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 с., 3 ч., 10 рис., 7 табл., 29 источников, 4 прил.

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, 10 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 данных 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 исследования 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

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 Ω является произвольным множеством элементарных событий, $\mathcal{F}$ — сигма-алгеброй подмножеств Ω, 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:

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

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, $t_1, t_2 \in \mathbb{T}$.

Определение 1.13. Случайный процесс $X(t), t \in \mathbb{T}$ называется гауссовским, если любые n его отсчетов $X(t_1), X(t_2), \ldots, X(t_n)$, где $t_1, t_2, \ldots, t_n \in \mathbb{T}$ имеют n-мерное нормальное распределение, то есть

$$F_n(\cdot) \equiv \Phi_n(\cdot) \forall n. \tag{1.4}$$

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

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

где χ_1^2 — случайная величина, распределенная по закону *хи-квадрат* с одной степенью свободы.

При этом

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

$$V\{X(t+h) - X(t)\}^2 = 2(2\gamma(h))^2. \tag{1.6}$$

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 $X(t), t \in \mathbb{Z}$.

Вариограмма процесса $X(t), 2\gamma(h)$, является неизвестной и, не нарушая общности, далее считаем $m(t) \equiv 0$, $V(t) \equiv \sigma^2, t \in \mathbb{Z}$.

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

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

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

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

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

Доказательство. Вычислим первый момент введённой 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 (1.5) 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** для вариограммы рассматриваемого процеса.

Найдём второй момент оценок 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 (1.6) 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):

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} =$$
$$= \frac{2}{(n-h_1)(n-h_2)}\sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2}(\gamma(t-h_2-s)+\gamma(t+h_1-s)-\gamma(t-s)-\gamma(t+h_1-s-h_2))^2 \qquad (2.6)$$

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*

$$\sum_{m=-\infty}^{+\infty} |\gamma(h)| < \infty, \tag{2.7}$$

то

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

$$= 2 \sum_{m=-\infty}^{+\infty} \gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2, \tag{2.8}$$

$$\lim_{n\to\infty} (n - h) V\{2\tilde{\gamma}(h)\} = 2 \sum_{m=-\infty}^{+\infty} \gamma(m - h) + \gamma(m + h) - 2\gamma(m))^2. \tag{2.9}$$

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

Доказательство. В (2.6) сделаем следующую замену переменных

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

Получим

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

$$= \frac{2}{(n - h_1)(n - h_2)} \sum_{t=1}^{n-h_1} \sum_{t-m=1}^{n-h_2} (\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2 \tag{2.10}$$

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

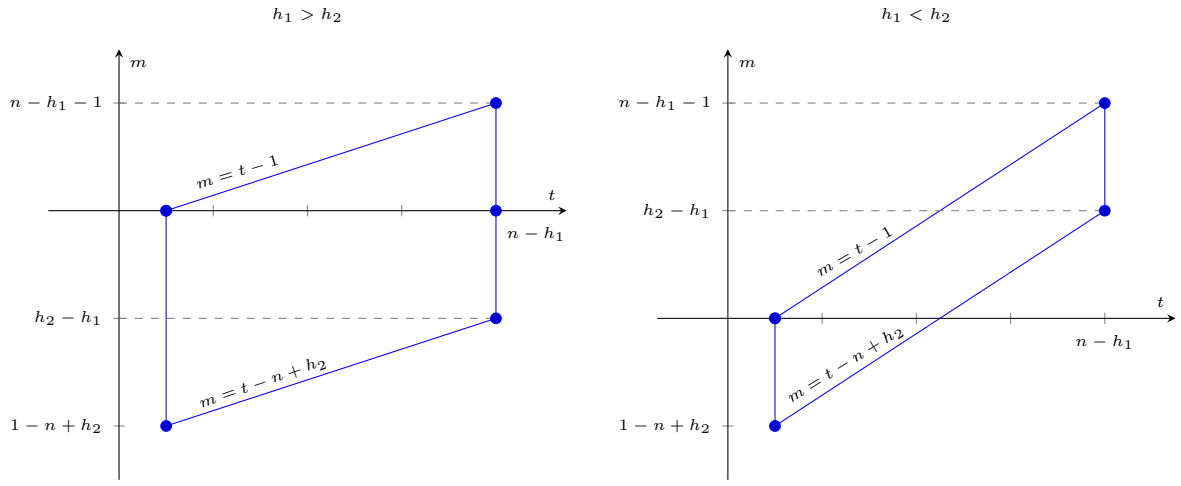

Рисунок 2.2.1 — Замена переменных

Рассмотрим первый случай: $h_1 > h_2$. Поменяем порядок суммирования в (2.10).

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

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

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

Найдём предел для случая $h_1 > h_2$:

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

Из предположения теоремы (2.7) ряд $\sum\limits_{m=-\infty}^{+\infty}\gamma(h)$, сходится абсолютно $\gamma(h)$ следовательно стремится к 0, при $n\to\infty$, быстрее, чем $\frac{1}{n}$. Тогда

$$\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2)$$

стремится к 0, при $m\to\infty$, быстрее, чем $\frac{1}{m}$, значит

$$\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2)\sim o(\frac{1}{m})$$
$$(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\sim o(\frac{1}{m^2})$$

Тогда ряд $\sum\limits_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2$ сходится при $n\to\infty$, а значит

$$\lim_{n\to\infty}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2=0\quad\text{и}$$

$$\lim_{n\to\infty}(n-h_2)\frac{2}{n-h_2}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{2}{n-h_1}\sum_{m=1}^{n-h_1-1}m(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\Big)=$$
$$=2\sum_{m=-\infty}^{+\infty}\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2\quad(2.11)$$

Рассуждая аналогично, получаем предел при $h_1 < h_2$:

$$\lim_{n\to\infty}(n-h_1)\frac{2}{n-h_1}\Big(\sum_{m=1-n+h_2}^{n-h_1-1}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2-$$
$$-\frac{2}{n-h_2}\sum_{m=1}^{n-h_2-1}m(\gamma(m+h_2)+\gamma(m-h_1)-\gamma(m)-\gamma(m-h_1+h_2))^2\Big)=$$
$$=2\sum_{m=-\infty}^{+\infty}\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2 \quad (2.12)$$

Тогда, объединяя вместе (2.11) и (2.12), получаем:

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

Нетрудно видеть, что если положить $h_1 = h_2 = h$, то

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

□

Глава 3
Анализ временного ряда в пакете R

3.1 Детерминированный подход

3.1.1 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 3.1.1.

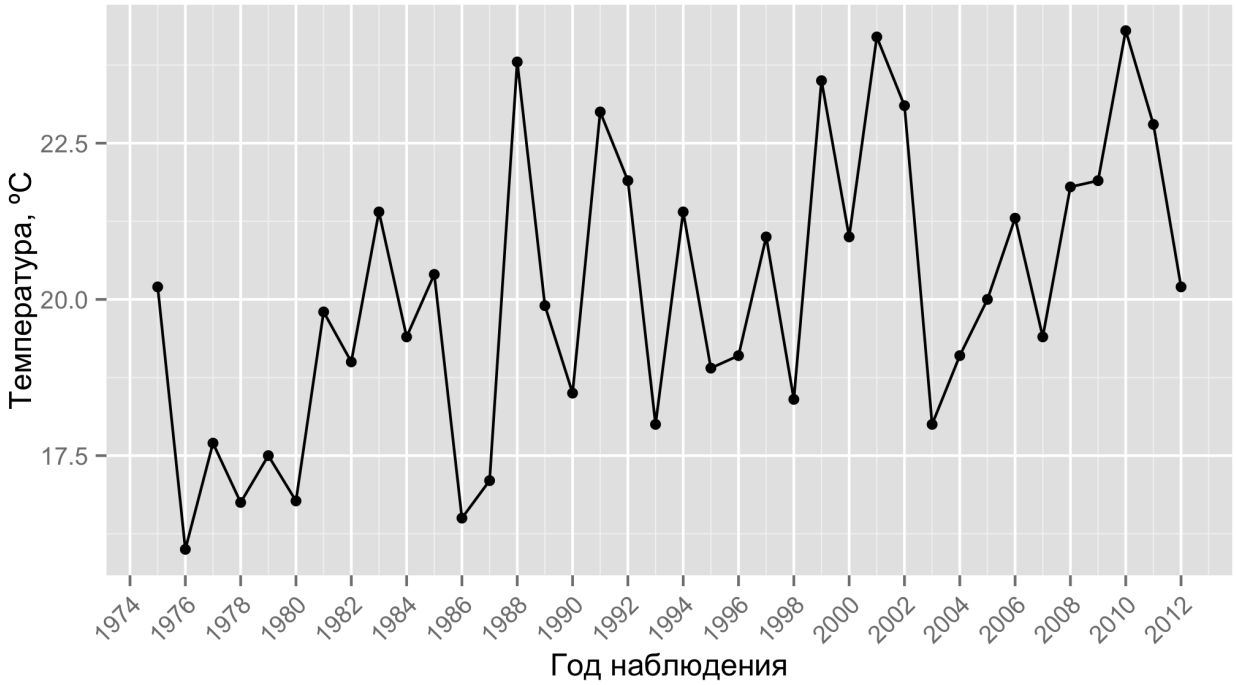

Рисунок 3.1.1 — График исходных данных

Следует отметить, что для непосредственного исследования были использованы наблюдения с 1975 по 2009 год. Наблюдения за 2010-2012 годы были намеренно исключены из исследования в целях дальнейшего оценивания результатов анализа и прогнозирования. Заметим, что работа, представленная в параграфах 3.1.1–3.1.3, была также проделана и для всей выборки. Так как поведение целой выборки сохранилось в уменьшенной, то, без потери общности, будем считать её исходной.

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 D листинге D.1. Данный модуль позволяет вычислять все рассмотренные в данной работе описательные статистики. Полученные результаты для исходных данных отображены в таблице 1.

	Значение
Среднее	19.88
Медиана	19.80
Нижний квартиль	18.20
Верхний квартиль	21.40
Минимум	16.00
Максимум	24.20
Размах	8.20
Квартильный размах	3.20
Дисперсия	4.92
Стандартное отклонение	2.22
Коэффициент вариации	24.75
Стандартная ошибка	0.37
Асимметрия	0.18
Ошибка асимметрии	0.40
Эксцесс	-0.79
Ошибка эксцесса	0.78

Таблица 1 — Описательные статистики для наблюдаемых температур.

Рассмотрим подробнее некоторые полученные статистики.

Как видно из таблицы, *средняя* температура в июле месяце за период с 1975 по 2009 составляет приблизительно 20°C.

Коэффициент вариации в нашем случае равен 24.75%. Из этого следует, что выборку можно считать однородной, так как полученное значение является меньшим 33% [10].

Коэффициент асимметрии — мера симметричности распределения. Полученное значение: 0.185. Данное значение говорит о незначительной правосторонней асимметрии распределения. То есть о том, что выборочное распределение можно считать близким к симметричному [12].

Коэффициент эксцесса в рассматриваемом случае равен -0.79. 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:

$$Z_{A_S} = \frac{A_S}{SES} = 0.465.$$

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:

$$Z_K = \frac{K}{SEK} = -1.02.$$

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 близость выборочного распределения к нормальному закону. 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, повзоляет сделать предположение о разновидности распределения. Для вычисления интервалов разбиения воспользуемся *формулой Стерджеса*. Из [16] количество интервалов разбиения рассчитывается по формуле:

$$k = \lceil log_2 n \rceil + 1 = \lceil log_2 35.00 \rceil + 1 = 7.00. \tag{3.1}$$

Так как по гистограмме можно визуально предположить близость выборочного распределения к нормальному распределению, нанесём на график кривую плотности нормального распределения. Построенная гистограмма отображена на рисунке 3.1.2. Проанализируем эту гистограмму. Во-первых, на ней наглядно представлена близость

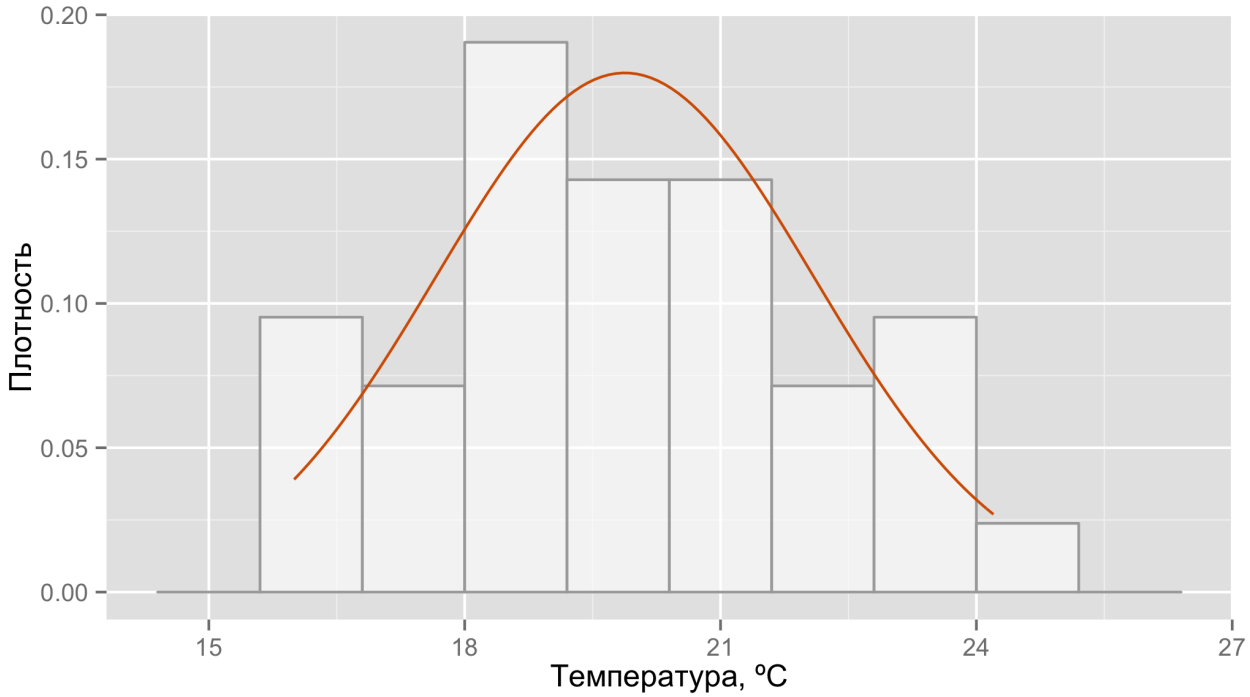

Рисунок 3.1.2 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения $\mathcal{N}(19.88, 4.92)$

выборочного распределения к нормальному с параметрами $\mathcal{N}(19.88, 4.92)$. Во-вторых, по этой гистограмме можно подтвердить или опровергнуть результаты, полученные на этапе вычисления описательных статистик.

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 стандартного 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 рисунок 3.1.3. На этом графике можно визуально обнаружить аномальное положение наблю-

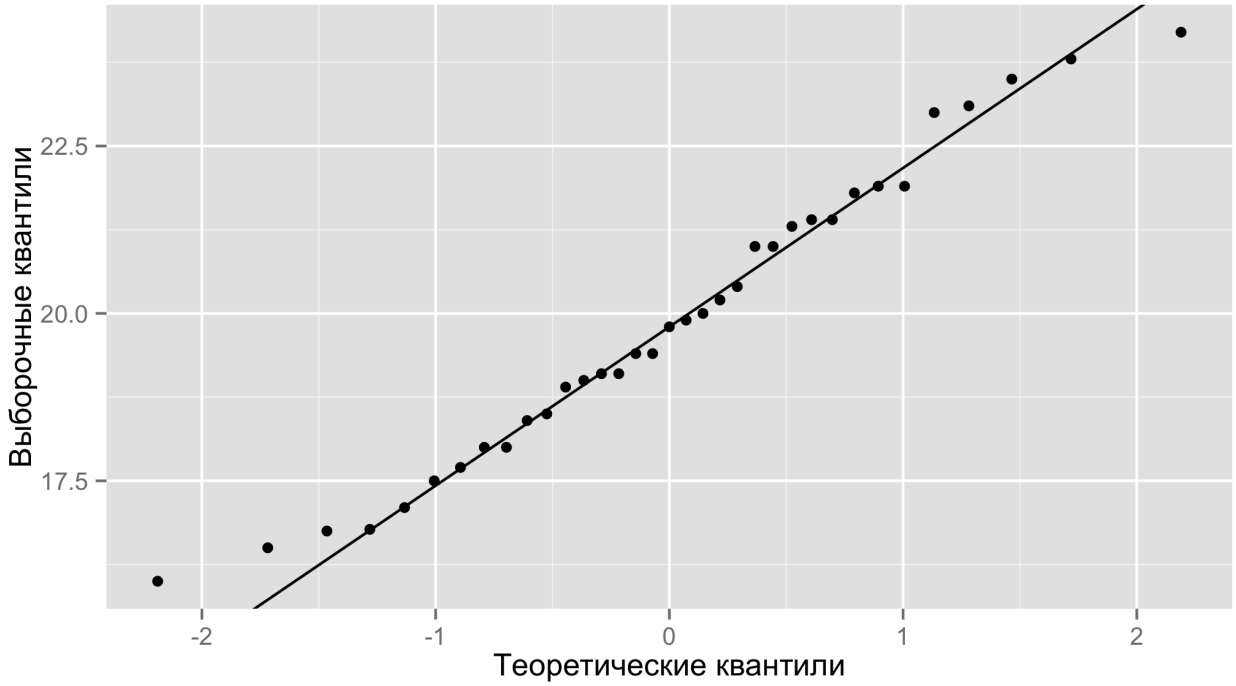

Рисунок 3.1.3 — График квантилей для наблюдаемых температур

даемых значений по отношению к нормальному распределению. В данном случае отклонения можно наблюдать на концах рассматриваемого промежутка. Остальные значения образуют отчетливую прямую. А значит, подтверждается предположение о нормальности выборочного распределения.

Далее следует проверить полученные результаты 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 сформулиировать 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 $p = 0.57 > 0.05$, а значит нулевая гипотеза не отвергается [18]. Следовательно опровергнуть предположение на основе данного теста нельзя.

Попробуем опровегнуть наше предположение на основе проверки критерия χ^2 Пирсона [19]. Для этого воспользуемся пакетом *nortest* и функцией *pearson.test*. Из полученных в **R** результатов, статистика χ^2 Пирсона $P = 2.80$. Вероятность ошибки $p = 0.83 > 0.05$, 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 χ^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 $D = 0.075$. Вероятность ошибки $p = 0.99 > 0.05$, а значит нулевую гипотезу отвергнуть нельзя. Следовательно опровергнуть предположение о нормальности, как и 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 $G = 1.950.89$, вероятность ошибки p-value $= 0.81$ — 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, наши 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 $\mathcal{N}(19.88, 4.92)$. Что подтверждается коэффициентами асимметрии и эксцесса из таблицы 1, а также результатами, полученными мной при исследовании в пакете **STATISTICA**. Следует также отметить, что эквивалентные результаты были получены и для всей выборки.

3.1.2 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.

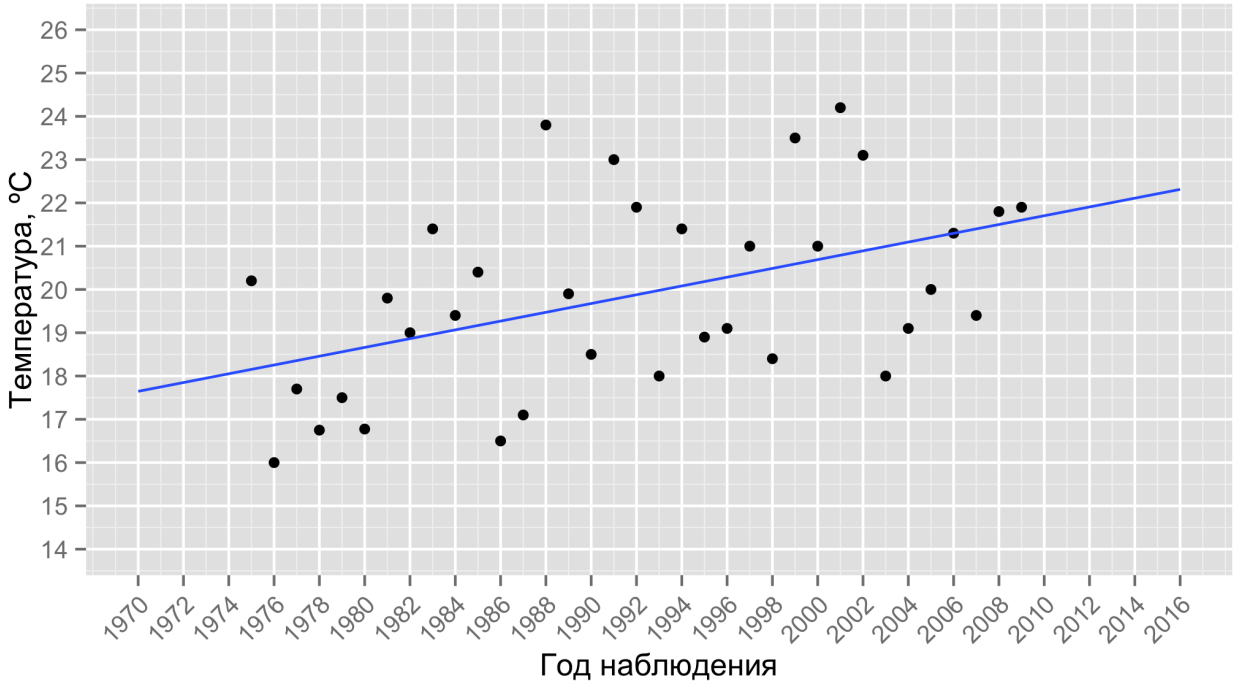

Рисунок 3.1.4 — Диаграмма рассеяния

Из рисунка 3.1.4 видно, что точки образуют своеобразное «облако», ориентированное по диагонали вверх, то есть присутствует некая зависимость между рассмтриваемыми переменными. Также, данная диаграмма наглядно показывает силу этой зависимости: так как точки не образуют чёткой формы, а разбросаны относительно диагонали, то можно говорить о наличии умеренной корреляции. То есть, нельзя сказать, что зависимость сильная, но и нельзя сказать, что связь между переменными отсутствует.

Проверим полученные результаты подробнее. Из расчётов в **R**, коэффициент корреляции $r_{xt} = 0.469$. Этим подтверждаются наши выводы из диаграмм рассеяния и концентрации о положительной корреляции, поскольку полученный коэффициент корреляции является положительным и присутствует умеренная зависимость: $r_{xt} \approx 0.5$.

Оценим значимость полученного выборочного коэффициента корреляции 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 $t = 3.05$, количество степеней свободы $df = 33$ и вероятность ошибки $p = 0.0045 < 0.05$, следовательно это говорит о том, что необходимо отвегнуть гипотезу $H_0 : r = 0$.

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

Следует также отметить, что аналогичный анализ, проведённый в пакете STATISTICA, подобным образом выявил зависимость между температурой воды и временем.

Следовательно, в рассматриваемом случае можно говорить о присутствии значимой зависимости между температурой воды в озере Баторино и временем.

3.1.3 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 файле данных наблюдаются через равные промежутки времени.

Большинство методов исследования временных рядов включает различные способы фильтрации шума, выделения сезонной и циклической составляющих, позволяющие увидеть регулярную составляющую более отчётливо.

Во временных рядах выделяют три составляющие:

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

2. *Циклические (K), сезонные (S) колебания* — это составляющие, которые проявляются как отклонения от основной тенденции развития изучаемого явления, и связанны с действие краткосрочных, систематических факторов развития. Циклические колебания состоят в том, что значения признака в течение какого-то времени возрастают, достигают определённого максимума, затем убывают, достигают определённого минимума, вновь возрастают до прежних значений и т.д. Эту составляющую можно выявить только по данным за длительные промежутки времени, например, в 10, 15 или 20 лет. Сезонные колебания — это колебания, периодически повторяющиеся в некоторое определённое время каждого года, месяца, недели, дня. Эти изменения отчётливо наблюдаются на графиках рядов динамики, содержащих данные за период не менее одного года.

3. *Нерегулярная случайная составляющая (ошибка) (E)*, 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 (X):

- Аддитивная модель: $X = T + K + S + E$;
- Мультипликативная модель: $X = T \times K \times S \times E$.

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 нашем временном ряде циклическая и сезонная составляющие отсутствуют. Построим график временного ряда.

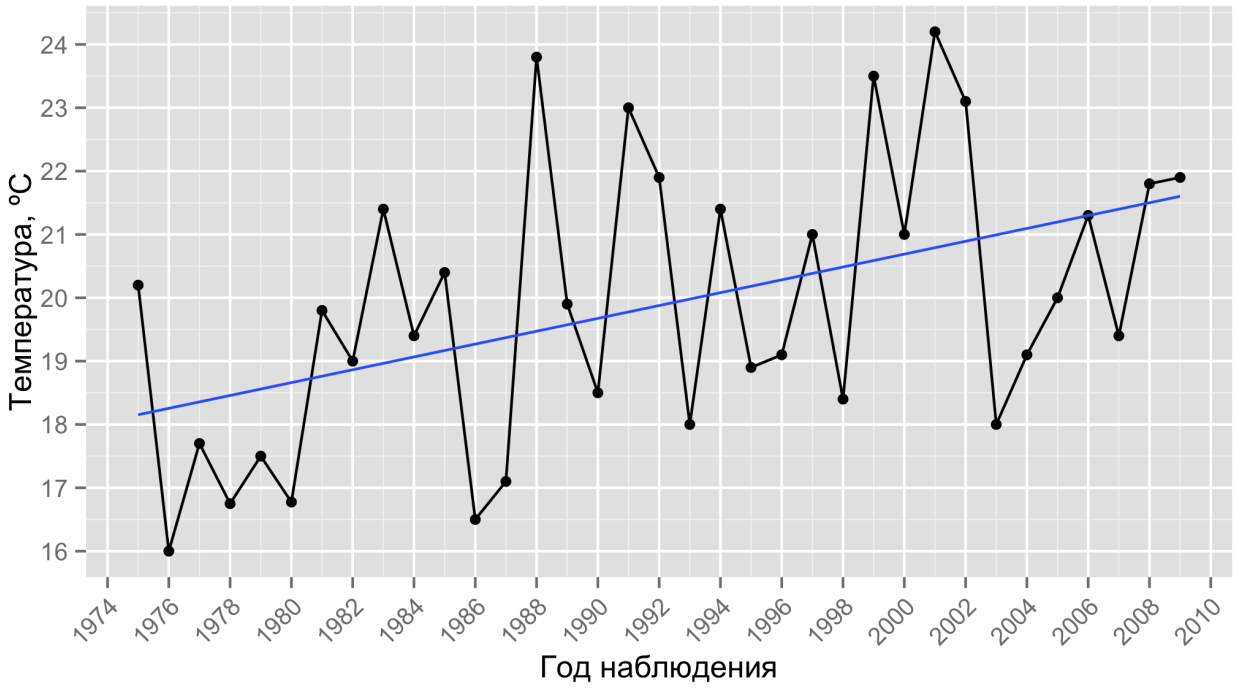

Рисунок 3.1.5 — График временного ряда с линией регрессии

На полученном графике можно заметить явно выраженный линейный рост значений со временем — он проиллюстрирован на графике прямой. Эта составляющая нашего временного ряда — тренд. Из этого следует, что уравнение тренда имеет вид:

$$x(t) = at + b.$$

Продолжая рассуждение, как наблюдение из графика, можно отметить, что не происходит увеличения амплитуды колебаний с течением времени. А значит, данная модель является аддитивной. Из всего вышесказанного можно заключить, что модель исходного временного ряда имеет вид:

$$X = T + E.$$

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(тренд), остатки после удаления тренда. Коэффицинты, полученные с помощью данной функции представлены в (3.2).

$$a = 0.1014, \quad b = 18.0521. \tag{3.2}$$

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

Таким образом получена линейная модель, описывающая тенденцию развития:

$$x(t) = at + b = 0.1014t + 18.0521 \tag{3.3}$$

На основе полученной линейной модели (3.3), построим ряд остатков(приложение С, таблица С.1), удалив тренд из исходного ряда. Полученный ряд графически представлен на рисунке В.1 в приложении В.

Проведём анализ полученной регрессионной модели. Для этого проверим значимость полученных коэффициентов регрессии и оценим адекватность регрессионной модели. Рассчитаем вспомогательные величины, воспользовавшись [23]. Дисперсия отклонения

$$\sigma_\varepsilon^2 \approx 3.823,$$

стандартные случайные погрешности параметров a, b:

$$\sigma_a \approx 0.029, \quad \sigma_b \approx 0.356.$$

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

$$T_a = 38.2, \quad T_b = 50.5.$$

Число степеней свободы $k = 36$, $t_{\text{кр}}(k, \alpha) = 2.028$.

- $|T_a| > t_{\text{кр}} \Rightarrow$ 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:

$$\overline{\sigma^2} \approx 1.44.$$

Остаточная дисперсия:

$$\overline{D} \approx 3.7.$$

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

$$F_{\text{крит}} \approx 14.01,$$

при степенях свободы $v_1 = 1, v_2 = 36, F_{\text{табл}}(v_1, v_2, \alpha) = 4.11$.

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

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

Рассчитаем коэффициент детерминации:

$$\eta_{x(t)}^2 \approx 0.275.$$

Проверим отклонение от линейности: $\eta_{x(t)}^2 - r_{xt}^2 \approx 0.0044 \leq 0.1$. Следовательно отклонение от линейности незначительно. Но при этом коэффициент детерминации оказался не достаточно 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. Вычисленые прогнозные значения на 2010-2012 годы для сравнения отображены на таблице 2:

Имеющееся отклонение прогнозов от реальных данных ещё раз подтверждает, что построенная модель временного ряда обладает невысокой точностью.

	Год	Актуальное	Прогнозное
1	2010	24.30	18.15
2	2011	22.80	18.25
3	2012	20.20	18.36

Таблица 2 — Сравнение прогнозных значений

3.1.4 Анализ остатков

Проанализируем временной ряд остатков. Для этого проверим свойства, которым должна удовлетворять нерегулярная составляющая ε:

1. $E(\varepsilon) = 0$.
2. Дисперсия ε 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	0.14
Нижний квартиль	-1.80
Верхний квартиль	1.28
Минимум	-2.99
Максимум	4.33
Размах	7.32
Квартильный размах	3.07
Дисперсия	3.84
Стандартное отклонение	1.96
Коэффициент вариации	0.00
Стандартная ошибка	0.33
Асимметрия	0.42
Ошибка асимметрии	0.40
Эксцесс	-0.77
Ошибка эксцесса	0.78

Таблица 3 — Описательные статистики остатков

Как видно из таблицы 3, среднее значение равно нулю. При этом коэффициенты асимметрии($A_S = 0.424$) и эксцесса($K = -0.773$) указывают на большее отклонение распределения остатков от нормального закона.

Построим гистограмму и график квантилей для проверки последних заключений. Построенная гистограмма (приложение В, рисунок В.2) наглядно демонстрирует полученные в таблице 3 коэффициенты асимметрии и эксцесса.

Для проверки нормальности построим график квантилей. На рисунке 3.1.6 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 $p = 0.12 > 0.05$, а значит нулевая

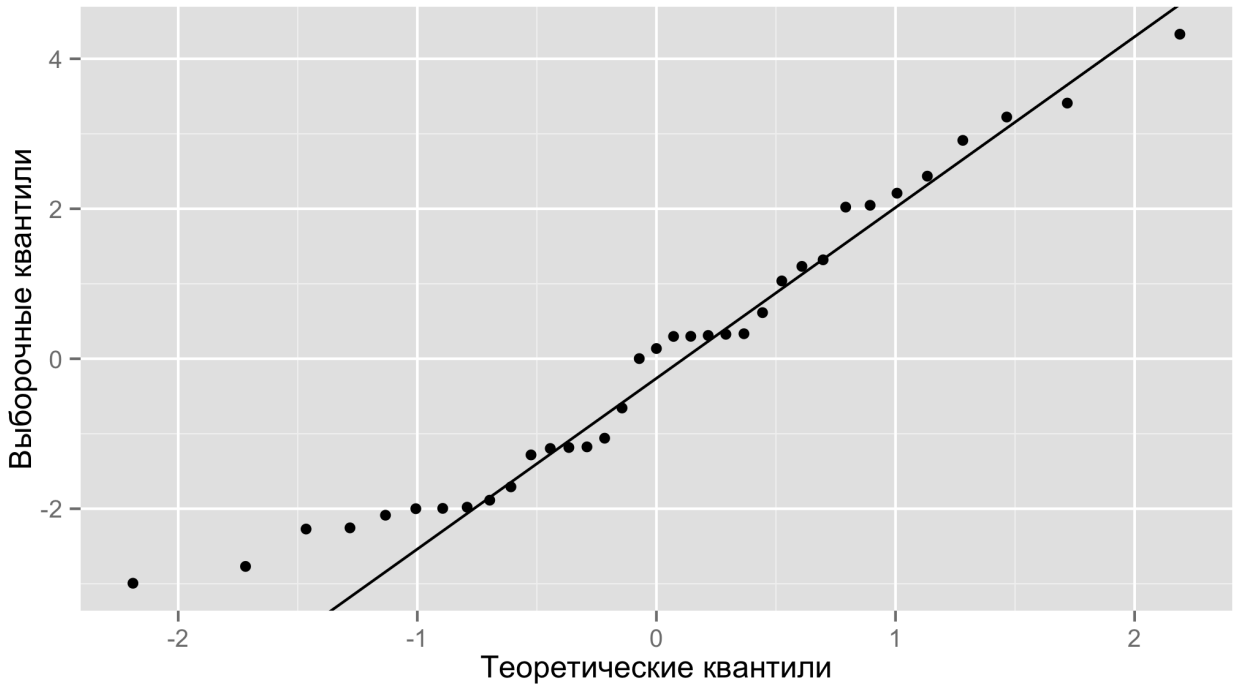

Рисунок 3.1.6 — График квантилей для остатков

гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста нельзя.

Проверим критерий χ^2 Пирсона. Из полученных в **R** результатов, статистика χ^2 Пирсона $P = 10.51$. Вероятность ошибки $p = 0.10 > 0.05$, а значит нулевая гипотеза не отвергается.Но при этом, это значение очень близко к 0.05. Проверим критерий: примем уровень значимости $\alpha = 0.05$, тогда из таблицы распределения χ^2 найдём критическое значение критерия $\chi^2_{\text{кр}}(\alpha, k) \approx 43.8$. Отсюда следует, что

$$\chi^2_{\text{набл}} < \chi^2_{\text{кр}},$$

где $\chi^2_{\text{набл}} = P = 10.51$. А значит, гипотезу о нормальности не отклоняем.

Построим график автокорреляционной функции для определения наличия взаимосвязей в ряде остатков (рисунок 3.1.7). На графике пунктирные линии разграничивают значимые и не значимые корреляции: значения, выходящие за линии, являются значимыми [14, с.376]. На представленном графике автокорреляционной функции можно заметить на лаге 15 значение, выходящее за интервал, обозначенный пунктирными линиями. Проверим значимость автокорреляций с помощью теста Льюнга-Бокса [14, с.377-378]. Данный тест проверяет наличие автокорреляций в исследуемом ряде. Используя возможности пакета **R** получили значения: статистика Льюнга–Бокса $X^2 = 0.075$ и вероятность ошибки $p = 0.78 > 0.05$ — это говорит о том, что тест не выявил значимых автокорреляций.

На рисунке 3.1.7 также можно заметить некоторое затухание всвязи 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 $DF = -3.27$, вероятность ошибки $p = 0.093 < 0.05$. Следовательно, необходимо принять альтернативную гипотезу о стационарности.

Полученная модель оказалась неоднозначной. С одной стороны, полученное значение коэффициента детерминации показало недостаточную точность полученной модели и не

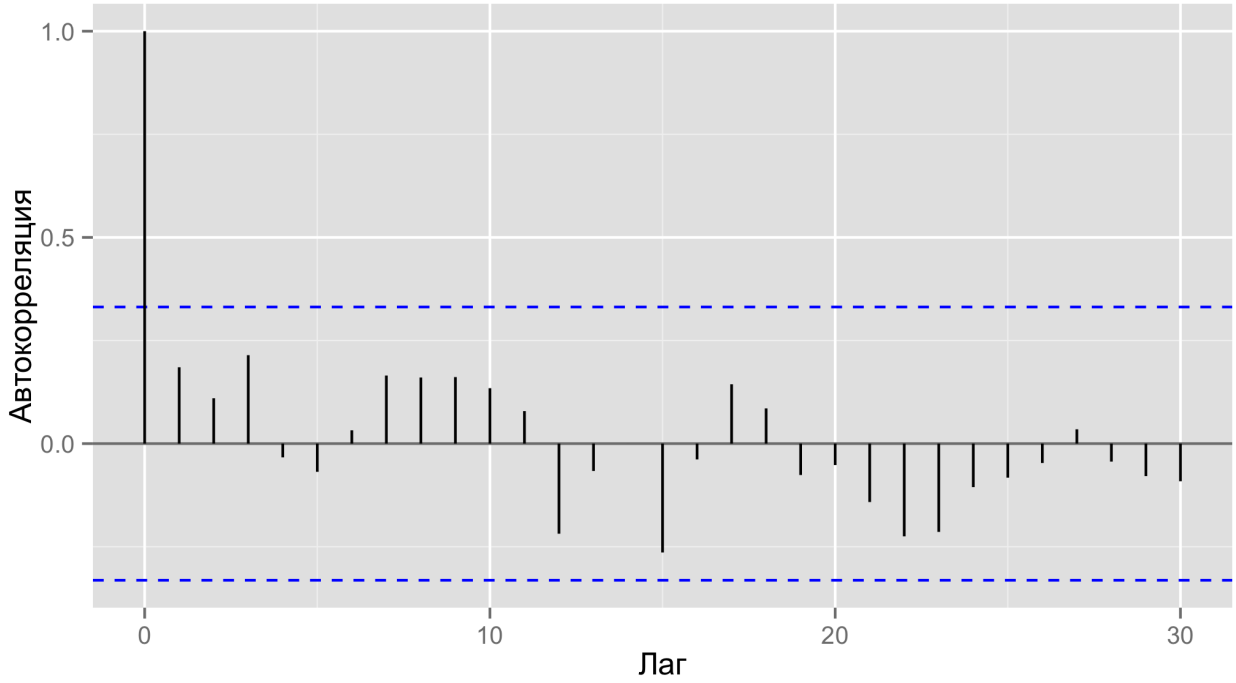

Рисунок 3.1.7 — График автокорреляционной функции

удалось достоверно показать нормальность ряда остатков. С другой стороны, была показана стационарность и отсутствие автокорреляции. Поэтому возникает необходимость строить модель другими методами.

3.2 Геостатистический подход

3.2.1 Вариограммный анализ. Кригинг.

В данной части работы для более объективного оценивания полученных прогнозов возьмем в качестве исследуемой выборки первые 32 значения исходных данных.

Традиционные детерминированные методы, 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 и определять значения на плотной сетке (в близких к измерениям точках). Следует также отметить, что данные измерений, как правило, дискретны и неоднородно распределены. 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 выводы были сделаны в результате проделанной в предыдущих частях данной работы. Отсюда следует, что необходимо использовать другие современные методы, позволяющие сделать более точные модели и выводы.

Для поставленной задачи в современных исследованиях хороших результатов позволяет добиться методы геостатистики, что подтверждается работами [26, 27]. Современная геостатистика — это широкий спектр статистических моделей и инструментов для анализа, обработки и представления пространственно-распределенной информации.

В рамках геостатистики, для получения наилучшей в статистическом смысле пространственной оценки используются модели из семейства кригинга (*kriging*) — наилучшего линейного несмещенного оценивателя (*Best Linear Unbiased Estimator — BLUE*). Кригинг

является “наилучшим” оценивателем в статистическом смысле — его оценка обладает минимальной дисперсией. Важным свойством кригинга является точное воспроизведение значений измерений в имеющихся точках (интерполяционные свойства). В отличие от многочисленных детерминированных методов оценка кригинга сопровождается оценкой ошибки интеполяции в каждой точке. Полученная ошибка позволяет охарактеризовать неопределенность интерполяционноий оценки данных при помощи доверительных интервалов.

В отличие от детерминированных методов, геостатистические оценки опираются на информацию о внутренней структуре данных, зависят от самих данных, т. е. являются адаптивными.

В различных геофизических явлениях выделяют свойство пространственной непрерывности: чем ближе две точки, тем ближе значение. Для оценки данного свойства построим диаграмму взаимного разброса пар точек (*h-scatterplot*), разделённых расстоянием h. Эта диаграмма позволяет увидеть пространственную непрерывность и проверить наличие корреляции в данных как качественно, так и количественно [28].

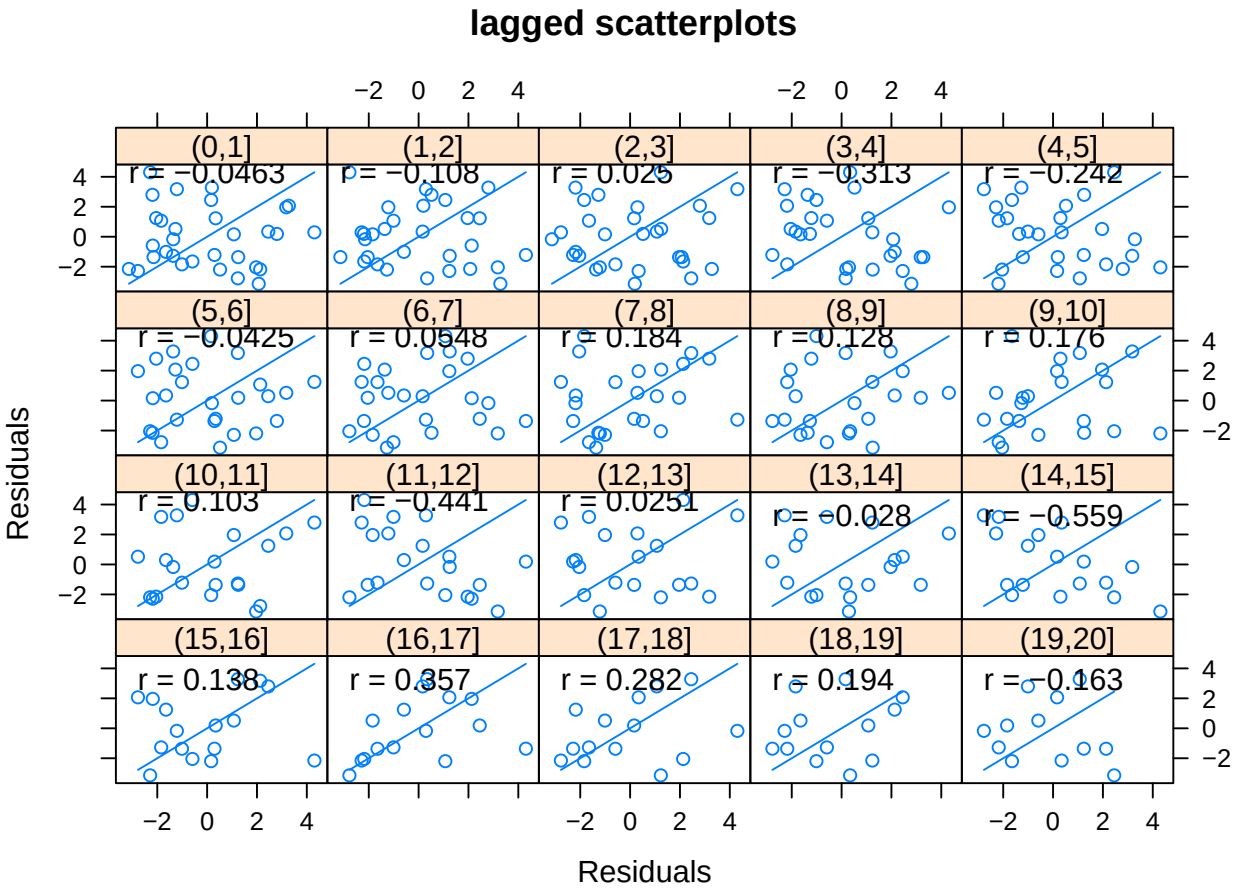

Рисунок 3.2.8 — Диаграмма взаимного разброса

Построенная диаграмма (рисунок 3.2.8) отображает поведение данных с увеличением лага. Следует отметить, что в классическом случае присутствия зависимости, поведение должно было быть следующим: на начальных графиках сильная концентрация, и с увеличением лага эта концетрация уменьшается. В нашем случае такого не наблюдается. Напротив, на всех лагах присутствует слабая зависимость. Что, вообще говоря, 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.

Центральная идея геостатистики состоит в использовании знаний о пространственной корреляции экспериментальных данных для построения пространственных оценок и интерполяций. Вариограмма — ключевой инструмент для оценки степени пространственной корреляции, имеющейся в данных, и для ее моделирования. Модель вариограммы является функцией, определяющей зависимость изменения исследуемой величины в пространстве 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 $h = x_i - x_j$, называемой *лагом*. Для близких точек разность значениеq функции в них обычно меньше и растет с увеличением расстояния между точками. Вычислив среднее значение квадратов разностей для каждого значения лага h (для каждого собранного класса пар измерений), можно получить дискретную функцию, называемую *экспериментальной вариограммой*. Вариограмма обычно характеризуется тремя значениями: эффект самородков(*nugget*), ранг(*range*) и порог(sill). Эффект самородков характеризуется разрывом вариограммы около нуля. Порог характеризует предельное значение вариограммы, на некотором расстоянии, называемом рангом, за которым последующие значения вариограммы становятся некоррелированными.

Также при построении вариограммы следует учитывать параметр максимального расстояния, для которого вычисляется вариограмма. Первоначальным параметром было выбрано следующее значение: $2n/3 = 21$ [29].

Построим экспериментальную вариограмму с помощью пакета *gstat* и функции *variogram*. С помощью этой функции можно построить экспериментальную вариограмму, основанную на классической оценки вариограммы и робастной оценки Кресси [29]. Построим экспериментальную вариограмму с помощью классической оценки.

Построенная вариограмма отображена на рисунке В.4 в приложении В. На представленном рисунке можно заметить, что на промежутке $[0; 1]$ не происходит роста значений вариограммы. Наоборот, наблюдается разрыв: первое значение находится значительно выше 0. При этом вариограмма не сильно выходит за пределы дисперсии переменной, которая равна 4.07. Более того, первые значения уже достигло порога. Что говорит о том, что вариограмма на первых значениях выходит на предельное значение, и последующие значения некоррелированы. Это, на самом деле, согласуется с нашими исходными данными, так как при анализе остатков было выявлено отстутсвие автокорреляций, и спецификой самих данных: наблюдение за каждый год, вообще говоря, не зависит от предшедствующего.

На основе этого делаем вывод о наличии эффекта самородков и делаем первоначальное предположение о равенстве порога 3.9.

На основе экспериментальной вариограммы построим модель вариограммы для дальнейшего использования на этапе кригинга. Моделью вариограммы может служить не каждая функция, а только та, для которой выполнено условие положительной определенности. Положительная определенность модели вариограммы гарантирует, что уравнения кригинга, построенные с использованием данной модели, имеют единственное устойчивое решение. Поэтому при моделировании используются только те функции, для которых положительная определенность установлена, а также их взвешенные линейные комбинации с неотрицательными весами, которые тоже будут являться положительно определенными. Модель вариограммы строится как линейная комбинация подходящих базисных моделей [28].

Для построения моделей вариограммы существует два подхода: вручную, т.е. визуально с ручным подбором параметров, и автоматическим подбором параметров с помощью специальных методов. И на практике построение модели вариограммы представляет собой итеративный процесс, на каждом шаге которого следует наилучшим образом подобрать параметры очередного модельного приближения. В различной литературе рекомендуется

строить моделей вручную, так как исследователь лучше знает специфику данных, чем различные методы оценивания. Попробуем построить модель вариограммы визуально.

Ранее было отмечено присутствие эффекта самородков. Другой, часто встречающейся моделью, является сферическая:

$$\gamma(|h|) = cSph_a(|h|) = \begin{cases} c(1.5|h|/a - 0.5(|h|/a)^3) & ,|h| \leq a, \\ c & ,|h| > a. \end{cases} \tag{3.4}$$

Возьмём эту модель в качестве базовой с помощью функции *vgm*, в качестве начального параметра возьмём порог, указанный ранее: 3.9. Далее воспользуемся функцией *fit.variogram* для подбора более точных значений указанной модели. Таким образом окончательная модель:

	Модель	Порог	Ранг
1	Nug	3.71	0.00
2	Sph	0.65	1.26

Таблица 4 — Модель вариограммы

И график полученной модели на рисунке 3.2.9 (пунктиром). На графике можно проследить все указанные ранее особенности: эффект самородков и порог.

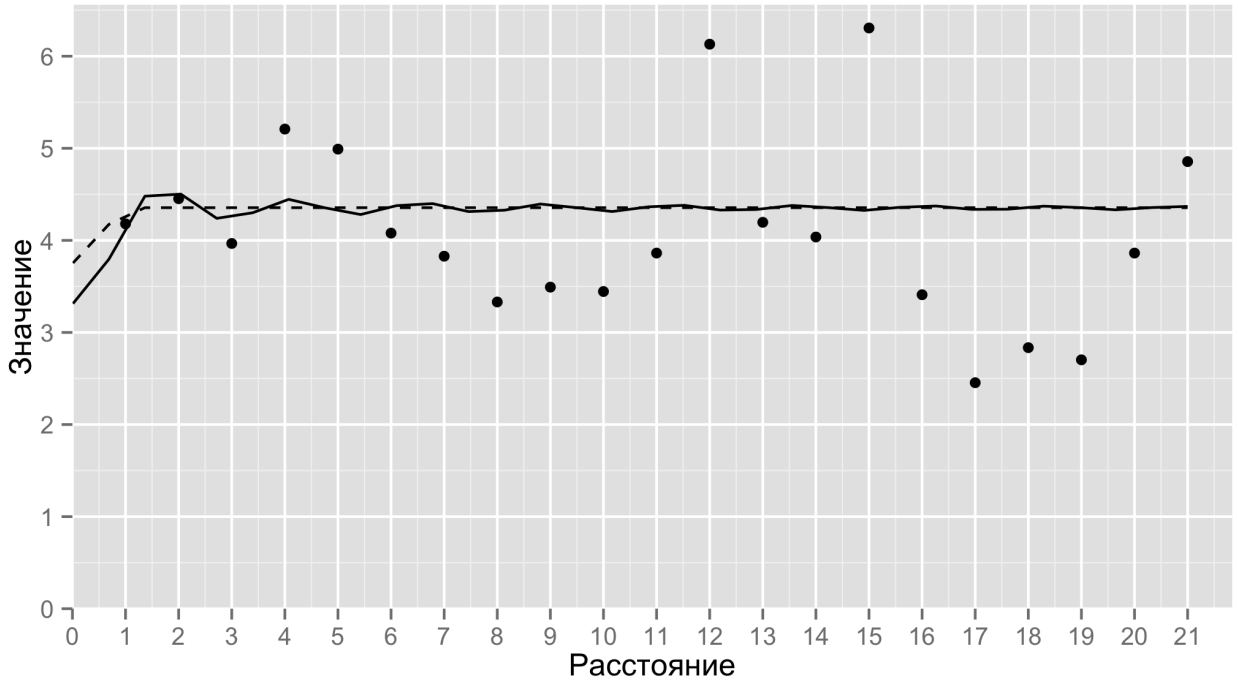

Рисунок 3.2.9 — Классические модели вариограммы

Задача геостатистики — оценить значения изучаемой пространственной переменной в произвольных точках области исследования на основе анализа ее значений, измеренных в ограниченном числе выборочных точек. По построенной модели вычислим оценки при помощи ординарного кригинга, реализованного функцией *krige*. Вычисленные значения отображены в таблице С.2 в приложении С. Оценку отклонения от истинных значений выразим с помощью среднеквадратической ошибки *(MSE)*. В данном случае $MSE = 2.62$.

Полученные значения оказались идентичными значению тренда, следовательно прогноз почти не изменился. Это говорит о том, что построенная модель не смогла уловить поведение исходных данных. По этой причине был использован второй вариант построения модели — с автоматическим подбором параметров.

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 3.2.9 сплошной линией и в приложении В на рисунке В.3 показан результат выполнения представленной ранее функции. Таким образом, наилучшей моделью вариограммы, построенной по классической оценке, стала линейная комбинация двух: эффект самородков с параметром 3.31 и модель с эффектом дыр (*Hole*) с параметрами: порог — 1.04, ранг — 0.379 изображенная в приложении В на рисунке В.5.

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

	Год	Наблюдаемое	Прогнозное	Тренд
1	2007	19.400	21.714	21.578
2	2008	21.800	21.578	21.687
3	2009	21.900	21.881	21.797
4	2010	24.300	21.876	21.906
5	2011	22.800	22.013	22.016
6	2012	20.200	22.171	22.126

Таблица 5 — Прогноз (классическая оценка)

некоторое поведение. Но в данном случае $MSE = 2.82$, что хуже предыдущего значения, а значит, прогноз ухудшился. Попробуем улучшить результат с помощью робастной оценки Кресси.

Модель вариограммы, представленная на рисунке В.6 в приложении В, является также линейной комбинацией двух базисных моделей: эффекта самородков с параметром 4.57 и волновая модель с параметрами: 0.559, 1.17. 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 $MSE = 2.35$, таким образом это значение близко к значе-

	Год	Наблюдаемое	Прогнозное	Тренд
1	2007	19.400	21.458	21.578
2	2008	21.800	21.748	21.687
3	2009	21.900	21.940	21.797
4	2010	24.300	22.027	21.906
5	2011	22.800	22.055	22.016
6	2012	20.200	22.089	22.126

Таблица 6 — Прогноз (робастная оценка)

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

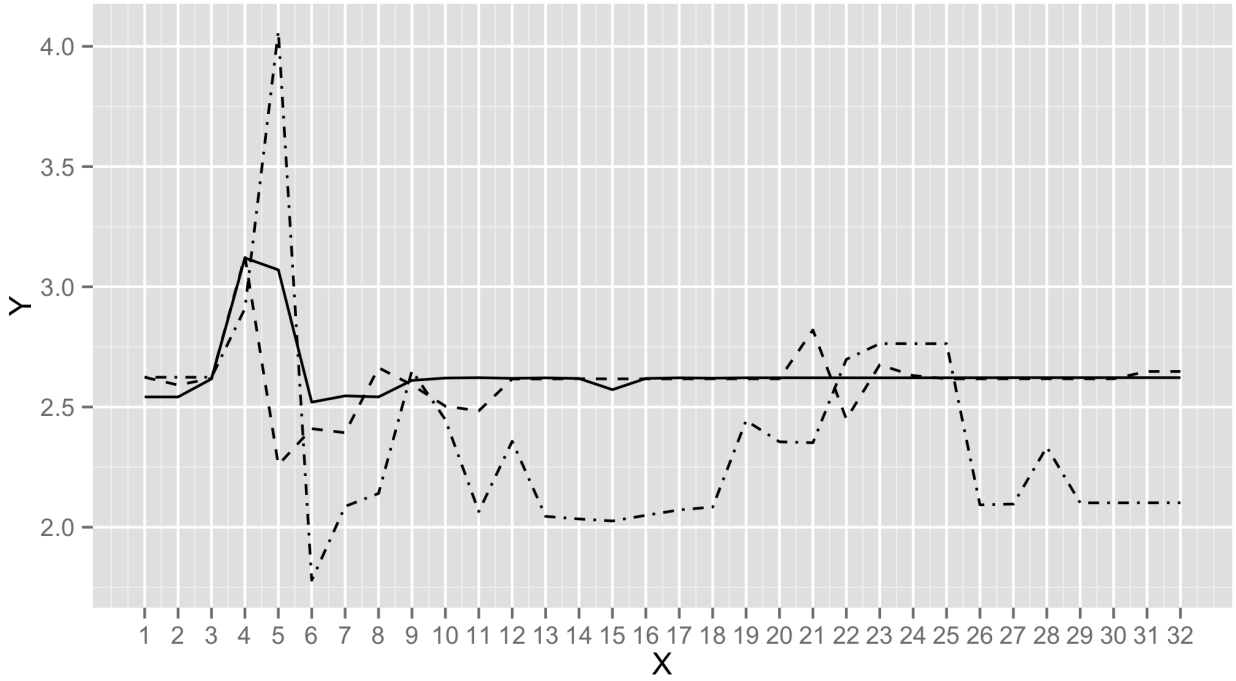

Рисунок 3.2.10 — Зависимость ошибки от максимального расстояния

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 3.2.10. 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	Наблюдаемое	Прогнозное	Тренд
1	2007	19.400	21.154	21.578
2	2008	21.800	21.626	21.687
3	2009	21.900	22.046	21.797
4	2010	24.300	22.302	21.906
5	2011	22.800	22.365	22.016
6	2012	20.200	22.290	22.126

Таблица 7 — Наилучший прогноз (робастная оценка)

оказалась равной $MSE = 1.78$. Что действительно является лучшим из полученных показателем.

Сравнительный анализ полученного прогноза представлен на графике В.7 в приложении В.

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 номральному 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.

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 В Графические материалы

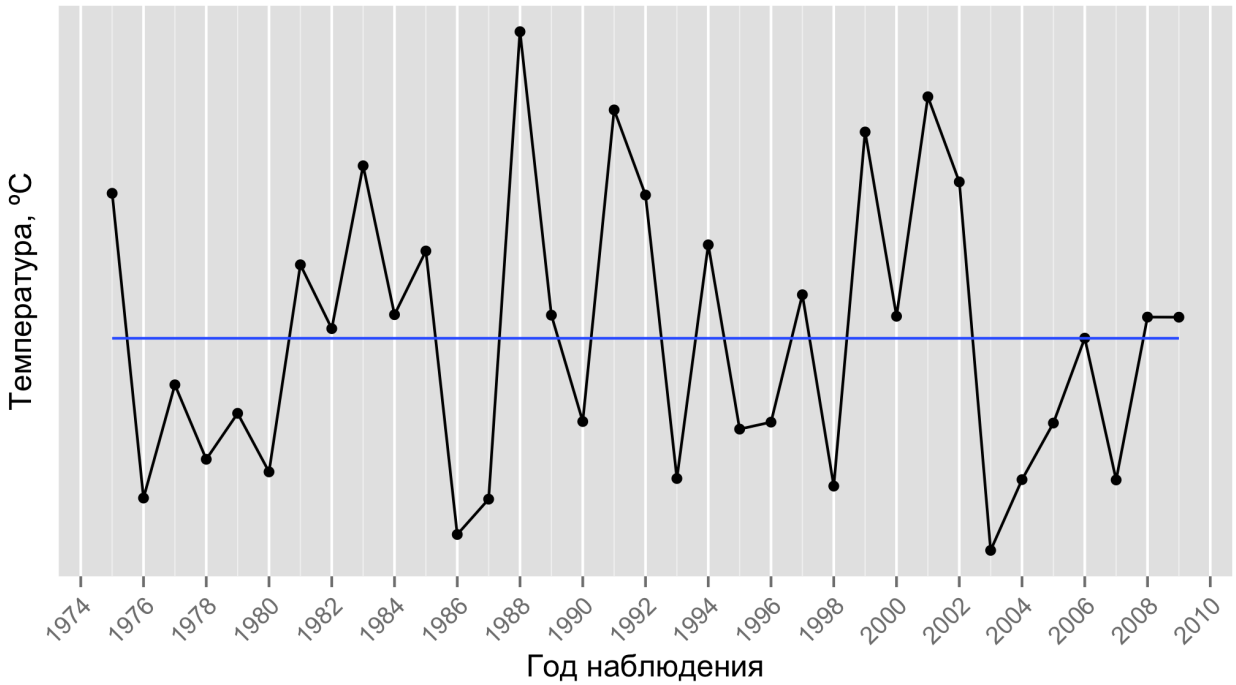

Рисунок В.1 — График ряда остатков

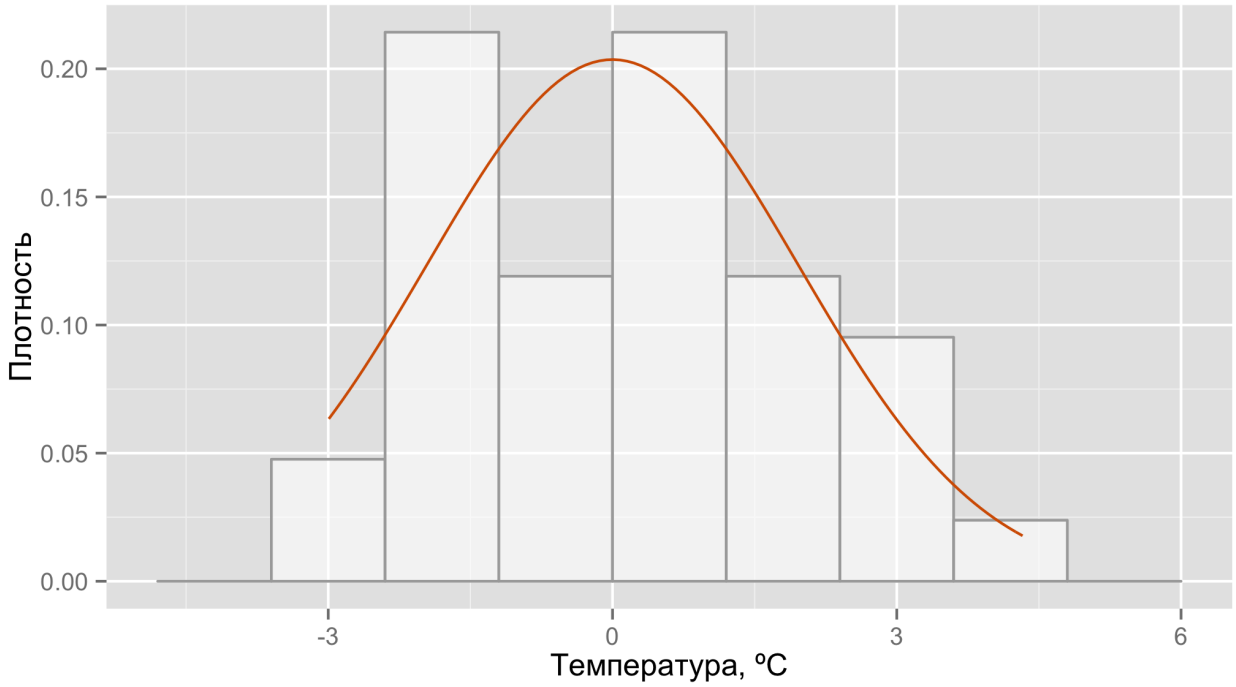

Рисунок В.2 — Гистограмма остатков с кривой плотности нормального распределения $\mathcal{N}(19.88, 4.92)$

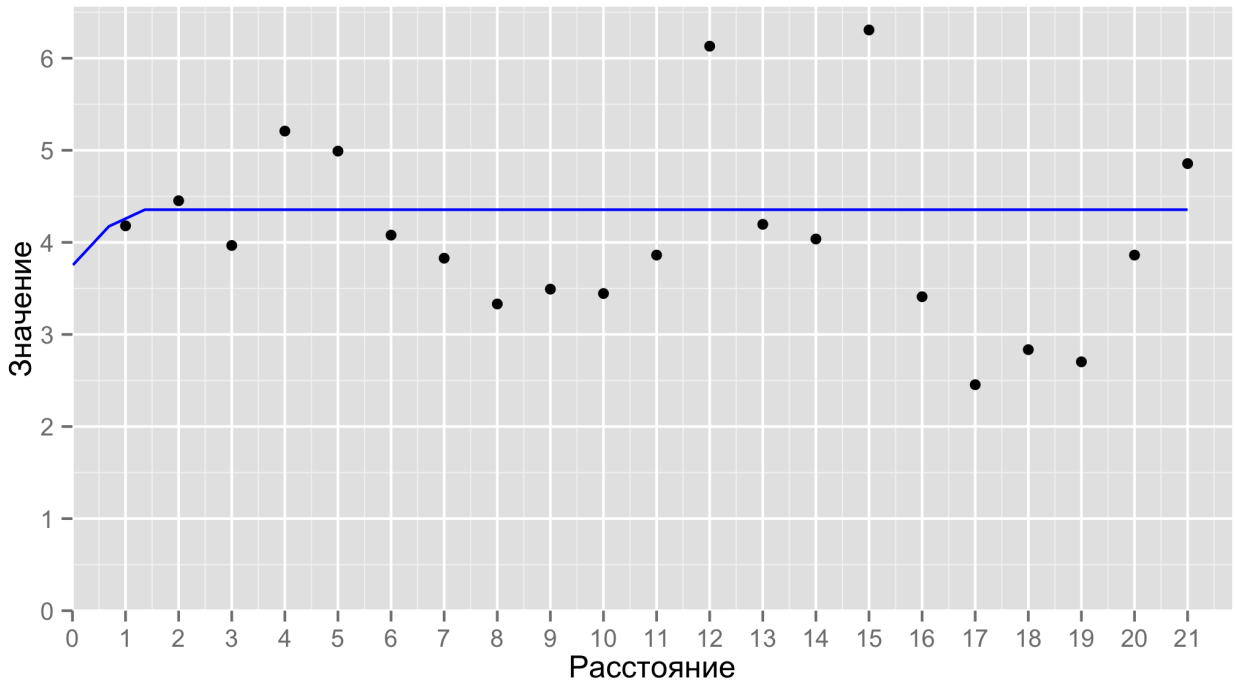

Рисунок В.3 — Экспериментальная и теоретическая вариограмма (сферическая модель)

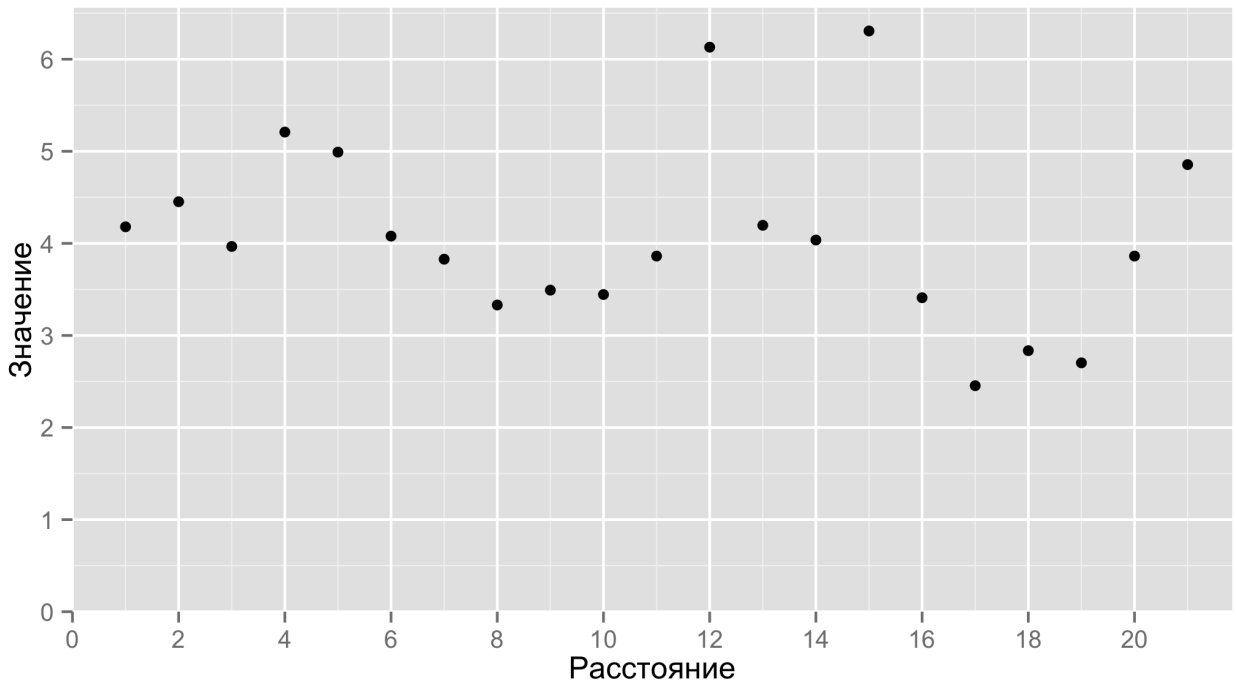

Рисунок В.4 — Экспериментальная вариограмма (классическая оценка)

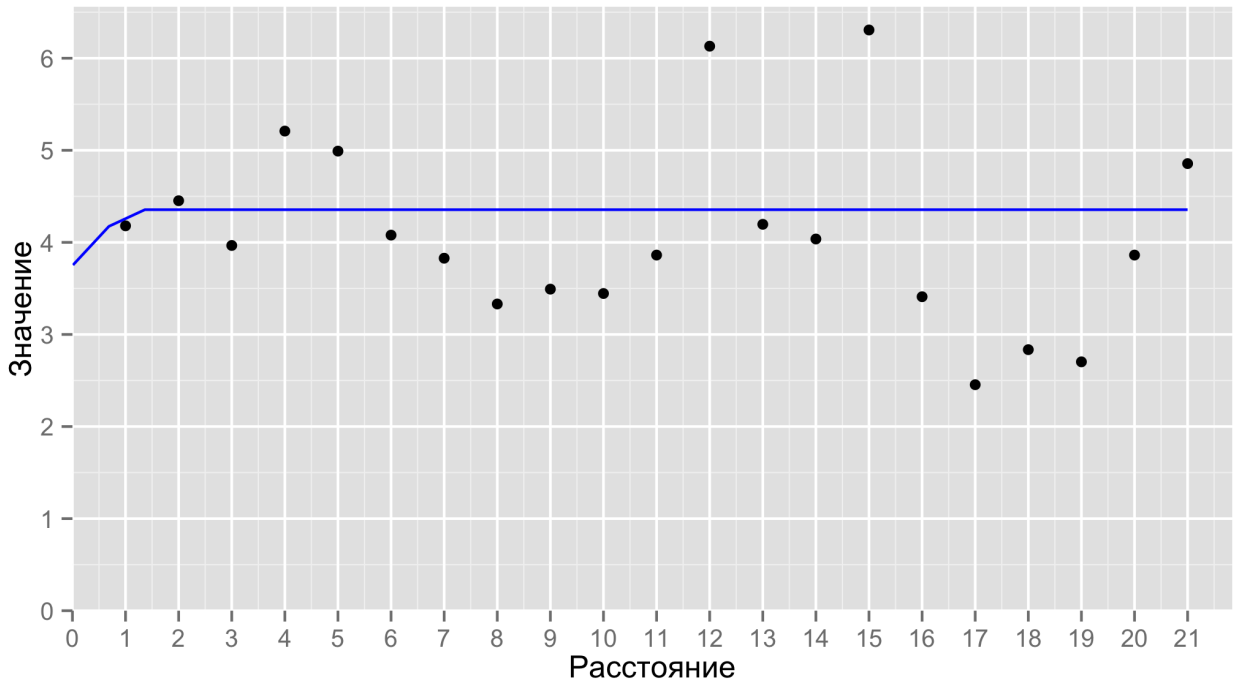

Рисунок В.5 — Экспериментальная и теоретическая вариограмма (классическая оценка)

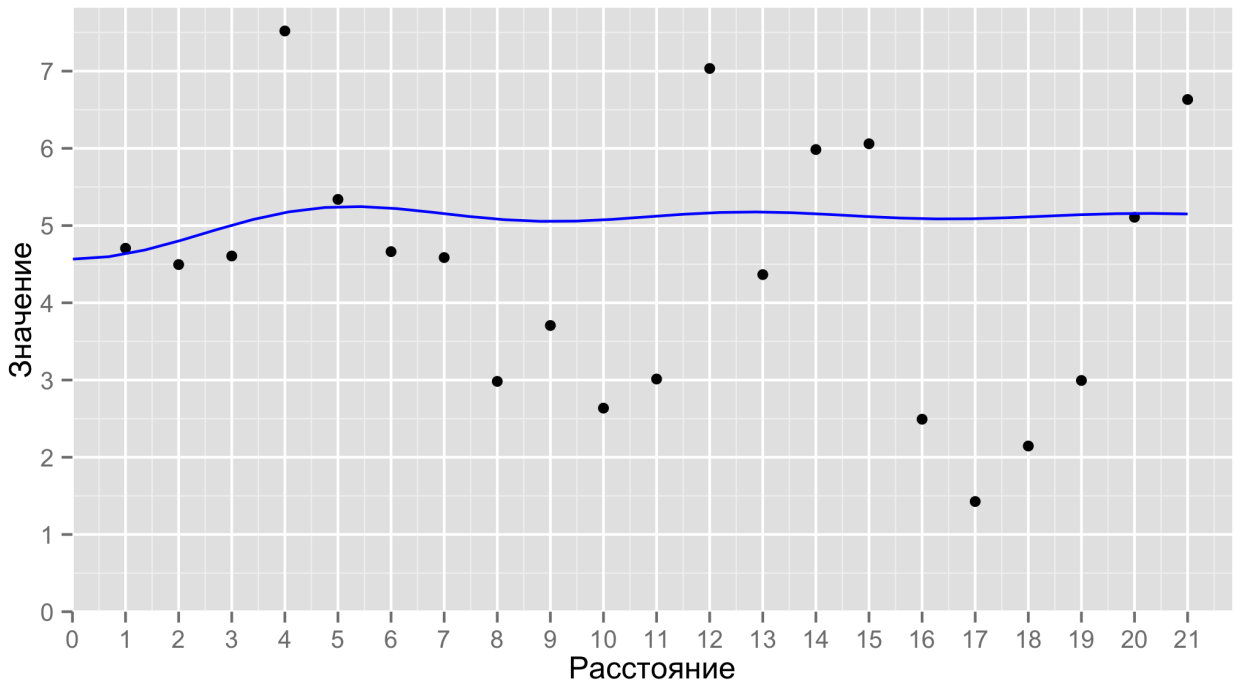

Рисунок В.6 — Экспериментальная и теоретическая вариограмма (робастная оценка)

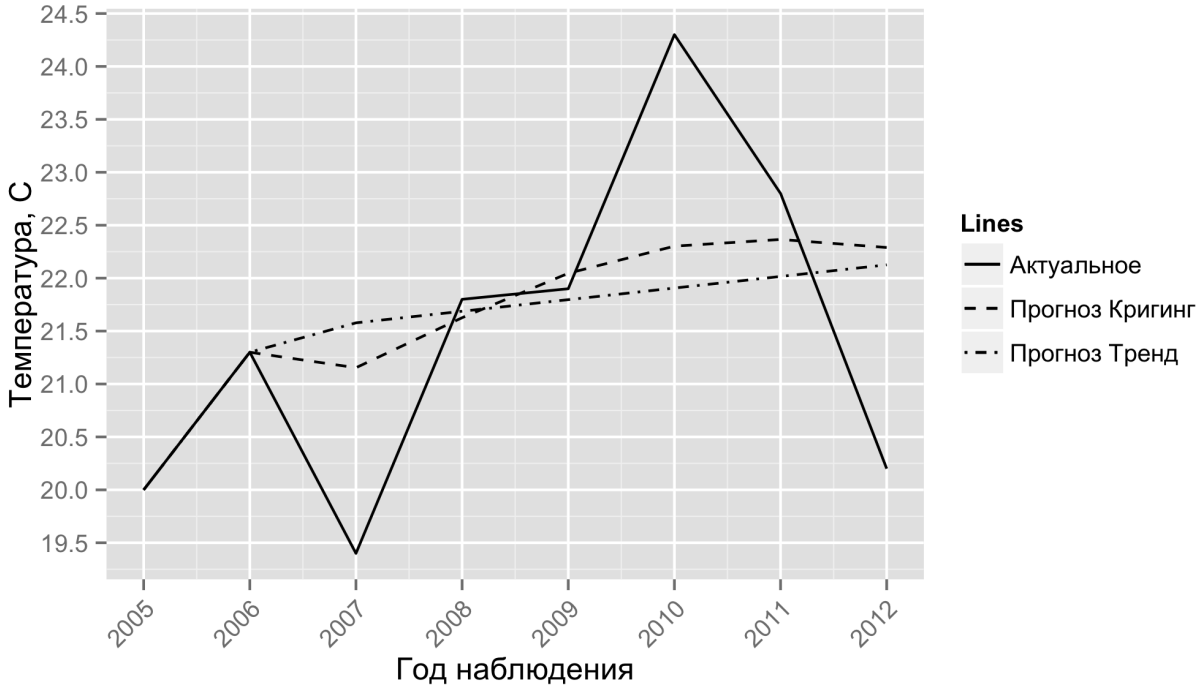

Рисунок В.7 — Сравнение прогнозных значений

Приложение C Результаты вычислений

	year	temperature
1	1975.00	2.05
2	1976.00	-2.25
3	1977.00	-0.66
4	1978.00	-1.71
5	1979.00	-1.06
6	1980.00	-1.89
7	1981.00	1.04
8	1982.00	0.14
9	1983.00	2.44
10	1984.00	0.33
11	1985.00	1.23
12	1986.00	-2.77
13	1987.00	-2.27
14	1988.00	4.33
15	1989.00	0.33
16	1990.00	-1.17
17	1991.00	3.22
18	1992.00	2.02
19	1993.00	-1.98
20	1994.00	1.32
21	1995.00	-1.28
22	1996.00	-1.18
23	1997.00	0.62
24	1998.00	-2.09
25	1999.00	2.91
26	2000.00	0.31
27	2001.00	3.41
28	2002.00	2.21
29	2003.00	-2.99
30	2004.00	-1.99
31	2005.00	-1.20
32	2006.00	0.00
33	2007.00	-2.00
34	2008.00	0.30
35	2009.00	0.30

Таблица C.1 — Временной ряд остатков.

	Год	Наблюдаемое	Прогнозное	Тренд
1	2007	19.400	21.577	21.578
2	2008	21.800	21.688	21.687
3	2009	21.900	21.798	21.797
4	2010	24.300	21.907	21.906
5	2011	22.800	22.017	22.016
6	2012	20.200	22.126	22.126

Таблица C.2 — Прогноз (сферическая модель)

Приложение D Код программ

```
# Descriptive statistics

# Function for getting all descriptive statistics
dstats.describe <- function(data, type="", locale=FALSE) {
  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), dstats.coef.var(data), 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", "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",
                   "Coefficient of Variance", "Standard Error", "Skewness",
                   "Std. Error Skewness", "Kurtosis", "Std. Error Kurtosis")
    descr.col <- c("Value")
  }
  df <- data.frame(stats, row.names=descr.row)
  colnames(df) <- descr.col

  df
}

dstats.mean <- function(data, ...) {
  mean(data, ...)
}

```

```
39 dstats.median <- function(data, ...) {
40   median(data, ...)
41 }
42
43 dstats.quartile.lower <- function(data, ...) {
44   quantile(data, ...)[[2]]
45 }
46
47 dstats.quartile.upper <- function(data, ...) {
48   quantile(data, ...)[[4]]
49 }
50
51 dstats.quartile.range <- function(data) {
52   dstats.quartile.upper(data) - dstats.quartile.lower(data)
53 }
54
55 dstats.min <- function(data, ...) {
56   min(data, ...)
57 }
58
59 dstats.max <- function(data, ...) {
60   max(data, ...)
61 }
62
63 dstats.range <- function(data) {
64   max(data) - min(data)
65 }
66
67 dstats.variance <- function(data, ...) {
68   var(data, ...)
69 }
70
71 dstats.std.dev <- function(data) {
72   sd(data)
73 }
74
75 dstats.coef.var <- function(data) {
76   mn <- mean(data)
77   if (abs(mn) > 1.987171e-15) {
78     (var(data) / mean(data)) * 100
79   } else
80     0
81 }
82
83 dstats.std.error <- function(data) {
84   sd(data) / sqrt(length(data))
85 }
86
87 dstats.skew <- function(data) {
88   n <- length(data)
89   mean <- mean(data)
90   (n * sum(sapply(data, FUN=function(x){(x - mean)^3}))) /
91     ((n - 1) * (n - 2) * dstats.std.dev(data)^3)
92 }
93
94 dstats.std.error.skew <- function(data) {
95   n <- length(data)
96   sqrt((6 * n * (n - 1)) / ((n - 2) * (n + 1) * (n + 3)))
97 }
98
```

```
99  dstats.test.skew <- function(data) {
100   dstats.skew(data) / dstats.std.error.skew(data)
101 }
102
103 dstats.kurtosis <- function(data) {
104   n <- length(data)
105   mean <- mean(data)
106   (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)) /
107     ((n - 1) * (n - 2) * (n - 3) * dstats.variance(data)^2)
108 }
109
110 dstats.std.error.kurtosis <- function(data) {
111   n <- length(data)
112   2 * dstats.std.error.skew(data) * sqrt((n^2 - 1) / ((n - 3) * (n + 5)))
113 }
114
115 dstats.test.kurtosis <- function(data) {
116   dstats.kurtosis(data) / dstats.std.error.kurtosis(data)
117 }
118
119 dstats.write <- function (data, type) {
120   WriteDescriptiveStatistic(expression=dstats.mean(data), type=type, name="mean"
        )
121   WriteDescriptiveStatistic(expression=dstats.variance(data), type=type, name="
        variance")
122   WriteDescriptiveStatistic(expression=paste(format(dstats.coef.var(data),
        nsmall=2, digits=4), "\\%"), type=type, name="coef-var")
123   WriteDescriptiveStatistic(expression=dstats.skew(data), type=type, name="skew"
        )
124   WriteDescriptiveStatistic(expression=dstats.kurtosis(data), type=type, name="
        kurtosis")
125   WriteDescriptiveStatistic(expression=dstats.test.skew(data), type=type, name="
        test-skew")
126   WriteDescriptiveStatistic(expression=dstats.test.kurtosis(data), type=type,
        name="test-kurtosis")
127 }
```

Листинг D.1: Описательные статистики

```
1  ## Cleaning up the workspace
2  rm(list=ls(all=TRUE))
3
4  ## Dependencies
5  library(ggplot2)  # eye-candy graphs
6  library(xtable)   # convert data to latex tables
7  library(outliers) # tests for outliers
8  library(tseries)  # adf test used
9  library(nortest)  # tests for normality
10 library(sp)       # spatial data
11 library(gstat)    # geostatistics
12 library(reshape2) # will see
13
14 ## Import local modules
15 source("R/lib/plot.R")       # useful functions for more comfortable plotting
16 source("R/lib/dstats.R")     # descriptive statistics module
17 source("R/lib/misc.R")       # some useful global-use functions
18 source("R/lib/draw.R")       # helpers for drawing
19 source("R/lib/write.R")      # helpers for writing
20 source("R/lib/ntest.R")      # tests for normality
21
```

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

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

```
    в", label="table:residuals_dstats"),
101         file="out/residual/dstats.tex")
102
103 ## Basic histogram for residuals / seems like the same as for non-residuals
104 plot.residuals.hist <- DrawHistogram(data=research.residuals, filename="residual
    /histogram.png")
105
106 ## Tests for normality
107 research.data.shapiro <- ntest.ShapiroWilk(data=research.residuals$temperature,
    type="residual", name="shapiro")
108 research.data.pearson <- ntest.PearsonChi2(data=research.residuals$temperature,
    type="residual", name="pearson")
109 research.data.ks      <- ntest.KolmogorovSmirnov(data=research.residuals$
    temperature, type="residual", name="ks")
110
111 ## Normal Quantile-Quantile plot for residuals
112 plot.residuals.qq <- DrawQuantileQuantile(data=research.residuals$temperature,
    filename="residual/quantile.png")
113
114 ## Auto Correlation Function plot
115 plot.residuals.acf <- DrawAutoCorrelationFunction(data=research.data$temperature
    , filename="residual/acf.png")
116
117 ## Box-Ljung and adf tests (some kind of stationarity and independence tests) //
     TODO: need to know exactly in theory what it is
118 research.residuals.box <- Box.test(research.residuals$temperature, type="Ljung-
    Box")
119 WriteTest(research.residuals.box$statistic, research.residuals.box$p.value,
    research.residuals.box$parameter[[1]], type="residual", name="ljung-box")
120
121 research.residuals.adf <- adf.test(research.residuals$temperature)
122 WriteTest(research.residuals.adf$statistic, research.residuals.adf$p.value, type
    ="residual", name="stationarity")
123
124 source("R/predictor.R")
```

Листинг D.2: Основной код программы

```
1  source("R/lib/afv.R")
2  source("R/lib/variogram.R")
3  source("R/lib/kriging.R")
4
5  ## Function definition: need to be moved into isolated place
6  ### Just definition of mean standard error // TODO: find out exact formula and
       describe each parameter
7  MSE <- function (e, N=1) {
8    sum(sapply(X=e, FUN=function(x) x**2)) / length(e)
9  }
10
11 # Completes trend values up to source observation number
12 computeTrend <- function (fit, future=0) {
13   c(sapply(c(1 : (src.nrows + future)), FUN=function(x) fit$coefficients[[1]] +
         x * fit$coefficients[[2]]))
14 }
15
16 kObservationNum <- 32
17
18 ## Form the data for research again
19 research.data <- src.data[0:kObservationNum, ]
20
21 research.data.fit <- lm(research.data$temperature ~ ConvertYearsToNum(research.
```

```
    data$year))
research.data.residuals <- research.data.fit$residuals
research.data.trend <- computeTrend(research.data.fit)

cutoff <- trunc(2 * kObservationNum / 3) # let it be "classical" value
#cutoff <- 2

# Draw H-Scatterplot
research.data.hscat <- DrawHScatterplot(research.data.residuals[1:
    kObservationNum], cutoff)

# Compute variogram manually with choosed model (best what i could found)
variogram.manual <- ComputeManualVariogram(research.data.residuals, cutoff=
    cutoff, file_modeled="figures/variogram/manual-model.png")

# Compute variogram with auto fit model using classical estimation
variogram.classical <- ComputeVariogram(data=research.data.residuals, x=
    ConvertYearsToNum(research.data$year), cressie=FALSE, cutoff=cutoff, width=
    FALSE,
                                        file_empirical="figures/variogram/
                                            classical-empirical.png",
                                        file_modeled="figures/variogram/
                                            classical-modeled.png")

WriteCharacteristic(variogram.classical$var_model[[2]][1], type="variogram",
    name="classical-nug")
WriteCharacteristic(variogram.classical$var_model[[2]][2], type="variogram",
    name="classical-psill")
WriteCharacteristic(variogram.classical$var_model[[3]][2], type="variogram",
    name="classical-range")

# Compute variogram with auto fit model using robust (cressie) estimation
variogram.robust <- ComputeVariogram(data=research.data.residuals, x=
    ConvertYearsToNum(research.data$year), cressie=TRUE, cutoff=cutoff, width=
    FALSE,
                                     file_empirical="figures/variogram/robust-
                                         empirical.png",
                                     file_modeled="figures/variogram/robust-
                                         modeled.png")

WriteCharacteristic(variogram.robust$var_model[[2]][1], type="variogram", name="
    robust-nug")
WriteCharacteristic(variogram.robust$var_model[[2]][2], type="variogram", name="
    robust-psill")
WriteCharacteristic(variogram.robust$var_model[[3]][2], type="variogram", name="
    robust-range")

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

kriging.manual    <- PredictWithKriging(research.data.residuals, x=
    ConvertYearsToNum(research.data$year), variogram_model=variogram.manual$var_
    model)
kriging.classical <- PredictWithKriging(research.data.residuals, x=
    ConvertYearsToNum(research.data$year), variogram_model=variogram.classical$
    var_model)
kriging.robust    <- PredictWithKriging(research.data.residuals, x=
    ConvertYearsToNum(research.data$year), variogram_model=variogram.robust$var_
    model)

```

```
58 prediction.manual <- data.frame("Год"=src.data$year[(kObservationNum + 1):src.
       nrows],
59    "Наблюдаемое"=src.data$temperature[(kObservationNum + 1):src.nrows],
60    "Прогнозное"=kriging.manual$var1.pred+research.data.trend[(kObservationNum +
          1):src.nrows],
61    "Тренд"=research.data.trend[(kObservationNum + 1):src.nrows])
62 print(xtable(prediction.manual, caption="Прогноз (сферическая модель)", label="
       table:prediction-manual", digits=c(0, 0, 3, 3, 3)),
63    file="out/variogram/prediction-manual.tex")
64
65 prediction.classical <- data.frame("Год"=src.data$year[(kObservationNum + 1):src
       .nrows],
66    "Наблюдаемое"=src.data$temperature[(kObservationNum + 1):src.nrows],
67    "Прогнозное"=kriging.classical$var1.pred+research.data.trend[(kObservationNum
          + 1):src.nrows],
68    "Тренд"=research.data.trend[(kObservationNum + 1):src.nrows])
69 print(xtable(prediction.classical, caption="Прогноз (классическая оценка)",
       label="table:prediction-classical", digits=c(0, 0, 3, 3, 3)),
70    file="out/variogram/prediction-classical.tex")
71
72 prediction.robust <- data.frame("Год"=src.data$year[(kObservationNum + 1):src.
       nrows],
73    "Наблюдаемое"=src.data$temperature[(kObservationNum + 1):src.nrows],
74    "Прогнозное"=kriging.robust$var1.pred+research.data.trend[(kObservationNum +
          1):src.nrows],
75    "Тренд"=research.data.trend[(kObservationNum + 1):src.nrows])
76 print(xtable(prediction.robust, caption="Прогноз (робастная оценка)", label="
       table:prediction-robust", digits=c(0, 0, 3, 3, 3)),
77    file="out/variogram/prediction-robust.tex")
78
79 res.manual <- CrossPrediction(src.data$temperature, src.data$year, research.data
       .trend, kriging.manual,    "figures/variogram/cross-prediction-manual.png")
80 res.classical <- CrossPrediction(src.data$temperature, src.data$year, research.
       data.trend, kriging.classical, "figures/variogram/cross-prediction-classical.
       png")
81 res.robust <- CrossPrediction(src.data$temperature, src.data$year, research.data
       .trend, kriging.robust,    "figures/variogram/cross-prediction-robust.png")
82
83 mse.manual    <- MSE(res.manual)
84 mse.classical <- MSE(res.classical)
85 mse.robust    <- MSE(res.robust)
86
87 WriteCharacteristic(mse.manual, type="variogram", name="manual-mse")
88 WriteCharacteristic(mse.classical, type="variogram", name="classical-mse")
89 WriteCharacteristic(mse.robust, type="variogram", name="robust-mse")
90
91 # Find best cutoff parameter
92 ComparePredictionParameters(research.data.residuals, research.data.trend,
       ConvertYearsToNum(research.data$year), filename="figures/variogram/parameter-
       comparison.png")
93
94
95 # Best prediction as we investigated is for robust kriging with cutoff=6. Let's
       make it!
96 variogram.robust.best <- ComputeVariogram(data=research.data.residuals, x=
       ConvertYearsToNum(research.data$year), cressie=TRUE, cutoff=6, width=FALSE,
97                                            file_empirical="figures/variogram/
                                                robust-best-empirical.png",
98                                            file_modeled="figures/variogram/robust
                                                -best-modeled.png")
```

```
99
100 kriging.robust.best <- PredictWithKriging(research.data.residuals, x=
        ConvertYearsToNum(research.data$year), variogram_model=variogram.robust.best$
        var_model)
101 res.robust.best <- CrossPrediction(src.data$temperature, src.data$year, research
        .data.trend, kriging.robust.best, "figures/variogram/cross-prediction-robust-
        best.png")
102 mse.robust.best <- MSE(res.robust.best)
103
104 prediction.robust.best <- data.frame("Год"=src.data$year[(kObservationNum + 1):
        src.nrows],
105   "Наблюдаемое"=src.data$temperature[(kObservationNum + 1):src.nrows],
106   "Прогнозное"=kriging.robust.best$var1.pred+research.data.trend[(
          kObservationNum + 1):src.nrows],
107   "Тренд"=research.data.trend[(kObservationNum + 1):src.nrows])
108 print(xtable(prediction.robust.best, caption="Наилучший прогноз (робастная оценк
        a)", label="table:prediction-robust-best", digits=c(0, 0, 3, 3, 3)),
109   file="out/variogram/prediction-robust-best.tex")
110
111 WriteCharacteristic(mse.robust.best, type="variogram", name="robust-best-mse")
112 ## TODO: form krige matrix for analysis
```

Листинг D.3: Вариограммный анализ