**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики и информатики**

**Кафедра т**е**ории вероятностей и математической статистики**

ВОДИЧ

Иван Юрьевич

**Сравнение различных подходов к прогнозированию временных рядов с регулярными периодическими компонентами на примере данных о температурном режиме озера**

Дипломная работа

|  |  |
| --- | --- |
|  | Научный руководитель: кандидат физ.-мат. наук доцент Т.В. Цеховая |

Допущен к защите

«\_\_» \_\_\_\_\_\_\_\_\_\_\_\_ 2017 г.

Зав. кафедрой теории вероятностей и математической статистики

доктор физико-математических наук, профессор Н.Н. Труш

Минск, 2017Белорусский государственный университет

Факультет прикладной математики и информатики

Кафедра теории вероятностей и математической статистики

Утверждаю

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Туш Н.Н.

**ЗАДАНИЕ НА ДИПЛОМНУЮ РАБОТУ**

Студенту 5 курса ПМ Водичу Ивану

1.Тема ***Сравнение различных подходов к прогнозированию временных рядов с регулярными периодическими компонентами на примере данных о температурном режиме озер***

Утверждена приказом ректора БГУ от \_\_\_\_\_\_\_\_\_\_\_\_\_№\_\_\_\_\_

2. Исходные данные к дипломной работе

2.1. <http://statsoft.ru/>

2.2. Шмойлова Р.А. Теория статистики. Учебник / Р.А. Шмойлова, В.Г. Минашкин, Н.А. Садовникова, Е.Б. Шувалова. М.: Финансы и статистика, 2011. – 656 с.

2.3. Палий И.А. Прикладная статистика. М.: Высшая школа, 2004.– 176с.

2.4. Харин Ю.С. Теория вероятностей, математическая и прикладная статистика/ Ю.С. Харин, Н.М. Зуев, Е. Е. Жук – Минск: БГУ, 2011. – 463 с.

2.5. База данных Учебно-научного центра БГУ «Нарочанская биологическая станция Г.Г. Винберга» .

3. Перечень подлежащих разработке вопросов или краткое содержание расчетно-пояснительной записки:

3.1. Обзор распространенных моделей и методов прогнозирования   временных рядов с регулярными периодическими компонентами.

3.2. Предварительный статистический анализ данных о температурном режиме озера Баторино (Беларусь).

3.3. Анализ и моделирование периодической компоненты реального временного ряда (тренд, ARIMA, сезонная компонента,  гармонический анализ, спектральный анализ).

3.4. Прогнозирование значений временного ряда на основе построенных моделей. Оценка точности полученных прогнозов.

4. Перечень графического материала (с точным указанием обязательных чертежей и графиков) .

5. Консультанты по дипломной работе с указанием относящихся к ним разделов  *Цеховая Т.В.*

6. Примерный график выполнения дипломной работы

10 апреля 2017 г. промежуточный отчет

24 апреля 2017г. промежуточный отчет

8 мая 2017 г. доклад о проделанной работе

7. Дата выдачи задания \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

8. Срок сдачи законченной дипломной работы \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель \_\_\_\_\_\_\_\_\_\_\_\_ Т.В. Цеховая

Подпись обучающегося \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата

АННОТАЦИЯ

Водич И.Ю. Сравнение различных подходов к прогнозированию временных рядов с регулярными периодическими компонентами на примере данных о температурном режиме озер: Дипломная работа/ Минск: БГУ, 2017. – 40 с.

Рассматривается задача анализа, моделирования и прогнозирования реальных временных рядов с регулярными периодическими компонентами на примере данных о температурном режиме озера Баторино (Беларусь).

**АНАТАЦЫЯ**

Водзіч І.Ю. Параўнанне розных падыходаў да прагназавання часавых шэрагаў з рэгулярнымі перыядычнымі кампанентамі на прыкладзе дадзеных аб тэмпературным рэжыме азёр: Дыпломная работа / Мінск: БДУ, 2017. - 40 с.

Разглядаецца задача аналізу, мадэлявання і прагназавання рэальных часовых шэрагаў з рэгулярнымі перыядычнымі кампанентамі на прыкладзе дадзеных аб тэмпературным рэжыме возера Баторына (Беларусь).

**ANNOTATION**

Vodzich I.Y. Comparison of different approaches to predicting time series with regular periodic components using data on the temperature regime of lakes: Diploma work / Minsk: BSU, 2017. - 40 p.

We consider the problem of analysis, modeling and forecasting of real time series of regular periodic components on the example of data about temperature Batorino lake (Belarus).

РЕФЕРАТ

Дипломная работа, 40 с., 9 рис., 13 табл., 13 источников.

***Ключевые слова****:* временные ряды, статистический анализ, прогнозирование, регулярные периодические компоненты, квадратический тренд, сезоны.

***Объект исследования*** – данные о температурном режиме озера.

***Цель исследования*** – построение математических моделей, описывающих динамику реальных временных рядов с регулярными периодическими компонентами.

***Методы исследования*** – теория вероятностей, математическая статистика, случайные процессы, математический анализ, геометрия и алгебра.

***В результате исследования*** построена математическая модель, описывающая поведение реальных временных рядов с регулярными периодическими компонентами.

***Область применения* –** любая область, где решается задача прогнозирования.

**Содержание**

[АННОТАЦИЯ 4](#_Toc484076645)

[РЕФЕРАТ 5](#_Toc484076646)

[Введение 7](#_Toc484076647)

[1. Основные определения 11](#_Toc484076648)

[2. Обзор моделей временного ряда 15](#_Toc484076649)

[3. Первичный анализ временного ряда 19](#_Toc484076650)

[4. Построение моделей временного ряда 27](#_Toc484076651)

[ЗАКЛЮЧЕНИЕ 31](#_Toc484076652)

[Список литературы 32](#_Toc484076653)

[ПРИЛОЖЕНИЕ А 34](#_Toc484076654)

[ПРИЛОЖЕНИЕ Б 36](#_Toc484076655)

# Введение

В настоящее время статистические приемы исследования применяются во всех науках о Земле и обществе. Данные экологической статистики используются в социально-экономическом анализе для оценки результатов мероприятий относительно снижения вредных выбросов в атмосферу, загрязненных стоков в естественные водные объекты, определения взаимосвязи качества окружающей среды и состояний здоровья населения, а также определения экономического убытка от загрязнения окружающей среды в связи со снижением урожайности сельскохозяйственных культур, ухудшением производительности в животноводстве, повышенным износом домов, сооружений и т. д. Вследствие постоянного роста требуемой точности прогнозирования растет сложность моделей и объемы обрабатываемых данных. Современное статистическое исследование в области экологии уже немыслимо без использования компьютерной техники для сбора, хранения и обработки данных.

Изучением озер занимается наука лимнология (озероведение). Все процессы, происходящие в озерах, — геологические, физические, химические, биологические — она изучает комплексно, во взаимосвязи с водосбором озера. В настоящее время важнейшей задачей изучения озер является исследование влияния антропогенного фактора на изменение природы озер, выработка научных основ охраны водоемов. Для решения этих вопросов необходимо глубокое изучение биологических и химических процессов, определяющих скорость образования органического вещества и его разрушения. Большой интерес представляет разработка проблемы управления озерными системами с использованием математических моделей и современной вычислительной техники.

Изучение термического режима озёр с целью прогнозирования их изменений во времени, а также влияния, как на внутриводные процессы, так и на окружающий водосбор — одна из основных задач. Знание тепловых процессов в водоёмах позволяет выявить причины многих явлений, происходящих в водной среде, и в атмосфере, а также проследить пути эволюции естественных водных экосистем. Озёра представляют собой сложную термическую гидроэкосистему. Повышение температуры воды уменьшает содержание кислорода, увеличивает продолжительность периода свободного ото льда и раннее вскрытие ледяного покрова на озёрных водоёмах.

При исследовании гидрологических характеристик водных объектов получают наиболее точные результаты и прогнозы. Продемонстрируем, что выбранная тема для исследования актуальна сейчас и имеет развитие. Приведём ниже примеры работ по лимнологии в нашей стране и за рубежом.

В [2] приведены климатические особенности местной территории, водный режим озёр Нарочь и Мястро, результаты комплексных гидроэкологических исследований Нарочанских озер, характеризующих их экологическое состояние, сведения о физико-химических и биологических показателях.

В [7] проанализированы данные о температурном режиме трех Нарочанских озер и температуре воздуха в Нарочанском регионе за последние пятьдесят лет. С помощью корреляционного анализа показано наличие слабых тенденций к многолетнему увеличению температуры для верхних горизонтов оз. Нарочь и поверхностных горизонтов оз. Мястро и Баторино.

В [3] дана общая характеристика водных ресурсов Национального парка «Нарочанский», а также краткие характеристики и описание географического положения, морфометрических и гидрологических параметров 48 водоемов, 68 водотоков, 16 родников, водных объектов искусственного происхождения, подземных вод. Для важнейших водоемов и водотоков приведены гидрохимическая, гидроэкологическая и гидробиологическая характеристики

Наиболее актуальные проблемы гидрологической науки и практики обсуждались на VII Всероссийском гидрологический съезде [6]. Среди значимых секций докладов были: «Опасные гидрологические явления — оценка, прогноз, снижение рисков», «Водные ресурсы и водный баланс водных объектов в условиях влияния антропогенной деятельности и климатических изменений», «Моделирование гидрологического цикла».

Последние научные прорывы в фундаментальных и прикладных разделах гидрологии опубликованы [9]. Например, приводится описание зависимости климата некоторой местности от её экологического состояния водных ресурсов и степенью антропогенного воздействия на них. Обсуждается, как научные достижения гидрологии связаны с признанием того, что глобальные климатические и экологические изменения реальны.

Температура является важной темой для изучения экологической адаптации на всех уровнях биологической организации [12]. Термодинамические свойства, которые лежат в основе биохимической кинетики и стабильности белков определяют общую температурную чувствительность скорости процессов.

Применение анализа временных рядов для моделирования гидрологических характеристик и процессов продемонстрированно в [8], [13]. Что обеспечивает использование статистической, вероятностной теории и стохастических процессов в гидрологии, исследовательский опыт которых, даёт понять различные аспекты в моделировании гидрологических рядов.

В настоящей работе исследуем измерения температуры воды на глубине 3 м озера Баторино. Это озеро расположено в Мядельском районе Минской области, входит в группу Нарочанских озёр. Площадь зеркала озера — 6,25 км², водосбора — 92,4 км², максимальная глубина — 5,5 м, длина береговой линии 15 км.

Актуальность темы исследования обусловлена необходимостью выявления закономерностей, определяющих формирование и развитие термической структуры в озёрах Беларуси.

Вычислительные процедуры статистического анализа являются трудоёмкими при их реализации. Поэтому эффективная статистическая обработка данных даже небольшого объёма невозможна без использования вычислительной техники. В настоящее время актуально развитие и использование компьютерных технологий обработки и анализа данных.

Статистический анализ полученной выборки и все расчёты будем проводить на компьютере с помощью пакета прикладных программ STATISTICA, среды RStudio. Это универсальные интегрированные системы, предназначенные для статистического анализа и обработки данных. Содержит многофункциональную систему для работы с данными, широкий выбор статистических модулей, специальный инструментарий для подготовки отчётов, мощную графическую систему для визуализации данных.

# Основные определения

Пусть  - вероятностное пространство, где  является произвольным множеством элементарных событий,  - сигма-алгеброй подмножеств , и  - вероятностной мерой.

**Определение 1.1.** *Действительным случайным процессом * называется семейство действительных случайных величин, заданных на вероятностном пространстве , где , , где Т – некоторое параметрическое множество.

**Определение 1.2.** Если , то ,  называют *случайным процессом с непрерывным временем*.

**Определение 1.3.** Если , то ,  называют *случайным процессом с дискретным временем.*

**Определение 1.4.** *n-мерной функцией распределения* случайного процесса , , называется функция вида:



где , , .

**Определение 1.5.** *Математическим ожиданием* случайного процесса , , называется функция вида:

 .

**Определение 1.6.** *Дисперсией* случайного процесса , , называется функция вида:

.

**Определение 1.7.** *Ковариационной функцией* случайного процесса , , называется функция вида:



**Определение 1.8.** *Корреляционной функцией* случайного процесса , , называется функция вида:

 .

Заметим, что для центрированного случайного процесса .

**Определение 1.9.** *Нормированной ковариацонной функцией* случайного процесса , , называется функция вида:

,

где ,  - случайный процесс.

**Определение 1.10.** *Вариограммой* случайного процесса ,  называется функция вида:



При этом функция  называется *семивариограммой*.

**Определение 1.11.** Случайный процесс , , называется *стационарным в узком смысле,* если , ,  выполняется соотношение:

.

**Определение 1.12.** Случайный процесс , , называется *стационарным в широком смысле,* если , и

1. , ;

2. , .

*Замечание* 1.1. Если случайный процесс , , является стационарным в узком смысле и , то он будет стационарным и в широком смысле, но не наоборот.

В дальнейшем стационарный в широком смысле процесс будем называть стационарным процессом.

**Определение 1.13.** Случайный процесс , , называется *внутренне стационарным*, если справедливы следующие равенства:



где  - вариограмма рассматриваемого процесса, .

Приведём основные характеристики стационарного в широком смысле случайного процесса в частотной области.

**Определение 1.14.** Спектральной плотностью стационарного процесса, , называется функция вида  при условии, что ряд справа абсолютно сходится .

**Определение 1.15.** Спектральной плотностью стационарного процесса, , называется функция вида  при условии .

**Определение 1.16.** Спектральной функцией стационарного случайного процесса, , называется функция вида , .

**Определение 1.17.** Спектральной функцией стационарного случайного процесса, , называется функция вида , .

Приведём в приложении A таблицы некоторых функций спектральной плотности и корреляционной функции [4].

# Обзор моделей временного ряда

Математические модели прогнозирования стремятся найти зависимость будущего значения от прошлого внутри самого процесса и на этой зависимости вычислить прогноз. Эти модели универсальны для различных предметных областей, то есть их общий вид не меняется в зависимости от природы временного ряда.

Рассматриваемый в настоящей работе класс временных рядов с регулярными периодическими компонентами является весьма распространенным.

Существует несколько математических моделей и методов, специально разработанных для прогнозирования временных рядов с регулярными периодическими компонентами, например, модель сезонного скользящего среднего (Хольт и Винтере, 1950-1960), сезонная модель ARIMA (Бокс и Дженкинс, 1970), метод гармонического анализа.

**Процесс авторегрессии.**Большинство временных рядов содержат элементы, которые последовательно зависят друг от друга. Такую зависимость можно выразить следующим уравнением:

 = http://statsoft.ru/home/textbook/graphics/xiblue.gif + http://statsoft.ru/home/textbook/graphics/lilphibl.gif1\* (t-1) + http://statsoft.ru/home/textbook/graphics/lilphibl.gif2\* (t-2) + http://statsoft.ru/home/textbook/graphics/lilphibl.gif3\* (t-3) + ... + http://statsoft.ru/home/textbook/graphics/errorblu.gif

Здесь:  
http://statsoft.ru/home/textbook/graphics/xiblue.gif-константа (свободный член),

http://statsoft.ru/home/textbook/graphics/lilphibl.gif1, http://statsoft.ru/home/textbook/graphics/lilphibl.gif2, http://statsoft.ru/home/textbook/graphics/lilphibl.gif3 , …  - параметры авторегрессии,

http://statsoft.ru/home/textbook/graphics/errorblu.gif- белый шум.

Каждое наблюдение есть сумма случайной компоненты (случайное воздействие, errorblu.gif (835 bytes)) и линейной комбинации предыдущих наблюдений.

**Процесс скользящего среднего.** В отличие от процесса авторегрессии, в процессе скользящего среднего каждый элемент ряда подвержен суммарному воздействию предыдущих ошибок. В общем виде это можно записать следующим образом:

 = µ + http://statsoft.ru/home/textbook/graphics/errorblu.gift - http://statsoft.ru/home/textbook/graphics/thetablu.gif1\*http://statsoft.ru/home/textbook/graphics/errorblu.gif(t-1) - http://statsoft.ru/home/textbook/graphics/thetablu.gif2\*http://statsoft.ru/home/textbook/graphics/errorblu.gif(t-2) - http://statsoft.ru/home/textbook/graphics/thetablu.gif3\*http://statsoft.ru/home/textbook/graphics/errorblu.gif(t-3) - ...

Здесь:

µ - константа,

http://statsoft.ru/home/textbook/graphics/thetablu.gif1, http://statsoft.ru/home/textbook/graphics/thetablu.gif2, http://statsoft.ru/home/textbook/graphics/thetablu.gif3 - параметры скользящего среднего,

http://statsoft.ru/home/textbook/graphics/errorblu.gif - белый шум.

Другими словами, текущее наблюдение ряда представляет собой сумму случайной компоненты   (случайное воздействие, errorblu.gif (835 bytes)) в данный момент и линейной комбинации случайных воздействий в предыдущие моменты времени.

**Модель ARIMA.**Общая модель, предложенная Боксом и Дженкинсом (1976) включает как параметры авторегрессии, так и параметры скользящего среднего. Именно, имеется три типа параметров модели: параметры авторегрессии (*p*), порядок разности (*d),*параметры скользящего среднего (*q*). В обозначениях Бокса и Дженкинса модель записывается как ARIMA (*p, d, q*). Например, модель (*0*, *1*,*2*) содержит 0 (нуль) параметров авторегрессии (*p*) и 2 параметра скользящего среднего (*q*), которые вычисляются для ряда после взятия разности с лагом 1..

Для модели ARIMA необходимо, чтобы ряд был стационарным, это означает, что его среднее постоянно, а выборочные дисперсия и автокорреляция не меняются во времени. Поэтому обычно необходимо брать разности ряда до тех пор, пока он не станет[стационарным](http://statsoft.ru/home/textbook/glossary/gloss_s.html#Stationary Series (in Time Series)) (часто также применяют логарифмическое преобразование для стабилизации дисперсии). Число разностей, которые были взяты, чтобы достичь стационарности, определяются параметром *d* . Для того чтобы определить необходимый порядок разности, нужно исследовать график ряда и автокоррелограмму. Сильные изменения уровня (сильные скачки вверх или вниз) обычно требуют взятия несезонной разности первого порядка (лаг=*1*). Сильные изменения наклона требуют взятия разности второго порядка. Сезонная составляющая требует взятия соответствующей сезонной разности. Если имеется медленное убывание выборочных коэффициентов автокорреляции в зависимости от лага, обычно берут разность первого порядка. Чрезмерное количество взятых разностейприводит к менее стабильным оценкам коэффициентов.

**Спектральный и гармонический анализ.** В спектральном анализе исследуются периодические модели данных. Цель анализа - разложить временные ряды с циклическими компонентами на несколько основных синусоидальных функций с определенной длиной волн. В результате успешного анализа можно обнаружить всего несколько повторяющихся циклов различной длины в интересующих вас временных рядах, которые, на первый взгляд, выглядят как случайный шум.

Цель спектрального анализа - разложить ряд на функции синусов и косинусов различных частот, для определения тех, появление которых особенно существенно и значимо. Один из возможных способов сделать это - решить задачу линейной множественной регрессии, где зависимая переменная -наблюдаемый временной ряд, а независимые переменные или регрессоры: функции синусов всех возможных (дискретных) частот. Такая модель линейной множественной регрессии может быть записана как:

= a0 + http://statsoft.ru/home/textbook/graphics/sigmablu.gif [ak\*cos(http://statsoft.ru/home/textbook/graphics/lambdabl.gifk\*t) + bk\*sin(http://statsoft.ru/home/textbook/graphics/lambdabl.gifk\*t)] (для k = 1 до q)

Следующее общее понятие классического гармонического анализа в этом уравнении - http://statsoft.ru/home/textbook/graphics/lambdabl.gif  -это круговая частота, выраженная в *радианах*в единицу времени, т.е. http://statsoft.ru/home/textbook/graphics/lambdabl.gif = 2\*http://statsoft.ru/home/textbook/graphics/pi.gif\*http://statsoft.ru/home/textbook/graphics/nu.gifk, где http://statsoft.ru/home/textbook/graphics/nu.gifk = k/q. Здесь важно осознать, что вычислительная задача подгонки функций синусов и косинусов разных длин к данным может быть решена с помощью множественной линейной регрессии. Заметим, что коэффициенты *ak* при косинусах и коэффициенты *bk* при синусах - это коэффициенты регрессии, показывающие степень, с которой соответствующие функции коррелируют с данными [заметим, что сами синусы и косинусы на различных частотах не коррелированы или, другим языком, ортогональны. Таким образом, мы имеем дело с частным случаем разложения по ортогональным полиномам.] Всего существует *q* различных синусов и косинусов; число функций синусов и косинусов не может быть больше числа данных в ряде. Не вдаваясь в подробности, отметим, если *n -*количество данных, то будет *n/2+1*функций косинусов и *n/2-1*функций синусов. Другими словами, различных синусоидальных волн будет столько же, сколько данных, и можно полностью воспроизвести ряд по основным  функциям. (Заметим, если количество данных в ряде нечетно, то последнее наблюдение обычно опускается. Для определения синусоидальной функции нужно иметь, по крайней мере, две точки: высокого и низкого пика.)

В итоге, спектральный анализ определяет корреляцию функций синусов и косинусов различной частоты с наблюдаемыми данными. Если найденная корреляция (коэффициент при определенном синусе или косинусе) велика, то можно заключить, что существует строгая периодичность на соответствующей частоте в данных.

# Первичный анализ временного ряда

Рассматриваемые данные ― температуры воды на глубине 3 м озера Баторино. Оно расположено в Мядельском районе Минской области. Входит в группу Нарочанских озёр. Характеристики озера Баторино: площадь зеркала — 6.25 км², площадь водосбора — 92.4 км², максимальная глубина — 5.5 м, длина береговой линии — 15 км.

Для изучения температуры озера рассмотрим выборку, полученную от учебно-научного центра "Нарочанская биологическая станция им. Г.Г. Винберга". Полученные данные представляют собой средние за месяц значения температуры с мая по октябрь во временном интервале 1979 – 2012 гг.. В результате имеем выборку размером 204 элемента.

Поставленная задача ― построить модель, которая описывает изменение температурного режима озера. Чтобы оценить точность полученной модели, исключим из выборки значения за последний год и спрогнозируем их.

***Пусть x(t)***, ***t = 1, 2, …, N, N = 198***― рассматриваемая выборка. Значения указанной выборки температуры (Приложение Б, Табл. 1).

Из графиков на рисунке 1 и рисунке 2 мы можем предположить, что ряд демонстрирует две особенности:

* присутствие линейного тренда.
* наличие сезонной компоненты с периодом 6.

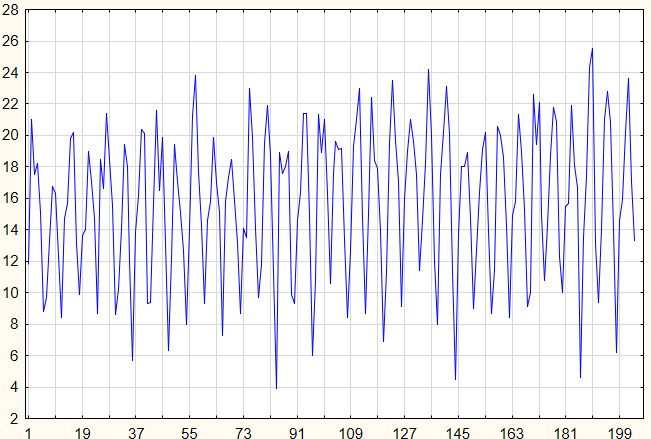


Рисунок 1 ― График исходной выборки.

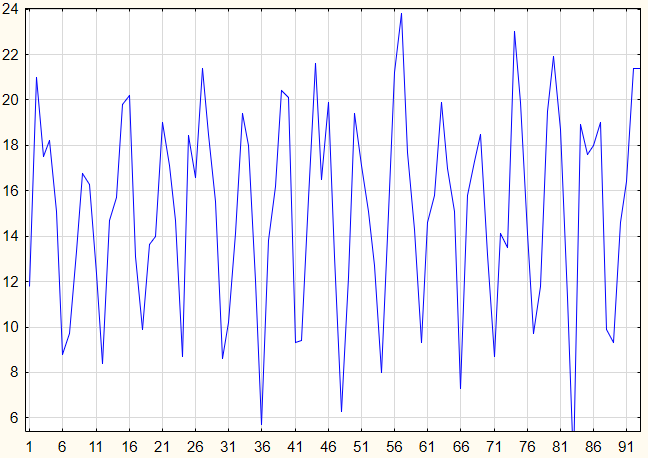


Рисунок 2 ― График половины выборки.

На начальном этапе обработки данных: вычислим наиболее часто используемые описательные статистики [10], укажем в (Приложение Б, Таблица 2), предоставим их:

Таблица 1 ― Основные описательные статистики.

|  |  |
| --- | --- |
| размах (range) | R = 21.60 |
| минимум (minimum) | min = 3.90 |
| максимум (maximum) | max = 25.50 |
| среднее арифметическое (mean) | 15.69 |
| медиана (median) | h = 15.88 |
| средний межквартильный размах (quartile range) | 6.75 |
| верхний квартиль Upper (Quartile) | Q3 = 19.40 |
| нижний квартиль Lower (Quartile) | Q1 = 12.65 |
| персентильный размах | P90 - P10 = 21.30– 9.00 |
| дисперсия(variance) | S 2 = 21.73 |
| среднее квадратическое отклонение (std. dev.) | S = 4.66­ |
| коэффициент вариации (coef. var.) | Cv = 29.71% |
| коэффициентом асимметрии (skewness) | Sk1 = -0.33 |
| коэффициентом эксцесса | Ex = -0.59 |

Выборка однородна, т.к. коэффициент вариации (***Cv = 29.7%****,*) меньше 33,3%. Левосторонняя небольшая асимметрия ***(-0.5 < Sk1 = -0.33 < -0.25***): левый хвост распределения длиннее правого. Отрицательный эксцесс плотности  ***(*Ex *= -*0.59*)***: значения величин не сосредоточены близко к средней, при этом плотность вероятности имеет в окрестности моды более низкую и плоскую вершину, чем плотность нормального закона.

Для физических явлений характерно нормальное распределение. Проверим исходную выборку на нормальность (Приложение Б, Таблица 4). Для проверки гипотезы о нормальности распределения стандартно используем критерии [10]:

* Хи-квадрат *(X2=22.7, p = 0.00239 < 0.05, df = 7 )* => принимаем гипотезу о нормальном распределении.
* Колмогорова-Смирнова (*p > 0.02),* значение d статистики не значительно, поэтому нельзя отвергнуть гипотезу о нормальном распределении,
* Лиллифорс достигнутый уровень значимости *(p < 0.05)* => принимаем гипотезу о нормальном распределении,
* Шапиро-Уилка *(W = 0.97935, p = 0.00425 < 0.05)* => принимаем гипотезу о нормальном распределении.

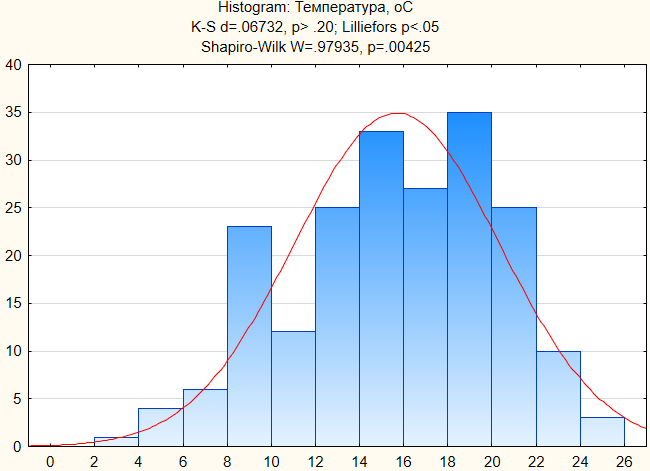


Рисунок 3 ― Гистограмма k = 12, проверка критериев.

На рисунке 3 с гистограммой (Приложение Б, Таблица 3) можно отметить сильную тенденцию группировки величин ряда вокруг среднего. Пожительные и отрицательные отклонения от центра равновероятны. Частота отклонений от среднего быстро падает, когда отклонения от центра становятся большими.

Проверим критерии на выбросы в выборке [10] (Приложение Б, Таблица 5). Тест Граббса (G = 2.53, p = 1) показал, что в выборке есть хотя бы один выброс. Построим диаграмму размаха (Рисунок 4). Наличие 2 выбросов в августе и октябре 2012 года и выброса в октябре 2010 года обусловлено более жаркими годами. Делаем вывод о корректности полученных данных.

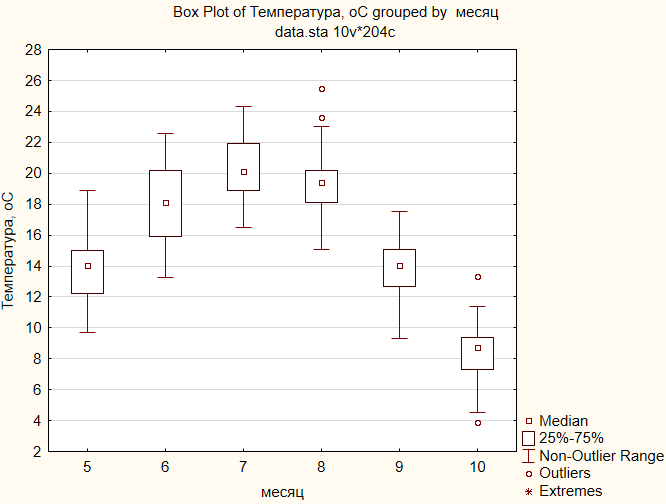


Рисунок 4 ― Диаграмма размаха.

Расширенный тест Дики – Фуллера (Augmented Dickey-Fuller test, ADF) используется для проверки временного ряда на стационарность(Таблица 2):

|  |  |  |
| --- | --- | --- |
| Dickey-Fuller | Lag order | p-value |
| -11.3533 | 6 | 0.01 |

Таблица 2 ― Расширенный тест Дики – Фуллера.

Поскольку p-значение (p-value) мало, то с уровнем значимости 0.01 данный ряд можно считать стационарным.

Вычислим нормированную ковариационную функцию(Приложение Б, Таблица 6). В таблице 1 приложения А приведены примеры нормированных ковариационных функций. Наш дискретных случай схож с непрерывным случаем для гармонического процесса.

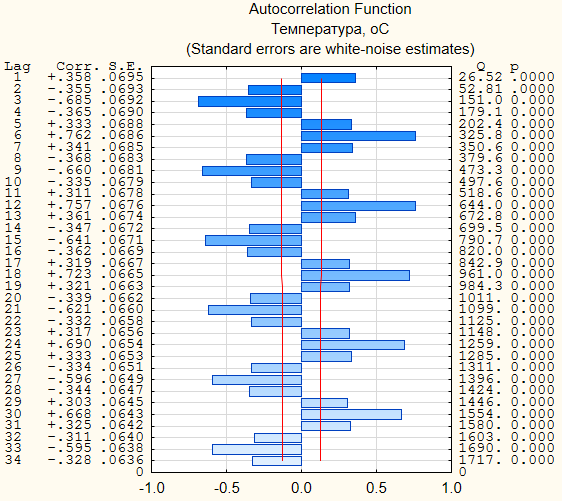


Рисунок 5 ― График нормированной ковариационной функции.

Не затухающий график нормированной ковариационной функции на рисунке 5, характеризует ряд: наличие тренда, сезонности и стационарности.

Рассмотрим ряд с регулярной сезонной составляющей. Для нахождения такой периодичности, воспользуемся спектральным анализом [10] (Приложение Б, Таблица 7). Построим графики периодограммы и спектральной плотности

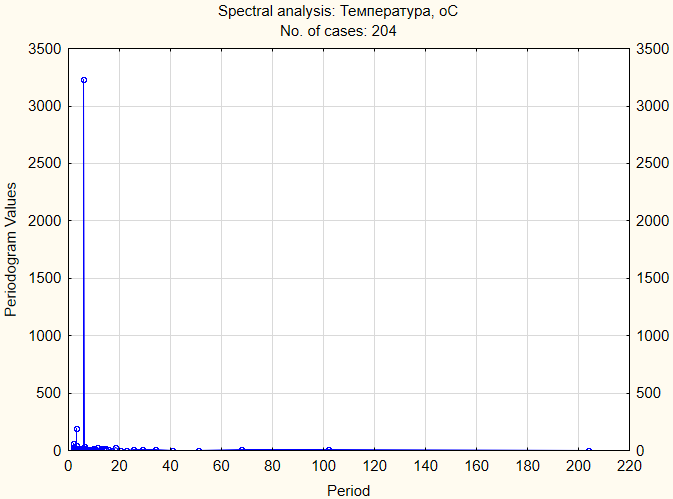


Рисунок 6 ― График для периодограммы.

На рисунке 6 с периодограммой: узкий высокий пик свидетельствует о наличии регулярного цикла. Значительно превосходит другие величина (≈ 3227) с периодом 6. График служит подтверждению того, что в ряду наблюдается тренд с периодом.

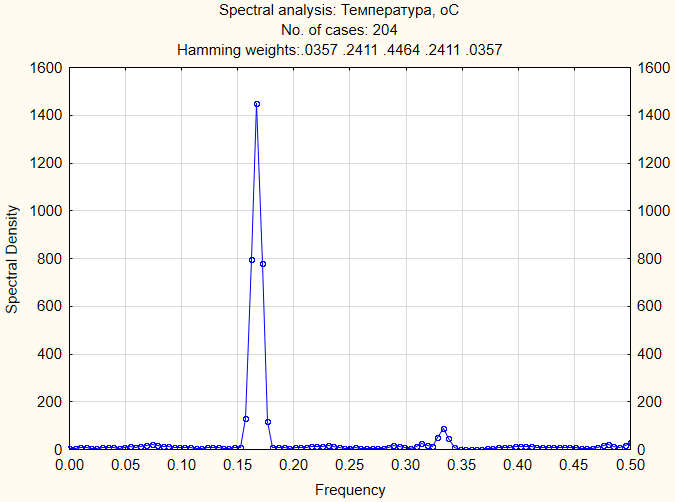


Рисунок 7 ― График спектральной плотности.

На рисунке 7 изображен график спектральной плотности. Наблюдается одно пиковое значение (≈1449.5). Полученная спектральная плотность и спектральная плотность гармонического процесса (Приложение А, Таблица 2) имеют схожий вид.

# Построение моделей временного ряда

Поскольку рассматриваемый ряд относится к стационарным в широком смысле случайным процессам представим его в виде аддитивной модели:

X(t) = u(t) +c(t) + e(t), t = 1, 2, …, N

где линейный тренд - u(t), сезонная компонента - c(t) и нерегулярная компонента - e(t).

Получили следующий линейный тренд:

u(t) = 17.04 + 0.0000065 t,

поскольку тренд близок к константе упростим модель, представим ряд в виде суммы тренд–циклической компоненты (uc(t)) и нерегулярной компоненты (e(t)).:

X(t) = uc(t) + e(t), t = 1, 2, …, N

Проведём гармонический анализ для выделения тренда (Приложение Б, Таблица 7). Для анализа выборки используем за 198 значений. Основные расчёты приведём в приложении. Ниже приведём таблицу 3 с наибольшими дисперсиями гармоник.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Номер гармоники | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Дисперсия | 2.11 | 2.157 | 25.89 | 0.29 | 0.91 | 0.68 | 0.49 | 0.13 |

Таблица 3 ― Дисперсии гармоник

Получили следующие модели циклического тренда:

1) 

2) 

По оценкам ошибок

1), 

2), 

можно сделать вывод о получении "хороших" краткосрочных и среднесрочных прогнозов, причем точность модели 1 выше. Добавление гармоник с меньшей дисперсией приводит к повышению погрешности модели. поэтому далее будем рассматривать модель 1.

Спрогнозируем последний год выборки и приведём график на рисунке 6. Значения рассчитаные моделью 1 и значения исходной выборки укажем в таблице 8 приложения.

Рисунок 1 ― График исходной выборки, полученной модели тренда и случайной компоненты

Построим циклический тренд используя метод ARIMA параметрами (p, d, q)(Ps, Ds, Qs) = (0, 0, 3)(0, 0, 3). В таблице 4 указаны полученные параметры модели. Используем оценки MAD и MSE для проверки точности модели. MAD = 2.09, MSE = 6.63 ― ошибки составляют около 15 % от фактических данных. Модель можно использовать для построения только краткосрочных прогнозов, т.к. при прогнозе на большой период значения модели сходятся к среднему. Запишем полученную формулу тренда:

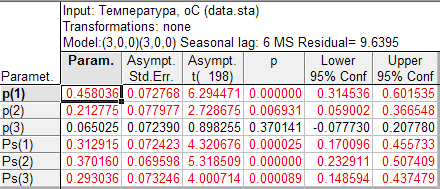


Таблица 4 ― Модель ARIMA

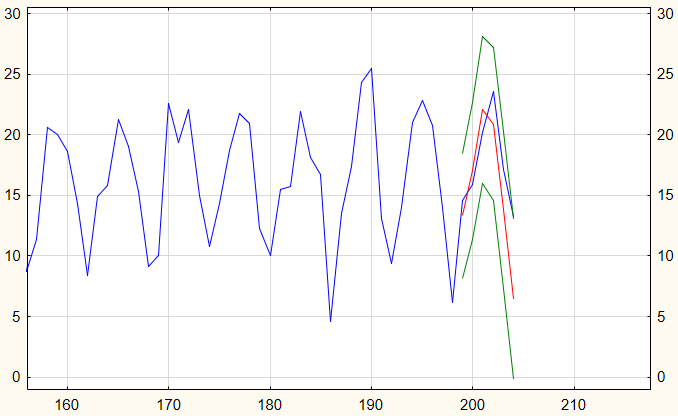


Рисунок 2 ― График исходной выборки и построенной модели на последний год.

Метод ARIMA дал большую точность для построенного прогноза, как видно по рисункам 1, 2 и оценкам ошибок. Этому способствует большая сложность модели и её применение для прогноза на короткий интервал времени. Гармоническая модель может использоваться для прогнозирования на более долгое время.

# ЗАКЛЮЧЕНИЕ

В дипломной работе было проведёно сравнение различных подходов к прогнозированию временных рядов с регулярной периодической компонентой с помощью прикладных программ. Были исследованы изменения температуры воды озера на глубине 3 м.

В качестве исследуемого объекта мы было выбрано озеро Баторино. Исследование было реализовано на основании данных, полученных из учебно-научного центра "Нарочанская биологическая станция им. Г.Г. Винберга ".

Для исследования были вычислены основные описательные характеристики, построены графики для визуальной оценки, вычислены основные статистические критерии. В результате этого было обнаружено: выборка однородна, имеется левосторонняя асимметрия и пологое распределение. Представленный ряд подчиняется нормальному закону распределения.

В результате спектрального и гармонического анализа выявлено наличие регулярной циклической компоненты ряда с периодом 6. Ряд имеет вид дискретного гармонического процесса.

Обоснована и построена аддитивная модель, как сумма тренд–циклической компоненты и нерегулярной компоненты. Построена модель методом гармонического анализа. Визуально по построенному прогнозу и полученным оценкам MAD и MSE получили вывод о валидности построенной модели. Гармонической моделью можно построить прогноз на длительный период времени.

Визуально по построенному прогнозу и полученным оценкам делаем вывод о более высокой точности у модели ARIMA. Модель возможно использовать для построения краткосрочных прогнозов.

# Список литературы

1. Анализ данных на компьютере [Книга] / авт. Тюрин Ю. Н. Макаров А.А.. - Москва : ИНФРА-М, 2003.
2. Бюллетень экологического состояния озер Нарочь, Мястро, Баторино [Книга] / авт. Т. В. Жукова Т. М. Михеева, Р. З. Ковалевская,Е. В. Лукьянова, Л. В. Никитина, О. А. Макаревич, И. В. Савич, и др.. - 2014.
3. Водные ресурсы Национального парка «Нарочанский» [Книга] / авт. А.Г. Аронов Т.И. Аронова, Б.П. Власов, О.С. Ежова, Т.В. Жукова, Р.З. Ковалевская,. - [б.м.] : Рифтур Принт, 2012.
4. Измерение и анализ случайных процессов[Книга] / авт. Бендат Дж., Пирсол А., - Москва : МИР, 1971.
5. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL [Книга] / авт. Вуколов Э. А.. - Москва : ФОРУМ, 2008.
6. СОСТОЯНИЕ И ЗАДАЧИ ОХРАНЫ И ИСПОЛЬЗОВАНИЯ ОЗЕР И ВОДОХРАНИЛИЩ [Конференция] / авт. Румянцев В.А. Филатов Н.Н. , Гронская Т.П. , Догановский А.М..
7. Температурный режим Нарочанских озер на фоне многолетних климатических изменений [Статья] / авт. Жукова Т. В. Радчикова, Н. П. Адамович, Б. В. Михеева, Т. М. Верес, Ю. К. Медвинский, А. Б. // Вестник БГУ. Серия 2. - Минск : БГУ, 2014 г.. - 2. с 25-35.
8. APPLIED MODELING OF HYDROLOGIC TIME SERIES [Книга] / авт. J.D. Salas J. W. Delleur, V.Yevjevich and W.L. Lane. - Michigan : BookCrafters, 1997.
9. Climatic Change, Aquatic Science, Multiple Shifts in Paradigms [Статья] / авт. Gerten Dieter // International Review of Hydrobiology. - 2008 г.. - 4-5 Страницы 395-638.
10. STATISTICA 6. Статистический анализ данных [Книга] / авт. Халавян А. А.. - Москва : БИНОМ, 2007.
11. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов [Книга] / авт. Боровиков В.. - Питер : ПИТЕР, 2003.
12. The American Naturalist [Журнал] / авт. Kingsolver Joel G.. - [б.м.] : The University of Chicago Press, 2009 г.. - Vol. 174, No. 6 (December), pp. 755-768.
13. Time series modelling of water resources and enviromental systems [Книга] / авт. K.W. Hipel A.I. McLeod. - Amsterdam-London-New York-Tokio : Elsevier, 1994.

#### **ПРИЛОЖЕНИЕ А**

Таблица 1 — некоторые нормированные ковариационные функции.

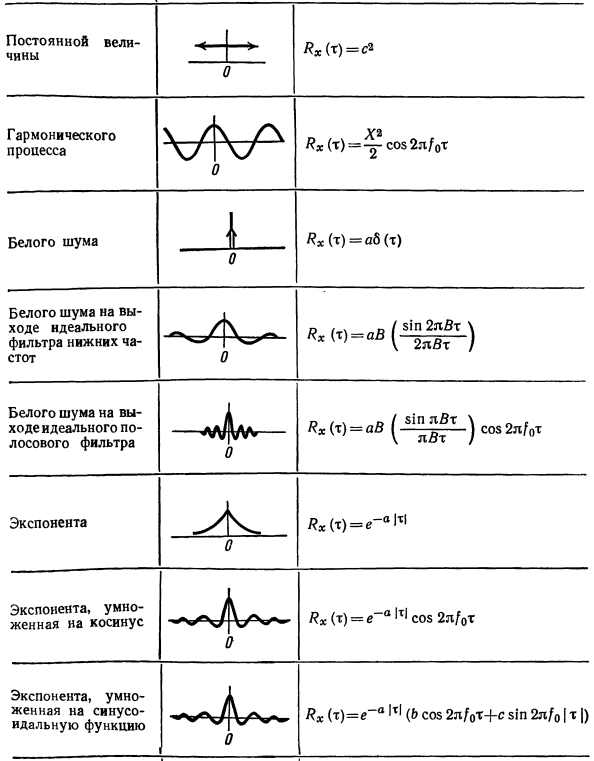
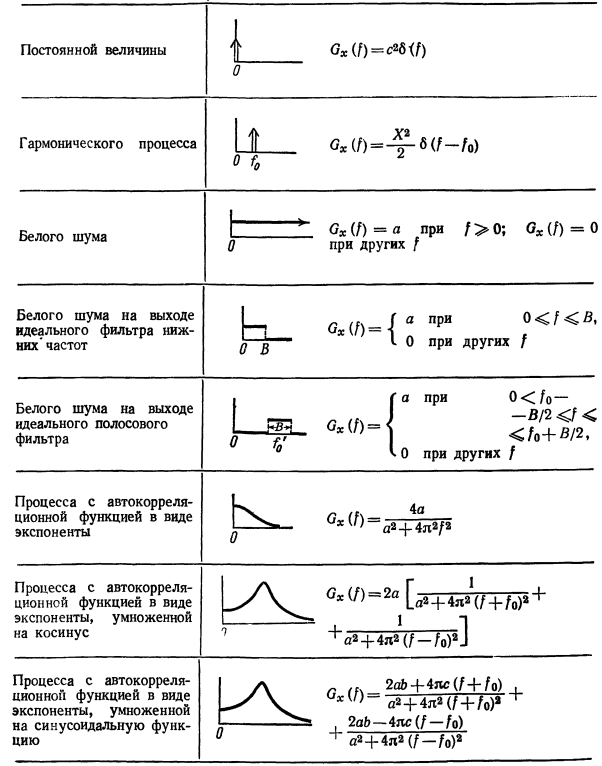


Таблица 2 — некоторые спектральные плотности

­

ПРИЛОЖЕНИЕ Б

Таблица 1 ― Измерения температуры оз. Баторино на глубине 3 м во временном интервале 1979 – 2012 гг. вычисленные средние за месяц значения с мая по октябрь

|  |  |  |
| --- | --- | --- |
| Номер | дата | Температура, оС |
| 1 | 15.05.1979 | 11.8 |
| 2 | 15.06.1979 | 21 |
| 3 | 14.07.1979 | 17.5 |
| 4 | 15.08.1979 | 18.2 |
| 5 | 10.09.1979 | 15.1 |
| 6 | 19.10.1979 | 8.8 |
| 7 | 24.05.1980 | 9.7 |
| 8 | 00.06.1980 | 13.2375 |
| 9 | 06.07.1980 | 16.775 |
| 10 | 18.08.1980 | 16.3 |
| 11 | 00.09.1980 | 12.35 |
| 12 | 16.10.1980 | 8.4 |
| 13 | 14.05.1981 | 14.7 |
| 14 | 16.06.1981 | 15.7 |
| 15 | 17.07.1981 | 19.8 |
| 16 | 17.08.1981 | 20.2 |
| 17 | 16.09.1981 | 13.1 |
| 18 | 13.10.1981 | 9.9 |
| 19 | 19.05.1982 | 13.65 |
| 20 | 21.06.1982 | 14 |
| 21 | 13.07.1982 | 19 |
| 22 | 17.08.1982 | 17.1 |
| 23 | 16.09.1982 | 14.7 |
| 24 | 15.10.1982 | 8.7 |
| 25 | 18.05.1983 | 18.45 |
| 26 | 16.06.1983 | 16.6 |
| 27 | 18.07.1983 | 21.4 |
| 28 | 24.08.1983 | 18.5 |
| 29 | 21.09.1983 | 15.55 |
| 30 | 17.10.1983 | 8.6 |
| 31 | 10.05.1984 | 10.2 |
| 32 | 14.06.1984 | 14.1 |
| 33 | 10.07.1984 | 19.4 |
| 34 | 23.08.1984 | 18 |
| 35 | 17.09.1984 | 12.05 |
| 36 | 18.10.1984 | 5.7 |
| 37 | 12.05.1985 | 13.8 |
| 38 | 14.06.1985 | 16.2 |
| 39 | 15.07.1985 | 20.4 |
| 40 | 14.08.1985 | 20.1 |
| 41 | 26.09.1985 | 9.3 |
| 42 | 16.10.1985 | 9.4 |
| 43 | 14.05.1986 | 15.65 |
| 44 | 17.06.1986 | 21.6 |
| 45 | 19.07.1986 | 16.5 |
| 46 | 14.08.1986 | 19.9 |
| 47 | 00.09.1986 | 13.1 |
| 48 | 20.10.1986 | 6.3 |
| 49 | 16.05.1987 | 12.2 |
| 50 | 17.06.1987 | 19.4 |
| 51 | 14.07.1987 | 17.1 |
| 52 | 12.08.1987 | 15.1 |
| 53 | 14.09.1987 | 12.7 |
| 54 | 19.10.1987 | 8 |
| 55 | 16.05.1988 | 14 |
| 56 | 09.06.1988 | 21.2 |
| 57 | 18.07.1988 | 23.8 |
| 58 | 15.08.1988 | 17.7 |
| 59 | 14.09.1988 | 14.3 |
| 60 | 17.10.1988 | 9.3 |
| 61 | 10.05.1989 | 14.6 |
| 62 | 12.06.1989 | 15.8 |
| 63 | 19.07.1989 | 19.9 |
| 64 | 11.08.1989 | 17 |
| 65 | 11.09.1989 | 15.1 |
| 66 | 16.10.1989 | 7.3 |
| 67 | 07.05.1990 | 15.8 |
| 68 | 12.06.1990 | 17.3 |
| 69 | 10.07.1990 | 18.5 |
| 70 | 00.08.1990 | 15.85 |
| 71 | 10.09.1990 | 13.2 |
| 72 | 15.10.1990 | 8.7 |
| 73 | 13.05.1991 | 14.1 |
| 74 | 03.06.1991 | 13.5 |
| 75 | 16.07.1991 | 23 |
| 76 | 19.08.1991 | 19.9 |
| 77 | 16.09.1991 | 14.1 |
| 78 | 21.10.1991 | 9.7 |
| 79 | 15.05.1992 | 11.8 |
| 80 | 16.06.1992 | 19.5 |
| 81 | 15.07.1992 | 21.9 |
| 82 | 24.08.1992 | 18.7 |
| 83 | 00.09.1992 | 11.3 |
| 84 | 21.10.1992 | 3.9 |
| 85 | 18.05.1993 | 18.9 |
| 86 | 07.06.1993 | 17.6 |
| 87 | 15.07.1993 | 18 |
| 88 | 11.08.1993 | 19 |
| 89 | 15.09.1993 | 9.9 |
| 90 | 12.10.1993 | 9.3 |
| 91 | 24.05.1994 | 14.6 |
| 92 | 13.06.1994 | 16.4 |
| 93 | 12.07.1994 | 21.4 |
| 94 | 12.08.1994 | 21.4 |
| 95 | 26.09.1994 | 14.9 |
| 96 | 19.10.1994 | 6 |
| 97 | 16.05.1995 | 10.8 |
| 98 | 12.06.1995 | 21.3 |
| 99 | 03.07.1995 | 18.9 |
| 100 | 01.08.1995 | 21 |
| 101 | 05.09.1995 | 16.3 |
| 102 | 16.10.1995 | 10.6 |
| 103 | 15.05.1996 | 17.6 |
| 104 | 10.06.1996 | 19.6 |
| 105 | 11.07.1996 | 19.1 |
| 106 | 12.08.1996 | 19.2 |
| 107 | 10.09.1996 | 13.4 |
| 108 | 15.10.1996 | 8.4 |
| 109 | 12.05.1997 | 12.6 |
| 110 | 11.06.1997 | 19.3 |
| 111 | 07.07.1997 | 21 |
| 112 | 11.08.1997 | 23 |
| 113 | 16.09.1997 | 14.8 |
| 114 | 15.10.1997 | 8.7 |
| 115 | 11.05.1998 | 15 |
| 116 | 16.06.1998 | 22.4 |
| 117 | 21.07.1998 | 18.4 |
| 118 | 17.08.1998 | 17.9 |
| 119 | 16.09.1998 | 13.9 |
| 120 | 27.10.1998 | 6.9 |
| 121 | 10.05.1999 | 11.3 |
| 122 | 09.06.1999 | 19.5 |
| 123 | 12.07.1999 | 23.5 |
| 124 | 17.08.1999 | 19.7 |
| 125 | 13.09.1999 | 17.1 |
| 126 | 00.10.1999 | 9.15 |
| 127 | 02.05.2000 | 15.9 |
| 128 | 06.06.2000 | 19 |
| 129 | 01.07.2000 | 21 |
| 130 | 08.08.2000 | 19.6 |
| 131 | 04.09.2000 | 17.5 |
| 132 | 04.10.2000 | 11.4 |
| 133 | 03.05.2001 | 14.6 |
| 134 | 11.06.2001 | 18.2 |
| 135 | 16.07.2001 | 24.2 |
| 136 | 13.08.2001 | 20 |
| 137 | 28.09.2001 | 12 |
| 138 | 00.10.2001 | 7.95 |
| 139 | 15.05.2002 | 17.5 |
| 140 | 12.06.2002 | 20.2 |
| 141 | 17.07.2002 | 23.1 |

Продолжение таблицы 1

|  |  |  |
| --- | --- | --- |
| 142 | 14.08.2002 | 20.1 |
| 143 | 23.09.2002 | 11.4 |
| 144 | 18.10.2002 | 4.5 |
| 145 | 12.05.2003 | 13.7 |
| 146 | 17.06.2003 | 18 |
| 147 | 16.07.2003 | 18 |
| 148 | 11.08.2003 | 18.9 |
| 149 | 09.09.2003 | 14.8 |
| 150 | 14.10.2003 | 9 |
| 151 | 13.05.2004 | 12.7 |
| 152 | 21.06.2004 | 16.4 |
| 153 | 19.07.2004 | 19.1 |
| 154 | 23.08.2004 | 20.2 |
| 155 | 20.09.2004 | 13.9 |
| 156 | 00.10.2004 | 8.7 |
| 157 | 17.05.2005 | 11.4 |
| 158 | 28.06.2005 | 20.6 |
| 159 | 25.07.2005 | 20 |
| 160 | 22.08.2005 | 18.6 |
| 161 | 21.09.2005 | 14.3 |
| 162 | 25.10.2005 | 8.4 |
| 163 | 15.05.2006 | 14.9 |
| 164 | 13.06.2006 | 15.8 |
| 165 | 18.07.2006 | 21.3 |
| 166 | 08.08.2006 | 19 |
| 167 | 12.09.2006 | 15.3 |
| 168 | 18.10.2006 | 9.1 |
| 169 | 10.05.2007 | 10 |
| 170 | 14.06.2007 | 22.6 |
| 171 | 17.07.2007 | 19.4 |
| 172 | 14.08.2007 | 22.1 |
| 173 | 11.09.2007 | 15 |
| 174 | 10.10.2007 | 10.8 |
| 175 | 21.05.2008 | 14.3 |
| 176 | 18.06.2008 | 18.7 |
| 177 | 17.07.2008 | 21.8 |
| 178 | 19.08.2008 | 20.9 |
| 179 | 16.09.2008 | 12.3 |
| 180 | 14.10.2008 | 10 |
| 181 | 12.05.2009 | 15.5 |
| 182 | 18.06.2009 | 15.7 |
| 183 | 21.07.2009 | 21.9 |
| 184 | 19.08.2009 | 18.1 |
| 185 | 10.09.2009 | 16.7 |
| 186 | 19.10.2009 | 4.6 |
| 187 | 04.05.2010 | 13.5 |
| 188 | 02.06.2010 | 17.4 |
| 189 | 12.07.2010 | 24.3 |
| 190 | 05.08.2010 | 25.5 |
| 191 | 06.09.2010 | 13.1 |
| 192 | 06.10.2010 | 9.4 |
| 193 | 11.05.2011 | 14 |
| 194 | 09.06.2011 | 21 |
| 195 | 13.07.2011 | 22.8 |
| 196 | 10.08.2011 | 20.8 |
| 197 | 26.09.2011 | 14 |
| 198 | 19.10.2011 | 6.2 |
| 199 | 15.05.2012 | 14.6 |
| 200 | 07.06.2012 | 15.9 |
| 201 | 04.07.2012 | 20.2 |
| 202 | 07.08.2012 | 23.6 |
| 203 | 06.09.2012 | 17.1 |
| 204 | 04.10.2012 | 13.3 |

Таблица 2 ― Описательные статистики.

|  |  |
| --- | --- |
| Valid N | 204 |
| % Valid obs. | 100.0000 |
| Mean | 15.69050 |
| Confidence | 15.04704 |
| Confidence | 16.33397 |
| Trimmed mean | 15.78648 |
| Winsorized mean | 15.73879 |
| Grubbs Test | 2.529518 |
| p-value | 1.000000 |
| Geometric | 14.86448 |
| Harmonic | 13.85150 |
| Median | 15.87500 |
| Mode | 21.00000 |
| Frequency | 5 |
| Sum | 3200.863 |
| Minimum | 3.900000 |
| Maximum | 25.50000 |
| Lower | 12.65000 |
| Upper | 19.40000 |
| Percentile | 9.000000 |
| Percentile | 21.30000 |
| Range | 21.60000 |
| Quartile | 6.750000 |
| Variance | 21.72647 |
| Std.Dev. | 4.661166 |
| Confidence SD | 4.248499 |
| Confidence SD | 5.163321 |
| Coef.Var. | 29.70693 |
| Standard | 0.326347 |
| Skewness | -0.330373 |
| Std.Err. | 0.170255 |
| Kurtosis | -0.589196 |
| Std.Err. | 0.338910 |

Таблица 3 ― Таблица частот для гистограммы k = 8.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Count | Cumulative Count | Percent | Cumulative Percent | Normal Expected | Cumulative Normal |
| 2.35<x<=5.45 | 3 | 3 | 1.47059 | 1.4706 | 2.41566 | 2.8471 |
| 5.45<x<=8.55 | 11 | 14 | 5.39216 | 6.8627 | 9.84308 | 12.6901 |
| 8.55<x<=11.65 | 29 | 43 | 14.21569 | 21.0784 | 26.25783 | 38.9480 |
| 11.65<x<=14.75 | 40 | 83 | 19.60784 | 40.6863 | 45.88708 | 84.8351 |
| 14.75<x<=17.85 | 43 | 126 | 21.07843 | 61.7647 | 52.55198 | 137.3870 |
| 17.85<x<=20.95 | 49 | 175 | 24.01961 | 85.7843 | 39.44531 | 176.8323 |
| 20.95<x<=24.05 | 26 | 201 | 12.74510 | 98.5294 | 19.40099 | 196.2333 |
| 24.05<x<=27.15 | 3 | 204 | 1.47059 | 100.0000 | 6.24987 | 202.4832 |
| Missing | 0 | 204 | 0.00000 | 100.0000 |  |  |

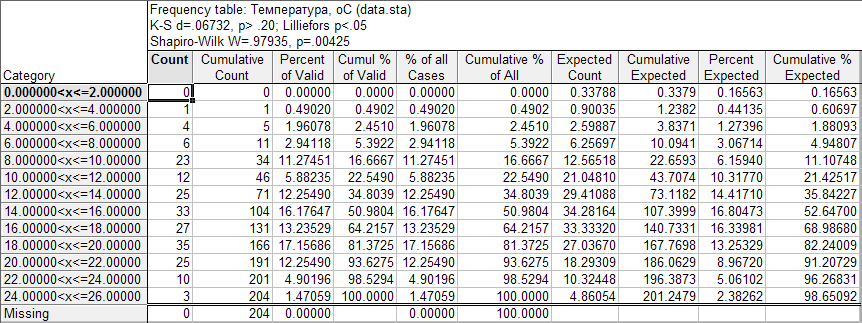
Таблица 4 ― Проверка на нормальность.  


Таблица 5 ― [Тест Граббса](https://ru.wikipedia.org/w/index.php?title=%D0%A2%D0%B5%D1%81%D1%82_%D0%93%D1%80%D0%B0%D0%B1%D0%B1%D1%81%D0%B0&action=edit&redlink=1)*.*

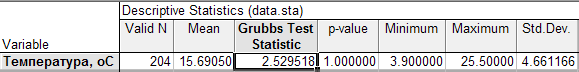


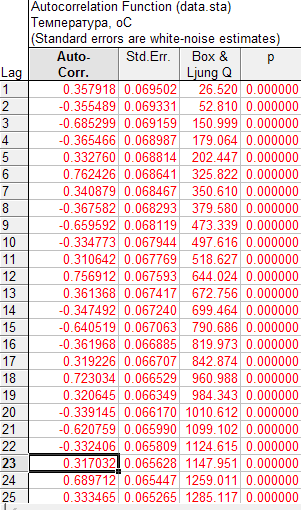
Таблица 6 ―Нормированная ковариационная функция

Таблица 7 ―Спектральный анализ  
  
