МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Павлов Александр Сергеевич

Анализ и статитистическая обработка временных рядов в пакете R

Отчет о прохождении преддипломной практики

Руководитель практики

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

Минск, 2015

Содержание

Введение

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

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

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

Среди представленных следует отметить работу [1], где в качестве объекта исследования рассматривается крупнейшее в мире озеро — Байкал, подробно изучается изменение климата в контексте данного озера в период с 1950 по 2012 гг.

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

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

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

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

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

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

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

Глава 1
Случайный процесс и его характеристики. Стационарность случайных процессов. Вариограмма

1.1 Случайный процесс. Стационарность

Для введения следующих понятий воспользуемся [6, 7].

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

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

Определение 1.2. 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 1.3. Если $\mathbb{T} = \mathbb{R}$, то $X(t), t \in \mathbb{T}$ называют *случайным процессом с непрерывным 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 $x_j \in \mathbb{T}, t_j \in \mathbb{T}, j = \overline{1, n}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2 Вариограмма и внутренне стационарный случайный процесс

Определение 1.11. Случайный процесс $X(t), t \in \mathbb{R} = (-\infty, +\infty)$, называется внутренне стационарным, если справедливы следующие равенства:

$$E\{X(t_1) - X(t_2)\} = 0, \quad V\{X(t_1) - X(t_2)\} = 2\gamma(t_1 - t_2),$$

где $2\gamma(t_1 - t_2)$ — вариограмма рассматриваемого процесса, $t_1, t_2 \in \mathbb{R}$. Заметим, что функция $\gamma(t), t \in \mathbb{R}$, — называется семивариограммой.

В дальнейшем рассматриваем случайные процессы с дискретным временем.

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

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

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

Заметим, что

$$E\{X(t+h) - X(t)\}^2 = 2\gamma(h), \tag{1.1}$$

$$V\{X(t+h) - X(t)\}^2 = 2(2\gamma(h))^2. \tag{1.2}$$

В качестве оценки вариограммы рассмотрим статистику вида:

$$2\tilde{\gamma}(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(X(t+h) - X(t))^2, \quad h = \overline{0, n-1}, \tag{1.3}$$

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

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

Рассмотрим внутренне стационарный гауссовский случайный процесс с дискретным временем $X(t), t \in \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$.

Не нарушая общности, далее считаем $m(t) \equiv 0$, $V(t) \equiv \sigma^2, t \in \mathbb{Z}$.

Вариограмма процесса $X(t)$,

$$2\gamma(h) = V\{X(t+h) - X(t)\},\ t, h \in \mathbb{Z},$$

является неизвестной.

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

В качестве оценки вариограммы рассмотрим статистику вида:

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

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

2.1 Математическое ожидание оценки вариограммы

Вычислим математическое ожидание введённой оценки (2.1):

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

Из равенства (1.1) 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) для вариограммы рассматриваемой модели является **несмещённой**.

2.2 Второй момент оценки вариограммы

Найдём ковариацию оценок вариограммы при различных значениях h:

$$\begin{aligned} cov\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= E\{(2\tilde{\gamma}(h_1) - E\{2\tilde{\gamma}(h_1)\})(2\tilde{\gamma}(h_2) - E\{2\tilde{\gamma}(h_2)\})\} = \\ &= E\{\tfrac{1}{n-h_1}\textstyle\sum_{t=1}^{n-h_1}((X(t+h_1) - X(t))^2 - E\{(X(t+h_1) - X(t))^2\}) \times \\ &\times \tfrac{1}{n-h_2}\textstyle\sum_{s=1}^{n-h_2}((X(s+h_2) - X(s))^2 - E\{(X(s+h_2) - X(s))^2\})\} = \\ &= \tfrac{1}{(n-h_1)(n-h_2)}\textstyle\sum_{t=1}^{n-h_1}\sum_{s=1}^{n-h_2} cov\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} \end{aligned} \tag{2.2}$$

Из свойства 1.1 корреляции получаем, что

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

Принимая во внимание (1.2) и предыдущее соотношение, из (2.2) получаем:

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

Далее воспользуемся леммой 1 из [8]:

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

Воспользовавшись леммой 3 из [8] и определением корреляционной функции, получаем соотношение

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

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

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

2.3 Новый раздел (без названия)

В (2.3) сделаем следующую замену переменных

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

Получим

$$cov\{2\tilde{\gamma}(h_1),2\tilde{\gamma}(h_2)\} = \\ = \frac{2}{(n-h_1)(n-h_2)}\sum_{t=1}^{n-h_1}\sum_{t-m=1}^{n-h_2}(\gamma(m-h_2)+\gamma(m+h_1)-\gamma(m)-\gamma(m+h_1-h_2))^2 \quad (2.4)$$

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

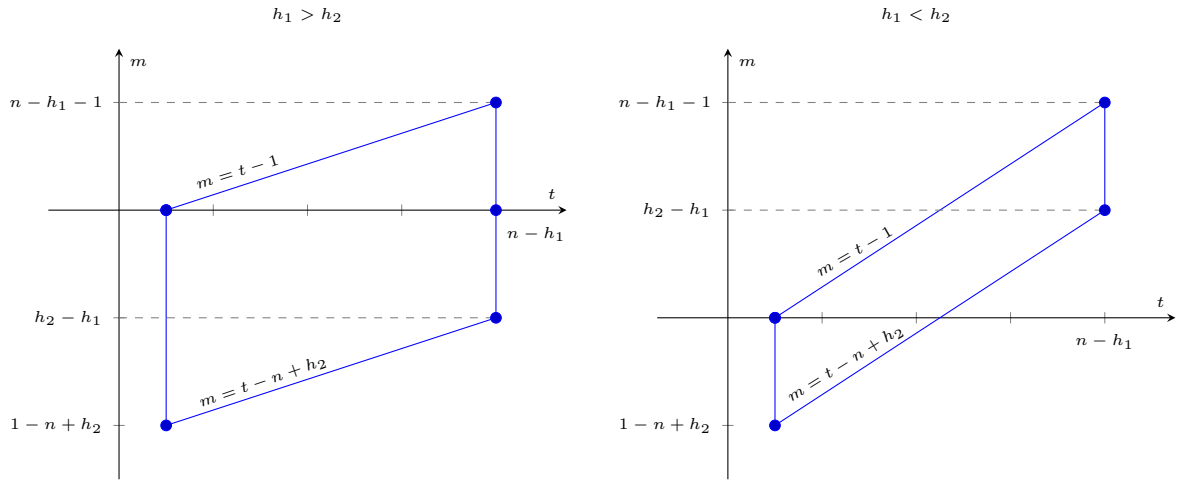

Рисунок 2.3.1 — Замена переменных

Рассмотрим первый случай: $h_1 > h_2$. Поменяем порядок суммирования в (2.4).

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

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

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

Преобразуем полученное выражение:

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

Приведем подобные:

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

Глава 3
Обработка временного ряда с помощью R

3.1 Вычисление основных описательных статистик

В качестве исходных данных примем выборку из полученной от учебно-научного центра базы данных, путём отбора наблюдений в июле месяце за период с 1975 по 2012 год. Выборка представлена в приложении А в таблице А.1. Графически исходные данные представлены на рисунке 3.1.1.

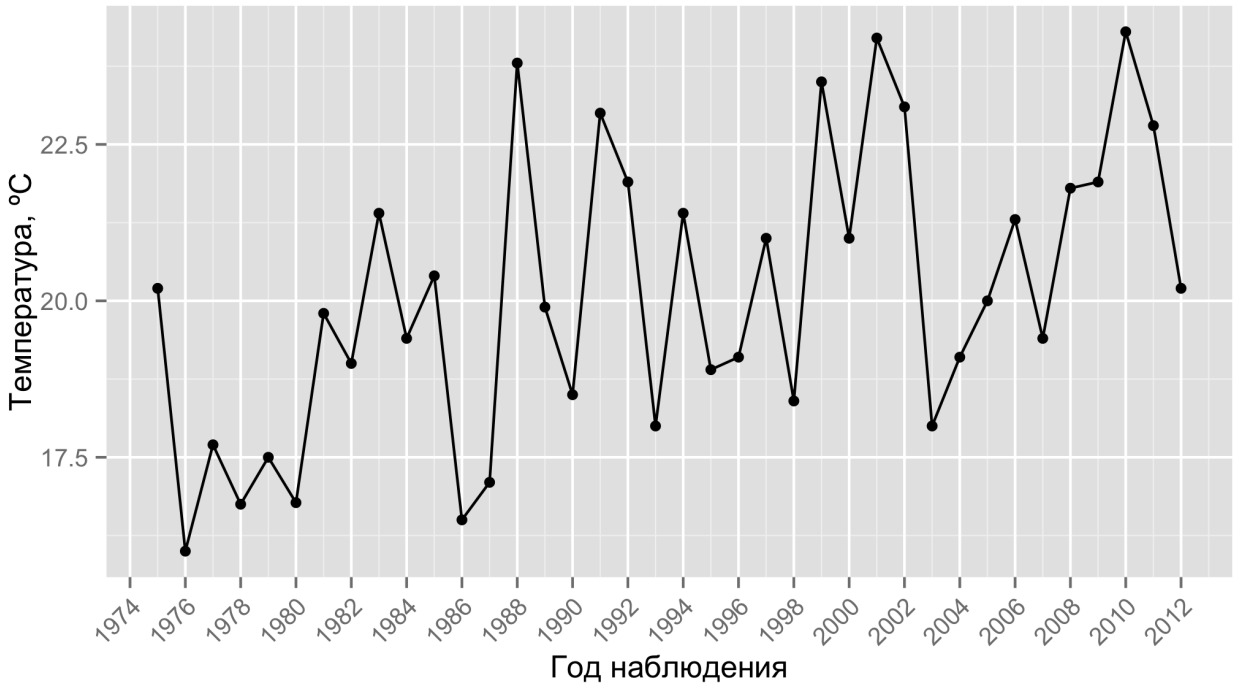

Рисунок 3.1.1 — График исходных данных.

Следует отметить, что для непосредственного исследования были использованы наблюдения с 1975 по 2009 год. Наблюдения за 2010-2012 годы были намеренно исключены из исследования в целях дальнейшего оценивания результатов анализа и прогнозирования. Заметим, что работа, представленная в параграфах 3.1–3.4, была также проделана и для всей выборки. Так как поведение целой выборки сохранилось в уменьшенной, то, без потери общности, будем считать её исходной.

Начнём исследование временного ряда с вычисления описательных статистик. **R** предоставляет в пакете *base* различные функции для расчетов базовых статистик. Также, в различных пакетах можно найти другие интересующие функции, как статистические, так и математические. Но в целях удобства, компактности и контроля за функциональностью на основе [9, 10] мной был написан модуль *dstats*, представленный в приложении С листинге С.1. Данный модуль позволяет вычислять все рассмотренные в данной работе описательные статистики. Полученные результаты для исходных данных отображены в таблице 1.

Рассмотрим подробнее некоторые полученные статистики.

Как видно из таблицы, *средняя* температура в июле месяце за период с 1975 по 2009

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

Таблица 1 — Описательные статистики для наблюдаемых температур.

составляет приблизительно 20°C. При этом *размах* температур равен 8.2°C, а *дисперсия* равна 4.91.

Стандартное отклонение оказалось равным приблизительно 2.21. Полученное значение не велико, а значит можно сказать, что среднее значение хорошо описывает выборку.

Коэффициент вариации в нашем случае равен 24.7%. Из этого следует, что выборку можно считать однородной, так как полученное значение является меньшим 33% [9].

Стандартная ошибка среднего значения равна 0.37.

Коэффициент асимметрии — мера симметричности распределения. Полученное значение: 0.18. Данное значение говорит о незначительной правосторонней асимметрии распределения. То есть о том, что выборочное распределение можно считать близким к симметричному [11].

Стандартная ошибка асимметрии равна 0.40.

Коэффициент эксцесса в рассматриваемом случае равен -0.85. Так как коэффициент эксцесса нормального распределения равен 0, то в данном случае можно говорить о пологости пика распределения выборки по отношению к нормальному распределению [11].

Стандартная ошибка коэффициента эксцесса равна 0.77.

С помощью тестовых статистик для коэффициента асимметрии и эксцесса [10, с.85-89], проверим значимость полученных значений для генеральной совокупности. Для этого в модуле *dstats* мной реализованы функции *dstats.test.skew* и *dstats.test.kurtosis*:

Полученная тестовая статистика для коэффициента асимметрии:

$$Z_{A_S} = \frac{A_S}{SES} = 0.4648153$$

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 [10, с.85].

Полученная тестовая статистика для коэффициента эксцесса:

$$Z_K = \frac{K}{SEK} = -1.015476.$$

Данное значение попадает под случай $|Z_K| \leq 2$, а значит, в данном случае выборочный коэффициент эксцесса не является значимым и нельзя ничего сказать о коэффициенте эксцесса генеральной совокупности [10, с.89].

Из полученных результатов следует отметить, что коэффициенты асимметрии и эксцесса, указывают на близость выборочного распределения к нормальному закону. Но при этом, из-за недостаточного объёма выборки, по этим коэффициентам нельзя судить о соответствующих коэффициентах генеральной совокупности.

3.2 Исследование статистических данных

В **R** можно найти различные пакеты, позволяющие строить разнообразные гистограммы, диаграммы рассеяния, вероятностные графики, линейные графики, диаграммы диапазонов, размахов, круговые диаграммы, столбчатые диаграммы, последовательные графики и т.д., позволяющие увидеть специфику данных. В процессе работы в контексте **R** использовались источники [12–14].

С помощью функции пакета *ggplot2* построим гистограмму для отображения вариационного ряда исходных данных [14]. Гистограммы позволяют увидеть, как распределены значения переменных по интервалам группировки, то есть как часто переменные принимают значения из различных интервалов. А также, что бывает более важным, повзоляет сделать предположение о разновидности распределения. Для вычисления интервалов разбиения воспользуемся *формулой Стерджеса*. Из [15] количество интервалов разбиения рассчитывается по формуле:

$$k = \lceil log_2 n \rceil + 1 = \lceil log_2 38 \rceil + 1 = 7. \tag{3.1}$$

Так как по гистограмме можно визуально предположить близость выборочного распределения к нормальному распределению, нанесём на график кривую плотности нормального распределения. Построенная гистограмма отображена на рисунке 3.2.2. Проанализируем эту гистограмму. Во-первых, на ней наглядно представлена близость выборочного распределения к нормальному с параметрами $\mathcal{N}(19.88, 4.91)$. Во-вторых, по этой гистограмме можно подтвердить или опровергнуть результаты, полученные на этапе вычисления описательных статистик в параграфе 3.1.

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

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

В процессе данной работы была написана функция *gqqqp*, с помощью которой построен рисунок 3.2.3. На этом графике можно визуально обнаружить аномальное положение

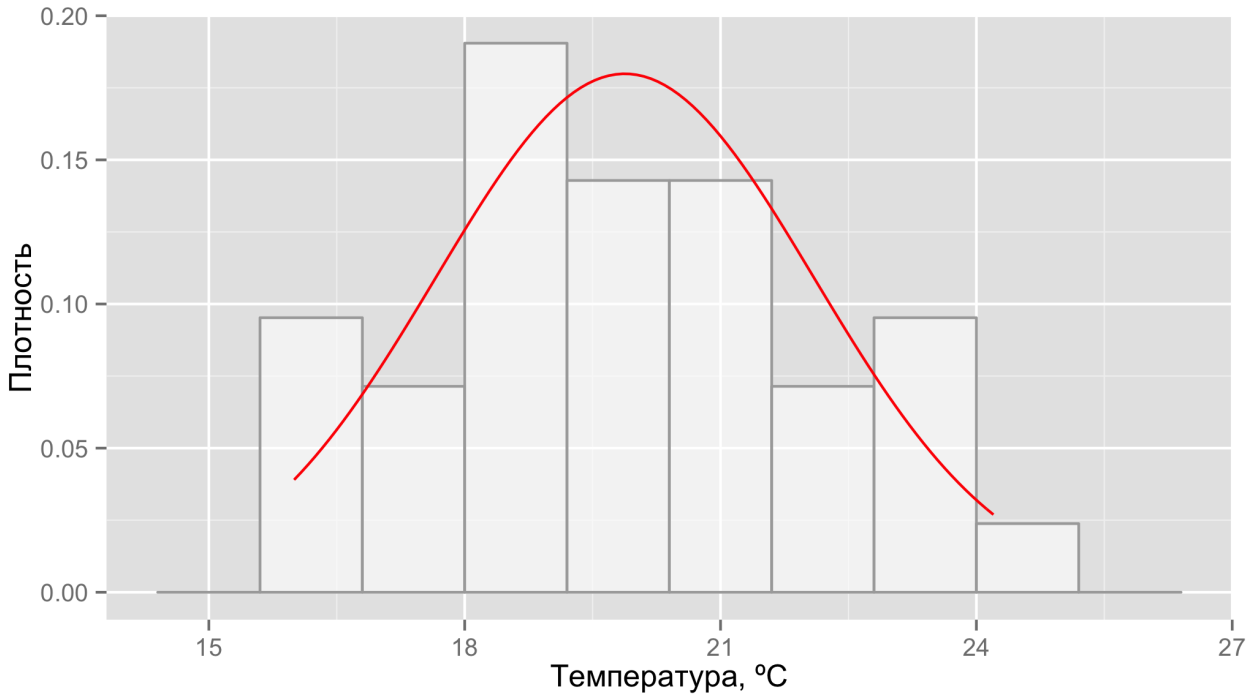

Рисунок 3.2.2 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения

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

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

В **R** реализованы практически все имеющиеся тесты на нормальность — либо в виде стандарных функций, либо в виде функций, входящих в состав отдельных пакетов. Примером базовой функции является *shapiro.test()*, при помощи которой можно выполнить широко используемый *тест Шапиро-Уилка* [16]:

```
Shapiro-Wilk normality test

data:  data
W = 0.9742, p-value = 0.5685
```

В полученных результатах W — статистика Шапиро-Уилка. Вероятность ошибки $p = 0.5685 > 0.05$, а значит нулевая гипотеза не отвергается [17]. Следовательно опровергнуть предположение на основе данного теста нельзя.

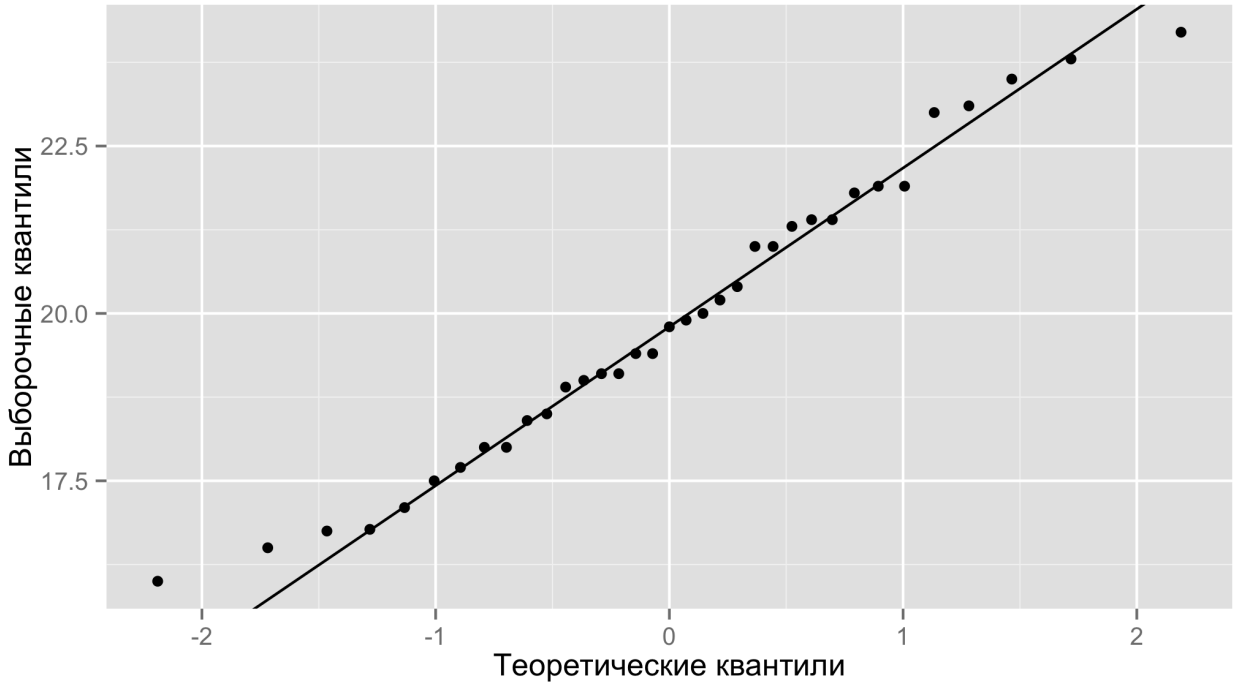

Рисунок 3.2.3 — График квантилей для наблюдаемых температур

Попробуем опровегнуть наше предположение на основе проверки критерия χ^2 Пирсона [18]. Для этого воспользуемся пакетом *nortest* и функцией *pearson.test*:

```
Pearson chi-square normality test

data:  data
P = 2.8, p-value = 0.8335
```

В полученных результатах P — статистика χ^2 Пирсона. Вероятность ошибки $p = 0.8335 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста также нельзя. Проверим критерий: примем уровень значимости $\alpha = 0.05$, тогда из таблицы распределения χ^2 найдём критическое значение критерия $P_{\text{кр}}(\alpha, k) = 43.8$. Отсюда следует, что

$$P < P_{\text{кр}}.$$

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

Воспользуемся для тех же целей критерием Колмогорова–Смирнова [19]. Как в предыдущем случае воспользуемся представленной в пакете *nortest* функцией *ks.test*:

```
Two-sample Kolmogorov-Smirnov test

data:  data and nsample
D = 0.0706, p-value = 0.995
alternative hypothesis: two-sided
```

Вероятность ошибки $p > 0.05$, а значит нулевую гипотезу отвергнуть нельзя. Следовательно опровергнуть предположение о нормальности, как и предыдущих случаях, также нельзя. Проверим критерий: примем так же уровень значимости $\alpha = 0.05$, тогда критическое значение $D_{\text{кр}}(\alpha) = 1.358$. Следовательно,

$$D < D_{\text{кр}}(\alpha),$$

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

На данном этапе по полученным ранее результатам возникли подозрения о выбросах в исходной выборке. Выявление таких аномальных значений важно, так как их наличие, как правило, сильно влияет на всю выборку, в частности, на коэффициент корреляции. Проверим наличие выбросов с помощью статистических критериев. Для этих целей воспользуемся критерием Граббса [20]. Данный основан на предположении о нормальности исходных данных. То есть, перед применением данного критерия необходимо убедиться, что данные могут быть в разумных пределах аппроксимированы нормальным распределением [21]. Поскольку ранее высказано предположение о нормальности, воспользуемся им для определения наличия выбросов.

Полученные результаты проверки критерия Граббса:

```
Grubbs test for one outlier

data:  research.data$temperature
G = 1.9487, U = 0.8850, p-value = 0.8103
alternative hypothesis: highest value 24.2 is an outlier
```

Данный результат (p-value $= 1$) однозначно говорит нам о том, что следует отклонить альтернативную гипотезу H_1 и принять гипотезу H_0. Другими словами, это говорит о том, что в исходной выборке нету выбросов. А значит выборка однородна. Значит, наши подозрения о выбросах не подтвердились проверкой критерия.

В соответствии с результатами проверки критериев и на основе построенных гистограммы и графика квантилей, можно сделать заключение о том, что распределение температуры воды озера Баторино в июле 1975–2009 годов является близким к нормальному закону распределения с параметрами $\mathcal{N}(19.88, 4.91)$. Что подтверждается коэффициентами асимметрии и эксцесса из таблицы 1, а также результатами, полученными мной при исследовании в пакете **STATISTICA**. Следует также отметить, что такие же результаты были получены и для всей выборки.

3.3 Корреляционный анализ

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

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

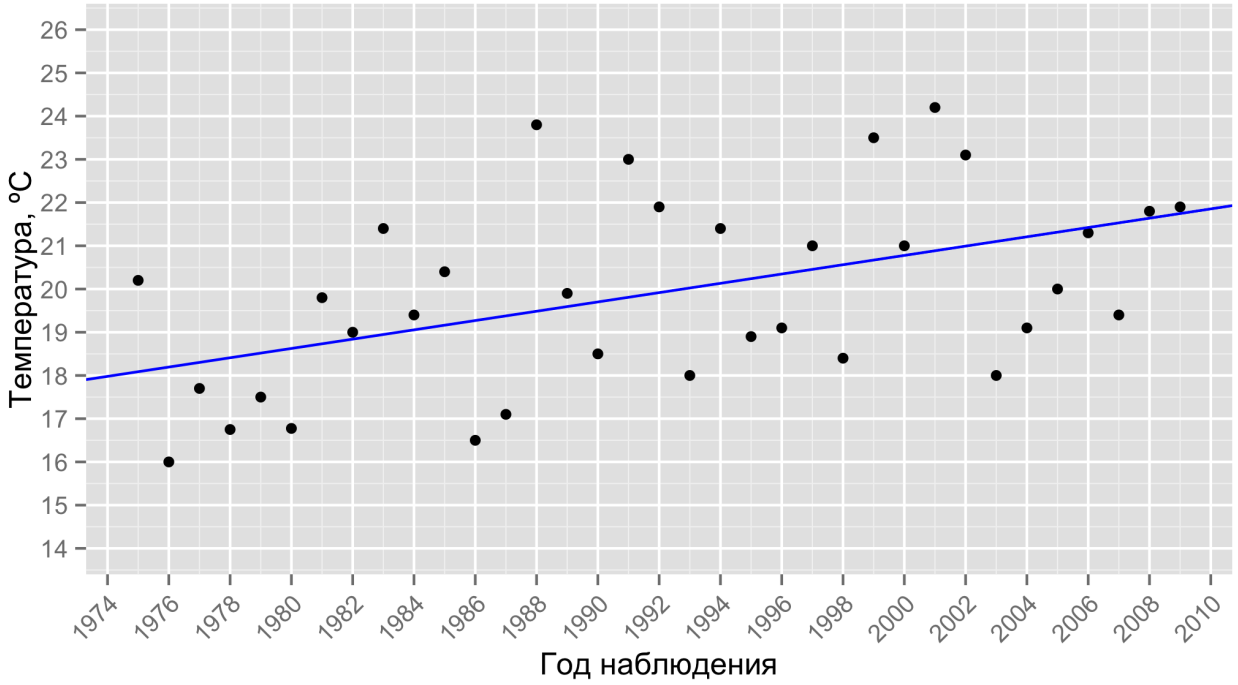

Рисунок 3.3.4 — Диаграмма рассеяния

Из рисунка 3.3.4 видно, что точки образуют своеобразное «облако», ориентированное по диагонали вверх, то есть присутствует некая зависимость между рассмтриваемыми переменными. Также, данная диаграмма наглядно показывает силу этой зависимости: так как точки не образуют чёткой формы, а разбросаны относительно диагонали, то можно говорить о наличии умеренной корреляции. То есть, нельзя сказать, что зависимость сильная, но и нельзя сказать, что связь между переменными отсутствует.

Проверим полученные результаты подробнее. Для начала построим корреляционную матрицу. Как видно из таблицы 2, коэффициент корреляции $r_{xt} = 0.47$. Этим подтвержда-

	Temperature	Date
Temperature	1.00	0.47
Date	0.47	1.00

Таблица 2 — Корреляционная матрица.

ются наши выводы из диаграмм рассеяния и концентрации о положительной корреляции, поскольку полученный коэффициент корреляции является положительным и присутствует умеренная зависимость: $r_{xt} \approx 0.5$.

Оценим значимость полученного выборочного коэффициента корреляции с помощью возможностей пакета **R** и описанного ранее в параграфе критерия значимости. Вычислим:

$$T_{\text{набл}} = \frac{r_{xt}\sqrt{n-2}}{\sqrt{1-r_{xt}^2}} \approx 3.05885.$$

Рассмотрим уровень значимости $\alpha = 0.05$. Число степеней свободы $k = n - 2 = 36$. Тогда из таблицы критических точек распределения Стьюдента $t_{\text{кр}}(\alpha, k) \approx 2.03$. Следовательно,

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

Значит нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности следует отклонить [9].

Пакет **R** предоставляет с помощью функции *cor.test* различные методы для проверки значимости выборочного коэффициента корреляции. Воспользуемся проверкой теста методом Пирсона:

```
Pearson's product-moment correlation

data:  research.data$temperature and c(1:kObservationNum)
t = 3.0471, df = 33, p-value = 0.004523
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1603947 0.6935396
sample estimates:
      cor
0.4685944
```

Как видно из полученных результатов $p - value < 0.05$, следовательно это говорит, о том, что необходимо отвегнуть гипотезу $H_0 : r = 0$.

Значит, нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности отвергаем и подтверждаем правильность полученных с помощью **R** результатов. Другими словами, выборочный коэффициент значимо отличается от нуля, т.е. температура воды и время при уровне значимости $\alpha = 0.05$ имеют зависимость.

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

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

3.4 Регрессионный анализ

Для введения последующих понятий анализа временных рядов воспользуемся [22].

В отличие от анализа случайных выборок, анализ временных рядов основывается на предположении, что последовательные значения в файле данных наблюдаются через равные промежутки времени.

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

Во временных рядах выделяют три составляющие:

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

2. *Циклические (K), сезонные (S) колебания* — это составляющие, которые проявляются как отклонения от основной тенденции развития изучаемого явления, и связанны с действие краткосрочных, систематических факторов развития. Циклические колебания состоят в том, что значения признака в течение какого-то времени возрастают, достигают определённого максимума, затем убывают, достигают определённого минимума, вновь возрастают до прежних значений и т.д. Эту составляющую можно выявить только по данным за длительные промежутки времени, например, в 10, 15 или 20 лет. Сезонные колебания — это колебания, периодически повторяющиеся в

некоторое определённое время каждого года, месяца, недели, дня. Эти изменения отчётливо наблюдаются на графиках рядов динамики, содержащих данные за период не менее одного года.

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

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

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

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

Аддитивной модели свойственно то, что характер циклических и сезонных колебаний остаётся постоянным.

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

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

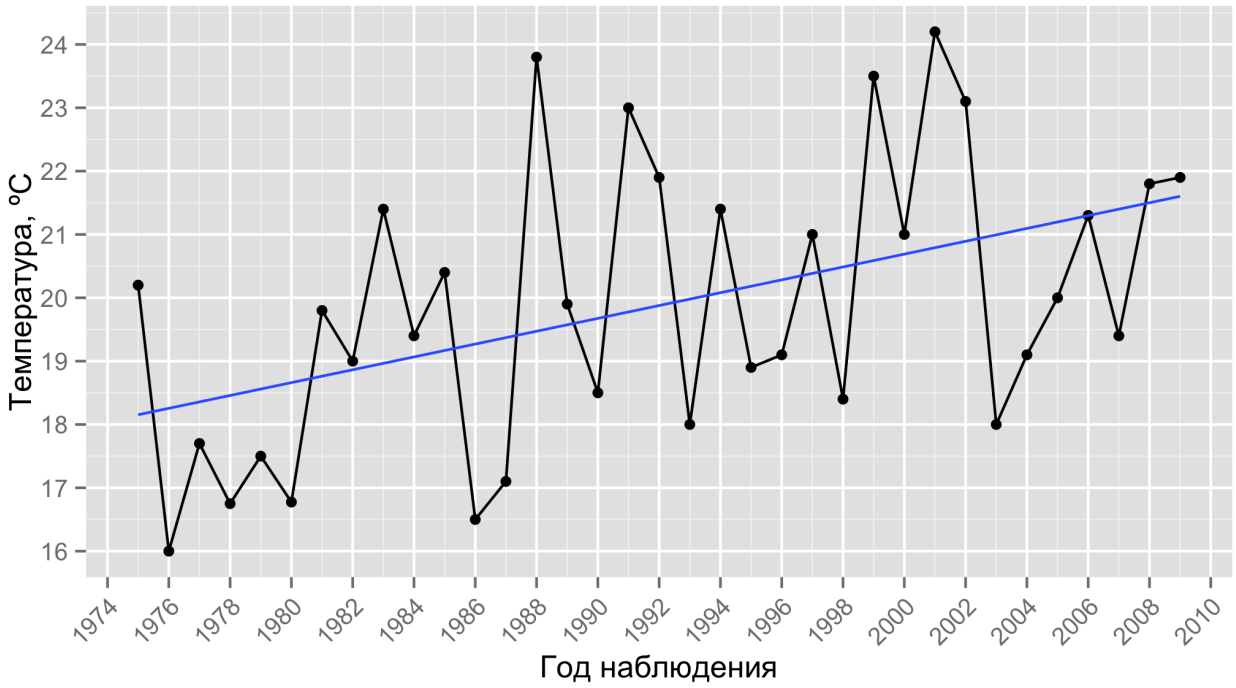

Рисунок 3.4.5 — График временного ряда с линией регрессии

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

$$x = at + b.$$

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

$$X = T + E.$$

В **R** реализованы функции, позволяющие подгонять линейные модели к исследуемым данным [23]. Одной из таких функций является *lm(Fitting Linear Model)* [12, с.178]. Она позволяет получить коэффициенты линии регрессии(тренд), остатки после удаления тренда. Коэффицинты, полученные с помощью данной функции представлены в (3.2).

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

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

Таким образом получена линейная модель, описывающая тенденцию развития:

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

На основе полученной линейной модели (3.3), построим ряд остатков(приложение В, таблица В.1), удалив тренд из исходного ряда. Полученный ряд представлен на рисунке 3.4.6.

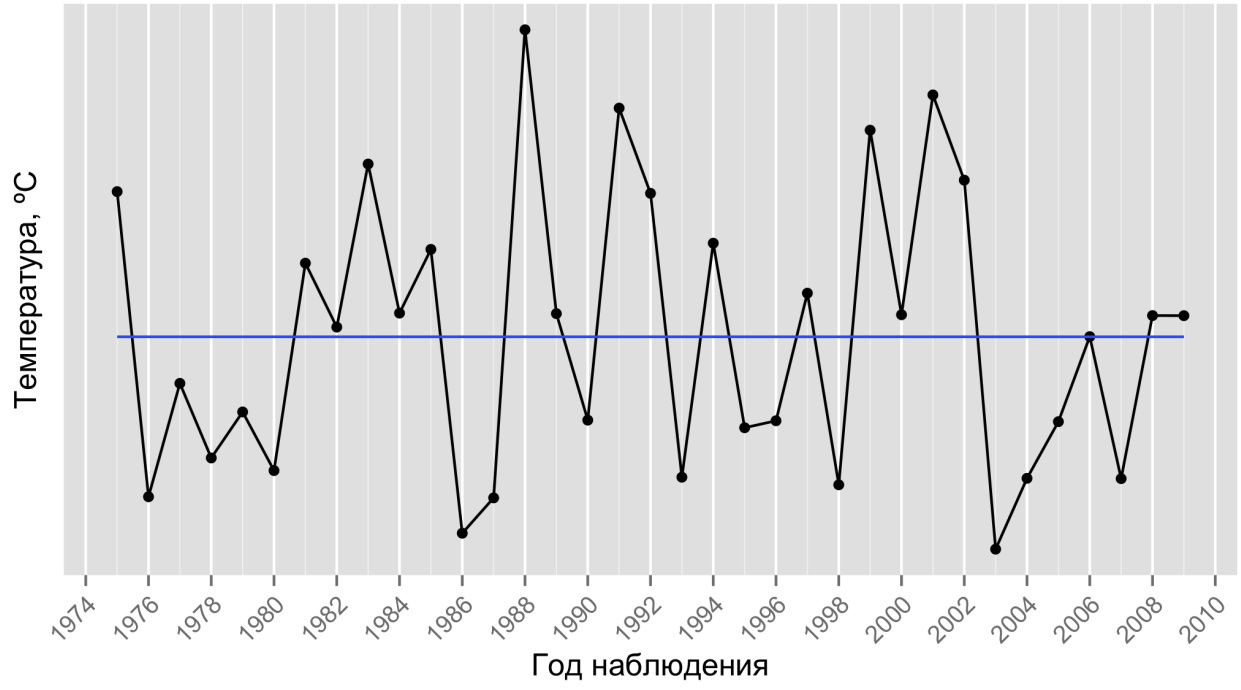

Рисунок 3.4.6 — График ряда остатков

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

Рассчитаем вспомогательные величины, воспользовавшись [22]. Дисперсия отклонения

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

стандартные случайные погрешности параметров a, b:

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

Воспользуемся критерием значимости коэффициентов линейной регрессии [9]. Примем уровень значимости $\alpha = 0.05$, тогда

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

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

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

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

Оценим адекватность полученной регрессионной модели. Дисперсия модели:

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

Остаточная дисперсия:

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

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

$$F_{\text{крит}} \approx 14.01,$$

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

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

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

Рассчитаем коэффициент детерминации:

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

Проверим отклонение от линейности: $\eta^2_{x(t)} - r^2_{xt} \approx 0.0044 \leq 0.1$. Следовательно отклонение от линейности незначительно. Но при этом коэффициент детерминации оказался не достаточно высоким(< 0.7), это говорит о том, что построенная регрессионная модель не описывает в достаточной мере поведение временного ряда. Это, в свою очередь, может значить, что изменение температуры зависит не только от времени, но ещё и от каких-то других, неучтённых, факторов.

Тем не менее, попробуем построить прогноз по полученной модели. Вычисленые прогнозные значения на 2010-2012 годы: 21.87 – 2010 год, 21.98 – 2011 год, 22.08 – 2012 год. Фактические данные на этот период: 24.3, 22.8, 20.2 соответственно. Имеющееся отклонение прогнозов от реальных данных еще раз подтверждает, что построенная модель временного ряда обладает невысокой точностью.

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

1. $E(\varepsilon) = 0$.
2. Дисперсия ε постоянна для всех значений.

	Значение
Среднее	-0.00
Медиана	0.14
Нижний квартиль	-1.80
Верхний квартиль	1.28
Минимум	-2.99
Максимум	4.33
Размах	7.32
Квартильный размах	3.07
Дисперсия	3.84
Стандартное отклонение	1.96
Коэффициент вариации	0.00
Стандартная ошибка	0.33
Асимметрия	0.42
Ошибка асимметрии	0.40
Эксцесс	-0.77
Ошибка эксцесса	0.78

Таблица 3 — Описательные статистики для остатков.

3. Остатки независимы и нормально распределены.

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

Как видно из таблицы 3, среднее значение равно нулю. При этом коэффициенты асимметрии($A_S \approx 0.37$) и эксцесса($K \approx -0.87$) указывают на отклонение распределения остатков от нормального закона.

Построим гистограмму и график квантилей для проверки последних заключений. Гистограмма наглядно демонстрирует полученные в таблице 3 коэффициенты асимметрии и эксцесса.

Для проверки нормальности построим график квантилей. На рисунке 3.4.8 можно заметить, что присутствуют скачки относительно нормального распределения. Наиболее явный из них — нижний хвост. Остальные — небольшие скачки по ходу линии нормального распределения. Проверим с помощью критерия Шапиро-Уилка, можно ли считать полученные остатки нормально распределёнными.

```
Shapiro-Wilk normality test

data:  data
W = 0.9513, p-value = 0.124
```

В полученных результатах W — статистика Шапиро-Уилка. Вероятность ошибки $p = 0.124 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста нельзя.

Проверим критерий χ^2 Пирсона:

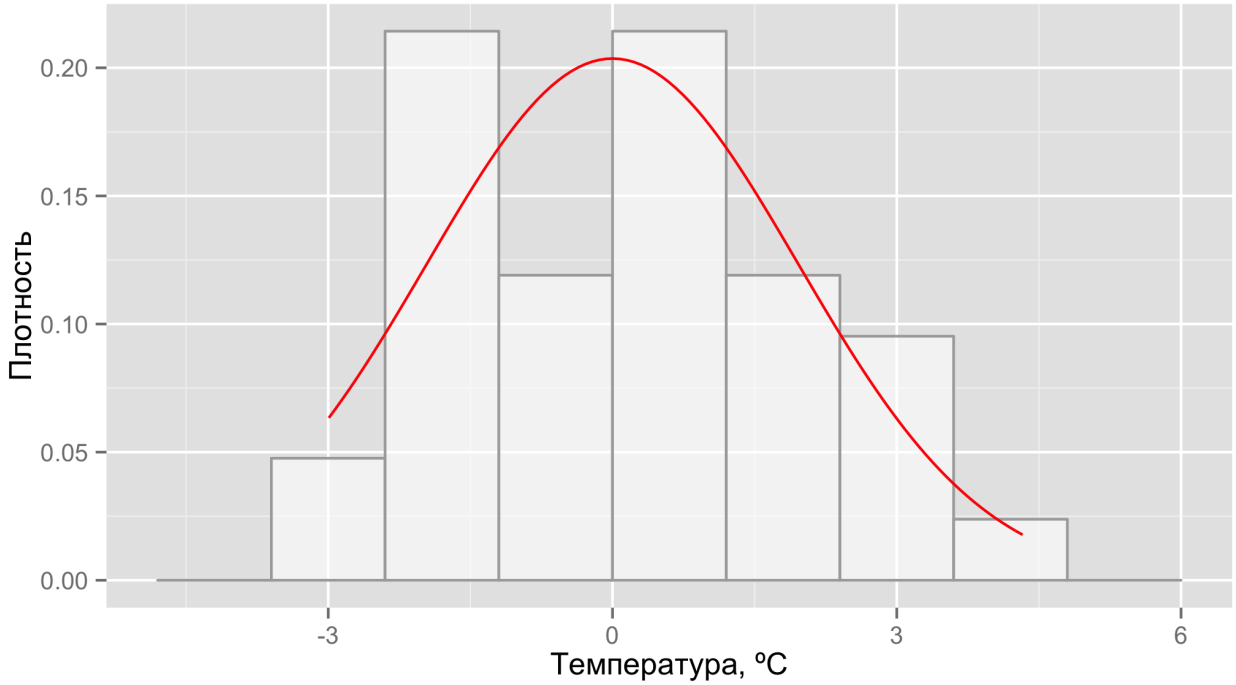

Рисунок 3.4.7 — Гистограмма остатков с кривой плотности нормального распределения

```
Pearson chi-square normality test

data:  data
P = 10.5143, p-value = 0.1046
```

В полученных результатах P — статистика χ^2 Пирсона. Вероятность ошибки $p = 0.1046 > 0.05$, а значит нулевая гипотеза не отвергается. Но при этом, это значение очень близко к 0.05. Проверим критерий: примем уровень значимости $\alpha = 0.05$, тогда из таблицы распределения χ^2 найдём критическое значение критерия $\chi^2_{\text{кр}}(\alpha, k) \approx 43.8$. Отсюда следует, что

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

где $\chi^2_{\text{набл}} = P = 10.5143$. А значит, гипотезу о нормальности не отклоняем.

Построим график автокорреляционной функции для определения наличия взаимосвязей в ряде остатков (рисунок 3.4.9). На графике пунктирные линии разграничивают значимые и не значимые корреляции: значения, выходящие за линии, являются значимыми [13, с.376]. На представленном графике автокорреляционной функции можно заметить на лаге 15 значение, выходящее за интервал, обозначенный пунктирными линиями. Проверим значимость автокорреляций с помощью теста Льюнга-Бокса [13, с.377-378]. Данный тест проверяет наличие автокорреляций в исследуемом ряде.

```
Box-Ljung test

data:  research.residuals$temperature
X-squared = 0.0754, df = 1, p-value = 0.7836
```

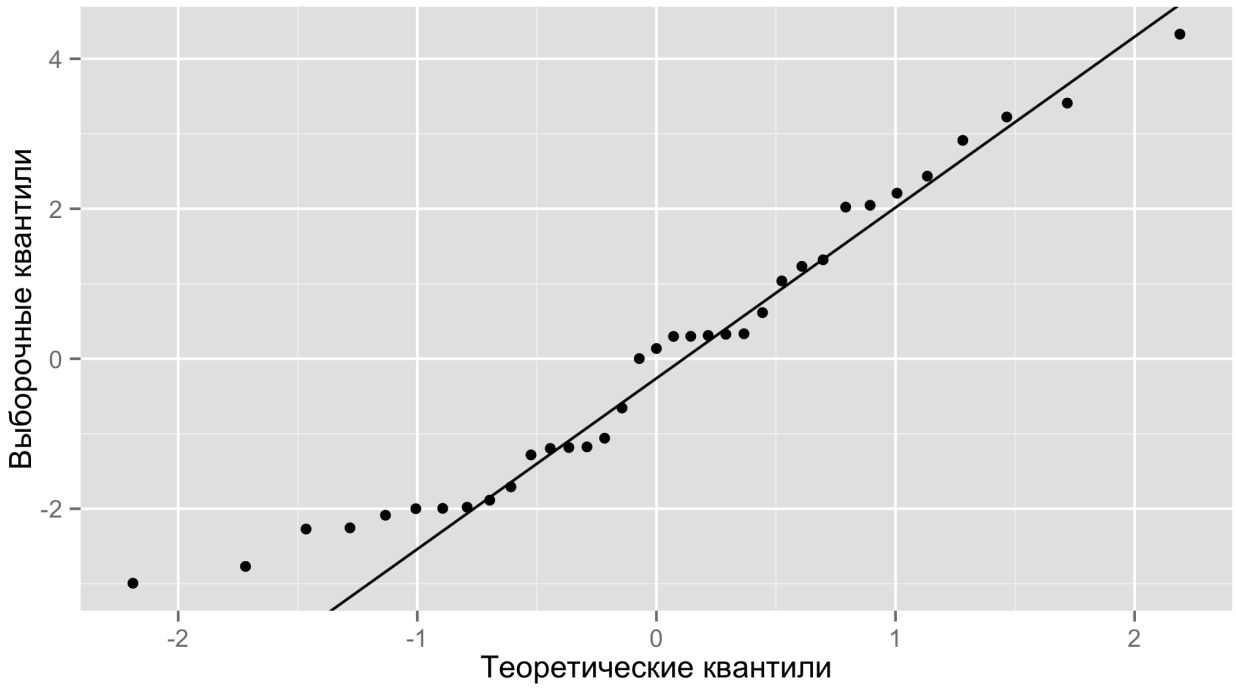

Рисунок 3.4.8 — График квантилей для остатков

В результатах теста $p > 0.05$, что говорит о том, что тест не выявил значимых автокорреляций.

На рисунке 3.4.9 также можно заметить затухание с увеличением лага. На основе этого можно сделать предположение о стационарности. Для проверки этого предположения воспользуемся расширенным тестом Дики–Фуллера(ADF) [24].

```
Augmented Dickey-Fuller Test

data:  research.residuals$temperature
Dickey-Fuller = -3.2695, Lag order = 3, p-value = 0.09261
alternative hypothesis: stationary
```

Как видно из результатов проверки теста, $p < 0.05$. Следовательно, необходимо принять альтернативную гипотезу о стационарности.

Полученная модель получилась неоднозначной, с одной стороны, полученное значение коэффициента детерминации показало недостаточную точность полученной модели и не удалось достоверно показать нормальность ряда остатков. С другой стороны, была показана стационарность и отсутствие автокорреляции. Поэтому возникает необходимость строить модель другими методами.

3.5 Вариограммный анализ. Кригинг.

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

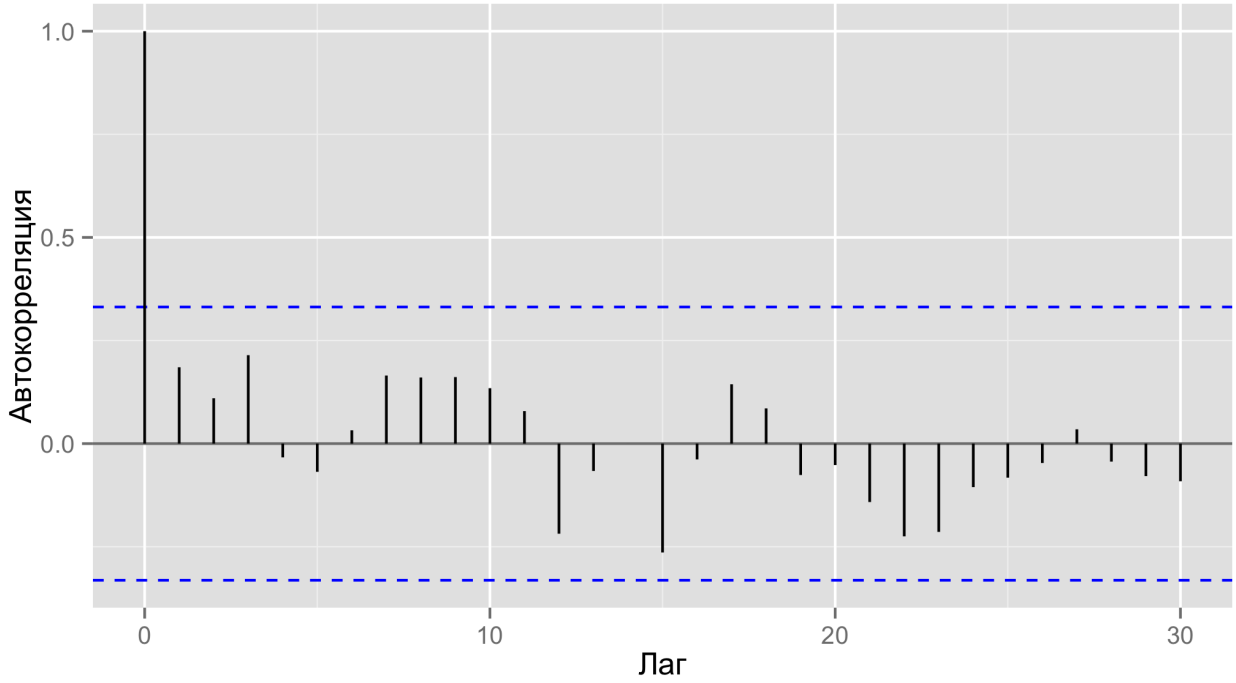

Рисунок 3.4.9 — График автокорреляционной функции

Традиционные детерминированные методы, широко используемые в задачах прогнозирования, в большинстве случаев на практике не позволяют в полной мере решить ту или иную задачу. В наиболее благоприятных вариантах исследований они позволяют оценивать значения в точках, в которых измерения не проводились и определять значения на плотной сетке (в близких к измерениям точках). Следует также отметить, что данные измерений, как правило, дискретны и пространственно неоднородно распределены. В свою очередь, анализ этих данных и его результаты в значительной мере зависят как от качества так и от количества исходных данных. И именно такие выводы были сделаны в результате проделанной в предыдущих частях данной работы. Отсюда следует, что необходимо использовать другие современные методы, позволяющие сделать более точные модели и выводы.

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

В рамках геостатистики, для получения наилучшей в статистическом смысле пространственной оценки используются модели из семейства кригинга (*kriging*) — наилучшего линейного несмещенного оценивателя (*Best Linear Unbiased Estimator — BLUE*). Кригинг является “наилучшим” оценивателем в статистическом смысле — его оценка обладает минимальной дисперсией. Важным свойством кригинга является точное воспроизведение значений измерений в имеющихся точках (интерполяционные свойства). В отличие от многочисленных детерминированных методов оценка кригинга сопровождается оценкой ошибки интеполяции в каждой точке. Полученная ошибка позволяет охарактеризовать неопределенность интерполяционноий оценки данных при помощи доверительных интервалов.

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

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

Также при построении вариограммы следует учитывать параметр максимального расстояния, для которого вычисляется вариограмма. Первоначальным параметром было выбрано следующее значение: $2n/3 = 23$. Построим экспериментальную вариограмму с помощью пакета “gstat” и функции “variogram”. С помощью этой функции можно построить экспериментальную вариограмму, основанную на классической оценки вариограммы и робастной оценки Кресси. Построим экспериментальную вариограмму с помощью классической оценки.

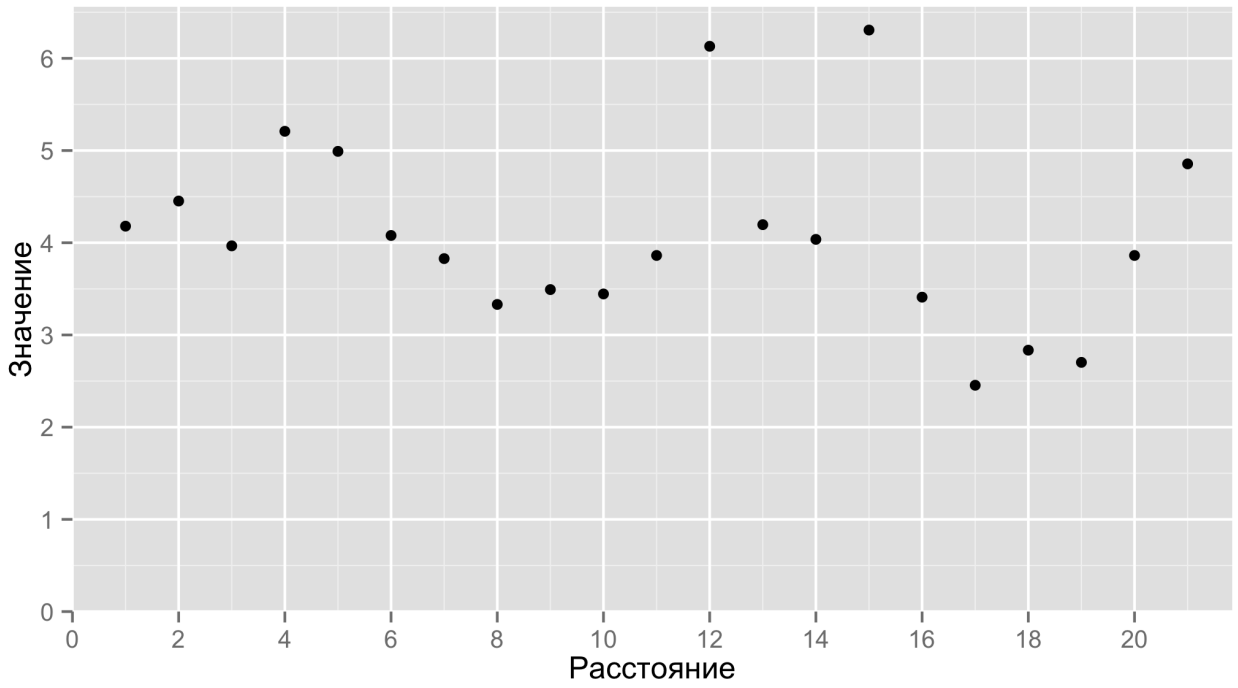

Рисунок 3.5.10 — Экспериментальная вариограмма

Построенная вариограмма отображена на рисунке 3.5.10. На представленном рисунке можно заметить, что на промежутке $[0; 1]$ не происходит роста значений вариограммы. Наоборот, наблюдается разрыв: первое значение находится значительно выше 0. При этом вариограмма не сильно выходит за пределы дисперсии переменной, которая равна 3.9.

Более того, первые значения уже достигло порога. Что говорит о том, что вариограмма на первых значениях выходит на предельное значение, и последующие значения некоррелированы. Это, на самом деле, согласуется с нашими исходными данными, так как при анализе остатков было выявлено отстутсвие автокорреляций и спецификой самих данных: наблюдение за каждый год, вообще говоря, не зависит от предшедствующего.

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

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

Для построения моделей вариограммы существует два подхода: вручную, т.е. визуально с ручным подбором параметров, и автоматическим подбором параметров с помощью специальных методов. И на практике построение модели вариограммы представляет собой итеративный процесс, на каждом шаге которого следует наилучшим образом подобрать параметры очередного модельного приближения. В различной литературе рекомендуется строить моделей вручную, так как исследователь лучше знает специфику данных, чем различные методы оценивания. Попробуем построить модель вариограммы визуально.

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

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

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

	Модель	Порог	Ранг
1	Nug	3.71	0.00
2	Sph	0.65	1.26

Таблица 4 — Модель вариограммы

И график полученной модели на рисунке 3.5.11 (пунктиром). На графике можно проследить все указанные ранее особенности: эффект самородков и порог.

Задача геостатистики — оценить значения изучаемой пространственной переменной в произвольных точках области исследования на основе анализа ее значений, измеренных в ограниченном числе выборочных точек. По построенной модели вычислим оценки при помощи ординарного кригинга, реализованного функцией “krige”. Вычисленные значения: $0.01228368, 0.01382626, 0.01382626, 0.01382626, 0.01382626, 0.01382626$. Оценку отклонения от истинных значений выразим с помощью среднеквадратической ошибки *(MSE)*. В данном случае $MSE = 2.636708$. Полученные значения оказались очень близкими к нулю и, начиная со второго значения значения не изменяются. Следовательно прогноз почти

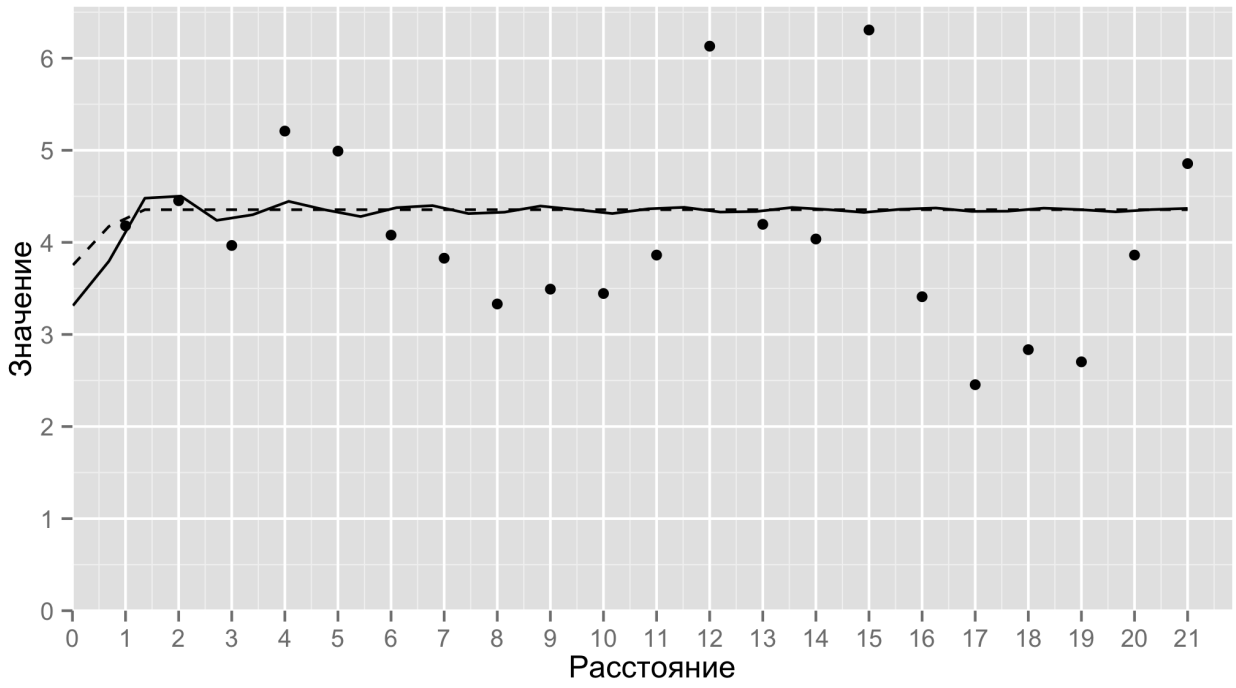

Рисунок 3.5.11 — Модели классической вариограммы

не изменился. Это говорит о том, что построенная модель не смогла уловить поведение исходных данных. По этой причине был использован второй вариант построения модели — автоматический.

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

На рисунке 3.5.11 сплошной линией показан результат выполнения представленной ранее функции. Таким образом, наилучшей моделью вариограммы, построенной по классической оценке, стала линейная комбинация двух: эффект самородков с параметром 3.130070 и модель с эффектом дырок (*Hole*) с параметрами: порог — 1.219741, ранг — 0.3686477.

Кригинг в этом случае построил следующие прогнозные значения: $0.123192458, -0.100531208, 0.130358294, -0.072544498, 0.076113473, -0.003595666$. Полученные значения отличаются от предыдущих, в них есть некоторое поведение. Но в данном случае $MSE = 2.8752$. А значит, прогноз ухудшился. Попробуем улучшить результат с помощью робастной оценки Кресси.

Модель вариограммы, представленная на рисунке 3.5.12, является также линейной комбинацией двух базисных моделей: эффекта самородков с параметром 4.6175154 и модель с эффектом дырок с параметрами: 0.4736754, 4.318585. Заметим, что эмпирическая вариограмма, построенная по робастной оценке, отличается от соотвествующей вариограмм, построенных по классической оценке. Появилось заметное поведение вариограммы, в отличие от предыдущей, где значения концентрировались около дисперсии выборки.

Результаты кригинга показали следующие прогнозные значениея: $-0.10273199, 0.07493855, 0.15517704, 0.13209470, 0.05161946, -0.02323001$. Среднеквадратическая ошибка $MSE = 2.368255$, что является наименьшей из полученных для

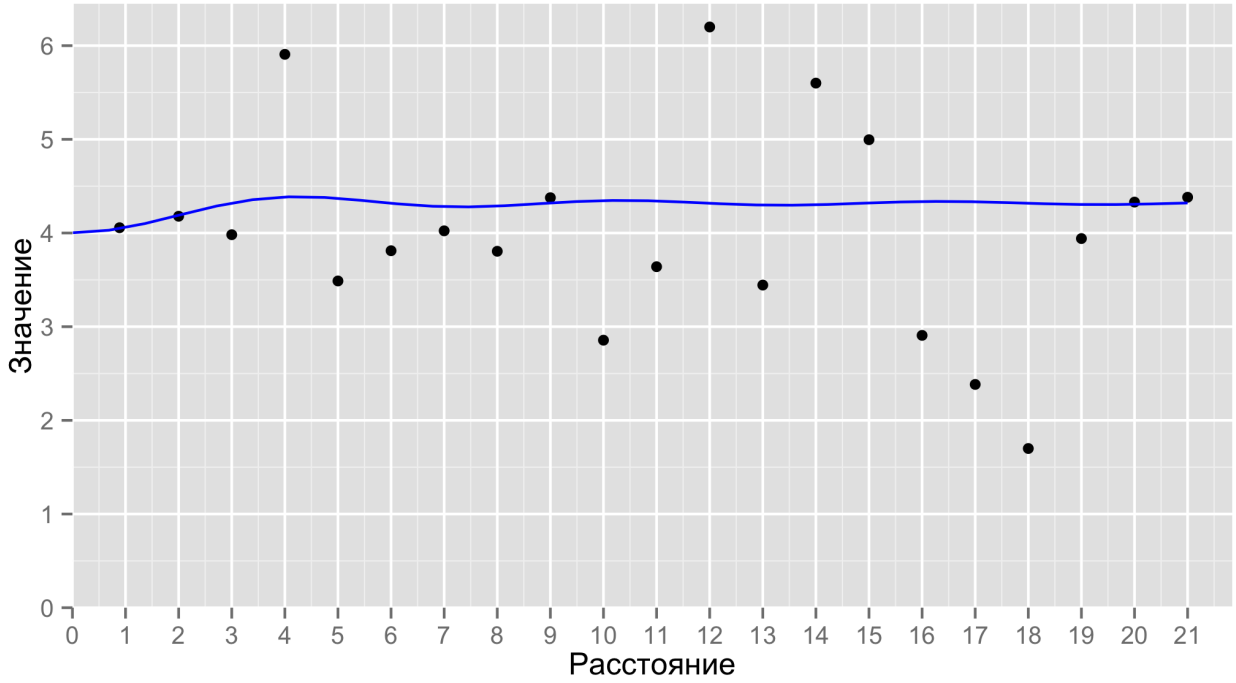

Рисунок 3.5.12 — Теоретическая вариограмма (робастная оценка)

заданных параметров. Таким образом, наилучшая оценка получена с помощью робастной оценки.

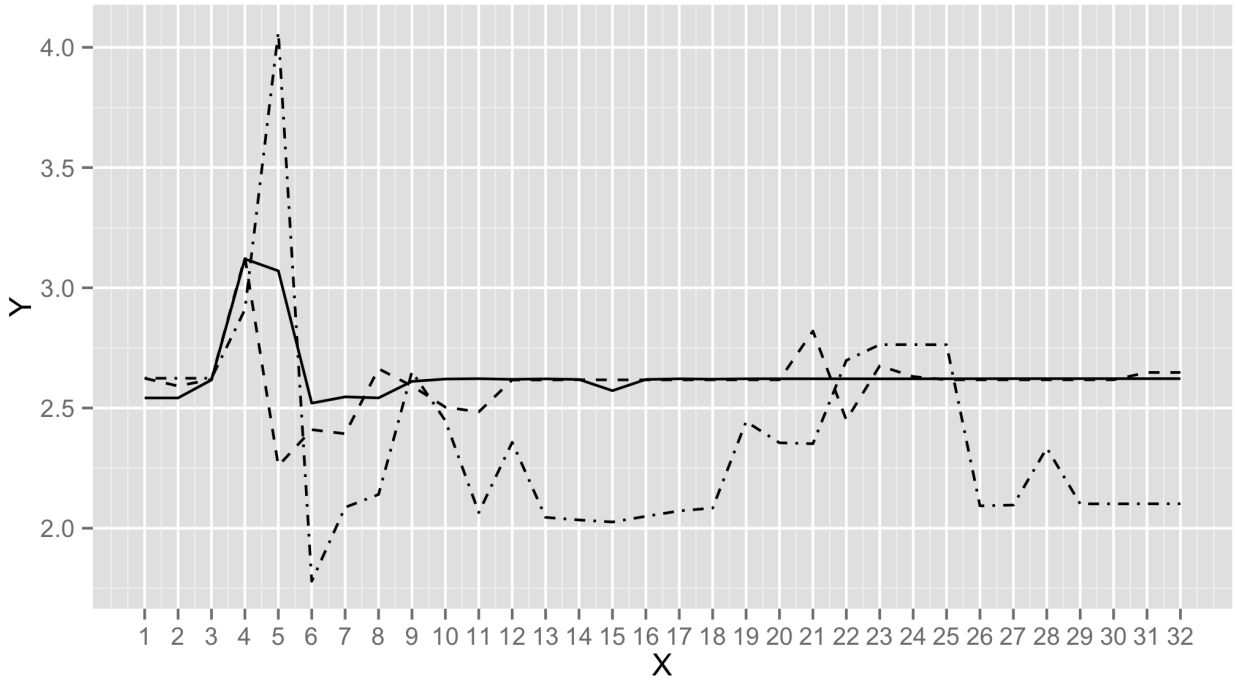

Рисунок 3.5.13 — Зависимость ошибки от максимального расстояния

Исследуем теперь поведение кригинга при различных параметрах максимального расстояния вариограммы. В качестве оценки качества полученного прогноза возьмем среднеквадратическую ошибку. Чем меньше ошибка — тем лучше прогноз. Для

этих целей реализована функция *checkDepByMSE*. Результат её работы на рисунке 3.5.13. На этом графике четко видно, что робастная оценка (пунктир-точка), в отличие от классической (пунктир) и модели, построенной вручную (сплошная), в большинстве случаев даёт более точные прогнозы. И наилучший при максимальном расстоянии равным 6. С этим параметром, наилучший прогноз составляют значения кригинга: $-0.39594965, -0.03720509, 0.26603554, 0.40557237, 0.353869380.16726908$. Среднеквадратическая ошибка составляет 1.805819.

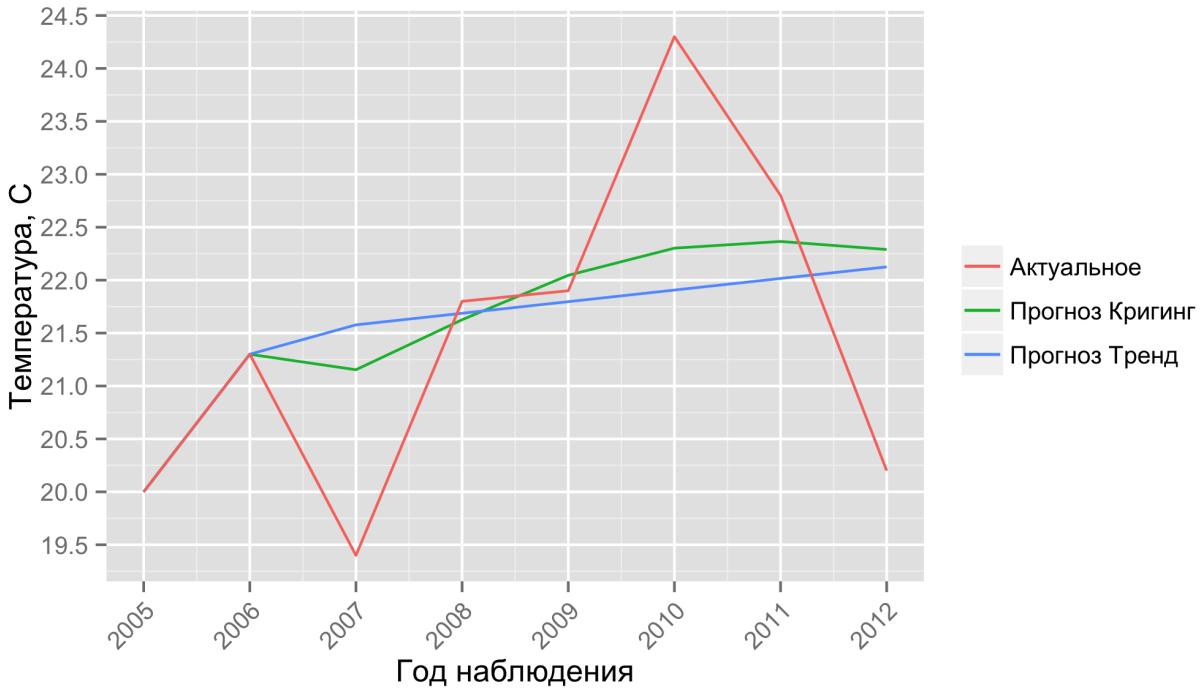

Рисунок 3.5.14 — Сравнение прогнозных значений

Сравнительный анализ полученного прогноза на графике 3.5.14.

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

Заключение

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

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

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

Литература

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

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

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

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

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

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

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

8. Т.В. Цеховая. Асимптотическое распределение оценки вариограммы. *Вестник БрГУ им. А.С. Пушкина*, №2(31):32 – 37, 2008.

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

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

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

12. Robert Kabacoff. R in Action. 2009.

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

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

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

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

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

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

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

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

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

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

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

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

ПРИЛОЖЕНИЕ А Исходные данные

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

Таблица А.1 — Исходные данные.

ПРИЛОЖЕНИЕ В Результаты вычислений

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

Таблица В.1 — Временной ряд остатков.

ПРИЛОЖЕНИЕ С Код программ

```
# Descriptive statistics

# Function for getting all descriptive statistics
dstats.describe <- function(data, locale=FALSE) {
  stats <- c(dstats.mean(data), dstats.median(data), dstats.quartile.lower(data),
             dstats.quartile.upper(data), dstats.min(data), dstats.max(data),
             dstats.range(data), dstats.quartile.range(data), dstats.variance(data),
             dstats.std.dev(data), dstats.coef.var(data), dstats.std.error(data),
             dstats.skew(data), dstats.std.error.skew(data), dstats.kurtosis(data),
             dstats.std.error.kurtosis(data))
  if (locale) {
    descr.row <- c("Среднее", "Медиана", "Нижний квартиль", "Верхний квартиль",
                   "Минимум", "Максимум", "Размах", "Квартильный размах",
                   "Дисперсия", "Стандартное отклонение", "Коэффициент вариации",
                   "Стандартная ошибка", "Асимметрия", "Ошибка асимметрии",
                   "Эксцесс", "Ошибка эксцесса")
    descr.col <- c("Значение")
  } else {
    descr.row <- c("Mean", "Median", "Lower Quartile", "Upper Quartile", "Range",
                   "Minimum", "Maximum", "Quartile Range", "Variance", "Standard Deviation",
                   "Coefficient of Variance", "Standard Error", "Skewness",
                   "Std. Error Skewness", "Kurtosis", "Std. Error Kurtosis")
    descr.col <- c("Value")
  }
  df <- data.frame(stats, row.names=descr.row)
  colnames(df) <- descr.col

  df
}

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

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

dstats.quartile.lower <- function(data, ...) {
  quantile(data, ...)[[2]]
}

dstats.quartile.upper <- function(data, ...) {
  quantile(data, ...)[[4]]
}

dstats.quartile.range <- function(data) {
  dstats.quartile.upper(data) - dstats.quartile.lower(data)
}

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

```
52    min(data, ...)
53  }
54
55  dstats.max <- function(data, ...) {
56    max(data, ...)
57  }
58
59  dstats.range <- function(data) {
60    max(data) - min(data)
61  }
62
63  dstats.variance <- function(data, ...) {
64    var(data, ...)
65  }
66
67  dstats.std.dev <- function(data) {
68    sd(data)
69  }
70
71  dstats.coef.var <- function(data) {
72    mn <- mean(data)
73    if (abs(mn) > 1.987171e-15) {
74      (var(data) / mean(data)) * 100
75    } else
76      0
77  }
78
79  dstats.std.error <- function(data) {
80    sd(data) / sqrt(length(data))
81  }
82
83  dstats.skew <- function(data) {
84    n <- length(data)
85    mean <- mean(data)
86    (n * sum(sapply(data, FUN=function(x){(x - mean)^3}))) /
87      ((n - 1) * (n - 2) * dstats.std.dev(data)^3)
88  }
89
90  dstats.std.error.skew <- function(data) {
91    n <- length(data)
92    sqrt((6 * n * (n - 1)) / ((n - 2) * (n + 1) * (n + 3)))
93  }
94
95  dstats.test.skew <- function(data) {
96    dstats.skew(data) / dstats.std.error.skew
97  }
98
99  dstats.kurtosis <- function(data) {
100   n <- length(data)
101   mean <- mean(data)
102   (n * (n + 1) * sum(sapply(data, FUN=function(x){(x - mean)^4})) - 3 * (sum(
        sapply(data, FUN=function(x){(x - mean)^2})))^2 * (n - 1)) /
103     ((n - 1) * (n - 2) * (n - 3) * dstats.variance(data)^2)
104 }
105
106 dstats.std.error.kurtosis <- function(data) {
107   n <- length(data)
108   2 * dstats.std.error.skew(data) * sqrt((n^2 - 1) / ((n - 3) * (n + 5)))
109 }
110
```

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

Листинг C.1: Описательные статистики

```
# Let's start from beginning.
# This file will be the master file of all diploma project's files (slaves).
# Content will be the same as for previous works (batorino analysis).
# Some thoughts for this investigation see in TODO.Rmd.
# Ideas for organizing further research see in ideas.Rmd

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

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

## Import local modules
source("R/lib/plot.R")       # useful functions for more comfortable plotting
source("R/lib/print.R")      # functions for print some data to files
source("R/lib/dstats.R")     # descriptive statistics module
source("R/lib/misc.R")       # some useful global-use functions
source("R/lib/draw.R")       # helpers for drawing
source("R/lib/ntest.R")      # tests for normality

## Read the data / pattern: year;temperature
path.data <- "data/batorino_july.csv" # this for future shiny support and may be
    choosing multiple data sources
src.nrows <- 38
src.data  <- read.csv(file=path.data, header=TRUE, sep=";", nrows=src.nrows,
    colClasses=c("numeric", "numeric"), stringsAsFactors=FALSE)

## Global use constants
kDateBreaks <- seq(min(src.data$year) - 5, max(src.data$year) + 5, by=2) # date
    points for graphs

## For the reason of prediction estimation and comparison, let cut observations
    number by 3
kObservationNum <- length(src.data[, 1]) - 3

## Source data as basic time series plot: points connected with line
plot.source <- DrawDataRepresentation(data=src.data, filename="source.png",
    datebreaks=kDateBreaks)

print(xtable(src.data, caption="Исходные данные.", label="table:source"),  table
    .placement="H",
      file="out/original/data.tex")

## Form the data for research
research.data <- src.data[0:kObservationNum, ]

# Getting descriptive statistics for temperature in russian locale
research.data.dstats <- dstats.describe(research.data$temperature, locale=TRUE)
```

```
50 print(xtable(research.data.dstats, caption="Описательные статистики для наблюдае
       мых температур.", label="table:dstats"),
51       file="out/original/dstats.tex")
52
53 ## Basic histogram based on Sturges rule (by default) with pretty output (also
       by default)
54 plot.data.hist <- DrawHistogram(data=research.data, filename="original/histogram
       .png")
55
56 ## Tests for normality
57 research.data.shapiro <- ntest.ShapiroWilk(data=research.data$temperature,
       filename="out/original/shapiro-test.tex")
58 research.data.pearson <- ntest.PearsonChi2(data=research.data$temperature,
       filename="out/original/pearson-test.tex")
59 research.data.ks      <- ntest.KolmogorovSmirnov(data=research.data$temperature,
        filename="out/original/ks-test.tex")
60
61 ## Normal Quantile-Quantile plot // TODO: check when it appears in text
62 plot.data.qq <- DrawQuantileQuantile(data=research.data$temperature, filename="
       original/quantile.png")
63
64 ## Scatter plot with regression line
65 plot.data.scatter <- DrawScatterPlot(research.data, filename="original/
       scatterplot.png", kDateBreaks);
66
67 ## Grubbs test for outliers
68 research.data.grubbs <- grubbs.test(research.data$temperature)
69 to.file(research.data.grubbs, "out/original/grubbs-test.tex")
70
71 ## Correlation matrix
72 research.data.cmatrix <- cor(cbind("Temperature"=research.data$temperature, "
       Date"=1:kObservationNum), method="pearson")
73 print(xtable(research.data.cmatrix, caption="Корреляционная матрица.", label="
       table:cmatrix"),
74       file="out/original/corr-matrix.tex")
75
76 ## Pearson's product-moment correlation test. Use time for y as numerical
77 research.data.ctest <- cor.test(research.data$temperature, c(1:kObservationNum),
        method="pearson")
78 to.file(research.data.ctest, "out/original/corr-test.tex")
79
80 ## Fitting linear model for researching data. It also compute residuals based on
        subtracted regression
81 research.data.fit <- lm(research.data$temperature ~ c(1:kObservationNum))
82
83 ## Time series (which is by default is research data) with trend line based on
       linear module estimate (lm)
84 plot.data.ts <- DrawTimeSeries(data=research.data, filename="original/time-
       series.png", datebreaks=kDateBreaks)
85
86 ## Next step is research residuals computed few lines above
87 research.residuals <- data.frame("year"=research.data$year, "temperature"=
       research.data.fit$residuals)
88 print(xtable(research.residuals, caption="Временной ряд остатков.", label="table
       :residuals"), table.placement="H",
89       file="out/residual/data.tex")
90
91 ## Residuals time series (data have gotten on computing step: fitting linear
       model)
92 plot.residuals.ts <- DrawTimeSeries(data=research.residuals, filename="residual/
```

```
    time-series.png", datebreaks=kDateBreaks)
93
94 ## Descriptive statistics for residuals
95 research.residuals.dstats <- dstats.describe(research.residuals$temperature,
    locale=TRUE)
96 print(xtable(research.residuals.dstats, caption="Описательные статистики для ост
    атков.", label="table:residuals_dstats"),
97       file="out/residual/dstats.tex")
98
99 ## Basic histogram for residuals / seems like the same as for non-residuals
100 plot.residuals.hist <- DrawHistogram(data=research.residuals, filename="residual
    /histogram.png")
101
102 ## Tests for normality
103 research.data.shapiro <- ntest.ShapiroWilk(data=research.residuals$temperature,
    filename="out/residual/shapiro-test.tex")
104 research.data.pearson <- ntest.PearsonChi2(data=research.residuals$temperature,
    filename="out/residual/pearson-test.tex")
105 research.data.ks      <- ntest.KolmogorovSmirnov(data=research.residuals$
    temperature, filename="out/residual/ks-test.tex")
106
107 ## Normal Quantile-Quantile plot for residuals
108 plot.residuals.qq <- DrawQuantileQuantile(data=research.residuals$temperature,
    filename="residual/quantile.png")
109
110 ## Auto Correlation Function plot // TODO: check the style
111 plot.residuals.acf <- DrawAutoCorrelationFunction(data=research.data$temperature
    , filename="residual/acf.png")
112
113 ## Box-Ljung and adf tests (some kind of stationarity and independence tests) //
     TODO: need to know exactly in theory what it is
114 research.residuals.box <- Box.test(research.residuals$temperature, type="Ljung-
    Box")
115 to.file(research.residuals.box, "out/residual/ljung-test.tex")
116 research.residuals.adf <- adf.test(research.residuals$temperature)
117 to.file(research.residuals.adf, "out/residual/stationarity-test.tex")
118
119
120 source("R/variogram.R")
```

Листинг С.2: Основной код программы

```
1 source("R/archive/variogram_analysis/afv.R")
2
3 ## Function definition: need to be moved into isolated place
4
5 ### Just definition of mean standard error // TODO: find out exact formula and
    describe each parameter
6 MSE <- function (e, N=1) {
7   sum(sapply(X=e, FUN=function(x) x**2)) / length(e)
8 }
9
10 CompareClassicalModels <- function(manual, classical, filename) {
11   # Arrange the data for the ggplot2 plot
12   # add the semivariance values of v2 to v1
13   Fitted1 <- data.frame(dist = seq(.01, max(manual$exp_var$dist), length =
       kObservationNum))
14   Fitted1$gamma <- variogramLine(manual$var_model, dist_vector = Fitted1$dist)$
       gamma
15   #convert the dataframes to a long format
16   Empirical1 <- melt(manual$exp_var, id.vars = "dist", measure.vars = c("gamma")
```

```
    )
  Modeled1 <- melt(Fitted1, id.vars = "dist", measure.vars = c("gamma"))

  Fitted2 <- data.frame(dist = seq(.01, max(classical$exp_var$dist), length =
      kObservationNum))
  Fitted2$gamma <- variogramLine(classical$var_model, dist_vector = Fitted2$dist
      )$gamma
  #convert the dataframes to a long format
  Empirical2 <- melt(classical$exp_var, id.vars = "dist", measure.vars = c("
      gamma"))
  Modeled2 <- melt(Fitted2, id.vars = "dist", measure.vars = c("gamma"))

  plot.modeled <- ggplot(Empirical1, aes(x = dist, y = value)) + geom_point() +
    geom_line(data = Modeled1, linetype="dashed") +
    geom_line(data = Modeled2) +
    labs(color="") +
    scale_y_continuous(expand=c(0,0),
      breaks=seq(0, 1.04 * max(manual$exp_var$gamma), 1),
      limits=c(min(0, 1.04 * min(manual$exp_var$gamma)), 1.04 * max(manual$exp_
          var$gamma))) +
    scale_x_continuous(expand=c(0,0),
      breaks=seq(0, 1.04 * max(manual$exp_var$dist), 1),
      limits=c(0, 1.04 * max(manual$exp_var$dist))) +
    xlab("Расстояние") + ylab("Значение")
  ggsave(plot=plot.modeled, file=filename, width=7, height=4)

  plot.modeled
}

### Missed complete understanding of this functionality, because it aren't used
    in further work. Seems like it used only for selection best parameters.
### Compares two predictions classical and robust in case of iterating through '
    cutoff' param based on MSE estimation.
#### todo: simpify this function, split it to several less complex functions
ComparePredictionParameters <- function (data, trend, x, y=rep(1,
    kObservationNum), width=1, filename) {
  lens <- 1:kObservationNum
  manualResult <- c()
  classicalResult <- c()
  robustResult <- c()

  spdata <- data.frame(cbind("x"=x, "y"=y, data))
  coordinates(spdata)=~x+y

  i <- 1
  for(l in lens) {
    variogram.manual    = ComputeManualVariogram(data, cutoff=l)
    variogram.classical = autofitVariogram(data~1, spdata, cutoff=l, cressie=
        FALSE, width=width)
    variogram.robust    = autofitVariogram(data~1, spdata, cutoff=l, cressie=
        TRUE, width=width)

    kriging.manual    <- PredictWithKriging(data, x=x, variogram_model=variogram
        .manual$var_model)
    kriging.classical <- PredictWithKriging(data, x=x, variogram_model=variogram
        .classical$var_model)
    kriging.robust    <- PredictWithKriging(data, x=x, variogram_model=variogram
        .robust$var_model)

    res.manual    <- CrossPrediction(src.data$temperature, src.data$year, trend,
```

```
          kriging.manual)
64      res.classical <- CrossPrediction(src.data$temperature, src.data$year, trend,
          kriging.classical)
65      res.robust    <- CrossPrediction(src.data$temperature, src.data$year, trend,
          kriging.robust)
66
67      manualResult[i]    <- MSE(e=res.manual)
68      classicalResult[i] <- MSE(e=res.classical)
69      robustResult[i]    <- MSE(e=res.robust)
70      i = i + 1 ## todo: find out how to avoid this construction
71    }
72
73    plot.check <- ggplot() +
74      geom_line(data=data.frame("X"=lens, "Y"=manualResult), aes(x=X,y=Y)) +
75      geom_line(data=data.frame("X"=lens, "Y"=classicalResult), aes(x=X,y=Y),
          linetype="dashed") +
76      geom_line(data=data.frame("X"=lens, "Y"=robustResult), aes(x=X,y=Y),
          linetype="dotdash") +
77      scale_x_continuous(breaks=lens)
78    ggsave(plot=plot.check, file=filename, width=7, height=4)
79  }
80
81  ### This comparison is worth than above one, the estimation of it's goodness is
      simpler // TODO: check, maybe it should be removed.
82  ### I don't see the difference and profit of this kind of comparison. Maybe it
      should be changed to more universal way (e.g. to pass estimation function).
83  ### Update. Now I feel the difference. Above case is better but may be both have
       rights to live together, will see.
84  ### Update of update. Hmm, this one compares only variogram calculations. The
      estimate based on sserr divided by length.
85  CompareVariogramParameters <- function (data, x, y=rep(1, kObservationNum),
      width) {
86    lens <- 1:kObservationNum
87    classicalResult <- c()
88    robustResult <- c()
89
90    spdata <- data.frame(cbind("x"=x, "y"=y, data))
91    coordinates(spdata) = ~x+y
92
93    i <- 1
94    for(l in lens) {
95      variogram.classical = autofitVariogram(data~1, spdata, cutoff=l, cressie=
          FALSE, width=width)
96      variogram.robust = autofitVariogram(data~1, spdata, cutoff=l, cressie=TRUE,
          width=width)
97      classicalResult[i] <- variogram.classical$sserr / l
98      robustResult[i] <- variogram.robust$sserr / l
99      i = i + 1
100   }
101
102   ggplot() +
103     geom_line(data=data.frame("X"=lens, "Y"=classicalResult), aes(x=X, y=Y,
          color="classic")) +
104     geom_line(data=data.frame("X"=lens, "Y"=robustResult), aes(x=X, y=Y, color="
          cressie")) +
105     scale_x_continuous(breaks=lens) +
106     scale_y_continuous(breaks=seq(1.04 * min(classicalResult, robustResult),
          1.04 * max(classicalResult, robustResult), 1))
107 }
108
```

```
109 ComputeManualVariogram <- function (data, cutoff, file=FALSE, file_modeled="") {
110   # Make fake second coordinate
111   p <- data.frame("X"=c(1:kObservationNum), "Y"=rep(1, kObservationNum))
112   coordinates(p) <- ~ X + Y
113   experimental_variogram <- variogram(data~1, p, width=1, cutoff=cutoff)
114
115   model.variog <- vgm(model="Sph", range=3.9, nugget=3.4)
116   fit.variog <- fit.variogram(experimental_variogram, model.variog)
117
118   if (file) {
119     # Arrange the data for the ggplot2 plot
120     # add the semivariance values of v2 to v1
121     Fitted <- data.frame(dist = seq(0.01, max(experimental_variogram$dist),
          length = kObservationNum))
122     Fitted$gamma <- variogramLine(fit.variog, dist_vector = Fitted$dist)$gamma
123     #convert the dataframes to a long format
124     Empirical <- melt(experimental_variogram, id.vars = "dist", measure.vars = c
          ("gamma"))
125     Modeled <- melt(Fitted, id.vars = "dist", measure.vars = c("gamma"))
126
127     plot.modeled <- ggplot(Empirical, aes(x = dist, y = value)) + geom_point()
          +
128       geom_line(data = Modeled, color='blue') +
129       scale_y_continuous(expand=c(0,0),
130                          breaks=seq(0, 1.04 * max(experimental_variogram$gamma),
                               1),
131                          limits=c(min(0, 1.04 * min(experimental_variogram$gamma
                               )), 1.04 * max(experimental_variogram$gamma))) +
132       scale_x_continuous(expand=c(0,0),
133                          breaks=seq(0, 1.04 * max(experimental_variogram$dist),
                               1),
134                          limits=c(0, 1.04 * max(experimental_variogram$dist))) +
135       xlab("Расстояние") + ylab("Значение")
136     ggsave(plot=plot.modeled, file=file_modeled, width=7, height=4)
137   }
138   print(xtable(data.frame("Модель"=fit.variog$model, "Порог"=fit.variog$psill, "
        Ранг"=fit.variog$range), caption="Модель вариограммы", label="table:manual_
        model"), table.placement="H",
139         file="out/variogram/manual-model.tex")
140   result = list(exp_var = experimental_variogram, var_model = fit.variog)
141 }
142
143 ## Calculates modeled variogram and creates plot of it.
144 ComputeVariogram <- function (data, x, y=rep(1, kObservationNum), file_empirical
      ="", file_modeled="", cressie, cutoff, width) {
145   spdata <- data.frame(cbind("x"=x, "y"=y, data))
146   coordinates(spdata) = ~x+y
147
148   variogram <- autofitVariogram(data~1, spdata, cutoff=cutoff, cressie=cressie,
        width=width)
149   if (nchar(file_empirical)) { ## here was another check: just <file>
150     # Arrange the data for the ggplot2 plot
151     # add the semivariance values of v2 to v1
152     Fitted <- data.frame(dist = seq(.01, max(variogram$exp_var$dist), length =
          kObservationNum))
153     Fitted$gamma <- variogramLine(variogram$var_model, dist_vector = Fitted$dist
          )$gamma
154     #convert the dataframes to a long format
155     Empirical <- melt(variogram$exp_var, id.vars="dist", measure.vars=c("gamma")
          )
```

```
156     Modeled <- melt(Fitted, id.vars="dist", measure.vars=c("gamma"))
157
158     plot.empirical <- ggplot(Empirical, aes(x=dist, y=value)) + geom_point() +
159       scale_y_continuous(expand = c(0, 0), breaks=seq(0, 7, 1), limits=c(min(0,
              1.04 * min(variogram$exp_var$gamma)), 1.04 * max(variogram$exp_var$
              gamma))) +
160       scale_x_continuous(expand = c(0, 0), breaks=seq(0, 1.04 * max(variogram$
              exp_var$dist), 2), limits=c(0, 1.04 * max(variogram$exp_var$dist))) +
161       xlab("Расстояние") + ylab("Значение")
162     ggsave(plot=plot.empirical, file=file_empirical, width=7, height=4)
163   }
164   if (nchar(file_modeled)) {
165     plot.modeled <- ggplot(Empirical, aes(x=dist, y=value)) + geom_point() +
166       geom_line(data=Modeled, color='blue') +
167       scale_y_continuous(expand=c(0, 0),
168                          breaks=seq(0, 1.04 * max(variogram$exp_var$gamma), 1),
169                          limits=c(min(0, 1.04 * min(variogram$exp_var$gamma)),
                                 1.04 * max(variogram$exp_var$gamma))) +
170       scale_x_continuous(expand=c(0, 0),
171                          breaks=seq(0, 1.04 * max(variogram$exp_var$dist), 1),
172                          limits=c(0, 1.04 * max(variogram$exp_var$dist))) +
173       xlab("Расстояние") + ylab("Значение")
174     ggsave(plot=plot.modeled, file=file_modeled, width=7, height=4)
175   }
176   #   plot(variogram)
177   variogram
178 }
179
180 ## Calculates kriging prediction based on passed varigram model
181 PredictWithKriging <- function (data, x, y=rep(1, kObservationNum), variogram_
      model, future=0) {
182   src_data <- data.frame(cbind("x"=x, "y"=y, data))
183   coordinates(src_data) = ~x+y
184
185   new_data <- data.frame("X"=c((kObservationNum + 1):(src.nrows + future)), "Y"=
        rep(1, src.nrows - kObservationNum + future))
186   coordinates(new_data) = ~X+Y
187
188   krige(data~1, src_data, new_data, model=variogram_model)
189 }
190
191 ## Compares predictions based on trend and kriging with actual values
192 CrossPrediction <- function (temperature, years, trend, kriging, file_prediction
      ="", future=0) {
193   prediction.trend <- data.frame("temperature"=c(temperature[(kObservationNum -
        1):kObservationNum], trend[(kObservationNum + 1):src.nrows]),
194                                  "year"=GetPredictionYears(years, src.nrows,
                                       future))
195
196   prediction.kriging <- data.frame("temperature"=c(temperature[(kObservationNum
        - 1):kObservationNum], trend[(kObservationNum + 1):src.nrows] + kriging$
        var1.pred),
197                                    "year"=GetPredictionYears(years, src.nrows,
                                         future))
198
199   actual <- data.frame("temperature"=temperature[(kObservationNum - 1):src.nrows
        ],
200                        "year"=GetPredictionYears(years, src.nrows, 0))
201
202   if (nchar(file_prediction)) {
```

```
203     plot.crossprediction <- ggplot() +
204       geom_line(data=prediction.kriging, aes(x=year, y=temperature, color="Прогноз Кригинг")) +
205       geom_line(data=prediction.trend, aes(x=year, y=temperature, color="Прогноз Тренд")) +
206       geom_line(data=actual, aes(x=year, y=temperature, colour="Актуальное")) +
207       labs(color="") +
208       scale_x_continuous(breaks=seq(min(actual$year), max(actual$year) + 5 + future, by=1)) + xlab("Год наблюдения") +
209       scale_y_continuous(breaks=seq(16, 28, .5)) + ylab("Температура, C") +
210       theme(axis.text.x = element_text(angle=45, hjust=1)) +
211       labs(color="")
212     ggsave(plot=plot.crossprediction, file=file_prediction, width=7, height=4)
213   }
214
215   prediction.kriging$temperature[3:(src.nrows-kObservationNum)] - actual$temperature[3:(src.nrows - kObservationNum)] ## what the heck? why 3?
216 }
217
218 ### once it was like this kObservationNum <- 32
219
220 ### src <- read.csv(file="data/batorino_july.csv", header=TRUE, sep=";", nrows=38, colClasses=c("numeric", "numeric"), stringsAsFactors=FALSE)
221
222 # Completes trend values to source observation number
223 computeTrend <- function (fit, future=0) {
224   c(sapply(c(1 : (src.nrows + future)), FUN=function(x) fit$coefficients[[1]] + x * fit$coefficients[[2]]))
225 }
226
227 kObservationNum <- 32
228
229 ## Form the data for research again
230 research.data <- src.data[0:kObservationNum, ]
231
232 research.data.fit <- lm(research.data$temperature ~ ConvertYearsToNum(research.data$year))
233 research.data.residuals <- research.data.fit$residuals
234 research.data.trend <- computeTrend(research.data.fit)
235
236 cutoff <- trunc(2 * kObservationNum / 3) # let it be "classical" value
237 #cutoff <- 2
238
239 # Compute variogram manually with choosed model (best what i could found)
240 variogram.manual <- ComputeManualVariogram(research.data.residuals, cutoff=cutoff, file=TRUE, file_modeled="figures/variogram/manual-model.png")
241
242 # Compute variogram with auto fit model using classical estimation
243 variogram.classical <- ComputeVariogram(data=research.data.residuals, x=ConvertYearsToNum(research.data$year), cressie=FALSE, cutoff=cutoff, width=FALSE,
244                                         file_empirical="figures/variogram/classical-empirical.png",
245                                         file_modeled="figures/variogram/classical-modeled.png")
246
247 # Compute variogram with auto fit model using robust (cressie) estimation
248 variogram.robust <- ComputeVariogram(data=src.data.residuals, x=ConvertYearsToNum(research.data$year), cressie=TRUE, cutoff=cutoff, width=FALSE,
```

```
249                                                 file_empirical="figures/variogram/robust-
                                                        empirical.png",
250                                                 file_modeled="figures/variogram/robust-
                                                        modeled.png")
251
252 models.comparison <- CompareClassicalModels(variogram.manual, variogram.
        classical, filename="figures/variogram/models-comparison.png")
253
254 kriging.manual    <- PredictWithKriging(research.data.residuals, x=
        ConvertYearsToNum(research.data$year), variogram_model=variogram.manual$var_
        model)
255 kriging.classical <- PredictWithKriging(research.data.residuals, x=
        ConvertYearsToNum(research.data$year), variogram_model=variogram.classical$
        var_model)
256 kriging.robust    <- PredictWithKriging(research.data.residuals, x=
        ConvertYearsToNum(research.data$year), variogram_model=variogram.robust$var_
        model)
257
258 mse.manual    <- MSE(CrossPrediction(src.data$temperature, src.data$year,
        research.data.trend, kriging.manual))
259 mse.classical <- MSE(CrossPrediction(src.data$temperature, src.data$year,
        research.data.trend, kriging.classical))
260 mse.robust    <- MSE(CrossPrediction(src.data$temperature, src.data$year,
        research.data.trend, kriging.robust))
261
262 res.ma <- CrossPrediction(src.data$temperature, src.data$year, research.data.
        trend, kriging.manual, "figures/variogram/cross-prediction-manual.png")
263 res.cl <- CrossPrediction(src.data$temperature, src.data$year, research.data.
        trend, kriging.classical, "figures/variogram/cross-prediction-classical.png")
264 res.ro <- CrossPrediction(src.data$temperature, src.data$year, research.data.
        trend, kriging.robust, "figures/variogram/cross-prediction-robust.png")
265
266 # Find best cutoff parameter
267 ComparePredictionParameters(research.data.residuals, research.data.trend,
        ConvertYearsToNum(research.data$year), filename="figures/variogram/parameter-
        comparison.png")
268
269 # Best prediction as we investigated is for robust kriging with cutoff=6. Let's
        make it!
270 variogram.robust.best <- ComputeVariogram(data=research.data.residuals, x=
        ConvertYearsToNum(research.data$year), cressie=TRUE, cutoff=6, width=FALSE,
271                                                 file_empirical="figures/variogram/
                                                        robust-best-empirical.png",
272                                                 file_modeled="figures/variogram/robust
                                                        -best-modeled.png")
273 kriging.robust.best <- PredictWithKriging(research.data.residuals, x=
        ConvertYearsToNum(research.data$year), variogram_model=variogram.robust.best$
        var_model)
274 mse.robust.best <- MSE(CrossPrediction(src.data$temperature, src.data$year,
        research.data.trend, kriging.robust.best))
275 res.ro.best <- CrossPrediction(src.data$temperature, src.data$year, research.
        data.trend, kriging.robust.best, "figures/variogram/cross-prediction-robust-
        best.png")
276
277 ## TODO: form krige matrix for analysis
```

Листинг С.3: Вариограммный анализ