

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра теории вероятностей и математической статистики

Павлов Александр Сергеевич

Анализ и прогнозирование гидрологических данных

Дипломная работа

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

Допущена к защите
«___» 2015 г.

Минск, 2015

АННОТАЦИЯ

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

АННАТАЦЫЯ

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

ANNOTATION

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

Реферат

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

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

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

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

В процессе работы проведён сравнительный анализ современных пакетов статистического анализа. При помощи пакета R вычислены и проанализированы описательные статистики, произведена подборка закона распределения, проведены 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 процессе работы алгоритмы исследования могут быть использованы для анализа других аналогичных данных.

Abstract

Bachelor's thesis, 35 pages, 3 chapters, 11 figures, 7 tables, 29 sources, 4 appendices.

TIME SERIES, R, DESCRIPTIVE 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 programming language were computed and analysed descriptive statistics, was performed distribution 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. Algorithms that were obtained during the research could be used for analysis other similar data.

Содержание

Введение	5
1 Случайный процесс и его характеристики	7
1.1 Случайный процесс. Стационарность	7
1.2 Вариограмма и внутренне стационарный случайный процесс	8
2 Оценка вариограммы внутренне стационарного гауссовского случайного процесса и ее свойства	9
2.1 Первые два момента оценки вариограммы	9
2.2 Асимптотическое поведение оценки вариограммы	11
3 Обзор реализованного программного обеспечения	17
3.1 Модуль первичного анализа	17
3.2 Модуль анализа остатков	19
3.3 Модуль вариограммного анализа	20
4 Анализ временного ряда в среде R	23
4.1 Детерминированный подход	23
4.1.1 Описательные статистики и первичный анализ данных	23
4.1.2 Корреляционный анализ	27
4.1.3 Регрессионный анализ	29
4.1.4 Анализ остатков	32
4.2 Геостатистический подход	34
4.2.1 Вариограммный анализ. Кригинг	34
Заключение	43
Список использованной литературы	45
Приложение А Исходные данные	46
Приложение Б Графические материалы	47
Приложение В Результаты вычислений	55
Приложение Г Исходный код	56

Введение

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

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

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

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

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

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

$$\begin{aligned} cov(t_1, t_2) &= cov\{X(t_1), X(t_2)\} = E\{(X(t_1) - m(t_1))(X(t_2) - m(t_2))\} = \\ &= \iint_{\mathbb{R}^2} (x_1 - m(t_1))(x_2 - m(t_2)) dF_2(x_1, x_2; t_1, t_2) \end{aligned}$$

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

При этом

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

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

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

Глава 2

Оценка вариограммы внутренне стационарного гауссовского случайного процесса и ее свойства

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

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

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

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

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

2.1 Первые два момента оценки вариограммы

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

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

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

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

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

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

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

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

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

Из равенства (1.4) получаем, что

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

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

Найдём второй момент оценки вариограммы при различных значениях h :

$$\begin{aligned}
\text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= E\{(2\tilde{\gamma}(h_1) - E\{2\tilde{\gamma}(h_1)\})(2\tilde{\gamma}(h_2) - E\{2\tilde{\gamma}(h_2)\})\} = \\
&= E\left\{\frac{1}{n-h_1} \sum_{t=1}^{n-h_1} ((X(t+h_1) - X(t))^2 - E\{(X(t+h_1) - X(t))^2\}) \times \right. \\
&\quad \times \left. \frac{1}{n-h_2} \sum_{s=1}^{n-h_2} ((X(s+h_2) - X(s))^2 - E\{(X(s+h_2) - X(s))^2\})\right\} = \\
&= \frac{1}{(n-h_1)(n-h_2)} \sum_{t=1}^{n-h_1} \sum_{s=1}^{n-h_2} \text{cov}\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} \tag{2.5}
\end{aligned}$$

Из свойства 1.1 корреляции получаем, что

$$\begin{aligned}
\text{cov}\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} &= \text{corr}\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} \times \\
&\quad \times \sqrt{V\{(X(t+h_1) - X(t))^2\} V\{(X(s+h_2) - X(s))^2\}}
\end{aligned}$$

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

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

Далее воспользуемся леммой 1 из [9]:

$$\begin{aligned}
\text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \\
&= \frac{2(2\gamma(h_1))(2\gamma(h_2))}{(n-h_1)(n-h_2)} \sum_{t=1}^{n-h_1} \sum_{s=1}^{n-h_2} (\text{corr}\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\})^2 = \\
&= \frac{2(2\gamma(h_1))(2\gamma(h_2))}{(n-h_1)(n-h_2)} \sum_{t=1}^{n-h_1} \sum_{s=1}^{n-h_2} \left(\frac{\text{cov}\{X(t+h_1) - X(t), X(s+h_2) - X(s)\}}{\sqrt{V\{X(t+h_1) - X(t)\} V\{X(s+h_2) - X(s)\}}} \right)^2
\end{aligned}$$

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

$$\begin{aligned}
\text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \\
&= \frac{2}{(n-h_1)(n-h_2)} \times \tag{2.6}
\end{aligned}$$

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

что и требовалось показать.

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

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

□

2.2 Асимптотическое поведение оценки вариограммы

Проанализируем асимптотическое поведение моментов второго порядка оценки (2.1).

Теорема 2.2. *Если имеет место соотношение*

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

то

$$\begin{aligned} & \lim_{n \rightarrow \infty} (n - \min\{h_1, h_2\}) \operatorname{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \\ & = 2 \sum_{m=-\infty}^{+\infty} \gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2, \end{aligned} \quad (2.9)$$

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

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

Доказательство. В (2.6) сделаем следующую замену переменных

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

Получим

$$\begin{aligned} & \operatorname{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \\ & = \frac{2}{(n-h_1)(n-h_2)} \sum_{t=1}^{n-h_1} \sum_{t-m=1}^{n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \end{aligned} \quad (2.11)$$

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

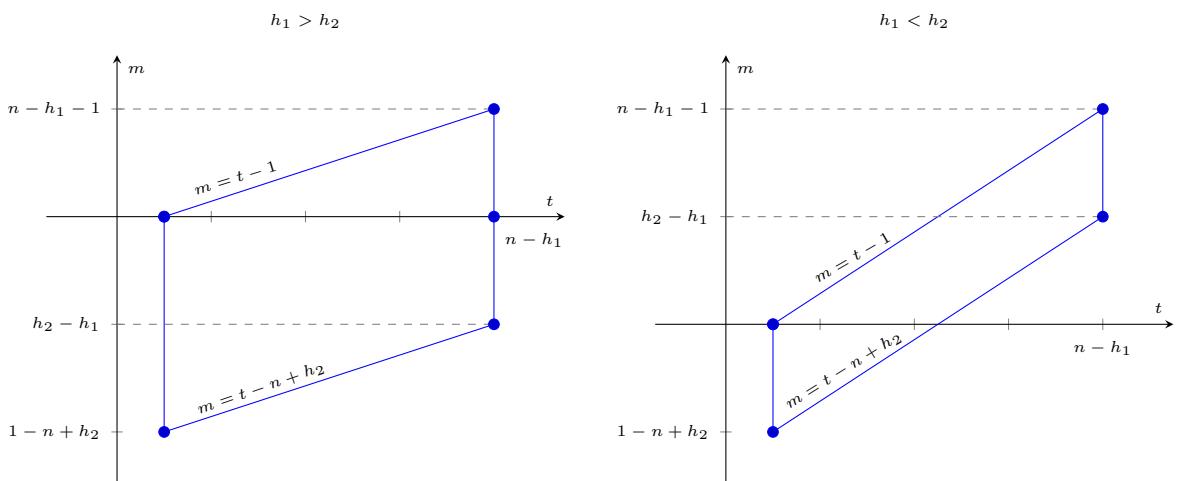


Рисунок 2.1 — Области суммирования после замены переменных

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

$$\begin{aligned} cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{(n-h_1)(n-h_2)} \times \\ &\times \left(\sum_{m=1-n+h_2}^{h_2-h_1-1} \sum_{t=1}^{m+n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\ &+ \sum_{m=h_2-h_1}^0 \sum_{t=1}^{n-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\ &\left. + \sum_{m=1}^{n-h_1-1} \sum_{t=m+1}^{n-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right) \end{aligned}$$

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

$$\begin{aligned} cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{(n-h_1)(n-h_2)} \times \\ &\times \left(\sum_{m=1-n+h_2}^{h_2-h_1-1} (m+n-h_2)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\ &+ (n-h_1) \sum_{m=h_2-h_1}^0 (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\ &\left. + \sum_{m=1}^{n-h_1-1} (n-h_1-m)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right) \end{aligned}$$

Вынесем $n - h_1$ из каждого слагаемого:

$$\begin{aligned} cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_2} \times \\ &\times \left(\sum_{m=1-n+h_2}^{h_2-h_1-1} \left(1 + \frac{h_1+m-h_2}{n-h_1}\right)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\ &+ \sum_{m=h_2-h_1}^0 (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\ &\left. + \sum_{m=1}^{n-h_1-1} \left(1 - \frac{m}{n-h_1}\right)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right) \end{aligned}$$

Раскроем скобки под знаками сумм:

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_2} \left(\sum_{m=1-n+h_2}^{h_2-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\
&+ \frac{1}{n-h_1} \sum_{m=1-n+h_2}^{h_2-h_1-1} (h_1+m-h_2)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\
&+ \sum_{m=h_2-h_1}^0 (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\
&+ \sum_{m=1}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \\
&\left. - \frac{1}{n-h_1} \sum_{m=1}^{n-h_1-1} m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right)
\end{aligned}$$

Приведём подобные:

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_2} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\
&+ \frac{1}{n-h_1} \sum_{m=1-n+h_2}^{h_2-h_1-1} (h_1+m-h_2)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \\
&- \frac{1}{n-h_1} \sum_{m=1}^{n-h_1-1} m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \left. \right)
\end{aligned}$$

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

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_2} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \right. \\
&- \frac{1}{n-h_1} \sum_{m=1}^{n-h_1-1} m(\gamma(-m-h_1) + \gamma(-m+h_2) - \gamma(-m-h_1+h_2) - \gamma(-m))^2 - \\
&- \frac{1}{n-h_1} \sum_{m=1}^{n-h_1-1} m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \left. \right)
\end{aligned}$$

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

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_2} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \right. \\
&- \frac{2}{n-h_1} \sum_{m=1}^{n-h_1-1} m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \left. \right) \quad (2.12)
\end{aligned}$$

Аналогично для случая $h_1 < h_2$:

$$\begin{aligned} \text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{(n-h_1)(n-h_2)} \times \\ &\times \left(\sum_{m=1-n+h_2}^0 \sum_{t=1}^{m+n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\ &+ \sum_{m=1}^{h_2-h_1} \sum_{t=m+1}^{m+n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\ &\left. + \sum_{m=h_2-h_1+1}^{n-h_1-1} \sum_{t=m+1}^{n-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right) \end{aligned}$$

Выражение под знаком суммы не зависит от t :

$$\begin{aligned} \text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{(n-h_1)(n-h_2)} \times \\ &\times \left(\sum_{m=1-n+h_2}^0 (m+n-h_2)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\ &+ (n-h_2) \sum_{m=1}^{h_2-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\ &\left. + \sum_{m=h_2-h_1+1}^{n-h_1-1} (n-h_1-m)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right) \end{aligned}$$

Вынесем $n-h_2$ из каждого слагаемого:

$$\begin{aligned} \text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_1} \times \\ &\times \left(\sum_{m=1-n+h_2}^0 \left(1 + \frac{m}{n-h_2}\right)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\ &+ \sum_{m=1}^{h_2-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\ &\left. + \sum_{m=h_2-h_1+1}^{n-h_1-1} \left(1 + \frac{h_2-h_1-m}{n-h_2}\right)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right) \end{aligned}$$

Раскроем скобки под знаками сумм:

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_1} \left(\sum_{m=1-n+h_2}^0 (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\
&\quad + \frac{1}{n-h_2} \sum_{m=1-n+h_2}^0 m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\
&\quad + \sum_{m=1}^{h_2-h_1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\
&\quad + \sum_{m=h_2-h_1+1}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\
&\quad \left. + \frac{1}{n-h_2} \sum_{m=h_2-h_1+1}^{n-h_1-1} (h_2-h_1-m)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right)
\end{aligned}$$

Приведём подобные:

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_1} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \right. \\
&\quad + \frac{1}{n-h_2} \sum_{m=1-n+h_2}^0 m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 + \\
&\quad \left. + \frac{1}{n-h_2} \sum_{m=h_2-h_1+1}^{n-h_1-1} (h_2-h_1-m)(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right)
\end{aligned}$$

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

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_1} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \right. \\
&\quad - \frac{1}{n-h_2} \sum_{m=0}^{n-h_2-1} m(\gamma(-m-h_2) + \gamma(-m+h_1) - \gamma(-m) - \gamma(-m+h_1-h_2))^2 - \\
&\quad \left. - \frac{1}{n-h_2} \sum_{m=1}^{n-h_2-1} m(\gamma(m-h_1) + \gamma(m+h_2) - \gamma(m-h_1+h_2) - \gamma(m))^2 \right)
\end{aligned}$$

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

$$\begin{aligned}
cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{n-h_1} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \right. \\
&\quad - \frac{2}{n-h_2} \sum_{m=1}^{n-h_2-1} m(\gamma(m+h_2) + \gamma(m-h_1) - \gamma(m) - \gamma(m-h_1+h_2))^2 \left. \right) \quad (2.13)
\end{aligned}$$

Далее, для доказательства (2.9) оценим разность, используя условие (2.8), выражение

(2.12) и лемму Кронекера [10]:

$$\begin{aligned}
& |(n-h_2)cov\{\gamma(\tilde{h}_1), \gamma(\tilde{h}_2)\} - \sum_{m=-\infty}^{+\infty} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2| \leq \\
& \leq \sum_{m=-\infty}^{n+h_2} |\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2)|^2 + \\
& + \sum_{m=n-h_1}^{+\infty} |\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2)|^2 + \\
& + \frac{1}{n-h_1} \sum_{m=0}^{n-h_1-1} |\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2)|^2 \rightarrow 0, \quad (2.14)
\end{aligned}$$

при $n \rightarrow \infty$.

Рассуждая аналогично, в силу сходимости ряда (2.8), выражения (2.13) и леммы Кронекера [10], получаем оценку разности для случая $h_1 < h_2$

$$\begin{aligned}
& |(n-h_1)cov\{\gamma(\tilde{h}_1), \gamma(\tilde{h}_2)\} - \sum_{m=-\infty}^{+\infty} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2| \leq \\
& \leq \sum_{m=-\infty}^{n+h_2} |\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2)|^2 + \\
& + \sum_{m=n-h_1}^{+\infty} |\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2)|^2 + \\
& + \frac{1}{n-h_2} \sum_{m=-n+h_2+1}^0 |m| |\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2)|^2 \rightarrow 0, \quad (2.15)
\end{aligned}$$

при $n \rightarrow \infty$.

Тогда, объединяя вместе полученные в (2.14) и (2.15) результаты, получаем требуемое предельное соотношение (2.9):

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

Нетрудно видеть, что если в (2.9) положить $h_1 = h_2 = h$, то получаем равенство для дисперсии оценки вариограммы (2.10). Действительно,

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

□

Следствие 2.1. Из теоремы 2 следует соотношение

$$\lim_{n \rightarrow \infty} V\{2\tilde{\gamma}(h)\} = 0, \quad h = \overline{0, n-1}$$

Замечание 2.1. В силу первого утверждения теоремы 1 и вышеприведённого следствия получаем, что оценка вариограммы $\tilde{\gamma}(h)$ является состоятельной в среднеквадратическом смысле для вариограммы $\gamma(h)$, $h \in \mathbb{Z}$

Глава 3

Обзор реализованного программного обеспечения

Для решения поставленной задачи, в рамках данной работы, было реализовано клиент-серверное приложение, позволяющее решать класс аналогичных по структуре задач. Для этого написаны несколько модулей, включающих в себя функционал, необходимый для решения конкретной подзадачи. Для удобства работы, каждый модуль имеет отдельные страницы, отвечающие за конкретные инструменты. Таким образом весь процесс работы в приложении разбивается на несколько этапов, на каждом из которых решается конкретная подзадача. В данной работе можно выделить три этапа: первичный анализ данных, анализ остатков и вариограммный анализ. Дальше в данной главе будет рассмотрены подробнее каждый из аспектов реализации.

Следует отметить, что каждая страница приложения имеет единый дизайн: экран можно условно поделить на панель выбора этапа анализа сверху и область исследования снизу. В свою очередь область исследования можно разделить также на две части: контрольная панель параметров и инструментов слева, и результаты вычислений и анализа справа.

3.1 Модуль первичного анализа

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 реализации данной программы на языке **R** в качестве опорной литературы использовались источники [11–13].

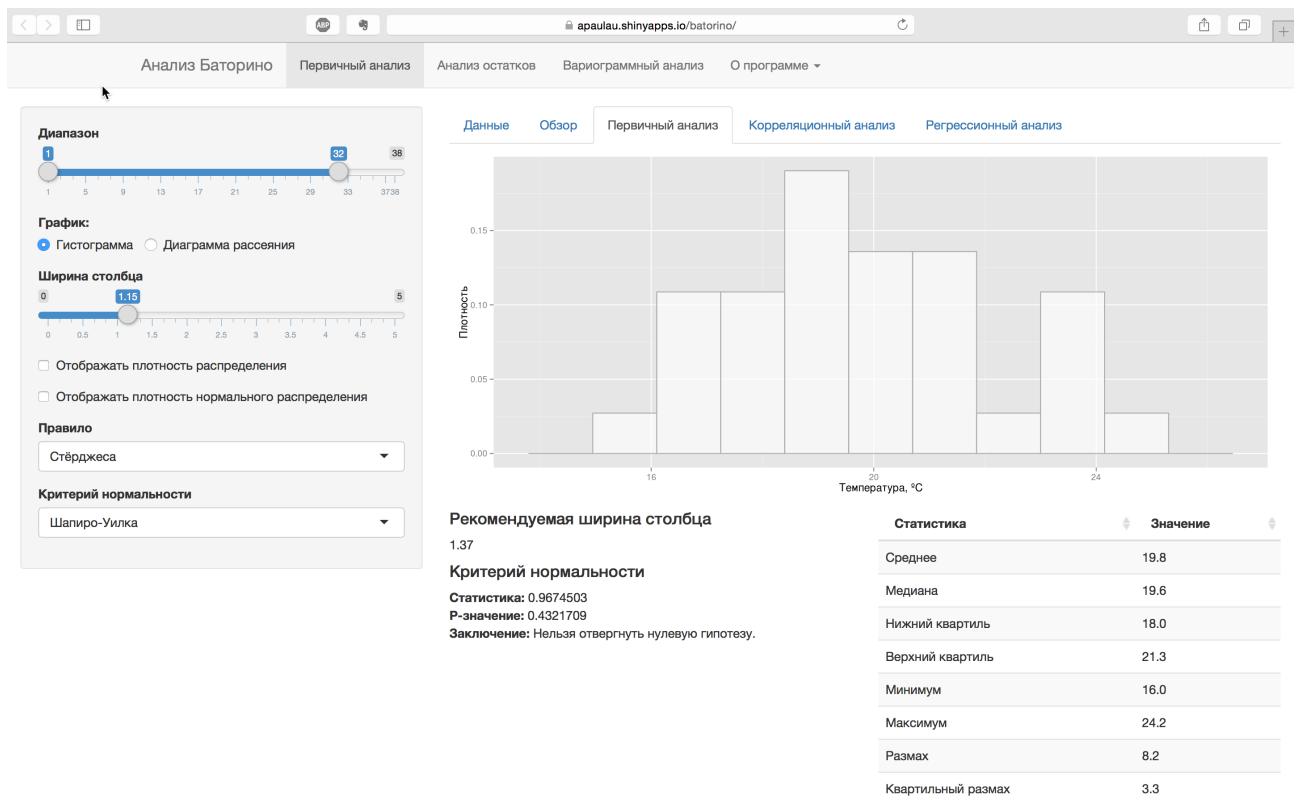


Рисунок 3.1 — Первичный анализ и описательные статистики

Данный модуль включает в себя возможности по просмотру и анализу непосредственно данных: графически и с помощью таблицы, позволяющей сортировать и производить поиск по определённому признаку. Непосредственно на рисунке 3.1 отображена вкладка первичного анализа. В которой представлены возможности по определению закона распределения исследуемых данных с помощью как проверки различными тестами, так и визуально, на гистограмме и графике квантилей. Контрольная панель позволяет изменять отображаемый в данный момент график, а также позволяет выбрать критерий нормальности. В случае выбора для отображения гистограммы, появляются управляющие элементы, позволяющие выбрать ширину столбца на гистограмме и правило по её вычислению (например, правило Стерджеса), отобразить плотность выборочного распределения и кривую нормального.

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 [14, 15] мной был написан модуль *dstats*, представленный в приложении Г листинге Г.1. Данный модуль позволяет вычислять все рассмотренные в данной работе описательные статистики, результат вычисления которых отображён на данной странице в виде таблицы.

Следующей вкладкой в данном модуле является корреляционный анализ. Данная страница позволяет оценить зависимость исследуемых данных с помощью диаграммы рассеяния, вычисляет коэффициент корреляции и с помощью критерия Стьюдента проверяет значимость вычисленного коэффициента, а также вычисляет для него доверительный интервал. Среди прочего, данная страница позволяет оценить наличие выбросов с помощью критерия Граббса.

Вкладка регрессионного анализа (рисунок 3.2) позволяет получить регрессионную модель по исследуемым данным. График временного ряда содержит также линию регрессии. Представленная страница демонстрирует возможности по анализу вычисленной модели:

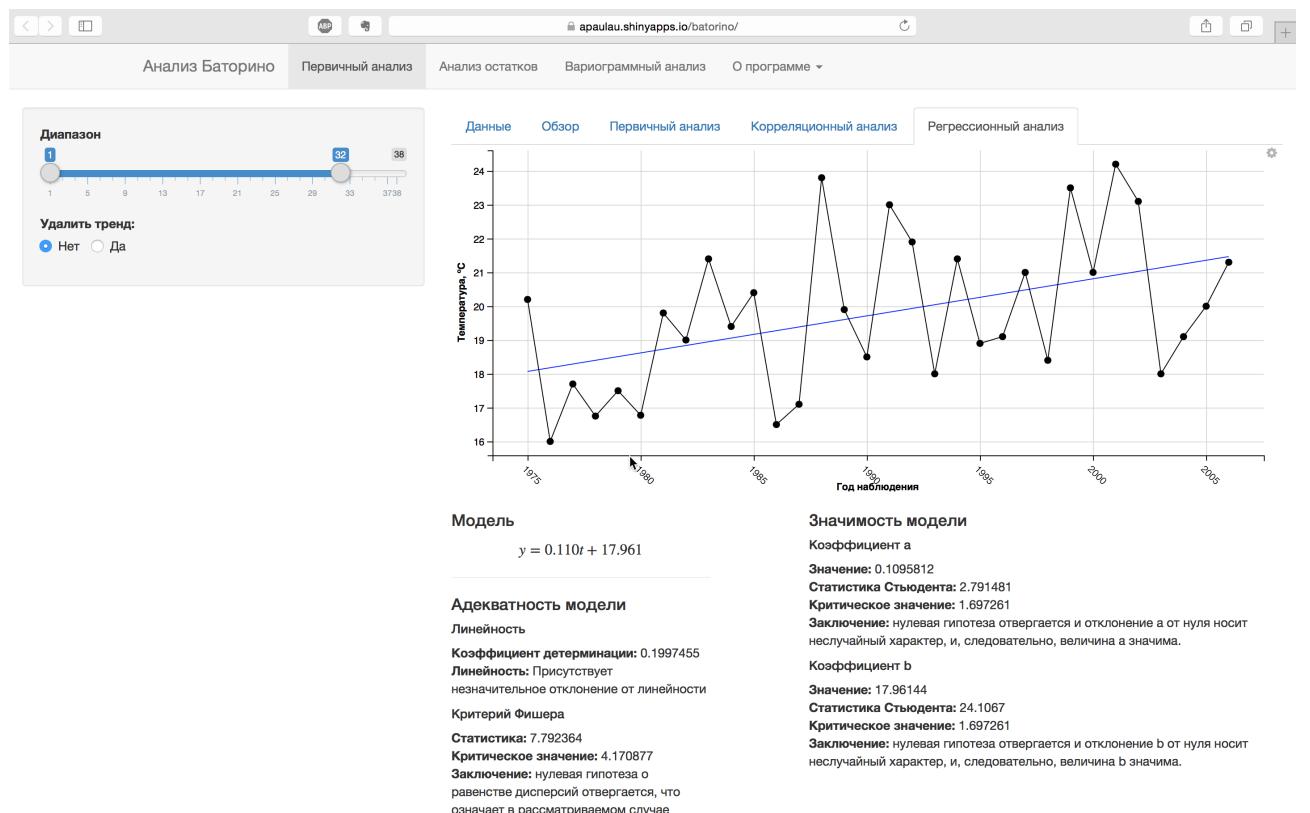


Рисунок 3.2 — Регрессионный анализ

определение значимости вычисленных коэффициентов, адекватность модели с помощью критерия Фишера и проверки линейности.

Инструменты, рассмотренные в рамках данного модуля, позволяют быстро получить информацию по исследуемым данным. А также сделать первые выводы и наметить шаги по дальнейшему исследованию. Заметим, что на каждом из этапов анализа и использования каждого из инструментов реализована возможность изменять объёмы выборки. Как снизу, так и сверху. Другими словами можно отбросить первые или последние наблюдения. Это позволяет быстро оценить, насколько влияют данные на результат в конкретном случае.

3.2 Модуль анализа остатков

Данный модуль является логическим продолжением рассмотренного ранее. После регрессионного анализа и удаления из исходного временного ряда тренда, основанного на регрессионном уравнении, получаем ряд остатков. Для его анализа реализованы возможности, которые включают в себя некоторые возможности предыдущего. Исключение составляют инструменты регрессионного и корреляционного анализов. Поскольку исследуемый на данном этапе временной ряд не имеет явных составляющих.

Таким образом данный модуль позволяет проверить остатки на нормальность как с помощью графиков квантилей и гистограммы, так и различными критериями: Шапиро-Уилка, χ^2 -Пирсона, Колмогорова-Смирнова. В дополнение к этому имеется возможность проанализировать описательные статистики, а также исследовать автокорреляционную функцию. Страница с таким инструментом представлена на рисунке 3.3. На рисунке про-

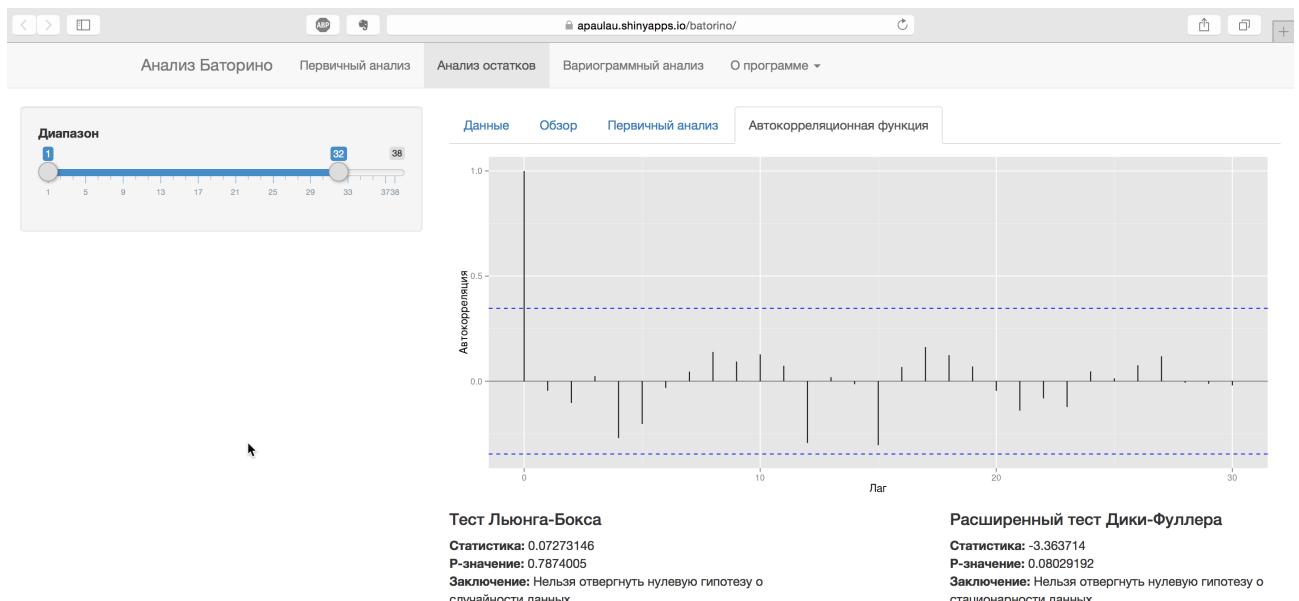


Рисунок 3.3 — Анализ автокорреляционной функции

демонстрирован график автокорреляционной функции, позволяющий визуально определить наличие автокорреляций в исследуемых данных. Также проверить наличие значимых автокорреляций позволяет проверка реализованного теста Льюнга-Бокса. В свою очередь,

расширенный тест Дики-Фуллера, также представленный на рассматриваемой странице, проверяет наличие стационарности в исследуемом случайному процессе.

В зависимости от результатов, полученных на рассмотренном этапе, можно либо закончить исследование, либо продолжить в модуле вариограммного анализа. Закончить исследование стоит в том случае, если ряд остатков полностью характеризуется как случайный шум, либо в случае, когда не выполняются условия для проведения следующего этапа.

3.3 Модуль вариограммного анализа

В данном модуле используются современные геостатистические методы и инструменты, которые, в рамках **R**, реализованы пакетом *gstat*. В этом пакете представлены функции для вычисления вариограмм, подбору моделей и параметров, интерполяции методами кригинга и методы валидации конечных результатов. Интерполяция методами кригинга подразумевает наличие подобранной модели вариограммы, поэтому в рассматриваемом модуле акцент сделан именно на подборе и анализе различных моделей вариограмм.

Начальный шаг состоит в подборе модели и её параметров к экспериментальной вариограмме. Для построения экспериментальной вариограммы присутствует возможность использовать две разновидности оценок вариограммы: рассмотренная в главе 2 оценка Матерона и робастная оценка Кресси-Хокинса. Для подбора модели вариограммы, в общем случае, существует два подхода: подбор визуально вручную, силами исследователя, и автоматическими методами. В данном модуле в полной мере реализованы оба подхода. В первом случае, изменение любого из параметров модели позволяет незамедлительно оценить эффект как на графике непосредственно вариограммы, так и по конечному прогнозу кригингом.

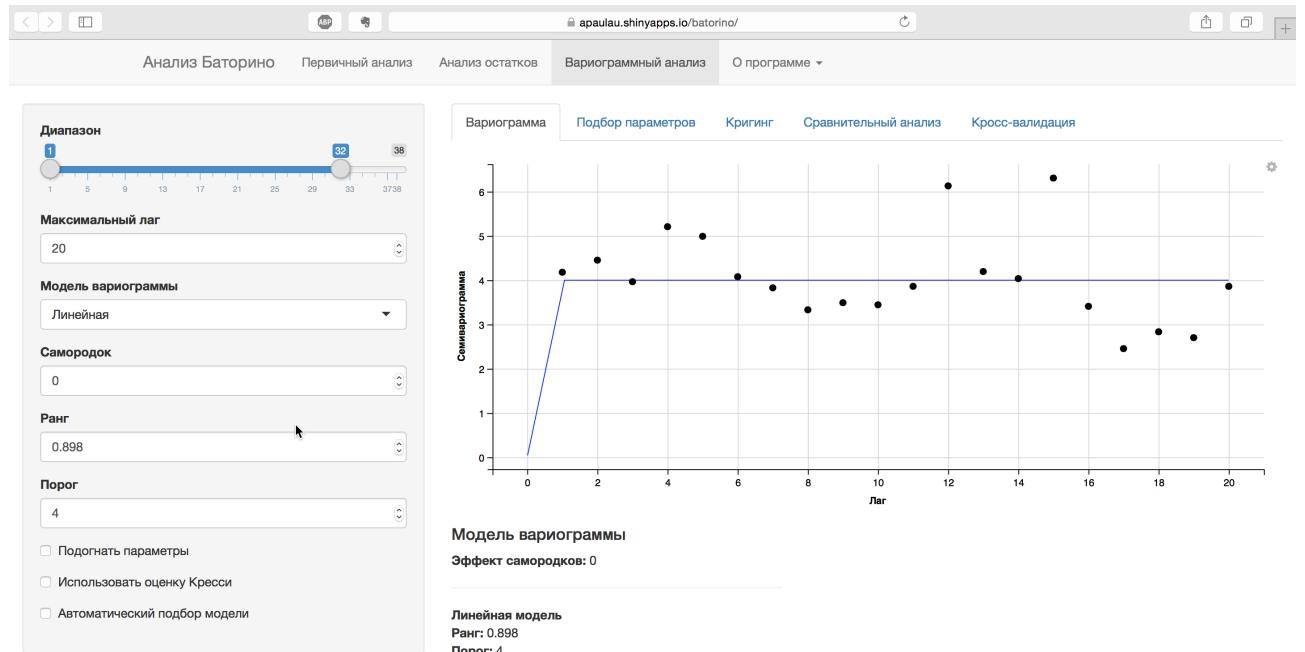


Рисунок 3.4 — Возможности по подбору модели вариограммы

На рисунке 3.4 изображён скриншот начального этапа вариограммного анализа. Ин-

струменты данной страницы позволяют выбрать модель из следующих: Линейная, Сферическая, Экспоненциальная, Гауссовская, Круговая, Бесселя, Пентасферическая, Волновая, Логарифмическая. А также задать к выбранной модели параметры. Заданные параметры считаются начальными, если выбрать опцию подгона методом наименьших квадратов. На этом шаге также можно воспользоваться реализованным в рамках данной работы алгоритмом автоматического подбора модели. Данная функциональность позволяет сразу перейти к вычислению прогнозных значений и не требует каких-либо прикладных знаний у пользователя. Алгоритм заключается в переборе всех представленных в пакете *gstat* моделей, и подборе параметров с помощью функции *fit.variogram* из того же пакета. Каждая итерация сопровождается оптимальным набором параметров для конкретной модели и невязкой. Выбор наилучшей модели осуществляется по минимальному значению невязки. Представленная страница позволяет оценить по графику вариограммы подобранные либо вручную, либо автоматически модель и параметры. Можно также проследить, как влияет тот или иной параметр на теоретическую вариограмму.

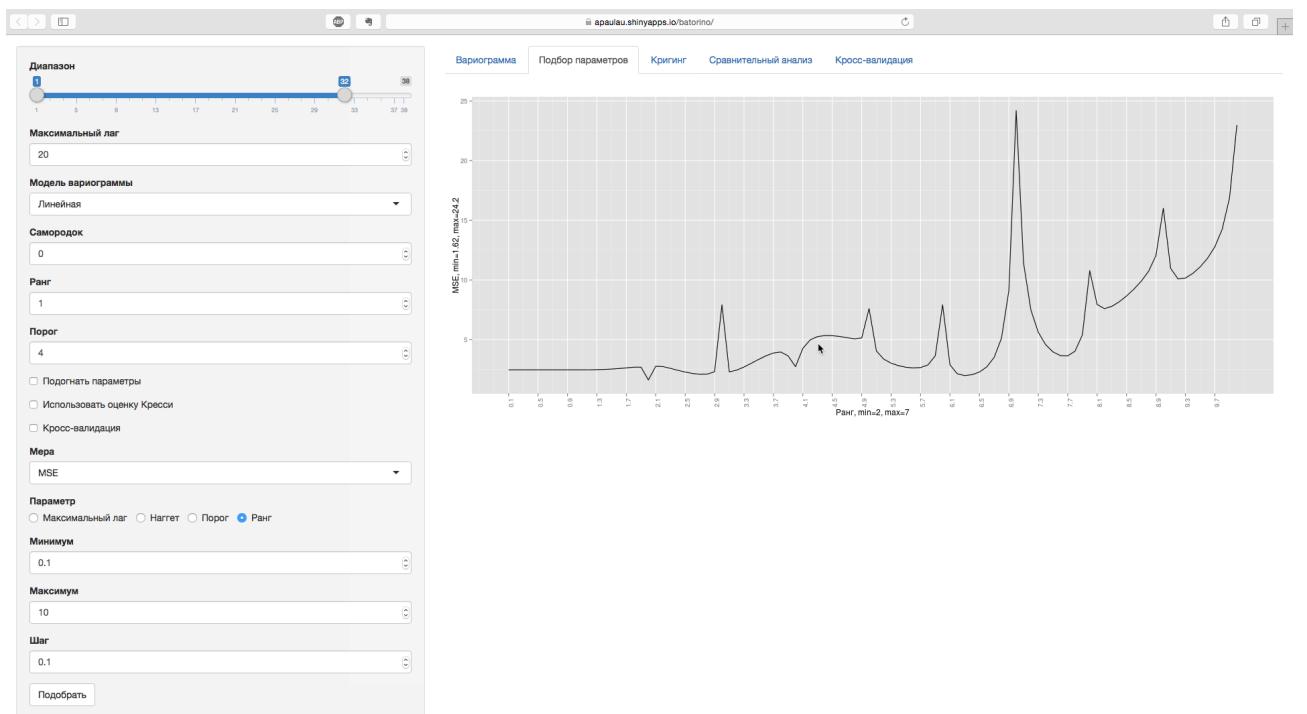


Рисунок 3.5 — Подбор параметров модели вариограммы

Следующая вкладка (рисунок 3.5) заключает в себя функциональность по подбору параметров. В большей мере это относится к ручному выбору. В общем случае, подбор осуществляется следующим образом:

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

В результате такого процесса получается ряд оценок моделей, зависящих от значения параметров. На их основе выбирается оптимальный. Затем процесс повторяется для другого параметра и так далее, пока не найдётся оптимальная модель. В реализованном приложении имеется два подхода по оценке качества построенной модели. Используя первый

подход, модель оценивается с помощью метода кросс-валидации. В данном случае он заключается в последовательном исключении одного из известных значений и построении интерполяции в этой точке по валидируемой модели. Таким образом получаем ряд интерполяций, который должен в идеальном случае воспроизводить поведение исследуемого ряда. Поэтому появляется возможность с помощью различных статистик оценивать конкретную модель вариограммы. С помощью таких статистик можно проследить, как изменяется качество модели при изменении какого-либо из параметров. Примером вычисляемых статистик являются среднеквадратическое отклонение и коэффициент корреляции между фактическими и вычисленными значениями. Таким образом появляется возможность построения графика зависимости качества модели от исследуемого параметра. Это в свою очередь позволяет найти оптимальное значение искомого параметра и использовать его для подбора остальных. При втором подходе, адаптивном, в исследуемых данных отдаётся предпочтение последним наблюдениям. Для этого отбрасывается некоторое количество значений для последующего обучения модели. Подбор параметров осуществляется по статистикам, рассчитанным по отклонениям прогнозных значений от наблюдаемых. Таким образом достигается наилучший прогноз в краткосрочной перспективе. Побочным результатом, реализованного, является возможность оценить поведение модели при изменении одного из параметров.

Страница кригинга (рисунок 3.6) является наглядной демонстрацией применения всего вышеописанного. На ней изображается график с наблюдаемыми значениями и прогнозными значениями, вычисленными кригингом и по регрессионной модели. Это позволяет

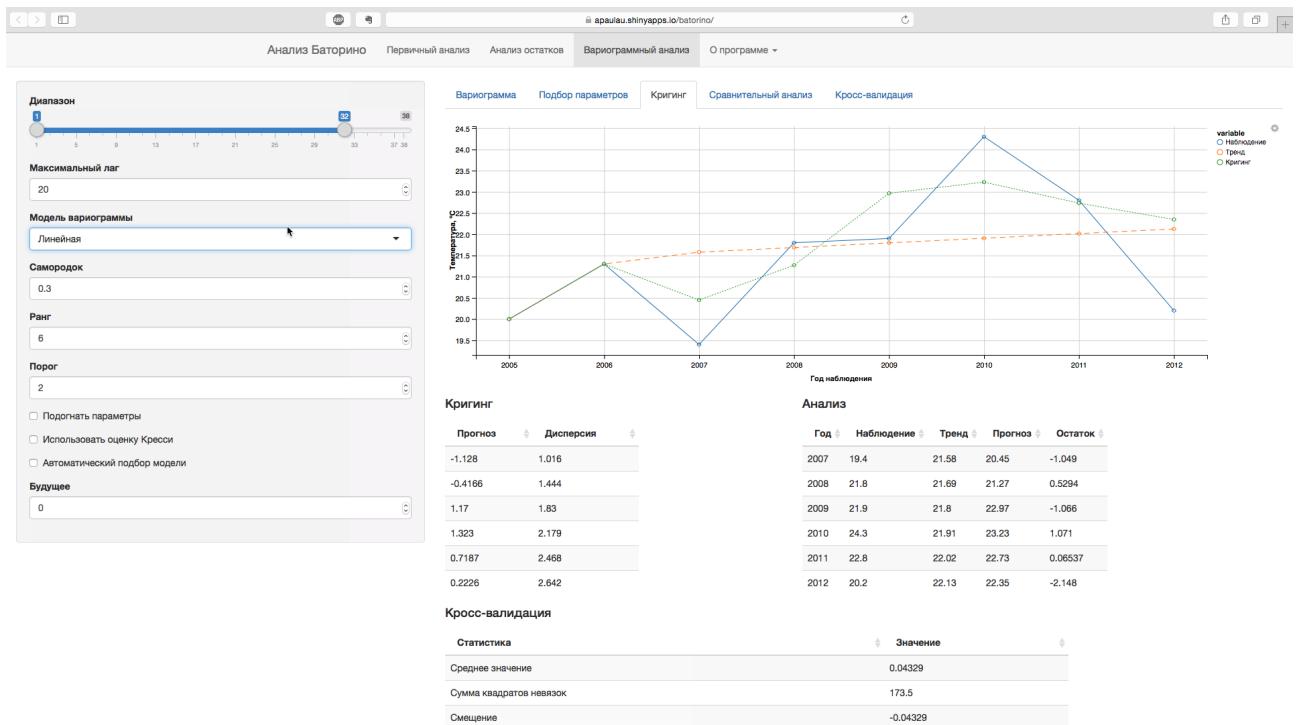


Рисунок 3.6 — Подбор параметров модели вариограммы

оценить полученную модель и сделать различные заключения. График также сопровождается вспомогательными таблицами с произведёнными в процессе расчётом. В первую очередь это результаты кригинга с ошибкой для каждого из значений. Также отображается табличный вариант данных, изображённых на графики. И последняя таблица показывает значения статистик после применения кросс-валидации, что сразу позволяет сравнить конкретную модель с другими.

Глава 4

Анализ временного ряда в среде R

В данной главе исследование проводится в программе, рассмотренной в главе 3. Такой подход позволяет быстро и наглядно рассмотреть и проанализировать различные группы данных. При этом инструменты анализа являются гибкими и легко расширяемыми. Что, в свою очередь, позволяет быстро реагировать под особенности определённой задачи.

4.1 Детерминированный подход

4.1.1 Описательные статистики и первичный анализ данных

В качестве исходных данных примем выборку из полученной от учебно-научного центра базы данных, путём отбора наблюдений в июле месяце за период с 1975 по 2012 год. Выборка представлена в приложении А в таблице А.1. Графически исходные данные представлены на рисунке 4.1.

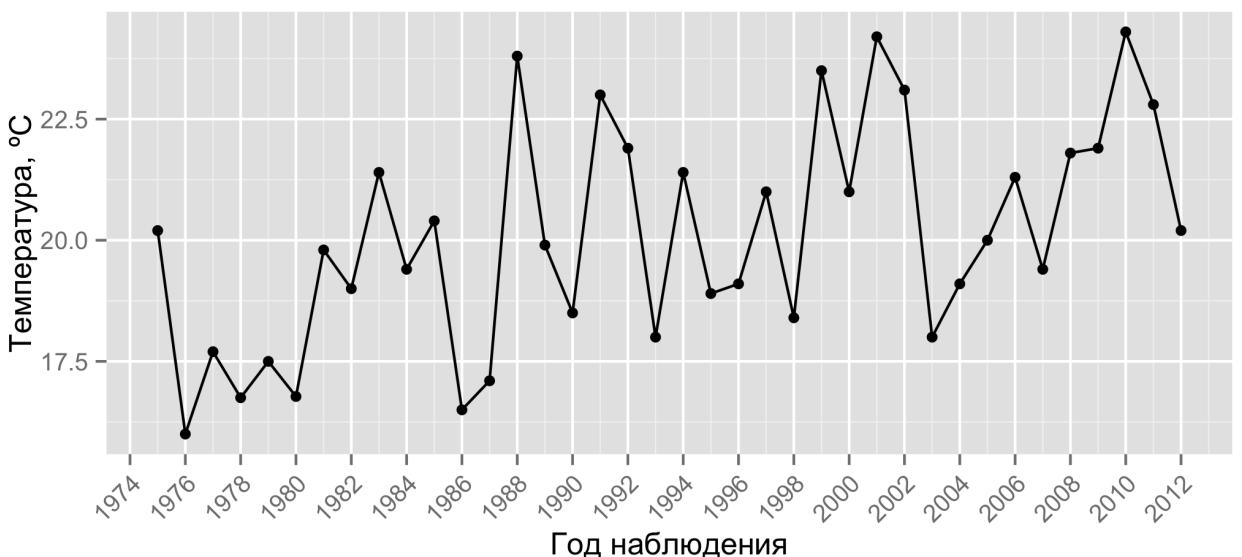


Рисунок 4.1 — График исходных данных

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

Начнём исследование временного ряда с вычисления описательных статистик. Полученные результаты для исходных данных отображены в таблице 4.1. Рассмотрим подробнее некоторые полученные статистики.

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

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

Таблица 4.1 — Описательные статистики для наблюдаемых температур.

Коэффициент вариации в нашем случае равен 25.92%. Из этого следует, что выборку можно считать однородной, так как полученное значение является меньшим 33% [14].

Коэффициент асимметрии — мера симметричности распределения. Полученное значение: 0.30. Данное значение говорит о незначительной правосторонней асимметрии распределения. То есть о том, что выборочное распределение можно считать близким к нормальному [16].

Коэффициент эксцесса в рассматриваемом случае равен -0.746. Так как коэффициент эксцесса нормального распределения равен 0, то в данном случае можно говорить о пологости пика распределения выборки по отношению к нормальному распределению [16].

С помощью тестовых статистик для коэффициента асимметрии и эксцесса [15, 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:

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

Данное значение попадает под случай $|Z_{A_S}| \leq 2$, а значит, выборочный коэффициент асимметрии не является значимым. Из чего, в свою очередь, следует, что по нему нельзя судить о коэффициенте асимметрии генеральной совокупности [15, с.85].

Полученная тестовая статистика для коэффициента эксцесса:

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

Данное значение попадает под случай $|Z_K| \leq 2$, а значит, в данном случае выборочный коэффициент эксцесса не является значимым и нельзя ничего сказать о коэффициенте эксцесса генеральной совокупности [15, с.89].

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

С помощью возможностей реализованной программы построим гистограмму для отображения вариационного ряда исходных данных [13]. Гистограммы позволяют увидеть, как распределены значения переменных по интервалам группировки, то есть как часто переменные принимают значения из различных интервалов. А также, что бывает более важным, позволяет сделать предположение о разновидности распределения. Для вычисления интервалов разбиения воспользуемся *формулой Стерджеса*. Из [17] количество интервалов разбиения рассчитывается по формуле:

$$k = \lceil \log_2 n \rceil + 1 = \lceil \log_2 32 \rceil + 1 = 6. \quad (4.1)$$

Так как по гистограмме можно визуально предположить близость выборочного распределения к нормальному распределению, нанесём на график кривую плотности нормального распределения (рисунок 4.2). Проанализируем эту гистограмму. Во-первых, на

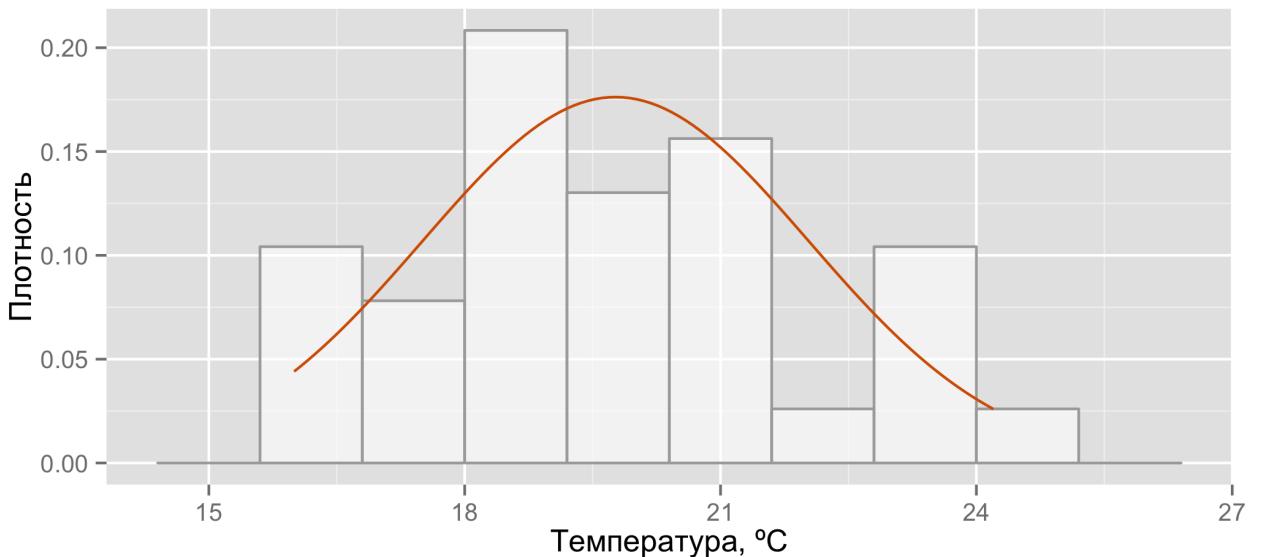


Рисунок 4.2 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения $\mathcal{N}(19.77, 5.12)$

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

Следует отметить согласованность полученных описательных статистик с полученной гистограммой. Во-первых, по коэффициенту асимметрии мы предположили о близости распределения к симметричному. Это подтверждается гистограммой: на ней можно заметить небольшую склонность 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 колокообразную форму.

Другим часто используемым графическим способом проверки характера распределения данных является построение т.н. *графиков квантилей* (*Q-Q plots*, *Quantile-Quantile plots*). На таких графиках изображаются квантили двух распределений — эмпирического (т.е. построенного по анализируемым данным) и теоретически ожидаемого нормального распределения. При нормальном распределении проверяемой переменной точки на графике квантилей должны выстраиваться в прямую линию, исходящую под углом 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 рамках реализованной программы построен график 4.3. На этом графике можно

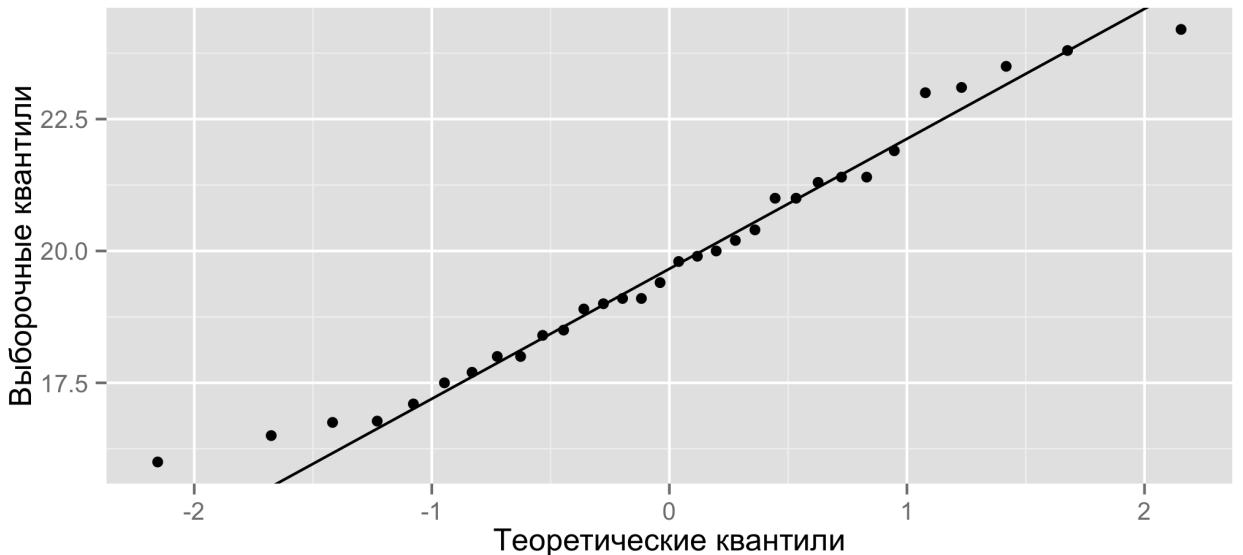


Рисунок 4.3 — График квантилей для наблюдаемых температур

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

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

В R реализованы практически все имеющиеся тесты на нормальность — либо в виде стандартных 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* [18]. Из полученных в R результатов, статистика Шапиро-Уилка $W = 0.97$. Вероятность ошибки $p = 0.43 > 0.05$, а значит нулевая гипотеза не отвергается [19]. Следовательно опровергнуть предположение на основе данного теста нельзя.

Попробуем опровергнуть наше предположение на основе проверки критерия χ^2 Пирсона [20]. Для этого воспользуемся пакетом `nortest` и функцией `pearson.test`. Из полученных в R результатов, статистика χ^2 Пирсона $P = 2.00$. Вероятность ошибки $p = 0.85 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста также нельзя. Проверим критерий: примем уровень значимости $\alpha = 0.05$, тогда из таблицы распределения 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 [21]. Как в предыдущем случае воспользуемся представленной в пакете *nortest* функцией *ks.test*. Из полученных в **R** результатов, статистика Колмогорова–Смирнова $D = 0.087$. Вероятность ошибки $p = 0.97 > 0.05$, а значит нулевую гипотезу отвергнуть нельзя. Следовательно опровергнуть предположение о нормальности, как и в предыдущих случаях, также нельзя. Проверим критерий: примем так же уровень значимости $\alpha = 0.05$, тогда критическое значение $D_{\text{кр}}(\alpha) = 1.358$. Следовательно,

$$D < D_{\text{кр}}(\alpha),$$

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

На данном этапе по полученным ранее результатам возникли подозрения о выбросах в исходной выборке. Выявление таких аномальных значений важно, так как их наличие, как правило, сильно влияет на всю выборку, в частности, на коэффициент корреляции. Проверим наличие выбросов с помощью статистических критериев. Для этих целей воспользуемся критерием Граббса [22]. 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 [23]. Поскольку ранее высказано предположение о нормальности, воспользуемся им для определения наличия выбросов.

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

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

4.1.2 Корреляционный анализ

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

Диаграммы рассеяния используются для визуального исследования зависимости между двумя переменными. Если переменные сильно связаны, то множество точек данных принимает определённую форму. С помощью таких диаграмм можно наглядно изучить знак коэффициента корреляции. Если точки на диаграмме расположены хаотически, то это говорит о независимости рассматриваемых переменных. Если с ростом переменной t возрастает переменная x то имеет место положительная корреляция. Если же с ростом переменной t переменная x убывает, то это указывает на отрицательную корреляцию.

Из рисунка 4.4 видно, что точки образуют своеобразное «облако», ориентированное по вверх, то есть присутствует некая зависимость между рассматриваемыми переменными.

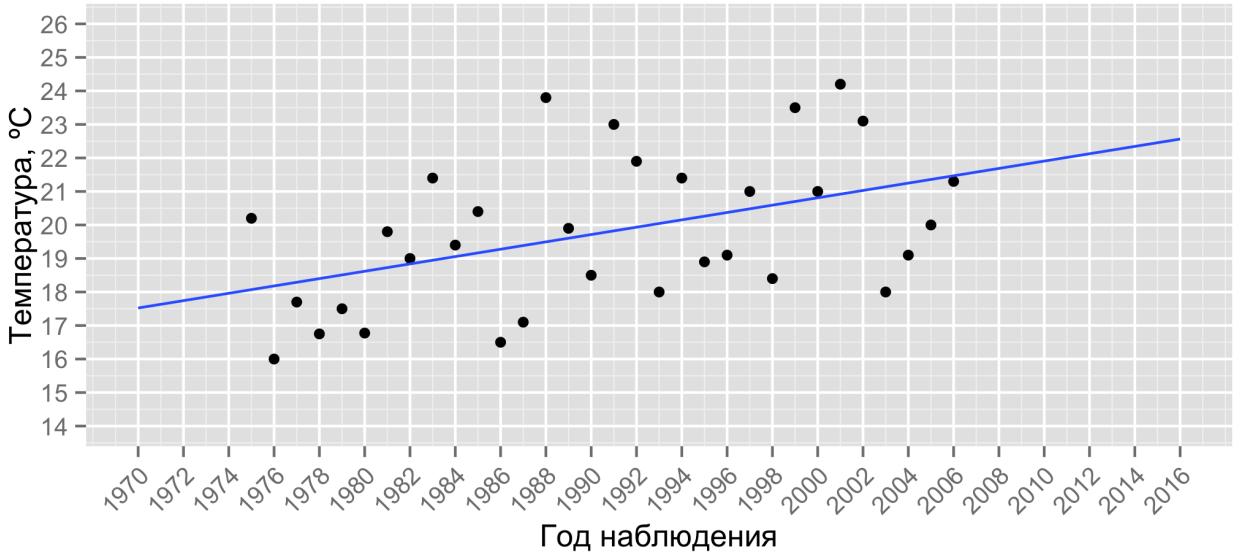


Рисунок 4.4 – Диаграмма рассеяния

Также, данная диаграмма наглядно показывает силу этой зависимости: так как точки не образуют чёткой формы, а разбросаны относительно линии, то можно говорить о наличии умеренной корреляции.

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

Проверим значимость полученного выборочного коэффициента корреляции с помощью критерия Стьюдента:

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

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

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

Значит нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности следует отклонить [14].

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

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

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

4.1.3 Регрессионный анализ

Для введения последующих понятий анализа временных рядов воспользуемся [24].

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

1. *Тренд (тенденция развития)* — эволюционная составляющая, которая характеризует общее направление развития изучаемого явления и связана с действием долговременных факторов развития.
2. *Циклические, сезонные колебания* — это составляющие, которые проявляются как отклонения от основной тенденции развития изучаемого явления, и связаны с действие краткосрочных, систематических факторов развития.
3. *Нерегулярная случайная составляющая (ошибка)*, являющаяся результатом действия второстепенных факторов развития.

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

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

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

где y, k, s, ε — тренд, циклическая, сезонная и нерегулярная составляющие соответственно.

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

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

При проведении корреляционного анализа, на графике 4.4 был замечен явно выраженный линейный рост значений со временем. Что впоследствии было подтверждено критериями. Из этого следует, что уравнение тренда имеет вид:

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

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

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

$$x = y + \varepsilon,$$

где y — тренд, ε — нерегулярная составляющая.

В **R** реализованы функции, позволяющие подгонять линейные модели к исследуемым данным [25]. Одной из таких функций является *lm(Fitting Linear Model)* [11, с.178]. Она позволяет получить коэффициенты линии регрессии. Таким образом, можно вычислить

одну из искомых компонент – тренд. И как следствие, после его удаления из исходных данных, получим нерегулярную составляющую $\varepsilon(t)$. Коэффициенты, полученные с помощью данной функции представлены в (4.2).

$$a = 0.11, \quad b = 18. \quad (4.2)$$

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

Таким образом получена линейная модель, описывающая тенденцию развития:

$$y(t) = at + b = 0.11t + 18 \quad (4.3)$$

На основе полученной линейной модели (4.3), построим ряд остатков, удалив тренд из исходного ряда. Полученный ряд представлен в приложении В в таблице В.1 и графически на рисунке 4.5.

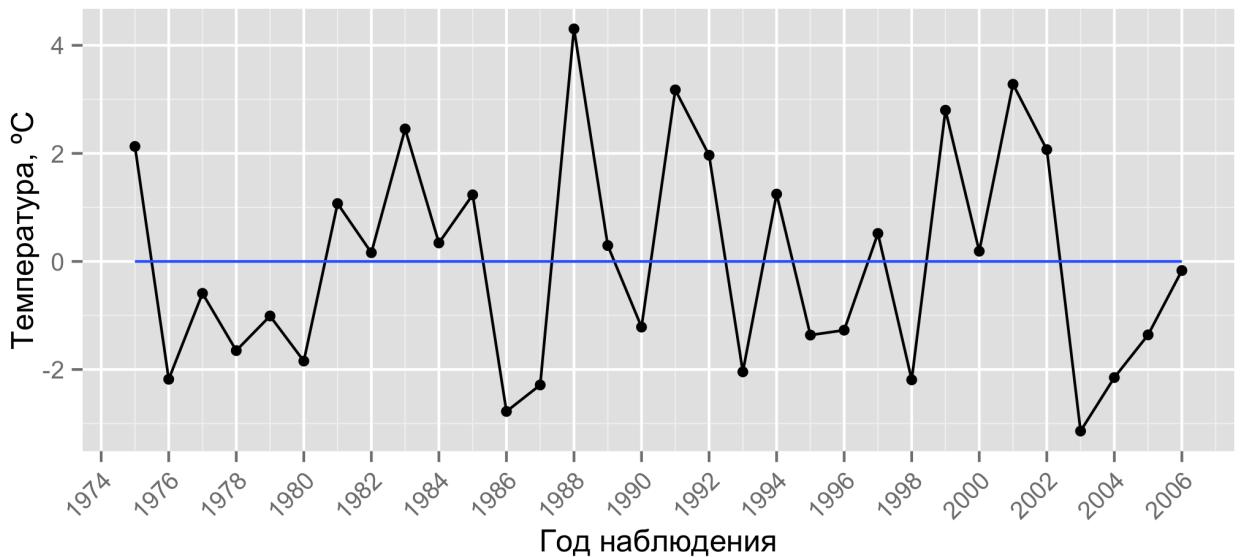


Рисунок 4.5 — Нерегулярная составляющая $\varepsilon(t)$

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

Рассчитаем вспомогательные величины, воспользовавшись [24]. Дисперсия отклонения

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

стандартные случайные погрешности параметров a, b :

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

Воспользуемся критерием значимости коэффициентов линейной регрессии [14]. Приемлем уровень значимости $\alpha = 0.05$, тогда

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

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

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

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

Оценим адекватность полученной регрессионной модели. Дисперсия модели:

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

Остаточная дисперсия:

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

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

$$F_{\text{крит}} \approx 7.79,$$

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

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

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

Рассчитаем коэффициент детерминации:

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

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

Тем не менее, попробуем построить прогноз по полученной модели. Вычисленные прогнозные значения на 2007-2012 годы для сравнения отображены в таблице 4.2:

	Год	Актуальное	Прогнозное	Ошибка
1	2007	19.40	18.07	1.33
2	2008	21.80	18.18	3.62
3	2009	21.90	18.29	3.61
4	2010	24.30	18.40	5.90
5	2011	22.80	18.51	4.29
6	2012	20.20	18.62	1.58

Таблица 4.2 — Сравнение прогнозных значений (тренда)

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

4.1.4 Анализ остатков

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

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

Вычислим описательные статистики для остатков. Полученные результаты проследим по таблице 4.3.

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

Таблица 4.3 — Описательные статистики остатков

Как видно из таблицы 4.3, среднее значение равно нулю. При этом коэффициенты асимметрии ($A_S = 0.38$) и эксцесса ($K = -0.905$) указывают на большее отклонение распределения остатков от нормального закона.

Построим гистограмму и график квантилей для проверки последних заключений. Построенная гистограмма (приложение Б, рисунок Б.1) наглядно демонстрирует полученные в таблице 4.3 коэффициенты асимметрии и эксцесса.

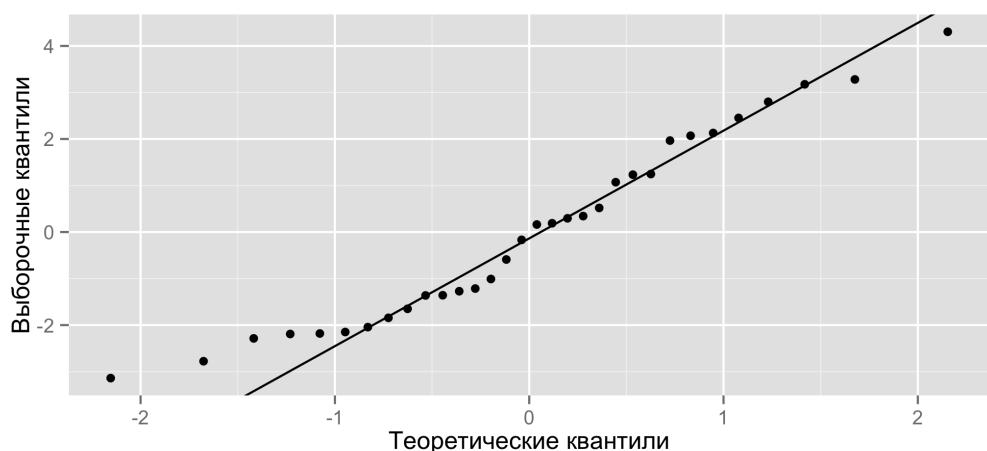


Рисунок 4.6 — График квантилей для остатков

Как и в случае исходных данных график квантилей позволяет наглядно оценить близость к нормальному распределению. На рисунке 4.6 можно заметить, что присутствуют отклонения относительно нормального распределения. Наиболее явный из них — нижний хвост. Остальные — небольшие скачки по ходу линии нормального распределения. Проверим с помощью критерия Шапиро–Уилка, можно ли считать полученные остатки нормально распределёнными. Из полученных в **R** результатов, статистика Шапиро–Уилка $W = 0.95$. Вероятность ошибки $p = 0.17 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста нельзя.

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

Построим график автокорреляционной функции для определения наличия взаимосвязей в ряде остатков (рисунок 4.7). На графике пунктирные линии разграничают значимые и не значимые корреляции: значения, выходящие за линии, являются значимыми [12, с.376]. На представленном графике автокорреляционной функции все значения

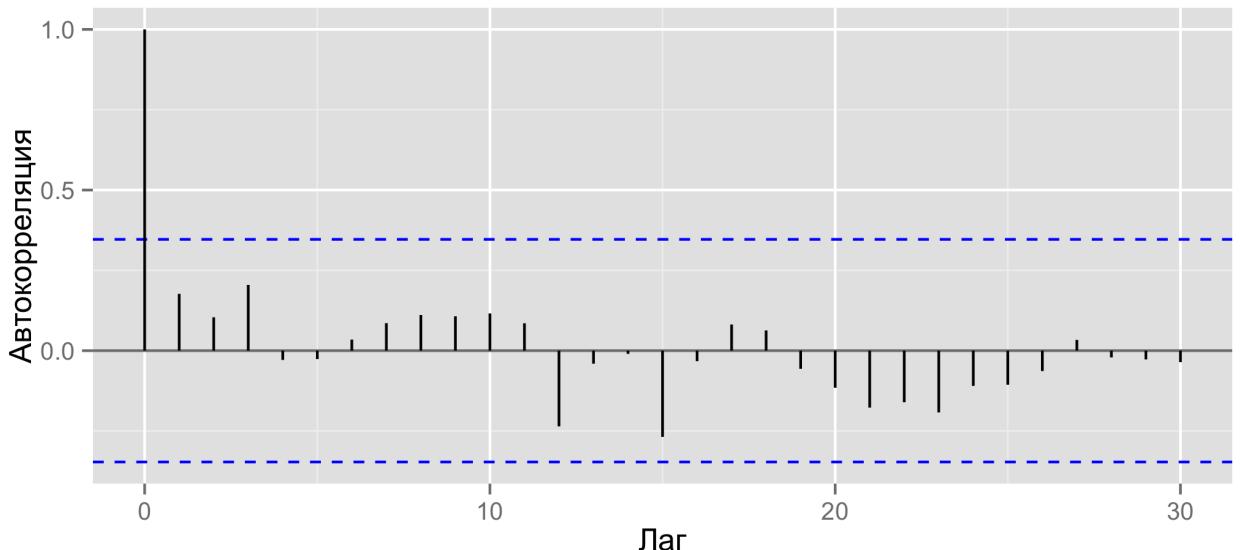


Рисунок 4.7 — График автокорреляционной функции

не выходят за интервал, обозначенный пунктирными линиями. Это означает, что в представленной автокорреляционной функции нету значимых автокорреляций. Проверим это замечание с помощью теста Льюнга–Бокса [12, с.377-378]. Данный тест позволяет проверить наличие автокорреляций в исследуемых данных. Используя возможности пакета **R** получили значения: статистика Льюнга–Бокса $X^2 = 0.073$ и вероятность ошибки $p = 0.79 > 0.05$ — это говорит о том, что тест не выявил значимых автокорреляций.

На рисунке 4.7 также можно заметить некоторое затухание значений автокорреляций с увеличением лага. На основе этого можно сделать предположение о стационарности. Для проверки этого предположения воспользуемся расширенным тестом Дики–Фуллера(ADF) [26]. Из результатов проверки теста, статистика Дики–Фуллера $DF = -3.36$, вероятность ошибки $p = 0.08 < 0.05$. Следовательно, при уровне значимости $\alpha = 0.05$ необходимо принять альтернативную гипотезу о стационарности.

Таким образом в результате анализа детерминированными методами выделены две составляющие исходной модели данных: тренд и нерегулярная составляющая. В ходе регрессионного анализа было показано, что модель, основанная на тренде, не позволяет

воспроизвести поведение исходного временного ряда. То есть нерегулярная составляющая $\varepsilon(t)$ является существенной и отвечает за это поведение. Для того, чтобы определить возможность её дальнейшего исследования проведен анализ остатков, в процессе которого показаны близость распределения к нормальному (с некоторыми отклонениями) и стационарность, при этом не выявлено значимых автокорреляций. Таким образом, это позволяет перейти к построению модели другими, современными статистическими методами интерполяции. Улучшение модели будет происходить за счёт суперпозиции модели, полученной на данном этапе, и найденной модели нерегулярной составляющей.

4.2 Геостатистический подход

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

В современных исследованиях аналогичного класса усилился интерес к геостатистическим моделям интерполяции, что подтверждается работами [27, 28]. Современная геостатистика — это широкий спектр статистических моделей и инструментов для анализа, обработки и представления пространственно-распределенной информации.

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

4.2.1 Вариограммный анализ. Кригинг.

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

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

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

Центральная идея геостатистики состоит в использовании знаний о корреляции экспериментальных данных для построения оценок и интерполяций. *Вариограмма* является ключевым инструментом для оценки степени корреляции, имеющейся в исследуемых данных, и для ее моделирования. Модель вариограммы является функцией, определяющей зависимость изменения исследуемой величины от расстояния. Следовательно, интерполяционная модель, основанная на такой корреляционной функции, будет отражать реальные явления, которые лежат в основе данных измерений. Все возможные пары точек могут быть рассортированы по классам в соответствии с разностью их координат

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

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

Для оценки поведения данных при увеличении лага построим диаграмму взаимного разброса пар точек (*h -scatterplot*), разделённых расстоянием h . Эта диаграмма позволяет проверить наличие корреляции в исследуемых данных как качественно, так и количественно [29]. Построенная диаграмма изображена на рисунке Б.2 в приложении Б. Следует отметить, что в классическом случае присутствия зависимости, поведение должно быть следующим: на графиках, соответствующим начальным лагам, должна присутствовать сильная корреляция, и с увеличением лага корреляция уменьшается. Это объясняется тем, что чем ближе находятся данными тем выше зависимость между ними и наоборот. В рассматриваемом случае такого не наблюдается. Напротив, на первом же лаге отсутствует корреляция, при этом можно наблюдать, что на некоторых лагах присутствует корреляция, на некоторых нет. Такое поведение свойственно так называемым беспороговым моделям вариограммы. Другими словами, моделям, в которых отсутствует ранг. Одной из таких моделей является линейная (4.4), с которой некоторые исследователи советуют начинать подбор модели. Аргументируется это тем, что, она является простейшей.

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

где b – параметр, отвечающий за угол наклона.

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

В некоторых источниках советуют при построении вариограммы учитывать параметр максимального расстояния, для которого вычисляется вариограмма, а также приводят рекомендацию по его подбору. Поэтому первоначальным параметром было выбрано значение, рассчитанное по такой рекомендации: $2n/3 = 20$ [30].

В общем случае процесс вариограммного анализа заключается в выполнении серии шагов. Первым шагом вычисляют экспериментальную вариограмму, затем, при начальных значениях порога и ранга подбирают теоретическую вариограмму и с помощью различных методов пробуют улучшить её качество. После получения удовлетворительной модели используют метод кригинга для вычисления прогнозных значений.

Экспериментальной вариограммой по сути является некоторая оценка вариограммы. Существует несколько известных оценок, каждая из которых имеет свои достоинства и недостатки. Для данного исследования были выбраны наиболее распространённые: оценка Матерона (2.1), введённая ранее в главе 2, и оценка Кресси-Хокинса [31, 32]:

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

для сравнения полученных результатов. Оценка Кресси-Хокинса является робастной и в теории позволяет учесть наличие выбросов в исследуемых данных [33].

Как уже было сказано ранее, в процессе анализа остатков, случайный процесс $\varepsilon(t)$ имеет распределение, близкое к нормальному, математическое ожидание равное 0, обладает

постоянной дисперсией, является стационарным и не имеет значимых автокорреляций. Это в полной мере удовлетворяет свойствам оценки Матерона, рассмотренной в главе 2. Поэтому сначала проведём исследования с её помощью, а затем сравним полученные результаты с результатами использования оценки Кресси-Хокинса, которая в теории должна учитывать наличие выбросов в выборке.

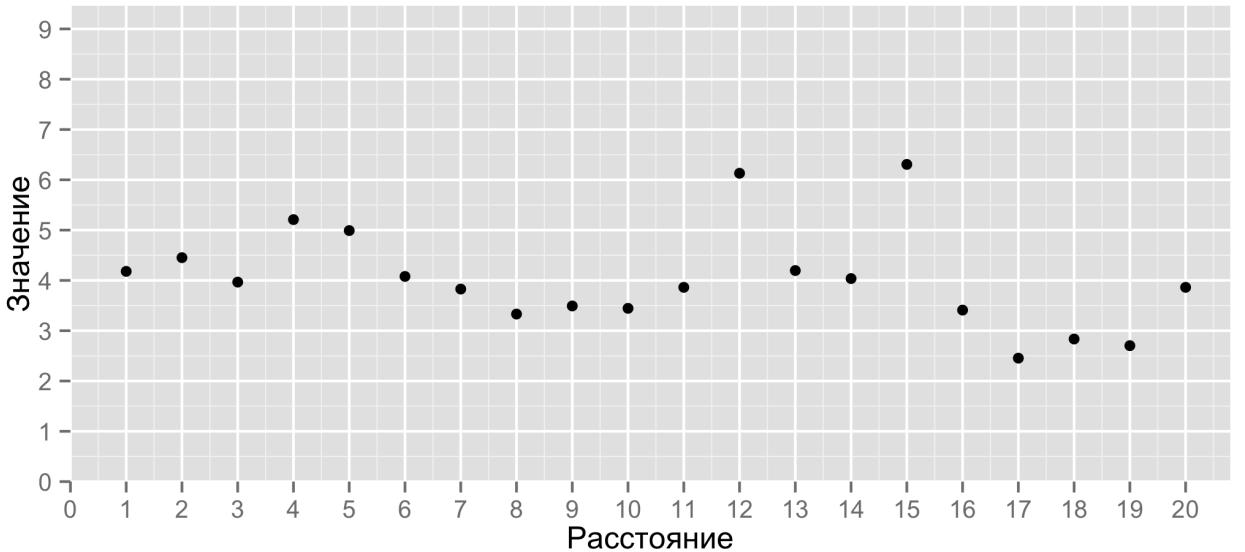


Рисунок 4.8 — Экспериментальная вариограмма (оценка Матерона)

Построенная вариограмма отображена на рисунке 4.8. При подборе моделей вариограммы, в общем случае, существует два пути: подбор силами исследователя, т.е. визуально с ручным выбором параметров, и автоматическим подбором параметров с помощью специальных методов и алгоритмов. На практике построение модели вариограммы представляет собой итеративный процесс, на каждом шаге которого следует наилучшим образом подобрать параметры очередного модельного приближения. В различных источниках рекомендуется строить модели вручную, так как исследователь лучше знает специфику данных, чем различные методы оценивания [34]. Далее будем строить модель вариограммы по рекомендации — визуально.

Вследствие выводов по диаграмме взаимного разброса, начнём подбор теоретической вариограммы (модели) с линейной. В целях приведения процесса отыскания оптимальных параметров выполним шаги упомянутые ранее подробно. Подбор начальных параметров для построения теоретической вариограммы осуществим визуально. Как уже было сказано, в беспороговых моделях ранг $a = 0$. Из уравнения (4.4) видно, что параметр отвечает за скорость роста значений, поэтому примем начальное значение $b = 4$. Вычисленная модель отображена на графике Б.3. Следует отметить, что данная модель после применения кригинга позволила получить прогнозные значения очень близкие к нулю, что не изменила прогноза, построенного по модели тренда (таблица 4.2).

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

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

с параметром $c = 4.04$. Объяснить такой результат можно тем, что значения вариограммы сразу достигают порогового значения приблизительно равному дисперсии и не имеют большого разброса относительно среднего. При этом также можно считать, что вывод,

сделанный по диаграмме взаимного разброса не совсем верен, поскольку это поведение можно объяснить стационарностью процесса. Поэтому метод подбора параметров, основывающийся на методе наименьших квадратов, производит такие результаты. Это следствие того, что данных подход не учитывает особенностей исследуемых данных. Поэтому результатов прогнозирования данная модель не улучшила.

Но при этом наличие порога объяснима видом вариограммы: по рисунку 4.8 можно видеть, что уже первые значения достигают уровня дисперсии. И отклонение от этого значения не велико. Это согласуется с исследуемыми исходными данными, так как при анализе остатков было выявлено отсутствие автокорреляций, и спецификой самих данных: значение температуры воды за определённый год слабо зависит от значения предшествующего. Из этого следует, что использование беспороговых моделей не обосновано. Поэтому попробуем улучшить результаты с помощью линейной модели с порогом:

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

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

Для подбора оптимальных параметров линейной модели с порогом будем проводить с помощью инструментов реализованной программы, рассмотренные в главе 3. Воспользуемся первым из описанных подходом; в качестве статистики будем использовать коэффициент корреляции между интерполированными и актуальными значениями. График зависимости значения ранга на качество модели отображён на рисунке 4.9. По рисунку

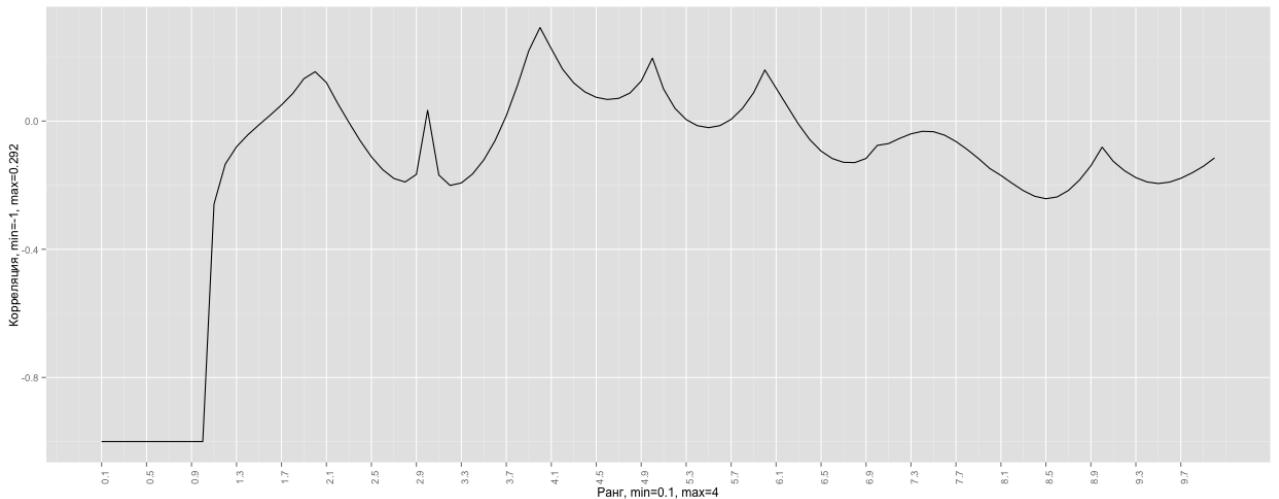


Рисунок 4.9 — Зависимость качества линейной модели от значения ранга

видно, что максимальное значение коэффициента корреляции $r = 0.292$ достигается при значении ранга $a = 4$, при этом среднеквадратическое отклонение от истинных значений $MSE = 7.931$. Таким образом получена модель $4 \cdot \text{Lin}(h, 4)$, ее график отображен на рисунке Б.5. Вычисленные по данной модели прогнозные значения можно проследить по таблице 4.4 и по графику Б.6 в приложении Б. Как можно видеть, прогнозные значения предсказали только поведение в первый год. Дальнейшие значения оказались далеки от истины, что объясняется значением среднеквадратической ошибки. Таким образом, данная модель неплохо себя показала при описании всех данных, что показал коэффициент корреляции, но при этом прогноз оказался не точным.

Для построения более точного прогноза воспользуемся аддитивным подходом по подбору параметров. График зависимости среднеквадратической ошибки от ранга отображен на рисунке 4.10. Из него видно, что оптимальным параметром для ранга является $a = 2$, с

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

Таблица 4.4 — Прогноз (линейная модель с порогом)

минимальной среднеквадратической ошибкой $MSE = 1.62$. График 4.10 прогнозных значений показывает, что данный подход позволил предсказать три первых значения. Что является хорошим результатом. При этом статистики по данной модели после проведения кросс-валидации оказались следующими: коэффициент корреляции $r = 0.152$, среднеквадратическая ошибка 18.69. Что говорит о том, что данная модель описывает всю выборку хуже чем предыдущая.

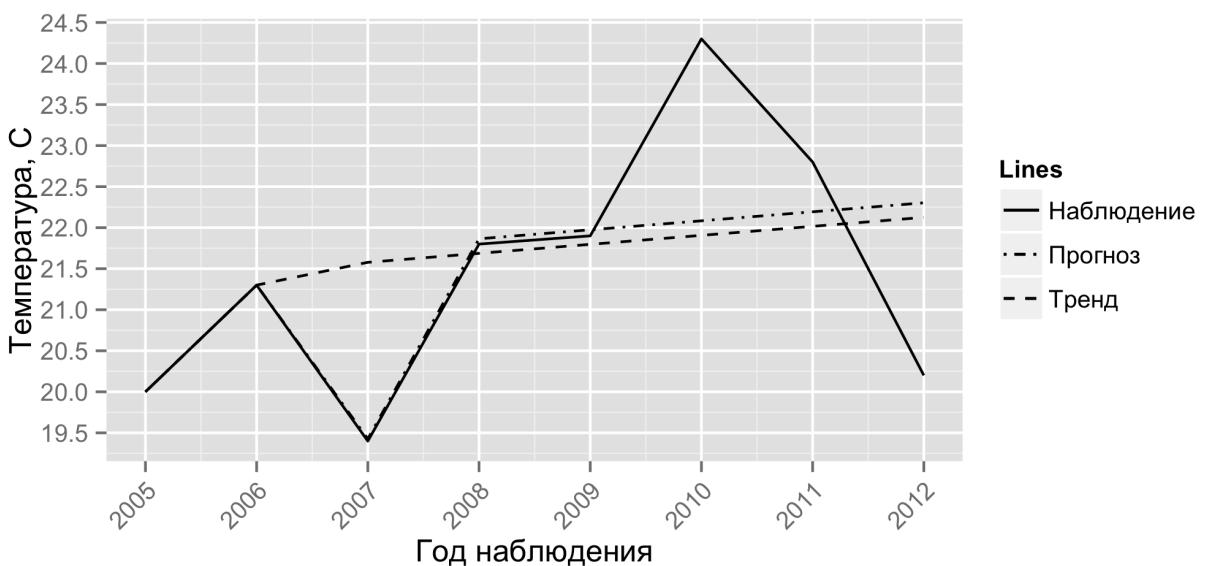


Рисунок 4.10 — Прогноз по модели $2 \cdot Lin(h, 2)$

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

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

пороговой моделью является сферическая:

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

Однако после подбора оптимальных параметров оказалось, что данная модель вписывается в исследуемые данные хуже, чем найденные ранее. При подборе параметров с помощью кросс-валидации наилучшей получилась модель $4Sph(h, 2.3)$ с показателями: коэффициент корреляции $r = -0.002$ и среднеквадратическим отклонением $MSE = 5.407$. В случае адаптивного подхода, оптимальными оказались параметры $c = 4, a = 6.9$ с эффектом самородков равным 0.9, при среднеквадратическом отклонении $MSE = 2.01$. Применив кросс-валидацию к этой модели, получаем следующие показатели качества: коэффициент корреляции $r = -0.009$ и среднеквадратическим отклонением $MSE = 5.396$. Графики вариограммы и прогнозных значений последней модели отображены на рисунках Б.9 и Б.10 в приложении Б соответственно. Можно сделать вывод, что как и линейная с порогом модель, сферическая не позволила описать поведение исследуемой выборки. Только в случае краткосрочного прогноза она проявила себя, предсказав характерное поведение исключённых значений, хоть и хуже предшествующей линейной с порогом модели. Похожее поведение можно объяснить видом их теоретических вариограмм. Это видно по графикам Б.8 и Б.9 в приложении Б.

Если обратить внимание на график экспериментальной вариограммы 4.8, то можно заметить некоторый периодический эффект в виде волны. Поэтому дальнейшей подбираемой моделью возьмем периодическую:

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

где c – порог, a – ранг.

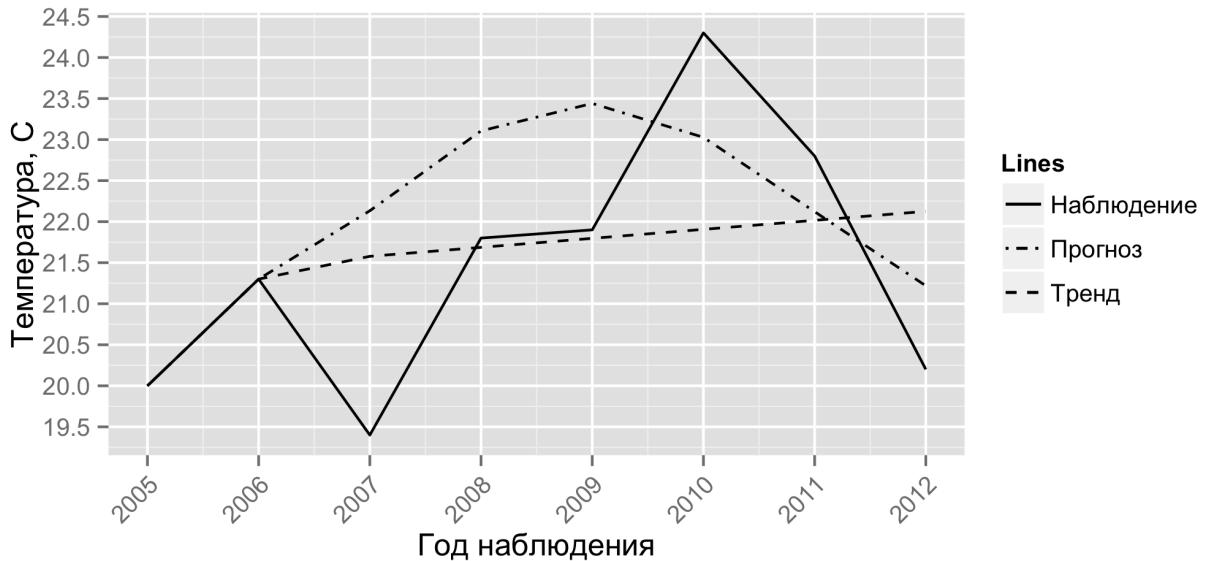


Рисунок 4.11 – Прогноз по модели $4 \cdot Per(h, 0.898)$

С помощью средств написанной программы подобрана модель $4 \cdot Per(h, 0.898)$, график вариограммы которой изображен на рисунке Б.11 в приложении Б. Показатели адекватности подобранный модели: коэффициент корреляции $r = 0.404$, среднеквадратическая ошибка $MSE = 4.369$. Следует отметить, что из всех подбираемых моделей, представленное значение коэффициента корреляции оказалось самым большим. Что говорит о том,

что данная модель наилучшим образом описывает исследуемые данные. Таблица В.2 в приложении В и график прогнозных значений по подобранный модели показывают, что прогноз получился не очень точный, но при этом следует принять во внимание упомянутый ранее эффект сглаживания. Данная модель оказалась наилучшей для описания всей выборки из всех проверенных. Следует отметить, что инструменты используемой программы позволяют относительно быстро проверить различные модели на практике.

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

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

Как было отмечено, существуют также автоматические методы подбора моделей и параметров специальными методами и алгоритмами. В рамках данной работы был реализован алгоритм, основанный на возможностях пакета *gstat*, позволяющий автоматически выбирать из заданных моделей и начальных параметров и наилучшую. Подбор параметров определённой модели вариограммы осуществляется методом наименьших квадратов, пример использования которого показан ранее. Данная процедура также вычисляет невязку полученной модели и экспериментальной вариограммы. Это позволяет основываясь на данном показателе выбирать наилучшим образом подходящей модели — по минимальному значению невязки. Следует отметить, что в данном случае параметр максимального расстояния, для которого вычисляется вариограмма, будет влиять на конечный выбор. Поскольку количество точек, по которым подбирается модель, будет влиять на метод наименьших квадратов. Код программы представлен в листинге Г.3.

Воспользуемся описанным способом подбора модели для оценки Матерона. Данная процедура, при принятом ранее максимальном значении лага равным 20, выбрала волновую модель (4.6) вариограммы с эффектом самородков: $3.03 + 1.011 \cdot Wav(h, 1.14)$.

$$\gamma(h) = c \cdot Wav(h, a) = 1 - \frac{a}{h} \cdot \sin\left(\frac{h}{a}\right), \quad (4.6)$$

где c — порог, a — ранг.

Графически подобранный модель и прогноз, построенный на её основе с применением кригинга, отображены на рисунках Б.13 и Б.14 соответственно. По которым видно, что результат получился значительно хуже найденных ранее. При этом показатели качества оказались равными: коэффициент корреляции $r = -0.2$ и среднеквадратическая ошибка $MSE = 4.155$.

Так как параметр максимального лага вариограммы влияет на подбор моделей, оценим качество модели, подобранных автоматически, в зависимости от него. Как и ранее, качество модели будем оценивать двумя подходами. Первый случай отображен на рисунке 4.12. Как видно из него, наибольший коэффициент корреляции соответствует оценке Матерона и максимальному расстоянию, равному 26. Для найденного параметра, наилучшая модель $3.46 + 0.5 \cdot Per(h, 2.67)$ с показателями качества: коэффициент корреляции $r = 0.196$, среднеквадратическим отклонением $MSE = 3.835$. График прогнозных значений отображен на рисунке Б.15. Таким образом автоматическим способом найдена модель, которая была получена ранее вручную. Отличие заключается только в значениях параметров. В случае автоматического подбора, модель хуже описывает поведение исходных

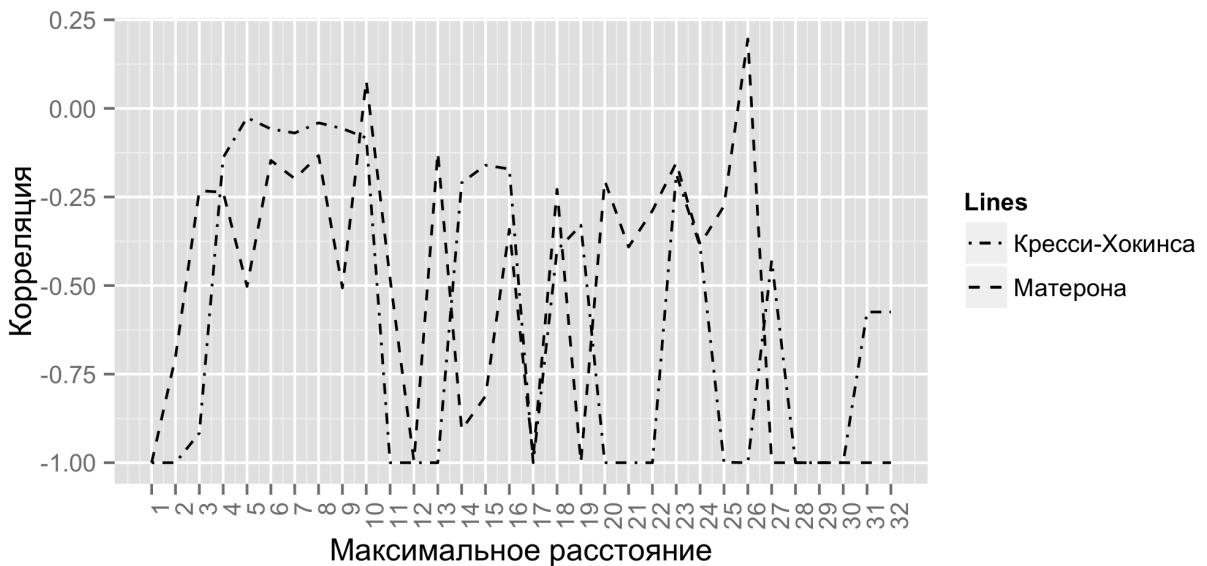


Рисунок 4.12 — Зависимость качества модели от значения максимального значения

данных. Но при этом следует учитывать затраты на поиск той или иной модели в обоих случаях.

Также проследим зависимость значения максимального расстояния при адаптивном подходе. График такой зависимости изображен на рисунке 4.13. В данном случае, оп-

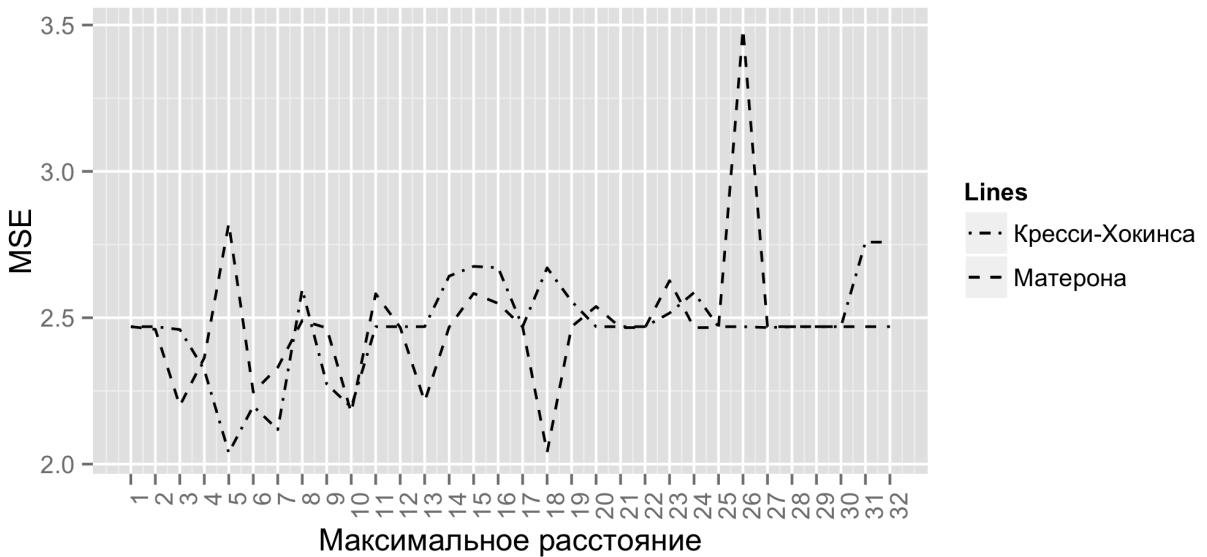


Рисунок 4.13 — Зависимость качества модели от значения максимального значения

тимальной моделью оказалась волновая модель с эффектом самородков $4.11 + 1.65 \cdot Wav(h, 3.59)$, полученная по оценке Кресси-Хокинса при значении максимального расстояния равного 5. И при почти равном значении среднеквадратической ошибки, периодическая модель с эффектом самородков $3.8 + 0.32 \cdot Per(h, 1.3)$ по оценке Матерона при значении максимально расстояния равного 18. Прогнозные значения первой и последней можно проследить по таблице В.3 в приложении В и таблице 4.5 соответственно. А также графически на рисунках Б.16 и Б.17. Как можно видеть, представленные прогнозные значения далеки от истины. Но при этом поведение исходных данных найденные модели уловили. Что является хорошим результатом, если учитывать специфичность рассматриваемой задачи.

Было показано, что метод наименьших квадратов не учитывает особенностей, которые может учесть исследовать. А так как в данной задаче может присутствовать множество неучтённых факторов и, то применение автоматического подбора моделей предсказуемо показывает результаты хуже. Данный метод подбора моделей может быть обоснованно использован в случае данных, изменение которых носят плавный характер. К примеру, уровень воды некоторого озера. Или, использовать для анализа данные, наблюдения в которых будут располагаться ближе, чем годовой промежуток. Так как температура воды в конкретном месяце скорее будет зависеть от температуры воды предыдущего месяца, чем от температуры воды год назад.

	Год	Наблюдение	Прогноз	Тренд	Ошибка
1	2007	19.400	21.170	21.578	-1.770
2	2008	21.800	21.553	21.687	0.247
3	2009	21.900	22.151	21.797	-0.251
4	2010	24.300	22.078	21.906	2.222
5	2011	22.800	21.665	22.016	1.135
6	2012	20.200	21.860	22.126	-1.660

Таблица 4.5 — Прогноз (периодическая модель)

Следует также отметить, что применение оценки Кресси-Хокинса дало результат лишь в случае автоматического подбора параметров. На визуальный подбор модели она никак не повлияла. Что обосновано ее назначением и показанным отсутствием выбросов в исследуемых данных.

Таким образом в результате вариограммного анализа были исследованы различные модели вариограмм, проанализированы оценки Матерона и Кресси-Хокинса, исследованы два подхода по оценке качества модели и подбору моделей и параметров. В результате применения к найденным моделям кrigинга построены наилучшая модель для описания исследуемых данных и модель для вычисления краткосрочных прогнозных значений. Следовательно можно сделать вывод о состоятельности этого подхода.

Заключение

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

В работе был проведён регрессионный анализ, в процессе которого была построена аддитивная модель временного 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 кrigинга 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 Evelyn Márcia Leão de Moraes Novo. Time series analysis of water surface temperature and heat flux components in the Itumbiara Reservoir (GO), Brazil. *Acta Limnologica Brasiliensis*, 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 БГУ им. А.С. Пушкина*, 2005.
10. А.Н. Ширяев. *Вероятность*. Наука, 1980.
11. Robert Kabacoff. *R in Action*. 2009.
12. Paul Teator. *R Cookbook (O'Reilly Cookbooks)*. O'Reilly Media, 1 edition, 2011 2011.
13. Winston Chang. *R graphics cookbook*. "O'Reilly Media, Inc. 2012.
14. Юзбашев М.М. Елисеева, И.И. *Общая теория статистики*. Москва : Финансы и статистика, 1995.
15. Duncan Cramer. *Basic statistics for social research: step-by-step calculations and computer techniques using Minitab*. Psychology Press, 1997.
16. M. G. Bulmer. *Principles of Statistics*. Dover Publications, 1979.
17. H. A. Sturges. The choice of a class interval. *American Statistical Association*, 21:65–66, 1926.
18. 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.
19. А.И. Кобзарь. *Прикладная математическая статистика*. М.: Физматлит, 2006.

20. В.Е. Гмурман. *Теория вероятностей и математическая статистика*. Москва : Высшая школа, 2003.
21. Метельский А.В. Микулик, Н.А. *Теория вероятностей и математическая статистика: Учеб. пособие*. Минск : Пион, 2002.
22. F. E. Grubbs. Sample criteria for testing outlying observations. *Ann. Math. Statistics*, 21:27–58, 1950.
23. Frank E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.
24. Стэнсфилд Р. Эддоус М. *Методы принятия решений*. Москва : Аудит, 1997.
25. Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2nd edition, 2006.
26. 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.
27. 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.
28. Eulogio Pardo-igu Zquia. 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.
29. 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.
30. Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2011.
31. Noel AC Cressie and Noel A Cassie. *Statistics for spatial data*, volume 900. Wiley New York, 1993.
32. Rudolf Dutter. On robust estimation of variograms in geostatistics. In Helmut Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods*, volume 109 of *Lecture Notes in Statistics*, pages 153–171. Springer New York, 1996.
33. Sueli Aparecida Mingoti and Gilmar Rosa. A note on robust and non-robust variogram estimators. *Rem: Revista Escola de Minas*, 61:87 – 95, 03 2008.
34. Е.А. Савельева and В.В. Демьянов. *Геостатистика: теория и практика*. Ин-т проблем безопасного развития атомной энергетики РАН. – М.: Наука, 2010.

Приложение А

Исходные данные

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

Таблица A.1 — Исходные данные.

Приложение Б

Графические материалы

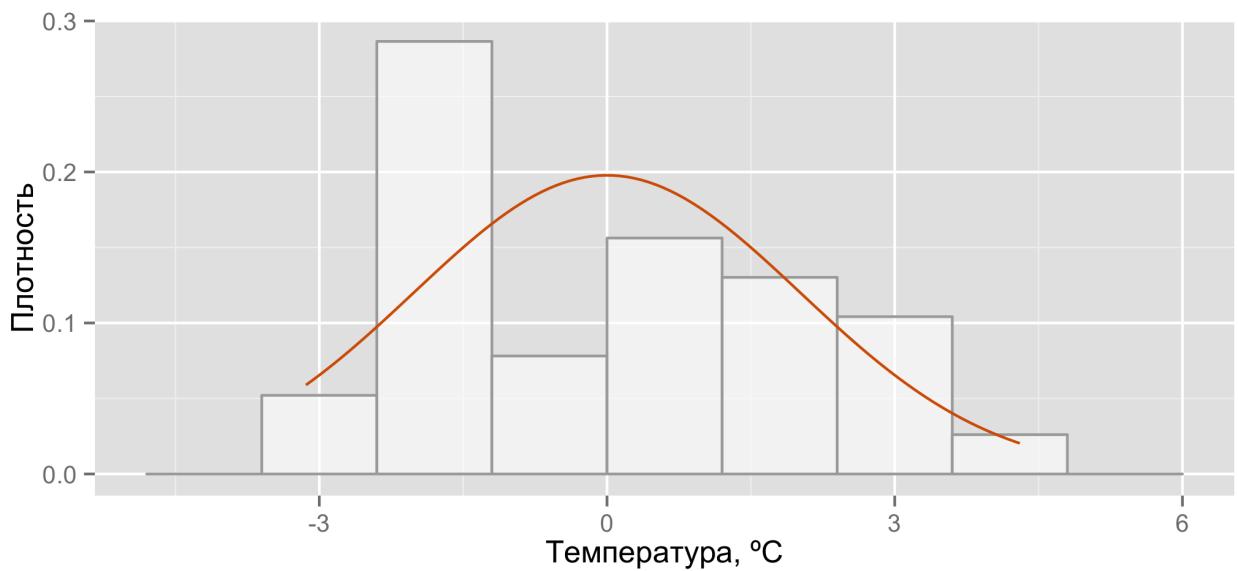


Рисунок Б.1 — Гистограмма остатков с кривой плотности нормального распределения $\mathcal{N}(19.88, 4.92)$

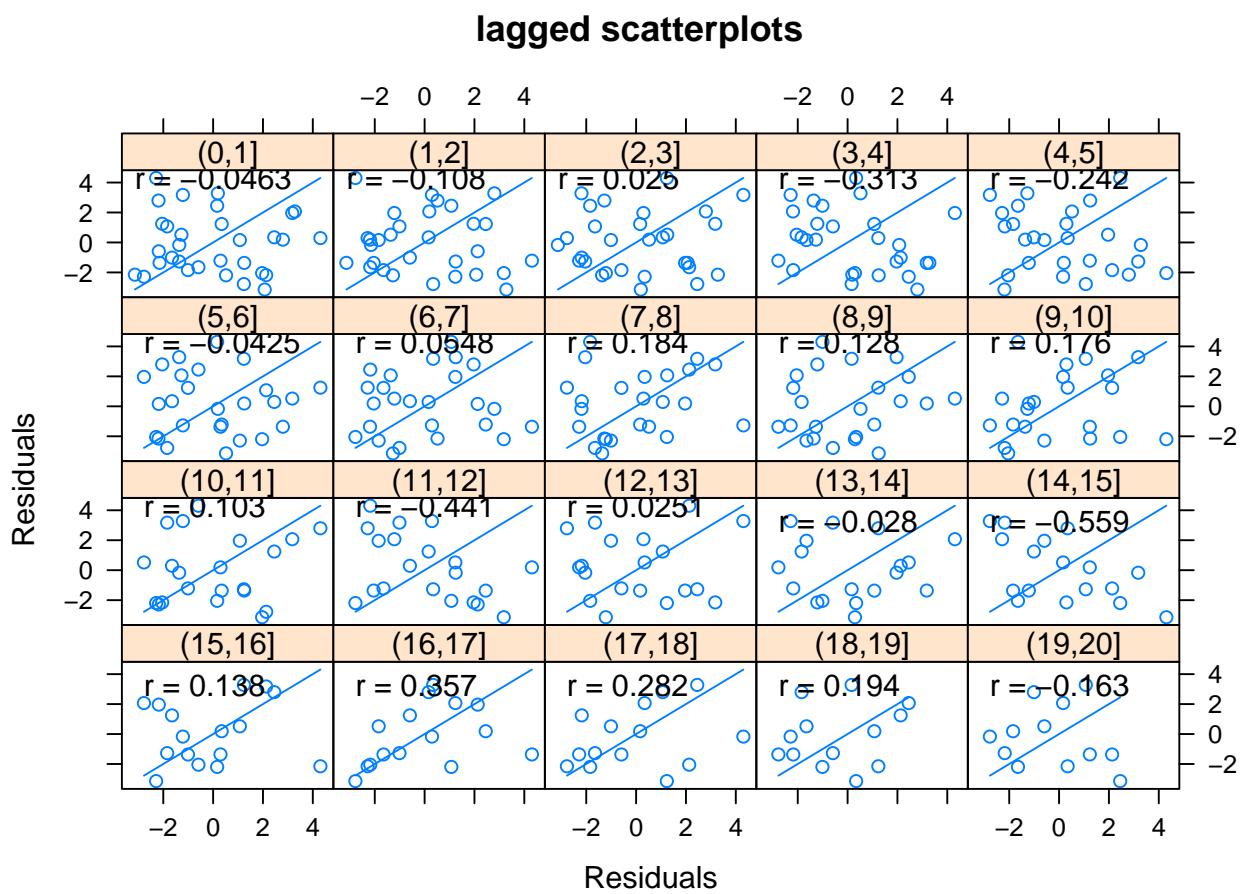


Рисунок Б.2 — Диаграмма взаимного разброса

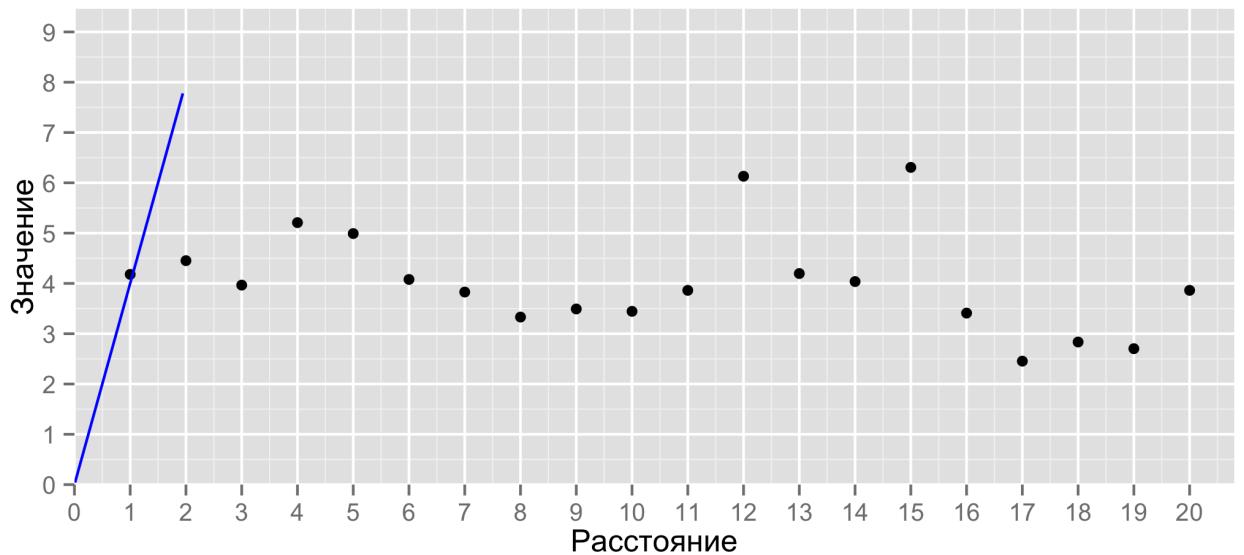


Рисунок Б.3 — Экспериментальная и теоретическая вариограмма $4 \cdot \text{Lin}(h, 0)$

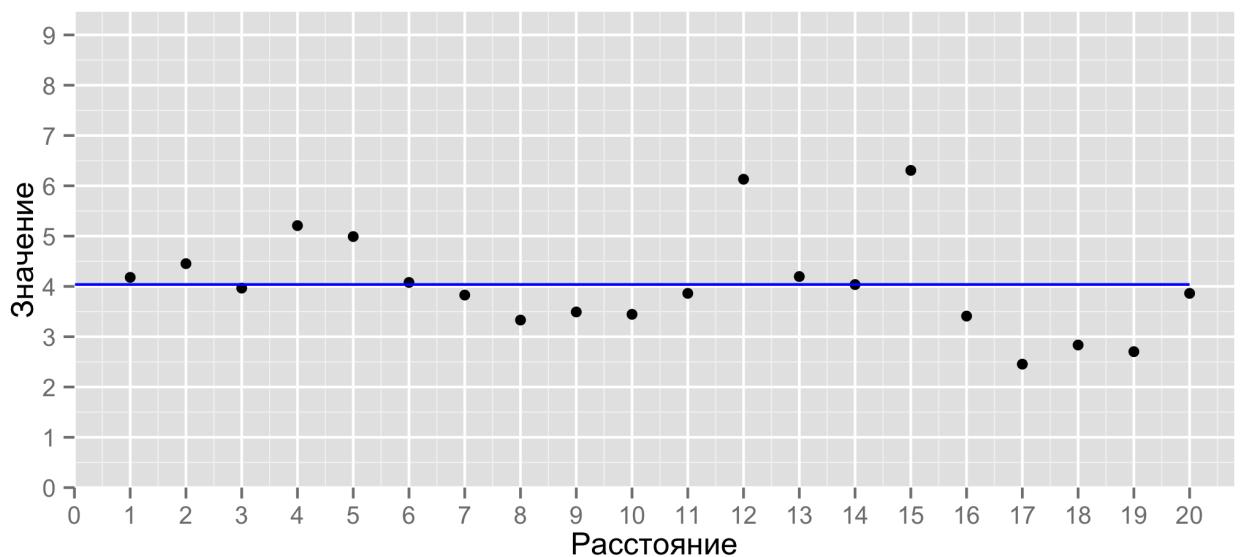


Рисунок Б.4 — Экспериментальная и теоретическая вариограмма $4.08 \cdot \text{Nug}(h)$

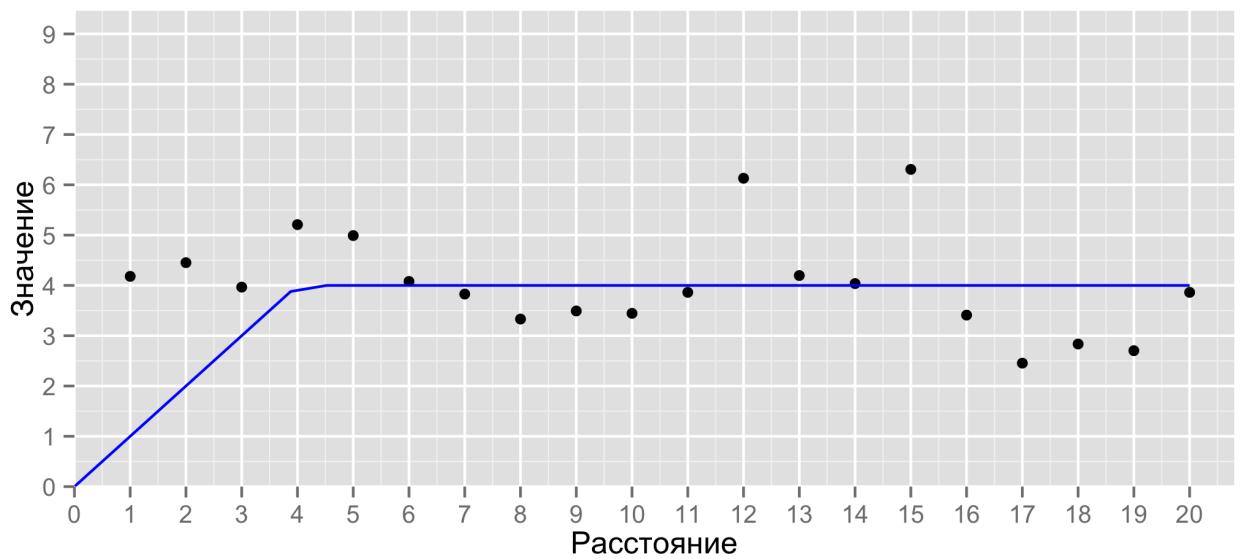


Рисунок Б.5 — Экспериментальная и теоретическая вариограмма $4 \cdot \text{Lin}(h, 4)$

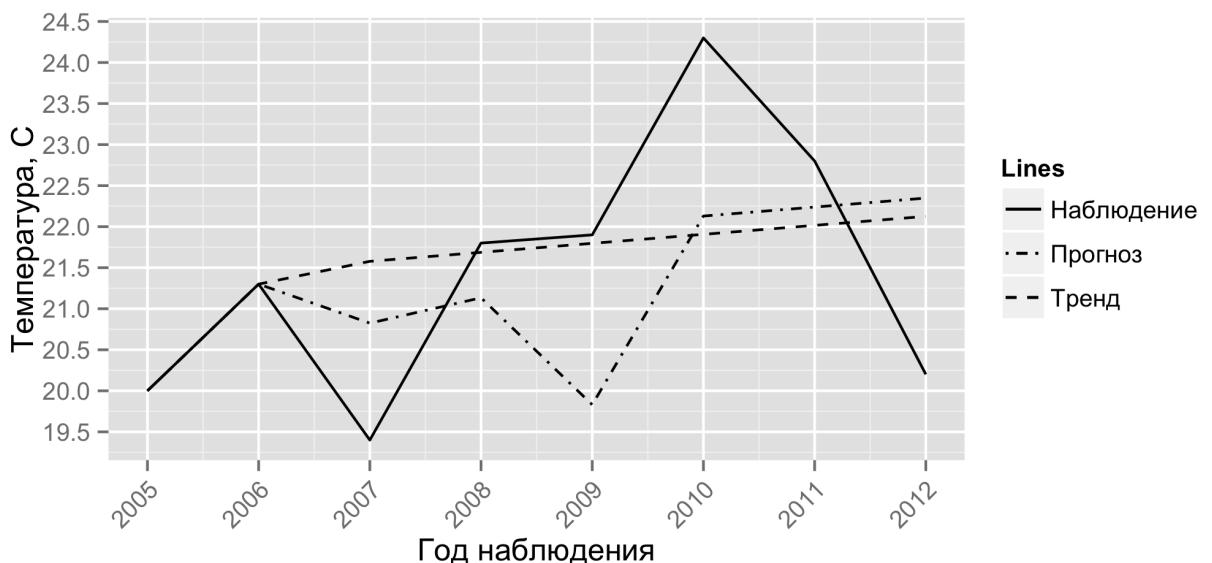


Рисунок Б.6 — Прогноз $4 \cdot \text{Lin}(h, 4)$

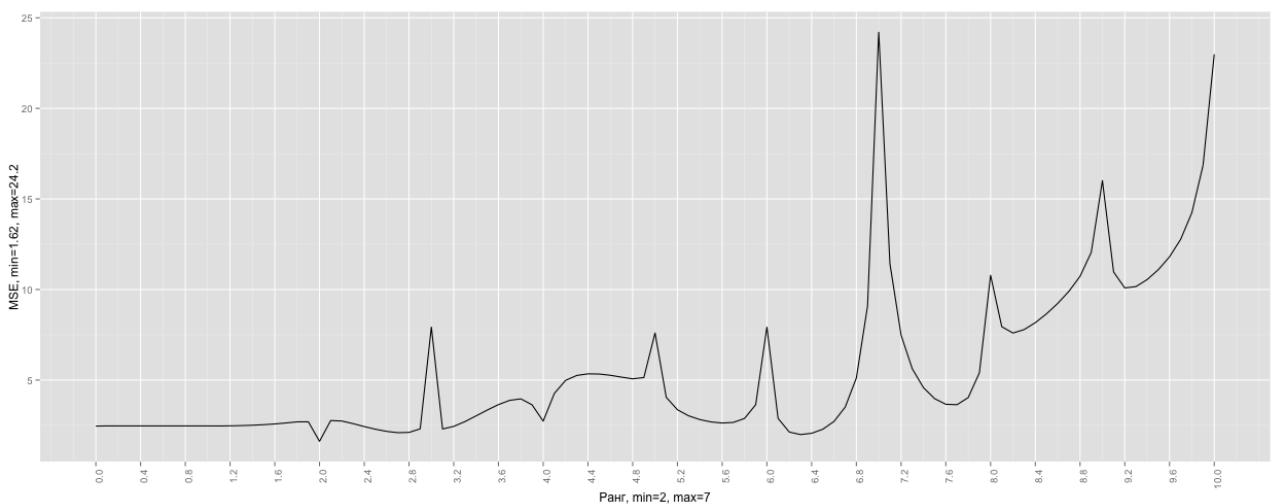


Рисунок Б.7 — Зависимость качества линейной модели от значения ранга

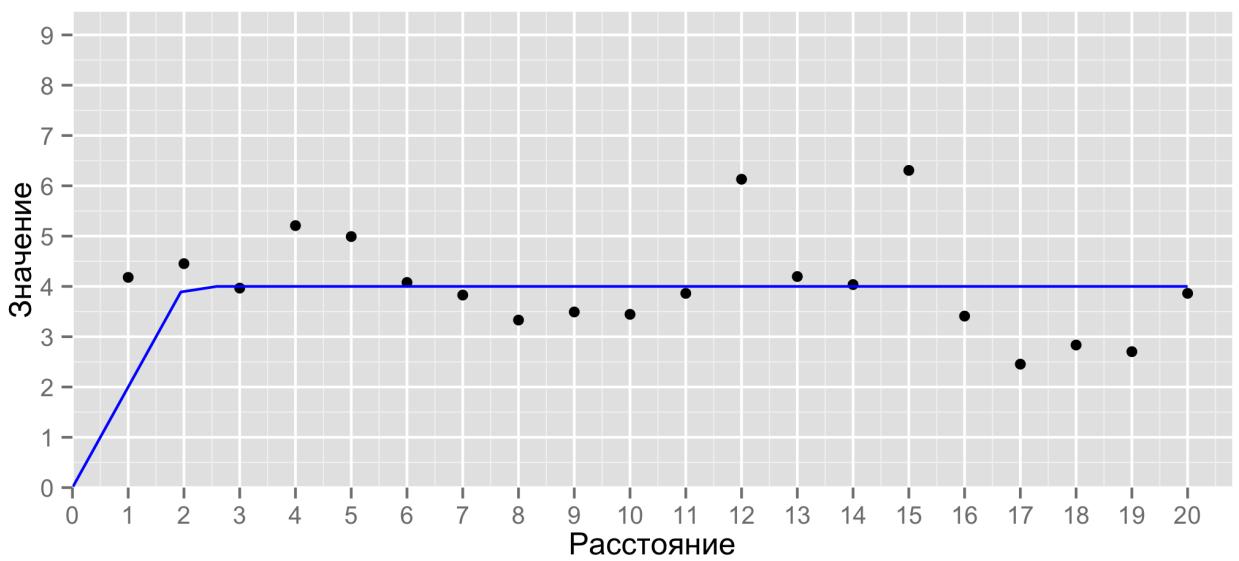


Рисунок Б.8 — Экспериментальная и теоретическая вариограмма $2 \cdot \text{Lin}(h, 2)$

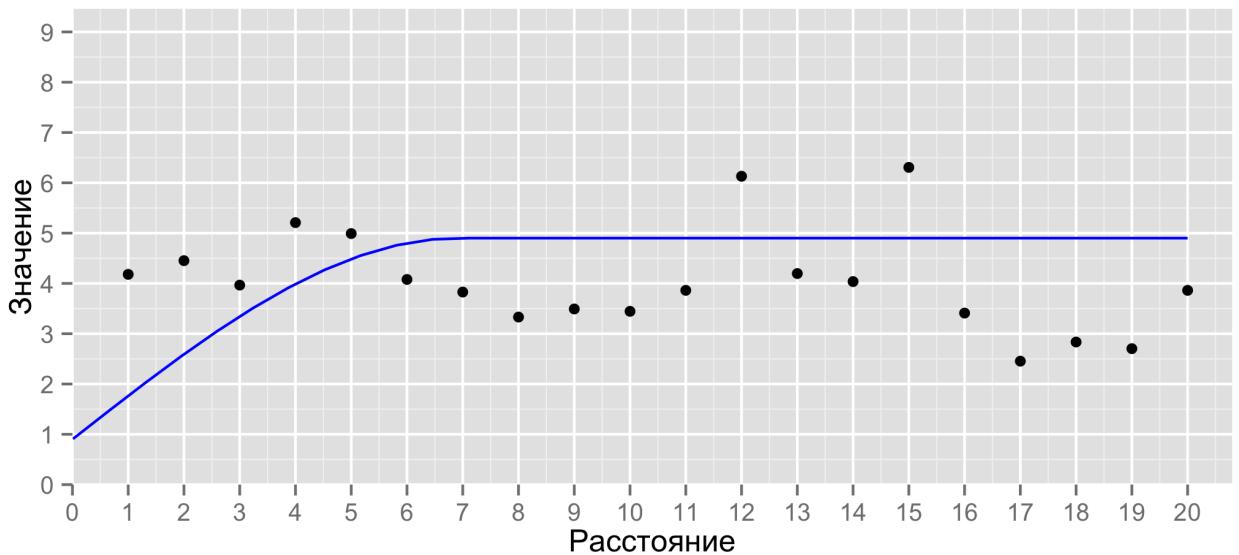


Рисунок Б.9 — Экспериментальная и теоретическая вариограмма $0.9 + 4 \cdot \text{Sph}(h, 6.9)$

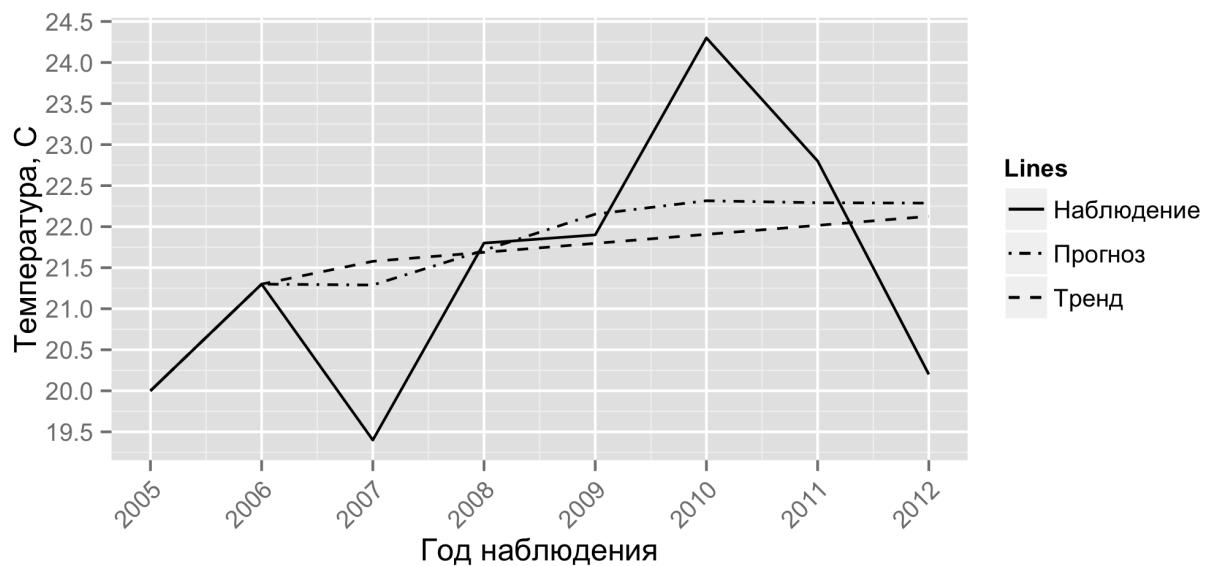


Рисунок Б.10 — Прогноз $0.9 + 4 \cdot Sph(h, 6.9)$

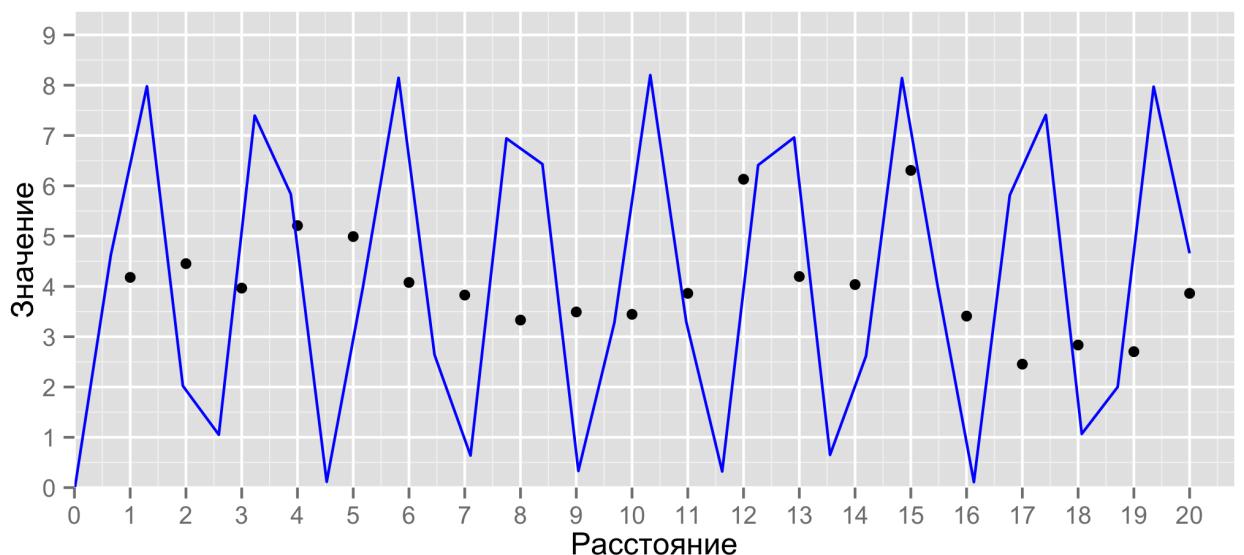


Рисунок Б.11 — Экспериментальная и теоретическая вариограмма $4 \cdot Per(h, 0.898)$

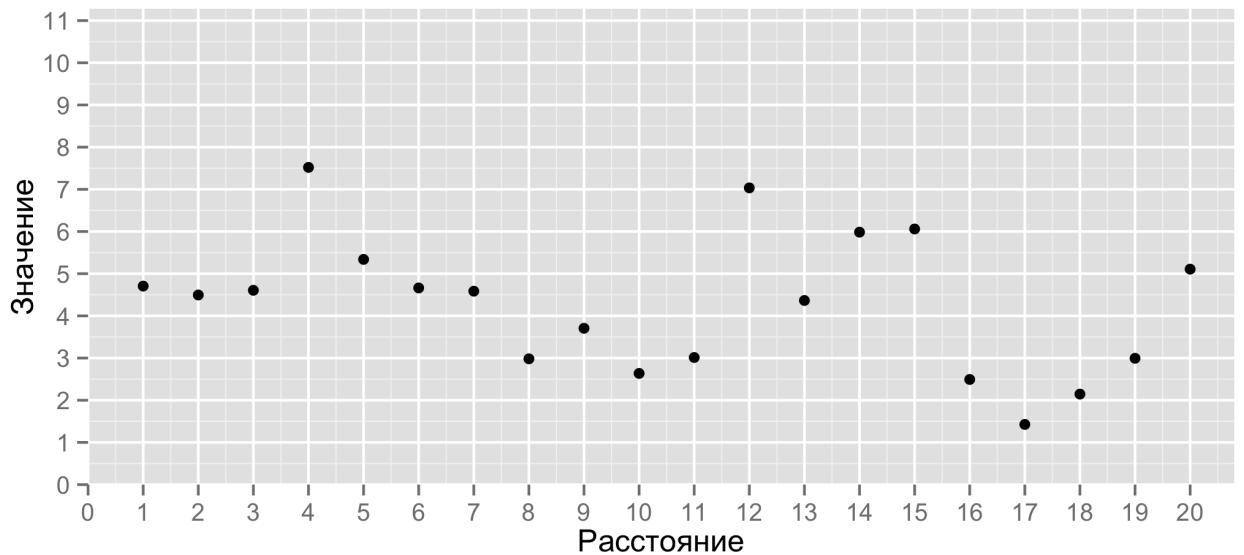


Рисунок Б.12 — Экспериментальная вариограмма (оценка Кресси-Хокингса)

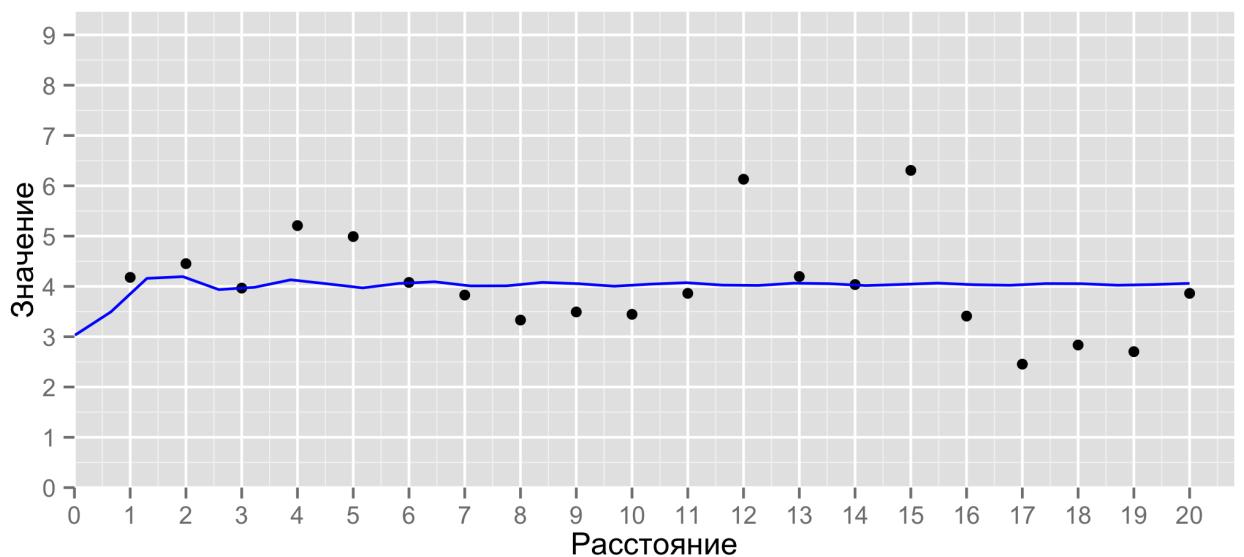


Рисунок Б.13 — Экспериментальная и теоретическая вариограмма $1.011 \cdot Wav(h, 1.14)$

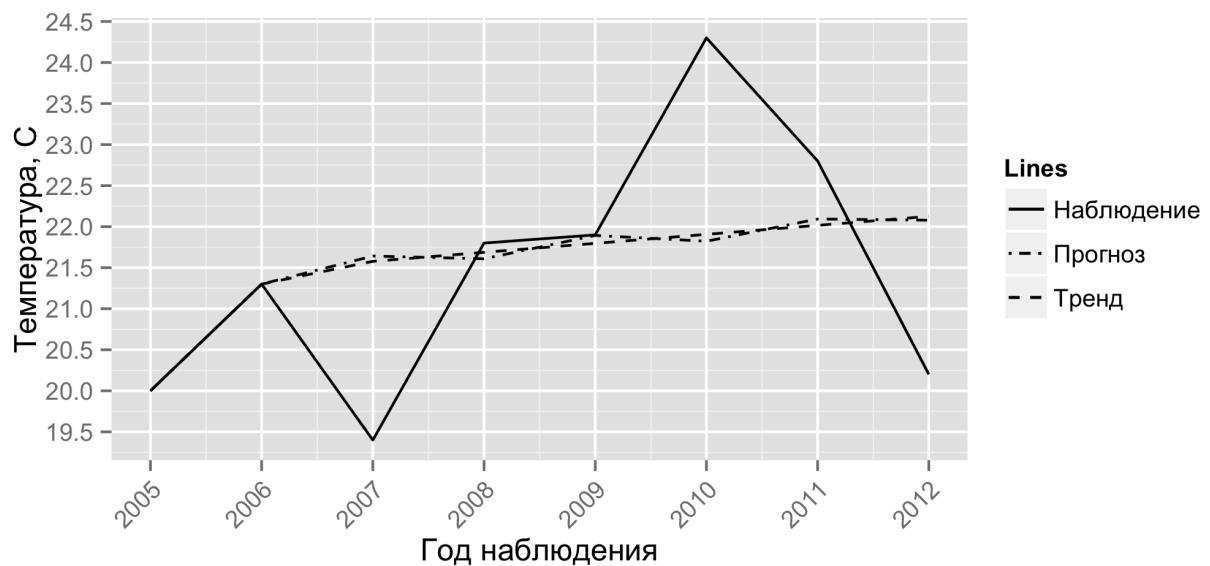


Рисунок Б.14 — Прогноз $1.011 \cdot Wav(h, 1.14)$

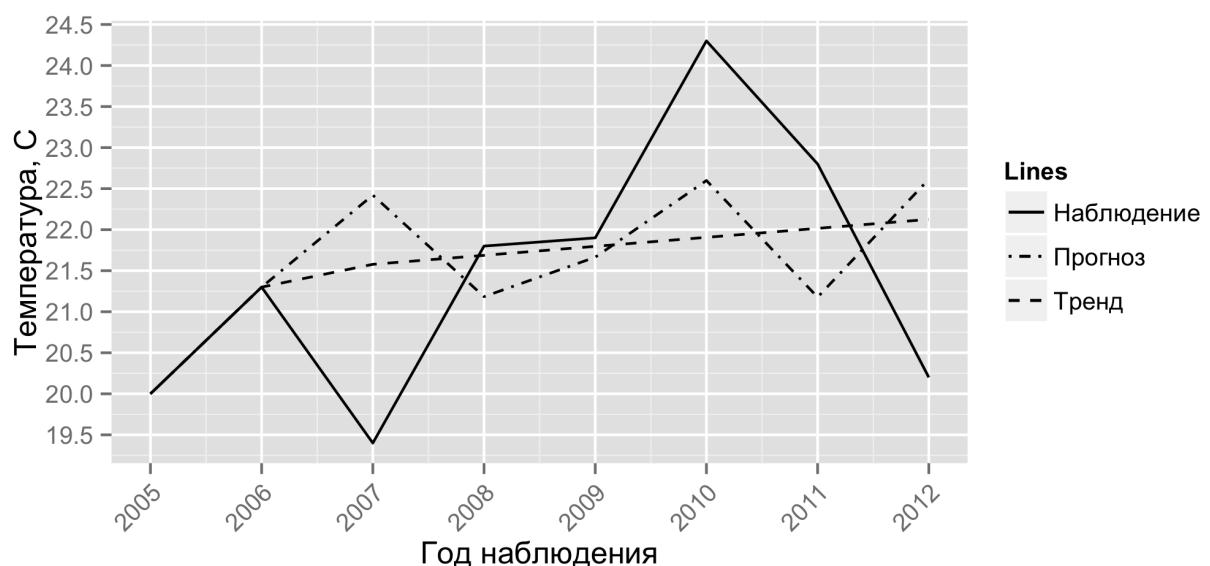


Рисунок Б.15 — Прогноз $3.46 + 0.5 \cdot Per(h, 2.67)$

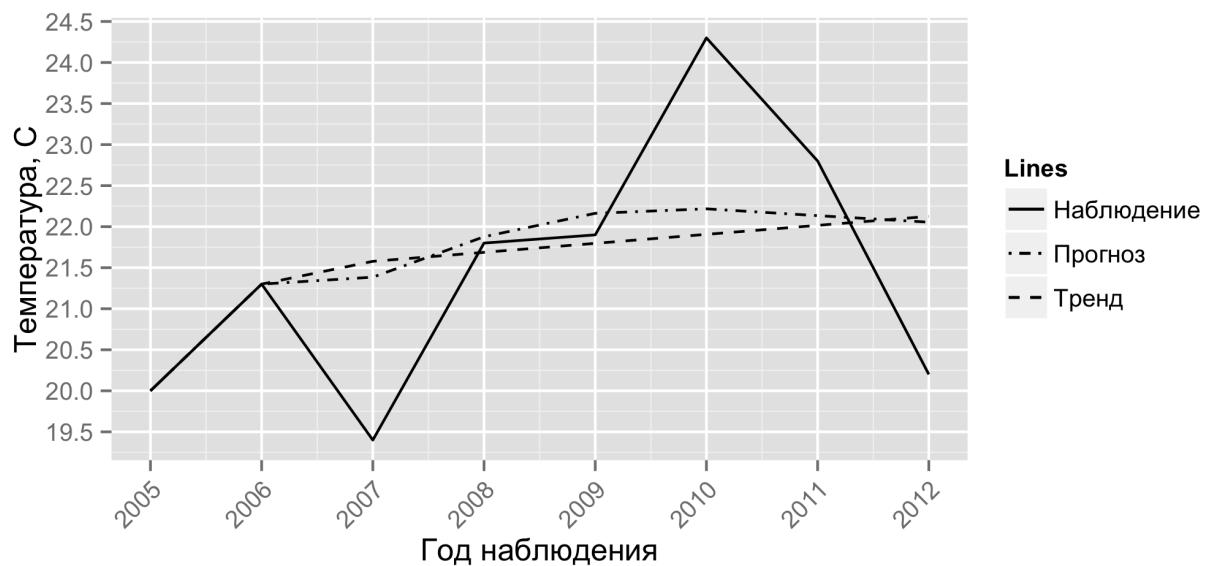


Рисунок Б.16 — Прогноз $4.11 + 1.65 \cdot \text{Wav}(h, 3.59)$

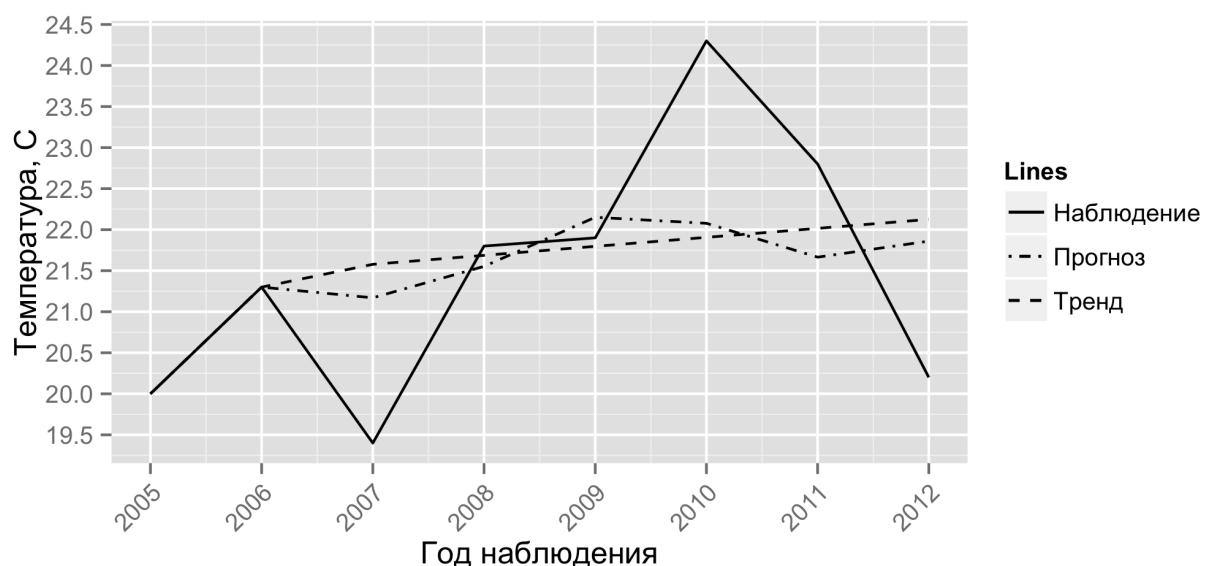


Рисунок Б.17 — Прогноз $3.8 + 0.32 \cdot \text{Per}(h, 1.3)$

Приложение В

Результаты вычислений

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

Таблица В.1 — Временной ряд остатков.

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

Таблица В.2 — Прогноз (периодическая модель)

	Год	Наблюдение	Прогноз	Тренд	Ошибка
1	2007	19.400	21.385	21.578	-1.985
2	2008	21.800	21.877	21.687	-0.077
3	2009	21.900	22.163	21.797	-0.263
4	2010	24.300	22.217	21.906	2.083
5	2011	22.800	22.134	22.016	0.666
6	2012	20.200	22.055	22.126	-1.855

Таблица В.3 — Прогноз (волнивая модель)

Приложение Г Исходный код

```

1  # Descriptive statistics
2
3  # Function for getting all descriptive statistics
4  dstats.describe <- function(data, type="", locale=FALSE, shiny=FALSE) {
5    cv <- dstats.coef.var(data)
6    stats <- c(dstats.mean(data), dstats.median(data), dstats.quartile.lower(data)
7      ,
8      dstats.quartile.upper(data), dstats.min(data), dstats.max(data),
9      dstats.range(data), dstats.quartile.range(data), dstats.variance(
10        data),
11        dstats.std.dev(data), if(!is.na(cv)){cv}, dstats.std.error(data),
12        dstats.skew(data), dstats.std.error.skew(data), dstats.kurtosis(
13          data),
14          dstats.std.error.kurtosis(data))
15
16  if(nchar(type)) {
17    dstats.write(data=data, type=type) ## TODO: need to improve — now it
18    computes two times the same things
19  }
20  if (locale) {
21    descr.row <- c("Среднее", "Медиана", "Нижний quartиль", "Верхний quartиль",
22    "Минимум", "Максимум", "Размах", "Квартильный размах",
23    "Дисперсия", "Стандартное отклонение", if(!is.na(cv)) {"Коэффициент вариации"},
24    "Стандартная ошибка", "Асимметрия", "Ошибка асимметрии",
25    "Эксцесс", "Ошибка эксцесса")
26    descr.col <- c("Значение")
27  } else {
28    descr.row <- c("Mean", "Median", "Lower Quartile", "Upper Quartile", "Range"
29      ,
30      "Minimum", "Maximum", "Quartile Range", "Variance", "Standard
31      Deviation",
32      if (!is.na(cv)) {"Coefficient of Variance"}, "Standard Error"
33      ,
34      "Skewness",
35      "Std. Error Skewness", "Kurtosis", "Std. Error Kurtosis")
36    descr.col <- c("Value")
37  }
38  if (!shiny) {
39    df <- data.frame(stats, row.names=descr.row)
40    colnames(df) <- descr.col
41  } else {
42    df <- data.frame(descr.row, sapply(stats, format, digits=2, scientific=FALSE
43      ,
44      nsmall=1))
45    colnames(df) <- c("Статистика", "Значение")
46  }

```

```

38 }   df
39 }
40
41 dstats.mean <- function(data, ...) {
42   m <- mean(data, ...)
43   if (m < .0000001) {
44     m <- 0
45   }
46   m
47 }
48
49 dstats.median <- function(data, ...) {
50   median(data, ...)
51 }
52
53 dstats.quartile.lower <- function(data, ...) {
54   quantile(data, ...) [[2]]
55 }
56
57 dstats.quartile.upper <- function(data, ...) {
58   quantile(data, ...) [[4]]
59 }
60
61 dstats.quartile.range <- function(data) {
62   dstats.quartile.upper(data) - dstats.quartile.lower(data)
63 }
64
65 dstats.min <- function(data, ...) {
66   min(data, ...)
67 }
68
69 dstats.max <- function(data, ...) {
70   max(data, ...)
71 }
72
73 dstats.range <- function(data) {
74   max(data) - min(data)
75 }
76
77 dstats.variance <- function(data, ...) {
78   var(data, ...)
79 }
80
81 dstats.std.dev <- function(data) {
82   sd(data)
83 }
84
85 dstats.coef.var <- function(data) {
86   mn <- mean(data)
87   if (abs(mn) > 1.987171e-15) {
88     (var(data) / mean(data)) * 100
89   } else
90     NA
91 }
92
93 dstats.std.error <- function(data) {
94   sd(data) / sqrt(length(data))
95 }
96
97 dstats.skew <- function(data) {

```

```

98 n <- length(data)
99 mean <- mean(data)
100 (n * sum(sapply(data, FUN=function(x){(x - mean)^3}))) / 
101 ((n - 1) * (n - 2) * dstats.std.dev(data)^3)
102 }
103
104 dstats.std.error.skew <- function(data) {
105   n <- length(data)
106   sqrt((6 * n * (n - 1)) / ((n - 2) * (n + 1) * (n + 3)))
107 }
108
109 dstats.test.skew <- function(data) {
110   dstats.skew(data) / dstats.std.error.skew(data)
111 }
112
113 dstats.kurtosis <- function(data) {
114   n <- length(data)
115   mean <- mean(data)
116   (n * (n + 1) * sum(sapply(data, FUN=function(x){(x - mean)^4}))) - 3 * (sum(
117     sapply(data, FUN=function(x){(x - mean)^2})))^2 * (n - 1) / 
118   ((n - 1) * (n - 2) * (n - 3) * dstats.variance(data)^2)
119 }
120
121 dstats.std.error.kurtosis <- function(data) {
122   n <- length(data)
123   2 * dstats.std.error.skew(data) * sqrt((n^2 - 1) / ((n - 3) * (n + 5)))
124 }
125
126 dstats.test.kurtosis <- function(data) {
127   dstats.kurtosis(data) / dstats.std.error.kurtosis(data)
128 }
129
130 dstats.write <- function (data, type) {
131   WriteDescriptiveStatistic(expression=dstats.mean(data), type=type, name="mean")
132   WriteDescriptiveStatistic(expression=dstats.variance(data), type=type, name="variance")
133   WriteDescriptiveStatistic(expression=paste(format(dstats.coef.var(data),
134     nsmall=2, digits=4), "\\"%), type=type, name="coef-var")
135   WriteDescriptiveStatistic(expression=dstats.skew(data), type=type, name="skew")
136   WriteDescriptiveStatistic(expression=dstats.kurtosis(data), type=type, name="kurtosis")
137   WriteDescriptiveStatistic(expression=dstats.test.skew(data), type=type, name="test-skew")
138   WriteDescriptiveStatistic(expression=dstats.test.kurtosis(data), type=type,
139     name="test-kurtosis")
140 }

```

Листинг Г.1: Описательные статистики

```

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 # Completes trend values up to source observation number
7 computeTrend <- function (fit, future=0) {
8   c(sapply(c(1 : (nrows + future)), FUN=function(x) fit$coefficients[[1]] + x *
9     fit$coefficients[[2]]))

```

```

10
11 # Computes prediction with passed parameters and saves all needed info and plots
12 processPrediction <- function (data, year, variogram, cressie, cutoff, name,
13   caption) {
14
15   prediction <- PredictWithKriging(data, x=ConvertYearsToNum(year), observations
16     =kObservationNum, variogram_model=variogram$var_model, nrows=nrows)
17   CrossPrediction(src$temperature, src$year, trend, prediction, name,
18     observations=kObservationNum, nrows=nrows)
19   residual <- ComputeKrigingResiduals(src$temperature, trend, prediction,
20     observations=kObservationNum, nrows=nrows)
21   mse <- MSE(residual)
22
23   prediction.compare <- data.frame("Год"=src$year[(kObservationNum + 1):nrows],
24     "Наблюдение"=src$temperature[(kObservationNum + 1):nrows],
25     "Прогноз"=prediction$var1.pred+trend[(kObservationNum + 1):nrows],
26     "Тренд"=trend[(kObservationNum + 1):nrows],
27     "Ошибка"=residual)
28   print(xtable(prediction.compare, caption=caption, label=paste0("table:", name,
29     "-prediction"), digits=c(0, 0, 3, 3, 3, 3)),
30     file=paste0("out/variogram/", name, "-prediction.tex"))
31
32   WriteCharacteristic(mse, type="variogram", name=paste0(name, "-mse"))
33
34   list(variogram=variogram, prediction=prediction, residual=residual, mse=mse)
35 }
36
37 trend <- computeTrend(sample.fit)
38 sample.residuals <- sample.fit$residuals
39
40 cutoff <- trunc(2 * kObservationNum / 3) # let it be "classical" value
41
42 # Draw H-Scatterplot
43 sample.hscat <- DrawHScatterplot(sample.residuals[1:kObservationNum])
44
45 lin.var1 <- ComputeManualVariogram(data=sample.residuals, x=sample$year, cressie
46   =FALSE, cutoff=20, model="Lin", name="lin", psill=4, range=0, nugget=0, fit=
47   FALSE)
48 lin.fit <- ComputeManualVariogram(data=sample.residuals, x=sample$year, cressie=
49   FALSE, cutoff=20, model="Lin", name="lin-fit", psill=4, range=0, nugget=0,
50   fit=TRUE)
51 lin.fit.cv <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
52   cressie=FALSE, cutoff=20, model="Lin", name="lin-fit-cv", psill=4, range=4,
53   nugget=0, fit=FALSE)
54 lin.fit.adapt <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
55   cressie=FALSE, cutoff=20, model="Lin", name="lin-fit-adapt", psill=4, range
56   =2, nugget=0, fit=FALSE)
57 lin.fit.cv.prediction <- processPrediction(data=sample.residuals, year=sample$year,
58   variogram=lin.fit.cv, cutoff=cutoff, name="lin-fit-cv", caption="Прогно
59   з (линейная модель с порогом)")
60 lin.fit.adapt.prediction <- processPrediction(data=sample.residuals, year=sample
61   $year, variogram=lin.fit.adapt, cutoff=cutoff, name="lin-fit-adapt", caption=
62   "Адаптивный прогноз (линейная модель с порогом)")
63 sph.fit.adapt <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
64   cressie=FALSE, cutoff=20, model="Sph", name="sph-fit-adapt", psill=4, range
65   =6.9, nugget=0.9, fit=FALSE)
66 sph.fit.adapt.prediction <- processPrediction(data=sample.residuals, year=sample
67   $year, variogram=sph.fit.adapt, cutoff=cutoff, name="sph-fit-adapt", caption=
68   "Адаптивный прогноз (сферическая модель)")
69 per.fit.cv <- ComputeManualVariogram(data=sample.residuals, x=sample$year,

```

```

cressie=FALSE, cutoff=20, model="Per", name="per-fit-cv", psill=4.1, range
=0.898, nugget=0.001, fit=FALSE)
49 per.fit.cv.prediction <- processPrediction(data=sample.residuals, year=sample$year,
      variogram=per.fit.cv, cutoff=cutoff, name="per-fit-cv", caption="Прогноз
      (периодическая модель)")

50
51 for.robust.only <- ComputeManualVariogram(data=sample.residuals, x=sample$year,
      cressie=TRUE, cutoff=20, model="Lin", name="robust", psill=0, range=0, nugget
      =0, fit=FALSE)

52
53 auto.class <- ComputeVariogram(data=sample.residuals, x=sample$year, name="auto-
      class-20", cressie=FALSE, cutoff=20)
54 auto.class.prediction <- processPrediction(data=sample.residuals, year=sample$year,
      variogram=auto.class, cutoff=20, name="auto-class-20", caption="Прогноз
      (волновая модель)")

55 auto.rob <- ComputeVariogram(data=sample.residuals, x=sample$year, name="auto-
      rob-20", cressie=TRUE, cutoff=20)
56 auto.rob.prediction <- processPrediction(data=sample.residuals, year=sample$year,
      , variogram=auto.rob, cutoff=20, name="auto-rob-20", caption="Прогноз (волнов
      ая модель)")

57 auto.class <- ComputeVariogram(data=sample.residuals, x=sample$year, name="auto-
      class-26", cressie=FALSE, cutoff=26)
58 auto.class.prediction <- processPrediction(data=sample.residuals, year=sample$year,
      , variogram=auto.class, cutoff=26, name="auto-class-26", caption="Прогноз
      (периодическая модель)")

59 # Compute prediction manually with choosed model ("best" what i found)
60 manual <- processPrediction(data=sample.residuals, year=sample$year, variog=
      ComputeManualVariogram, cressie=FALSE, cutoff=cutoff, name="manual", caption=
      "Прогноз (сферическая модель)")

61
62 cv.cutoff <- ComparePredictionParameters(sample.residuals, trend,
      ConvertYearsToNum(sample$year), filename="figures/variogram/auto-corr-cutoff.
      png", observations=kObservationNum, nrows=nrows, adapt=FALSE)
63 adapt.cutoff <- ComparePredictionParameters(sample.residuals, trend,
      ConvertYearsToNum(sample$year), filename="figures/variogram/auto-mse-cutoff.
      png", observations=kObservationNum, nrows=nrows)

64
65 auto.rob.adapt <- ComputeVariogram(data=sample.residuals, x=sample$year, name="
      auto-rob-5", cressie=TRUE, cutoff=5)
66 auto.rob.prediction <- processPrediction(data=sample.residuals, year=sample$year,
      , variogram=auto.rob.adapt, cutoff=5, name="auto-rob-5", caption="Прогноз (во
      лновая модель)")

67 auto.class.adapt <- ComputeVariogram(data=sample.residuals, x=sample$year, name=
      "auto-class-18", cressie=FALSE, cutoff=18)
68 auto.class.prediction <- processPrediction(data=sample.residuals, year=sample$year,
      , variogram=auto.class.adapt, cutoff=18, name="auto-class-18", caption="П
      рогноз (периодическая модель)")

69
70
71 # Compute prediction with auto fit model using classical estimation
72 classical <- processPrediction(data=sample.residuals, year=sample$year, cressie=
      FALSE, cutoff=cutoff, name="classical", caption="Прогноз (классическая оценка
      )")

73
74 # Compute prediction with auto fit model using robust (cressie) estimation
75 robust <- processPrediction(data=sample.residuals, year=sample$year, cressie=
      TRUE, cutoff=cutoff, name="robust", caption="Прогноз (робастная оценка)")

76
77 models.comparison <- CompareClassicalModels(manual$variogram, classical$variogram,
      filename="figures/variogram/models-comparison.png")

```

```

78 # Find best cutoff parameters
79 cutoff <- ComparePredictionParameters(sample.residuals, trend, ConvertYearsToNum
80   (sample$year), filename="figures/variogram/parameter-comparison.png",
81   observations=kObservationNum, nRows=nRows)
82 manual.best <- processPrediction(data=sample.residuals, year=sample$year,
83   variog=ComputeManualVariogram, cressie=FALSE, cutoff=cutoff$manual, name="manual-best",
84   caption="Наилучший прогноз (сферическая модель)")
85 classical.best <- processPrediction(data=sample.residuals, year=sample$year,
86   cressie=FALSE, cutoff=cutoff$classical, name="classical-best", caption="Наилучший прогноз (классическая оценка)")
87 robust.best <- processPrediction(data=sample.residuals, year=sample$year,
88   cressie=TRUE, cutoff=cutoff$robust, name="robust-best", caption="Наилучший прогноз (робастная оценка)")

```

Листинг Г.2: Вариограммный анализ

```

1 # This function automatically fits a variogram to input_data
2 autofitVariogram = function(formula, input_data,
3   test_models = c("Nug", "Exp", "Sph", "Gau", "Cir", "Lin", "Bes", "Pen", "Per",
4     "Wav", "Hol", "Log", "Spl"),
5   kappa=c(0.05, seq(0.2, 2, 0.1), 5, 10), GLS.model=NA,
6   fix.values=c(NA,NA,NA), start_vals=c(NA,NA,NA),
7   cutoff, width=1, cressie, verbose=FALSE, ...) {
8
9   # If you specify a variogram model in GLS.model the Generalised Least Squares
10  # sample variogram is constructed
11  if(!is(GLS.model, "variogramModel")) {
12    experimental_variogram = variogram(formula, input_data, cutoff=cutoff, width=
13      width, cressie=cressie, ...)
14  } else {
15    g = gstat(NULL, "bla", formula, input_data, model=GLS.model, set=list(gls=1))
16    experimental_variogram = variogram(g, cutoff=cutoff, width=width, cressie=
17      TRUE, ...)
18
19  # set initial values
20  if(is.na(start_vals[1])) { # Nugget
21    initial_nugget = min(experimental_variogram$gamma)
22  } else {
23    initial_nugget = start_vals[1]
24  }
25  if(is.na(start_vals[2])) { # Range
26    diagonal = spDists(t(bbox(input_data)))[1,2] # 0.35 times the length of the
27    # central axis through the area
28    initial_range = 0.1 * diagonal # 0.10 times the length of the central axis
29    # through the area
30  } else {
31    initial_range = start_vals[2]
32  }
33  if(is.na(start_vals[3])) { # Sill
34    initial_sill = mean(c(max(experimental_variogram$gamma), median(experimental
35    # Determine what should be automatically fitted and what should be fixed
36    # Nugget

```

```

36 | if(!is.na(fix.values[1])) {
37 |   fit_nugget = FALSE
38 |   initial_nugget = fix.values[1]
39 | } else {
40 |   fit_nugget = TRUE
41 | }
42 |
43 | # Range
44 | if(!is.na(fix.values[2])) {
45 |   fit_range = FALSE
46 |   initial_range = fix.values[2]
47 | } else {
48 |   fit_range = TRUE
49 | }
50 |
51 | # Partial sill
52 | if(!is.na(fix.values[3])) {
53 |   fit_sill = FALSE
54 |   initial_sill = fix.values[3]
55 | } else {
56 |   fit_sill = TRUE
57 | }
58 |
59 | getModel <- function(psill, model, range, kappa, nugget, fit_range, fit_sill,
60 |   fit_nugget) {
61 |   if(model == "Pow") {
62 |     if(is.na(start_vals[1])) nugget = 0
63 |     if(is.na(start_vals[2])) range = 1      # If a power mode, range == 1 is a
64 |           better start value
65 |     if(is.na(start_vals[3])) sill = 1
66 |   }
67 |   if(model == "Nug") {
68 |     if(is.na(start_vals[2])) range = 0
69 |
70 |     obj = try(fit.variogram(experimental_variogram,
71 |       model = vgm(psill=psill, model=model, range=range,
72 |         nugget=nugget, kappa = kappa),
73 |       fit.ranges = c(fit_range), fit.sills = c(fit_nugget, fit_sill),
74 |       debug.level=0, fit.method = 6),
75 |       silent=TRUE)
76 |     if("try-error" %in% class(obj)) {
77 |       #print(traceback())
78 |       if (verbose) {
79 |         warning("An error has occurred during variogram fitting. Used:\n",
80 |             "\tnugget:\t", nugget,
81 |             "\n\tmodel:\t", model,
82 |             "\n\tpsill:\t", psill,
83 |             "\n\trange:\t", range,
84 |             "\n\tkappa:\t", ifelse(kappa == 0, NA, kappa),
85 |             "\n\tas initial guess. This particular variogram fit is not taken into\n\taccount. \nGstat error:\n", obj)
86 |       }
87 |       return(NULL)
88 |     } else return(obj)
89 |
90 |   # Automatically testing different models, the one with the smallest sums-of-
91 |   squares is chosen
92 |   SSerr_list = c()

```

```

92 vgm_list = list()
93 counter = 1
94
95 for(m in test_models) {
96   if(m != "Mat" && m != "Ste") {           # If not Matern and not Stein
97     model_fit = getModel(initial_sill - initial_nugget, m, initial_range,
98                           kappa = 0, initial_nugget, fit_range, fit_sill, fit_nugget)
99   if(!is.null(model_fit)) { # skip models that failed
100     vgm_list [[counter]] = model_fit
101     SSerr_list = c(SSerr_list, attr(model_fit, "SSErr"))
102   }
103   counter = counter + 1
104 } else {                                     # Else loop also over kappa values
105   for(k in kappa) {
106     model_fit = getModel(initial_sill - initial_nugget, m, initial_range, k,
107                           initial_nugget, fit_range, fit_sill, fit_nugget)
108     if(!is.null(model_fit)) {
109       vgm_list [[counter]] = model_fit
110       SSerr_list = c(SSerr_list, attr(model_fit, "SSErr"))
111     }
112   }
113 }
114
115 # Check for negative values in sill or range coming from fit.variogram
116 # and NULL values in vgm_list, and remove those with a warning
117 strange_entries = sapply(vgm_list, function(v) any(c(v$psill, v$range) < 0) |
118   is.null(v))
119 if(any(strange_entries)) {
120   if(verbose) {
121     print(vgm_list [strange_entries])
122     cat("^^^ ABOVE MODELS WERE REMOVED ^^^\n\n")
123   }
124   SSerr_list = SSerr_list [!strange_entries]
125   vgm_list = vgm_list [!strange_entries]
126 }
127
128 if(verbose) {
129   cat("Selected:\n")
130   print(vgm_list [[which.min(SSerr_list)]])
131   cat("\nTested models, best first:\n")
132   tested = data.frame("Tested models" = sapply(vgm_list, function(x) as.
133     character(x[2,1])),
134     kappa = sapply(vgm_list, function(x) as.character(x[2,4])),
135     "SSerror" = SSerr_list)
136   tested = tested[order(tested$SSerror), ]
137   print(tested)
138 }
139
140 result = list(exp_var = experimental_variogram, var_model = vgm_list [[which.
141   min(SSerr_list)]], sserr = min(SSerr_list, na.rm=TRUE))
142 class(result) = c("autofitVariogram", "list")
143
144 return(result)
145 }
```

Листинг Г.3: Автоматический подбор моделей