БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Анализ и прогнозирование гидрологических данных

Дипломная работа

Александр Сергеевич Павлов

Научный руководитель: Цеховая Татьяна Вячеславовна

Факультет прикладной математики и информатики

Кафедра теории вероятностей и математической статистики

Минск, 2015

Содержание

Постановка задачи

1. Предварительный статистический анализ гидроэкологических данных озера Баторино;
2. Вариограммный анализ временного ряда: построение оценок семивариограммы, подбор моделей семивариограммы;
3. Исследование статистических свойств оценки семивариограммы гауссовского случайного процесса;
4. Прогнозирование значений временного ряда с помощью интерполяционного метода Кригинг;
5. Исследование точности прогноза в зависимости от оценки семивариограммы и модели семивариограммы, лежащих в основе метода Кригинг.

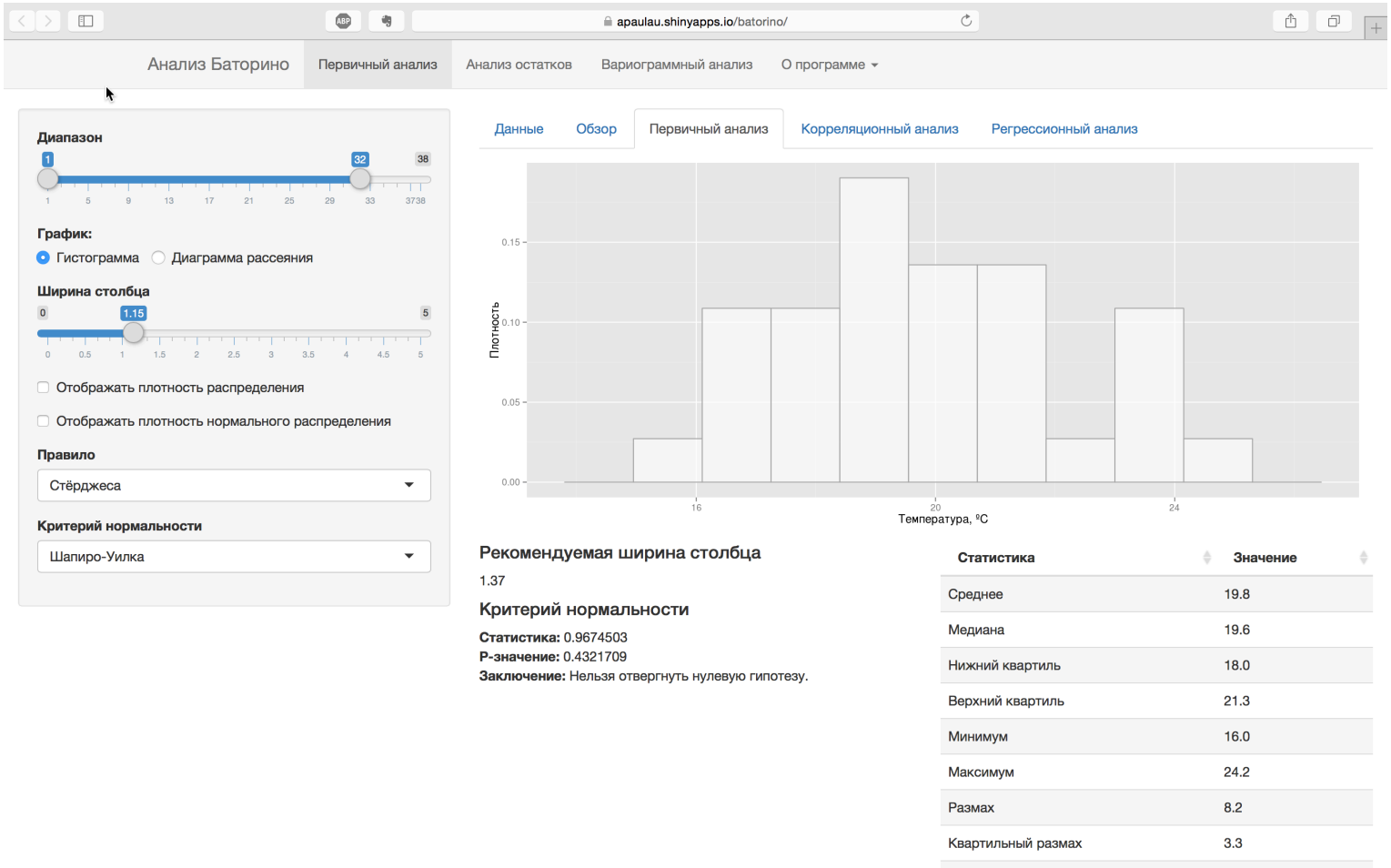
Обзор реализованного программного обеспечения

Особенности

- Доступно с любого устройства, имеющего доступ в Интернет, по адресу apaulau.shinyapps.io/batorino;
- Реализовано на языке программирования **R**;
- Логически разделено на три модуля;
- Имеет простой, быстро расширяемый гибкий интерфейс;
- Широкие графические возможности;
- Проверка тестов и критериев;
- Мгновенный отклик на изменение параметров;
- Быстрая проверка различных моделей.

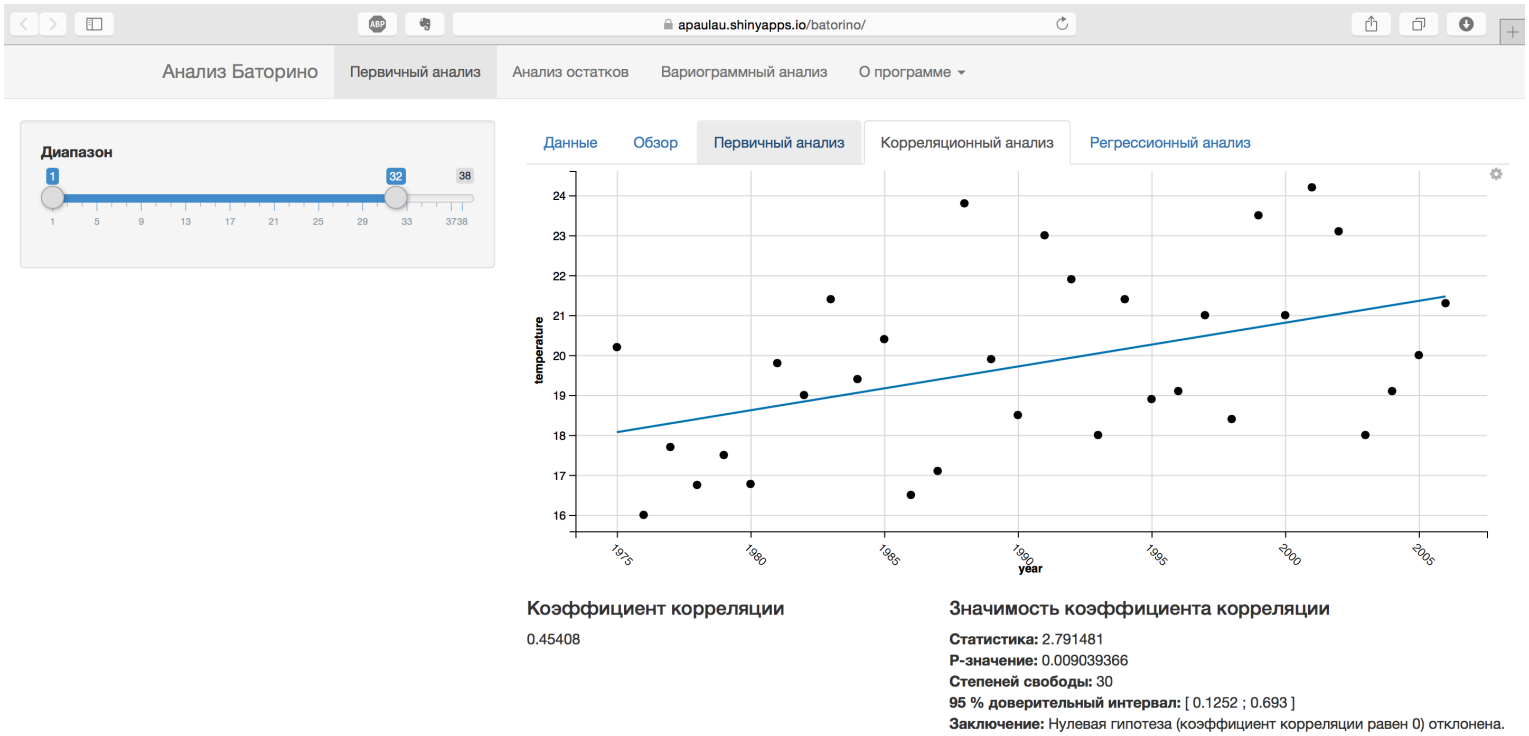
Модуль предварительного анализа

Первичный анализ и описательные статистики



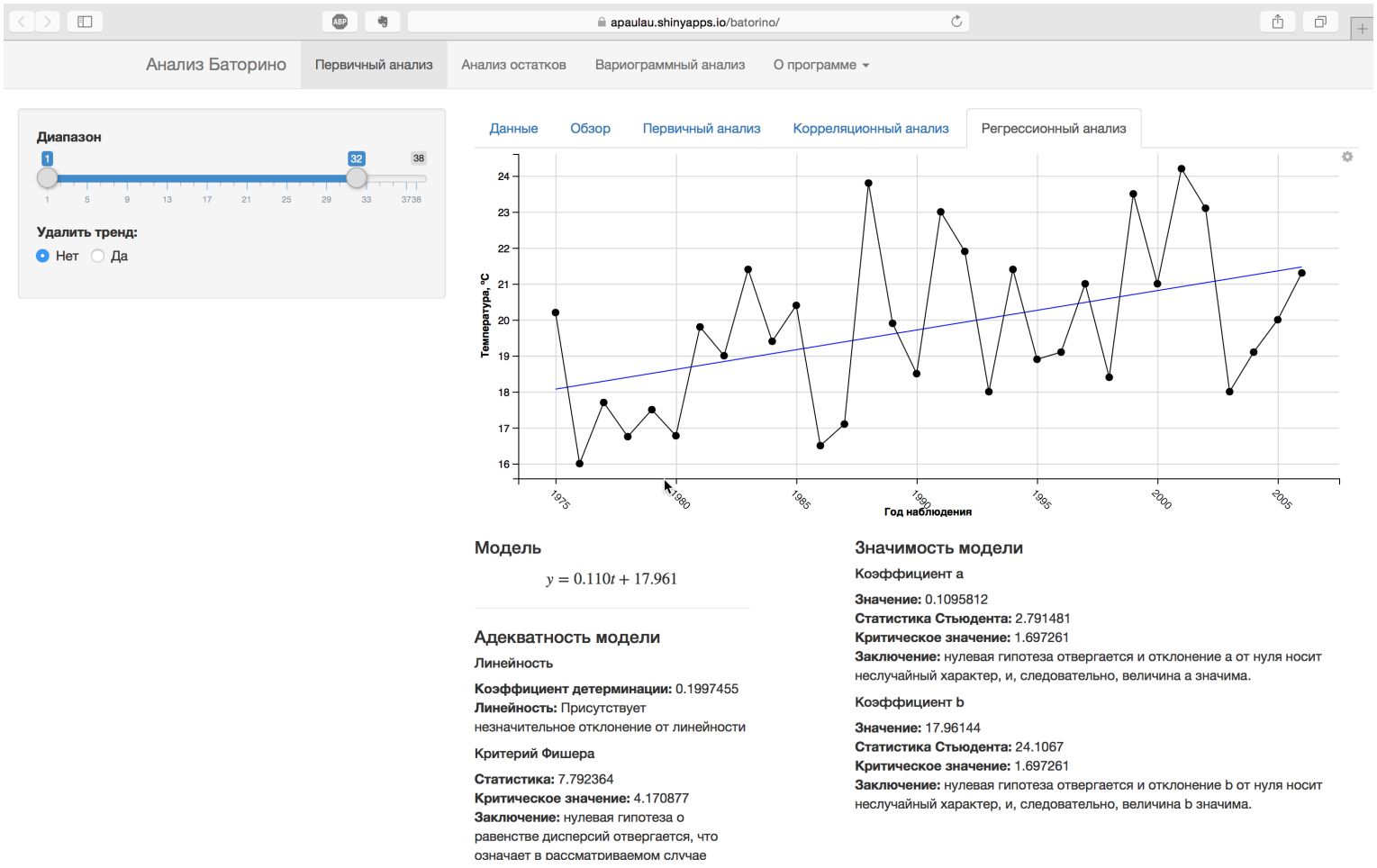
Модуль предварительного анализа

Корреляционный анализ



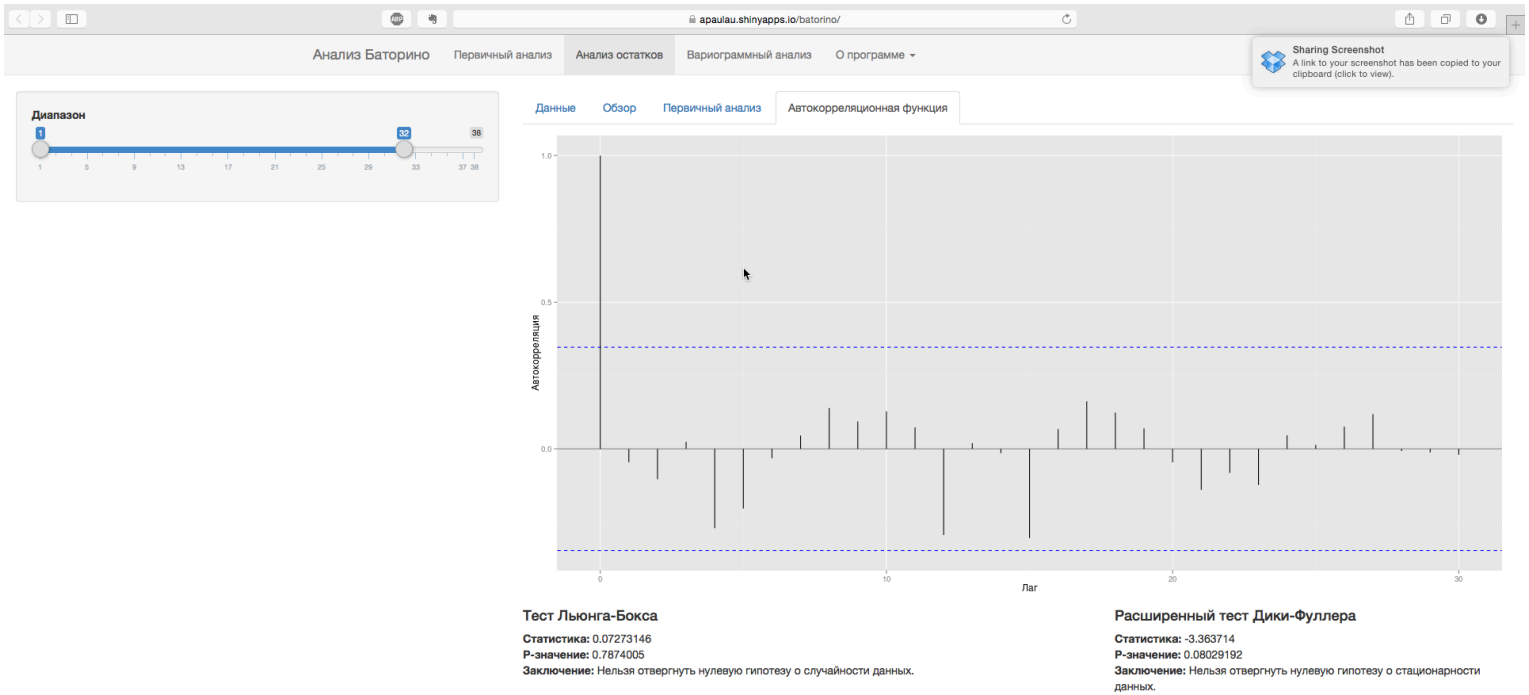
Модуль предварительного анализа

Регрессионный анализ



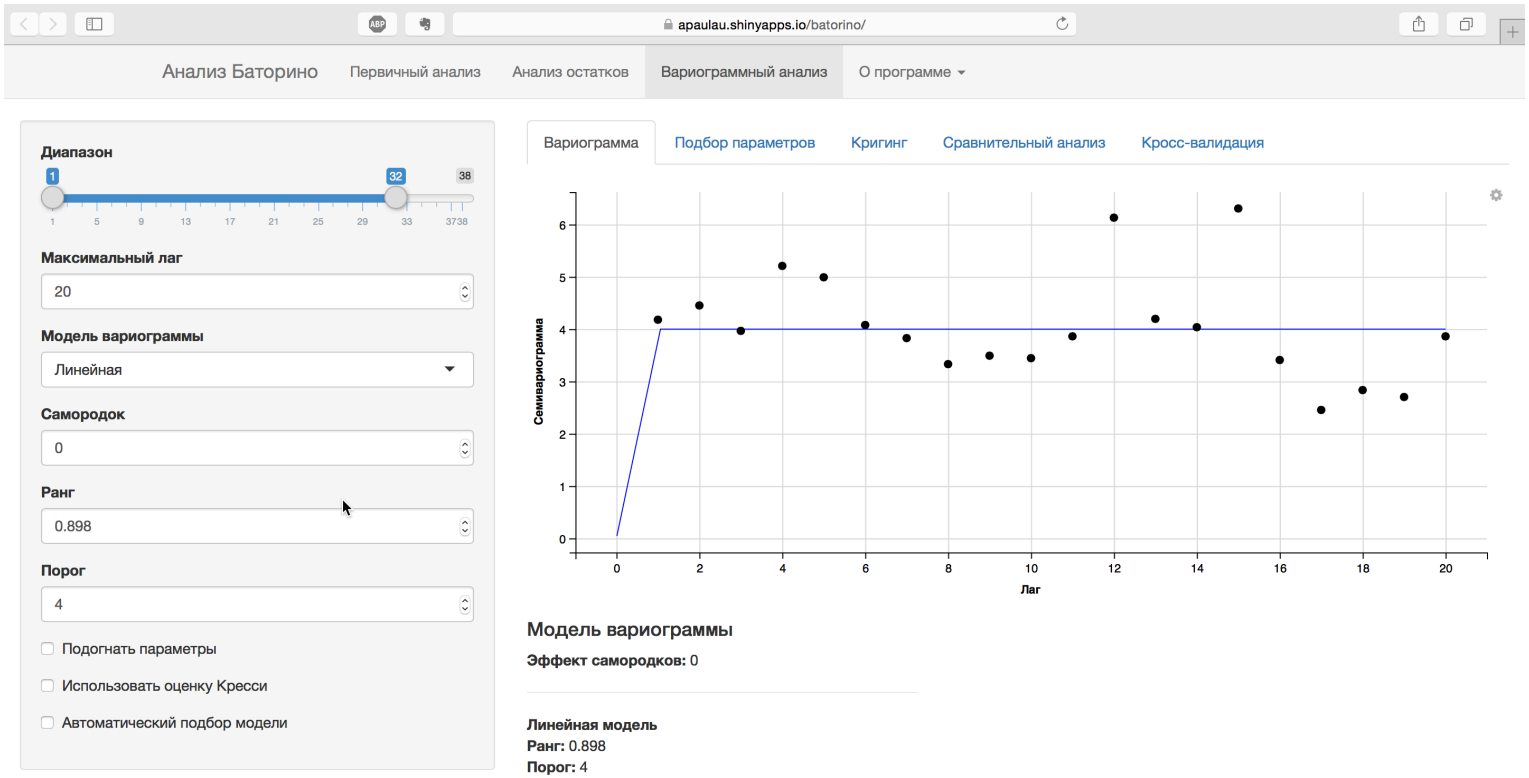
Модуль анализа остатков

Автокорреляционная функция



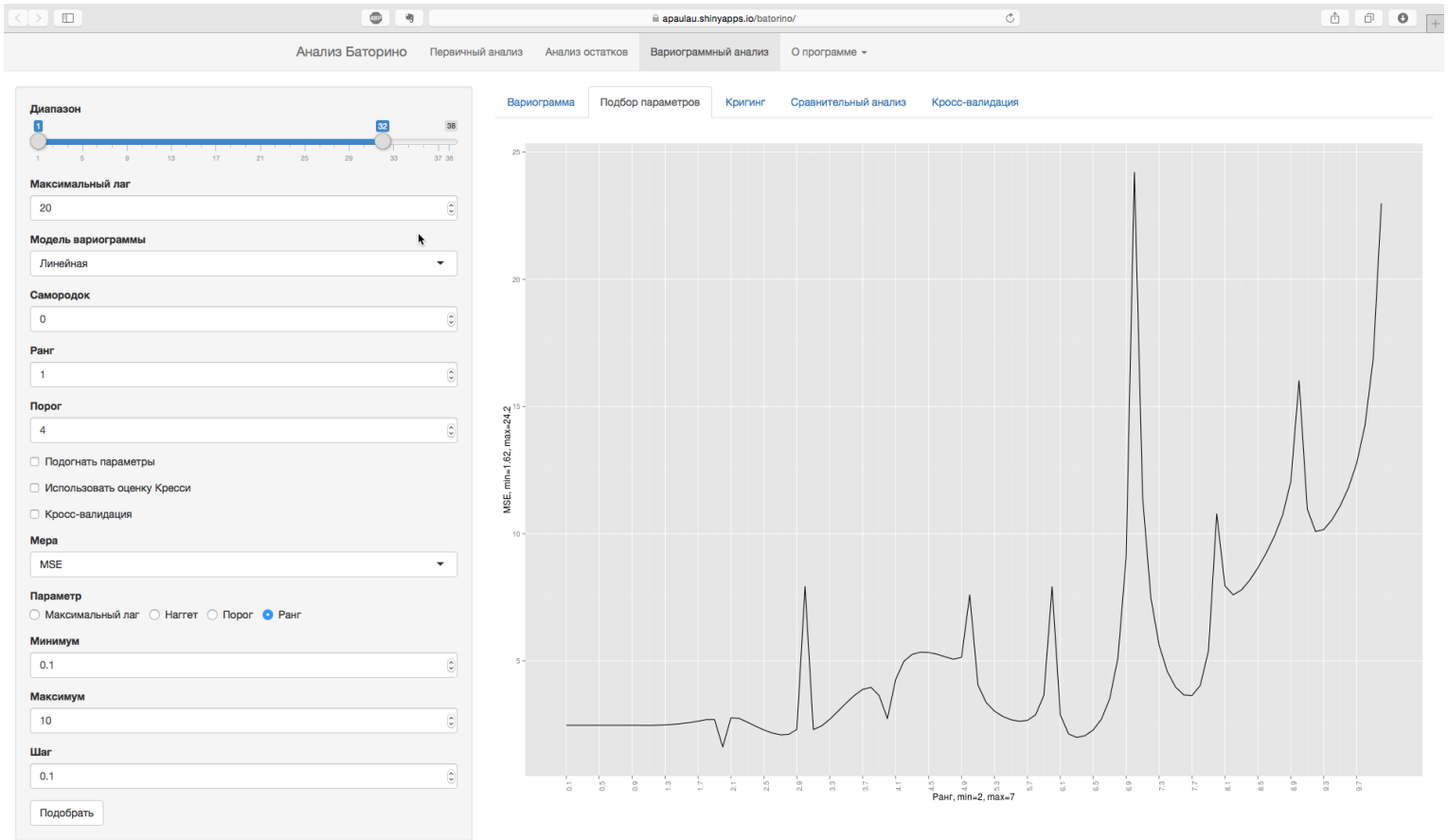
Модуль вариограммного анализа

Возможности по подбору модели вариограммы



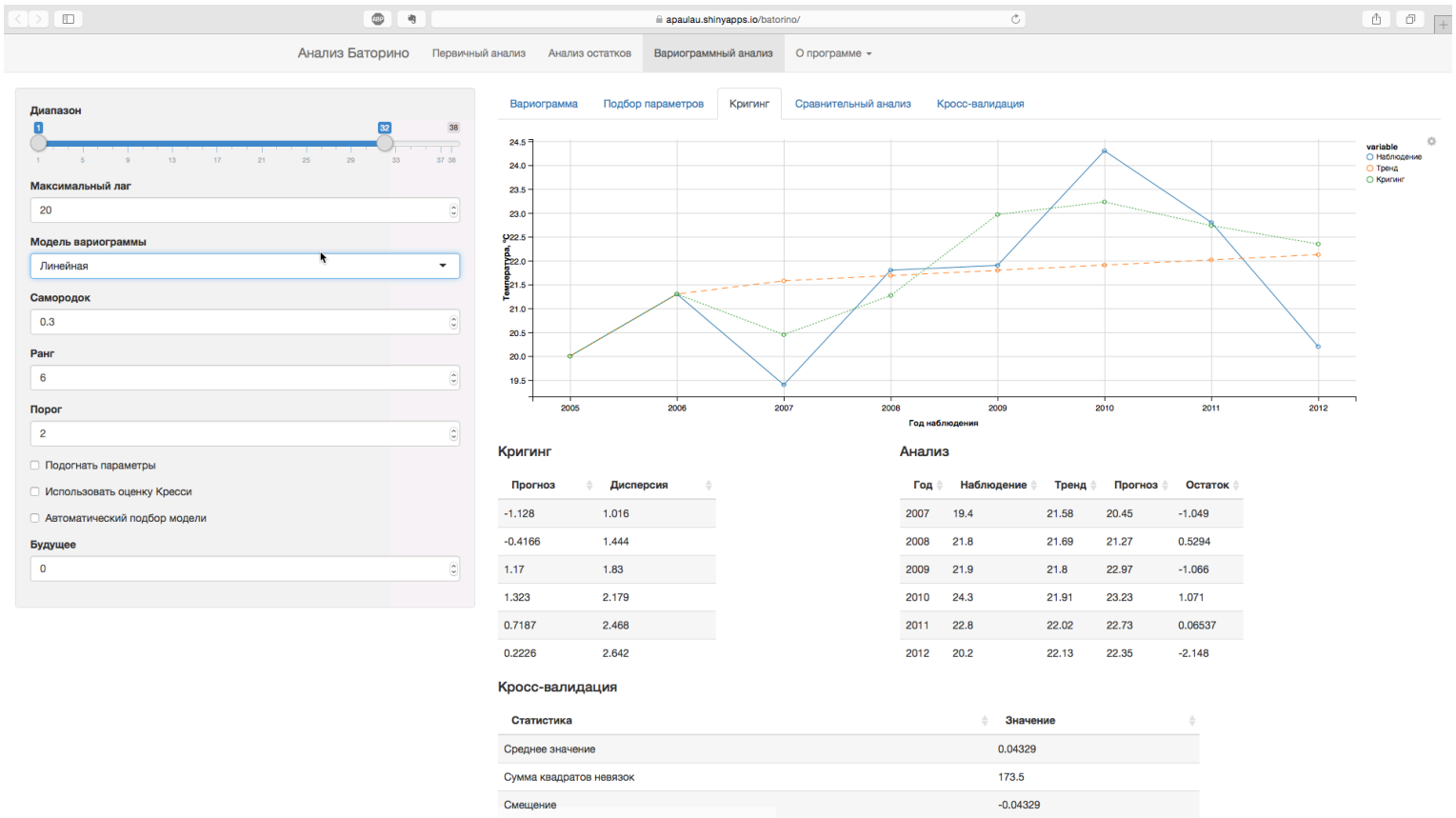
Модуль вариограммного анализа

Подбор параметров модели вариограммы



Модуль вариограммного анализа

Сравнение прогнозных значений



Детерминированный подход

Исходные данные

Данные получены от учебно-научного центра «Нарочанская биологическая станция им. Г.Г.Винберга».

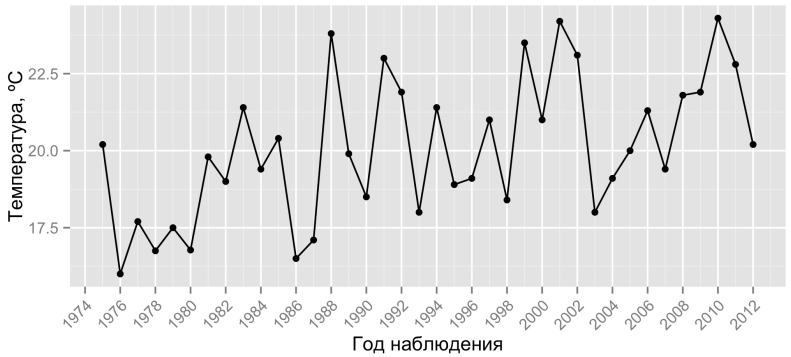


Рис.: Исходные данные

Исходные данные: выборка $X(t)$, $t = \overline{1, n}$, $n = 38$, $X(t)$ — значение средней температуры воды оз. Баторино в июле месяце для каждого года в период с 1975 по 2012 годы.

Детерминированный подход

Проверка на нормальность

- Коэффициент асимметрии $0.30 \Leftrightarrow$ распределение скошено вправо;
- Коэффициент эксцесса $-0.746 \Leftrightarrow$ пик кривой распределения пологий относительного нормального.

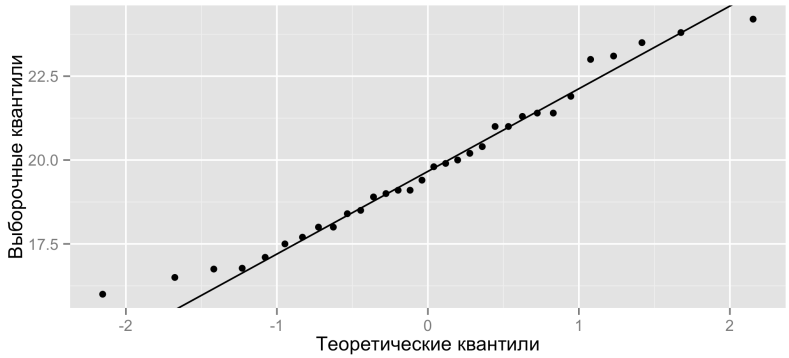


Рис.: График квантилей

Выборочное распределение близко к нормальному $\mathcal{N}(19.77, 5.12)$(визуально, критерии Шапиро-Уилка, χ^2-Пирсона и Колмогорова-Смирнова).

Детерминированный подход

Корреляционный анализ

- Выбросы в исходных данных отсутствуют (критерий Граббса);
- Выборочный коэффициент корреляции $r_{xt} = 0.454$ — при уровне значимости $\alpha = 0.05$ является значимым.

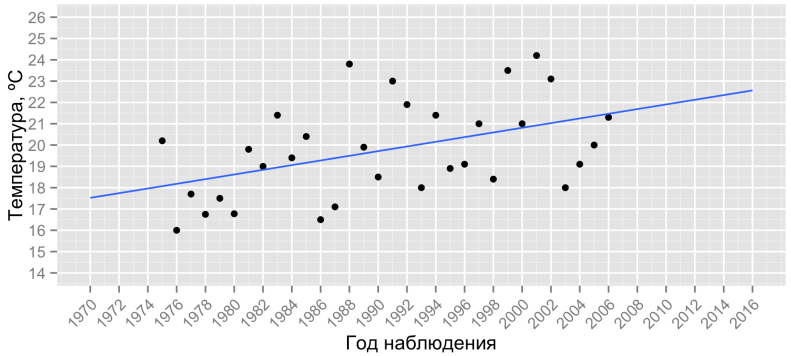


Рис.: Диаграмма рассеяния

Детерминированный подход

Регрессионный анализ: регрессионная модель

Исследуемый временной ряд является аддитивным,

$$X(t) = y(t) + \varepsilon(t); \qquad (1)$$

$y(t)$ — тренд, $\varepsilon(t)$ — нерегулярная составляющая.

Найденная модель тренда:
$y(t) = at + b =$
$0.1014t + 18.0521.$

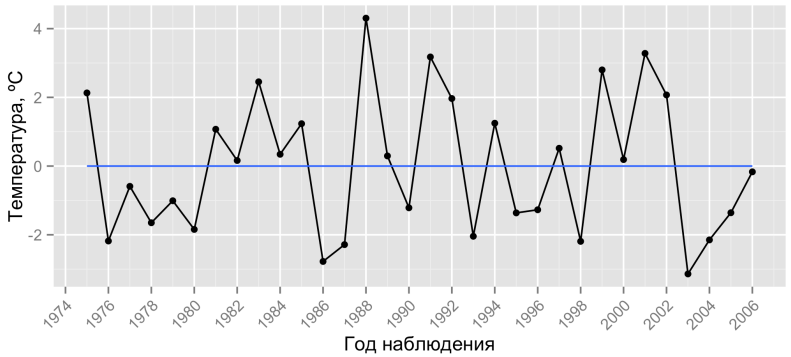


Рис.: Ряд остатков $\varepsilon(t)$

Детерминированный подход

Регрессионный анализ: качество регрессионной модели

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

Таблица: Сравнение прогнозных значений (модель $y(t)$)

- Коэффициенты регрессионной модели значимы (критерий Стьюдента, $\alpha = 0.05$);
- Модель адекватна (F-критерий Фишера, $\alpha = 0.05$);
- Точность модели невысока (поскольку коэффициент детерминации $\eta^2_{x(t)} = 0.275$).

Детерминированный подход

Анализ остатков

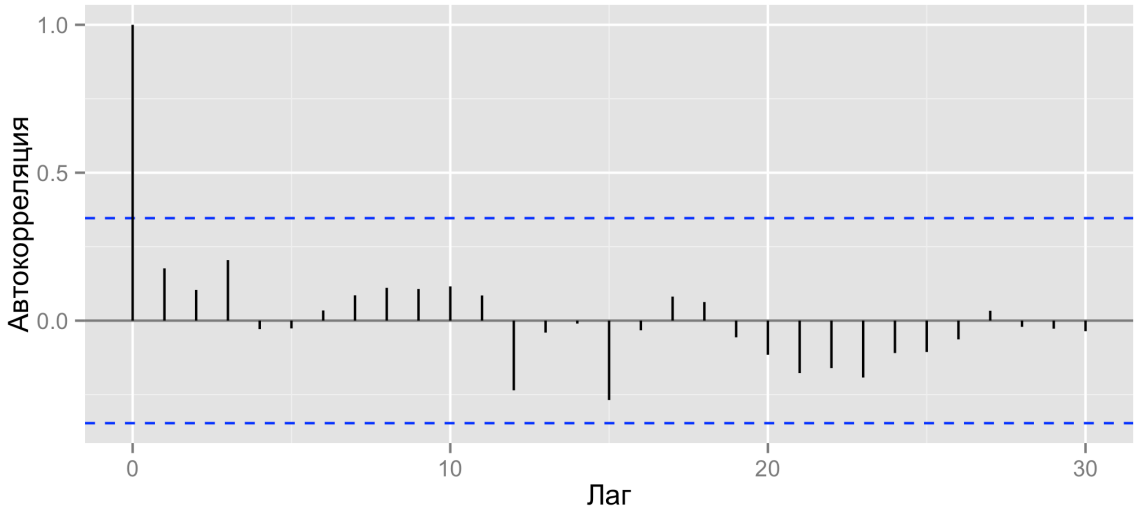


Рис.: Автокорреляционная функция

- Выборочное распределение близко к $\mathcal{N}(0.00, 4.07)$;
- Значимые автокорреляции отсутствуют;
- Значения имеют небольшую амплитуду и имеют тенденцию к затуханию $\Leftrightarrow$ ряд стационарен в широком смысле.

Геостатистический подход

Оценка вариограммы

Пусть $X(t),\ t \in \mathbb{Z}$ — стационарный в широком смысле гауссовский случайный процесс с дискретным временем, нулевым математическим ожиданием и постоянной дисперсией.

Определение 1

Вариограмма случайного процесса $X(t), t \in \mathbb{Z}$:

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

При этом функция $\gamma(h), h \in \mathbb{Z}$, называется *семивариограммой*.

Оценка вариограммы (Матерон):

$$2\tilde{\gamma}(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(X(t+h) - X(t))^2, \quad h = \overline{0, n-1}. \qquad (3)$$

Теорема 2

Для оценки $2\tilde{\gamma}(h)$ имеют место следующие соотношения:

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

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

$$+\gamma(t+h_1-s)-\gamma(t-s)-\gamma(t+h_1-s-h_2))^2,$$

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

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

Теорема 3

Если имеет место соотношение $\sum_{h=-\infty}^{+\infty} |\gamma(h)| < +\infty$, *то*

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

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

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

Следствие 4

Из теоремы 2 следует соотношение

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

Следствие 5

В силу показанной в теореме 1 несмещённости оценки и вышеприведённого следствия получаем, что оценка вариограммы $2\tilde{\gamma}(h)$ является состоятельной в среднеквадратическом смысле для вариограммы $\gamma(h), h \in \mathbb{Z}$.

Вариограммный анализ

График экспериментальной вариограммы

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

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

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

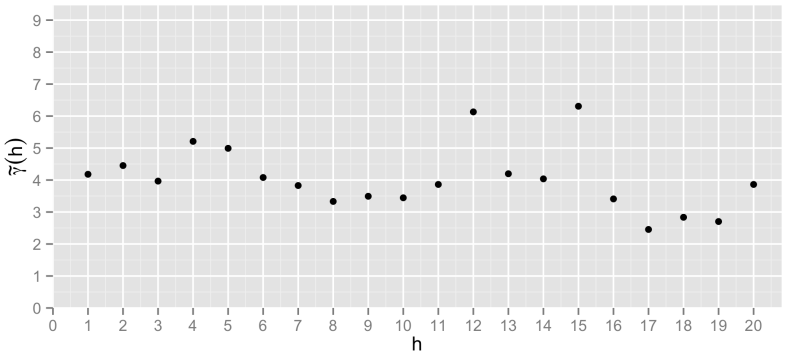


Рис.: Оценка семивариограммы Матерона

Для оценки качества модели используются

- коэффициент корреляции $r_{\varepsilon\varepsilon^*}$;
- Среднеквадратическая ошибка (n — объём выборки):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon(t_i) - \varepsilon^*(t_i))^2. \qquad (4)$$

Вариограммный анализ

Линейная модель

Общий вид модели:

$$\widehat{\gamma}(h) = c_0 + Lin(h) =$$

$$= \begin{cases} c_0 + b \cdot h, & h > 0, \\ c_0, & h \leq 0, \end{cases} \qquad (5)$$

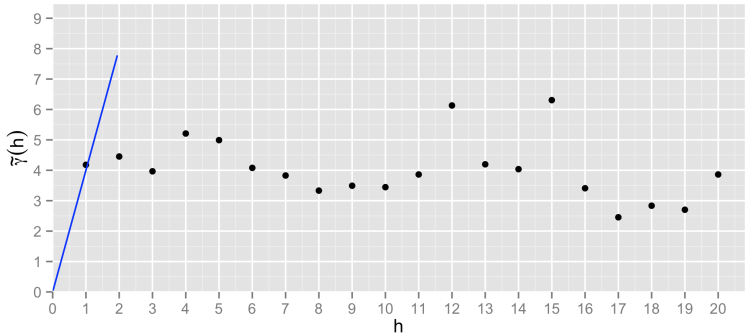
где b – параметр, отвечающий за угол наклона, c_0 — эффект самородков.

Подобранная модель:

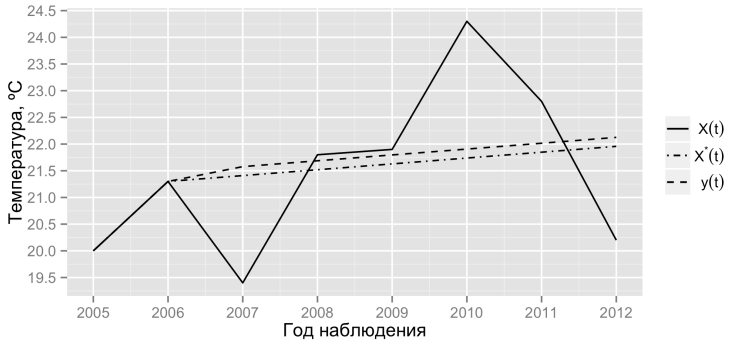
$$\widehat{\gamma}_1(h) = Lin(h), \quad b = 4 \qquad (6)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = -0.09129, \quad MSE = 6.324$$



Модель семивариограммы $\widehat{\gamma}_1(h)$



Прогноз по модели $\widehat{\gamma}_1(h)$

Вариограммный анализ

Чистый эффект самородков

Общий вид модели:

$$\widehat{\gamma}(h) = c \cdot Nug(h) = \\ = \begin{cases} 0, & h = 0, \\ c, & h \neq 0, \end{cases} \qquad (7)$$

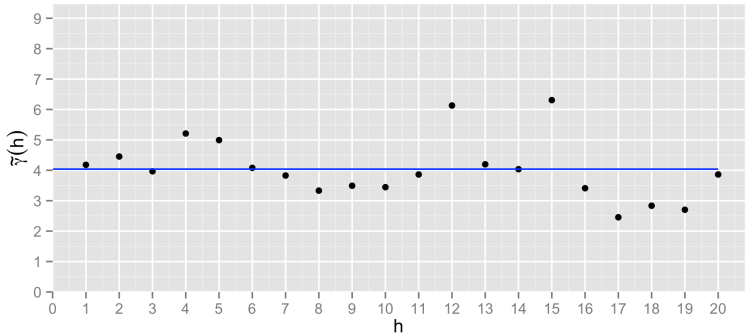
где b – параметр, отвечающий за угол наклона, c_0 — эффект самородков.

Подобранная модель:

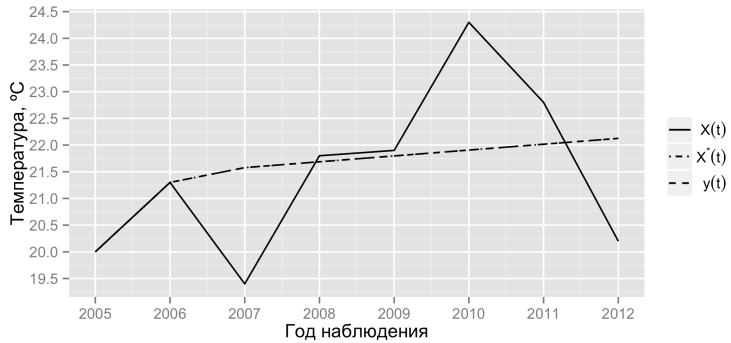
$$\widehat{\gamma}_2(h) = 4.04 \cdot Nug(h). \qquad (8)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = -1, \quad MSE = 4.199$$



Модель семивариограммы $\widehat{\gamma}_2(h)$



Прогноз по модели $\widehat{\gamma}_2(h)$

Вариограммный анализ

Линейная модель с порогом

Общий вид модели:

$$\widehat{\gamma}(h) = c_0 + c \cdot Lin(h, a) = \\ = \begin{cases} c_0 + c \cdot \frac{h}{a}, & 0 \le h \le a, \\ c_0 + c, & h > a, \end{cases} \qquad (9)$$

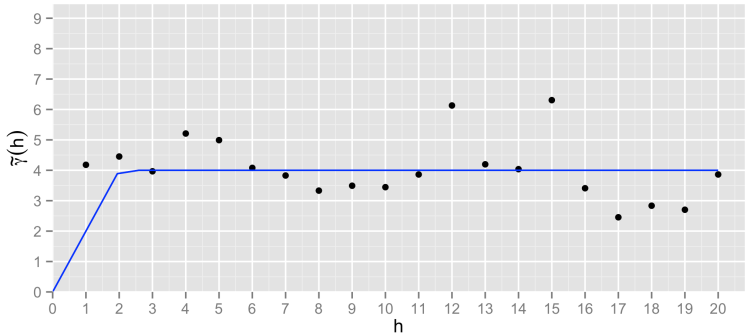
где c_0 – эффект самородков, c – порог, a – ранг.

Подобранная модель:

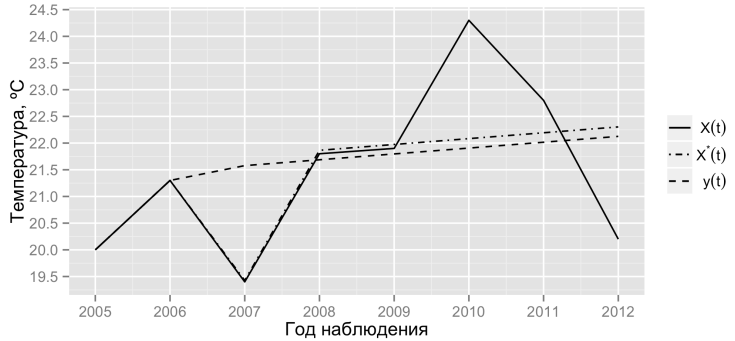
$$\widehat{\gamma}_4(h) = 4 \cdot Lin(h, 2). \qquad (10)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = 0.152, \quad MSE = 18.69$$



Модель семивариограммы $\widehat{\gamma}_4(h)$



Прогноз по модели $\widehat{\gamma}_4(h)$

Вариограммный анализ

Сферическая модель

Общий вид модели:

$$\widehat{\gamma}(h) = c_0 + c \cdot Sph(h, a) =$$

$$= \begin{cases} c_0 + c \cdot (\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3), & h \le a, \\ c_0 + c, & h \ge a, \end{cases} \quad (11)$$

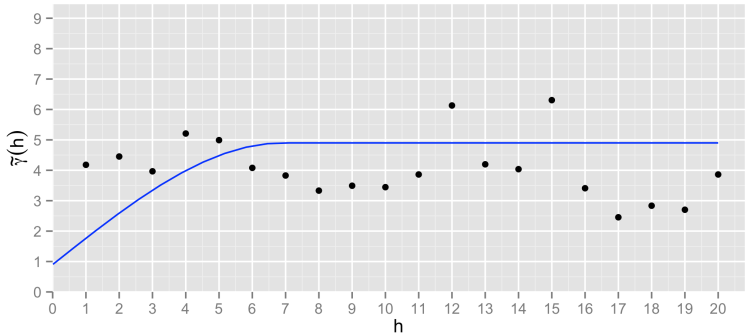
где c_0 – эффект самородков, c – порог, a – ранг.

Подобранная модель:

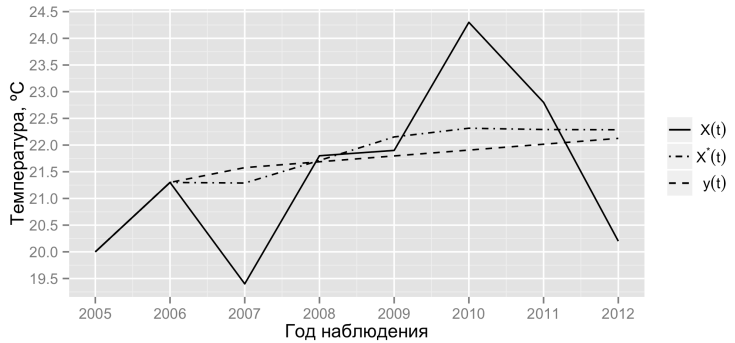
$$\widehat{\gamma}_5(h) = 0.9 + 4Sph(h, 6.9), \quad (12)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = -0.009, \quad MSE = 5.396$$



Модель семивариограммы $\widehat{\gamma}_5(h)$



Прогноз по модели $\widehat{\gamma}_5(h)$

Вариограммный анализ

Периодическая модель

Общий вид модели:

$$\widehat{\gamma}(h) = c_0 + c \cdot Per(h, a) = \quad (13)$$
$$= 1 - cos(\frac{2\pi h}{a}),$$

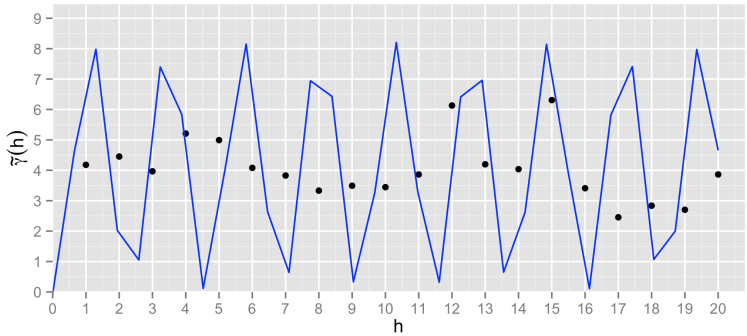
где c_0 – эффект самородков, c – порог, a – ранг.

Подобранная модель:

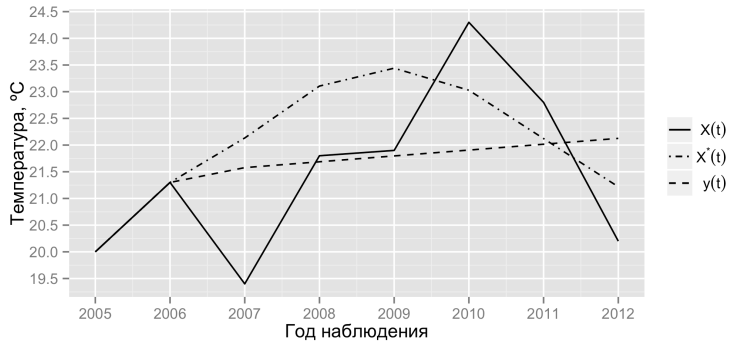
$$\widehat{\gamma}_6(h) = 4 \cdot Per(h, 0.898), \quad (14)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = 0.404, \quad MSE = 4.369$$



Модель семивариограммы $\widehat{\gamma}_6(h)$



Прогноз по модели $\widehat{\gamma}_6(h)$

Автоматический подход

Волновая модель

Общий вид модели:

$$\widehat{\gamma}(h) = c_0 + c \cdot Wav(h, a) = \quad (15)$$
$$= 1 - \frac{a}{h} \cdot sin(\frac{h}{a}),$$

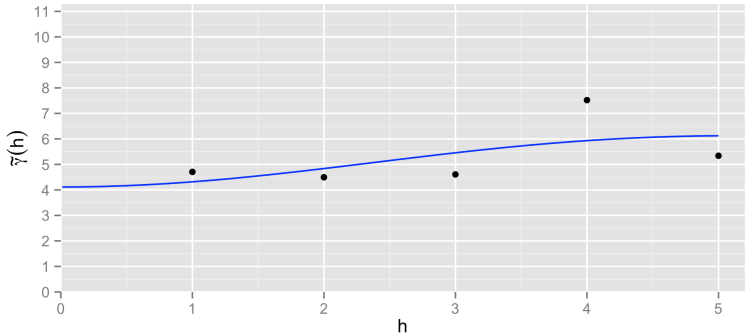
где c_0 – эффект самородков, c – порог, a – ранг.

Подобранная модель:

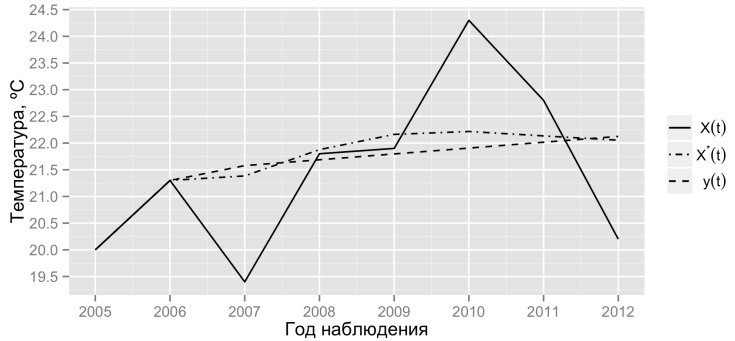
$$\widehat{\gamma}_9(h) = 4.11 + 1.65 \cdot Wav(h, 3.59), \quad (16)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = -1, \quad MSE = 4.20$$



Модель семивариограммы $\widehat{\gamma}_9(h)$



Прогноз по модели $\widehat{\gamma}_9(h)$

Автоматический подход

Периодическая модель

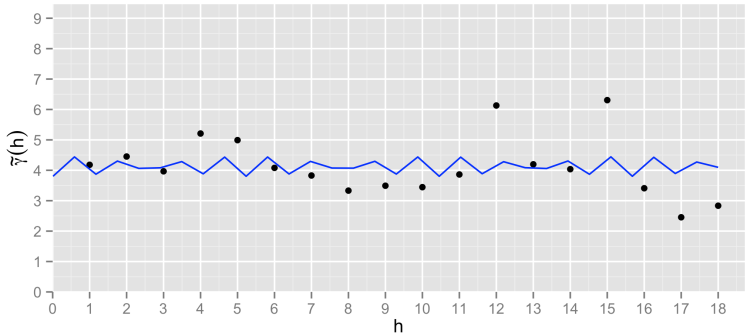
Модель семивариограммы вида (13).

Подобранная модель:

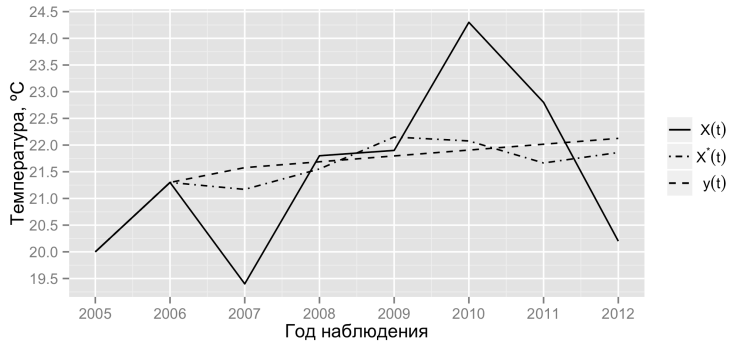
$$\widehat{\gamma}_{10}(h) = 3.8 + 0.32 \cdot Per(h, 1.3) \tag{17}$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = -0.15, \quad MSE = 5.22$$



Модель семивариограммы $\widehat{\gamma}_{10}(h)$



Прогноз по модели $\widehat{\gamma}_{10}(h)$

Заключение

- Проведён предварительный статистический анализ данных:
  - показана близость выборочного распределения к нормальному $\mathcal{N}(19.77, 5.12)$;
  - показана умеренная положительная зависимость температуры от времени;
  - построена регрессионная модель и вычислен ряд остатков;
- Выполнен вариограммный анализ:
  - рассмотрены два подхода по подбору моделей семивариограмм;
  - визуальным подходом построены наилучшие модели: линейная модель с порогом (10) и периодическая (14);
  - автоматическим подходом построены модели: волновая (16) и периодическая (17);

Заключение (продолжение)

- По различным моделям построены прогнозные значения методом Кригинг. Проанализирована зависимость точности прогноза от оценки вариограммы и модели;
- Исследованы статистические свойства оценки семивариограммы гауссовского случайного процесса. Показана несмещённость и состоятельность в среднеквадратическом смысле оценки вариограммы (3);
- Реализовано программное обеспечение, позволяющее решать класс задач, аналогичных исходной.

Список использованных источников


Cressie N.
Statistics for Spatial Data.
New York. — Wiley, 1991.


А.А. Савельев, С.С. Мухарамова, А.Г. Пилюгин, Н.А. Чижикова
Геостатистический анализ данных в экологии и природопользовании (с применением пакета R)
Казань: Казанский университет, 2012.


Robert H. Shumway, David S. Stoffer
Time series and Its Applications: With R Examples (Springer Texts in Statistics).
Springer Science+Business Media, LLC 2011, 3d edition, 2011.


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

Спасибо за внимание!

БЕЛАРУСКІ ДЗЯРЖАЎНЫ УНІВЕРСІТЭТ

Анализ и прогнозирование гидрологических данных

Александр Сергеевич Павлов

Научный руководитель: Цеховая Татьяна Вячеславовна

Факультет прикладной математики и информатики

Кафедра теории вероятностей и математической статистики

Минск, 2015