TEMPUS-TACIS Project CD_JEP 21242-00/Ukr
«Education Development on Environmentally Safe Energetics»

ПРОЕКТ ТЕМПУС-ТАСИС:
"Развитие образования в области экологически безопасной энергетики"

Проект относится к высшему образованию в области окружающей среды применительно к альтернативной энергетике.

Цель Проекта - развитие в Украине образования для подготовки менеджеров регионального развития альтернативной энергетики и энергоэкологии.

В Украине базовыми вузами-партнерами являются:
- Харьковский национальный университет имени В.Н.Каразина (вуз-координатор)
- Таврический национальный университет имени В.И.Вернадского.

Зарубежными вузами-партнерами Проекта являются:
- Политехнический университет Каталонии (вуз-координатор)
- Научно-технологический институт университета Манчестера
- Национальный институт прикладных научных исследований Ренна

Серия учебных пособий предназначена для обеспечения обучения магистров и кандидатов наук по специальностям «Экономическая и социальная география» и «Экология и рациональное природопользование» для специализаций:
- природоохранный менеджмент и мониторинг,
- экоэнергетика и устойчивое развитие.

Заявки на учебные пособия направлять через Интернет по указанным адресам

www.univer.kharkov.ua

EDUCATION DEVELOPMENT ON ENVIRONMENTALLY SAFE ENERGETICS

РАЗВИТИЕ ОБРАЗОВАНИЯ В ОБЛАСТИ ЭКОЛОГИЧЕСКИ БЕЗОПАСНОЙ ЭНЕРГЕТИКИ

Education and Culture

Tacis

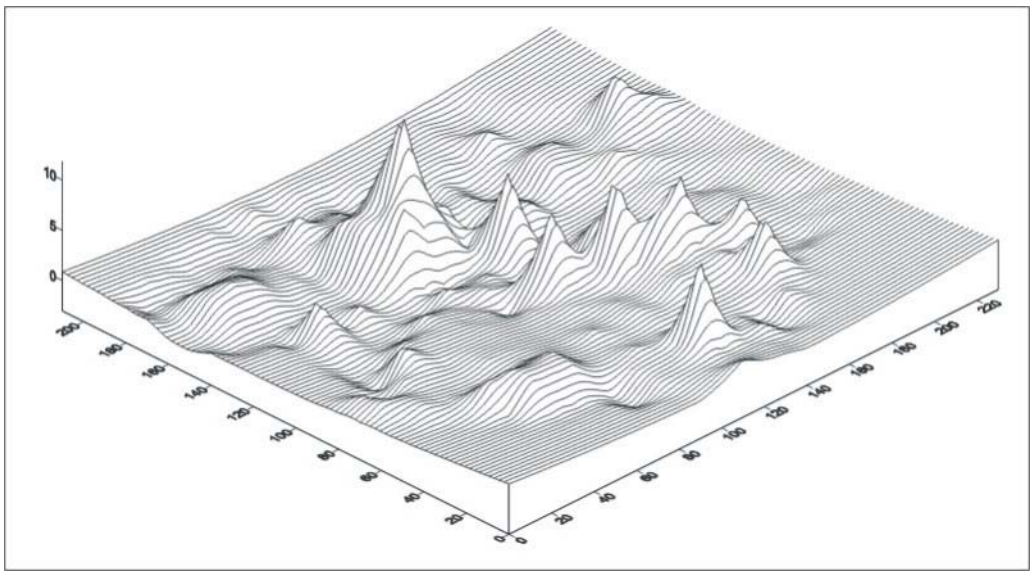
Статистические методы в прикладных географических исследованиях

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРИКЛАДНЫХ ГЕОГРАФИЧЕСКИХ ИССЛЕДОВАНИЯХ

Учебно-методическое пособие

Автор-составитель: Третьяков А.С.

Под редакцией проф. И.Г. Черванева



Харьков – 2004

Третьяков А.С. Статистические методы в прикладных географических исследованиях: Учебно-методическое пособие. Научный редактор: проф. И.Г. Черванев – Х.: Шрифт, 2004. – 96 с.

В учебном пособии приведены методы математической статистики, которые наиболее часто применяются в прикладной географии. Порядок освещения методов соответствует очередности их использования при проведении географических исследований: создание выборочной совокупности, ее анализ, выявление и оценка зависимости между выборками, многомерные методы анализа. Приведены примеры использования описанных методов.

Входит в состав серии учебно-методических пособий, изданных при финансовой поддержке Европейского Союза (международный проект Темпус-Тасис CD-JEP 21242-2000/Ukr)

Редакционная коллегия серии учебно-методических пособий: проф. И.Г.Черванев (председатель), профессоры В.А.Боков, А.К.Кузин (Украина), И. Клемеш (Великобритания, Научно-технологический институт университета г. Манчестер), С. Моншо (Национальный институт прикладных наук г.Ренн, Франция), Л. Пуиженер (Политехнический университет Каталонии, Испания).

© Третьяков А.С., 2004

© Харьковский национальный университет имени В.Н. Каразина, 2004

Содержание.

ТЕМА 1. СБОР ДАННЫХ.

Источником материала для прикладного географического исследования могут быть собственные экспериментальные исследования, аналитические данные других исследователей, географические карты специальные и общего назначения, фондовые материалы, литературные источники.

Проведем анализ возможных источников информации в сфере альтернативной энергетики.

Собственные экспериментальные исследования.

Рассмотрим возможность проведения собственного эксперимента, для определения потенциала производства энергии и, соответственно для выделения территорий, развитие альтернативной энергетики на которых представляется наиболее предпочтительным. Так, для определения ветрового либо солнечного потенциалов данной территории возможно при использовании метеорологических приборов, работа с которыми была изучена студентами в курсе «Метеорология с основами климатологии». Проведение экспериментов по измерению гидрологических показателей, необходимых для расчета гидроэнергетического потенциала также возможно при использовании методик, подробно изложеных в курсе «Гидрология».

Фондовые материалы.

Фондовые материалы, по мнению автора, являются оптимальным источником для проведения исследований. Работая с подобного рода данными, мы получаем наиболее достоверные результаты, так как фондовые материалы – это своего рода первоисточник для всех исследований. Возвращаясь к пункту 1.1. хотелось бы заметить, что неспосредственное измерение параметров на территории, имеющей значительную площадь в большинстве слуаев является довольно трудоемкой задачей. Поэтому собственные исследования могут служить лишь средством проверки достоверности экстраполяции данных более общего порядка. Так измерив среднемесячную скорость ветра в конкретной точке мы можем сравнить полученное значение с картой распределения скоростей ветра, постоенной по данным Госкомгидромет.

Основным препятствием при использовании фондовых данных является отсутствие открытого доступа к ним. Однако исследователь, по мнению автора, должен стремится к использованию именно этого вида данных.

Географические карты.

Географические карты довольно часто используются при анализе потенциала производства альтернативной энергии. Особенно это касается рассмотрения проблемы на наиболее общем уровне (общемировой, уровень группы государств, отдельно взятой страны). Так при помощи мировой карты радиационного баланса мы можем выделить территории, потенциал развития солнечной энергетики будет наибольшим; анализ карты распределения продуктивности биомассы [22] позволит выделить районы, наиболее перспективные для развития производства энергии из древесины.

Наличие региональных атласов для определенного района (например, [6]) позволит провести более детилизированное исследование.

Аналитические данные других исследователей.

При использовании данного вида данных студенту-географу, занимающемуся проблемами развития альтернативной энергетики придется столкнутся прежде всего с проблемой того, что большинство подобных оценок потенциала для той или иной территории проводилось, в основном, представителями инженерных специальностей, что предполагает некоторую специфику изложения материала. Однако благодаря данному виду источников информации можно получить эмпирические формулы для определения объема производства электроэнергии той или иной установки, расчет оптимальных технических показателей оптимальной установки для данной территории и, наконец, выбор оптимальной

технологии производства энергии из уже существующих вариантов установок. В подобного рода работах довольно часто встречаются уже проведенные оценки потенциала ветроэнергетического, солнечноэнергетического и др потенциалов. Эти данные, во-первых, могут служить материалом для сравнения с данными, полученными в проводимом студентом исследовании. Во-вторым, особенностью резельтатов оценок, изложенных в подобного рода работах, как правило, представлены в виде таблиц. Построение карт по этим данным с последующим анализом пространственного аспекта полученной картины уже можно считать прведением небольшого исследования.

Статистические сборники.

Статистические сборники, могут содержатьв себе информацию различного уровня детализации, в зависимости от их назначения. Так, хотелось бы заметить такие издания, как украинские статистические ежегодники (например [18], [21]), результаты переписей населения, отчеты Госкомстата и другних ведомств, (которые конечно, можно отнести и к фондовым материалам), а также другие издания, посвященные, как правило, определенным встречам, конференциям, семинарам и др.

Задание. *Опираясь на описанные виды и способы получения исходных данных, продумайте цель Вашего будущего исследования. Определившись с направлением Вашей работы, обдумайте, какие данные будут Вам необходимы для достижения поставленной в исследовании цели. Попробуйте добыть необходимые данные (в случае затруднений проконсультируйтесь в преподавателем).*

ТЕМА 2. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ПОЛУЧЕННЫХ КОЛИЧЕСТВЕННЫХ ДАННЫХ.

2.1 Общие сведения.

Одна из важнейших задач статистической обработки – установление или выявление таких параметров, которые в компактной форме достаточно полно характеризуют свойства исследуемой генеральной совокупности.

Генеральной совокупностью называют совокупность всех возможных наблюдений, которые могли бы быть проведены в соответствии с целью исследования [23]. Общее число членов генеральной совокупности называют *объемом генеральной совокупности.* Число членов в генеральной совокупности может быть конечным или бесконечным. Например, конечным числом членов (элементов) генеральной совокупности являются все свалки твердых бытовых отходов Харьковской области. Бесконечным числом членов может быть скоростей ветра в г. Харькове, величины которой колеблются по месяцам, годам, столетиям и т. д. В непрерывной генеральной совокупности можно вычленить дискретные промежутки, характеризующие определенное десятилетие или столетие, которые принимаются за генеральную совокупность.

Исследование объекта, т.е. генеральной совокупности, практически не проводят полностью. С целью экономии времени и средств прибегают к подбору характерных ключей или точек, пространственных или временных ограничений, которые принято называть выборкой из генеральной совокупности.

Выборочной совокупностью, или выборкой, называется совокупность N наблюдений, полученных с целью характеристики генеральной совокупности. Число членов выборочной совокупности называют *объемом выборки.* Выборочная совокупность дает оценку параметров, которые представляют собой константы, характеризующие распределение в генеральной совокупности [23].

Самым сложным является определение количества наблюдений в исследованиях для получения надежного представления о характере изменчивости признака в генеральной совокупности. Если объект исследуется впервые, то определить объем наблюдений практически очень трудно.

Чаще всего объем выборки N определяют по следующей формуле:

$$N = \frac{\sigma^2}{m^2} \qquad (1)$$

где m^2 – ошибка среднего арифметического; σ^2 – среднее квадратическое отклонение *(см. параграфы 2.4.1 и 2.4.2)*.

Задачей определения объема выборочной совокупности является получение достоверной информации о генеральной совокупности путем расчета минимального, но объективного количества наблюдений. Объем выборки не дает 100%-ную информацию о генеральной совокупности, но выборочные параметры могут служить приближенными оценками генеральных параметров (средней арифметической, варьирования и др.).

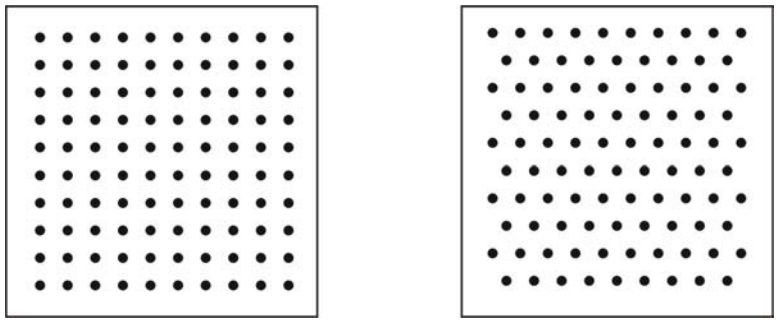
Таким образом, приемы теории вероятности и математической статистики позволяют по результатам анализа выборки характеризовать всю генеральную совокупность с известной степенью достоверности. При этом определяются не параметры генеральной совокупности, а только пределы, в которых они заключаются.

2.2 Правила составления выборок

Основным требованием к составлению выборок при проведении географических исследований, является их репрезентативность. *Репрезентативная выборка* должна по возможности наиболее полно и точно характеризовать генеральную совокупность. Это достигается определенными правилами составления.

Систематическая выборка. Рассмотрим для начала случай, когда данные генеральной совокупности не являются «пространственно привязанными», т.е. мы имеем дело с простым списком, таблицей. В указанном случае при составлении систематической выборки выбирается определенный интервал, «шаг отбора». Согласно выбранному интервалу в выборку включаются лишь варианты, отстоящие друг от друга на заданный интервал. Таким образом, в выборку будет включен каждый n-й элемент генеральной совокупности.

При работе с данными, имеющими пространственную привязку (например, при съемке данных с карты) мы должны построить регулярную сеть с определенным шагом, которая равномерно покрывала бы использованную территорию. Для данных целей можно использовать либо регулярную сеть квадратов (рис.1, а), либо регулярную сеть равносторонних треугольников (рис.1, б).



а б

Рис.1. Методы построения систематической выборки по карте.

Процедура систематической выборки очень проста. Однако при применении этого способа существует реальная опасность преувеличить долю признаков, которые повторяются через регулярные промежутки, что приведет к искажению результата. Так, известно, что большая часть свалок ТБО в Харьковской области невелики по площади и объему, при проведении систематической выборки существует опасность того, что в нее попадут лишь небольшие свалки и полигоны. Соответственно, подобная выборка не была бы полностью репрезентативной. Именно по этой причине систематическая выборка, рассмотренная в этом разделе, оценивается как весьма ненадежная.

Простая случайная выборка. В целях преодоления проблем, связанных с применением систематической выборки, выделение вариант из генеральной совокупности можно проводить случайным образом. Основным преимуществом этого метода является то, что при случайном отборе все объекты имеют одинаковую возможность попасть в выборку.

Наиболее часто при случайной выборке элементов применяется способ отбора, основанный на таблице случайных чисел (приложение 1), которые можно получить при помощи таблицы случайных чисел, генерирование которой возможно при помощи ЭВМ. Процедуру такой выборки легче понять, если представить, что машина играет десятигранной фишкой, бросая ее 4 раза в каждой серии. Каждый раз выпадает одна цифра, от 0 до 9, которая и используется для записи очередного четырехзначного числа. Так как каждая цифра выпадает случайным образом, из них можно составлять таблицы по горизонтали или по вертикали, причем случайный» характер числовой последовательности сохраняется [20].

Случайную выборку можно проводить как для элементов, имеющий пространственную привязку, так и при отсутствии таковой. Так, в случае конечного числа дискретных данных в генеральной совокупности, представленных в форме таблицы, либо карты, то для использования таблицы случайных чисел мы должны выполнить следующие операции:

1. определить общее число элементов в выборке;
2. определить количество разрядов в полученном числе элементов генеральной совокупности;
3. используя таблицу случайных чисел для полученного количества разрядов и рассчитав по формуле (1) необходимое количество элементов выборки, проводим, непосредственно, выборку из генеральной совокупности.

Например, если бы для отбора 9 элементов из совокупности, содержащей 900 членов, мы, имея три разряда в числе 900, использовали таблицу случайных трехзначных чисел. Выбор мог бы пасть на элементы с номерами 236, 008, 397, 626, 814, 220, 740, 049, 616 в приведенном порядке, в то время как при систематической выборке отобранными оказались бы номера 100, 200 300, 400, 500, 600, 700, 800, 900.

Рассмотрим случай, когда генеральная совокупность не имеет конечного числа элементов. Примером такой совокупности может быть любая физическая (рельеф) либо статическая (потенциал поля расселения) поверхность. В подобном случае в качестве «отправной точки для использования таблицы случайных чисел может быть построенная на карте координатная решетка. Так, мы можем определить наибольшее значение координаты для каждой из координатных осей и затем, при помощи таблицы случайных чисел получаем координаты отдельно по X и Y для каждого элемента выборки. Пример подобной выборки показан на рис.2.

Преимущество случайной выборки состоит в возможности быть сравнительно уверенным – даже без предварительного знания структуры совокупности – в том, что выборка отобразит большинство вариаций совокупности. Но в определенных условиях этот метод оказывается несостоятельным. Например, не зная числа элементов генеральной совокупности, мы не можем прибегнуть к случайной выборке. По этой причине часто делается предварительная выборка для получения общего представления о величине и свойствах исходной совокупности. Кроме того, всегда сохраняется возможность, хотя и призрачная, что законы вероятности будут работать против нас, а не на нас и приведут нас к приближенным или неверным результатам, подобно тому, как систематическая выборка может включить в выборку элементы, не типичные для исходной совокупности.

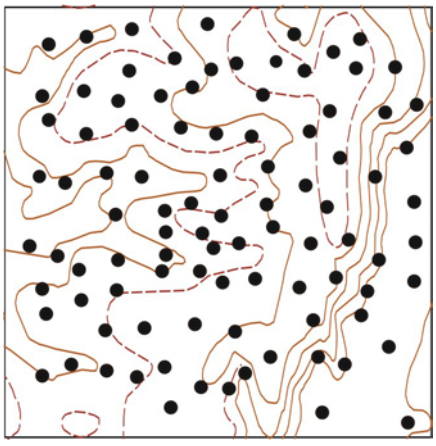


Рис.2. Пример случайной выборки. Для оценки ветрового потенциала данной территории необходимо провести измерения скоростей ветра. Для того чтобы для точек с разной абсолютной высотой вероятность попадания в выборку была одинаковой, выборка точек проводится случайным образом.

Стратифицированная выборка является единственным способом учета всех вариаций генеральной совокупности при отборе из нее элементов. Суть метода состоит в том, что

генеральная совокупность делится на части и затем выборка производится независимо из каждой части. Например, нам необходимо определить ветроэнергетический потенциал территории, часть которой лежит на правом берегу и имеет высокую степень овражно-балочного расчленения, а другая – на левом и представляет собой слаборасчлененную равнину. Известно, что по одним из основных показателей эффективности работы ветроэнергетической станции (ВЭС) является количество часов работы установки, т.е. количество часов, когда ВЭС может генерировать энергию. Одним из основных факторов, влияющих на данную характеристику, является стабильность ветрового потока над определенной территорией. В качестве отправной точки для оценки стабильности ветрового потока над данной территорией воспользуемся законом факторной относительности Маккавеева-Черванева [9] согласно которому формы рельефа, независимо от своей размерности не одновременно и по-разному реагируют на одни и те же внешние воздействия. Значит, на сложно устроенной поверхности при одинаковом внешнем влиянии будет происходить усложнение структуры и процессов взаимодействия, что не отвечает характеру внешнего воздействия. Таким образом, на территории, которая характеризуется высоким значением овражно-балочной расчлененности, будет наблюдаться менее стойкий по скорости и направлению ветровой поток, нежели над однородной территорией [3].

Предположим, что в рамках поставленной задачи нам необходимо провести измерения скорости ветра на данной территории. Для этого разделим территорию на два района: левый и правый берега и определим точки замеров согласно принципу формирования стратифицированной выборки.

Логичным предположением является то, что группы, на которые будет расчленена генеральная совокупность, не будут составлять равные ее доли. В этом случае для получения репрезентативной выборки необходимо, чтобы соотношение количества элементов, выбранных из той или иной группы, распределялось в соответствии с долями этих групп в генеральной совокупности. Например, территория, для которой мы рассчитываем ветроэнергетический потенциал, на 30% представляет собой сильно расчлененную поверхность (правый берег реки), а на 70% - равнину (левый берег). Предположим, мы собираемся сделать выборку по 20 точкам. Тогда по пропорции мы получаем, что на правом берегу мы должны взять 6, а на левом – 14 точек. Результат проведения стратифицированной выборки представлен на рис.3.

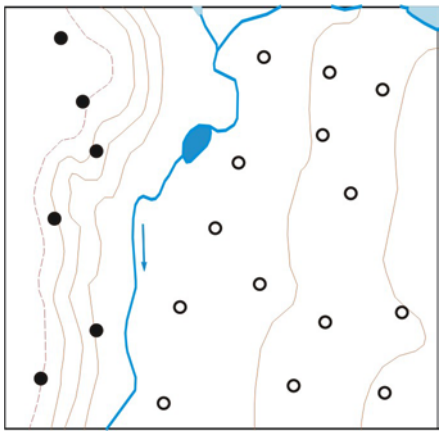


Рис.3. Стратифицированная выборка. Черными кружками показаны токи замеров скорости ветра на правом берегу; белыми – на левом.

Согласно с [20] метод стратифицированной выборки особенно хорош для географических исследований.

Вопросы:

1. Что называется генеральной совокупностью?

2. С какой целью определяется выборочная совокупность?

3. Какие виды выборок вы знаете? В чем состоит их принцип?

4. Дайте короткую сравнительную характеристику рассмотренных видов выборок.

Задание. *По данным, отобранным Вами в предыдущем разделе, проведите различные виды выборок. Определите объем выборки, необходимый для характеристики данной генеральной совокупности. Сравните и словесно опишите выборки, полученные разными способами.*

2.3. Обработка выборочной совокупности.

Случайные величины, представленные рядом количественных показателей, образуют *статистическую (выборочную) совокупность.* Каждый член этой совокупности называют *вариантой,* или *датой.* Число вариант в совокупности называют *объемом совокупности.*

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.* В вариационном ряду всегда есть максимальная и минимальная варианты. Разность между ними составит *размах варьирования,* или *амплитуду изменчивости* [23].

В вариационном ряду могут встречаться варианты, которые резко отличаются от остальных значений в выборке. При проведении эксперимента, наличие подобных значений может быть следствием неправильной работы прибора, ошибкой в расчетах и т.д. Подобные значения, которые вызывают сомнение у исследователя, определяются как *артефакт* и подлежит исключению из выборочной совокупности, т.е. подлежит выбраковке.

Если данные совокупности представлены в виде вариационного ряда, то сомнение вызывают крайние значения. К примеру, в вариационном ряду 12, 16, 22, 26, 29, 150 вызывает сомнение максимальное значение, а в ряду 24, 56, 78, 80, 84, 87 – минимальное. Варианты, вызывающие сомнение можно принять за артефакт и исключить (выбраковать) из обработки, однако 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 (приложение 2). Если $\tau_ф >= \tau_т$, где $\tau_т$ – табличное критическое значение случайной величины при объеме выборки N и уровне значимости а, то соответствующие значения вариант $x_1(x_n)$ допустимо отбросить как артефакт. Значения τ для вызывающей сомнение величины вычисляются по следующим формулам. Для наименьшего значения переменной величины в вариационном ряду (x_1) формула имеет вид:

$$\tau_1 = (x_2 - x_1)/(x_{n-1} - x_1), \qquad (2.1)$$

для максимального значения переменной в вариационном ряду (x_n):

$$\tau_n = (x_n - x_{n-1})/(x_n - x_2). \qquad (2.2)$$

Пример. Проанализируем два вариационных ряда (табл.1)

Табл.1.

X	12	16	22	26	29	150
Y	24	56	78	80	84	87

В выборке X сомнение вызывает наибольшая, а в выборке Y – наименьшая варианты. Проведем анализ на наличие артефактов. Для выборочной совокупности X τ_n равняется (150-29)/(29-12)=7.117, что значительно больше табличного значения $\tau_ф$ (0.669 при вероятности 95% и 0.805 для вероятности 99%). Для выборки Y расчет имеет следующий вид: (56-24)/(87-56)=1.032, что также больше теоретического значения [23].

Применение метода выбраковки возможно и к данным статистической отчетности в тех случаях, когда необходимо охарактеризовать общую ситуацию посредством уравнений регрессии и т.д.

После проведения анализа вариационного ряда на репрезентативность приступают к статистической обработке полученных результатов. Решение одной и той же задачи будет зависеть от объема выборки. *Малые выборки* образуют *невзвешенный вариационный ряд.* При их обработке производят обычные арифметические действия (сложение, вычитание, умножение и деление). *Большие выборки* составляют *взвешенный вариационный ряд.*

В работе со взвешенными выборками возникает необходимость объединения близких по значению вариант в классы (разряды, ступени). Такая группировка вариант облегчает последующие расчеты, однако вносит неточность в получаемые результаты, так как при обработке данных варианты заменяются средними значениями классов. Неточность в таких случаях невелика, и ею можно пренебречь. Вследствие того, что на сегодняшний день применение компьютеров приобрело широкое распространение, рекомендуется работать по алгоритму невзвешенного вариационного ряда, т. е. не производить разбивку большой выборки на классы. Однако взвешенные вариационные ряды составляют с целью построения шкалы балльной оценки, установления типа распределения обрабатываемых данных, если он неизвестен (нормальное, логнормальное и др.).

При составлении взвешенного вариационного ряда принимается следующий порядок действий. Сначала определяется величина классового интервала i, которая зависит от принятого числа классов k и объема выборки N:

$$i = (x_{\max} - x_{\min}) / k \qquad (3)$$

Число классов в зависимости от объема выборки определяется формулой:

$$k = 1 + 3{,}3 \ \lg N \qquad (4)$$

Исходя из формулы (4), можно рекомендовать следующее число классов в зависимости от объема выборки (табл.2):

Табл.2.

N	30–50	50–100	100–400	400–1000	1000–2000
k	4–6	6–8	8–9	9-11	11–12

Величина классового интервала должна быть одинаковой на протяжении всего вариационного ряда, Границы классов выбираются такими, чтобы каждая варианта могла быть отнесена только к одному классу. Например, правильная граница классов: 5–9, 10–14 или 5,5–9,4, 9,5–14,4; неправильная граница классов: 5–10, 10–15 или 5,9–9,5, 9,5–14,5. Первый и последний классы могут быть неполными. Границы классов, желательно выбирать так, чтобы крайние варианты, x_{min} и x_{max} по возможности оказались ближе к середине интервала своего класса [23].

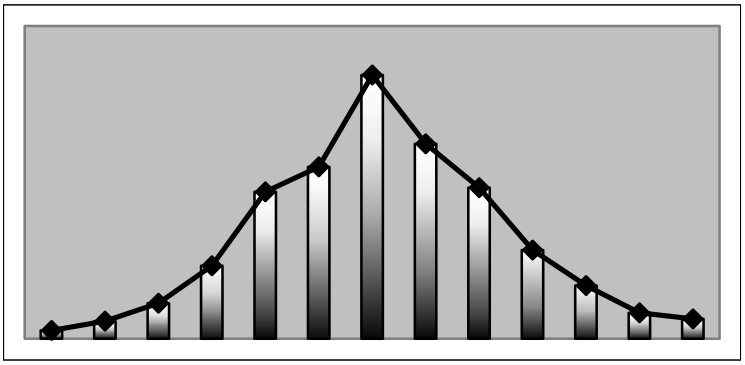


Рис.4. Полигон (график с точками) и гистограмма (столбцы) как средство графического отображения вариационного ряда.

Вариационный ряд может быть представлен графически в виде полигона (кривая распределения частот) или гистограммы (рис.4). При построении вариационной кривой по оси абсцисс откладываются значения вариант или середин классов, 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 сравниваемых выборок должна быть одинаковой. Если объем выборок не одинаков, все частоты должны быть выражены в процентах от объема выборки по каждой совокупности отдельно.

Вопросы.

1. *Что такое статистическая выборочная совокупность?. Как называются члены этой совокупности?*
2. *Дайте определение понятию «вариационный ряд». В чем состоит отличие взвешенного и невзвешенного вариационных рядов? Опишите методику составления взвешенного вариационного ряда.*
3. *Для чего необходима процедура выбраковки артефактов? Что такое артефакт?*

Задание. *На основе данных отобранных Вами для проведения исследования, проведите следующие операции:*

а) постройте вариационный ряд и определите амплитуду варьирования данных в вариационном ряду.

б) проведите анализ на наличие артефактов в Вашем вариационном ряду.

в) Постройте полигон и гистограмму распределения для Вашего вариационного ряда.

2.4. Основные выборочные параметры.

Характеристику генеральной совокупности дают по параметрам, полученным на основании выборки. Основные выборочные параметры подразделяют на три группы. Первую группу образуют *показатели среднего положения,* или *центральной тенденции.* К ним относятся мода, медиана, различные виды средних. Они выражаются именованными величинами, т. е. сохраняют размерность признака. Вторую группу образуют *показатели разнообразия признака* (разброса, изменчивости): среднее квадратическое отклонение, квадрат отклонений, коэффициент вариации. Третью группу образуют *показатели формы, распределения:* показатели асимметрии и эксцесса. Рассмотрим все три группы показателей.

2.4.1 Показатели среднего положения.

Для того чтобы получить достаточно обоснованное представление о генеральной совокупности на основании выборки, необходимо использовать наиболее характерные параметры признака. К ним относятся показатели среднего положения, которые можно разделить на следующие группы:

- непараметрические, т. е. значение которых не зависит от частотного распределения (мода, медиана);
- параметрические, значение которых тесно связано с формой распределения частот более точные (средние величины: арифметическое, гармоническое, квадратическое, кубическое, геометрическое).

2.4.1.1.Непараметрические показатели среднего.

Мода (Mo) 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,* и определяется по формуле:

$$Mo = x_m + i\left(\frac{f_2 - f_1}{2f_2 - f_1 - f_3}\right), \qquad (5)$$

где x_m.–меньший предел модального класса: i–классовый интервал; f_1 – частота класса, предшествующего модальному; f_2–частота модального класса; f_3–частота класса, следующего за модальным.

При *полимодальном (многовершинном)* распределении вариационный ряд имеет несколько значений моды (рис. 5).

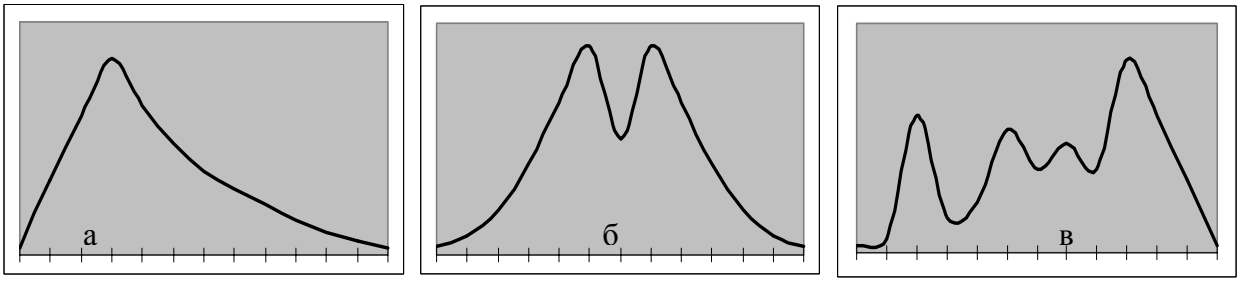


Рис. 5. Одномодальное (а), бимодальное (б) и полимодальное (в) распределения.

Медиана (Me) представляет собой среднюю варианту в ранжированном вариационном ряду, которая делит его на две равные по числу вариант 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 группировке вариационного ряда в классы медиану определяют по следующей формуле:

$$Me = x_{Me} + i\left(\frac{0{,}5N - \sum f}{f_{Mo}}\right), \qquad (6)$$

где x_{Me} – начало класса, в котором находится медиана; N–объем выборки; Σf–сумма частот всех классов, предшествующих модальному классу; f_{Mo} – частота модального класса.

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, а также в случаях, когда распределение не является нормальным.

2.4.1.2. Параметрические характеристики среднего.

Как было указано выше, к параметрическим характеристикам среднего относится довольно большое количество величин. Однако наиболее часто используемым является среднее арифметическое.

Среднее арифметическое (M, x) представляет собой величину, сумма положительных и отрицательных отклонений от которой равна нулю. Оно является основной характеристикой статистической совокупности. Для невзвешенного вариационного ряда среднее арифметическое вычисляется по формуле:

$$M = \frac{\sum x}{N}, \qquad (7)$$

где Σx - сумма всех вариант совокупности, N – ее объем.

Среднее арифметическое выборки характеризует среднее арифметическое генеральной совокупности, абсолютная и точная величина которого нам неизвестна. Для точности определения выборочных параметров необходимо установить величину ошибок репрезентативности. Ошибку среднего арифметического выборки обозначают индексом m_M. Если m_M=0, величина выборочной совокупности равна величине генеральной совокупности. Ошибка среднего арифметического выборки рассчитывается по формуле:

$$m_M = \sigma / \sqrt{N}, \qquad (8.1)$$

где σ–среднее квадратическое отклонение.

Если объем генеральной совокупности известен, то ошибка вычисляется по формуле:

$$m_M = \sigma \sqrt{\frac{N_Г - N}{N(N_Г - 1)}}, \qquad (8.2)$$

где $N_г$ – объем генеральной совокупности; N – объем выборки из нее.

Ошибку среднего арифметического можно вычислить, используя сумму квадратов от среднего $\Sigma(x_i - M)^2$:

$$m_M = \sqrt{\frac{\sum (x_i - M)^2}{N(N-1)}}. \qquad (8.3)$$

Если объем генеральной совокупности известен, то ошибка среднего арифметического вычисляется по формуле:

$$m_M = \sigma \sqrt{\frac{N_Г - N}{N(N_Г - 1)}}, \qquad (8.4)$$

где $N_Г$–объем генеральной совокупности; N–объем выборки из нее.

Пригодность среднего арифметического выборки для характеристики среднего арифметического генеральной совокупности определяется путем установления достоверности. *Достоверность* – это априорное убеждение в осуществимости некоторого явления, исключающее всякое сомнение. Достоверность характеризует реализуемость некоторого события, подтверждая его осуществимость высокими значениями уровней вероятности (P=0,95; 0,99). Достоверность среднего арифметического оценивают по критерию Стьюдента:

$$t_ф = M / m_M \qquad (9)$$

Расчетный критерий Стьюдента $t_ф$ сопоставляют с его табличным значением t_T (приложение 3). В случае, если $t_ф > t_T$, то значение показателя достоверно.

Для $t_ф$ необходимо знать значение таких показателей, как число степеней свободы v и уровень вероятности, или доверительную вероятность P.

Числом степеней свободы v считается число независимых отклонений отдельных вариант от среднего. Из всех отклонений независимыми считаются все, кроме последнего, величина которого уже определена остальными отклонениями, и поэтому оно не будет независимым. Число степеней для критерия Стьюдента определяется по формуле:

$$v = N - 1 \qquad (10)$$

Основным критерием выбора границ возможных значений признака является *степень вероятности P*, которой должны соответствовать эти границы. Выбор вероятности определяется конкретными задачами исследования и степенью точности выводов. Вероятность, с которой устанавливаются возможные значения переменной величины, получила название *доверительной вероятности p.* Наиболее часто доверительную вероятность представляют три уровня вероятности: 0,95, 0,99 и 0,999. Они могут быть представлены и в процентах: 95, 99 и 999%. Уровни доверительной вероятности показывают процент объема выборочной совокупности, значениям которых можно доверять и которыми можно уверенно пользоваться при установлении определенных закономерностей. Например, из 95% выборочной совокупности лишь 5% не подтверждают искомую закономерность. Эти 5% составляют *уровень значимости a,* показывающий процент числа вариант, значения которых не подтверждают искомую закономерность. Очевидно, что доверительной вероятности P=0,95, или 95% соответствует уровень значимости a=0,05, или 5%; для P=0,99 a =0,01, для P= 0,999 a==0,001. Таким образом, при установлении доверительной вероятности уровень значимости выражает ту вероятность, которой в данном случае решено пренебречь.

Пример расчета характеристик среднего.

На 2001 г. по областям Украины насчитывалось следующее количество свалок и полигонов твердых бытовых отходов составляло (табл.3) [10].

Для определения моды данной совокупности построим для нее вариационный ряд (табл. 4). С его помощью можно определить, что в данной совокупности имеется две моды (24 и 32), однако этого не достаточно для того, чтобы утверждать, что распределение, характеризующее данное распределение, является бимодальным. Для более объективного анализа построим гистограмму распределения частот данной совокупности данных (рис.6):

1. Согласно формуле (4), число классов, на которые необходимо разбить данную совокупность, равняется 1+3.3*lg25=5.61, что при округлении равняется 6 классам.
2. Величина классового интервала, согласно формуле (3), будет равна (61-11)/6=8.5.
3. Согласно формуле определения моды при разбиении вариационного ряда на классы (5), мода равна 20+8.5((12-2)/(2*12-2-7)) = 25.2. Исходя из рис.6, класс, содержащий в себе второе значение моды (32) не является модальным.

Для определения медианы воспользуемся построенным вариационным рядом (табл.4). Исходя из того, что ряд составляют 25 элементов, значит медианой буден значение элемента №13, т.е. медиана для данной совокупности равняется 27 местам захоронения ТБО. Медиану можно определить также по сгруппированному вариационному ряду (рис.6) по формуле (7): 19.5+8.5((0.5*25-2)/12) = 26.94.

Табл.3.

Области	Количество свалок и полигонов Тбо
АРК	28
Винницкая	35
Волынская	24
Днепропетровская	33
Донецкая	62
Житомирская	24
Закарпатская	18
Запорожская	27
Ивано-Франковская	29
Киевская	30
Кировоградская	20
Луганская	45
Львовская	48
Николаевская	20
Одесская	39
Полтавская	32
Ровненская	23
Сумская	22
Тернопольская	24
Харьковская	32
Херсонская	22
Хмельницкая	32
Черкасская	21
Черновицкая	11
Черниговская	26

Табл.4

Количество свалок и полигонов Тбо
11
18
20
20
21
22
22
23
24
24
24
26
27
28
29
30
32
32
32
33
35
39
45
48
62

Среднее арифметическое для данной совокупности имеет значение 727/25 = 29.08.

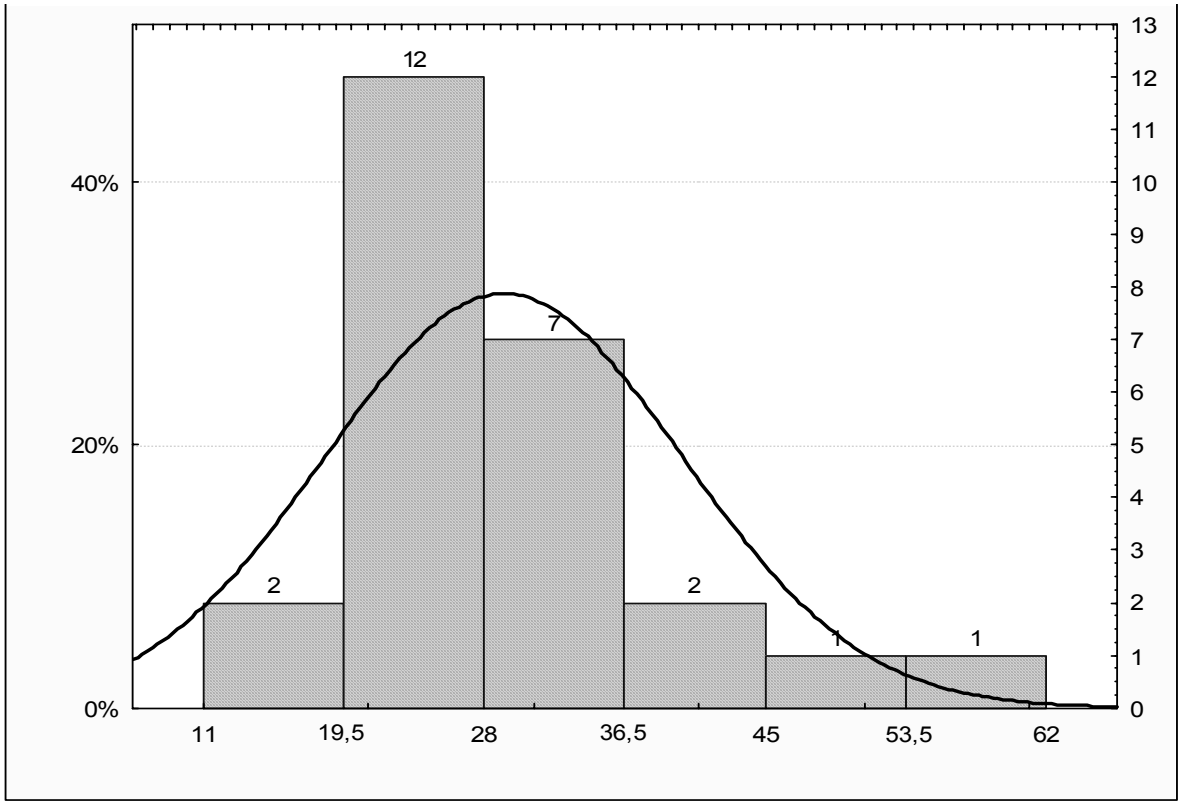


Рис.6. Гистограмма распределения частот (показатель количество свалок по областям Украины на 2001г.). Линией показана кривая распределения; цифрами над столбцами гистограммы – частоты каждого класса.

2.4.2. Показатели разнообразия признаков.

В каждой совокупности варианты отклоняются от среднего значения. Поэтому для изучаемой статистической выборки недостаточно определить лишь среднее значение, необходимы показатели, характеризующие степень разнообразия совокупности. К показателям разнообразия признаков относятся максимальная и минимальная величины в вариационном ряду (лимит), амплитуда варьирования, среднее квадратическое отклонение, квадрат отклонений от среднего, коэффициент вариации. Эти показатели признаков характеризуют различную степень и особенности разброса.

Разность между максимальной и минимальной вариантами характеризует *амплитуду варьирования.* Чем ближе минимальные и максимальные варианты к среднему и чем меньше амплитуда варьирования, тем меньше степень разнообразия между переменными в вариационном ряду, тем надежнее характеризуют статистические показатели искомую закономерность. Более точно степень разнообразия признака можно характеризовать другими показателями.

Среднее квадратическое (стандартное) отклонение, или *сигма* (σ), показывает степень рассеяния значений статистической совокупности около среднего значения. Среднее квадратическое отклонение определяется для невзвешенного ряда по формуле:

$$\sigma = \sqrt{\frac{\sum (x_i - M)^2}{N-1}}, \qquad (11)$$

где x_i – индивидуальная варианта совокупности; $x_i–M$ – отклонение от среднего индивидуальных вариант;

Ошибка среднего квадратического отклонения определяется по формуле:

$$m_\sigma = \sigma / \sqrt{2(N-1)}. \qquad (12)$$

Точность вычисления ошибок среднего квадратического отклонения и среднего арифметического можно проверить приближенно при помощи соотношения: σ/M = 0,70711. Если соотношение окажется близким к 0,7, то полученные результаты вычислений следует считать репрезентативными. В противном случае необходимо проверить расчет. При получении тех же результатов приходим к выводу, что изучаемое явление не соответствует закону нормального распределения и его оценку следует проводить с использованием непараметрических показателей [23].

Средний квадрат отклонений, или дисперсия (σ^2)–показатель, характеризующий степень рассеяния значений переменных около среднего значения. Средний квадрат отклонений можно вычислить путем возведения в квадрат показателя среднего квадратического отклонения или определить по формуле:

$$\sigma^2 = \sum (x_i - M)^2 / (N - 1) \qquad (13)$$

Средний квадрат отклонений выражается в тех же единицах, что и соответствующие показатели среднего положения. Форма записи исходных данных для расчета σ^2 такая же, как и для σ.

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

$$\sigma^2_{общ} = \sum (N-1)\sigma_i^2 / \left(\sum N_i - k\right)$$

где σ^2– дисперсия индивидуальной выборки; k – число частных выборок.

Пример. Рассчитаем показатели разнообразия признаков для описанного нами показателя количества полигонов ТБО по областям Украины за 2001 г.

Стандартное отклонение равно $\sqrt{2779{,}84/(25-1)}$ = 10,76.

В соответствии с полученным значениям найдем дисперсию как квадрат среднего квардратического отклонения: 10.76^2 = 115,83.

Основные показатели среднего, и разнообразия признака мы можем рассчитать ошибки среднего арифметического и стандартного отклонения.

Так, ошибка среднего арифметического равна $10{,}76/\sqrt{24}$ = 2,2. Как видно, несмотря на то, что мы не проводили выборку, т.е. рассматривали всю генеральную совокупность, получено ненулевое значение ошибки среднего арифметического. Это можно объяснить тем, что описанные показатели являются параметрическими, т.е. адекватную информацию несут в случае распределения, близкого к нормальному. Таким образом в данном случае мы имеем некоторое отклонение от нормального распределения.

Проверим достоверность среднего арифметического по критерию Стьюдента: $t_ф$ = 13,22, то значительно больше t_T (3,75) при уровне доверительной вероятности 0,999.

Ошибка стандартного отклонения равна $10{,}76/\sqrt{2*(24-1)}$ = 1,55. Учитывая тот факт, что значение t_T в данном случае будет тем же, что и при определении достоверности среднего арифметического, можно сразу сказать, что значение σ также можно считать достоверным.

Табл.5

	x_i	x_i-M	$(x_i$-$M)^2$
	28	-1,08	1,1664
	35	5,92	35,0464
	24	-5,08	25,8064
	33	3,92	15,3664
	62	32,92	1083,726
	24	-5,08	25,8064
	18	-11,08	122,7664
	27	-2,08	4,3264
	29	-0,08	0,0064
	30	0,92	0,8464
	20	-9,08	82,4464
	45	15,92	253,4464
	48	18,92	357,9664
	20	-9,08	82,4464
	39	9,92	98,4064
	32	2,92	8,5264
	23	-6,08	36,9664
	22	-7,08	50,1264
	24	-5,08	25,8064
	32	2,92	8,5264
	22	-7,08	50,1264
	32	2,92	8,5264
	21	-8,08	65,2864
	11	-18,08	326,8864
	26	-3,08	9,4864
Σ	727	4,26E-14	2779,84
M =	29,08		

2.4.3. 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. 7).

При получении асимметричной кривой следует проверить асимметричность распределения. Если асимметричность не будет доказана, то рассматриваемое распределение относят к симметричному.

Для проверки асимметричности распределения вычисляют коэффициент асимметрии, его ошибку, затем на основании показателя достоверности устанавливают вид кривой распределения.

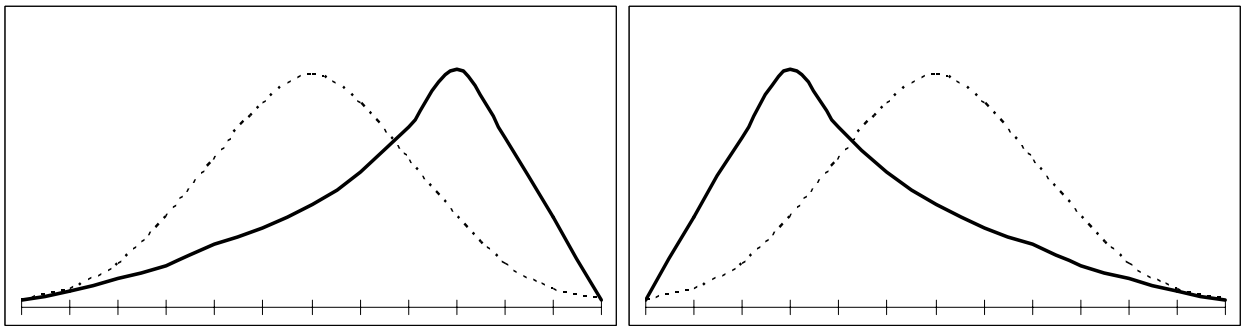


Рис.7. Положительная и отрицательная асимметрия распределения (пунктиром показана форма нормального распределения).

Коэффициент асимметрии находят по формулам:

$$K_{as} = (M - Mo)/\sigma \;\; или \;\; K_{as} = (M - Me)/\sigma \qquad (14)$$

Коэффициент асимметрии также необходимо проверять на достоверность с помощью критерия Стьюдента:

$$t = K_{as} / m_{as} \qquad (15)$$

Ошибка коэффициента асимметрии определяется по формуле:

$$m_{as} = \sqrt{6/(N+3)} \qquad (16)$$

Эксцесс кривой распределения (E) имеет место в тех случаях, когда большинство вариант совокупности сосредоточено около среднего арифметического. Тогда эмпирическая кривая распределения отклоняется от нормальной теоретической кривой у ее вершины и количественно выражается показателем эксцесса (рис. 8).

Положительный эксцесс представлен кривой островершинной (эксцессивной, или лептокуртичной) (сплошная линия на рис.8), отрицательный–плосковершинной (депрессивной, или платикуртичной) (штриховая линия на рис.8). При сильном отрицательном эксцессе кривая может приобрести вид двухвершинной.

Показатель эксцесса определяется по формуле:

$$E = \left[\sum (x - M)^4 / N\sigma^4\right] - 3 \qquad (17)$$

Ошибка коэффициента эксцесса вычисляется следующим образом:

$$m_E = 2\sqrt{6/(n+5)} \qquad (18)$$

Оценка достоверности показателя эксцесса производится аналогично оценке показателя асимметрии по критерию Стьюдента:

$$t = \frac{E}{m_E}. \qquad (19)$$

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

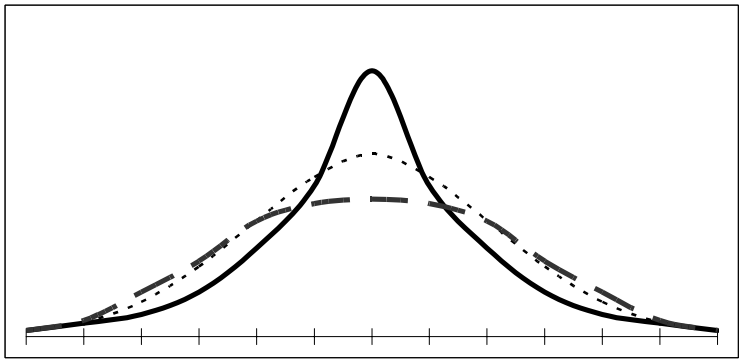


Рис.8. Положительный и отрицательный эксцесс (см. объяснения в тексте). Пунктиром показана форма нормального распределения.

Асимметрия и эксцесс эмпирических кривых указывают иногда на важные особенности объекта исследования (например, на изменение признака в ходе дифференциации природно-экологических условий в ландшафте). В таких случаях изучение степени и характера асимметрии и эксцесса вариационных кривых может быть самостоятельной задачей при проведении исследовательских работ [23].

Пример. Рассчитаем коэффициент асимметрии и эксцесса для данных, использованных в предыдущих примерах.

Коэффициент асимметрии для данной совокупности равен: (29,08 – 25,17)/10,76 = 0,3633. Полученный результат говорит о незначительной положительной асимметрии рассматриваемого распределения.

Определим ошибку коэффициента асимметрии: $\sqrt{6/25+3} = 0,4629$.

Проверим достоверность коэффициента при помощи критерия Стьюдента: t_ϕ = 0,785, а наименьшее значение t_T (при уровне доверительной вероятности 0,95) составляет 2,06. Значит, исследуемое распределение можно считать симметричным.

Определим коэффициент эксцесса. Подставляя данные в формулу (17) получаем значение, равное 4,5513 (значительный положительный эксцесс). Ошибка коэффициента составляет 0,8944. t_ϕ = 5,089, t_T= 3,75 (p=0,999*)*. Таким образом, распределение частот показателя количества свалок и полигонов ТБО по областям Украины имеет значительный положительный эксцесс.

Вопросы.

1. Какие группы показателей относятся к основным выборочным параметрам?

2. Охарактеризовать непараметрические показатели среднего.

3. Какие виды параметрических показателей среднего положения Вы знаете? Что такое среднее арифметическое. Как проводится проверка достоверности Среднего арифметического?

4. Объясните цель использования показателей разнообразия признаков.

5. Дайте определение понятию положительная и отрицательная асимметрия и положительный/отрицательный эксцесс.

Задание. *В выбранных Вами совокупностях данных провести развернутую статистическую характеристику (определение показателей среднего, разброса, асимметрии и эксцесса) Провести анализ Каждого их полученных показателей на достоверность).*

2.5. Методы установления различий между выборками

Проведение прикладных географических исследований предполагает не только изучение строения, развития, закономерностей распространения исследуемых объектов, 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.

Достоверность различий между генеральными совокупностями *(N_1, N_2...)* может быть определена с помощью следующих критериев достоверности: критерия Стьюдента *(t),* наименьшей существенной разности (НСР), критерия соответствия *(χ^2),* критерия Фишера (*F*), критерия Колмогорова *(λ).*

Каждый из методов может применяться лишь при определенных условиях, которые задаются целью исследования. Несоблюдение указанных условий при решении задач по математической статистике может привести к ошибочным выводам.

2.5.1. Критерий Стьюдента.

Сравнение выборочных совокупностей по критерию Стьюдента *t* позволяет утверждать с некоторой долей уверенности сходство или различие между средними выборок по разнице между ними с использованием формулы:

$$t = d / m_d , \qquad (20)$$

где *d* – разность между средними *(M_1–M_2); m_d* – ошибка разности средних.

При расчете разницы между средними из большей величины вычитают меньшую независимо от нумерации выборочных совокупностей. С методической точки зрения весьма важным для исследователя является установление типа выборочной совокупности. От этого будет зависеть применение соответствующих формул при расчете степени свободы и ошибки разности между средними выборочных совокупностей.

Выделяют три типа сравниваемых статистических совокупностей:

- независимые с одинаковым объемом выборок *(N_1=N_2)*;
- независимые с разным объемом выборок *(N_1=N_2)*;
- сопряженные только с одинаковым объемом выборок *(N_1=N_2)*.

Независимые статистические совокупности могут быть получены на одной или нескольких точках, но при одинаковых условиях проведения эксперимента. В качестве примера можно привести такие данные, измерение скорости ветра в зимние месяцы в г. Харькове в течение нескольких лет и установление достоверных различий между этими показателями по годам исследований. В этом случае условия наблюдения одинаковы. Поэтому при установлении степени свободы в каждом независимом эксперименте выборочные совокупности суммируются.

Сопряженные статистические совокупности, как и независимые, однозначны по смыслу, их получают при проведении исследований на одной или в нескольких точках, но в разных условиях. Например, измерение скорости ветра на высоте 18м и 35м. В этом случае с увеличением высоты уменьшается влияние подстилающей поверхности, т.е. поток становится более ламинарным (измерения производятся в одних и тех же точках, но в разных

условиях). Степень свободы в каждом рассматриваемом эксперименте при использовании сопряженных выборок определяется по числу пар сравниваемых выборок (M_i) [23].

2.5.1.1.Расчет критерия Стьюдента для независимых статистических совокупностей.

При одинаковом объеме выборок в случаях независимых статистических совокупностей производят следующие расчеты. Вычисляют средние в сравниваемых выборках M_1 и M_2. Затем находят ошибки средних для . каждой выборки в отдельности по формулам (8.1-8.4), определяют разность между средними $d=M_1 - M_2$. Ошибку разности между средними вычисляют по формуле:

$$m_d = \sqrt{m_1^2 + m_2^2}\ , \qquad (21)$$

где m_1 – ошибка среднего арифметического первой выборки; m_2 – ошибка среднего арифметического второй выборки.

Критерий Стьюдента определяют по формуле (9). Число степеней свободы устанавливают следующим образом:

$$\nu = N_1 + N_2 - 2 \qquad (22)$$

Сопоставляя $t_ф$ и t_T, устанавливают или отвергают с некоторой долей уверенности различия между средними арифметическими выборок.

Пример. Для развития отрасли использования биогаза свалок и полигонов ТБО в качестве энергетического ресурса важным показателем является объем свалок ТБО. В связи с этим проанализируем объем ТБО, накопленный на свалках по отобранным районам Харьковской области. Интегральной характеристикой района по указанному показателю может быть частотное распределение объемов ТБО захороненного на полигонах данного района. Следовательно, при проведении анализа сходства совокупностей данных по отдельным административным районам специалистам по региональному развитию альтернативной энергетики, можно провести районирование области по принципу единства подходов относительно развития биоэнергетики (т.е. несколько административных районов могут объединяться в единый таксон).

Для проведения анализа воспользуемся критерием Стьюдента.

В Лозовском и Краснокутском районах Харьковской области имеется по 4 действующих свалки ТБО, а также по одной запроектированной. Объемы ТБО на указанных свалках приведены в табл.6.

Табл.6.

районы	объемы ТБО (м.куб)				
Краснокутский	35040	33630,9	0	60692,36	34 32
Лозовской	299355,9	400000	1375 0	22000	0

Данные совокупности относятся к независимым совокупностям и имеют равный объем. Работа по определению достоверности различия между выборками состояла из следующих этапов:

1. Для каждой совокупности определены среднее арифметическое и стандартное отклонения (M_1 и M_2).
2. Вычисляем разность между средними (из большего вычитаем меньшее) ($M_2 – M_1$).
3. По формуле (8.1) рассчитаны ошибки средних арифметических (m_1 и m_2).
4. По формуле (21) определяем ошибку разности средних m_d.
5. Находим значение $t_ф$ для критерия Стьюдента ($t_ф$= 1,42).
6. Находим значение количества степеней свободы (5+5-2=8) и сопоставляем полученное значение $t_ф$ с табличными значениями критерия Стьюдента (t_T), которое даже

при уровне p=0,95 составляет 2,31. Следовательно, различие между исследуемыми совокупностями несущественно, т.е. мы имеем право объединить по описываемому признаку в единый район.

Результаты всех этапов вычислений представлены в табл.7.

Табл.7.

	M	σ	m	m_d	M_1-M_2	$t_ф$
Красноку тский	265 59,05	251 35,95	1,26 E+08	85070,013 67	120 462,1	1,41 6035
Лозовско й	147 021,2	188 554,3	7,11 E+09			

При разном объеме выборок в сравниваемых совокупностях порядок вычислений критерия Стьюдента такой же, как и при установлении достоверности в независимых выборках с одинаковым числом наблюдений. Различие состоит лишь в вычислении ошибки разности средних, которая определяется по формуле:

$$m_d = \sqrt{\frac{\sum\left(x_{i_1} - M_1\right)^2 + \sum\left(x_{i_2} - M_2\right)^2}{\left(N_1 + N_2 - 2\right)} \frac{N_1 + N_2}{N_1 \cdot N_2}}, \qquad (23)$$

где $\Sigma(x_{i1} - M_1)^2$–сумма квадратов отклонений от среднего для первой выборки; $\Sigma(x_{i2} - M_2)^2$– второй выборки; N_1, N_2 – количество вариант в первой и второй выборках соответственно.

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

$$\nu = \frac{1}{u^2 / \left(N_1 - 1\right) + \left(1 - u\right)^2 / \left(N_2 - 1\right)}, \qquad (24)$$

где $u = m_1^2 / (m_1^2 + m_2^2)$; m_1, m_2 – ошибка среднего первой и второй выборок соответственно.

Пример. Рассмотрим случай с независимыми выборками, имеющими различный объем. Так, в Харьковском районе имеется 11 свалок и полигонов ТБО, а в Балаклейском – 7.

1. Рассчитаем суммы квадратов отклонений от среднего$(x_i - M)^2$.
2. Проведем расчет констант, необходимых в расчете $(N_1+N_2-2 = 17,\ N_1+N_2/\ N_1*N_2 = 84)$.
3. Рассчитаем ошибку разности средних (m_d) по формуле (23).
4. Найдем значение $t_ф$ для критерия Стьюдента ($t_ф$= 0,23).
5. Рассчитаем количество степеней свободы по формуле (24). Для этого вычислим ошибки средних арифметических для каждой совокупности (υ = 12).
6. Сопоставим полученные значения критерия Стьюдента с табличными. (t_T), даже при уровне p=0,95 составляет 2,18. Значит, как и в предыдущем примере, различия между выборками можно признать, как несущественные.

Расчеты приведены в табл.8.

Табл.8.

Район	Объем, м.куб	$(x_i - M)$	$(x_i - M)^2$		u	m_d	$t_ф$
Харьковски й	19000	- 285362	81431314571	m_1	0,3276 43	68254 5	0, 219608
	358405,9	54044, 17	2920772761	66962,721 77		M_1-M_2	
	630035,9	325674 ,2	1,06064E+11	σ_1	1-u	14989 2,1	
	653655,9	349294	1,22006E+11	222090,22	0,6723		

		,2		3	57		
	200000	-104362	10891369819				
	18000	12800	163840000		υ		
	200000	-104362	10891369819		11,61724		
	204985	-99376,7	9875733637				
	417455,9	113094,2	12790292230				
	56840,31	-247521	61266851296				
	417455,9	113094,2	12790292230				
	476505,9	172144,2	29633616700				
M_1	304361,7258	Σ =	4,60726E+11				
Балаклейский	720000	-565530	3,19825E+11	m_2			
	7500	146969,6	21600077181	95925,19757			
	82181,04	72288,61	5225642723				
	5200	149269,6	22281427558	σ_2			
	61120	93349,65	8714156622	253794,2172			
	61102,95	93366,7	8717340135				
	144183,54	10286,11	105804000,2				
M_2	154469,6471	Σ =	3,86469E+11				

2.5.1.2. Сопряженные выборочные совокупности.

При установлении различий между сопряженными выборками алгоритм тот же, что и для независимых наблюдений. Вычисление ошибки разности средних в этом случае производится по формулам:

$$m_d = \sqrt{\frac{\sum \left(d_i - \bar{d}\right)^2}{N_\Pi \left(N_\Pi - 1\right)}}; \quad m_d = \sqrt{\frac{\sum d_i^2 - \left(\sum d_i\right)^2 / N_\Pi}{N_\Pi \left(N_\Pi - 1\right)}}, \qquad (25)$$

где d_i – разность между индивидуальными сопряженными вариантами в выборках; $\bar{d}$ – разность между средними сопряженных выборок; N_n – число сопряженных пар в сопряженных выборках.

Число степеней свободы находят по равенству $\nu = N_П - 1$.

Если при проведении эксперимента пренебречь сопряженностью выборок и обработку статистических показателей проводить по независимым наблюдениям, то получим противоположный вывод, т. е. различие будет признано несущественным. Поэтому необходимо подбирать такой способ обработки выборочных совокупностей, который соответствовал бы условиям проведения опыта.

2.5.2. Наименьшая существенная разность.

Достоверность различий между двумя выборками может быть проверена по наименьшей существенной разности (НСР). Наименьшая существенная разность показывает то минимальное различие между средними, начиная с которого при выбранном уровне вероятности средние сравниваемые показатели существенно отличаются друг от друга. Величина критерия НСР выражается в тех же единицах, что и сравниваемые средние выборочных совокупностей, и определяется по формуле:

$$HCP = t_T m_d, \qquad (26)$$

где *ma*–ошибка разности средних; *ti*–табличное значение критерия Стьюдента при выбранном значении уровня вероятности.

Если разность между сравниваемыми средними в условиях эксперимента больше или равна величине НСР при Р=0,95 или 0,99, то различие существенно. Если разность между средними меньше НСР, то различие обусловлено случайными факторами и признается недостоверным.

Пример. Рассчитаем показатель НСР для данных использованных в предыдущем примере (табл.8.). Табличное значение критерия Стьюдента для 12 степеней свободы и *p*=0,95, соответствует 2,18. Ошибка разности средних m_d равняется 682545. Согласно формуле (26) НСР равно 2,18*682545 = 1487948,034. Согласно табл.8, Разность средних составляет 149892,1, что значительно меньше критерия НСР. Следовательно, различие между выборками является несущественным.

2.5.3. Критерий Фишера.

Сравниваемые совокупности могут отличаться не только по величине средних, но и по другим параметрам распределения случайных величин, в частности по дисперсиям. В таких случаях при установлении достоверности различия между совокупностями лучше использовать критерий Фишера F (положительное асимметричное распределение). Расчет критерия Фишера производится по формуле:

$$F = \sigma_1^2 / \sigma_2^2, \qquad (27)$$

где σ_1^2 по абсолютной величине должна быть больше, чем σ_2^2. Если величина расчетного критерия Фишера F_Φ не превышает величины приведенного в таблице F_T (приложение 6), то различие между сравниваемыми дисперсиями считается недостоверным. При $F_\phi > F_T$ эти дисперсии достоверно различны, а различие сравниваемых генеральных совокупностей признается неодинаковым. Степень свободы рассчитывается для сравниваемых совокупностей отдельно по формуле $v=N-1$.

Пример. Для расчетов воспользуемся данными предыдущего примера. Так, дисперсия по совокупности данных по Харьковскому району σ_1 составляет 222090,223; по Балаклейскому району (σ_2) 253794,2172. Рассчитаем фактическое значение критерия Фишера по формуле (27). Так как совокупность по Балаклейскому району больше, поставим ее в числитель. Получили $F=(253794{,}2172)^2/(253794{,}2172)^2= 1{,}142752768$. Табличное значение критерия Фишера для степеней свободы 6 и 11 (p= 0,95) составляет 3,09. Следовательно, в данном случае достоверного различия между выборками не наблюдается.

2.5.4. Критерий хи-квадрат.

Количественное изучение явлений требует создания гипотез, с помощью которых можно объяснить эти явления. Чтобы проверить гипотезу, нужно получить ряд опытных данных и сопоставить их с теоретически ожидаемыми согласно гипотезе. Совпадение может служить основанием для принятия гипотезы и подтверждения ее правильности. Степень несоответствия фактических наблюдений теоретически ожидаемым результатам может быть различной. Отсюда возникает задача статистической оценки разницы между расчетными и

теоретически ожидаемыми данными. Для этой цели используется критерий хи-квадрат (χ^2), или критерий соответствия, который рассчитывается по формуле:

$$\chi^2 = \frac{(\varphi - \varphi')^2}{\varphi'}, \qquad (28)$$

где φ, φ' – число наблюдений в опыте фактическое и теоретически ожидаемое.

Значения хи-квадрат могут быть только положительными и возрастать от нуля до бесконечности. Если расчетные значения хи-квадрат превышают табличные (приложение 4), то гипотеза о независимости признаков отвергается. Если $\chi^2_\phi < \chi^2_T$,то признаки можно считать независимыми. Степень свободы при проверке гипотезы о нормальном распределении вычисляется по формуле:

$$\nu = k - 3, \qquad (29)$$

где k – число классов.

Достоверность расчетных данных можно также оценить по формуле:

$$D = \left(\chi^2 - \nu\right) / \sqrt{2\nu} \geq 3 \qquad (30)$$

Различие считается достоверным, если D>3. При обработке данных по условиям применения критерия хи-квадрат требуется, чтобы частота в каждом классе была не менее пяти [23].

2.5.5. Критерий Колмогорова.

Как и уже рассмотренный нами критерий хи-квадрат, критерий А.Н. Колмогорова λ также относится к приемам непараметрической статистики и служит для оценки сходства или различия между выборками. Он определяется по формуле:

$$\lambda = \frac{\left|\Sigma f_н - \Sigma f'_н\right| \max}{\sqrt{n}}, \qquad (31)$$

где $\Sigma f_н$ и $\Sigma f_н'$ – суммы накопленных частот эмпирического и теоретического распределений; $\left|\Sigma f_н - \Sigma f'_н\right| \max$ - максимальная разность накопленных частот для одного и того же интервала (без учета знака).

Критерий λ не требует знания числа степеней свободы (что облегчает процесс его вычисления), он имеет следующие стандартные значения для разных степеней вероятности (табл.9):

Табл.9.

Вероятность P	0,90	0,95	0,99	0,999
Значение λ	1,244	1,358	1,627	1,950

Если расчетное значение λ меньше первой или второй табличной величины, то различие между сравниваемыми распределениями можно считать несущественным [1].

Вопросы.

1. *На какие три типа необходимо делить исследуемые выборочные совокупности для выбора соответствующей методики расчета критерия Стьюдента?*
2. *Объясните связь критерия НСР с критерием Стьюдента.*
3. *В чем заключается принципиальное отличие критерия Фишера от предыдущих рассмотренных критериев?*
4. *Для чего служит критерий хи-квардат?*

5. *В чем заключается большая простота использования критерия Колмогорова по сравнению с критерием хи-квардат?*

Задание. *Отберите для исследования на достоверность различия две или более выборок, проведение анализа на достоверность различия между которыми имело бы смысл (проконсультируйтесь с преподавателей по поводу адекватности применения тех или иных данных). Рассчитайте для Ваших выборок критерий Стьюдента, НСР, критерий Фишера, хи-квадрат и критерий Колмогорова. Одинаковый ли результат дали все эти критерии?.*

2.6.Нормальное распределение. Анализ на нормальность.

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 прикладных географических исследованиях при описании некоторых статистических совокупностей их можно представить в виде *нормального распределения.*

Нормальное распределение по К. Пирсону, или распределение Гаусса–Лапласа, имеет место среди природных процессов и явлений. В природной 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, а другие – к уменьшению. В то же время встречаемость вариант, 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.9). Это означает, что большие отклонения от средней встречаются реже, чем малые [23].

При нормальном распределении среднее, мода и медиана совпадают. Кривая плотности не пересекает оси абсцисс, что подтверждает вероятность существования неограниченно больших отклонений. Уравнение нормального распределения можно записать в нескольких модификациях. Наиболее часто используемая форма следующая:

$$f' = \frac{N_i}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x_i - M)^2}{2\sigma^2}} \qquad (32)$$

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

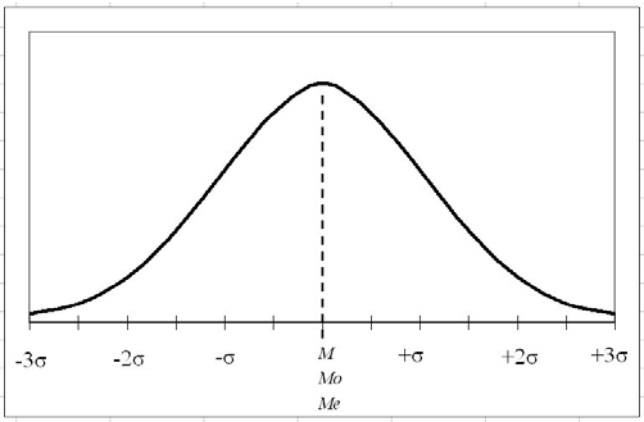


Рис.9. Кривая плотности нормального распределения.

Одной из наиболее существенных причин необходимости проверки распределения на нормальность для той или иной совокупности данных является, прежде всего, ограниченность сферы применения многих статистических приемов, относящихся, прежде всего, к приемам параметрической статистики, так как последние правомерно использовать лишь для данных, которые описываются нормальным распределением.

Проверки распределения на нормальность можно осуществить разными способами. Прежде всего, это критерий хи-квадрат, изложенный в предыдущем пункте. Действительно, зная формулу, описывающую нормальное распределение, а также имея конкретные данные по исследуемой совокупности, мы можем оценить достоверность различия между фактическими и теоретическими данными.

Однако существуют и другие, более простые способы проверки распределения на нормальность Так, известно, что на интервал ±σ от среднего арифметического должно приходиться 68,27% от общего количества вариант в выборке; в интервал ±2σ жолжно попадать 95,45%; в интервал ±3σ – 99,73; в интервал ±4σ – 99,99%. В соответствии с этим для проверки на нормальность некоторой выборки необходимо вычислить среднее арифметическое и стандартное отклонение и затем найти долю вариант, попадающих в те или иные интервалы. При работе с малыми выборками целесообразно использовать лишь первые три интервала, однако необходимо проверять по каждому из них, так как распределение может иметь положительный или отрицательный эксцесс.

Вопросы:

1. *Что такое нормальное распределение? Приведите его основные характеристики.*
2. *Каким образом можно проверить, является ли ваше распределение нормальным или нет?*

Задание. *На основе построенного Вами взвешенного вариационного ряда, постройте гистограмму распределения частот. Поведите сигма-анализ на нормальность.*

ТЕМА 3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ.

При проведении прикладных географических исследований зачастую необходимо определить степень взаимозависимости между теми или иными показателями. Так, к примеру, исследуя размещение твердых бытовых отходов на свалках Харьковской области можно сразу сказать, что существует определенная зависимость между площадью той или иной свалки с ее объемом. Однако подобный вывод не может быть использован в дальнейших этапах исследования, так как он не имеет доказательства. Именно в этом случае используется корреляционный анализ.

Корреляционным анализом называется математический метод, позволяющий установить направление и тесноту взаимосвязи между определенными явлениями.

Корреляционный анализ широко применяется в прикладных географических исследованиях, однако необходимо иметь в виду, что доказательство математической связи должно опираться на реальную зависимость между явлениями, так как иногда можно установить несуществующие корреляции. Примером, может быть достоверная положительная корреляция между площадью сельскохозяйственных угодий и количеством предприятий черной металлургии.

По форме корреляционная связь бывает *линейной* и *нелинейной (криволинейной),* по направлению – *прямой* и *обратной,* по величине – от 0 до ±1, по количеству коррелируемых признаков *парной* и *множественной.* По типу представления данных также выделяют *корреляцию между признаками, выраженными количественными или качественными данными*. Для расчета линейной корреляции используются коэффициенты корреляции, для нелинейных зависимостей – корреляционные отношения, рассмотрение которых по ряду причин не вошло в состав данного пособия.

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);
- субпричинную, когда один фактор выступает как отдельная причина сопряженного изменения признака (связь биомассы с количеством осадков);
- взаимоупреждающую, когда причина и следствие, находясь в устойчивой взаимной связи, последовательно влияют друг на друга.

Вопросы:

1. *Что такое корреляционный анализ?*
2. *Какие формы корреляционной связи Вы знаете?*
3. *Назовите основные виды парной корреляционной связи.*

3.1. Парная корреляция.

В практической работе по установлению парной корреляции между признаками и явлениями необходимо придерживаться следующей последовательности: 1) на основании исследований определяют, существует ли связь между рассматриваемыми признаками; 2) если связь между явлениями и признаками существует, устанавливают форму, направление и тесноту связи, используя график взаимозависимости между признаками. Для построения этого графика составляются сопряженные вариационные ряды, в которых следует определить аргумент *X* и функцию *Y*.

Табл.10.

№ п\п	X	Y
1	49,1	215,4
2	38,1	128
3	24	80,5
4	32,8	176,5
5	62	260
6	24	101,7
7	18	45,1
8	29,8	149
9	29	93
10	36,2	187
11	20	64,1
12	48	157,3
13	48	186

№ п\п	X	Y
14	20	110
15	42,3	203
16	32	118,5
17	23	134,6
18	51,7	186
19	24	83,4
20	46	170
21	22	77
22	34,3	112
23	21	105,2
24	11	67
25	26	129

Затем график в виде точек наносятся сопряженные варианты (рис.10)

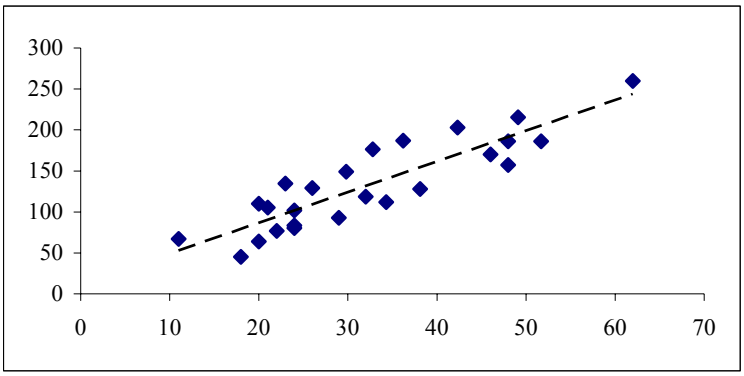


Рис.10. Пример прямолинейной корреляционной связи, построенной по данным табл.10.

По виду полученного графика можем утверждать, является зависимость прямолинейной (рис.10) либо криволинейной (рис.11):

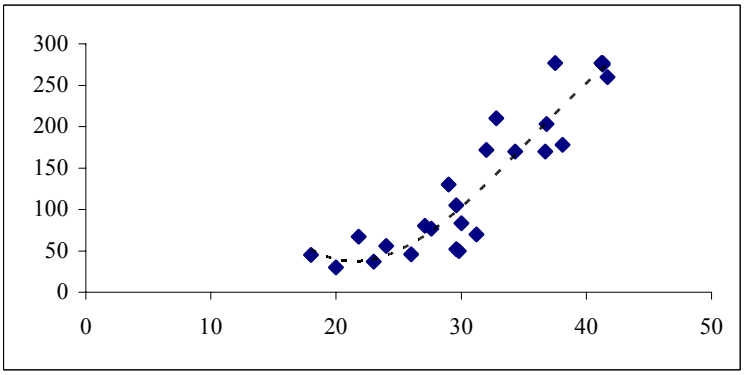


Рис.11. Пример криволинейной корреляционной связи.

Степень рассеяния частот или вариант относительно линии регрессии на графике указывает ориентировочно на тесноту связи: чем меньше рассеяние, тем сильнее связь (рис.12)

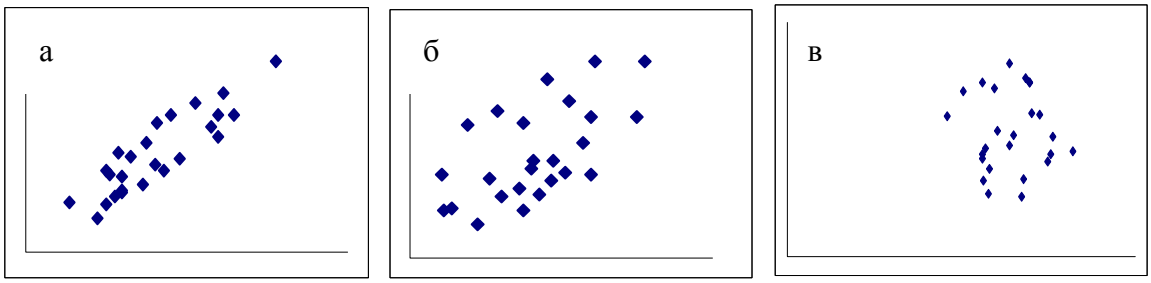


Рис.12. Точечные графики, соответствующие коэффициентам корреляции 0,88(а), 0,59(б) и –0,03(в).

Коэффициент корреляции Пирсона.

Если зависимость между признаками на графике указывает на линейную корреляцию, рассчитывают *коэффициент корреляции Пирсона r,* который используется в случае распределений, близких к нормальному. Он позволяет оценить тесноту связи переменных величин, а также выяснить, какая доля изменений признака обусловлена влиянием основного фактора, какая – влиянием других факторов. При положительной зависимости величина коэффициента корреляции изменяется от 0 до +1, при отрицательной - от 0 до -1. Если $r = 0$, то связь между признаками отсутствует. Принято считать, что при r <0,5 корреляционная зависимость слабая, при r -=0,50,7 - средняя, при r = 0,7-0,99 - сильная.

Коэффициент корреляции Пирсона рассчитывается по формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{n\sigma_x\sigma_y}, \qquad (33)$$

где M_x и M_y – средние арифметические для явлений X и Y, σ_x σ_y – средние квадратические отклонения для явлений X и Y, n – число сопряженных пар данных.

Достоверность вычисленного коэффициента корреляции можно установить через критерий Стьюдента. При использовании критерия Стьюдента для доказательства достоверности r вначале рассчитывают стандартную (квадратическую) ошибку коэффициента корреляции по формуле:

$$m_r = \sqrt{(1-r^2)/(n-2)}, \qquad (34)$$

где n – число сопряженных пар данных в сравниваемых выборочных совокупностях.

Значение коэффициента корреляции записывают с учетом его ошибки: $r \pm m_r$. Затем вычисляют критерий Стьюдента для коэффициента корреляции:

$$t_r = r / m_r \qquad (35)$$

Критерий Стьюдента можно также рассчитать иначе:

$$t_r = r\sqrt{n-2} / \sqrt{1-r^2}\ . \qquad (36)$$

Если $t_ф > t_T$, то корреляционная связь существенна, при $t_ф < t_T$-недостоверна.

Пример. Так как в данных статистической отчетности в целом и в данных по энергетике в частности крайне редко можно найти совокупности данных, отвечающие закону нормального распределения, то воспользуемся гипотетическими совокупностями данных. Предположим, что исследованиями установлено, что на значения показателя Y влияет показатель X. Необходимо доказать достоверность установленной зависимости. Исходные данные составляют следующий сопряженный вариационный ряд:

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

Построим график, указывающий на существование зависимости между исследуемыми показателями (рис.13):

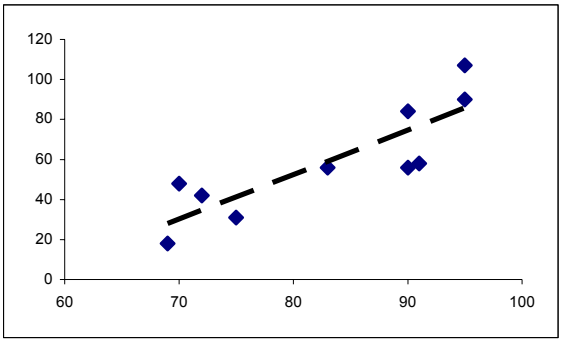


Рис.13. График зависимости между показателями *x* и *y*.

Построенный график наглядно показывает наличие положительной линейной зависимости между данными показателями.

Рассчитываем коэффициент корреляции по формуле (33). Для большего удобства составим таблицу расчета коэффициента корреляции Пирсона (табл.11):

Табл.11.

№ п\п	*x*	*y*	x-M	y-M	(x-M)(y-M)
	83	56	0	-3	0
	72	42	-11	-17	187
	69	18	-14	-41	574
	90	84	7	25	175
	90	56	7	-3	-21
	95	107	12	48	576
	95	90	12	31	372
	91	58	8	-1	-8
	75	31	-8	-28	224
	70	48	-13	-11	143
M_{xy}	83	59		Σ	2222
σ_{xy}	10,54093	27,45501			
n	10		$n*\sigma_x*\sigma_y$	2894,013	
r	0,77				

Определим достоверность полученного коэффициента корреляции при помощи критерия Стьюдента по формуле (35):

$t_r = 0,77*(\sqrt{10-2})/\sqrt{1-0,77^2} = 3,4,$

что больше значения t_t (3.36) при ν=8 и P=0,99. Значит, зависимость между рассматриваемыми показателями достоверна.

Вопросы:

1. *Имеет ли коэффициент Пирсона широкое применение в географии? В чем причина этого ограниченния?*

Задание. *Для отобранных Вами совокупностей рассчитайте коэффициент корреляции по Пирсону (даже в случае несоответствия Вашего частотного распределения нормальному – в этом случае Ваши расеты будут носить характер получения практических навыков расчета данного показателя).*

3.2. Ранговая корреляция.

В прикладных социально- и экономико-географических исследованиях большая часть совокупностей данных имеет распределения частот, сильно отличающиеся от нормального. В этом случае мы не имеем право использовать приемы так называемой *параметрической статистики***,** т.е. работать непосредственно с конкретными значениями того или иного показателя. С другой стороны, некоторые географические объекты могут обладать признаками, лишенными точной количественной оценки, но позволяющими сравнивать их по качественным оценкам. В этом случае используются приемы *непараметрической статистики***,** одним из которых является *ранжирование***,** смысл которого заключается в том, что совокупность объектов упорядочивают, устанавливая порядковый номер (ранг) каждого из них по данному признаку. 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.

Применительно к корреляционному анализу приемом ранжирования показателей оперирует *ранговая корреляция*. Ранговую корреляцию можно применять для всех упорядоченных признаков (например, экспертные оценки, баллы, бонитеты). Объем сопряженных 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
N_2	1	2	3	4	5

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
N_2	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.

Наиболее широкое применение для расчета зависимости *(x, y)* получили коэффициенты ранговой корреляции Спирмена и Кендалла.

Коэффициент ранговой корреляции Спирмена.

Коэффициент ранговой корреляции Спирмена r_s имеет несложный порядок расчета, поэтому в естественных науках ему отдается предпочтение. Ранжирование производится по описанному принципу. В случае наличия двух или более одинаковых значений берут среднее арифметическое из двух соседних рангов, которые они должны занять, и каждому значению приписывается полученное среднее арифметическое (например, если в выборке сть два значения, равные 181,2, их они должны поделить между собой 14 и 15 ранги, то каждому из них пишется ранг (14+15)/2=14,5). Подобная операция делается и при большем количестве равных значений.

Коэффициент Спирмена представляет собой следующее соотношение:

$$r_S = 1 - \frac{6\Sigma(d^2)}{N^3 - N} = 1 - \frac{6\Sigma(x' - y')^2}{N^3 - N}, \qquad (37)$$

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* количество сопряженных пар.

Достоверность рангового коэффициента корреляции Спирмена можно установить аналогично достоверности коэффициента корреляции Пирсона, либо по таблице критических значений коэффициента ранговой корреляции Спирмена (табл.12.) [12].

Табл.12.

Длина рядов	Довер. вероятность (Р)	
	0,95	0,99
5	0,900	1,000
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,746
12	0,506	0,712
14	0,456	0,645
16	0,425	0,601
18	0,399	0,564
20	0,377	0,534
22	0,359	0,508
24	0,343	0,485
26	0,329	0,465
28	0,317	0,448
30	0,306	0,432

Пример. Исследованием установлено, что существует зависимость между потреблением бензина (тыс т) (х) и потреблением дизельного топлива (тыс т) (у) по областям Украины за 2000г. Анализ распределения показал, что распределение частот данных совокупностей отлично от нормального. Построим график сопряженных пар данных показателей (рис.14).

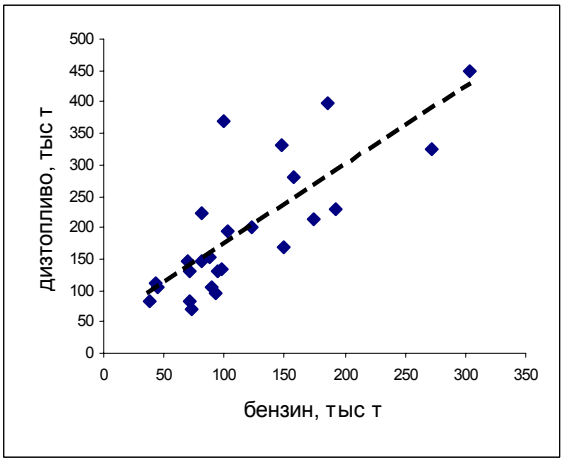


Рис.14. Взаимозависимость между показателями потребления бензина (тыс т) (х) и потребления дизельного топлива (тыс т) (у) по областям Украины за 2000г.

Построенный график показывает наличие положительной корреляции между данными показателями. Рассчитаем коэффициент ранговой корреляции Спирмена. Для этого составим следующую таблицу (табл.13):

Табл.13.

X	x'	y	y'	x'-y'	$(x'-y')^2$
81	17	224,5	8	9	81
103,3	10	195,8	11	-1	1
71,9	20	82,6	23	-3	9
303,3	1	448,1	1	0	0
185,1	4	399	2	2	4
97,4	12	133,6	16	-4	16
92,6	14	95	22	-8	64
174,3	5	214,3	9	-4	16
99,4	11	370,8	3	8	64
271,4	2	326,3	5	-3	9
42,5	24	111,8	19	5	25
123,1	9	200,5	10	-1	1
148,6	7	170,3	12	-5	25
69,1	22	147,8	14	8	64
147,3	8	332,4	4	4	16
158,3	6	281,1	6	0	0
44,9	23	103,8	21	2	4
95,2	13	129	18	-5	25
37,7	25	82	24	1	1
192,1	3	228,6	7	-4	16
71,7	21	130	17	4	16
80,8	18	146,3	15	3	9
87,5	16	153,7	13	3	9
73,4	19	69,4	25	-6	36
88,8	15	106	20	-5	25
				Σ=	536

Полученное значение суммы подставим в формулу (37):

r_s= 1- (6*536/19656)=0.84.

Определим достоверность полученного коэффициента по табл.13. Для N=24 и P=0,99 минимальное достоверное значение = 0,485. 0.84>0.485, значит, взаимосвязь между данными показателями является существенной.

Коэффициент ранговой корреляции рассчитывается главным образом тогда, когда нужно выявить приближенную тесноту связи. Однако он представляет весьма значительную ценность для географов, так как нам нередко приходится иметь дело с данными о многих природных и общественных явлениях, которые выражены в рангах и баллах [13].

Вопросы:

1. *Что такое ранговая корреляция. Ее отличие от обычной линейной парной корреляции.*
2. *В чем заключается суть ранжирования?*
3. *Каков порядок проверки достоверности полученного коэффициента корреляции?*

Задание. *По данным, использованным для расчета коэффициента корреляции Пирсона, рассчитайте коэффициент ранговой корреляции Спирмена. Велико ли различие между полученными значениями. Чем Вы можете это объяснить?.*

3.3. Коэффициент прямолинейной корреляции в случае качественных данных.

Как уже упоминалось, сравниваемые выборочные совокупности могут быть представлены как количественными, так и качественными данными. Зачастую сталкиваясь с необходимостью оценить взаимосвязь между данными качественного характера, исследователь попадает в затруднительную ситуацию относительно адекватности переведения качественной информации в количественную. Одним из выходов является проведение ранжирования и последующего вычисления рангового коэффициента корреляции. Однако нередко приходится иметь дело в данными «одного уровня», т.е. данными, которые нельзя расставить согласно их приоритетности. В этом случае целесообразным является применение *коэффициента прямолинейной корреляции в случае качественных данных* [14].

Как правило, на практике работают с описательными признаками, имеющие несколько разновидностей каждая. В подобном случае первым этапом для вычисления значения корреляции следует составить так называемую *корреляционную таблицу,* где где X_1, $X_2 \ldots X_n$ обозначают разновидности одного признака, а Y_1, Y_2, $\ldots Y_n$ - разновидности другого. При наличии такой таблицы коэффициент корреляции находят по формуле:

$$r = \sqrt{\frac{\varphi^2}{1+\varphi^2}} \sqrt{\frac{m \cdot n}{(m-1)(n-1)}}, \qquad (38)$$

где m – число разновидностей признака X, n - число разновидностей признака Y, φ^2 – коэффициент связи, рассчитываемый по формуле:

$$\varphi^2 = \xi^2 - \frac{(m-1)(n-1)}{\sum f}, \qquad (39)$$

где Σf – общее число наблюдений, а ξ^2 – число, алгоритм получения которого приведен ниже (по[14]).

Исследовано состояние 130 природно-территориальных комплексов – X_i и установлена степень антропогенной нагрузки – Y_j. Относительно показателя X были приняты 4 градации: X_1- плохое состояние, X_2- среднее состояние, X_3 - хорошее и X_4 -очень хорошее. Для явления Y условно приняты также 4 разновидности: Y_1 - фактическое отсутствие антропогенной нагрузки (условия заповедников, отдаленных таежных и т.п. районов), Y_2 - низкая антропогенная нагрузка, Y_2 - средняя и Y_3 - высокая. Необходимо определить степень связи уровня антропогенной нагрузки с состоянием ландшафта. Составим таблицу (Табл.14) частот отдельных комбинаций разновидностей признаков:

Вычисление коэффициента корреляции проводится по следующему алгоритму:

1. Каждую частоту возводим в квадрат и записываем в соответствующем поле таблицы (первое сверху число в скобках в каждой графе).

2. Полученные квадраты делим на сумму всех частот соответствующего столбца (например, для j ,= 1-25 : 30 = 0,833; 25. : 30= = 0.833; 400: 30 = 13,333 и т.д.). Результаты записывают в соответствующем поле таблицы (нижнее число в скобках в каждой графе).

3. Складываем полученные частные каждой строки (например, для i = 1 – 2.500+2.500+20.000 = 25.000 и т.д.)

4. Полученные таким образом итоги делим на общее количество соответствующей разновидности признака X (например, 25,000/40=0,625; 11,458/30=0,382 и т.д.). Полученные частные заносим в соответствующие строки столбца «частное».

5. Полученные частные складываем

(0,625+0,382+0,486=1,875)

6. Находим коэффициент ξ^2, вычитая из итога единицу. (в примере ξ^2 = 1,875-1=0,875.

7. После этого по формуле (39) находим коэффициент связи:

$$\varphi^2 = 0{,}875 - \frac{(4-1)(4-1)}{130} = 0{,}806$$

8. Затем по формуле (38) вычисляем коэффициент корреляции:

$$r = \sqrt{\frac{0.806}{1+0.806}} * \sqrt{\frac{4.4}{(4-1)(4-1)}} = 0.77$$

достоверность полученной связи можно определить с помощью таблицы значимостей коэффициентов корреляции, приведенной в приложении 5.

Табл.14

антроп. нагр. / сост. ландшафта	Y_1	Y_2	Y_3	Y_4	ИТОГО	ЧАСТНОЕ
X_1	-	10 (100) (2.500)	10 (100) (2.500)	20 (400) (20,000)	40 (25000)	0,625
X_2	5 (25) (0,833)	5 (25) (0,625)	20 (400) (10,000)	-	30 (11.458)	0,382
X_3	5 (25) (0,833)	20 (400) (10,000)	5 (25) (0,625)	-	30 (11,458)	0,382
X_4	20 (400) (13,333)	5 (25) (0,625)	5 (25) (0,625)		30 (14,583)	0,486
ИТОГО	30	40	40	20	130	1,875 (ξ^2+1)

Вопросы:

1. *Для каких данных применяется данных коэффициент? Приведите пример.*
2. *На основе какого ключевого показателя построен расчет этого коэффициента?*

Задание. *Проведите исследование по расчету данного коэффициента корреляции. Для этого вы можете взять две карты (к примеру, карта условий жизни населения и степени загрязнения) для одной и той же территории и затем по* ***одной и той же*** *сети точек (желательно, регулярной – см правила составления систематической выборки) снять значения характеристик. Далее, используя методику, изложенную выше, Вычислите коэффициент корреляции между качественными признаками.*

3.4. Коэффициент множественной корреляции.

Иногда в определенных исследованиях необходимо оценить степень общей взаимосвязи между некоторым показателем X и влияющими на него факторами (Y,Z…,N). В этом случае используется коэффициент множественной корреляции *R,* формула которого для трех коррелируемых величин имеет вид [1]:

$$R = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{1 - r_{yz}^2}}, \qquad (40)$$

где r_{xy}, r_{xz}, r_{yz} – коэффициенты парной корреляции.

Пример. При анализе ситуации в сфере обращения с твердыми бытовыми отходами в Украине: использовались следующие показатели: количество мусора, тыс. м. куб (Z); количество свалок (Y); средняя стоимость вывоза одной тонны отходов, грн (X). В

процессе исследования была поставлена цель определить основной фактор, влияющий на «ценообразование» вывоза мусора на свалки. Была определены парные коэффициенты корреляции (табл.15.), однако проверка на значимость полученных коэффициентов корреляции показала, что влияние каждого отдельного показателя на стоимость вывоза является незначительной.

Табл.15.

	X	Y	Z
X	1	-0,01	0,28
Y	-0,01	1	0,57
Z	0,28	0,57	1

Поэтому был рассчитан коэффициент множественной корреляции по формуле (40):

$$R = \sqrt{\frac{-0{,}01^2 + 0{,}28^2 - 2*(-0{,}01)*0{,}28*0{,}57}{1-0{,}57}} = 0{,}35$$

Согласно приложению 5, для количества элементов в выборке, равному 25 минимально значимым коэффициентом корреляции является 0,38, что больше полученного значения (0,35). Отсюда можно сделать вывод, что выбранные показатели даже совместно не оказывают существенное влияние на стоимость вывоза 1 тонны твердых бытовых отходов.

3.5. Частный коэффициент корреляции.

В ряде случаев требуется выяснить, не является ли связь между двумя какими-либо явлениями X и Y обусловленной влиянием третьего явления Z, либо наоборот, связь между некоторыми двумя показателями невозможно выявить из-за сильного влияния третьего показателя [1]. В этом случае целесообразно использовать частный коэффициент корреляции, который позволяет оценить связь между двумя интересующими нас явлениями при исключении третьего. Формула частного коэффициента корреляции имеет вид:

$$r_{xy/z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}}, \qquad (41)$$

где r_{xy}, r_{xz}, r_{yz} – коэффициенты парной корреляции.

Пример. Используя данные предыдущего примера попробуем оценить степень взаимосвязи между показателями X иY при исключении влияния Z по формуле (41):

$$r_{xy/z} = \frac{(-0{,}01) - 0{,}28*0{,}57}{\sqrt{\left(1-0{,}28^2\right)\left(1-0{,}57^2\right)}} = -0{,}27.$$

Как можно увидеть, что практически полное отсутствие связи между показателями X иY при определении парного коэффициента корреляции значительно возросло при исключении влияния показателя Z.

Вопросы:

1. *Для каких целей вычисляется множественный коэффициент корреляции?*
2. *Чего желает добиться исследователь, рассчитывая частный коэффициент корреляции?*

Задание. *Для отобранных Вами данных рассчитайте частный и множественный коэффициент корреляции.*

ТЕМА 4. РЕГРЕССИОННЫЙ АНАЛИЗ.

По сравнению с корреляционным анализом, рассмотренным в предыдущем разделе, регрессионный анализ отчасти является уже и методом моделирования, так как исходит из предпосылки, что существует полная (функциональная) зависимость между показателями, и, следовательно, позволяет выразить одни признаки 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:

- путем построения эмпирических линий;
- путем составления уравнения и затем - построения теоретических линий регрессии;
- с помощью коэффициента регрессии.

Регрессионный анализ тесно связан с корреляционным, так как уравнение регрессии, построенное для слабо коррелирующих признаков, не будет достоверным. Поэтому регрессионный анализ является своего рода *продолжением* корреляционного. Следовательно, уравнение наиболее точно выражает зависимость между двумя переменными *(x,y),* если корреляция между ними близка к единице.

Существует два способа составления уравнений регрессии:

- *способ координат* точек, с использованием двух-трех точек, расположенных на эмпирической линии (желательно в начале, середине и конце ее). Этот способ используется в тех случаях, когда расчет не требует большой точности (наиболее часто – при составлении линий линейной регрессии.
- *способ наименьших квадратов*, более точный, так как, для составления уравнения регрессии привлекаются все сопряженные наблюдения. Рассмотрим наиболее простые способы составления уравнений регрессии.

4.1.Линейная регрессия.

Уравнение линейной регрессии в общей форме записывается следующим образом:

$$y = kx + b\,, \qquad (42)$$

где y - значение зависимой переменной; x - значение независимой переменной; k - коэффициент, 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.

Используя указанные два способа, найдем построим уравнение линейной регрессии для 2 сопряженных вариационных рядов гипотетических данных., сопряженный вариационный ряд которых имеет вид (Табл.16)

Табл.16.

49,1	14,6	29,8	36,2	14,4	48	51,7	30,2	40,8	22	34,3	21	11	31,4
215,4	45,1	149	130	64,1	157,3	186	83,4	170	77	112	105,2	67	129

Пример. Коэффициент корреляции для отобранных данных, составляет 0,91, что говорит о сильной положительной взаимосвязи, что говорит о том, что мы сможем получить достоверное уравнение регрессии.
Вначале для решения поставленной задачи используем способ координат точек.

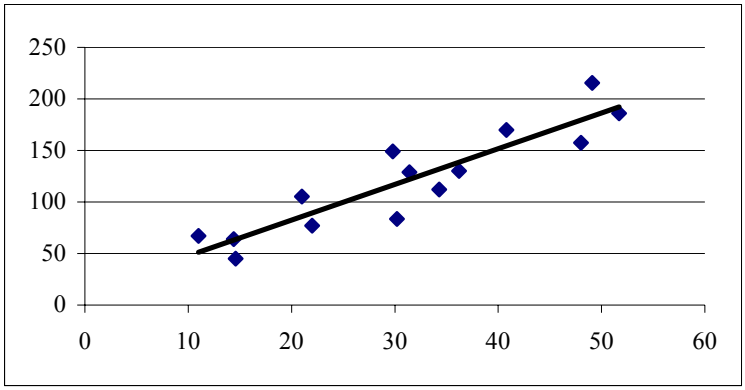


Рис.15. Точечный график для сопряженного вариационного ряда.

1. Результаты наблюдений наносим на график, затем проводим прямую так, чтобы число точек по обе стороны линии было одинаковым (рис. 15).
2. Для расчета параметров а и Ъ 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: да=69, </$_t$ = = 18; *$_8$=95, «/в=90. Подставляя значения переменных в общее уравнение прямой, получаем систему уравнений:

$$\begin{cases} 64{,}1 = k \cdot 14.4 + b \\ 186 = k \cdot 51.7 + b \end{cases} = \begin{cases} b = -14.4k + 64.1 \\ 186 = 51.7k - 14.4k + 64.1 \end{cases} = \begin{cases} b = -14.4k + 64.1 \\ 186 = 37.3k + 64.1 \end{cases} =$$

$$\begin{cases} b = -14.4k + 64.1 \\ 37.3k = 121.9 \end{cases} = \begin{cases} b = 17.01 \\ k = 3.27 \end{cases}$$

Следовательно, уравнение регрессии в данном случае будет иметь вид:

$$y = 3.27x + 17.01$$

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 *k* и *b* составляется следующая система уравнений:

$$\begin{cases} \sum y = k\sum x + bn\,; \\ \sum xy = k\sum x^2 + b\sum x.' \end{cases} \qquad (43)$$

где *n* – количество пар в сопряженном вариационном ряду.

Для вычисления коэффициентов *k,x* уравнения регрессии составим таблицу (Табл.17). Затем решим систему уравнений, подставив в формулу (43) соответствующие значения:

$$\begin{cases} 1690.5 = 434.5k + 15b \\ 60457 = 15791k + 434.5b \end{cases} = \begin{cases} b = 112.7 - 28.97k \\ 60457 = 15791k + 48968.15 - 12587.465k \end{cases} = \begin{cases} b = 9.28 \\ k = 3.59 \end{cases}$$

В итоге получаем уравнение регрессии для нашей совокупности данных:

$$y = 3.59x + 9.28$$

Как можно увидеть, уравнения, полученные обоими способами, имеют некоторые расхождения. Поэтому проверим достоверность полученных уравнений

Табл.17.

	x	y	x^2	xy
1	49,1	215,4	2410,81	10576,14
2	14,6	45,1	213,16	658,46
3	29,8	149	888,04	4440,2
4	36,2	130	1310,44	4706
5	14,4	64,1	207,36	923,04
6	48	157,3	2304	7550,4
7	51,7	186	2672,89	9616,2
8	30,2	83,4	912,04	2518,68
9	40,8	170	1664,64	6936
10	22	77	484	1694
11	34,3	112	1176,49	3841,6
12	21	105,2	441	2209,2
13	11	67	121	737
14	31,4	129	985,96	4050,6
Σ	434,5	1690,5	15791,83	60457,52

Проверка на достоверность уравнения регрессии может быть осуществлена многими способами. Среди них выделим коэффициент выравнивания точности линии r_1, критерий χ^2 и критерий Колмогорова λ. Проведем проверку на достоверность полученных нами результатов при помощи первых двух критериев.

Коэффициент выравнивания точности линии r_1.

Он отражает степень приближения (соответствия) фактических данных наблюдения вероятным. Этот коэффициент определяется следующим образом:

$$r_1 = \sqrt{\frac{\sum \alpha^2 - \sum \beta^2}{\sum \alpha^2}} = \sqrt{\frac{\sum \left(y_\phi - M_\phi\right)^2 - \sum \left(y_\phi - y_\text{в}\right)^2}{\sum \left(y_\phi - M_\phi\right)^2}}, \qquad (44)$$

где $(y_\phi - M_\phi) = \alpha$ - отклонение индивидуальных вариант от среднего арифметического по y; $(y_\phi - y_в) = \beta$ - отклонение индивидуальных экспериментальных (теоретических) вариант от расчетных по уравнению. Принято считать, что если r_1>0.95, то уравнение регрессии соответствует точному положению линии на графике. При r_1<0.95 необходимо найти другую математическую зависимость [23].

Пример. Рассчитаем данный коэффициент для нашего уравнения регрессии, полученного методом наименьших квадратов. Для этого составим таблицу расчета данных для определения коэффициента точности выравнивания линии (табл.18). Определим искомый коэффициент по формуле (44):

$$r_1 = \sqrt{\frac{33378 - 528.47}{33378}} = 0{,}83.$$

Полученное нами значение коэффициента говорит о необходимости подбора другой математической зависимости между данными показателями, т.е. полученное уравнение не недостоверно.

Табл 18.

	$y_ф$	$y_в$	$y_ф$-$M_ф$	$(y_ф$-$M_ф)^2$	$y_ф - y_в$	$(y_ф - y_в)^2$
1	215,4	185,549	94,65	8958,623	29,851	891,0822
2	45,1	61,694	-75,65	5722,923	-16,594	275,3608
3	149	116,262	28,25	798,0625	32,738	1071,777
4	130	139,238	9,25	85,5625	-9,238	85,34064
5	64,1	60,976	-56,65	3209,223	3,124	9,759376
6	157,3	181,6	36,55	1335,903	-24,3	590,49
7	186	194,883	65,25	4257,563	-8,883	78,90769
8	83,4	117,698	-37,35	1395,023	-34,298	1176,353
9	170	155,752	49,25	2425,563	14,248	203,0055
10	77	88,26	-43,75	1914,063	-11,26	126,7876
11	112	132,417	-8,75	76,5625	-20,417	416,8539
12	105,2	84,67	-15,55	241,8025	20,53	421,4809
13	67	48,77	-53,75	2889,063	18,23	332,3329
14	129	122,006	8,25	68,0625	6,994	48,91604
	120,75			33378		5728,447

Критерий хи-квадрат.

При изложении описания данного критерия мы не приводили примеров его практического применения. Поэтому определим достоверность полученного уравнения при помощи критерия хи-квадрат. Составим следующую таблицу (Табл.19):

Табл.19.

y’	y-y’	(y-y’)	$(y-y')^2/y$
185,549	29,851	891,0822	4,802409
61,694	-16,594	275,3608	4,463333
116,262	32,738	1071,777	9,218632
139,238	-9,238	85,34064	0,612912
60,976	3,124	9,759376	0,160053
181,6	-24,3	590,49	3,251597
194,883	-8,883	78,90769	0,404898
117,698	-34,298	1176,353	9,994671
155,752	14,248	203,0055	1,303389
88,26	-11,26	126,7876	1,436524
132,417	-20,417	416,8539	3,148039
84,67	20,53	421,4809	4,977925
48,77	18,23	332,3329	6,81429
122,006	6,994	48,91604	0,400931
		$\chi^2=$	50,9896

Фактического значение больше теоретического (50,9896>31.264 при P=0,999) для данного числа степеней свободы.

4.2. 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

$$y = k / x + b, \qquad (45)$$

где x - аргумент; y - функция; k 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. Для этих же параметров по способу наименьших квадратов используется система уравнений:

$$\begin{cases} \sum xy = kn + b\sum x; \\ \sum x^2 y = k\sum x + b\sum x^2. \end{cases} \quad (46)$$

Порядок расчета достоверности аналогичен примерам, приведенным для случая линейной регрессии.

4.3. Параболическая зависимость.

Общее уравнение параболы n-го порядка имеет вид:

$$y = ax^n + bx^{n-1} + cx^{n-2} + \ldots + kx + l \quad (47)$$

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

$$y = ax^2 + bx + c \quad (48)$$

Формула для расчета коэффициентов параболического уравнения регрессии имеет вид:

$$\begin{cases} \sum y = a\sum x^2 + b\sum x + cn; \\ \sum xy = a\sum x^3 + b\sum x^2 + c\sum x; \\ \sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2. \end{cases} \quad (49)$$

Вопросы:

1. *Что такое регрессионный анализ? В чем состоит смысл применения?*
2. *Способ координат. Методика расчета коэффициентов линейного уравнения регрессии по методу координат.*
3. *В чем заключается особенность использования способа наименьших квадратов?*

Задание. *Используя данные, по которым был рассчитан коэффициент корреляции в предыдущем разделе, выполните следующие действия:*

а) Постройте график взаимозависимости между явлениями, по которому определите (предварительно) форму линии регрессии;

б) Рассчитайте уравнение регрессии и оцените достоверность полученного уравнения.

ИТОГОВОЕ ЗАДАНИЕ1 .

По данным Харьковского отделения госкомгидромет проведена оценка ветрового потенциала Харьковской области (рис.16). Необходимо провести более глубокий анализ территории области с точки зрения перспективности размещения ветровых установок.

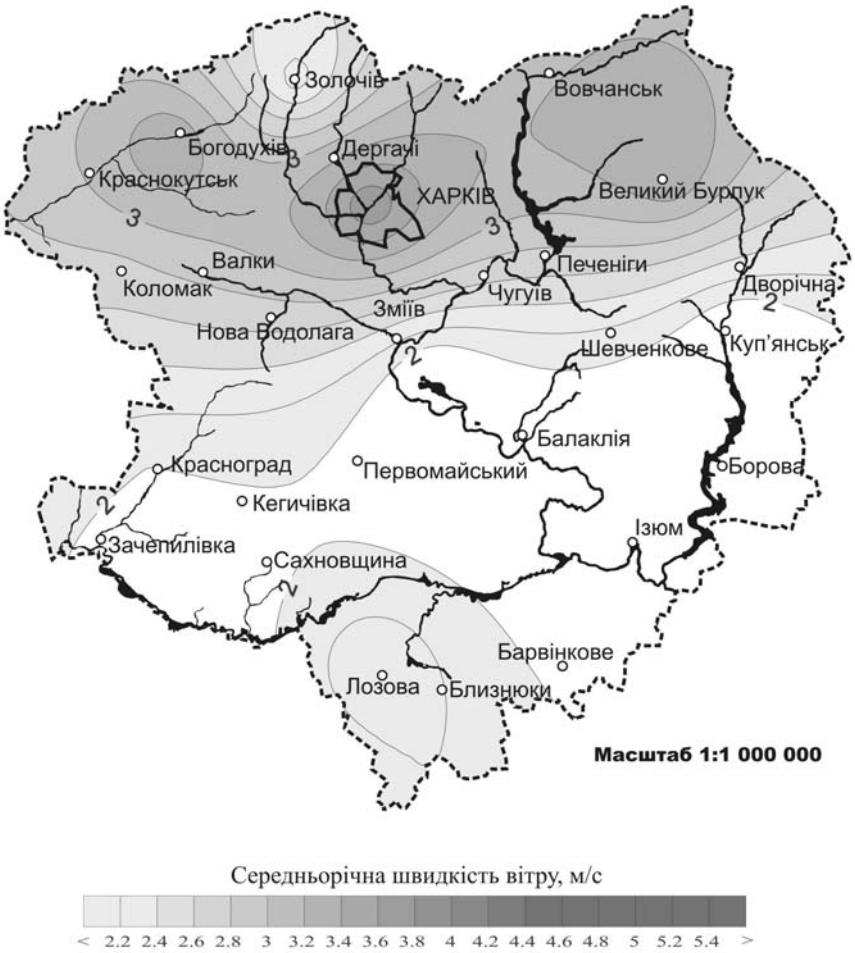


Рис.16. Распределение скоростей ветра по территории Харьковской области на высоте 18 м.

В табл. приведены средние за трехлетний период скорости ветра на каждый месяц по метеостанциям Харьковской области.

Необходимо определить (метод определения выбрать самостоятельно из выученных Вами):

1. *Среднюю скорость ветра для Харьковской области (общую и отдельно по каждому месяцу);*
2. *Определить моду для Харьковской области в целом (т.е. по используя все данные, приведенные в табл.);*
3. *Построить распределение частот для Харьковской области в целом (используя среднемесячные значения скоростей ветра, рассчитанные в пункте 1 данного задания), провести анализ на нормальность;*
4. *Провести районирование территории, используя критерии различия между выборками (необходимо определить любой из выученных критериев различия для всех возможных пар метеостанций – построить матрицу данного критерия);*
5. *Рассчитать коэффициенты корреляции между всеми возможными парами метеостанций и оценить степень их взаимосвязанности (построить корреляционную матрицу).*

Табл. 20

Назв метеостанции	январь	февраль	Март	апрель	май	июнь	июль	август	сентябрь	октябрь	ноябрь	декабрь
Золочев	2,2	2,7	2,8	2	2,3	2	1,5	1,5	1,6	2,1	2,5	2,4
Богодухов	3,8	4,8	4,4	3,2	3,5	3,1	2,6	2,7	2,9	2,3	3,6	3,5
Вел Бурлук	3,4	4,2	4,1	3	3,6	3,1	2,8	2,6	3	3,3	3,9	3,6
Коломак	3,1	3,7	3,4	2,5	2,7	2,4	2	1,9	2,2	2,8	3,1	3
Харьков	3,9	4,6	4,4	3,7	3,7	3,6	3,2	3,2	3,5	3,6	4,1	3,9
Купянск	1,8	2,3	2,2	1,3	1,9	1,6	1,3	1	1,4	1,5	1,8	1,7
Комсомольское	2	2,3	2,1	1,6	1,4	1,4	1,2	1,1	1,3	1,4	2	2
Красноград	2,1	2,6	2,5	1,9	1,8	1,7	1,5	1,5	1,6	1,9	2,3	2,5
Изюм	1,8	2,2	2,1	1,4	1,9	1,7	1,5	1,2	1,5	1,7	1,8	1,8
Лозовая	2,5	3	3	2,4	2,4	2,1	1,7	1,4	2,1	2,3	2,8	2,7

<u>ИТОГОВОЕ ЗАДАНИЕ2.</u> *На основе проведенных Вами расчетов, сделайте общие выводы относительно результатов ваших исследований. Опираясь на помощь преподавателя, напишите научную статья и подготовьте Ваше выступление.*

ТЕМА 5. ТАКСОНОМИЧЕСКИЙ (КЛАСТЕРНЫЙ) АНАЛИЗ.

Широко представленной в прикладных географических исследованиях проблемой является группировка территориальных единиц по комплексу показателей. Конечной целью подобного исследования является построение типологических синтетических карт. Нередко по результатам подобных исследований становится возможным проведение оценки территории по степени соответствия условиям поставленной в рамках исследования задачи.

Наибольшее распространение в решении подобного рода задач получил *таксономический анализ.* В рамках этого вида анализа исследуются многопараметрические географические объекты, которые по определенным алгоритмам классифицируются на группы (таксоны, кластеры) на основе подобия их внутренней структуры, представленной посредством системы показателей.

В качестве характеристики степени сходства между объектами в таксономическом анализе используется акт называемое «таксономическое расстояние», определяющее степень удаленности данных объектов друг от друга в многомерном математическом пространстве. Наиболее часто эти расстояния вычисляют по формуле Пифагора, используя Евклидову метрику пространства [12]:

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2}\ ,\ i = 1,2,3,\ldots, n,\ k = 1,2,3,\ldots, n. \qquad (50)$$

В качестве исходного материала при проведении таксономического анализа используется таблица, в которой в строках записываются данные по каждой точке измерения, а в столбцах – значения данного показателя по всей точкам.

Пример. В предыдущих разделах автор уже касался вопроса анализа ситуации в сфере твердых бытовых отходов по областям Украины. С помощью таксономического анализа выделим группы областей, имеющих сходную ситуацию в сфере обращения с ТБО. Целью подобного исследования является дифференциация областей Украины с точки зрения степени сходства в сфере методов управления отходами.

В рамках данного исследования использовались такие показатели:

1. *Объем вывезенного мусора, тыс. м. куб;*
2. *Количество свалок и полигонов ТБО;*
3. *Общая площадь свалок и полигонов;*
4. *Средняя стоимость вывоза 1 м куб ТБО, грн.*

Таблица исходных показателей будет иметь следующий вид (Табл. 21).

Табл.21.

Области \ Пок-ли*	1	2	3	4
Крым	2389	28	215,4	5
Винницкая	352	35	97,7	6,4
Волынская	105	24	80,5	4,64
Днепропетровская	2818	33	176,5	14,2
Донецкая	2750	62	260	5,8
Житомирская	485	24	101,7	2,62
Закарпатская	403	18	45,1	5,9
Запорожская	1555	27	210	3,29
Ивано-Франковская	428	29	60,6	14
Киевская	3514	30	187	15,6
Кировоградская	326	20	64,1	4
Луганская	1290	45	157,3	3,85
Львовская	1610	48	135,1	4,6
Николаевская	523	20	110	10,2
Одесская	3400	39	275	4,65
Полтавская	986	32	118,5	4,6
Ровненская	510	23	134,6	9,1

Сумская	483	22	55,9	5,57
Тернопольская	495	24	83,4	5,6
Харьковская	1405	32	102	10
Херсонская	320	22	66	6,8
Хмельницкая	860	32	77	12
Черкасская	807	21	105,2	4,49
Черновицкая	447	11	67	6,59
Черниговская	520	26	129	3,9

**Здесь и далее номера столбцов в таблице соответствуют порядковым номерам показателей в списке, приведенном выше таблицы.*

Важным условием достоверности результатов, полученных в таксономическом анализе, является «ортогональность» данных, на основе которых проводится исследование, т.е. исходные показатели не должны иметь высоких значений корреляции между собой.

Пример. Построим корреляционную матрицу для показателей, представленных в табл.21. (Табл.22):

Табл.22. Пример корреляционной матрицы.

Показатели*	1	2	3	4
1	1	0,574203	**0,861915**	0,277396
2	0,574203	1	**0,630266**	-0,01303
3	**0,861915**	**0,630266**	1	-0,0498
4	0,277396	-0,01303	-0,0498	1

Как можно увидеть, показатель общей площади свалок и полигонов (№3) заметно коррелирует с показателями: объема вывезенного мусора (№1) и числа полигонов (№2) (высокие значения коэффициентов корреляции выделены жирным шрифтом). Это повлекло за собой исключение показателя №3 из последующих расчетов.

Влияние искажений, вызванных высокой взаимозависимостью переменных, можно также, устранить, предварительно взвесив их по компонентным нагрузкам, выделенным с помощью компонентного анализа. Однако, так как данный вид анализа выходит за рамки настоящего пособия, мы предлагаем читателю изучить данный вопрос самостоятельно (компонентный анализ и его использование в географических исследованиях, рассмотрено, например, в [19]*).*

Следующей проблемой, с которой сталкивается исследователь при проведении таксономического (кластерного) анализа, является то, что в большинстве случаев анализируемые показатели, представляют собой несоизмеримые величины (т.е. одни показатели могут быть представлены в кубометрах, в другие в киловаттах на квадратный километр и т.д.). В этом случае возникает необходимость в процедуре стандартизации или нормировки. Существует довольно большое количество способов подобного преобразования, однако мы остановимся на рассмотрении способа *нормировки по дисперсиям*, как наиболее часто использующемся в таксономическом анализе:

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \qquad (51)$$

де x_{ij} – исходные показатели; $\bar{x}_j$ - среднее арифметическое для исходного показателя ; x_{ij} - нормированные значения исходных показателей; σ_j -среднее квадратическое отклонение..

Нормированные показатели образуют таблицу нормированных показателей, идентичную таблице исходных показателей. Следует также отметить, что при нормировании мы не теряем информации относительно взаимозависимости внутри вариационного ряда, мы лишь переводим значения из абсолютных в относительные показатели.

Пример. Проведем нормирование показателей, приведенных в табл.21. Так как показатель №3 был исключен, то номированию будут подлежать лишь показатели №1, №2 и №4. Процедура нормирования будет состоять из двух этапов:

1. Рассчитаем все необходимые для нормирования показатели среднего арифметического (7) и стандартного отклонения (11) (табл.23):

Табл.23.

№ показателя	1	2	4
Среднее	1151,24	29,08	6,936
Стандартное отклонение	1028,817536	10,76227981	3,679695413

2. По формуле (51) рассчитаем нормированные значения для каждого показателя (Табл.24):

Табл.24.

Области / Пок-ли*	1	2	4
Крым	1,20309	-0,10035	-0,52613
Винницкая	-0,77685	0,550069	-0,14566
Волынская	-1,01693	-0,47202	-0,62396
Днепропетровская	1,620073	0,364235	1,974076
Донецкая	1,553978	3,058831	-0,30872
Житомирская	-0,64758	-0,47202	-1,17292
Закарпатская	-0,72728	-1,02952	-0,28155
Запорожская	0,392451	-0,19327	-0,99084
Ивано-Франковская	-0,70298	-0,00743	1,919724
Киевская	2,296578	0,085484	2,354543
Кировоградская	-0,80212	-0,84369	-0,79789
Луганская	0,134873	1,47924	-0,83866
Львовская	0,44591	1,757992	-0,63484
Николаевская	-0,61064	-0,84369	0,88703
Одесская	2,185771	0,921738	-0,62125
Полтавская	-0,16061	0,271318	-0,63484
Ровненская	-0,62328	-0,56494	0,588092
Сумская	-0,64952	-0,65785	-0,37123
Тернопольская	-0,63786	-0,47202	-0,36307
Харьковская	0,246652	0,271318	0,832678
Херсонская	-0,80796	-0,65785	-0,03696
Хмельницкая	-0,28308	0,271318	1,376201
Черкасская	-0,3346	-0,75077	-0,66473
Черновицкая	-0,68451	-1,67994	-0,09403
Черниговская	-0,61356	-0,28618	-0,82507

Все точки измерения показателей представляются в виде точек в многомерном *m*-мерном пространстве, в котором координатами служат нормированные исходные показатели (в нашем примере мы имеем четырехмерное пространство). Таксономические расстояния, соединяющие каждую пару точек (d_{ik}), отражают различия по комплексу показателей территориальных единиц. Поэтому их можно положить в основу дифференциации (т.е. деления на группы) исходной совокупности точек они рассчитываются по формуле, аналогичной (50), однако координатами в которой являются нормированные показатели:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2}\ ,\ i = 1,2,3,\ldots, n,\ k = 1,2,3,\ldots, n. \qquad (52)$$

все рассчитанные расстояния образуют квадратную симметричную матрицу расстояний (о терминах *симметричная* и *квадратная матрица* см. радел «факторный анализ»).

Пример. Матрица расстояний для нашего примера имеет размерность, согласно числу точек измерения (в нашем случае, административных областей) - 26X26 (см. Приложение 7). *Для удобства обозначим области латинскими буквами (табл.25).*

Табл.25.

Крым	A
Винницкая	B
Волынская	C
Днепропетровская	D
Донецкая	E
Житомирская	F
Закарпатская	G
Запорожская	H
Ивано-Франковская	I
Киевская	J
Кировоградская	K
Луганская	L
Львовская	M
Николаевская	N
Одесская	O
Полтавская	P
Ровненская	Q
Сумская	R
Тернопольская	S
Харьковская	T
Херсонская	U
Хмельницкая	V
Черкасская	W
Черновицкая	X
Черниговская	Y

Рассмотрим далее два метода проведения таксономического анализа, наиболе часто встречающиеся в прикладных географических исследованиях: метод «вроцлавской таксономии» и агломеративно-иерархический метод (или метод Берри). Следует отметить, что все изложенные выше этапы работы характерны для обоих этих методов. Различия состоят лишь в методике представления результатов анализа.

5.1 «Вроцлавская таксономия»

Согласно[11], вроцлавский таксономический метод позволяет исследовать внутреннюю структуру совокупности индивидуальных точек. Осуществляется это путем объединения близлежащих индивидуальных точек в группы; нахождения точек или ареалов, которые являются характерными для данной совокупности индивидуальных точек; деления генеральной совокупности на части.

Характерной особенностью многомерной статистической модели «вроцлавской таксономии» является то, что она очень проста и удобна для расчетов, при которых не требуется сложная вычислительная техника.

По результатам расчетов по методу «вроцлавской таксономии» строится дендрит дендрит, который затем делится на части, т.е. проводится группировка исходных точек.

Построение дендрита.

Дендрит строится на базе данных матрицы таксономических расстояний по следующему алгоритму: из первого столбца выбирается наименьшее (но не равное 0) расстояние.

Соответствующие номера столбца и строки определяются номерами двух исследуемых точек, которые наносятся на диаграмму и соединяются линией. Длина линии пропорциональна расстоянию между точками. Соединение точек производится до тех пор, пока все точки не будут сведены в одну систему. Необходимо подчеркнуть, что в процессе построения дендрита линия, соединяющая пункты, не должна замыкаться в цепь.

Объясним данный этап на примере. Итак, мы имеем матрицу расстояний 25Х25 и нам необходимо построить по этой матрице таксономический дендрит. Его построение идет по следующему алгоритму:

1. *Находим в 1 колонке наименьшее расстояние. Таковым является расстояние АЕ.*
2. *Отмечаем расстояние ЕА, т.к. расстояния, обратные выбранным, исключаются.*
3. *Откладываем на графике (в масштабе) расстояние АЕ.*
4. *Находим в колонке В наименьшее значение. Таковым будет расстояние ВК.*
5. *Аналогично пункту 2, исключаем значение КВ.*
6. *Откладываем (в масштабе) отрезок ВК.*

Процедура повторяется подобным образом для каждого столбца. На определенном этапе разрозненные ветви графа будут соединяться в единый дендрит. В конце концов все ветви графа станут единым целым, а также все точки будут обозначены. На этом этапе процедура построения заканчивается. В описываемом примере процедура прекращается на колонке Х.

Хотелось ее раз повторить, что необходимо следить за тем, чтобы ветви не образовывали циклов – это противоречит условиям построения графа.

Для большей наглядности на матрице, используемой в примере, отрезки, участвующие в построении дендрита, показаны укрупненным полужирным курсивом; обратные им отрезки отмечены полужирным зачеркнутым шрифтом; ячейки, которые содержат отрезки, образующие циклы, отмечены серым фоном (см. Приложение 7).

В результате проведенных построений был получен следующий дендрит (рис.17).

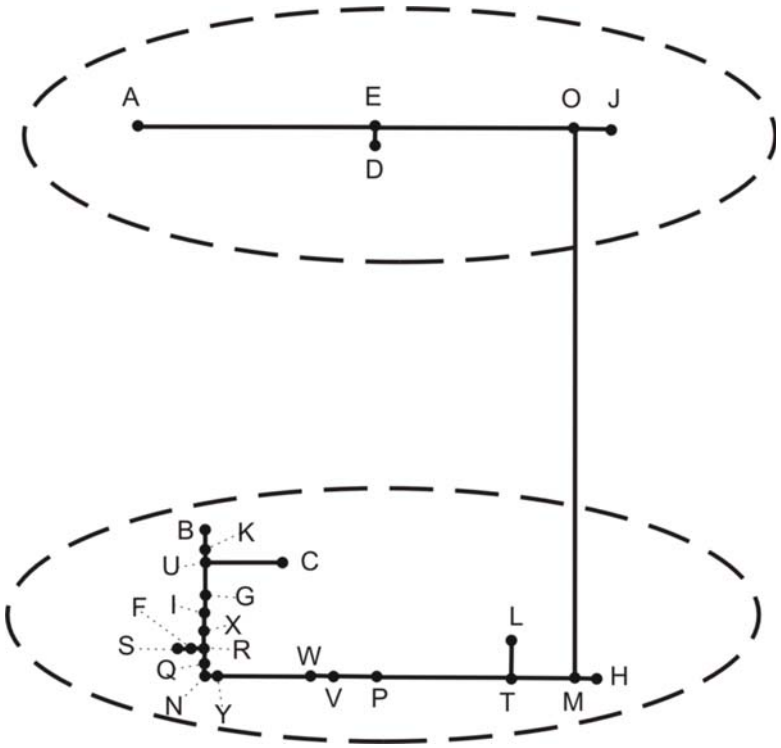


Рис.17. Группировка областей Украины по ситуации в области обращения с ТБО по методу «вроцлавской таксономии».

Деление дендрита на части.

Наиболее простым способом выделения групп по построенному дендриту, является выделение наиболее длинных отрезков. Однако зачастую полученная конфигурация довольно сложна, поэтому отбирают некоторое условное расстояние, затем все точки отстоящие друг от друга на расстояние, меньшее обозначенного, объединяются в группы (рис.17).

5.2 Агломеративно-иерархический метод.

Этот метод является сходным с описанным выше методом «вроцлавской таксономии». Он характеризуется значительно большим объемом вычислений, однако примечателен тем, что широко используется при автоматизации процесса расчетов.

Агломеративно-иерархический метод таксономического анализа предполагает аналитическое построение и графическое изображение процесса группировки территориальных единиц в виде «дерева объединений», тогда как граф, строящийся по методу «вроцлавской таксономии», показывает степень близости территориальных единиц по анализируемому комплексу показателей. Этот метод также основан на построении матрицы таксономических расстояний [12].

Алгоритм построения «дерева объединений» состоит из следующих этапов:

1. Из матрицы расстояний выбирают отрезок минимальной длины и объединяют территориальные единицы, которые данное расстояние символично соединяет, считая в дальнейших вычислениях их как одну точку.
2. Для вновь образованной единицы рассчитываются среднеарифметические значения исходных показателей (в расчете учитываются значения, характеризующие две исходные точки).
3. Строится новая матрица таксономических расстояний (для *n-1* территориальных единиц).

Процесс объединения территориальных единиц продолжается до тех пор, пока все единицы не сольются в один таксон.

Пример. Построим «древо объединений» для нашей матрицы таксономических расстояний. Оно будет иметь следующий вид(рис.18):

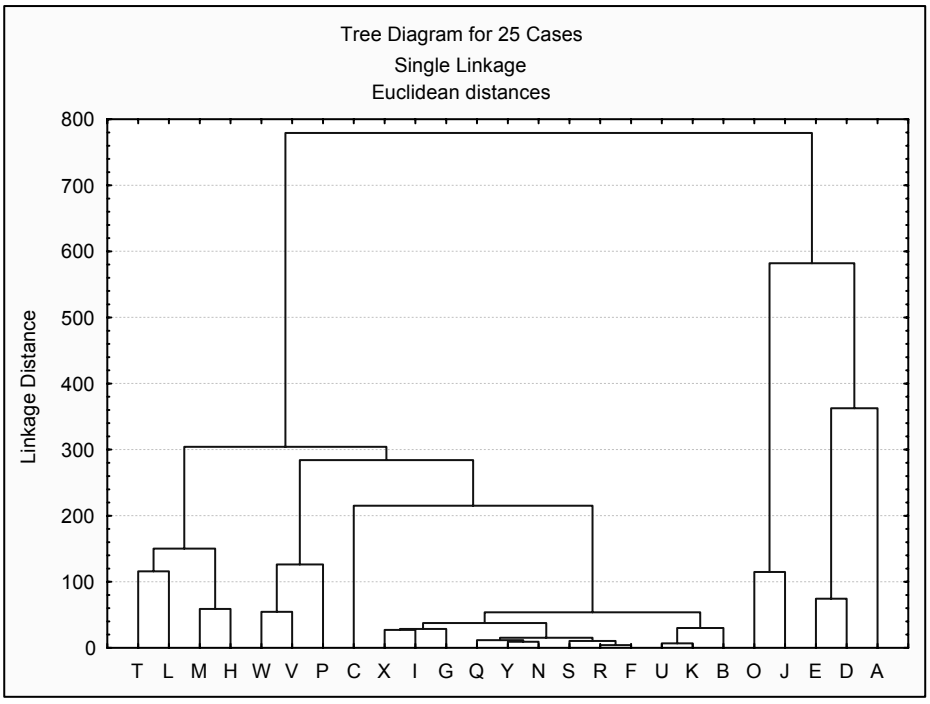


Рис.18. Пример «дерева объединений»

В результате проведенного исследования бала произведено деление областей Украины на две группы: группа №1 характеризуется значительной напряженностью в сфере обращения с ТБО (большие объемы ТБО, высокая цена вывоза, и т.д.). Наоборот, группа №2 характеризуется слабой напряженностью, т.к. показатели, которые были положены в основу их группировки, в данных областях имеют незначительные величины.

Следует также отметить, что в 1 группе, несмотря на малое количество в ней областей -5 – расстояния между отдельными областями довольно велико (это особенно видно на «вроцлавском» дендрите). Напротив, большинство областей из группы№2 практически «сливаются» в одно. Это значит, что ситуация в областях с незначительной напряженностью в сфере обращения с ТБО позволяет использовать единую стратегию управления бытовыми отходами. И наоборот, каждая из областей с напряженной ситуацией требует, своего рода, индивидуального подхода.

Вопросы:

1. *При решении каких задач наиболее часто используется таксономический анализ?*
2. *Что принимается за меру различия или сходства показателей в таксономическом анализе?*
3. *Каков алгоритм расчетов таксономических расстояний в таксономическом анализе?*
4. *Какими положительными качествами характеризуется «вроцлавский» метод проведения таксономического анализа?*
5. *Что положительного и отрицательного, на Ваш взгляд, есть в агломеративно-иерархическом методе Б. Берри?*

Задания:

1. *Дана следующая матрица расстояний (табл.26):*

Табл.26.

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	0,863	1,447	3,347	3,622	3,527	3,974	3,878	4,596	1,656	2,847	4,172
B	0,863	0	1,379	3,168	3,715	3,424	3,713	3,843	4,327	1,822	3,088	3,908
C	1,447	1,379	0	2,866	3,18	3,47	3,527	3,798	3,392	1,377	2,229	3,457
D	3,347	3,168	2,866	0	1,801	1,519	2,898	3,325	3,457	2,535	3,942	3,794
E	3,622	3,715	3,18	1,801	0	2,188	3,212	2,974	4,103	2,288	1,924	3,193
F	3,527	3,424	3,47	1,519	2,188	0	4,079	4,118	4,902	2,905	3,537	3,877
G	3,974	3,713	3,527	2,898	3,212	4,079	0	1,495	3,069	3,189	3,253	4,47
H	3,878	3,843	3,798	3,325	2,974	4,118	1,495	0	4,248	2,98	2,935	4,446
I	4,596	4,327	3,392	3,457	4,103	4,902	3,069	4,248	0	3,951	3,775	3,875
J	1,656	1,822	1,377	2,535	2,288	2,905	3,189	2,98	3,951	0	1,342	2,914
K	2,847	3,088	2,229	3,942	1,924	3,537	3,253	2,935	3,775	1,342	0	2,921
L	4,172	3,908	3,457	3,794	3,193	3,877	4,47	4,446	3,875	2,914	2,921	0

Используя «вроцлавскую» методику, построить дендрит для данной совокупности точек. Опишите алгоритм проведения исследования.

2. *На рис.18 для выделения 2 групп областей был выбран уровень распознавания 700. На каком уровне необходимо провести линию, чтобы получить теперь в каждой из выделенных групп по 3 подгруппы?*
3. *Выделите обозначенные подгруппы на «вроцлавском» дендрите (рис.17).*

ТЕМА 6. ФАКТОРНЫЙ АНАЛИЗ.

Аналогично таксономическому анализу, рассмотренному в предыдущем параграфе, факторный анализ также относится к многомерным видам статистического анализа. Основным предположением факторного анализа является утверждение о том, что явления в определенной области исследования, несмотря на свою разнородность и изменчивость, могут быть описаны относительно небольшим количеством функциональных единиц, параметров или факторов [15].

Факторный анализ не ограничивается сопоставлением изменений, лежащих на поверхности явлений, которые, как правило, представлены определенными показателями. Он стремится обнаружить основные влияния, лежащие в основе этих явлений, которыми и являются факторы. Например, имея данные по 33 цензовым округам 2 провинций Канады (13 различных показателей) В.М, Жуковская [7] выявила 3 фактора, лежащие в основе типологизации сельского хозяйства для данной территории:

1. фактор товарной специализации на производстве пшеницы;
2. фактор уровня развития капиталистических отношений в сельском хозяйстве;
3. фактор роста населения за счет миграции.
4. На основе выделенных факторов были построены карты, показывающие степень влияния (нагрузки) того или иного фактора в каждом округе.

В факторном анализе в качестве критерия сходства или различия между переменными используется корреляция между переменными. Это является существенным отличием от таксономического анализа, где основным критерием является «таксономическое» расстояние [19].

6.1 Основные понятия корреляционного анализа.

Прежде, чем приступить к изложению основных элементов факторного анализа, рассмотрим основным математические понятия, которые будут встречаться в процессе изложения. Речь пойдет о понятии матриц, дисперсии и корреляции.

А) Матрицы.

Безусловно, читатель изучал данный раздел математики в рамках курса высшей математики. Поэтому ограничимся лишь основными определениями, которые необходимы для дальнейшего изложения материала.

Матрицей называется прямоугольная или квадратная таблица чисел, рассматриваемая безотносительно к тому, что именно представляют собой эти числа и существуют ли между ними какие-то заранее определенные закономерности. Вертикальный ряд чисел, расположенных в матрице одно над другим, называется *столбцом*, горизонтальный ряд чисел – *строкой*. Матрица, в которой число строк равно числу столбцов, называется *квадратной*. В тех случаях, когда нужно обозначить какие-либо элементы матрицы, им приписываются соответствующие индексы, первый из которых указывает номер строки, а второй – номер столбца, в котором находится данный элемент.

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ 1 & a_{11} & a_{12} & a_{13} \\ 2 & a_{21} & a_{22} & a_{23} \\ 3 & a_{31} & a_{32} & a_{33} \end{array}$$

Рис.19. Квадратная матрица 3×3.

Таким образом, в матрице, показанной на рис.19, символ a_{12} обозначает элемент, находящийся на пересечении первой строки и второго столбца. Вся матрица обозначается *A*.

О матрице, имеющей m строк и n столбцов говорят, что ее порядок составляет $m \times n$. Квадратная матрица $n \times n$ имеет порядок n.

Важным понятием, которое встречается в факторном анализе, является *транспонирование матрицы* – процесс преобразования матрицы, в ходе которого строки исходной матрицы A становятся столбцами. В результате подобного преобразования возникает *транспонированная* матрица A'. Приведем пример транспонирования матрицы:

$$\begin{Vmatrix} 5 & 8 \\ 9 & 1 \\ 7 & 0 \\ 2 & 1 \end{Vmatrix} \qquad \begin{Vmatrix} 5 & 9 & 7 & 2 \\ 8 & 1 & 0 & 1 \end{Vmatrix}$$

A A'

Рис.20. Транспонирование матрицы A.

Если матрица A квадратная и совпадает с транспонированной к ней матрице, то матрица *A симметрична*.

В факторном анализе часто встречаются такие виды матриц, как диагональная, скалярная и диагональная. *Диагональной* называется такая квадратная матрица, в которой отличны от нуля только элементы, лежащие на главной диагонали (линии которая соединяет верхний левый и нижний правый углы матрицы). *Скалярная матрица* – это такая диагональная матрица, все элементы главной диагонали которой равны между собой. Частным случаем скалярной матрицы является *единичная матрица*, элементы главной диагонали у которой равны 1. Единичная матрица является аналогом единицы в арифметике и обозначается символом *I*. Примеры диагональной, скалярной и единичной матриц приведены на рис.21:

$$\begin{Vmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{Vmatrix} \qquad \begin{Vmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{Vmatrix} \qquad \begin{Vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{Vmatrix}$$

а б в

Рис.21. Примеры диагональной (а), скалярной (б) и единичной (в) матриц.

Вследствие широкого применения в факторном анализе, особого освещения также заслуживает вопрос практических правил умножения матриц.

Умножение матриц имеет ряд отличий от обычного умножения в арифметике. В первую очередь это выражается в том, том, что при умножении матриц не действует *закон коммутативности*, гласящий, что произведение не зависит от порядка, в котором стоят множители, т.е. $AB \neq BA$.

Для умножения матрицы A на матрицу B необходимо выполнение следующего условия: *матрица A должна иметь столько столбцов, сколько строк в матрице B.* Сам процесс умножения матриц исходит из правила «строка на столбец». Это правило означает, что каждый элемент матрицы-произведения представляет собой сумму произведений соответствующих элементов *строки* первой матрицы на соответствующие элементы *столбца* второй матрицы (рис.22).

$$\begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{23} \end{Vmatrix} \times \begin{Vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{Vmatrix} =$$

$$= \begin{Vmatrix} (a_{11}b_{11} + a_{12}b_{21}) & (a_{11}b_{12} + a_{12}b_{22}) & (a_{11}b_{13} + a_{12}b_{23}) & (a_{11}b_{14} + a_{12}b_{24}) \\ (a_{21}b_{11} + a_{22}b_{21}) & (a_{21}b_{12} + a_{22}b_{22}) & (a_{21}b_{13} + a_{22}b_{23}) & (a_{21}b_{14} + a_{22}b_{24}) \\ (a_{31}b_{11} + a_{32}b_{21}) & (a_{31}b_{12} + a_{32}b_{22}) & (a_{31}b_{13} + a_{32}b_{23}) & (a_{31}b_{14} + a_{32}b_{24}) \end{Vmatrix}$$

Рис.22. Умножение матриц.

Здесь следует также отметить одну особенность единичной матрицы: если квадратную матрицу *A* умножить на единичную матрицу *I* того же порядка, что и матрица *A*, то, независимо от порядка, в котором стоят сомножители, получим матрицу *A,* т.е.: *AI=IA=A*

В рамках рассмотрения процесса умножения матриц особого внимания заслуживает понятие *обратной матрицы.* Так, *если матрица A квадратная и невырожденная, то существует такая обратная матрица* A^{-1}*, что выполняется следующие условие:* $AA^{-1} = I$.

Обратная матрица в определенном смысле аналогична обратному числу в арифметике:

$$x \cdot \frac{1}{x} = 1 .$$

Б) Дисперсия.

В предыдущих разделах книги уже рассматривалось понятие «дисперсия», поэтому остановимся на основных особенностях рассмотрения дисперсии в факторном анализе.

Так необходимо различать отдельные компоненты дисперсии – полная дисперсия σ^2 переменной *x* может быть разбита на три основные компоненты:

- *общая дисперсия* – часть полной дисперсии, которая коррелирует с другими переменными или является общей для ряда переменных;
- *специфичная дисперсия* – часть полной дисперсии, которая присуща только этой переменной (т.е. обусловлена некими ее индивидуальными характеристиками);
- *дисперсия, обусловленная ошибкой* – часть полной дисперсии, которая является случайной, вызванной ошибками в процессе выборки, неточностью инструментов наблюдения.

Общая и специфичная дисперсия в сумме образуют, так называемую, *надежную дисперсию.*

Согласно терминологии теории дисперсии, одну из основных задач факторного анализа можно определить как исследование полной дисперсии для определения числа и видов «общих дисперсий», которые обусловливают корреляции в данной совокупности переменных (считается, что т.н. общая дисперсия каждой переменной состоит из нескольких не коррелированных частей соответствующих каждому фактору).

Полную дисперсию переменной *x* можно представить как сумму ее компонент в виде формулы:

$$\sigma_x^2 = \sigma_{x1}^2 + \sigma_{x2}^2 + ...\sigma_{xn}^2 + \sigma_{xs}^2 + \sigma_{xb}^2 , \quad (53)$$

где $\sigma_{x1}^2 ... \sigma_{xn}^2$ - элементы «общих дисперсий», σ_{xs}^2. – специфичная дисперсия и σ_{xb}^2 - дисперсия, обусловленная ошибкой, *n* – число компонент, составляющих общую дисперсию. Если обе стороны этого уравнения разделить на σ_x^2, то получится:

$$1 = \omega_{x1}^2 + \omega_{x2}^2 + ... + \omega_{xn}^2 + s_x^2 + b_x^2 , \quad (54)$$

где $\sum_k \omega_{xk}^2 = \sum_k \frac{\sigma_{xk}^2}{\sigma_x^2}$ - общая дисперсия; $s_x^2 = \frac{\sigma_{xs}^2}{\sigma_x^2}$ - специфичная дисперсия; $b_x^2 = \frac{\sigma_{xb}^2}{\sigma_x^2}$ - дисперсия, обусловленная ошибкой.

Левая сторона уравнения становится равной 1. Это означает, что полная дисперсия теперь равна 1, а все ее составляющие оказались выраженными, как доли полной дисперсии. Другими словами, (54) исходит из нормализованных соотношений, выраженных в единицах стандартного отклонения.

Общая дисперсия, составляющая ядро факторного анализа, может состоять из *n* компонент (от ω_{x1}^2 до ω_{xn}^2), каждый элемент которой представляет собой ту ее часть, которую можно приписать *n* общих факторов. В случае некоррелированных факторов нагрузка фактора *m* равна квадратному корню из того элемента общей дисперсии, который можно приписать влиянию фактора *m*.

Факторная нагрузка – это своего рода «мера наполнения» данного показателя – это своего рода «мера наполнения» данного показателя определенным фактором. Факторная нагрузка имеет форму коэффициента корреляции между данным показателем и некоторым фактором. Чем выше эта корреляция, тем в большей степени тест «наполнен» данным фактором и тем в большей степени является определяющей мерой для данного фактора.

Так, возвращаясь к описанию интерпретации термина «дисперсия» в рамках факторного анализа, перейдем к изложению двух основных понятий, связанных с делением дисперсии на составляющие, а именно понятия *общность и характерность*.

- Общность данной переменной – это та часть ее дисперсии, которая обуславливается общими для нескольких переменных факторами. В рамках факторного анализа общность является основным объектом рассмотрения, так как объясняет влияние некоторый общих факторов. Обозначается h_x^2

$$h_x^2 = \omega_{x1}^2 + \omega_{x2}^2 + ... + \omega_{xn}^2$$

- Характерность данной переменной, соответственно, та часть общей дисперсии, которая связана лишь с данной переменной и присуща только ей. В состав характерности входят специфичная дисперсия и дисперсия, обусловленная ошибкой. Таким образом, та часть дисперсии, которая входит в состав характерности, может быть объяснена лишь влиянием *специфичных* факторов, присущих *только* данной переменной. Обозначается u_x^2

$$u_x^2 = s_x^2 + b_x^2$$

Соответственно, опираясь на уравнение (54), мы можем записать:

$$1 = h_x^2 + u_x^2$$

В) Корреляция.

Как уже указывалось в начале данного параграфа, в факторном анализе корреляция является критерием сходства или различия между отдельными показателями. Читатель уже должен иметь представление о том, что такое корреляция на основе параграфа 3, поэтому мы не будем заострять внимание на основах корреляционного анализа.

Первичным «материалом» для факторного анализа является *корреляционная матрица* (рис.23), элементы которой являются коэффициентами корреляции между всеми переменными данной совокупности. Эта матрица является симметрической, так как коэффициенты по одну сторону от главной диагонали уже характеризуют все имеющиеся взаимозависимости между показателями. На главной диагонали этой матрицы находятся единицы, поскольку корреляция каждой переменной с самой собой равна +1.

Матрица корреляции, у которой на главной диагонали стоят +1, называется полной корреляционной матрицей (рис.23).

$$\begin{Vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{Vmatrix} = R_1$$

Рис.23. Общий вид полной корреляционной матрицы.

Однако, необходимо отметить, что, помещая на главную единицы, мы учитываем полную дисперсию каждой переменной, представленной в матрице, т.е. принимается во внимание влияние не только общих, но и специфичных (характерных) факторов.

Если же на главной диагонали корреляционной матрицы находятся элементы h_x^2, соответствующие общностям и относящиеся лишь к общей части дисперсии, то учитывается влияние лишь общих факторов, т.е. отбрасывается специфичность и дисперсия ошибок.

Матрицы корреляции, в которой элементы главной диагонали соответствуют общностям, называется *редуцированной* и обозначается *R* (рис.24)

$$\begin{Vmatrix} h_x^2 & r_{12} & r_{13} \\ r_{21} & h_x^2 & r_{23} \\ r_{31} & r_{32} & h_x^2 \end{Vmatrix} = R$$

Рис.24. Редуцированная корреляционная матрица.

Как уже говорилось основной целью *расчетной* части факторного анализа, является определение факторных нагрузок, о которых мы уже говорили выше. Совокупность факторных нагрузок записывается в виде матрицы, столбцы которой состоят из нагрузок данного фактора применительно ко всем переменным данной совокупности, а строки – из факторных нагрузок данной переменной. Такая матрица называется *факторной матрицей.* Факторная матрица также бывает полной и редуцированной. Так элементы полной факторной матрицы (которая обозначается F_1 соответствуют полной единичной дисперсии каждой переменной из данной совокупности. Если нагрузки на общие факторы обозначить через *c,* а нагрузки специфичных факторов – через *g*, то полную факторную матрицу можно представить в следующем виде (рис.25):

		Общие факторы		Специфичные факторы			
		I	II	1	2	3	4
Переменные	1	c_{11}	c_{12}	g_1			
	2	c_{21}	c_{22}		g_2		
	3	c_{31}	c_{32}			g_3	
	4	c_{41}	c_{42}				g_4

$= F_1$

Рис.25. Полная факторная матрица для 4 переменных с выделенными двумя общими факторами.

Как можно увидеть, показанная полная факторная матрица состоит из двух частей, соответствующих общим и специфичным факторам.

Столбец факторной матрицы характеризует фактор и его влияние на все переменные. Строка характеризует переменную и ее наполненность различными факторами, другими словами, ***факторную структуру переменной***. При анализе только первой части данной матрицы мы имеем дело с матрицей, которая характеризует лишь общую часть дисперсии каждой переменной. Такая матрица называется редуцированной факторной матрицей (*F)*.

Так как в факторном анализе основное внимание уделяется общим факторам, то мы в дальнейшем будем использовать главным образом *редуцированную корреляционную и редуцированную факторную матрицы.*

В основе процесса выделения факторов лежит уравнение для некоррелированных факторов:

$$R = F \cdot F', \qquad (55)$$

которое звучит следующим образом: ***редуцированная корреляционная матрица R равна произведению редуцированной факторной матрицы F на транспонированную F':***

$$\begin{Vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \end{Vmatrix} \times \begin{Vmatrix} c_{11} & c_{21} & c_{31} & c_{41} \\ c_{12} & c_{22} & c_{32} & c_{42} \end{Vmatrix} = \begin{Vmatrix} h_1^2 & r_{12} & r_{13} & r_{14} \\ r_{21} & h_2^2 & r_{23} & r_{24} \\ r_{31} & r_{32} & h_3^2 & r_{34} \\ r_{41} & r_{42} & r_{43} & h_4^2 \end{Vmatrix}$$

F F' R

Следствием данной теоремы является уравнение, имеющее основополагающее значение в практике выделения факторов:

$$r_{ab} = r_{aC_1} r_{bC_1} + r_{aC_2} r_{bC_2} + ... + r_{aC_n} r_{bC_n} . \quad (56)$$

Другими словами, корреляция между переменными *a* и *b* (r_{ab}) в случае *n* некоррелированных факторов *C*, общих для обеих переменных, равна сумме произведений нагрузок каждого из факторов на эти переменные (корреляция r_{aC_1} и r_{bC_1} - нагрузки соответствующих факторов на переменные *a* и *b*.

Приведенное уравнение читатель может легко вывести из матричной интерпретации указанной теоремы.

6.2 Геометрическая интерпретация некоторых элементов теории факторного анализа.

Помимо рассмотренных выше алгебраических выкладок, обязательным элементом теории факторного анализа является геометрическая интерпретация основных зависимостей. Это имеет большое значение по двум причинам [15].

1. на определенных этапах процедуры факторного анализа графическая интерпретация является важным элементом, без которого невозможно довести процедуру до конца.
2. графическое изображение некоторых зависимостей позволяет лучше их понять «не математикам», в число которых входят и географы.

Корреляция. Каждую из двух переменных, для которых определяется коэффициент корреляции, можно представить как вектор с обозначенными концами. Если предположить, что оба вектора равны 1, (т.е. полной дисперсии), формула определения коэффициента корреляции будет следующей:

$$r_{12} = \cos\alpha_{12}, \quad (57)$$

Где r_{12} - коэффициент корреляции между переменными 1 и 2; α_{12} - угол между векторами 1 и 2.

Пример. Два единичных вектора наклонены друг к другу под углом 60^0. Соответственно, косинус этого угла составит 0.5, а это значит, что, согласно формуле (57) r=0.5. Таким образом, графически коэффициент корреляции между переменными a и b, равный .5 можно графически показать в виде двух векторов, наклоненных друг к другу под углом 60^0 (рис.26).

Одновременно с этим мы можем увидеть, что отрезок Oa – не что иное, как прямоугольная проекция вектора a на вектор b. Таким образом, ***скалярное произведение (в нашем случае – коэффициент корреляции) двух единичных векторов равно проекции одного из них на другой.***

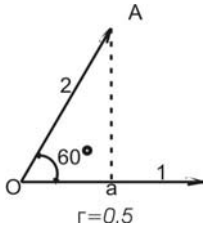


Рис.26. Графическая интерпретация коэффициента корреляции в случае двух единичных векторов.

Исходя из описанной интерпретации корреляции, можно сказать, что корреляционная матрица в геометрическом отношении представляет собой систему векторов, у которой их

длина соответствует элементам главной диагонали, а углы – остальным элементам корреляционной матрицы, называемой *конфигурацией векторов*. Рассмотрим подробнее вопрос длины векторов. Для этого обратимся к формуле (57). Можно увидеть, что данная формула представляет собой частный случай для единичных отрезков формулы следующего вида:

$$r_{12} = h_1 \cdot h_2 \cdot \cos\alpha_{12}, \qquad (58)$$

где $h_1 u h_2$ - длины векторов переменных 1 и 2.

В соответствии с формулой (58), в случае полной корреляционной матрицы все длины будут равны между собой и равны 1. В случае, когда на главную диагональ помещаются значения общностей, по формуле (58) мы получаем, что длины соответствующих векторов будут равны общностям.

Следующим этапом, заслуживающим внимания, является геометрическая интерпретация редуцированной факторной матрицы.

В качестве примера воспользуемся примером из [15] – возьмем редуцированную факторную матрицу *F* с двумя факторами и четырьмя переменными (табл.27).

Табл.27.

	I	II
1	0,7	0,3
2	0,9	0,0
3	0,4	0,6
4	0,6	0,3

Мы уже знаем, что все элементы этой матрицы являются факторными нагрузками или корреляциями переменных и факторов. Например, коэффициент 0.70 в первом столбце и первой строке представляет собой нагрузку фактора 1 у переменной 1. Для графической иллюстрации факторных нагрузок, необходимо, во-первых, изобразить векторы, соответствующие переменным, а во-вторых, векторы, соответствующие факторам (рис.27).

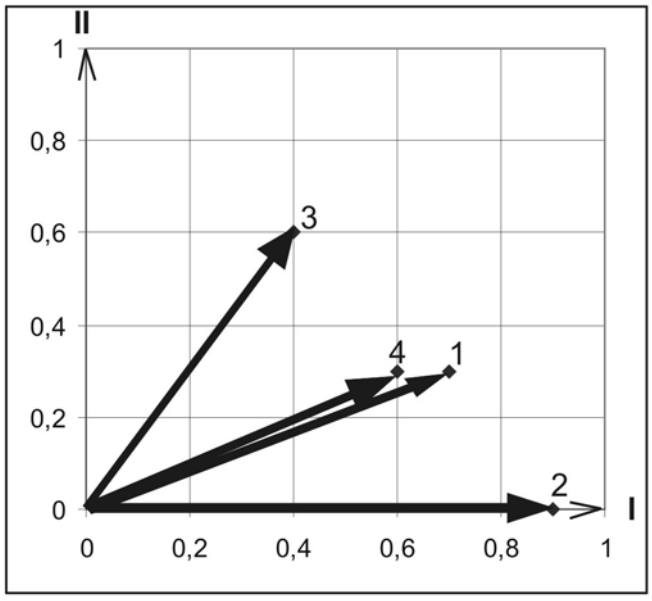


Рис.27. Факторная структура.

Векторы, соответствующие факторам, играют особую роль – так как по определению они не коррелированны, то по формуле (57) они представляют собой взаимно перпендикулярные векторы, образуя систему отсчета, применительно к которой, используя соответственные факторные нагрузки, определяется положение векторов, соответствующих переменным.

С графической интерпретацией редуцированной факторной матрицы связан термин

факторная структура – объединение конфигурации векторов, соответствующих переменным и векторов факторов, образующих систему координат.

Таким образом, количество факторов определяет размерность пространства, в котором размещаются векторы переменных. Нередко количество факторов, а, следовательно, и количество осей координат становится большим 3. В этом случае пространственное изображение носит чисто символический характер – оси показываются по 2 или 3, на них изображаются проекции векторов переменных на эти оси и не учитывается влияние других факторов.

Не вдаваясь в подробные объяснения (читатель может ознакомиться с подробными обоснованиями в монографиях, посвященных конкретно факторному анализу), хотелось заметить, что постоянной в факторном анализе есть только корреляционная структура. Положений же векторов, соответствующих факторам, может быть бесчисленное множество. Этот вывод дал возможность проведения такой процедуры, как вращение системы координат, которая будет рассмотрена ниже.

6.3 Центроидный метод факторного анализа.

Мы уже знаем, что «отправной точкой» расчетов факторного анализа является корреляционная матрица, а его целью – факторная матрица. С геометрической точки зрения это означает наложение на конфигурацию векторов соответствующей системы координат. Существует довольно много способов конкретного выполнения этой операции (метод главный осей, метод главных компонент, центроидный метод и т.д.). Однако рассмотрение всех методов выходит за рамки данного пособия, поэтому мы остановимся на рассмотрении лишь наиболее общего – *центроидного метода выделения факторов.*

Рассмотрим основные положения центроидного метода. Итак, согласно основному уравнению факторного анализа (56) становится возможным определение корреляционной матрицы на основе факторной. В практике факторного анализа, как правило, решается обратная задача: на основе известных корреляций рассчитываются факторные нагрузки. Найдем способ решения данной задачи. Пусть существует лишь один фактор C_1 и известные коэффициенты корреляции переменной *a* с шестью другими переменными *b, c, d, e, f, g*. Согласно уравнению (56) каждый из коэффициентов корреляции может быть записан в следующем виде:

$$\begin{aligned} r_{ab} &= r_{aC_1} \times r_{bC_1} \\ r_{ac} &= r_{aC_1} \times r_{cC_1} \\ r_{ad} &= r_{aC_1} \times r_{dC_1} \\ r_{ae} &= r_{aC_1} \times r_{eC_1} \\ r_{af} &= r_{aC_1} \times r_{fC_1} \\ r_{ag} &= r_{aC_1} \times r_{gC_1} \end{aligned} \qquad (59)$$

Правая сторона всей уравнений содержит корреляцию r_{aC_1}. Это значит, что при суммировании относительно большого числа таких произведений в столбце разница между коэффициентами корреляции переменных *b, c, d…n* с фактором C_1 становится незначительной по сравнению с суммой всех произведений. Получается число, являющееся в нашем случае суммой шести коэффициентов r_{aC_1}. С учетом этого можно сформулировать теорему:

Средняя корреляция переменной со всеми другими переменными, рассчитанная из суммы всех корреляций в столбце, пропорциональна корреляции этой переменной с общим фактором. На практике такая корреляция рассчитывается путем суммирования элементов столбца и деления полученной суммы на корень квадратный из суммы всех столбцов матрицы.

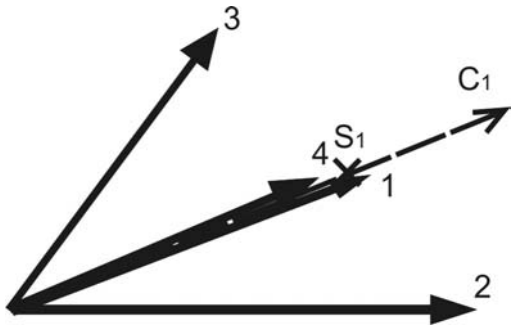


Рис.28. Центроид системы векторов. Использована корреляционная конфигурация векторов из примера графической интерпретации факторных нагрузок.

При попытке геометрической интерпретации данной задачи (рис.28.) мы видим, что ось первого фактора проходит через центр пучка векторов, через эго центральный пункт ***(центроид)***, т.е. через его центр тяжести. В данном контексте центр тяжести понимается следующим образом: если представить концы векторов в виде металлических шариков, а сами векторы в виде невесомых прутиков, то ось пройдет через центр тяжести системы шариков. На рис.28 этот центр обозначен S_1.

Анализ начинается с поиска первого фактора.

Условимся, мы уже имеем полную корреляционную матрицу R_1. однако, как уже было сказано, в факторном анализе наиболее часто оперируют с редуцированной корреляционной матрицей R. Поэтому первым этапом нахождения факторных нагрузок для первого фактора является преобразование полной корреляционной матрицы в редуцированную, т.е. на главной диагонали корреляционной матрицы необходимо записать значения общности h_x^2. Однако, определение значения данного показателя является проблематичным вследствие невозможности определения его экспериментальным путем. Поэтому величины h_x^2 определяются в некоторой степени произвольно. Наиболее простым способом является записывание на главной диагонали модуля наибольшего коэффициента корреляции в столбце. Последовательность дальнейших действий по определению нагрузок первого фактора такова:

1. Суммируются элементы каждого столбца, включая значение общности. Сумма записывается под столбцами в строке Σr, для контроля в последнем столбце матрицы записываются суммы коэффициентов корреляции по строкам.
2. Складываются все суммы столбцов; получающаяся величина, обозначается буквой T;
3. Суммы столбцов делятся на $\sqrt{T}$, результате чего определяются нагрузки первого фактора:

$$C_{1a} = \frac{\sum r_a}{\sqrt{T}};$$

4. Рассчитанные таким способом величины C_1 записываются в последней строке таблицы;
5. Определяется дополнительно величина $\frac{1}{\sqrt{T}}$. Она служит критерием правильности расчетов – произведение $T \cdot \frac{1}{\sqrt{T}} = \sqrt{T}$.

В результате мы получаем редуцированную факторную матрицу , состоящую из n строк и одного столбца пока мы имеем лишь один фактор.

Пример расчета нагрузок первого фактора. Возьмем ***гипотетическую*** *корреляционную матрицу, для 6 переменных (из [15]) (табл.28):*

Табл.28.

	X_1	X_2	X_3	X_4	X_5	X_6	$\sum r$
X_1	**0,4**	0,299	0,4	0,297	0,116	0,232	1,744
X_2	0,299	**0,568**	0,568	0,534	0,432	0,154	2,555
X_3	0,4	0,568	**0,568**	0,487	0,436	0,071	2,53
X_4	0,297	0,534	0,487	**0,545**	0,545	0,092	2,5
X_5	0,116	0,432	0,436	0,545	**0,545**	0,016	2,058
X_6	0,232	0,154	0,071	0,092	0,016	**0,232**	0,765
$\sum r$	1,744	2,555	2,53	2,5	2,058	0,765	
C_1	0,5	0,733	0,726	0,717	0,59	0,219	

$T =$	12,152
$\sqrt{T} =$	3,48598
$1/\sqrt{T} =$	0,28686
$T \cdot (1/\sqrt{T}) =$	3,48592

Следующим этапом является определение факторных нагрузок других факторов. Выражаясь языком геометрии, мы пока определили лишь одну ось и теперь надо найти остальные. Процедура определения нагрузок второго фактора состоит из следующих этапов:

1. Так как лишь некоторая часть общей дисперсии относится к первому фактору, необходимо рассчитать новые коэффициенты корреляции, отражающие ту часть общей дисперсии, которая может быть отнесена на счет других факторов – расчет «остатков». Расчет проводится, исходя из формулы определения коэффициента корреляции на основе факторных нагрузок:

$r_{ab} = r_{aC_1} \times r_{bC_1}$.

   Затем из «исходного» коэффициента корреляции отнимаем полученное по этой формуле значение.
2. Составляем матрицу первых «остатков». Критерием правильности полученных значений является то, что сумма элементов каждого столбца не должна превышать 0,10.

Проблема 1: положительные и отрицательные «остатки» уравновешиваются, вследствие чего мы не можем приступить непосредственно к расчету нагрузок второго фактора, потому что сумма всех столбцов(табл.29) практически равны нулю (эта особенность одновременно является критерием правильности расчетов – суммы по столбцам в матрице остатков не должны превышать 0,1). Это объясняется тем, что мы провели первую центроидную ось через центр тяжести группы точек, являющихся концами векторов. Вычитая часть дисперсии, которая обусловлена 1 фактором, мы «устраняем» ось фактора 1 из графика, т.е. начало координат перемещается вдоль оси 1 фактора в центр тяжести.

Табл.29.

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	**0,15**	**-0,067**	**0,037**	**-0,061**	**-0,179**	**0,123**
X_2	**-0,067**	**0,031**	**0,036**	**0,009**	**0**	**-0,006**
X_3	**0,037**	**0,036**	**0,041**	**-0,033**	**0,008**	**-0,088**
X_4	**-0,061**	**0,009**	**-0,033**	**0,031**	**0,122**	**-0,065**
X_5	**-0,179**	**0**	**0,008**	**0,122**	**0,197**	**-0,145**
X_6	**0,123**	**-0,006**	**-0,088**	**-0,065**	**-0,145**	**0,185**
	0,003	**0,003**	**0,001**	**0,003**	**0,003**	**0,004**

Матрица первых остатков (т.е. после выделения первого фактора).

Для решения данной проблемы необходимо провести процедуру обращения алгебраических знаков в матрице остатков, после чего проводится определение факторных нагрузок по второму фактору согласно методике определения нагрузок первого фактора.

Вследствие того, что на сегодняшний день процедура расчета факторных нагрузок полностью автоматизирована (например, факторный анализ можно проводить при помощи программы Statistica) в рамках данного пособия мы не станем излагать процедуру обращения

знаков. Читатель может подробно ознакомиться в алгоритмом действий, например, в книге [15].

Проблема 2: На каком этапе следует прекратить дальнейшее выделение факторов? Существует довольно большое количество критериев, дающих ответ на данный вопрос. Самый простой из критериев, дающий удовлетворительные результаты при небольшом числе наблюдений, требует, чтобы произведение двух наибольших факторных нагрузок по данному фактору было больше величины $\frac{1}{\sqrt{N}}$, где N – есть число единиц наблюдения. Если это требование не соблюдается, дальнейшее извлечение факторов следует прекратить [7].

Другим, более сложным в расчете критерием является критерий Саундерса. Расчет этого критерия состоит из следующих этапов:

1. возводим в квадрат и складываем остатки, полученные после выделения *к*-го фактора, опуская элементы главной диагонали. Полученную величину умножаем на $\frac{2n}{(n-1)}$ (*n*– число переменных). Обозначим вычисленную величину *A*.
2. Делим разницу между числом переменных и числом уже выделенных факторов на число переменных и результат возводим в квадрат. Обозначим эту величину *B*.
3. Возводим в квадрат все факторные нагрузки, включая нагрузки к-го фактора и суммируем их (число факторных нагрузок равно $k \times n$). Результат вычитаем из *n*, и полученную величину снова возводим в квадрат. Результат делим на *N* (на число единиц наблюдения в исходной совокупности). Обозначим эту величину С.

Если окажется, что $A < B \times C$, выделение факторов прекращается.

Табл.30. Пример редуцированной факторной матрицы для трех факторов.

	I	II	III
1	**0,5**	**0,365**	**0,145**
2	**0,733**	**-0,118**	**0,075**
3	**0,726**	**-0,095**	**0,312**
4	**0,717**	**-0,22**	**-0,155**
5	**0,59**	**-0,404**	**-0,194**
6	**0,219**	**0,365**	**-0,099**

Итак, пусть мы имеем *n* выделенных факторов (в примере – 3 фактора – табл.30). Как же говорилось, для наилучшей интерпретации результатов нам необходимо графически изобразить полученные нами значения факторных нагрузок. Однако здесь мы стакиваемся с проблемой – мы можем вычертить максимум трехмерную систему координат, а количество ортогональных факторов (т.е. осей координат) может быть гораздо больше. Для выхода из сложившейся ситуации применяют способ изображения одновременно лишь двух осей. Следует заметить, что данный способ применяют даже в случае трех факторов, т.к. трехмерное построение редко дает нам полное (и главное, поддающееся интерпретации) представление о пространственном размещении векторов.

На рис.29 показан пример изображения трехмерно рисунка на трех его двумерных проекциях (значения факторных нагрузок рассчитаны на основе данных табл.30). Следует заметить, что по мере роста числа факторов число необходимых рисунков будет увеличиваться в соответствии с формулой:

$$m = \frac{n \cdot (n-1)}{2}, \qquad (60)$$

где *m* – число рисунков, а *n* – число факторов. Рисунки вычерчиваются таким образом, что факторные нагрузки из каждой строки факторной матрицы принимаются в качестве координат соответствующего вектора, причем всегда одновременно берется два столбца. Удобно представлять векторы точками их концов. При этом имеется в виду, что вектор

имеет длину от начала координат до этой точки.

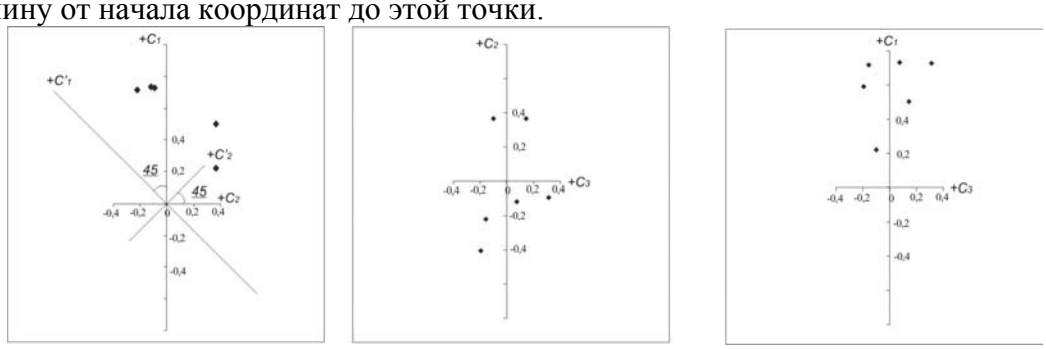


Рис.29. Изображение факторных нагрузок трех факторов при помощи двумерных проекций *(штриховой линией показано положение осей координат C_1 и C_2 после вращения – см. ниже)*.

6.4 Вращение системы координат.

Рассмотрим теперь вопрос вращения системы координат. Пусть в результате расчетов центроидным методом определения факторных нагрузок мы получили совокупность нагрузок шести переменных (или набор их проекций на перпендикулярные оси координат). Как уже отмечалось, данная совокупность проекций не является единственно возможной. Вспомним, почему это так. Как известно, в факторном анализе мы имеем дело с двумя основными элементами, один из которых является постоянным, а второй – переменным. Постоянным элементом является система или *конфигурация векторов*, соответствующая переменным, так как углы между векторами определены матрицей корреляции, которая является «отправным элементом» факторного анализа. Если обратится к графической иллюстрации, то можно утверждать, что единственным постоянным элементом будет система стрел, направленных в разные стороны и выходящих из одной точки. Соответственно, изменяющимся элементом будет система координат, накладываемая на конфигурацию векторов. Этот элемент является неопределенным в том смысле, что его можно вращать вокруг точки, которая представляет собой начало координат. Таким образом, мы получаем теоретически бесконечное число возможных положений осей координат. Каждое из этих положений дает, очевидно, другую совокупность проекций на оси координат, т.е. другое множество факторных нагрузок.

По окончании процесса выделения факторов мы получаем некую определенную совокупность нагрузок, которая соответствует определенному положению систем координат. Это положение зависит от метода, использованного для расчета факторов. Таких методов, как уже было отмечено выше, существует несколько. Так как основная конфигурация векторов представляет собой неизменный элемент, то ее проекции на различно расположенные оси могут взаимно преобразовываться, являясь в этом смысле эквивалентными с той лишь оговоркой, что при этом *не изменяется начало координат, а только вращается система осей вокруг этой точки, представляющую собой в то же самое время начальную точку конфигурации векторов.*

Вращая таким образом систему координат, можно изменить набор факторных нагрузок. Эта операция в процедуре факторного анализа носит название *вращения* или *оборота*, и соответственно первый набор «сырых» факторных нагрузок, полученных по окончании процесса выделения факторов, называется *исходным*.

Определение истинного положения системы координат, дающее наиболее эффективные результаты, представляет собой одну из наиболее трудных и тонких проблем факторного анализа. Правильное осуществление процесса вращения требует определенных знаний как в области самого факторного анализа, так и в той области прикладных географических знаний, где используется факторный анализ.

Существует несколько концепций определения оптимального положения осей факторов. Рассмотрим некоторые из них:

1. «простая структура» – вращение осей определяется: 1) целесообразностью получения максимального числа больших факторных нагрузок; 2) целесообразностью получения наибольшего числа нулевых или относительно близких к нулю нагрузок, которое возможно для данной системы переменной. То есть мы должны добиться, чтобы каждая из осей характеризовалась максимальной нагрузкой одних переменных и минимальной – других, т.е. чтобы векторы переменных находились как можно более близко к осям факторов.
2. при вращении можно руководствоваться стремлением к согласованности результатов с достижениями исследований, выполненных другими методами;
3. если в конфигурации векторов существуют большие «пучки» корреляций, то можно стремиться провести оси координат через их центр;
4. можно стремиться к определению совокупностей факторных нагрузок, согласующихся с какими-либо общими предположениями данной отрасли науки;
5. можно руководствоваться стремлением к согласованности результатов с факторами, определенными в ранее выполненных работах по факторному анализу.

Итак рассмотрим непосредственно организацию процесса вращения системы координат в фактором анализе.

Основными задачами, которые ставятся перед исследователем при вращении, являются следующие:

- определить наилучшее положение осей координат, которое дает наиболее явную простую структуру;
- точно рассчитать новые проекции (факторные нагрузки) векторов на новые оси координат.

Первая задача решается как алгебраически, так и графически, причем графическому способу принадлежит главенствующая роль. Только с помощью графиков можно наглядно представить, как перемещаются отдельные группы точек (концов векторов) относительно вращаемых осей, что облегчает поиск описанной выше *простой структуры*.

Вторую задачу лучше решать при помощи алгебраических и тригонометрических вычислений.

После построения графика и определения направления и угла вращения возникает проблема: как определить новые проекции переменных на оси факторов? Опишем, как осуществляется данный этап расчетов.

Так как оси координат перемещаются на некоторый угол относительно их исходного положения, нужно прежде всего установить, каким образом вообще можно количественно определить направление любой оси в пространстве. Если рассматривать вращаемую ось как вектор, то из аналитической геометрии известно, что его положение может быть определено тремя углами, заключенными между этим вектором и тремя взаимно перпендикулярными осями какой-либо системы координат (описан случай трехмерного пространства). Соответственно, если мы имеем *n* взаимно перпендикулярных осей, то положение вектора будет определяться при помощи *n* углов).

Известно, что любые углы могут быть представлены их косинусами. Косинусы, определяющие положение вектора относительно системы координат, называются *направляющими косинусами*.

Таким образом, *положение новых осей координат определятся при помощи соответствующего количества направляющих косинусов относительно исходной системы координат.*

При вращении нас интересует перемещение осей координат относительно их исходного положения. Поэтому положение перемещаемых осей можно определить при помощи направляющих косинусов, рассчитанных с учетом *исходного* положения осей координат, которое трактуется как система отсчета.

Рассмотрим трехмерный случай. Обозначим исходные оси через C_1, C_2 и C_3, а их новые положения после вращения C_1', C_2', C_3' соответственно. Эти новые положения можно

определить при помощи матрицы направляющих косинусов. В общем виде матрица направляющих косинусов имеет вид (Табл.31):

Табл.31. Матрица направляющих косинусов

	C_1'	C_2'	C_3'
C₁	$X_{C_1C_1'}$	$X_{C_1C_2'}$	$X_{C_1C_3'}$
C₂	$X_{C_2C_1'}$	$X_{C_2C_2'}$	$X_{C_2C_3'}$
C₃	$X_{C_3C_1'}$	$X_{C_3C_2'}$	$X_{C_3C_3'}$

где направляющие косинусы обозначены через *X*.

Элементы такой матрицы представляют собой направляющие косинусы углов, заключенных между *новым положением* осей и системой отсчета, образующей исходные оси.

Важно отметить, что сумма квадратов косинусов каждого столбца матрицы равна 1 при предположении, что длина вращаемой оси равна 1. Для двумерных систем то очевидно из теоремы Пифагора, однако отмеченное свойство присуще также трехмерному и *n*-мерному пространству. Как правило, предполагается, что каждая ось системы координат представляет собой полную дисперсию данного фактора, равную 1, а потому имеет длину, также равную 1.

Таким образом, новое положение осей координат относительно исходного может быть определено при помощи матрицы направляющих косинусов, количество столбцов (или строк в которой определяет размерность пространства.

Матрица, содержащая направляющие косинусы новых осей относительно исходных, а также столбцы и строки которой являются нормализованными (т.е. сумма квадратов содержащихся в них направляющих косинусов будет равна 1 называется матрицей трансформации или λ- матрицей.

Процедура расчета новых проекций заключается в умножении матрицы исходных факторных нагрузок на матрицу трансформации *λ*.

Итак, непосредственно после того, как мы определили направление и угол поворота осей необходимо рассчитать матрицу *λ*.

Однако первоначально мы должны составить матрицу трансформации для *исходного положения*. Строго говоря, последняя не является еще трансформирующей, так как воссоздает *исходное положение* факторов. Это особая форма матрицы *λ*. Она строится с учетом таких соображений (покажем правила ее построения на примере первого столбца): направляющий косинус для фактора C_1 относительно самого себя (т.е. фактора C_1) равен 1, так как ось C_1 сама с собой образует угол 0º. Также, направляющий косинус для C_1 относительно C_2 составит 0, так как угол между ними равен 90º. Аналогично направляющий косинус для C_1 относительно C_3 будет равен 0. Подобным образом составляются и другие столбцы. В результате (в случае трех факторов) мы получаем следующую матрицу (табл.32):

Табл.32. Первая матрица трансформации *λ*.

	C₁	***C₂***	***C₃***
C₁	1,0	0	0
C₂	0	1,0	0
C₃	0	0	1,0

Характерным свойством данной матрицы является то, что если умножить исходную редуцированную факторную матрицу *F* на такую матрицу *λ*, то получим ту же матрицу *F* (данное свойство автор рекомендует проверить, проведя умножение данных матриц и проверив изложенные выводы).

На рис.29 Были изображены три проекции системы шести векторов и было установлено, что нам необходимо произвести вращение проекции C_1C_2 на угол 45º против часовой стрелки. Рассчитаем матрицу λ для данного примера.

Так, расчер новых направляющих косинусов в данном случае осуществляется по формуле:

$$(C_1') = (C_1) - (tg45°) \cdot (C_2), \tag{61}$$

где (C_1) – столбец направляющих косинусов для (C_1'). Аналогично поступают в отношении других осей. Тангенс, на который умножается величина C_2, имеет отрицательный знак, так как мы движемся в направлении **от оси C_2**. Далее процедура расчета новых факторных нагрузок после поворота состоит из следующих этапов:

1. Переписываем по горизонтали столбец косинусов для C_1 в матрицу λ с аналогичным столбцом для C_2 (значения берутся из исходной матрицы трансформации).
2. Столбец C_2 умножаем на $-tg45°$, что составляет –1, и результат прибавляем к C_1.

$$\begin{aligned} C_2 &= \quad 0 \quad 1 \quad 0 \\ C_1 &= \quad 1 \quad 0 \quad 0 \\ (-tg45°) \cdot (C_2) &= \quad 0 \quad -1 \quad 0 \\ C_1 - (tg45°)(C_2) &= \quad 1 \quad -1 \quad 0 \end{aligned}$$

3. Нормализуем значения полученных направляющих косинусов с учетом того, что длины осей равны 1 и поэтому квадраты направляющих косинусов должны в сумме давать 1. Нормализация совокупности чисел заключается в возведении каждого числа в квадрат, расчете суммы этих квадратов и делении каждого числа на корень квадратный из этой суммы. В случае трех чисел *a,b,c*(т.к. в каждой строке или столбце мы имеем лишь три числа, т.к. выделены только три фактора) нормализованные величины составят:

$$\frac{a}{\sqrt{a^2+b^2+c^2}}; \frac{b}{\sqrt{a^2+b^2+c^2}}; \frac{c}{\sqrt{a^2+b^2+c^2}}. \tag{62}$$

таким образом, нормализованные направляющие косинусы для C_1 составят:

$0.71 \quad -0.71 \quad 0$

4. Аналогично рассчитываем направляющие косинусы для оси C_2. Так как ось C_2 движется к оси C_1, то теперь формула расчета направляющих косинусов будет иметь вид:

$(C_2') = (C_2) + (tg45°) \cdot (C_1)$.

5. Сделав подстановку и нормализовав полученные значения, имеем для C_2:

$0{,}71 \quad 0{,}71 \quad 0$

6. Из рассчитанных столбцов составляем новую матрицу трансформации. Ее столбец C_3' остается тем же, что и в исходной матрице трансформации, так как положение этой оси не менялось.
7. Умножая редуцированную факторную матрицу исходных нагрузок на полученную матрицу трансформации, получаем первую матрицу повернутых факторов:

	C_1	C_2	C_3
1	0,5	0,365	0,145
2	0,733	-0,118	0,075
3	0,726	-0,095	0,312
4	0,717	-0,22	-0,155
5	0,59	-0,404	-0,194
6	0,219	0,365	-0,99

F

$\times$

	C_1'	C_2'	C_3'
C_1	0,71	0,71	0
C_2	-0,71	0,71	0
C_3	0	0	1

λ_1

$=$

	C_1'	C_2'	C_3'
1	0,096	0,614	0,145
2	0,604	0,436	0,075
3	0,582	0,448	0,312
4	0,665	0,353	-0,155
5	0,705	0,132	-0,194
6	-0,104	0,414	-0,099

F_1'

Процедура вращения осей является крайне важной, так как практически во всех случаях интерпретация первоначальной факторной матрицы крайне затруднена. Особенно это касается изложенного здесь центроидного метода определения факторных нагрузок, основным преимуществом которого является сравнительная простота расчетов.

Для «уточнения» факторных нагрузок (т.е. положения осей факторов относительно корреляционной структуры), определенных центроидным методом в частности, существует ряд методов (метод максимального правдоподобия и т.д.) Однако изложения этих методик выходит за рамки данного пособия, поэтому читатель может ознакомиться с ними самостоятельно в других литературных источниках (например, [19]).

6.5 Пример использования факторного анализа в прикладных географических исследованиях.

В процессе изложения сущности данного метода мы использовали лишь абстрактные числовые примеры. Это объясняется, в первую очередь, тем, что «реальные» примеры при детальном изложении, как правило, являются довольно громоздкими (например, корреляционные матрицы 35 на 35, многократные вращения координат и т.д.). Однако для большего уяснения сферы применения факторного анализа приведем пример его применения в географических исследованиях. В качестве подобного примера была отобрана работа [16], посвященная применению факторного анализа при оценке природных условий жизни населения Забайкалья.

Для характеристики влияния природных условий на жизнь населения отбирались такие показатели, которые обычно используют при проектировании жилищ, выборе одежды, устройстве мест массового отдыха, в здравоохранении и т.д. Для составления хорошо читаемой оценочной карты количество показателей должно быть ограничено. Эти показатели должны отображать основные географические особенности оцениваемой территории. Например, для территории Забайкалья были отобраны следующие показатели:

I. Климат: 1)солнечная радиация, 2)зимняя эффективная температура, 3)продолжительность отопительного периоды, 4)суровость погоды января, 5)нормально-эффективная температура июля, 6)число дней в году с сильным ветром, 7)число дней без осадков, 8)многолетняя мерзлота грунтов.
II. Рельеф и геофизические условия: 9)абсолютная высота местности,10)крутизна склонов, 11)лавинная опасность, 12)сейсмичность.
III. Воды 13)ресурсы поверхностных вод, 14)химический состав питьевых вод, 15)лечебные минеральные воды.
IV. Почвы: 16)естественное, биологическое самоочищение почв, 17) предпосылки биогеохимических эндемий.
V. Растительность: 18) пригодность лесных ландшафтов для массового отдыха.
VI. Животный мир: 19) предпосылки болезней с природной очаговостью.

Отобранные показатели оценивались в условных единицах, в качестве которых были приняты контуры типов ландшафтов, выделенные на ландшафтной карте. При разработке шкал баллов был применен следующий принцип: в 5 баллов оценивалось наиболее благоприятное для Забайкалья значение каждого показателя; в 1 балл – наименее благоприятное.

Далее при помощи центроидного метода были определены факторные нагрузки, которые были уточнены при помощи методы максимального правдоподобия. Затем, при помощи процедуры вращения была достигнута оптимальная для целей исследования факторная структура (Табл.33).

Табл.33.

№п\п	1	2	3	4	5	6	7	8	9	10
Фактор 1	0,643	0,0831	0,314	0,823	-0,006	-0,439	0,206	0,764	0,004	0,013
Фактор 2	0,352	-0,29	0,895	-0,314	0,768	-0,013	0,795	0,25	0,734	0,591
Фактор 3	-0,139	0,047	0,061	0,106	0,05	-0,349	-0,115	0,169	0,433	0,364

№п\п	11	12	13	14	15	16	17	18	19
Фактор 1	0,131	0,199	0	-0,34	0,178	0,386	-0,337	0,029	-0,486
Фактор 2	0,741	0,752	0,29	-0,524	0,048	0,681	-0,632	-0,127	-0,691
Фактор 3	-0,154	-0,448	0,705	0,138	0,335	0,201	-0,043	0,004	-0,072

Так, фактор 1 имеет наибольшие нагрузки показателей, связанных с климатом; фактор 2 – с рельефом, почвами и растительностью. Фактор 3 имеет наименьшую ценность в синтетической оценке всего комплекса природных условий, так как существенное влияние на него имеют показатели №9, 12 и 13 (интерпретация такого фактора является затруднительной).

Следующим этапом было представление полученных результатов в виде взвешенных сумм (весами служат факторные нагрузки). Для этого матрицу оценочных баллов умножили на транспонированную матрицу окончательных факторных нагрузок (Табл.33). По полученным значениям взвешенных сумм была построена шкала общей синтетической оценки, в соответствии с которой была проведена группировка территориальных единиц. В результате были выделены территории с различной степенью благоприятности природных условий для жизни населения (рис.30).

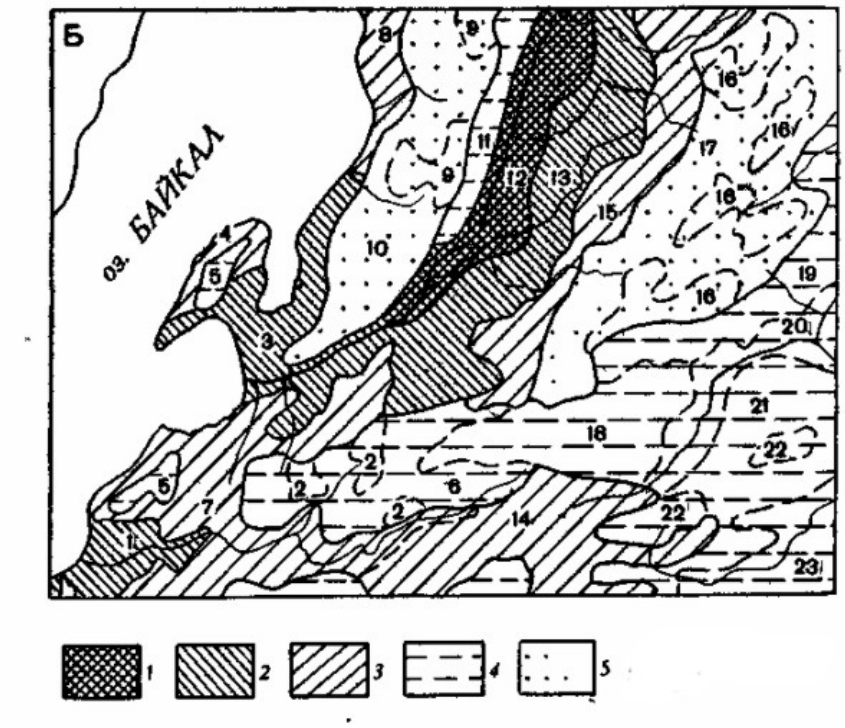


Рис.31. Фрагмент результирующей карты оценки природных условий территории Забайкалья для жизни населения. *Условные обозначения:*

- *территории про степени благоприятности (в баллах): 1 – благоприятные (36,4 – 26,4); 2 – условно благоприятные (26,3 – 18,3); 3 – недостаточно благоприятные (13,2 – 12,0); 4 – мало благоприятные (11,7 – 6,4); неблагоприятные (6,3 – 0).*
- *Цифры на карте – порядковые номера учитываемых территориальных единиц.*

Вопросы:

1. *При решении каких задач применяется факторный анализ?*
2. *Назовите основные виды матриц, использующихся в фактором анализе.*
3. *В чем разница между полной и редуцированной факторной (и корреляционной матрицами)?*
4. *Как интерпретируется дисперсия в факторном анализе? Что такое общность и характерность (специфичность)?*
5. *Что такое факторная структура?*
6. *В чем состоит определение факторных нагрузок при помощи центроидного метода?*
7. *Назовите основные критерии вращения системы координат.*
8. *Что такое направляющие косинусы?*

Задания.

1. *Работа с литературными источниками. В сборниках научных статей, посвященных применению математических методов в географии* ***самостоятельно*** *найти и проработать примеры применения факторного анализа в географии.*
2. *Дана полная корреляционная матрица (Табл.34):*

Табл.34.

	A	B	C	D	E	F	G	H	I	J	K	L
A	1	0,863	1,447	3,347	3,622	3,527	3,974	3,878	4,596	1,656	2,847	4,172
B	0,863	1	1,379	3,168	3,715	3,424	3,713	3,843	4,327	1,822	3,088	3,908
C	1,447	1,379	1	2,866	3,18	3,47	3,527	3,798	3,392	1,377	2,229	3,457
D	3,347	3,168	2,866	1	1,801	1,519	2,898	3,325	3,457	2,535	3,942	3,794
E	3,622	3,715	3,18	1,801	1	2,188	3,212	2,974	4,103	2,288	1,924	3,193
F	3,527	3,424	3,47	1,519	2,188	1	4,079	4,118	4,902	2,905	3,537	3,877
G	3,974	3,713	3,527	2,898	3,212	4,079	1	1,495	3,069	3,189	3,253	4,47
H	3,878	3,843	3,798	3,325	2,974	4,118	1,495	1	4,248	2,98	2,935	4,446
I	4,596	4,327	3,392	3,457	4,103	4,902	3,069	4,248	1	3,951	3,775	3,875
J	1,656	1,822	1,377	2,535	2,288	2,905	3,189	2,98	3,951	1	1,342	2,914
K	2,847	3,088	2,229	3,942	1,924	3,537	3,253	2,935	3,775	1,342	1	2,921
L	4,172	3,908	3,457	3,794	3,193	3,877	4,47	4,446	3,875	2,914	2,921	1

А) На основе построить редуцированную корреляционную матрицу;

Б) Рассчитать факторные нагрузки первого фактора;

В) Построить матрицу остатков корреляции.

3. *На рис.29 изображено направление первого поворота осей координат.*

А) Рассчитать матрицу λ для следующего поворота осей (угол и направление – по вашему выбору), и на их основе.

Б) Построить матрицу факторных нагрузок полученных после проведенного вращения. Приблизились ли Вы к «простой структуре», по сравнению с факторной матрицей, которая имелась до вращения координат?

ПРИМЕРЫ ПРИКЛАДНЫХ ГЕОГРАФИЧЕСКИХ ИССЛЕДОВАНИЙ С ИСПОЛЬЗОВАНИЕМ СТАТИЧЕСКИХ МЕТОДОВ АНАЛИЗА.

1. Анализ ситуации в сфере обращения с твердыми бытовыми отходами (ТБО) по областям Украины.

Система обращения с твердыми бытовыми отходами нашей страны характеризуется значительными объемами образования и накопления последних. Так, только в городах ежегодно образуется 40 млн. м3 ТБО. Отходы поступают в основном из трех источников: жилищного сектора, бюджетных организаций (учебных заведений, детских организаций, больниц, общественных учреждений и др.).

Согласно с [4], 90% ТБО вывозится на 656 свалок, которые расположены в 10-20 км от городов. В среднем на одного жителя Украины приходится 0,8-1,0 кг ТБО на сутки [17].

Для анализа ситуации в сфере ТБО по областям Украины автором были проведены такие работы:

- по данным [10] была построена карта ситуации с сфере обращения с ТБО об областям Украины;
- с помощью кластерного анализа области были сгруппированы в определенные таксономические группы;
- были рассчитаны оценочные показатели, дающие возможность оценить эффективность функционирования системы обращения с ТБО по областям Украины.

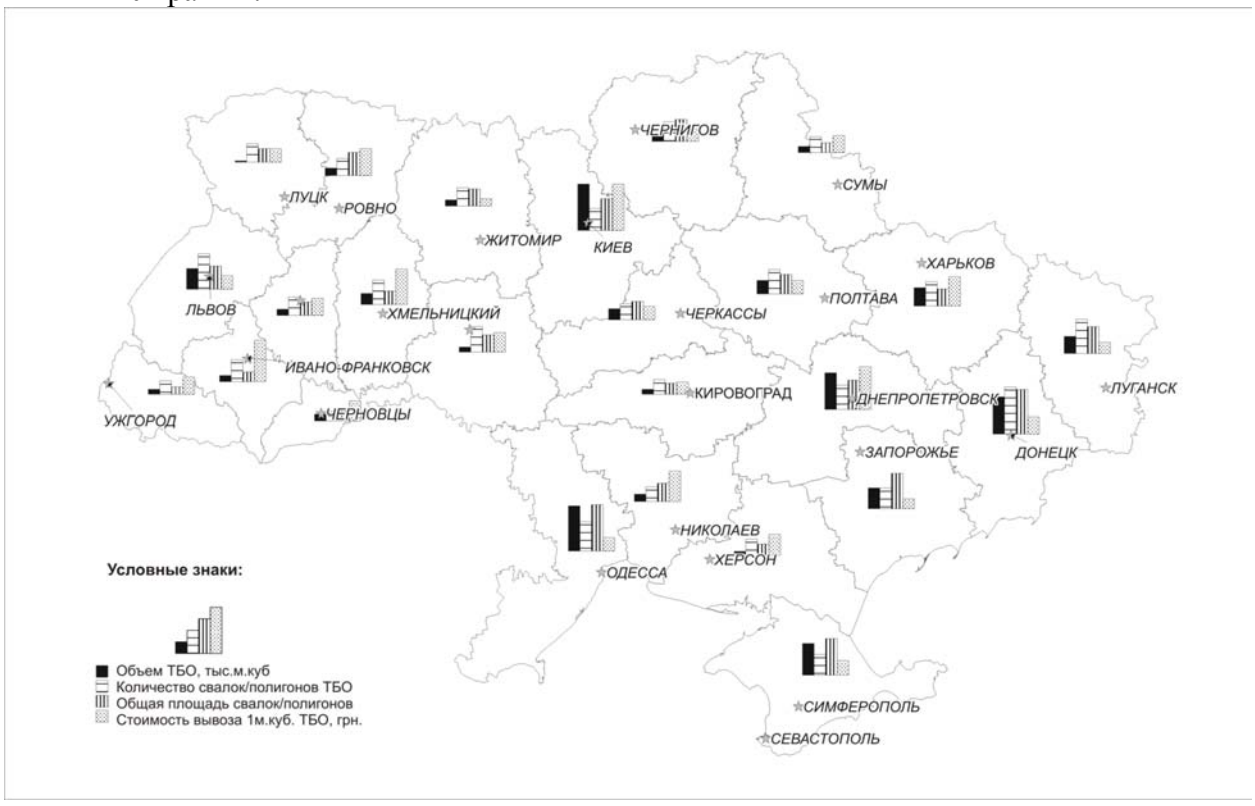


Рис.32. Характеристика ситуации в сфере обращения с ТБО по перечню показателей (см. легенду карты на рисунке).

На рис. 32. представлена карта, иллюстрирующая ситуацию в сфере обращения с ТБО по Украине по таким показателям, как:

1. количество вывезенного мусора, тыс. м3;
2. количество свалок и полигонов ТБО;
3. Общая площадь свалок и полигонов;

4. Средняя стоимость 1 м3 мусора, грн.

На основе данных показателей был проведен кластерный анализ. (см. Тему 5).

По результатам проведенной группировки была построена карта, представленная на рис.33.

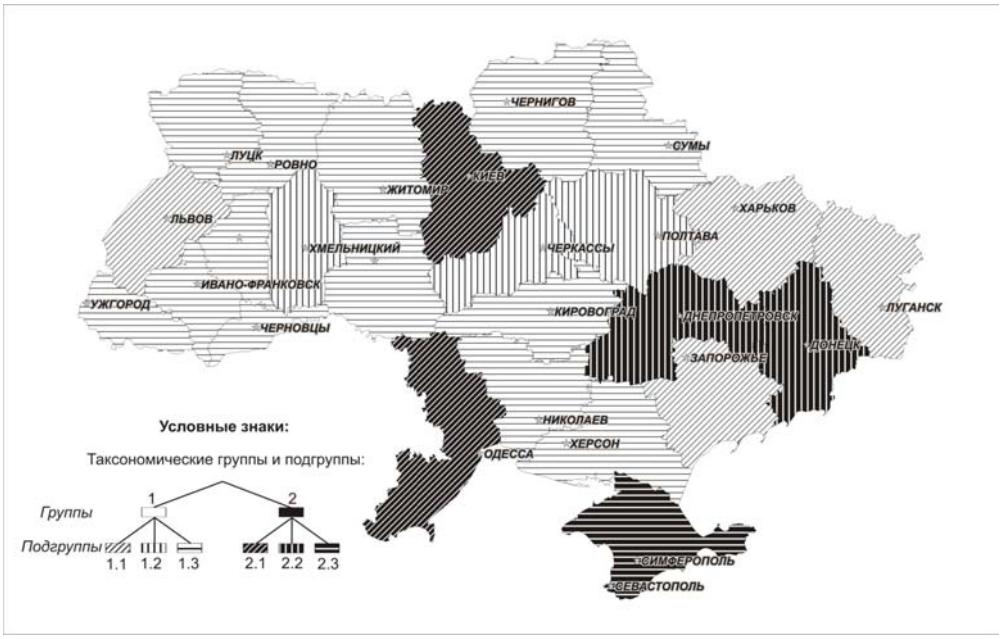


Рис.33. Группировка областей по степени сходства ситуации в сфере обращения с ТБО, реализованная при помощи таксономического анализа (картографическое представление результатов).

Проанализируем полученные результаты. На уровне распознавания 700 мы получили 2 группы, где в первую вошли 20 областей; во вторую – 4 области и Республика Крым. Первая та вторая группы, в свою очередь, разделяются на три подгруппы каждая.

Группа 1. К ее составу входят области, которые имеют удовлетворительное и относительно удовлетворительное состояние ситуации из области ТБО.

Подгруппа 1.1. Включает Харьковскую, Луганскую, Львовскую и Запорожскую области. Характеризуются максимальными для 1 группы значениями показателей, которые исследуются. Все области имеются высокий уровень индустриального развития. Определяющим показателем в данной группе является большое количество свалок.

Подгруппа 1.2. Включает Хмельницкую, Черкасскую и Полтавскую области. Характеризуются средними для 1 группы значениями. На картодиаграммах (рис.32) нельзя выделить показатель, который можно было бы назвать определяющим для данной группы (исключение – Хмельницкая область с очень высоким уровнем себестоимости вывоза единицы массы ТБО).

Подгруппа 1.3. Входят Волынская, Ровенская, Житомирская, Черниговская, Сумская, Закарпатская, Ивано-Франковская, Тернопольская, Черновицкая, Винницкая, Кировоградская, Николаевская та Херсонская области. Характеризуются удовлетворительной ситуацией в области ТБО.

Группа 2. В ее состав входят области, которые характеризуются достаточно напряженной ситуацией в сфере ТБО.

Группа 2.1. Входят Киевская та Одесская области. Определяющим показателем являются наибольшие в Украине объемы образования ТБО.

Группа 2.2. Входят Днепропетровская та Донецкая области – центры горнодобывающей промышленности. Характеризуются очень высокими значениями объемов ТБО и площади свалок ТБО.

Группа 2.3 В ее состав входит Автономная Республика Крым. По значениям величин (согласно рис.32) данная группа имеет сходство с группой 2.2., однако по абсолютным величинам несколько им уступает. Определяющими показателями являются высокие значения объемов ТБО и площади свалок ТБО.

Следующим этапом работы был расчет оценочных показателей, по которым можно бы было провести более детальный анализ функционирования системы обращения с ТБО по областям Украины. Для этого в исследование было введено два показателя:

1. Степень напряженности в сфере поведения ТБО. Для получения данного показателя были выполнены такие этапы работы:
   - Методом простого простого ранжирования [2] каждой области по каждому показателю присваивался соответствующий оценочный балл от 1 до 25, который характеризовал «место» области по данному показателю в общей совокупности данных по областям по всем областям (причем 1 место имела область с наименьшим значением).
   - Была определена сумма баллов по каждой области.
   - Полученные баллы были разбиты на 5 интервалов, и методом картограмм были отображены на карте (рис. 34.).
2. Показатель оптимальности функционирования системы размещения ТБО. Сущность данного показателю заключается в том, что оптимальным является размещение максимальных объемов ТБО на минимальной площади, что дает возможность «локализовать» систему защиты окружающей среды от влияния ТБО и значительно упростить систему обслуживания свалок. Расчет данного показателя заключается в определении отношения объемов образования ТБО к общей площади свалок по области (рис.34.).

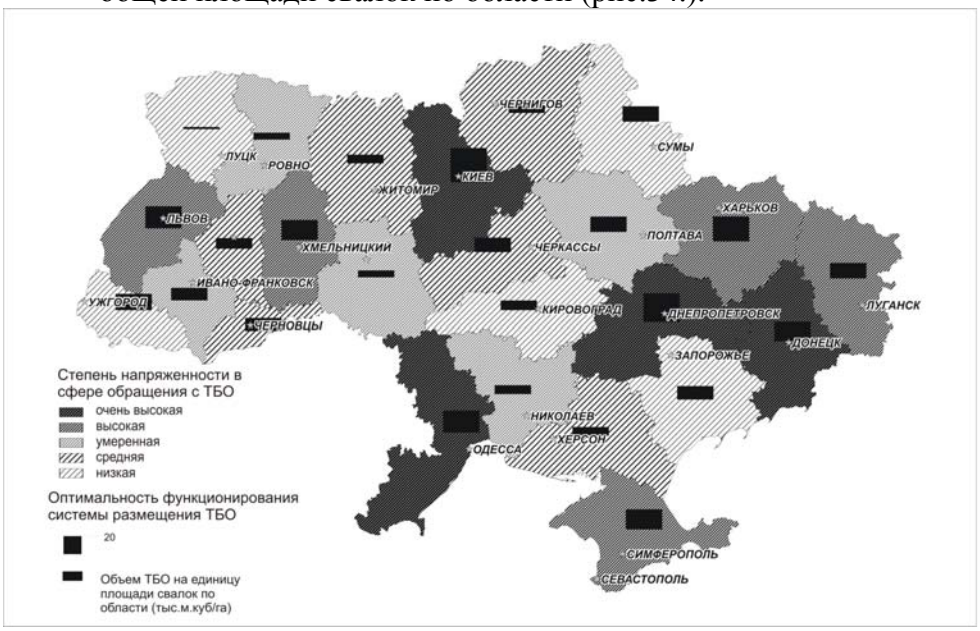


Рис.34. Оценочные показатели, характеризующие ситуацию в сфере ТБО.

За полученными результатами можно сделать определенные выводы. Так, в целом, показатель состояния ситуации в сфере ТБО по областям Украины в целом подтверждает распределение на группы по результатам кластерного анализа, а также подтверждает выводы, сделанные по результатам группировки.

По второму показателю мы видим, что наименее эффективная система обращения ТБО присуща областям Украины, которые характеризуются наименее напряженной ситуацией в сфере ТБО, что делает их мало перспективными с точки зрения развития биогазовой отрасли биоэнергетики.

Обоснуем данное заключение. Так, данные области характеризуются малыми объемами накопленных отходов, которые размещены на значительных площадях. Следовательно, необходимо строительство либо большого количества маломощных станций по переработке биогаза, либо большой по протяженности системы транспортировки биогаза

и дальнейшей его централизованной переработке. Безусловно, рентабельность функционирования подобной системы требует дополнительных экономических расчетов, однако даже на первый взгляд можно увидеть, что ее внедрение в областях с «более концентрированным» размещением бóльших объемов ТБО является долее перспективным.

Напротив, в областях с наибольшими объемами отходов системы управления твердыми бытовыми отходами работают более эффективно (рис.34.).

Таким образом, области с наибольшими значениями накопленных объемов ТБО одновременно характеризуются и наибольшей «компактностью» их размещения, что упрощает и удешевляет организацию системы по сбору и утилизации свалочного газа.

2. Оценка объемов накопленных ТБО на свалках и полигонах Харьковской области при помощи регрессионного анализа.

Харьковская область на данный момент имеет значительный биоэнергетический потенциал, который, в первую очередь, определяется объемом твердых бытовых отходов, накопленных на территории области. Вследствие этого на сегодняшний день достаточно актуальной является задача расчета общего энергетического потенциала твердых бытовых отходов Харьковской области.

Как свидетельствует [8], реальная нагрузка ТБО на территорию Харьковской области оценивается лишь с некоторой степенью приближенности, так как сведения относительно накопления ТБО не точны, а, в большинстве случаев, совсем отсутствуют. Так, даже с небольшой точностью определения объемов отходов по свалкам данные относительно объемов накопленных отходов отсутствуют приблизительно для половины исследованных свалок ТБО. Однако, для всех свалок (кроме запроектированных), известны данные относительно площадей, занимаемых той или иной свалкой, а также данные относительно процента ее заполнения.

Таким образом, необходимо определить ориентировочные объемы отходов для всех свалок и полигонов Харьковской области. Решение данной задачи возможно с помощью изложенного выше регрессионного анализа - используя данные по объемам ТБО и площади, занимаемой ими, выведем уравнение зависимости объема ТБО от площади того или иного полигона (свалки).

Для определения максимальных объемов ТБО для каждой свалки, необходимо рассчитать объем, соответствующий 100% заполнению данной свалки, исходя из указанной доли ее заполнения.

Определим характер зависимости между объемом ТБО и площадью свалки ТБО на свалках. Для этого, прежде всего, необходимо знать правомерность использования методов параметрической статистики, т.е. соответствует ли распределение исследуемых величин нормальному или нет. Построить гистограмму частот выборочных совокупностей исследуемых величин – площади полигонов, га и объемов ТБО, м.куб (рис.35).

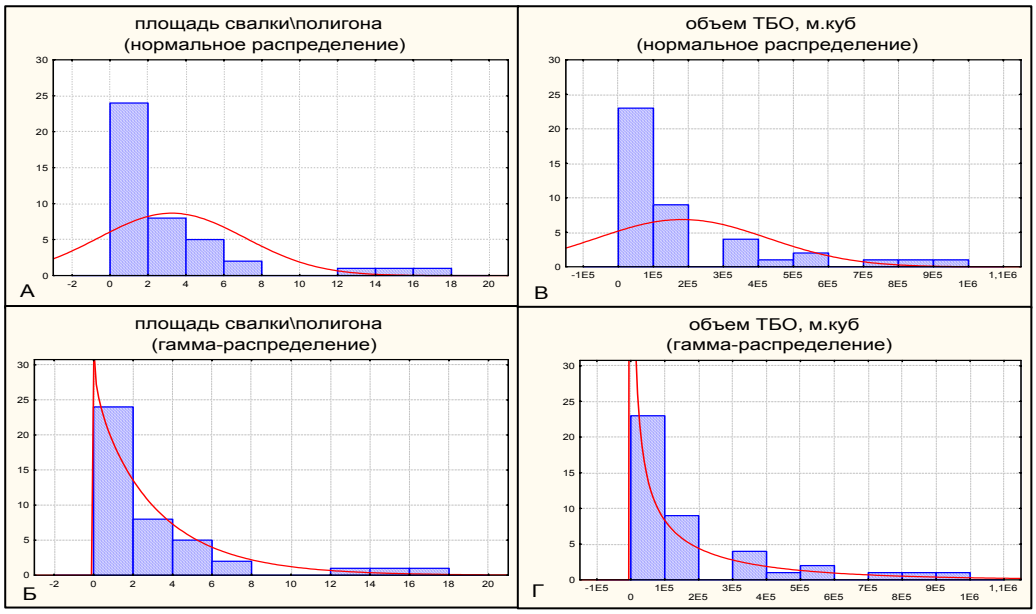


Рис.35. Анализ распределений исследуемых величин. Столбчатой диаграммой показаны гистограммы частот; линией показана кривая распределения.

Даже при визуальном анализе гистограммы частот можно сказать, что исследуемые распределения отличаются от нормального. Подтвердим данное утверждение при помощи

критерия хи-квадрат. Так, на рис.35 А,В представлены возможные кривые нормального распределения, а рис. 35 Б,Д – кривые гамма распределения для каждого из показателей (см. подписи на графиках). Фактические значения хи-квадрат ля каждого из четырех случаев, приведены в табл.35:

Табл.35.

	Площадь свалки, га	Объем ТБО, $м^3$
Нормальное распределение	50,47	1,88
Гамма-распределение	67,33	2,09
Число степеней свободы*	1	1
Табличное значение χ^2 (P=0,95)	3,841	3,841

** так как необходимо, чтобы в каждый интервал входило не менее 5 значений, то в каждой гистограмме было произведено объединение интервалов.*

Из данной таблицы следует, что распределения исследуемых выборочных совокупностей значительно отличаются от нормального. Поэтому для оценки зависимости между площадью и объемом ТБО используем коэффициент ранговой корреляции Спирмена, который равняется 0,69, что, согласно табл. 12, является достоверной величиной. Построим график взаимозависимости площадь/объем (рис.36):

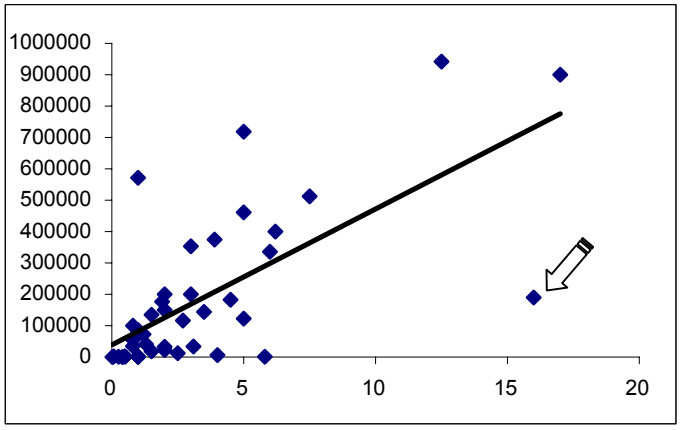


Рис.36. Характер взаимосвязи между площадью (ось абсцисс) и объемом (ось ординат) свалок ТБО Харьковской области. Прямая на графике – линия регрессии. Стрелка указывает на точку, предполагаемую, как артефакт.

Рассмотрим полученную совокупность точек с точки зрения наличия или отсутствия артефактов. Статистический анализ (см. выше) показал, что данная выборочная совокупность не содержит артефактов. Однако, анализируя рис.36 Можно увидеть, что точка, отвечающая свалке пгт. Бабаи (16 га, 190000 $м^3$) значительно «отстоит» от остальной совокупности данных (обозначен стрелкой). Подобное отличие данной точки от остальной совокупности, скорее всего, определяется некоторыми специфическими факторами ее организации, и, безусловно, заслуживает более глубокого анализа на локальном уровне исследования.

Рассмотрим, как изменится взаимосвязь между переменными в случае исключения свалки пгт. Бабаи из данной совокупности. Так, коэффициент корреляции Спирмена практически не изменил своего значения – 0,686, т.е. имеем даже некоторое ослабление зависимости. Однако подойдем к данной проблеме с другой стороны, рассчитав взаимозависимость между величинами при помощи коэффициента корреляции Пирсона. Так, в первом случае данный коэффициент практически соответствует значению коэффициента корреляции по Спирмену (0,687). Однако после исключения артефакта его значение стало равным 0,79, т.е. исключение данной точки из последующих расчетов влечет за собой значительное усиление положительной корреляционной зависимости.

Итак, предположим, что значения исследуемых величин в точке, соответствующей свалке около пгт Бабаи, является артефактом. Однако мы не можем исключить ее без более глубокого анализа данной проблемы, так как, во-первых, согласно статистической проверке,

данная точка по своим характеристикам не является артефактом, а, во-вторых, целесообразность ее исключения обоснована при помощи показателя, применение которого не является полностью адекватным соответствующим распределениям величин.

Предположим, что оптимальным уравнением для функционального описания полученной зависимости, является линейное уравнение регрессии. Рассчитаем его коэффициенты для каждой из двух совокупностей – до и после исключения артефакта.

Полученные уравнения имеют следующий вид:

$$y_1 = 43349x + 38212.9$$

$$y_2 = 59050x + 4105.9 \quad ,$$

где y_1 – уравнение регрессии для изначальной выборочной совокупности, а y_2 – уравнение регрессии после исключения точки-артефакта.

Анализ соответствия теоретической зависимости реальным значениям при помощи критерия Колмогорова, показал, что оба уравнения регрессии соответствуют фактическим данным. Однако величина достоверности аппроксимации R^2 значительно выше у уравнения 2 (0,471 и 0,639 для первого и второго уравнения соответственно).

Автором также был предложен другой вид анализа, суть которого заключается в суммировании модулей всех отклонений от линии регрессии и сравнении полученных сумм. Результаты расчетов показали, что сумма отклонений второго уравнения на 1 млн. $м^3$ меньше, чем у уравнения 1. Другими словами, это означает, что погрешность первого уравнения соизмерима с крупнейшим полигоном ТБО в Харьковской области, находящимся в районе пгт Дергачи. Таким образом, мы обосновали исключение свалки в пгт Бабаи, как артефакта.

Следующим этапом данной работы является определение оптимальной формы зависимости между исследуемыми величинами.

Для совокупности исследуемых данных были построены разных порядков (рис.37 а-е):

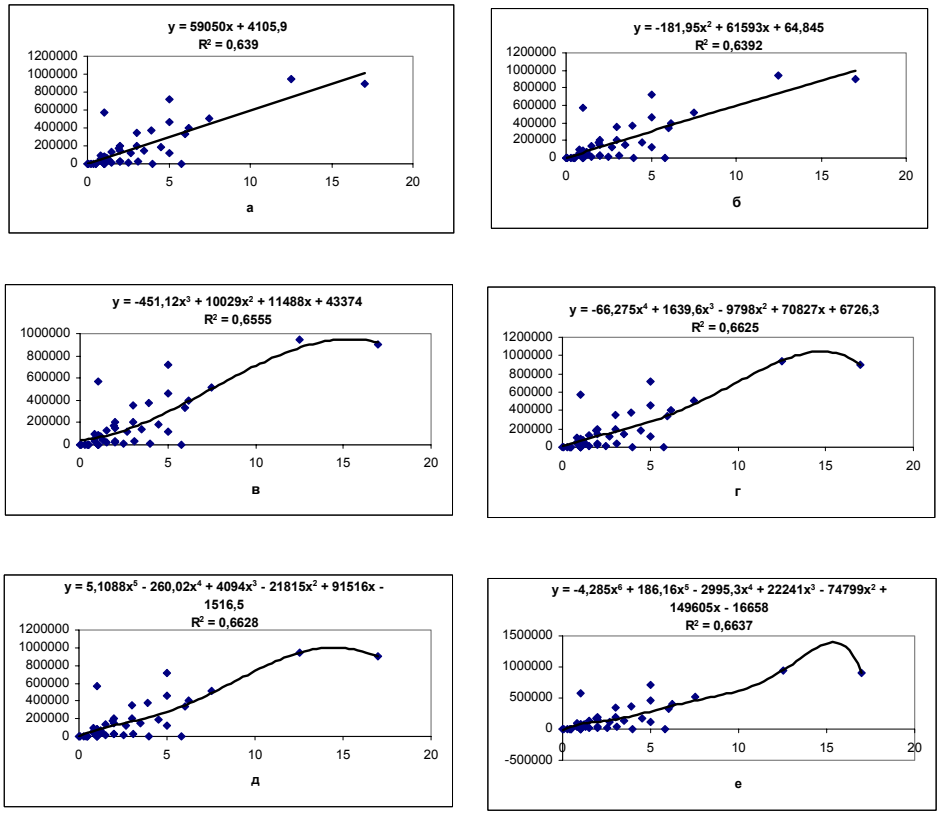


Рис.37. Уравнения регрессии разных порядков для показателей площади/объема свалок. (а-е – аппроксимации 1-6 порядков соответственно). Ось абсцисс – площадь валки, га; ось ординат – объем ТБО, м. куб. В верхней части графиков представлены соответственные уравнения регрессии, а также значения показателя R^2

На графиках рис.37. можно увидеть, что, начиная с уравнения третьего порядка, на линии регрессии начинает появляться точка максимума (экстремум). Такая тенденция, по мнению автора, не отражает реальной ситуации, так как это значило бы, что существует некоторое критическое значение площади свалки, при котором ее объем будет максимальным.

Следовательно, для описания исследуемой зависимости мы должны выбрать одно из уравнений низших порядков. Для выбора оптимального решения автор использовал упоминавшийся ранее показатель R^2. Значения R^2 для всех построенных трендов отображены на рис 38.

Рис.38. Значение R^2 для построенных трендов.

Согласно рис.38. можем сделать вывод, что значения R^2 практически не изменяются при переходе от 1-го ко 2-му порядку уравнения регрессии. Рост показателя R^2 при более высоких порядках уравнения не отражает реальной ситуации (см. выше).

Следовательно, для вычисления объема ТБО на остальных полигонах мы использовали линейное уравнение:

$$y = 59050x + 4105{,}9,$$

где y – объем ТБО, а x – площадь полигона.

Таким образом, мы получили возможность оценки потенциальных объемов производства энергии на основе использования биогаза свалок Харьковской области. В качестве иллюстративного материала на рис.39. Приведены карты объемов накопленных отходов на территории Харьковской области на основе имеющихся статистических данных и на основе данных, рассчитанных с помощью полученного уравнения – благодаря полученному уравнению были определены объемы накопленных объемов отходов более чем на вдвое большем количестве свалок.

Полученные в ходе эксперимента данные могут быть использованы и в других видах деятельности, связанной с обращением с ТБО. Например, по полученным данным мы имеем возможность рассчитать нагрузки ТБО на той или иной территории, что будет характеризовать степень негативного воздействия ТБО на окружающую среду.

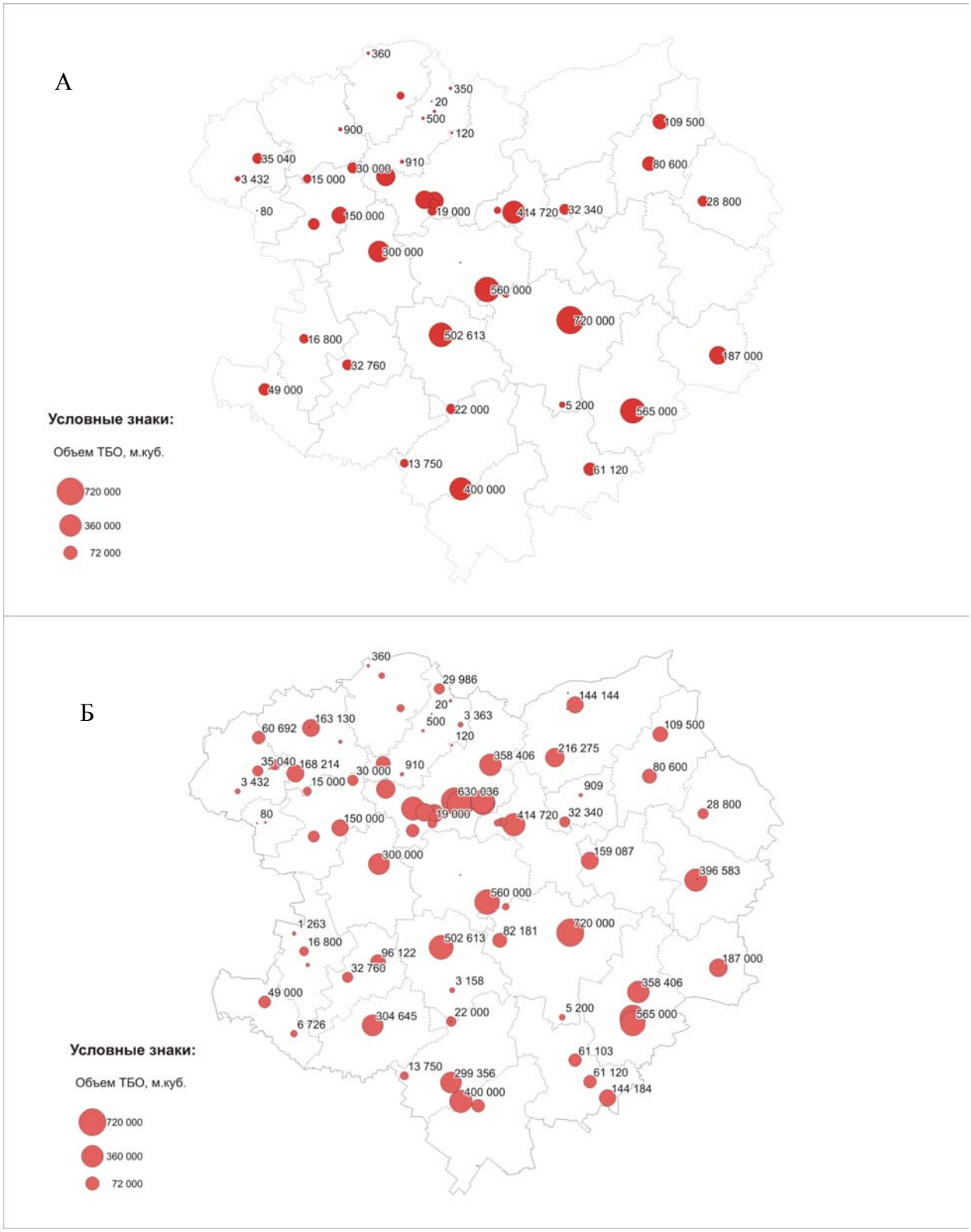


Рис.39. Объемы твердых бытовых отходов, накопленных на свалках Харьковской области (А – по статистическим данным; Б – по статистическим данным , а также полученные расчетным путем).

Литература:

1. Берлянт А.М. Картографический метод исследования. – М.: изд-во Моск. Ун-та, 1978. – 257 с.
2. Бешелев С.Д., Гурвич Ф.Г. Математико-статистические методы экспертных оценок. – 2-е изд., перераб. и доп. – М.: Статистика, 1980. – 263 с.
3. Варивода А.В., Варивода Е.А. Оценка природного и технически доступного ветроэнергетического потенциала Украины относительно выбора участков для планирования ветроэлектростанций// Ученые записки ТНУ. Серия: География. – 2003. – Т.16. - №1. – С. 40-46.
4. Гелетуха Г.Г., Марценюк З.А. Обзор технологий добычи и использования биогаза на свалках и полигонах твердых бытовых отходов и перспективы развития на Украине.// Экотехнологии и ресурсосбережение.-1999.-№4.
5. Географічна енциклопедія України: В 3-х т./ Під. ред. Маринина О.М. та ін. – К.: „Українська радянська енциклопедія” ім. М.П. Бажана, 1989. – Т.1: А-Ж. – 416с.
6. Екологічний атлас Харківської області (електронна версія) Х., УкрНДІЕП, 2001.
7. Жуковская В.М. Опыт применения методов многофакторного анализа для характеристики сельского хозяйства степных провинций Канады // Количественные методы исследования в экономической географии. - М.: Изд-во Моск. ун-та, 1964. – с. 122-166.
8. Звіт по проекту “Розробка обласної програми поводження з твердими побутовими відходами. Обстеження, обґрунтування та розробка першочергових заходів щодо створення чи облаштування місць складування та захоронення твердих побутових відходів. – Держбуд України. НВО Укрекогеобуд УкрНДІІНТВ ХД ІГВ, 2001.
9. Землезнавство: Підручник / Багров М.В., Боков В.О., Черваньов І.Г.; За ред. П.Г. Шищенка . – К.: Либідь, 2000. – 464 с.
10. Комбинированное производство тепловой и электрической энергии на основе мусороперерабатывающих комплексов / А.П. Игнатенко, В.А. Гинайло, В.Я. Помохин, В.Д. Семенко // Энергетика и электрификация. – 2001. - №1. – с. 37-43.
11. Левиньский С. Таксономические методы в региональных исследованиях // Региональная наука о размещении производительных сил. – Новосибирск-Иркутск, вып III. – с. 149-159.
12. Математико-картографическое моделирование в географии/ Жуков В.Т., Сербенюк С.Н., Тикунов В.С. Под ред. проф. Салищева К.А. – М.: Мысль, 1980. – 224с.
13. Математические методы в географии. Голиков А.П., Черванев И.Г., Трофимов А.М. – Х.: Вища шк., Изд-во при Харьк. ун-те, 1986. – 144с.
14. Некос В.Е., Снопик Л.М. Численный анализ в природоохранных исследованиях. Учебное пособие. – Харьков, РИГ ХГУ, 1984. – 122с.
15. Окунь Я. Факторный анализ. – М.: Статистика, 1974. – 200с.
16. Сербенюк С.Н., Шкурков В.В. Разработка синтетических карт оценки условий жизни населения с применением факторного анализа // Синтетические карты населения и экономики. – М.: Изд-во Моск. ун-та, 1972 – с.114-133.
17. Сигал И.Я., Кирилюк Н.И., Домбровская Э.П. Проблема мусоросжигания в Украине // Экотехнологии и ресурсосбережение. – 1997. - №1. – с 64-68.
18. Статистичний щорічник України за 2001 рік. За редакцією Осауленка О.Г. – К.: Техніка, 2002. – 648с.
19. Тикунов В.С. Моделирование в социально-экономической картографии. – М.: Изд-во МГУ, 1985. – 280с.
20. Тойн П., Ньюби П. Методы географических исследований. Выпуск 1. Экономическая география. Перевод с англ. – М.: изд-во «Прогресс», 1977. – 272 с.
21. Україна в цифрах у 2000 році. Щорічний статистичний довідник. – К.: Техніка, 2002. – 286 с.

22. Хаггет П.. География: Синтез современных знаний/ Перев. с англ. – М.: Прогресс, 1979. – 684с.
23. Чертко Н.К. Математические методы в физической географии: Учеб. Пособие для геогр. спец. Вузов. – Мн.: изд-во «Университетское», 1987. – 151 с.

ПРИЛОЖЕНИЯ

Приложение 1

Случайные числа

3393	6270	4228	6069	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	3883
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	650P	7156	5634	9703	6221
0905	6986	9396	3975	9255	0537	2479	4589	0562	5345
1420	0470	8679	2328	3939	1292	0406	5428	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

Приложение 2.

Значения критерия τ в зависимости от объема выборки *N* и уровня значимости *a*

N	*a*		*N*	*a*	
	0,05	0,01		0,05	0,01
4	0,955	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

Приложение 3

Значения критерия Стьюдента *t* при различных уровнях вероятности

υ	Уровни вероятности *P*			υ	Уровни вероятности *P*		
	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	4,59	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	2,62	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	00	1,96	2,58	3,29

Приложение 4

Значение критерия хи-квадрат

Число степеней свободы υ	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

Приложение 5

Минимальные существенные значения коэффициентов корреляции при различных уровнях значимости и числах степеней свободы (υ= N_n-2)

υ	*P*		υ	*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	0,47
5	0,75	0,87	28	0,36	0,46
6	0,71	0,83	29	0,36	0,46

Продолжение приложения 5

υ	*P*		υ	*P*	
	0,95	0,99		0.95	0,99
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	0,35
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

Приложение 6

Значение 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	300	∞
3	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74	8,71	8,69	8,66	8,62	8,6	8,58	8,57	8,56	8,54	8,54	8,53
	26,49	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,92	26,83	26,69	26,5	26,41	26,35	26,27	26,23	26,18	26,14	26,12
4	6,59	6,39	6,26	6,16	6,09	6,04	6	5,96	5,91	5,87	5,84	5,8	5,74	5,71	5,7	5,68	5,66	5,65	5,64	5,63
	16,69	15,98	15,52	15,21	14,98	14,8	14,66	14,54	14,37	14,24	14,15	14,02	13,83	13,74	13,69	13,61	13,57	13,52	13,48	13,46
5	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68	4,64	4,6	4,56	4,5	4,46	4,44	4,42	4,4	4,38	4,37	4,37
	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,89	9,77	9,68	9,55	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02
6	4,76	4,53	4,39	4,28	4,21	4,15	4,1	4,06	4	3,96	3,92	3,87	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67
	9,78	9,15	8,75	8,47	8,26	8,1	7,98	7,87	7,72	7,6	7,52	7,39	7,23	7,14	7,09	7,02	6,99	6,94	6,9	6,88
7	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57	3,52	3,49	3,44	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23
	8,45	7,85	7,46	7,19	7	6,84	6,71	6,62	6,47	6,35	6,27	6,07	5,9	5,85	5,78	5,75	5,7	5,67	5,66	5,65
8	4,07	3,84	3,69	3,58	3,5	3,44	3,39	3,34	3,28	3,23	3,2	3,15	3,08	3,05	3,03	3	2,98	2,96	2,94	2,93
	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,67	5,56	5,48	5,36	5,2	5,11	5,06	5	4,96	4,91	4,88	4,86
9	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07	3,02	2,98	2,93	2,86	2,82	2,8	2,77	2,76	2,73	2,72	2,71
	6,99	6,42	6,06	5,8	5,62	5,47	5,35	5,26	5,11	5	4,92	4,8	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31
10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91	2,86	2,82	2,77	2,7	2,67	2,64	2,62	2,59	2,56	2,55	2,54
	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,71	4,6	4,52	4,41	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91
11	3,59	3,36	3,2	3,09	3,01	2,95	2,9	2,86	2,78	2,74	2,7	2,65	2,57	2,53	2,5	2,47	2,45	2,42	2,41	2,4
	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,4	4,29	4,21	4,1	3,94	3,86	3,8	3,74	3,7	3,66	3,62	3,6
12	3,49	3,26	3,11	3	2,92	2,85	2,8	2,76	2,69	2,64	2,6	2,54	2,46	2,42	2,4	2,36	2,35	2,32	2,31	2,3
	5,95	5,41	5,06	4,82	4,65	4,5	4,39	4,3	4,16	4,05	3,98	3,86	3,7	3,61	3,56	3,49	3,46	3,41	3,38	3,36
13	4,41	4,18	3,02	2,92	2,84	2,77	2,72	2,67	2,6	2,55	2,51	2,46	2,38	2,34	2,32	2,28	2,26	2,24	2,22	2,21
	5,74	5,2	4,86	4,62	4,44	4,3	4,19	4,1	3,96	3,85	3,78	3,67	3,51	3,42	3,37	3,3	3,27	3,21	3,18	3,16
14	3,34	3,11	2,96	2,85	2,77	2,7	2,65	2,6	2,53	2,48	2,44	2,39	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,8	3,7	3,62	3,51	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3
15	3,29	3,06	2,9	2,79	2,7	2,64	2,59	2,55	2,48	2,43	2,39	2,33	2,25	2,21	2,18	2,15	2,12	2,1	2,08	2,07
	5,42	4,89	4,56	4,32	4,14	4	3,89	3,8	3,67	3,56	3,48	3,36	3,2	3,12	3,07	3	2,97	2,92	2,89	2,87
16	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,37	2,33	2,28	2,2	2,16	2,13	2,09	2,07	2,04	2,02	2,01

	5,29	4,77	4,44	4,2	4,03	3,89	3,78	3,69	3,55	3,45	3,37	3,25	3,1	3,01	2,96	2,89	2,86	2,8	2,77	2,75
17	3,2	2,96	2,81	2,7	2,62	2,55	2,5	2,45	2,38	2,33	2,29	2,23	2,15	2,11	2,08	2,04	2,02	1,99	1,87	1,96
	5,18	5,67	4,34	4,1	3,93	3,79	3,68	3,52	3,4	3,35	3,27	3,16	3	2,92	2,86	2,79	2,76	2,7	2,67	2,65
50	2,79	2,56	2,4	2,29	2,2	2,13	2,07	2,02	1,95	1,9	1,85	1,78	1,69	1,63	1,6	1,55	1,52	1,48	1,46	1,44
	4,2	3,72	3,41	3,18	3,02	2,88	2,87	2,7	2,56	2,46	2,39	2,26	2,1	2	1,94	1,86	1,82	1,76	1,71	1,68
200	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87	1,8	1,74	1,69	1,62	1,52	1,45	1,42	1,35	1,28	1,24	1,17	1,11
	3,88	3,41	3,11	2,9	2,74	2,6	2,5	2,41	2,28	2,17	2,09	1,97	1,79	1,69	1,62	1,53	1,48	1,39	1,33	1,28
∞	2,6	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,75	1,69	1,64	1,57	1,46	1,4	1,35	1,28	1,24	1,17	1,11	1
	3,78	3,32	3,02	2,8	2,64	2,51	2,41	2,32	2,18	2,07	1,99	0,87	1,69	1,59	1,52	1,41	1,36	1,25	1,15	1,09

Прмечание: Верхнее значение для каждой ν_2 (степени свободы для меньшей дисперсии) соответствует уровню *P=0,95,* нижнее - *P=0,99*

Приложение 7

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	2037	2284	429	~~363~~	1904	1986	834	1961	1125	2063	1099
B	2037	0	247	2466	2398	134	54	1203	77	3162	~~30~~	938
C	2284	247	0	2713	2645	380	298	1450	323	3409	221	1185
D	429	2466	2713	0	~~74~~	2333	2415	1263	2390	696	2492	1528
E	***363***	2398	2645	***74***	0	2265	2347	1196	2322	765	2424	1460
F	1904	134	380	2333	2265	0	82	1070	58	3029	159	805
G	1986	54	298	2415	2347	82	0	1152	~~28~~	3111	77	887
H	834	1203	1450	1263	1196	1070	1152	0	1127	1959	1229	266
I	1961	77	323	2390	2322	58	***28***	1127	0	3086	103	862
J	1125	3162	3409	696	765	3029	3111	1959	3086	0	3188	2224
K	2063	***30***	221	2492	2424	159	77	1229	103	3188	0	964
L	1099	938	1185	1528	1460	805	887	266	862	2224	964	0
M	779	1258	1505	1208	1140	1125	1207	***59***	1182	1904	1284	320
N	1866	172	418	2295	2227	39	120	1032	96	2991	197	767
O	1011	3048	3295	582	***650***	2915	2997	1845	2972	***115***	3074	2110
P	1403	634	881	1832	1764	501	583	569	558	2528	660	304
Q	1879	158	405	2308	2240	26	107	1045	82	3004	184	780
R	1906	132	378	2335	2267	***4***	80	1072	56	3031	157	807
S	1894	143	390	2323	2255	~~10~~	92	1060	68	3019	169	795
T	984	1053	1300	1413	1345	920	1002	150	977	2109	1079	***116***
U	2069	35	***215***	2498	2430	165	~~83~~	1235	108	3194	***7***	970
V	1529	508	755	1958	1890	375	457	695	432	2654	534	430
W	1582	455	702	2011	1943	322	404	748	379	2707	481	484
X	1942	98	342	2371	2304	40	45	1108	***27***	3067	121	844
Y	1869	168	415	2298	2230	35	117	1035	93	2994	194	770

	M	N	O	P	Q	R	S	T	U	V	W	X	Y
A	779	1866	1011	1403	1879	1906	1894	984	2069	1529	1582	1942	1869
B	1258	172	3048	634	158	132	143	1053	35	508	455	98	168
C	1505	418	3295	881	405	378	390	1300	~~215~~	755	702	342	415
D	1208	2295	582	1832	2308	2335	2323	1413	2498	1958	2011	2371	2298
E	1140	2227	~~650~~	1764	2240	2267	2255	1345	2430	1890	1943	2304	2230
F	1125	39	2915	501	26	~~4~~	***10***	920	165	375	322	40	35
G	1207	120	2997	583	107	80	92	1002	***83***	457	404	45	117
H	~~59~~	1032	1845	569	1045	1072	1060	150	1235	695	748	1108	1035
I	1182	96	2972	558	82	56	68	977	108	432	379	~~27~~	93
J	1904	2991	~~115~~	2528	3004	3031	3019	2109	3194	2654	2707	3067	2994
K	1284	197	3074	660	184	157	169	1079	~~7~~	534	481	121	194
L	320	767	2110	304	780	807	795	~~116~~	970	430	484	844	770
M	0	1087	***1790***	624	1100	1127	1115	~~206~~	1290	750	803	1164	1090
N	1087	0	2877	463	13	40	29	882	203	337	302	77	~~9~~
O	~~1790~~	2877	0	2414	2890	2917	2905	1995	3080	2540	2593	2953	2880
P	624	463	2414	0	476	503	491	***419***	666	~~126~~	179	539	466
Q	1100	13	2890	476	0	***27***	15	895	190	350	297	64	~~12~~
R	1127	40	2917	503	~~27~~	0	28	922	163	377	324	***38***	37
S	1115	29	2905	491	15	28	0	910	175	365	312	50	25
T	***206***	882	1995	~~419~~	895	922	910	0	1085	545	598	958	885
U	1290	203	3080	666	190	163	175	1085	0	540	487	127	200
V	750	337	2540	***126***	350	377	365	545	540	0	~~55~~	414	340
W	803	302	2593	179	297	324	312	598	487	***55***	0	360	~~287~~
X	1164	77	2953	539	64	~~38~~	50	958	127	414	360	0	75
Y	1090	***9***	2880	466	***12***	37	25	885	200	340	***287***	75	0