

Белорусский государственный университет

Кафедра теории вероятностей и математической статистики

Утверждаю

Заведующий кафедрой _____ Труш Н.Н.

Дата

ЗАДАНИЕ НА ДИПЛОМНУЮ РАБОТУ

Обучающемуся(студенту) *Павлову А.С..*

1. Тема дипломной работы

Анализ и прогнозирование гидрологических данных

Утверждена приказом ректора БГУ от _____ №_____

2. Исходные данные к дипломной работе

- (а) Cressie N. Statistics for Spatial Data. — New York. — Wiley, 1991. — 900 р
- (б) Савельев А.А. Геостатистический анализ данных в экологии и природопользовании (с применением пакета R) / Савельев А.А., Мухарамова С.С., Пилюгин А.Г., Чижикова Н.А. — Казань: Казанский университет, 2012 — 120 с.
- (в) Robert H. Shumway, David S. Stoffer. Time series and Its Applications: With R Examples (Springer Texts in Statistics). Springer Science+Business Media, LLC 2011, 3d edition, 2011. — 576 р.
- (г) Paul Teator. R Cookbook (O'Reilly Cookbooks). O'Reilly Media, 1 edition, 2011. — 438 р.
- (д) База данных характеристик водной системы озера Баторино (Нарочанская биологическая станция им. Г.Г.Винберга).

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

С использованием среды программирования R осуществить:

- (а) предварительный статистический анализ гидроэкологических данных озера Баторино.
- (б) вариограммный анализ временного ряда: построение оценок семивариограммы, подбор моделей семивариограммы.

- (в) исследование статистических свойств оценки вариограммы гауссского случайного процесса.
- (г) прогнозирование значений временного ряда с помощью интерполяционного метода кригинг. Исследование точности прогноза в зависимости от оценки вариограммы и модели вариограммы, лежащих в основе метода кригинг.
4. Перечень графического материала (с точным указанием обязательных чертежей и графиков)
5. Консультанты по дипломной работе с указанием относящихся к ним разделов *Цеховая Т.В.*
6. Примерный график выполнения дипломной работы
- 31 марта 2015 г. промежуточный отчет
 - 28 апреля 2015 г. промежуточный отчет
 - 12 мая 2015 г. доклад о проделанной работе
7. Дата выдачи задания _____
8. Срок сдачи законченной дипломной работы _____

Руководитель _____ Т.В. Цеховая

Подпись обучающегося _____

Дата

Реферат

Дипломная работа, 57 страниц, 20 рисунков, 7 таблиц, 35 источников, 4 приложения

ВРЕМЕННОЙ РЯД, ПРОГНОЗИРОВАНИЕ, R, ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ, КОРРЕЛЯЦИОННЫЙ АНАЛИЗ, РЕГРЕССИОННЫЙ АНАЛИЗ, ВАРИОГРАММА, КРИГИНГ, КРОСС-ВАЛИДАЦИЯ.

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

Цель работы — с помощью современного языка программирования R осуществить анализ, обработку и прогнозирование реального временного ряда.

В процессе работы реализовано веб-приложение, позволяющее решать класс аналогичных поставленной задач. В данном приложении вычислены и проанализированы описательные статистики, подобран закона распределения, проведены корреляционный и регрессионный анализы, исследован ряд остатков, построены и проанализированы различные модели вариограмм и на их основе вычислены прогнозные значения временного ряда.

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

Реализованное программное обеспечение и предложенные алгоритмы могут использоваться для решения задач, аналогичных рассматриваемой в работе.

Данная работа может быть продолжена для получения модели, более точно описывающей поведение исходного временного ряда. Программное обеспечение и алгоритмы могут быть усовершенствованы в процессе дальнейших исследований, для решения других задач.

Рэферат

Дыпломная работа, 57 старонак, 20 малюнкаў, 7 табліц, 35 крыніц, 4 прыкладання

ЧАСОВЫ ШЭРАГ, ПРАГНАЗАВАННЕ, R, АПІСАЛЬНЫЯ СТАТЫСТИКІ, КАРЭЛЯЦЫЙНЫ АНАЛІЗ, РЭГРЕСІЙНЫ АНАЛІЗ, ВАРЫЯГРАММА, КРЫГІНГ, КРОСС-ВАЛІДАЦЫЯ

Аб'ектам даследавання з'яўляюцца назіранні за тэмпературай вады ў возеры Баторына ў перыяд з 1975 па 2012 гг.

Мэта працы — з дапамогай сучаснай мовы праграмавання R ажыццяўіць аналіз і прагназаванне рэальнага часовага шэрага.

У працэсе працы рэалізаваны вэб-дадатак, якое дазваляе вырашаць клас аналагічных пастаўленай задач. У дадзеным дадатку вылічаны і прааналізаваны апісальныя статыстыкі, падабраны закон размеркавання, праведзен карэліяцыйны і рэгресійны аналізы, даследаван шэраг рэшткаў, пабудаваны і прааналізаваны розныя мадэлі варыяграмм і на іх аснове вылічаны прагнозныя значэнні часовага шэрага.

Атрыманыя вынікі могуць быць выкарыстаны для далейшых даследаванняў у розных прыкладных галінах навукі: біялогіі, хіміі, гідралогіі, — а таксама, для аналізу экалагічнай сітуацыі ў Нарачанскім парку і іншых рэгіёнах.

Рэалізаванае праграмнае забеспечэнне і прапанаваныя алгарытмы могуць выкарыстоўвацца для вырашэння задач, аналагічных разгледзенай у працы.

Дадзеная праца можа быць працягнутая для атрымання мадэлі, больш дакладна апісвае паводзіны зыходнага часовага шэрагу. Праграмнае забеспечэнне і алгарытмы могуць быць удасканалены ў працэсе далейшых даследаванняў, для вырашэння іншых задач.

Abstract

Bachelor's thesis, 57 pages, 20 figures, 7 tables, 35 sources, 4 appendices.

TIME SERIES, PREDICTION, R, DESCRIPTIVE STATISTICS,
CORRELATIONAL ANALYSIS, REGRESSION ANALYSIS,
VARIOGRAMM, KRIGING, CROSS-VALIDATION.

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

Research purpose — with help of modern programming language R perform analysis and forecasting real time series.

During the research was implemented rich web-application that allows to solve problems similar to researched within current thesis. With help of implemented application and R programming language were computed and analysed descriptive statistics, was performed distribution analysis and fitting, were conducted correlation and regression analysis, was performed research of residual time series, variogram models and based on them time series 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.

Implemented web-application and suggested algorithms could be used in case of solving problems with similar with this research problem.

This research could be continued in case of getting model that will be more accurate in describing source time series. Software and algorithms that were obtained during the research could be enhanced during further research for solving different problems.

Содержание

Введение	6	
1 Вспомогательные определения	9	
1.1 Случайный процесс и его основные характеристики	9	
1.2 Вариограмма случайного процесса	10	
2 Оценка вариограммы гауссовского случайного процесса	12	
2.1 Первые два момента оценки вариограммы	12	
2.2 Асимптотическое поведение моментов второго порядка оценки вариограммы	14	
3 Обзор реализованного программного обеспечения	22	
3.1 Модуль первичного анализа	22	
3.2 Модуль анализа остатков	24	
3.3 Модуль вариограммного анализа	25	
4 Анализ временного ряда в среде R	30	
4.1 Детерминированные методы	30	
4.1.1 Описательные статистики и первичный анализ данных	30	
4.1.2 Корреляционный анализ	35	
4.1.3 Регрессионный анализ	37	
4.1.4 Анализ остатков	41	
4.2 Геостатистические методы	43	
4.2.1 Визуальный подход	44	
4.2.2 Автоматический подход	51	
Заключение	56	
Список использованных источников	58	
Приложение А	Исходные данные	61
Приложение Б	Графические материалы	62
Приложение В	Результаты вычислений	68
Приложение Г	Код программы	70

Введение

Работа посвящена обработке, исследованию и статистическому анализу реального временного ряда. 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, собранной в процессе каких-либо наблюдений, не всегда позволяет раскрыть те или иные причины и следствия, имеющие место в конкретном случае. Для выявления всех скрытых закономерностей и свойств объекта, за которым проводилось наблюдение, необходимо выполнить 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 вышеизложенными 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 раз-

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

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

В [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 статье [5] автор проводит исследование на предмет выявления антропогенного влияния на качество воды в крупнейшем озере в Гондурасе — Йоджоа.

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

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

- **R** является специализированным языком программирования для статистической обработки данных
- На сегодняшний день **R** — один из самых популярных в статистической среде инструментов анализа данных, имеющий широкую пользовательскую аудиторию, развитую систему поддержки
- Пакет постоянно развивается и дополняется современнейшими средствами, моделями и алгоритмами

- Бесплатен, свободно распространяется и доступен для всех популярных операционных систем
- Обладает развитыми возможностями для работы с графикой

Глава 1

Вспомогательные определения

1.1 Случайный процесс и его основные характеристики

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:

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

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

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

Определение 1.9. Нормированной корреляционной функцией называется функция вида

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

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

Определение 1.10. 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. $R(t_1, t_2) = R(t_1 - t_2), t_1, t_2 \in \mathbb{T}$.

Определение 1.11. 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 1.1. 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 случайного процесса

Определение 1.12. 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 1.13. 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 1.2. Если $X(t)$, $t \in \mathbb{T}$, стационарный в широком смысле случайный процесс с ковариационной функцией $R(t)$, $t \in \mathbb{T}$, и семивариограммой $\gamma(t)$, $t \in \mathbb{T}$, то

$$\gamma(t) = R(0) - R(t), \quad t \in \mathbb{T}.$$

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 гауссовского случайного процесса

Рассмотрим стационарный в широком смысле гауссовский случайный процесс с дискретным временем $X(t)$, $t \in \mathbb{Z}$, нулевым математическим ожиданием, постоянной дисперсией и неизвестной вариограммой $2\gamma(h)$, $h \in \mathbb{Z}$.

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:

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

$$\begin{aligned} cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= \frac{2}{(n-h_1)(n-h_2)} \times \\ &\times \sum_{t=1}^{n-h_1} \sum_{s=1}^{n-h_2} (\gamma(t-h_2-s) + \gamma(t+h_1-s) - \gamma(t-s) - \gamma(t+h_1-s-h_2))^2, \end{aligned} \quad (2.3)$$

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

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 **несмешённой** оценкой вариограммы.

Найдём второй момент оценки вариограммы.

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

Из определения 1.9 нормированной ковариационной функции получаем, что

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

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

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

Далее воспользуемся леммой 1 из [9]:

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

Воспользовавшись леммой 3 из [9], получаем соотношение (2.3):

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

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 моментов второго порядка оценки вариограммы

Исследуем асимптотическое поведение моментов второго порядка оценки (2.1).

Теорема 2.2. *Если имеет место соотношение*

$$\sum_{h=-\infty}^{+\infty} |\gamma(h)| < +\infty, \quad (2.6)$$

то

$$\lim_{n \rightarrow \infty} (n - \min\{h_1, h_2\}) cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \\ = 2 \sum_{m=-\infty}^{+\infty} \gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2, \quad (2.7)$$

$$\lim_{n \rightarrow \infty} (n - h) V\{2\tilde{\gamma}(h)\} = 2 \sum_{m=-\infty}^{+\infty} \gamma(m-h) + \gamma(m+h) - 2\gamma(m))^2. \quad (2.8)$$

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

Доказательство. В (2.3) сделаем следующую замену переменных

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

Получим

$$cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \frac{2}{(n-h_1)(n-h_2)} \times \\ \times \sum_{t=1}^{n-h_1} \sum_{t-m=1}^{n-h_2} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \quad (2.9)$$

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

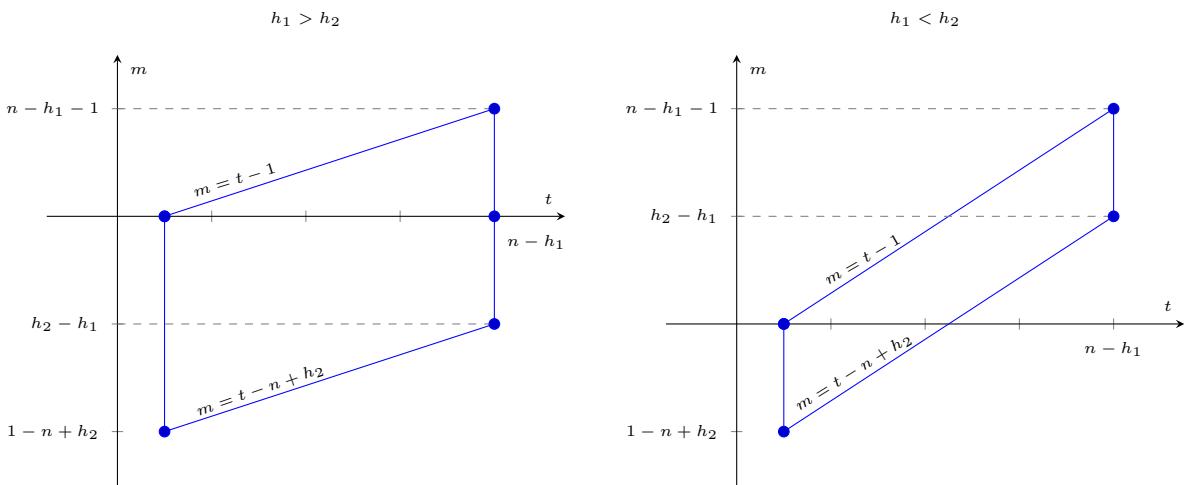


Рисунок 2.1 — Области суммирования после замены переменных

Рассмотрим первый случай: $h_1 > h_2$. Поменяем порядок суммирования в (2.9).

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

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:

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

Раскроем скобки под знаками сумм:

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

Приведём подобные:

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

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

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

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

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

$$\begin{aligned}
& cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} = \\
& = \frac{2}{n-h_2} \left(\sum_{m=1-n+h_2}^{n-h_1-1} (\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 - \right. \\
& \left. - \frac{2}{n-h_1} \sum_{m=1}^{n-h_1-1} m(\gamma(m-h_2) + \gamma(m+h_1) - \gamma(m) - \gamma(m+h_1-h_2))^2 \right). \\
& \quad (2.10)
\end{aligned}$$

Аналогично для случая $h_1 < h_2$:

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

Выражение под знаком суммы не зависит от t :

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

Вынесем $n - h_2$ из каждого слагаемого:

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

Раскроем скобки под знаками сумм:

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

Приведём подобные:

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

$$\times (\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2).$$

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

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

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

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

Далее, для доказательства (2.7) оценим разность, используя условие (2.6), выражение (2.10) и лемму Кронекера [10]:

$$\begin{aligned} & |(n - h_2)cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} - \\ & - 2 \sum_{m=-\infty}^{+\infty} (\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2| \leq \\ & \leq 2 \sum_{m=-\infty}^{n+h_2} |\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2)|^2 + \\ & + 2 \sum_{m=n-h_1}^{+\infty} |\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2)|^2 + \\ & + \frac{2}{n - h_1} \sum_{m=0}^{n-h_1-1} |\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2)|^2 \rightarrow 0, \end{aligned} \tag{2.12}$$

при $n \rightarrow \infty$.

Рассуждая аналогично, в силу сходимости ряда (2.6), выражения (2.11) и леммы Кронекера [10], получаем оценку разности для случая $h_1 < h_2$

$$\begin{aligned}
& |(n - h_1)cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} - \\
& - 2 \sum_{m=-\infty}^{+\infty} (\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2))^2| \leq \\
& \leq 2 \sum_{m=-\infty}^{n+h_2} |\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2)|^2 + \\
& + 2 \sum_{m=n-h_1}^{+\infty} |\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2)|^2 + \\
& + \frac{2}{n - h_2} \sum_{m=-n+h_2+1}^0 |m| |\gamma(m - h_2) + \gamma(m + h_1) - \gamma(m) - \gamma(m + h_1 - h_2)|^2 \rightarrow 0,
\end{aligned} \tag{2.13}$$

при $n \rightarrow \infty$.

Тогда, объединяя вместе полученные в (2.12) и (2.13) результаты, получаем требуемое предельное соотношение (2.7):

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

Нетрудно видеть, что если в (2.7) положить $h_1 = h_2 = h$, то получаем равенство для дисперсии оценки вариограммы (2.8). 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 2.2 следует соотношение*

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

Замечание 2.1. В силу показанной в теореме 2.1 несмещённости оценки и вышеприведённого следствия получаем, что оценка вариограммы $2\tilde{\gamma}(h)$ является состоятельной в среднеквадратическом смысле для вариограммы $2\gamma(h)$, $h \in \mathbb{Z}$.

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. Далее в этой главе будут 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. Любое изменение параметров контрольной панели сразу же отображается в качестве результата в области исследования.

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].

Представленный модуль включает в себя возможности по просмотру и анализу данных: графически и с помощью таблицы, позволяющей сортировать и производить поиск по определённому признаку. 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 с помощью гистограммы и графика квантилей. Контрольная панель позволяет изменять отображаемый в данный момент график, а также позволяет выбрать критерий нормальности. В случае выбора для отображения гистограммы, по-

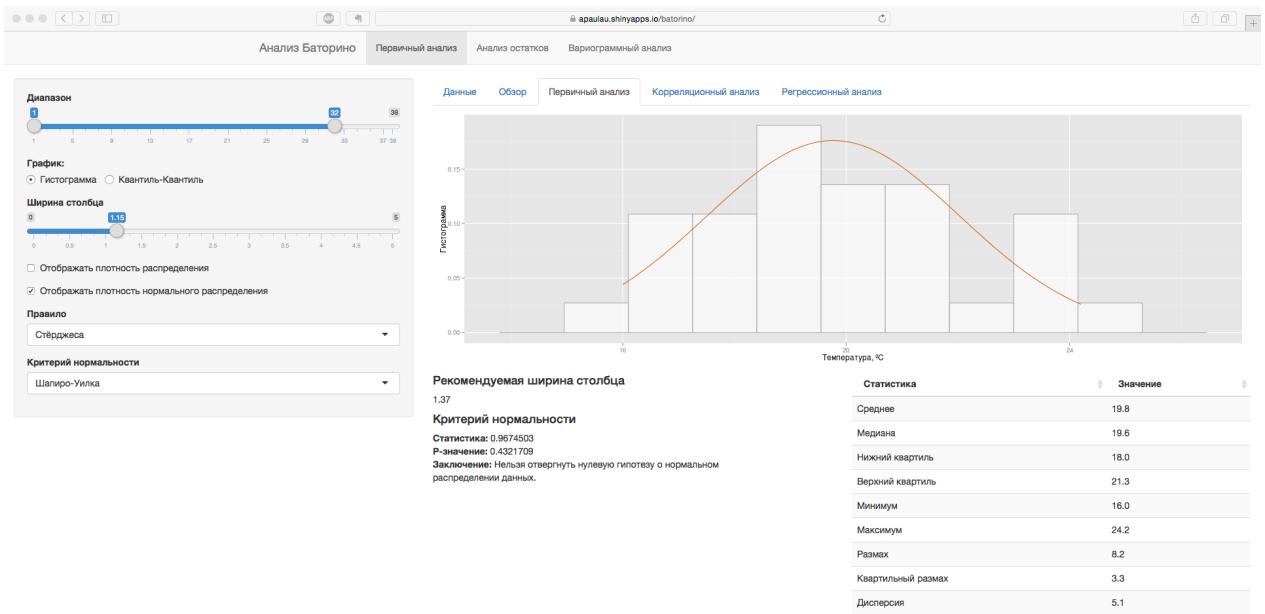


Рисунок 3.1 — Первичный анализ и описательные статистики

являются управляющие элементы, позволяющие выбрать ширину столбца гистограммы и правило по её вычислению (например, правило Стерджеса), отобразить плотность выборочного распределения и кривую нормального распределения.

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] был разработан 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 рисунке 3.1 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 коэффициента корреляции, а также вычисляет для него границы доверительного интервала. Среди прочего, данная страница содержит проверку на 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: определение значимости вычисленных коэффициентов, адекватность модели с помощью критерия Фишера и проверки линейности.

Инструменты, рассмотренные в рамках данного модуля, позволяют

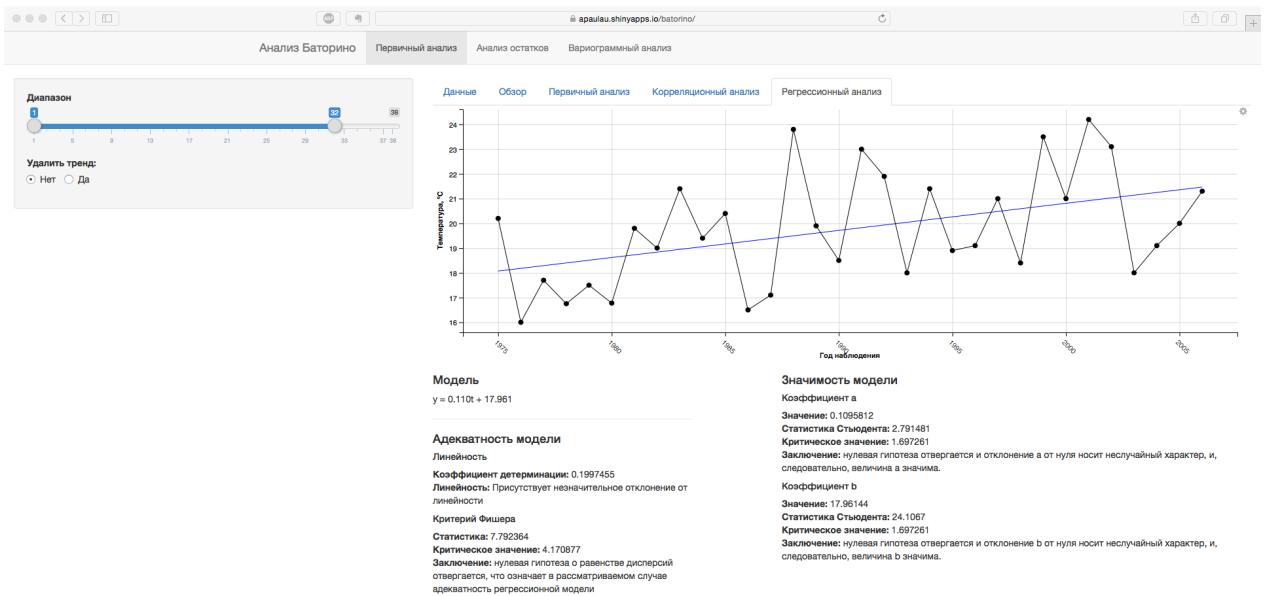


Рисунок 3.2 — Регрессионный анализ

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, отбрасывая первые или последние элементы. Это позволяет быстро оценить, насколько влияют данные на результат в конкретном случае.

3.2 Модуль анализа остатков

Модуль анализа остатков 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 модуля. Исключение составляют инструменты регрессионного и корреляционного анализов. Поскольку исследуемый на данном этапе временной ряд представляет собой ошибку.

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 продемонстрирован график автокорреляционной функции, позволяющий визуально определить наличие автокорреляций в исследуемых данных. Также проверить наличие значимых автокорреляций позволяет реа-

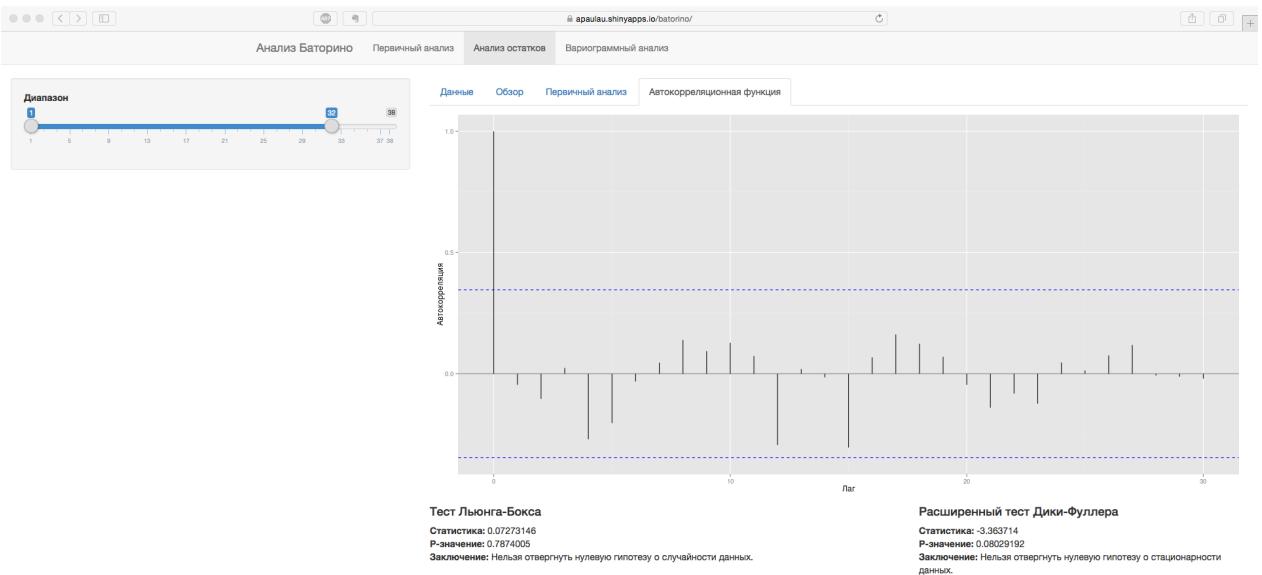


Рисунок 3.3 — Анализ автокорреляционной функции

лизованный тест Льюнга-Бокса. В свою очередь, расширенный тест Дики-Фуллера, также представленный на рассматриваемой странице, проверяет наличие стационарности в исследуемом временном ряду.

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 модель удовлетворительного качества, 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 ряда моделей вариограмм, поэтому в рассматриваемом модуле акцент сделан именно на подборе и анализе моделей вариограмм, наиболее подходящих для экспериментальных данных.

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 [16]. Для подбора модели вариограммы, в общем случае, существует два подхода: подбор визуально 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 семивариограммы, так и по конечному прогнозу кригингом.

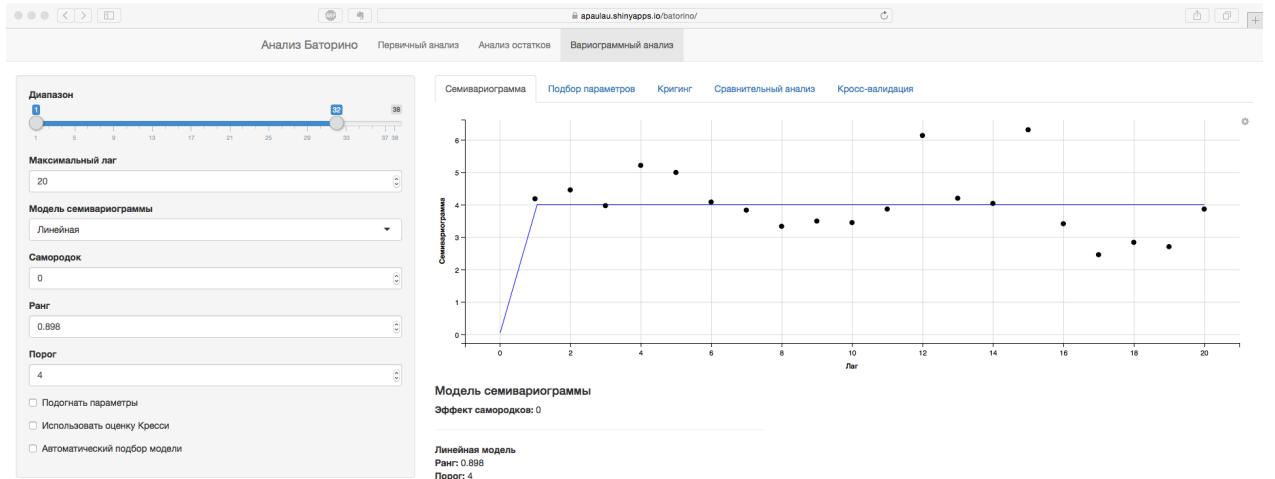


Рисунок 3.4 — Возможности по подбору модели семивариограммы

На рисунке 3.4 изображён скриншот начального этапа вариограммного анализа. Инструменты 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 между экспериментальной семивариограммой и моделью. Выбор наилучшей модели осуществляется по минимальному значению невязки. Представленная страница позволяет оценить по графику семивариограммы подобранные либо вручную, либо автоматически модель и параметры.

При использовании той или иной модели интерполяции крайне важно правильно подобрать значения модельно-зависимых параметров. Для кригинга такими параметрами являются параметры модели семивариограммы. Для проверки качества модели в дальнейшем используется кроссвалидация. Кросс-валидация является наиболее простым и часто использующимся подходом при сравнении результатов, получаемых различными

методами или одним и тем же методом, но с различными параметрами. В данном случае процесс кросс-валидации заключается в последовательном исключении одного значения температуры из исследуемых данных $\varepsilon(t)$ и построении интерполяции в этой точке по валидируемой модели. Таким образом, получаем ряд интерполяций $\varepsilon^*(t)$, который должен в идеальном случае воспроизводить поведение исследуемого ряда. По ряду $\varepsilon^*(t)$, с помощью различных статистик можно оценивать качество конкретной модели семивариограммы. С помощью таких статистик можно проследить, как изменяется качество модели при изменении какого-либо из параметров. В данной работе используются следующие статистики [17]:

1. Сумма квадратов невязок:

$$S = \sum_{i=1}^n (\varepsilon(t_i) - \varepsilon^*(t_i))^2,$$

где n — объём выборки.

2. Коэффициент эффективности:

$$E = \frac{S}{\sum_{i=1}^n (\varepsilon(t_i) - \bar{\varepsilon})},$$

где $\bar{\varepsilon}$ — среднее значение исследуемых данных.

3. Среднее абсолютных значений:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\varepsilon(t_i)|,$$

где n — объём выборки.

4. Среднеквадратическая ошибка

$$MSE = \frac{1}{n} \sum_{i=1}^n (\varepsilon(t_i) - \varepsilon^*(t_i))^2, \quad (3.1)$$

где n — объём выборки.

5. Коэффициент корреляции $r_{\varepsilon\varepsilon^*}$.

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 вид модели семивариограммы, начальные значения параметров и статистика

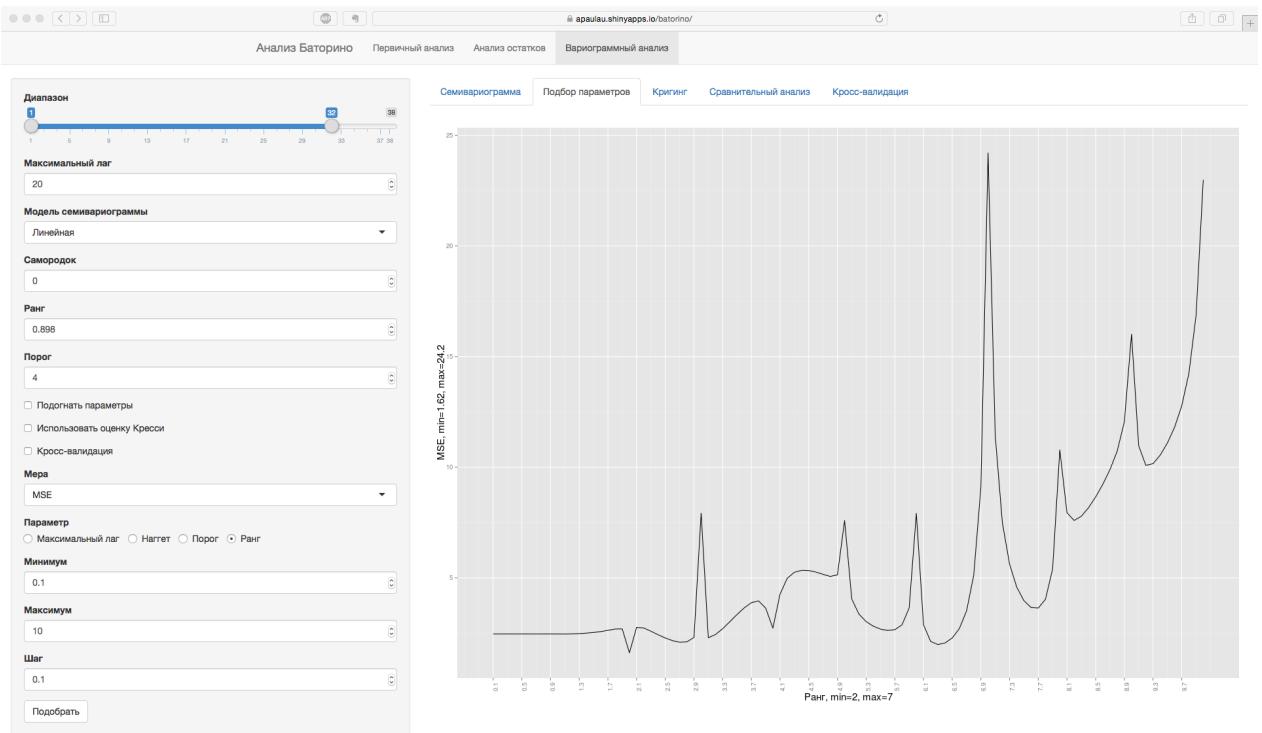


Рисунок 3.5 — Подбор параметров модели семивариограммы

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

В результате такого процесса получается ряд оценок моделей, из которых выбирается оптимальная. 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, перекрёстный, модель оценивается с помощью описанного ранее метода кросс-валидации. При втором подходе, адаптивном, в исследуемых данных отдаётся предпочтение последним наблюдениям. Для этого из исходных данных исключается некоторое количество значений и модель строится по оставшимся. Подбор параметров осуществляется по показателям качества, основанным на отклонении вычисленных значений от исключенных. Таким образом, как будет показано в главе 4, достигается наилучший прогноз в краткосрочной перспективе.

Таким образом на данной странице можно оценить поведение модели при изменении какого-либо из параметров и для каждого подобрать оптимальное значение.

Страница кригинга (рисунок 3.6) является наглядной демонстрацией применения всего вышеописанного. На ней изображается график с наблю-

даемыми значениями и прогнозными значениями, вычисленными по линейной регрессионной модели и кригингом. Это позволяет оценить полученные

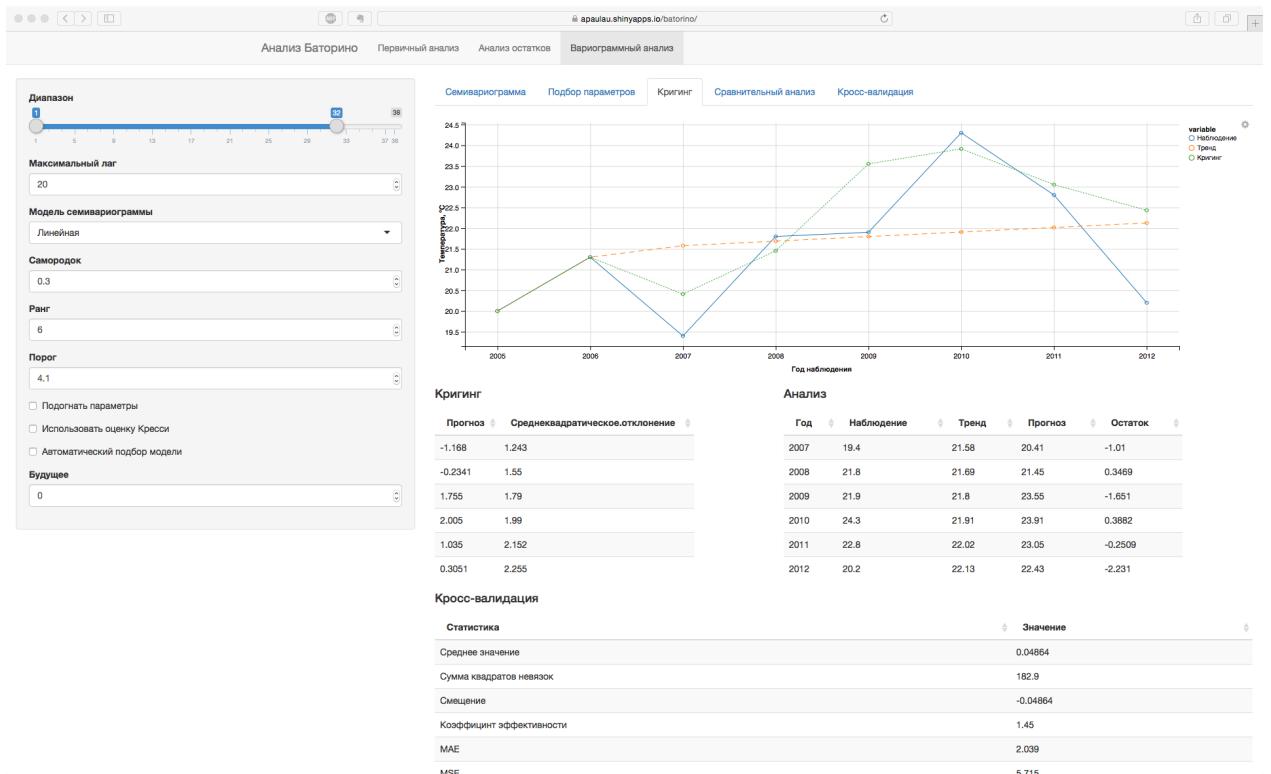


Рисунок 3.6 – Сравнение прогнозных значений

ную модель и сделать конкретные заключения. Приводятся вспомогательные таблицы с произведёнными в процессе расчётами. В первую очередь это результаты кригинга со среднеквадратической ошибкой для каждого из значений. Также отображается табличный вариант данных, изображённых на графике. И последняя таблица показывает значения показателей качества 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 Детерминированные методы

4.1.1 Описательные статистики и первичный анализ данных

В качестве исследуемых данных примем выборку объема $N = 38$ 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. График исследуемых данных изображён на рисунке 4.1.

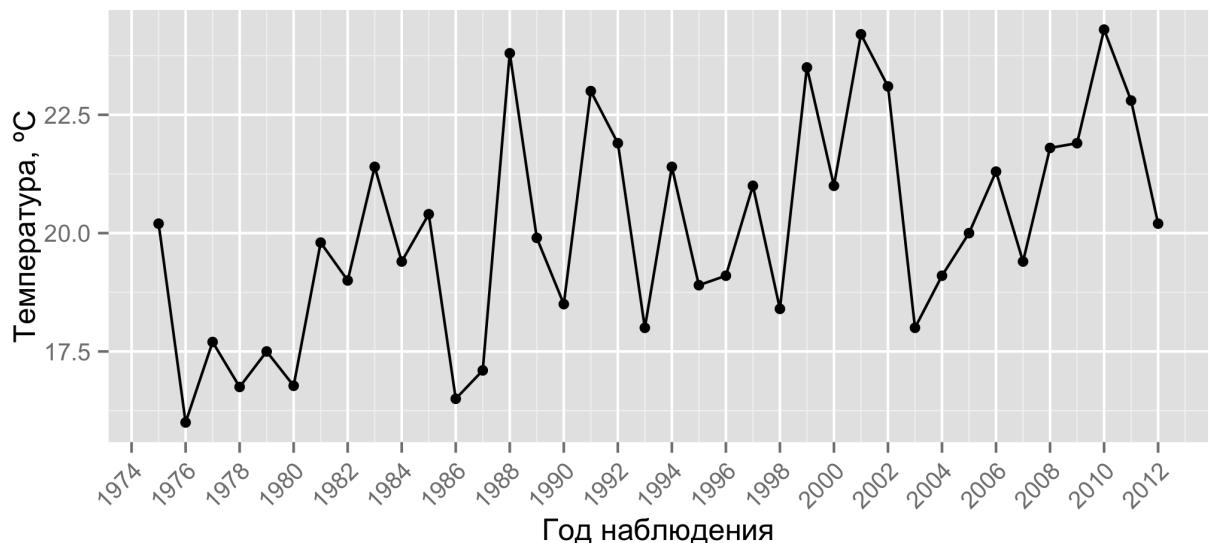


Рисунок 4.1 — График исходных данных

Следует отметить, что для непосредственного изучения 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 прогнозирования. Заметим, что работа, представленная в параграфах 4.1.1–4.1.3, была также проделана и для всей

выборки объёма $N = 38$. 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.

Таблица 4.1 — Описательные статистики для наблюдаемых температур.

	Значение
Среднее	19.77
Медиана	19.60
Нижний quartиль	18.00
Верхний quartиль	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

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

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. Такое значение говорит о незначительной правосторонней асимметрии распределения. То есть о том, что выборочное распределение можно считать близким к нормальному [18].

Коэффициент эксцесса в рассматриваемом случае равен -0.746. Так как коэффициент эксцесса нормального распределения равен 0, то в этом случае можно говорить о пологости пика распределения выборки по отношению кциальному распределению [18].

С помощью тестовых статистик для коэффициента асимметрии и эксцесса [15, с.85-89], проверим значимость полученных значений для гене-

ральной 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, 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 небольшого 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 виде распределения. Для вычисления интервалов разбиения воспользуемся *формулой Стерджеса*. Из [19] 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 с параметрами $\mathcal{N}(19.77, 5.12)$ (рисунок 4.2). Проанализируем эту гистограмму. Во-первых, на ней можно заметить небольшую скошенность вправо, что подтверждается показателем асимметрии, полученным на этапе вычисления описательных статистик. Во-вторых, коэффициент эксцесса в таблице 4.1 указывал на пологость пика распределения, что подтверждается кривой плотности — она имеет чуть более растянутую колоколообразную форму. Таким образом, представленная гистограмма показывает близость выборочного распределения кциальному с параметрами $\mathcal{N}(19.77, 5.12)$.

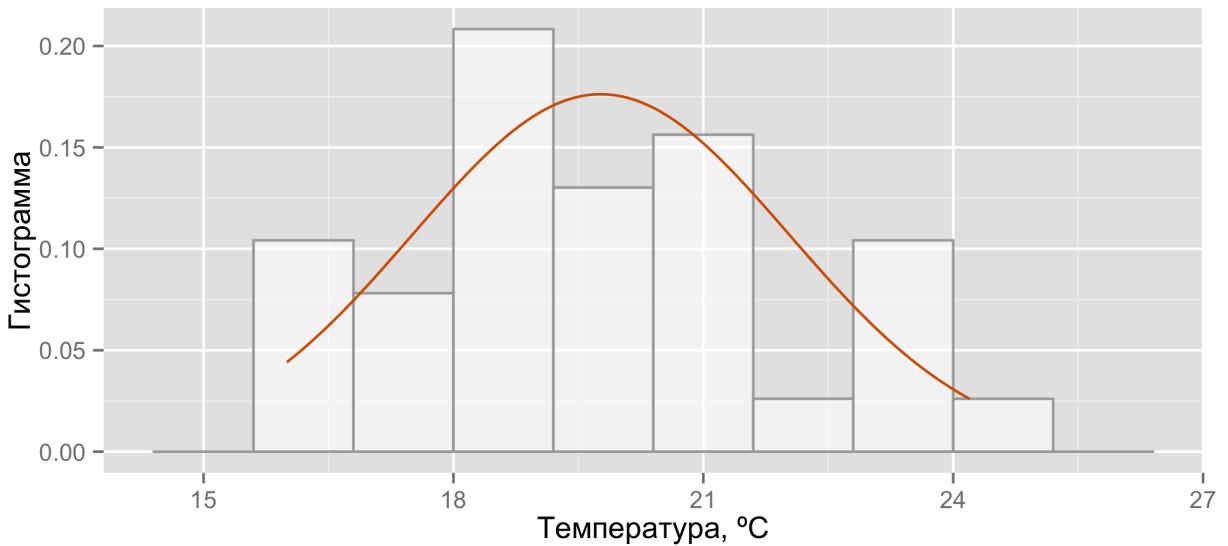


Рисунок 4.2 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения $\mathcal{N}(19.77, 5.12)$

Другим часто используемым графическим способом проверки характера распределения данных является построение т.н. *графиков квантилей*. 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 визуально обнаружить нетипичное 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 при-

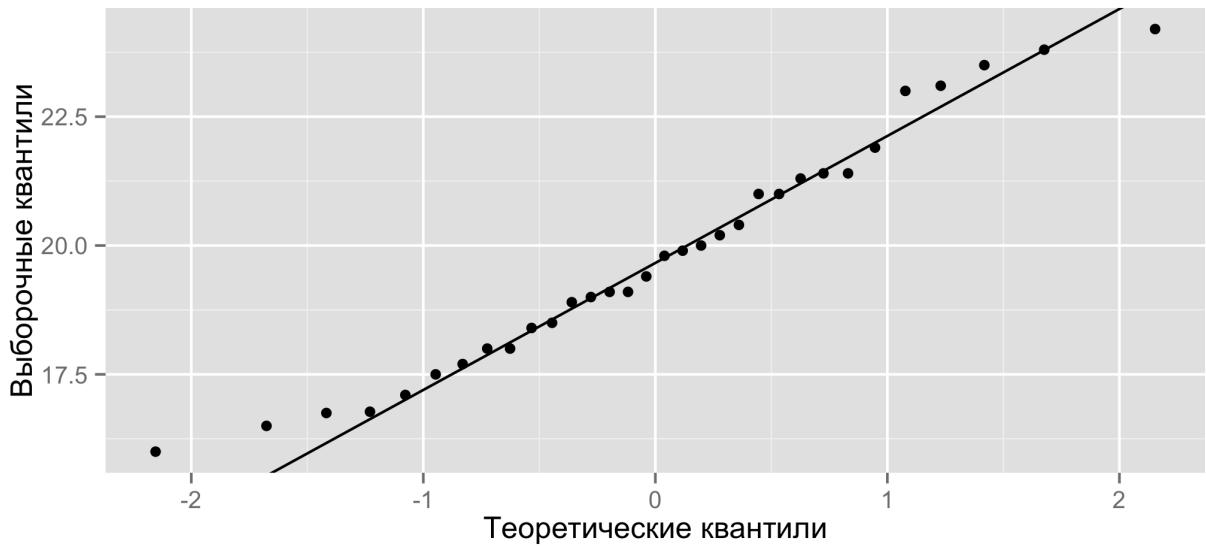


Рисунок 4.3 — График квантилей для наблюдаемых температур

нятого 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* [20]. Из полученных в **R** результатов, статистика Шапиро-Уилка $W = 0.97$. Вероятность ошибки $p = 0.43 > 0.05$, а значит нулевая гипотеза не отвергается [21]. Следовательно опровергнуть предположение на основе данного теста нельзя.

Попробуем опровергнуть наше предположение на основе проверки критерия χ^2 Пирсона [22]. 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 теста χ^2 Пирсона 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 [23]. Как в предыдущем случае воспользуемся представленной в пакете *nortest* функцией *ks.test*. Из полученных в **R** результатов, статистика Колмогорова–Смирнова $D = 0.098$. Вероятность ошибки $p = 0.92 > 0.05$, 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 присутствии выбросов в исходной выборке. Выявление таких аномальных значений важно, так как их наличие, как правило, сильно влияет на выборочные характеристики, в частности, на коэффициент корреляции. Проверим наличие выбросов с помощью статистических критериев. Для этих целей воспользуемся критерием Граббса [24], который 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 [25]. 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 о наличии в выборке выброса. 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 1975–2006 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 направление зависимости. Если точки на диаграмме расположены хаотически, то это говорит о независимости рассматриваемых переменных. Если с ростом переменной t возрастает переменная X то имеет место прямая зависимость. Если же с ростом переменной t переменная X убывает, то это указывает на обратную зависимость.

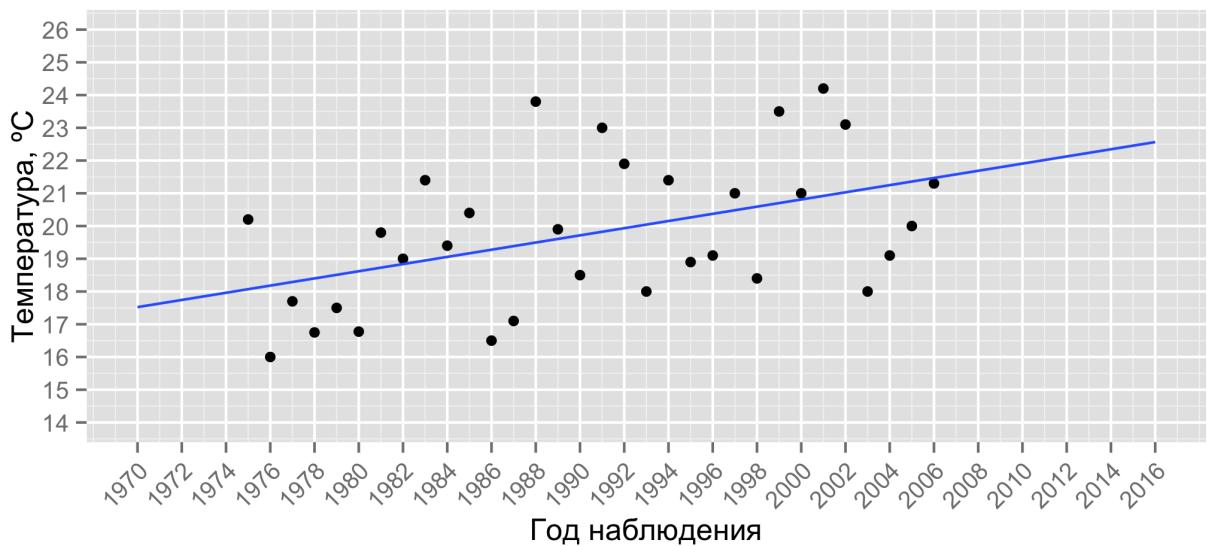


Рисунок 4.4 — Диаграмма рассеяния

Из рисунка 4.4 видно, что точки образуют своеобразное «облако», ориентированное вверх, то есть присутствует прямая зависимость между рассматриваемыми переменными. Также, 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. В качестве нулевой гипотезы принимается следующая “коэффициент корреляции генеральной совокупности равен нулю”.

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

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 о равенстве нулю коэффициента корреляции генеральной совокупности.

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 прямую умеренную 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 [26].

Во временных рядах выделяют четыре составляющие:

1. *Тренд (тенденция развития)* y — эволюционная составляющая, которая характеризует общее направление развития изучаемого явления и связана с действием долговременных факторов развития.
2. *Циклические k , сезонные колебания s* — 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)* ε , 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: $x = y \times k \times s \times \varepsilon$.

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

По причине того, что в данном случае мы рассматриваем среднюю температуру июля месяца каждого года 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 $a, b \in \mathbb{R}$ – коэффициенты модели, $t = \overline{1, n}$, n – объем выборки.

Продолжая рассуждение, как наблюдение из графика 4.4, можно отметить, что не происходит увеличения амплитуды колебаний с течением времени. А значит, искомая модель является аддитивной. Из всего вышеизложенного 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 [27]. 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).

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

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

Таким образом получена линейная модель, описывающая тенденцию развития:

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

На основе построенной 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.

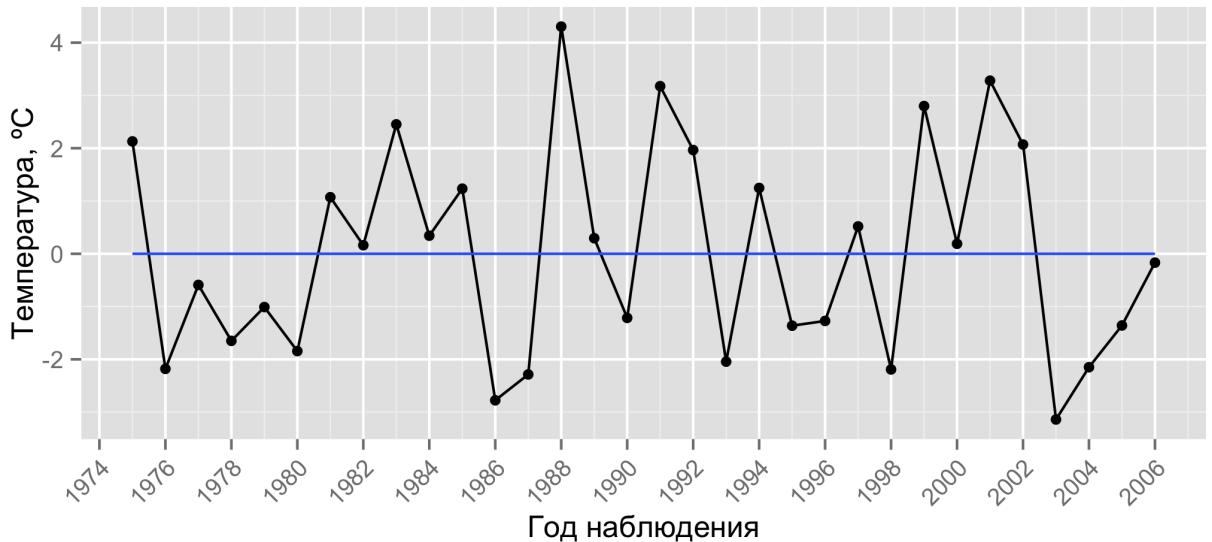


Рисунок 4.5 — Нерегулярная составляющая $\varepsilon(t)$

Проведём анализ регрессионной модели (4.3). Для этого проверим значимость коэффициентов регрессии и оценим адекватность построенной регрессионной модели.

Рассчитаем вспомогательные величины, воспользовавшись [26]. 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]. Примем уровень значимости $\alpha = 0.05$, тогда

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

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

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

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

Оценим адекватность полученной регрессионной модели. Дисперсия модели:

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

Остаточная дисперсия:

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

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

$$F_{\text{крит}} \approx 7.79,$$

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

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

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

Рассчитаем коэффициент детерминации:

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

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, построим прогноз по полученной модели. Вычисленные прогнозные значения на 2007-2012 годы для сравнения отображены в таблице 4.2: Имеющееся отклонение прогнозов от реальных данных ещё раз

Таблица 4.2 – Сравнение прогнозных значений (модель $y(t)$)

	$X(t)$	$y(t)$	$X(t) - y(t)$
2007	19.400	18.071	1.329
2008	21.800	18.181	3.619
2009	21.900	18.290	3.610
2010	24.300	18.400	5.900
2011	22.800	18.509	4.291
2012	20.200	18.619	1.581

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 составляющую $\varepsilon(t)$. 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.

Таблица 4.3 — Описательные статистики остатков

	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

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 по сравнению с исходным случаем.

Построим гистограмму и график квантилей для проверки последних заключений. Построенная гистограмма (приложение Б, рисунок Б.1) демонстрирует скошенность вправо и пологость пика, отображённой кривой нормального распределения, что согласуются с полученными в таблице 4.3 коэффициентами асимметрии и эксцесса.

На рисунке 4.6 можно заметить, что присутствуют отклонения относительно нормального распределения. Наиболее явный из них — нижний

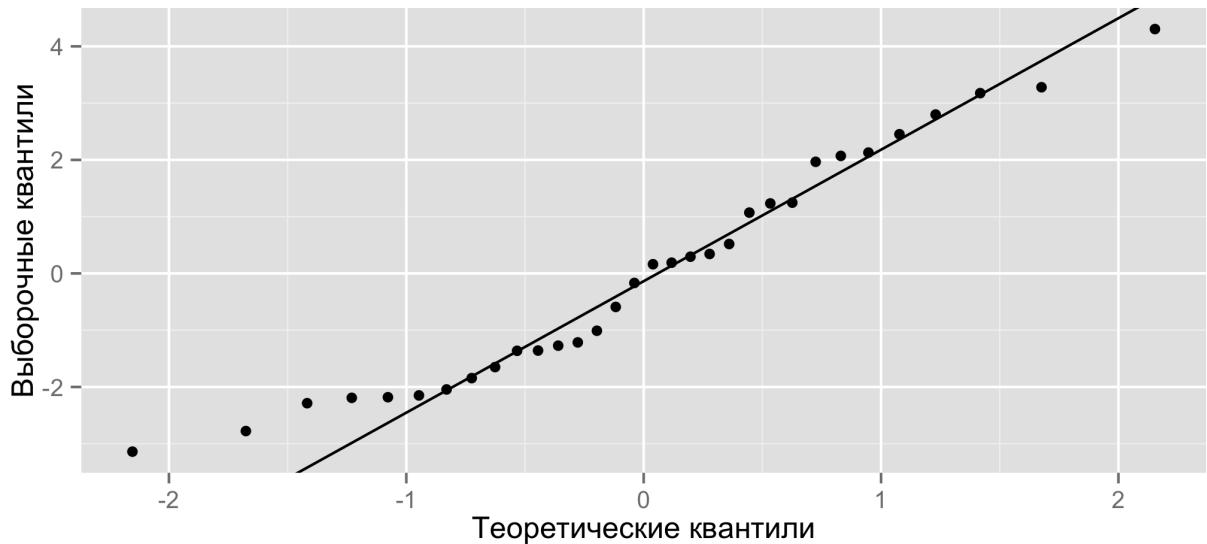


Рисунок 4.6 — График квантилей для остатков

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, сформулированная ранее, 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 разграничивают значимые и не значимые корреляции: значения, выходящие за линии, являются значимыми [12, с.376]. На представленном графике автокорреляционной функции все значения 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 резкое затухание значений автокорреляций с увеличением лага. На основе этого можно сделать предположение о стационарности в широком смысле. Для проверки этого предположения воспользуемся расширенным тестом Дики–Фуллера (ADF) [28]. Из результатов проверки теста, статистика Дики–Фуллера $DF = -3.36$, веро-

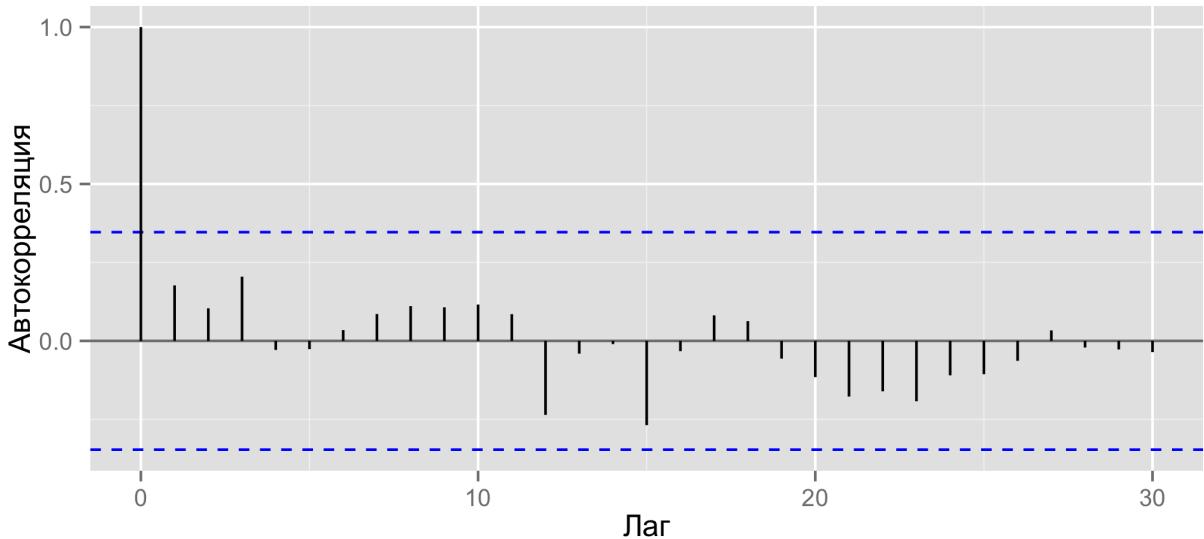


Рисунок 4.7 — График автокорреляционной функции

ятность ошибки $p = 0.04 < 0.05$. 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 точно воспроизвести поведение исходного временного ряда. То есть нерегулярная составляющая $\varepsilon(t)$ является существенной и отвечает за это поведение. Был проведен анализ остатков, в процессе которого показаны близость распределения к нормальному (с некоторыми отклонениями) и стационарность в широком смысле, 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 искомой модели для нерегулярной составляющей.

4.2 Геостатистические методы

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 выводы были сделаны в результате проведённого в предыдущем разделе. А также сделан вывод о

необходимости использования современных методов.

В современных исследованиях аналогичного класса задач усилился интерес к геостатистическим моделям интерполяции, что подтверждается работами [29, 30]. 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 Визуальный подход

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

Прогнозные значения $X^*(t)$ вычисляются по формуле:

$$X^*(t) = y(t) + \varepsilon^*(t),$$

где $y(t)$ — тренд, $\varepsilon^*(t)$ — значения, вычисленные с помощью кригинга.

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 семивариограммы 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. Разность координат наблюдений

$$h = \varepsilon(i) - \varepsilon(j), \quad i, j = \overline{1, n}, \quad i \neq j,$$

называется *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 семивариограммой*. Модель семивариограммы является непрерывной функцией, описывающей экспериментальную. Обычно эта функция характеризуется двумя параметрами: *рангом* a и *порогом* c . Порог характеризует

пределное значение семивариограммы, на некотором расстоянии, называемом рангом, за которым последующие значения становятся некоррелированными. Также в некоторых моделях может присутствовать так называемый эффект самородков. Он характеризует разрыв в значениях около нуля и в геостатистике такой эффект связывают с погрешностями измерений [17].

Для изучения поведения данных при увеличении лага построим диаграмму взаимного разброса наблюдений (*h-scatterplot*), разделённых расстоянием h . Эта диаграмма позволяет проверить наличие корреляции в исследуемых данных как качественно, так и количественно [17]. Построенная диаграмма изображена на рисунке Б.2 в приложении Б. Следует отметить, что на первом же лаге присутствует незначительная корреляция, при этом на некоторых лагах коэффициент корреляции выше. Такое поведение свойственно так называемым беспороговым моделям семивариограммы. 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 [17].

$$\hat{\gamma}(h) = c_0 + Lin(h) = \begin{cases} c_0 + b \cdot h, & h > 0, \\ c_0, & h \leq 0, \end{cases} \quad (4.4)$$

где b — параметр, отвечающий за угол наклона, c_0 — эффект самородков.

Выводы по диаграмме Б.2 в рассматриваемом случае вполне обосновано спецификой исследуемых данных: рассматривается средняя температура воды за один определённый месяц в течение нескольких лет. Ко всему прочему, это подтверждается результатами проведённого ранее анализа остатков, в котором мы выяснили, что распределение ряда остатков является близким к нормальному и значения некоррелируемы и независимы.

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

Экспериментальной семивариограммой по сути является некоторая оценка семивариограммы. Существует несколько известных оценок, каждая из которых имеет свои достоинства и недостатки. Для исследования были выбраны наиболее распространённые: оценка Матерона (2.1), введённая ранее в главе 2, и оценка Кресси-Хокинса [16, 31]:

$$2\tilde{\gamma}(h) = \frac{1}{n-h} \left(\sum_{t=1}^{n-h} |X(t+h) - X(t)|^{\frac{1}{2}} \right)^4 / (0.457 + \frac{0.494}{n-h} + \frac{0.045}{(n-h)^2}), \quad h = \overline{0, n-1}. \quad (4.5)$$

Данная оценка является робастной и в теории позволяет учесть наличие выбросов в исследуемых данных [32].

Как уже было сказано ранее, в процессе анализа остатков, $\varepsilon(t)$ обладает свойствами нерегулярной составляющей, указанными в разделе 4.1.3. В свою очередь, случайный процесс из главы 2 также удовлетворяет этим свойствам. Поэтому сначала проведём исследования с помощью оценки Матерона (2.1), а затем сравним полученные результаты с результатами использования оценки Кресси-Хокинса. Построенная оценка семивариограммы (Матерона) отображена на рисунке 4.8.

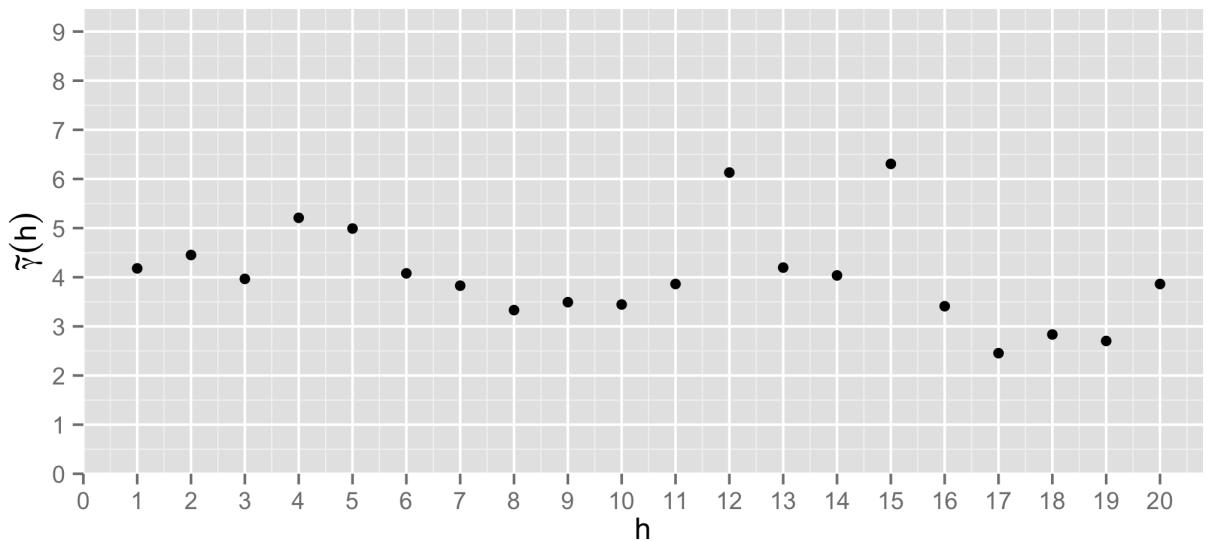


Рисунок 4.8 — Оценка семивариограммы Матерона

На практике построение модели семивариограммы представляет собой итеративный процесс, на каждом шаге которого следует наилучшим образом подобрать параметры очередного модельного приближения. При этом существует два пути подбора моделей семивариограммы: подбор силами исследователя, т.е. визуально с ручным выбором параметров, и автоматическим подбором параметров с помощью специальных методов и алгоритмов. В различных источниках рекомендуется строить модели вручную, так как исследователь лучше знает специфику данных и может учесть это при подборе параметров [33]. Исследуем оба подхода: визуальный и автоматический подбор параметров.

Следует также отметить, что в некоторых источниках советуют при построении модели семивариограммы учитывать параметр максимального расстояния, до которого вычисляются значения семивариограммы, а также приводят рекомендацию по его подбору. В связи с этим, первоначальным параметром было выбрано значение, рассчитанное по рекомендации [34]:

$$h_{\max} = 2n/3 = 20 \quad (4.6)$$

При таком значении можно наблюдать все особенности семивариограммы и не рассматривать отдаленные значения.

Вследствие выводов по диаграмме взаимного разброса, начнём подбор модели семивариограммы с линейной модели (4.4). Подбор начальных параметров осуществим визуально. Из уравнения (4.4) видно, что параметр b отвечает за угол наклона прямой, поэтому примем начальное значение $b = 4$, приблизительно равное дисперсии. Построенная модель, вида

$$\hat{\gamma}_1(h) = \text{Lin}(h), \quad b = 4,$$

отображена на графике Б.3. Следует отметить, что данная модель после применения кригинга позволила получить значения очень близкие к нулю, что не изменило прогноза, построенного по модели тренда (таблица 4.2).

Далее подберём модель семивариограммы с помощью возможностей пакета *gstat*. В результате его использования получаем модель $\hat{\gamma}_2(h)$ с чистым эффектом самородков

$$\hat{\gamma}(h) = c \cdot \text{Nug}(h) = \begin{cases} 0, & h = 0, \\ c, & h \neq 0, \end{cases} \quad (4.7)$$

с параметром $c = 4.04$. Объяснить такой результат можно тем, что автоматический подбор параметров из пакета *gstat* основан на методе наименьших квадратов. А поскольку значения семивариограммы сразу достигают порогового значения, приблизительно равному дисперсии, то эффект самородков $\hat{\gamma}_2(h)$, изображённый на рисунке Б.4, оказывается наилучшей моделью. Но при этом данная модель не учитывает особенностей исследуемых данных, поэтому результатов прогнозирования она не улучшила (таблица В.2, приложение В). Другими словами, рассмотренный подход не учитывает поведение оценки семивариограммы около нуля, поскольку в исходных данных нет информации о ближайших к исследуемому месяцам.

Наличие порога объяснимо видом оценки семивариограммы: на рисунке 4.8, как уже было замечено, уже первые значения достигают уровня дисперсии. И отклонение от этого значения не велико. Это согласуется с результатами, полученными при анализе остатков. Из этого следует, что использование только беспороговых моделей не обосновано. При этом следует учесть то, что не учёл автоматический подбор параметров: поведение $\tilde{\gamma}(h)$ около нуля. Поэтому воспользуемся линейной моделью с порогом (4.8), которая является комбинацией моделей (4.4) и (4.7) [35].

$$\hat{\gamma}(h) = c_0 + c \cdot \text{Lin}(h, a) = \begin{cases} c_0 + c \cdot \frac{h}{a}, & 0 \leq h \leq a, \\ c_0 + c, & h > a, \end{cases} \quad (4.8)$$

где c_0 – эффект самородков, c – порог, a – ранг.

Для подбора оптимальных параметров модели семивариограммы (4.8) воспользуемся инструментами реализованной программы, рассмотренной в

главе 3. Воспользуемся первым из описанных подходом — перекрёстным. В качестве статистики, для оценки качества модели, будем использовать коэффициент корреляции между $\varepsilon(t)$ и $\varepsilon^*(t)$. График зависимости значения ранга на качество модели отображён на рисунке 4.9. По рисунку видно, что

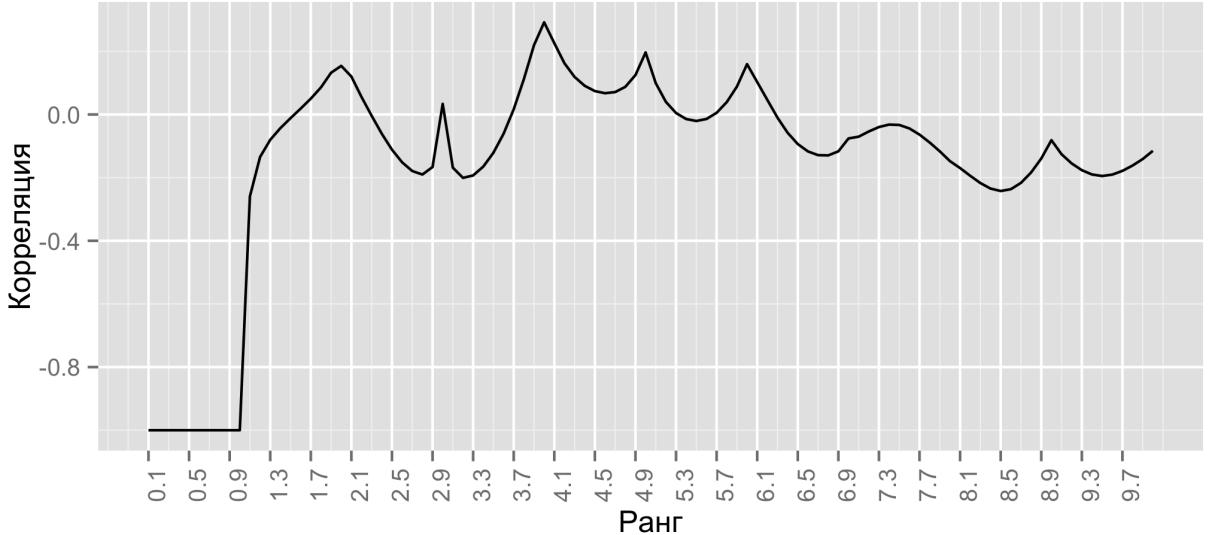


Рисунок 4.9 — Зависимость качества линейной модели от значения ранга

максимальное значение коэффициента корреляции $r_{\varepsilon\varepsilon^*} = 0.292$ достигается при значении ранга $a = 4$. Таким образом получена модель

$$\hat{\gamma}_3(h) = 4 \cdot \text{Lin}(h, 4), \quad (4.9)$$

её график отображен на рисунке Б.5. Вычисленные по представленной модели прогнозные значения можно проследить по таблице 4.4 и по графику Б.6 в приложении Б. Прогнозные значения оказались не очень точными, что объясняется высоким значением среднеквадратической ошибки (3.1) ($MSE = 7.931$). При этом, коэффициент корреляции $r_{\varepsilon\varepsilon^*} = 0.292$ выше, чем аналогичные коэффициенты корреляции уже построенных моделей. Поэтому можно сделать вывод о том, что модель $\hat{\gamma}_3(h)$ может использоваться для описания исходных данных, в частности, для вычислений пропущенных наблюдений.

Таблица 4.4 — Прогнозные значения (модель $\hat{\gamma}_3(h)$)

	$X(t)$	$X^*(t)$	$y(t)$	$X(t) - X^*(t)$
2007	19.400	20.824	21.578	-1.424
2008	21.800	21.133	21.687	0.667
2009	21.900	19.831	21.797	2.069
2010	24.300	22.129	21.906	2.171
2011	22.800	22.239	22.016	0.561
2012	20.200	22.348	22.126	-2.148

Для построения более точного прогноза воспользуемся адаптивным подходом по подбору параметров, который введён в главе 3. График зависимости среднеквадратической ошибки от ранга при таком подходе отображен на рисунке Б.7. Оптимальным параметром для ранга, из этого графика, является значение $a = 2$, с минимальной среднеквадратической ошибкой $MSE = 1.62$.

$$\hat{\gamma}_4(h) = 4 \cdot \text{Lin}(h, 2). \quad (4.10)$$

Построенная модель семивариограммы (4.10) изображена на рисунке Б.8. График 4.10 и таблица В.3 прогнозных значений показывают, что алап-

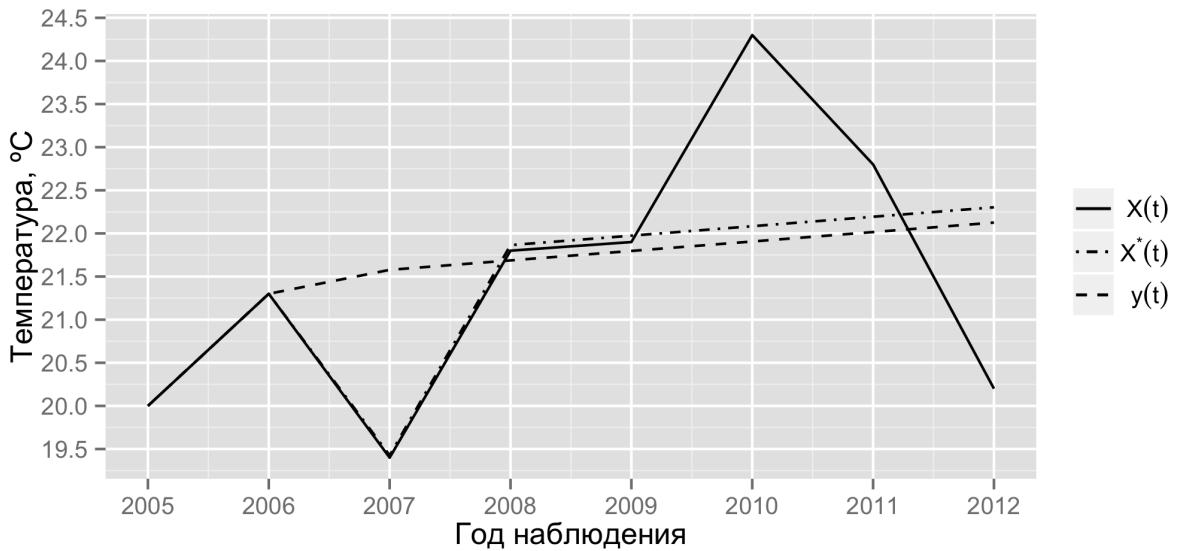


Рисунок 4.10 — Прогноз по модели $\hat{\gamma}_4(h)$

тивный подход позволил получить значения за 2007–2009 годы очень близкие к исходным. Что является хорошим результатом. При этом статистики по данной модели после проведения кросс-валидации оказались равными: коэффициент корреляции $r_{\varepsilon\varepsilon^*} = 0.152$, среднеквадратическая ошибка $MSE = 18.69$. Что говорит о том, что модель $\hat{\gamma}_4(h)$ описывает всю выборку хуже чем предыдущая. Таким образом модель, полученная адаптивным подбором параметров хорошо себя показывает для вычисления краткосрочных значений, в свою очередь перекрёстный метод может применяться для нахождения модели, описывающей наилучшим образом исходные данные.

Модель $\hat{\gamma}_4(h)$, полученная адаптивным подходом, 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 [33]:

$$\hat{\gamma}(h) = c_0 + c \cdot Sph(h, a) = \begin{cases} c_0 + c \cdot \left(\frac{3}{2} \frac{h}{a} - \frac{1}{2} \left(\frac{h}{a}\right)^3\right), & h \leq a, \\ c_0 + c, & h \geq a. \end{cases}$$

Однако после подбора оптимальных параметров оказалось, что данная модель вписывается в исследуемые данные хуже, чем найденные ранее. При перекрёстном подборе параметров наилучшей получилась модель

$$\hat{\gamma}_5(h) = 4Sph(h, 2.3), \quad (4.11)$$

с показателями качества: коэффициент корреляции $r_{\varepsilon\varepsilon^*} = -0.002$ и среднеквадратической ошибкой $MSE = 5.407$. В случае адаптивного подхода, оптимальными оказались параметры $c = 4, a = 6.9$ с эффектом самородков $c_0 = 0.9$, со среднеквадратической ошибкой прогнозных и истинных значений $MSE = 2.01$. Применив кросс-валидацию к этой модели, получаем следующие показатели качества: коэффициент корреляции $r_{\varepsilon\varepsilon^*} = -0.009$ и среднеквадратической ошибкой $MSE = 5.396$. Графики семивариограммы и прогнозных значений последней модели отображены на рисунках Б.9 и Б.10 в приложении Б соответственно. Можно сделать вывод, что как и модель $\hat{\gamma}_4(h)$ (4.10), сферическая модель $\hat{\gamma}_5(h)$ не позволила описать поведение исследуемой выборки. Только в случае краткосрочного прогноза она проявила себя, предсказав характерное поведение исключённых значений, хоть и хуже модели $\hat{\gamma}_4(h)$. Полученный результат является следствием вида их моделей семивариограммы. На их графиках Б.8 и Б.9 в приложении Б можно видеть, что они похожи: линейное поведение до значения порога, и прямая линия, отвечающая за некоррелированные значения, на его уровне. Как следствие, похожие результаты.

Если обратить внимание на график экспериментальной семивариограммы 4.8, то можно заметить некоторый периодический эффект в виде волны. Поэтому дальнейшей подбираемой моделью возьмем периодическую [35]:

$$\hat{\gamma}(h) = c_0 + c \cdot Per(h, a) = 1 - \cos\left(\frac{2\pi h}{a}\right), \quad (4.12)$$

где c_0 – эффект самородков, c – порог, a – ранг.

С помощью перекрёстного метода в написанной программе подобрана модель

$$\hat{\gamma}_4(h) = 4 \cdot Per(h, 0.898), \quad (4.13)$$

график семивариограммы которой изображен на рисунке Б.11 в приложении Б. Показатели качества подобранный модели: коэффициент корреляции $r_{\varepsilon\varepsilon^*} = 0.404$, 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.4 в приложении В и график прогнозных значений 4.11 по подобранный модели (4.13) показывают, что прогноз получился не очень точный, но при этом следует принять во внимание, что модель $\hat{\gamma}_4(h)$ подбиралась для

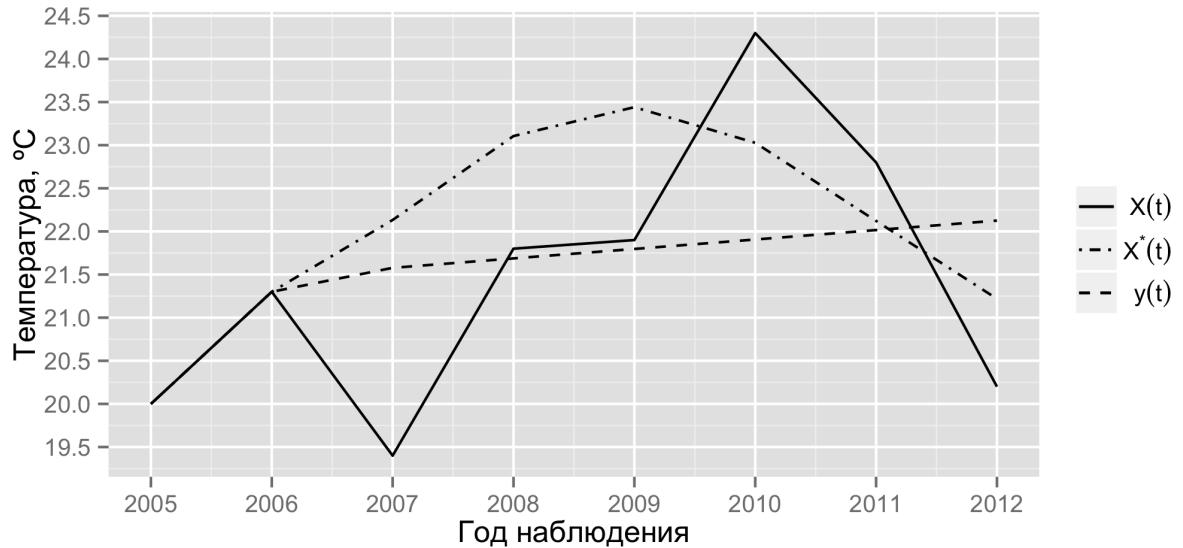


Рисунок 4.11 — Прогноз по модели $\hat{\gamma}_6(h)$

описания всей выборки, не учитывая локальное поведение в последних наблюдениях. Модель $\hat{\gamma}_6(h)$ оказалась наилучшей для описания всей выборки из всех проверенных.

Следует также упомянуть использовавшуюся оценку семивариограммы Кресси-Хокинса (4.5). Она изображена на рисунке Б.12 в приложении Б. Данная модель не сильно отличается от оценки Матерона (2.1), поэтому на результаты подбора моделей визуальным подходом она не повлияла.

Таким образом, с помощью инструментов реализованной программы проанализированы различные модели семивариограмм и влияние их параметров на результаты прогноза. Подбор осуществлён двумя методами: перекрёстным и адаптивным. С их помощью найдены наилучшие модели: перекрёстным методом подбора найдена периодическая модель $\hat{\gamma}_6(h)$, адаптивным — линейная с порогом $\hat{\gamma}_4(h)$. Как следствие, метод перекрёстный метод подбора параметров следует использовать для решения задач описания исследуемых данных, адаптивный — для решения задач по построению точных краткосрочных прогнозов.

4.2.2 Автоматический подход

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 наилучшую. Подбор параметров определённой модели семивариограммы осуществляется методом наименьших квадратов, пример использования которого показан ранее (для модели $\gamma_2(h)$). Подбор параметров сопровождается невязкой модели и значений семивариограммы. Это позволяет, основываясь на этом показателе,

выбирать наилучшим образом подходящей модели — по минимальному значению невязки. Следует отметить, что в рассматриваемом случае параметр максимального расстояния h_{\max} , для которого вычисляется семивариограмма, будет проявлять себя при выборе модели, поскольку количество точек, по которым подбирается модель, будет влиять на метод наименьших квадратов. Исходный код функции, реализующей автоматический подбор моделей, представлен в листинге Г.2.

Воспользуемся автоматическим методом подбора модели для оценки Матерона. Данная процедура, при принятом ранее максимальном значении лага (4.6), выбрала волновую модель (4.14) семивариограммы с эффектом самородков:

$$\hat{\gamma}(h) = c_0 + c \cdot Wav(h, a) = 1 - \frac{a}{h} \cdot \sin\left(\frac{h}{a}\right), \quad (4.14)$$

где c_0 — эффект самородков, c — порог, a — ранг.

Обозначим полученную модель

$$\hat{\gamma}_7(h) = 3.03 + 1.011 \cdot Wav(h, 1.14), \quad (4.15)$$

графически подобранная 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 результатов по моделям, найденных ранее. При этом показатели качества оказались равными: коэффициент корреляции $r_{\varepsilon\varepsilon^*} = -0.2$ и среднеквадратическая ошибка $MSE = 4.155$. Что подтверждает выводы по графикам.

Так как параметр h_{\max} влияет на подбор моделей семивариограммы, оценим качество моделей, подобранных автоматически, в зависимости от него. Как и ранее, качество модели будем оценивать двумя подходами: перекрёстным и адаптивным. Первый случай отображен на рисунке 4.12. Как видно из него, наибольший коэффициент корреляции соответствует оценке Матерона и максимальному расстоянию $h_{\max} = 26$. Для найденного значения h_{\max} , наилучшей моделью является периодическая (4.12):

$$\hat{\gamma}_8(h) = 3.46 + 0.5 \cdot Per(h, 2.67),$$

с показателями качества: коэффициент корреляции $r_{\varepsilon\varepsilon^*} = 0.20$, среднеквадратической ошибкой $MSE = 3.835$. 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 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, оптимальной моделью оказалась волновая модель с эф-



Рисунок 4.12 — Зависимость качества модели от максимального расстояния

фектом самородков

$$\hat{\gamma}_9(h) = 4.11 + 1.65 \cdot Wav(h, 3.59),$$

полученная по оценке Кресси-Хокинса при значении максимального расстояния $h_{\max} = 5$. И при почти равном значении среднеквадратической ошибки, периодическая модель с эффектом самородков

$$\hat{\gamma}_{10}(h) = 3.8 + 0.32 \cdot Per(h, 1.3)$$

по оценке Матерона при значении максимально расстояния $h_{\max} = 18$. Прогнозные значения первой и последней можно проследить по таблице В.5 в приложении В и таблице 4.5 соответственно. А также графически на рисунках Б.16 и Б.17.

Таблица 4.5 — Прогнозные значения (модель $\hat{\gamma}_{10}(h)$)

	$X(t)$	$X^*(t)$	$y(t)$	$X(t) - X^*(t)$
2007	19.400	21.170	21.578	-1.770
2008	21.800	21.553	21.687	0.247
2009	21.900	22.151	21.797	-0.251
2010	24.300	22.078	21.906	2.222
2011	22.800	21.665	22.016	1.135
2012	20.200	21.860	22.126	-1.660

Как можно видеть, представленные прогнозные значения не очень точны. Но при этом особенности поведения исходных данных найденные мо-



Рисунок 4.13 — Зависимость качества модели от максимального расстояния

дели 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 исследуемой задаче может присутствовать множество неучтённых факторов, 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 в рассматриваемый месяц год назад.

Для сравнения всех построенных моделей визуальным и автоматическим методами по показателям качества, использовалась сводная таблица 4.6.

Таблица 4.6 — Сводная таблица показателей качества моделей семивариограмм

	$\hat{\gamma}_1(h)$	$\hat{\gamma}_2(h)$	$\hat{\gamma}_3(h)$	$\hat{\gamma}_4(h)$	$\hat{\gamma}_5(h)$	$\hat{\gamma}_6(h)$	$\hat{\gamma}_7(h)$	$\hat{\gamma}_8(h)$	$\hat{\gamma}_9(h)$	$\hat{\gamma}_{10}(h)$
S	202.36	134.38	253.80	598.03	172.66	108.19	132.97	122.72	141.62	167.00
E	1.60	1.07	2.01	4.74	1.37	0.86	1.05	0.97	1.12	1.32
MAE	2.22	1.76	2.42	4.00	2.01	1.42	1.72	1.67	1.74	1.98
MSE	6.32	4.20	7.93	18.69	5.40	3.38	4.16	3.84	4.43	5.22
$r_{\varepsilon\varepsilon^*}$	-0.09	-0.04	0.29	0.15	-0.09	0.40	-0.20	0.20	-0.03	-0.15

В свою очередь, в таблице 4.7 наглядно представлены полученные с помощью каждой из моделей прогнозные значения.

Таблица 4.7 — Сводная таблица реальных $X(t)$ и прогнозных $X_i^*(t), i = \overline{1, 10}$ значений

Год	$X(t)$	$X_1^*(t)$	$X_2^*(t)$	$X_3^*(t)$	$X_4^*(t)$	$X_5^*(t)$	$X_6^*(t)$	$X_7^*(t)$	$X_8^*(t)$	$X_9^*(t)$	$X_{10}^*(t)$
2007	19.40	21.41	21.58	20.82	19.43	21.29	22.13	21.64	22.42	21.38	21.17
2008	21.80	21.52	21.69	21.13	21.86	21.71	23.11	21.61	21.18	21.88	21.55
2009	21.90	21.63	21.80	19.83	21.97	22.15	23.44	21.89	21.66	22.16	22.15
2010	24.30	21.74	21.91	22.13	22.08	22.32	23.03	21.82	22.60	22.22	22.08
2011	22.80	21.85	22.02	22.24	22.19	22.29	22.12	22.09	21.18	22.13	21.66
2012	20.20	21.96	22.13	22.35	22.30	22.29	21.22	22.08	22.62	22.05	21.86

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 семивариограмм, проанализированы оценки Матерона и Кресси-Хокинса, исследованы два подхода по оценке качества модели и подбору моделей и параметров. Как следствие, в рамках рассматриваемой задачи, можно рекомендовать использование перекрёстный метод подбора параметров для вычисления прогнозных значений и адаптивный метод для описания исследуемых данных. И соответствующие модели, найденные этими методами: линейной модели с порогом $\hat{\gamma}_3(h)$ вида (4.8) и периодической модели $\hat{\gamma}_6(h)$ вида (4.12). Следует также отметить, что автоматический метод подбора моделей в рассматриваемых условиях показал результаты хуже, чем визуальный подбор. Но в условиях, когда нужно быстро построить модель семивариограммы, данный метод может оказаться очень полезным. Также, автоматический метод может использоваться для данных, изменение которых имеет более плавный характер. В других случаях, следуют применять визуальных подход по выбору моделей и параметров, поскольку он позволяет наиболее точно учесть все факторы и, как следствие, получить наилучший результат.

Заключение

В дипломной работе исследована температура воды озера Баторино (Беларусь). Исследование проводилось на основе данных, полученных от учебно-научного центра «Нарочанская биологическая станция им. Г.Г.Винберга», состоящих из наблюдений, в период с 1975 по 2012 год в июле месяце.

Были вычислены и проанализированы описательные статистики, проведена проверка на нормальность. Обнаружено, что распределение температуры воды в озере Баторино близко к нормальному закону распределения с параметрами $\mathcal{N}(20.08, 5.24)$. Отклонение от нормальности характеризуется полученными коэффициентами асимметрии и эксцесса, которые свидетельствуют о небольшой скошенности вправо и более пологой колоколообразной форме относительно нормального закона распределения. В результате проведённого корреляционного анализа вычислен коэффициент корреляции и доказана его значимость. Выявлена умеренная прямая зависимость между температурой воды и временем.

Проведён регрессионный анализ, в процессе которого была построена аддитивная модель временного ряда и найдён линейный тренд. Вычислен коэффициент детерминации, проверена значимость регрессионных коэффициентов критерием Стьюдента и адекватность модели критерием Фишера. Построенная детерминированными методами линейная регрессионная модель оказалась значимой и адекватной, но при этом описывает поведение временного ряда лишь частично. Что говорит о наличии некоторых неучтённых данной моделью факторов.

Как следствие удаления тренда из исходных данных получен ряд остатков. В результате его анализа была также показана близость распределения кциальному. С помощью проверки тестов Дики-Фуллера и Льюнга-Бокса показана стационарность в широком смысле и отсутствие автокорреляций в ряде остатков.

Использованы также современные геостатистические методы анализа реального временного ряда. Исследованы свойства оценки вариограммы гауссовского случайногопроцесса: найдены первые два момента, изучено асимптотическое поведение моментов второго порядка. Доказана несмещённость и состоятельность в среднеквадратическом смысле оценки вариограммы.

В рамках вариogramмного анализа были изучены оценки семивариограммы Матерона и Кресси-Хокингса, исследованы различные модели семивариограмм, рассмотрены два подхода по оценке качества модели — перекрёстный и адаптивный, и два метода по подбору моделей и параметров — визуальный и автоматический. Таким образом, наилучшими моделями для семивариограммы временного ряда, представляющего собой ряд остатков,

оказались: линейная с порогом $\hat{\gamma}_3(h)$ вида (4.8) и периодическая $\hat{\gamma}_6(t)$ вида (4.12). Показаны преимущества по использованию перекрёстного метода подбора параметров для описания исследуемых данных и адаптивного метода для вычисления прогнозных значений. На основе построенных моделей семивариограмм с помощью интерполяционного метода кригинг вычислены прогнозные значения временного ряда. Изучена зависимость точности прогноза от оценки вариограммы и модели.

В процессе исследования показано, что в рассматриваемых условиях результаты, полученные автоматическим методом подбора моделей, оказались хуже, чем полученные визуальным подбором. Но при этом, автоматический подбор моделей можно использовать без глубоких знаний в вариограммном анализе, тогда как для визуального подбора требуется применение знаний по статистическому анализу временных рядов и знаний геостатистических методов. Сделаны выводы о возможности использования каждого из них: автоматический подбор следует использовать в случаях, когда нужно быстро получить результат и когда данные имеют плавный характер изменений. В остальных случаях, предпочтительнее использовать визуальный подбор параметров с учётом всех характерных особенностей исследуемых данных.

Результаты дипломной работы получены в среде программирования **R**. В частности, на её основе было реализовано программное обеспечение, позволяющее решать не только данную задачу, но и более широкий класс аналогичных задач в таких областях, как биология, гидрология, природопользование, экология и других. Разработанное приложение имеет простой, быстро расширяемый и гибкий интерфейс, поэтому может быть изменено и дополнено с учётом потребностей конкретного исследования. Для ознакомления, приложение доступно в сети интернет по адресу «apaulau.shinyapps.io/batorino».

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Influence of Long-Distance Climate Teleconnection on Seasonality of Water Temperature in the World's Largest Lake - Lake Baikal, Siberia // PLoS ONE. — 2011. — 02. — Vol. 6, no. 2. — P. e14688.
2. Influence of water temperature on rainbow smelt spawning and early life history dynamics in St. Martin Bay, Lake Huron / T.P. O'Brien, W.W. Taylor, A.S. Briggs, E.F. Roseman // Journal of Great Lakes Research. — 2012. — dec. — Vol. 38, no. 4. — P. 776–785.
3. Subehi L., Fakhrudin M. Preliminary study of the changes in water temperature at pond Cibuntu // Journal of Ecology and the Natural Environment. — 2011. — March. — Vol. 3(3). — P. 72–77.
4. Time series analysis of water surface temperature and heat flux components in the Itumbiara Reservoir (GO), Brazil / Enner Herenio de Alcântara, José Luiz Stech, João Antônio Lorenzzetti, Evelyn Márcia Leão de Moraes Novo // Acta Limnologica Brasiliensis. — 2011. — 09. — Vol. 23. — P. 245 – 259.
5. Mira Chokshi. Temperature analysis for lake Yojoa, Honduras. — 2006.
6. Бриллинджер Д. Временные ряды. Обработка данных и теория. — Мир, 1980.
7. Труш Н.Н. Асимптотические методы статистического анализа временных рядов. — Белгосуниверситет, 1999.
8. Матерон Ж. Основы прикладной геостатистики. — М.: Мир, 1968.
9. Цеховая Т.В. Первые два момента оценки вариограммы гауссовского случайного процесса // Вестник БрГУ им. А.С. Пушкина. — 2005.
10. Ширяев А.Н. Вероятность. — Наука, 1980.
11. Kabacoff Robert. R in Action. — 2009.
12. Teator Paul. R Cookbook (O'Reilly Cookbooks). — 1 edition. — O'Reilly Media, 2011. — 2011. — ISBN: 0596809158.
13. Chang Winston. R graphics cookbook. — " O'Reilly Media, Inc.", 2012.
14. Елисеева И.И. Юзбашев М.М. Общая теория статистики. — Москва : Финансы и статистика, 1995.

15. Cramer Duncan. Basic statistics for social research: step-by-step calculations and computer techniques using Minitab. — Psychology Press, 1997.
16. Cressie Noel AC, Cassie Noel A. Statistics for spatial data. — Wiley New York, 1993. — Vol. 900.
17. Геостатистический анализ данных в экологии и природопользовании (с применением пакета R) / А.А. Савельев, С.С. Мухарамова, А.Г. Пилюгин, Н.А. Чижикова. — Казанский университет, 2012.
18. Bulmer M. G. Principles of Statistics. — Dover Publications, 1979. — ISBN: 0486637603.
19. Sturges H. A. The choice of a class interval // American Statistical Association. — 1926. — Vol. 21. — P. 65–66.
20. Shapiro S. S., Francia R. S. An Approximate Analysis of Variance Test for Normality // Journal of the American Statistical Association. — 1972. — Vol. 67, no. 337. — P. 215–216.
21. Кобзарь А.И. Прикладная математическая статистика. — М.: Физматлит, 2006.
22. Гмурман В.Е. Теория вероятностей и математическая статистика. — Москва : Высшая школа, 2003.
23. Микулик Н.А. Метельский А.В. Теория вероятностей и математическая статистика: Учеб. пособие. — Минск : Пион, 2002.
24. Grubbs F. E. Sample criteria for testing outlying observations // Ann. Math. Statistics. — 1950. — Vol. 21. — P. 27–58.
25. Grubbs Frank E. Procedures for Detecting Outlying Observations in Samples // Technometrics. — 1969. — Vol. 11, no. 1. — P. 1–21.
26. Эддоус М. Стэнсфилд Р. Методы принятия решений. — Москва : Аудит, 1997.
27. Shumway Robert H., Stoffer David S. Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics). — 2nd edition. — Springer, 2006. — ISBN: 0387293175.
28. Dickey David A., Fuller Wayne A. Distribution of the Estimators for Autoregressive Time Series With a Unit Root // Journal of the American Statistical Association. — 1979. — Vol. 74, no. 366. — P. 427–431.

29. Ahmed Shakeel, De Marsily Ghislain. Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity // Water Resources Research. — 1987. — Vol. 23, no. 9. — P. 1717–1737.
30. Zquia Eulogio Pardo-igu. Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography // Int. J. Climatol. — 1998. — Vol. 18. — P. 1031–1047.
31. Dutter Rudolf. On Robust Estimation of Variograms in Geostatistics // Robust Statistics, Data Analysis, and Computer Intensive Methods / Ed. by Helmut Rieder. — Springer New York, 1996. — Vol. 109 of Lecture Notes in Statistics. — P. 153–171.
32. Mingoti Sueli Aparecida, Rosa Gilmar. A note on robust and non-robust variogram estimators // Rem: Revista Escola de Minas. — 2008. — 03. — Vol. 61. — P. 87 – 95.
33. Савельева Е.А., Демьянин В.В. Геостатистика: теория и практика. — Ин-т проблем безопасного развития атомной энергетики РАН. — М.: Наука, 2010.
34. Cressie Noel, Wikle Christopher K. Statistics for spatio-temporal data. — John Wiley & Sons, 2011.
35. Pebesma Edzer J. Gstat user's manual // Dept. of Physical Geography, Utrecht University, Utrecht, The Netherlands. — 2001.

Приложение А

Исходные данные

Таблица А.1 — Исходные данные

Год	Температура, °C
1975	20.20
1976	16.00
1977	17.70
1978	16.75
1979	17.50
1980	16.77
1981	19.80
1982	19.00
1983	21.40
1984	19.40
1985	20.40
1986	16.50
1987	17.10
1988	23.80
1989	19.90
1990	18.50
1991	23.00
1992	21.90
1993	18.00
1994	21.40
1995	18.90
1996	19.10
1997	21.00
1998	18.40
1999	23.50
2000	21.00
2001	24.20
2002	23.10
2003	18.00
2004	19.10
2005	20.00
2006	21.30
2007	19.40
2008	21.80
2009	21.90
2010	24.30
2011	22.80
2012	20.20

Приложение Б

Графические материалы

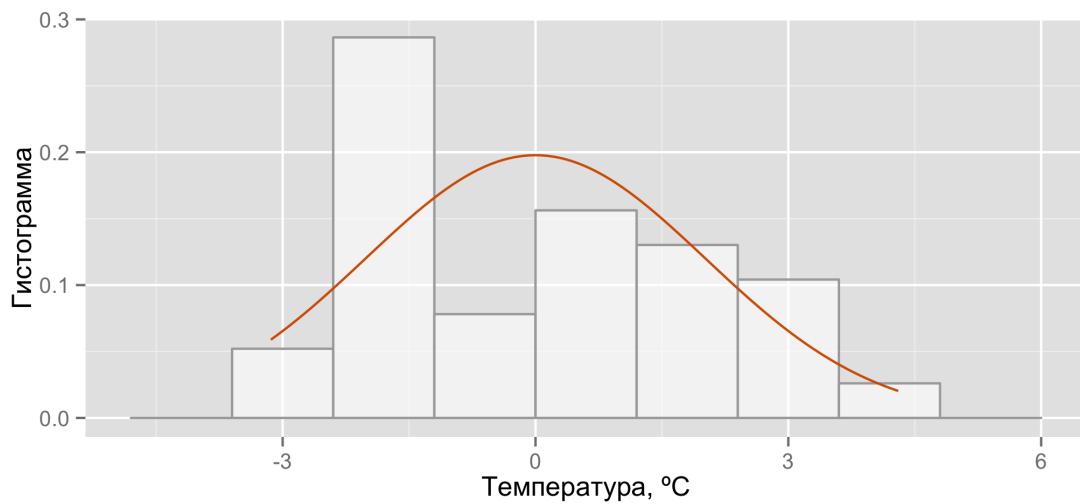


Рисунок Б.1 — Гистограмма остатков с кривой плотности нормального распределения $\mathcal{N}(19.88, 4.92)$

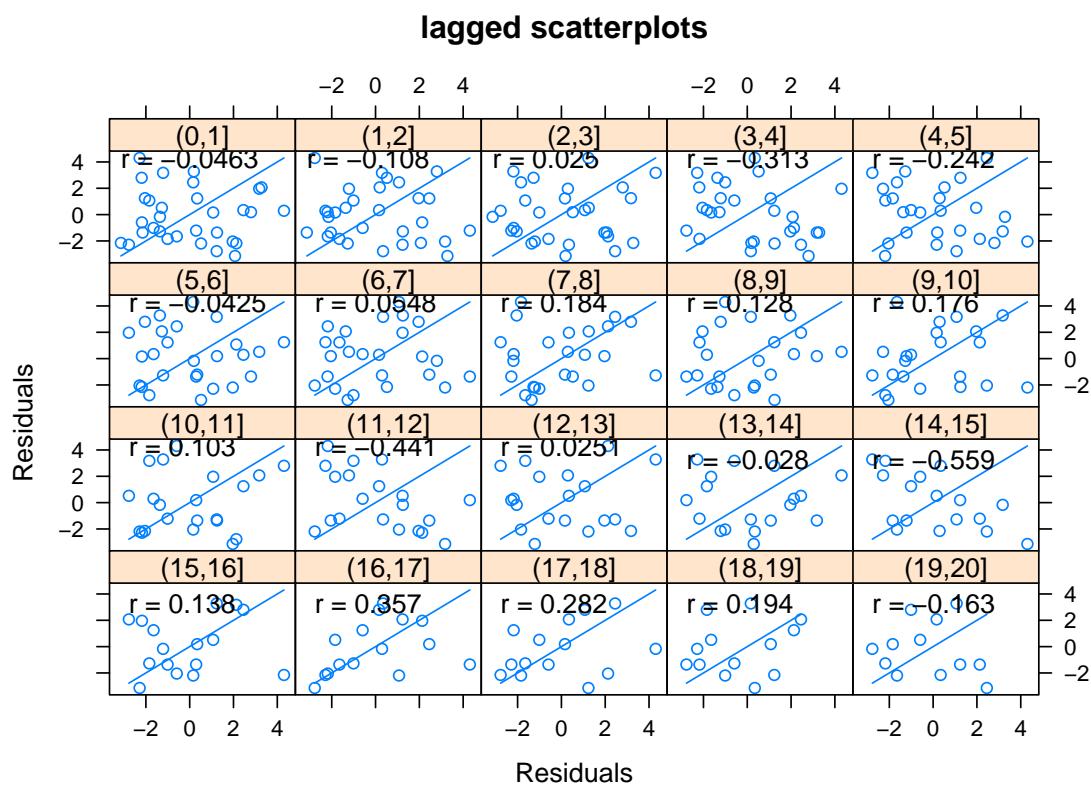


Рисунок Б.2 — Диаграмма взаимного разброса

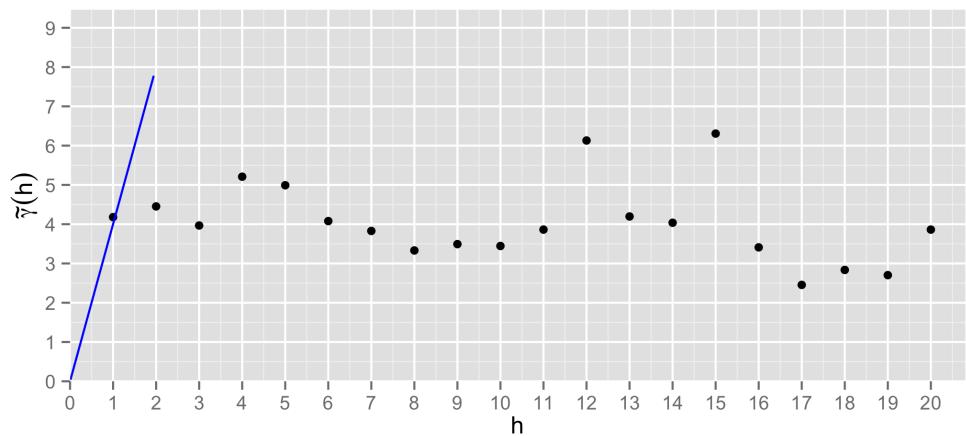


Рисунок Б.3 — Семивариограмма и оценка $\hat{\gamma}_1(h)$

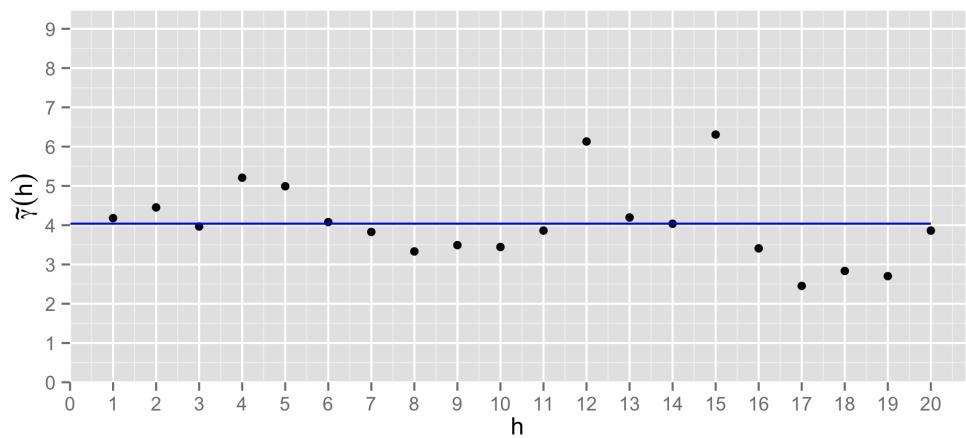


Рисунок Б.4 — Семивариограмма и оценка $\hat{\gamma}_2(h)$

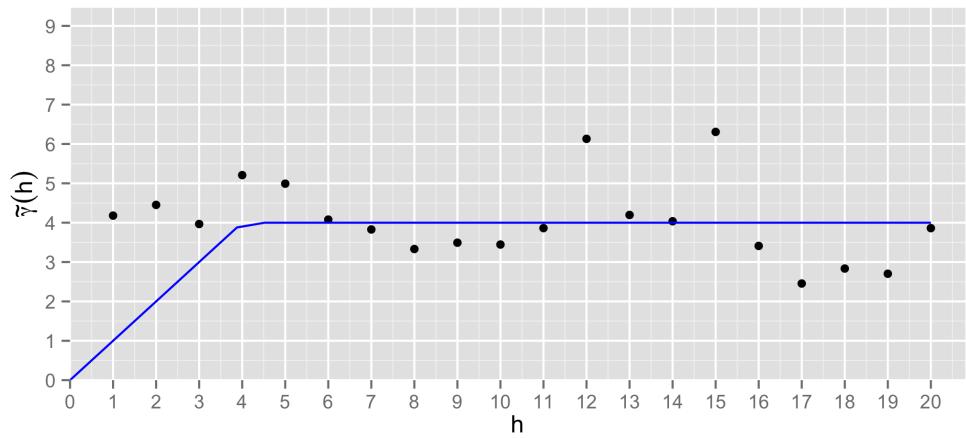


Рисунок Б.5 — Семивариограмма и оценка $\hat{\gamma}_3(h)$

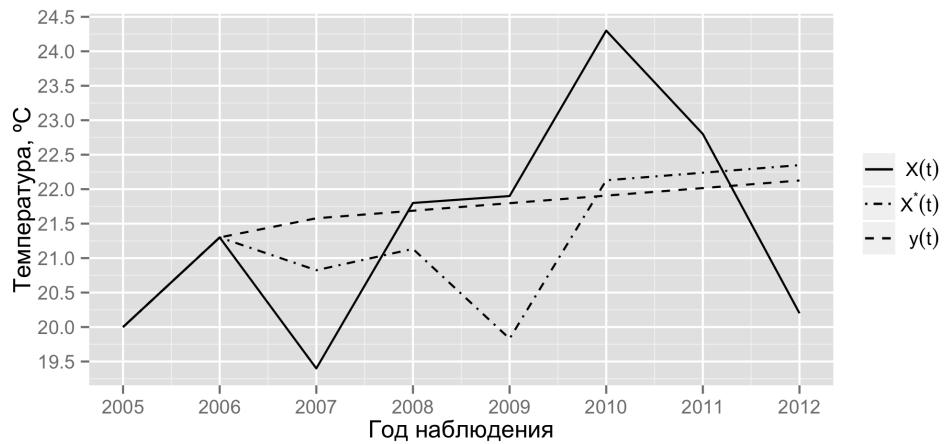


Рисунок Б.6 — Прогноз (модель $\hat{\gamma}_3(h)$)

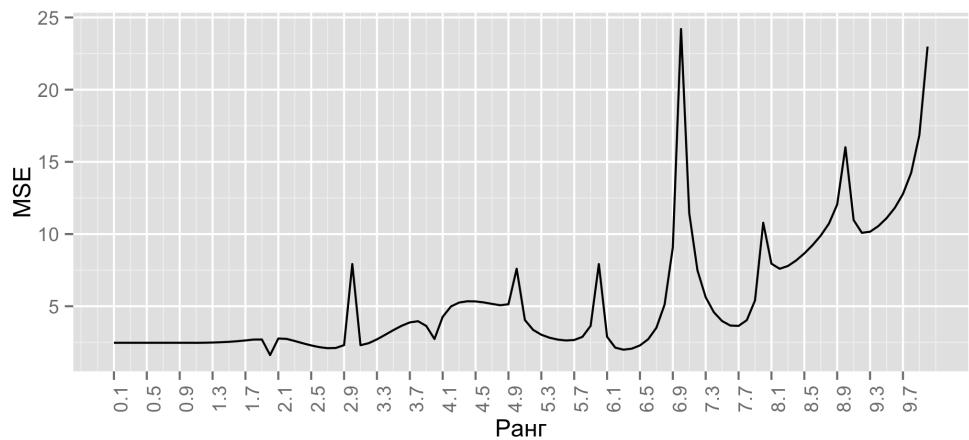


Рисунок Б.7 — Зависимость качества линейной модели от значения ранга

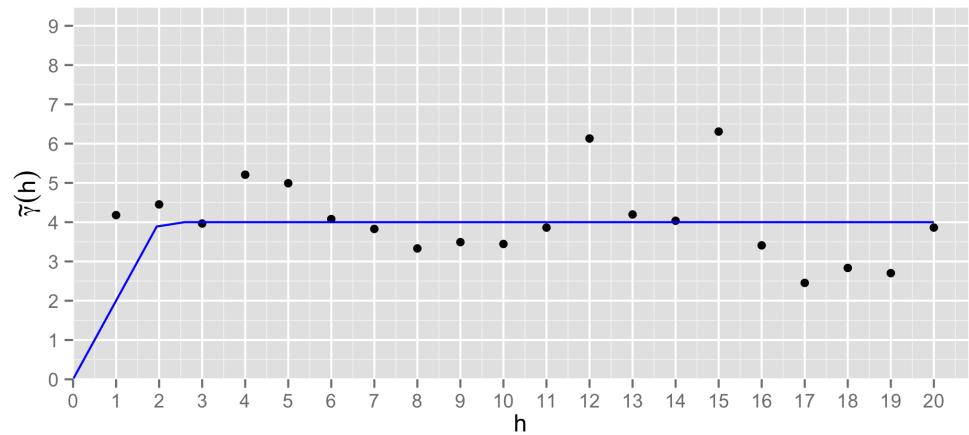


Рисунок Б.8 — Семивариограмма и оценка $\hat{\gamma}_4(h)$

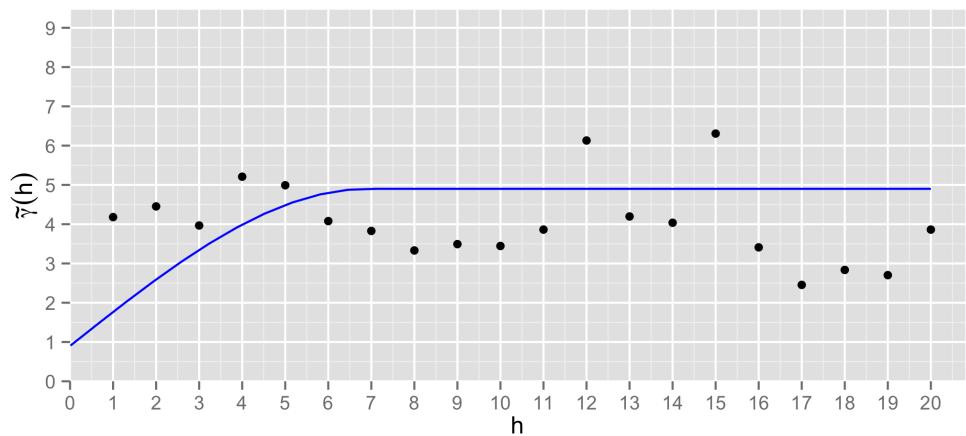


Рисунок Б.9 — Семивариограмма и оценка $\widehat{\gamma}_5(h)$

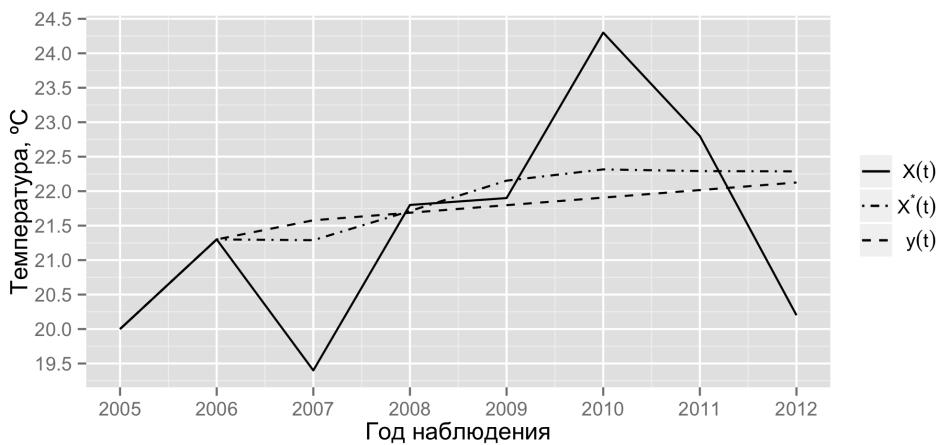


Рисунок Б.10 — Прогноз (модель $\widehat{\gamma}_5(h)$)

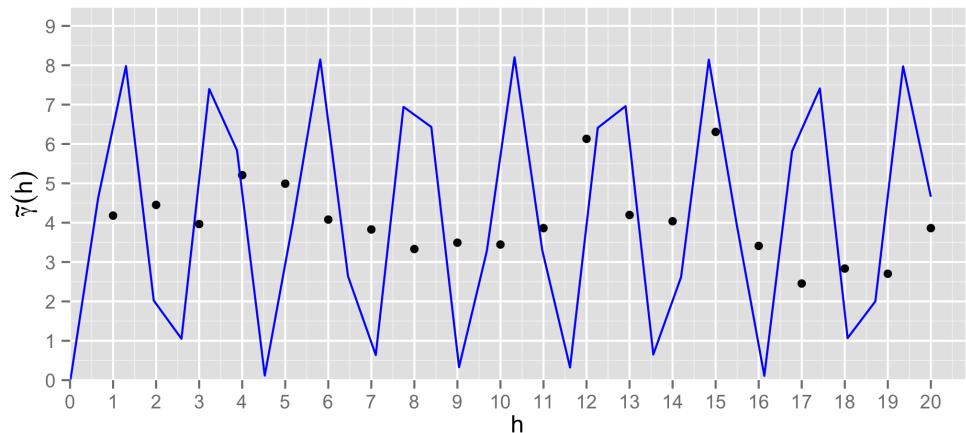


Рисунок Б.11 — Семивариограмма и оценка $\widehat{\gamma}_6(h)$

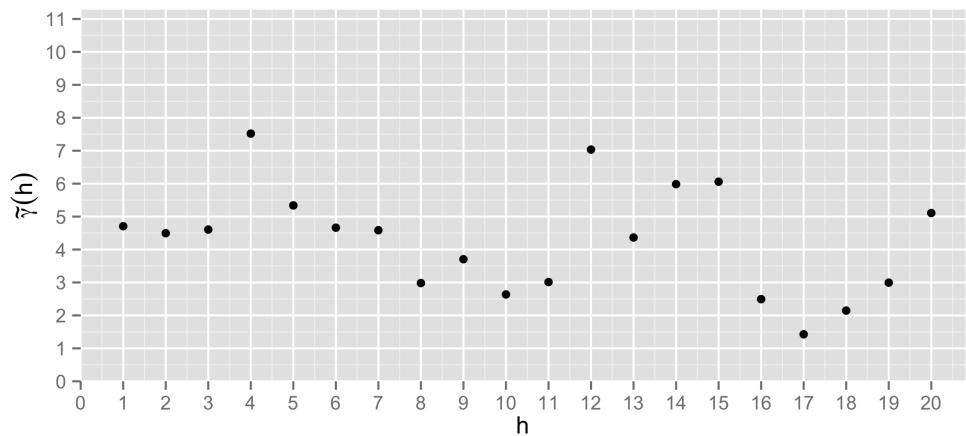


Рисунок Б.12 – Оценка семивариограммы Кресси-Хокинса

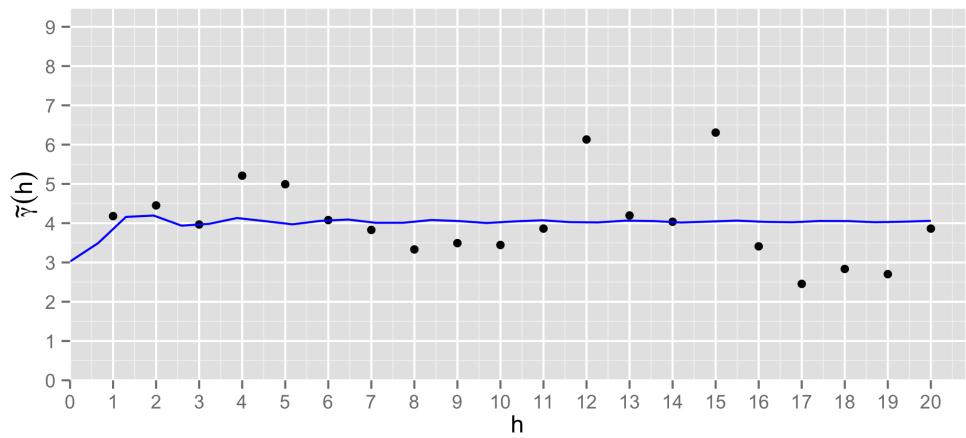


Рисунок Б.13 – Семивариограмма и оценка $\hat{\gamma}_7(h)$

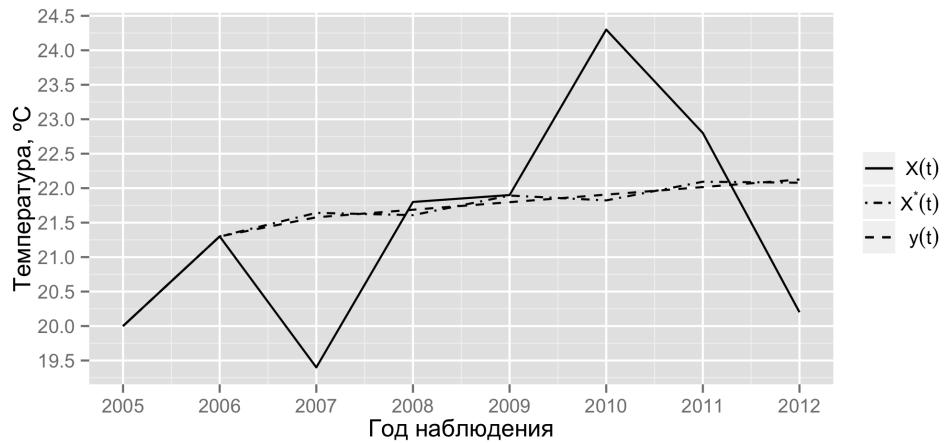


Рисунок Б.14 – Прогноз (модель $\hat{\gamma}_7(h)$)

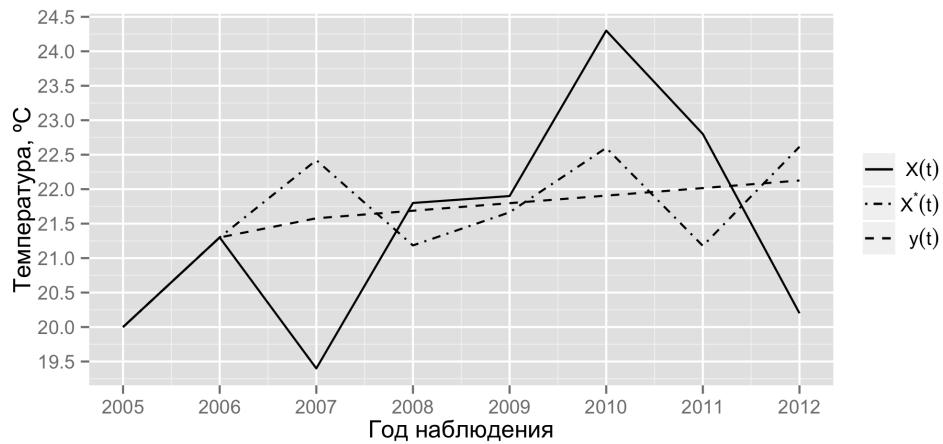


Рисунок Б.15 — Прогноз (модель $\hat{\gamma}_8(h)$)

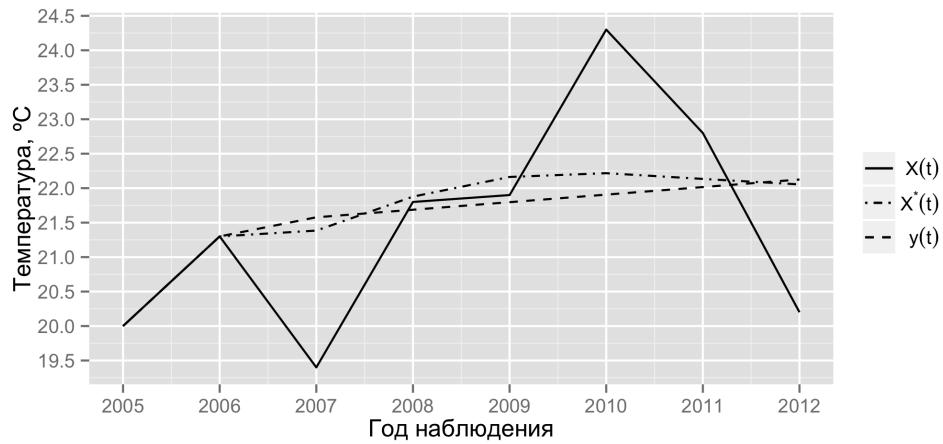


Рисунок Б.16 — Прогноз (модель $\hat{\gamma}_9(h)$)

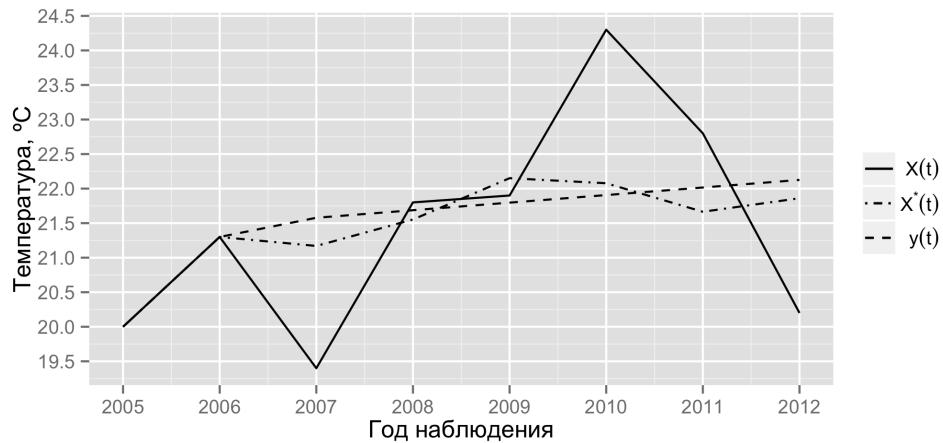


Рисунок Б.17 — Прогноз (модель $\hat{\gamma}_{10}(h)$)

Приложение В

Результаты вычислений

Таблица В.1 — Временной ряд остатков

Год	Температура, °C
1975	2.13
1976	-2.18
1977	-0.59
1978	-1.65
1979	-1.01
1980	-1.84
1981	1.07
1982	0.16
1983	2.45
1984	0.34
1985	1.23
1986	-2.78
1987	-2.29
1988	4.30
1989	0.29
1990	-1.21
1991	3.18
1992	1.97
1993	-2.04
1994	1.25
1995	-1.36
1996	-1.27
1997	0.52
1998	-2.19
1999	2.80
2000	0.19
2001	3.28
2002	2.07
2003	-3.14
2004	-2.15
2005	-1.36
2006	-0.17

Таблица В.2 — Прогнозные значения (модель $\hat{\gamma}_2(h)$)

	$X(t)$	$X^*(t)$	$y(t)$	$X(t) - X^*(t)$
2007	19.400	21.578	21.578	-2.178
2008	21.800	21.687	21.687	0.113
2009	21.900	21.797	21.797	0.103
2010	24.300	21.906	21.906	2.394
2011	22.800	22.016	22.016	0.784
2012	20.200	22.126	22.126	-1.926

Таблица В.3 — Прогнозные значения (модель $\hat{\gamma}_4(h)$)

	$X(t)$	$X^*(t)$	$y(t)$	$X(t) - X^*(t)$
2007	19.400	19.427	21.578	-0.027
2008	21.800	21.864	21.687	-0.064
2009	21.900	21.974	21.797	-0.074
2010	24.300	22.084	21.906	2.216
2011	22.800	22.193	22.016	0.607
2012	20.200	22.303	22.126	-2.103

Таблица В.4 — Прогнозные значения (модель $\hat{\gamma}_6(h)$)

	$X(t)$	$X^*(t)$	$y(t)$	$X(t) - X^*(t)$
2007	19.400	22.133	21.578	-2.733
2008	21.800	23.106	21.687	-1.306
2009	21.900	23.441	21.797	-1.541
2010	24.300	23.028	21.906	1.272
2011	22.800	22.122	22.016	0.678
2012	20.200	21.219	22.126	-1.019

Таблица В.5 — Прогнозные значения (модель $\hat{\gamma}_9(h)$)

	$X(t)$	$X^*(t)$	$y(t)$	$X(t) - X^*(t)$
2007	19.400	21.385	21.578	-1.985
2008	21.800	21.877	21.687	-0.077
2009	21.900	22.163	21.797	-0.263
2010	24.300	22.217	21.906	2.083
2011	22.800	22.134	22.016	0.666
2012	20.200	22.055	22.126	-1.855

Приложение Г

Код программы

```
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(
7     data),
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     ,
23     "Минимум", "Максимум", "Размах", "Квартильный размах",
24     "Дисперсия", "Стандартное отклонение", if(!is.na(cv)) {"Коэффициент вариации"},
25     "Стандартная ошибка", "Асимметрия", "Ошибка асимметрии",
26     "Эксцесс", "Ошибка эксцесса")
27     descr.col <- c("Значение")
28   } else {
29     descr.row <- c("Mean", "Median", "Lower Quartile", "Upper Quartile", "Range",
30     "Minimum", "Maximum", "Quartile Range", "Variance", "Standard Deviation",
31     if (!is.na(cv)) {"Coefficient of Variance"}, "Standard Error", "Skewness",
32     "Std. Error Skewness", "Kurtosis", "Std. Error Kurtosis")
33     descr.col <- c("Value")
34   }
35   if (!shiny) {
36     df <- data.frame(stats, row.names=descr.row)
37     colnames(df) <- descr.col
38   } else {
39     df <- data.frame(descr.row, sapply(stats, format, digits=2, scientific=
40       FALSE, nsmall=1))
41     colnames(df) <- c("Статистика", "Значение")
42   }
43
44   df
45
46   dstats.mean <- function(data, ...) {
47     m <- mean(data, ...)
48     if (m < .0000001) {
49       m <- 0
50     }
51     m
52   }
53 }
```

```

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=
132     "mean")
133   WriteDescriptiveStatistic(expression=dstats.variance(data), type=type, name=
134     "variance")
135   WriteDescriptiveStatistic(expression=paste(format(dstats.coef.var(data),
136     nsmall=2, digits=4), "\\\%"), type=type, name="coef-var")
137   WriteDescriptiveStatistic(expression=dstats.skew(data), type=type, name=
138     "skew")
139   WriteDescriptiveStatistic(expression=dstats.kurtosis(data), type=type, name=
140     "kurtosis")
141   WriteDescriptiveStatistic(expression=dstats.test.skew(data), type=type, name=
142     "test-skew")
143   WriteDescriptiveStatistic(expression=dstats.test.kurtosis(data), type=type,
144     name="test-kurtosis")
145 }

```

Листинг Г.1: Описательные статистики

```

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",
4     "Per", "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 specifiy a variogram model in GLS.model the Generelised Least
10  # Squares sample variogram is constructed
11  if(!is(GLS.model, "variogramModel")){
12    experimental_variogram = variogram(formula, input_data, cutoff=cutoff,
13      width=width, cressie=cressie, ...)
14  } else {
15    g = gstat(NULL, "bla", formula, input_data, model=GLS.model, set=list(gls
16      =1))
17    experimental_variogram = variogram(g, cutoff=cutoff, width=width, cressie=
18      TRUE, ...)
19  }

```

```

16 # set initial values
17 if(is.na(start_vals[1])) { # Nugget
18   initial_nugget = min(experimental_variogram$gamma)
19 } else {
20   initial_nugget = start_vals[1]
21 }
22 if(is.na(start_vals[2])) { # Range
23   diagonal = spDists(t(bbox(input_data)))[1,2] # 0.35 times the length of
24   # the central axis through the area
25   initial_range = 0.1 * diagonal # 0.10 times the length of the central
26   # axis through the area
27 } else {
28   initial_range = start_vals[2]
29 }
30 if(is.na(start_vals[3])) { # Sill
31   initial_sill = mean(c(max(experimental_variogram$gamma), median(
32     experimental_variogram$gamma)))
33 } else {
34   initial_sill = start_vals[3]
35 }
36
37 # Determine what should be automatically fitted and what should be fixed
38 # Nugget
39 if(!is.na(fix.values[1])) {
40   fit_nugget = FALSE
41   initial_nugget = fix.values[1]
42 } else {
43   fit_nugget = TRUE
44 }
45
46 # Range
47 if(!is.na(fix.values[2])) {
48   fit_range = FALSE
49   initial_range = fix.values[2]
50 } else {
51   fit_range = TRUE
52 }
53
54 # Partial sill
55 if(!is.na(fix.values[3])) {
56   fit_sill = FALSE
57   initial_sill = fix.values[3]
58 } else {
59   fit_sill = TRUE
60 }
61
62 getModel <- function(psill, model, range, kappa, nugget, fit_range, fit_sill,
63   , fit_nugget) {
64   if(model == "Pow") {
65     if(is.na(start_vals[1])) nugget = 0
66     if(is.na(start_vals[2])) range = 1 # If a power mode, range == 1 is a
67     # better start value
68     if(is.na(start_vals[3])) sill = 1
69   }
70   if(model == "Nug") {
71     if(is.na(start_vals[2])) range = 0
72   }
73
74   obj = try(fit.variogram(experimental_variogram,
75     model = vgm(psill=psill, model=model, range=range,

```

```

71     nugget=nugget ,kappa = kappa) ,
72     fit.ranges = c(fit_range) , fit.sills = c(fit_nugget , fit_sill) ,
73     debug.level=0, fit.method = 6) ,
74     silent=TRUE)
75 if("try-error" %in% class(obj)) {
76   #print(traceback())
77   if (verbose) {
78     warning("An error has occured during variogram fitting. Used:\n",
79             "\tnugget:\t", nugget,
80             "\n\tmodel:\t", model,
81             "\n\tpsill:\t", psill ,
82             "\n\trange:\t", range ,
83             "\n\tkappa:\t", ifelse(kappa == 0, NA, kappa) ,
84             "\n  as initial guess. This particular variogram fit is not taken
85             into account. \nGstat error:\n", obj)
86   }
87   return(NULL)
88 } else return(obj)
89 }

# Automatically testing different models, the one with the smallest sums-of-
# squares is chosen
90 SSerr_list = c()
91 vgm_list = list()
92 counter = 1

93
94 for(m in test_models) {
95   if(m != "Mat" && m != "Ste") {           # If not Matern and not Stein
96     model_fit = getModel(initial_sill - initial_nugget, m, initial_range,
97                           kappa = 0, initial_nugget, fit_range, fit_sill, fit_nugget)
98     if(!is.null(model_fit)) { # skip models that failed
99       vgm_list[[counter]] = model_fit
100      SSerr_list = c(SSerr_list , attr(model_fit , "SSErr"))
101    }
102    counter = counter + 1
103  } else {                                # Else loop also over kappa values
104    for(k in kappa) {
105      model_fit = getModel(initial_sill - initial_nugget, m, initial_range,
106                            k, initial_nugget, fit_range, fit_sill, fit_nugget)
107      if(!is.null(model_fit)) {
108        vgm_list[[counter]] = model_fit
109        SSerr_list = c(SSerr_list , attr(model_fit , "SSErr"))
110      }
111      counter = counter + 1
112    }
113  }
114}

# Check for negative values in sill or range coming from fit.variogram
# and NULL values in vgm_list , and remove those with a warning
115 strange_entries = sapply(vgm_list , function(v) any(c(v$psill , v$range) < 0
116                           | is.null(v)))
117 if(any(strange_entries)) {
118   if(verbose) {
119     print(vgm_list [strange_entries])
120     cat("^^^ ABOVE MODELS WERE REMOVED ^^^\n\n")
121   }
122   SSerr_list = SSerr_list [!strange_entries]
123   vgm_list = vgm_list [!strange_entries]
124 }

```

```

126
127 if(verbose) {
128   cat("Selected:\n")
129   print(vgm_list [[which.min(SSerr_list)]])
130   cat("\nTested models, best first:\n")
131   tested = data.frame("Tested models" = sapply(vgm_list , function(x) as.
132     character(x[2,1])),
133     kappa = sapply(vgm_list , function(x) as.character(x[2,4])),
134     "SSerror" = SSerr_list)
135   tested = tested[order(tested$SSerror), ]
136   print(tested)
137 }
138 result = list(exp_var = experimental_variogram , var_model = vgm_list [[which.
139   min(SSerr_list)]], sserr = min(SSerr_list , na.rm=TRUE))
140 class(result) = c("autofitVariogram" , "list")
141 return(result)
142 }
```

Листинг Г.2: Автоматический подбор моделей