**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ**

**Кафедра теории вероятностей и математической статистики**

**Павлов Александр Сергеевич**

# Анализ и прогнозирование гидрологических данных

Дипломная работа

**Научный руководитель:**
*Цеховая Татьяна Вячеславовна*
*доцент кафедры ТВиМС*
*канд. физ.-мат. наук*

**Допущена к защите**
«___» _____ 2015 г.

Минск, 2015

## АННОТАЦИЯ

В курсовом проекте исследована одна из важнейших характеристик любого водоёма — температура воды. проведёны корреляционный и регрессионный анализы, проанализирован ряд остатков, построены модели вариограмм и на их основе вычислены прогнозные значения временного ряда наблюдений с 1975 по 2012 гг. для озера Баторино.

## АННАТАЦЫЯ

У курсавым праекце даследавана адна з найважнейшых характарыстык любога вадаёма — тэмпература вады. Вылічаны апісальныя статыстыкі, прааналізаваны закон размеркавання, праведзены карэляцыйны і рэгрэсійны аналіз, прааналізаваны шэраг рэшткаў, пабудаваны мадэлі варыаграм і на іх аснове вылічаны прагнозныя значэнні часовага шэрагу назіранняў з 1975 па 2012 гг. для возера Баторына.

## ANNOTATION

One of the most important characteristics of any pond — the water temperature — was investigated in the course project. Descriptive statistics were calculated, the distribution was analysed, the correlation and regression analyses were conducted, variogram models and based on them prediction values of time series of observations from 1975 to 2012 for Lake Batorino were computed.

# Реферат

Дипломная работа 35 страниц, 3 главы, 11 рисунков, 7 таблиц, 29 источников, 4 приложения

ВРЕМЕННЫЕ РЯДЫ, R, ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ, КОРРЕЛЯЦИОННЫЙ АНАЛИЗ, РЕГРЕССИОННЫЙ АНАЛИЗ, АНАЛИЗ ОСТАТКОВ, ВАРИОГРАММА, КРИГИНГ.

*Объектом исследования* являются наблюдения температуры воды в озере Баторино в период с 1975 по 2012 гг.

*Цель работы* — анализ, обработка и прогнозирование в современном пакете прикладных программ для статистического анализа R.

В процессе работы проведён сравнительный анализ современных пакетов статистического анализа. При помощи пакета R вычислены и проанализированы описательные статистики, произведена подборка закона распределения, проведёны корреляционный и регрессионный анализы, проанализирован ряд остатков, построены модели вариограмм и на их основе вычислены прогнозные значения.

Полученные результаты могут быть использованы для дальнейшего исследований в различных прикладных областях науки: биологии, химии, гидрологии, — а также, для анализа экологической ситуации в Нарочанском парке и других регионах.

Данная работа может быть продолжена для получения модели, более точно описывающей поведение исходного временного ряда. Полученные в процесса работы алгоритмы исследования могут быть использованы для анализа других аналогичных данных.

# Abstract

Diploma thesis, 35 pages, 3 chapters, 11 figures, 7 tables, 29 sources, 4 appendixes.

TIME SERIES, R, DISCRIPTIONAL STATISTICS, CORRELATIONAL ANALYSIS, REGRESSION ANALYSIS, RESIDUAL ANALYSIS, VARIOGRAMM, KRIGING.

*Object of research* is water temperature observations of Batorino lake in period from 1975 till 2012.

*Research purpose* — analysis, processing and forecasting in modern software package for statistical analysis — R.

During the research was performed comparative analysis of modern packages for statistical research. With help of R package were computed and analysed descriptional statistics, was performed destribution analysis and fitting, were conducted correlational and regression analysis, was performed analysis of residual time series, variogram models and based on them prediction values were computed.

Results of this research could be used for further researches in various applied areas of science: biology, chemistry, hydrology, — and also for analysis of ecology situation at the Narochansky park and other regions.

This research could be continued in case of getting model that will be more accurate in describing source time series. Algorythms that were obtained during the research could be used for analysis other similar data.

# Содержание

# Введение

Работа посвящена обработке, исследованию и статистическому анализу реальных временных рядов. В современных условиях выбор этой направленности соответствует необходимости в проведении анализа наблюдений, полученных в течение длительного времени, с математической и, в частности, статистической точки зрения. Часто наличие даже большого количества информации, полученной в процессе каких-либо наблюдений, не всегда позволяет раскрыть те или иные причины и следствия, имеющие место в конкретном случае. Для выявления всех скрытых проблем и свойств объекта, за которым проводилось наблюдение, необходимо провести всесторонний анализ полученной информации. В свою очередь, математический аппарат и его конкретные прикладные части могут позволить не только проанализировать сложившуюся ситуацию, но и постараться дать некоторый прогноз по состоянию объекта в будущем.

В качестве материала для исследования в данной работе используется база данных с реальными наблюдениями, зафиксированными на озёрах, входящих в Нарочанский национальный парк, за период с 1955 по 2012 годы, полученная от учебно-научного центра «Нарочанская биологическая станция им. Г.Г.Винберга». В представленной базе присутствуют данные, полученные в ходе наблюдений за озёрами Баторино, Нарочь и Мястро. Из них для исследования было выбрано озеро Баторино. Данное озеро является уникальным природным объектом, изучение которого позволит решать экологические проблемы не только в региональном, но и глобальном масштабе. Оно располагается у самой границы города Мядель и, вместе с вышеназванными Нарочью и Мястро, а также озерами Белое и Бледное, входит в состав Нарочанской озёрной группы.

В данной работе исследуемым показателем озера Баторино было выбрано значение температуры воды. Температура воды принадлежит к числу наиболее важных и фундаментальных характеристик любого водоёма. Её изменение во времени является одним из главных факторов, отражающих изменения в окружающей среде. Также нужно отметить, что свойства воды непосредственно зависят от температуры, что делает исследование температуры воды еще более актуальной задачей. Данная характеристика оказывает сильное влияние на плотность воды, растворимость в ней газов, минеральных и органических веществ, в том числе кислорода. Растворимость кислорода и насыщенность воды этим газом — одни из важнейших характеристик для условий обитания в воде живых организмов. В частности, от температуры воды в значительной мере зависит жизнедеятельность рыб: их распределение в водоёме, питание, размножение. К тому же, температура тела рыб, как правило, не превышает температуры окружающей их воды. В то же время, любой водоём как экосистема является средой обитания различных, отличных от рыб, организмов. И поэтому отслеживание всех изменений и влияние этих изменений на их жизнь является крайне важным не только в экологическом смысле, но и в биологическом. Как следствие вышесказанного, изменение температуры с течением времени следует считать одним из важнейших индикаторов изменений, происходящих в экосистеме озера. А исследование данного показателя, в свою очередь, является важнейшим в исследовании различных проблем, возникающих в экосистемах водоёмов. В подтверждение актуальности исследования данной темы можно привести научные работы [1–5], имеющие аналогичное направление.

Среди представленных следует отметить работу [1], где в качестве объекта исследова-

ния рассматривается крупнейшее в мире озеро — Байкал, подробно изучается изменение климата в контексте данного озера в период с 1950 по 2012 гг.

В работе [2] исследуется температура воды Великих озёр в Северной Америке, а также исследуется влияние, оказываемое изменением температуры на рыб, обитающих в этих озёрах.

В [3] исследуется влияние гидрологических, метеорологических и топологических параметров на изменение температуры воды в озере Цибунту (Индонезия) на основе данных с 2008 по 2009 год.

В работе [4] анализируется временной ряд температуры поверхности воды и потоки тепла водоема Итумбиара (Бразилия) в целях улучшения понимания изменений как следствие находящейся там гидроэлектростанции.

В последней упомянутой работе [5] автор исследует на предмет выявления антропогенного влияния на качество воды в крупнейшем озере в Гондурасе — Йоджоа.

В настоящее время, в условиях глобального потепления и крайне нестабильной климатической ситуации, наблюдения за состоянием озёрных экосистем представляют особую ценность как с научной, так и с практической стороны, поскольку только на основе таких наблюдений возможно выделить последствия антропогенного воздействия на фоне изменения природных факторов. А также получить некоторые заключения по экологической обстановке в определенной области.

Основным инструментом анализа данных в работе является пакет **R**. Такой выбор был обусловлен тем, что

- **R** является специализированным языком программирования для статистической обработки данных
- На сегодняшний день **R** — один из самых популярных в статистической среде инструментов анализа данных, имеющий широкую пользовательскую аудиторию, развитую систему поддержки
- Пакет постоянно развивается и дополняется современнейшими средствами, моделями и алгоритмами
- Бесплатен, свободно распространяется и доступен для всех популярных операционных систем
- Обладает развитыми возможностями для работы с графикой

Первичный анализ был также выполнен в пакете **STATISTICA**.

# Глава 1
# Случайный процесс и его характеристики

## 1.1 Случайный процесс. Стационарность

Для введения следующих понятий воспользуемся [6, 7].

Пусть $(\Omega, \mathcal{F}, P)$ — вероятностное пространство, где $\Omega$ является произвольным множеством элементарных событий, $\mathcal{F}$ — сигма-алгеброй подмножеств $\Omega$, и $P$ — вероятностной мерой.

**Определение 1.1.** *Действительным случайным процессом* $X(t) = X(\omega, t)$ называется семейство действительных случайных величин, заданных на вероятностном пространстве $(\Omega, \mathcal{F}, P)$, где $\omega \in \Omega, t \in \mathbb{T}$, где $\mathbb{T}$ — некоторое параметрическое множество.

При $\omega = \omega_0$, $t \in \mathbb{T}$ $X(\omega_0, t)$ является неслучайной функцией временного аргумента и называется *траекторией случайного процесса.*

При $t = t_0$, $\omega \in \Omega$, $X(\omega, t_0)$ является случайной величиной и называется *отсчетом случайного процесса.*

**Определение 1.2.** Если $\mathbb{T} = \mathbb{R} = (-\infty; +\infty)$, или $\mathbb{T} \subset \mathbb{R}$, то $X(t), t \in \mathbb{T}$ называют *случайным процессом с непрерывным временем.*

**Определение 1.3.** Если $\mathbb{T} = \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$, или $\mathbb{T} \subset \mathbb{Z}$, то говорят, что $X(t), t \in \mathbb{T}$, — *случайный процесс с дискретным временем.*

**Определение 1.4.** *n-мерной функцией распределения случайного процесса* $X(t), t \in \mathbb{T}$, называется функция вида

$$F_n(x_1, \dots, x_n; t_1, \dots, t_n) = P\{X(t_1) < x_1, \dots, X(t_n) < x_n\},$$

где $x_j \in \mathbb{R}, t_j \in \mathbb{T}, j = \overline{1, n}$.

**Определение 1.5.** *Математическим ожиданием* случайного процесса $X(t), t \in \mathbb{T}$, называется функция вида

$$m(t) = E\{X(t)\} = \int\limits_{\mathbb{R}} x \, dF_1(x; t), t \in \mathbb{T}.$$

**Определение 1.6.** *Дисперсией* случайного процесса $X(t), t \in \mathbb{T}$ называется функция вида:

$$V(t) = V\{X(t)\} = E\{X(t) - m(t)\}^2 = \int\limits_{\mathbb{R}} (x - m(t))^2 \, dF_1(x; t).$$

**Определение 1.7.** *Ковариационной функцией* случайного процесса $X(t), t \in \mathbb{T}$ называется функция вида:

$$cov(t_1, t_2) = cov\{X(t_1), X(t_2)\} = E\{(X(t_1) - m(t_1))(X(t_2) - m(t_2))\} = \\ = \iint\limits_{\mathbb{R}^2} (x_1 - m(t_1))(x_2 - m(t_2)) \, dF_2(x_1, x_2; t_1, t_2)$$

**Определение 1.8.** *Корреляционной функцией* случайного процесса $X(t), t \in \mathbb{T}$ называется функция вида:

$$corr\{X(t_1), X(t_2)\} = E\{X(t_1)X(t_2)\} = \iint\limits_{\mathbb{R}^2} x_1 x_2 \, dF_2(x_1, x_2; t_1, t_2)$$

*Замечание* 1.1. Имеет место следующее соотношение, связывающее ковариационную и корреляционную функции:

$$corr\{X(t_1), X(t_2)\} = \frac{cov\{X(t_1), X(t_2)\}}{\sqrt{V\{X(t_1)\}V\{X(t_2)\}}},$$

где $X(t), t \in \mathbb{T}$, — случайный процесс.

**Определение 1.9.** Случайный процесс $X(t), t \in \mathbb{T}$, называется *стационарным в широком смысле*, если $\exists E\{x^2(t)\} < \infty, t \in \mathbb{T}$, и

1. $m(t) = E\{x(t)\} = m = const, t \in \mathbb{T}$;
2. $cov(t_1, t_2) = cov(t_1 - t_2), t_1, t_2 \in \mathbb{T}$.

**Определение 1.10.** Случайный процесс $X(t), t \in \mathbb{T}$, называется *стационарным в узком смысле*, если $\forall n \in \mathbb{N}$, $\forall t_1, \ldots, t_n \in \mathbb{T}$, $\forall \tau, t_1 + \tau, \ldots, t_n + \tau \in \mathbb{T}$ выполняется соотношение:

$$F_n(x_1, \ldots, x_n; t_1, \ldots, t_n) = F_n(x_1, \ldots, x_n; t_1 + \tau, \ldots, t_n + \tau).$$

*Замечание* 1.2. Если случайный процесс $X(t), t \in \mathbb{T}$, является стационарным в узком смысле и $\exists E\{x^2(t)\} < \infty, t \in \mathbb{T}$, то он будет стационарным и в широком смысле, но не наоборот.

## 1.2 Вариограмма и внутренне стационарный случайный процесс

**Определение 1.11.** *Вариограммой* случайного процесса $X(t), t \in \mathbb{T}$ называется функция вида

$$2\gamma(h) = V\{X(t+h) - X(t)\}, \ t, h \in \mathbb{T}. \tag{1.1}$$

При этом функция $\gamma(h), h \in \mathbb{T}$, называется *семивариограммой*.

**Определение 1.12.** Случайный процесс $X(t), \ t \in \mathbb{T}$, называется *внутренне стационарным*, если справедливы следующие равенства:

$$E\{X(t_1) - X(t_2)\} = 0, \tag{1.2}$$

$$V\{X(t_1) - X(t_2)\} = 2\gamma(t_1 - t_2), \tag{1.3}$$

где $2\gamma(t_1 - t_2)$ — вариограмма рассматриваемого процесса, $t_1, t_2 \in \mathbb{T}$.

*Замечание* 1.3. Если $X(t), t \in \mathbb{T}$, — гауссовский случайный процесс, то

$$(X(t+h) - X(t))^2 = 2\gamma(h)\chi_1^2,$$

где $\chi_1^2$ — случайная величина, распределенная по закону *хи-квадрат* с одной степенью свободы.

При этом

$$E\{X(t+h) - X(t)\}^2 = 2\gamma(h), \tag{1.4}$$

$$V\{X(t+h) - X(t)\}^2 = 2(2\gamma(h))^2. \tag{1.5}$$

В дальнейшем в данной работе будем рассматривать случайные процессы с дискретным временем.

# Глава 2
# Оценка вариограммы внутренне стационарного гауссовского случайного процесса и ее свойства

Рассмотрим внутренне стационарный гауссовский случайный процесс с дискретным временем $X(t),\ t \in \mathbb{Z}$, нулевым математическим ожиданием, постоянной дисперсией и неизвестной вариограммой.

Наблюдается процесс $X(t),\ t \in \mathbb{Z}$, и регистрируются наблюдения $X(1), X(2), \ldots, X(n)$ в последовательные моменты времени $1, 2, \ldots, n$.

В качестве оценки вариограммы рассмотрим статистику, предложенную Матероном [8]:

$$2\tilde{\gamma}(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(X(t+h) - X(t))^2, \quad h = \overline{0, n-1}, \tag{2.1}$$

при этом положим $\tilde{\gamma}(-h) = \tilde{\gamma}(h), h = \overline{0, n-1}$; $\tilde{\gamma}(h) = 0, |h| \geq n$.

## 2.1 Первые два момента оценки вариограммы

Найдем выражения для первых двух моментов оценки вариограммы (2.1).

**Теорема 2.1.** *Для оценки $2\tilde{\gamma}(h)$, представленной равенством (2.1), имеют место следующие соотношения:*

$$E\{2\tilde{\gamma}(h)\} = 2\gamma(h), \tag{2.2}$$

$$cov(2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)) =$$

$$= \frac{2}{(n-h_1)(n-h_2)}\sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2}(\gamma(t-h_2-s)+\gamma(t+h_1-s)-\gamma(t-s)-\gamma(t+h_1-s-h_2))^2, \tag{2.3}$$

$$V\{2\tilde{\gamma}(h)\} = \frac{2}{(n-h)^2}\sum_{t,s=1}^{n-h}(\gamma(t-h-s)+\gamma(t+h-s)-2\gamma(t-s))^2, \tag{2.4}$$

*где $\gamma(h), h \in \mathbb{Z}$, — семивариограмма процесса $X(t), t \in \mathbb{Z}$, $h, h_1, h_2 = \overline{0, n-1}$.*

*Доказательство.* Вычислим первый момент оценки (2.1), используя свойства математического ожидания:

$$E\{2\tilde{\gamma}(h)\} = E\{\frac{1}{n-h}\sum_{t=1}^{n-h}(X(t+h) - X(t))^2)\} = \frac{1}{n-h}\sum_{t=1}^{n-h}E\{(X(t+h) - X(t))^2\}.$$

Из равенства (1.4) получаем, что

$$E\{2\tilde{\gamma}(h)\} = \frac{1}{n-h}\sum_{t=1}^{n-h}2\gamma(h) = 2\gamma(h).$$

Таким образом, оценка (2.1) является **несмещённой** оценкой вариограммы.

Найдём второй момент оценки вариограммы при различных значениях $h$:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = E\{(2\tilde{\gamma}(h_1) - E\{2\tilde{\gamma}(h_1)\})(2\tilde{\gamma}(h_2) - E\{2\tilde{\gamma}(h_2)\})\} =$$

$$= E\{\frac{1}{n-h_1}\sum_{t=1}^{n-h_1}((X(t+h_1) - X(t))^2 - E\{(X(t+h_1) - X(t))^2\}) \times$$

$$\times \frac{1}{n-h_2}\sum_{s=1}^{n-h_2}((X(s+h_2) - X(s))^2 - E\{(X(s+h_2) - X(s))^2\})\} =$$

$$= \frac{1}{(n-h_1)(n-h_2)}\sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2} cov\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} \qquad (2.5)$$

Из свойства 1.1 корреляции получаем, что

$$cov\{(X(t+h_1)-X(t))^2, (X(s+h_2)-X(s))^2\} = corr\{(X(t+h_1)-X(t))^2, (X(s+h_2)-X(s))^2\} \times$$
$$\times \sqrt{V\{(X(t+h_1) - X(t))^2\}V\{(X(s+h_2) - X(s))^2\}}$$

Принимая во внимание (1.5) и предыдущее соотношение, из (2.5) получаем:

$$cov\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} = \frac{2(2\gamma(h_1))(2\gamma(h_2))}{(n-h_1)(n-h_2)} \times$$
$$\times \sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2} corr\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\}$$

Далее воспользуемся леммой 1 из [9]:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} =$$

$$= \frac{2(2\gamma(h_1))(2\gamma(h_2))}{(n-h_1)(n-h_2)}\sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2}(corr\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\})^2 =$$

$$= \frac{2(2\gamma(h_1))(2\gamma(h_2))}{(n-h_1)(n-h_2)}\sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2}(\frac{cov\{X(t+h_1) - X(t), X(s+h_2) - X(s)\}}{\sqrt{V\{X(t+h_1) - X(t)\}V\{X(s+h_2) - X(s)\}}})^2$$

Воспользовавшись леммой 3 из [9] и определением корреляционной функции, получаем соотношение (2.3):

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} =$$
$$= \frac{2}{(n-h_1)(n-h_2)} \times \qquad (2.6)$$

$$\times \sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2}(\gamma(t-h_2-s) + \gamma(t+h_1-s) - \gamma(t-s) - \gamma(t+h_1-s-h_2))^2, \qquad (2.7)$$

что и требовалось показать.

Отсюда нетрудно получить соотношение (2.4) для дисперсии оценки вариограммы $2\tilde{\gamma}(h)$, если положить $h_1 = h_2 = h$:

$$V\{2\tilde{\gamma}(h)\} = \frac{2}{(n-h)^2}\sum_{t,s=1}^{n-h}(\gamma(t-h-s) + \gamma(t+h-s) - \gamma(t-s) - \gamma(t+h-s-h))^2 =$$

$$= \frac{2}{(n-h)^2}\sum_{t,s=1}^{n-h}(\gamma(t-h-s) + \gamma(t+h-s) - 2\gamma(t-s))^2.$$

□

## 2.2 Асимптотическое поведение оценки вариограммы

Проанализируем асимптотическое поведение моментов второго порядка оценки (2.1).

**Теорема 2.2.** *Если имеет место соотношение*

$$\sum_{h=-\infty}^{+\infty} |\gamma(h)| < +\infty, \tag{2.8}$$

*то*

$$\lim_{n\to\infty} (n - \min\{h_1, h_2\}) cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} =$$

$$= 2 \sum_{m=-\infty}^{+\infty} \gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2, \tag{2.9}$$

$$\lim_{n\to\infty} (n - h) V\{2\tilde{\gamma}(h)\} = 2 \sum_{m=-\infty}^{+\infty} \gamma(m - h) + \gamma(m + h) - 2\gamma(m))^2. \tag{2.10}$$

*где* $\gamma(h), h \in \mathbb{Z}$, — *семивариограмма процесса* $X(t), t \in \mathbb{Z}$, $h, h_1, h_2 = \overline{0, n-1}$.

*Доказательство.* В (2.6) сделаем следующую замену переменных

$$t = t, \quad m = t - s.$$

Получим

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} =$$

$$= \frac{2}{(n - h_1)(n - h_2)} \sum_{t=1}^{n-h_1} \sum_{t-m=1}^{n-h_2} (\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2 \tag{2.11}$$

Таким образом, в зависимости от $h_1$ и $h_2$, возможны два случая: $h_1 > h_2$ и $h_1 < h_2$.

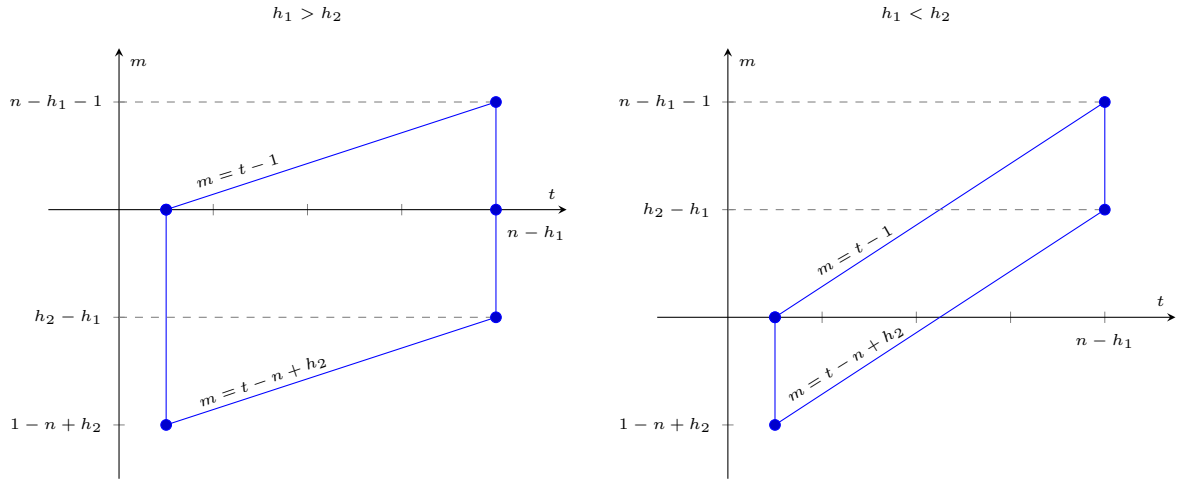


Рисунок 2.2.1 — Области суммирования после замены переменных

Рассмотрим первый случай: $h_1 > h_2$. Поменяем порядок суммирования в (2.11).

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{(n-h_1)(n-h_2)} \times$$
$$\times \Big( \sum_{m=1-n+h_2}^{h_2-h_1-1} \sum_{t=1}^{m+n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=h_2-h_1}^{0} \sum_{t=1}^{n-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=1}^{n-h_1-1} \sum_{t=m+1}^{n-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \Big)$$

Заметим, что выражение под знаком суммы не зависит от $t$, получим:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{(n-h_1)(n-h_2)} \times$$
$$\times \Big( \sum_{m=1-n+h_2}^{h_2-h_1-1} (m+n-h_2)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ (n-h_1) \sum_{m=h_2-h_1}^{0} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=1}^{n-h_1-1} (n-h_1-m)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \Big)$$

Вынесем $n - h_1$ из каждого слагаемого:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_2} \times$$
$$\times \Big( \sum_{m=1-n+h_2}^{h_2-h_1-1} (1 + \frac{h_1+m-h_2}{n-h_1})(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=h_2-h_1}^{0} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=1}^{n-h_1-1} (1 - \frac{m}{n-h_1})(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \Big)$$

Раскроем скобки под знаками сумм:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_2}\Big(\sum_{m=1-n+h_2}^{h_2-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\frac{1}{n-h_1}\sum_{m=1-n+h_2}^{h_2-h_1-1}(h_1+m-h_2)(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\sum_{m=h_2-h_1}^{0}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\sum_{m=1}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{1}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)$$

Приведём подобные:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_2}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\frac{1}{n-h_1}\sum_{m=1-n+h_2}^{h_2-h_1-1}(h_1+m-h_2)(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{1}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)$$

Во втором слагаемом сделаем замену переменных $m = -(m+h_1-h_2)$, получим:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_2}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{1}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(-m-h_1)+\gamma(-m+h_2)-\gamma(-m-h_1+h_2)-\gamma(-m))^2-$$
$$-\frac{1}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)$$

По определению семивариограммы, $\gamma(-h) = \gamma(h)$, тогда

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_2}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{2}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)$$

Аналогично для случая $h_1 < h_2$:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{(n-h_1)(n-h_2)} \times$$
$$\times \Big( \sum_{m=1-n+h_2}^{0} \sum_{t=1}^{m+n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=1}^{h_2-h_1} \sum_{t=m+1}^{m+n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=h_2-h_1+1}^{n-h_1-1} \sum_{t=m+1}^{n-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \Big)$$

Выражение под знаком суммы не зависит от $t$:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{(n-h_1)(n-h_2)} \times$$
$$\times \Big( \sum_{m=1-n+h_2}^{0} (m+n-h_2)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ (n-h_2) \sum_{m=1}^{h_2-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=h_2-h_1+1}^{n-h_1-1} (n-h_1-m)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \Big)$$

Вынесем $n - h_2$ из каждого слагаемого:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_1} \times$$
$$\times \Big( \sum_{m=1-n+h_2}^{0} (1 + \frac{m}{n-h_2})(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=1}^{h_2-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 +$$
$$+ \sum_{m=h_2-h_1+1}^{n-h_1-1} (1 + \frac{h_2-h_1-m}{n-h_2})(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \Big)$$

Раскроем скобки под знаками сумм:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_1}\Big(\sum_{m=1-n+h_2}^{0}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\frac{1}{n-h_2}\sum_{m=1-n+h_2}^{0} m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\sum_{m=1}^{h_2-h_1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\sum_{m=h_2-h_1+1}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\frac{1}{n-h_2}\sum_{m=h_2-h_1+1}^{n-h_1-1}(h_2-h_1-m)(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)$$

Приведём подобные:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_1}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\frac{1}{n-h_2}\sum_{m=1-n+h_2}^{0} m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2+$$
$$+\frac{1}{n-h_2}\sum_{m=h_2-h_1+1}^{n-h_1-1}(h_2-h_1-m)(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)$$

Во втором слагаемом сделаем замену переменных $m=-m$, в третьем $m=m-h_1+h_2$, получим:

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_1}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{1}{n-h_2}\sum_{m=0}^{n-h_2-1} m(\gamma(-m-h_2)+\gamma(-m+h_1)-\gamma(-m)-\gamma(-m+h_1-h_2))^2-$$
$$-\frac{1}{n-h_2}\sum_{m=1}^{n-h_2-1} m(\gamma(m-h_1)+\gamma(m+h_2)-\gamma(m-h_1+h_2)-\gamma(m))^2\Big)$$

По определению семивариограммы, $\gamma(-h)=\gamma(h)$, тогда

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{n-h_1}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{2}{n-h_2}\sum_{m=1}^{n-h_2-1} m(\gamma(m+h_2)+\gamma(m-h_1)-\gamma(m)-\gamma(m-h_1+h_2))^2\Big)$$

**Дальше начинаются рассуждения!!!**

Нужно доказать: при $h_1 > h_2$

$$\lim_{n\to\infty}(n-h_2)cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = 2\sum_{m=-\infty}^{+\infty}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2 \tag{2.12}$$

Изначально для поиска предела нужно домножать $cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\}$ на $(n-h_2)$, потому что эта ковариация была расписана в виде произведения, а предел произведения искать мы не можем.

Дальше смотрим:

$$\begin{aligned}
&\lim_{n\to\infty}(n-h_2)\frac{2}{n-h_2}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2- \\
&\quad-\frac{2}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big) = \\
&\quad=\lim_{n\to\infty}2\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2- \\
&\quad-\frac{2}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big) = \\
&\quad=2\sum_{m=-\infty}^{+\infty}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2- \\
&\quad-4\lim_{n\to\infty}\frac{1}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2
\end{aligned}$$

Последнее равно стоит потому, что в первой сумме от n зависели только пределы суммирования, и они стали бесконечностями. Дальше нужно работать с выражением $-4\lim_{n\to\infty}\frac{1}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2$ и показать, что оно равно 0.

Тут, кажется, можно использовать несколько способов (но ни один у меня не вышел).

Используем способ, как в доказательстве теоремы о свойствах выборочного среднего для временного ряда (Харин-Зуев-Жук, с 409 или ЭУМК по ТВиМС на с. 1840). Идея состоит в том, что если ряд из гамма сходится абсолютно, то любая его часть ограничена и с некоторого места даже мала. Это нас более чем устраивает, но в нашем пределе гамма стоит в квадрате, да еще и домножается на $m$.

Второй способ заключается в том, что мы берем не просто предел, а предел от модуля того же выражения, можем везде его по максимуму раскрывать и ставить знаки $\leq$. И снова пытаться действовать, как в той теореме. И вопрос возникает такой же - у нас выражения с гамма под квадратом и есть сомножитель.

На странице 411 из Харин Зуев Жук (страница 1842 и 1843 ЭУМК) есть интересные сходимости для рядов, влекомые нашим условием. Может, может получиться что-то с ними.

□

# Глава 3
# Обзор реализованного программного обеспечения

Здесь я буду описывать возможности реализованного приложения. Говорить, что может и насколько это удобно.

# Глава 4
# Анализ временного ряда в среде R

## 4.1 Детерминированный подход

### 4.1.1 Описательные статистики и первичный анализ данных

В качестве исходных данных примем выборку из полученной от учебно-научного центра базы данных, путём отбора наблюдений в июле месяце за период с 1975 по 2012 год. Выборка представлена в приложении А в таблице А.1. Графически исходные данные представлены на рисунке 4.1.1.

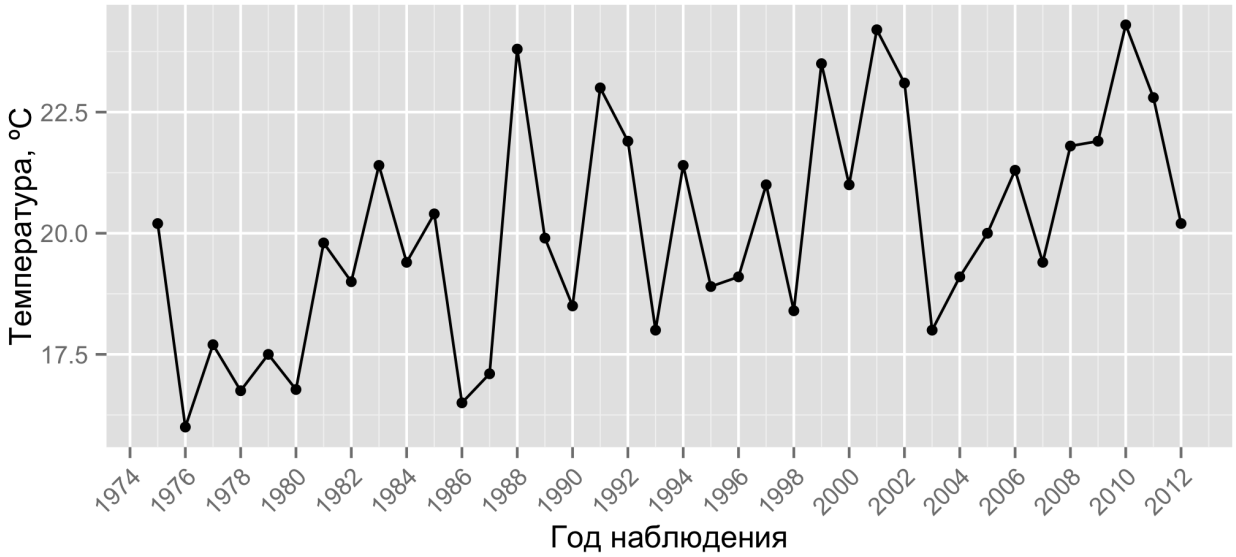


Рисунок 4.1.1 — График исходных данных

Следует отметить, что для непосредственного исследования в данном разделе были использованы наблюдения с 1975 по 2006 год. Наблюдения за 2007-2012 годы были намеренно исключены из исследования в целях дальнейшего оценивания результатов анализа и прогнозирования. Заметим, что работа, представленная в параграфах 4.1.1–4.1.3, была также проделана и для всей выборки. Так как поведение целой выборки сохранилось в уменьшенной, то, без потери общности, будем считать её исходной. Обозначим её $x(t), t = \overline{1,n}$, где $n$ — объём выборки, в данном случае равный 32.

Начнём исследование временного ряда с вычисления описательных статистик. **R** предоставляет в пакете *base* различные функции для расчетов базовых статистик. Также, в различных пакетах можно найти другие интересующие функции, как статистические, так и математические. Но в целях удобства, компактности и контроля за функциональностью на основе [10, 11] мной был написан модуль *dstats*, представленный в приложении Г листинге D.1. Данный модуль позволяет вычислять все рассмотренные в данной работе описательные статистики. Полученные результаты для исходных данных отображены в таблице 1.

Рассмотрим подробнее некоторые полученные статистики.

Как видно из таблицы, *средняя* температура в июле месяце за период с 1975 по 2006 составляет приблизительно 20°C.

| | Значение |
|---|---|
| Среднее | 19.77 |
| Медиана | 19.60 |
| Нижний квартиль | 18.00 |
| Верхний квартиль | 21.33 |
| Минимум | 16.00 |
| Максимум | 24.20 |
| Размах | 8.20 |
| Квартильный размах | 3.33 |
| Дисперсия | 5.12 |
| Стандартное отклонение | 2.26 |
| Коэффициент вариации | 25.92 |
| Стандартная ошибка | 0.40 |
| Асимметрия | 0.30 |
| Ошибка асимметрии | 0.41 |
| Эксцесс | -0.75 |
| Ошибка эксцесса | 0.81 |

Таблица 1 — Описательные статистики для наблюдаемых температур.

*Коэффициент вариации* в нашем случае равен 25.92%. Из этого следует, что выборку можно считать однородной, так как полученное значение является меньшим 33% [10].

*Коэффициент асимметрии* — мера симметричности распределения. Полученное значение: 0.30. Данное значение говорит о незначительной правосторонней асимметрии распределения. То есть о том, что выборочное распределение можно считать близким к нормальному [12].

*Коэффициент эксцесса* в рассматриваемом случае равен $-0.746$. Так как коэффициент эксцесса нормального распределения равен 0, то в данном случае можно говорить о пологости пика распределения выборки по отношению к нормальному распределению [12].

С помощью тестовых статистик для коэффициента асимметрии и эксцесса [11, с.85-89], проверим значимость полученных значений для генеральной совокупности. Для этого в модуле *dstats* мной реализованы функции *dstats.test.skew* и *dstats.test.kurtosis*:

Полученная тестовая статистика для коэффициента асимметрии:

$$Z_{A_S} = \frac{A_S}{SES} = 0.723.$$

Данное значение попадает под случай $|Z_{A_S}| \leq 2$, а значит, выборочный коэффициент асимметрии не является значимым. Из чего, в свою очередь, следует, что по нему нельзя судить о коэффициенте асимметрии генеральной совокупности [11, с.85].

Полученная тестовая статистика для коэффициента эксцесса:

$$Z_K = \frac{K}{SEK} = -0.922.$$

Данное значение попадает под случай $|Z_K| \leq 2$, а значит, в данном случае выборочный коэффициент эксцесса не является значимым и нельзя ничего сказать о коэффициенте эксцесса генеральной совокупности [11, с.89].

Из полученных результатов следует отметить, что коэффициенты асимметрии и эксцесса, указывают на некоторое отклонение выборочного распределения от нормального закона. Но при этом, из-за недостаточного объёма выборки, по этим коэффициентам нельзя судить о соответствующих коэффициентах генеральной совокупности.

В **R** можно найти различные пакеты, позволяющие строить разнообразные гистограммы, диаграммы рассеяния, вероятностные графики, линейные графики, диаграммы диапазонов, размахов, круговые диаграммы, столбчатые диаграммы, последовательные графики и т.д., позволяющие увидеть специфику данных. В процессе работы в пакете **R** использовались источники [13–15].

С помощью функции пакета *ggplot2* построим гистограмму для отображения вариационного ряда исходных данных [15]. Гистограммы позволяют увидеть, как распределены значения переменных по интервалам группировки, то есть как часто переменные принимают значения из различных интервалов. А также, что бывает более важным, позволяет сделать предположение о разновидности распределения. Для вычисления интервалов разбиения воспользуемся *формулой Стерджеса*. Из [16] количество интервалов разбиения рассчитывается по формуле:

$$k = \lceil log_2 n \rceil + 1 = \lceil log_2 32 \rceil + 1 = 6. \tag{4.1}$$

Так как по гистограмме можно визуально предположить близость выборочного распределения к нормальному распределению, нанесём на график кривую плотности нормального распределения (рисунок 4.1.2). Проанализируем эту гистограмму. Во-первых,

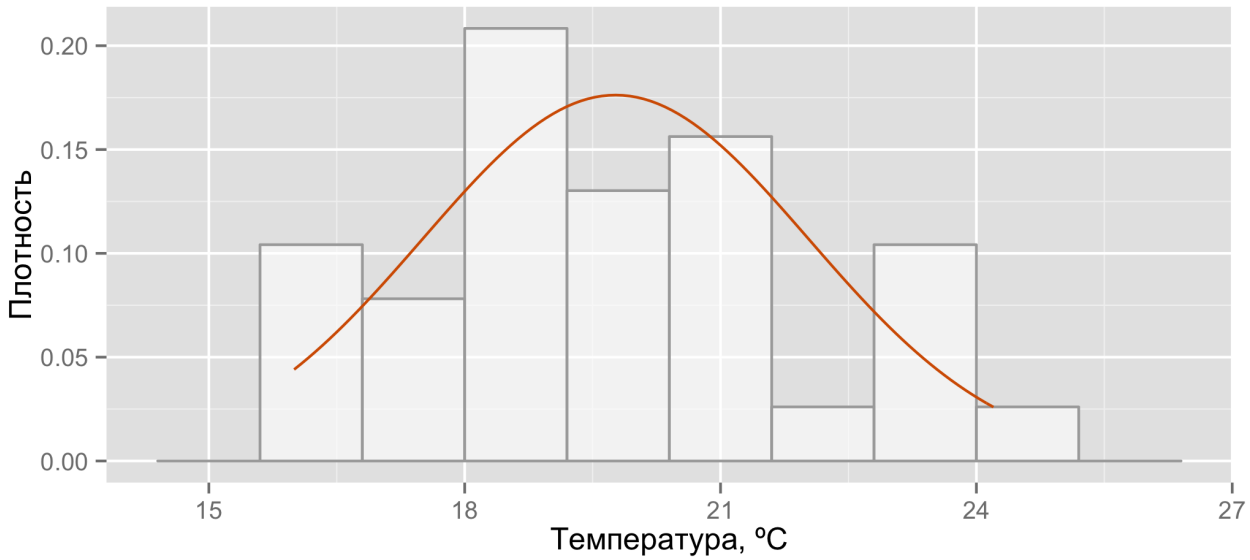


Рисунок 4.1.2 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения $\mathcal{N}(19.77, 5.12)$

на ней наглядно представлены показатели асимметрии и эксцесса, полученные на этапе вычисления описательных статистик. Таким образом показывает отношение выборочного распределения к нормальному с параметрами $\mathcal{N}(19.77, 5.12)$.

Следует отметить согласованность полученных описательных статистик с полученной гистограммой. Во-первых, по коэффициенту асимметрии мы предположили о близости распределения к симметричному. Это подтверждается гистограммой: на ней можно заметить небольшую скошенность вправо, что также согласовывается со знаком коэффициента. Во-вторых, коэффициент эксцесса указывал на пологость пика распределения. Данное заключение подтверждается кривой плотности — она имеет чуть более растянутую колоколообразную форму.

Другим часто используемым графическим способом проверки характера распределения данных является построение т.н. *графиков квантилей* (*Q-Q plots*, *Quantile-Quantile*

*plots*). На таких графиках изображаются квантили двух распределений — эмпирического (т.е. построенного по анализируемым данным) и теоретически ожидаемого нормального распределения. При нормальном распределении проверяемой переменной точки на графике квантилей должны выстраиваться в прямую линию, исходящую под улом 45 градусов из левого нижнего угла графика. Графики квантилей особенно полезны при работе с небольшими по размеру совокупностями, для которых невозможно построить гистограммы, принимающие какую-либо выраженную форму.

В ходе данной работы была написана функция *ggqqp*, с помощью которой построен график 4.1.3. На этом графике можно визуально обнаружить аномальное положение наблю-

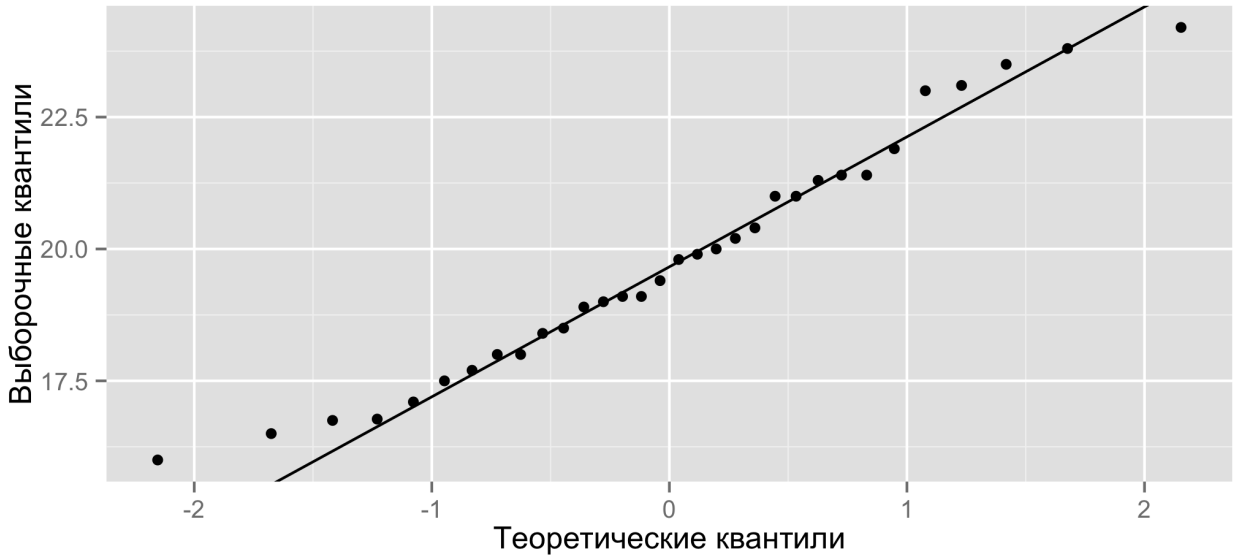


Рисунок 4.1.3 — График квантилей для наблюдаемых температур

даемых значений по отношению к нормальному распределению. В данном случае отклонения можно наблюдать на концах рассматриваемого промежутка. Остальные значения образуют отчетливую прямую. Это следует интерпретировать как близость выборочного распределения к нормальному с параметрами $\mathcal{N}(19.77, 5.12)$.

Далее следует проверить полученные результаты и предположения с помощью некоторых формальных тестов. Существует целый ряд статистических тестов, специально разработанных для проверки нормальности выборочного распределения. В общем виде проверяемую при помощи этих тестов нулевую гипотезу можно сформулировать следующим образом: “Анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение”. Если получаемая при помощи того или иного теста вероятность ошибки $P$ оказывается меньше некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется.

В **R** реализованы практически все имеющиеся тесты на нормальность — либо в виде стандарных функций, либо в виде функций, входящих в состав отдельных пакетов. Примером базовой функции является *shapiro.test()*, при помощи которой можно выполнить широко используемый *тест Шапиро-Уилка* [17]. Из полученных в **R** результатов, статистика Шапиро-Уилка $W = 0.97$. Вероятность ошибки $p = 0.43 > 0.05$, а значит нулевая гипотеза не отвергается [18]. Следовательно опровергнуть предположение на основе данного теста нельзя.

Попробуем опровергнуть наше предположение на основе проверки критерия $\chi^2$ Пирсона [19]. Для этого воспользуемся пакетом *nortest* и функцией *pearson.test*. Из полученных в **R** результатов, статистика $\chi^2$ Пирсона $P = 2.00$. Вероятность ошибки $p = 0.85 > 0.05$,

а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста также нельзя. Проверим критерий: примем уровень значимости $\alpha = 0.05$, тогда из таблицы распределения $\chi^2$ найдём критическое значение критерия $P_{\text{кр}}(\alpha, k) = 43.8$. Отсюда следует, что

$$P < P_{\text{кр}}.$$

А значит, нулевую гипотезу при уровне значимости $\alpha = 0.05$ не отвергаем и подтверждаем сделанный вывод на основании вычисленной вероятности ошибки.

Воспользуемся для тех же целей критерием Колмогорова–Смирнова [20]. Как в предыдущем случае воспользуемся представленной в пакете *nortest* функцией *ks.test*. Из полученных в **R** результатов, статистика Колмогорова-Смирнова $D = 0.087$. Вероятность ошибки $p = 0.97 > 0.05$, а значит нулевую гипотезу отвергнуть нельзя. Следовательно опровергнуть предположение о нормальности, как и в предыдущих случаях, также нельзя. Проверим критерий: примем так же уровень значимости $\alpha = 0.05$, тогда критическое значение $D_{\text{кр}}(\alpha) = 1.358$. Следовательно,

$$D < D_{\text{кр}}(\alpha),$$

и подтверждаем сделанные ранее заключения: нельзя отвергнуть нулевую гипотезу о нормальности выборочного распределения.

На данном этапе по полученным ранее результатам возникли подозрения о выбросах в исходной выборке. Выявление таких аномальных значений важно, так как их наличие, как правило, сильно влияет на всю выборку, в частности, на коэффициент корреляции. Проверим наличие выбросов с помощью статистических критериев. Для этих целей воспользуемся критерием Граббса [21]. Данный основан на предположении о нормальности исходных данных. То есть, перед применением данного критерия необходимо убедиться, что данные могут быть в разумных пределах аппроксимированы нормальным распределением [22]. Поскольку ранее высказано предположение о нормальности, воспользуемся им для определения наличия выбросов.

Полученные результаты проверки критерия Граббса: статистика $G = 1.96$, вероятность ошибки $p$-value $= 0.72$ — что однозначно говорит нам о том, что следует отклонить альтернативную гипотезу $H_1$ и принять гипотезу $H_0$. Другими словами, это говорит о том, что в исходной выборке нету выбросов. А значит выборка однородна. Таким образом, подозрения о выбросах не подтвердились проверкой критерия.

В соответствии с результатами проверки критериев и на основе построенных гистограммы и графика квантилей, можно сделать заключение о том, что распределение температуры воды озера Баторино в июле 1975–2009 годов является близким к нормальному закону распределения с параметрами $\mathcal{N}(19.77, 5.12)$. При этом, обнаружены отклонения от нормальности, описываемые коэффициентами асимметрии и эксцесса. Следует также отметить, что эквивалентные результаты были получены и для всей выборки, до исключения последних наблюдений. При этом, отклонение от нормальности было менее выраженным. Таким образом, отклонение от нормальности можно считать следствием потери информации при исключении наблюдений из исходной выборки.

### 4.1.2 Корреляционный анализ

Исследуем теперь зависимость температуры воды от времени, построив диаграмму рассеяния и вычислив коэффициент корреляции соответствующих переменных.

Диаграммы рассеяния используются для визуального исследования зависимости между двумя переменными. Если переменные сильно связаны, то множество точек данных

принимает определённую форму. С помощью таких диаграмм можно наглядно изучить знак коэффициента корреляции. Если точки на диаграмме расположены хаотически, то это говорит о независимости рассматриваемых переменных. Если с ростом переменной $t$ возрастает переменная $x$ то имеет место положительная корреляция. Если же с ростом переменной $t$ переменная $x$ убывает, то это указывает на отрицательную корреляцию.

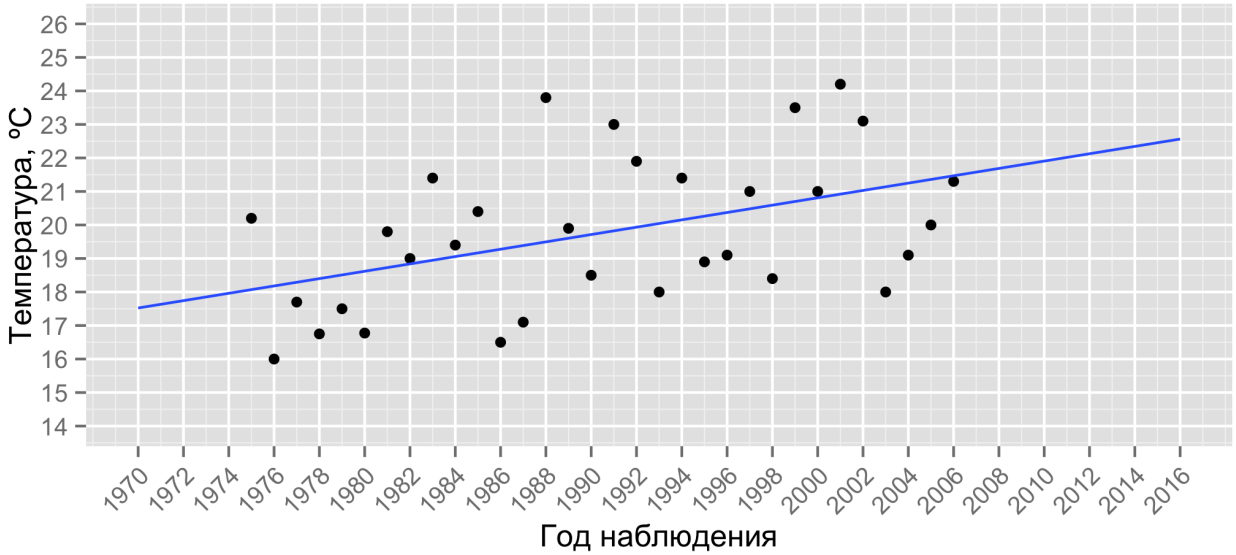


Рисунок 4.1.4 — Диаграмма рассеяния

Из рисунка 4.1.4 видно, что точки образуют своеобразное «облако», ориентированное по вверх, то есть присутствует некая зависимость между рассматриваемыми переменными. Также, данная диаграмма наглядно показывает силу этой зависимости: так как точки не образуют чёткой формы, а разбросаны относительно линии, то можно говорить о наличии умеренной корреляции.

Проверим полученные результаты подробнее. Из расчётов в **R**, коэффициент корреляции $r_{xt} = 0.454$. Этим подтверждаются наши выводы из диаграммы рассеяния о положительной корреляции, поскольку полученный коэффициент корреляции является положительным и присутствует умеренная зависимость: $r_{xt} \approx 0.5$.

Проверим значимость полученного выборочного коэффициента корреляции с помощью критерия Стьюдента:

$$T_{\text{набл}} = \frac{r_{xt}\sqrt{n-2}}{\sqrt{1-r_{xt}^2}} \approx 3.13.$$

Рассмотрим уровень значимости $\alpha = 0.05$. Число степеней свободы $k = n - 2 = 30$. Тогда из таблицы критических точек распределения Стьюдента $t_{\text{кр}}(\alpha, k) \approx 1.70$. Следовательно,

$$T_{\text{набл}} > t_{\text{кр}}(\alpha, k).$$

Значит нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности следует отклонить [10].

Также оценим значимость с помощью возможностей пакета **R** и функции *cor.test*. Представленная функция позволяет с помощью различных методов выполнять проверки значимости выборочного коэффициента корреляции. Воспользуемся проверкой теста методом Пирсона. Из результатов её выполнения статистика $t = 2.79$, количество степеней свободы $df = 30$ и вероятность ошибки $p = 0.009 < 0.05$, следовательно это говорит о том, что необходимо отвергнуть гипотезу $H_0 : r = 0$.

Результаты обоих подходов в проверке значимости совпали. Другими словами, выборочный коэффициент значимо отличается от нуля, т.е. температура воды и время при уровне значимости $\alpha = 0.05$ имеют зависимость.

Следовательно, в рассматриваемом случае можно говорить о присутствии значимой корреляции между температурой воды в озере Баторино и временем. Что говорит росте температуры окружающей среды с момента начала наблюдений.

### 4.1.3 Регрессионный анализ

Для введения последующих понятий анализа временных рядов воспользуемся [23].

В отличие от анализа случайных выборок, анализ временных рядов основывается на предположении, что последовательные значения в исходных данных наблюдаются через равные промежутки времени. Во временных рядах выделяют три составляющие:

1. *Тренд (тенденция развития)* — эволюционная составляющая, которая характеризует общее направление развития изучаемого явления и связанна с действием долговременных факторов развития.

2. *Циклические, сезонные колебания* — это составляющие, которые проявляются как отклонения от основной тенденции развития изучаемого явления, и связанны с действие краткосрочных, систематических факторов развития.

3. *Нерегулярная случайная составляющая (ошибка)*, являющаяся результатом действия второстепенных факторов развития.

Первые два типа компонент представляют собой детерминированные составляющие. Случайная составляющая образована в результате суперпозиции некоторого числа внешних факторов.

По типу взаимосвязи вышеперечисленных составляющих ряда динамики можно построить следующие модели временных рядов:

- Аддитивная модель: $x = y + k + s + \varepsilon$;

- Мультипликативная модель: $x = y \times k \times s \times \varepsilon$,

где $y, k, s, \varepsilon$ — тренд, циклическая, сезонная и нерегулярная составляющие соответственно.

Аддитивной модели свойственно то, что характер циклических и сезонных колебаний остаётся постоянным. В мультипликативной модели характер циклических и сезонных колебаний остаётся постоянным только по отношению к тренду (т.е. значения этих составляющих увеличиваются с возрастанием значений тренда).

По причине того, что в данном случае мы рассматриваем один месяц в году на протяжении длительного периода, будем считать, что в рассматриваемом временном ряде циклическая и сезонная составляющие отсутствуют.

При проведении корреляционного анализа, на графике 4.1.4 был замечен явно выраженный линейный рост значений со временем. Что впоследствии было подтверждено критериями. Из этого следует, что уравнение тренда имеет вид:

$$y(t) = at + b,$$

где $a, b \in \mathbb{R}, t = \overline{0, n-1}$ – некоторые коэффициенты, $n$ — объем выборки.

Продолжая рассуждение, как наблюдение из графика, можно отметить, что не происходит увеличения амплитуды колебаний с течением времени. А значит, искомая модель

является аддитивной. Из всего вышесказанного можно заключить, что модель исходного временного ряда имеет вид:

$$x = y + \varepsilon,$$

где $y$ – тренд, $\varepsilon$ - нерегулярная составляющая.

В **R** реализованы функции, позволяющие подгонять линейные модели к исследуемым данным [24]. Одной из таких функций является *lm(Fitting Linear Model)* [13, с.178]. Она позволяет получить коэффициенты линии регрессии. Таким образом, можно вычислить одну из искомых компонент – тренд. И как следствие, после его удаления из исходных данных, получим нерегулярную составляющую $\varepsilon(t)$. Коэффициенты, полученные с помощью данной функции представлены в (4.2).

$$a = 0.11, \quad b = 18. \tag{4.2}$$

Следует отметить, что в пакете **STATISTICA** похожая процедура была проведена для всей выборки с помощью инструмента *Trend Subtract*, результаты которой согласуются с полученными в **R** коэффициентами.

Таким образом получена линейная модель, описывающая тенденцию развития:

$$y(t) = at + b = 0.11t + 18 \tag{4.3}$$

На основе полученной линейной модели (4.3), построим ряд остатков, удалив тренд из исходного ряда. Полученный ряд представлен в приложении В в таблице В.1 и графически на рисунке 4.1.5.

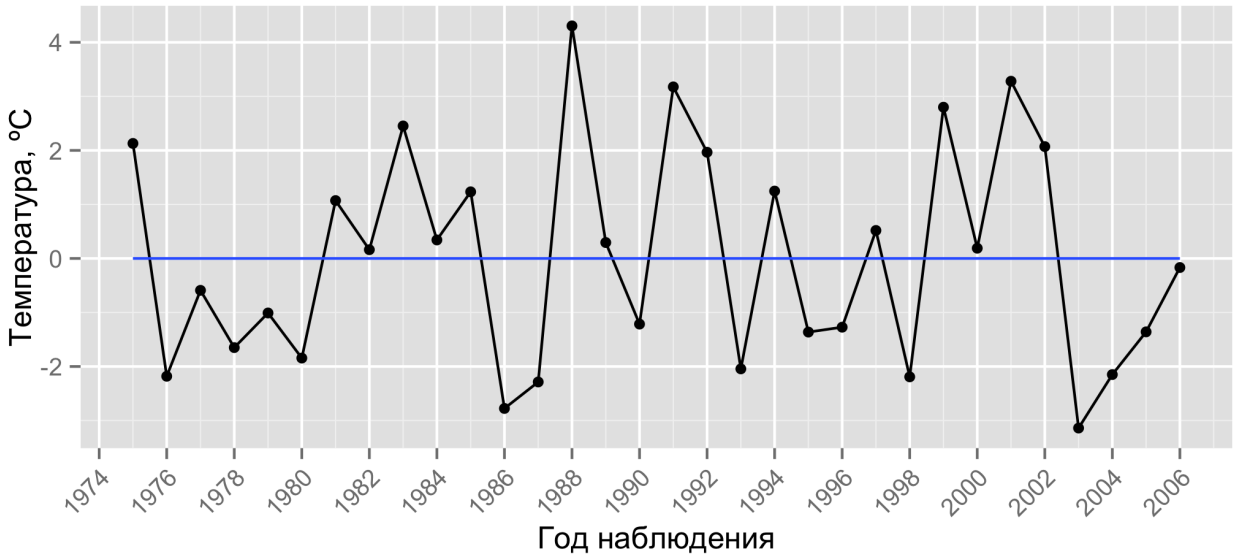


Рисунок 4.1.5 — Нерегулярная составляющая $\varepsilon(t)$

Проведём анализ полученной регрессионной модели. Для этого проверим значимость полученных коэффициентов регрессии и оценим адекватность вычисленной регрессионной модели.

Рассчитаем вспомогательные величины, воспользовавшись [23]. Дисперсия отклонения

$$\sigma_\varepsilon^2 \approx 4.07,$$

стандартные случайные погрешности параметров $a, b$:

$$\sigma_a \approx 0.0393, \quad \sigma_b \approx 0.745.$$

Воспользуемся критерием значимости коэффициентов линейной регрессии [10]. Примем уровень значимости $\alpha = 0.05$, тогда

$$T_a = 2.79, \quad T_b = 24.1.$$

Число степеней свободы $k = 30$, $t_{\text{кр}}(k, \alpha) = 1.7$.

- $|T_a| > t_{\text{кр}} \Rightarrow$ коэффициент $a$ значим.
- $|T_b| > t_{\text{кр}} \Rightarrow$ коэффициент $b$ значим.

Следовательно, при уровне значимости $\alpha = 0.05$, коэффициенты линейной регрессии являются значимыми.

Оценим адекватность полученной регрессионной модели. Дисперсия модели:

$$\overline{\sigma^2} \approx 1.02.$$

Остаточная дисперсия:

$$\overline{D} \approx 3.94.$$

Воспользуемся F-критерием Фишера. Пусть уровень значимости $\alpha = 0.05$,

$$F_{\text{крит}} \approx 7.79,$$

при степенях свободы $v_1 = 1, v_2 = 30, F_{\text{табл}}(v_1, v_2, \alpha) = 4.17$.

$$F_{\text{крит}} > F_{\text{табл}}.$$

Следовательно, при уровне значимости $\alpha = 0.05$, регрессионная модель является адекватной.

Рассчитаем коэффициент детерминации:

$$\eta^2_{x(t)} \approx 0.2.$$

Проверим отклонение от линейности: $\eta^2_{x(t)} - r^2_{xt} \approx -0.00644 \leq 0.1$. Следовательно отклонение от линейности незначительно. Но при этом коэффициент детерминации оказался не высоким($< 0.7$), это говорит о том, что построенная регрессионная модель не описывает в достаточной мере поведение временного ряда. Это, в свою очередь, может значить, что изменение температуры зависит не только от времени, но ещё и от каких-то других, неучтённых, факторов.

Тем не менее, попробуем построить прогноз по полученной модели. Вычисленные прогнозные значения на 2007-2012 годы для сравнения отображены в таблице 2:

| | Год | Актуальное | Прогнозное |
|---|---|---|---|
| 1 | 2007 | 19.40 | 18.07 |
| 2 | 2008 | 21.80 | 18.18 |
| 3 | 2009 | 21.90 | 18.29 |
| 4 | 2010 | 24.30 | 18.40 |
| 5 | 2011 | 22.80 | 18.51 |
| 6 | 2012 | 20.20 | 18.62 |

Таблица 2 — Сравнение прогнозных значений (тренда)

Имеющееся отклонение прогнозов от реальных данных ещё раз подтверждает, что построенная модель временного ряда обладает невысокой точностью. И поэтому необходимо её улучшать другими методами.

### 4.1.4 Анализ остатков

Проанализируем полученную на этапе регрессионного анализа нерегулярную составляющая $\varepsilon$. Для этого проверим свойства, которым она должна удовлетворять:

1. Математическое ожидание $\varepsilon$ равно 0;
2. Дисперсия $\varepsilon$ постоянна для всех значений;
3. Остатки независимы и нормально распределены.

Вычислим описательные статистики для остатков. Полученные результаты проследим по таблице 3.

| | Значение |
|---|---|
| Среднее | 0.00 |
| Медиана | -0.00 |
| Нижний квартиль | -1.70 |
| Верхний квартиль | 1.43 |
| Минимум | -3.14 |
| Максимум | 4.30 |
| Размах | 7.44 |
| Квартильный размах | 3.13 |
| Дисперсия | 4.07 |
| Стандартное отклонение | 2.02 |
| Стандартная ошибка | 0.36 |
| Асимметрия | 0.38 |
| Ошибка асимметрии | 0.41 |
| Эксцесс | -0.90 |
| Ошибка эксцесса | 0.81 |

Таблица 3 — Описательные статистики остатков

Как видно из таблицы 3, среднее значение равно нулю. При этом коэффициенты асимметрии($A_S = 0.38$) и эксцесса($K = -0.905$) указывают на большее отклонение распределения остатков от нормального закона.

Построим гистограмму и график квантилей для проверки последних заключений. Построенная гистограмма (приложение Б, рисунок Б.1) наглядно демонстрирует полученные в таблице 3 коэффициенты асимметрии и эксцесса.

Как и в случае исходных данных график квантилей позволяет наглядно оценить близость к нормальному распределению. На рисунке 4.1.6 можно заметить, что присутствуют отклонения относительно нормального распределения. Наиболее явный из них — нижний хвост. Остальные — небольшие скачки по ходу линии нормального распределения. Проверим с помощью критерия Шапиро-Уилка, можно ли считать полученные остатки нормально распределёнными. Из полученных в **R** результатов, статистика Шапиро–Уилка $W = 0.95$. Вероятность ошибки $p = 0.17 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста нельзя.

Проверим критерий $\chi^2$ Пирсона. Из полученных в **R** результатов, статистика $\chi^2$ Пирсона $P = 7.00$. Вероятность ошибки $p = 0.22 > 0.05$, а значит нулевая гипотеза не отвергается.

Построим график автокорреляционной функции для определения наличия взаимосвязей в ряде остатков (рисунок 4.1.7). На графике пунктирные линии разграничивают значимые и не значимые корреляции: значения, выходящие за линии, являются значимыми [14,

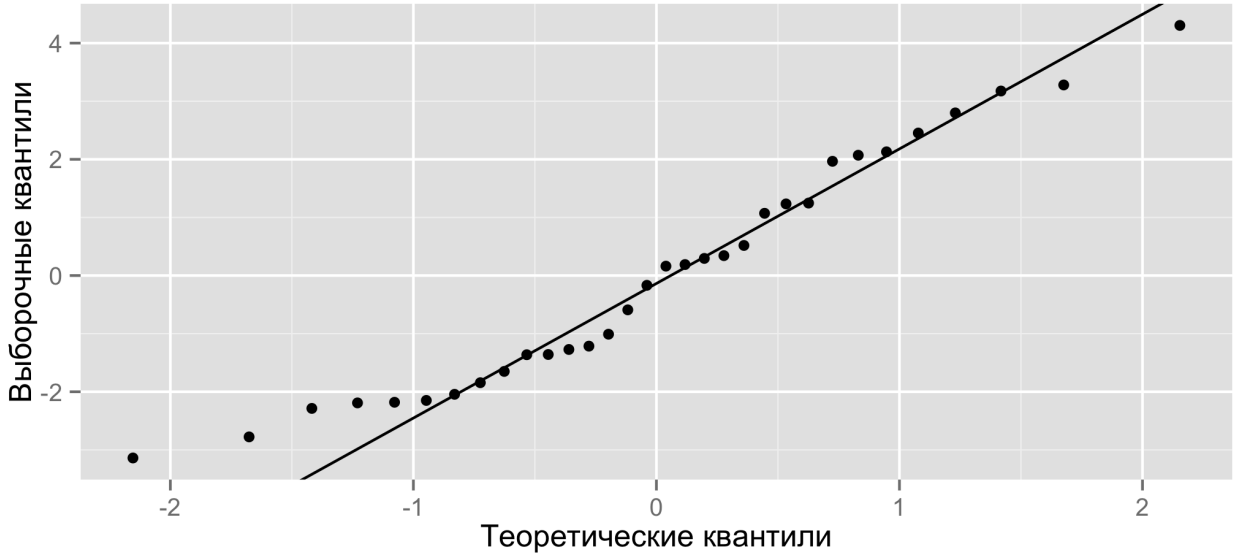


Рисунок 4.1.6 — График квантилей для остатков

с.376]. На представленном графике автокорреляционной функции все значения не выходят

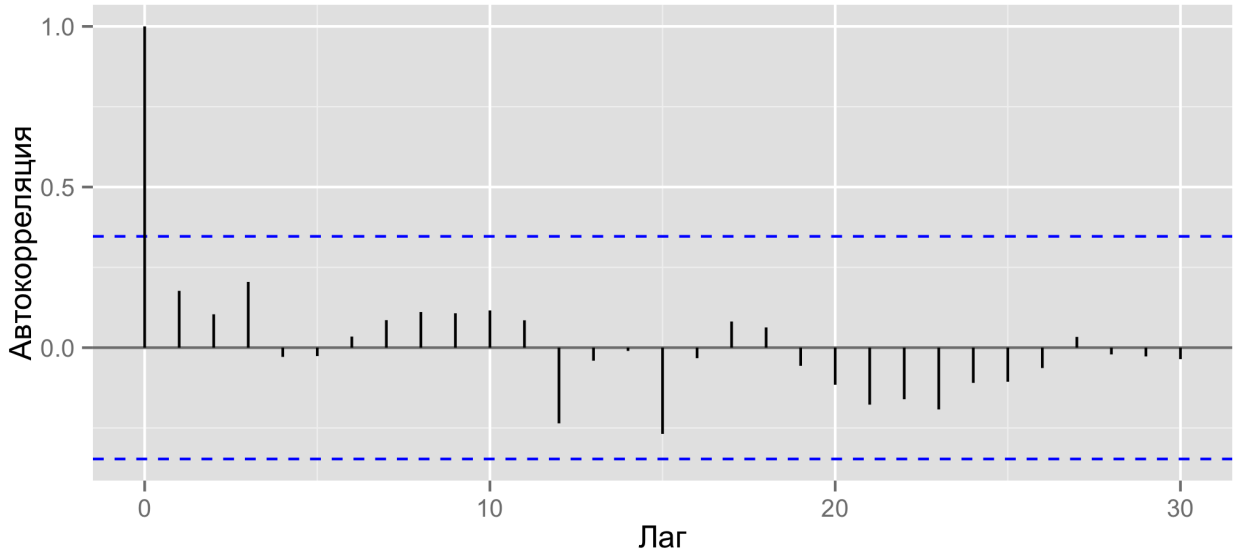


Рисунок 4.1.7 — График автокорреляционной функции

за интервал, обозначенный пунктирными линиями. Это означает, что в представленной автокорреляционной функции нету значимых автокорреляций. Проверим это замечание с помощью теста Льюнга-Бокса [14, с.377-378]. Данный тест позволяет проверить наличие автокорреляций в исследуемых данных. Используя возможности пакета **R** получили значения: статистика Льюнга–Бокса $X^2 = 0.073$ и вероятность ошибки $p = 0.79 > 0.05$ — это говорит о том, что тест не выявил значимых автокорреляций.

На рисунке 4.1.7 также можно заметить некоторое затухание значений автокорреляций с увеличением лага. На основе этого можно сделать предположение о стационарности. Для проверки этого предположения воспользуемся расширенным тестом Дики–Фуллера(ADF) [25]. Из результатов проверки теста, статистика Дики-Фуллера $DF = -3.36$, вероятность ошибки $p = 0.08 < 0.05$. Следовательно, при уровне значимости $\alpha = 0.05$ необходимо

принять альтернативную гипотезу о стационарности.

Таким образом в результате анализа детерминированными методами выделены две составляющие исходной модели данных: тренд и нерегулярная составляющая. В ходе регрессионного анализа было показано, что модель, основанная на тренде, не позволяет воспроизвести поведение исходного временного ряда. То есть нерегулярная составляющая $\varepsilon(t)$ является существенной и отвечает за это поведение. Для того, чтобы определить возможность её дальнейшего исследования проведен анализ остатков, в процессе которого показаны близость распределения к нормальному (с некоторыми отклонениями) и стационарность, при этом не выявлено значимых автокорреляций. Таким образом, это позволяет перейти к построению модели другими, современными статистическими методами интерполяции. Улучшение модели будет происходить за счёт суперпозиции модели, полученной на данном этапе, и найденной модели нерегулярной составляющей.

## 4.2 Геостатистический подход

Традиционные детерминированные модели интерполяции, широко используемые в задачах прогнозирования, в большинстве случаев на практике не позволяют в полной мере решить ту или иную задачу. В наиболее благоприятных вариантах исследований они позволяют оценивать значения в точках, в которых измерения не проводились. В свою очередь, анализ этих данных и его результаты в значительной мере зависят как от качества так и от количества исходных данных. И именно такие результаты были получена в результате проведенного в предыдущем разделе исследования. А также сделан вывод о необходимости использования современных методов исследования.

В современных исследованиях аналогичного класса задач усилился интерес к геостатистическим моделям интерполяции, что подтверждается работами [26, 27]. Современная геостатистика — это широкий спектр статистических моделей и инструментов для анализа, обработки и представления пространственно-распределенной информации.

В частности, широкое распространение получили модели из семейства *кригинга*. Преимущество данного семейства перед детерминированными методами в том, что они позволяют получить наилучшую в статистическом смысле оценку — несмещенную оценку с минимальной дисперсией, при этом оценка кригинга сопровождается оценкой ошибки интерполяции в каждой точке. Полученная ошибка позволяет охарактеризовать неопределенность интерполяционной оценки данных при помощи доверительных интервалов.

### 4.2.1 Вариограммный анализ. Кригинг.

В последующем исследовании в качестве объекта анализа будем использовать нерегулярную составляющую $\varepsilon(t)$. Поэтому исследуемой выборкой будем считать остатки, полученные на этапе регрессионного анализа и представленные в приложении В в таблице В.1.

Прогнозные значения будем вычислять как сумму значений по модели тренда (4.3) $y(t)$ и вычисленным с помощью кригинга значений $k(t)$:

$$x^*(t) = y(t) + k(t).$$

Центральная идея геостатистики состоит в использовании знаний о корреляции экспериментальных данных для построения оценок и интерполяций. *Вариограмма* является ключевым инструментом для оценки степени корреляции, имеющейся в исследуемых данных, и для ее моделирования. Модель вариограммы является функцией, определяющей зависимость изменения исследуемой величины от расстояния. Следовательно, интерполяционная модель, основанная на такой корреляционной функции, будет отражать реальные

явления, которые лежат в основе данных измерений. Всевозможные пары точек могут быть рассортированы по классам в соответствии с разностью их координат

$$h = x_i - x_j, \quad i, j = \overline{1, n}, \quad i \neq j,$$

называемой *лагом*. Для близких точек разность значений функции в них обычно меньше и растет с увеличением расстояния между точками. Вычислив среднее значение квадратов разностей для каждого значения лага $h$, можно получить дискретную функцию, называемую *экспериментальной вариограммой*. Вариограмма обычно характеризуется двумя параметрами: *рангом* и *порогом*. Порог характеризует предельное значение вариограммы, на некотором расстоянии, называемом рангом, за которым последующие значения вариограммы становятся некоррелированными.

Для оценки поведения данных при увеличении лага построим диаграмму взаимного разброса пар точек (*h-scatterplot*), разделённых расстоянием $h$. Эта диаграмма позволяет проверить наличие корреляции в исследуемых данных как качественно, так и количественно [28]. Построенная диаграмма изображена на рисунке Б.2 в приложении Б. Следует отметить, что в классическом случае присутствия зависимости, поведение должно быть следующим: на графиках, соответствующим начальным лагам, должна присутствовать сильная корреляция, и с увеличением лага корреляция уменьшается. Это объясняется тем, что чем ближе находятся данными тем выше зависимость между ними и наоборот. В рассматриваемом случае такого не наблюдается. Напротив, на первом же лаге отсутствует корреляция, при этом можно наблюдать, что на некоторых лагах присутствует корреляция, на некоторых нет. Такое поведение свойственно так называемым беспороговым моделям вариограммы. Другими словами, моделям, в которых отсутствует ранг. Одной из таких моделей является линейная (4.4), с которой некоторые исследователи советуют начинать подбор модели. Аргументируется это тем, что, она является простейшей.

$$\gamma(h) = Lin(h) = \begin{cases} b \cdot h, & h > 0, \\ 0, & h \leq 0, \end{cases} \tag{4.4}$$

где $b$ – параметр, отвечающий за угол наклона.

Поведение диаграммы в рассматриваемом случае вполне обосновано спецификой исследуемых данных: рассматривается температура воды за один определённых месяц в течение нескольких лет. Ко всему прочему, это подтверждается результатами проведённого ранее анализа остатков, в котором мы выяснили, что ошибка распределение ряда остатков является близким к нормальному и значения некоррелируемы и независимы.

В некоторых источниках советуют при построении вариограммы учитывать параметр максимального расстояния, для которого вычисляется вариограмма, а также приводят рекомендацию по его подбору. Поэтому первоначальным параметром было выбрано значение, рассчитанное по такой рекомендации: $2n/3 = 20$ [29].

В общем случае процесс вариограммного анализа заключается в выполнении серии шагов. Первым шагом вычисляют экспериментальную вариограмму, затем, при начальных значениях порога и ранга подбирают теоретическую вариограмму и с помощью различных методов пробуют улучшить её качество. После получения удовлетворительной модели используют метод кригинга для вычисления прогнозных значений. Существует два способа подбора модели. Визуальный, с учетом специфики данных, и методы, основанные на подборе параметров для определённой модели с помощью обычного метода наименьших квадратов.

Экспериментальной вариограммой по сути является некоторая оценка вариограммы. Существует несколько известных оценок, каждая из которых имеет свои достоинства и недостатки. Для данного исследования были выбраны наиболее распространённые: оценка

Матерона (2.1), введённая ранее в главе 2, и оценка Кресси-Хокинса [30, 31]:

$$2\tilde{\gamma}(h) = \frac{1}{n-h}(\sum_{t=1}^{n-h} |X(t+h) - X(t)|^{\frac{1}{2}})^4 / (0.457 + \frac{0.494}{n-h} + \frac{0.045}{(n-h)^2}), \quad h = \overline{0, n-1},$$

для сравнения полученных результатов.

Проведённый анализ остатков выявил стационарность, независимость и близость к нормальному распределению. Из условий накладываемых на вариограмму (ссылку на теорию, свою и не очень), следует возможность её применения. Как уже было сказано ранее, случайный процесс $\varepsilon(t)$ имеет распределение, близкое к нормальному $\mathcal{N}(0.00, 4.07)$, математическое ожидание равное 0 и обладает постоянной дисперсией. Это в полной мере удовлетворяет свойствам оценки Матерона, полученным в главе 2. Поэтому сначала проведём исследования с её помощью, а затем сравним полученные результаты с результатами использования оценки Кресси-Хокинса.

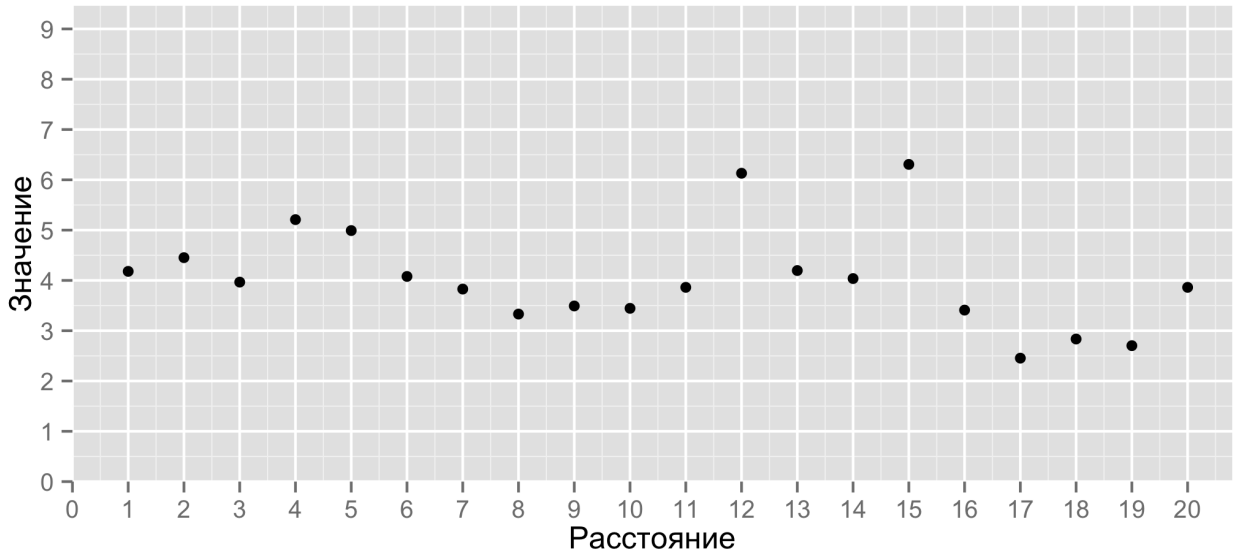


Рисунок 4.2.8 — Экспериментальная вариограмма (оценка Матерона)

Построенная вариограмма отображена на рисунке 4.2.8. При подборе моделей вариограммы, в общем случае, существует два пути: подбор силами исследователя, т.е. визуально с ручным выбором параметров, и автоматическим подбором параметров с помощью специальных методов и алгоритмов. На практике построение модели вариограммы представляет собой итеративный процесс, на каждом шаге которого следует наилучшим образом подобрать параметры очередного модельного приближения. В различных источниках рекомендуется строить модели вручную, так как исследователь лучше знает специфику данных, чем различные методы оценивания [32]. Далее будем строить модель вариограммы по рекомендации — визуально.

Вследствие выводов по диаграмме взаимного разброса, начнём подбор теоретической вариограммы (модели) с линейной. В целях приведения процесса отыскания оптимальных параметров выполним шаги упомянутые ранее подробно. Подбор начальных параметров для построения теоретической вариограммы осуществим визуально. Как уже было сказано, в беспороговых моделях ранг $a = 0$. Из уравнения (4.4) видно, что параметр отвечает за скорость роста значений, поэтому примем начальное значение $b = 4$. Вычисленная модель отображена на графике Б.3. Следует отметить, что данная модель после применения кригинга позволила получить прогнозные значения очень близкие к нулю, что не изменила прогноза, построенного по модели тренда (таблица 2).

Попробуем подобрать параметры, основываясь на полученной модели, с помощью возможностей пакета *gstat*. В результате получаем модель с чистым эффектом самородков

$$\gamma(h) = c \cdot Nug(h) = \begin{cases} 0, & h = 0, \\ c, & h \neq 0, \end{cases}$$

с параметром $c = 4.04$. Объяснить такой результат можно тем, что значения вариограммы сразу достигают порогового значения приблизительно равному дисперсии и не имеют большого разброса относительно среднего. При этом также можно считать, что вывод, сделанный по диаграмме взаимного разброса не совсем верен, поскольку это поведение можно объяснить стационарностью процесса. Поэтому метод подбора параметров, основывающийся на методе наименьших квадратов, производит такие результаты. Это следствие того, что данных подход не учитывает особенностей исследуемых данных. Поэтому результатов прогнозирования данная модель не улучшила.

Но при этом наличие порога объяснима видом вариограммы: по рисунку 4.2.8 можно видеть, что уже первые значения достигают уровня дисперсии. И отклонение от этого значения не велико. Это согласуется с исследуемыми исходными данными, так как при анализе остатков было выявлено отсутствие автокорреляций, и спецификой самих данных: значение температуры воды за определённый год слабо зависит от значения предшествующего. Из этого следует, что использование беспороговых моделей не обосновано. Поэтому попробуем улучшить результаты с помощь линейной модели с порогом:

$$\gamma(h) = c \cdot Lin(h, a) = \begin{cases} c \cdot \frac{h}{a}, & 0 \leq h \leq a, \\ c \cdot h, & h > a, \end{cases} \tag{4.5}$$

где $c$ – порог, $a$ – ранг. Данная модель применима только к одномерным данным, что является рассматриваемым случаем.

Для подбора оптимальных параметров линейной модели с порогом будем проводить с помощью инструментов реализованной программы. Подбор осуществляется следующим образом:

- задаются начальные значения параметров
- выбирается параметр для подбора, диапазон поиска и шаг итерации
- на каждом шаге кригингом вычисляются прогнозные значения
- на основе полученных значений строится статистика

В результате такого процесса получается ряд оценок моделей, зависящих от значения параметров. На их основе выбирается оптимальный. Затем процесс повторяется для другого параметра и так далее, пока не найдётся оптимальная модель. В реализованном приложении имеется два подхода по оценке качества построенной модели. Используя первый подход, модель оценивается с помощью метода кросс-валидации. В данном случае он заключается в последовательном исключении одного из известных значений и построении интерполяции в этой точке по валидируемой модели. Таким образом получаем ряд интерполяций, который должен в идеальном случае воспроизводить поведение исследуемого ряда. Поэтому появляется возможность с помощью различных статистик оценивать конкретную модель вариограммы. С помощью таких статистик можно проследить, как изменяется качество модели при изменении какого-либо из параметров. Это в свою очередь позволяет найти оптимальное значение искомого параметра и использовать его для подбора остальных. При втором подходе, адаптивном, в исследуемых данных отдаётся предпочтение последним наблюдениям. Для этого отбрасывается некоторое количество

значений для последующего обучения модели. Подбор параметров осуществляется по статистикам, рассчитанным по отклонениям прогнозных значений от наблюдаемых. Таким образом достигается наилучший прогноз в краткосрочной перспективе.

Воспользуемся первым подходом, в качестве статистики будем использовать коэффициент корреляции между интерполированными и актуальными значениями. График зависимости значения ранга на качество модели отображён на рисунке 4.2.9. По рисунку

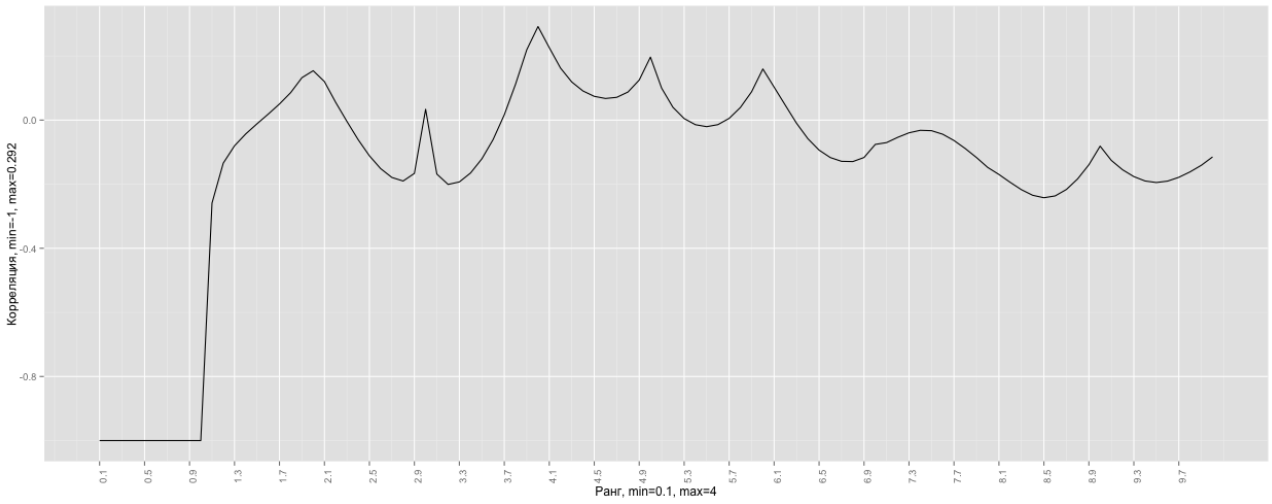


Рисунок 4.2.9 — Зависимость качества линейной модели от значения ранга

видно, что максимальное значение коэффициента корреляции $r = 0.292$ достигается при значении ранга $a = 4$, при этом среднеквадратическое отклонение от истинных значений $MSE = 7.931$. Таким образом получена модель $4 \cdot Lin(h, 4)$, ее график отображен на рисунке Б.5. Вычисленные по данной модели прогнозные значения можно проследить по таблице 4 и по графику Б.6 в приложении Б. Как можно видеть, прогнозные значения

| | Год | Наблюдение | Прогноз | Тренд | Ошибка |
|---|---|---|---|---|---|
| 1 | 2007 | 19.400 | 20.824 | 21.578 | -1.424 |
| 2 | 2008 | 21.800 | 21.133 | 21.687 | 0.667 |
| 3 | 2009 | 21.900 | 19.831 | 21.797 | 2.069 |
| 4 | 2010 | 24.300 | 22.129 | 21.906 | 2.171 |
| 5 | 2011 | 22.800 | 22.239 | 22.016 | 0.561 |
| 6 | 2012 | 20.200 | 22.348 | 22.126 | -2.148 |

Таблица 4 — Прогноз (линейная модель с порогом)

предсказали только поведение в первый год. Дальнейшие значения оказались далеки от истины, что объясняется значением среднеквадратической ошибки. Таким образом, данная модель неплохо себя показала при описании всех данных, что показал коэффициент корреляции, но при этом прогноз оказался не точным.

Для построения более точного прогноза воспользуемся адаптивным подходом по подбору параметров. График зависимости среднеквадратической ошибки от ранга отображен на рисунке 4.2.10. Из него видно, что оптимальным параметром для ранга является $a = 2$, с минимальной среднеквадратической ошибкой $MSE = 1.62$. График 4.2.10 прогнозных значений показывает, что данный подход позволил предсказать три первых значения. Что является хорошим результатом. При этом статистики по данной модели после проведения кросс-валидации оказались следующими: коэффициент корреляции $r = 0.152$, среднеквад-

ратическая ошибка 18.69. Что говорит о том, что данная модель описывает всю выборку хуже чем предыдущая.

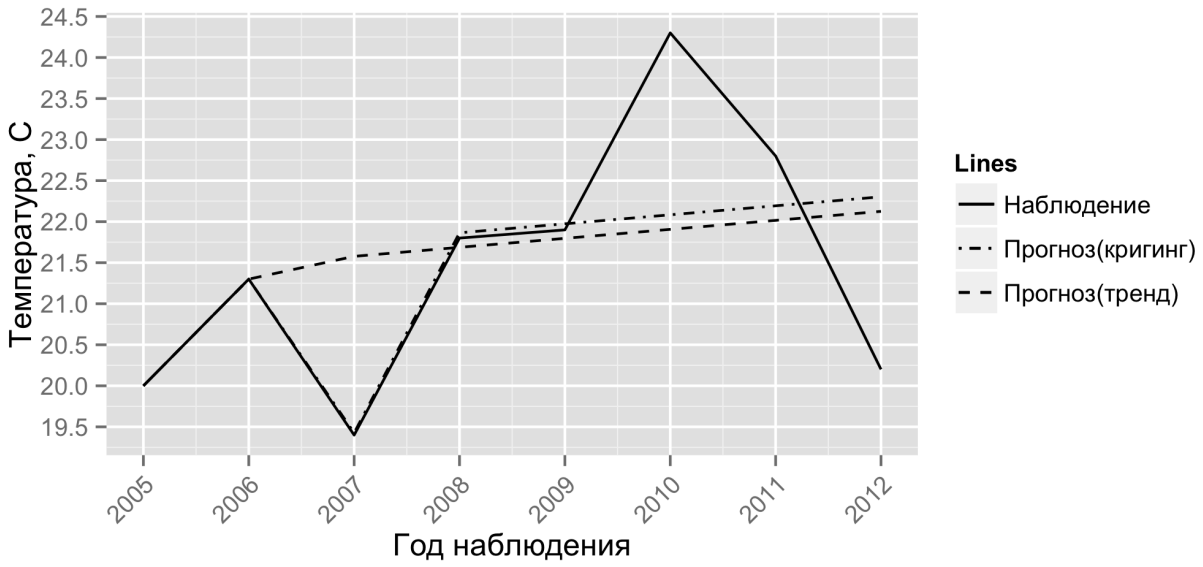


Рисунок 4.2.10 — Прогноз по модели $2 \cdot Lin(h, 2)$

Таким образом можно сделать выводы о преимуществах продемонстрированных подходов. Недостаток первого подхода заключается в том, что модель, описывающая поведение исследуемых данных сглаживает локальное поведение, вследствие чего прогнозные значения могут получиться не всегда точными. В рамках рассматриваемой задачи, когда данные имеют сложную структуру, потеря информации о локальных изменениях в значениях, влечёт ухудшение прогноза. Во втором же случае, в отличие от первого, описывается поведение не всей выборки, а только те, которые интересны больше всего — последние наблюдения, так как они в большей мере влияют на будущие значения. Но в этом случае учитывается в меньшей степени поведение данных в целом. Поэтому в зависимости от поставленных целей можно использовать один из описанных подходов.

Модель $4Lin(h, 2)$, полученная с помощью первого подхода, как было упомянуто ранее, описывает исходные данные не очень точно. Поэтому есть необходимость в поиске моделей, дающих лучшие результаты. Одной из самых распространённых и часто используемой пороговой моделью является сферическая:

$$\gamma(h) = c \cdot Sph(h, a) = \begin{cases} c \cdot (\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3), & h \le a, \\ c, & h \ge a. \end{cases}$$

Однако после подбора оптимальных параметров оказалось, что данная модель вписывается в исследуемые данные хуже, чем найденные ранее. При подборе параметров с помощью кросс-валидации наилучшей получилась модель $4Sph(h, 2.3)$ с показателями: коэффициент корреляции $r = -0.002$ и среднеквадратическим отклонением $MSE = 5.407$. В случае адаптивного подхода, оптимальными оказались параметры $c = 4, a = 6.9$ с эффектом самородков равным 0.9, при среднеквадратическом отклонении $MSE = 2.01$. Применив кросс-валидацию к этой модели, получаем следующие показатели качества: коэффициент корреляции $r = -0.009$ и среднеквадратическим отклонением $MSE = 5.396$. Графики вариограммы и прогнозных значений последней модели отображены на рисунках Б.9 и Б.10 в приложении Б соответственно. Можно сделать вывод, что как и линейная с порогом модель, сферическая не позволила описать поведение исследуемой выборки. Только в

случае краткосрочного прогноза она проявила себя, предсказав характерное поведение исключённых значений, хоть и хуже предшествующей линейной с порогом модели. Похожее поведение можно объяснить видом их теоретических вариограмм. Это видно по графикам Б.8 и Б.9 в приложении Б.

Если обратить внимание на график экспериментальной вариограммы 4.2.8, то можно заметить некоторый периодический эффект в виде волны. Поэтому дальнейшей подбираемой моделью возьмем периодическую:

$$\gamma(h) = c \cdot Per(h, a) = 1 - cos(\frac{2\pi h}{a}),$$

где $c$ – порог, $a$ – ранг.

С помощью средств написанной программы подобрана модель $4 \cdot Per(h, 0.898)$, график вариограммы которой изображен на рисунке Б.11 в приложении Б. Показатели адекватности подобранной модели: коэффициент корреляции $r = 0.404$, среднеквадратическая ошибка $MSE = 4.369$. Следует отметить, что из всех подбираемых моделей, представленное значение коэффициента корреляции оказалось самым большим. Что говорит о том, что данная модель наилучшим образом описывает исследуемые данные. Таблица В.2 в приложении В и график прогнозных значений по подобранной модели показывают, что

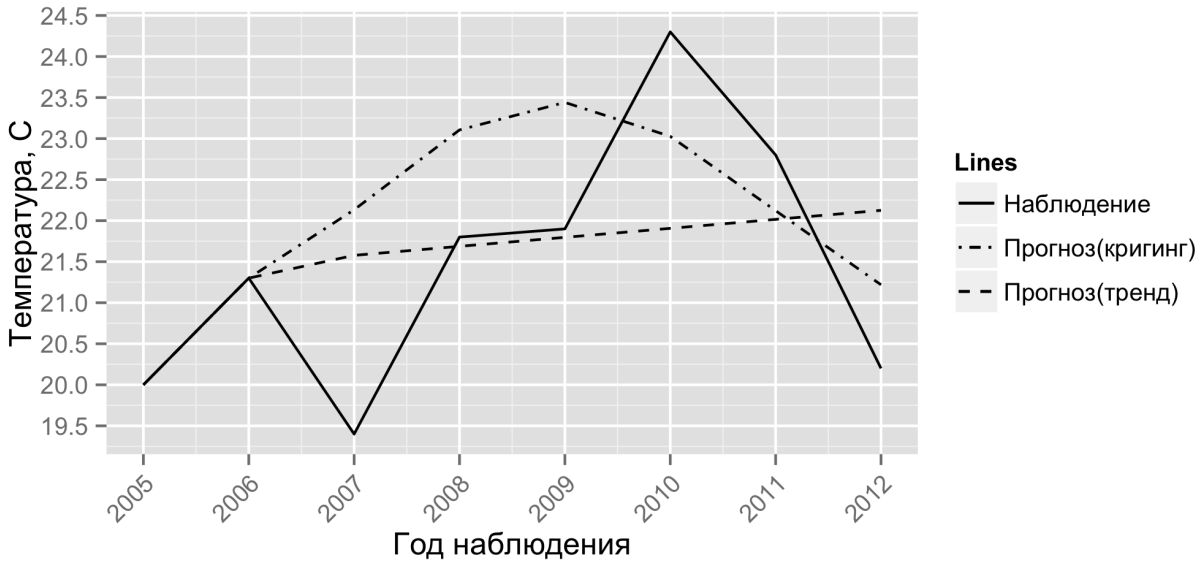


Рисунок 4.2.11 — Прогноз по модели $4 \cdot Lin(h, 0.898)$

прогноз получился не очень точный, но при этом следует принять во внимание упоминаемый ранее эффект сглаживания. Данная модель оказалась наилучшей для описания всей выборки из всех проверенных.

Таким образом найдены модели, которые в каждом случае ведут себя наилучшим образом. Периодическая с параметрами $4 \cdot Lin(h, 0.898)$ для описания исходных данных и линейная с порогом $2 \cdot Lin(h, 2)$ для построения краткосрочных прогнозов.

Следует отметить, что также обнаружено, что параметр максимального расстояния при котором вычисляется вариограмма, при фиксированной модели вариограммы никак не влияет на прогноз. При этом в этом случае на прогнозные значения не будет влиять и оценка Кресси-Хокинса, изображённая на рисунке Б.12 в приложении Б. Поскольку экспериментальная вариограмма это начальный шаг, по которой визуально подбирается теоретическая вариограмма. При описанных ранее подходах, эти факторы никак не влияют на конечный результат.

Полученные значения оказались идентичными значению тренда, следовательно прогноз почти не изменился. Это говорит о том, что построенная модель не смогла уловить поведение исходных данных.

Для построения модели вариограммы была реализована возможность автоматического подбора модели на основе функции *fit.variogram*. Суть этого подхода заключается в следующем: при заданных начальных условиях (эффект самородков, ранг, порог), для всех возможных базисных моделей подгонялись их параметры, для этих моделей вычислялись сумма квадратов ошибок, и на основе этого показателя выбиралась наиболее эффективная модель. Код программы представлен в листинге D.2.

На рисунке **[DELETED]** сплошной линией и в приложении Б на рисунке **??** показан результат выполнения представленной ранее функции. Таким образом, наилучшей моделью вариограммы, построенной по классической оценке, стала линейная комбинация двух: эффект самородков с параметром 3.45 и модель с эффектом дыр (*Hole*) с параметрами: порог — 0.816, ранг — 1.71 изображенная в приложении Б на рисунке **??**.

Методом простого кригинга в этом случае были построены прогнозные значения, отображенные в таблице 5. Полученные значения отличаются от предыдущих, в них появилось

| | Год | Наблюдение | Прогноз | Тренд |
|---|---|---|---|---|
| 1 | 2007 | 19.400 | 21.576 | 21.578 |
| 2 | 2008 | 21.800 | 21.691 | 21.687 |
| 3 | 2009 | 21.900 | 21.801 | 21.797 |
| 4 | 2010 | 24.300 | 21.910 | 21.906 |
| 5 | 2011 | 22.800 | 22.020 | 22.016 |
| 6 | 2012 | 20.200 | 22.129 | 22.126 |

Таблица 5 — Прогноз (классическая оценка)

некоторое поведение. Но в данном случае $MSE = 2.47$, что хуже предыдущего значения, а значит, прогноз ухудшился. Попробуем улучшить результат с помощью робастной оценки Кресси.

Модель вариограммы, представленная на рисунке Б.12 в приложении Б, является также линейной комбинацией двух базисных моделей: эффекта самородков с параметром 0 и волновая модель с параметрами: 0, 0. Заметим, что эмпирическая вариограмма, построенная по робастной оценке, отличается от соотвествующей вариограмм, построенных по классической оценке. Появилось заметное поведение вариограммы, в отличие от предыдущей, где значения концентрировались около дисперсии выборки.

Результаты применения кригинга показали прогнозные значения, указанные в 6. Среднеквадратическая ошибка $MSE = 2.47$, таким образом это значение близко к значе-

| | Год | Наблюдение | Прогноз | Тренд |
|---|---|---|---|---|
| 1 | 2007 | 19.400 | 21.578 | 21.578 |
| 2 | 2008 | 21.800 | 21.687 | 21.687 |
| 3 | 2009 | 21.900 | 21.797 | 21.797 |
| 4 | 2010 | 24.300 | 21.906 | 21.906 |
| 5 | 2011 | 22.800 | 22.016 | 22.016 |
| 6 | 2012 | 20.200 | 22.126 | 22.126 |

Таблица 6 — Прогноз (робастная оценка)

нию, полученному вручную. Таким образом, использование робастной оценки улучшило результат применения кригинга.

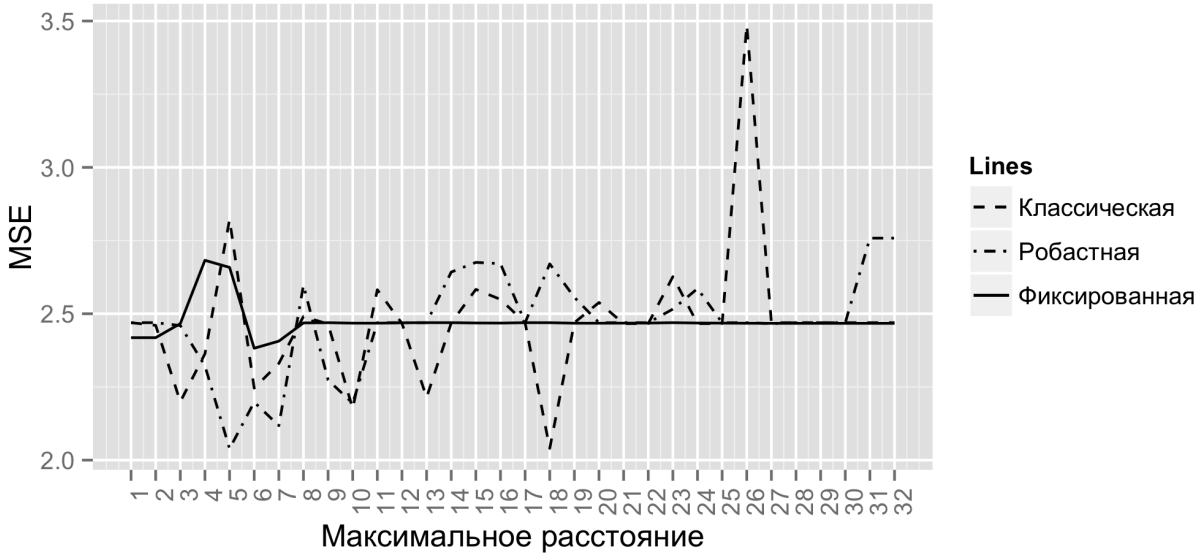


Рисунок 4.2.12 — Зависимость ошибки от максимального расстояния

Исследуем теперь поведение прогнозных значений, полученных с помощью кригинга, при различных параметрах максимального расстояния вариограммы. В качестве оценки качества полученного прогноза возьмем среднеквадратическую ошибку. Чем меньше ошибка — тем лучше прогноз. Для этих целей реализована функция *ComparePredictionParameters*. Результат её работы на рисунке 4.2.12. На этом графике отчетливо видно, что робастная оценка (пунктир-точка), в отличие от классической (пунктир) и модели, построенной вручную (сплошная), в большинстве случаев даёт более точные прогнозы. И наилучший при максимальном расстоянии равным 6. С этим параметром, наилучший прогноз составляют значения кригинга из 7. Среднеквадратическая ошибка

| | Год | Наблюдение | Прогноз | Тренд |
|---|---|---|---|---|
| 1 | 2007 | 19.400 | 21.385 | 21.578 |
| 2 | 2008 | 21.800 | 21.877 | 21.687 |
| 3 | 2009 | 21.900 | 22.163 | 21.797 |
| 4 | 2010 | 24.300 | 22.217 | 21.906 |
| 5 | 2011 | 22.800 | 22.134 | 22.016 |
| 6 | 2012 | 20.200 | 22.055 | 22.126 |

Таблица 7 — Наилучший прогноз (робастная оценка)

оказалась равной $MSE = 2.04$. Что действительно является лучшим из полученных показателем.

Сравнительный анализ полученного прогноза представлен на графике Б.13 в приложении Б.

Таким образом в результате вариограммного анализа были исследованы различные модели вариограмм, оценки, проведены два подхода по вычислению. В результате кригинга построена наилучшая модель прогнозных значений. Которая в свою очередь имеет погрешность в пределах стандартного отклонения. Следовательно данная модель является хорошим вариантом для построения прогнозных значений.

# Заключение

В представленной работе бы проведён сравнительный анализ современных пакетов прикладных программ для статистического анализа. Из них как инструмент исследования был выбран язык программирования **R**, по причине его доступности и предоставления огромного числа пакетов. С помощью этого пакета была исследована важнейшая характеристика любого водоёма — температура воды. Исследование проводилось на основе данных, полученных из наблюдений за озером Баторино, в период с 1975 по 2012 год в июле месяце. Для этого были вычислены и проанализированы описательные статистики, проведена проверка на нормальность, проведён визуальный анализ. В результате указанной части работы было обнаружено, что распределение температуры воды в озере Баторино близко к нормальному закону распределения с параметрами $\mathcal{N}(20.08, 5.24)$. Отклонение от нормальности отмечается полученными коэффициентами асимметрии и эксцесса. Исследуемое распределение имеет небольшую скошенность вправо и более растянутую колоколообразную форму относительно нормального закона распределения. В результате проведённого корреляционного анализа была выявлена умеренная зависимость между температурой воды и временем: был обнаружен рост температуры с течением времени.

В работе был проведён регрессионный анализ, в процессе которого была построена аддитивная модель временого ряда, найдён тренд, и, как следствие удаления тренда из построенной модели, был получен ряд остатков. Построенная детерминированными методами линейная регрессионная модель оказалась значимой и адекватной, но при этом описывает поведение временного ряда лишь частично. В результате анализа ряда остатков было выявлено отклонение распределения от нормальности. Что говорит о наличии некоторых неучтённых данной моделью факторов, затрудняющих дальнейшее исследование классическими методами. Следует также отметить стационарность и отсутствие автокорреляций в ряде остатков. Эти результаты говорят о постоянстве вероятностных свойств с течением времени, а также об отсутствии зависимостей между наблюдениями.

Так как представленные в данной работе классические методы анализа временных рядов в этом случае оказались недостаточными для полноценного исследования, то следующим этапом стало использование современных геостатистических методов. В процессе чего были построены различные вариограммы, подобраны модели этих вариограмм. С помощью кригинга был осуществлён прогноз значений и их анализ. Найден наилучший прогноз для исходных данных.

# Литература


1. Stephen L. Katz, Stephanie E. Hampton, Lyubov R. Izmest'eva, and Marianne V. Moore. Influence of long-distance climate teleconnection on seasonality of water temperature in the world's largest lake - lake Baikal, Siberia. *PLoS ONE*, 6(2):e14688, 02 2011.

2. T.P. O'Brien, W.W. Taylor, A.S. Briggs, and E.F. Roseman. Influence of water temperature on rainbow smelt spawning and earlylife history dynamics in st. martin bay, lake huron. *Journal of Great Lakes Research*, 38(4):776–785, dec 2012.

3. L. Subehi and M Fakhrudin. Preliminary study of the changes in water temperature at pond cibuntu. *Journal of Ecology and the Natural Environment*, 3(3):72–77, March 2011.

4. Enner Herenio de Alcântara, José Luiz Stech, João Antônio Lorenzzetti, and Evlyn Márcia Leão de Moraes Novo. Time series analysis of water surface temperature and heat flux components in the Itumbiara Reservoir (GO), Brazil. *Acta Limnologica Brasiliensia*, 23:245 – 259, 09 2011.

5. Chokshi Mira. Temperature analysis for lake Yojoa, Honduras. Master's thesis, Massachusetts Institute of Technology, Dept. of Civil and Environmental Engineering, 2006.

6. Д. Бриллинджер. *Временные ряды. Обработка данных и теория.* Мир, 1980.

7. Н.Н. Труш. *Асимптотические методы статистического анализа временных рядов.* Белгосуниверситет, 1999.

8. Ж. Матерон. *Основы прикладной геостатистики.* М.: Мир, 1968.

9. Т.В. Цеховая. Первые два момента оценки вариограммы гауссовского случайного процесса. *Вестник БрГУ им. А.С. Пушкина*, 2005.

10. Юзбашев М.М. Елисеева, И.И. *Общая теория статистики.* Москва : Финансы и статистика, 1995.

11. Duncan Cramer. *Basic statistics for social research: step-by-step calculations andcomputer techniques using Minitab.* Psychology Press, 1997.

12. M. G. Bulmer. *Principles of Statistics.* Dover Publications, 1979.

13. Robert Kabacoff. R in Action. 2009.

14. Paul Teetor. *R Cookbook (O'Reilly Cookbooks).* O'Reilly Media, 1 edition, 2011 2011.

15. Winston Chang. *R graphics cookbook.* "O'Reilly Media, Inc. 2012.

16. H. A. Sturges. The choice of a class interval. *American Statistical Association*, 21:65–66, 1926.

17. S. S. Shapiro and R. S. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, 1972.

18. А.И. Кобзарь. *Прикладная математическая статистика.* М.: Физматлит, 2006.

19. В.Е. Гмурман. *Теория вероятностей и математическая статистика.* Москва : Высшая школа, 2003.

20. Метельский А.В. Микулик, Н.А. *Теория вероятностей и математическая статистика: Учеб. пособие.* Минск : Пион, 2002.

21. F. E. Grubbs. Sample criteria for testing outlying observations. *Ann. Math. Statistics*, 21:27–58, 1950.

22. Frank E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

23. Стэнсфилд Р. Эддоус М. *Методы принятия решений.* Москва : Аудит, 1997.

24. Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2nd edition, 2006.

25. David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, 1979.

26. Shakeel Ahmed and Ghislain De Marsily. Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, 23(9):1717–1737, 1987.

27. Eulogio Pardo-igu Zquiza. Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography. *Int. J. Climatol*, 18:1031–1047, 1998.

28. А.А. Савельев, С.С. Мухарамова, А.Г. Пилюгин, and Н.А. Чижикова. *Геостатистический анализ данных в экологии и природопользовании (с применением пакета R).* Казанский университет, 2012.

29. Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data.* John Wiley & Sons, 2011.

30. Noel AC Cressie and Noel A Cassie. *Statistics for spatial data*, volume 900. Wiley New York, 1993.

31. Rudolf Dutter. On robust estimation of variograms in geostatistics. In Helmut Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods*, volume 109 of *Lecture Notes in Statistics*, pages 153–171. Springer New York, 1996.

32. Е.А. Савельева and В.В. Демьянов. *Геостатистика: теория и практика.* Ин-т проблем безопасного развития атомной энергетики РАН. – М.: Наука, 2010.

# Приложение А Исходные данные

| | year | temperature |
|---|---|---|
| 1 | 1975.00 | 20.20 |
| 2 | 1976.00 | 16.00 |
| 3 | 1977.00 | 17.70 |
| 4 | 1978.00 | 16.75 |
| 5 | 1979.00 | 17.50 |
| 6 | 1980.00 | 16.77 |
| 7 | 1981.00 | 19.80 |
| 8 | 1982.00 | 19.00 |
| 9 | 1983.00 | 21.40 |
| 10 | 1984.00 | 19.40 |
| 11 | 1985.00 | 20.40 |
| 12 | 1986.00 | 16.50 |
| 13 | 1987.00 | 17.10 |
| 14 | 1988.00 | 23.80 |
| 15 | 1989.00 | 19.90 |
| 16 | 1990.00 | 18.50 |
| 17 | 1991.00 | 23.00 |
| 18 | 1992.00 | 21.90 |
| 19 | 1993.00 | 18.00 |
| 20 | 1994.00 | 21.40 |
| 21 | 1995.00 | 18.90 |
| 22 | 1996.00 | 19.10 |
| 23 | 1997.00 | 21.00 |
| 24 | 1998.00 | 18.40 |
| 25 | 1999.00 | 23.50 |
| 26 | 2000.00 | 21.00 |
| 27 | 2001.00 | 24.20 |
| 28 | 2002.00 | 23.10 |
| 29 | 2003.00 | 18.00 |
| 30 | 2004.00 | 19.10 |
| 31 | 2005.00 | 20.00 |
| 32 | 2006.00 | 21.30 |
| 33 | 2007.00 | 19.40 |
| 34 | 2008.00 | 21.80 |
| 35 | 2009.00 | 21.90 |
| 36 | 2010.00 | 24.30 |
| 37 | 2011.00 | 22.80 |
| 38 | 2012.00 | 20.20 |

Таблица А.1 — Исходные данные.

# Приложение Б Графические материалы

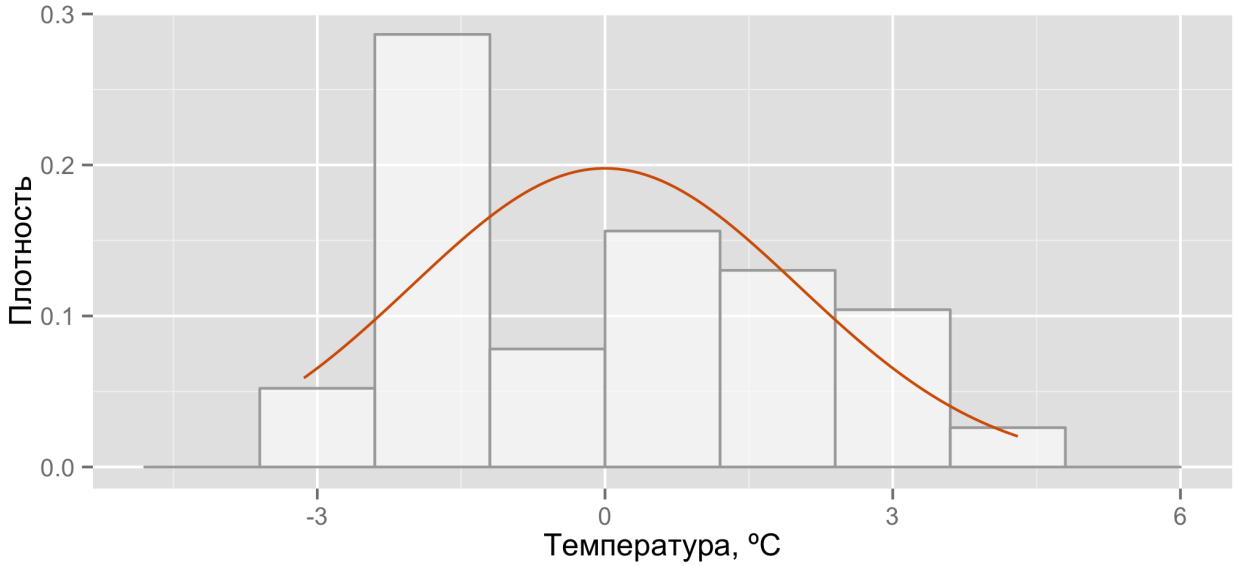


Рисунок Б.1 — Гистограмма остатков с кривой плотности нормального распределения $\mathcal{N}(19.88, 4.92)$

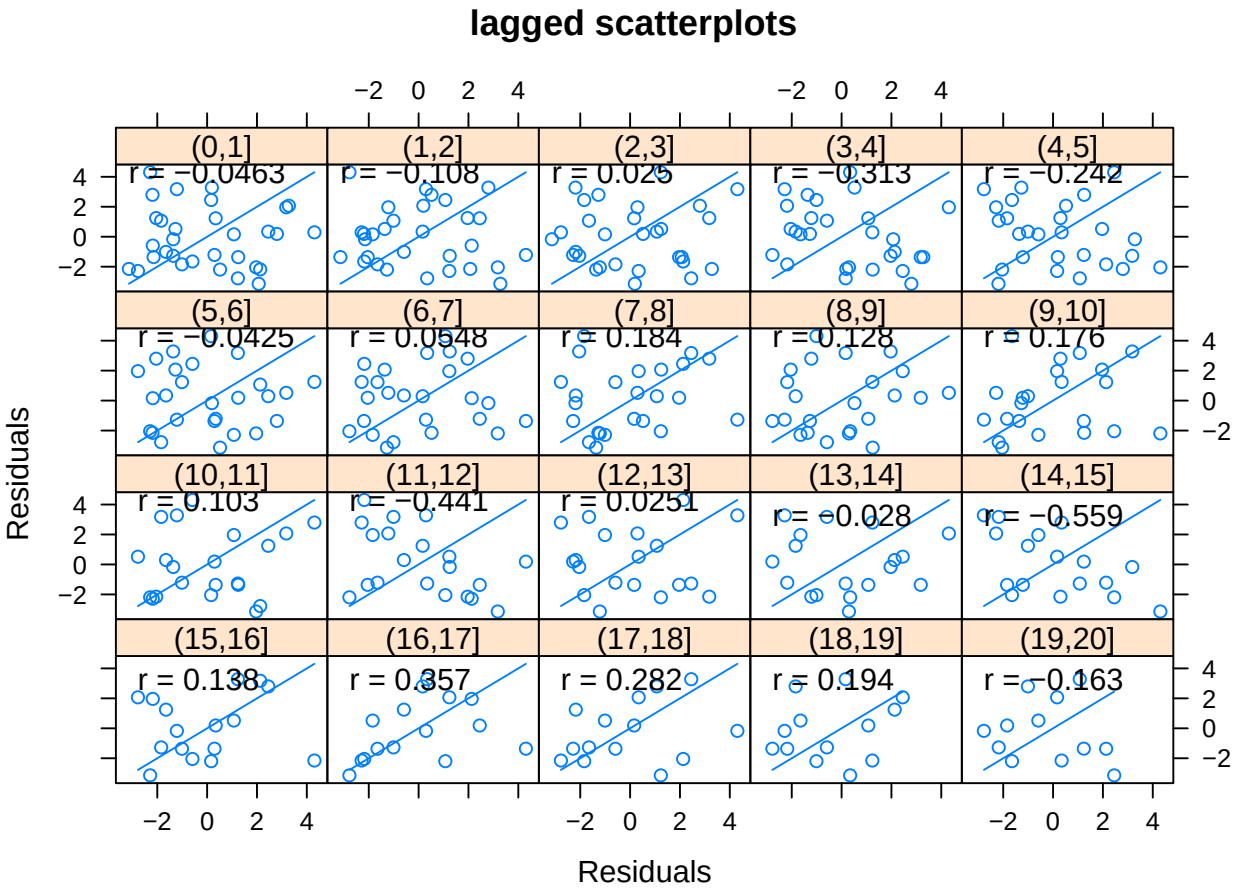


Рисунок Б.2 — Диаграмма взаимного разброса

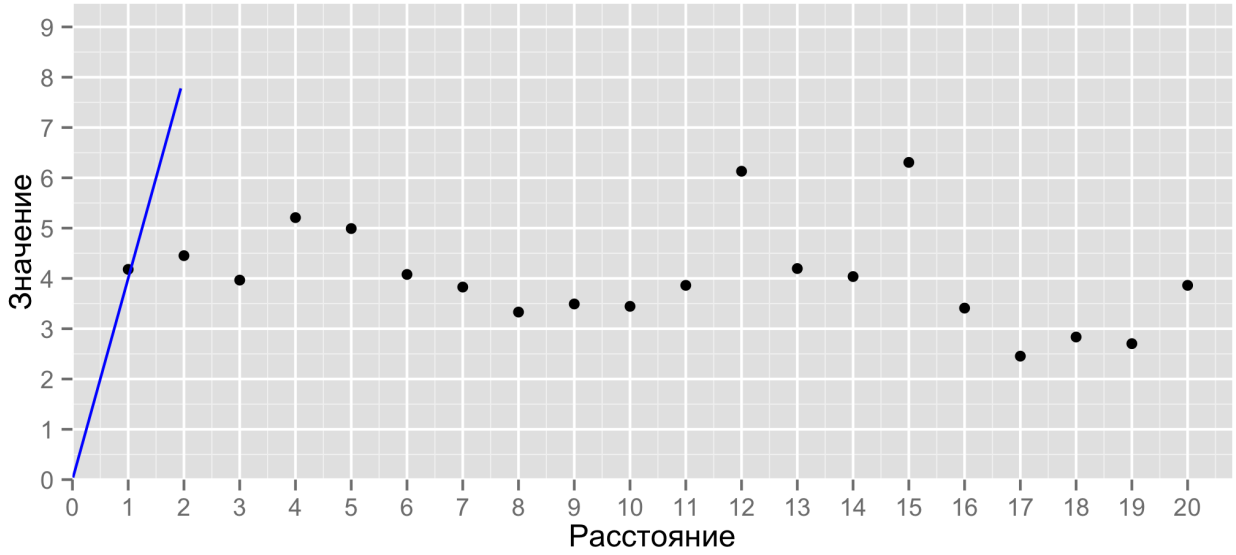


Рисунок Б.3 — Экспериментальная и теоретическая вариограмма $4 \cdot Lin(h, 0)$

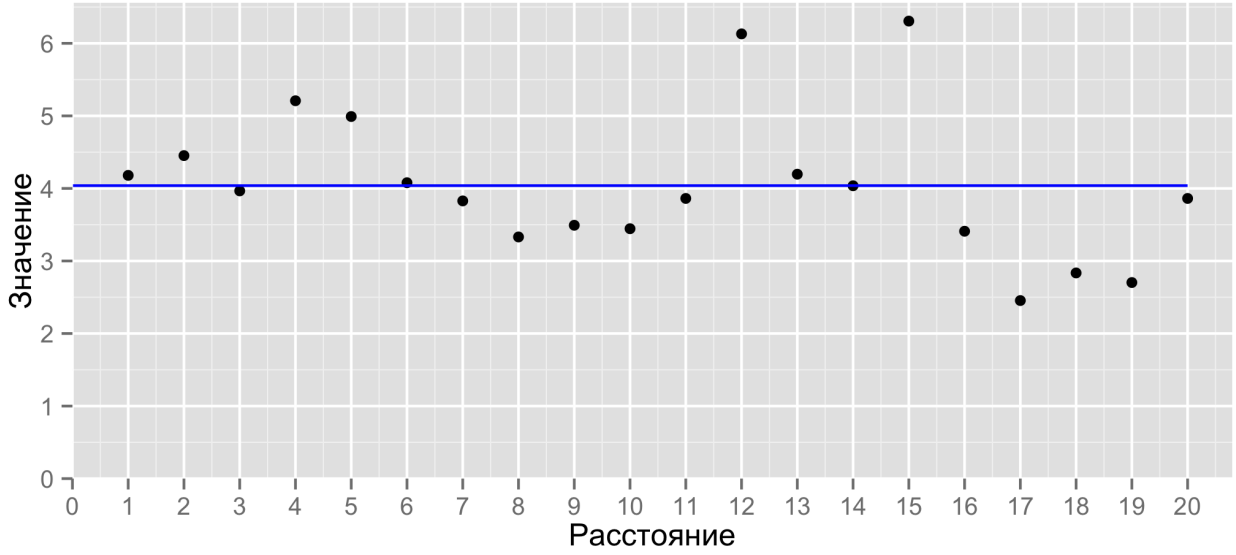


Рисунок Б.4 — Экспериментальная и теоретическая вариограмма $4.08 \cdot Nug(h)$

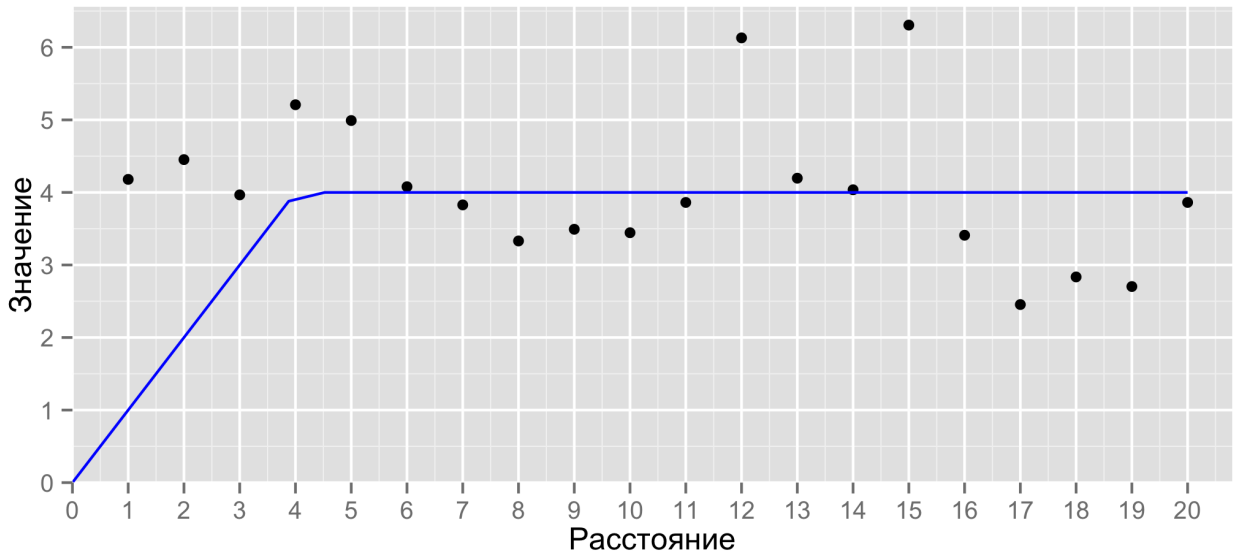


Рисунок Б.5 — Экспериментальная и теоретическая вариограмма $4 \cdot Lin(h, 4)$

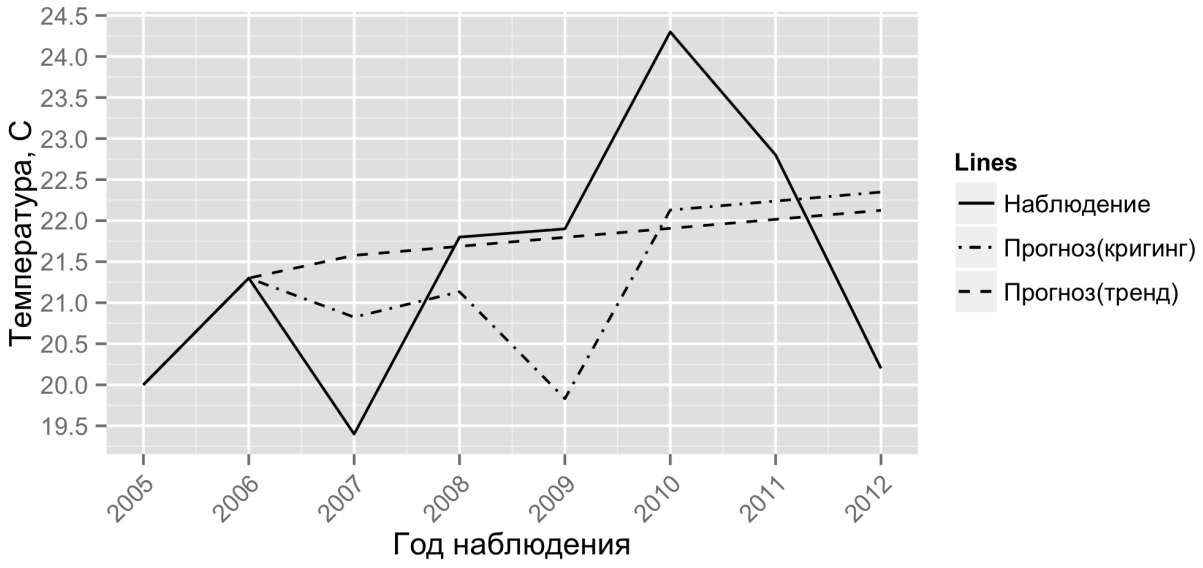


Рисунок Б.6 — Прогноз $4 \cdot Lin(h, 4)$

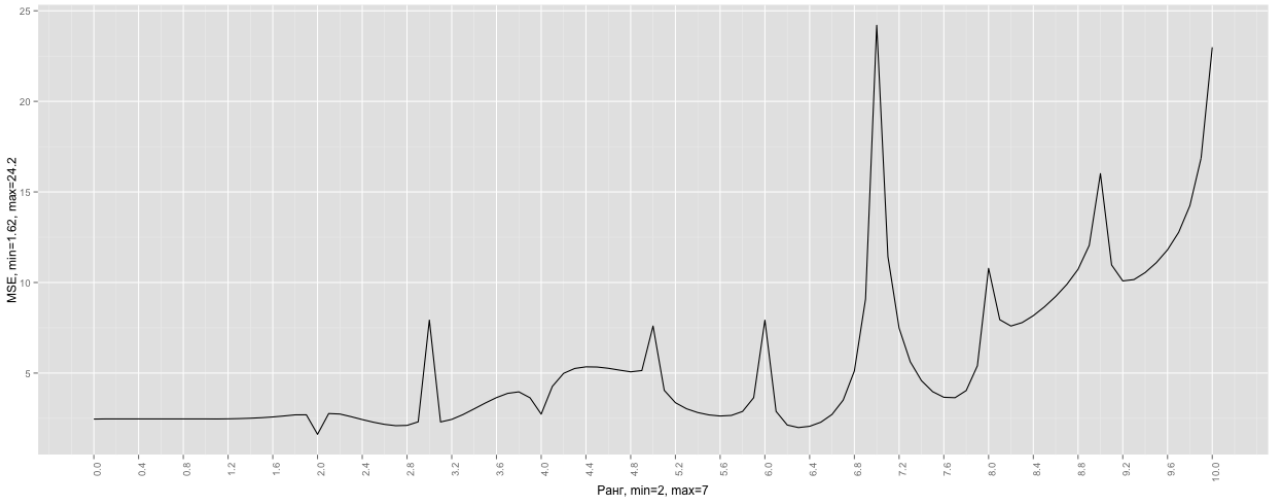


Рисунок Б.7 — Зависимость качества линейной модели от значения ранга

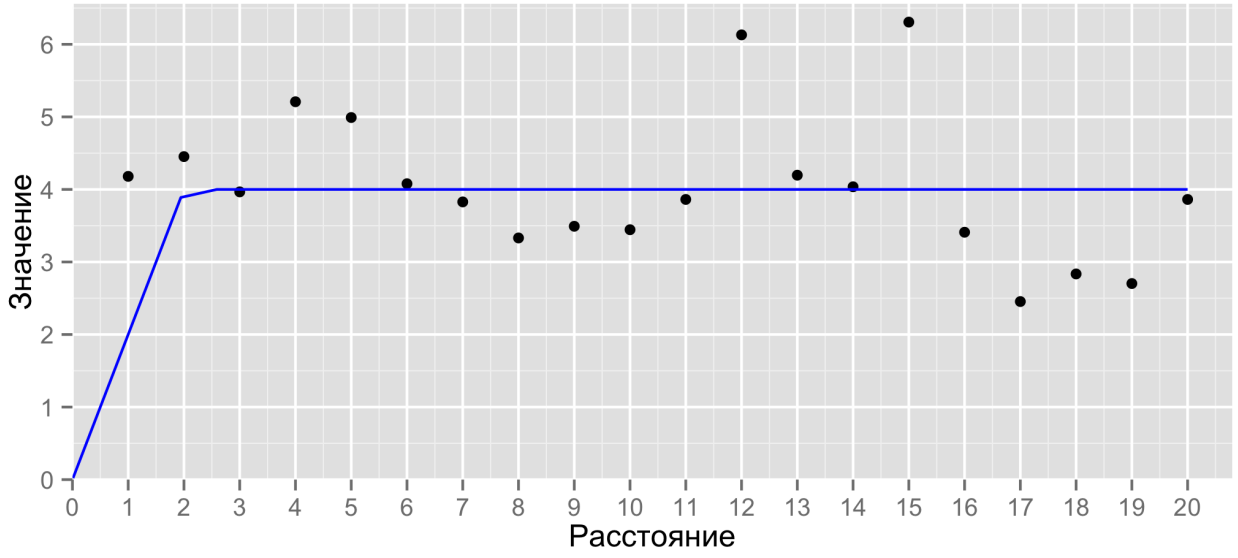


Рисунок Б.8 — Экспериментальная и теоретическая вариограмма $2 \cdot Lin(h, 2)$

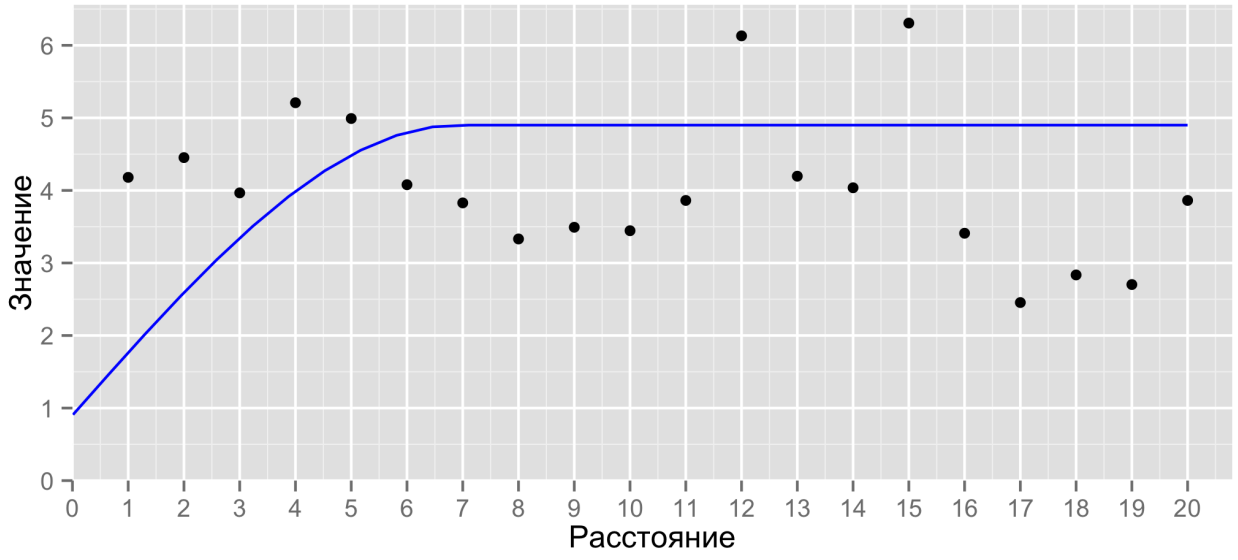


Рисунок Б.9 — Экспериментальная и теоретическая вариограмма $0.9 + 4 \cdot Sph(h, 6.9)$

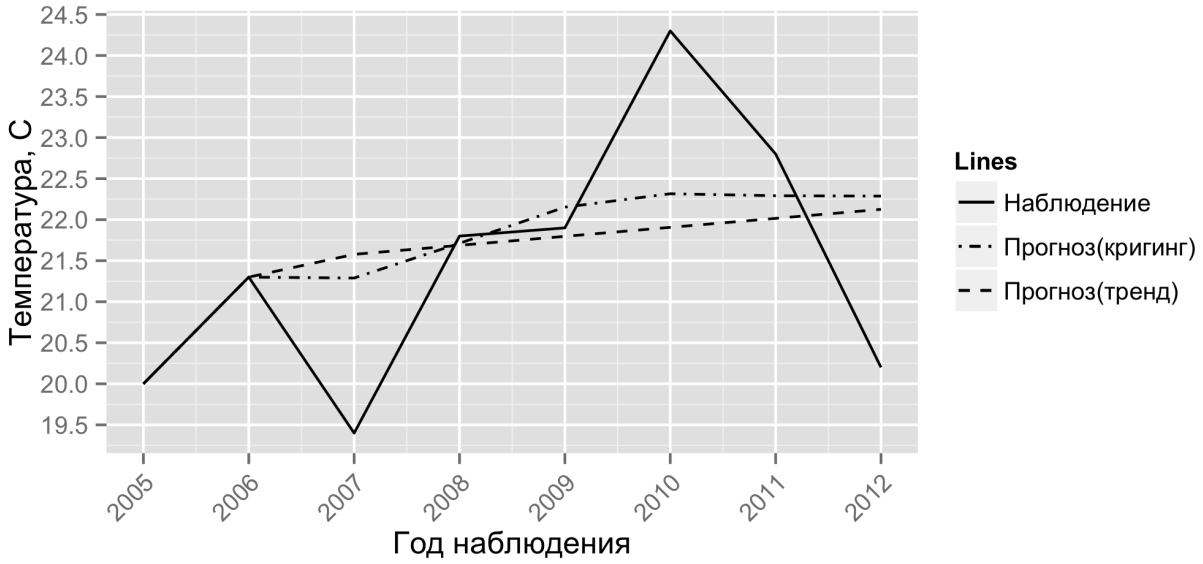


Рисунок Б.10 — Прогноз $0.9 + 4 \cdot Sph(h, 6.9)$

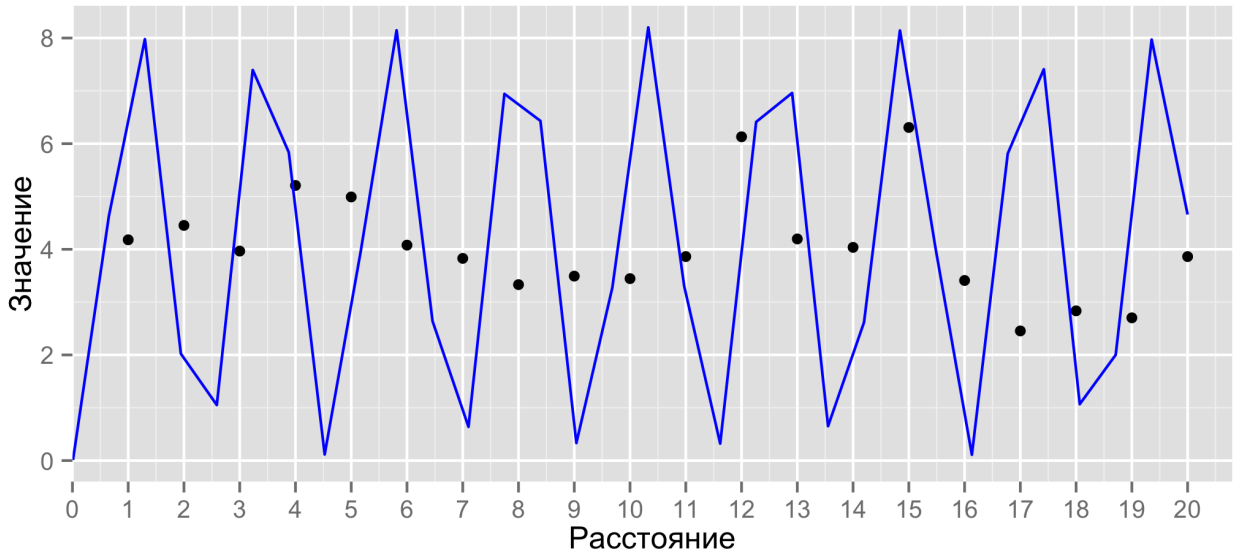


Рисунок Б.11 — Экспериментальная и теоретическая вариограмма $4 \cdot Per(h, 0.898)$

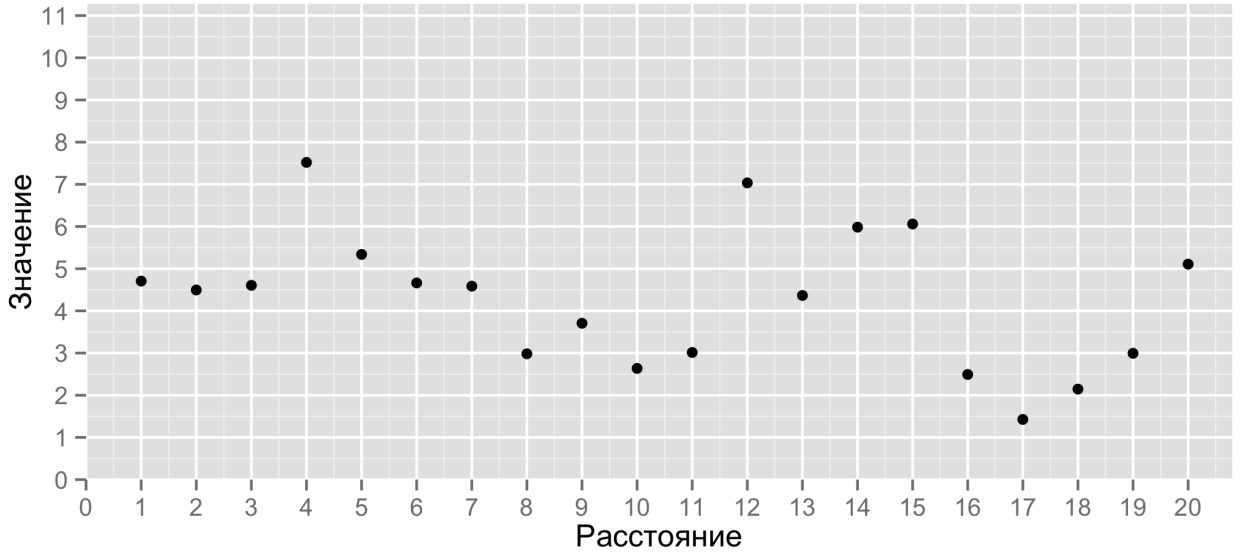


Рисунок Б.12 — Экспериментальная вариограмма (оценка Кресси-Хокинса)

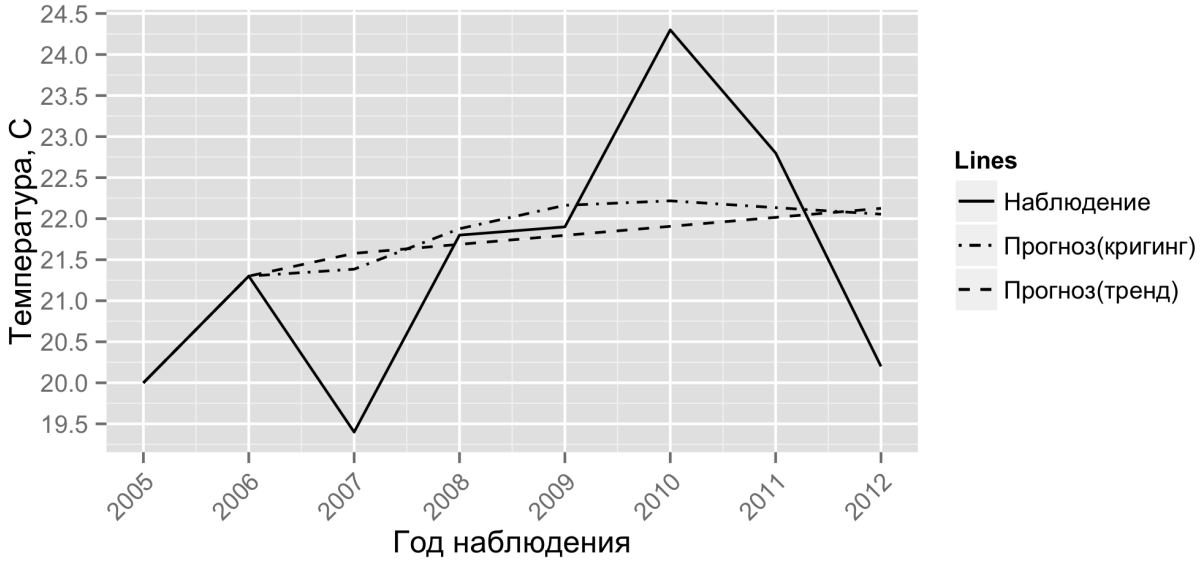


Рисунок Б.13 — Сравнение прогнозных значений (робастная оценка)

# Приложение В Результаты вычислений

| | year | temperature |
|---|---|---|
| 1 | 1975.00 | 2.13 |
| 2 | 1976.00 | -2.18 |
| 3 | 1977.00 | -0.59 |
| 4 | 1978.00 | -1.65 |
| 5 | 1979.00 | -1.01 |
| 6 | 1980.00 | -1.84 |
| 7 | 1981.00 | 1.07 |
| 8 | 1982.00 | 0.16 |
| 9 | 1983.00 | 2.45 |
| 10 | 1984.00 | 0.34 |
| 11 | 1985.00 | 1.23 |
| 12 | 1986.00 | -2.78 |
| 13 | 1987.00 | -2.29 |
| 14 | 1988.00 | 4.30 |
| 15 | 1989.00 | 0.29 |
| 16 | 1990.00 | -1.21 |
| 17 | 1991.00 | 3.18 |
| 18 | 1992.00 | 1.97 |
| 19 | 1993.00 | -2.04 |
| 20 | 1994.00 | 1.25 |
| 21 | 1995.00 | -1.36 |
| 22 | 1996.00 | -1.27 |
| 23 | 1997.00 | 0.52 |
| 24 | 1998.00 | -2.19 |
| 25 | 1999.00 | 2.80 |
| 26 | 2000.00 | 0.19 |
| 27 | 2001.00 | 3.28 |
| 28 | 2002.00 | 2.07 |
| 29 | 2003.00 | -3.14 |
| 30 | 2004.00 | -2.15 |
| 31 | 2005.00 | -1.36 |
| 32 | 2006.00 | -0.17 |

Таблица В.1 — Временной ряд остатков.

| | Год | Наблюдение | Прогноз | Тренд | Ошибка |
|---|---|---|---|---|---|
| 1 | 2007 | 19.400 | 22.133 | 21.578 | -2.733 |
| 2 | 2008 | 21.800 | 23.106 | 21.687 | -1.306 |
| 3 | 2009 | 21.900 | 23.441 | 21.797 | -1.541 |
| 4 | 2010 | 24.300 | 23.028 | 21.906 | 1.272 |
| 5 | 2011 | 22.800 | 22.122 | 22.016 | 0.678 |
| 6 | 2012 | 20.200 | 21.219 | 22.126 | -1.019 |

Таблица В.2 — Прогноз (периодическая модель)

# Приложение Г Код программ

```
# Descriptive statistics

# Function for getting all descriptive statistics
dstats.describe <- function(data, type="", locale=FALSE, shiny=FALSE) {
  cv <- dstats.coef.var(data)
  stats <- c(dstats.mean(data), dstats.median(data), dstats.quartile.lower(data),
              dstats.quartile.upper(data), dstats.min(data), dstats.max(data),
              dstats.range(data), dstats.quartile.range(data), dstats.variance(data),
              dstats.std.dev(data), if(!is.na(cv)){cv}, dstats.std.error(data),
              dstats.skew(data), dstats.std.error.skew(data), dstats.kurtosis(data),
              dstats.std.error.kurtosis(data))

  if(nchar(type)) {
    dstats.write(data=data, type=type) ## TODO: need to improve — now it computes two times the same things
  }
  if (locale) {
    descr.row <- c("Среднее", "Медиана", "Нижний квартиль", "Верхний квартиль",
                   "Минимум", "Максимум", "Размах", "Квартильный размах",
                   "Дисперсия", "Стандартное отклонение", if(!is.na(cv)) {"Коэффициент вариации"},
                   "Стандартная ошибка", "Асимметрия", "Ошибка асимметрии",
                   "Эксцесс", "Ошибка эксцесса")
    descr.col <- c("Значение")
  } else {
    descr.row <- c("Mean", "Median", "Lower Quartile", "Upper Quartile", "Range",
                   "Minimum", "Maximum", "Quartile Range", "Variance", "Standard Deviation",
                   if (!is.na(cv)) {"Coefficient of Variance"}, "Standard Error", "Skewness",
                   "Std. Error Skewness", "Kurtosis", "Std. Error Kurtosis")
    descr.col <- c("Value")
  }
  if (!shiny) {
    df <- data.frame(stats, row.names=descr.row)
    colnames(df) <- descr.col
  } else {
    df <- data.frame(descr.row, sapply(stats, format, digits=2, scientific=FALSE, nsmall=1))
    colnames(df) <- c("Статистика", "Значение")
  }

  df
}

dstats.mean <- function(data, ...) {
  m <- mean(data, ...)
  if (m < .0000001) {
    m <- 0
  }
  m
}

dstats.median <- function(data, ...) {
```

```
50    median(data, ...)
51  }
52
53  dstats.quartile.lower <- function(data, ...) {
54    quantile(data, ...)[[2]]
55  }
56
57  dstats.quartile.upper <- function(data, ...) {
58    quantile(data, ...)[[4]]
59  }
60
61  dstats.quartile.range <- function(data) {
62    dstats.quartile.upper(data) - dstats.quartile.lower(data)
63  }
64
65  dstats.min <- function(data, ...) {
66    min(data, ...)
67  }
68
69  dstats.max <- function(data, ...) {
70    max(data, ...)
71  }
72
73  dstats.range <- function(data) {
74    max(data) - min(data)
75  }
76
77  dstats.variance <- function(data, ...) {
78    var(data, ...)
79  }
80
81  dstats.std.dev <- function(data) {
82    sd(data)
83  }
84
85  dstats.coef.var <- function(data) {
86    mn <- mean(data)
87    if (abs(mn) > 1.987171e-15) {
88      (var(data) / mean(data)) * 100
89    } else
90      NA
91  }
92
93  dstats.std.error <- function(data) {
94    sd(data) / sqrt(length(data))
95  }
96
97  dstats.skew <- function(data) {
98    n <- length(data)
99    mean <- mean(data)
100   (n * sum(sapply(data, FUN=function(x){(x - mean)^3}))) /
101     ((n - 1) * (n - 2) * dstats.std.dev(data)^3)
102 }
103
104 dstats.std.error.skew <- function(data) {
105   n <- length(data)
106   sqrt((6 * n * (n - 1)) / ((n - 2) * (n + 1) * (n + 3)))
107 }
108
109 dstats.test.skew <- function(data) {
```

```
  dstats.skew(data) / dstats.std.error.skew(data)
}

dstats.kurtosis <- function(data) {
  n <- length(data)
  mean <- mean(data)
  (n * (n + 1) * sum(sapply(data, FUN=function(x){(x - mean)^4})) - 3 * (sum(
      sapply(data, FUN=function(x){(x - mean)^2})))^2 * (n - 1)) /
    ((n - 1) * (n - 2) * (n - 3) * dstats.variance(data)^2)
}

dstats.std.error.kurtosis <- function(data) {
  n <- length(data)
  2 * dstats.std.error.skew(data) * sqrt((n^2 - 1) / ((n - 3) * (n + 5)))
}

dstats.test.kurtosis <- function(data) {
  dstats.kurtosis(data) / dstats.std.error.kurtosis(data)
}

dstats.write <- function (data, type) {
  WriteDescriptiveStatistic(expression=dstats.mean(data), type=type, name="mean"
      )
  WriteDescriptiveStatistic(expression=dstats.variance(data), type=type, name="
      variance")
  WriteDescriptiveStatistic(expression=paste(format(dstats.coef.var(data),
      nsmall=2, digits=4), "\\%"), type=type, name="coef-var")
  WriteDescriptiveStatistic(expression=dstats.skew(data), type=type, name="skew"
      )
  WriteDescriptiveStatistic(expression=dstats.kurtosis(data), type=type, name="
      kurtosis")
  WriteDescriptiveStatistic(expression=dstats.test.skew(data), type=type, name="
      test-skew")
  WriteDescriptiveStatistic(expression=dstats.test.kurtosis(data), type=type,
      name="test-kurtosis")
}
```

Листинг D.1: Описательные статистики

```
## Cleaning up the workspace
rm(list=ls(all=TRUE))

## Dependencies
library(ggplot2)  # eye-candy graphs
library(xtable)   # convert data to latex tables
library(outliers) # tests for outliers
library(tseries)  # adf test used
library(nortest)  # tests for normality
library(sp)       # spatial data
library(gstat)    # geostatistics
library(reshape2) # will see

## Import local modules
source("R/lib/plot.R")       # useful functions for more comfortable plotting
source("R/lib/dstats.R")     # descriptive statistics module
source("R/lib/misc.R")       # some useful global-use functions
source("R/lib/draw.R")       # helpers for drawing
source("R/lib/write.R")      # helpers for writing
source("R/lib/ntest.R")      # tests for normality
source("R/lib/regr.R")
source("R/lib/measures.R")
```

```
23
24 ## Read the data / pattern: year;temperature
25 path.data <- "data/batorino_july.csv" # this for future shiny support and may be
       choosing multiple data sources
26 nrows <- 38
27 src  <- read.csv(file=path.data, header=TRUE, sep=";", nrows=nrows, colClasses=c
       ("numeric", "numeric"), stringsAsFactors=FALSE)
28
29 ## Global use constants
30 kDateBreaks <- seq(min(src$year) - 5, max(src$year) + 5, by=2) # date points for
       graphs
31
32 ## For the reason of prediction estimation and comparison, let cut observations
       number by 3
33 kObservationNum <- length(src[, 1]) - 6
34 WriteCharacteristic(expression=kObservationNum, type="original", name="n")
35
36 ## Source data as basic time series plot: points connected with line
37 plot.source <- DrawDataRepresentation(data=src, filename="source.png",
       datebreaks=kDateBreaks)
38
39 print(xtable(src, caption="Исходные данные.", label="table:source"),  table.
       placement="H",
40         file="out/original/data.tex")
41
42 ## Form the data for research
43 sample <- src[0:kObservationNum, ]
44
45 # Getting descriptive statistics for temperature in russian locale
46 sample.dstats <- dstats.describe(sample$temperature, type="original", locale=
       TRUE)
47 print(xtable(sample.dstats, caption="Описательные статистики для наблюдаемых тем
       ператур.", label="table:dstats"),
48         file="out/original/dstats.tex")
49
50 # Compute Sturges rule for output
51 WriteCharacteristic(expression=nclass.Sturges(sample$temperature), type="
       original", name="sturges")
52
53 ## Basic histogram based on Sturges rule (by default) with pretty output (also
       by default)
54 plot.data.hist <- DrawHistogram(data=sample, filename="original/histogram.png")
55
56 ## Tests for normality
57 sample.shapiro <- ntest.ShapiroWilk(data=sample$temperature, type="original",
       name="shapiro")
58 sample.pearson <- ntest.PearsonChi2(data=sample$temperature, type="original",
       name="pearson")
59 sample.ks      <- ntest.KolmogorovSmirnov(data=sample$temperature, type="
       original", name="ks")
60
61 ## Normal Quantile-Quantile plot // TODO: check when it appears in text
62 plot.data.qq <- DrawQuantileQuantile(data=sample$temperature, filename="original
       /quantile.png")
63
64 ## Scatter plot with regression line
65 plot.data.scatter <- DrawScatterPlot(sample, filename="original/scatterplot.png"
       , kDateBreaks);
66
67 ## Grubbs test for outliers
```

```
68 sample.grubbs <- grubbs.test(sample$temperature)
69 WriteTest(sample.grubbs$statistic[1], sample.grubbs$p.value, type="original",
       name="grubbs")
70
71 ## Compute correlation for output
72 sample.correlation <- cor(x=sample$year, y=sample$temperature)
73 WriteCharacteristic(sample.correlation, type="original", name="correlation")
74
75 WriteTest(sample.correlation * sqrt(kObservationNum - 2)/(1 - sample.correlation
       ^2), 0, qt(1 - 0.05, kObservationNum - 2), type="original", name="student")
76
77 ## Pearson's product-moment correlation test. Use time for y as numerical
78 sample.ctest <- cor.test(sample$temperature, c(1:kObservationNum), method="
       pearson")
79 WriteTest(sample.ctest$statistic, sample.ctest$p.value, sample.ctest$parameter
       [[1]], type="original", name="correlation")
80
81 ## Fitting linear model for researching data. It also compute residuals based on
        subtracted regression
82 sample.fit <- lm(sample$temperature ~ c(1:kObservationNum))
83
84 linear <- function(x, a, b) a * x + b
85 sample.residuals.prediction.trend <- data.frame("Год"=src$year[(kObservationNum
       + 1):nrows],
86                                                 "Актуальное"=src$temperature[(
                                                       kObservationNum + 1):nrows
                                                       ],
87                                                 "Прогнозное"=sapply(X=
                                                       ConvertYearsToNum(src$year
                                                       [(kObservationNum + 1):
                                                       nrows]), FUN=linear, a=
                                                       sample.fit$coefficients
                                                       [[2]], b=sample.fit$
                                                       coefficients[[1]]))
88 print(xtable(sample.residuals.prediction.trend, caption="Сравнение прогнозных зн
       ачений (тренда)", label="table:prediction_trend", digits=c(0, 0, 2, 2)),
89   file="out/residual/prediction-trend.tex")
90
91 ## Time series (which is by default is research data) with trend line based on
       linear module estimate (lm)
92 plot.data.ts <- DrawTimeSeries(data=sample, filename="original/time-series.png",
        datebreaks=kDateBreaks)
93
94 ## Next step is research residuals computed few lines above
95 sample.residuals <- data.frame("year"=sample$year, "temperature"=sample.fit$
       residuals)
96 print(xtable(sample.residuals, caption="Временной ряд остатков.", label="table:
       residuals"), table.placement="H",
97       file="out/residual/data.tex")
98
99 sign <- regr.significance(sample$temperature, write=TRUE)
100 adeq <- regr.adequacy(sample$temperature, write=TRUE)
101
102 ## Residuals time series (data have gotten on computing step: fitting linear
       model)
103 plot.residuals.ts <- DrawTimeSeries(data=sample.residuals, filename="residual/
       time-series.png", datebreaks=kDateBreaks)
104
105 ## Descriptive statistics for residuals
106 sample.residuals.dstats <- dstats.describe(sample.residuals$temperature, type="
```

```
    residual", locale=TRUE)
print(xtable(sample.residuals.dstats, caption="Описательные статистики остатков"
    , label="table:residuals_dstats"),
      file="out/residual/dstats.tex")

## Basic histogram for residuals / seems like the same as for non-residuals
plot.residuals.hist <- DrawHistogram(data=sample.residuals, filename="residual/
    histogram.png")

## Tests for normality
sample.shapiro <- ntest.ShapiroWilk(data=sample.residuals$temperature, type="
    residual", name="shapiro")
sample.pearson <- ntest.PearsonChi2(data=sample.residuals$temperature, type="
    residual", name="pearson")
sample.ks      <- ntest.KolmogorovSmirnov(data=sample.residuals$temperature,
    type="residual", name="ks")

## Normal Quantile-Quantile plot for residuals
plot.residuals.qq <- DrawQuantileQuantile(data=sample.residuals$temperature,
    filename="residual/quantile.png")

## Auto Correlation Function plot
plot.residuals.acf <- DrawAutoCorrelationFunction(data=sample$temperature,
    filename="residual/acf.png")

## Box-Ljung and adf tests (some kind of stationarity and independence tests) //
     TODO: need to know exactly in theory what it is
sample.residuals.box <- Box.test(sample.residuals$temperature, type="Ljung-Box")
WriteTest(sample.residuals.box$statistic, sample.residuals.box$p.value, sample.
    residuals.box$parameter[[1]], type="residual", name="ljung-box")

sample.residuals.adf <- adf.test(sample.residuals$temperature)
WriteTest(sample.residuals.adf$statistic, sample.residuals.adf$p.value, type="
    residual", name="stationarity")

source("R/predictor.R")
```

Листинг D.2: Основной код программы

```
source("R/lib/afv.R")
source("R/lib/variogram.R")
source("R/lib/kriging.R")

## Function definition: need to be moved into isolated place
# Completes trend values up to source observation number
computeTrend <- function (fit, future=0) {
  c(sapply(c(1 : (nrows + future)), FUN=function(x) fit$coefficients[[1]] + x *
      fit$coefficients[[2]]))
}

# Computes prediction with passed parameters and saves all needed info and plots
processPrediction <- function (data, year, variogram, cressie, cutoff, name,
    caption) {

  prediction <- PredictWithKriging(data, x=ConvertYearsToNum(year), observations
      =kObservationNum, variogram_model=variogram$var_model, nrows=nrows)
  CrossPrediction(src$temperature, src$year, trend, prediction, name,
      observations=kObservationNum, nrows=nrows)
  residual <- ComputeKrigingResiduals(src$temperature, trend, prediction,
      observations=kObservationNum, nrows=nrows)
  mse <- MSE(residual)
```

```
18
19    prediction.compare <- data.frame("Год"=src$year[(kObservationNum + 1):nrows],
20      "Наблюдение"=src$temperature[(kObservationNum + 1):nrows],
21      "Прогноз"=prediction$var1.pred+trend[(kObservationNum + 1):nrows],
22      "Тренд"=trend[(kObservationNum + 1):nrows],
23      "Ошибка"=residual)
24    print(xtable(prediction.compare, caption=caption, label=paste0("table:", name,
          "-prediction"), digits=c(0, 0, 3, 3, 3, 3)),
25      file=paste0("out/variogram/", name, "-prediction.tex"))
26
27    WriteCharacteristic(mse, type="variogram", name=paste0(name, "-mse"))
28
29    list(variogram=variogram, prediction=prediction, residual=residual, mse=mse)
30  }
31
32  trend <- computeTrend(sample.fit)
33  sample.residuals <- sample.fit$residuals
34
35  cutoff <- trunc(2 * kObservationNum / 3) # let it be "classical" value
36
37  # Draw H-Scatterplot
38  sample.hscat <- DrawHScatterplot(sample.residuals[1:kObservationNum])
39
40  lin.var1 <- ComputeManualVariogram(data=sample.residuals, x=sample$year, cressie
      =FALSE, cutoff=20, model="Lin", name="lin", psill=4, range=0, nugget=0, fit=
      FALSE)
41  lin.fit <- ComputeManualVariogram(data=sample.residuals, x=sample$year, cressie=
      FALSE, cutoff=20, model="Lin", name="lin-fit", psill=4, range=0, nugget=0,
      fit=TRUE)
42  lin.fit.cv <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
      cressie=FALSE, cutoff=20, model="Lin", name="lin-fit-cv", psill=4, range=4,
      nugget=0, fit=FALSE)
43  lin.fit.adapt <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
      cressie=FALSE, cutoff=20, model="Lin", name="lin-fit-adapt", psill=4, range
      =2, nugget=0, fit=FALSE)
44  lin.fit.cv.prediction <- processPrediction(data=sample.residuals, year=sample$
      year, variogram=lin.fit.cv, cutoff=cutoff, name="lin-fit-cv", caption="Прогно
      з (линейная модель с порогом)")
45  lin.fit.adapt.prediction <- processPrediction(data=sample.residuals, year=sample
      $year, variogram=lin.fit.adapt, cutoff=cutoff, name="lin-fit-adapt", caption=
      "Адаптивный прогноз (линейная модель с порогом)")
46  sph.fit.adapt <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
      cressie=FALSE, cutoff=20, model="Sph", name="sph-fit-adapt", psill=4, range
      =6.9, nugget=0.9, fit=FALSE)
47  sph.fit.adapt.prediction <- processPrediction(data=sample.residuals, year=sample
      $year, variogram=sph.fit.adapt, cutoff=cutoff, name="sph-fit-adapt", caption=
      "Адаптивный прогноз (сферическая модель)")
48  per.fit.cv <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
      cressie=FALSE, cutoff=20, model="Per", name="per-fit-cv", psill=4.1, range
      =0.898, nugget=0.001, fit=FALSE)
49  per.fit.cv.prediction <- processPrediction(data=sample.residuals, year=sample$
      year, variogram=per.fit.cv, cutoff=cutoff, name="per-fit-cv", caption="Прогно
      з (периодическая модель)")
50
51  for.robust.only <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
      cressie=TRUE, cutoff=20, model="Lin", name="robust", psill=0, range=0, nugget
      =0, fit=FALSE)
52
53  # Compute prediction manually with choosed model ("best" what i found)
54  manual <- processPrediction(data=sample.residuals, year=sample$year, variog=
```

```
    ComputeManualVariogram, cressie=FALSE, cutoff=cutoff, name="manual", caption=
    "Прогноз (сферическая модель)")

# Compute prediction with auto fit model using classical estimation
classical <- processPrediction(data=sample.residuals, year=sample$year, cressie=
    FALSE, cutoff=cutoff, name="classical", caption="Прогноз (классическая оценка
    )")

# Compute prediction with auto fit model using robust (cressie) estimation
robust <- processPrediction(data=sample.residuals, year=sample$year, cressie=
    TRUE, cutoff=cutoff, name="robust", caption="Прогноз (робастная оценка)")

models.comparison <- CompareClassicalModels(manual$variogram, classical$
    variogram, filename="figures/variogram/models-comparison.png")

# Find best cutoff parameters
cutoff <- ComparePredictionParameters(sample.residuals, trend, ConvertYearsToNum
    (sample$year), filename="figures/variogram/parameter-comparison.png",
    observations=kObservationNum, nrows=nrows)

manual.best    <- processPrediction(data=sample.residuals, year=sample$year,
    variog=ComputeManualVariogram, cressie=FALSE, cutoff=cutoff$manual, name="
    manual-best", caption="Наилучший прогноз (сферическая модель)")
classcial.best <- processPrediction(data=sample.residuals, year=sample$year,
    cressie=FALSE, cutoff=cutoff$classical, name="classical-best", caption="Наилу
    чший прогноз (классическая оценка)")
robust.best    <- processPrediction(data=sample.residuals, year=sample$year,
    cressie=TRUE, cutoff=cutoff$robust, name="robust-best", caption="Наилучший пр
    огноз (робастная оценка)")
```

Листинг D.3: Вариограммный анализ