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** **2**

**1 Случайный процесс и его характеристики. Стационарность случайных процессов. Вариограмма** **4**
1.1 Случайный процесс. Стационарность . . . . . 4
1.2 Вариограмма и внутренне стационарный случайный процесс . . . . . 5

**2 ?? Теория ??** **6**
2.1 Оценка вариограммы гауссовского случайного процесса . . . . . 6

**3 Обработка временного ряда с помощью R** **9**
3.1 Вычисление основных описательных статистик . . . . . 9
3.2 Исследование статистических данных . . . . . 11
3.3 Корреляционный анализ . . . . . 14

**Заключение** **17**

**Литература** **19**

# 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 $\Omega$ является произвольным множеством, $\mathcal{F}$ — сигма-алгеброй подмножеств $\Omega$, и $P$ — вероятностной мерой.

**Определение 1.1.** Попробуй сделать через дефинишены, должно получаться красивее и можно будет на них ссылаться

*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.

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

Если $\mathbb{T} = \mathbb{R}$, то $X(t), t \in \mathbb{T}$ называют *случайным процессом с непрерывным временем.*

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

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

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

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

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

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

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

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

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

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

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

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:

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

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. $cov(t_1, t_2) = cov(t_1 - t_2), t_1, t_2 \in \mathbb{T}$.

*Замечание* 1.1. Если случайный процесс $X(t), t \in \mathbb{T}$, является стационарным в узком смысле и $\exists E\{x^2(t)\} < \infty, t \in \mathbb{T}$, то он будет стационарным и в широком смысле, но не наоборот.

## 1.2 Вариограмма и внутренне стационарный случайный процесс

Случайный процесс $X(t), t \in \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$, 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, $t_1, t_2 \in \mathbb{Z}$.

В дальнейшем рассматриваем случайные процессы с дискретным временем.

Пусть $X(t), t \in \mathbb{Z}$ — внутренне стационарный гауссовский случайный процесс с нулевым математическим ожиданием, дисперсией $\sigma^2$ и неизвестной вариограммой.

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

Заметим, что

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

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

# Глава 2
# ?? Теория ??

## 2.1 Оценка вариограммы гауссовского случайного процесса

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

Вычислим математическое ожидание введённой оценки

$$E\{2\tilde{\gamma}(h)\} = \tfrac{1}{n-h}\sum_{t=1}^{n-h} E\{(X(t+h)-X(t))^2\} =$$
$$= [\text{так как процесс является внутренне стационарным}] =$$
$$= \tfrac{1}{n-h}\sum_{t=1}^{n-h} 2\gamma(h) = 2\gamma(h).$$

Таким образом оценка является несмещённой.

Далее, найдём ковариацию:

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

По определению, $cov\{a,b\} = corr\{a,b\}\sqrt{V\{a\}V\{b\}}$, тогда

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

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

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

Далее воспользуемся леммой 1 из [8]:

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

Воспользовавшись леммой 3 из [8], получаем соотношение

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

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 $h_1 < h_2$.

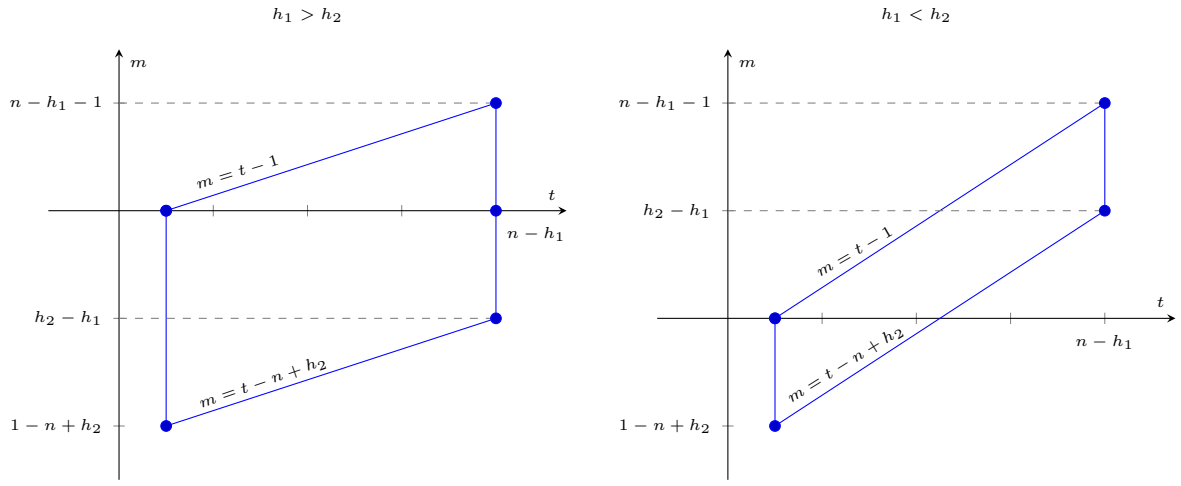


Рисунок 2.1.1 — Замена переменных

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 **??** в таблице **??**. Графически исходные данные представлены на рисунке 3.1.1.

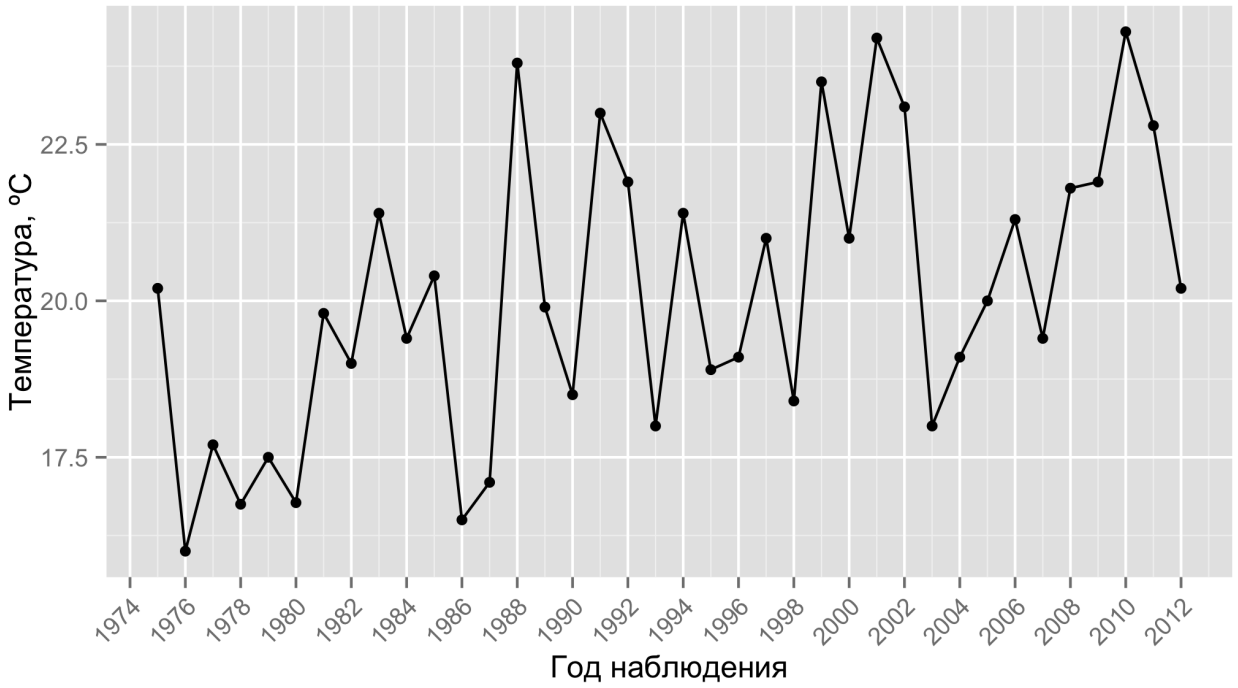


Рисунок 3.1.1 — График исходных данных.

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–**??**, 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 **??** листинге **??**. 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

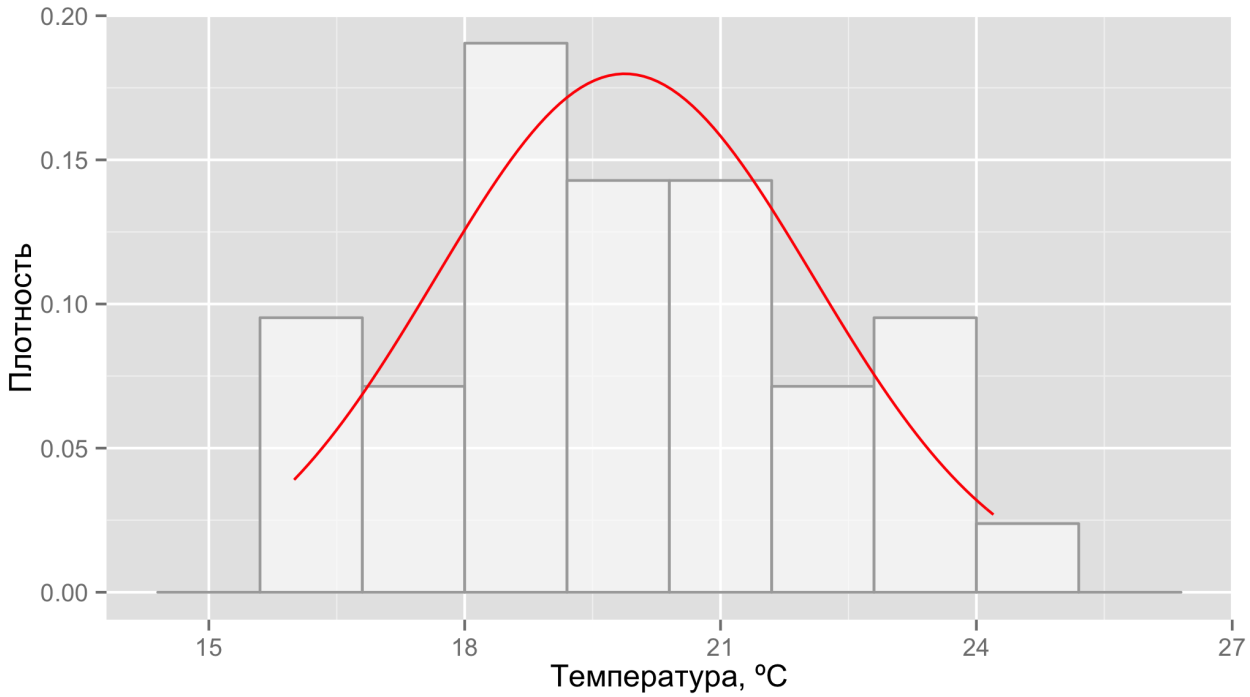


Рисунок 3.2.2 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения

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.

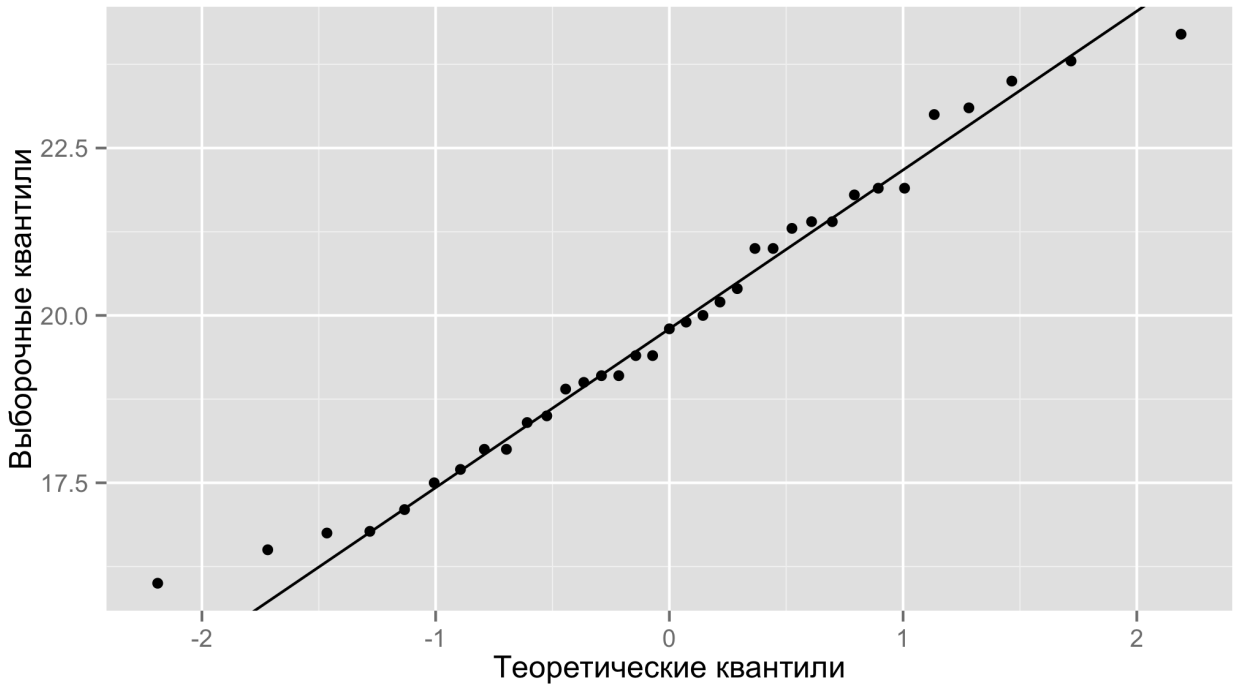


Рисунок 3.2.3 — График квантилей для наблюдаемых температур

Попробуем опровегнуть наше предположение на основе проверки критерия $\chi^2$ 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 $\chi^2$ 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 $\chi^2$ найдём критическое значение критерия $P_{\text{кр}}(\alpha, k) = 43.8$. Отсюда следует, что

$$P < P_{\text{кр}}.$$

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

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

```
Two-sample Kolmogorov-Smirnov test

data:  data and nsample
D = 0.0663, p-value = 0.9979
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.

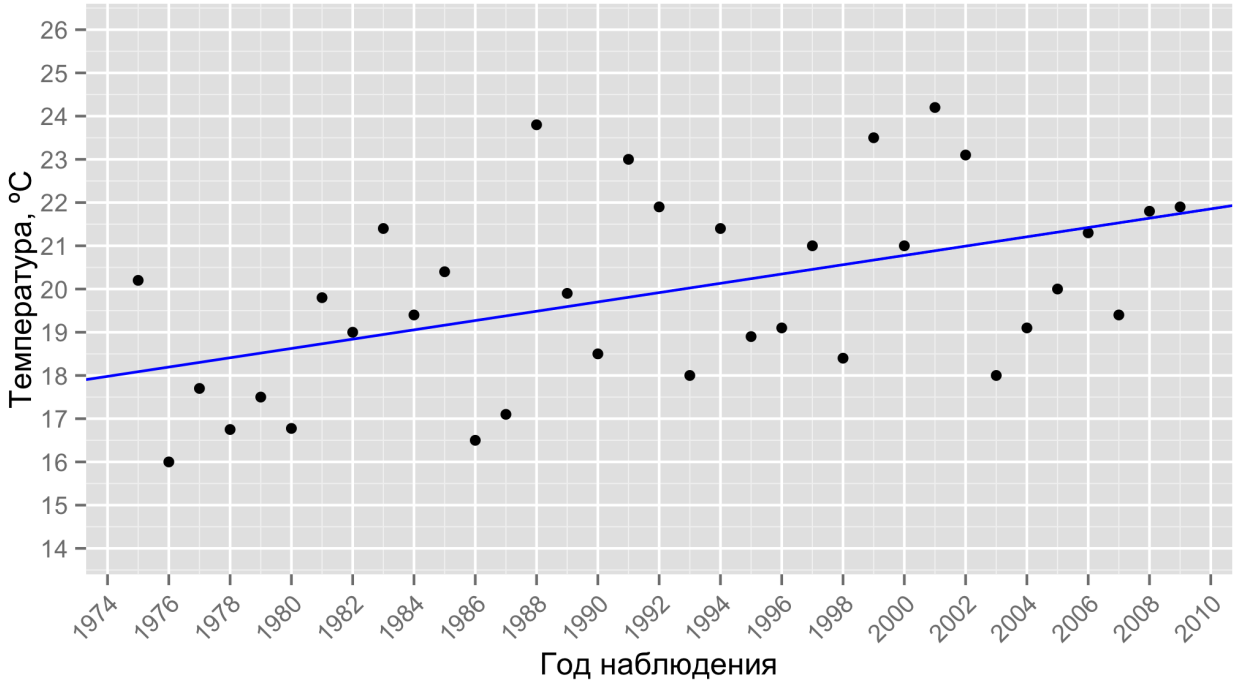


Рисунок 3.3.4 — Диаграмма рассеяния

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 из таблицы **??**, 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.

# 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.