

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

Павлов Александр Сергеевич

Анализ и статистическая 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

<b>Введение</b>	<b>2</b>
<b>1 Определения и вспомогательные результаты</b>	<b>4</b>
1.1 Случайный процесс . . . . .	4
1.2 Вариограмма . . . . .	5
<b>2 ?? Теория ??</b>	<b>6</b>
2.1 Оценка вариограммы гауссовского случайного процесса . . . . .	6

## 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 [1–3], 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 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

## Определения и вспомогательные результаты

### 1.1 Случайный процесс

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

*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  $\mathbb{T} = \mathbb{Z} = 0, \pm 1, \pm 2, \dots$ , 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

*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.

Далее будем рассматривать случайный процесс с дискретным временем.

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

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

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

## 1.2 Вариограмма

## Глава 2

### ?? Теория ??

## 2.1 Оценка вариограммы гауссовского случайного процесса

В качестве оценки вариограммы рассмотрим статистику вида:

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

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

Вычислим математическое ожидание введённой оценки

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

Таким образом оценка является несмещённой.

Далее, найдём ковариацию:

$$\begin{aligned} \text{cov}\{2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)\} &= E\{(2\tilde{\gamma}(h_1) - E\{2\tilde{\gamma}(h_1)\})(2\tilde{\gamma}(h_2) - E\{2\tilde{\gamma}(h_2)\})\} = \\ &= E\left\{\frac{1}{n-h_1} \sum_{t=1}^{n-h_1} ((X(t+h_1) - X(t))^2 - E\{(X(t+h_1) - X(t))^2\}) \times \right. \\ &\quad \times \left. \frac{1}{n-h_2} \sum_{s=1}^{n-h_2} ((X(s+h_2) - X(s))^2 - E\{(X(s+h_2) - X(s))^2\})\right\} = \\ &= \frac{1}{(n-h_1)(n-h_2)} \sum_{t=1}^{n-h_1} \sum_{s=1}^{n-h_2} \text{cov}\{(X(t+h_1) - X(t))^2, (X(s+h_2) - X(s))^2\} \quad (2.2) \end{aligned}$$

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

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

Принимая во внимание  $V\{X(t+h) - X(t)\}^2 = 2(2\gamma(h))^2$  [ссылка на источник, либо упоминание раньше] и предыдущее соотношение, из (2.2) получаем:

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

Далее воспользуемся леммой 1 [Брест2005]:

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

Воспользовавшись леммой 3 [Брест2005], получаем соотношение

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

В (2.3) сделаем следующую замену переменных

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

Получим

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

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

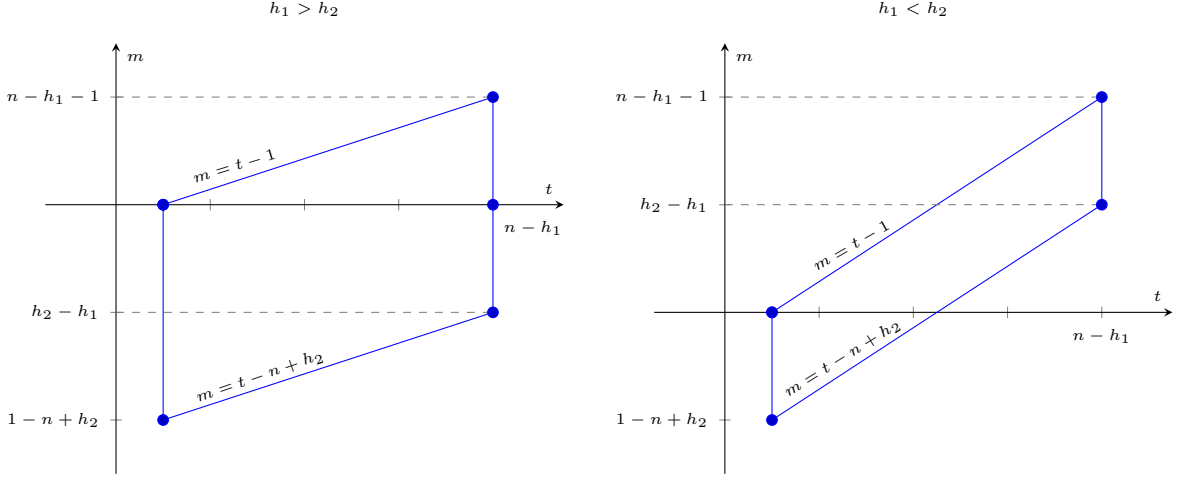


Рисунок 2.1.1 — Замена переменных

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

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

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

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

Разделим каждое слагаемое на общий член  $n-h_1$ :

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



## Литература

1. Stephen L. Katz, Stephanie E. Hampton, LyubovR. Izmet'seva, 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. W.W. Taylor A.S. Briggs O'Brien, T.P. and E.F. Roseman. Influence of water temperature on rainbow smelt spawning and early life 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.