Министерство образования и науки Российской Федерации
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

С.М. Гордеева

ПРАКТИКУМ

по дисциплине

«СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ»

Рекомендовано Учебно-методическим объединением в области гидрометеорологии в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности «Океанология»

РГГМУ
Санкт-Петербург
2010

УДК 551.46(07)

Гордеева С.М. Практикум по дисциплине «Статистические методы обработки и анализа гидрометеорологической информации» – СПб.: РГГМУ, 2010. – 74 с.

Рецензент: Г.В. Алексеев, д-р геогр. наук, проф. (ААНИИ)

Ответственный редактор: Л.Н. Карлин, д-р физ.-мат. наук, проф., РГГМУ

Приводятся краткие теоретические сведения о статистических методах, даются рекомендации по выполнению практических работ. Рассмотрена технология расчетов в табличном процессоре Microsoft Excel.

Предназначено студентам гидрометеорологических специальностей.

ПРЕДИСЛОВИЕ

Практикум составлен в соответствии с программой дисциплины «Статистические методы обработки и анализа гидрометеорологической информации», которая является одной из основных в подготовке специалистов гидрометеорологического профиля. Она включает следующие группы статистических методов: первичная обработка наблюдений, корреляционный и регрессионный анализ, анализ временных рядов, непараметрический анализ.

В практикуме приводятся краткие теоретические сведения о статистических методах, даются рекомендации по выполнению всех практических работ. В качестве инструмента для статистических расчетов предлагается табличный процессор Microsoft Excel. В соответствии с этим выполнены примеры и анализ соответствующей информации.

Вводные понятия статистики

Совокупность наблюдений (измерений) какой-либо характеристики объекта представляется в виде ***статистического ряда***. Статистический ряд записывается в виде столбца. Если одновременно измерялось несколько характеристик объекта, получается несколько рядов. Совокупность статистических рядов можно записать как *матрицу наблюдений (матрицу исходных данных)*, где каждая строка представляет собой *случай* наблюдений, а столбец — *признак*. Часто столбцы в статистической матрице называют *переменные*.

Пример матрицы наблюдений (исходных данных):

Случаи	Переменные (признаки)	
	рост, см	вес, кг
1	156	48
2	185	95
3	173	81
4	167	69
5	161	55
6	170	85
…	…	…

Вообще весь мыслимо возможный набор значений характеристики природного объекта бесконечен. Эта совокупность называется ***генеральной совокупностью***. В силу ограниченности возможностей наблюдения (измерения) нам приходится иметь дело с ***выборкой***, которая является частью генеральной совокупности.

Особым случаем статистического ряда является *временной ряд*. Здесь в качестве случаев выступают *моменты времени*, в которые производится измерение характеристики (переменной). В отличие от обычного статистического ряда, строки которого можно переставлять в пределах матрицы, при этом информация не изменится, для временного ряда *последовательность* случаев принципиальна и изменение ее ведет к «порче» информации.

Для временного ряда **обязательно** должны быть представлены атрибуты:

– *что* – смысл наблюдаемой характеристики;

– *где* – географические координаты, где производились наблюдения;

– *когда* – период времени, в течение которого велись наблюдения;

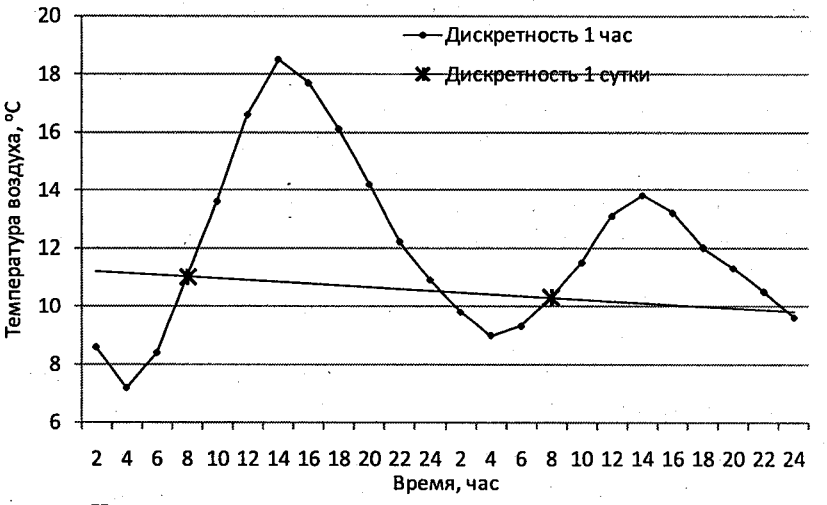
– *дискретность* измерений.

Дискретность представляет собой интервал времени, через который проводились измерения и через который представлены случаи. Например, измерения проводились через 2 с., через 3 ч., через сутки или через месяц и т.п.

В зависимости от дискретности, мы теряем информацию о поведении характеристики в то время, когда наблюдений не проводилось, см. как пример рисунок: наблюдения за температурой воздуха (T_a) на многосуточной станции проводились с дискретностью 2 ч. (ряд 1). Было отмечено, что в первые сутки была холодная ночь и очень теплый день, а во вторые сутки, наоборот, была относительно теплая ночь и холодный день. Если бы мы измеряли T_a только в 8 ч. утра (дискретность 1 сутки, ряд 2), то отметили бы, что в первые и во вторые сутки температура воздуха была почти одинаковая. Таким образом, при переходе к большей дискретности мы потеряли информацию о суточном ходе T_a.

Выбор интервала дискретности связан с масштабом изменчивости исследуемого природного процесса. Например, для исследо-

вания турбулентности дискретность измерений должна быть менее часа, для климатических процессов – не меньше одного года, для исследования сезонного хода – декадная (10 дней) или месячная, для синоптического масштаба – срочная (через 4–6 часов) или суточная.



Изменчивость температуры воздуха на многосуточной станции

Вводные понятия технологии

Под *технологией статистических расчетов* подразумевается способ использования некоторого программного обеспечения, предназначенного для решения статистических задач. Чаще всего для этих целей используют специализированные статистические пакеты, такие как STATISTICA, SPSS, STADIA и др.

В данном практикуме будет рассмотрена технология расчетов с помощью табличного процессора Microsoft Excel, который входит в состав пакета Microsoft Office и, как правило, установлен на каждом компьютере под управлением ОС Windows.

Для проведения статистической обработки информации Excel включает в себя программную надстройку «***Пакет анализа***» и библиотеку из 78 статистических функций.

Для активации «Пакета анализа» нужно выполнить следующее:

– Запустить Excel. Появится окно активного листа новой или существующей книги Excel;

– В меню «Сервис» выбрать пункт «Надстройки». Раскроется окно со списком доступных надстроек.

– В этом списке найти «Пакет анализа». Поставить рядом с ним «галку». Если она уже стоит, ничего делать не надо. Нажать «ОК».

– В меню «Сервис» появится команда **«Анализ данных»**.

Если в процессе указанной процедуры возникли дополнительные запросы со стороны системы (компьютера), значит для установки «Пакета анализа» потребуется дистрибутив Microsoft Office.

Кроме того, простые расчеты лучше всего выполнять с помощью *статистических функций*, найти которые можно следующим образом:

– В меню «Вставка» выбрать «функция»;

– В области «категория» выбрать «статистические»;

– В открывшемся списке выбрать нужную функцию, сверяя ее смысл по текстовому описанию, приводимому внизу окна.

В конце каждой работы рассматриваются технологические особенности ее выполнения с помощью табличного процессора Microsoft Excel.

ПРАКТИЧЕСКАЯ РАБОТА 1

Первичные статистики и эмпирическая функция распределения

Теоретическая часть

Для характеристики статистических рядов используются *показатели первичной статистики.*

Они разбиваются на несколько групп:

1. Показатели положения (например, среднее арифметическое, медиана, мода);
2. Показатели рассеяния (например, дисперсия, стандартное отклонение, размах и др.);
3. Эмпирическая функция распределения (ЭФР) и характеристики ее формы.

Показатели положения

Среднее арифметическое значение статистического ряда ($\bar{x}$). Характеризует *центр тяжести* исследуемой характеристики или *точку ее равновесия* при различных колебаниях. Рассчитывается по формуле

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

где N – длина статистического ряда (количество значений в нем).

Медианой (Me) называется значение признака, приходящегося на середину ранжированного (упорядоченного по возрастанию) ряда. *Главное свойство* медианы заключается в том, что сумма абсолютных отклонений членов ряда от медианы есть величина наименьшая:

$$\sum_{i=1}^{N} |x_i - Me| = min.$$

Для коротких статистических рядов ($N < 30$) медиану используют вместо среднего арифметического значения.

***Модой** (Мо)* называется наиболее часто встречающаяся в статистическом ряду величина.

Показатели разброса

Дисперсия** (D)* и связанное с ней ***стандартное (или среднее квадратическое) ***отклонение** (σ)* характеризуют *среднее* рассеяние значений ряда от среднего арифметического значения.

Рассчитываются по формулам соответственно:

$$D = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \quad \sigma = \sqrt{D}.$$

***Размах вариации** R* характеризует максимальный разброс значений ряда:

$$R = max - min\,.$$

Рассмотренные характеристики можно наглядно представить на графике (рис. 1.1).

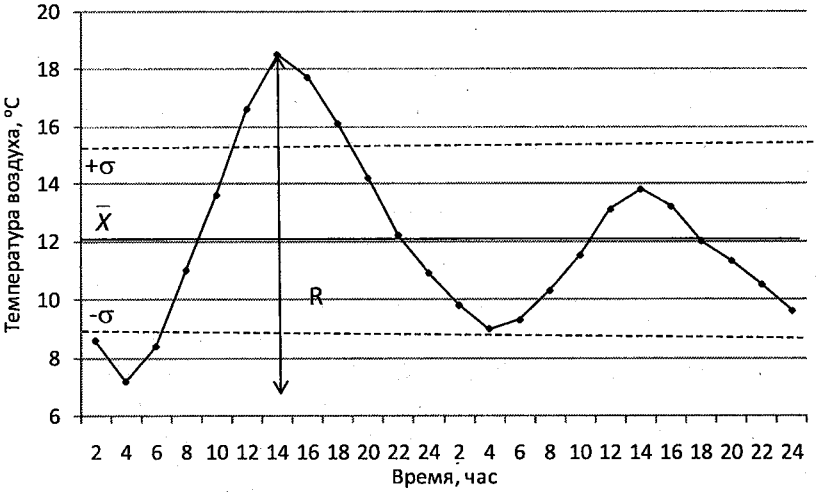


Рис. 1.1. Изменчивость температуры воздуха на многосуточной станции

***Коэффициент вариации** (C)* – безразмерный показатель изменчивости характеристики в исследуемой выборке.

$$C = \frac{\sigma}{\bar{x}} \cdot 100\%.$$

Коэффициент вариации позволяет оценить, велика или нет изменчивость ряда: если он составляет более 33% – изменчивость значительна, а если меньше – изменчивость мала, выборка может считаться однородной.

Коэффициент вариации также используется для сравнения изменчивости двух выборок с разными единицами измерения.

Показатели, характеризующие закон распределения

Функция распределения показывает соотношение между возможными значениями случайной величины и вероятностями их появления. ***Эмпирической*** (полученной опытным путем) ***функцией распределения*** (ЭФР) называется функция распределения, рассчитанная по выборке. В гидрометеорологии часто рассчитывается функция, называемая «повторяемость», по сути, это и есть ЭФР.

Для расчета ЭФР предварительно формируется несколько интервалов изменчивости переменной в исследуемом статистическом ряду, а затем рассчитывается количество значений переменной, попадающих в каждый интервал (*частота*).

Для изображения ЭФР применяется *гистограмма*, где по оси абсцисс откладываются значения интервалов, а частоты представлены прямоугольниками, построенными на соответствующих интервалах и имеющими высоту, пропорциональную частоте.

Кроме того, рассчитывается ***интегральная ЭФР*** путем последовательного суммирования частот (вероятности) на всех интервалах. Она характеризует вероятность появления величины, меньше заданной. Интегральная ЭФР в значениях вероятности изображается в виде графика кривой, монотонно возрастающей от 0 до 1.

Пример ЭФР представлен на рис. 1.2.

По гистограмме можно легко определить ***медиану.*** Для этого нужно найти абсциссу, соответствующую ординате «интегральной вероятности» 50%.

Кроме того, по гистограмме определяются ***моды*** как локальные максимумы ЭФР («вершины холмов»). Мода бывает одна, две или несколько. Соответственно, распределение является одномодальным, двухмодальным и многомодальным. Моды характеризуют наиболее устойчивые состояния характеристики.

20 - 10.

26,03 27,26 28,49

26,38 27,26. 27,84

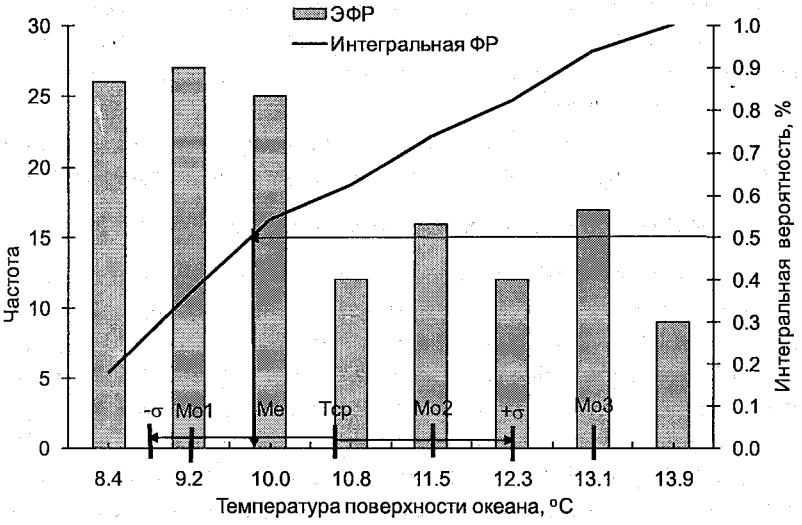


Рис. 1.2. Эмпирическая функция распределения среднемесячных значений температуры поверхности океана в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Например, на гистограмме на рис. 1.2. выделяются три моды. Первая мода ($Mo_1 = 9{,}2$ °С) характеризует относительно холодное состояние поверхности океана, две другие моды – относительно теплое состояние (соответственно, 11,5 °С и 13,1 °С). Причем, чаще вода бывает холоднее (чем теплее) среднего.

ЭФР может иметь *«хвосты»*, т.е. некоторое небольшое число наблюдений значительно больше (положительный «хвост») или меньше (отрицательный «хвост») среднего значения.

Характеристики формы ЭФР

Асимметрия характеризует симметричность ЭФР относительно среднего значения и рассчитывается по формуле:

$$As = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \bar{x})^3.$$

При полной симметрии ЭФР относительно среднего значения $As = 0$. Если $As > 0$, то ЭФР обладает положительным «хвостом» и основная масса наблюдений (а также медиана) меньше среднего

значения. Если $As < 0$, то ЭФР обладает отрицательным «хвостом» и основная масса наблюдений (а также медиана) больше среднего значения.

Эксцесс характеризует островершинность распределения и рассчитывается по формуле:

$$Ex = \left[\frac{1}{N\sigma^4} \sum_{i=1}^{N} (x_i - \bar{x})^4 \right] - 3.$$

Если $Ex > 0$, то ЭФР является относительно островершинной. Если $Ex < 0$, то ЭФР является относительно плосковершинной и распределение стремится к случайному.

Расчетная часть

Исходные данные

Среднемесячная температура поверхности океана (ТПО) в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Порядок выполнения работы

1. Рассчитать ряд ТПО с другой дискретностью – ряд среднегодовых значений.

Далее для каждого из двух рядов (среднемесячных и среднегодовых значений):

2. Рассчитать описательные статистики (см. технологию выполнения работы, п.1):
   - среднее арифметическое значение;
   - дисперсию;
   - среднее квадратическое (стандартное) отклонение;
   - коэффициент вариации;
   - минимум, максимум;
   - размах вариации;
   - асимметрию;
   - эксцесс.

3. Построить график временного хода рядов ТПО. На график нанести среднее значение, стандартное отклонение, размах вариации (пример на рис. 1.1).

4. Рассчитать ЭФР (см. технологию выполнения работы, п. 2) Сформировать таблицу (см. пример табл. 1.1). Построить гистограмму и кривую интегральной ЭФР (см. технологию выполнения

работы, п. 3). На гистограмму нанести среднее значение и стандартное отклонение (пример на рис. 1.2).

5. По рисунку ЭФР определить медиану и моду (моды), нанести их на гистограмму (пример на рис. 1.2).

6. Проанализировать полученные результаты для каждого ряда отдельно, а также в сравнении друг с другом соответствующих характеристик.

Пример вычислений приводится в табл. 1.1 и рис. 1.2.

Технология выполнения работы

1. Все первичные статистики можно рассчитать в «Пакете анализа» Excel в модуле «Описательные статистики». В качестве «входного интервала» выделить исходный статистический ряд. Поставить «галку» в поле «итоговая статистика». Нажать «ОК».

2. Расчет эмпирической функции распределения:

– определить количество интервалов, адекватное для расчета ЭФР.

Число интервалов $k_{max} = 5lg\,N$;

– округлить до целого;

– рассчитать размах интервала $\Delta X_k = \dfrac{X_{max} - X_{min}}{k_{max}}$;

– определить границы интервалов

$(C_1, C_2), (C_2, C_3),...,(C_{kmax}, C_{kmax+1})$, где $C_1=min$, $C_2 = C_1+\Delta X_k$,

$C_3 = C_2+\Delta X_k$ и т.д. так что $C_{kmax+1}=max$;

– рассчитать середину каждого интервала x_k ;

– оценить частоту (повторяемость) m_k как число членов выборки, попавших в каждый интервал. Гистограмма распределения частоты представляет собой эмпирическую функцию распределения. Частоту m_k можно рассчитать в «Пакете анализа» Excel в модуле «Гистограмма». В качестве «входного интервала» выделить исходный статистический ряд. В качестве «интервала карманов» взять графу табл.1.1 «интервалы «до». Нажать «ОК»;

– рассчитать накопленную частоту m' путем последовательной суммы m для каждого интервала: $m'_1=m_1$, $m'_2=m_1+m_2$, $m'_3 = m_1+m_2+m_3$ и т.д.;

– рассчитать вероятность появления значения характеристики в конкретном интервале $p = m/N$, где N – длина ряда;

– рассчитать интегральную вероятность (обеспеченность) появления значения характеристики в конкретном интервале $p' = m'/N$, где N – длина ряда.

3. Рисунок с совмещением гистограммы и графика интегральной вероятности выполняется на основе диаграммы Excel: «нестандартные» – «график/гистограмма 2».

Таблица 1.1

Расчет эмпирической функции распределения среднемесячных значений температуры поверхности океана в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Номер интервала	Интервалы		$\bar{x}_k$	Частота, m_k	Накопленная частота m'	Вероятность, $p=m/N$	Интегральная вероятность, $p'=m'/N$
	от	до					
1	8,04	8,82	8,43	26	26	0,181	0,181
2	8,82	9,59	9,20	27	53	0,188	0,368
3	9,59	10,37	9,98	25	78	0,174	0,542
4	10,37	11,15	10,76	12	90	0,083	0,625
5	11,15	11,92	11,53	16	106	0,111	0,736
6	11,92	12,70	12,31	12	118	0,083	0,819
7	12,70	13,47	13,09	17	135	0,118	0,938
8	13,47	14,25	13,86	**9**	144	0,063	1,000

7 27,55 | 4 | 0,97 частота
8 27,68 | 3 | 1,00

ПРАКТИЧЕСКАЯ РАБОТА 2

Проверка соответствия эмпирической функции распределения нормальному закону

Теоретическая часть

На исследуемую характеристику влияет большое количество разнообразных факторов. Например, температура поверхности океана (ТПО) меняется под влиянием радиационного баланса, испарения, турбулентного обмена с атмосферой, адвекции и др. Однако влияние их не одинаково по величине, а зависит от расположения точки на земном шаре. Кроме того, при изменении дискретности информации (т.е. при переходе к другому *масштабу осреднения*) некоторые факторы могут уменьшать или увеличивать свое влияние.

Тем не менее, может существовать такая дискретность и пространственное положение точки, что влияние всех факторов будет равнозначно, мало и случайным образом разнонаправленно. Тогда, несмотря на любые изменения, характеристика в конечном итоге будет стремиться к состоянию равновесия (т.е. к своему среднему значению). Если для такого ряда построить функцию распределения, то она будет иметь правильную колоколообразную форму (рис. 2.1) и называться ***нормальным законом***.

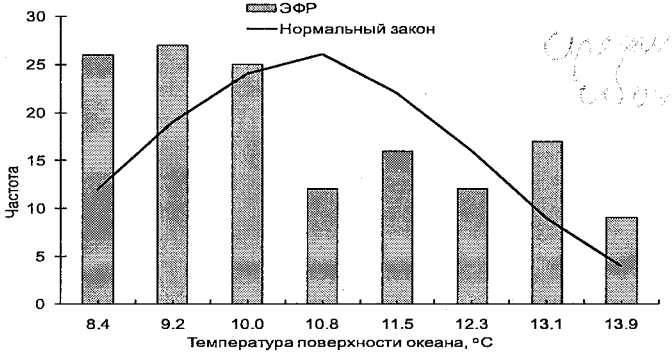


Рис. 2.1. Эмпирическая функция и соответствующий ей нормальный закон распределения среднемесячных значений температуры поверхности океана в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Кроме того, ***нормальный закон*** имеет особое значение в статистике, так как он является *предельным законом*, к которому приближаются другие законы распределения при часто встречающихся условиях.

Функция *плотности вероятности нормального закона распределения* может быть выражена формулой:

$$f(\bar{x},\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\bar{x})^2}{2\sigma^2}},$$

где x – значение случайной величины; $\bar{x}$ и σ – среднее значение и стандартное отклонение исходного ряда, соответственно.

Нормальному закону свойственно равенство среднего значения, моды и медианы, а также симметричность относительно среднего. Кроме того, $As = Ex = 0$.

Форма кривой функции нормального закона и ее расположение на горизонтальной оси зависит от среднего значения и стандартного отклонения исходного ряда.

ЭФР, как правило, отличаются от нормального закона в большей или меньшей степени. Однако всегда нужно ответить на вопрос: являются ли эти различия значительными или ими можно пренебречь и считать нашу ЭФР нормальным законом?

Для объективного ответа привлекается ***статистический критерий*** χ^2 (хи-квадрат). Он характеризует совокупность относительных различий ЭФР и теоретического распределения (в частности, нормального закона) и может быть рассчитан на основе интервалов ЭФР (см. работу 1) по формуле:

$$\chi^2=\sum_{k=1}^{k_{max}}\frac{(m_k-n_k)^2}{n_k}$$

где m_k – частота эмпирической функции распределения в k-том интервале; k_{max} – количество интервалов; n_k – частота нормального закона для k-того интервала.

Если различий нет, т.е. ЭФР абсолютно совпадает с функцией нормального закона распределения, то $\chi^2 = 0$.

В других случаях необходимо задать точность, с которой мы хотим определить соответствие ЭФР нормальному закону. Эта точность определяется *доверительной вероятностью* и задается исследователем в зависимости от задачи. Для гидрометеорологи-

ческих рядов, как правило, доверительная вероятность определяется: $p = 95\%$. Тогда соответствующая ошибка при принятии решения составляет 5%. Это *уровень значимости* α.

В зависимости от уровня значимости около нуля формируется некоторый интервал значений χ^2, в пределах которого делается вывод о соответствии ЭФР нормальному закону. Граница интервала определяется как $\chi^2_{кр}$ (хи-квадрат критическое), которое определяется по статистическим таблицам в зависимости от *уровня значимости* α и *числа степеней свободы* $\nu = k - 3$, где k – количество интервалов ЭФР.

Таким образом, если рассчитанный по формуле χ^2 будет меньше, чем $\chi^2_{кр}$, то статистически он близок к нулю и ЭФР соответствует нормальному закону. А если χ^2 будет больше, чем $\chi^2_{кр}$, тогда различия между ЭФР и нормальным законом статистически значимы и ЭФР не соответствует нормальному закону.

На основании этого статистического вывода можно сделать физический вывод: если ЭФР соответствует нормальному закону, значит на характеристику влияет большое число малых равнозначных факторов. Если ЭФР не соответствует нормальному закону, тогда среди факторов, влияющих на характеристику, есть один или несколько преобладающих (рассуждения о факторах см. выше).

Расчетная часть

Исходные данные

Среднемесячная температура поверхности океана (ТПО) в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Порядок выполнения работы

В продолжение таблицы расчета эмпирической функции распределения (табл. 1.1):

1. Для каждого интервала рассчитать плотность вероятности нормального закона распределения $f(x_k, \bar{x}, \sigma)$ по формуле (2.1), где x_k – середина интервала, $\bar{x}$ – среднее значение выборки, σ – стандартное отклонение выборки (см. технологию выполнения работы, п.1).

2. Для каждого интервала перевести значения плотности вероятности нормального закона в значения соответствующих *частот нормального закона*:

$$n_k^* = f(x_k, \bar{x}, \sigma) \cdot \Delta X_k \cdot N,$$

где ΔX_k – размах интервала; N – длина исходного ряда.

3. Частоты нормального закона n_k* округлить до целых значений. Получится n_k.

4. Рассчитать эмпирический критерий χ^2 по формуле (2.2).

Пример расчетов приведен в табл. 2.1.

5. Проверить условие $\chi^2 \geq \chi^2_{кр}$ при числе степеней свободы $\nu = k - 3$ и уровне значимости $\alpha = 5\%$ (см. технологию выполнения работы, п.2). Если это условие выполняется, то гипотеза о соответствии эмпирического и теоретического распределений отвергается; расхождение между ними носит неслучайный характер. Следовательно, на исследуемую характеристику влияют некоторые преобладающие неслучайные факторы.

6. Построить совмещенный график ЭФР и нормального закона распределения (пример на рис. 2.1)

Технология выполнения работы

1. Плотность вероятности нормального закона распределения можно рассчитать с помощью функции Excel «нормрасп», где «интегральная» = 0.

2. $\chi^2_{кр}$ можно рассчитать с помощью функции Excel «хи2обр».

3. Рисунок с совмещением гистограммы и графика нормального закона выполняется на основе диаграммы Excel: «нестандартные» – «график/гистограмма 2».

Таблица 2.1

Проверка соответствия ЭФР среднемесячных значений температуры поверхности океана в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг. нормальному закону распределения

Номер интервала	Интервалы		x_k	Частота, m_k	$f(x_k, \bar{x}, \sigma)$	n_k*	n_k	$\frac{(m_k - n_k)^2}{n_k}$
	от	до						
1	8,04	8,82	8,43	26	0,107	11,966	12	16,33
2	8,82	9,59	9,20	27	0,168	18,814	19	3,37
3	9,59	10,37	9,98	25	0,217	24,222	24	0,04
4	10,37	11,15	10,76	12	0,228	25,538	26	7,54
5	11,15	11,92	11,53	16	0,197	22,048	22	1,64
6	11,92	12,70	12,31	12	0,139	15,587	16	1,00
7	12,70	13,47	13,09	17	0,081	9,024	9	7,11
8	13,47	14,25	13,86	9	0,038	4,278	4	6,25
							χ^{2*}	43,28
							$\chi^2_{кр}(0,05;5)$	11,07

ПРАКТИЧЕСКАЯ РАБОТА 3

Проверка статистических гипотез. Оценка стационарности временного ряда

Теоретическая часть

Под ***статистической гипотезой*** понимают всякое высказывание о генеральной совокупности, проверяемое по выборке.

Проверяемую статистическую гипотезу принято называть ***нулевой гипотезой*** и обозначать H_0. Противоречащую ей гипотезу – ***альтернативной*** и обозначать H_1.

Поскольку при проверке статистических гипотез приходится иметь дело со статистическим материалом, то, отвергая или принимая нулевую гипотезу, всегда рискуем совершить ошибку. Ошибку, заключающуюся в том, что нулевая гипотеза отвергается, тогда как она в действительности верна, называют ***ошибкой первого рода.*** Ошибку, состоящую в том, что нулевая гипотеза принимается, тогда как она в действительности неверна, называют ***ошибкой второго рода.***

Проверка статистических гипотез осуществляется с помощью различных ***критериев***, т.е. величин, значения которых можно вычислить на основе выборки.

Множество значений критерия разбивается на две области: ***доверительную*** (область принятия нулевой гипотезы) ***и критическую*** (область, где нулевая гипотеза отвергается) (рис. 3.1).

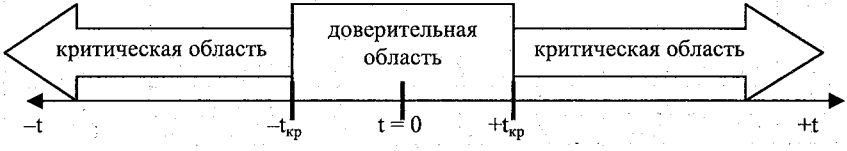


Рис. 3.1. Области существования критерия Стьюдента

Критическая область выбирается таким образом, чтобы вероятность совершить ошибку первого рода не превосходила заранее заданный ***уровень значимости***.

Уровень значимости α задается исследователем и для гидрометеорологических рядов обычно выбирается равным 0,01, 0,05, 0,10.

Таким образом, критическое значение критерия и, соответственно, размер критической области зависят от α. Причем, если учесть, что α показывает вероятность попадания в критическую область, то при больших значениях α критическая область больше, а доверительная – меньше (рис. 3.2).

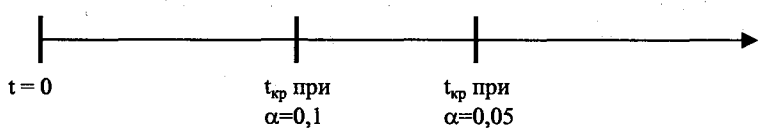


Рис. 3.2. Положение критического значения критерия в зависимости от разных уровней значимости α

Порядок проверки статистической гипотезы

1. Формулируется нулевая гипотеза H_0;
2. Формулируется альтернативная гипотеза H_1;
3. Выбирается статистический критерий и рассчитывается его эмпирическое значение;
4. Задается уровень значимости α и определяется критическое значение критерия;
5. Сравниваются эмпирическое и критическое значения критерия. Если эмпирическое значение больше критического (по модулю), т.е. эмпирическое значение критерия попадает в критическую область при заданном α, тогда нулевая гипотеза отвергается.

Доверительным интервалом называется интервал значений статистической характеристики, соответствующий доверительной области значений некоторого статистического критерия. Таким образом, чтобы определить доверительный интервал необходимо привлечь теорию проверки статистических гипотез.

Определение доверительного интервала для математического ожидания

Напомним, что мы имеем дело с выборкой из генеральной совокупности (ГС). Естественно предположить, что статистические характеристики нашей выборки свойственны и её генеральной совокупности, например, среднее значение выборки и среднее значение ГС (т.е. математическое ожидание m_x) и т.п.

Таким образом, можно сформулировать нулевую гипотезу $\boldsymbol{H_0}$: $m_x = \bar{x}$; соответственно, альтернативная ей гипотеза $\boldsymbol{H_1}$: $m_x \neq \bar{x}$.

Для проверки гипотезы используется статистический критерий Стьюдента.

Так как задача состоит в определении доверительной области статистического критерия, то достаточно найти ее границы, т.е. критическое значение критерия Стьюдента $t_{кр}$ по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $\nu = N–1$, где N – длина ряда.

Тогда доверительный интервал для m_x определится как

$$\bar{x} - \Delta x < m_x < \bar{x} + \Delta x,$$

где

$$\Delta x = t_{кр} \frac{\sigma}{\sqrt{N}},$$

здесь σ – стандартное отклонение исходного ряда; N – длина ряда.

Определение доверительного интервала для дисперсии ГС

Аналогично математическому ожиданию определяется доверительный интервал для дисперсии генеральной совокупности.

$$\boldsymbol{H_0}: \mu_2 = s^2; \boldsymbol{H_1}: \mu_2 \neq s^2;$$

Для проверки гипотезы выбирается критерий χ^2.

Определяются нижняя χ_1^2 и верхняя χ_2^2 границы доверительной области критерия и тогда доверительный интервал для дисперсии генеральной совокупности определяется как

$$s^2 \cdot \Delta D_1 < \mu_2 < s^2 \cdot \Delta D_2,$$

где s^2 – дисперсия выборки,

$$\Delta D_1 = \frac{N}{\chi_1^2}; \Delta D_2 = \frac{N}{\chi_2^2}.$$

χ_1^2 определяется по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $\nu = N - 1$; χ_2^2 определяется по уровню значимости $(1 - \alpha) = 0{,}95$ и числу степеней свободы $\nu = N - 1$; где N – длина ряда.

Проверка гипотезы о равенстве средних значений

Часто возникает ситуация, когда необходимо сравнить характеристики двух качественно одинаковых выборок, например, средний рост студентов в двух группах. Естественно, числовые значения всегда будут различаться. Но являются ли эти различия значительными? Ответить на этот вопрос можно с помощью теории статистических гипотез.

Сформулируем нулевую гипотезу H_0: $\bar{x}_1 = \bar{x}_2$; Альтернативную, H_1: $\bar{x}_1 \neq \bar{x}_2$.

Для проверки гипотезы выберем t-критерий Стьюдента и рассчитаем его эмпирическое значение:

$$t^* = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{N_1 D_1 + N_2 D_2}} \sqrt{\frac{N_1 N_2 (N_1 + N_2 - 2)}{N_1 + N_2}}$$

где D_1 и D_2 – дисперсии двух частей выборки соответственно; N_1 и N_2 – длины соответствующих частей ряда.

Определим критическое значение критерия Стьюдента $t_{кр}$ по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $\nu = N_1 + N_2 - 2$, где N_1 и N_2 – длины соответствующих частей ряда.

Сравним эмпирическое и критическое значения критерия. Если эмпирическое значение больше критического (по модулю), нулевая гипотеза отвергается. Следовательно, различия в средних значениях двух выборок статистически значимы (при заданном α).

Проверка гипотезы о равенстве дисперсий

Так же как и средние значения, можно сравнить степень изменчивости характеристики в двух выборках, т.е. их дисперсию.

Сформулируем нулевую гипотезу $\mathbf{H_0}$: $D_1 = D_2$; альтернативную $\mathbf{H_1}$: $D_1 \neq D_2$.

Для проверки гипотезы используется F-критерий Фишера. Рассчитаем его эмпирическое значение

$$F^* = \frac{D_1}{D_2} \text{ или } \frac{D_2}{D_1} \quad (F^* > 1),$$

где D_1 и D_2 – дисперсии двух выборок соответственно.

0,05; 275; 203

Определим критическое значение критерия Фишера $F_{кр}$ по уровню значимости α=0,05 и числам степеней свободы $\nu_1 = N_1 - 1$ и $\nu_2 = N_2 - 1$, где N_1 и N_2 – длины соответствующих частей ряда.

Сравним эмпирическое и критическое значение критерия. Если эмпирическое значение больше критического, нулевая гипотеза отвергается. Следовательно, различия в дисперсии (т.е. в степени изменчивости) двух выборок статистически значимы (при заданном α).

Понятие о стационарности

Под ***стационарностью*** понимают неизменность характеристики во времени. Соответственно, если характеристика испытывает изменения во времени, то можно говорить о ***нестационарности***.

Нестационарность во времени может быть трех типов:

1) ***нестационарность по математическому ожиданию***, когда среднее значение характеристики за какой-либо период времени значительно отличается от её среднего значения за другой период;

2) ***нестационарность по дисперсии***, когда средняя изменчивость характеристики за какой-либо период времени значительно отличается от средней изменчивости за другой период;

3) ***нестационарность по автокорреляционной функции*** (АКФ), когда в разные периоды времени у характеристики отмечается различная частотная структура.

Одной из первых задач анализа временных рядов является оценка стационарности имеющегося ряда наблюдений, потому что большинство последующих методов анализа требует, чтобы исследуемый ряд был стационарным.

Оценить стационарность можно на основании теории проверки статистических гипотез.

Для этого временной ряд разбивается на части (периоды времени), для каждой из которых отдельно рассчитываются простые статистики (среднее, дисперсия), а потом попарно проводится проверка равенства средних и дисперсии частей рядов (см.выше). Если различия в средних окажутся статистически значимыми, следовательно, ряд нестационарен по мат.ожиданию. Если различия дисперсий окажутся статистически значимыми, ряд нестационарен по дисперсии.

93/С 94 г.

Расчетная часть

Исходные данные

Среднемесячные и среднегодовые значения ТПО в точке 20° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Порядок выполнения работы

Для каждого из двух рядов отдельно:

1. Определить доверительный интервал для математического ожидания.
2. Определить доверительный интервал для дисперсии генеральной совокупности.
3. Преобразовать статистический ряд: разбить его на две части.
4. Проверить гипотезу о равенстве средних значений двух частей ряда. Сделать вывод о стационарности по математическому ожиданию.
5. Проверить гипотезу о равенстве дисперсий двух частей ряда. Сделать вывод о стационарности по дисперсии.
6. Построить график временной изменчивости для ряда среднегодовых значений, на который нанести: отметку разбиения ряда на части, отдельно для каждой части – среднее значение и стандартное отклонение (пример на рис. 3.3).

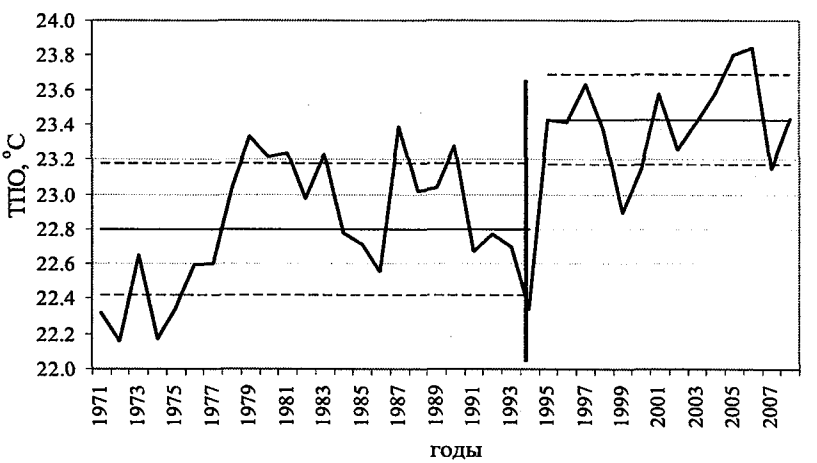


Рис. 3.3. Межгодовой ход ТПО в в точке 20° с.ш. 60° з.д. и оценки среднего и стандартного отклонения за разные периоды времени.

Технология выполнения работы

1. $t_{кр}$ можно рассчитать с помощью функции Excel «стьюдраспобр».

2. χ^2 можно рассчитать с помощью функции Excel «хи2обр».

3. $F_{кр}$ можно рассчитать с помощью функции Excel «Fраспобр».

ПРАКТИЧЕСКАЯ РАБОТА 4

Корреляционный анализ

Теоретическая часть

Между двумя переменными может существовать взаимосвязь, которая бывает ***функциональная*** **и** ***стохастическая*** *(вероятностная).*

Для оценки тесноты и направления связи между изучаемыми переменными пользуются показателем *корреляции.*

Коэффициент корреляции r характеризует степень тесноты **линейной** зависимости.

Линейная зависимость двух случайных величин заключается в том, что при возрастании одной величины другая имеет тенденцию возрастать (или убывать) по линейному закону.

Любую зависимость двух переменных можно представить на графике, где в поле двух координат, соответствующих значениям двух переменных, проставлены точки (рис. 4.1). Совокупность точек образует ***облако точек.***

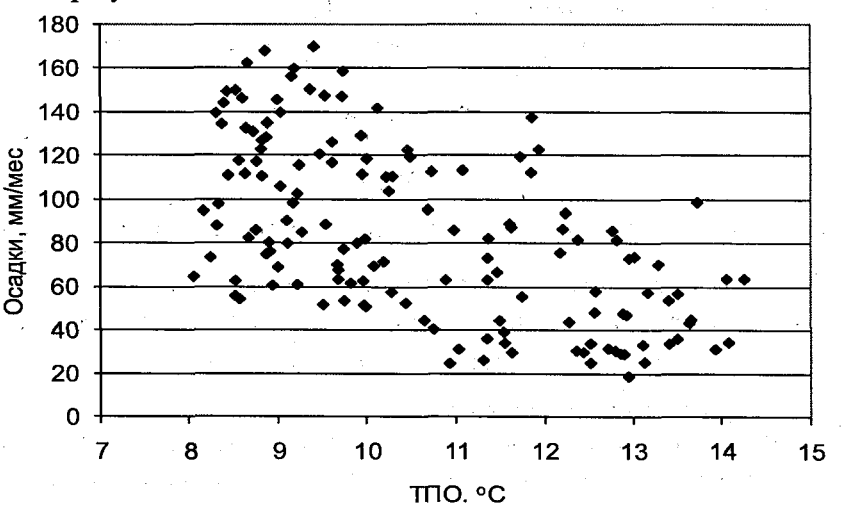


Рис. 4.1. График связи осадков и ТПО в точке 55° с.ш. 60° з.д. за период с 1971 по 1985 гг.

В общем случае облако точек представляет собой ***эллипс рассеяния***, большая ось которого расположена под наклоном к осям. Эксцентриситет эллипса (соотношение большой и малой осей) отражает качество зависимости между двумя переменными.

Если эллипс вырожден в прямую линию (присутствует только большая ось), связь между переменными является ***функциональной***, т.е. одному значению первой переменной соответствует одно и только одно значение второй переменной. В этом случае $|r| = 1$.

Если эллипс вырожден в круг (большая и малая ось эллипса равны), связь между переменными является ***абсолютно случайной***, т.е. одному значению первой переменной соответствует любое значение второй переменной, и $|r| = 0$.

В остальных случаях связь является ***стохастической***, т.е. одному значению первой переменной соответствует некоторое значение второй переменной с определенной вероятностью, $0 < |r| < 1$.

Характер связи определяется по соотношению значений двух переменных. Если при возрастании одной величины другая имеет тенденцию возрастать – это *прямая зависимость*, тогда $r > 0$. Если при возрастании одной величины другая имеет тенденцию убывать – это *обратная зависимость*, тогда $r < 0$.

Коэффициент корреляции может быть рассчитан по формуле:

$$r = \frac{\sum_{i=1}^{N}\left[(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]}{N\sigma_1\sigma_2},$$

где x_{i1} и x_{i2} – значения первой и второй переменных соответственно; $\bar{x}_1$ и $\bar{x}_2$ – средние значения первой и второй переменных соответственно; σ_1 и σ_2 – стандартные отклонения первой и второй переменных соответственно; N – длина рядов первой и второй переменных.

Корреляционная матрица

Если необходимо рассчитать коэффициенты корреляции для нескольких переменных (больше двух) во всех сочетаниях друг с другом, то набор получившихся коэффициентов корреляции можно записать в виде матрицы, которая называется ***корреляционной***.

Например, для четырех переменных *T, S, P, Si* – 16 коэффициентов корреляции представлены в виде корреляционной матрицы ***R***:

	T	*S*	*P*	*Si*
T	1,0	r_{12}	r_{13}	r_{14}
S	r_{21}	1,0	r_{23}	r_{24}
P	r_{31}	r_{32}	1,0	r_{34}
Si	r_{41}	r_{42}	r_{43}	1,0

Если связь стохастическая, то рассчитанная величина *r* может быть большой или маленькой, что отражает *степень связи*. Чтобы ее оценить, необходимо выполнить **проверку коэффициента корреляции на значимость.**

Сформулируем нулевую гипотезу H_0: $r = 0$; альтернативную H_1: $r \neq 0$.

Для проверки этой гипотезы выбирается критерий Стьюдента, выборочное значение которого рассчитывается по формуле:

$$t^* = \frac{|r|}{\sigma_r}, \quad \sigma_r = \frac{1-r^2}{\sqrt{N-2}},$$

где σ_r – средняя квадратическая погрешность расчета коэффициента корреляции.

Далее определяется критическое значение $t_{кр}$ (α, ν), где уровень значимости α принимается равным 5%, а число степеней свободы $\nu = N-2$, где *N*–длина ряда.

Сравниваем t^* с $t_{кр}$.

Если $t^* > t_{кр}$ нулевая гипотеза отвергается, коэффициент корреляции *значим*, т.е. между двумя переменными существует статистически значимая прямая (или обратная, в зависимости от знака) связь.

Если $t^* < t_{кр}$ предполагается, что нет оснований отвергнуть нулевую гипотезу, т.е. коэффициент корреляции *незначим*, т.е. между двумя переменными статистически значимая линейная связь отсутствует.

Примечание: Всегда нужно помнить, что наличие значимой корреляции между двумя переменными не определяет причинно-следственных отношений между ними, т.е. не значит, что одна переменная зависит от другой!

Расчетная часть

Исходные данные

1. Среднемесячная и среднегодовая ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

2. Среднемесячные и среднегодовые значения гидрометеорологических характеристик в широтной зоне умеренных широт Северной Атлантики (45 – 60° ю.ш.) за период с 1971 по 1985 гг.

Порядок выполнения работы

Для каждого из рядов среднемесячных и среднегодовых значений характеристик отдельно:

1. Рассчитать корреляционную матрицу для всей группы гидрометеорологических характеристик, включая ТПО.

2. Все коэффициенты корреляции в матрице проверить на значимость. Значимые коэффициенты корреляции в матрице выделить (цветом, шрифтом и т.п.).

3. Сделать вывод о наличии линейных статистических связей ТПО с другими характеристиками, указать направление связей, охарактеризовать существующие связи с точки зрения физических процессов.

4. Построить два графика связи: 1) ТПО и другая переменная – с максимальным коэффициентом корреляции (по модулю); 2) ТПО и другая переменная – с минимальным коэффициентом корреляции (по модулю).

Технология выполнения работы

1. Расчет корреляционной матрицы:

– подготовить матрицу исходных данных, где столбцы представляют собой различные гидрометеорологические характеристики, включая ТПО, меняющиеся со временем (по строкам) синхронно;

– в меню «Сервис» выбрать «Анализ данных» и далее модуль «Корреляция»;

– в качестве «входного интервала» выбрать *всю* подготовленную матрицу;

– «группирование» – по столбцам;

– нажать «ОК».

2. $t_{кр}$ можно рассчитать с помощью функции Excel «стьюдраспобр».

ПРАКТИЧЕСКАЯ РАБОТА 5

Парная линейная регрессия

Теоретическая часть

На графике связи между двумя переменными (рис. 5.1) облако точек можно аппроксимировать прямой линией вдоль большой оси эллипса рассеяния (см. работу 4). Тогда для этих двух переменных можно сформулировать ***модель линейной регрессии***. Например, для рис. 5.1:

$$Tw_{XII} = a\,Tw_{XI} + b + \varepsilon,$$

где a и b – коэффициенты регрессии; ε – отклонения от прямой.

В общем виде ***уравнение линейной регрессии*** может быть записано как

$$y = ax + b + \varepsilon,$$

здесь y – называется ***зависимой переменной***; x – ***независимой переменной***; a – ***коэффициент регрессии***; b – ***свободный член***.

Коэффициент регрессии ***a*** представляет собой тангенс угла наклона линии регрессии к оси абсцисс и определяется по формуле:

$$a = r\frac{\sigma_y}{\sigma_x},$$

где r – коэффициент корреляции переменных, входящих в регрессию; σ_y и σ_x – стандартные отклонения зависимой и независимой переменной соответственно.

Свободный член ***b*** представляет собой расстояние от начала координат до пересечения оси ординат уравнением регрессии и рассчитывается:

$$b = \bar{y} - a \cdot \bar{x},$$

где a – коэффициент регрессии; $\bar{y}$ и $\bar{x}$ – средние значения зависимой и независимой переменных соответственно.

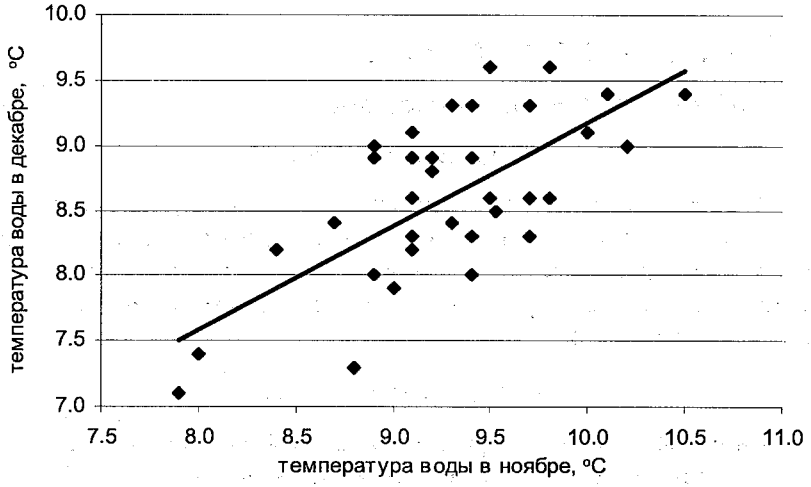


Рис. 5.1. График связи ТПО в декабре и ноябре
за период с 1971 по 1985 гг. в точке 50° с.ш. 60° з.д.

Качество модели линейной регрессии определяется по ***параметрам (характеристикам) модели линейной регрессии:***

1. Коэффициенты регрессии и их значимость.

Оценка значимости коэффициентов регрессии производится на основе проверки гипотезы:

$\boldsymbol{H_0}$: $\boldsymbol{a} = 0$; $\boldsymbol{H_0}$: $\boldsymbol{b} = 0$;

$\boldsymbol{H_1}$: $\boldsymbol{a} \neq 0$; $\boldsymbol{H_1}$: $\boldsymbol{b} \neq 0$.

Для проверки рассчитывается критерий Стьюдента:

$$t_a^* = \frac{|a|}{\sigma_a}; \quad t_b^* = \frac{|b|}{\sigma_b}, \quad \sigma_a = \frac{\sigma_y(1-r^2)}{\sigma_x(\sqrt{N-1})}, \quad \sigma_b = \frac{\sigma_y\sqrt{1-r^2}}{\sqrt{N-1}}.$$

где σ_a и σ_b – стандартные случайные погрешности расчета коэффициентов $\boldsymbol{a}$ и $\boldsymbol{b}$; r – коэффициент корреляции переменных, входящих в регрессию; σ_y и σ_x – стандартные отклонения зависимой и независимой переменной соответственно; N – длина ряда.

Критическое значение $t_{кр}$ (α,ν) определяется по уровню значимости α (принимается равным 5%) и числу степеней свободы $\nu = N - \mathbf{2}$, где N – длина ряда.

Сравниваем t^* с $t_{кр}$.

Если $t^* > t_{кр}$ нулевая гипотеза отвергается, соответствующий коэффициент регрессии значим.

Если $t^* < t_{кр}$ нет оснований отвергнуть нулевую гипотезу, соответствующий коэффициент регрессии незначим.

2. *Коэффициент детерминации* r^2 показывает долю дисперсии исходного ряда, которая описывается моделью регрессии и равен квадрату коэффициента корреляции.

3. *Адекватность регрессионной модели исходным данным.*

Для оценки адекватности необходимо выполнить следующие расчеты:

1) для каждого момента времени определить *вычисленные по уравнению регрессии значения* температуры воды $(\hat{y})$;

2) рассчитать ***дисперсию модели*** $\hat{y}$, характеризующую изменчивость линии регрессии относительно среднего значения $\bar{y}$

$$D_{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2;$$

3) для каждого момента времени рассчитать ***остатки регрессии***

$$\varepsilon_i = y_i - \hat{y}_i;$$

4) рассчитать ***дисперсию остатков*** ε, характеризующую отклонение уравнения регрессии от результатов наблюдений y

$$D_{\varepsilon} = \frac{1}{N-2}\sum_{i=1}^{N}(\varepsilon_i^{\,2});$$

5) оценить ***адекватность*** регрессионной модели. Для этого выдвигаем нулевую гипотезу о равенстве дисперсий $\boldsymbol{H_0}$: $D_{\hat{y}} = D_{\varepsilon}$ и альтернативную ей гипотезу $\boldsymbol{H_1}$: $D_{\hat{y}} \neq D_{\varepsilon}$. Для проверки используется критерий Фишера $F^* = \frac{D_{\hat{y}}}{D_{\varepsilon}} \cdot N$,

который сравнивается с $\boldsymbol{F}_{кр}$ при заданном уровне значимости α (α = 0,05) и степенях свободы $\nu_1 = 1$, $\nu_2 = N - 2$. Если $F^* > \boldsymbol{F}_{кр}$, то нулевая гипотеза о равенстве дисперсий отвергается, что означает в рассматриваемом случае адекватность регрессионной модели.

4. ***Стандартная (среднеквадратическая) ошибка модели*** $\sigma_\varepsilon = \sqrt{D_\varepsilon}$.

Модель считается качественной, если выполняются следующие условия:

1. Все коэффициенты регрессии значимы.

2. Коэффициент детерминации больше 0,70. Это свидетельствует о том, что независимых переменных *достаточно* для описания дисперсии исходного ряда.

3. Модель должна быть адекватна.

4. Стандартная ошибка модели σ_ε должна быть меньше стандартного отклонения ряда зависимой переменной σ_y.

Расчетная часть

Исходные данные

1. Среднемесячная и среднегодовая ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

2. Среднемесячные и среднегодовые значения гидрометеорологических характеристик в широтной зоне умеренных широт Северной Атлантики (45 – 60° ю.ш.) за период с 1971 по 1985 гг.

Порядок выполнения работы

1. В корреляционной матрице (см. работу 4) выбрать наибольший (по модулю) коэффициент корреляции для ТПО. Выбрать из исходных данных два соответствующих ему ряда (один из них – ТПО).

2. Определить по смыслу зависимую и независимую переменные.

3. Рассчитать коэффициенты регрессионной модели (см. технологию выполнения работы, п.1).

4. Определить основные параметры модели линейной регрессии модели (см. технологию выполнения работы, п.1):

– коэффициент детерминации r^2;

– стандартную ошибку модели σ_ε;

– расчетный критерий Фишера F^*.

5. Оценить (см. технологию выполнения работы, п.2):

– значимость коэффициентов регрессии;

– адекватность модели по критерию Фишера.

6. Сделать анализ регрессионной модели (пример анализа см. ниже).

7. Построить совмещенный график фактических y и вычисленных по уравнению регрессии $\hat{y}$ значений температуры воды (рис. 5.2).

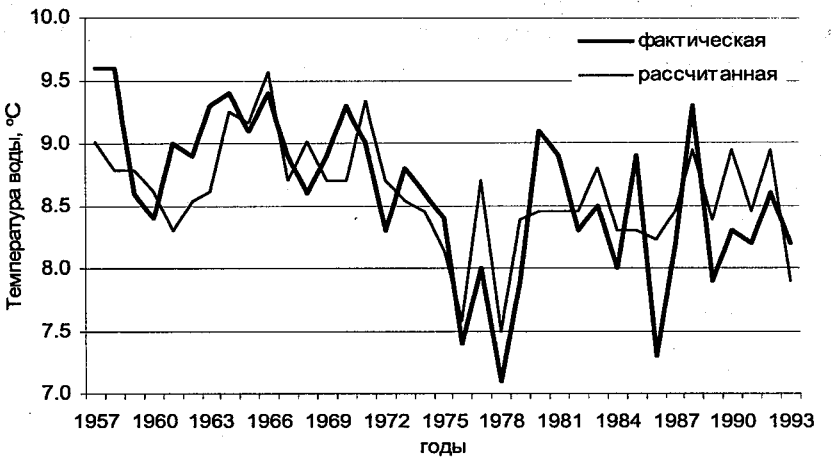


Рис. 5.2. Фактические и рассчитанные по уравнению регрессии значения ТПО в точке 60° с.ш. 60° з.д.

Технология выполнения работы

1. Все характеристики регрессии можно рассчитать в пакете «анализ данных» в модуле «регрессия».

2. $t_{кр}$ можно рассчитать с помощью функции Excel «стьюдраспобр».

$F_{кр}$ можно рассчитать с помощью функции Excel «Fраспобр».

Пример анализа регрессионной модели

Для рядов среднемесячных значений ТПО и ветра было получено уравнение регрессии

$$\text{ТПО} = 0{,}35W + 5{,}64 + \varepsilon.$$

Коэффициенты регрессии были проверены на значимость при уровне значимости 0.05 и все оказались значимыми ($t_a = 5{,}62$; $t_b = 8{,}51$; $t_{кр} = 1{,}96$). Коэффициент детерминации показал, что моделью описывается 82,3% дисперсии исходного ряда и свидетель-

ствует о том, что данных о ветре (независимых переменных) достаточно для описания изменчивости ТПО (зависимой переменной). Проверка адекватности по критерию Фишера при уровне значимости 5% показала, что модель адекватна ($F^* = 156{,}2$ при $F_{кр} = 4{,}01$). Стандартная ошибка модели σ_ε составила 0,35 °С, что меньше чем стандартное отклонение ряда ТПО ($\sigma_y = 1{,}63$ °С). Таким образом, очевидно, что полученная модель линейной регрессии имеет хорошее качество.

ПРАКТИЧЕСКАЯ РАБОТА 6

Нелинейная регрессия

Теоретическая часть

Иногда на графике связи между двумя переменными очевидно, что облако точек располагается не вдоль прямой линии, а вдоль кривой линии (т.е. представляет собой проекцию не «огурца», а «банана»), тогда его нужно аппроксимировать ***нелинейной*** (т.е. криволинейной) функцией.

Вообще нелинейных функций существует огромное множество и заранее трудно сказать, какая из них будет отражать связь между двумя переменными наилучшим образом. Поэтому, как правило, рассчитывается несколько различных моделей и из них выбирается наилучшая.

Эта задача наименее трудоемко решается в рамках ***алгоритма линеаризации*** нелинейных моделей:

1. Формулируется нелинейная модель в общем виде;
2. Путем математических преобразований она преобразуется в линейную модель.
3. Находятся все характеристики модели линейной регрессии, в том числе коэффициенты (см. работу 5).
4. Путем обратных преобразований эти коэффициенты преобразуются в коэффициенты нелинейной модели и на их основе формулируется конкретное уравнение нелинейной регрессии.
5. Оценивается качество модели нелинейной регрессии.

Примеры.

Квадратичная модель

1. В общем виде: $y = ax^2 + b + \varepsilon$.

2. Преобразуется в линейную модель путем замены $x^2 = x^*$, получим $y = ax^* + b + \varepsilon$.

3. Находим коэффициенты a и b линейной модели. Например, получим $a = 0{,}16$, $b = 3{,}58$.

4. Так как в п.2. никаких преобразований для коэффициентов не производилось, то и обратное преобразование не производится. Следовательно модель нелинейной регрессии будет формулироваться как $y = 0{,}16x^2 + 3{,}58 + \varepsilon$.

Экспоненциальная модель

1. В общем виде формулируется $y = b \cdot e^{ax} + \varepsilon$.

2. Для преобразования в линейную модель данное уравнение необходимо логарифмировать: $ln(y) = ln(b) + ax + \varepsilon$. Произведем замену $ln(y) = y^*$, $ln(b) = c$.

Тогда линеаризованное уравнение: $y^* = ax + c + \varepsilon$.

3. Находим коэффициенты a и c линейной модели. Например, получим $a = 0{,}21$, $c = 1{,}23$.

4). Сделаем обратное преобразование. Коэффициент a в п.2 не преобразовывался, поэтому не изменяется. Чтобы получить коэффициент b: $b = e^c$.

Тогда модель нелинейной регрессии будет формулироваться как $y = 3{,}42 \cdot e^{0{,}21x} + \varepsilon$..

Степенная модель

1. В общем виде формулируется $y = b \cdot x^a + \varepsilon$.

2. Для преобразования в линейную модель данное уравнение необходимо логарифмировать: $ln(y) = ln(b) + a\,ln(x)$. Произведем замену $ln(y) = y^*$, $ln(b) = d$, $ln(x) = x^*$. Тогда линеаризованное уравнение: $y^* = ax^* + d + \varepsilon$.

3. Находим коэффициенты a и d линейной модели. Например, получим $a = 0{,}46$, $d = 2{,}12$.

4). Сделаем обратное преобразование. Коэффициент a в п.2 не преобразовывался, поэтому не изменяется. Чтобы получить коэффициент b: $b = e^d$.

Тогда модель нелинейной регрессии будет формулироваться, например, $y = 8{,}33 \cdot x^{0{,}46} + \varepsilon$.

Основные ***характеристики качества*** нелинейной модели – это аналог коэффициента детерминации и среднеквадратическая ошибка.

Аналог коэффициента детерминации *численно* равен коэффициенту детерминации для линеаризованного уравнения. Квадратный корень из него является аналогом коэффициента корреляции и называется «***корреляционное отношение»:***

$$\eta = \sqrt{\frac{D_{\tilde{y}}}{D_y}},$$

где $D_{\tilde{y}}$ – дисперсия модели; D_y – дисперсия исходного зависимого ряда.

Таким образом, для оценки *линейной* связи двух переменных используется коэффициент корреляции, а для оценки *нелинейной связи* двух переменных – корреляционное отношение.

Так как линейная связь является частным случаем нелинейной, величина корреляционного отношения больше (или равна) величины коэффициента корреляции. Чем больше разница между коэффициентом корреляции и корреляционным отношением, тем больше степень нелинейности в зависимости двух переменных.

Другой характеристикой качества нелинейной модели является среднеквадратическая ошибка:

$$\sigma_\varepsilon = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\varepsilon_i^{\,2})}.$$

Чтобы сделать вывод о хорошем качестве модели, необходимо, чтобы корреляционное отношение достаточно велико, а ошибка достаточно мала.

Расчетная часть

Исходные данные

Значения удельной влажности q и температуры воздуха T_a в январе 1951 г. на меридиональном разрезе вдоль 180° долготы в Тихом океане.

Порядок выполнения работы

1. Сформулировать (написать уравнение) в общем виде модель линейной регрессии зависимости удельной влажности q от T_a.

2. Сформулировать в общем виде модели нелинейной регрессии и линеаризовать их для следующих зависимостей:

– квадратичной;

– экспоненциальной;

– степенной.

3. Подготовить исходные данные: для каждой модели сформировать ряды зависимой и независимой переменных (см. технологию выполнения работы,п.1).

4. Для каждой из 4-х моделей рассчитать все характеристики: коэффициенты модели, коэффициент детерминации или корреляционное отношение, стандартную ошибку модели (см. технологию выполнения работы, п.2).

5. Для каждой модели:

– написать уравнения регрессии в линеаризованном и нелинейном видах (с реальными коэффициентами);

– по этим уравнениям рассчитать вычисленные значения влажности;

– построить график связи влажности q с T_a. Нанести на него вычисленные значения влажности. См.пример на рис. 6.1

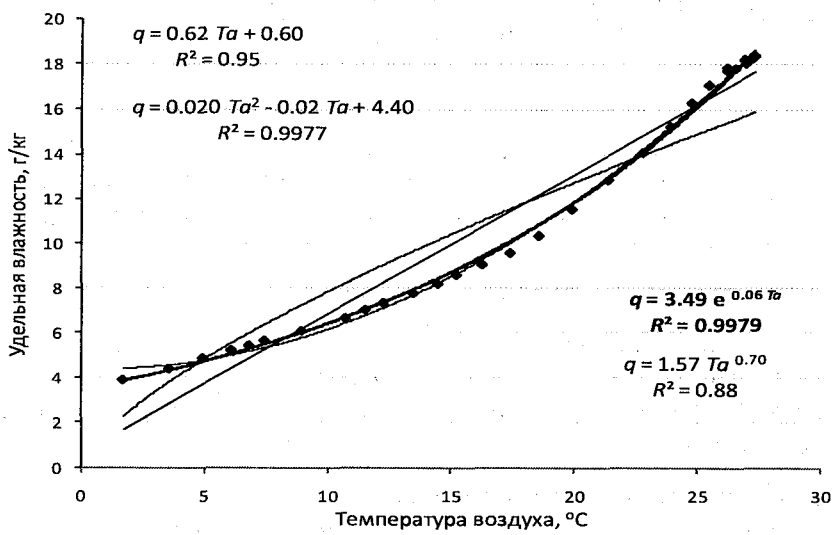


Рис. 6.1. Фактические и рассчитанные по различным моделям значения удельной влажности в январе 1951 г. на меридиональном разрезе вдоль 180° долготы в Тихом океане.

6. Выбрать наилучшую модель. Для нее построить график вычисленных и фактических значений в зависимости от широты.

Технология выполнения работы

1. Для последущих расчетов нужно создать дополнительные ряды:

– для квадратичной модели: ряд x^2;

– для экспоненциальной модели: ряд *ln(y)*;

– для квадратичной модели: ряды *ln(x)* и *ln(y)*

2. Для определения всех характеристик моделей нужно воспользоваться модулем «регрессия» пакета «анализ данных», где в качестве зависимой переменной будет выступать преобразованные ряды *y*, а в качестве независимой переменной – преобразованные ряды *x*.

Тогда коэффициент детерминации полученной линейной модели будет тождественен корреляционному отношению соответствующей нелинейной модели, стандартная ошибка будет та же, а коэффициенты нелинейной модели получатся обратным преобразованием (см. теоретическую часть, примеры).

ПРАКТИЧЕСКАЯ РАБОТА 7

Множественная линейная регрессия

Теоретическая часть

Чаще всего, любая физическая характеристика связана не с одной, а с *несколькими* другими физическими характеристиками. Тогда для определения этой связи можно рассчитать ***множественную линейную регрессию (МЛР).***

Тогда общее уравнение связи переменных может быть сформулировано в виде:

$$Y = a_1X_1 + a_2X_2 + \ldots a_mX_m + b + \varepsilon,$$

где Y – зависимая переменная; X_i – i-тая независимая переменная ($i = 1 \div m$); a_i – коэффициент регрессии при i-той переменной; b – свободный член; ε – остатки модели.

Y и X_i формируют в виде ***матрицы исходных данных***: $(m+1)$столбцов на N строк.

Нахождение коэффициентов линейной регрессии методом МНК требует решения системы из m линейных уравнений с m неизвестными. Это является достаточно трудоемкой процедурой и хорошо проработано в специализированных статистических программных пакетах.

Качество модели множественной линейной регрессии, так же как и для простой линейной регрессии (см. работу 5), определяется по ***параметрам (характеристикам) МЛР.***

1. Коэффициенты регрессии и их значимость

Оценка значимости коэффициентов регрессии производится на основе проверки гипотезы H_0: $a_i = 0$; $b = 0$ при H_1: $a_i \neq 0$; $b \neq 0$.

Для проверки рассчитываются критерии Стьюдента:

$$t^*_{ai} = \frac{|a_i|}{\sigma_{ai}};\ t^*_b = \frac{|b|}{\sigma_b},$$

где σ_{ai} и σ_b – стандартные случайные погрешности расчета коэффициентов a_i и b (рассчитываются совместно с расчетом коэффи-

циентов).

Критическое значение $t_{кр}$ (σ,, ν) определяется по уровню значимости σ (принимается равным 5%) и числу степеней свободы $\nu=N-m-1$, где N–длина ряда, m – количество независимых переменных.

Сравниваем t^* с $t_{кр}$ (по модулю).

Если $t^* > t_{кр}$ нулевая гипотеза отвергается, соответствующий коэффициент регрессии значим.

Если $t^* < t_{кр}$ – нет оснований отвергнуть нулевую гипотезу, соответствующий коэффициент регрессии незначим.

2. Адекватность регрессионной модели

Для оценки адекватности необходимо выполнить следующие расчеты:

– для каждого момента времени определить *вычисленные по уравнению регрессии значения* температуры воды ($\hat{y}$);

– рассчитать ***дисперсию рассчитанных по модели*** $\hat{y}$, характеризующую изменчивость линии регрессии относительно среднего значения $\bar{y}$:

$$D_{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2;$$

– для каждого значения ряда рассчитать ***остатки регрессии:***

$$\varepsilon_i = y_i - \hat{y}_i;$$

– рассчитать ***дисперсию остатков*** ε, характеризующую отклонение уравнения регрессии от результатов наблюдений y:

$$D_{\varepsilon} = \frac{1}{N-m-1}\sum_{i=1}^{n}(\varepsilon_i^{\,2});$$

– оценить ***адекватность*** регрессионной модели. Для этого выдвигаем нулевую гипотезу о равенстве дисперсий $\boldsymbol{H_0}$: $D_{\hat{y}} = D_{\varepsilon}$. Альтернативную $\boldsymbol{H_1}$: $D_{\hat{y}} \neq D_{\varepsilon}$. Для проверки используется критерий Фишера $F^* = \frac{D_{\hat{y}}}{D_{\varepsilon}} \cdot \frac{N}{m}$, который сравнивается с $\boldsymbol{F}_{кр}$ при заданном уровне значимости α (α = 0,05) и степенях свободы $\nu_1 = m$,

$\nu_2 = N - m - 1$. Если $F^* > F_{кр}$, то нулевая гипотеза о равенстве дисперсий отвергается, что означает в рассматриваемом случае адекватность регрессионной модели.

3. Коэффициент детерминации R^2

R^2 показывает долю дисперсии исходного ряда, которая описывается моделью регрессии и представляет собой квадрат коэффициента множественной корреляции.

Коэффициент множественной корреляции $R = \sqrt{\frac{D_y}{D_y}}$

4. Стандартная ошибка модели $\sigma_\varepsilon = \sqrt{D_\varepsilon}$

Модель считается качественной, если выполняются следующие условия:

1. Все коэффициенты регрессии значимы.
2. Коэффициент детерминации больше 0,70. Это свидетельствует о том, что независимых переменных достаточно для описания дисперсии исходного ряда.
3. Модель должна быть адекватна по F-критерию.
4. Стандартная ошибка модели должна быть меньше стандартного отклонения ряда зависимой переменной ***y***.

Пошаговый алгоритм

Как правило, если независимых переменных много (больше 5), то полная модель МЛР (в которой учитываются все независимые переменные), несмотря на возможное хорошее качество, будет достаточно громоздкой. Тогда требуется выполнить поиск моделей МЛР приемлемого качества, но с меньшим количеством предикторов.

В частности, эта задача выполняется в рамках ***пошагового алгоритма***.

Рассмотрим алгоритм пошаговой регрессии ***методом исключения***.

Шаг 1.

1. Рассчитывается полная модель МЛР (коэффициенты регрессии и все характеристики качества).
2. Находится самый малый (по модулю) критерий Стьюдента t^* для оценки значимости коэффициентов регрессии.

3. Определяется независимая переменная, которая имеет этот коэффициент регрессии.

Шаг 2.

1. Из матрицы исходных данных убирается столбец, определенный в п.3 предыдущего шага.

2. МЛР пересчитывается с числом независимых переменных, меньше на одну. Определяются новые коэффициенты модели и характеристики качества.

3. Находится самый малый (по модулю) критерий Стьюдента t^* для оценки значимости коэффициентов регрессии.

4. Определяется независимая переменная, которая имеет этот коэффициент регрессии.

Шаг 3. Повторить шаг 2.

Это продолжается до тех пор, пока не останется одна независимая переменная. Таких шагов будет m.

Качество модели рассматривается *на каждом шаге,* и из m моделей выбирается наилучшая. Наилучшей считается та, у которой: 1) наименьшая стандартная ошибка модели; 2) все коэффициенты модели значимы; 3) коэффициент детерминации больше 0,70.

Расчетная часть

Исходные данные

1. Среднемесячная и среднегодовая ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

2. Среднемесячные и среднегодовые значения гидрометеорологических характеристик в широтной зоне умеренных широт Северной Атлантики (45 – 60° ю.ш.) за период с 1971 по 1985 гг.

Порядок выполнения работы

1. Рассчитать полную модель МЛР зависимости ТПО от других гидрометеорологических характеристик и все ее параметры (см. технологию выполнения работы, п.1).

2. С помощью пошаговой регрессии методом исключения переменных рассчитать все остальные модели МЛР и их параметры.

3. Сформировать таблицу, в которой указать характеристики качества всех моделей и их критические значения (табл. 7.1). Значения характеристик, отражающие *хорошее* качество модели, в таблице выделить.

4. Выбрать оптимальную модель связи ТПО с другими гидрометеорологическими характеристиками. Доказать ее оптимальность.

5. Для оптимальной модели рассчитать модельные значения ТПО и построить график временной изменчивости фактических и модельных значений ТПО.

Технология выполнения работы.

1. Все коэффициенты регрессии и характеристики модели можно найти с помощью надстройки Excel «Пакет анализа» (Меню «сервис» – «Анализ данных» – «регрессия»).

2. На каждом шаге пошагового моделирования в отдельную таблицу (например, табл. 7.1) выписываются характеристики качества полученных моделей. Из анализа таблицы можно выбрать модель оптимального качества.

Таблица 7.1

Характеристики моделей множественной линейной регрессии, полученные пошаговым моделированием методом исключения переменных

Шаг модели	Кол-во независимых переменных	t^*_{min}	$t_{кр}$	F^*	$F_{кр}$	R^2	σ_ε	σ_y
1 2 ...								

480 − 1 − 6 = 473

ПРАКТИЧЕСКАЯ РАБОТА 8

Анализ тренда временного ряда

Теоретическая часть

Для применения большинства методов анализа временного ряда одним из основных требований к ряду является его *стационарность,* т.е. неизменность его основных статистических характеристик во времени. В частности, это касается его среднего значения и дисперсии. Поэтому на первом этапе анализа временного ряда оценивается его стационарность, и если она не выявляется, ряд преобразовывают к стационарному виду.

В случае нестационарности ряда среднее значение и/или дисперсия для частей выборки могут меняться скачкообразно (см. пример в работе 3), а могут иметь непрерывный характер изменения. В последнем случае говорят, что *ряд имеет тренд* (по математическому ожиданию или по дисперсии).

С другой стороны, в генеральной совокупности (ГС) может существовать некоторое колебание, и, выбирая из ГС выборку, мы можем это колебание не захватить всей длиной выборки (рис. 8.1). Тогда оно будет отражаться в нашей выборке как тренд.

Таким образом, ***тренд*** отражает наличие во временной изменчивости исследуемой характеристики длиннопериодного колебания с периодом, существенно превышающим длину выборки.

Исходя из этого соображения, тренд может быть линейным (рис.8.1 выборки 1,3) или квадратичным (рис.8.1, выборки 2,4), в зависимости от того, на какую часть длиннопериодного колебания попала выборка.

Сформулируем уравнение ***линейного тренда***:

$$y = a_1 t + a_0 + \varepsilon ,$$

и нелинейного (квадратичного) тренда:

$$y = a_2 t^2 + a_1 t + a_0 + \varepsilon ,$$

где t – время.

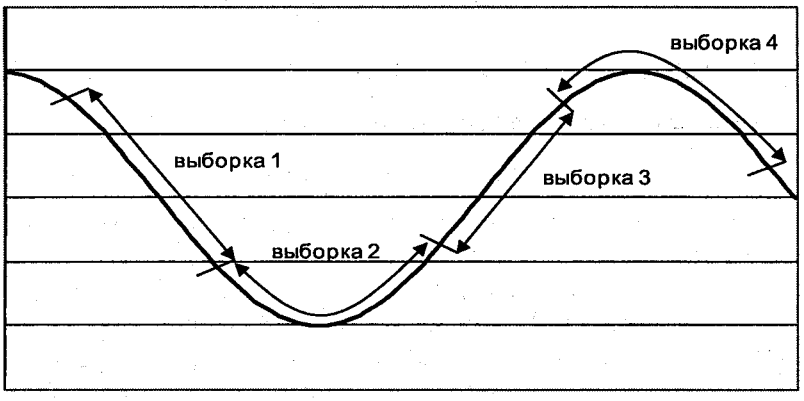


Рис. 8.1. Генеральная совокупность и выборки из нее

Основными характеристиками тренда являются:

1. ***Коэффициент детерминации*** r^2 (для линейного тренда) или его нелинейный аналог η^2, характеризующий ***вклад тренда*** в общую дисперсию ряда. Вклад может быть *значительным* или *незначимым*. Это определяется на основании проверки на значимость коэффициента корреляции r или η.

В случае их незначимости считается, что *тренда нет*.

2. ***Величина тренда*** – изменение характеристики по *линейному* тренду за определенный промежуток времени. Для среднегодовых рядов, как правило, величина тренда рассчитывается за 10 лет, для среднемесячных – за год.

Величина тренда равна коэффициенту a_1 линейного тренда и имеет размерность характеристики y за единицу дискретности. Например, для ряда среднегодовых значений $a_1 = 0{,}07$ °С/год. Тогда величина тренда $\boldsymbol{Tr} = 0{,}7$ °С/10 лет.

Если в исследуемой выборке и линейный и нелинейный тренды значимы, тогда при анализе предпочтение отдают нелинейному тренду, если он вносит значительно (более чем на 5%) больший вклад в дисперсию выборки, или линейному – в обратном случае.

Для дальнейшего статистического анализа из временного ряда ***тренд должен быть удален.*** Это решается простым вычитанием из исходного ряда значений, рассчитанных по тренду, т.е.

$$dy_i = y_i - (a_1 t_i + a_0) \quad \text{или} \quad dy_i = y_i - (a_2 t^2_i + a_1 t_i + a_0)$$

Расчетная часть

Исходные данные

Среднемесячные и среднегодовые значения ТПО в точке 20° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Порядок выполнения работы

1. Сформировать дополнительный исходный ряд времени $t_i = i$, где $i = 1, 2, 3 \ldots N$, где N – длина исходной реализации, а также ряд t^2_i

2. Рассчитать характеристики линейного тренда как линейной регрессии (работа 5): коэффициенты регрессии, коэффициент детерминации, коэффициент корреляции. Рассчитать характеристики нелинейного тренда как множественной регрессии (работы 6, 7): коэффициенты регрессии и корреляционное отношение, его квадрат.

3. Оценить значимость каждого вида тренда путем оценки значимости коэффициента корреляции r или корреляционного отношения η (работа 4).

Если r значим, это означает, что тренд неслучайным образом отличается от нуля и вносит определенный вклад в формирование изменчивости исходного ряда. Если r не значим – тренда нет.

4. Выявить, какой из трендов (линейный или нелинейный) предпочтительнее.

5. Рассчитать величину тренда, если предпочтен линейный тренд.

6. Построить график временного хода исходного ряда. На график нанести рассчитанные значения и линейного и нелинейного трендов. Пример на рис. 8.2.

7. Проанализировать полученные результаты. Привести уравнения трендов, оценку их значимости. Оценить их вклад в дисперсию исходного ряда. Указать характер тренда (положительный или отрицательный, т.е. рост или падение температуры воды) и возможные физические причины его формирования. Привести величину тренда.

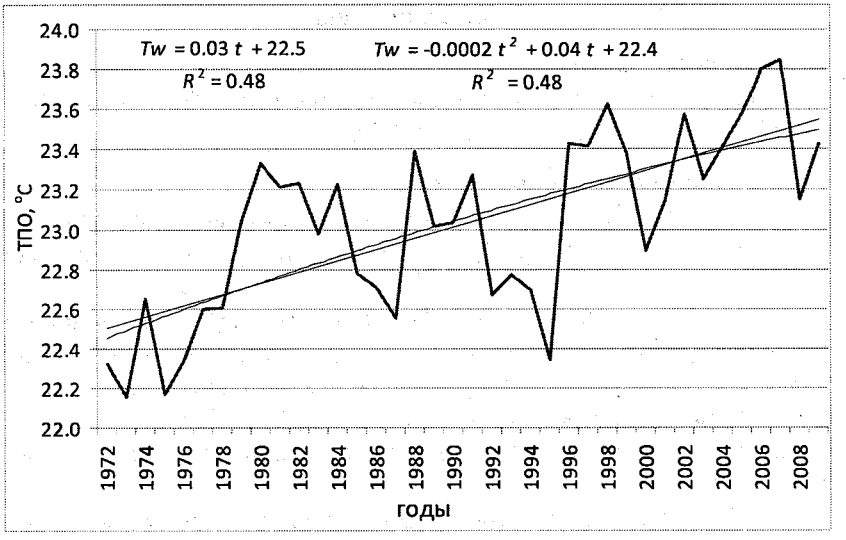


Рис. 8.2. Межгодовая изменчивость ТПО в точке 20° с.ш. 60° з.д. и ее линейный и нелинейный тренды.

ПРАКТИЧЕСКАЯ РАБОТА 9

Автокорреляционный анализ и авторегрессия 1 порядка

Теоретическая часть

Анализ внутренней структуры исследуемого процесса проводится на основе ***автокорреляции***.

Автокорреляция – это корреляция статистического ряда самого с собой при разных сдвигах во времени.

Например (рис. 9.1) взят исходный ряд X. Можно сделать его точную копию – ряд X_0 и рассчитать между парой этих рядов коэффициент корреляции r_0. Теперь сдвинем ряд X_0 относительно ряда X на одно число (ряд X_1) и уравняем длины рядов X и X_1, обрезав на одно число ряд X с начала, а ряд X_1 – с конца. И снова рассчитаем коэффициент корреляции между этими рядами – r_1. И так можно сдвигать ряды далее, на каждом сдвиге уравнивая их длины и рассчитывая коэффициенты корреляции.

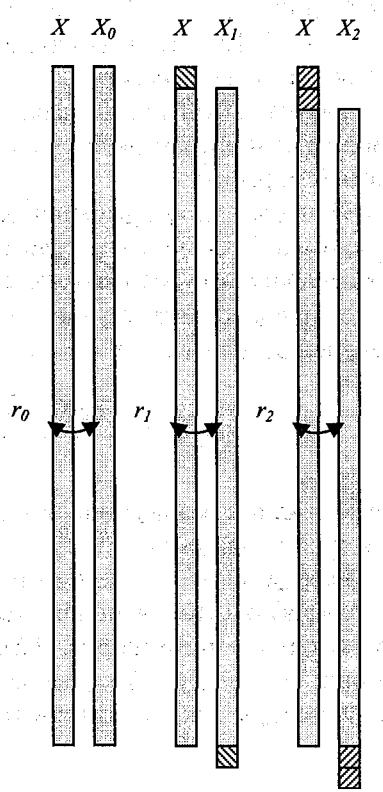


Рис. 9.1. Схема расчета коэффициентов автокорреляции

Если все коэффициенты корреляции нанести на график зависимости r от τ (сдвига) и соединить кривой, получим ***автокорреляционную функцию*** (АКФ) (рис. 9.2).

Временной ряд при расчетах можно сдвигать в любую

сторону (вперед или назад). Но, так как исходный ряд один и тот же $r_1 = r_{-1}$, т.е. автокорреляционная функция симметрична.

Коэффициент автокорреляции r(τ) на каждом сдвиге также можно рассчитать по формуле

$$r(\tau) = \frac{1}{\sigma_x^2 (N-1-\tau)} \sum_{i=1}^{N-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x}),$$

где N – длина реализации, τ – сдвиг, который меняется от 1 до максимума.

Максимальное количество сдвигов (длина автокорреляционной функции) зависит от длины ряда N: если ряд короткий (N порядка 30) $\tau_{max} = N/3$; если ряд длинный (N порядка 1000) $\tau_{max} = N/10$. Так что, чем длиннее ряд, тем меньшую долю составляет количество сдвигов.

Так как АКФ представляет собой совокупность коэффициентов корреляции, каждый из них должен быть проверен на значимость (см. работу 4).

Однако, для сокращения работы, можно на основе проверки по критерию Стьюдента нулевой гипотезы $\boldsymbol{H_0}$: $r = 0$ путем решения квадратного уравнения относительно r рассчитать критическое значение $r_{кр}$, соответствующее $t_{кр}$ при уровне значимости α и числе степеней свободы $\nu = N - \tau - 1$

$$r_{кр}(\tau) = \frac{-\sqrt{N-\tau-1} + \sqrt{N-\tau-1+4t_{кр}^2}}{2t_{кр}},$$

где N – длина реализации, τ – сдвиг АКФ; $t_{кр}$ $(\alpha, \nu = N - \tau - 1)$ – критерий Стьюдента.

Рассчитанные для каждого сдвига значения $r_{кр}$ также наносятся на график АКФ в положительной и отрицательной области r (пунктирные линии на рис. 9.2). Тогда все значения АКФ, превышающие (по модулю) $r_{кр}$, являются *значимыми,* а значения АКФ, меньшие $r_{кр}$, случайно отличаются от нуля.

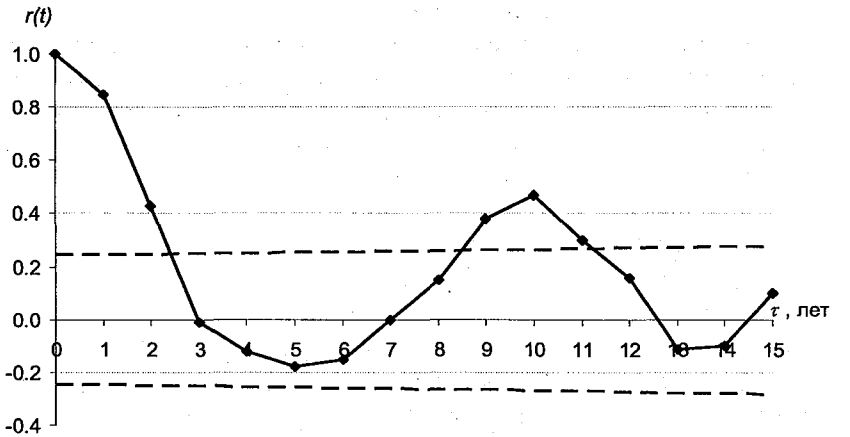


Рис. 9.2. Автокорреляционная функция межгодовой изменчивсти ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Анализ АКФ

При анализе АКФ можно сделать следующие выводы:

1. Инерционность процесса

Инерционность процесса определяется по радиусу корреляции. **Радиус корреляции** $\tau^2_{кор}$ – это сдвиг при первом пересечении функцией нуля. Инерционность процесса показывает, насколько долго характеристика сохраняет свое предыдущее состояние (например, если сильно потеплело, как долго это тепло будет сохраняться). В связи с этим процессы бывают малоинерционные ($\tau^2_{кор}$ = 1–2) или инерционные ($\tau^2_{кор}$ > 4–5). Например, на рис. 9.2 инерционность составляет 3 года.

2. Периодичность *процесса*

Если в исследуемом процессе присутствуют циклические колебания, то они отражаются на АКФ в виде локальных максимумов. Сколько выявляется *значимых локальных максимумов*, столько и периодических составляющих присутствует в процессе. Тогда ***период циклического колебания*** определяется по сдвигу АКФ, соответствующему локальному максимуму. Например, на рис. 9.2 выявляется один локальный максимум, характеризующий наличие в исследуемом процессе колебания с периодом 10 лет.

Если в АКФ отмечаются значимые локальные максимумы на *кратных сдвигах*, в таком случае делается вывод о наличии в исходном ряду *гармоники* с периодом, равным сдвигу первого локального максимума. Например, если локальные максимумы на сдвигах 6, 12, 18 мес и т.д., то делается вывод о наличии квазигармонического колебания с периодом 6 мес.

3. *Тип процесса*

По форме АКФ можно определить тип процесса. В классификации процессов выявляются следующие:

– «*белый шум*». Характеризует абсолютно случайный процесс. Значения АКФ равны нулю на всех сдвигах, кроме $\tau = 0$;

– *простая цепь Маркова I порядка.* Характеризует процесс, для которого свойственна связь со своим предыдущим состоянием. АКФ имеет значимый коэффициент автокорреляции на сдвиге $\tau = 1$;

– «*красный шум*». Характеризует процесс с высокой инерционностью (более 5 единиц дискретности). Может отражать наличие в исходном ряду периодичности с периодом, сравнимым с длиной ряда. (Служит признаком того, что в исходном ряду, возможно, не удален тренд). В АКФ отмечается радиус корреляции > 5;

– *циклический (квазигармонический) процесс.* Характеризует процесс с устойчивым ярко выраженным периодическим колебанием.

Процессы иногда могут иметь смешанный тип (например, простая цепь Маркова + квазигармонический процесс).

4. Возможность автопрогноза

Если АКФ не представляет собой «белый шум», на ней отмечаются *значимые* коэффициенты автокорреляции на сдвигах, отличных от 0. Следовательно, технически *можно сделать автопрогноз*. Тогда ***заблаговременность автопрогноза*** определится, как сдвиг, при котором коэффициент автокорреляции значим.

Если таких коэффициентов несколько, то существует возможность делать автопрогноз *с разной заблаговременностью*. Чем больше заблаговременность, тем ценнее прогноз.

С другой стороны качество прогноза будет зависеть от величины коэффициента автокорреляции – чем больше r, тем лучше прогноз. Как правило, два требования к автопрогнозу (лучшее ка-

чество и большая заблаговременность) противоречат друг другу. В этих случаях предпочтение отдается качеству.

Например, по АКФ на рис. 9.2 определяется, что формально автопрогноз можно делать с заблаговременностью 1, 2, 9 и 10 лет. Однако, наибольший коэффициент автокорреляции находится на сдвиге 1 год. Следовательно, оптимальнее всего делать автопрогноз с заблаговременностью 1 год.

Авторегрессия 1 порядка (АР1).

Автопрогноз любой заблаговременности выполняется на основе модели парной линейной регрессии, которая в этом случае называется ***авторегрессией 1 порядка.*** АР1 формулируется в общем виде:

$$x_{i+k} = a_1 x_i + a_0 + \varepsilon,$$

где k – заблаговременность прогноза.

Для нахождения коэффициентов и параметров регрессии берутся два ряда: x и x_k, сформированные по принципу на рис. 9.1 (т.е. исходный ряд и его аналог, сдвинутый на k чисел, обрезанные соответственно, один с начала, другой – с конца). Тогда ряд x (обрезанный с начала) будет *зависимой переменной*, а ряд x_k (обрезанный с конца) – *независимой переменной.*

Далее определяются коэффициенты, все параметры регрессии и оценивается ее качество (см. работу 5).

Например, конкретная модель автопрогноза ТПО с заблаговременностью 1 год (для рис. 9.2) формулируется так:

$$\text{ТПО}_{i+1} = 0{,}25 \cdot \text{ТПО}_i + 4{,}16.$$

Для задач оценки качества прогнозов исходный ряд предварительно делят на две части. Первая часть значительной длины называется ***зависимой выборкой*** и используется для расчета АКФ и АР1. Вторая, небольшая часть в конце ряда, называется ***независимой выборкой*** и используется для проверки полученной модели на данных, как бы неизвестных ранее.

Тогда дополнительно к оценке качества полученной модели АР1 добавляется еще одна – **стандартная ошибка независимого прогноза**, которая рассчитывается по формуле:

$$\sigma_{\varepsilon н} = \sqrt{\frac{1}{N_H}\sum_{i=1}^{N_H}(\varepsilon^2{}_i)} \qquad \varepsilon_i = x_{факт} - x_{пр},$$

где N_H – длина независимой выборки; $x_{факт}$ – фактические значения независимой выборки; $x_{пр}$ – соответствующие тем же моментам времени, рассчитанные по модели значения.

Стандартная ошибка независимого прогноза сравнивается со стандартным отклонением зависимой выборки (так же как и стандартная ошибка модели). Если $\sigma_{\varepsilon н} < \sigma_y$ считается, что внутренняя структура изменчивости характеристики не меняется со временем и модель можно использовать для прогноза.

Расчетная часть

Исходные данные

Среднемесячные и среднегодовые значения ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Порядок выполнения работы

Часть 1. Расчет и анализ АКФ

1. Из исходного ряда удалить тренд, если он значим (работа 8).

2. Разбить исходный ряд на две выборки – зависимую и независимую. В качестве независимой выборки взять часть ряда за 4–5 последних лет. Значения ряда за предыдущее время будут зависимой выборкой.

Для зависимой выборки:

1. Рассчитать автокорреляционную функцию (АКФ).

2. Рассчитать уровни значимости АКФ.

3. Построить совмещенный график АКФ и уровней значимости.

4. По графику определить радиус корреляции АКФ; периодичность.

5. Сделать выводы о внутренней структуре процесса: оценить его инерционность, периодичность, тип процесса.

Часть 2. Построение и оценка качества модели AP1

6. По графику определить возможность автопрогноза и его максимальную заблаговременность.

7. Сформулировать уравнение авторегрессии 1 порядка (AP1).

8. Сформировать ряды, необходимые для расчета модели АР1.

9. Рассчитать все характеристики модели АР1. Оценить ее качество (т.е. качество зависимого прогноза) (см. работу 5).

Для независимой выборки:

10. Рассчитать по уравнению модели АР1 значения характеристики в моменты времени независимой выборки.

11. Определить стандартную ошибку независимого прогноза. Оценить качество модели на независимой выборке.

12. Сделать общий вывод о возможности и качестве автопрогноза с рассмотренной заблаговременностью.

13. Рассчитать значения исследуемой характеристики (например, ТПО) по модели для зависимой и независимой выборок. Построить совмещенный график фактических и рассчитанных по модели значений характеристики. Проанализировать его.

ПРАКТИЧЕСКАЯ РАБОТА 10

Взаимнокорреляционный анализ (кросскорреляция)

Теоретическая часть

Физические процессы могут протекать *синхронно,* испытывая колебания параметра *одновременно,* но чаще всего – *асинхронно,* когда изменение одного параметра отразится на другом параметре через некоторое время, которое определяется *запаздыванием.*

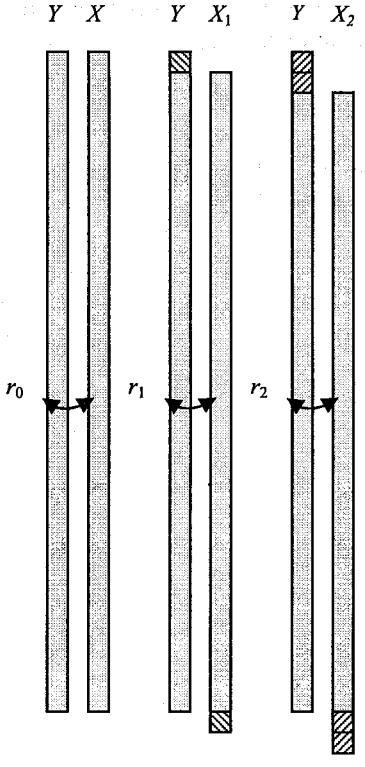


Рис.10.1. Схема расчета коэффициентов ВКФ

Например, если выпал дождь, то уровень в реке поднимется не сразу, воде требуется некоторое время, чтобы до реки «добежать».

Когда мы рассчитываем *коэффициент парной корреляции,* мы оцениваем связь параметров при *их синхронном* взаимодействии, и в большинстве случаев, когда между характеристиками существует *асинхронное* взаимодействие, коэффициент корреляции мал. В результате мы полностью теряем представление о связи двух характеристик.

Чтобы разрешить эту проблему, т.е. определить наличие связи двух характеристик при *асинхронном* взаимодействии и применяется взаимнокорреляционный анализ (кросскорреляция).

Взаимная корреляция (кросскорреляция) – это корреляция двух статистических рядов друг с другом при разных сдвигах во времени.

Принцип расчета взаимнокорреляционной функции (ВКФ) аналогичен расчету АКФ (работа 9).

Например (рис. 10.1), взяты два исходных ряда X и Y. Рассчитаем между парой этих рядов коэффициент корреляции r_0. Теперь сдвинем ряд X относительно ряда Y на 1 число вперед (ряд X_1) и уравняем длины рядов Y и X_1, обрезав на 1 число ряд Y с начала, а ряд X_1 – с конца. И снова рассчитаем коэффициент корреляции между этими рядами – r_1. И так можно сдвигать ряды далее, на каждом сдвиге уравнивая их длины и рассчитывая коэффициенты корреляции.

Так как ВКФ (в отличие от АКФ) *несимметрична*, ряд X нужно сдвигать и в другую сторону (назад) на то же количество сдвигов (отрицательное направление сдвигов).

Если все коэффициенты взаимной корреляции нанести на график зависимости r от τ (сдвига) и соединить плавной кривой, получим ***взаимнокорреляционную функцию*** (рис. 10.2).

Коэффициент взаимной корреляции $r(\tau)$ на каждом сдвиге также можно рассчитать по формуле

$$r(\pm\tau)=\frac{1}{\sigma_x\sigma_y(N-1-|\tau|)}\sum_{i=1}^{N-|\tau|}\left[(x_{\circledcirc}-\bar{x})(y_{i\pm\tau}-\bar{y})\right],$$

где N – длина реализации; τ – сдвиг, который меняется от $-\tau_{max}$ до τ_{max}.

Максимальное количество сдвигов τ_{max} (длина ВКФ) зависит от длины ряда N: если ряд короткий (N порядка 30–50) – $\tau_{max} = N/3$; если ряд длинный (N порядка 1000–10000) $\tau_{max} = N/10$, т.е. чем длиннее ряд, тем меньшую долю составляет количество сдвигов.

Так как ВКФ представляет собой совокупность коэффициентов корреляции, то их можно проверить на значимость или рассчитать ***уровень значимости ВКФ*** на каждом сдвиге, аналогично АКФ (работа 9).

Тогда все значения ВКФ, превышающие (по модулю) уровень значимости являются *значимыми,* а значения ВКФ между уровнями значимости статистически равны нулю.

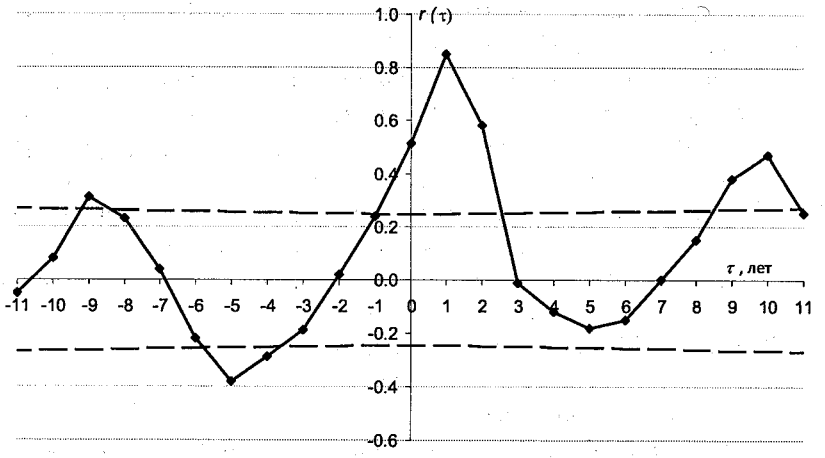


Рис. 10.2. Взаимнокорреляционная функция стока р. Волга и летних осадков на её водосборе

Анализ ВКФ

1. *Оценка асинхронной связи между двумя характеристиками.* Если на ВКФ присутствует *один* значимый коэффициент корреляции на сдвиге $\tau = 0$, эти две характеристики связаны *синхронно.*

Если на ВКФ присутствуют один или несколько значимых коэффициентов корреляции на сдвигах, *отличных от нуля,* значит эти две характеристики связаны *асинхронно.*

Если на ВКФ нет значимых коэффициентов корреляции, эти две характеристики линейно не связаны.

2. *Направление запаздывания.* Если асинхронные связи обнаружены, определяется *направление запаздывания:* какая характеристика изменяется раньше, а какая позже. Этому соответствуют разные направления сдвигов ВКФ (положительное или отрицательное). Например, при способе расчета на рис. 10.1, положительному направлению сдвигов ВКФ соответствует более раннее наступление характеристики x, а затем, через какой-то сдвиг – наступление характеристики y, т.е. $(y_{i+k} = f(x_i))$ или прогноз ряда y. Тогда отрицательному направлению сдвигов ВКФ соответствует, наоборот, $x_{i+k} = f(y_i)$ или прогноз ряда x.

3. *Возможность прогноза*. Если в ВКФ отмечаются *значимые* коэффициенты корреляции на сдвигах, отличных от 0, то теоретически, *можно сделать прогноз*. Тогда ***заблаговременность прогноза*** определится, как сдвиг, при котором коэффициент корреляции значим.

Если таких коэффициентов несколько, то существует возможность делать прогноз *с разной заблаговременностью*. Чем больше заблаговременность, тем ценнее прогноз.

С другой стороны, качество прогноза будет зависеть от величины коэффициента корреляции: чем больше r, тем лучше прогноз. Как правило, два требования к прогнозу (лучшее качество и большая заблаговременность) противоречат друг другу. В этих случаях предпочтение отдается качеству.

Кроме того, имеет значение, в какой области сдвигов (положительной или отрицательной) находятся эти значимые коэффициенты корреляции. Иногда прогноз не имеет смысла из физических соображений.

Например, по ВКФ на рис. 10.2 определяется, что прогноз можно делать с заблаговременностью 1, 2, 9 и 10 лет (положительное направление) – зависимость стока от осадков и с заблаговременностью 5 и 9 лет (отрицательное направление) – зависимость осадков от стока реки. Очевидно, что рассматривать прогноз осадков в зависимости от стока реки нет физического смысла.

Из оставшихся коэффициентов корреляции наибольший находится на сдвиге 1 год. Следовательно, оптимальнее всего делать прогноз стока реки в зависимости от осадков с заблаговременностью 1 год.

Модель прогноза.

Основанный на ВКФ прогноз любой заблаговременности выполняется на основе модели парной линейной регрессии.

Уравнение прогностической модели формулируется в общем виде:

$$y_{i+k} = a_1 \cdot x_i + a_0 + \varepsilon,$$

где k – заблаговременность прогноза.

Для нахождения коэффициентов и параметров регрессии берутся два ряда: y и x, сформированные по принципу на рис. 10.1

(т.е. два исходных ряда, сдвинутые друг относительно друга на k чисел, и обрезанные соответственно один с начала, другой – с конца). Тогда ряд y (обрезанный с начала) будет *зависимой переменной*, а ряд x (обрезанный с конца) – *независимой переменной*.

Далее определяются коэффициенты, все параметры регрессии и оценивается ее качество (см. работу 5).

Например, конкретная модель прогноза стока с заблаговременностью 1 год (для рис. 10.2) формулируется так:

$$S_{i+1} = 0{,}05 * P_i + 2516{,}2,$$

где S – сток Волги (км^3/год); P – летние осадки (мм) на её водосборе.

Расчетная часть

Исходные данные

1. Среднемесячная и среднегодовая ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

2. Среднемесячные и среднегодовые значения гидрометеорологических характеристик в широтной зоне умеренных широт Северной Атлантики (45 – 60° ю.ш.) за период с 1971 по 1985 гг.

Порядок выполнения работы

Разбить исходные ряды на две выборки: зависимую и независимую. В качестве независимой выборки взять часть ряда за 4–5 последних лет. Значения ряда за предыдущее время будут зависимой выборкой.

Для зависимой выборки:

1. Рассчитать ВКФ для ТПО и R.

2. Рассчитать уровни значимости ВКФ.

3. Построить совмещенные графики ВКФ и уровней значимости.

4. Определить прогностическое направление ТПО = $f(R)$ и в этом направлении выявить оптимальную заблаговременность прогноза ТПО.

4. Сформулировать уравнение регрессии (модели) ТПО= $f(R)$ с оптимальной заблаговременностью.

5. Сформировать ряды (с учетом сдвигов), необходимые для расчета модели.

6. Рассчитать все характеристики модели (см. работу 5): коэффициенты модели и 4 параметра качества.

7. На основе полученных коэффициентов написать уравнение прогностической модели. Оценить ее качество (т.е. качество зависимого прогноза).

Для независимой выборки:

8. Рассчитать по уравнению прогностической модели значения ТПО в моменты времени независимой выборки.

9. Определить стандартную ошибку независимого прогноза (см. работу 9). Оценить качество независимого прогноза.

10. Сделать общий вывод о возможности и качестве прогноза ТПО = $f(R)$ с рассмотренной заблаговременностью.

11. Рассчитать значения исследуемой характеристики (например, ТПО) по модели для зависимой и независимой выборок. Построить совмещенный график фактических и рассчитанных по модели значений характеристики для зависимой и независимой выборок одновременно. Проанализировать его.

ПРАКТИЧЕСКАЯ РАБОТА 11

Гармонический анализ Фурье и спектр

Теоретическая часть

Во внутренней структуре временного ряда могут отмечаться одна или несколько циклических составляющих. Наиболее ярко выраженные из них отражаются в АКФ. Менее выраженные могут в АКФ не проявиться.

Если возникает задача выявить в исследуемом процессе только циклические составляющие, можно воспользоваться ***спектральным или гармоническим анализом***, не рассчитывая предварительно АКФ.

В основе ***гармонического анализа*** лежит идея, что любой ряд можно разложить без остатка в ряд Фурье, т.е. на конечное число гармоник. **Гармониками** называются тригонометрические функции, имеющие периоды, кратные длине ряда, т.е. каждая гармоника целое число раз «укладывается» в длину исходного ряда.

Формула гармоники:

$$G_k = A_k \cos(\omega_k \cdot t - \varphi_k), \qquad \omega_k = 2\pi/T_k$$

где k – номер гармоники; A_k – амплитуда k-той гармоники; ω_k – частота k-той гармоники; T_k – период k-той гармоники; φ_k – фаза k-той гармоники; t – время.

Полное разложение в ряд Фурье предполагает определение гармоник количеством $N/2$.

Гармоники, как правило, нумеруются. Гармоника № 1 (G_1) имеет период равный длине ряда N, далее период гармоники уменьшается: G_2 имеет период $N/2$; G_3 – $N/3$ и т.д. Последняя гармоника имеет период равный 2 единицы дискретности, соответствующая ей частота называется ***частотой Найквиста***.

Характеристики гармоники определяются на основании *коэффициентов Фурье* a_k *и* b_k:

$$a_k = \frac{2}{N}\sum_{i=1}^{N}\left[x_i \sin(\varpi_k \cdot t_i)\right]; \quad b_k = \frac{2}{N}\sum_{i=1}^{N}\left[x_i \cos(\varpi_k \cdot t_i)\right],$$

где x_i – исходный ряд; N – длина ряда; ω_k – частота k-той гармоники; t_i – ряд времени $t_i = i$, $i = 1, 2, 3....N$.

Характеристики гармоники определяются по формулам:

Амплитуда гармоники $A_k = \sqrt{a_k^2 + b_k^2}$;

Фаза гармоники $\varphi_k = arctg\frac{a_k}{b_k} \pm \pi$;

Дисперсия гармоники $D_k = \frac{A_k^2}{2}$;

Вклад гармоники в общую дисперсию ряда: $V_k = \frac{D_k}{D_y}$,

где D_y – дисперсия исходного ряда.

Амплитуда гармоники показывает наибольшее отклонение характеристики от среднего значения на данном периоде.

Фаза гармоники может быть определена в единицах дискретности (времени): $\varphi_{kt} = \frac{\varphi_k T_k}{2\pi}$. Фаза характеризует время наступления максимума на данном периоде.

Вклад дисперсии гармоники в общую дисперсию ряда, по сути, является аналогом коэффициента детерминации r^2, поэтому точно так же может быть проверен на значимость (см. работу 8). При проверке можно сделать вывод *о значительном вкладе в дисперсию ряда* или *незначимости гармоники.*

Если потом соответствующие характеристики подставить в формулу гармоники, то можно рассчитать её значения на каждый момент времени.

Современные статистические программы («Statistica», «SPSS» и т.п.) рассчитывают разложение Фурье методом БПФ (быстрого преобразования Фурье) при этом теряя информацию о некоторых характеристиках гармонических составляющих. По сути, в качестве результата там представляется зависимость дисперсии гармонических составляющих от их частоты, и называется это **периодограммой (или спектром)**. Тогда гармоники, имеющие большую дисперсию (и соответственно, амплитуду), будут представлены «пиками» на соответствующем графике спектра.

В большинстве исследований представляется достаточным выявление частот (и соответственно, периодов) для которых отмечаются пики периодограммы.

Однако, часто возникает необходимость *восстановить* исходный процесс по нескольким значимым гармоникам, тогда необходимо прибегнуть к процедуре *гармонического анализа на основе рассчитанной периодограммы.*

Тогда по таблицам и графикам периодограмм, полученных методом БПФ нужно определить периоды для пиков спектра, а потом для этих периодов провести гармонический анализ: рассчитать характеристики гармоник, сформулировать для них уравнения, рассчитать ряды этих гармоник (в зависимости от времени) и путем сложения этих рядов получить восстановленный ряд.

Расчетная часть

Исходные данные

1. Среднемесячная и среднегодовая ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1985 гг.

Порядок выполнения работы

Из исходного ряда удалить тренд (если он значим) или среднее значение (если тренд незначим). Получится ряд *отклонений.*

Для ряда среднемесячных отклонений.

2. По формулам рассчитать характеристики **годовой и полугодовой** гармоник: амплитуды, фазы, дисперсии гармоник, вклад в дисперсию исходного ряда.

3. Оценить значимость гармоник.

4. Восстановить ряд ТПО по формулам гармоник как сумму значимых гармоник.

5. Построить совмещенный график исходных и восстановленных значений ТПО.

Для ряда среднегодовых отклонений.

6. В программе «Statistica» рассчитать периодограмму для ряда отклонений от тренда (см. технологию выполнения работы). Рисунок периодограммы и ему соответствующую таблицу перенести в Excel. По таблице и рисунку определить периоды, соответствующие «пикам» периодограммы.

7. Для каждого из этих периодов отдельно по формулам рассчитать характеристики гармоники: амплитуды, фазы, дисперсии гармоник, вклад в дисперсию исходного ряда.

8. Оценить значимость выбранных гармоник.

9. Восстановить ряд ТПО по формулам гармоник как сумму значимых гармоник.

10. Построить совмещенный график исходных и восстановленных значений ТПО.

Технология выполнения работы

Построение периодограммы в программе «Statistica»

После открытия программы вызвать меню «File»–«New». Согласиться на предлагаемый вариант. В открывшееся окно чистой таблицы из Excel через буфер обмена вставить свой ряд. Его имя будет «Var1», если он стоит в первой колонке.

Далее в меню «Statistics» – «Advanced …» – « Time series…».

Нажать на кнопку «Variable» и выбрать имя своего столбца «Var1».

Далее (внизу) «Spectral (Furier) analysis». В открывшемся окне – «ОК».

В открывшемся окне с результатами расчета нажать кнопку «Summary». Получится таблица с результатами разложения Фурье. Ее нужно выделить. В меню «Edit»–«Save with headers». Вставить в Excel.

Вернуть окно с результатами (внизу слева). На вкладке «Quick» нажать «Periodogram». Получится рисунок периодограммы. Рисунок можно сохранить как часть экрана, если нажать комбинацию Alt-F3 и крестиком выбрать область сохранения в буфер. Вставить в Excel.

Рисунок можно нарисовать в Excel и по таблице, полученной на предыдущем шаге. Для этого рисуется столбец «Periodogram» как функция от столбца «Frequency» или «Period».

ПРАКТИЧЕСКАЯ РАБОТА 12

Объективный анализ океанологических полей

Теоретическая часть

Статистическая выборка может представлять собой *временной ряд* какой-то характеристики, а может отражать распределение этой характеристики в геометрическом (или географическом) пространстве в один момент времени. В последнем случае выборка называется ***полем*** характеристики.

Поле представляет собой информацию о характеристике в разных точках пространства, в общем случае, расположенных как угодно.

Одной из ***задач объективного анализа*** является определение закономерности, по которой характеристика проявляется в разных точках поля, чтобы можно было её определить в любой другой точке.

Для решения этой задачи можно применить *методы пространственной интерполяции.*

Кроме того, её можно решить *на основе статистического подхода.* Если рассматривать геофизические характеристики, то их распределение в пространстве географических координат (широты и долготы) имеет некоторые известные закономерности. Например, ТПО в океанах зависит от широты: чем севернее, тем холоднее. С другой стороны, эта закономерность нарушается системами океанской циркуляции так, что на западных границах океанов ТПО теплее, а на восточных – холоднее, т.е. ТПО зависит и от долготы. Таким образом, распределение ТПО зависит от географических координат места.

Чтобы наиболее полно выразить эту закономерность, её можно сформулировать в виде модели множественной нелинейной регрессии, которая представляет собой двумерный полином 2-го порядка:

$$\text{ТПО} = a_1 \varphi + a_2\lambda + a_3\varphi^2 + a_4\lambda^2 + \alpha_5\varphi\lambda + a_0 + \varepsilon,$$

где φ – широта места; λ – долгота места.

Коэффициенты модели и характеристики её качества можно определить обычными статистическими методами (см. работы 5 и

6), а также упростить модель, подобрав наилучшую на основе пошаговой линейной регрессии (работа 7).

Тогда, для того чтобы узнать значение ТПО в какой-нибудь точке с координатами φ_l и λ_l, достаточно подставить значения этих координат в полученную модель.

Расчетная часть

Исходные данные

Среднегодовая температура поверхности океана (ТПО) в 1971 г. в 456 точках на акватории Северной Атлантики в сетке 1×1 градус.

Порядок выполнения работы

Разбить исходный ряд на две выборки: зависимую и независимую. В качестве независимой выборки выбрать из всей совокупности точек в пространстве 1/3 часть приблизительно в «шахматном» порядке. Оставшиеся 2/3 совокупности точек будут зависимой выборкой.

Для зависимой выборки:

1. Рассчитать полную модель множественной нелинейной регрессии (полином 2 порядка) связи ТПО с географическими координатами.

2. Рассчитать все характеристики модели (работа 5). Оценить ее качество (т.е. качество восстановления поля ТПО).

3. Написать уравнение модели на основе полученных коэффициентов.

Для независимой выборки:

4. Рассчитать по уравнению модели значения ТПО в точках независимой выборки.

5. Определить стандартную ошибку независимой оценки (работа 9).

6. Сделать общий вывод о возможности и качестве объективного анализа ТПО на основе регрессионной модели.

7. Построить карту распределения ТПО на акватории Северной Атлантики.

8. Построить карту распределения ошибок независимой оценки на акватории Северной Атлантики. Проанализировать её, учитывая карту распределения ТПО.

Технология выполнения работы.

Для построения карт можно пользоваться пакетом «Surfer».

ПРАКТИЧЕСКАЯ РАБОТА 13

Анализ малых выборок

Теоретическая часть

Если длина статистических рядов меньше 25–30 значений, считается, что это ***малая выборка.***

Для малых выборок статистические оценки являются неэффективными. Поэтому для статистического анализа малых выборок применяются специально разработанные, так называемые *непараметрические методы оценивания.*

В частности, невозможно рассчитать ЭФР на основе интервального оценивания, так как количество интервалов стремится к нулю. Поэтому для исследования ЭФР используется ***квантильный анализ***.

Квантиль вероятности p_k характеризует собой такое значение характеристики, при которой вероятность появления значений меньше неё равна p_k.

Чтобы рассчитать квантили необходимо произвести ранжирование ряда по возрастанию и каждому значению ряда поставить в соответствие значение его интегральной вероятности:

$$p_i = i/N,$$

где i – номер значения ряда в ранжированном ряду; N – длина ряда.

Квантилем вероятности 10% (децилем), например, будет значение ряда с $p_i = 0{,}1$.

В квантильном анализе чаще всего используются квантили с вероятностями 25%, 50%, 75%, которые делят всю функцию распределения на четыре части и поэтому называются ***квартилями***.

Квартилей три:

– первый квартиль – с вероятностью 25 %;

– второй квартиль – с вероятностью 50 % (медиана);

– третий квартиль – с вероятностью 75 %.

Также часто рассчитывают ***интерквартильное (межквартильное) расстояние*** $x_{0,75}–x_{0,25}$.

На основе квартилей рисуют ***«ящик с усами»,*** что является схематическим изображением ЭФР (рис. 13.1). «Ящик» рисуется по первому и третьему квартилям, внутри него отчерчивается линией медиана, а «усы» отражают минимальное и максимальное значения ряда.

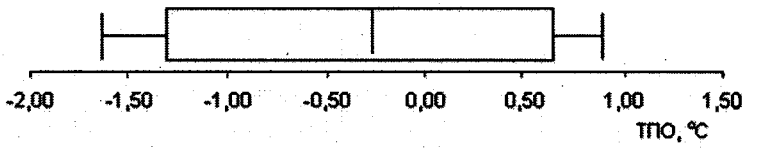


Рис. 13.1. Квантильный анализ изменчивости ТПО в точке 50° с.ш. 60° з.д. в 1971 г.

Анализ *«ящика с усами»* проводится по аналогии с ЭФР: медиана характеризует центр тяжести, её расположение в межквартильном промежутке отражает асимметрию распределения, а длина «усов» отражает эксцесс (плосковершинность) распределения (см. работу 1).

Среди непараметрических методов оценки тесноты связи наиболее частое применение находят *ранговые коэффициенты Спирмена и Кендалла.* Эти коэффициенты используются как для количественных рядов, так и для рядов с качественными признаками.

Ранг – это порядковый номер значения *ранжированного ряда.* ***Ранжирование*** – упорядочение значений ряда по возрастанию или убыванию.

Если два значения ряда одинаковы, то их ранги также одинаковы и равны средней арифметической между порядковыми номерами этих значений в ранжированном ряду. Такие ранги называются ***связными***. Например, строки 9 и 10 в табл. 13.1.

Ранговый коэффициент корреляции Спирмена рассчитывается по формуле

$$\rho_{x/y} = 1 - \frac{6\sum d_i^2}{N(N^2-1)},$$

где d_i^2 – квадрат разности рангов; N – число наблюдений.

Значимость коэффициента корреляции Спирмена проверяется на основе t-критерия Стьюдента. Расчетное значение критерия определяется по формуле

$$t^* = \rho_{x/y}\sqrt{\frac{N-2}{1-\rho^2_{x/y}}}.$$

Значение коэффициента корреляции считается значимым, если $t^* > t_{кр}(\alpha, \nu = N-2)$.

Расчет коэффициента корреляции Спирмена лучше всего делать с помощью таблицы.

Например (табл. 13.1), исходные значения ТПО в двух точках были ранжированы по возрастанию, на основе этого были определены ранги для каждого числа и рассчитан коэффициент корреляции Спирмена, который составил 0,58. Он был проверен на значимость при уровне значимости 0,05 и оказался значимым, так как $t^*(2{,}46) > t_{кр}(2{,}18)$.

Таблица 13.1

Расчет коэффициента корреляции Спирмена

Номер	ТПО в точке 1	Та в точке 2	Ранги (ряд 1)	Ранги (ряд 2)	Разность рангов d_i	d_i^2
1	0,62	28,90	11	7	4	16
2	0,89	28,78	13	5	8	64
3	0,60	28,85	10	6	4	16
4	0,13	29,04	9	9	0	0
5	–0,41	29,20	7	12	–5	25
6	–1,01	28,91	6	8	–2	4
7	–1,42	28,69	4	4	0	0
8	–1,59	28,23	3	2	1	1
9	–1,70	28,15	1,5	1	0,5	0,25
10	–1,70	28,54	1,5	3	–1,5	2,25
11	–1,12	29,15	5	10	–5	25
12	–0,17	29,29	8	14	–6	36
13	0,70	29,16	12	11	1	1
14	0,91	29,24	14	13	1	1
					Сумма	191,5

Расчетная часть

Исходные данные

1. Среднегодовая ТПО в точке 50° с.ш. 60° з.д. за период с 1971 по 1980 гг.

2. Среднегодовые значения температуры воздуха в широтной зоне умеренных широт Северной Атлантики (45 – 60° ю.ш.) за период с 1971 по 1980 гг.

Порядок выполнения работы

1. Рассчитать медиану и другие квартили распределения для обоих рядов отдельно (см. технологию выполнения работы). Нарисовать «ящики с усами».

2. Сделать анализ «ящиков с усами» как эмпирических распределений.

3. Рассчитать коэффициент ранговой корреляции Спирмена между двумя рядами. Проверить его на значимость. Сделать вывод о наличии или отсутствии связи между этими двумя характеристиками.

Технология выполнения работы

Квартили можно рассчитать с помощью функции Excel «квартиль», где входными параметрами служат исходный ряд и код квартиля (1 – минимум, 2 – первый квартиль, 3 – второй квартиль, 4 – третий квартиль, 5 – максимум)

ЛИТЕРАТУРА

1. *Малинин В.Н.* Статистические методы анализа гидрометеорологической информации. Учебник. – СПб, изд. РГГМУ, 2008. – 408 с.
2. *Вайновский П.А., Малинин В.Н.* Методы обработки и анализа гидрометеорологической информации. Ч.1. Одномерный анализ. – Л. изд. ЛГМИ, 1991.– 136 с.
3. *Вайновский П.А., Малинин В.Н.* Методы обработки и анализа гидрометеорологической информации. Ч.2. Многомерный анализ. – СПб. изд. РГГМИ, 1992.– 96 с.
4. *Вуколов В.И.* Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. –М.: ФОРУМ-ИНФРА-М, 2004. – 462 с.
5. *Макарова Н.В., Трофимец В.Я.* Статистика в Excel.– М.: Финансы и статистика, 2002.–365 с.

СОДЕРЖАНИЕ

Предисловие 3
Вводные понятия статистики 3
Вводные понятия технологии. 5
Практическая работа 1. Первичные статистики и эмпирическая функция распределения 7
Практическая работа 2. Проверка соответствия эмпирической функции распределения нормальному закону 14
Практическая работа 3. Проверка статистических гипотез. Оценка стационарности временного ряда 18
Практическая работа 4. Корреляционный анализ 25
Практическая работа 5. Парная линейная регрессия 29
Практическая работа 6. Нелинейная регрессия 35
Практическая работа 7. Множественная линейная регрессия 40
Практическая работа 8. Анализ тренда временного ряда 45
Практическая работа 9. Автокорреляционный анализ и авторегрессия 1 порядка 49
Практическая работа 10. Взаимнокорреляционный анализ (кросскорреляция) 56
Практическая работа 11. Гармонический анализ Фурье и спектр 62
Практическая работа 12. Объективный анализ океанологических полей 66
Практическая работа 13. Анализ малых выборок 68
Литература 72

Учебное издание

Светлана Михайловна Гордеева

ПРАКТИКУМ

по дисциплине

"СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ"

Учебное пособие

Редактор И.Г. Максимова
Компьютерная верстка Н.И. Афанасьевой

ЛР № 020309 от 30.12.96

Подписано в печать 27.10.10. Формат 60×90 1/16. Гарнитура Times New Roman.
Бумага офсетная. Печать офсетная. Уел. печ. л. 4,8. Тираж 250 экз. Заказ № 59/10
РГГМУ, 195196, Санкт-Петербург, Малоохтинский пр., 98.
ЗАО «НПП «Система», 197045, Санкт-Петербург, Ушаковская наб., 17/1.