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 Определения и вспомогательные результаты **4**

1.1 Случайный процесс 4

1.2 Вариограмма 5

2 ?? Теория ?? **6**

2.1 Оценка вариограммы гауссовского случайного процесса 6

Заключение **9**

Литература **10**

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
Определения и вспомогательные результаты

1.1 Случайный процесс

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.

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:

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

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.

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

1.2 Вариограмма

Пусть $X(t), t \in \mathbb{Z}$ — внутренне стационарный гауссовский случайный процесс с нулевым математическим ожиданием, дисперсией σ^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

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

Таким образом оценка является несмещённой.

Далее, найдём ковариацию:

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

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

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

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

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

Далее воспользуемся леммой 1 из [8]:

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

Воспользовавшись леммой 3 из [8], получаем соотношение

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

В (2.3) сделаем следующую замену переменных

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

Получим

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

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

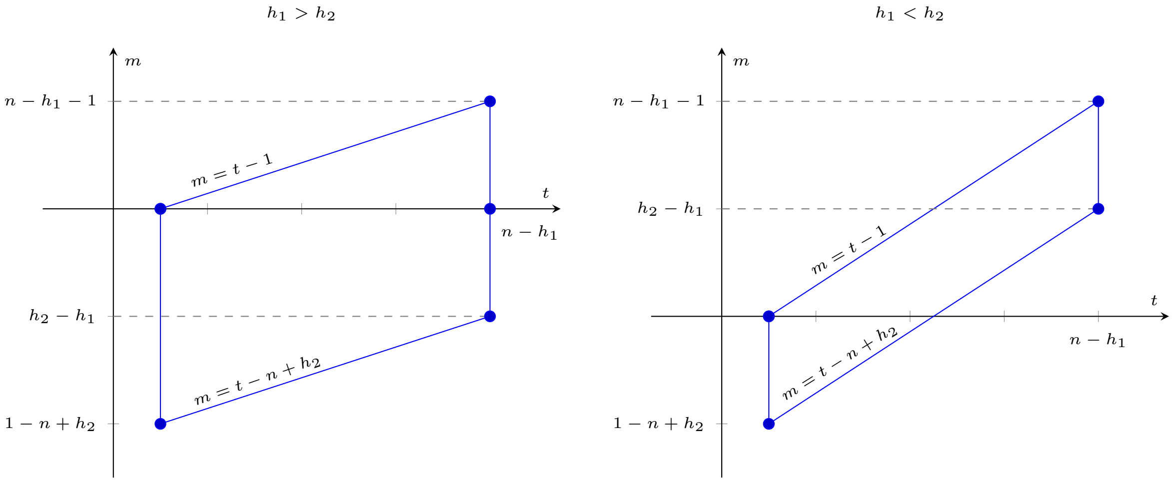

Рисунок 2.1.1 — Замена переменных

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

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

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

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

Преобразуем полученное выражение:

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

Приведем подобные:

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

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.