

Анализ и прогнозирование гидрологических данных

Павлов Александр Сергеевич

Научный руководитель: Цеховая Татьяна Вячеславовна

Кафедра Теории Вероятностей и Математической Статистики
Факультет Прикладной Математики и Информатики
Белорусский Государственный Университет

Минск, 2015

- 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 Кrigинг.
- ⑤ Исследование точности прогноза в зависимости от оценки вариограммы и модели семивариограммы, лежащих в основе метода Кrigинг.



① Обзор реализованного программного обеспечения

Модуль предварительного анализа

Модуль анализа остатков

Модуль вариограммного анализа

② Детерминированный подход

Проверка на нормальность

Корреляционный анализ

Регрессионный анализ

Анализ остатков

③ Геостатистический подход

Введение

Вариограммный анализ

Автоматический подход



- 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

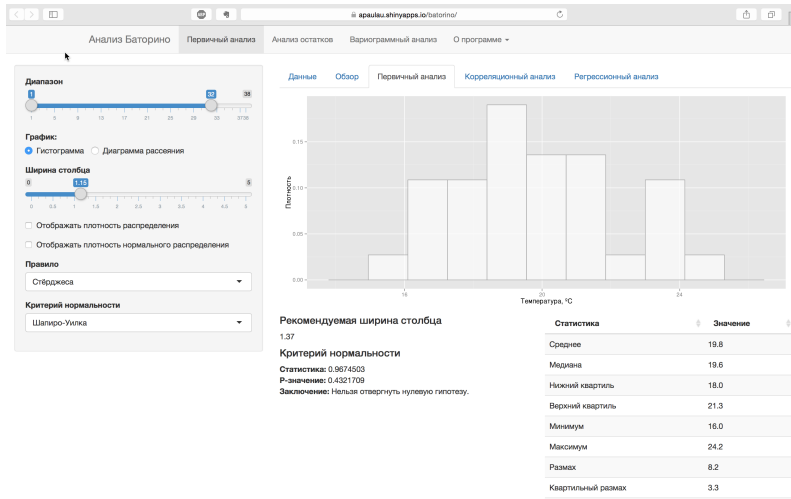


Рис. 1: Первичный анализ и описательные статистики



Модуль предварительного анализа

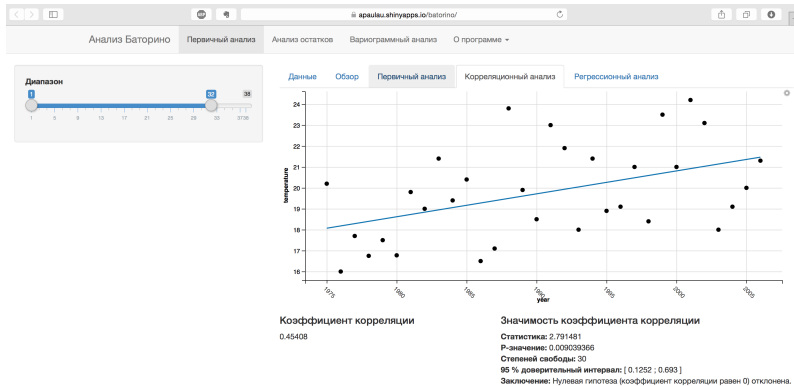


Рис. 2: Корреляционный анализ



Модуль предварительного анализа

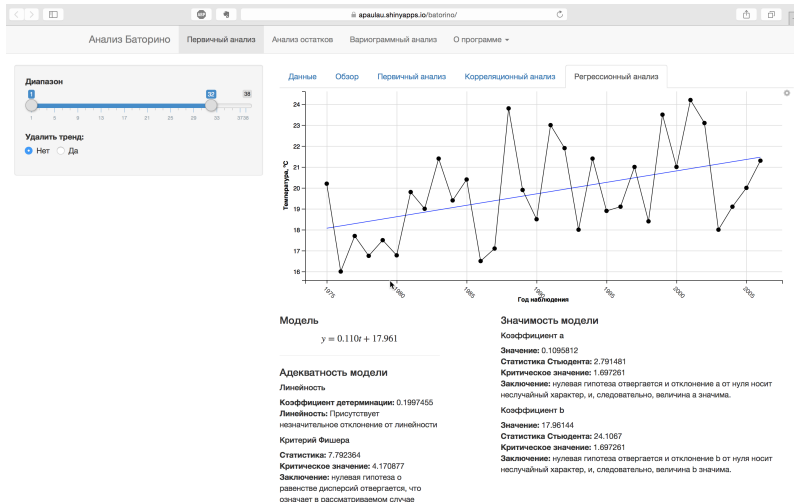


Рис. 3: Регрессионный анализ



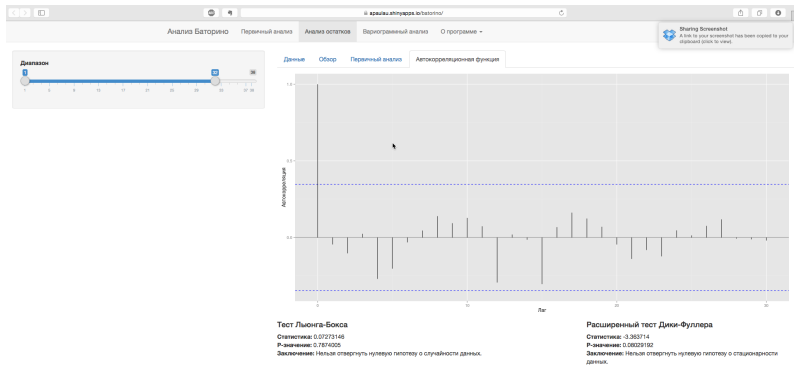


Рис. 4: Автокорреляционная функция



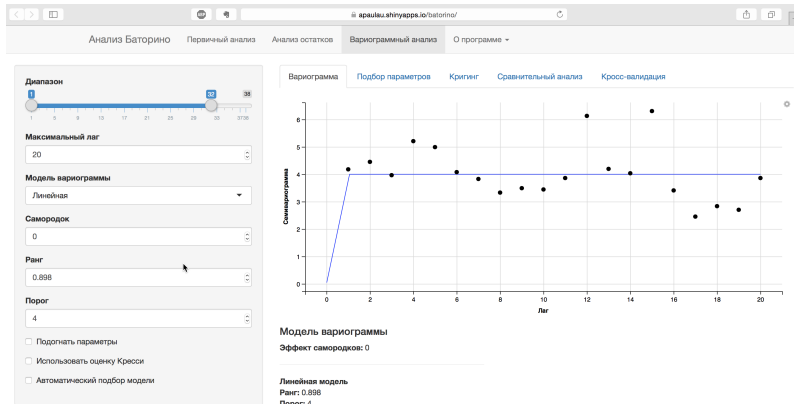


Рис. 5: Возможности по подбору модели вариограммы



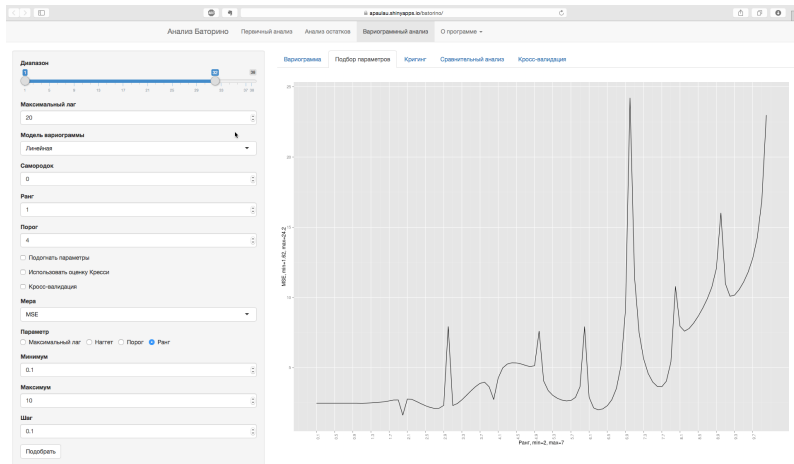


Рис. 6: Подбор параметров модели вариограммы



Модуль вариограммного анализа

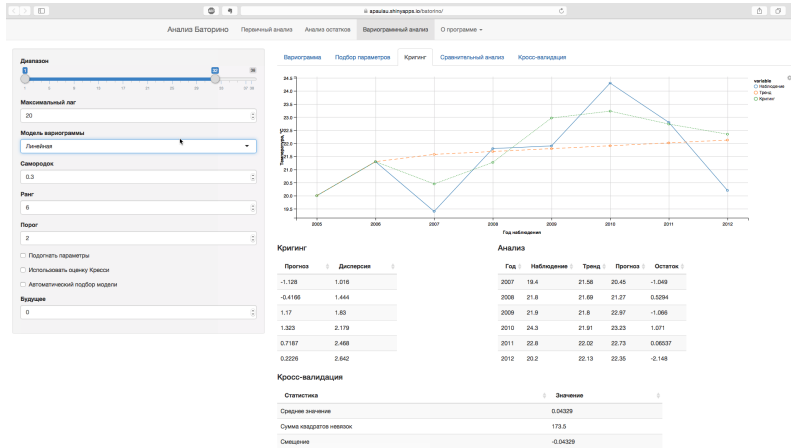


Рис. 7: Сравнение прогнозных значений



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

Исходные данные представляют собой выборку $X(t)$, $t = \overline{1, n}$, $n = 38$, состоящую из значений средней температуры воды в июле месяце каждый год в период с 1975 по 2012 годы.

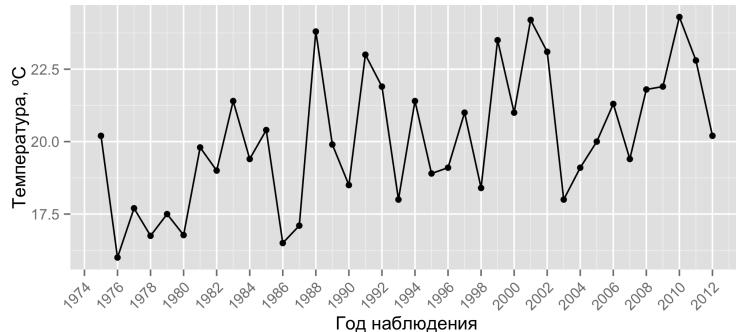


Рис. 8: Исходные данные



По описательным статистикам выборочное распределение характеризуется небольшой скошенностью вправо (коэффициент асимметрии 0.30) и пологостью пика кривой распределения (коэффициент эксцесса -0.746) относительного нормального.

Визуально и проверкой критериев Шапиро-Уилка, χ^2 -Пирсона и Колмогорова-Смирнова была показана близость выборочного распределения к нормальному с параметрами $\mathcal{N}(19.77, 5.12)$.

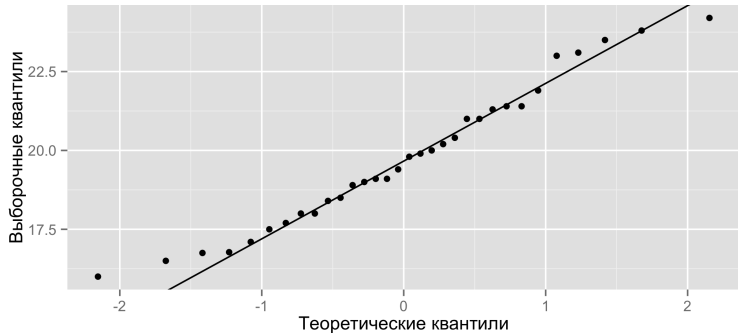


Рис. 9: График квантилей



С помощью критерия Граббса показано отсутствие выбросов в исходных данных.

Вычислен выборочный коэффициент корреляции:
 $r_{xt} = 0.454$.

При уровне значимости $\alpha = 0.05$ доказана его значимость.

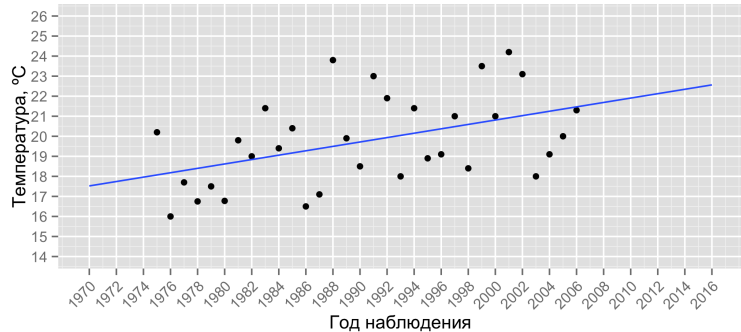


Рис. 10: Диаграмма рассеяния



Выявлено, что исследуемый временной ряд является аддитивным:

$$X(t) = y(t) + \varepsilon(t), \quad (1)$$

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

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

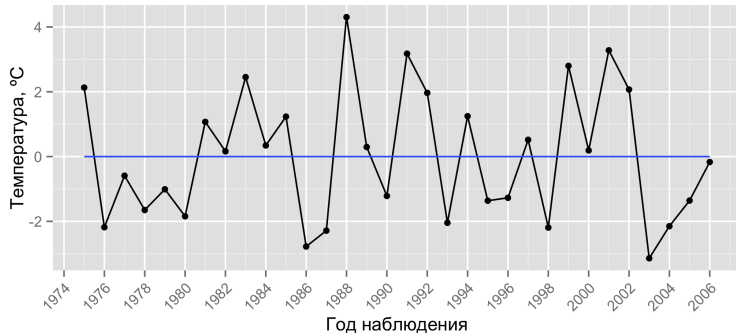


Рис. 11: Ряд остатков $\varepsilon(t)$



- С помощью критерия Стьюдента, при уровне значимости $\alpha = 0.05$, доказана значимость коэффициентов регрессионной модели
- F-критерий Фишера при уровне значимости $\alpha = 0.05$ показал адекватность модели
- Точность модели невысока, поскольку коэффициент детерминации $\eta_{x(t)}^2 = 0.275$

	$X(t)$	$y(t)$	$X(t) - y(t)$
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 1: Сравнение прогнозных значений (модель $y(t)$)



Визуально и проверкой тестов показана близость выборочного распределения к нормальному $\mathcal{N}(0.00, 4.07)$.

По графику и тестом Льюнга-Бокса сделано заключение об отсутствии значимых автокорреляций.

Также было отмечено, что значения имеют небольшую амплитуду и имеют тенденцию к затуханию. Это говорит о стационарности в широком смысле, что показал расширенный тест Дики-Фуллера.

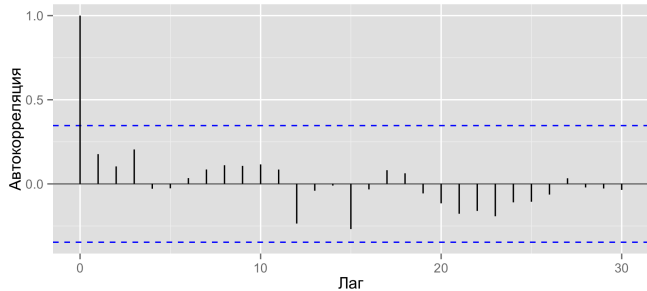


Рис. 12: Автокорреляционная функция



Рассматривается стационарный в широком смысле гауссовский случайный процесс с дискретным временем $X(t)$, $t \in \mathbb{Z}$, нулевым математическим ожиданием, постоянной дисперсией и неизвестной вариограммой $2\gamma(h)$, $h \in \mathbb{Z}$.

Определение 1

Вариограммой случайного процесса $X(t)$, $t \in \mathbb{Z}$, называется функция вида

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

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

В качестве оценки вариограммы рассматривается статистика, предложенная Матероном:

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



Теорема 1

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

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

$$\text{cov}(2\tilde{\gamma}(h_1), 2\tilde{\gamma}(h_2)) =$$

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

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

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



Теорема 2

Если имеет место соотношение

$$\sum_{h=-\infty}^{+\infty} |\gamma(h)| < +\infty, \text{ то}$$

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

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

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



Следствие 1

Из теоремы 2 следует соотношение

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

Следствие 2

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 $2\gamma(h)$, $h \in \mathbb{Z}$.



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

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

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

Для оценки качества модели
используются

- коэффициент корреляции $r_{\varepsilon\varepsilon^*}$
- Среднеквадратическая ошибка

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

где n — объём выборки

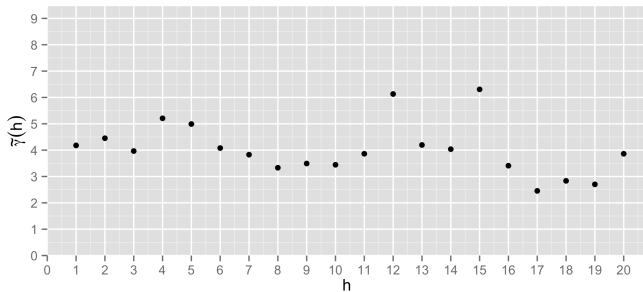


Рис. 13: Оценка семивариограммы Матерона



$$\hat{\gamma}(h) = c_0 + \text{Lin}(h) = \begin{cases} c_0 + b \cdot h, & h > 0, \\ c_0, & h \leq 0, \end{cases} \quad (5)$$

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

Подобранная модель:

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

Показатели качества

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

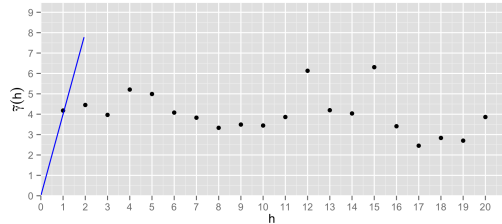


Рис. 14: Модель семивариограммы $\hat{\gamma}_1(h)$

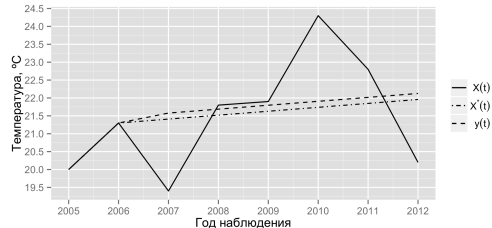


Рис. 15: Прогноз по модели $\hat{\gamma}_1(h)$



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

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

Подобранная модель:

$$\hat{\gamma}_2(h) = 4.04 \cdot \text{Nug}(h). \quad (8)$$

Показатели качества

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

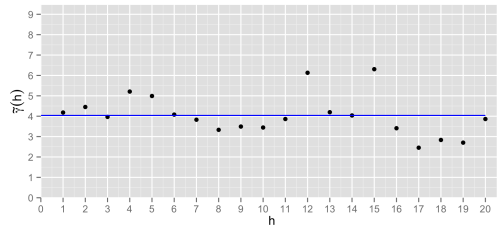


Рис. 16: Модель семивариограммы $\hat{\gamma}_1(h)$

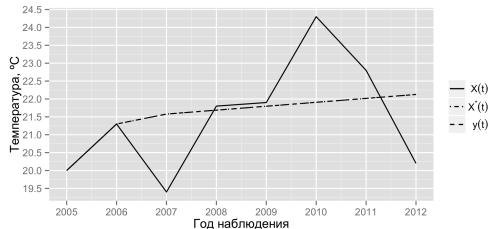


Рис. 17: Прогноз по модели $\hat{\gamma}_1(h)$



$$\begin{aligned}\hat{\gamma}(h) &= c_0 + c \cdot \text{Lin}(h, a) = \\ &= \begin{cases} c_0 + c \cdot \frac{h}{a}, & 0 \leq h \leq a, \\ c_0 + c, & h > a, \end{cases} \quad (9)\end{aligned}$$

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

Подобранная модель:

$$\hat{\gamma}_4(h) = 4 \cdot \text{Lin}(h, 2). \quad (10)$$

Показатели качества

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

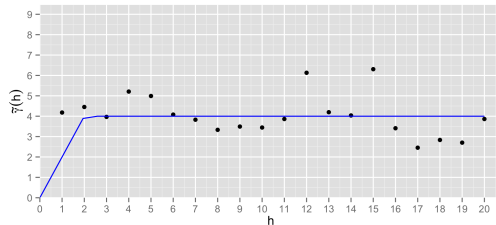


Рис. 18: Модель семивариограммы $\hat{\gamma}_4(h)$

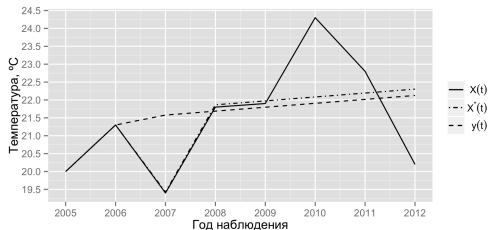


Рис. 19: Прогноз по модели $\hat{\gamma}_4(h)$



$$\hat{\gamma}(h) = c_0 + c \cdot Sph(h, a) = \begin{cases} c_0 + c \cdot \left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right), & h \leq a, \\ c_0 + c, & h \geq a, \end{cases} \quad (11)$$

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

Подобранная модель:

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

Показатели качества

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

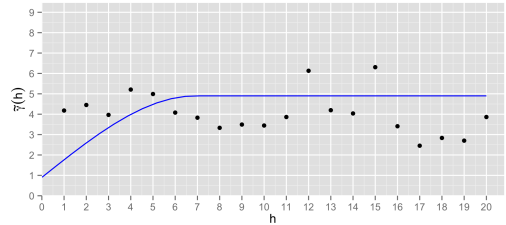


Рис. 20: Модель семивариограммы $\hat{\gamma}_5(h)$

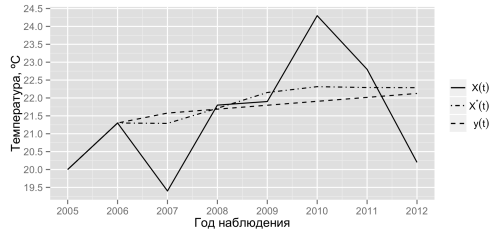


Рис. 21: Прогноз по модели $\hat{\gamma}_5(h)$



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

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

Подобранная модель:

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

Показатели качества

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

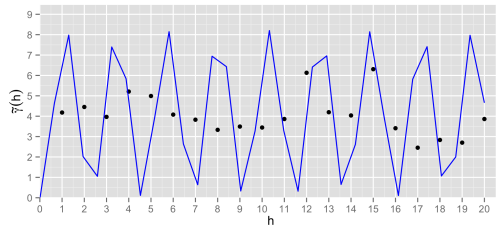


Рис. 22: Модель семивариограммы $\hat{\gamma}_6(h)$

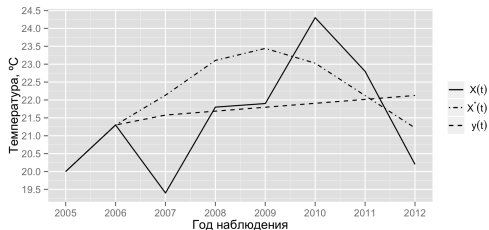


Рис. 23: Прогноз по модели $\hat{\gamma}_6(h)$



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

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

Подобранная модель:

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

Показатели качества

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

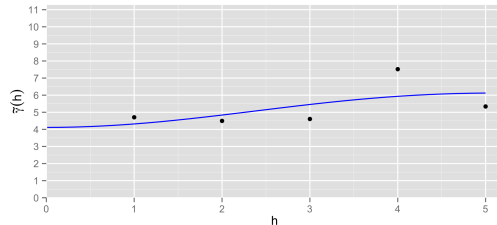


Рис. 24: Модель семивариограммы $\hat{\gamma}_9(h)$

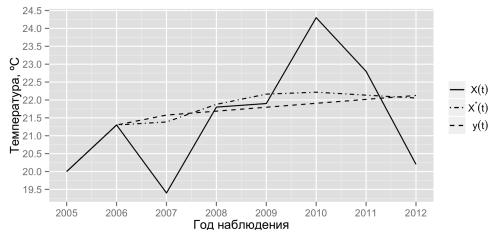


Рис. 25: Прогноз по модели $\hat{\gamma}_9(h)$



Модель семивариограммы вида (13).

Подобранная модель:

$$\hat{\gamma}_{10}(h) = 3.8 + 0.32 \cdot \text{Per}(h, 1.3) \quad (17)$$

Показатели качества

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

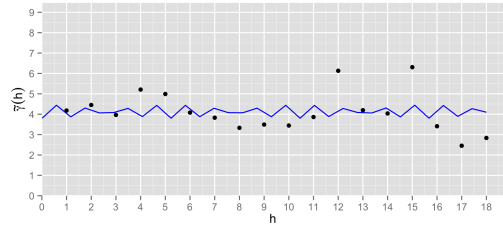


Рис. 26: Модель семивариограммы $\hat{\gamma}_{10}(h)$

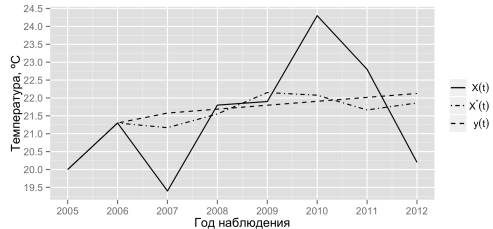


Рис. 27: Прогноз по модели $\hat{\gamma}_{10}(h)$



- ① 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
- ② 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).
- ③ По различным моделям построены прогнозные значения методом Кригинг.
Проанализирована зависимость точности прогноза от оценки вариограммы и модели.
- ④ Исследованы статистические свойства оценки семивариограммы гауссовского случайного процесса. Показана несмещённость и состоятельность в среднеквадратическом смысле оценки вариограммы (2)
- ⑤ Реализовано программное обеспечение, позволяющее решать класс задач, аналогичных исходной





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!

