм числа x с основанием b в логарифм с основанием a, используется формула

$$\log_a x = \log_b x / \log_b a. \qquad (4.1)$$

Форма представления вероятности для опыта, имеющего K равновероятных исходов, имеет следующий вид:

исход опыта $A_1 \quad A_2 \quad \ldots \quad A_K$.
вероятность $1/K \quad 1/K \ldots \quad 1/K$.

Поскольку общая неопределенность опыта равна lg K, то каждый отдельный исход, имеющий вероятность $1/K$, вносит неопределенность, равную $(1/K)$ lg $K = (-1/K)$ *lg* $1/K$. Аналогично этому для опыта α мера неопределенности вытекает из таблицы вероятности:

исход опыта $A_1 \quad A_2 \quad \ldots \quad A_K$,
вероятность $P(A_1) \quad P(A_2) \quad \ldots \quad P(A_K)$

и равна

$$-P(A_1)\lg P(A_1) - P(A_2)\lg P(A_2) - \ldots - P(A_K)\lg P(A_K).$$

Приведенное выражение называют *энтропией опыта* α и обозначают через H (α).

Энтропия характеризуется следующими свойствами. Ее величина не принимает отрицательных значений. Так как $0 \le P(A) \le 1$, то lg $P(A)$ не может быть положительным, а $-P(A_1)\lg P(A_1)$ – отрицательным. При $P \to 0$ произведение $P \cdot \lg P$ убывает, поэтому $\lim\limits_{P \to 0}(-P \lg P) = 0$.

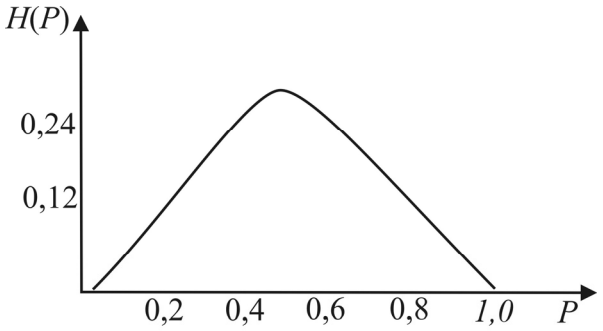


Рис. 4.1. Значение функции $-P \lg P$

Если $P(A_i)$ представляет собой большую величину (близкую к единице), то член $P(A_i)\lg P(A_i)$ будет невелик, так как при $P \to 1$ lg $P \to 0$. В области между вероятностями $P = 0{,}2$ и $P = 0{,}6$ функция P lg P принимает наибольшие значения, и соответствующая кривая меняется на графике сравнительно плавно (рис. 4.1). Поэтому в данной ситуации существенные изменения вероятности мало отражаются на величине энтропии.

П р и м е р. Предположим, что для г. Минска вероятность того, что 1 июля выпадут осадки, равна 0,4, а вероятность того, что дождя не будет, – 0,6 (опыт α_1); вероятность того, что в г. Минске 1 ноября пройдет дождь – 0,65, вероятность того, что выпадет снег, – 0,15 и вероятность того, что 1 ноября вовсе не будет осадков, – 0,2 (опыт α_2). В какой из двух указанных дней погоду следует считать более неопределенной?

Опыты α_1 и α_2 по выяснению состояния погоды представим следующим образом:

Опыт α_1

исход опыта	дождь	без дождя
вероятность	0,40	0,60

Опыт α_2

исход опыта	дождь	снег	без дождя
вероятность	0,65	0,15	0,20

Энтропия обоих опытов равна

$$H(\alpha_1) = -0{,}4 \lg 0{,}4 - 0{,}6 \lg 0{,}6 \approx 0{,}292;$$

$$H(\alpha_2) = -0{,}65 \lg 0{,}65 - 0{,}15 \lg 0{,}15 - 0{,}2 \lg 0{,}2 \approx 0{,}385.$$

Поскольку величина энтропии в опыте α_2 больше величины энтропии в опыте α_1, погоду 1 ноября в г. Минске следует считать более неопределенной, чем 1 июля. При этом учитывается процент случаев, когда прогноз оправдывается: вероятность $P = 0{,}4 = 40$ %, $P = 0{,}6 = 60$ % и т. д.

При сравнении опытов заключение представляет интерес для оценки качества предсказания явлений.

4.1. Показатели неопределенности объектов

Многие объекты и процессы в ландшафте характеризуются неоднородностью полученных данных. Для ее оценки лучше всего подходит показатель меры неопределенности, или показатель энтропии. Его можно рассчитывать для системы, которая принимает различные состояния с установленными вероятностями.

Показатель энтропии определяется вероятностями всех элементарных событий данного поля и рассчитывается по формуле

$$H_{(P_1,P_2,\ldots,P_K)} = -\sum_{i=1}^{n} P_i \log_2 P_i, \qquad (4.2)$$

где P_1, P_2 ... – вероятности данного поля, которые можно заменить частостями распределений; $\log_2$ – двоичный логарифм вероятности; n – число классов совокупности; P_i – вероятность отдельных исходов опытов.

Показатель энтропии одиночного события выражается через логарифм его вероятности:

$$H_i = -\log_2 P_i.$$

При использовании критерия энтропии можно объективно решать вопрос о наличии полезной информации, заключенной в опыте. Группы, выделяемые в эксперименте, рассматриваются с точки зрения теории вероятности как поле, состоящее из независимых событий. Например, для получения репрезентативных данных при анализе образцов почв с целью оценки обеспеченности растений элементами питания следует провести серию экспериментов в разное время вегетационного периода. Это обусловлено различной степенью их потребления из почвы в разные фазы роста и развития. Соответственно будет меняться и содержание химических элементов в почве. Не учитывая последнего, можно сделать ошибочные выводы.

П р и м е р. Предположим, что лучшим временем для отбора почвенных образцов является период с 21 апреля. В отобранных образцах почв определяется содержание подвижной формы интересующего нас элемента питания (например бора). Отобран 431 образец в указанный интервал времени на определенном участке. При распределении образцов по классам были получены частоты (табл. 4.1). Затем по ним рассчитаны частости и натуральные логарифмы частостей. Перемножая показатели и затем суммируя произведения, имеем величину энтропии в нитах (H = 1,6108 нит). Для перевода в биты делим ее на ln 2: H = 1,6108 : 0,69315 = 2,324 битов. Аналогично вычисляется величина энтропии для других ландшафтных условий. Получив ряд показателей энтропии, делаем выводы о наиболее полезной информативности определенного периода. Чем меньше величина энтропии, тем информативнее период, так как сни-

жение энтропии приводит к увеличению упорядоченности. Предположим, что получен ряд показателей энтропии (в битах):

Период	май	июнь	июль	август
Энтропия	2,450	2,225	2,135	2,057

Отсюда следует, что наиболее информативен отбор почвенных образцов в случае H = 2,057 битов (в августе). Этот период характеризуется наименьшим энтропийным показателем.

Показатель энтропии можно использовать при анализе развития явлений – от беспорядка к организованности (например, смена стадий развития речной системы, формирование сквозной речной долины).

Таблица 4.1

Расчет показателя энтропии для установления оптимального времени отбора образцов

Границы класса, дни	Середина класса, x	Частота, f	Частость, $\omega = f/N$	ln ω	– ω ln ω
1–5	3	5	0,0116	–4,46541	0,0518
6–10	8	38	0,0882	–2,42815	0,2142
11–15	13	120	0,2784	–1,27870	0,3560
16–20	18	160	0,3712	–0,99101	0,3679
21–25	23	68	0,1578	–1,84643	0,2914
26–30	28	20	0,0464	–3,07046	0,1425
31–35	33	12	0,0278	–3,58272	0,0996
36–40	38	6	0,0139	–4,27587	0,0594
41–45	43	1	0,0023	–6,07485	0,0140
46–60	48	1	0,0023	–6,07485	0,0140
	$\sum = 431$		0,9999		1,6108

4.2. Применение информационного анализа в картографии

При составлении карты необходимо использовать методы оценки объема информации (оценка абсолютного объема содержания карты). Визуальная оценка карты, например «богатое содержание», «малосодержательна», не несет в себе элементов достоверности. Дать объективную оценку нагрузки карты можно с помощью информационного анализа. Для этой цели вводится понятие *информационная емкость карты* – количественная мера объема содержания карты, выражающая в условных единицах общее количество информации, которое можно получить. Информационная емкость может быть выражена в легенде карты отдельным условным знаком (в битах).

Кроме оценки абсолютного объема содержания карты, важна степень полноты отображения исследуемого явления (отношение объема содержания карты к ее структурной модели, считающейся условно полной). Та часть информационной емкости карты, которая отображает ее тематическое содержание, названа *специальной информационной емкостью карты* (количество отображаемых показателей и их градаций и число характеризуемых ими географических объектов).

Для расчета специальной информационной емкости J_S рекомендуется использовать при определенных условиях следующие формулы, в которых применяются двоичные логарифмы.

1. На карте нанесен один вид объектов, характеризующихся одним показателем при числе объектов N и градаций D:

$$J_S = \log_2 N D = \log_2 N + \log_2 D.$$

2. Для N объектов одного вида приведено два показателя с числом градаций D_1 и D_2.

$$J_S = \log_2 N D_1 D_2 = \log_2 N + \log_2 D_1 + \log_2 D_2.$$

Если число показателей равно m и число градаций каждого показателя $D_1, D_2, \ldots, D_i, \ldots, D_m$, то формула примет вид

$$J_S = \log_2 (N_1 \prod_{i=1}^{m} D_i) = \log_2 N_1 + \sum_{i=1}^{m} \lg D_i.$$

3. На карте выделено два вида объектов N_1, N_2, каждый из них имеет по одному показателю, число градаций соответственно D_1 и D_2:

$$J_S = \log_2 (N_1 D_1 + N_2 D_2).$$

4. На карте выделено два вида объектов N_1, N_2, по каждому объекту приведено два показателя. Число градаций по показателям составляет соответственно по две:

$$J_S = \log_2 (N_1 A_1 A_2 + N_2 B_1 B_2),$$

где A_1, A_2 и B_1, B_2 – первая и вторая градации первого и второго вида объектов соответственно.

Имеется оригинальный способ применения информационных функций при анализе карт с использованием натуральных логарифмов для характеристики неоднородности картографического изображения. Предположим, на участке карты показано n районов (ареалов). Требуется определить и выразить количественную меру их неоднородности, или степень разнообразия картографического содержания. При наличии на карте лишь одного участка показатель неоднородности равен нулю ($H = 0$). При увеличении числа ареалов неоднородность участка карты увеличи-

вается, и показатель H будет возрастать. Если число районов на участке карты остается постоянным, то неоднородность картографического изображения будет зависеть от площади S_i каждого района. Неоднородность достигает максимума (H = max), если их площади равны между собой. Всем перечисленным условиям отвечает функция энтропии для дискретного распределения:

$$H(\alpha) = -\sum_{i=1}^{n} S_i \ln S_i. \qquad (4.3)$$

Показатель энтропии может быть вычислен для явлений, имеющих на картах абсолютную или относительную числовую характеристику (например, количество осадков) и не имеющих никакой количественной характеристики (границы распространений форм рельефа). Для подсчета показателя абсолютной энтропии необходимо определить вероятность наличия каждого района на карте, т. е. отношения площади каждого контура (S_i) к площади всех контуров n на карте:

$$P_i = S_i / \sum_{i=1}^{n} S_i. \qquad (4.4)$$

Вероятность того, что на карте будут помещены участки с максимальной площадью, будет тем больше, чем мельче масштаб.

Показатель относительной энтропии удобен для сравнительной характеристики неоднородности картографического изображения на двух или более разных участках. Вычисляется по формуле

$$H(\alpha)_r = -\sum_{i=1}^{n} P_i \ln P_i / \ln n. \qquad (4.5)$$

Приведем конкретный пример. Для двух участков поверхности с кратерным расчленением (C_1, C_2) измерены диаметры кольцевых структур и подсчитаны их вероятности по формуле (4.4). Вычисленные значения энтропии для участков C_1 и C_2 соответственно равны: $H(C_1)$ = 1,114; $H(C_2)$ = 0,738; $H(C_1)$ = 0,71; $H(C_2)$ = 0,53. Отсюда следует, что второй участок C_2 более однороден по кратерному расчленению поверхности, чем первый C_1, так как показатели абсолютной энтропии по участку C_2 меньшие, чем по участку C_1.

Информационные показатели предлагается также использовать и для оценки степени взаимного соответствия явлений на картах разного содержания. Пусть на одной карте изображено явление Z, состоящее из n ареалов или градаций $Z_1, Z_2, \ldots, Z_i, \ldots, Z_n$ с вероятностями $P_1, P_2, \ldots, P_i, \ldots, P_n$. На другой карте отражено явление L, имеющее m ареалов $l_1, l_2, \ldots, l_j, \ldots, l_m$

с вероятностями P_1, P_2, ..., P_j, ..., P_n. Если эти явления независимы, то их совместная энтропия равна сумме индивидуальных энтропий:

$$H(Z+L) = H(Z) + H(L) = -\sum_{i=1}^{n} P_i \ln P_i - \sum_{j=1}^{m} P_j \ln P_j. \quad (4.6)$$

Если некоторые из ареалов Z_i совпадают с ареалами l_j, то энтропия системы ZL выразится следующим образом:

$$H(ZL) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \ln P_{ij}, \quad (4.7)$$

где P_{ij} – вероятность совпадения ареалов.

Энтропия независимых событий всегда больше энтропии зависимых событий: $H(Z+L) > H(ZL)$, причем разность $T(ZL)$ служит показателем взаимного соответствия явлений Z и L и отражает уменьшение неопределенности за счет внутренних ограничений в системе ZL:

$$T(ZL) = H(Z + L) - H(Z \cdot L). \quad (4.8)$$

Взаимосвязь можно оценить отношением, которое называется информационным коэффициентом:

$$K(ZL) = [T(ZL) / H(ZL)] \cdot 100. \quad (4.9)$$

Информационный коэффициент изменяется в пределах от 0 до 100 %. Если $K(ZL) = 0$, то явления Z и L не связаны между собой. При $K(ZL) = 100$ % имеет место однозначное функциональное состояние между явлениями, т. е. P_i, P_j и P_{ij} равны между собой и $H(Z) = H(L) = H(ZL)$. Тогда $T(ZL) = H(Z + L) - H(ZL) = H(ZL)$, значит, $K(ZL) = 100$ %.

П р и м е р. Предположим, нам необходимо сравнить связь контуров почв (Z) и растительности (L) для одного и того же района, но нанесенных на отдельные

Таблица 4.2

Решетка для вычисления информационных показателей

L	Z				nL
	1	2	3	…	$n_L P_L - P_L \ln P_L$
1	3	9		…	14
	0,005	0,010		…	0,020
	0,027	0,046		…	0,084
2			1		9
			0,002		0,020
			0,012		0,084
3	…	…	…	…	…
$n_Z P_Z - P_Z \ln P_Z$	5 0,008 0,038	14 0,02 0,084	… … …	… … …	630 1,000 5,400

специальные карты. На обе карты помещаем квадратную точечную палетку. Пусть всего на участке разместилось 630 точек. В каждой из них отмечены номера почвенного и растительного контуров. Для расчета показателей составляется информационная решетка (табл. 4.2). В каждой клетке таблицы, образованной пересечением строк и столбцов, проставлено по 3 показателя: 1) количество точек, попавших одновременно в пределы *i*-го (почвенного) и *j*-го (растительного) контуров (верхнее число); 2) величина вероятности P_{ij} (среднее число); 3) произведение $P_{ij}\ln P_{ij}$ (нижнее число). Результаты суммируются для контуров растительности и почвенных контуров. После вычисления информационных функций получены результаты:

$$H(Z) = -\sum_{i=1}^{n} P_i \ln P_i = 2,060;$$

$$H(L) = -\sum_{j=1}^{m} P_j \ln P_j = 1,882;$$

$$H(Z)_r = 2,060 : \ln 12 = 0,83;$$

$$H(L)_r = 1,882 : \ln 13 = 0,73;$$

$$H(Z + L) = H(Z) + H(L) = 3,192;$$

$$H(ZL) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \ln P_{ij} = 3,456;$$

$$T(ZL) = H(\mathrm{Z} + \mathrm{L}) - H(Z \cdot L) = 0,486;$$

$$K(ZL) = [T(ZL) / H(ZL)] \cdot 100 = 14,1\ \%.$$

Таким образом, значение абсолютной и относительной энтропии для явления *L* меньше, чем для *Z*, несмотря на увеличение числа градаций. Карта растительности обладает большей однородностью ($H(L)$ = 1,882), чем почвенная ($H(Z)$ = 2,060).

Глава 5

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Основоположниками теории корреляции считаются английские биометрики Ф. Гальтон (1822–1911) и К. Пирсон (1857–1936). Термин «корреляция» означает соотношение, соответствие. Представление о корреляции как о взаимозависимости случайных переменных величин лежит в основе статистической теории корреляции – изучение зависимости вариации признака от окружающих условий. Одни признаки выступают в роли *влияющих* (*факторных*), другие – на которые влияют – *результативных*. Зависимости между признаками могут быть *функциональными и корреляционными.* Функциональные связи характеризуются полным соответствием между изменением факторного признака и изменением результативной величины. Каждому значению признака-фактора соответствует определенное значение результативного признака. В корреляционных связях между изменением факторного и результативного признаков нет полного соответствия. В сложном взаимодействии находится сам результативный признак. Поэтому результаты корреляционного анализа имеют значение в данной связи, а интерпретация этих результатов в общем виде требует построения системы корреляционных связей. Они характеризуются множеством причин и следствий, и с их помощью устанавливается тенденция изменения результативного признака при изменении величины факторного признака. Например, на производительность труда влияют факторы степени совершенствования техники и технологии, уровень механизации и автоматизации труда, специализации производства, текучесть кадров и т. д.

В природе и обществе явления и события протекают по характеру корреляционной связи, когда при изменении величины одного признака существует тенденция изменения другого признака. Корреляционная связь – это частный случай статистической связи. *Корреляционный анализ используется при установлении тесноты зависимости между явлениями, процессами, объектами.*

Целью исследования часто бывает установление взаимосвязи (корреляции) между признаками. Знание зависимости дает возможность решать кардинальную задачу любого исследования – возможность предвидеть, прогнозировать развитие ситуации при изменении влияющего фактора. С помощью корреляции можно дать лишь формальную оценку взаимосвязей. Поэтому прежде чем приступать к вычислению коэффициентов корреляции между любыми признаками, следует теоретически установить, имеется ли между этими признаками взаимосвязь. Ведь формально статистика может доказать несуществующие связи, например между высотой здания в городе и урожайностью пшеницы в фермерских хозяйствах.

Связь между явлениями (корреляция) определяется путем постановки опытов, статистического анализа. Корреляцию не следует отождествлять с причинностью. Однако необходимо иметь в виду, что доказательство математической связи должно опираться на реальную зависимость между явлениями. Например, минерализация воды понижается с севера на юг Беларуси, в этом же направлении понижается содержание питательных веществ в почве. Между рассматриваемыми показателями может быть получена положительная достоверная зависимость. Однако степень минерализации воды не определяет оптимальное содержание питательных веществ в почве. Иначе в ландшафтах пустынь плодородие было бы максимальным, так как здесь максимальная минерализация воды (почвенно-грунтовые воды солоноватые), а это противоречит истине. Поэтому проведение подобной связи в ландшафтах пустынь бессмысленно.

Любой показатель связи служит приближенной оценкой рассматриваемой зависимости и не является гарантией существования жесткой (функциональной) соподчиненности. Отсутствие жесткой зависимости в природе и обществе способствует саморегуляции процессов, явлений, систем.

По направлению связь может быть *прямой* и *обратной*; по характеру – *функциональной* или *статистической* (*корреляционной*); по величине – *слабой, средней* или *сильной*; по форме – *линейной* и *нелинейной*; по количеству коррелируемых признаков – *парной и множественной*.

Функциональная зависимость характерна для геометрических форм, технических систем, когда каждому значению одного признака соответствует точное значение другого. Это пример взаимосвязи площади прямоугольника и длины его одной из сторон. Такая зависимость полная, или исчерпывающая.

Выделяют несколько видов парной корреляционной связи:

• параллельно-соотносительную, или ассоциативную, когда оба признака изменяются сопряженно, частично под действием общих причин и

следствий (приуроченность растительности и почв к определенным формам рельефа; развитие промышленности и рост населения к сырьевым ресурсам);

• субпричинную, когда один фактор выступает как отдельная причина сопряженного изменения признака (связь биомассы с количеством осадков; рост населения и рождаемости);

• взаимоупреждающую, когда причина и следствие, находясь в устойчивой взаимной связи, последовательно влияют друг на друга (влажность воздуха и осадки).

Если на признак влияет несколько факторов, то приходится оценивать множественную корреляцию. *Множественная корреляция* служит основой выявления связей между признаками, но требует строгой нормальности и прямолинейности распределения, поэтому использование ее может быть затруднено. С ростом числа переменных объем вычислительных работ увеличивается пропорционально квадрату числа переменных. В этом случае труднее оценивать значимость результатов, так как увеличиваются ошибки коэффициентов корреляции. Практически в таких случаях ограничиваются изучением лишь главных факторов. Однако характер влияния главных факторов на признак более детально и точно исследуют путем факторного анализа.

В практической работе по установлению корреляции между признаками и явлениями необходимо придерживаться следующей последовательности:

• на основании проведенных исследований предварительно определяют, существует ли связь между рассматриваемыми признаками;

• если связь между ними существует, устанавливают ее форму, направление и тесноту, используя график.

Вначале составляются сопряженные вариационные ряды, в которых следует определить аргумент x и функцию y:

x	10	12	16	18	21	23	25	30
y	2	4	5	7	8	9	9	10

По сопряженным вариантам строится график, который помогает установить вид зависимости между аргументом и функцией. От формы корреляционной связи зависит дальнейшая обработка экспериментальных или статистических данных. Линейная зависимость предполагает вычисление коэффициента корреляции r, а нелинейная – корреляционного отношения η (рис. 5.1). Степень рассеяния частот или вариант относительно линии регрессии на графике указывает ориентировочно на тесноту связи: чем меньше рассеяние, тем сильнее связь (рис. 5.2).

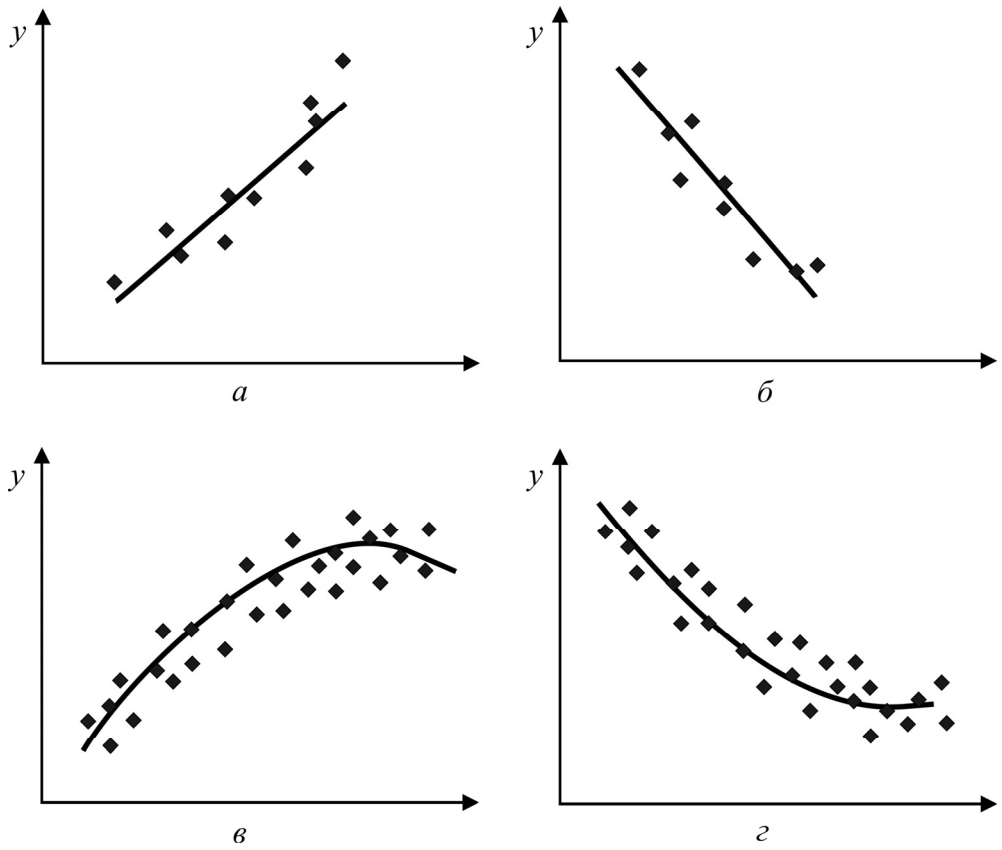


Рис. 5.1. Форма корреляционной связи:
а – прямая линейная; *б* – обратная линейная;
в – параболическая; *г* – гиперболическая

Корреляционный анализ решает следующие задачи:

- установление направления и формы связи,
- оценка тесноты связи,
- оценка репрезентативности статистических оценок взаимосвязи,
- определение величины детерминации (доли взаимовлияния) коррелируемых факторов.

Для оценки связи используют следующие численные критерии (коэффициенты) корреляционной связи:

- коэффициент корреляции (r) при линейной зависимости,
- корреляционное отношение (η) при нелинейной зависимости,
- коэффициенты множественной регрессии,
- ранговые коэффициенты линейной корреляции Пирсона или Кендэла.

5.1. Линейная корреляция

Для установления формы зависимости по исходным (x, y) строится график. В случае линейной зависимости вычисляется коэффициент корреляции (r), при нелинейной – корреляционное отношение (η). В зависи-

мости от величины разброса точек на графике можно предварительно установить форму (см. рис. 5.1) и тесноту (см. рис. 5.2) связи.

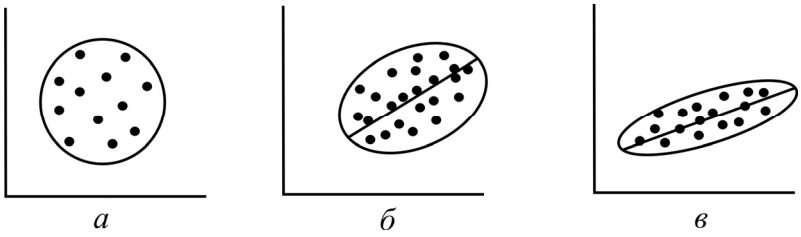


Рис. 5.2. Степень рассеяния частот и величина связи:
а – $r \approx 0$; *б* – $r \approx 0{,}5$; *в* – $r \approx 0{,}7$

Линия регрессии по координатам точек на графике проводится таким образом, чтобы точки в равном количестве находились по обе стороны линии. Более точное значение r получаем расчетным способом как при прямой (r от 0 до 1), так и при обратной (r от 0 до –1) зависимости:

$$r = \frac{\sum (x_i - M_x)(y_i - M_y)}{\sqrt{\sum (x_i - M_x)^2 \sum (y_i - M_y)^2}}, \qquad (5.1)$$

где $(x_i - M_x)$, $(y_i - M_y)$ – отклонения значений индивидуальных вариант x_i и y_i от их средних значений M_x и M_y.

Более простой алгебраический расчет коэффициента вариации с учетом объема выборки (n):

$$r = \frac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \dfrac{\left(\sum y_i\right)^2}{n}\right)}}. \qquad (5.2)$$

Исходные данные и суммы по ним получаем из представленной формы:

x_i	x_i^2	$x_i - M_x$	y_i	y_i^2	$y_i - M_y$	xy	$(x_i - M_x)(y_i - M_y)$

Принимается следующая характеристика тесноты корреляционной связи: если r (η) = 0 ± 0,4, то связь считается слабой; от ± 0,4 до ± 0,7 – средняя; от ± 0,7 до ± 1 – сильная; $r = \pm 1$ и $\eta = 1$ – связь функциональная.

Достоверность вычисленного коэффициента корреляции может быть установлена двумя путями: сравнением с табличным значением r (прил. 7); через критерий Стьюдента. Если $r_{выч} > r_{табл}$, то влияние фактора на признак достоверно; если меньше табличного – недостоверно.

При использовании критерия Стьюдента для доказательства достоверности r вначале рассчитывают стандартную ошибку коэффициента корреляции:

$$m_r = \sqrt{(1-r^2)/(N_n-2)}, \qquad (5.3)$$

где N_n – число сопряженных пар в сравниваемых выборках.

Значение коэффициента корреляции записывают с учетом его ошибки и уровня значимости: $r_{0,95\ (0,99)} \pm m_r$. Затем вычисляют критерий Стьюдента для коэффициента корреляции:

$$t_r = r / m_r. \qquad (5.4)$$

Критерий Стьюдента можно рассчитать иначе:

$$t_r = r\sqrt{N_n - 2} / \sqrt{1-r^2}. \qquad (5.5)$$

Если вычисленный критерий Стьюдента больше табличного (прил. 4), то зависимость существенна, если меньше – не достоверна. Приближенная оценка статистической достоверности r осуществляется исходя из того, что абсолютное значение r должно превышать ошибку (m_r) в два и более раза.

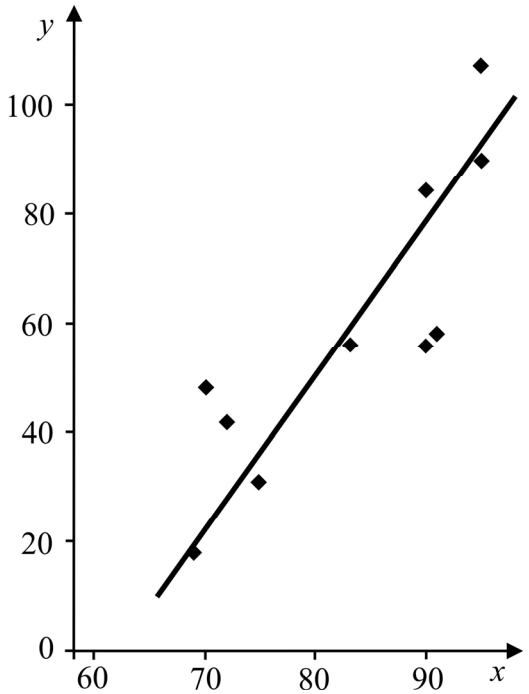


Рис. 5.3. Зависимость содержания подвижного марганца (y) от гидролитической кислотности (x)

П р и м е р. Исследованиями установлено, что на содержание подвижного марганца в почве влияет реакция среды. Необходимо доказать достоверность установленной зависимости. Получены следующие исходные данные (x – гид-

ролитическая кислотность, мэкв на 100 г почвы; y – содержание подвижного марганца, мг/кг почвы):

x	83	72	69	90	90	95	95	91	75	70
y	56	42	18	84	56	107	90	58	31	48

Вначале строим график (рис. 5.3), который указывает на существование между исследуемыми показателями положительной линейной зависимости, что требует вычисления коэффициента корреляции. Для этого проводим расчет данных по таблице исходных данных (5.1). Необходимые суммарные результаты подставляем в формулу (5.1) и вычисляем коэффициент корреляции:

$$r = 2302 / \sqrt{1000 \cdot 6854} = 0{,}88.$$

Таблица 5.1

Исходные данные для расчета коэффициента корреляции

x_i	$x_i - M_x$	$(x_i - M_x)^2$	y_i	$y_i - M_y$	$(y_i - M_y)^2$	$(x_i - M_x) \cdot (y_i - M_y)$
69	–14	196	18	–42	1764	558
70	–13	169	48	–12	144	156
72	–11	121	42	–18	324	198
75	–8	64	31	–29	841	232
83	0	0	56	–4	16	0
90	7	49	84	24	576	168
90	7	49	56	–4	16	–28
91	8	64	68	8	64	64
95	12	144	90	30	900	360
95	12	144	107	47	2209	564
$\sum = 830$ $M_x = 83$	$\sum = 0$	$\sum = 1000$	$\sum = 600$ $M_y = 60$	$\sum = 0$	$\sum = 6854$	$\sum = 2302$

Поскольку $r_{выч} = 0{,}88 > r_{табл} = 0{,}77$ при $P = 0{,}99$ и $\nu = 8$, то зависимость между содержанием подвижного марганца и гидролитической кислотностью достоверная линейная положительная.

Определим также достоверность зависимости с использованием критерия Стьюдента t по формуле (5.5): $t_r = 0{,}88 \cdot \sqrt{10-2} / \sqrt{1-0{,}88^2} = 5{,}27$.

Поскольку $t_{выч} = 5{,}27 > t_{табл} = 3{,}36$ при $\nu = 8$ и $P = 0{,}99$ (см. прил. 4), то зависимость между данными показателями доказана (достоверна).

В рассмотренном примере оба критерия подтвердили достоверную линейную положительную зависимость между содержанием подвижного марганца и гидролитической кислотностью.

Таким образом, достоверность связи устанавливается путем сравнения r (η) расчетного (фактического) и r (η) теоретического (табличного). Если $r(\eta)_{выч} > r(\eta)_{табл}$ при учете степени свободы (ν) вариационных рядов и уровня вероятности $P = 0{,}95$ и $0{,}99$, то зависимость между призна-

ками доказана без учета величины r (η). Регрессионный анализ обычно является продолжением корреляционного в случае, если r (η) $\geq \pm 0{,}7$.

Коэффициент детерминации (*причинности*) R^2 (D^2) – это коэффициент корреляции, возведенный в квадрат, например $R^2 = r^2 = 0{,}2^2 = 0{,}04$. С помощью коэффициента детерминации можно установить долю влияния анализируемого факторного признака на результативный признак. В случае, когда $R^2 = 0{,}04$, можно утверждать, что доля влияющего фактора (x) на признак (y) составляет 4 %. Следовательно, на долю других факторов приходится 96 % влияния.

5.2. Нелинейная корреляция

Зависимость между признаками не всегда выражается в виде прямой линии. Если рассеяние точек на графике приближается к кривой линии (см. рис.5.1, *в, г*), то зависимость устанавливается с использованием корреляционного отношения (η), величина которого изменяется только от 0 до 1. Для него теоретические значения приводятся отдельно в таблице или находятся при перерасчете его в критерий Стьюдента. При нелинейной корреляции вычисляется корреляционное отношение (η).

Для установления формы связи иногда используется *критерий криволинейности* в случаях, когда кривая линия мало отличается от прямой. Существует несколько способов оценки степени криволинейности. Рассмотрим два из них.

Первый способ менее точный. Оценка степени криволинейности определяется по разности коэффициента корреляции и корреляционного отношения использованием неравенства: $\eta^2 - r^2 \geq 0{,}1$. Корреляция считается криволинейной, если полученный результат соответствует этому неравенству. Предварительно следует рассчитать между сравниваемыми выборками r и η.

Второй способ оценки степени криволинейности связан с применением критерия Стьюдента:

$$t = 0{,}5\sqrt{\frac{N}{(\eta^2 - r^2)^{-1} - 2 + (\eta^2 + r^2)}} \geq 3.$$

Если $t_{выч} < 3$ или $t_{выч} < t_{табл}$, то рассматриваемая связь несущественно отклоняется от прямолинейной, поэтому относим ее к линейной. В других случаях связь между признаками относят к криволинейной и рассчитывается корреляционное отношение.

Корреляционное отношение, как и коэффициент корреляции, используется для оценки прямой и обратной зависимости между признаками.

Оценка прямой нелинейной зависимости между признаками. Прямая нелинейная зависимость определяется как параболическая. Расчет корреляционного отношения производится по формуле с использованием функции y:

$$\eta = \sqrt{\frac{n\sum(\bar{y} - M_y)^2}{\sum(y - M_y)^2}}, \tag{5.6}$$

где $\bar{y}$ – среднее арифметическое частных групп по y_i; n – число вариант в частной группе; $\bar{y} - M_y$ – отклонение общего среднего (M_y) от средних арифметических частных групп ($\bar{y}$).

Ошибка корреляционного отношения независимо от способа расчета вычисляется следующим образом:

$$m_\eta = \sqrt{(1-\eta^2)/(N_{nap} - 2)}. \tag{5.7}$$

Критерий Стьюдента определяется с использованием η:

$$t_\eta = \eta / m_\eta. \tag{5.8}$$

Если $t_{выч} > t_{табл}$, то корреляционное отношение признается достоверным.

П р и м е р. Следует установить, существует ли зависимость между температурой воздуха (x, °С) и упругостью водяного пара (y, мбар) по шести метеорологическим постам Беларуси исходя из следующих данных:

x_i	14,7	14,9	15,3	15,6	16,0	16,7
y_i	13,3	13,7	14,2	14,5	14,7	14,6

При построении графика получена кривая, близкая к параболе (рис. 5.4).

По исходным данным (табл. 5.2) рассчитываем корреляционное отношение между x и y. Выборку разбиваем на частные группы по значениям y. Их должно быть не менее трех. В нашем примере выделены две частные группы для сокращения расчета. Для частных групп рассчитываются средние ($\bar{y}$) и отклонение их от общей средней для выборки (M_y), а также отклонения индивидуальных вариант выборки (y_i) от общей средней (M_y). Сумму отклонений в квадрате из табл. 5.2 заносим в формулу (5.6) и вычисляем η.

$$\eta_y = \sqrt{(3 \cdot 0{,}40)/1{,}92} = 0{,}79.$$

Ошибку корреляционного отношения находим по формуле (5.7):

$$m_\eta = \sqrt{(1-(0{,}78)^2)/(6-2)} = 0{,}31.$$

Достоверность результатов определяем по критерию Стьюдента (5.8):

$$t_\eta = 0{,}78 / 0{,}31 = 2{,}51.$$

Поскольку $t_\eta = 2{,}51 < t_{табл} = 2{,}78$ при $P = 0{,}95$ для $\nu = 4$ (см. прил. 4), то значение корреляционного отношения следует признать недоказанным, а зависимость между температурой воздуха и упругостью водяного пара положительна, но не достоверна.

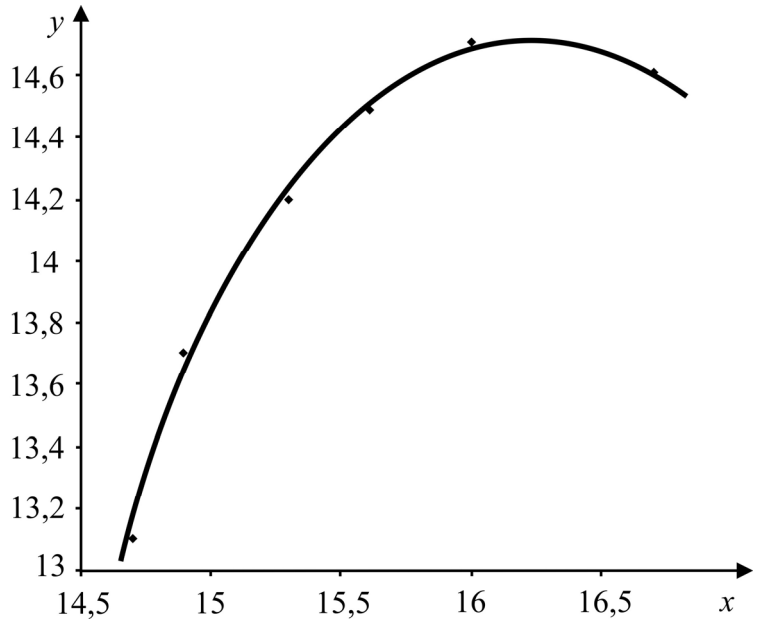


Рис. 5.4. Кривая зависимости упругости водяного пара (x) от температуры воздуха

Таблица 5.2

Исходные данные по упругости водяного пара

y_i	$\sum y_i$ по группам	$\bar{y}$, среднее по группам	$\bar{y} - M_y$	$(\bar{y} - M_y)^2$	$y_i - M_y$	$(y_i - M_y)^2$
I группа						
13,1					–1,03	1,06
13,7	41,0	13,7	–0,43	0,18	–0,43	0,18
14,2					0,07	0,005
II группа						
14,5					0,37	0,14
14,7	43,8	14,6	0,47	0,22	0,57	0,32
14,6					0,47	0,22
$\sum = 84{,}8$ $M_y = 14{,}13$			$\sum = 0{,}04$	$\sum = 0{,}40$	$\sum = 0{,}02$	$\sum = 1{,}92$

Оценка обратной нелинейной зависимости между признаками. Алгоритм вычисления и доказательств при расчете корреляционного от-

ношения обратной нелинейной (гиперболической) зависимости аналогичен алгоритму прямой нелинейной зависимости. Различие состоит в том, что в качестве исходных вариант используется выборка со значениями x.

Для нелинейной обратной (гиперболической) зависимости корреляционное отношение определяется с использованием аргумента x по формуле (5.9), условные обозначения в которой аналогичны формуле (5.6):

$$\eta_x = \sqrt{\frac{n\sum(\bar{x}_{\text{гр}} - M_x)^2}{\sum(x_i - M_x)^2}}. \qquad (5.9)$$

Расчет производится по влияющему фактору (x_i) после составления таблицы по форме и получения необходимых сумм:

x_i	$\sum x_i$ по группам	$\bar{x}_{\text{гр}}$	$\bar{x}_{\text{гр}} - M_x$	$(\bar{x}_{\text{гр}} - M_x)^2$	$x_i - M_x$	$(x_i - M_x)^2$

Расчетные величины η по x сравнивают с табличными (теоретическими) для степени свободы ($\nu = N_{\text{пар}} - 1$) и $P = 0{,}95$ и $0{,}99$. Если расчетная величина больше табличной, то можно утверждать с уверенностью о наличии достоверной зависимости между признаком и фактором.

Для всех коэффициентов можно рассчитать их ошибки: $r \pm m_r$; $\eta \pm m_\eta$.

При расчете η с использованием выборочных вариант x и y можно также применить следующие формулы с известными обозначениями:

$$\eta_x = \sqrt{\frac{\sum(x_i - M_x)^2 - (x_i - \bar{x}_{\text{гр}})^2}{\sum(x_i - M_x)^2}}, \qquad (5.10)$$

$$\eta_y = \sqrt{\frac{\sum(y_i - M_y)^2 - (y_i - \bar{y}_{\text{гр}})^2}{\sum(y_i - M_y)^2}}. \qquad (5.11)$$

5.3. Частная (парциальная) корреляция

В практических целях часто приходится выявлять взаимодействие нескольких факторов. Производится комбинационная группировка собранного материала, которая требует большого числа наблюдений. Можно использовать специальные статистические методы. С помощью этих методов производится последовательная элиминация влияния одних факторов и выделение результатов влияния других факторов. К таким методам относится метод частной корреляции. Элиминация – это исключение неизвестного из системы уравнений.

В ходе вычисления коэффициентов частной корреляции для трех признаков последовательно элиминируется влияние одного из призна-

ков: x_3, x_2, x_1; последовательно выявляется взаимосвязь в чистом виде: x_1 и x_2, x_1 и x_3, x_3 и x_2.

Элиминирование влияния третьего признака (x_3) и выявление связи между x_1 и x_2 производится по формуле:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}. \quad (5.12)$$

Аналогично производится элиминирование влияния второго признака (x_2) и выявление связи между x_1 и x_3:

$$r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}. \quad (5.13)$$

Затем проводится элиминирование влияния первого признака (x_1) и выявление взаимосвязи x_2 и x_3:

$$r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{13}^2)(1-r_{12}^2)}}. \quad (5.14)$$

П р и м е р. Оценить взаимосвязь фактора длительности рабочего времени с компьютером, усталости (число ошибок в тексте) и производительности труда (количество набранных страниц текста) (табл. 5.3).

Рассчитав коэффициенты корреляции Пирсона $r_{12} = 0{,}4$; $r_{13} = -0{,}7$; $r_{23} = -0{,}4$, можно сделать выводы о влиянии длительности рабочего времени (r_{12}) на появление усталости, а также увеличения продолжительности работы (r_{13}) на снижение производительности труда. Между увеличением усталости и снижением производительности труда обнаружена обратная статистическая связь (r_{23}).

Таблица 5.3

Исходные данные для расчета коэффициентов частной корреляции

Время работы, часы	Число ошибок	Число страниц текста
4	5	4
1	6	6
3	6	2
3	6	6
5	6	4
2	7	3
1	8	1
5	8	3
6	9	1
6	9	1

Используя формулы коэффициентов частной корреляции, произведем их расчет:

$$r_{12,3} = \frac{0,4 - (-0,7)(-0,4)}{\sqrt{(1-(-0,7)^2)(1-(-0,4)^2)}} = 0,2,$$

$$r_{13,2} = \frac{-0,7 - 0,4(-0,4)}{\sqrt{(1-(0,4)^2)(1-(-0,4)^2)}} = -0,7,$$

$$r_{23,1} = \frac{-0,4 - 0,4(-0,7)}{\sqrt{(1-(-0,7)^2)(1-(-0,4)^2)}} = -0,1.$$

Таблица 5.4

Итоговые значения коэффициентов корреляции

r_{12}	0,4	$r_{12,3}$	0,2
r_{13}	–0,7	$r_{13,2}$	–0,7
r_{23}	–0,4	$r_{23,1}$	–0,1

Анализ коэффициентов в табл. 5.4 показывает, что при устранении фактора продолжительности труда, произошел сдвиг показателя $r_{23,1} = -0,1$ (связь между усталостью и производительностью труда исчезла). Снижение выработки продукции (набранный текст) к концу рабочего дня связано в первую очередь не с нарастанием усталости, а с какими-то другими причинами.

5.4. Понятие о множественной корреляции

Метод множественной корреляции применяется в случаях, когда необходимо установить совокупное влияние всего комплекса факторов на результативный признак. Величина коэффициента множественной корреляции изменяется от 0 до 1. Его можно вычислить с использованием коэффициентов частной линейной корреляции по формуле:

$$R_{1,23} = \sqrt{1-(1-r_{12}^2)(1-r_{13}^2)} = \sqrt{1-(1-0,4^2)(1-(-0,7)^2)} = 0,75.$$

По коэффициенту $R = 0,75$ определяется коэффициент детерминации R^2 $(R_D) = 0,75^2 = 0,56$. Он показывает, что доля совместного влияния второго и третьего признаков составляет 56 %.

5.5. Оценка различий коэффициентов корреляции

Решение задач по оценке различий между коэффициентами корреляции возникает иногда в случае, если обе выборки принадлежали к одной генеральной совокупности.

П р и м е р. Требуется оценить статистическую достоверность различий между коэффициентами $r_1 = 0{,}45$; $r_2 = 0{,}58$. Число наблюдений в первой и второй группах составили соответственно $N_1 = 74$ и $N_2 = 50$.

По таблице величин Z [$r = \varphi(Z)$] (прил. 8) значения коэффициентов корреляции переводятся в соответствующие им величины $Z_1 = 0{,}48$, $Z_2 = 0{,}66$.

Оценка производится по критерию Стьюдента:

$$t = |z_2 - z_1| / \sqrt{(N_1 + N_2) / [(N_1 - 3)(N_2 - 3)]}; \qquad (5.15)$$

$$t = |0{,}66 - 0{,}48| / \sqrt{(74 + 50) / [(74 - 3)(50 - 3)]} = 1{,}09.$$

Число степеней свободы равно $N_1 + N_2 - 4 = 74 + 50 - 4 = 120$. При уровне значимости $\alpha = 0{,}05$ критическая величина критерия Стьюдента составляет 1,98 (см. прил. 4), что больше вычисленного (1,09). Поэтому различия между r_1 и r_2 следует признать статистически недостоверными.

Следует иметь в виду при анализе коэффициентов корреляции: чем больше r, тем меньшие различия между ними становятся значимыми. Если для $r = 0{,}14$ и 0,24 (разница между ними в 0,1) может быть статистически не значимой, то для $r = 0{,}80$ и 0,90 (разница 0,1) может оказаться значимой.

5.6. Ранговая корреляция

В географических исследованиях иногда приходится обрабатывать быстро и с наименьшими затратами фактический материал, даже если получаются менее точные результаты. В некоторых случаях работают с качественной информацией или с громоздкими цифрами. В таких случаях для установления зависимости между признаками используется ранговая корреляция.

Процесс упорядочения вариант по какому-либо признаку (например, увеличение или уменьшение количества населения по районам) называют ранжированием. Каждому члену ранжированного ряда присваивается *ранг.* Для обозначения рангов, как правило, используются числа в пределах единиц и десятков, например: 1, 2, 3, ..., *n.* Первой варианте или группе вариант присваивается ранг 1, второй варианте или группе – 2 и т. д. Следует иметь в виду, что одни и те же варианты в зависимости от цели группировки могут иметь различные ранги. Величина ранга не позволяет нам судить о том, насколько близко друг к другу расположены на шкале измерения различные варианты совокупности или качественные признаки.

Ранговую корреляцию можно применять для всех упорядоченных признаков (например, экспертные оценки, баллы, бонитеты). Объем со-

пряженных выборок должен быть не менее пяти. Коэффициент ранговой корреляции характеризуется следующими свойствами.

1. Если ранжированные варианты выборочных совокупностей имеют один и тот же ранг независимо от цели ранжирования, то коэффициент корреляции должен быть равен +1, т. е. существует полная положительная функциональная зависимость:

N_1	1	2	3	4	5	6	7
N_2	1	2	3	4	5	6	7

2. Если ранги вариант в сравниваемых рядах выборочных совокупностей расположены в обратной последовательности, то коэффициент корреляции равен –1, т. е. будет иметь место полная обратная функциональная зависимость:

N_1	1	2	3	4	5	6	7
N_2	7	6	5	4	3	2	1

3. В других случаях коэффициент ранговой корреляции имеет значения между +1 и –1, что больше соответствует фактической связи между признаками.

Для расчета зависимости (x, y) существуют следующие коэффициенты ранговой корреляции: коэффициент неупорядоченности $r_н$ и коэффициент Спирмена $r_с$. Коэффициент ранговой корреляции Спирмена рассчитать легче, чем коэффициент неупорядоченности, поэтому в естественных науках предпочтение отдается $r_с$. Коэффициент Спирмена представляет собой следующее соотношение:

$$r_с = 1 - \frac{6\sum(x' - y')^2}{N_n^3 - N_n}, \text{ или } r_с = 1 - \frac{6\sum(d^2)}{N_n^3 - N_n}, \qquad (5.16)$$

где d – разность между сопряженными рангами; x' – величины рангов, заменяющие фактические варианты или качественные признаки по аргументу x; y' – величины рангов, заменяющие фактические варианты или качественные признаки по функции y; N_n – количество сопряженных пар.

Достоверность полученного рангового коэффициента можно установить аналогично достоверности коэффициента корреляции (прил. 9).

П р и м е р. Следует дать эстетическую оценку ландшафта для обоснования выбора зоны отдыха. Предложено сравнить пять видов ландшафта (аргумент x), имеющих свои преимущества с точки зрения чистоты и влажности воздуха, насыщенности полезными фитонцидами, характеризующихся разнообразием рельефа, растительности, наличием рек и водоемов.

Исходя из имеющихся показателей, расположим виды ландшафта с учетом возрастающей оздоровительной и эстетической их роли (табл. 5.5). Соответст-

венно этому, видам ландшафта присваиваются ранги по возрастающей величине. Для получения необходимых показателей при расчете рангового коэффициента корреляции составляем табл. 5.6. Вычисляем разность между парными рангами (x'–y'), которые возводим в квадрат и суммируем. Результаты используются для расчета рангового коэффициента корреляции по формуле (5.16):

$$r_c = 1 – [6 \cdot 1 : (125 – 5)] = 0{,}95.$$

Таблица 5.5

Оценка ландшафта для рекреационной цели

Вид ландшафта	Ранг x'	Самочувствие отдыхающих	Ранг y'
Плоский пониженный, со смешанным лесом на суглинистых почвах	1	удовлетворительное	1
Слегка волнистый, с ельником на суглинистых почвах	2	удовлетворительное	1
Всхолмленный, с сосново-лиственным лесом и водоемом на песчаных почвах	3	хорошее	3
Пересеченный, с сосновым лесом на песчаных и супесчаных почвах	3	хорошее	3
Слегка пересеченный, с сосново-можжевелловым лесом на песчаных и супесчаных почвах	4	отличное	4

Таблица 5.6

Расчет рангового коэффициента корреляции

x'	y'	x'–y'	$(x'–y')^2$
1	1	0	0
2	1	1	1
3	3	0	0
3	3	0	0
4	4	0	0
			$\sum 1$

Поскольку ранговый коэффициент корреляции $r_c = 0{,}95 > r_т = 0{,}80$ при $P = 0{,}90$ для $\nu = 4$ (прил. 9), можно сделать вывод, что влияние изучаемых типов ландшафта на самочувствие отдыхающих достоверно и положительно.

Глава 6

РЕГРЕССИОННЫЙ АНАЛИЗ

Логическим продолжением корреляционного анализа является регрессионный анализ, который развивает и углубляет представление о корреляционной связи. Если корреляционный анализ позволяет установить лишь форму и тесноту зависимости между случайными переменными, то регрессионный анализ математически описывает выявленную зависимость, т. е. дает возможность численно оценить одни параметры через другие. Составив и решив уравнения регрессии, можно произвести выравнивание эмпирических линий регрессии, т. е. моделировать наблюдаемую зависимость путем подбора функции, график которой представляет собой теоретическую линию регрессии. Если подобранная функция отражает сущность процесса или явления, то возможно прогнозирование значений признака за пределами сделанных наблюдений. Подобно корреляции, регрессия может быть *парной* (простой) и *множественной*, по форме связи – *линейной* и *нелинейной*, по зависимости – *односторонней* (изменяется лишь один признак под влиянием другого) и *двусторонней* (изменяются оба признака под воздействием друг друга).

Регрессия выражается несколькими способами: построением эмпирических линий, составлением уравнения и затем – построением теоретических линий регрессии, а также с помощью коэффициента регрессии. Уравнение наиболее точно выражает зависимость между двумя переменными (x, y), если корреляция между ними близка к единице.

Регрессионный анализ возможен при наличии всего лишь нескольких пар сопряженных наблюдений, но при условии сильных связей между признаками ($r \geq 0{,}7$). Для вывода уравнения линейной регрессии достаточно двух пар наблюдений. Обычно рядом с уравнением регрессии приводится коэффициент корреляции или корреляционного отношения, например: $y = 0{,}1106x + 0{,}298$, $r_{0,95} = 0{,}75$ (это обусловлено практическим использованием уравнения регрессии). Из приведенных равенств вытекает, что влияние аргумента (x) на функцию (y) достаточно сильное. Поэтому, имея в своем распоряжении данные по аргументу, можно по формуле уравнения регрессии вычислить значение функции, не прибегая к полевым наблюдениям.

Точки эмпирических линий регрессии ($\bar{x}_{гр}, \bar{y}_{гр}$) определяются как взвешенные средние арифметические, для невзвешенных рядов – как средние малых групп выборки. Вычислив координаты точек, наносим их на график и соединяем прямой; в результате получаются эмпирические линии регрессии (рис. 6.1). По графическому изображению можно предварительно сделать заключение о характере связи. При полном отсутствии связи эмпирические линии располагаются параллельно осям графика. При полной связи между x, y (r =1) линии регрессии на графике, построенные по точкам эмпирических линий регрессии, совместятся.

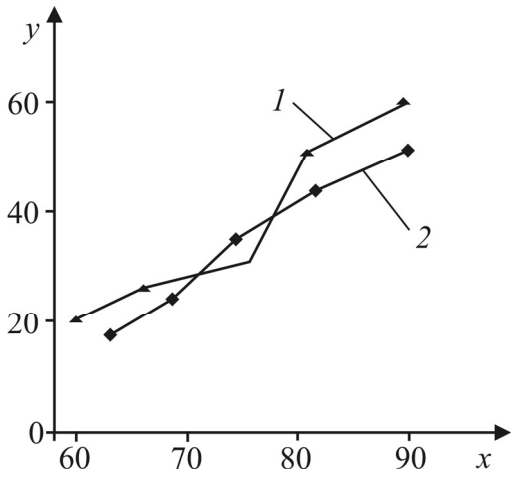


Рис. 6.1. Эмпирические линии регрессии по $\bar{x}_{гр}, \bar{y}_{гр}$

Существует два способа составления уравнений регрессии: а) способ координат точек, с использованием двух-трех точек, расположенных на эмпирической линии (желательно в начале, середине и конце ее),– для тех случаев, когда расчет не требует большой точности; б) способ наименьших квадратов, более точный, так как для составления уравнения регрессии привлекаются все сопряженные наблюдения. Рассмотрим наиболее простые способы составления уравнений регрессии.

6.1. Линейная зависимость

Линейная регрессия на графике изображается в виде прямой так, чтобы точки эмпирической линии располагались по обе стороны ее и по возможности ближе к ней.

Известно следующее уравнение линейной регрессии:

$$y = ax + b, \qquad (6.1)$$

где y – значение зависимой переменной (признак); x – значение независимой переменной (фактор, влияющий на признак); a – коэффициент

регрессии, показывающий степень зависимости между переменными (может быть также выражен тангенсом угла наклона линии регрессии к оси абсцисс); b – ордината линии, показывающая смещение начала прямой относительно начала координат.

Определим двумя способами неизвестные параметры a и b. Используем для этого пример нахождения линейной корреляции (см. п. 5.1).

П р и м е р. Следует установить, как влияет гидролитическая кислотность (x_i, мэкв. на 100 г почвы) на содержание подвижного марганца (y_i, мг/кг почвы). В результате аналитических работ получены следующие данные:

x_i	69	70	72	75	83	90	91	95	95
y_i	18	48	42	31	56	84	68	90	107

Для решения поставленной задачи используем *способ координат точек.* Результаты наблюдений наносим на график, затем проводим прямую так, чтобы число точек по обе стороны линии было одинаковым (рис. 6.2). Для расчета параметров a и b выбираем две точки, которые находятся на прямой или рядом с ней (одну в начале и одну в конце). Используем координаты точек 1-й и 8-й: $x_1 = 69$, $y_1 = 18$; $x_8 = 95$, $y_8 = 90$. Подставляя значения переменных в общее уравнение прямой, получаем систему уравнений:

$$\begin{cases} 18 = 69a + b; \\ 90 = 95a + b. \end{cases}$$

Решаем эту систему относительно a и b: $b = 18 - 69a$; $90 = 95a + (18 - -69a)$; $72 = 26a$; $a = 2{,}76$ (или $\text{tg} = 70^{\circ}06'$); $b = 18 - 69 \cdot 2{,}76 = -173{,}07$. Получив количественное значение параметров a и b, связь между x и y можно выразить конкретным уравнением регрессии:

$$y = 2{,}76x - 173{,}07, \quad r_{0,99} = 0{,}87.$$

Это уравнение можно использовать для расчета содержания марганца, если имеются данные по гидролитической кислотности (с учетом заданных условий).

Приведенное выше уравнение регрессии можно получить также способом наименьших квадратов, используя координаты всех точек. Этот способ заключается в построении такой линии на графике, чтобы сумма квадратов отклонений от нее до точек эмпирической линии регрессии была наименьшей. Для определения параметров a и b составляется система уравнений:

$$\begin{cases} \sum y = a\sum x + bn; \\ \sum xy = a\sum x^2 + b\sum x. \end{cases} \qquad (6.2)$$

Систему уравнений выводим следующим образом. Подставляем в общее уравнение прямой (6.1) все имеющиеся значения по гидролитиче-

ской кислотности (x) и содержанию подвижного марганца (y), суммируем правые и левые части и получаем первое уравнение:

$$\begin{aligned} y_1 &= ax_1 + b; \\ y_2 &= ax_2 + b; \\ &\cdots\cdots\cdots\cdots \\ \frac{y_n &= ax_n + b}{\sum y = a\sum x + bn}. \end{aligned} \qquad (6.3)$$

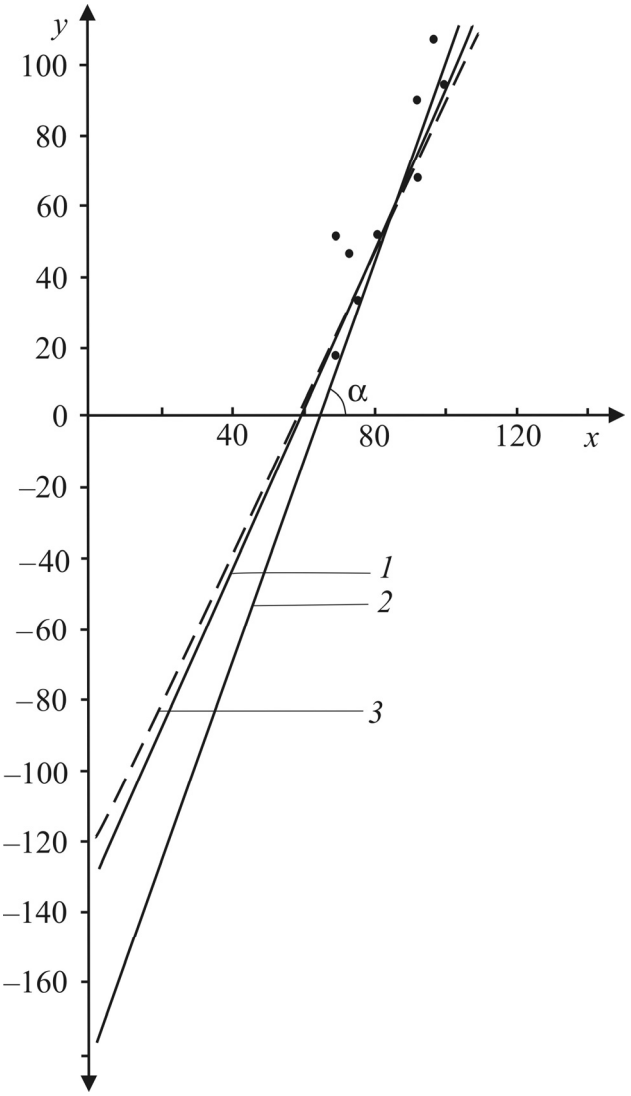


Рис. 6.2. Сравнение местоположения эмпирических линий *1*, *2* с теоретической *3* по зависимости содержания подвижного марганца y от гидролитической кислотности x ($\angle\alpha = 70^{o}06' = \mathrm{tg}_x\ 2{,}76$):
для эмпирических линий (*1*) $y = 2{,}30x - 130{,}9$;
для (*2*) $y = 2{,}76x - 173{,}0$; $r_{0{,}99} = 0{,}87$

Затем каждое исходное уравнение из (6.3) умножаем на соответствующее значение x; просуммировав правые и левые части, получим второе уравнение:

$$x_1 y_1 = a x_1^2 + b x_1;$$
$$x_2 y_2 = a x_2^2 + b x_2;$$
$$\ldots\ldots\ldots\ldots$$
$$\frac{x_n y_n = a x_n^2 + b x_n}{\sum xy = a\sum x^2 + b\sum x}.$$

Для расчета параметров a и b составляем табл. 6.1. Полученные данные подставляем в систему уравнений (6.2):

$$\begin{cases} 600 = 830a + 10b; \\ 52102 = 69890a + 830b. \end{cases}$$

Решая систему, находим искомые параметры: a = 2,30 (tg = 66°30'); b = –130,9. Подставив полученные показатели в искомое уравнение регрессии, находим $y = 2{,}30x - 130{,}9$, $r_{0,99} = 0{,}87$.

Хотя значения параметров a и b, рассчитанные двумя способами, близки между собой, второй способ (наименьших квадратов) более точно определяет положение линии регрессии.

Таблица 6.1

Расчет данных для уравнения линейной зависимости

x	y	xy	x^2	$y'=ax+b$	Расчет критерия χ^2		
					$y-y'$	$(y-y')^2$	$\frac{(y-y')^2}{y'}$
69	18	1242	4761	27,8	–9,8	96,04	3,45
70	48	3360	4900	30,1	17,9	320,41	10,64
72	42	3024	5184	34,7	7,3	53,29	1,54
75	31	2325	5625	41,6	–10,6	112,36	2,70
83	56	4648	6889	60,0	–4,0	16,00	0,27
90	84	7560	8100	76,1	7,9	62,41	0,82
90	56	5040	8100	76,1	–20,1	404,01	5,31
91	68	6188	8281	78,4	–10,4	108,16	1,38
95	90	8550	9025	87,6	2,4	5,76	0,07
95	107	10165	9025	87,6	19,4	376,36	4,30
$\sum = 830$	600	52102	69860				$\chi^2 = 30{,}47$

Кроме того, коэффициенты a и b для уравнения регрессии также могут быть рассчитаны на основе исходных данных (x, y) по формулам, которые обеспечивают наименьший квадрат отклонений этих точек от линии регрессии (метод наименьших квадратов):

$$b = \frac{\sum x^2 \sum y^2 - \sum x \sum xy}{n_{\text{пар}} \sum x^2 - \left(\sum x\right)^2}, \qquad a = \frac{n_{\text{пар}} \sum xy - \sum x \sum y}{\sum x^2 - \left(\sum x\right)^2}.$$

После составления уравнения регрессии и определения параметров a и b производим расчет точек y' теоретической линии регрессии. Для этого в уравнение регрессии поочередно подставляем значения x. Степень совпадения теоретической и эмпирической линий регрессии можно проверить, используя критерий хи-квадрат. Цифровые показатели для $(y - y')^2/y'$ (см. табл. 6.1) суммируем и получаем $\chi^2 = 30{,}47$. Поскольку $\chi_{\text{ф}}^2 = 30{,}47 > \chi_{\text{т}}^2 = 21{,}66$ при $P = 0{,}99$ для $\nu = 9$, то можно указать на недостаточное соответствие теоретической линии регрессии эмпирическому ряду. Составленные уравнения регрессии можно проверить на точность зависимости между переменными (x, y) не только по критерию хи-квадрат, но и по коэффициенту точности выравнивания линии r_1, отражающему степень приближения (соответствия) фактических данных наблюдения к вероятным. Этот коэффициент определяем следующим образом:

$$r_1 = \sqrt{\frac{\sum \alpha^2 - \sum \beta^2}{\sum \alpha^2}} = \sqrt{\frac{\sum (y_{\text{ф}} - M_{\text{ф}})^2 - \sum (y_{\text{ф}} - y_{\text{в}})^2}{\sum (y_{\text{ф}} - M_{\text{ф}})^2}}, \qquad (6.4)$$

где $(y_{\text{ф}} - M_{\text{ф}}) = \alpha$ – отклонение индивидуальных вариант от общего среднего арифметического по y; $(y_{\text{ф}} - y_{\text{в}}) = \beta$ – отклонение индивидуальных экспериментальных вариант по y от расчетных по уравнению.

На основании исходных данных, полученных в табл. 6.2, используя формулу (6.4), имеем:

$$r_1 = \sqrt{(6806 - 1554{,}8) : 6806} = 0{,}88.$$

Принято считать: если $r_1 > 0{,}95$, то уравнение регрессии соответствует более точному положению линии на графике. При $r_1 < 0{,}95$ необходимо найти другую математическую зависимость. В приведенном примере $r_1 = 0{,}88 < 0{,}95$, поэтому следует подобрать другую математическую зависимость. Такие же выводы получены при проверке на точность зависимости между переменными по критерию хи-квадрат. Оба критерия оценки (χ^2, r_1) на точность выравнивания линии уравнения регрессии используются и для других форм регрессионной зависимости.

Таблица 6.2

Расчет данных для определения точности выравнивания линии

y		α-отклонения		β-отклонения	
$y_ф$	$y_в$	$y_ф - M_ф$	$(y_ф - M_ф)^2$	$y_ф - y_в$	$(y_ф - y_в)^2$
18	27,8	–42	1764	–9,8	96,04
48	30,1	–12	144	17,9	320,41
42	34,7	–18	324	7,3	53,29
31	41,6	29	841	–10,6	112,36
56	60,0	–4	16	–4,0	16,00
84	76,1	24	576	7,9	62,41
56	76,1	–4	16	–20,1	404,01
68	78,4	4	16	–10,4	108,16
90	87,6	30	900	2,4	5,76
107	87,6	47	2209	19,4	376,36
$M_ф$=60			$\sum = 6806$		$\sum = 1554,80$

Ошибку уравнения регрессии можно определить по формуле

$$m = \sqrt{\frac{\sum (y - y')^2}{n - k}} = \sqrt{\frac{\sum (y_ф - y_в)^2}{n - k}},$$

где n – число точек линии регрессии (см. рис. 6.2); k – число коэффициентов в уравнении регрессии (два плюс свободный член уравнения).

6.2. Гиперболическая зависимость

При проведении исследований может быть установлена нелинейная зависимость между аргументом и функцией, представляющая собой на графике кривую в виде гиперболы. Общее уравнение регрессии для гиперболической зависимости имеет вид

$$y = a/x + b, \qquad (6.5)$$

где x – аргумент; y – функция; a и b – коэффициенты, величину которых следует установить.

Расчет сводится к следующему. Чтобы установить вид зависимости между функцией и аргументом, по исходным данным строится график. Затем при вычислении параметров a и b по способу координат точек подбираются две точки, расположенные на кривой или около нее по методу, описанному для линейной регрессии (см. п. 6.1). Для этих

же параметров по способу наименьших квадратов используется система уравнений

$$\begin{cases} \sum xy = an + b\sum x; \\ \sum x^2 y = a\sum x + b\sum x^2. \end{cases} \qquad (6.6)$$

Эта система получена в результате умножения на x и x^2 исходных уравнений по x и y:

$$x_1 y_1 = a + bx_1; \qquad x_1^2 y_1 = ax_1 + bx_1^2;$$
$$x_2 y_2 = a + bx_2; \qquad x_2^2 y_2 = ax_2 + bx_2^2;$$
$$\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots\ldots$$
$$\frac{x_n y_n = a + b\sum x_n}{\sum xy = an + b\sum x}. \qquad \frac{x_n^2 y_n = ax_n + b\sum x_n^2}{\sum x^2 y = a\sum x + b\sum x^2}.$$

П р и м е р. Установим зависимость между температурой воздуха в июле *(x*, °С) и относительной влажностью воздуха (*y*, %) по следующим исходным данным:

x_i	14,7	14,9	15,3	15,6	16,0	16,7
y_i	80	78	76	75	74	73,7

При построении графика видно, что зависимость между функцией и аргументом гиперболическая, поэтому используем общее уравнение гиперболы. Для расчета параметров a и b по способу координат точек используем данные первой и шестой пары наблюдений: $x_1 = 14{,}7$, $y_1 = 80$; $x_в = 16{,}7$, $y_в = 73{,}7$. Подставляем эти данные в общее уравнение (6.5), предварительно преобразовав его: $xy = a + bx$. Получим систему уравнений

$$\begin{cases} 1176 = a + 14{,}7b; \\ 1230{,}8 = a + 16{,}7b. \end{cases}$$

Таблица 6.3

Расчет данных для уравнения нелинейной зависимости

x	y	xy	x^2	x^2y
14,7	80,0	1176,0	216,09	17287,2
14,9	78,0	1162,2	222,01	17316,78
15,3	76,0	1162,8	234,09	17790,84
15,6	75,0	1170,0	243,36	18252,00
16,0	74,0	1184,0	256,00	18994,00
16,7	73,7	1230,79	278,89	20554,19
$\sum = 93{,}2$	456,7	7085,79	1450,44	110195

Решаем систему относительно a и b: $a = 773{,}22$; $b = 27{,}4$. В результате конкретное уравнение регрессии для гиперболической зависимости по способу координат точек будет иметь вид $y = 773{,}22/x + 27{,}4$; $\eta_{0{,}99} = 0{,}84$.

Для установления параметров a и b по способу наименьших квадратов по уравнению (6.6) предварительно проводим соответствующие вычисления (табл. 6.3). Полученные данные подставляем в уравнение (6.6):

$$\begin{cases} 7086 = 6a + 93{,}2b; \\ 22615{,}9 = 93{,}2a + 1450{,}44b. \end{cases}$$

Делим первое уравнение на b, второе уравнение – на 93,2 и освобождаемся от коэффициентов при неизвестном a. Затем вычитаем второе уравнение из первого и определяем b. Подставив значение b в одно из уравнений, вычисляем a. Искомое уравнение регрессии примет вид $y = 484597{,}4/x - 31280$; $\eta_{0{,}95} = 0{,}84$.

Коэффициент точности выравнивания линии r_1 по формуле (6.4) рассчитываем таким же образом, как в п. 6.1.

6.3. Параболическая зависимость

Общее уравнение параболы n-го порядка имеет вид

$$y = ax^n + bx^{n-1} + cx^{n-2} + \ldots + kx + l.$$

Если ограничиться второй ступенью независимой переменной величины x, будем иметь частный случай параболы второго порядка:

$$y = ax^2 + bx + c. \qquad (6.7)$$

П р и м е р. Для решения конкретной задачи используем данные примера из п. 5.2, где было рассчитано прямое корреляционное отношение ($\eta = 0{,}78$) и доказана его достоверность. На графике зависимость между температурой воздуха (x) и упругостью водяного пара (y) имеет вид параболы.

Для расчета коэффициентов a, b, c способом координат точек используем координаты 2-й, 4-й и 6-й точек:

$$x_2 = 14{,}9; \quad x_4 = 15{,}6; \quad x_6 = 16{,}7;$$
$$y_2 = 13{,}7; \quad y_4 = 14{,}5; \quad y_6 = 14{,}6.$$

Подставляя значения координат точек в общее уравнение параболы второго порядка (6.7), получаем систему уравнений, которую решаем относительно a, b, c:

$$\begin{cases} 13{,}7 = 222{,}01a + 14{,}9b + c; \\ 14{,}5 = 243{,}36a + 15{,}6b + c; \\ 14{,}6 = 278{,}89a + 16{,}7b + c. \end{cases}$$

В результате $a = 0{,}066$; $b = -0{,}19$; $c = 1{,}94$. С помощью уравнения параболы (6.7) имеем следующую зависимость между переменными: $y = 0{,}066x^2 - 0{,}19x + 1{,}94$, $\eta_{0{,}95} = 0{,}78$.

Для вычисления коэффициентов a, b, c по способу наименьших квадратов используется общее уравнение параболы второго порядка. Подставив в формулу (6.7) все имеющиеся данные и просуммировав правые и левые части уравнений, получаем первое уравнение системы:

$$y_1 = ax_1^2 + bx_1 + c;$$

$$y_2 = ax_2^2 + bx_2 + c;$$

$$\dots\dots\dots\dots\dots$$

$$\frac{y_n = ax_n^2 + bx_n + c}{\sum y = a\sum x^2 + b\sum x + cn}.$$

Второе и третье уравнения системы определяем путем умножения соответственно на x и x^2 исходного общего уравнения параболы второго порядка. В результате имеем систему трех уравнений:

$$\begin{cases} \sum y = a\sum x^2 + b\sum x + cn; \\ \sum xy = a\sum x^3 + b\sum x^2 + c\sum x; \\ \sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2. \end{cases} \qquad (6.8)$$

Конкретные данные для уравнения (6.8) рассчитаны по табл. 6.4.

П р и м е р. Решаем систему (6.8) относительно a, b, c:

$$84{,}8 = 1450{,}44a + 93{,}2b + 6c;$$

$$1319{,}18 = 22615{,}93a + 1450{,}44b + 93{,}2c;$$

$$20560{,}10 = 353321{,}18a + 22615{,}93b + 1450{,}44c.$$

Таблица 6.4

Расчет данных для уравнения параболической зависимости

x	y	xy	x^2	x^3	x^2y	x^4
14,7	13,1	192,57	216,09	3176,52	2830,78	46694,89
14,9	13,7	204,13	222,01	3307,95	3041,54	49288,44
15,3	14,2	217,26	234,09	3581,58	3324,08	54798,13
15,6	14,5	226,20	243,36	3796,42	3528,72	59224,09
16,0	14,7	235,20	256,00	4096,00	3763,20	65536,00
16,7	14,6	243,82	278,89	4657,46	4071,79	77779,63
$\sum = 93,2$	84,8	1319,18	1450,44	22615,93	20560,10	353321,18

Имеем параметры a, b, c: $a = -0{,}014$; $b = 1{,}13$; $c = -0{,}93$. Таким образом, уравнение параболы 2-го порядка, полученное по способу наименьших квадратов, примет следующий вид: $y = -0{,}014x^2 + 1{,}13x - 0{,}093$. Сравним уравне-

ния параболы, полученные двумя способами, подставив в эти уравнения одно из значений x (14,7 °С):

$$y = 0{,}066x^2 - 0{,}19x + 1{,}94 = 14{,}26 - 2{,}79 + 1{,}94 = 13{,}41$$

по способу координат точек;

$$y = -0{,}014x^2 + 1{,}13x - 0{,}093 = 3{,}02 + 16{,}61 - 0{,}093 = 13{,}49$$

по способу наименьших квадратов.

6.4. Множественная регрессия

Если при установлении зависимости между признаками используется больше одной независимой переменной, то применяют *множественный регрессионный анализ.* Проведение такого анализа возможно в следующих условиях: распределение зависимой переменной при различных значениях независимых должно быть близко к нормальному; дисперсия зависимой переменной при разных значениях признаков x должна считаться одинаковой. С увеличением числа признаков и в случаях нелинейной множественной регрессии необходимо использовать ЭВМ. Поэтому рассмотрим простой вариант множественной линейной регрессии без применения ЭВМ, когда один признак зависит от двух факторов. Общее уравнение линейной множественной регрессии имеет вид

$$y = a + bx + cz. \tag{6.9}$$

Для вычисления параметров a, b, c составляется следующая система уравнений:

$$\begin{cases} \sum y = an + b\sum x + c\sum z; \\ \sum xy = a\sum x + b\sum x^2 + c\sum xz; \\ \sum yz = a\sum z + b\sum xz + c\sum z^2. \end{cases} \tag{6.10}$$

Соответствие между теоретическими (y') и эмпирическими (y) значениями признака устанавливают с помощью критериев хи-квадрат или Стьюдента (см. п. 1.6). При необходимости ошибка уравнения линейной множественной регрессии определяется по формуле:

$$m = \sqrt{\frac{\sum a_y^2 - \left(b\sum a_y a_x + c\sum a_y a_z\right)}{n-k}}, \tag{6.11}$$

где a, b, c – значения параметров уравнения множественной регрессии; n – число сопряженных значений вариант; k – число коэффициентов уравнения регрессии *(*a, b, c плюс свободный член).

Другие параметры для (6.11) вычисляются по формулам:

$$\sum a_y^2 = \sum y^2 - nM_y^2; \tag{6.12}$$

$$\sum a_y a_x = \sum xy - nM_y M_x; \tag{6.13}$$

$$\sum a_y a_z = \sum yz - nM_y M_z. \tag{6.14}$$

П р и м е р. При изучении зависимости между биомассой трав (y, г/м^2) в агроландшафте, с одной стороны, температурой (x, °С) и количеством атмосферных осадков (z, мм) – с другой, установлена прямая односторонняя зависимость y от x и z. С практической точки зрения целесообразно составить уравнение множественной регрессии, которое можно было бы использовать для прогноза биомассы по температуре и количеству выпавших осадков. Данные по x, y, z представляют собой средние многолетние показатели за период вегетации (май, июнь):

y	300	350	370	420	450	500
x	14,5	15,0	15,6	17,2	18,5	19,3
z	82	95	105	120	130	140

Вычисленные в табл. 6.5 показатели подставляем в систему уравнений (6.10):

$$\begin{cases} 2390 = 6a + 100{,}09b + 672c; \\ 40571 = 100{,}09a + 1689{,}19b + 11423c; \\ 275600 = 672a + 11423b + 77674c. \end{cases}$$

Таблица 6.5

Расчет данных для уравнения линейной множественной регрессии

y	x	z	y^2	x^2	z^2	xy	xz	yz
300	14,5	82	90000	210,25	6724	4350	1189	24600
350	15,0	95	122500	225,00	9025	5250	1425	33250
370	15,6	105	136900	243,36	11025	5772	1638	38850
420	17,2	120	176400	295,84	14400	7224	2064	50400
450	18,5	130	202500	342,25	16900	8325	2405	58500
500	19,3	140	250000	372,49	19600	9650	2702	70000
$\sum = 2390$ $M_y = 398{,}33$	100,1 $M_x = 16{,}68$	672 $M_z = 112$	978300	1689,19	77674	40571	11423	275600

Решаем систему уравнений относительно a, b, c. Получаем $a = -3{,}26$; $b = 5{,}01$; $c = 2{,}84$. Подставляем значения a, b, c в общую формулу уравнения множественной регрессии (6.9):

$$y = -3{,}26 + 5{,}01x + 2{,}84z. \tag{6.15}$$

Затем находим теоретические значения y'. Для этого подставляем в формулу (6.15) экспериментальные данные по x и z и заносим в табл. 6.6 для расчета критерия хи-квадрат. Поскольку $\chi^2_{\text{ф}} = 0{,}602 < \chi^2_{\text{т}} = 11{,}1$ при $P = 0{,}95$ для $\nu = 5$, то можно сделать вывод, что расчет биомассы по данным температуры (x) и осадкам (z) достаточно точный.

Таблица 6.6

Расчет данных для критерия хи-квадрат

y	y'	$y - y'$	$(y - y')^2$	$\frac{(y - y')^2}{y'}$
300	302,2	–2,26	5,11	0,017
350	341,7	8,31	69,06	0,202
370	373,1	–3,09	9,55	0,026
420	423,7	–3,71	13,76	0,032
450	458,6	–8,62	74,30	0,162
500	491	8,97	80,46	0,164
		$\sum = -0{,}4$	252,241	$\chi^2 = 0{,}603$

Для определения ошибки уравнения линейной множественной регрессии показатели рассчитываем по формулам (6.12 – 6.14):

$$\sum a_y^2 = 978300 - 6 \cdot 398{,}33^2 = 29299{,}4;$$

$$\sum a_y a_x = 40571 - 6 \cdot 398{,}33 \cdot 16{,}68 = 706{,}14;$$

$$\sum a_y a_z = 275600 - 6 \cdot 398{,}33 = 7922{,}3.$$

Затем подставляем полученные значения в формулу (6.11):

$$m = \sqrt{\frac{26299{,}4 - (5{,}01 \cdot 706{,}14 + 2{,}84 \cdot 7922{,}3)}{6 - 3}} = 9{,}35 \text{ г/м}^2.$$

Таким образом, прогнозируя урожай биомассы трав за период вегетации по температуре и осадкам, мы рискуем ошибиться в среднем на 9,35 г/м^2, т. е. на 2,3 %.

Уравнения регрессии широко используются в научных исследованиях и в практических целях.

Глава 7

ФАКТОРНЫЙ АНАЛИЗ

7.1. Сущность и возможности применения

При изучении взаимного влияния многих процессов и явлений в последнее время все чаще обращаются к методам многомерного статистического анализа, в частности, факторного анализа. Методы многомерного статистического анализа требуют применения сложной вычислительной техники.

Факторный анализ основывается на использовании статистических знаний (вычислении стандартных отклонений, знании корреляционного и регрессионного анализов). В большинстве случаев исследуется система корреляций, отраженных в корреляционной матрице. Факторный анализ представляет собой ветвь математической статистики, цель которого – разработка моделей, понятий и методов, позволяющих анализировать и интерпретировать массивы экспериментальных данных независимо от их физической природы. Анализ данных включает краткое описание распределения объектов, установление взаимоотношения процессов и явлений, отражающихся в виде *параметров.*

Используемый набор моделей и методов предназначен для «сжатия» информации, содержащейся в корреляционной матрице. В основе различных моделей факторного анализа лежит следующая гипотеза: параметры – это косвенные характеристики объекта или явления и представляют в совокупности тот или иной фактор. В связи с этим задача факторного анализа состоит в том, чтобы показать наблюдаемые параметры в виде линейных комбинаций факторов. Изменение фактора не всегда одинаково отражается на параметрах, поэтому среди последних могут быть выделены группы, реагирующие на каждый из факторов порознь. Параметры, входящие в одну и ту же группу, сильно коррелируют между собой; параметры, входящие в разные группы, слабо коррелируют между собой. Задача выявления факторов понимается как разбиение параметров на группы таким образом, чтобы можно было описать взаимоотношения между параметрами.

Разработано несколько вариантов факторного анализа с использованием коэффициентов только линейной корреляции (нелинейная корреляция вызывает затруднения при обработке материала). Наиболее употребительны при этом *метод главных компонент, метод главных факторов* и *центроидный метод*. Определение главных компонент и главных факторов производится с помощью ЭВМ.

Наиболее типичной формой представления данных является *матрица*. Это прямоугольная (или квадратная) таблица чисел, вертикальный ряд которой (столбец) обозначается индексом *j*, горизонтальный (строка) – индексом *i*. Любой элемент матрицы обозначается символом *a* с индексами, первый указывает номер строки, второй – номер столбца, которым соответствует данный элемент (в общем виде a_{ij}). Матрица обозначается прописной буквой (*A*, *B* и т. д.). О матрице, имеющей *m* строк и *n* столбцов, говорят, что ее порядок составляет $m \cdot n$. Квадратная матрица $n \cdot n$ имеет порядок *n*. В общем виде матрица записывается следующим образом:

$$A = \begin{vmatrix} a_{11}a_{12}a_{13} \ldots a_{1n} \\ a_{21}a_{22}a_{23} \ldots a_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ a_{m1}a_{m2}a_{m3} \ldots a_{mn} \end{vmatrix}$$

В факторном анализе с использованием правил матричной алгебры часто встречается операция умножения матриц. Для того чтобы умножить матрицу *A* на матрицу *B*, необходимо следующее условие: матрица *A* должна иметь столько столбцов, сколько строк в матрице *B*. Сам процесс умножения протекает по правилу «строка на столбец». Это правило означает, что каждый элемент матрицы произведения представляет собой сумму произведений элементов строки первой матрицы на соответствующие элементы столбца второй матрицы, например:

$$A = \begin{vmatrix} a_{11}a_{12} \\ a_{21}a_{22} \\ a_{31}a_{32} \end{vmatrix} \times B = \begin{vmatrix} b_{11}b_{12}b_{13} \\ b_{21}b_{22}b_{23} \end{vmatrix} = C = \begin{vmatrix} (a_{11}b_{11} + a_{12}b_{21})(a_{11}b_{12} + a_{12}b_{22})(a_{11}b_{13} + a_{12}b_{23}) \\ (a_{21}b_{11} + a_{22}b_{21})(a_{21}b_{12} + a_{22}b_{22})(a_{21}b_{13} + a_{22}b_{23}) \\ (a_{31}b_{11} + a_{32}b_{21})(a_{31}b_{12} + a_{32}b_{22})(a_{31}b_{13} + a_{32}b_{23}) \end{vmatrix}$$

Матрица-произведение будет иметь всегда столько строк, сколько их было в первой матрице, и столько столбцов, сколько их было во второй матрице: $(p \cdot q) \cdot (q \cdot r) = (p \cdot r)$.

Существует ряд математических методов, которые по информации, заложенной в матрице, позволяют провести классификацию объектов. Такие методы объединены в многомерный анализ, а при наличии одной строки в матрице – в одномерный анализ.

В факторном анализе используются следующие виды матриц: *диагональная* (в ней отличны от нуля только элементы, лежащие на главной диагонали), *скалярная* (все элементы диагональной матрицы равны между собой), *единичная* (все элементы главной диагонали равны единице), *обратная* (аналогична обратному числу в арифметике).

Элементами исходной матрицы в факторном анализе являются коэффициенты корреляции. В ходе анализа вычисляется также общая дисперсия σ^2, указывающая, в каких границах находятся значения параметров, которые характеризуют фактор. Кроме общей дисперсии, в анализе учитывается факторная дисперсия (общность) и специфическая дисперсия, связанная с некоторой переменной и характеризующая только ее. Дисперсию, обусловленную ошибкой, стремятся свести к минимуму.

В итоге составляется факторная матрица. Элементы столбцов матрицы представляют собой *факторные нагрузки*, или *коэффициенты факторного отображения*, выраженные коэффициентами корреляции данной переменной с данным фактором. Таким образом, коэффициенты факторного отображения характеризуют фактор и его влияние на все параметры.

Результат факторного анализа можно также выразить в виде графика, который наглядно иллюстрирует полученные выводы. Каждую из двух связанных друг с другом переменных можно изобразить как вектор, т. е. отрезок прямой, имеющий определенную длину и направление. Величина корреляции между переменными равна произведению абсолютных величин обоих векторов на косинус угла между ними: $r_{1,2} = h_1 h_2 \cos \alpha_{1,2}$, где $r_{1,2}$ – коэффициент корреляции; h_1 – длина вектора, соответствующая переменной 1; h_2 – длина вектора, соответствующая переменной 2; $\cos \alpha_{1,2}$ – угол между векторами h_1 и h_2.

7.2. Последовательность операций

На конкретном примере рассмотрим один из методов факторного анализа. На основе выборки по 395 ландшафтам в пределах водораздельного пространства была получена исходная информация о восьми параметрах агроландшафта. Они включают: 1) органические удобрения; 2) минеральные удобрения; 3) известь; 4) пестициды; 5) содержание гумуса в пахотном горизонте; 6) реакцию среды; 7) влажность почвы; 8) содержание физической глины. Следует определить, какова роль этих параметров в эволюции агроландшафтов.

Первый этап. Производится вычисление коэффициентов корреляции между всеми изучаемыми параметрами (табл. 7.1). Корреляционная мат-

Таблица 7.1

Корреляционная матрица *R* для восьми параметров агроландшафта

Параметры	1	2	3	4	5	6	7	8
1. Органические удобрения	1							
2. Минеральные удобрения	0,846	1						
3. Известь	0,805	0,881	1					
4. Пестициды	0,859	0,826	0,801	1				
5. Гумус	0,473	0,376	0,380	0,436	1			
6. Реакция почвы	0,398	0,326	0,319	0,329	0,762	1		
7. Влажность почвы	0,301	0,277	0,237	0,327	0,730	0,583	1	
8. Физическая глина	0,382	0,415	0,345	0,365	0,629	0,577	0,539	1

Примечание. В столбцах приведены параметры, аналогичные указанным в строках

рица *R* симметрична, поэтому достаточно заполнить лишь ее половину до линии диагонали. Если параметр коррелирует сам с собой, коэффициент корреляции равен единице.

Второй этап. Для описания параметров используется линейная модель (параметры выражаются через скрытые гипотетические факторы линейно). Основная модель факторного анализа может быть записана в виде формулы:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jm}F_m + d_j u_{ji},$$

где z_j – параметр, F_1 – фактор; a_{ji} – приближение (коэффициент) факторного отображения (нагрузки). Первый член правой части равенства показывает долю первого фактора в исследуемых явлениях, второй – долю второго фактора, последний – долю независимого фактора (остаток). Чем больше величина коэффициента факторного отображения при факторе, тем больше роль данного фактора в рассматриваемом явлении.

Для выражения общей дисперсии определяется факторная дисперсия, или значение общности (σ_i^2) для каждого диагонального параметра. Наиболее простой способ ее установления – вычисление первого центроидного фактора (табл. 7.2). На главную диагональ корреляционной матрицы помещают максимальные значения коэффициентов корреляции каждого столбца. Отношение квадрата суммы элементов каждого столбца к сумме всех элементов матрицы составит факторную дисперсию для столбца *j*:

$$\sigma_j^2 = \frac{\left(\sum_{i=1}^{n} r_j\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}}, \qquad (7.1)$$

где $\sum r_i$ – суммарный коэффициент корреляции по столбцу; $\sum\sum r_{ij}$ – сумма восьми суммарных коэффициентов корреляции.

Подставив данные в формулу (7.1), имеем первую факторную дисперсию: $\sigma_i^2 = 4{,}923^2 / 35{,}411 = 0{,}684$. Аналогично проводим расчет дисперсии по остальным столбцам табл. 7.2. Полученные данные помещаем по главной диагонали редуцированной корреляционной матрицы R^x (табл. 7.3). Если рассчитанные коэффициенты корреляции мало отличаются от исходных, значит, модель хорошо описывает экспериментальные данные. Однако максимальный коэффициент $r_1 = 0{,}859$ (см. табл. 7.2) отличается от рассчитанного $r_1 = 0{,}684$ (см. табл. 7.3).

Таблица 7.2

Корреляционная матрица *R* с приближенными значениями общностей

Номер параметра	1	2	3	4	5	6	7	8
1	**0,859**	0,846	0,805	0,859	0,473	0,398	0,301	0,382
2	0,846	**0,881**	0,881	0,826	0,376	0,326	0,277	0,415
3	0,805	0,881	**0,881**	0,801	0,380	0,319	0,237	0,345
4	0,859	0,826	0,801	**0,859**	0,436	0,329	0,327	0,365
5	0,473	0,376	0,380	0,436	**0,762**	0,762	0,730	0,629
6	0,398	0,326	0,319	0,329	0,762	**0,762**	0,583	0,577
7	0,301	0,277	0,237	0,327	0,730	0,583	**0,730**	0,539
8	0,382	0,415	0,345	0,365	0,629	0,577	0,539	**0,629**
$\sum r_j$	4,923	4,828	4,649	4,802	4,548	4,056	3,724	3,881
							$\sum\sum r_{ij} = 35{,}411$	

Третий этап. Проводим группировку параметров с целью определения факторов. Восемь параметров образуют две группы (см. табл. 7.1): первые четыре параметра характеризуют химическую мелиорацию почв (первый фактор), остальные – их плодородие (второй фактор).

Таблица 7.3

Редуцированная корреляционная матрица R^x

Номер параметра	1	2	3	4	5	6	7	8	$\sum r_i^{(1)}$	$a_{ij}^{(1)}$
1	**0,684**	0,846	0,805	0,849	0,473	0,398	0,301	0,382	4,738	1,000
2	0,846	**0,658**	0,881	0,826	0,376	0,326	0,277	0,415	4,605	0,971
3	0,805	0,881	**0,610**	0,801	0,380	0,319	0,237	0,345	4,378	0,924
4	0,859	0,826	0,801	**0,651**	0,436	0,329	0,327	0,365	4,595	0,969
5	0,473	0,376	0,380	0,436	**0,584**	0,762	0,730	0,629	4,370	0,922
6	0,398	0,326	0,319	0,329	0,762	**0,464**	0,583	0,577	3,758	0,793
7	0,301	0,277	0,237	0,327	0,730	0,583	**0,391**	0,539	3,391	0,715
8	0,382	0,415	0,345	0,365	0,629	0,577	0,539	**0,425**	3,677	0,776

Четвертый этап. Находим первое приближение факторного отображения. Предполагается, что полученные факторы не коррелируют между собой. Для каждой строки матрицы R^x вычисляем сумму коэффициентов корреляции

$$\sum r_{i1} = r_{i1} + r_{i2} + \ldots r_{ij}, \tag{7.2}$$

где $\sum r_1$ = 0,654 + 0,846 + 0,805 + 0,849 + 0,473 + 0,398 + 0,301 + 0,382 = = 4,738 (см. табл. 7.3). Результаты записываем в предпоследний столбец редуцированной корреляционной матрицы. Каждую сумму $\sum r_i$ делим на максимальное значение (в нашем примере максимальная $\sum r_1 = 4{,}738$). Имеем первое приближение $a_{ij}^{(1)}$) так называемого факторного отображения: $a_{11}^{(1)}$ = 4,738 / 4,738 = 1,000; $a_{21}^{(1)}$ = 4,605 / 4,738 = 0,971. Результаты вносим в последний столбец редуцированной корреляционной матрицы. Эти числа не применяются непосредственно в качестве элементов собственного вектора матрицы.

Пятый этап. Возводим редуцированную матрицу (см. табл. 7.3) в квадрат. Для этого необходимо каждое число возвести в квадрат в первом столбце матрицы и суммировать результаты:

$$(0{,}684)^2 + (0{,}846)^2 + (0{,}805)^2 + (0{,}859)^2 + (0{,}473)^2 +$$
$$+ (0{,}398)^2 + (0{,}301)^2 + (0{,}382)^2 = 3{,}188.$$

Получаем первый элемент матрицы R^2 (табл. 7.4). Поскольку квадрат симметричной матрицы есть также симметричная матрица, то вычисляем диагональные элементы и элементы выше (или ниже) диагонали. Для контроля выполненных вычислений суммируем элементы строк $\sum r_i^{(2)}$ матрицы R^2, например: 3,188 + 3,450 + 3,301 + 3,325 + 2,577 + 2,185 + 1,903 + + 2,179 = 22,11.

Затем определяем сумму элементов строк с помощью формулы:

$$T_i^{(2)} = \sum_{i\|1}^{n} r_i^{(2)} r_{i/j}, \tag{7.3}$$

где $\sum r_i^{(2)}$ – сумма элементов строк матрицы;

$$r_{i/j} = \sum_{i=1}^{n} r_1 / \sum_{j=1}^{m} r_j.$$

Приведем пример расчета указанных выше показателей: $r_{1/1}$ = = 4,738 : 4,923 = 0,962; $T_1^{(2)}$ = 22,11 · 0,962 = 22,26 (см. табл. 7.4). Значение $T_i^{(2)}$ должно соответствовать величине $r_i^{(2)}$, т. е. каждому значению полученной суммы.

Таблица 7.4

Квадрат корреляционной матрицы

Номер параметра	1	2	3	4	5	6	7	8	$\sum r_i^{(2)}$	$T_i^{(2)}$	$a_{ij}^{(2)}$
1	3,188	3,450	3,301	3,325	2,577	2,185	1,903	2,179	22,11	22,26	1
2	3,450	3,475	2,322	3,333	2,471	2,093	1,815	2,115	22,07	22,07	0,986
3	3,301	3,322	3,181	3,188	2,341	1,983	1,718	2,003	21,03	21,03	0,94
4	3,325	3,333	3,188	3,213	2,461	2,084	1,810	2,085	21,49	21,49	0,961
5	2,577	2,471	2,341	2,461	2,966	2,550	2,780	2,372	20,01	20,01	0,894
6	2,185	2,093	1,983	2,084	2,550	2,200	1,965	2,041	17,10	17,10	0,764
7	1,903	1,815	1,718	1,810	2,278	1,965	1,765	1,820	15,07	15,07	0,673
8	2,176	2,115	2,003	2,085	2,372	2,041	1,820	1,925	16,54	16,54	0,739

Величина $a_{ij}^{(2)}$ представляет собой второе приближение чисел, которое получаем путем деления каждой из сумм $T_i^{(2)}$ на максимальную величину в данном столбце (см. табл. 7.4). Показатель $a_{ij}^{(2)}$ сравниваем с соответствующими значениями первого приближения $a_{ij}^{(1)}$ (см. табл. 7.3) (различие между ними должно быть < 0,005). Различие $a_{i3}^{(1)}$ и $a_{i3}^{(2)}$ результатов превышает 0,005, поэтому возведение корреляционной матрицы в квадрат производится до тех пор, пока собственный вектор не перестанет изменяться. Перед очередной операцией возведения матрицы R в степень вычисляются значения $T_i^{(e)}$ и $a_{ij}^{(e)}$ последующей матрицы R^e. Если коэффициенты a_{ij} последующей матрицы совпадают с коэффициентами предыдущей матрицы с достаточной точностью, то нет необходимости вычислять остальные элементы матрицы R^e. В нашем случае максимальное различие между $a_{7j}^{(8)}$ и $a_{7j}^{(4)}$ составляет всего 0,004, поэтому элементы R^e можно не вычислять (табл. 7.5).

Таблица 7.5

Показатели четвертой и восьмой степени корреляционной матрицы

Номер параметра	$T_i^{(4)}$	$a_{ij}^{(4)}$	$T_i^{(8)}$	$a_{ij}^{(8)}$
1	447,9	1,000	176,7	1,000
2	443,0	0,989	174,8	0,989
3	422,4	0,943	166,7	0,943
4	430,7	0,961	19,9	0,961
5	390,9	0,872	153,7	0,869
6	333,6	0,744	131,2	0,742
7	293,6	0,655	115,4	0,651
8	324,0	0,723	127,5	0,721

Шестой этап. Вычисляем коэффициенты при первом факторе F_1. Найденное восьмое приближение чисел $a_{7j}^{(8)}$ (см. табл. 7.5) представляет собой вектор и умножается на R. Значение R_{q1}, соответствующее $a_{11}^{(8)} = = 1{,}000$, является первым корнем характеристического уравнения λ_k. Далее рассчитываем коэффициенты b_{i1} при первом факторе F_1 (табл. 7.6), которые учитывают максимально возможную долю суммарной общности:

$$b_{i1} = a_{i1}\sqrt{\lambda_1} / \sqrt{\sum a_{i1}^{(2)}}, \qquad (7.4)$$

где $a_{i1} = a_{11}^{(8)}$; $\sum a_{i1}^2 = \sum a_{i1}^{2(2)}$; $\lambda = \sum_{i=1}^{n} b_{i1}^2$; $b_{i1} = 0{,}85827$.

Получим искомые коэффициенты b_{i1} при F_1 в факторном отображении. Сумма вкладов первого фактора в суммарную общность должна быть равна первому характеристическому корню:

$$\sum_{i=1}^{n} b_{i1}^2 = \lambda_1. \qquad (7.5)$$

В нашем примере $\lambda_1 = 4{,}4556$, $\sum_{i=1}^{8} b_{i1}^2 = 4{,}455$; следовательно, результаты являются удовлетворительными.

Таблица 7.6

Квадрат корреляционной матрицы

Номер параметра	$a_{i1}^{(8)}$	R_{q1}	b_{i1}
1	1,000	4,455	0,858
2	0,989	4,408	0,849
3	0,943	4,208	0,810
4	0,961	4,285	0,825
5	0,869	3,875	0,747
6	0,742	3,307	0,637
7	0,653	2,909	0,561
8	0,721	3,214	0,619
			D_1=5,905

Седьмой этап. Проводим поиск фактора, который учитывал бы максимум остаточной общности. Для этого после учета F_1 необходимо построить матрицу R_1 используя коэффициенты первого фактора. По строкам табл. 7.7 рассчитываются суммы элементов E_{i1}. Например, E_{11} = 0,736 + + 0,728 + 0,695 + 0,708 + 0,641 + 0,547 + 0,481 + 0,531 = 5,067. Результаты сравниваем с произведениями $b_{i1}D_1$, где $D_1 = \Sigma b_{i1} = 5{,}905$ (см. табл. 7.6).

Таблица 7.7

Матрица произведений $\widetilde{R}_1(a_{i1}a_{j1})$

Номер параметра	1	2	3	4	5	6	7	8	E_{i1}	$b_{i1}D_1$
1	0,736	0,728	0,695	0,708	0,641	0,547	0,481	0,531	5,067	5,067
2	0,728	0,721	0,688	0,700	0,634	0,541	0,476	0,526	5,014	5,014
3	0,695	0,688	0,656	0,668	0,605	0,516	0,454	0,501	4,783	4,784
4	0,708	0,700	0,668	0,681	0,616	0,526	0,463	0,511	4,873	4,782
5	0,641	0,634	0,605	0,616	0,558	0,476	0,419	0,462	4,411	4,412
6	0,547	0,541	0,516	0,526	0,476	0,406	0,357	0,394	3,763	3,762
7	0,481	0,476	0,454	0,463	0,419	0,357	0,315	0,347	3,312	3,313
8	0,531	0,526	0,501	0,511	0,462	0,394	0,347	0,383	3,655	3,656

Первые остаточные коэффициенты корреляции (табл. 7.8) равны разности соответствующих элементов матриц R^x и R_1 (см. табл. 7.3 и 7.7). Суммы элементов матрицы R_1, полученной по строкам, должны быть равны разности соответствующих сумм матриц R^x и R_1.

После выполнения необходимых операций по первому фактору и получения соответствующих показателей (табл. 7.9) переходим к вычислению элементов матрицы по второму фактору, сводные сведения по которым приведены в табл. 7.10. В итоге получаем коэффициенты факторного отображения и общности (табл. 7.11), по которым делаем соответствующие выводы.

Таблица 7.8

Матрица первых остаточных коэффициентов корреляции R_1

Номер параметра	1	2	3	4	5	6	7	8	$\sum r_{i1}$	$a_{i1}^{(1)}$
1	0,118	0,118	0,110	0,151	–0,168	–0,149	–0,180	–0,149	–0,149	–0,608
2	0,118	0,176	0,193	0,126	–0,258	–0,215	–0,199	–0,111	–0,170	–0,693
3	0,110	0,193	0,177	0,133	–0,225	–0,197	–0,217	–0,156	–0,182	–0,742
4	0,151	0,126	0,133	0,102	–0,180	–0,197	–0,136	–0,146	–0,147	–0,600
5	–0,168	–0,258	–0,225	–0,180	0,312	0,286	0,311	0,167	0,245	1,000
6	–0,149	–0,215	–0,197	–0,197	0,286	0,281	0,226	0,183	0,218	0,889
7	–0,180	–0,199	–0,217	–0,136	0,311	0,226	0,206	0,192	0,203	0,828
8	–0,149	–0,111	–0,156	–0,146	0,167	0,183	0,192	0,196	0,176	0,718

Таблица 7.9

Этапы вычисления приближенных значений коэффициентов

Номер параметра	$\sum r_{ij}$	E_{i1}	$\sum r_{i1}$	$a_{i1}^{(1)}$	$\lambda_1 \sum r_1$	$\sum r_1^{(2)}$	$\sum r_{i1}^{(2)}$	$a_{i1}^{(2)}$
1	4,918	5,067	–0,149	–0,608	22,37	22,57	–0,20	–0,57
2	4,844	5,014	–0,170	–0,693	22,07	22,33	–0,26	–0,73
3	4,601	4,783	–0,182	–0,742	21,04	21,30	–0,26	–0,73
4	4,726	4,873	–0,147	–0,600	21,50	21,70	–0,20	–0,57
5	4,656	4,411	0,245	1,000	20,02	19,65	0,37	1,00
6	3,981	3,763	0,218	0,889	17,10	16,76	0,34	0,89
7	3,521	3,312	0,203	0,828	15,07	14,75	0,32	0,84
8	3,831	3,655	0,176	0,718	16,54	16,28	0,26	0,68

Таблица 7.10

Вычисление коэффициентов при факторе F_2

Номер параметра	a_{i2}	R_{1q2}	a_{i2}	R_{1q2}	a_{i2}	R_{1q2}	a^2_{i2}	b_{i2}
1	–0,57	–0,865	–0,580	–0,8856	–0,5851	–0,8852	–0,5851	–0,328
2	–0,73	–1,100	–0,737	–1,1152	–0,7368	–1,1148	–0,7369	–0,414
3	–0,73	–1,097	–0,735	–1,1118	–0,7345	–1,1112	–0,6345	–0,412
4	–0,57	–0,902	–0,605	–0,9129	–0,6031	–0,9129	–0,6034	–0,339
5	1,00	1,492	1,000	1,5136	1,0000	1,5129	1,0000	0,561
6	0,89	1,348	0,903	1,3666	0,9029	0,3660	1,9029	0,507
7	0,84	1,300	0,871	1,3144	0,8684	0,3142	1,8687	0,488
8	0,68	0,987	0,662	1,0050	0,6610	0,0001	1,6610	0,371

Таблица 7.11

Этапы вычисления приближенных значений коэффициентов

Параметры	Коэффициенты факторного отображения			Общность	
	b_{i1}	b_{i2}	u_i	исходная	вычисленная
1. Органические удобрения	0,858	–0,328	0,395	0,854	0,844
2. Минеральные удобрения	0,849	–0,414	0,328	0,897	0,892
3. Известь	0,810	–0,412	0,417	0,833	0,826
4. Пестициды	0,825	–0,339	0,452	0,783	0,796
5. Гумус	0,747	0,567	0,375	0,870	0,873
6. Реакция почвы	0,637	0,507	0,581	0,687	0,663
7. Влажность почвы	0,561	0,488	0,669	0,521	0,553
8. Физическая глина	0,619	0,371	0,692	0,579	0,521
Сумма				6,024	5,968
Вклад факторов	4,455	1,511			
Процент от суммарной исходной общности	74,00	25,10			99,10

Поскольку коэффициенты при первом факторе (b_{i1}) положительные и достаточно велики, можно утверждать, что роль первого фактора (химическая мелиорация) в эволюции агроландшафтов весьма существенна. Второй фактор (b_{i2}) (плодородие почв) относится к биполярным, так как имеет одинаковое число положительных и отрицательных нагрузок: коэффициенты со знаком плюс соответствуют параметрам, отражающим степень плодородия почв, со знаком минус – параметрам, отражающим химическую мелиорацию. Таким образом, эволюция агроландшафтов обусловлена прежде всего химической мелиорацией почв. Параметры плодородия почв формируются под воздействием первого фактора комплексной химической мелиорации и в эволюции агроландшафтов выполняют второстепенную роль. Из всех параметров наибольший удельный вес в эволюции агроландшафтов занимают органические удобрения (b_{i1} = 0,858). Коэффициенты факторного отображения второго фактора, выраженные отрицательными числами, характерны для показателей, описывающих степень химизации почв. Это позволяет интерпретировать полученные данные как дефицит химических мелиорантов для рассматриваемых конкретных условий, что отражается отрицательно на прогрессивной эволюции агроландшафтов.

Изложенные выводы подтверждаются данными рис. 7.1. Судя по размещению коэффициентов факторного отображения, параметры *1–4* (химические мелиоранты) расположены компактно в пространстве, что указывает на их важную совместную роль в эволюции агроландшафтов.

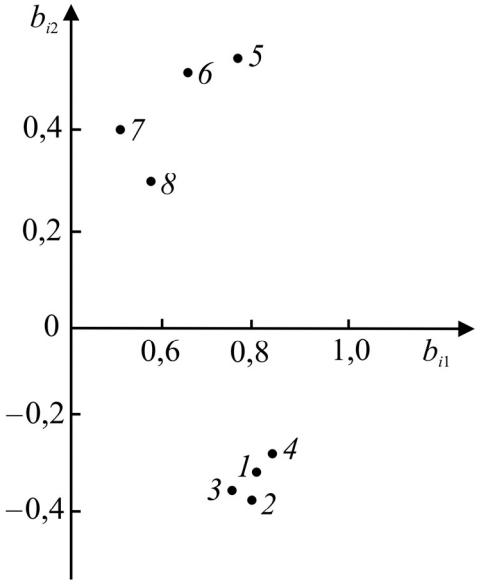


Рис. 7.1. Распределение коэффициентов факторного отображения

Между параметрами *5–8*, отражающими степень плодородия почв, связь слабее и соответственно слабее их влияние на эволюцию агроландшафтов.

Метод факторного отображения используется при решении ряда других географических задач: для целей инженерно-географического районирования и количественной оценки влияния природных условий на производство; для организации отдыха и т. д. По матрицам значений факторов можно составлять картосхемы, на основе которых осуществляется территориальный анализ выражения важнейших факторов.

Глава 8

МЕТОДЫ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Линейные модели активно используются в экономике и экономической географии как достаточно эффективные в ряде ситуаций. Линейная функция (тройное правило) самая удобная, простая, хорошо разработанная математическая модель.

Линейность – это свойство математических выражений и функций. Выражение типа $ax + by$, где x, y – переменные величины, a, b – постоянные числа, называется линейным относительно переменных x, y. Если переменных больше двух (x_1, x_2, ..., x_n), линейное выражение относительно них имеет вид:

$$a_1x_1 + a_2x_2 + \ldots a_nx_n .$$

В линейное выражение все переменные входят в первой степени и никакие переменные не перемножаются.

Линейное программирование – это совокупность методов решения экстремальных задач, в которых цель (критерий оптимальности) и условия (ограничения) заданы уравнениями и неравенствами первой степени. Программирование используется в данной ситуации как планирование, линейное – означает, что ищется экстремум линейной целевой функции при линейных ограничениях (линейных уравнениях, линейных неравенствах). Однако вычислительные средства при решении задач этого класса играют существенную роль в повышении эффективности их приложений.

Для решения задач с применением линейного программирования эффективны следующие:

- составление смеси продукции предполагает выбор наиболее экономичного топлива, пищевых продуктов и т. д.;
- задачи производства – подбор наиболее выгодной производственной программы выпуска одного или нескольких видов продукции при использовании некоторого числа ограниченных источников сырья;
- задачи распределения, или транспортные задачи;
- комбинированные задачи – производство товаров в разных местах, задачи производства и распределения объединяют в единую задачу.

Разработан ряд алгоритмов, среди которых наиболее известны *симплексный* и *распределительный методы*. Наиболее эффективен метод *эллипсоидов* (*графический*). Оба метода базируются на последовательном улучшении первоначального плана путем повторения вычислений (интераций). После каждой интерации значение целевой функции улучшается. Процесс повторяется до получения оптимального плана, а полученный план проверяется на оптимальность простыми критериями.

Симплекс-метод более универсален, так как позволяет решать задачи, условия которых выражены в различных единицах измерения. В задачах, решаемых распределительным методом (транспортные задачи), все переменные должны иметь одну и ту же единицу измерения. Транспортные задачи являются специальной разновидностью симплекс-метода.

Землеустроительные задачи, решаемые методами линейного программирования, должны удовлетворять следующим требованиям:

- их решение не должно быть однозначным;
- иметь определенную целевую функцию, для которой ведется поиск максимального и минимального значения;
- иметь условия ограничения, формирующие область допустимых решений задачи.

8.1. Составные части общей модели линейного программирования

Все модели линейного программирования состоят из стандартных составных частей: совокупность основных переменных, линейные ограничения (условия), целевая функция, определяющая критерий оптимальности задачи.

Совокупность основных переменных характеризует размеры землепользований, площади, объемы производства, затраты материальных, трудовых, финансовых ресурсов.

Система линейных ограничений (*условий*) определяет область допустимых значений основных переменных. Каждое отдельное условие отражает реальное ограничение (нормы внесения удобрений, выполнение контрольных цифр бизнес-плана и т. д.).

Целевая функция представляется показателем, который обобщенно характеризует один из аспектов деятельности хозяйства данной землеустроительной задачи, например, чистый доход, валовую продукцию и т. д.

Критерий оптимальности в зависимости от условий задачи требует максимизации или минимизации целевой функции при заданных ограничениях.

В общем виде формализованная модель линейного программирования, построенная для решения задачи, в которой выделено N основных переменных $x_1, \ldots, x_N$ и M ограничений, будет иметь вид целевой функции:

$$Z = (c_1x_1 + c_2x_2 + \ldots + c_Nx_N) \rightarrow \max \text{ (min)}. \qquad (8.1)$$

Система ограничений:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1N}x_N \therefore b_1; \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2N}x_N \therefore b_2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{M1}x_1 + a_{M2}x_2 + \ldots + a_{MN}x_N \therefore b_M, \end{cases}$$

где знак $\therefore$ означает или $\leq$, или $\geq$, или $=$; константы $b_1 \ldots b_M$ в правых частях ограничений предполагаются неотрицательными; требование неотрицательности основных переменных: $x_1 \geq 0, x_2 \geq 0, \ldots, x_N \geq 0$.

Совокупность ограничений и требований неотрицательности основных переменных определяет область допустимых значений задачи.

В кратком виде вышеприведенная развернутая запись будет следующей:

$$Z = \sum_{j=1}^{N} (c_j x_j) \rightarrow \max \text{ (min)}; \ \sum_{j=1}^{N} (a_{ij} x_j) \therefore b_i;$$

$$x_j \geq 0, \ \ j = 1, \ldots, N. \qquad (8.2)$$

8.2. Распределительная модель линейного программирования

Распределительный метод среди задач линейного программирования получил распространение из-за упрощения расчетов, точности вычислений и снижения затрат времени на ввод исходной информации. Метод предложили А. Толстой и Л. В. Конторович в 1939–1940 гг. Первоначально он применялся в задачах, связанных с транспортировкой грузов, их распределением между поставщиком и потребителем, поэтому получил название «транспортная задача». Применяется при решении ряда землеустроительных задач: распределение севооборотов и угодий по участкам, размещение культур на землях различных категорий, перераспределение участков между хозяйствами для экономии транспортных затрат и др.

Суть распределительной задачи следующая. Задано m источников ресурса (производители продукции, базы с готовой продукцией) и n пунктов его потребления. Запасы ресурса в источниках составляют A_i, i=1, …, m, потребности – B_j, j=1, …, n. Стоимость транспортировки еди-

ницы ресурса от i-го источника к j-му потребителю C_{ij}. Количество ресурса, транспортируемого от i-го источника к j-му потребителю X_{ij}. Требуется определить такие значения X_{ij}, при которых общие транспортные расходы будут минимальны.

При сбалансированности, когда общий спрос на запас ресурса у поставщиков и общий спрос на него у потребителя равны, задачу называют *закрытой*:

$$\sum_{i=1}^{m} A_i = \sum_{j=1}^{n} B_j. \tag{8.3}$$

Если баланс не выдерживается, то транспортная задача является *открытой*:

$$\sum_{i=1}^{m} A_i < \sum_{j=1}^{n} B_j, \text{ или } \sum_{i=1}^{m} A_i > \sum_{j=1}^{n} B_j. \tag{8.4}$$

При наличии баланса модель транспортной задачи формулируется следующим образом.

Целевая функция:

$$Z = \sum C_{ij} X_{ij} \to \min(\max). \tag{8.5}$$

Условия. Ограничения по запасам:

$$\sum_{j=1}^{n} X_{ij} = A_i, i = 1, ..., m. \tag{8.6}$$

Ограничения по потребностям:

$$\sum_{j=1}^{m} X_{ij} = B_j, \ \ j = 1, ..., n. \tag{8.7}$$

Условие баланса:

$$\sum_{i=1}^{m} A_i = \sum_{j=1}^{n} B_j. \tag{8.8}$$

Условие неотрицательности:

$$X_{ij} \geq 0, i = 1, \ldots, m, j = 1, \ldots, n. \tag{8.9}$$

Особенности распределительных транспортных задач следующие:

- условия задачи описываются уравнениями (в симплекс-методе описываются и неравенствами);
- все переменные выражаются в одних и тех же единицах измерения;
- во всех уравнениях коэффициенты при переменных равны единице;
- каждая переменная встречается только в двух уравнениях системы ограничений: в одном по строке (по запасам) и в одном по столбцу (по потребностям).

Целевая функция Z выражает суммарные расходы на транспортировку грузов. Ограничения по запасам и по потребностям означают, что сумма ресурса, забираемого из i-го источника, должна быть равна запасу ресурса в нем, как и сумма ресурса, доставляемого j-му потребителю, должна быть равна его потребности.

Величина C_{ij} может выражать транспортные расходы (минимизация) или прибыль от транспортных операций (максимизация) и другие показатели.

П р и м е р землеустроительной задачи, решаемой транспортным методом. При землеустроительном обследовании в хозяйстве выделено 5 участков с различным плодородием, которые пригодны для трансформирования. Площади участков 250, 100, 520, 310 и 130 га. По проекту на них намечается разместить кормовой севооборот площадью 600 га, полевой – 560 га, улучшенные сенокосы – 150 га. Необходимо распределить севообороты и угодья по участкам так, чтобы получить максимальный чистый доход.

Матрицу исходных данных строим, как в табл. 8.1.

Таблица 8.1

Исходные данные для землеустроительной задачи

Угодья и севообороты	b_j / a_i	Чистый доход при размещении на участке, руб/га (C_{ij})				
		1 пастбище, 250 га	2 пашня, 100 га	3 пашня, 520 га	4 пашня, 310 га	5 сенокосы, 130 га
Кормовой	600 га	800	1100	800	600	440
Полевой	560 га	1000	1800	2000	2200	2000
Улучшенные сенокосы	150 га	550	440	380	300	700

На «транспортном» языке эта задача может быть описана следующим образом. «Ресурсы» в источниках (A_i) – площади севооборотов и улучшенных сенокосов; «потребности в ресурсах» (B_j) – площади участков; «прибыль от транспортных операций» (C_{ij}) – чистый доход с единицы площади; «транспортируемый ресурс» (X_{ij}) – часть площади i-го севооборота или угодья, размещаемого на j-м участке; максимальная целевая функция (F) – чистый доход хозяйства от рационального размещения и трансформации угодий; $\sum a_i = \sum b_j$. Чистый доход проставляется в правом верхнем углу каждой клетки (C_{ij}, руб/га). Дальнейшее решение задачи проводится с использованием метода потенциалов.

Для решения транспортных задач разработан ряд методов: функционала, потенциала, дельта-метод, лямбда-задача. Используются модифицированные модели: транспортно-производственная, многоэтапная, многопродуктовая.

Вначале рассмотрим основные правила работы с матрицей, составление и перемещение по цепи и расчет необходимых параметров.

8.3. Правила работы с матрицей

Расположение элемента (числа) в матрице строго фиксировано. Строку обозначают буквой *i,* столбец – *j*, элемент матрицы – a_{ij} (где *i* – номер строки, *j* – номер столбца). Запись $a_{12,7}$ показывает, что данный элемент расположен в 12-й строке и 7-м столбце матрицы. Цифры, указывающие строку и столбец, до 10 не разделяют запятой (a_{23}).

Матрицу обозначают заглавной буквой (*A*, *B*, *C*):

$$A \begin{bmatrix} a_{11}a_{12}\ldots a_{1n} \\ a_{21}a_{22}\ldots a_{2n} \\ \ldots\ldots\ldots\ldots \\ a_{m1}a_{m2}\ldots a_{mn} \end{bmatrix}.$$

В матрице число строк равно *m*, столбцов – *n*. В сокращенном виде матрицу записывают $A=(a_{ij})$.

Размер матрицы определяют путем произведения *m* на *n*. Запись $\sum_{i=1}^{m} a_{ij}$ означает, что в матрице из чисел *a* необходимо просуммировать все числа матрицы по столбцам.

Матрицу можно транспонировать, т. е. перемещать элементы матрицы так, что ее строки становятся столбцами, а столбцы – строками. При большинстве вычислений (кроме умножения матрицы на матрицу) не имеет значения, что считать в ней строками, а что столбцами.

Вектор в матрице представляет собой упорядоченную последовательность элементов или ряд, состоящий из некоторого количества элементов. Поэтому вектором можно считать любую строку или любой столбец матрицы. Если размер матрицы $m \cdot n$, то она состоит либо из *m* векторов, в каждом из которых по *n* элементов, либо из *n* векторов, в каждом из которых по *m* элементов.

При решении транспортной задачи используются следующие обозначения:

i – индекс поставщика ($i = 1, 2, \ldots, m$);

j – индекс потребителя ($j = 1, 2, \ldots, n$);

a_i – мощность *i*-го поставщика;

b_j – спрос *j*-го потребителя;

C_{ij} – затраты на перевозку продукции от *i*-го поставщика *j*-му потребителю;

X_{ij} – количество продукции, которое необходимо перевезти от i-го поставщика j-му потребителю.

Условия транспортной модели приведены выше в составных частях общей модели линейного программирования. Совокупные затраты на перевозку сводятся к минимуму целевой функции. Исходная информация для решения транспортной задачи представлена в матрице:

Потребители		B_1	B_2	……	B_n
Поставщики	b_j / a_i	b_1	b_2	.....	b_n
A_1	a_1	C_{11} X_{11}	C_{21} X_{12}	.....	C_{1n} X_{1n}
A_2	a_2	C_{21} X_{11}	C_{22} X_{22}	.....	C_{2n} X_{2n}
…….	……	……	……	……	……
A_m	a_m	C_{m1} X_{m1}	C_{m2} X_{m2}		C_{mn} X_{mn}

В транспортной задаче $m \cdot n > (m + n - 1)$ можно составить множество планов перевозок. Такие планы называют *допустимыми.*

В табл. 8.2 дана запись исходных данных задачи по трем поставщикам и четырем потребителям. Указаны мощности поставщиков (a_i) и спросы потребителей (b_j), в правом верхнем углу клетки – затраты на перевозку единицы груза (C_{ij}).

Таблица 8.2

Исходные данные транспортной задачи

Поставщики, их мощности (a_i)		Потребители и их спрос (b_j)			
		B_1 **40**	B_2 **25**	B_3 **15**	B_4 **20**
A_1	**30**	5	4	1	2
A_2	**50**	1	2	3	4
A_3	**20**	3	2	5	1

По исходным данным табл. 8.2 могут быть составлены следующие допустимые планы перевозок (табл. 8.3).

В допустимом плане C_{ij} обводятся кружком в случаях наличия в таких клетках поставок (X_{ij}), поэтому клетки таблиц с поставками условимся называть *клетками с кружками.*

В табл. 8.3,*а* число кружков 5, т. е. меньше, чем $m + n - 1$; в табл. 8.3,*б* число кружков 6 (равно $m + n - 1$); в табл. 8.3,*в* кружков 7 (больше, чем $m + n - 1$). В методе потенциалов число кружков в допустимом плане должно быть равно $m + n - 1$ (вариант *б*) и они должны быть расположены в порядке *вычеркиваемой комбинации.*

Таблица 8.3

Допустимые планы перевозок грузов

а

a_i \ b_j	**40**	**25**	**15**	**20**
30			25 (4)	5 (1)
50	40 (1)		10 (3)	
20				20 (1)

б

a_i \ b_j	**40**	**25**	**15**	**20**
30			15 (1)	15 (2)
50	40 (1)	5 (2)		5 (4)
20		20 (2)		

в

a_i \ b_j	**40**	**25**	**15**	**20**
30			15 (1)	15 (2)
50	25 (1)	20 (2)		5 (4)
20	15 (3)	5 (2)		

Вычеркиваемая комбинация получается в случае, если каждый кружок – единственный в своем столбце или строке, и тогда он может быть вычеркнут. Такому условию соответствует распределение в табл. 8.3,*б*. Последовательность вычеркивания следующая: 40; 15; 20; 15; 5; 5 или 20; 40; 15; 5; 15; 5.

Существует несколько способов составления допустимого (базисного) плана: северо-западного угла, поисков наименьшего элемента в столбце,

наименьшего элемента в строке, наименьшего элемента в матрице. Роль наименьшего элемента выполняет C_{ij} (цифры в правом верхнем углу клеток матрицы).

Способ северо-западного угла более сложный, и в случае большой матрицы не рекомендуется его использование. Если столбцов меньше, используют поиски наименьшего C_{ij} в столбцах; если строк мало – способ поиска наименьшего элемента в строках; если матрица большая, проводится поиск наименьшего элемента в клетках матрицы.

Проведем распределение поставок перечисленными выше способами при одинаковых исходных данных и для сравнения вычислим их функционалы $Z = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}X_{ij} \to \min$.

Способ северо-западного угла. В матрице $m + n - 1 = 7$. Поставки распределяем по диагонали независимо от величины C_{ij}: $30^5 - 25^2 - 15^5 - 10^3$. Остаток поставок распределяем между потребителями с учетом минимальных значений C_{ij}: $5^3, 5^4$.

Таблица 8.4

b_j / a_i	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	30 (5)	2	3	4
A_2 30	5 (4)	25 (2)	2	0 (1)
A_3 20	5 (3)	3	15 (5)	2
A_4 10	2	4	6	10 (3)

Все потребители получили необходимый объем продукции.

Функционал при таком способе распределения имеет величину:

$$Z = (30 \cdot 5) + (25 \cdot 2) + (15 \cdot 5) + (10 \cdot 3) + (5 \cdot 4) + (5 \cdot 3) = 320.$$

Способ поиска наименьшего C_{ij} в столбце. Поставки распределяются последовательно по столбцам с учетом наименьших C_{ij}. Получаем следующую последовательность:

$$10^2 - 20^3 - 10^4 - 25^2 - 15^2 - 5^1 - 5^4.$$

Расчет функционала:

$$Z = (10 \cdot 4) + (20 \cdot 3) + (10 \cdot 2) + (25 \cdot 2) + (15 \cdot 2) + (5 \cdot 4) + (5 \cdot 1) = 225.$$

Полученный функционал ($Z = 225$) указывает, что допустимый план по способу наименьшего элемента в столбце более оптимальный, чем по способу северо-западного угла ($Z = 320$).

Таблица 8.5

a_i \ b_j	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	5	(2) 25	3	(4) 5
A_2 30	(4) 10	2	(2) 15	(1) 5
A_3 20	(3) 20	3	5	2
A_4 10	(2) 10	4	6	3

Способ поиска наименьшего элемента C_{ij} в строке. Поставки распределяем последовательно сверху вниз по строкам с учетом наименьших величин C_{ij}: $25^2 – 10^1 – 20^3 – 10^2 – 15^2 – 5^5$.

Таблица 8.6

a_i \ b_j	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	(5) 5	(2) 25	3	4
A_2 30	(4) 5	2	(2) 15	(1) 10
A_3 20	(3) 20	3	5	2
A_4 10	(2) 10	4	6	3

Получаем функционал:

$$Z = (5 \cdot 5) + (25 \cdot 2) + (5 \cdot 4) + (15 \cdot 2) + (10 \cdot 1) + (20 \cdot 3) + (10 \cdot 2) = 215.$$

По данному способу получен функционал меньше, чем в предыдущем.

Способ поиска наименьшего элемента C_{ij} в матрице. Поставки распределяются начиная с поиска наименьших величин C_{ij} в матрице:

$$10^1 – 15^2 – 25^2 – 10^2 – 20^3 – 5^4 – 5^5.$$

На основании распределения поставок получаем функционал:

$$Z = (5 \cdot 5) + (25 \cdot 2) + (5 \cdot 4) + (15 \cdot 2) + (10 \cdot 1) + (20 \cdot 3) + (10 \cdot 2) = 215.$$

Сопоставляя величины функционала (Z), полученные в результате составления базисного плана, делаем вывод: наименьший функционал (215), а значит, и наиболее оптимальное первоначальное распределение

Таблица 8.7

a_i \ b_j	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	(5) 5	(2) 25	3	4
A_2 30	(4) 5	2	(2) 15	(1) 10
A_3 20	(3) 20	3	5	2
A_4 10	(2) 10	4	6	3

поставок получено в нашем примере по способу наименьшего элемента в матрице и строке.

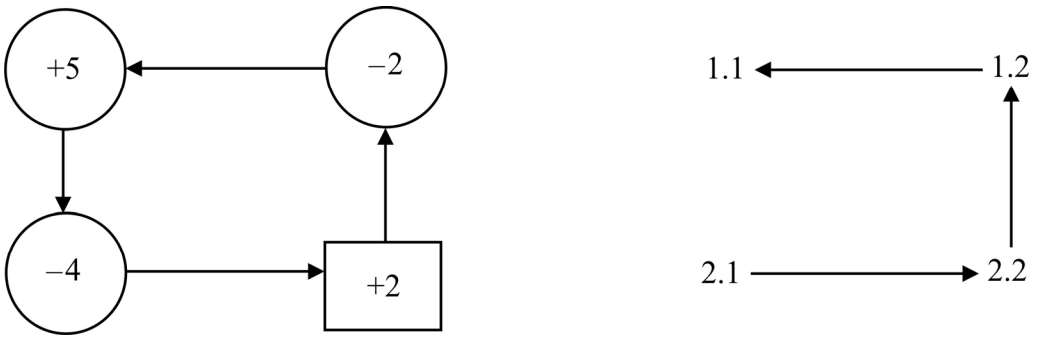
Во всех случаях распределения поставок клеток с кружками в матрицах меньше или равно $m + n - 1$, комбинации кружков вычеркиваемые, поэтому распределение поставок выполнено по установленным правилам.

Изменение базисного допустимого плана. Для получения оптимального плана транспортной задачи следует выполнить условие минимизации Z:

$$Z = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}X_{ij} \to \min$$

путем изменения базисного допустимого плана. Для этого перемещаем меньшую поставку с большим C_{ij} в кружке в клетку, где нет поставки, а значение C_{ij} без кружка меньшее. Произведем перемещения в предыдущей матрице, составленной по способу наименьшего C_{ij} в строке.

Для реализации правила цепи, по которой должна перемещаться поставка, переместим поставку в клетке 2.1, равную 5, в клетку 2.2, где нет поставки. Перемещение проводим в направлении клеток, где есть поставки, и там же делаем повороты под прямым углом, пока цепь не замкнется. В нашем случае цепь имеет следующую форму:



При перемещении поставки в вершинах цепи должны чередоваться плюсы и минусы. В клетке 2.2, куда вносим поставку, должен быть плюс, и ее обозначаем квадратом. Алгебраическая сумма C_{ij} по перемещаемым клеткам дает представление об увеличении функционала при получении положительной суммы или уменьшении – при получении отрицательной: (+5) + (–2) + (+2) + (–4) = + 1. При указанном перемещении поставки базисный допустимый план ухудшился, так как алгебраическая сумма равна + 1.

Других вариантов перемещения поставки по цепи в матрице произвести не можем, так как придется перемещать большие поставки из клеток с меньшей C_{ij} в клетки с большей C_{ij}, т. е. увеличивать функционал.

Если бы по условиям задачи необходимо было свести функционал к максимуму $Z = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}X_{ij} \to \max$, то такие перемещения улучшили бы функционал.

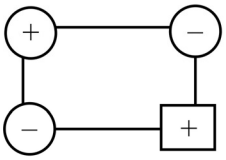
В клетках, где имеются поставки, при прохождении поставки по цепи их величина увеличивается (+) или уменьшается (–) на величину перемещаемой поставки (в нашем случае + или – 5).

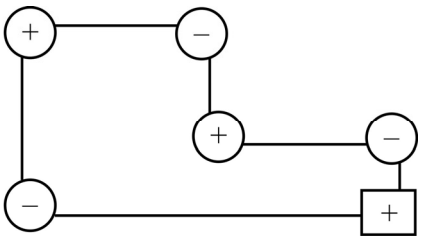
После неудачного перемещения матрица примет вид:

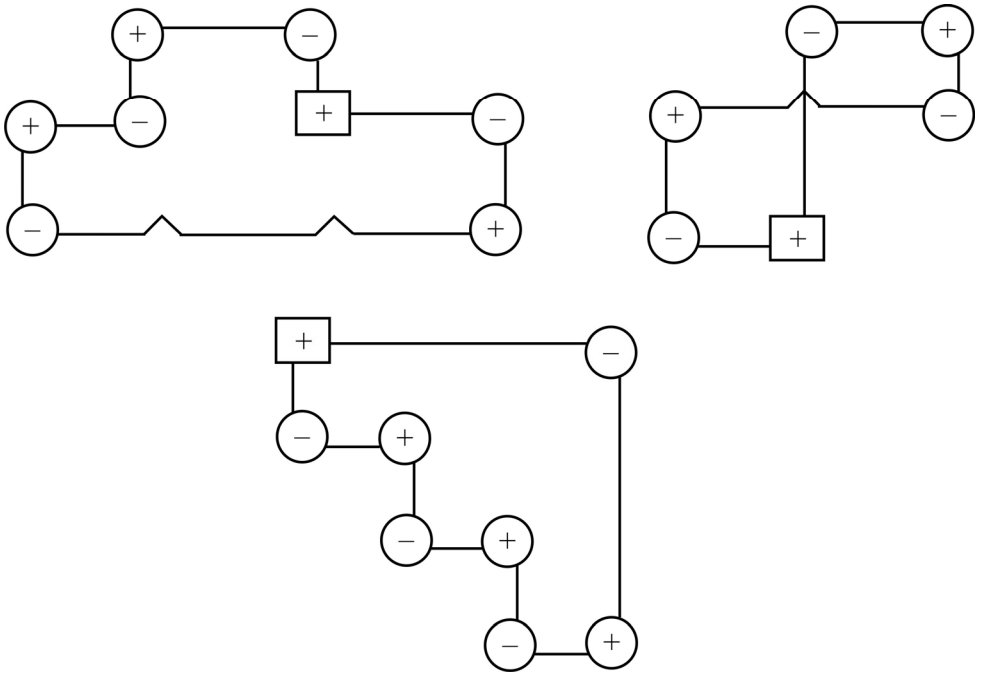
Таблица 8.8

a_i \ b_j	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	(5) 10 ← - - - (2) 20		3	4
A_2 30	4 ↓ - - - → 5 (2)		15 (2)	10 (1)
A_3 20	20 (3)	3	5	2
A_4 10	10 (2)	4	6	3

Замкнутая цепь может иметь различную прямоугольную форму:







Правила построения цепи:

- Цепь должна быть замкнутым многоугольником.
- В цепи четное число вершин.
- Все углы цепи прямые.
- Отрезки цепи могут проходить через клетки матрицы, не являющиеся вершинами данной цепи, хотя в них могут содержаться поставки.
- Положительными (плюсовыми) вершинами будут клетки, в которых при перераспределении по цепи поставки увеличиваются.
- Отрицательными (минусовыми) вершинами являются клетки, в которых поставки при перераспределении уменьшаются.
- В цепи положительные вершины чередуются с отрицательными, количество их равно между собой.
- Вершина-квадрат, куда вносится поставка в ходе перераспределения, всегда положительная.
- При перераспределении поставок по цепи можно двигаться только по горизонтали или вертикали, изменяя направление только в вершинах цепи.
- Клетки, пересекаемые отрезками цепи, вершинами не являются, но в цепи отражаются в виде изломанной линии: ⎯⎯⎯⎯
- Алгебраическая сумма чисел в вершинах цепи C_{ij} (–2 + 3 – 4 + 1 = –2) показывает, на сколько может измениться значение функционала, если внести в вершину-квадрат поставку, равную 1. Сумму называют *характеристикой цепи*. Минусовая сумма указывает на уменьшение величины

функционала, плюсовая – на увеличение. Эта величина равна произведению характеристики цепи (–2) на величину поставки (X_{ij}), которую мы перемещаем по цепи (5), т. е. $–2 \cdot 5 = –10$.

Вырождение матрицы – случаи в математике, которые являются исключением из общего правила. В транспортной задаче вырождение бывает в случаях, когда в допустимом плане поставок число клеток с кружками может быть меньше или больше, чем $m + n – 1$ (см. матрицу составления плана по способу северо-западного угла, где число клеток с кружками $< m + n – 1$, т. е. 6). В данном случае для решения проблемы вырождения в свободную клетку записывают нулевую поставку в порядке вычеркиваемой комбинации (клетка 2.4).

Если число кружков или число поставок в клетках больше $m + n – 1$, как в матрице *а*, необходимо выбрать наименьшую поставку (5) в клетке 3.2 и перераспределить ее по цепи, как в матрице *б*.

Равенство мощностей и спросов (закрытая задача) создает потенциальную возможность вырождения, но не всегда приводит к нему.

Таблица 8.9

а

a_i \ b_j	45	25	15	20
30			15	15
50	30	20		5
20	15	5		

б

a_i \ b_j	45	25	15	20
30			15	15
50	20	25		5
20	20			

8.4. Метод потенциалов

Метод разработан в 1940 г. академиком Л. В. Конторовичем. В 1951 г. американским ученым Дж. Б. Данцигом предложен распределительный метод (МОДИ), аналогичный методу потенциалов. В обоих методах при

проверке допустимого плана на оптимальность определяются *потенциалы* (числа), с помощью которых вычисляются *характеристики клеток без кружков* (в них нет поставок).

Обозначив потенциалы строк через U_i, потенциалы столбцов через V_j, показатели C_{ij} в клетках с поставками и кружками через $\overline{C_{ij}}$, характеристики клеток без кружков (без поставок) через E_{ij}, получим следующие соотношения:

метод потенциалов	метод МОДИ	
$u_i = v_j - \bar{c}_{ij}$;	$u_i = \bar{c}_{ij} - v_j$;	(8.10)
$v_j = \bar{c}_{ij} + u_i$;	$v_j = \bar{c}_{ij} - u_i$;	(8.11)
$\bar{c}_{ij} = v_j - u_i$;	$\bar{c}_{ij} = v_j + u_i$;	(8.12)
$E_{ij} = c_{ij} - (v_j - u_i)$.	$E_{ij} = c_{ij} - (v_j + u_i)$.	(8.13)

Каждый показатель $\bar{c}_{ij}$ (в клетке матрицы он находится в кружке) должен быть равен разнице потенциалов своих столбцов и строк. Определение потенциала можно начинать с любой строки или столбца. Первый потенциал по величине выбирается произвольно (лучше определение начинать с нуля). Величины других потенциалов определяются с использованием предложенных выше формул (при первом вычислении применяется выбранный нами потенциал).

Рассмотрим пример решения транспортной задачи, предложенный В. С. Михеевой (1981). Базисный допустимый план составлен способом наименьшего элемента в столбце, его первоначальный функционал 555 (табл. 8.4).

Вначале рассчитаем потенциалы строк и столбцов по методу потенциалов с использованием формул. Произвольно выбранную величину потенциала выбираем в том столбце или строке, где наибольшее количество клеток с кружками. В нашем примере – это третья строка. В качестве потенциала для нее возьмем число 0. По формуле (8.11) определяем потенциалы (v_j) первого ($v_j = \bar{c}_{ij} + u_i = 0 + 2 = 2$), третьего (0 + 6 = 6), четвертого (0 + 4 = 4), пятого (0 + 2 = 2) столбцов. Зная потенциалы по четырем столбцам, можно вычислить потенциалы строк (u_i) по формуле (8.10): первой ($u_i = v_j - \bar{c}_{ij} = 2 - 1 = 1$), четвертой (6 – 2 = 4). По полученным потенциалам новых строк вычисляем потенциалы новых столбцов, а по ним – новых строк (см. табл. 8.10).

Определив потенциалы строк и столбцов, вычисляют характеристики клеток (E_{ij}) без кружков (в них нет поставок) по формуле (8.13). Приведем расчет характеристики клетки 1.2: $E_{ij} = c_{ij} - (v_j - u_i) = 2 - (7 - 1) = -4$.

Таблица 8.10

Базисный допустимый план (матрица 1)

a_i \ b_j	50	85	35	25	20	u_i
30	(1) **30**	2 *–4*	3 *–2*	4 *1*	5 *4*	1
40	4 *6*	(3) **40**	2 *0*	3 *3*	4 *6*	4
70	(2) **20**	5 *–2*	(6) **5**	(4) **25**	(2) **20**	0
75	4 *6*	(3) **45**	(2) **30**	1 *1*	1 *3*	4
v_j	2	7	6	4	2	

Аналогично рассчитываем E_{ij} (показаны курсивом) для других клеток без кружков.

Среди вычисленных характеристик клеток (курсив в клетках) отрицательные величины получены в клетках матрицы 1.2, 1.3, 3.2 (их величины соответственно: – 4, –2, –2), поэтому составленный первичный базовый план не оптимален. Проводят следующее перераспределение поставок по правилам цепи.

Выбираем клетку с наибольшей отрицательной абсолютной величиной характеристики (E_{12}), равную – 4. К клетке 1.2 строится цепь по перемещению наименьшей поставки 5 из клетки 3.3, так как функционал стремится к минимуму. Путь перемещения следующий: 3.3 (–5)→4.3 (+5)→4.2 (–5)→1.2 (+5)→1.1 (–5)→3.1 (+5)→3.3. К имеющейся поставке в клетке прибавляется или отнимается 5 с целью сохранения баланса между поставщиками и потребителями.

Получаем новую матрицу с измененными поставками (табл. 8.11). В ней повторяем алгоритм расчетов, как в табл. 8.10: рассчитываем потен-

Таблица 8.11

Результаты первого перераспределения поставок (матрица 2)

a_i \ b_j	50	85	35	25	20	u_i
30	(1) **25**	(2) 5	3 *2*	4 *1*	5 *4*	1
40	4 *2*	(3) **40**	2 *0*	3 *–1*	4 *2*	0
70	(2) **25**	5 *2*	6 *4*	(4) **25**	(2) **20**	0
75	4 *2*	(3) **40**	(2) **35**	1 *–3*	1 *–1*	0
v_j	2	3	2	4	2	

циалы строк и столбцов, характеристику клеток без поставок, производим перераспределение поставок с использованием другой минимальной поставки 25 в клетке 3.4. Другая минимальная поставка 25 в клетке 3.1 перемещаться не может по цепи, так как в свободной клетке 4.4 с максимальной отрицательной абсолютной характеристикой (–3) ее следует вычитать из несуществующей поставки.

Поставка 25 в клетке 3.4 перемещается по цепи: 3.4 (–25)→3.1 (+25)→1.1 (–25)→1.2 (+25)→4.2 (–25)→4.4 (+25)→3.4 (цепь замыкается). Расчет потенциалов и характеристики в новой матрице показал, что распределение не оптимально. Получена отрицательная характеристика –1 в клетке 4.5. Следует произвести очередное перераспределение минимальной поставки, равной 0, в клетке 1.1. Здесь получена нулевая поставка, так как из прежней поставки в клетке 25 следовало вычесть перераспределяемую 25. В таких ситуациях допускается наличие нулевой поставки, чтобы не нарушались правила перемещения поставки по цепи. Результаты распределения представлены в матрице 3 (табл. 8.12).

Таблица 8.12

Результаты второго распределения поставок (матрица 3)

a_i \ b_j	50	85	35	25	20	u_i
30	(1) **0**	(2) 30	3 *2*	4 *4*	5 *4*	1
40	4 *2*	(3) **40**	2 *0*	3 *2*	4 *2*	0
70	(2) **50**	5 *2*	6 *4*	4 *3*	(2) **20**	0
75	4 *2*	(3) **15**	(2) **35**	(1) **25**	1 *–1*	0
v_j	2	3	2	1	2	

Таблица 8.13

Результаты третьего распределения поставок (матрица 4)

a_i \ b_j	50	85	35	25	20	u_i
30	(1) **0**	(2) 30	3 *2*	4 *4*	5 *5*	1
40	4 *2*	(3) **40**	2 *0*	3 *2*	4 *4*	0
70	(2) **50**	5 *2*	(6) *4*	4 *3*	(2) **20**	–1
75	4 *2*	(3) **15**	(2) **35**	① **25**	(1) **0**	0
v_j	1	3	2	1	1	

В ней снизилось отрицательное значение E_{ij} до –1. План приблизился к оптимальному и требуется его усовершенствовать. Проводим очередное перераспределение поставок. Минимальную нулевую поставку из клетки 1.1 перемещаем в клетку 4.5, где отрицательная характеристика клетки. Прибавление и вычитание нуля по цепи не изменяет величины поставок в клетках и не нарушает правил построения цепи. В новой матрице 4 (табл. 8.13) после перерасчетов v_j, u_i, E_{ij} получены все *положительные характеристики* цепи при стремлении функционала к *минимуму*, поэтому план распределения поставок *оптимальный*, величина $Z = 460$. По сравнению с базовым планом функционал снизился на 95 единиц.

В задачах при стремлении функционала к *максимуму* план распределения поставок или иного показателя считается *оптимальным*, если в матрице получены *отрицательные характеристики* в клетках.

8.5. Дельта-метод Аганбегяна

Для решения закрытых и открытых транспортных задач А. Г. Аганбегян (1961) разработал дельта-метод для ручной обработки. Исходные данные используем из табл. 8.14. В каждом столбце этой таблицы находим минимальное значение c_{ij} и обводим его кружком. Если в столбце несколько равных по значению c_{ij}, выбираем любой из них (обычно первый сверху).

Вычисляем в каждом столбце приросты затрат (ΔC_{ij}) для строки как разницу между элементом c_{ij} строки и минимальным значением c_{ij} в столбце: $\Delta C_{ij} = c_{ij} - c_{ij\,\min}$. Для первого столбца значения ΔC_{ij} следующие (сверху вниз): 1 – 1 = 0; 4 – 1 = 3; 2 – 1 = 1; 4 – 1 = 3. Аналогично вычисляем ΔC_{ij} для других столбцов. В дальнейшем используем матрицу со значениями ΔC_{ij} в правом верхнем углу (табл. 8.14). В нее заносим поставки по столбцам, равные величине потребителя, в клетки с нулевыми значениями в кружках.

Таблица 8.14

Рабочая матрица прироста затрат

a_i \ b_j	50	85	35	25	20	d_i
30	**50** (0)	**85** (0)	1	3	4	*–105*
40	3	1	**35** (0)	2	3	*5*
70	1 ↓	3	4	3	1	*70*
75	3	1	0	**25** (0)	**20** (0)	*30*

В дальнейшем производится расчет баланса (d_i), по которому определяем избыток или недостаток строк: $d_1 = 30 – (50 + 85) = –105$. Аналогично производится расчет баланса для последующих строк. Отрицательный баланс сложился лишь в первой строке. Значит, план распределения поставок неоптимальный.

Производится перераспределение поставок из строк с минусовым балансом в строки с плюсовым балансом и учетом минимального значения прироста затрат (ΔC_{ij}). У нас минимальные значения $\Delta C_{ij} = 1$ в трех клетках с положительным, но разным балансом в строках (3.1; 2.2; 4.2). Так как d_3 (70) больше, чем d_4 (30) и d_2 (5), выбирается клетка 3.1 для перемещения поставки (50) из соответствующего ей столбца 1. В строку с нулевым балансом поставка не перемещается.

В клетке 3.1 новой матрицы (табл. 8.15) обводим кружком ΔC_{ij}. В эту клетку (указано стрелкой) вносим поставку 50 из клетки 1.1, т. е. наименьшую из строки с отрицательным балансом –105. На величину 50 уменьшится отрицательный баланс первой строки и составит ($d_1 = 30 – 85 = – 55$), а также положительный баланс третьей строки ($d_3 = 70 – 50 = 20$).

Таблица 8.15

Первый вариант перемещения поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	(0) **85**	1	3	4	*–55*
40	3	1	(0) **35**	2	3	*5*
70	(1) **50**	3	4	3	1	*20*
75	3	1 ↓	0	(0) **25**	(0) **20**	*30*

После первого перемещения поставки отрицательный баланс в первой строке сохранился. Необходимо продолжить перемещение поставки из минусовой строки в плюсовую по описанному выше алгоритму. Следует из клетки 1.2 переместить поставку в клетку 4.2 величиной не более d_4 (30). В результате перемещения новая матрица примет вид, как в табл. 8.16.

Второе перемещение поставки не привело к исчезновению отрицательного баланса первой строки ($d_1 = –25$). Следует переместить поставку из первой строки в клетке 1.2, равную этой величине d_1, в строку с положительным потенциалом. В табл. 8.16 положительный потенциал имеют вторая и третья строка. Их суммарная величина соответствует величине отрицательного баланса первой строки. Поэтому из поставки клетки 1.2 (55) сначала перемещаем 5 в клетку 2.2, так как прирост затрат здесь наименьший и новая матрица примет вид табл. 8.17.

Таблица 8.16

Второе перемещение поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	(0) **55**	1	3	4	*–25*
40	3	1 ↓	(0) **35**	2	3	*5*
70	(1) **50**	3	4	3	1	*20*
75	3	(1) **30**	0	(0) **25**	(0) **20**	*0*

Таблица 8.17

Третье перемещение поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	(0) **50**	1	3	4	*–20*
40	3	(1) **5**	(0) **35**	2	3	*0*
70	(1) **50**	3 ↓	4	3	1 ↑	*20*
75	3	(1) **30**	0 →	(0) **25**	(0) **20**	*0*

Затем переместим поставку 20 из клетки 1.2 в третью строку с положительным балансом 20 в клетку 3.5 с наименьшим приростом затрат ($\Delta C_{ij} = 1$). Оптимальный путь перемещения поставки 20 указан стрелками. В дельта-задаче цепь открытая. При перемещении поставки по цепи сохраняется чередование плюсов и минусов с изменением величин поставок на поворотах цепи под прямым углом, как и в методе потенциалов. Новая матрица примет вид табл. 8.18, где нет отрицательного баланса, все значения его по строкам нулевые.

Таблица 8.18

Четвертое перемещение поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	(0) **30**	1	3	4	*0*
40	3	(1) **5**	(0) **35**	2	3	*0*
70	(1) **50**	3	4	3	(1) **20**	*0*
75	3	(1) **50**	0	(0) **25**	(0) **0**	*0*

При перемещении поставки в другие клетки увеличивается прирост затрат. Функционал будет стремиться к максимуму вместо минимума. Следовательно, получен оптимальный план размещения поставок с использованием дельта-метода.

Проверка решения. В дельта-методе поиск клетки в плюсовой строке, к которой следует строить цепь, не формализован и опирается на мыслительную способность человека. Поэтому могут быть допущены ошибки при выборе наиболее выгодных цепей, и в результате будет получен допустимый план вместо оптимального.

В оптимальном варианте распределения поставок (табл. 8.18) можно рассчитать потенциалы строк и столбцов, а также характеристики клеток без кружков, чтобы убедиться в отсутствии отрицательных характеристик, как в методе потенциалов, но несколько иным способом. Для той плюсовой строки, которая на последнем шаге вычислительных операций превратилась в нулевую, берется потенциал, равный нулю. Для минусовой строки, которая на последнем шаге вычислительных операций превратилась также в нулевую, берется потенциал, равный приросту затрат на этом последнем шаге, т. е. алгебраическая сумма цепи: – 0 (клетка 1.2) + + 1 (4.2) + (– 0 (4.5)) + 1 (3.5) = + 2. В нашем примере третья строка имеет потенциал, равный нулю, а первая с отрицательным балансом – плюс 2. Расчет потенциалов остальных строк и столбцов осуществляется по формулам Конторовича (8.16–8.19). Для проверки правильности решения можно пользоваться матрицей со значениями $\overline{C_{ij}}$ или ΔC_{ij}.

Проще проверить оптимальность плана по дельта-методу вычислением функционала и сравнением его с функционалом табл. 8.7 ($Z = 460$):

$$Z = \overline{C_{ij}} \cdot X_{ij} = 2 \cdot 50 + 2 \cdot 30 + 3 \cdot 5 + 3 \cdot 50 + 2 \cdot 35 + 1 \cdot 25 + 2 \cdot 20 = 460.$$

Величина $\overline{C_{ij}}$ для соответствующих клеток табл. 8.18 взята из табл. 8.7. Оба метода распределения поставок показали один и тот же результат функционала, равный 460.

Отличие построения цепей в дельта-методе:

- цепь строится незамкнутая;
- цепь начинается в клетке с кружком (с поставкой), которая находится в минусовой строке; в этой клетке поставка уменьшается, и она становится отрицательной вершиной цепи;
- перемещение поставки в конец открытой цепи производится, как в методе потенциалов, с чередованием положительных и отрицательных вершин;
- в этом методе не требуется количества кружков (клеток с поставками), равного $m + n - 1$;

• в исходном плане число кружков равно числу столбцов, и лишь в ходе решения появляются новые клетки с кружками (поставками);

• в незамкнутой цепи вершинами бывают клетки без кружков (без поставок); они положительны, так как в них вносится поставка;

• характеристика незамкнутой цепи рассчитывается как алгебраическая сумма показателей ΔC_{ij} или $\overline{C_{ij}}$ в ее вершинах; так как при распределении поставок по цепи функционал увеличивается, характеристика цепи всегда положительная; она показывает, насколько увеличивается функционал, если передвинуть по цепи поставку, равную 1, из минусовой строки в плюсовую.

8.6. Модификация моделей транспортных задач

Транспортные задачи могут быть открытыми, учитывать время транспортировки продукции, затраты на производство единицы продукции, многоэтапными, многопродуктовыми. Все они, как и закрытая транспортная задача, являются частным случаем более сложной лямбда-задачи. Рассмотрим некоторые из них.

8.6.1. Открытая транспортная задача

Транспортная задача, в которой суммарная мощность поставщиков не совпадает с суммарным спросом потребителей, называется открытой. В связи с этим условия модели записываются как: $\sum a_i > \sum b_j$ или $\sum a_i < \sum b_j$. Для решения открытой транспортной задачи могут применяться методы: потенциалов, дельта-метод, МОДИ.

При решении задачи методом потенциалов или МОДИ проводятся следующие дополнительные мероприятия. Если суммарные мощности поставщиков превышают суммарные мощности потребителей, в матрицу исходных данных следует ввести дополнительный столбец – *фиктивный потребитель* (*B*) со спросом, равным небалансу: $b_{n+1} = \sum a_i - \sum b_j$. Показатели c_{ij} в столбце фиктивного потребителя должны быть *одинаковыми* по величине, которая устанавливается произвольно (любая величина, обычно проставляют 0).

Если суммарный спрос потребителей превышает суммарную мощность поставщиков, необходимо ввести в матрицу дополнительную строку – *фиктивного поставщика* (*A*), мощность которого должна быть

равна небалансу: $a_{m+1} = \sum b_j - \sum a_i$. Показатели c_{ij} этой строки должны быть *одинаковыми* и произвольными (обычно нулевые).

При составлении базисного допустимого плана и в процессе вычислительных операций в матрице должно содержаться число поставок (клеток с кружками), равное $m + n - 1$. Они должны находиться в порядке вычеркиваемой комбинации. Учитываются фиктивные строки и столбцы.

При использовании дельта-метода фиктивные поставщики или потребители не вводятся. Задача решается с нарушением баланса строк и столбцов.

8.6.2. Максимизация целевой функции

Многие экономгеографические задачи требуют максимизации функции (повышение производительности труда, прибыли и т. д.): $Z = \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} X_{ij} \to \max$. При использовании метода потенциалов, МОДИ в базисном допустимом плане поставки размещаются в клетках с *наибольшим значением* c_{ij} . Перераспределение поставок производится с учетом построения цепи к клеткам с *наибольшей положительной характеристикой*. Оптимальным будет такой план перераспределения поставок, в котором характеристики клеток без кружков будут отрицательными и нулевыми.

Решение задач дельта-методом следует начинать с распределения поставок в клетки, в которых показатели c_{ij} имеют максимальные величины. Транспортная задача с максимизацией функции может решаться по способу ее минимизации при условии придания всем c_{ij} отрицательных значений. Получив оптимальный план, необходимо рассчитать значение целевой функции, используя c_{ij} до их преобразования в отрицательные величины.

Решение транспортных задач может происходить при условиях ограничения поставок или потребления: «не меньше, чем» (обязательные поставки) и «не больше, чем»). Конечный результат решения таких задач не достигает оптимальных условий, поэтому их следует преобразовать в закрытую задачу.

8.6.3. Ограничения по времени транспортировки продукции

Ограничения в транспортную задачу вводят при учете *времени транспортировки продукции.* Для этого в искомом оптимальном плане не должны быть такие перевозки между поставщиками и потребителями, временная продолжительность которых больше заданной величины.

П р и м е р. Требуется решить задачу на минимум совокупных затрат на транспортировку продукции. Длительность перевозки не может превышать 4 ч. В матрицу *а* (табл. 8.19) вводится дополнительно время перевозки продукции (f_{ij}) и размещается в левом верхнем углу клетки.

Таблица 8.19

Учет времени перевозки продукции

а					*б*				
a_i \ b_j	**17**	**10**	**35**	**23**	a_i \ b_j	**17**	**10**	**35**	**23**
30	7 3	2 3	2 2	5 1	**30**	M	2 3	(2) **30**	M
20	3 2	2 1	3 2	2 2	**20**	3 2	2 1	3 2	(2) **20**
15	1 2	6 2	4 3	2 4	**15**	(2) **10**	M	(3) **5**	2 4
20	2 1	4 1	5 3	1 2	**20**	(1) **7**	(1) **10**	M	(2) **3**

При максимальном времени перевозки продукции 4 ч запрещаются поставки в клетки, где время указано больше: 1,1 (7 ч), 1,4 (5 ч), 3,2 (6 ч), 4,3 (5 ч). После этого задача может решаться любым методом. Ее итоговое решение (оптимальный план) представлено в табл. 8.19 *б*.

Иногда введение ограничений приводит к невозможности построить даже единственно допустимый план, который в таком случае был бы оптимальным. Значит, исходная информация противоречит условиям содержательной математической постановки задачи, которая по этой причине не имеет решения.

8.6.3. Транспортно-производственная задача

В географических исследованиях, посвященных вопросам определения границ зон сбыта продукции или рациональных связей по прикреплению потребителей к поставщикам, должны учитываться не только транспортные, но и производственные затраты. Такие задачи получили название транспортно-производственных. В качестве $c_{ij} = S_i + t_{ij}$ выступают транспортно-производственные затраты, т. е. S_i – затраты на производство единицы продукции (себестоимость, цена единицы продукции или приведенные удельные затраты) i-м поставщиком; t_{ij} – затраты на перевозку продукции между i-м поставщиком и j-м потребителем. Если увеличить или уменьшить на одну и ту же величину все показатели c_{ij} в матрице или в строке, или в столбце, то свойства матрицы не изменятся. Суммарные мощности поставщиков равны суммарному спросу потребителей. Следовательно, какой бы ни была стоимость производства,

потребители для удовлетворения своего спроса возьмут продукцию у всех поставщиков. От каких поставщиков получит каждый потребитель продукцию, зависит от транспортных затрат.

Решение открытой транспортно-производственной задачи должно учитывать показатель S_i, например себестоимость продукции. При суммарной мощности поставщиков, предположим, на 20 единиц превышающих суммарный спрос потребителей, у последних появляется свобода выбора в получении продукции от более выгодных поставщиков, поэтому оптимальный план может быть экономически более эффективным.

Модель транспортно-производственной задачи при введении дополнительных условий можно использовать для оптимизации развития и размещения промышленного производства, получить ответ, где должны располагаться новые промышленные объекты.

Для решения этих закрытых и открытых задач используются рассмотренные методы функционала, потенциала.

8.6.4. Многоэтапная транспортная задача

В современных условиях перевозка продукции от поставщика к потребителю осуществляется двумя путями: *поставщик → потребитель* (наиболее экономически выгодный) и *поставщик → база → потребитель* (требует больше транспортных и иных затрат). Поставка продукции через базу к потребителю требует построения модели многоэтапной транспортной задачи, в которой за критерий оптимальности обычно принимается минимальное значение совокупных транспортных затрат. Способ решения транспортных задач с двумя и более этапами предложен американским ученым А. Орденом. Впоследствии его назвали *способом фиктивной диагонали*.

План перевозки между поставщиками и складами и план перевозки между складами и потребителями не зависят друг от друга. Решаются две самостоятельные транспортные задачи раздельно и в любом порядке.

Если суммарная мощность складов больше суммарной мощности поставщиков, то необходимо осуществлять единый расчет, чтобы получить экономически более эффективный план многоэтапных перевозок. Рассмотрим построение матрицы в двухэтапной задаче (табл. 8.20) при следующих условиях:

$$\sum D_p > \sum A_i, \quad \sum A_i = \sum B_j.$$

При различных возможных вариантах использования емкостей складов другими могут быть варианты перевозок грузов между складами и потребителями. В матрице (см. табл. 8.22) в вектор поставщиков попадают истинные поставщики (A_i) и склады (D_p), так как склады выступают

по отношению к истинным (конечным) потребителям (B_j) как поставщики. В вектор потребителей попадают истинные потребители и склады, получающие продукцию от поставщиков. По этой причине матрица состоит из четырех блоков.

Элементами первого (I) блока матрицы (левого верхнего прямоугольника) (см. табл. 8.20) являются затраты на перевозку грузов между поставщиками и складами. Во втором (II) блоке (правом верхнем прямоугольнике) все клетки содержат запреты (З), так как поставщики передают свою продукцию сначала на склад и прямых связей с потребителями не имеют. Элементами четвертого (IV) блока (правого нижнего прямоугольника) являются затраты на перевозку грузов от складов к потребителям. В третьем (III) блоке (левом нижнем прямоугольнике) склады не поставляют продукцию складам, поэтому во всех клетках, за исключением диагональных, проставляются запреты (М). Запись поставок в фиктивную диагональ будет символизировать недоиспользованную емкость складов.

Фиктивная диагональ вводится для того, чтобы связать I и IV блоки. Суммарный размер поставок в блоках I и III по каждому столбцу равен емкости соответствующего склада. Суммарный размер поставок в блоках III и IV по каждой строке равен емкости склада.

Таблица 8.20

Форма записи исходных данных в четырехблочную матрицу

Поставщики и их мощности	Потребители и их спрос						
	D_1 50	D_2 50	D_3 50	B_1 20	B_2 25	B_3 15	B_4 30
A_1 55	**I** 7	5	4	**II** M	M	M	M
A_2 35	2	3	4	M	M	M	M
D_1 50	**III** 0	M	M	**IV** 7	5	3	5
D_2 50	M	0	M	3	4	5	6
D_3 50	M	M	0	10	9	8	7

Решение задач по блочным матрицам не отличается от алгоритма транспортных задач. Имеются лишь различия в составлении базисного плана. Его построение надо начинать с распределения поставок в одном из двух блоков I или IV. Затем следует определить, где осталась неиспользованная часть емкости складов, и записать «поставки» в соответствующие клетки фиктивной диагонали. С учетом этих «поставок» можно переходить к построению плана распределения поставок в оставшийся блок – IV или I. Требование к числу кружков, равному $m + n - 1$, расположенных в порядке вычеркиваемой комбинации, предъявляется к матрице в целом.

8.6.5. Многопродуктовая транспортная задача

Все рассмотренные транспортные задачи относятся к числу однопродуктовых. Однако иногда возникает необходимость составления базисного плана перевозок взаимозаменяемых видов продукции. Такой вопрос следует решать как единую задачу, так как в ней различные продукты могут приравниваться друг к другу через переводные коэффициенты. Решение задачи данной модели не имеет принципиальных отличий от решения закрытой однопродуктовой задачи. Существуют лишь специфические методические приемы обработки исходной информации, которые необходимо знать, чтобы подготовить матрицу для выполнения расчетов.

П р и м е р. Потребителю необходимо поставить взаимозаменяемое топливо: торф, бурый уголь. Необходимое условие: суммарная потребность в торфе и буром угле, выраженная в единицах условного топлива, будет полностью удовлетворена. Известно, что 1 т условного топлива равна 7000 ккал, 1 т торфа – 2800 ккал, 1 т бурого угля – 4200 ккал. Отсюда переводной коэффициент по теплотворной способности топлива (калорийный эквивалент) для торфа равен 2800 / 7000 = 0,4, для бурого угля – 0,6.

В табл. 8.21 представлены мощности и спросы по торфу в тоннах и показан оптимальный план перевозки с функционалом равным 13 980 (F_1), в табл. 8.22 представлены эти же данные по бурому углю с функционалом 10 620 (F_2). По двум планам объем грузооборота равен $F_1 + F_2$ = 24 600 т/км. У поставщиков A_1 и A_2 имеется торф и бурый уголь, у поставщика A_3 – только торф, у поставщика A_4 – только бурый уголь. В обеих таблицах расстояния между поставщиками A_1 и A_2 и потребителями одинаковые, так как оба вида топлива будут перевозиться по одним и тем же транспортным путям.

Таблица 8.21

Мощности и спрос по торфу

b_j / a_i	B_1 100	B_2 180	B_3 120
A_1 150	**12** **100**	72	**60** **50**
A_2 75	48	**24** **5**	**48** **70**
A_3 175	72	**36** **175**	60

Используя коэффициенты теплотворной способности торфа (0,4) и бурого угля (0,6) и данные A_i, B_j, x_{ij}, c_{ij} табл. 8.21, 8.22, производим перерасчет и составляем табл. 8.23, в которой данные указаны в условных (перерасчетных) единицах. Приводим ниже пояснения связанные с перерасчетом.

Таблица 8.22

Мощности и спрос по бурому углю

a_i \ b_j	B_1 60	B_2 210	B_3 125
A_1 130	**12** **60**	72	**60** **70**
A_2 100	48	**24** **45**	**48** **55**
A_4 165	36	**12** **165**	72

1. Расчет спроса потребителей в условных единицах (у. е.) проведем на примере B_1 (см. табл. 8.21, 8.22). Спрос потребителя B_1 на торф равен 100 т, на бурый уголь – 60 т. Используя переводные коэффициенты, рассчитываем его потребность в условном топливе:

$$B_1 = 100 \cdot 0{,}4 + 60 \cdot 0{,}6 = 76.$$

2. Мощность поставщиков (A_i) в условных единицах дается отдельно для каждого вида топлива, потому что она выступает как ограничение на возможный размер поставок k-го вида продукта, который находится у i-го поставщика. Ее получают путем умножения величины мощности поставщика на переводной коэффициент по торфу и по бурому углю отдельно:

$A_1 = 150 \cdot 0{,}4 = 60$ (по торфу); $A_1 = 130 \cdot 0{,}6 = 78$ (по бурому углю).

3. Показатели расстояний (c_{ij} – *правый верхний угол в клетках матриц, полужирный шрифт*) в условных единицах получают путем деления их на переводные коэффициенты:

$c_{11} = 12 / 0{,}4 = 30$ (по торфу); $c_{11} = 12 / 0{,}6 = 20$ (по бурому углю).

После перевода всех показателей матрицы в условные единицы, как показано в пунктах 1–3, получаем новую матрицу, в которой проводим перераспределение условных единиц до получения оптимального варианта (табл. 8.23).

Таблица 8.23

Оптимальный вариант распределения поставок в условных единицах

a_i, у. е. \ b_j, у. е.			B_1 76	B_2 198	B_3 123
A_1	торф	60	**30** **60**	150	120
	бурый уголь	78	**20** **16**	100	**80** **62**
A_2	торф	30	150	**60** **29**	**90** **1**
	бурый уголь	60	100	40	**60** **60**
A_3	торф	70	180	**90** **70**	150
A_4	бурый уголь	99	60	**20** **99**	100

Функционал оптимального плана, выраженный в условных единицах в табл. 8.23, составляет 20790. По сравнению с суммарным потенциалом предыдущих таблиц по торфу и бурому углю, объем транспортной работы в последнем варианте с условными единицами снизился на 3810 единиц, что дает экономию по объему грузооборота более, чем на 15 %.

8.6.6. Лямбда-задача

Алгоритм транспортных задач по методам решения значительно проще, чем лямбда-задача. Ее называют распределительная или обобщенная транспортная задача. В ее модели отражается более широкий круг практических задач, богатых по содержанию. Способ ее решения предложили американские математики А. Фергюссон, Дж. Данциг (1955). Позже ее разрабатывали российские ученые А. Л. Брудно, У. Х. Малков, А. Г. Аганбегян и др.

Алгоритм У. Х. Малкова строго формализован и реализован в машинных программах и дает возможность преодолеть трудности решения лямбда-задач, но очень сложный. А. Г. Аганбегян предложил операторскую схему дельта-метода решения лямбда-задач, которую можно использовать для расчетов вручную. Основные принципы этого метода изложены выше в п. 8.5 применительно к транспортной задаче. Решение лямбда-задачи дельта-методом также сложно. В ходе вычислительных операций возможно частое допущение ошибок, поэтому итоговое решение следует проверять. Лямбда-задача открытая, и в ходе ее решения всегда останется хотя бы одна избыточная (плюсовая) строка. Потенциал (оценка) этой строки принимается равным нулю, а расчет потенциалов других строк и столбцов, характеристик клеток без кружков осуществляется по следующим формулам:

$$u_i = \overline{\lambda}_{ij}\ (v_j\ - \overline{c}_{ij});$$

$$v_j = \overline{c}_{ij}\ + u_i / \ \overline{\lambda}_{ij};$$

$$\overline{c}_{ij} = v_j - u_i / \overline{\lambda}_{ij};$$

$$E_{ij} = c_{ij} - (v_j - \ u_i / \overline{\lambda}_{ij}).$$

Показатель $\overline{\lambda}_{ij}$ размещается в левом верхнем углу клеток матрицы и выполняет важную роль в оптимизации условий задачи, включается в формулу при расчете функционала: $Z = \sum c_{ij} \cdot \overline{\lambda}_{ij} \cdot x_{ij} \to \min\ (\max)$. Для той минусовой строки, которая на последнем шаге вычислительных операций превратилась в нулевую, рекомендуется взять потенциал, равный приросту затрат на последнем шаге.

Глава 9

МЕТОДЫ ТЕОРИИ ГРАФОВ

Основоположником теории графов является швейцарский математик Л. Эйлер (1736). В дальнейшем теории графов уделялось малое внимание. Начало бурного развития и практического применения теории графов было положено венгерским математиком Д. Кенигом, который опубликовал в 1936 г. монографию «Теория конечных и бесконечных графов». Российский академик Л. В. Канторович разработал метод решения транспортных задач для их сетевой постановки. В настоящее время имеется большое количество работ по теории графов, включая прикладное направление.

Теория графов используется в исследованиях по экономической географии *с целью глубокого познания внутренних взаимосвязей в пространственных структурах и закономерностей их развития.*

Простые и лаконичные формы графов неразрывно связаны с глубинной сущностью отображаемых явлений и процессов, позволяют вскрыть неточности, допущенные в ходе теоретических построений. Их можно использовать в классификации объектов.

9.1. Элементы теории графов

Фигура, состоящая из точек (вершин) и соединяющих их линий (ребер), называется *графом* (рис. 9.1). Вершины A •—• B называются *смежными* (*связанными*). Граф называется *связным*, если любая пара его вершин связана. Граф может состоять только из вершин (нуль-граф). Расположению вершин, длине и форме ребер или дуг не придается значения. Существенно лишь то, какие вершины соединены. Ребра (дуги) графов указывают на соответствие между вершинами в графе. Граф может быть представлен геометрически в виде определенной фигуры или в виде матрицы, в которой для каждой вершины записывается число связанных с ней ребер (дуг).

Нумерованные кружки (см. рис. 9.1) в графах служат его *вершинами*, которые соединены *ребрами* – неориентированными линиями (*h, i*). Вершина называется *четной*, если в ней сходится четное число ребер, и *нечетной*, если число всех сходящихся в ней ребер нечетное.

Ориентированное ребро называют *дугой* (*a,e, f, g, j* – ↑), которая может быть входящей в вершину (1 – *g*) и выходящей из нее (1 – *a, e, f*). Ребра могут быть *инцидентны* вершине, если они являются одним из ее концов, а вершина – инцидентна каждому из входящих в нее ребер.

Каждое ребро (дуга) может соединять только две вершины. Если ребро соединяет вершину с ней же самой, то его называют *петлей* (*b, c, d, k*). Она имеет овальную форму (0). Это цикл (контур) единичной длины, т. е. образованный одним ребром (дугой), связывающим вершину саму с собой.

Вершины *3* и *5* изолированы, так как они не имеют ни одного инцидентного им ребра (дуги). Ни одно ребро не соединяет такую вершину с другой. Вершину *3* можно назвать голой, желая подчеркнуть, что при ней нет даже петель, как в вершине *5*. Рассмотренный граф содержит конечное множество вершин, но бесконечное множество (континуум) ребер (дуг).

Маршрут представлен в ориентированном графе *путем*, в неориентированном – *цепью* (∟), если каждое ребро графа встречается в нем не более одного раза. Вершины и цепи могут повторяться несколько раз.

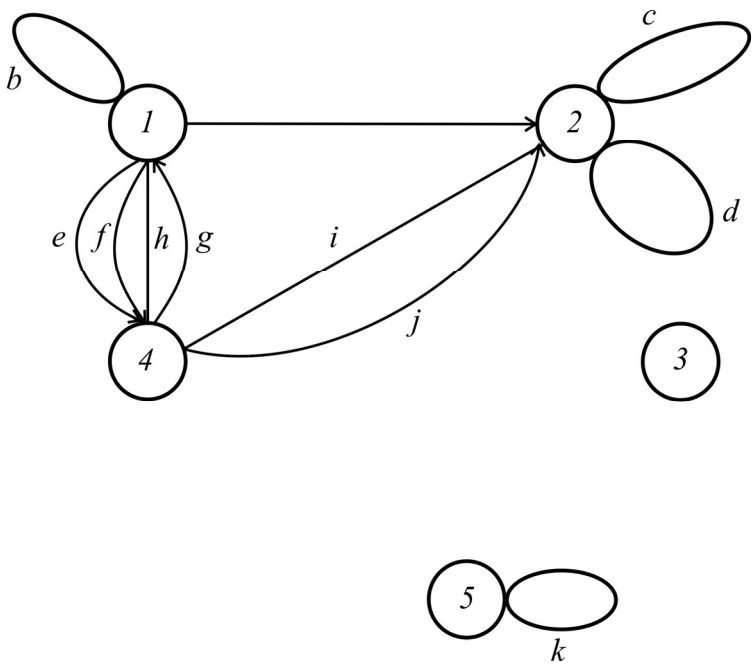


Рис. 9.1. Элементы теории графов

Цепь, начальная и конечная вершины которой совпадают, называется *контуром* (ориентированный граф) и *циклом* (⌂) (неориентированный граф). Они имеют форму треугольника, многоугольника.

Элементарные пути, цепи, циклы и контуры называют *гамильтоновыми*, простые – *эйлеровыми.* В *элементарные* формы графов вершины не включаются более одного раза, в *простые* – дуги (ребра) не включаются более одного раза. Длина цепи (пути) или цикла (контура) есть число ребер (дуг), которые их образуют.

Число ребер, сходящихся в вершине графа, называется *степенью* (*порядком*) s (G, x) *вершины* x в графе $G = (X, U)$ или *числом ребер, инцидентных этой вершине*. При *изоморфизме* двух графов соответствующие друг другу вершины должны иметь одинаковую степень вершин. Упорядоченную систему степени его вершин называют вектором степеней графа G и кратко обозначают s (G).

Обыкновенным графом $G = (X, U)$ называется упорядоченная пара множеств: конечного непустого X, элементы которого называют вершинами графа G, и подмножества $U \subseteq \overline{X}$, элементы которого называются ребрами этого графа. Граф называется *конечным*, если множество его ребер конечно. Граф интерпретируется как *сеть*, а его вершины – *узлы.* Если линии, соединяющие вершины, не имеют ориентации, то граф называется *неориентированным* (рис. 9.2, *а*), при наличии стрелок на линиях граф считают *ориентированным, или орграфом* (см. рис. 9.2, *б, в*). Может быть и *смешанный* граф.

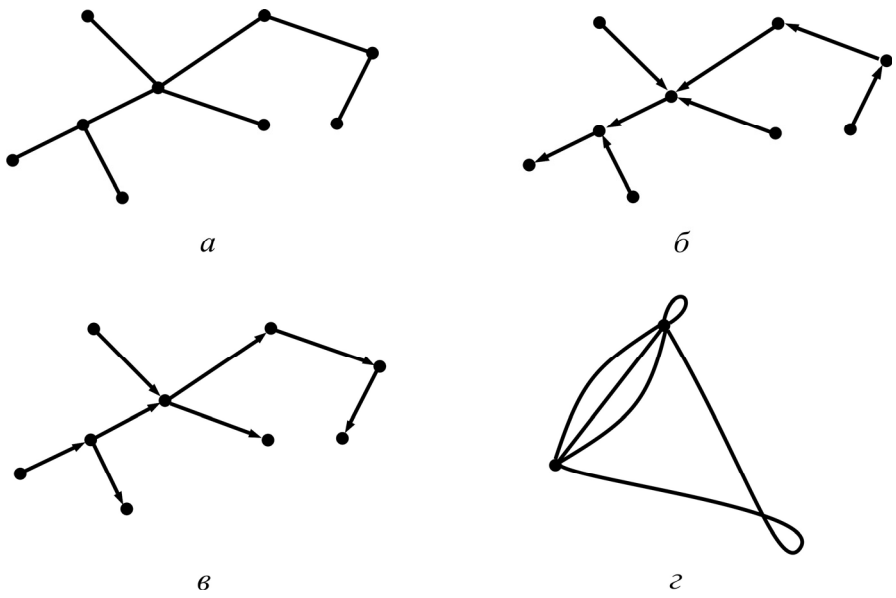


Рис. 9.2. Виды графов:
а – неориентированный граф-дерево; *б* – входящее дерево;
в – исходящее дерево; *г* – псевдограф

Псевдограф содержит петли и кратные ребра (рис. 9.2, *г*).

Важный класс графов составляют *«деревья».* Это связный граф, который имеет не менее двух вершин и не имеет циклов (см. рис. 9.2). Ребра графа-дерева называют *ветвями.* Дерево, все ветви которого имеют общую вершину, называют *лагранжевым деревом.* Корнем дерева может быть любая вершина, которую выбирают за начальную точку.

Среди ориентированных деревьев различают *входящее* и *выходящее* дерево (см. рис. 9.2, *б, в*). Входящее дерево может быть моделью производственной системы, показывающей, что при изготовлении одного конечного продукта используется несколько видов промежуточной продукции, получаемой из различных видов сырья. Выходящее дерево может быть моделью пространственной системы производства, где за начальную точку принимается стадия добычи комплексного сырья, при переработке которого получают несколько конечных продуктов. *Лесом* называется несвязный граф, все связные компоненты которого являются деревьями.

Сумма степеней всех вершин неориентированного графа является четным числом, так как каждое ребро соединяет две вершины. Следовательно, число ребер m в графе $G(X)$ равно половине суммы степеней всех его вершин: $m = 1/2 \sum_{i=1}^{n} p(x_i)$, где i – индекс вершины, n – число вершин. Эта формула справедлива в случае наличия петель, если они рассматриваются как двойные ребра.

Граф $G(X)$ называется однородным, если степень всех его вершин одинакова. Понятие «*однородный граф степени r*» означает, что каждая вершина данного графа имеет степень, равную r. В однородных графах степени r число ребер равно: $m = (1/2)\, n \cdot r$. Примером однородных графов являются правильные многогранники: тетраэдр, куб, октаэдр.

Под *цикломатическим числом* понимается число *независимых* циклов графа. К *независимым* относятся циклы, не имеющие ни одного общего ребра с другими циклами графа. К *зависимым* относятся циклы, у которых имеются общие ребра.

Матрица графов строится следующим образом. Ряды и столбцы матрицы представлены вершинами графа. В каждый рядок или столбец вносится количество инцидентных ребер для каждой вершины или кратчайших расстояний между вершинами и т. д. Затем производятся соответствующие расчеты степеней, индексов.

9.2. Топологический анализ сетей

Теория графов позволяет исследовать топологический анализ транспортных и экономико-географических сетей: *доступность, связность, форму и структуру*. Имеется ряд показателей, описывающих эти сети. Их называют *мерами* (количественные показатели, характеризующие явление или процесс).

Показатели доступности. Построение графа, моделирующего доступность транспортной сети, заключается в следующем. Все точки пересечения дорог принимаются за *фиктивные* вершины. Это понятие условное, поскольку эти точки правомерно считать вершинами, как и населенные пункты. Содержательная интерпретация их может быть представлена различием условных обозначений в графе: ● – фактическая вершина (населенный пункт); ○ – фиктивная вершина точек пересечения дорог. Использование фиктивных ребер приводит к существенным погрешностям.

Меры доступности применяются для оценки транспортно-географического положения. К ним относят *число Кенига, индекс оптимальной связности вершин, индекс центральности Бавелаша, Бошама.*

Если граф небольшой, эти меры можно рассчитать по графическому изображению. Для большого графа строится матрица и расчетные операции производят с помощью матричной алгебры.

Пусть будет следующий граф с населенными пунктами (рис. 9.3). К нему составим матрицу кратчайших расстояний – по количеству инцидентных ребер, соединяющих вершины (табл. 9.1), вычислим меры доступности и дадим оценку оптимальной связности вершин.

Четыре индекса (*Si, Ki*, *Ba*, *Bi*) в табл. 9.1 характеризуют степень доступности вершин. Они претендуют на центральное положение в графе. Индексы *Si, Ki* – абсолютные, *Ba*, *Bi* – относительные. Абсолютный индекс доступности *Si* рассчитывается как сумма инцидентных ребер к каждой вершине по строкам: $Si_i = \sum x_i$. Для первой строки матрицы эта сумма равна: 0 + 1 + 2 + 3+ + 3 + 2 + 4 + 3 = 18. По абсолютным индексам центральное положение занимают объекты с наименьшими их величинами, т. е. вершины *2*, *3*, *4*, *6* с индексом *Si*, равным 12 (в матрице все индексы центрального положения и сами вершины выделены полужирным курсивом).

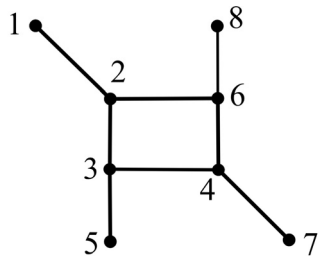


Рис. 9.3. Граф для оценки местоположения объектов

Таблица 9.1

Матрица кратчайших расстояний между вершинами и индексы доступности вершин

№ вершин	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Si*	*Ki*	*Ba*	*Bi*
1	0	1	2	3	3	2	4	3	18	4	6,66	0,38
2	1	0	1	2	2	1	3	2	***12***	***3***	***10,0***	***0,58***
3	2	1	0	1	1	2	2	3	***12***	***3***	***10,0***	***0,58***
4	3	2	1	0	2	1	1	2	***12***	***3***	***10,0***	***0,58***
5	3	2	1	2	0	3	3	4	18	4	6,66	0,38
6	2	1	2	1	3	0	2	1	***12***	***3***	***10,0***	***0,58***
7	4	3	2	1	3	2	0	3	18	4	6,66	0,38
8	3	2	3	2	4	1	3	0	18	4	6,66	0,38
									$\sum = 120$			

Второй абсолютный индекс число Кенига (*Ki*) – это наибольший по величине элемент (x_i) строки матрицы: $Ki_i = x_i$ (max). Его минимальное значение, равное трем, для вершин *2*, *3*, *4*, *6* указывает на их центральное положение (степень их доступности).

Индекс Бавелаша (*Ba*) определяется как отношение суммарного значения индекса $\sum S_i$ к величине индекса Si_i каждой строки: $\sum S_i$ / Si_i. Максимальное значение индекса Бавелаша указывает на высокую степень доступности вершин. Ими являются вершины *2*, *3*, *4*, *6*, имеющие *Ba* = 10.

Относительный индекс Бошама (*Bi*) рассчитывается по следующей формуле: $Bi = (n - 1) / Si_i$, где *n* – общее число вершин на транспортной сети без одной (в нашем случае 8 – 1 = 7). Величину 7 делят на значение Si_i каждой строки. Максимальное значение индекса Бошама (0,58) для вершин *2*, *3*, *4*, *6*, который определяет их центральное положение.

Таким образом, как абсолютные, так и относительные индексы указали на центральное положение второй, третьей, четвертой, шестой вершин графа и матрицы.

Показатели связности. К мерам связности относятся следующие топологические параметры: α-, β-, γ-индексы. Индексы принимают наибольшие значения в случаях насыщения сети контактами.

Индекс α представляет собой отношение цикломатического числа графа ($m - n + \acute{\imath}$) к максимально возможному числу циклов в этом графе ($2n - 5$):

$$\alpha = (m - n + \acute{\imath}) / (2n - 5),$$

где *m* – число ребер, *n* – число вершин, *í* – число связных компонент графа; для связного графа цикломатическое число будет $m - n + 1$.

Индекс α характеризует избыточность связей в сетке. Его значения варьируют в пределах от 0 до 1, при умножении на 100 – в процентах. Избыточность связей можно оценить по цикломатическому числу, но его нельзя использовать для сравнения связей в различных сетях.

Индекс β представляет собой отношение числа ребер *m* сети к числу ее вершин *n*. Чем больше ребер связывают одно и то же число вершин, тем больше циклов в сети, тем сложнее ее структура и выше связность. Значения индекса колеблются в пределах от 0 до 3. В несвязных графах и деревьях величина индекса меньше единицы. При значении β = 1 граф имеет только один цикл, при изменении от 1 до 3 графы имеют более одного цикла.

Индекс γ представляет собой отношение числа ребер *m* к их максимально возможному количеству в сети, которое в плоских графах равно $3 \cdot (n - 2)$, где *n* – число вершин. Величина индекса в графе колеблется от 0 до 1. Он характеризует полноту связей в цепи.

Показатели формы графа. Меры формы сетей связаны с определением топологического диаметра графа. Диаметр графа (δ) представляет собой топологическую длину, которая равна числу ребер в кратчайшей цепи, соединяющей две самые отдаленные друг от друга вершины (рис. 9.4, *а*). Если на ребрах указаны конкретные расстояния, то такие помеченные модели графа более содержательны (рис. 9.4, *б*).

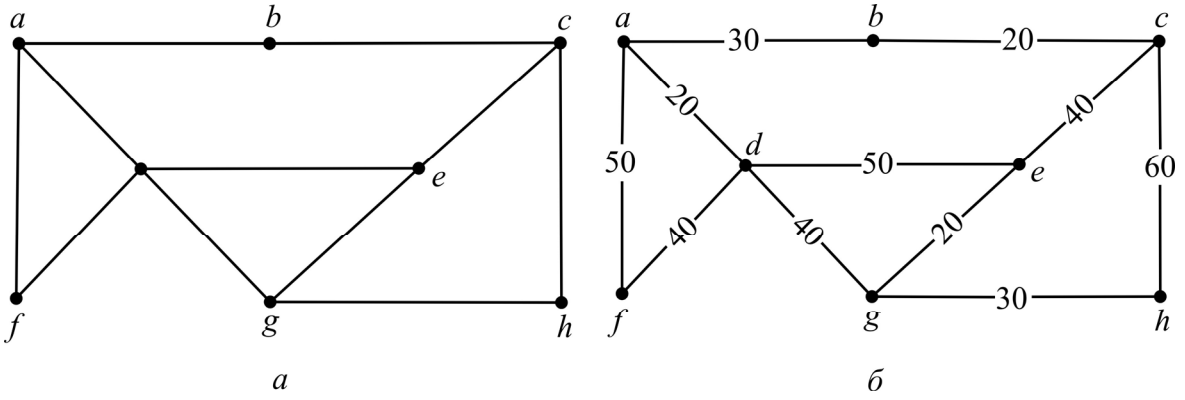


Рис. 9.4. Меры формы сетей:
а – с учетом числа ребер (δ = 6; $\pi^{(r)}$ = 1,33); *б* – с учетом реальных расстояний

Топологическая мера формы ($\pi^{(r)}$) с учетом общего числа ребер в графе (*m*) и его диаметра (δ) – топологической длины – определяется по формуле: $\pi^{(r)} = m / \delta = 11 / 3 = 3{,}6$ (см. рис. 9.4, *а*). В этом графе восемь топологических диаметров, равных 3, связывающих различные пары вершин. По мере увеличения числа ребер в сети улучшаются связи между ее вершинами, топологический диаметр уменьшается, значение меры

формы увеличивается. Это означает, что улучшается форма сети. Она становится более компактной.

Для графов, в которых указано расстояние в определенных единицах измерения, мера формы определяется таким же способом, как и при учете количества ребер, но с учетом протяженности всей сети графа в километрах (Д) и длины топологического диаметра в километрах ($Т_д$):

$$\pi = Д / Т_д.$$

Если топологических диаметров в графе несколько, их реальная длина в километрах будет различной. На рис. 9.4, *б* также восемь топологических диаметров δ, равных 3. На рис. 9.5 эти диаметры выделены жирной линией, а реальные длины топологических диаметров различны. Для таких графов топологическая мера формы π определяется с учетом средней длины топологического диаметра Т. Последний определяется по формуле: $Т = \sum Т_д / p$, где p – число топологических диаметров сети.

Топологические диаметры выделены жирными линиями на рис. 9.5. Выделено три диаметра по 130 км, два – по 140, два – по 130, один – 160 км. Средний диаметр $Т_д$ = 128,75. Общая протяженность этой сети 500 км. Отсюда индекс π = 500 / 128,75 = 3,88. Результаты показывают, что индекс для оценки меры формы, полученный с учетом длины ребер в километрах, более точный по сравнению с индексом $\pi^{(r)}$.

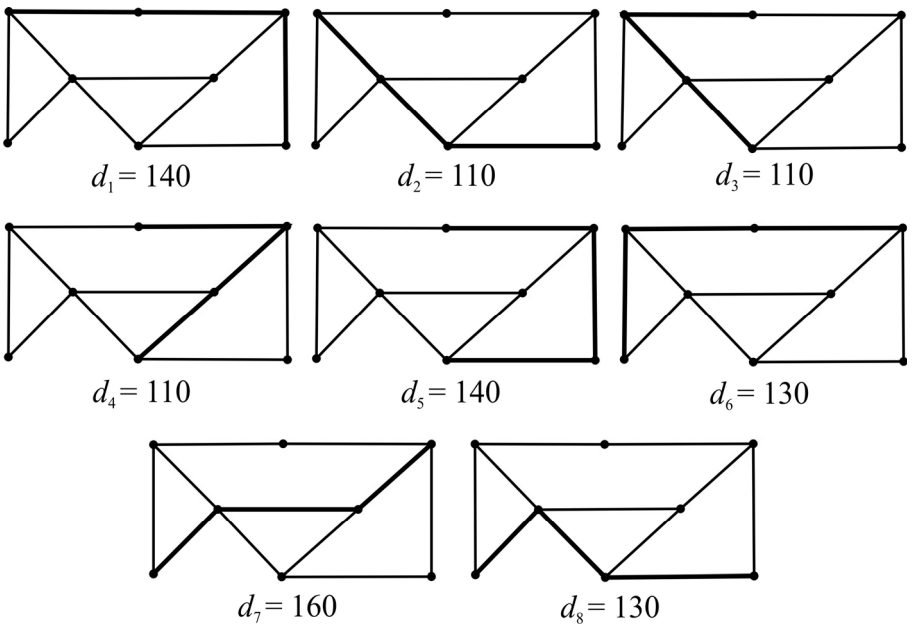


Рис. 9.5. Варианты топологических диаметров графа

Высокие значения π-индекса указывают на компактную территорию, охватываемую графом.

Структурные параметры сетей. Выявления параметров территориальных структур основываются на топологических параметрах, среди которых следует отметить меры *интеграции, униполярности и централизации.* Эти меры основаны на суммах расстояний.

Интегрирование в обществе, промышленности, политике в современных условиях играет существенную роль. Для географических сетей важно выявить степень интеграции. Предложен следующий расчет меры интеграции: $S = 1/2 \sum Si_i$. Сеть следует считать интегрированной, если все ее вершины имеют приблизительно равные значения мер интеграции. Параметр интеграции характеризует меру центральности на множестве вершин.

Униполярность указывает на наличие вершины в графе, которая изолирована от других вершин, т. е. характеризуется минимальным значением индекса оптимальной связности *Si.* Параметр униполярности относится к вершине, имеющей минимальное значение связности: $V = Si_i \min$.

Иногда в сети встречаются группы вершин, которые резко отличаются между собой по величине индекса оптимальной связности. Такую сеть можно рассматривать как централизованную. Меру централизации можно рассчитать следующим образом: $H = \sum (Si_i - S_{i\min})$ или $H = 2S - nV$.

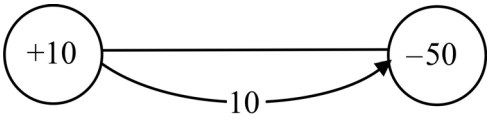
Рассмотренные меры, характеризующие структурные параметры, дают возможность выявить особенности количественной и качественной структуры сетей, отличающихся закономерностями формирования, развития и функционирования.

9.3. Сетевые постановки транспортных задач

Транспортные задачи, рассмотренные в теме по линейному программированию, можно решать с использованием методов теории графов. Они имеют преимущества, так как в ходе поисков оптимального плана поставок одновременно выбираются наиболее рациональные пути их перевозок. В сетевой постановке транспортных задач существуют непосредственные связи между пунктами и отсутствуют косвенные связи.

Сетевая постановка закрытой транспортной задачи. Граф должен представлять ориентированное дерево. Построение его можно начинать с любой вершины. Если начальная вершина положительная (+), то продукция из нее вывозится и она является началом дуги (стрелки), которая должна входить в одну из смежных вершин. При построении графа с отрицательной вершины (–), в которую ввозится продукция, она должна быть концом дуги (стрелки), выходящей из любой смежной вершины.

Стрелка вдоль ребра символизирует превращение его в дугу. На стрелке указывается величина перемещаемой продукции:



Составление базисного плана. В качестве начальной вершины выбрана положительная вершина *a* – поставщик, имеющий 70 единиц продукции (рис. 9.6). От нее направлены две стрелки: а) в отрицательную вершину *d*, которой требуется 50 единиц продукции; б) в вершину *f* (–20) перемещаем 20 единиц продукции.

Исчерпав возможности поставщика *a*, переходим к распределению продукции поставщика *b* (+90). Из вершины *b* отправляем в вершину *c* всю продукцию (90 ед.). В ней оставляем (вершине *c*) необходимые 55 единиц, а лишнюю часть (35 ед.) передаем в вершину *h*, куда необходимо поставить 75 единиц продукции. Недостающую продукцию 40 единиц должны получить от соседнего поставщика *g* , который имеет 15 единиц. Этого недостаточно, поэтому вершину *h* пополняем недостающими единицами из вершины *e* (+25), пройдя через вершину *g*. В результате проведенных операций все поставщики отправили продукцию всем потребителям.

Однако мы получили пока несвязный граф (имеется пробел между стрелками). Он состоит из двух связных компонент. Число стрелок в графе (рис. 9.6) равно 6, а должно быть 7, так как в графе 8 вершин. Следовательно, на одном из ребер ставим стрелку любого направления с нулевой поставкой продукции, чтобы образовать контур. Для этого подходит соединение вершин *a* – *b*, а также *d* – *e* и не подходит соединение *c* – *e*, *d* – *f*, чтобы не образовать контуры.

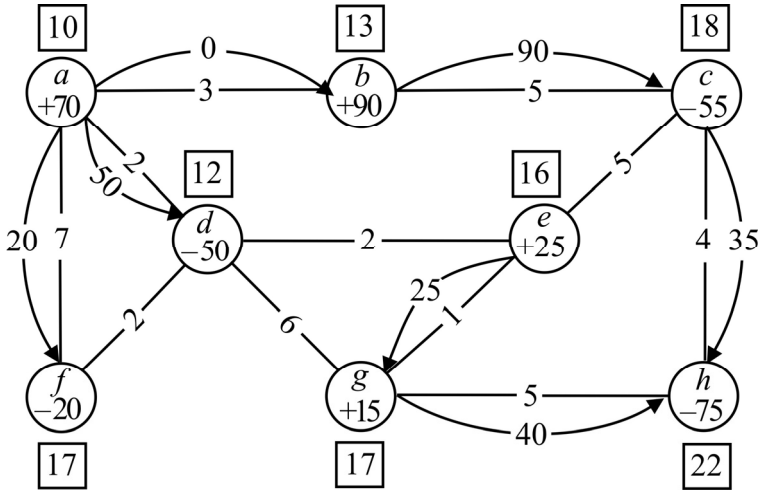


Рис. 9.6. Базисный план закрытой транспортной задачи в сетевой постановке

При базисном распределении поставок, соблюдая правила, надо стремиться к размещению стрелок на ребрах с меньшими значениями c_{ij} (они указаны посередине ребра), что трудно сделать в задачах большой размерности.

Допустимый план на оптимальность проверяется методом потенциалов. Для этого вначале любой вершине присваивается любая величина потенциала (λ). Однако его величина должна быть большая по сравнению с c_{ij} ребер графа, чтобы не получить отрицательных потенциалов вершин и не усложнять работу.

В вершине *a* потенциал устанавливаем равным $\lambda_a = 10$. Двигаясь по дугам со стрелками, вычисляем потенциалы других вершин сети с учетом направления стрелок. Если стрелка выходит из вершины, то к ее потенциалу прибавляется величина c_{ij} ребра; если стрелка входит в вершину, то с ее потенциала вычитается c_{ij} соответствующего ребра. Потенциалы вершин (λ) заключаем в прямоугольник у вершины.

Функционал можно рассчитывать с использованием c_{ij} или λ:

$$Z = \sum(\lambda_i \cdot x_{ij}) = 10 \cdot 70 + 13 \cdot 90 + 18 \cdot (-55) + 12 \cdot (-50) + 16 \cdot 25 + 17 \cdot (-20) + + 17 \cdot 15 + 22 \cdot (-75) = -1055.$$

$$Z = \sum(c_{ij} \cdot x_{ij}) = 3 \cdot 0 + 90 \cdot 5 + 7 \cdot 20 + 2 \cdot 50 + 1 \cdot 25 + 4 \cdot 35 + 5 \cdot 40 = 1055.$$

Величины функционалов получены одинаковые, но с противоположными знаками. Следует проверить план на оптимальность.

Базисный план на оптимальность проверяется путем расчета характеристики (E_{ij}) ребер, не имеющих дуг (стрелок) (см. рис. 9.6). Любому ребру сетки соответствует два потенциала вершин, которые им соединяются. Следует из большего потенциала вершины вычесть меньший потенциал и полученную разность вычесть из c_{ij} ребра, например:

$$E_{df} = c_{df} - (\lambda_f - \lambda_d) = 2 - (17 - 12) = -3.$$

В нашем графе план неоптимальный, так как имеет два ребра с отрицательными характеристиками E_{ij}: ребро $d - e$ (–2) и $d - f$ (–3). Поэтому производим перераспределение поставок. Для этого на ребро с наибольшей отрицательной $E_{df} = -3$ ставится стрелка, которая имеет направление от вершины с меньшим потенциалом к вершине с большим потенциалом $d \rightarrow f$.

Размер поставки на новой стрелке зависит от следующих обстоятельств. Новая стрелка привела к образованию псевдоконтура $a \rightarrow d \rightarrow f \leftarrow a$, что требует изъятия одной стрелки для соблюдения правила: число вершин минус единица равно числу стрелок в графе. Кроме того, следует разрушить псевдоконтур, которого не должно быть в графе. В возникшем псевдоконтуре выбираем стрелку противоположного направления новой стрелке $d \rightarrow f$. Выбираем, если есть несколько, ту стрелку, которая имела наименьшую величину поставки ($af = 20$) до появления новой стрелки $d \rightarrow f$.

Минимальную поставку (20) перераспределяем по псевдоконтуру следующим образом: она прибавляется на стрелках, имеющих одинаковое направление с новой стрелкой, и вычитается из поставок на стрелках, которые имеют противоположное направление новой стрелке.

В рассматриваемом псевдоконтуре стрелка $a \rightarrow d$ имеет одинаковое направление с новой $d \rightarrow f$. На $a \rightarrow d$ прибавляется поставка 20 к прежней 50, и получается новая поставка 70. Излишек поставки 20, образовавшийся в вершине d, передается вершине f по новой стрелке $d \rightarrow f$. Прежняя поставка 20 между вершинами $a \rightarrow f$ убирается вместе со стрелкой (рис. 9.7). В результате перемещения минимальной поставки 20 по псевдоконтуру ликвидирован сам контур, потребители получили необходимую продукцию.

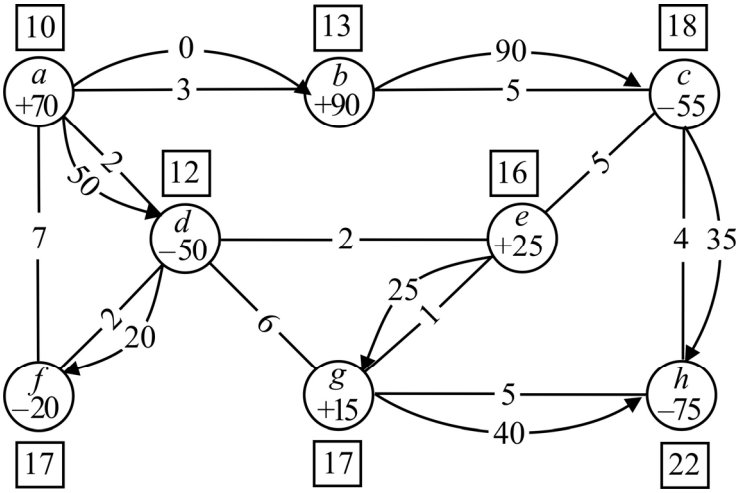


Рис. 9.7. Первое перераспределение поставки (20) по псевдоконтуру

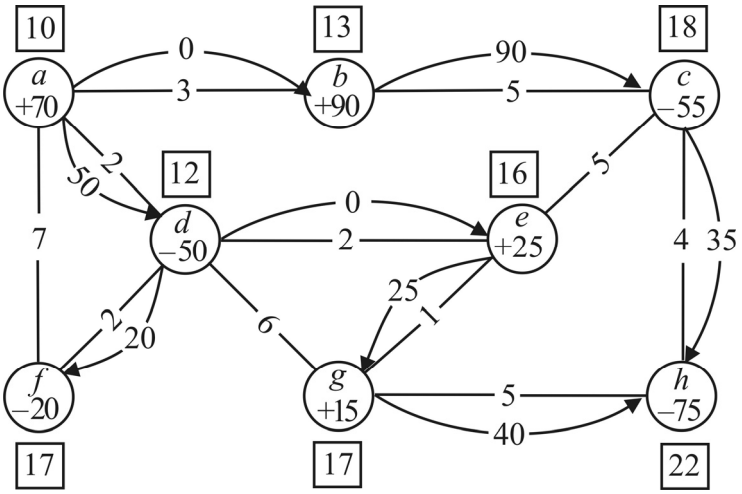


Рис. 9.8. Второе перераспределение поставки (0) по псевдоконтуру

Новый план стал более оптимален, так как величина функционала уменьшилась на 100 (см. рис. 9.7). Однако при расчете потенциалов вершин и характеристики ребер получена отрицательная величина –2 между вершинами d – e. Значит, необходимо произвести новое перераспределение продукции.

Для этого на ребро с отрицательной характеристикой ставим стрелку $d \to e$. Образуется псевдоконтур $a \to d \to e \to g \to h \leftarrow c \leftarrow b \leftarrow a$, в котором одна из встречных стрелок $a \to b$ имеет минимальную поставку, равную нулю. Ее перераспределяем по псевдоконтуру, как описано выше, и получаем новый допустимый план (рис. 9.8).

Расчет потенциалов вершин и характеристики ребер без стрелок показывает, что ребра не имеют отрицательных характеристик. Таким образом, получен оптимальный вариант без изменения функционала, так как перемещалась нулевая поставка.

9.4. Сетевая постановка открытой транспортной задачи

Открытая транспортная задача решается аналогично закрытой задаче. Ее необходимо превратить в закрытую путем введения в граф (сеть) дополнительной вершины (фиктивного потребителя или фиктивного поставщика) с величиной спроса, равной небалансу. Фиктивный поставщик или потребитель должен быть соединен ребрами с одинаковыми и высокими значениями c_{ij} со всеми поставщиками (потребителями).

Ребрам, инцидентным фиктивной вершине, присваиваются высокие значения c_{ij}. Это необходимо, чтобы априори сделать неэффективным

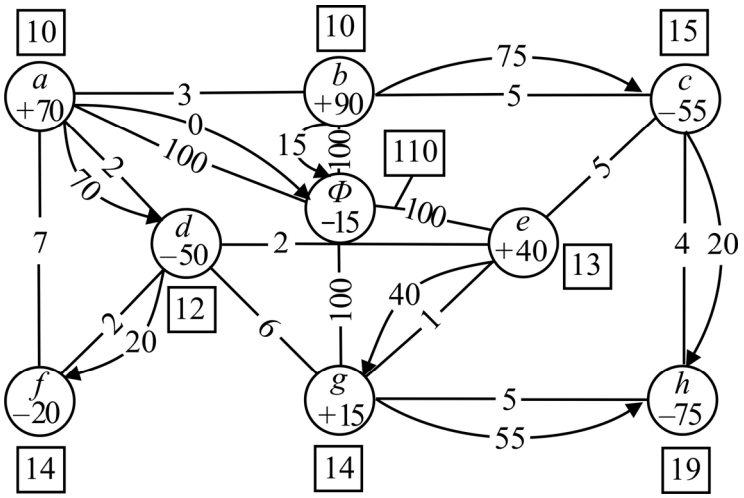


Рис. 9.9. Открытая транспортная задача

использование фиктивной вершины (Ф = –15) как промежуточного пункта, поскольку в реальной сети его нет (рис. 9.9).

В примере открытой задачи (см. рис. 9.9) суммарная мощность поставщиков превышает суммарный спрос потребителей на 15 единиц продукции (Ф = –15). Фиктивный потребитель в вершине Ф соединен со всеми поставщиками ребрами, показатели c_{ij} которых равны 100. При расчете функционала не учитывается фиктивная вершина. Необходимо из показателя мощности вершины (b_i), из которой исходит стрелка к фиктивной вершине, сначала вычесть величину символической поставки этой стрелки (15), а затем эту разность умножить на потенциал реально существующей вершины ($\lambda_b = 10$). Функционал оптимального плана:

$$F = 10 \cdot 70 + 10 \cdot (90 - 15) + 15 \cdot (-55) + 12 \cdot (-50) + 13 \cdot 40 + \\ + 14 \cdot (-20) + + 14 \cdot 15 + 19 \cdot (-75) = -950.$$

9.5. Транспортно-производственная задача

Задачи с учетом производственных затрат могут быть открытыми и закрытыми. При их решении имеются некоторые отличия от транспортных задач в матричной постановке. В сетевой постановке нельзя прибавлять затраты на единицу продукции к транспортным затратам, так как заранее неизвестно, показатели производственных затрат какого поставщика и к каким ребрам необходимо прибавлять.

Для представления ситуации в сетевой постановке используем рис. 9.8 как исходную основу. Пусть поставщик на вершине *c* имеет производственные затраты 35 единиц, а поставщик на вершине *d* – 45 единиц. На графе (см. рис.9.8) каждая положительная вершина заменяется нулевой мощностью и к ней добавляется ребро под названием «ус», на конце которого ставится поставка этой нулевой вершины, а на ребре – производственная затрата.

На рис. 9.10 отражены «усы», производственные затраты на ребрах и мощности нулевых вершин на конце «усов». Место вершины *c* заняла *нулевая вершина* ***r*,** а вершины *d* – *нулевая вершина* ***u***. К нулевым вершинам добавлены тупиковые отрезки (усы) с производственными затратами на ребрах, равными 35 и 45. На концах усов величины поставок 120 и 60.

Способ решения транспортно-производственной задачи на сети тот же, что и транспортно-производственной задачи матрицы.

При решении транспортно-производственной задачи открытой модели итоговые оптимальные планы по сети и матрице могут значительно отличаться друг от друга.

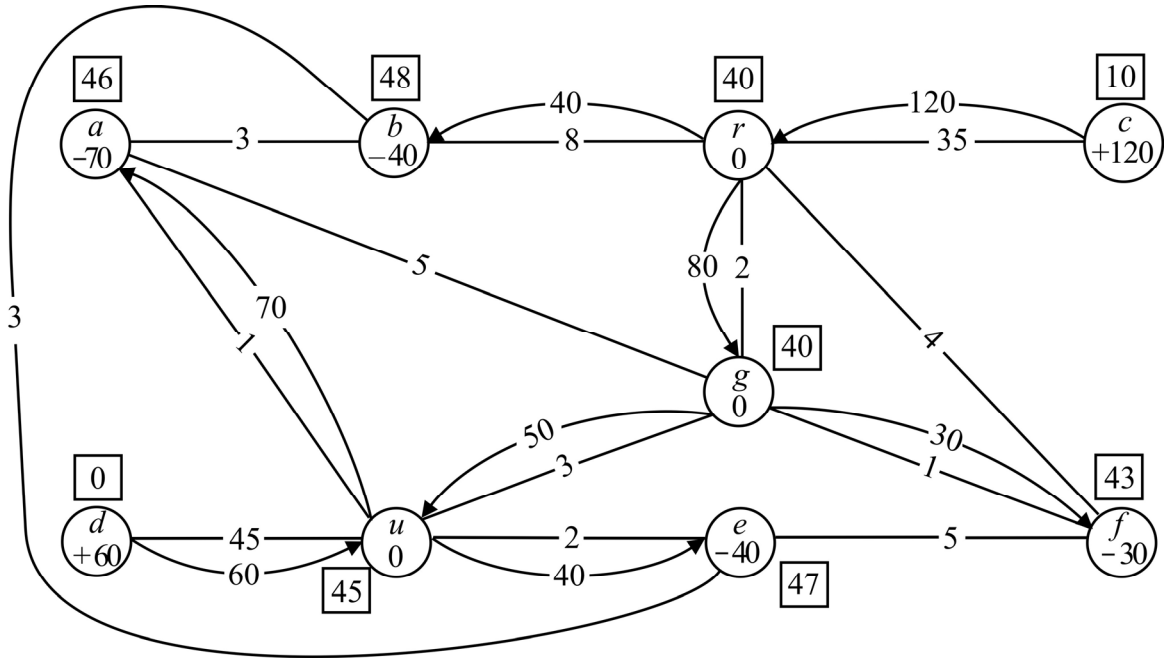


Рис. 9.10. Граф транспортно-производственной задачи

При проверке допустимого плана на оптимальность с помощью потенциалов должны отсутствовать контуры, а граф – быть в форме дерева.

9.6. Классификация с использованием графов

В научных направлениях по мере накопления информации проводится ее обобщение, группировка и классификация. Содержание классификаций зависит от критериев или признаков, которые используются авторами. Математические методы можно использовать для классификации объектов по наиболее типичным признакам.

В зависимости от методического подхода и используемых признаков классификации делят на естественные и искусственные (вспомогательные).

Естественная классификация раскрывает внутренние закономерности в развитии классифицируемых объектов и служит целям познания окружающего мира. Знание о том, к какому классу принадлежит объект, дает возможность судить о его свойствах. Познавательное значение искусственной классификации ограничено. Она создается для облегчения поиска того или иного индивидуального объекта среди других в классификационной схеме. Сложные классификации часто совмещают свойства естественной и искусственной классификаций.

Под классификацией понимается разработка способов и приемов построения классификационных схем. Она строится по следующим формальным правилам:

- на каждом этапе классификации (деления множества на подмножества) должен сохраняться один классификационный признак;
- классификация должна быть исчерпывающей, т. е. объединение подмножеств должно составить делимое множество;
- получаемые в результате деления подмножества должны исключать друг друга;
- классификация должна быть непрерывной, без скачков; на каждом этапе деления множества на подмножества последние должны быть ближайшими видами делимого множества.

Выбор методических приемов построения классификации зависит от характера того конкретного множества изучаемых объектов, которые подлежат классификации, от их количества и полноты имеющихся знаний о них. Ниже рассмотрим основные типы классификационных схем, структуру которых наиболее удобно отобразить в форме графов.

Иерархическая классификация. Иерархия представляет собой отношение подчиненности между объектами разных порядков. В ней отражаются и отношения соподчиненности объектов.

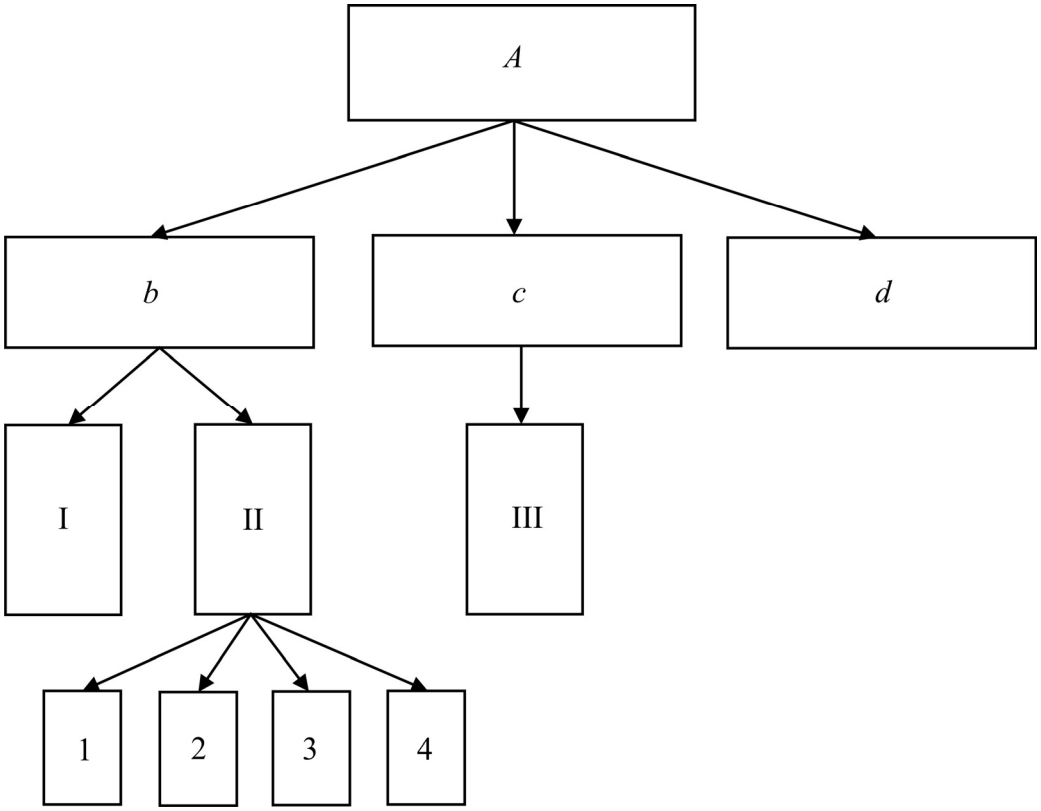


Рис. 9.11. Общая схема иерархической классификации

На рис. 9.11 иерархическая классификация представляет собой выходящее дерево графа, корень которого – множество классифицируемых объектов М. В ней выделено три этапа. На первом этапе выделены группы $М_1$, $М_2$ ... на основании признака П. Это ряд первого уровня классификации. На втором этапе каждая группа первого уровня по признаку $П_2$ делится на ряды второго уровня $М_{11}$, $М_{12}$ .... На третьем этапе каждая из классификационных групп второго уровня может делиться по признаку $П_3$ на более дробные группировки, которые образуют классификационные ряды третьего уровня $М_{111}$, $М_{112}$ …. Количество этапов классификации определяет ее глубину. С увеличением широты классификации уменьшается ее глубина. Глубина и широта классификации на каждом этапе может быть различной. Упорядочение групп в классификационном ряду может производиться на основе количественного или качественного признака.

Дихотомическая классификация. Дихотомия – это последовательное деление целого на две несовпадающие части. Количество этапов деления зависит от специфики классифицируемого объекта. Примером может быть деление клетки только на две части. Классификационная схема изображена на рис. 9.12.

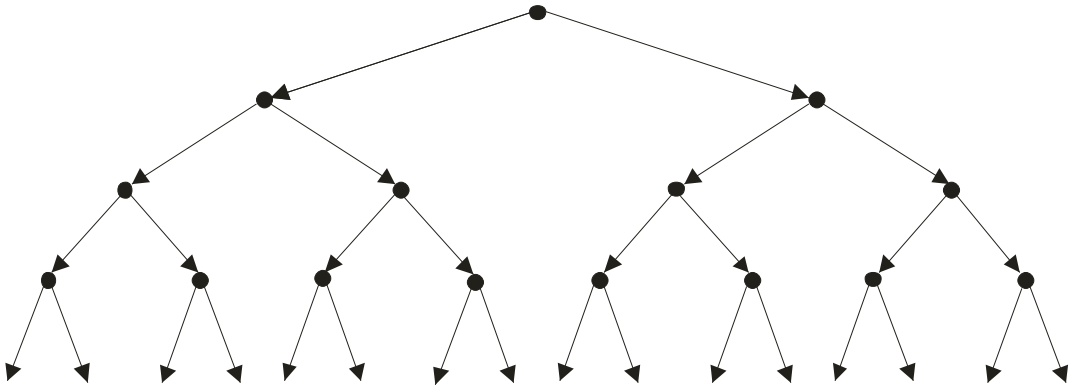


Рис. 9.12. Граф дихотомической классификации

Дихотомия понимается в географии как выделение части из целого, когда объект делится на что-то ведущее и все остальное. Например, выделение городов-миллионеров и все остальные города, города стотысячники и все остальные и т. д.

Таксономическая классификация. Построение графа таксономизации аналогично построению его в иерархической классификации. Различие состоит в выборе классификационных признаков. В таксономической классификации объектов сходство устанавливается по *совокупности признаков*, поэтому ее называют *многопризнаковой.* Трудность классификации состоит в выборе комплексного классификационного признака на каждом этапе. В географии в таксономической классификации ис-

пользуются иерархически соподчиненные *таксоны* (зона, ареал, район и др.). Принцип таксономической классификации заключается в том, что по совокупности некоторых признаков таксоны сходные, по совокупности ряда других отличаются между собой. Совокупность признаков для выделения таксона подбирается на каждом этапе (уровне) классификации.

Рассмотрим таксономическую классификацию с учетом несложной ситуации (табл. 9.2). В Республике Беларусь – шесть областей. Они имеют сходство по развитию одних направлений в сельском хозяйстве и отличия по другим. Следует провести таксономизацию областей в сельскохозяйственном направлении. Среди ведущих признаков отобраны: выращивание зерновых (признак под номером 1), картофеля (2), сахарной свеклы (3), льна (4), кукурузы на зерно (5). Исходные данные поместим в табл. 9.2. В ней знаком плюс отмечено наличие данного признака у определенного таксона.

Таблица 9.2

Выращивание сельскохозяйственных культур в разрезе областей

Признаки / Таксоны	1	2	3	4	5
Витебская (В)	+	+		+	
Могилевская (Мо)	+	+		+	
Минская (Мн)	+	+		+	
Гродненская (Гр)	+	+	+	+	
Брестская (Б)	+	+	+		+
Гомельская (Гм)	+	+	+		+

Визуально по табл. 9.2 все множество таксонов (областей) можно разделить на две группы первого порядка: В, Мо, Мн, Гр и Б, Гм. В дальнейшей таксономизации группа Б и Гм не делится, так как все признаки у них повторяются. В группе первого порядка выделяем две подгруппы второго порядка по наличию сходных признаков: В, Мо, Мн и Гр. Схематически форма графа будет, как показано на рис. 9.13:

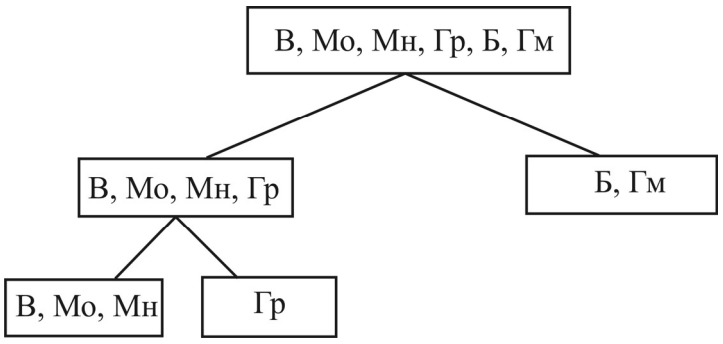


Рис. 9.13. Граф таксономической классификации

В классификации множество называется *монотетическим*, если оно объединяет полностью однородные таксоны, например, Гр. Множество Б, Гм политетическое, так как в него входят таксоны, однородные лишь по ведущим признакам.

Многоаспектная (фасетная) классификация. Географические объекты характеризуются множеством признаков, например, хозяйство республики. В таком случае нельзя создать единую систему их классификации, поэтому создают многоаспектную классификацию (рис. 9.14). Она заключается в параллельной классификации одного исходного множества (хозяйство республики) объектов по признакам, которые соответствуют различным целям: промышленность, сельское хозяйство и т. д. В результате выполняем самостоятельные фасетные классификации, количество которых определено практической целью.

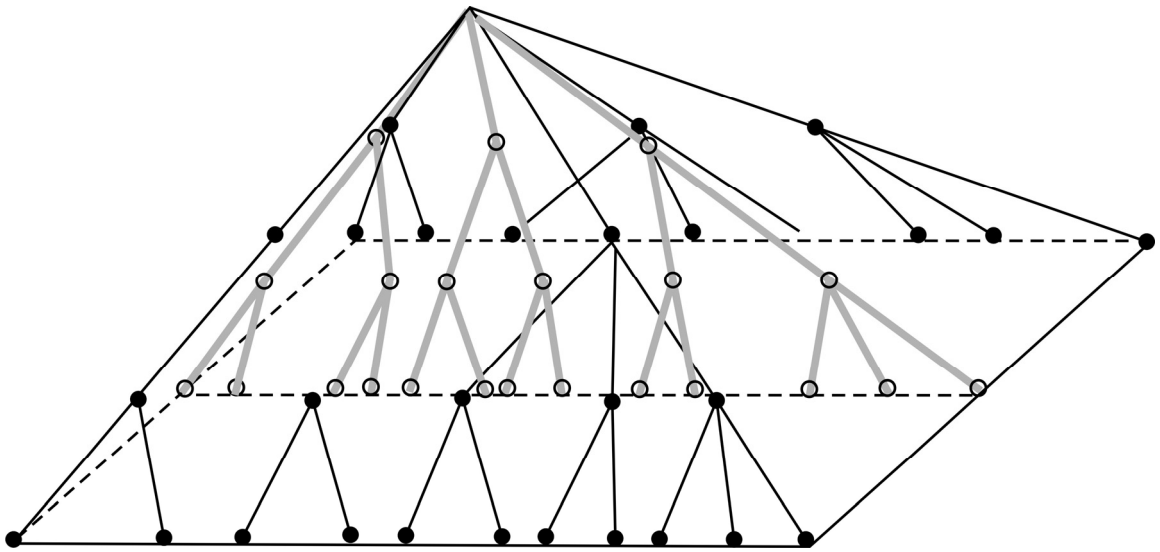


Рис. 9.14. Графическое отображение многоаспектной классификации

Общее у них – множество классифицируемых объектов. Сложный по содержанию фасет будет иметь различную глубину классификационных ветвей и различную широту классификационных рядов.

Таким образом, теоретико-графовый подход позволяет изучать экономико-географические объекты с неизвестной метрикой пространства. Использование помеченных графов расширяет возможности количественного анализа явлений и процессов. Аппарат теории графов связан с теорией множеств, теорией отношений, матричной алгеброй, математическим программированием. К сожалению, он недостаточно используется в решении ряда проблем экономической географии: при районировании, исследовании территориально-производственных комплексов и промышленных узлов, при планировании различных линий (электропередач, нефте- и газопроводов, каналов связей и др.) производственной инфраструктуры.

Глава 10

ДИНАМИЧЕСКИЕ РЯДЫ

Ряд расположенных в хронологической последовательности значений статистических показателей представляет собой *временной* (*динамический*) *ряд*.

Статистические показатели, характеризующие изучаемый объект, называют *уровнями ряда*. В динамическом ряду они могут быть *абсолютными, относительными или средними величинами.* Ряды динамики, представленные за определенный промежуток времени, называются *интервальными*. В результате суммирования уровней интервального динамического ряда получаем *накопленные итоги.* Вследствие многих обстоятельств однородность величин, составляющих динамический ряд, может нарушаться, и таким образом изменяется сопоставимость уровней динамического ряда. Если каждый уровень динамического ряда сравнивается с одним и тем же предшествующим уровнем, как правило, первоначальным – это сравнение с *первоначальной базой*. Если сравнение проводится с предшествующим уровнем – это сравнение с *переменной базой*.

Для представления модели динамического ряда используется *аналитическое выравнивание ряда динамики*. Закономерно изменяющийся уровень изучаемого показателя оценивается как функция времени. В табл. 10.1 приводятся различные виды трендовых моделей, наиболее часто используемые для аналитического выравнивания.

Выбор формы кривой определяет результаты экстраполяции тренда. Один из наиболее распространенных приемов сглаживания уровней первоначального ряда динамики – это *метод скользящей средней*.

Выполнить прогноз по уравнению тренда можно путем экстраполяции тенденции, наблюдавшейся в прошлом. Уровень динамического ряда ($\hat{y}$), полученный в результате экстраполяции, используется для определения прогнозного значения на будущее.

Наличие зависимости между последующими и предшествующими уровнями динамического ряда называют *автокорреляцией*, а построение модели зависимости будущих значений рассматриваемого показателя от прошлых его значений называется *авторегрессией.*

Таблица 10.1

Виды трендовых моделей

Название функции	Описание функции
Линейная	$\hat{Y}_t = b_0 + b_1 t$
Парабола второго порядка	$\hat{Y}_t = b_0 + b_1 t + b_2 t^2$
Кубическая парабола	$\hat{Y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$
Показательная	$\hat{Y}_t = b_0 \cdot b_1 t$
Экспоненциальная	$\hat{Y}_t = b_0 \cdot e^{b_1^t}$
Модифицированная экспонента	$\hat{Y}_t = b_0 + b_1 \cdot b_2^t$
Кривая Гомперца	$\hat{Y}_t = b_0 \cdot b^{b_1^t}$
Логистическая кривая	$\hat{Y}_t = \frac{b_0}{1 + b_1 e^{-b_2 t}}$
Логарифмическая парабола	$\hat{Y}_t = b_0 b_1^t b_2^{t^2}$
Гиперболическая	$\hat{Y}_t = b_0 + b_1 \cdot (1 / t)$

Ряд исследований проводится длительное время (мониторинг), чтобы выявить тенденцию или закономерность развития и прогнозирования какого-либо процесса или явления. Для оценки таких событий используют динамические ряды (тренд-анализ). Они представляют собой однородные статистические величины, показывающие изменение явления или процесса во времени. С помощью тренд-анализа характеризуются тенденции изменения явления во времени, подбираются статистические модели, описывающие эти изменения, производится поиск промежуточных значений путем интерполяции, предсказание результатов значений в перспективе (экстраполяция).

Динамические ряды бывают *простые* (описание одного явления), *сложные* (описание нескольких явлений), *производные* (составленные из средних или относительных величин), *моментные* (оценка события за определенный момент времени), *интервальные* (анализ явления за год, полгода, месяц).

Для создания линии тренда по данным диаграммы применяются регрессионный анализ, описывающий взаимодействие между переменными. Следует лишь выбрать один из шести способов аппроксимации данных: линейная, логарифмическая, полиномиальная, степенная, экспоненциальная, скользящая средняя.

10.1. Показатели динамического ряда

На первом этапе статистической обработки динамических рядов анализируются основные тенденции (*тренд*) изменения явления во времени. Используется графическое изображение, которое дает исчерпывающую информацию. Вычисляется комплекс специальных показа-

телей, позволяющих дать количественную оценку динамики анализируемого явления.

Абсолютный прирост или *убыль* характеризуют изменение явления в единицу или интервал времени. Вычитают из данных последующего периода данные предыдущего. Если ряд возрастает, то прирост считается положительным.

Темп роста или снижения – соотношение в процентах последующего уровня к предыдущему, умноженное на 100. Положительный прирост имеет показатель более 100%, отрицательный – менее 100%.

Темп прироста показывает, на сколько процентов увеличился или уменьшился уровень явления. Отражает относительную скорость изменения явления от одного отрезка времени к другому. Вычисляется путем деления абсолютного прироста на предыдущий уровень либо вычитанием из показателя темпа роста 100. При положительном приросте показатель больше нуля, при отрицательном – меньше нуля.

Абсолютное значение 1% прироста характеризует значение или стоимость 1 % прироста изучаемого явления. Может вычисляться делением абсолютного прироста на темп прироста или делением показателя предыдущего уровня на 100. «Стоимость» 1 % темпа роста и прироста в различных совокупностях разная.

П р и м е р. Число районов г. Минска с высоким уровнем загрязнения атмосферного воздуха в 2004 г. было 4, в 2005 г. стало 8. Темп роста составил 200 %. В г. Новополоцке таких районов в 2004 г. было 10, а в 2005 г. стало 15. Темп роста составил 50 %. Однако в первом случае число неблагополучных районов увеличилось на 4, во втором – на 5. Это говорит о том, что даже в одном динамическом ряду значение 1 % роста и темпа прироста может существенно отличаться на разных отрезках времени.

Показатель наглядности характеризует динамику явления в процентах относительно исходного уровня, который принимается за 100. В отличие от других показателей стоимость одного процента здесь остается неизменной. Однако динамика изменения исходных данных от одного промежутка времени к другому становится менее выразительной

Существуют различные варианты вычисления показателей динамики. Они отличаются набором исходных данных и трудоемкостью вычислений (табл. 10.2).

Приведем примеры расчета показателей, представленных в табл. 10.2. Абсолютный прирост в 1986 и 1987 годах:

$$А_{86=} У_{86} – У_{85} = 90,2 – 65,8 = 24,4; \quad А_{87} = У_{87} – У_{86} = 67,4 – 90,2 = –22,8.$$

Таблица 10.2

Уровень производства промышленной продукции (ПП) предприятия

Год	Уровень ПП	Абсолютный прирост	Темп роста	Темп прироста	1% прироста	Показатель наглядности
	У	А	Т	Р	П	Н
1985	65,8					100,0
1986	90,2	24,4	137,1	37,1	0,7	137,1
1987	67,4	–22,8	74,7	–25,3	0,9	102,1
1988	94,3	26,9	139,9	39,9	9,7	143,3
1989	55,4	–38,9	58,7	–41,3	0,9	84,2
1990	45,1	–10,3	81,4	–18,6	0,6	68,5
1991	48,2	3,1	106,9	6,9	0,5	73,3

Темп роста в 1986 и 1987 годах:

$$Т_{86} = (У_{86} / У_{85}) \cdot 100 = (90,2 / 65,8) \cdot 100 = 137,1;$$

$$Т_{87} = (У_{87} / У_{86}) \cdot 100 = (67,4 / 90,2) \cdot 100 = 74,7.$$

Темп прироста в 1986 и 1987 годах:

первый способ расчета: $Р_{86} = (А_{86} / У_{85}) \cdot 100 = (24,4 / 65,8) \cdot 100 = 37,1;$

$$Р_{87} = (А_{87} / У_{86}) \cdot 100 = (–22,8 / 90,2) \cdot 100 = –5,3;$$

второй способ расчета: $Р_{86} = Т_{86} – 100 = 137,1 – 100 = 37,1;$

$$Р_{87} = Т_{87} – 100 = 74,4 – 100 = –25,3.$$

Абсолютное значение 1 % прироста в 1986 и 1987 годах:

первый способ расчета: $П_{86} = У_{85} / 100 = 65,8 / 100 = 0,66;$

$$П_{87} = У_{86} / 100 = 90,2 / 100 = 0,9;$$

второй способ расчета: $П_{86} = А_{86} / Р_{86} = 24,4 / 37,1 = 0,7;$

$$П_{87} = А_{87} / Р_{87} = –22,8 / –25,3 = 0,9.$$

Показатель наглядности прироста в1986 и 1987 годах по сравнению с 1985 г.:

$$Н_{86} = (У_{86} / У_{85}) \cdot 100 = (90,2 / 65,8) \cdot 100 = 137,1;$$

$$Н_{87} = (У_{87} / У_{85}) \cdot 100 = (67,4 / 65,8) \cdot 100 = 102,4.$$

Вычисление средних. Расчет средней в моментном ряду с равными промежутками между датами:

$$М = (1/2\ У_{85} + У_{86} + У_{87} + \ldots + 1/2\ У_{91}) / n,$$

где n – число анализируемых наблюдений.

Средний уровень в моментном ряду с неравными промежутками между датами:

$$М = (1/2\ У_{85} \cdot t_{85} + У_{86} \cdot t_{86} + \ldots + 1/2\ У_{91} \cdot t_{91}) / (t_{85} + t_{86} + \ldots + t_{91}),$$

где t – число дней в году.

Средний уровень в интервальном ряду: $М = (У_{85} + У_{86} + \ldots + У_{91}) / n$.

Средний абсолютный прирост: $М = (А_{85} + А_{86} + \ldots + А_{91}) / n$.

Средний темп прироста (среднее хронологическое) вычисляется в виде среднего геометрического: $М_г = \sqrt[n]{P_{85} \cdot P_{86} \cdot \ldots \cdot P_{91}}$.

Динамический характер всех используемых показателей может принимать самые разнообразные формы. Например, абсолютные приросты могут быть стабильными, а темпы роста (прироста) при этом увеличиваться или уменьшаться.

10.2. Сглаживание динамических рядов

Углубленный анализ временных рядов требует использования более сложных методик математической статистики. При наличии в динамических рядах значительной случайной ошибки (шума) применяют один из двух простых приемов – *сглаживание* или *выравнивание* путем укрупнения интервалов и вычисления групповых средних. Этот метод позволяет повысить наглядность ряда, если большинство «шумовых» составляющих находятся внутри интервалов. Однако, если «шум» не согласуется с периодичностью, распределение уровней показателей становится грубым, что ограничивает возможности детального анализа изменения явления во времени.

Более точные характеристики получаются, если используют *скользящие средние* – широко применяемый способ для сглаживания показателей среднего ряда. Он основан на переходе от начальных значений ряда к средним в определенном интервале времени. В этом случае интервал времени при вычислении каждого последующего показателя как бы скользит по временному ряду.

Применение скользящего среднего полезно при неопределенных тенденциях динамического ряда или при сильном воздействии на показатели циклически повторяющихся выбросов (резко выделяющиеся варианты или интервенция).

Чем больше интервал сглаживания, тем более плавный вид имеет диаграмма скользящих средних. При выборе величины интервала сглаживания необходимо исходить из величины динамического ряда и содержательного смысла отражаемой динамики. Большая величина динамического ряда с большим числом исходных точек позволяет использо-

вать более крупные временные интервалы сглаживания (5, 7, 10 и т. д.). Если процедура скользящего среднего используется для сглаживания несезонного ряда, то чаще всего величину интервала сглаживания принимают равной 3 или 5.

Приведем пример вычисления скользящего среднего числа хозяйств с высокой урожайностью (более 30 ц/га) (табл. 10.3).

Таблица 10.3

Сглаживание динамического ряда укрупнением интервалов и скользящим средним

Учетный год	Число хозяйств с высокой урожайностью	Суммы за три года	Скользящие за три года	Скользящие средние
1982	84			90,0
1983	94	270	90,0	89,7
1984	92			88,7
1985	83			87,3
1986	91	262	87,3	87,0
1987	88			86,7
1988	82			83,0
1989	90	249	83,0	82,3
1990	77			82,3
1991	80			82,6
1992	90	248	82,7	82,7
1993	78			

Примеры вычисления скользящего среднего:

1982 г. (84 + 94 + 92) / 3 = 90,0;
1983 г. (94 + 92 + 83) / 3 = 89,7;
1984 г. (92 + 83 + 91) / 3 = 88,7;
1985 г. (83 + 91 + 88) / 3 = 87,3.

Составляется график. На оси абсцисс указываются годы, на оси ординат – число хозяйств с высокой урожайностью. Указывают координаты числа хозяйств на графике и соединяют полученные точки ломаной линией. Затем указываются координаты скользящей средней по годам на графике и соединяются точки плавной полужирной линией.

Более сложным и результативным методом является сглаживание (выравнивание) рядов динамики с помощью различных *функций аппроксимации.* Они позволяют формировать плавный уровень общей тенденции и основную ось динамики.

Наиболее эффективный метод сглаживания с использованием математических функций – *простое экспоненциальное сглаживание.* Его

применяют для учета всех предшествующих наблюдений ряда по формуле:

$$S_t = \alpha \cdot X_t + (1 - \alpha) \cdot S_{t-1},$$

где S_t – каждое новое сглаживание в момент времени t; S_{t-1} – сглаженное значение в предыдущий момент времени t –1; X_t – фактическое значение ряда в момент времени t; α – параметр сглаживания.

Если α = 1, то предыдущие наблюдения полностью игнорируются; при величине α = 0 игнорируются текущие наблюдения; значения α между 0 и 1 дают промежуточные результаты. Изменяя значения этого параметра, можно подобрать наиболее приемлемый вариант выравнивания. Выбор оптимального значения α осуществляется путем анализа полученных графических изображений исходной и выравненной кривых либо на основе учета суммы квадратов ошибок (погрешностей) вычисленных точек. Практическое использование этого метода следует проводить с использованием ЭВМ в программе MS Excel. Математическое выражение закономерности динамики данных можно получить с помощью *функции экспоненциального сглаживания*.

10.3. Выравнивание по способу наименьших квадратов

Предлагаемый способ – один из самых эффективных. Суть его следующая: из бесконечного числа линий, которые могли бы быть теоретически проведены между точками, изображающими исходный ряд, выбирается только одна прямая, которая имела бы наименьшую сумму квадратов отклонений исходных (эмпирических) точек от этой теоретической прямой. Выравнивание проводят по уравнению прямой $y = a + bt$ или по уравнению параболы второго порядка $y = a + bt + ct^2$. В основе выбора параболы для выравнивания лежит предположение о том, что не скорость динамики, а ускорение является постоянной величиной. В качестве постоянных величин выступают a, b, c порядкового номера какого-либо периода t. После расчета постоянных величин a и b известным способом получаем следующее уравнение прямой, по которому вычисляем ряд выравнивания y^1 (табл. 10.4):

$$y^1 = 18{,}748 + 1{,}8382\ t;\ R^2 = 0{,}4047.$$

Показателем правильности выбора того или иного уравнения служит коэффициент R^2. Чем ближе его значение к единице, тем больше соответствие фактического и выравненного распределений.

Современные программы статистической обработки позволяют получать различные теоретические кривые в автоматическом режиме. По результатам можно проводить экстраполяцию или интерполяцию рядов.

Таблица 10.4

Выравнивание динамического ряда по способу наименьших квадратов

Номер года	Фактический уровень	Отклонение от центра	Расчетные параметры уравнений	Произведение *yd*	Ряд выравнивания
t	*y*	*d*	d^2	*yd*	y^1
1	16,5	–7	49	–115,5	20,6
2	14,3	–6	36	–85,8	22,4
3	44,0	–5	25	–220,0	24.3
4	35,6	–4	16	–142,4	26,1
5	30,4	–3	9	–91,2	27,9
6	32,4	–2	4	–64,8	29,8
7	22,5	–1	1	–22,5	31,6
8	28,8	0	0	0	33,5
9	15,2	1	1	15,2	35,3
10	42,0	2	4	84,0	37,1
11	26,6	3	9	79,8	39,0
12	42,6	4	16	170,4	40,8
13	51,3	5	25	256.5	42,6
14	46,2	6	36	277,2	44,5
15	53,4	7	49	373,8	46,3
Итого	501,8		280,0	514,7	

П р и м е р. Дать прогноз на следующий шестнадцатый год (см. табл. 10.4) с использованием уравнения регрессии: $У_{16} = 18{,}768 + 1{,}832 \cdot 16 = 48{,}06$.

Достоверность статистического прогноза зависит от степени интерации взаимосвязи явлений, которая обеспечивает сохранение механизма формирования явления и инерционность характера динамики (темп, направление, устойчивость) на протяжении длительного времени. Экстраполяция на очень большой период времени вперед или назад резко снижает точность прогноза при R^2 меньше 0,6.

Глава 11

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ГЕОГРАФИИ

Моделирование выполняет роль необходимого инструмента в географической науке. В современных условиях распространено математическое моделирование, которое направлено на эффективное использование имеющейся информации. Оно формирует представление о процессах или явлениях функционирования сложной системы для определенного этапа ее развития.

На планетарном уровне модели классифицируются на *имитационные, концептуальные и промежуточные.* Имитационные модели составляются для представления динамики изменения явлений, например климата. Концептуальные, или методические, модели предназначены для демонстрации правдоподобности процессов и формируются на основе общего понимания обратных связей. Модели промежуточной сложности необходимы для имитации взаимодействия среди процессов в природной системе.

Термин «модель» произошел от латинского *modulus* – образец, норма, мера. Модель представляет частный случай *аналогии* – важного метода научного познания. Исследователи стремятся к объяснению неизвестного через известное, понятное. Например, топографо-геодезическая карта дает представление о рельефе.

В географии различают следующие основные модели: словесные, картографические, структурные, графические, математические, натурные. Модели могут быть также комбинированными: математико-картографическими, математико-графическими и др.

Словесные модели представляют собой описание геосистемы с помощью средств языка.

Картографические модели – это географические карты вместе с нанесенной на них ситуацией определенного содержания и назначения. Использование в географических исследованиях результатов математи-

ческого анализа и отражение их на карте приводит к созданию математико-картографической модели (например, отражение коэффициентов корреляции на карте в виде изокоррелят, характеризующих пространственную зависимость между двумя переменными).

Структурные модели (схемы) весьма часто применяются при классификации объектов, систем, процессов по определенному признаку или для передачи последовательности процессов при изучении генезиса, эволюции объекта или системы (например, классификация ландшафтов на местном, региональном или глобальном уровне, представление о смене элементарных природных процессов при гумификации и минерализации органического вещества). Соподчиненность отдельных структурных элементов при составлении модели выражается не только в виде линий и геометрических фигур, но и с включением словесной модели. Так формируется структурно-словесная модель.

Графическая модель представляет собой график с нанесенными на него результатами исследований в виде точек, линий и с помощью других способов отображения. Графическая модель может сочетаться с математической с помощью уравнения, характеризующего изображение. Такая модель называется математико-графической.

Математические модели представляют собой абстрактное описание объектов, явлений или процессов с помощью знаков (символов). Они имеют вид уравнений или неравенств, формул. Применяются в случаях, когда иное моделирование затруднено или невозможно.

Все модели отражают наиболее существенные стороны объекта, способны замещать его, давать информацию о предполагаемом поведении или изменяющихся условиях ($y = ax^2$, где x – переменная). Таким образом, модель служит средством познания оригинала и отражает наиболее важные его свойства.

Натурная модель представляет собой имитацию природного объекта или явления в виде макета.

По характеру отражения системы или процесса выделяют соответственно *статические* и *динамические модели*. И те и другие бывают *детерминированными* и *стохастическими*. Статистические детерминированные модели характеризуют структуру без развития ее во времени. Статистические стохастические модели учитывают возможные варианты состояния системы в определенный момент. Динамические детерминированные модели отражают определенное направление развития системы. Динамические стохастические модели воспроизводят структуру, связь и процесс развития системы с учетом вероятностей колебания фак-

торов, оказывающих влияние на динамику этого процесса. Такие модели являются оптимальными и идеальными для изучения всех сложных геосистем. Идеальная модель не должна быть слишком сложной или слишком простой.

Ценность любой модели определяется достигнутым в ней уровнем обобщения. Поэтому модели изменяются и уточняются по мере поступления новых данных. Для достижения высокого уровня обобщения при построении модели требуется высокое качество отбора используемой информации. Хорошая модель может наталкивать исследователя на новые проблемы, выдвижение гипотез, на сбор, упорядочение и выявление необходимой информации. Модель выполняет также конструктивную функцию как ступень на пути создания теории и познания законов.

Следует иметь в виду, что при создании модели нельзя полностью устранить недостатки, которые обусловлены самой методикой упрощения. Это может привести к несоответствию модели и оригинала, стать причиной неточностей в интерпретации исследуемого явления и ошибок в прогнозе. Поэтому при построении модели необходимо использовать лишь объективно отобранный и проверенный материал.

Моделирование представляет собой процесс воспроизводства модели объекта, явления или процесса с целью решения поставленной задачи определенными методическими приемами для контроля за результатами исследования и их реализацией.

Объект является физическим (материальным) телом и изучается при помощи геометрической модели или реже цифровой математической модели. *Явление* – это внешние свойства и признаки предмета, постигаемые через ощущение, восприятие, представление (форма, размер, цена отражают объективно действующий экономический закон стоимости). *Процесс* выражается через ход, развитие явления, последовательную смену состояния объекта (производительность труда).

С точки зрения кибернетики (от греч. – рулевой) объектами моделирования являются *системы* – относительно обособленные и упорядоченные совокупности, обладающие особой связностью и целесообразно взаимодействующими частями, способными реализовать определенные функции.

Состояние системы, ее составных частей и происходящие в ней процессы выражает *информация.* Имея представление о системе на основе информации можно *управлять* системой в ходе целенаправленного воздействия с целью обеспечить ее контролируемое поведение при изменяющихся условиях.

11.1. Математическое моделирование природных и общественных процессов

С помощью математического моделирования можно решать задачи в области географии: проводить классификацию, районирование, прогнозирование. Практически нет таких областей географии, где бы не строились математические модели различной сложности.

Процесс математического моделирования включает пять стадий: формализацию, реализацию, обработку модели, интерпретацию результатов, проверку. При *формализации* составляется географическая модель. При этом устанавливается цель исследования, определяются моделируемые свойства, способ идентификации и ограничения объема информации и измерения его свойств. *Реализация* (построение) модели предполагает выражение системы аксиом на выбранном языке. *Обработка модели* включает экспериментальные действия: анализ, разделение на подмодели, учет частных свойств, синтез. *Интерпретация результатов* состоит в том, что полученные в ходе обработки модели новые знания переносятся на оригинал. *Проверка модели* заключается в интерпретации результатов, анализе правильности преобразований, сопоставлении полученных результатов с реальными данными. Последнее положение относится к проверке эмпирической модели.

Математическое моделирование позволяет количественно выражать географические закономерности в виде различных моделей, которые дают возможность ответить на вопросы, почему именно так развивается система, что станет с ней при изменении обстановки. Модель позволяет также обнаружить недостатки эмпирических исследований, их слабые стороны.

Сложная математическая модель обычно строится географом совместно с математиком. Однако при этом явление упрощают, оставляя ведущие факторы и причины, которые выявляются с использованием статистического, корреляционного, факторного и других рассмотренных видов анализа. В процессе моделирования интуиция и опыт специалиста играют определяющую роль.

Специфика математической модели в географии заключается в моделировании как отдельных компонентов географической среды, так и комплекса элементов, составляющих ландшафт. Рассмотрим пример математического моделирования с использованием простой модели.

П р и м е р. Известно, что в результате ураганов ветровалу подвержены в большей степени древесные породы, имеющие поверхностную корневую систему (ель), породы с мягкой древесиной (береза, осина, липа), а также раз-

реженный лесной массив. Это необходимо учитывать при искусственном возобновлении леса. Требуется найти общую характеристику, по которой можно было бы судить о защитных свойствах различных массивов леса, т. е. определить толщу леса, необходимую для защиты от ураганных ветров. Древесную толщу (Т) выражаем через показатели густоты леса (N) и толщины деревьев (d): $\mathrm{T} = N + d$.

Процесс моделирования включает нахождение зависимости между древесной толщей и расстоянием от опушки леса (т. е. эпицентра урагана) $L_т$, густотой леса и толщиной деревьев.

При дальности видимости в лесу $L_в$ защитный слой в 1 см от эпицентра урагана будет образован на расстоянии, равном:

$$\Delta L = L_в/d. \qquad (11.1)$$

Для создания толщи леса (Т) потребуется расстояние

$$L_т = \Delta L \mathrm{T}. \qquad (11.2)$$

Определим дальность видимости в лесу:

$$L_в = 10^6 / Nd. \qquad (11.3)$$

Подставляя в формулу (11.2) значение ΔL из (11.1), а затем $L_в$ из (11.3), имеем

$$L_т = \mathrm{T} \cdot 10^6 / Nd^2. \qquad (11.4)$$

На основании этой формулы вычисляем расстояние $L_т$, при котором образуется толща Т для различных N и d, т. е. для любого леса. Например, если лес имеет толщу деревьев d = 20 см, густоту N = 765 деревьев на 1 га при защитной толще Т = 25 см, то по формуле (11.4) вычисляем расстояние ($L_т$):

$$L_т = 25 \cdot 10^6 / 765 \cdot 20^2.$$

Аналогично рассчитываем защитную толщу на определенном расстоянии, подставляя различные по величине параметры в формулу (11.4).

Глава 12

ГЕОГРАФИЧЕСКОЕ ПОЛЕ

Одномерный (скалярный) анализ связан с понятиями «поле» и «статистическая поверхность». В математическом понятии *географическое поле* – это такое разделение по земной поверхности количественной оценки, когда каждая ее точка характеризуется конкретной величиной (скаляром). Геометрическое место точек, каждая из которых представлена скаляром географического поля, определяет его *статистическую поверхность*. К каждой точке поверхности, несущей определенную информацию, восстанавливается перпендикуляр, на котором откладывается отрезок, который соответствует величине информации для данной точки. Вершины перпендикуляров объединяются плавной кривой линией. Полученную поверхность называют статистической, или скалярным полем (рис. 12.1).

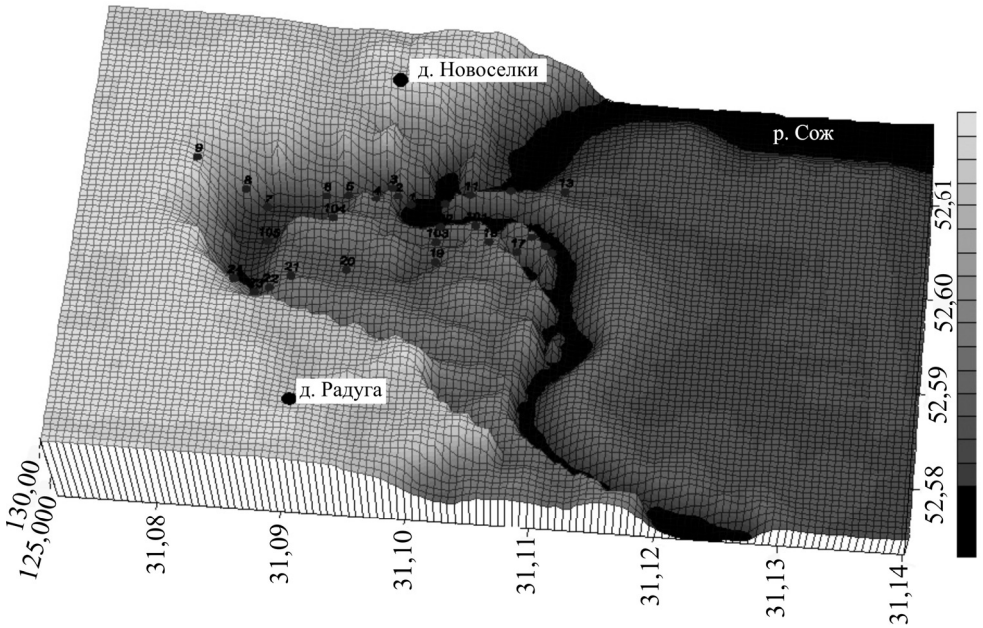


Рис. 12.1. Трехмерная модель рельефа поймы р. Сож (Т. А. Тимофеева, 2006)

Скаляры, или одномерные величины, могут представлять числа одноразового измерения, средних величин, коэффициентов корреляции, вычисленные значения определенной функции и др. Скалярное поле можно представить в виде картографической модели. Полученная таким способом карта статистической поверхности – это *образно-знаковая модель географического поля*. Поле можно изображать разными способами. Наиболее часто употребляется способ изолиний, например поле густоты населения представляется изодендами. На картах любое явление отражается элементами статистической поверхности – «низинами», «горами», «хребтами», «пиками», «впадинами» и т. д.

На картах математической поверхности можно проводить математические операции сложения, вычитания, умножения, деления. Для этого поверхность должна быть представлена способом изолиний.

12.1. Операции над статистическими поверхностями

Скалярное поле математически можно изобразить как функцию трех переменных координат точки $P(x, y, z)$.

При сложении двух или более статистических поверхностей необходимо сложить значения точек z с одинаковыми координатами x и y: $z = f(x, y)$.

При вычитании двух статистических поверхностей необходимо из большей величины z вычесть меньшую величину: $z_3 = z_1 - z_2$.

Практически эти операции выполняются следующим образом. Совмещают картосхемы и на пересечении изолиний совмещенных карт производят сложение или вычитание их величин. Получив новые точки и их значение, соединяют одновысотные точки и получают новые изолинии суммарной или разностной поверхности.

Пример. Имеем две картограммы со статистической поверхностью – сельскохозяйственных и лесных ресурсов, единицы измерения одинаковы. Если сложить эти статистические поверхности, получим новую картосхему суммарной обеспеченности каждой точки территории данными ресурсами. Если вычесть из статистической поверхности общего количества населения статистическую поверхность с трудовыми ресурсами, получим закономерности распределения нетрудоспособного населения на территории.

Существует три способа умножения или деления статистических поверхностей. Первый способ – непосредственное умножение или деление значений пересекающихся изолиний на совмещенных картосхемах. Второй способ – использование вспомогательных логарифмических кривых

на картосхемах. Он применяется в случаях, когда при совмещении поверхностей изолинии не пересекаются. Третий способ используется при условии, когда статистические поверхности имеют сложный рельеф. В таких случаях наносят регулярную квадратную или треугольную сетку точек и в каждой из них вычисляют значения *z* исходных поверхностей путем умножения или деления. Затем между точками проводят линейную интерпретацию и наносят соответствующие изолинии.

Статистические поверхности можно дифференцировать или интегрировать. Дифференцирование поверхности – это определение скорости падения ее «рельефа» в какой-либо точке (градиента точек). Поле высот трансформируется в поле градиентов. Поле покрывают сеткой, определяют угол падения (подъема) поверхности и по таблицам находят тангенс угла. Это модуль (градиент) узловой точки. Полученная новая поверхность будет показывать степень крутизны поверхности в каждой точке. Процесс интегрирования противоположный дифференцированию.

Статистические поверхности позволяют унифицировать визуализацию количественной информации, выполнить количественный анализ закономерностей, объективизировать районирование.

12.2. Методика составления карт изокоррелят

Математические методы в географии начали использоваться в 60-х гг. XX столетия. Однако отражение математических моделей на географическом поле происходит лишь в 1990-х гг. Картографический метод исследования является одним из основных методов познания ландшафтов. Он включает постановку исследовательских задач, всестороннее изучение картографической модели как источника информации, разработку методов, средств и алгоритмов для извлечения, обработки и преобразования картографической информации, оценку точности и достоверности результатов исследования.

Карта как модель обладает особым набором свойств, отвечающих специфической сути географии, и не может быть заменена математическими моделями, для построения которых она служит источником информации. Внедрение математики в географию происходит также через математизацию картографической модели и методов исследования по картам.

География характеризуется противоречивостью представлений о поле. Б. Л. Гуревич и Ю. Г. Саушкин дают следующее определение поля: «Если в точках некоторой рассматриваемой области пространства (трехмерного, например, в тропосфере; двумерного, например, на участке земной поверхности; одномерного, скажем, вдоль полотна железной до-

роги) данная величина (например, температура – скаляр, скорость и направление ветра – векторы) имеет в данный момент времени строго определенное значение, то говорят о поле данной величины» (цит. по кн. Г. И. Сачка и Т. В. Цурковой). Используются два способа представления информации о поле: *изолинейный* и *картограммный*. Создание и анализ серий карт полей – это важный способ описания многогранности географических явлений, исследования их в статике, динамике и взаимосвязях.

Концепция поля позволяет более осознанно подходить к анализу комплексов географических показателей, отображению одного комплекса характеристик в другом, прогнозу во временном и пространственном аспектах. Рассчитанные по картам полей или иным способом значения коэффициентов корреляции дают возможность построить карту в *изокоррелятах* – линиях равных значений коэффициента корреляции. Карта изокоррелят удовлетворяет всем требованиям, предъявляемым к картам статистических поверхностей, и является графической моделью корреляционного поля географических параметров.

Построение карт изокоррелят может быть выполнено различными способами, нами применялся графический подход к определению корреляции по картам в изолиниях. Численное значение коэффициента корреляции двух величин оценивается как косинус угла α между направлениями наибольших скатов (градиентами) – $r_\alpha = \cos \alpha$. Эти направления определяются по изолиниям при совмещении двух карт как перпендикуляры к ним в точке пересечения. Угол α равен углу между изолиниями в данной точке. Корреляция положительная, если скаты однонаправлены, и отрицательная при разнонаправленности скатов.

Приемы математической статистики в последние десять лет широко используются и выделяются в самостоятельный картографо-статистический метод. Особенностью нынешнего этапа применения статистики в картографии является стремление модифицировать применение статистических показателей так, чтобы они соответствовали требованиям картографического метода.

Впервые в Беларуси Г. И. Сачком и Т. В. Цурковой выполнена работа по использованию математических методов в картографии «Математико-картографическое моделирование природных условий Белоруссии» (1984). В ней использован корреляционный анализ, графическая корреляция карт, методы главных компонент, таксономии, пространственной корреляционной функции многомерного поля.

Одни и те же данные служат для получения разных факторных структур при помощи тех же вычислительных процедур. Для обработки выборки используется техника *R*, в которой попарно коррелируют переменные при возможно большем числе наблюдений. Если коэффициенты

корреляции рассчитываются для пары наблюдений по всем значениям переменных, то этот способ расчетов носит название техники *Q*. Здесь подвергается факторизации матрица выборочных данных, транспонированная по отношению к использованной в *R*.

Методика использования фактического материала. Информационную основу исследований составляют мелкомасштабные карты обзорного характера. Они служат источником информации малой точности. Многомерные методы позволяют выделить и исследовать главную ее часть. Доля отбрасываемой информации, связанной с различного рода ошибками, составляет 30 – 40 %. Для получения количественной информации использовались карты рельефа, климата, вод, почв и растительности Беларуси. При необходимости дополнительной информации по отдельным параметрам использовались как карты специального назначения (глубины и густоты расчленения рельефа и др.), так и количественные параметры из литературных и фондовых источников. Среди многочисленных параметров, характеризующих каждый из компонентов ландшафта, выбирались наиболее информативные и ведущие в формировании рельефа, климата, вод, почв, растительности.

Информационную основу составили 20 мелкомасштабных обзорных карт. Значениями переменных являются количественные показатели, снятые с карты по квадратной сетке с шагом 100 км в 40 точках.

Методика составления карт изокоррелят. Определение связей между компонентами – одна из важнейших задач ландшафтоведения, являющаяся основным направлением в изучении структуры и функционирования геосистем и в то же время использующаяся при оценке однородности природных комплексов, их классификации, прогнозирования. Степень связи – предмет изучения корреляционного анализа, формы связи – регрессионного анализа. Нами принята следующая оценка корреляционной связи: тесная $| r | \geq 0{,}7$, средняя $| r | \geq 0{,}4 - 0{,}7$, слабая $| r | \leq 0{,}4$.

Отобранные количественные параметры наносились на картографическую основу по всей республике. В каждой точке получали два коррелируемых между собой параметра. Объем выборок принимался не менее 10. После вычисления коэффициента корреляции с использованием программного пакета «Statistica» составлялась новая картографическая основа с нанесенными величинами коэффициентов корреляции. Затем проводилась интерполяция по полученным величинам коэффициентов корреляции (r). На основе полученных изолиний выявлялись общие закономерности изменения величин корреляции в пространственном аспекте.

Карты изокоррелят дают представление об изменении в пространстве силы связи двух переменных и предназначены для изучения пространственной структуры ландшафтов. Они характеризуют географиче-

ское поле. Концепция поля позволяет более осознанно подходить к анализу географических показателей, исследованию взаимосвязей между разнородными показателями, прогнозу явлений и процессов. В ряде задач возникает необходимость изучения обобщенных зависимостей между двумя множествами случайных величин с совместным распределением. Общая теория корреляции многомерных выборок не разработана. Создан частный метод – каноническая корреляция между двумя векторами. Он представляет собой общий случай линейной множественной регрессии. Каноническая корреляция может быть применена для оценки одного множества случайных величин по другому. Нами метод использован для выявления факторов, формирующих пространственные связи.

Пространственные изменения параметров корреляции в рельефе. Рельеф Беларуси находится над уровнем моря в среднем на 159 м с колебаниями высот от 85 до 346 м. Большая часть низинных и равнинных пространств охватывает 3/5 территории и расположена на высоте 100–200 м над уровнем моря. Холмистые районы, занимающие 1/3 территории республики, поднимаются выше 200 м.

Наибольшей густотой расчленения рельефа обладают холмисто-моренные возвышенности. Большей густотой расчленения рельефа характеризуются платообразные районы лесовидных отложений в пределах Оршанско-Могилевского и отчасти Минского плато (600–1000 м). Среднюю густоту расчленения рельефа (100–1500 м) имеют Нарочано-Вилейская, Неманская и Суражская низины, Центрально-Березинская и Барановичская равнины. Наибольшее однообразие рельефа характерно для Белорусского Полесья с густотой расчленения рельефа 1500–3000 м и в пределах центральной части Дисненской и Верхне-Березинской низин (1500–4200 м).

В пределах республики глубина расчленения рельефа колеблется в пределах от 0 до 80 м. Низинные участки с незначительным волнистым рельефом имеют минимальную глубину расчленения (0–7 м). Это часть районов Полесья, Центрально-Березинской равнины, Верхне-Березинской низины. Большая глубина расчленения рельефа характерна для холмисто-моренных возвышенностей (6–20 м) при максимальной глубине около 80 м. Моренные возвышенности Поозерья имеют небольшую глубину расчленения – 6–12 м. В пределах Ошмянской и Минской возвышенностей и юго-западной части Белорусской гряды глубина расчленения рельефа составляет 12–18 м. Моренные гряды юга Беларуси имеют среднюю глубину расчленения 6–9 м, исключение составляет Мозырская гряда (12 м).

Наличие районов с высокими значениями коэффициента корреляции ($r \geq 0{,}6$) указывает на непосредственную связь двух переменных. Сопряженность может проявляться в наличии тенденции изменения корреля-

рованности, которая может совпадать с тенденцией изменения какой-либо переменной. Последнюю можно рассматривать как основную в паре или комплексе коррелированных переменных.

Максимальная величина коэффициента корреляции, согласно рис. 12.2, между абсолютной высотой и глубиной расчленения рельефа ($r \geq 0{,}9$) характерна для территории Гродненской области. К востоку и юго-востоку корреляционная связь понижается и достигает минимума ($r \geq 0{,}4$) в пределах центральной части Беларуси. В восточной и юго-восточной частях республики корреляционная связь увеличивается до $r \geq 0{,}6$. Средний коэффициент корреляции для Беларуси между абсолютной высотой и глубиной расчленения рельефа составляет 0,57.

Корреляция абсолютной высоты с густотой расчленения рельефа в соответствии с рис. 12.3 минимальная на северо-западе и юго-востоке ($r \geq -0{,}2$). Максимальные значения корреляции в центральной части республики ($r \geq -0{,}4 - 0{,}6$). Отрицательная корреляция для республики в среднем составляет $-0{,}45$. Густота расчленения рельефа отрицательно коррелирует с большинством геоморфологических характеристик.

Рис. 12.2. Карта изокоррелят между абсолютной высотой и глубиной расчленения рельефа

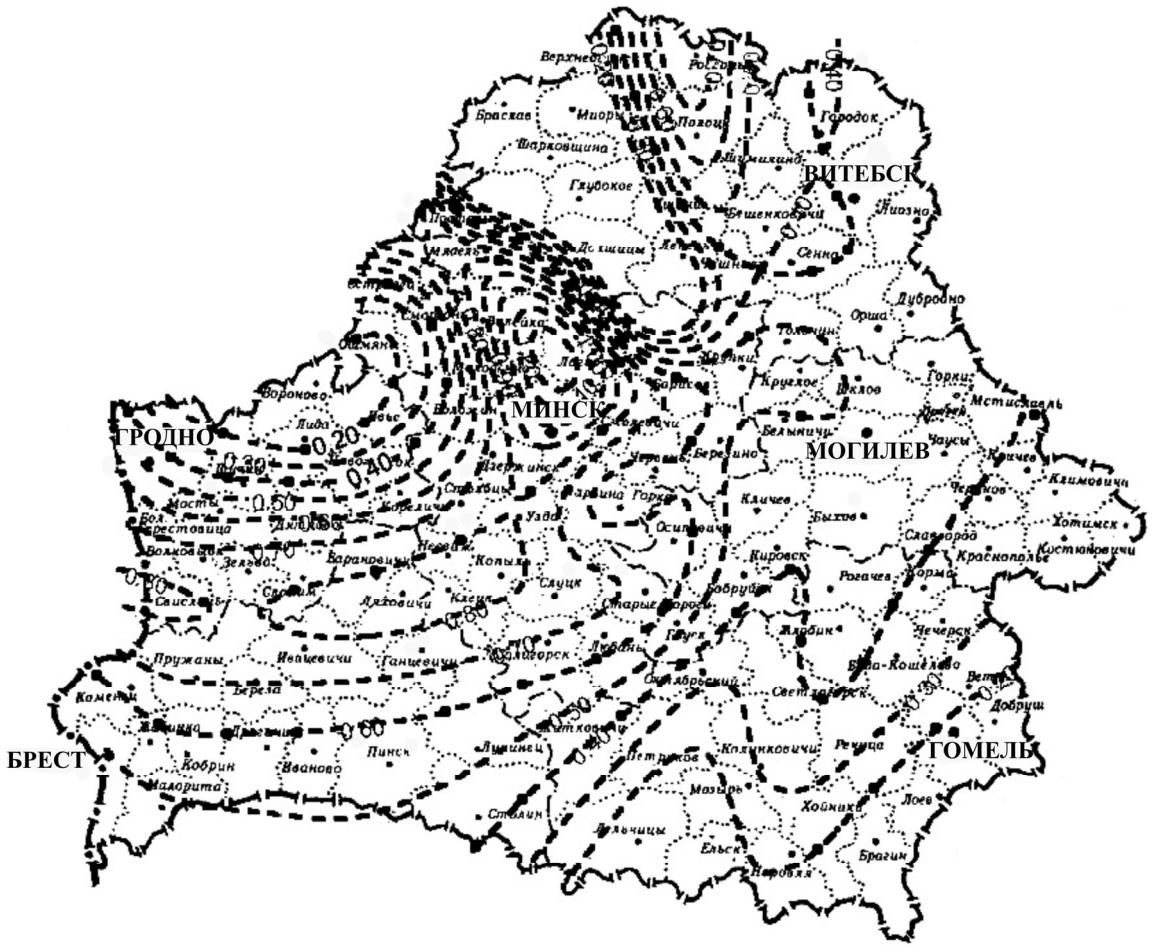


Рис. 12.3. Карта изокоррелят между абсолютной высотой и густотой расчленения рельефа

Пространственные изменения параметров корреляции климата. Климат Беларуси умеренно континентальный и формируется под воздействием западных атлантических воздушных масс. В результате совместного воздействия климатических параметров формируется своеобразный тепловой режим, который характеризуется постепенным понижением температуры с юго-запада на северо-восток. Сумма активных температур постепенно увеличивается с северо-востока (2300 °С) на юго-запад (2800 °С). Количество выпадающих осадков определяется циклонической деятельностью. Осадки конвективного происхождения образуются редко. Циклоническая деятельность убывает с северо-запада на юго-восток, поэтому в этом направлении уменьшается общее количество осадков. В засушливые годы выпадает около 300 мм осадков в год, во влажные – около 1000 мм. Влажные годы повторяются чаще, чем засушливые.

Исследование изменчивости в пространстве климатических условий выполнено количественными методами. Анализу были подвергнуты мелкомасштабные карты по климатическим параметрам, опубликован-

ные в Национальном атласе (2002). Климатические условия изучались по 11 параметрам. Взаимосвязи оценивались с помощью парных коэффициентов корреляции, образующих корреляционную матрицу.

Важными показателями, на наш взгляд, являются сумма годовых осадков и сумма активных температур за вегетационный период.

На рис. 12.4 отражена пространственная корреляция между ними. Средняя корреляция отрицательная и составляет – 0,48.

Максимальные величины коэффициентов корреляции отмечены в Поозерье, в средней части Предполесья и на Новогрудской возвышенности. Их значения колеблются от 0,6 до 0,7. В восточном, западном и южном направлениях величины коэффициентов корреляции постепенно понижаются до 0,4. Такие значения отмечены для юго-восточной части республики и крайней южной части Полесья. Крайняя западная часть Беларуси характеризуется величиной коэффициента корреляции 0,5. Чем выше контраст между температурой и осадками, тем меньшая величина коэффициента корреляции.

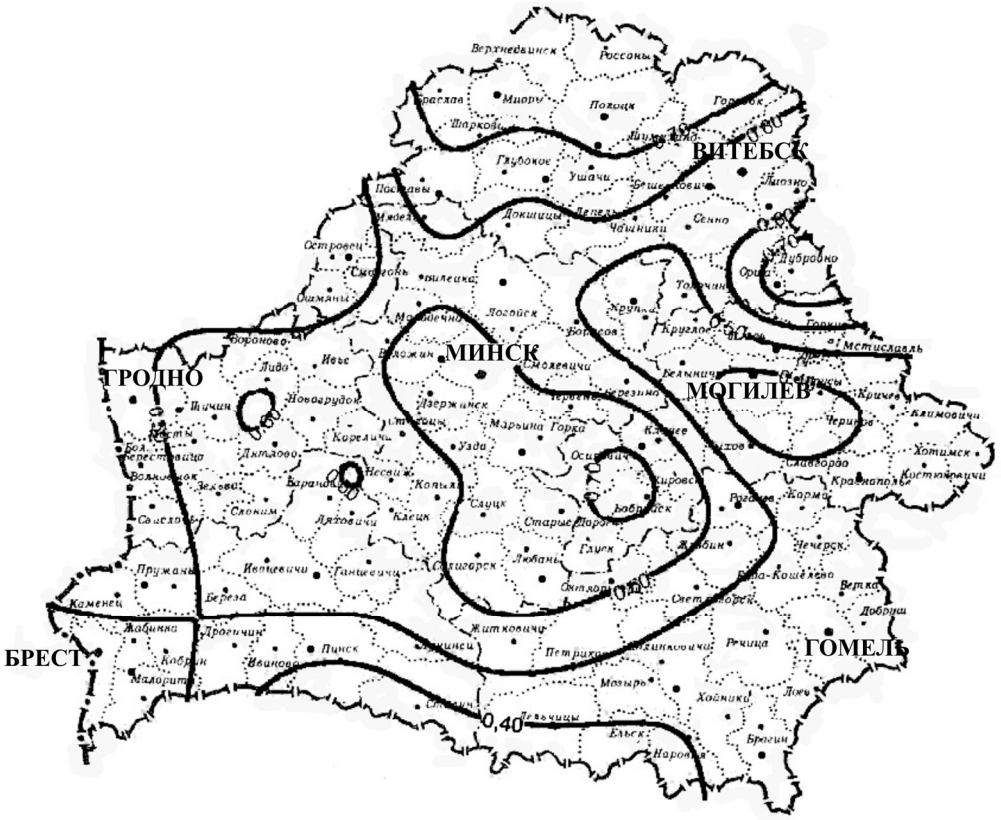


Рис. 12.4. Карта изокоррелят между суммой активных температур вегетационного периода и годовым количеством осадков

ЛИТЕРАТУРА

Основная

Зыков, *А. А*. Основы теории графов / А. А. Зыков. М., 2004.

Калинина, *В. Н.* Математическая статистика : учеб. / В. Н. Калинина, В. Ф. Панкин. 4-е изд., испр. М., 2002.

Колеснев, *В. И.* Экономико-математические методы и моделирование в землеустройстве. Практикум : учеб. пособие / В. И. Колеснев, И. В. Шафранская. Минск, 2007.

Михеева, *В. С.* Математические методы в экономической географии : в 2 ч. / В. С. Михеева. – М., 1981, 1983. Ч. 1 : Применение методов линейного программирования. 1981 ; Ч. 2 : Приложения теории графов. 1983.

Пузаченко, *Ю. Г.* Математические методы в экологических и географических исследованиях : учеб. пособие / Ю. Г. Пузаченко. М., 2004.

Статистика : учеб. пособие / под ред. М. Р. Ефимовой. М., 2000.

Чертко, *Н. К*. Математические методы в физической географии : учеб. пособие для геогр. спец. вузов / Н. К. Чертко. Минск, 1987.

Шикин, *Е. В*. Математические методы и модели в управлении : учеб. пособие / Е. В. Шикин, А. Г. Чхартишвили. М., 2000.

Дополнительная

Боровиков, *В. П.* Statistica : Статистический анализ и обработка данных в среде Windows / В. П. Боровиков. 2-е изд. М., 1998.

Боровиков, *В. П.* Программа STATISTICA для студентов и инженеров / В. П. Боровиков. М., 2001.

Математические методы в географии / Ю. Р. Архипов [и др.]. Казань, 1976.

Оре, *О*. Графы и их применение / О. Оре ; пер. с англ. 2-е изд., доп. М., 2002.

Сачок, *Г. И.* Математико-картографическое моделирование природных условий Белоруссии / Г. И. Сачок, Т. В. Цурканова. Минск, 1984.

Тикунов, *В. С*. Классификация в географии / В. С. Тикунов. М. ; Смоленск, 1997.

Тикунов, *В. С*. Моделирование в картографии / В. С. Тикунов. М., 1997.

Тюрин, *Ю. Н.* Статистический анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. М., 1998.

ПРИЛОЖЕНИЕ

1. Таблица достаточно больших чисел

P	Ошибка опыта *p*, %									
	10	9	8	7	6	5	4	3	2	1
0,75	33	40	51	67	91	132	206	367	827	3308
0,80	41	50	64	83	114	164	256	456	1026	4105
0,85	51	63	80	105	143	207	323	575	1295	5180
0,90	67	83	105	138	187	270	422	751	1690	6763
0,91	71	88	112	146	199	287	449	798	1796	7185
0,92	76	94	119	156	212	306	478	851	1915	7662
0,93	82	101	128	167	227	328	512	911	2051	8207
0,94	88	109	138	180	245	353	552	981	2210	8843
0,95	96	118	150	195	266	384	600	1067	2400	9603
0,96	105	130	164	215	292	421	659	1171	2636	10544
0,965	111	137	173	226	308	444	694	1234	2778	11112
0,970	117	145	183	240	327	470	735	1308	2943	11773
0,975	125	155	196	256	348	502	784	1395	3139	12559
0,980	135	167	211	276	375	541	845	1503	3382	13529
0,985	147	182	231	301	410	591	924	1643	3697	14791
0,990	165	204	259	338	460	663	1036	1843	4146	16587
0,995	196	243	307	402	547	787	1288	2188	4924	19698
0,999	270	334	422	552	751	1082	1691	3009	6767	27069

2. Случайные числа

3393	6270	4228	6909	9407	1865	8549	3217	2351	8410
9108	2330	2157	7416	0398	6173	1703	8132	9065	6717
7891	3590	2502	5945	3402	0491	4328	2365	6175	7695
9085	6307	6910	9174	1753	1797	9229	3422	9861	8357
2638	2908	6368	0398	5495	3283	0031	5955	6544	38383
1313	8338	0623	8600	4950	5414	7131	0134	7241	0651
3897	4202	3814	3505	1599	1649	2784	1994	5775	1406
4380	9543	1646	2815	8415	9120	8062	2421	6161	4634
1618	6309	7909	0874	0401	4301	4517	9197	3350	0434
4858	4676	7363	9141	6133	0549	1972	3461	7116	1496
5354	9142	0847	5393	5416	6505	7156	5634	9703	6221
0905	6986	9396	3975	9255	0537	2479	4589	0562	5345
1420	0470	8679	2328	3939	1292	0406	5528	3789	2882
3218	9080	6604	1813	8209	7039	2086	3369	4437	3798
9697	8431	4387	0622	6893	8788	2320	9358	5904	9539
0912	4964	0502	9683	4636	2861	2876	1273	7870	2030
4636	7072	4868	0601	3894	7182	8417	2367	7032	1003
2515	4734	9897	6761	5636	2949	3979	8650	3430	0635
5964	0412	5012	2369	6461	0678	3693	2928	3740	8047
7848	1523	7904	1521	1455	7089	8094	9872	0898	7174
5182	2571	3643	0707	3434	6818	5729	8615	4298	4129
8438	8325	9886	1805	0226	2310	3675	5058	2515	2388
8166	6349	0319	5436	6838	2460	6433	0644	7428	8556
9158	8263	6504	2562	1160	1526	1816	9690	1215	9590
6061	3525	4048	0382	4224	7148	8256	6526	5340	4064

3. Значение критерия τ в зависимости от объема выборки *N* и уровня значимости α

N	α		*N*	α	
	0,05	0,01		0,05	0,01
4	0,95	0,991	17	0,359	0,460
5	0,807	0,916	18	0,349	0,449
6	0,669	0,805	19	0,341	0,439
7	0,610	0,740	20	0,334	0,430
8	0,544	0,683	21	0,327	0,421
9	0,512	0,635	22	0,320	0,414
10	0,477	0,597	23	0,314	0,407
11	0,450	0,566	24	0,309	0,400
12	0,428	0,541	25	0,304	0,394
13	0,410	0,520	26	0,299	0,389
14	0,395	0,502	27	0,295	0,383
15	0,381	0,486	28	0,291	0,378
16	0,369	0,472	29	0,287	0,374
			30	0,283	0,369

4. Значения критерия Стьюдента *t* при различных уровнях значимости

ν	Уровни вероятности			ν	Уровни вероятности		
	0,95	0,99	0,999		0,95	0,99	0,999
2	4,30	9,93	31,60	21	2,08	2,83	3,82
3	3,18	5,84	12,94	22	2,07	2,82	3,79
4	2,78	4,60	8,61	23	2,07	2,81	3,77
5	2,57	4,03	6,86	24	2,06	2,80	3,75
6	2,45	3,71	5,96	25	2,06	2,79	3,73
7	2,37	3,50	5,41	26	2,06	2,78	3,71
8	2,31	3,36	5,04	27	2,05	2,77	3,69
9	2,26	3,25	4,78	28	2,05	2,76	3,67
10	2,23	3,17	4,49	29	2,04	2,76	3,66
11	2,20	3,11	4,44	30	2,04	2,75	3,65
12	2,18	3,06	4,32	40	2,02	2,70	3,55
13	2,16	3,01	4,22	50	2,01	2,68	3,50
14	2,15	2,98	4,14	60	2,00	2,66	3,46
15	2,13	2,95	4,07	80	1,99	2,64	3,42
16	2,12	2,92	4,02	100	1,98	2,63	3,39
17	2,11	2,90	3,97	120	1,98	2,63	3,37
18	2,10	2,88	3,92	200	1,97	2,60	3,34
19	2,09	2,86	3,88	500	1,96	2,59	3,31
20	2,09	2,85	3,85	∞	1,96	2,58	3,29

5. Критические значения *F* (критерия Фишера)*

ν_2**	ν_1 – степени свободы для большей дисперсии																		
	3	4	5	6	7	8	9	10	12	14	16	20	30	40	50	75	100	200	500	∞
3	9,28 26,46	9,12 28,71	9,01 28,24	8,94 27,91	8,88 27,67	8,84 27,34	8,81 27,34	8,78 27,23	8,74 27,05	8,71 26,92	8,69 26,83	8,66 26,69	8,62 26,50	8,60 26,41	8,58 26,35	8,57 26,27	8,56 26,23	8,54 26,18	8,54 26,14	8,53 26,12
4	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,00 14,66	5,96 14,54	5,96 14,54	5,91 14,37	5,87 14,24	5,84 14,15	5,80 14,02	5,74 13,83	5,71 13,74	5,70 13,69	5,68 13,61	5,66 13,57	5,65 13,52	5,64 13,48	5,63 13,46
5	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,45	4,78 10,15	4,71 10,05	4,74 10,05	4,68 9,89	4,64 9,77	4,60 9,68	4,56 9,55	4,50 9,38	4,46 9,29	4,44 9,24	4,42 9,17	4,40 9,13	4,38 9,07	4,37 9,04	4,37 9,02
6	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,10 7,98	4,06 7,87	4,06 7,87	4,00 7,72	3,96 7,60	3,92 7,52	3,87 7,39	3,81 7,23	3,77 7,14	3,75 7,09	3,72 7,02	3,71 6,99	3,69 6,94	3,68 6,90	3,67 6,88
7	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 7,00	3,68 6,71	3,63 6,62	3,63 6,62	3,57 6,47	3,52 6,35	3,49 6,27	3,44 6,07	3,38 5,90	3,34 5,85	3,32 5,78	3,29 5,75	3,28 5,70	3,25 5,67	3,24 5,66	3,23 5,65
8	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,19	3,39 5,91	3,34 5,82	3,34 5,82	3,28 5,67	3,23 5,56	3,20 5,48	3,15 5,36	3,08 5,20	3,05 5,11	3,03 5,06	3,00 5,00	2,98 4,96	2,96 4,91	2,94 4,88	2,93 4,86
9	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,62	3,18 5,35	3,13 5,26	3,13 5,26	3,07 5,11	3,02 5,00	2,98 4,92	2,93 4,80	2,86 4,64	2,82 4,56	2,80 4,51	2,77 4,45	2,76 4,41	2,73 4,36	2,72 4,33	2,71 4,31
10	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,21	3,02 4,95	2,97 4,85	2,97 4,85	2,91 4,71	2,86 4,60	2,82 4,52	2,77 4,41	2,70 4,25	2,67 4,17	2,64 4,12	2,62 4,05	2,59 4,01	2,56 3,96	2,55 3,93	2,54 3,91
11	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,88	2,90 4,63	2,86 4,54	2,86 4,54	2,78 4,40	2,74 4,29	2,70 4,21	2,65 4,10	2,57 3,94	2,53 3,86	2,50 3,80	2,47 3,74	2,45 3,70	2,42 3,66	2,41 3,62	2,40 3,60
12	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,92 4,65	2,80 4,39	2,76 4,30	2,76 4,30	2,69 4,16	2,64 4,05	2,60 3,98	2,54 3,86	2,46 3,70	2,42 3,61	2,40 3,56	2,36 3,49	2,35 3,46	2,32 3,41	2,31 3,38	2,30 3,36
13	3,41 5,74	3,18 5,20	3,02 4,86	2,92 4,62	2,84 4,44	2,72 4,19	2,67 4,10	2,67 4,10	2,60 3,96	2,55 3,85	2,51 3,78	2,46 3,67	2,38 3,51	2,34 3,42	2,32 3,37	2,28 3,30	2,26 3,27	2,24 3,21	2,22 3,18	2,21 3,16
14	3,34 5,56	3,11 5,03	2,96 4,69	2,85 4,46	2,77 4,28	2,65 4,03	2,60 3,94	2,60 3,94	2,53 3,80	2,48 3,70	2,44 3,62	2,39 3,51	2,31 3,34	2,27 3,26	2,24 3,21	2,21 3,14	2,19 3,11	2,16 3,06	2,14 3,02	2,13 3,00
15	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,70 4,14	2,59 3,89	2,55 3,80	2,55 3,80	2,48 3,67	2,43 3,56	2,39 3,48	2,33 3,36	2,25 3,20	2,21 3,12	2,18 3,07	2,15 3,00	2,12 2,97	2,10 2,92	2,08 2,89	2,07 2,87
16	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,54 3,78	2,49 3,69	2,49 3,69	2,42 3,55	2,37 3,45	2,33 3,37	2,28 3,25	2,20 3,10	2,16 3,01	2,13 2,96	2,09 2,89	2,07 2,86	2,04 2,80	2,02 2,77	2,01 2,75
50	2,79 4,20	2,56 3,72	2,40 3,41	2,29 3,18	2,20 3,02	2,07 2,87	2,02 2,70	2,02 2,70	1,95 2,56	1,90 2,46	1,85 2,39	1,78 2,26	1,69 2,10	1,63 2,00	1,60 1,94	1,55 1,86	1,52 1,82	1,48 1,76	1,46 1,71	1,44 1,68
200	2,65 3,88	2,41 3,41	2,26 3,11	2,14 2,90	2,05 2,73	1,92 2,50	1,87 2,41	1,87 2,41	1,80 2,28	1,74 2,17	1,69 2,09	1,62 1,97	1,52 1,79	1,45 1,69	1,42 1,62	1,35 1,53	1,32 1,48	1,26 1,39	1,22 1,33	1,19 1,28
∞	2,60 3,78	2,37 3,32	2,21 3,02	2,09 2,80	2,01 2,64	1,88 2,41	1,83 2,32	1,83 2,32	1,75 2,18	1,69 2,07	1,64 1,99	1,57 1,87	1,46 1,69	1,40 1,59	1,35 1,52	1,28 1,36	1,24 1,36	1,17 1,25	1,11 1,15	1,00 1,09

Примечание. * В числителе – для $F_{0,95}$, в знаменателе – для $F_{0,95}$. ** Степени свободы для меньшей дисперсии.

6. Значения критерия χ^2 (Пирсона)

Степень свободы, ν	Уровни вероятности, *P*		
	0,95	0,99	0,999
1	3,841	6,635	10,827
2	5,991	9,210	13,815
3	7,815	11,345	16,268
4	9,488	13,277	18,465
5	11,070	15,086	20,517
6	12,592	16,812	22,457
7	14,067	18,475	24,322
8	15,507	20,090	26,125
9	16,919	21,666	27,877
10	18,307	23,209	29,588
11	19,675	24,725	31,264
12	21,026	26,217	32,909
13	22,362	27,688	34,528
14	23,685	29,141	36,123
15	24,996	30,578	37,697
16	26,296	32,000	39,252
17	27,587	33,409	40,790
18	28,869	34,805	42,312
19	30,144	36,191	43,820
20	31,410	37,566	45,315
21	32,671	38,932	46,797
22	33,924	40,289	48,268
23	35,172	41,638	49,728
24	36,415	42,980	51,179
25	37,652	44,314	52,620
26	38,885	45,642	54,052
27	40,113	46,963	55,476
28	41,337	48,278	56,893
29	42,557	49,588	58,302
30	43,773	50,892	59,703

7. Минимальные существенные значения коэффициентов корреляции

ν	Уровень вероятности (*P*)		ν	Уровень вероятности (*P*)	
	0,95	0,99		0,95	0,99
3	0,94	0,99	26	0,37	0,48
4	0,84	0,93	27	0,37	0,48
5	0,75	0,87	28	0,36	0,46
6	0,71	0,83	29	0,36	0,46
7	0,67	0,80	30	0,35	0,45
8	0,63	0,77	35	0,33	0,42
9	0,60	0,74	40	0,30	0,39
10	0,58	0,71	45	0,29	0,37
11	0,55	0,68	50	0,27	0,36
12	0,53	0,66	60	0,25	0,33
13	0,51	0,64	70	0,23	0,30
14	0,50	0,62	80	0,22	0,28
15	0,48	0,61	90	0,21	0,27
16	0,47	0,59	100	0,20	0,25
17	0,46	0,58	125	0,17	0,23
18	0,44	0,56	150	0,16	0,21
19	0,43	0,56	200	0,14	0,18
20	0,42	0,54	300	0,11	0,15
21	0,41	0,53	400	0,10	0,13
22	0,40	0,52	500	0,09	0,12
23	0,40	0,51	700	0,07	0,10
24	0,39	0,50	900	0,06	0,09
25	0,38	0,49	1000	0,06	0,09

8. Соотношение между *r* и *z′* для *z′* значений от 0 до 5*

z′	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0690	0,0798	0,0898
0,1	0,0997	0,1096	0,1194	0,1293	0,1391	0,1489	0,1587	0,1684	0,1781	0,1878
0,2	0,1974	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821
0,3	0,2913	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714
0,4	0,3800	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542
0,5	0,4621	0,4700	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	0,5370	0,5441	0,5511	0,5581	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980
0,7	0,6044	0,6107	0,6169	0,6231	0,6291	0,6352	0,6411	0,6469	0,6527	0,6584
0,8	0,6640	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	07114
0,9	0,7163	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574
1,0	0,7616	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969
1,1	0,8005	0,8041	0,8076	08110	08144	0,8178	0,8210	0,8243	0,8275	0,8306
1,2	0,8337	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591
1,3	0,8617	0,8643	0,8668	0,8693	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832
1,4	0,8854	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033
1,5	0,9052	0,9069	0,9087	09104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9202
1,6	0,9217	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9342

z'	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,7	0,9354	0,9367	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458
1,8	0,9468	0,9478	0,9498	0,9488	0,9508	0,9518	0,9527	0,9536	0,9545	0,9554
1,9	0,9562	0,9571	0,9579	0,9587	0,9595	0,9603	0,9611	0,9619	0,9626	0,9633
2,0	0,9640	0,9647	0,9654	0,9661	0,9668	0,9674	0,9680	0,9687	0,9693	0,9699
2,1	0,9705	0,9710	0,9716	0,9722	0,9727	0,9732	0,9738	0,9743	0,9748	0,9753
2,2	0,9757	0,9762	0,9767	0,9771	0,9776	0,9780	0,9785	0,9789	0,9793	0,9797
2,3	0,9801	0,9805	0,9809	0,9812	0,9816	0,9820	0,9823	0,9827	0,9830	0,9834
2,4	0,9837	0,9840	0,9843	0,9846	0,9849	0,9852	0,9855	0,9858	0,9861	0,9863
2,5	0,9866	0,9869	0,9871	0,9874	0,9876	0,9879	0,9881	0,9884	0,9886	0,9888
2,6	0,9890	0,9892	0,9895	0,9897	0,9899	0,9901	0,9903	0,9905	0,9906	0,9908
2,7	0,9910	0,9912	0,9914	0,9915	0,9917	0,9919	0,9920	0,9922	0,9923	0,9925
2,8	0,9926	0,9928	0,9929	0,9931	0,9932	0,9933	0,9935	0,9936	0,9937	0,9938
2,9	0,9940	0,9941	0,9942	0,9943	0,9944	0,9945	0,9946	0,9947	0,9949	0,9950
3,0	0,9951									
4,0	0,9993									
5,0	0,9999									

Примечание. * Цифры таблицы являются значениями коэффициента корреляции r, соответствующими значениям z', указанным слева и сверху таблицы.

9. Значения коэффициента корреляции рангов Спирмена для двусторонних пределов уровня значимости α

n \ α	0,20	0,10	0,05	0,02	0,01	0,002
4	0,8000	0,8000				
5	0,7000	0,8000	0,9000	0,9000		
6	0,6000	0,7714	0,8286	0,8857	0,9429	
7	0,5357	0,6786	0,7450	0,8571	0,8929	0,9643
8	0,5000	0,6190	0,7143	0,8095	0,8571	0,9286
9	0,4667	0,5833	0,6833	0,7667	0,8167	0,9000
10	0,4424	0,5515	0,6364	0,7333	0,7818	0,8667
11	0,4182	0,5273	0,6091	0,7000	0,7455	0,8364
12	0,3986	0,4965	0,5804	0,6713	0,7273	0,8182
13	0,3791	0,4780	0,5549	0,6429	0,6978	0,7912
14	0,3626	0,4593	0,5341	0,6220	0,6747	0,7670
15	0,3500	0,4429	0,5179	0,6000	0,6536	0,7464
16	0,3382	0,4265	0,5000	0,5824	0,6324	0,7265
17	0,3260	0,4118	0,4853	0,5637	0,6152	0,7083
18	0,3148	0,3994	0,4716	0,5480	0,5975	0,6904
19	0,3070	0,3895	0,4579	0,5333	0,5825	0,6737
20	0,2977	0,3789	0,4451	0,5203	0,5684	0,6586
21	0,2909	0,3688	0,4351	0,5078	0,5545	0,6455
22	0,2829	0,3597	0,4241	0,4963	0,5426	0,6318

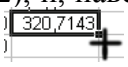
n \ α	0,20	0,10	0,05	0,02	0,01	0,002
23	0,2767	0,3518	0,4150	0,4852	0,5306	0,6186
24	0,2704	0,3435	0,4061	0,4748	0,5200	0,6070
25	0,2646	0,3362	0,3977	0,4654	0,5100	0,5962
26	0,2588	0,3299	0,3894	0,4564	0,5002	0,5856
27	0,2540	0,3236	0,3822	0,4481	0,4915	0,5757
28	0,2490	0,3175	0,3749	0,4401	0,4828	0,5660
29	0,2443	0,3113	0,3685	0,4320	0,4744	0,5567
30	0,2400	0,3059	0,3620	0,4251	0,4665	0,5479

10. Алгоритм вычисления основных показателей описательной статистики и критерия Стьюдента в Microsoft Office Excel 2003

Решение рассмотрим на примере двух выборок, в которых приведены площади фермерских хозяйств в Брестской и Гомельской областях. Первоначально набираем в ячейках A2:A3 названия областей, в B2:H2 и B3:H3 цифры площадей для каждой области (рис. 1).

	A	B	C	D	E	F	G	H
1	Площадь фермерских хозяйств							
2	Брестская обл.	300	305	315	320	330	335	340
3	Гомельская обл	180	175	190	185	187	197	200

Рис. 1. Исходные данные для расчетов

Основными статистическими показателями, характеризующими данные выборки, являются: *среднее* (*M*), *медиана* (Me), *наименьшее, наибольшее, коэффициент вариации* (*V*), *среднеквадратическое отклонение* (σ), *дисперсия* (σ^2). Среднее (*M*) находится следующим образом: выполняем команду *Функция* из меню *Вставка* (или нажимаем на иконку $\mathbf{f_x}$ на панели инструментов), далее в категориях *Статистические* выбираем функцию СРЗНАЧ (рис. 2), сворачиваем появившееся окно нажатием на кнопку  напротив поля **Число 1.** Выделяем ячейки со значениями площадей для первой области (B2:H2), разворачиваем окно, нажав на эту же кнопку, и жмем *OK*. Для второй области можно не делать описанную выше процедуру, а воспользоваться функцией автозаполнения: выделяем ячейку с найденным средним значением для первой области (I2), и, наведя курсор на правый край клетки I2 до превращения курсора в «крестик»: 

, удерживая левую кнопку мыши, растягиваем выделение на нижележащую клетку (I3), в которой появится значение для второй области.

Аналогичным способом находим медиану (команда МЕДИАНА(B2:H2)), наименьшее =МИН(B2:H2) и наибольшее =МАКС(B2:H2) значения, коэффициент вариации =СТАНДОТКЛОН(B2:H2)/СРЗНАЧ(B2:H2)*100, среднеквадратическое отклонение =СТАНДОТКЛОН(B2:H2) и дисперсию =ДИСП(B2:H2). При помощи автозаполнения производим расчет для второй области. MS Excel может производить вычисления при наборе функции вручную, при этом следует помнить, что команды набираются на русском языке, а буквенные обозначения ячеек – латинские.

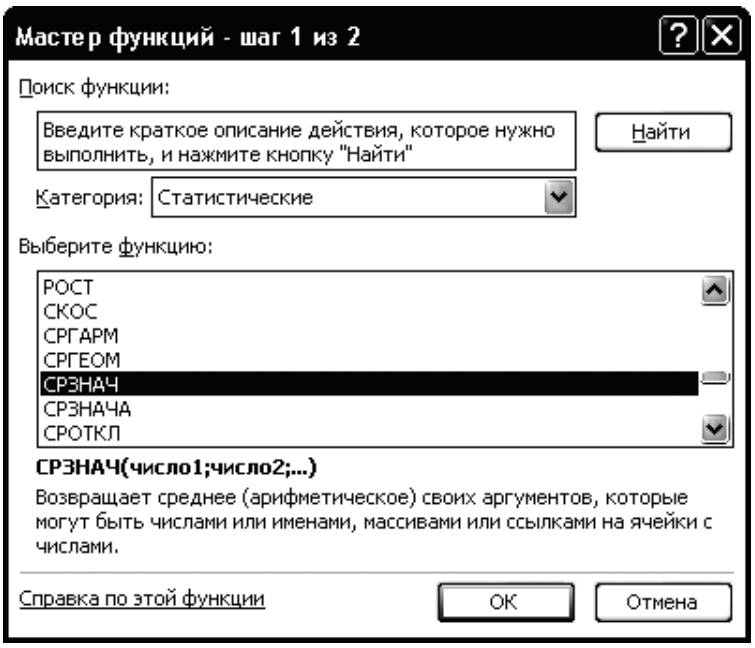


Рис. 2. Окно выбора функции

Расчет базовых статистических показателей может производиться с использованием надстройки (опции) «*Пакет анализа*», которая позволяет оперативно получить значения показателей описательной статистики. По умолчанию эта опция не установлена, поэтому для ее активации необходимо с помощью команды *Надстройки* из меню *Сервис* открыть окно диалога «*Надстройки*» и в нем установить флажок для компоненты «*Пакет анализа*». После нажатия *ОК* меню *Сервис* будет дополнено командой *Анализ данных*.

Для расчета показателей выполняем команду *Анализ данных* из меню *Сервис*, в диалоговом окне *Анализ данных* выбираем *Описательная статистика*, в поле «*Входной интервал*» указываем наш (клетки A2:H3), в поле «*группирование*» выбираем «*по строкам*», ставим галочку у «*Метки в первом столбце*» в «*Параметрах вывода*», выбираем «*Выходной интервал*» и указываем там ячейку B5 или другую свободную, отмечаем параметры «*Итоговая статистика*» и «*Уровень надежности*» (значение можно изменять, в нашем случае указываем 95 %), нажимаем *ОК*.

Нахождение сходства или отличия между двумя выборками с помощью *t*-теста (критерия Стьюдента). Выбор конкретной команды зависит от типа выборки (зависимая/независимая) и от величин дисперсий. Так, для **независимой** выборки с **различными дисперсиями** выполняются следующие действия: *Сервис – Анализ данных – Двухвыборочный t-тест с различными дисперсиями – ОК*. Для **независимой** выборки с **одинаковыми дисперсиями** алгоритм следующий: *Сервис – Анализ данных – Двухвыборочный t-тест с одинаковыми дисперсиями – ОК*, для **сопряженной** выборки: *Сервис – Анализ данных – Парный двухвыборочный t-тест для средних – ОК*.

В поле «*интервал переменной* 1» указываем интервал значений для первой области (A2:H2), в поле «*интервал переменной* 2» – интервал значений для второй области (A3:H3), ставим галочку у окна «*Метки*», далее выбираем «*Выходной интервал*» и указываем там ячейку G5 (или другую свободную), нажимаем *ОК*.

В полученных данных *df* – число степеней свободы; *t-статистика* – критерий Стьюдента (фактический); *t-критическое двухстороннее* – критерий Стьюдента (таб-

личный). На основании соотношения *t-статистики* (берется по модулю) и *t-критического двухстороннего* делается вывод о достоверности различия выборок.

11. Алгоритм проведения однофакторного дисперсионного анализа в Microsoft Office Excel 2003

Рассмотрим с помощью дисперсионного анализа влияние внесения удобрений на урожайность сельскохозяйственных культур по различным вариантам опыта. В MS Excel набираем исходные данные из индивидуального задания по образцу, показанному на рис. 3:

	A	B	C	D	E
1	Влияние удобрений на урожай с/х культур			Повторности	
2	Варианты	I	II	III	IV
3	фон	30,9	28	30,1	28,4
4	фон+100	33,9	36,3	34,3	36
5	фон+200	44,6	44,3	45,6	46,6
6	фон+300	51,1	48,8	50,4	51,4

Рис. 3. Исходные данные

Для анализа используем надстройку «*Пакет анализа*». Для ее активации необходимо с помощью команды *Надстройки* из меню *Сервис* открыть окно диалога «*Надстройки*» и в нем установить флажок для компоненты «*Пакет анализа*». После нажатия кнопки *ОК* меню *Сервис* будет дополнено командой *Анализ данных* (если надстройка вызывалась ранее и не отключалась, то этот пункт можно пропустить).

Для расчета показателей выполняем последовательность команд *Сервис – Анализ данных – Однофакторный дисперсионный анализ – ОК,* в поле «*Входной интервал*» указываем наш интервал (A3:E6 для приведенного примера), ставим галочки напротив показателей *по строкам* и *метки в первом столбце*; в «*Параметрах вывода*» выбираем «*Выходной интервал*» и указываем там ячейку на этом же листе, значение *Альфа* оставляем прежним, равным 0,05, нажимаем *ОК*.

Результаты дисперсионного анализа будут состоять из двух таблиц. В первой таблице для каждого столбца исходной таблицы, в которых располагаются анализируемые группы, приведены числовые параметры: количество чисел (счет), суммы по строкам, средние дисперсии по строкам. Во второй части результатов *MS Excel* использует следующие обозначения: *SS* – сумма квадратов; *df* – степени свободы; *MS* – средний квадрат (дисперсия); *F* – *F*-статистика Фишера (фактическое значение); *P-значение* – значимость критерия Фишера (критерий является значимым, если величина данного параметра менее 0,05); *F-критическое* – критическое (табличное) значение *F*-статистики при $P = 0{,}05$. Путем сравнения *F* и *F-критического* делаем вывод. Для данного примера эти значения будут соответственно 252,646 и 3,490, поэтому положительное влияние удобрений на урожайность доказано.

Если сделать дисперсионный анализ для повторностей опыта (действия аналогичны первоначальным, только вместо показателя *по строкам* выставляется значение *по столбцам* и интервал меняется на B2:E6), то $F < F$-*критического*, что и ожидалось, поскольку изменения фактора внутри повторности не происходило.

12. Алгоритм проведения корреляционного и регрессионного анализов в Microsoft Office Excel 2003

Проверим зависимость между балом пашни (x) и урожайностью многолетних трав (y), для чего набираем в ячейках A2:K3 следующие данные:

x	43	42	38	36	33	45	40	45	36	32
y	33,2	18,6	28,4	26,5	30,9	31,8	32,4	30,6	26,8	24,4

Строим точечную диаграмму: выделяем набранную таблицу (ячейки A2:K3) и жмем на пиктограмму  на панели инструментов или *Вставка – Диаграмма,* в закладке *Стандартные* выбираем *Точечная* и первый сверху из имеющихся примеров, жмем *Далее,* в закладке *Диапазон данных* отмечаем *Ряды, в строках – Далее.* В закладке *Заголовки* в окошке *Ось X (категорий)* набираем «*Балл пашни*», в окошке *Ось Y (значений)* – «*Урожайность многолетних трав*», в закладке *Легенда* снимаем галочку с показателя «*Добавить легенду» – Далее – Поместить диаграмму на имеющемся листе – Готово*.

Добавляем линию тренда, для чего кликаем на маркере точки данных правой клавишей и выбираем пункт *Добавить линию тренда* (см. рис. 4).

В закладке *Тип* выбирается *Линейная,* в закладке *Параметры* отмечаются пункты *показывать уравнение на диаграмме* и *поместить на диаграмму величину достоверной аппроксимации – ОК*. В итоге будет построена линия тренда и составлено уравнение линейной регрессии. Находим артефакты – значения, которые сильно отдалены от линии тренда и не вписываются в общую картину (рис. 5). Более правильно выявлять артефакт через расчеты, приведенные в п. 1.2 данного пособия. Удаляем эти значения из таблицы данных (в указанном примере очищаются от цифр ячейки C2:C3), MS Excel автоматически пересчитает уравнение регрессии. В некоторых случаях (при нелинейной корреляции) можно использовать другие варианты линий тренда, например, логарифмическую, степенную или экспоненциальную.

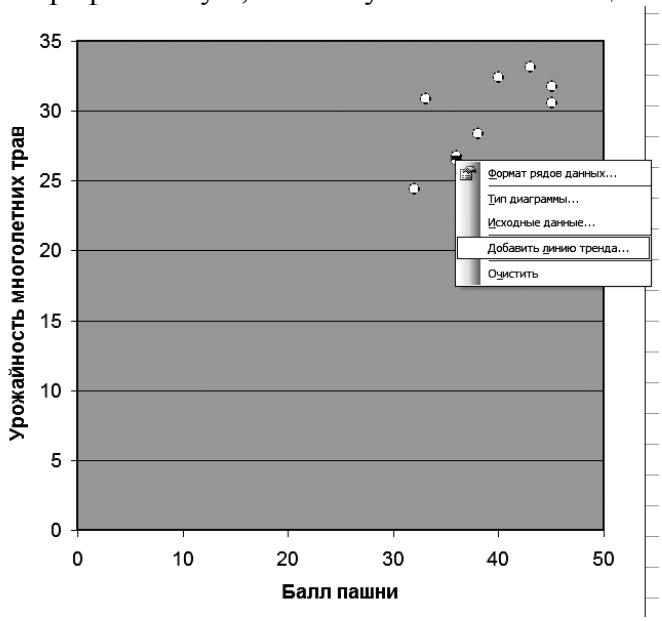


Рис. 4

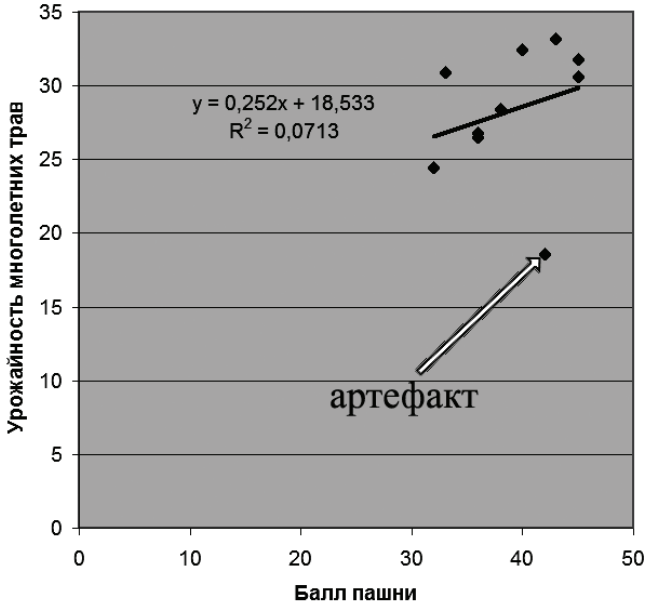


Рис. 5

Рассчитываем коэффициент корреляции установив курсор в клетку B5, используя команду КОРРЕЛ: *Вставка – Функция* (или иконка $\mathbf{f_x}$ на панели инструментов) – выбираем в категориях *Статистические* функцию КОРРЕЛ – сворачиваем появившееся окно нажатием на кнопку напротив поля *Массив 1*. Выделяем ячейки со значениями ***x*** (B2:K2), далее в поле *Массив 2* выделяем ячейки со значениями ***y*** (B3:K3), разворачиваем окно, нажав на эту же кнопку и жмем *OK*.

Оцениваем значимость коэффициента корреляции (***r***) по критерию Стьюдента по формуле $t_r = \sqrt{N-2} / \sqrt{1-r^2}$ и сравниваем с табличным (критическим) значением, если фактическое значение больше критического, то корреляционная связь существенна, если меньше – недостоверна (вид формул на рис. 6).

	A	B
1	Между балом пашни и урожайностью	многолетних трав
2	x	43
3	y	33,2
4		
5	Коэф. корреляции	=КОРРЕЛ(B2:K2;B3:K3)
6	Критерий Стьюдента	=B5*КОРЕНЬ(СЧЁТ(B2:K2)-2)/КОРЕНЬ(1-B5*B5)
7	Критическое значение критерия Стьюдента	=СТЬЮДРАСПОБР(0,05;СЧЁТ(B2:K2)-2)

Рис. 6

Регрессионный анализ проводится с помощью надстройки «*Пакет анализа*», последовательность команд *Сервис – Анализ данных – Регрессия,* в поле «*Входной интервал*» указываем значения для *Y* и *X* (A3:K3 и A2:K2 соответственно), в «*Параметрах вывода*» выбираем «*Выходной интервал*» и указываем там ячейку на этом же листе, отмечаем параметры «*Уровень надежности*» (значение можно изменять, в нашем случае указываем 95 %) и «*Метки*», нажимаем *OK*. Если удалялся артефакт, то необходимо скопировать первоначальные значения в другие ячейки, поскольку значения во входном интервале должны быть непрерывными.

13. Алгоритм проведения кластерного анализа в Statsoft Statistica 6.0

Проведем кластерный анализ для областей Беларуси по показателям внесения удобрений и урожайности ряда сельскохозяйственных культур.

Допускается выполнение работы по двум вариантам (на выбор пользователя):

а) Набор исходных данных в MS Excel. Открыть MS Excel. Набрать исходные данные, указанные в табл. 1, в ячейках диапазона A1:F6 Листа 1. Сохранить введенные данные и закрыть файл.

Таблица 1

16,6	212	27,3	175	38,9	12,4
13,9	193	15,7	156	40,1	11,8
16,3	226	25,3	186	28,6	13,9
13,5	240	29,4	178	43,5	15,4
11,6	205	25,9	193	33,6	10,3
15,5	231	27,5	185	32,5	14,4

Запустить программу **Statistica** (через *Пуск – Все программы* или ярлык на рабочем столе), открыть в ней набранный в Excel файл (*File – Open* или через пиктограмму  на панели инструментов, в появившемся окне укажите путь к файлу с вышеуказанной таблицей, не забудьте выбрать в окне *«Тип файлов» Excel files (.xls)*). Далее в появившемся диалоговом окне выбираем *Import selected sheet to a Spreadsheet*, затем в следующем окне выбираем *Лист 1 – OK*, в следующем окне ничего не изменяем и сразу жмем *OK*.

б) Подобную таблицу можно сразу создать путем набора в программе **Statistica**, пример **а** показывает на возможность импорта данных из MS Excel.

Переименовать в **Statistica** строки последовательно: *Брестская, Витебская, Гомельская, Гродненская, Могилевская, Минская*, для чего нужно дважды щелкнуть по ним левой клавишей мышки, а столбцы (Var 1, Var 2 и т. д.) – в поле *Name*, после двойного щелчка левой клавиши мыши соответственно набираем: *органич. удобр., т/га; минерал. удобр., кг/га; зерновые, ц/га; картофель, ц/га; кормовые травы, ц/га; зернобобовые, ц/га*.

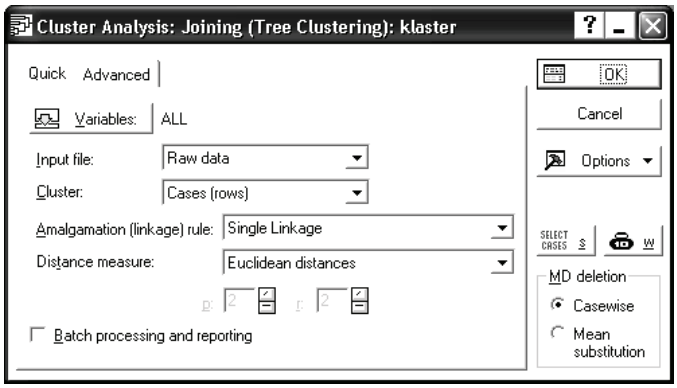


Рис. 7

Проводим кластерный анализ, для чего выполняем следующие действия: *Statistics – Multivariate Exploratory Techniques – Cluster analysis – Joining tree clustering*

(оно выбрано по умолчанию) – *OK*. В следующем диалоговом окне выбираем закладку *Advanced* – жмем на кнопку *Variables,* там отмечаем все переменные (выделяем левой клавишей мыши при нажатой клавише *Shift* или просто щелкнем на *Select All*) – *OK*. В полях *Input file* ставим *Raw data, Kluster – Cases (rows), Amalgamation (linkage) rule – Single Linkage, Distance Measure – Euclidean distances.* Если ваши параметры соответствуют представленным на рис. 7, то жмем *OK*.

Далее в появившемся окне нажимаем *Summary.* Появится дендрограмма с разбитием данных на кластеры. После этого нажимаем на *Joining result:имя файла* (слева в самом низу программного окна). Там на закладке *Advanced* выбираем по очереди показатели: *Distance matrix, Descriptive statistics* и *Matrix.* Так же можно выбрать вертикальное расположение древа (показатель *Vertical icicle plot*). Полученный график и таблицы используются для интерпретации данных анализа.

14. Алгоритм проведения факторного анализа в Statsoft Statistica 6.0

С помощью факторного анализа оценим плодородие почв в Минском районе под влиянием природных и агротехногенных факторов, для чего набираем в программе **Statistica** табл. 2:

Таблица 2

8	2,7	3,5	0,3	2,5	4,6	65	21
11	2,9	4,6	0,2	2,4	4,4	63	27
13	3	4,7	0,1	2,3	4,5	64	26
7	1,9	3,1	0,3	2,1	4,5	54	23
8	2,1	3,6	0,4	2,6	4,7	42	24
9	2,3	4,2	0,1	2,7	5,1	43	24
14	2,7	4,1	0,5	2,8	5	60	22
13	2,8	4,5	0,7	2	4,4	52	23
12	2,6	4,7	0,2	2,2	4,5	47	25
14	2,4	4,8	0,4	2,4	4,6	42	26
9	2,2	3,9	0,6	2,6	4,9	51	27

Переименовываем столбцы в **Statistica** (Var 1, Var 2 и т. д.), для чего нужно дважды щелкнуть по ним левой клавишей мышки и набрать в поле *Name* соответственно: *органические удобрения, т/га; минеральные удобрения, ц/га; дозы извести, т/га; пестициды, кг/га; гумус, т/га; гидролитическая кислотность* (*Н*)*, мэкв/100 г; влажность почвы, %; физическая глина, %.*

Проводим факторный анализ, для чего выполняем следующие действия: *Statistics – Multivariate Exploratory Techniques – Factor analysis – OK.* В следующем диалоговом окне жмем на *Variables,* там отмечаем все переменные (выделяем левой клавишей мыши при нажатой клавише *Shift* или просто щелкаем на *Select All*) – *OK*. В поле *Input file* ставим *Raw data,* в поле *MD deletion – Casewise* (выставлено по умолчанию) и жмем *OK*.

В следующем окне переходим на закладку *Advanced*, где по умолчанию выбраны *Principal components*, а значение *Max. no. of factors* равно 2. Если выбраны другие значения, то устанавливаем вышеуказанные и жмем *OK.*

В полученном окне, на закладке *Quick* жмем на *Eigenvalues*. В получившейся таблице *Eigenvalues (Factors)* приведены: 1) собственные значения факторов, которые были выделены; 2) процент объясненной дисперсии; 3) кумулятивные собственные значения и 4) кумулятивный процент объясненной дисперсии. В нашем случае выделялось два фактора.

После этого возвращаемся в диалоговое окно *Factor Analysis Results: factor.* (слева в самом низу программного окна), где на закладке *Loadings* выбираем в окне *Factor rotation* показатель *Varimax raw,* после чего нажимаем на кнопку *Summary: Factor loadings* и *Plot of loadings, 2D.* На закладке *Explained Variance* нажимаем по очереди на *Scree plot, Communalities.*

Далее переходим на закладку *Descriptives* и нажимаем на кнопку *Review correlations, means, standard deviations*, в новом окне на закладке *Quick* поочередно нажимаем на кнопки *Means & SD* и *Correlations*. Вернуться в окно *Factor Analysis Results: factor* можно, нажав на *Cancel*.

Полученные график и таблицы используются для интерпретации данных анализа.

15. Решение задачи на оптимальность

Требуется обосновать оптимальные размеры отраслей фермерского хозяйства, позволяющие сохранить плодородие пашни и получить максимум прибыли.

Исходная информация.

1. Фермер имеет 40 + К пашни, 1300 + 10К чел.-дн. годового труда, 600 + 5К ц единиц кормов с пастбищ и сенокосов.

2. Расход ресурсов и выход продукции на единицу отрасли приведен в табл. 3:

Таблица 3

Показатели	Зерновые	Картофель	Многолетние травы	Коровы
Площадь пашни, га	1	1	1	40
Затраты труда, чел.-дн.	10	32	3	24
Баланс гумуса, т/га	– 0,9	– 1,6	+0,5	
Расход кормов, ц. к. ед.				50
Выход кормов, ц к. ед.	10	15	25	
Выход навоза от коровы, т				9
Прибыль, у. д. е.	50 + 0,5К	60 – 0,5К		100 – К

3. Коэффициент перевода органического удобрения в гумус 0,1.

4. Площадь зерновых должна быть не менее 10 га.

Решите задачу симплексным методом и проведите анализ.

Покажем первоначальные условия задачи на листе Excel в виде рабочего листа «Оптимизация».

Решение подобной задачи возможно при помощи надстройки «*Поиск решения*», для ее активации необходимо с помощью команды *Надстройки* из меню *Сервис* открыть окно диалога «*Надстройки*» и в нем установить флажок для компоненты «*Поиск решения*». После нажатия *ОК* меню *Сервис* будет дополнено командой *Поиск решения*.

В меню *Сервис* Выбираем команду *Поиск решения*. Установить целевую ячейку, которая должна принимать максимальное, минимальное или конкретное значение, в нашем случае это ячейка F10. Ставим отметку тип «*максимальное значение*». В поле *изменяя ячейки* указываем диапазоны ячеек, оптимальные значения которых требуется найти (B3, C3, D3, E4). Вводим условия ограничения, для чего здесь же вызываем диалоговое окно «*Ограничение*», щелкнув по *добавить*. В диалоговом окне *добавление ограничения* в окошке *ссылка на ячейку* вносим адрес ячейки с функцией *ограничения*, где указывается число или адрес ячейки, содержащей ограничения (табл. 4). Между ними проставить знаки <= или =>. После ввода всех ограничений выбирают «*ОК*».

	A	B	C	D	E	F	G
1	Расчет оптимальной программы развития сельского хозяйства						
2	Показатели	Зерновые	Картофель	Мн. травы	Коровы	Итого	Имеется
3	Площадь, га					=СУММ(B3:D3)	40
4	Поголовье коров, гол						
5	Затраты труда, чел.-дн.	=B3*10	=C3*32	=D3*3	=E4*24	=СУММ(B5:E5)	1300
6	Выход кормов, ц к.ед.	=B3*10	=C3*15	=D3*25		=СУММ(B6:D6)+600	
7	Расход кормов, ц к.ед.				=E4*50	=E4*50	
8	Баланс гумуса, т	=B3*0,9	=C3*1,6	=D3*0,5		=B8+C8-D8	
9	Выход навоза				=E4*9		
10	Прибыль, у.д.е.	=B3*50	=C3*60		=E4*100	=СУММ(B10;C10;E10)	

Рис. 8

Таблица 4

Ограничения	Описание
B3 : D3 ≥ 0	Площадь посева сельскохозяйственных культур не может быть отрицательной.
E4 ≥ 0	Поголовье коров не может принимать отрицательные значения.
F3 ≤ G3	Общая площадь посева сельхозкультур не должна быть больше площади пашни.
F5 ≤ G5	Затраты труда на возделывание сельхозкультур в растениеводстве и животноводстве не могут превышать имеющиеся ресурсы труда.
F7 ≤ F6	Расход кормов в животноводстве не должен превышать выхода кормов с отраслей растениеводства с учетом их заготовки на сенокосах и пастбищах.
B3 ≥ 10	Площадь зерновых культур должна быть не менее 10 га.
F8 ≤ E9	Вынос (минерализация) гумуса с урожаем сельхозкультур не должен превышать его поступления с отрасли животноводства.

Появляется диалоговое окно «*Поиск решения*», в нем для решения задачи щелкаем по кнопке *выполнить*. После завершения расчетов появится диалоговое окно «*Результаты поиска решений*». В нем помечаем пункт «*сохранить найденное решение»* и указываем необходимый тип отчета (*результаты, устойчивости, пределы)*. Далее нажимаем *ОК* для сохранения результата.

Если решение неверно, то появляется:

- значения целевой ячейки не сходятся;
- поиск не может найти подходящее решение;
- условия для линейной модели не удовлетворительны и др.

При положительном решении выбрать «*Сохранить сценарий*», при отрицательном – «*Восстановить исходные данные*».

ОГЛАВЛЕНИЕ

Учебное издание

Чертко Николай Константинович
Карпиченко Александр Александрович

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ГЕОГРАФИИ

Учебно-методическое пособие

Редактор *Н. Ф. Акулич*
Художник обложки *Т. Ю. Таран*
Технический редактор *Т. К. Раманович*
Корректор *Т. С. Петроченко*
Компьютерная верстка *Е. М. Товчковой*

Подписано в печать 05.05.2009.
Формат 60×84/16. Бумага офсетная.
Гарнитура Таймс. Печать офсетная.
Усл. печ. л. 11,62. Уч.-изд. л. 13,53.
Тираж 100 экз. Зак. 596.

Белорусский
государственный университет.
ЛИ № 02330/0494425 от 08.04.2009.
220030, Минск,
проспект Независимости, 4.

Отпечатано
с оригинала-макета заказчика.
Республиканское
унитарное предприятие
«Издательский центр
Белорусского
государственного университета».
ЛП № 02330/0494178 от 03.04.2009.
220030, Минск,
ул. Красноармейская, 6.