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

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 вариограммы гауссовского случайного процесса.
4. Прогнозирование значений временного ряда с помощью интерполяционного метода кригинг.
5. Исследование точности прогноза в зависимости от оценки вариограммы и модели вариограммы, лежащих в основе метода кригинг.

Содержание

Исходные данные

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

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.

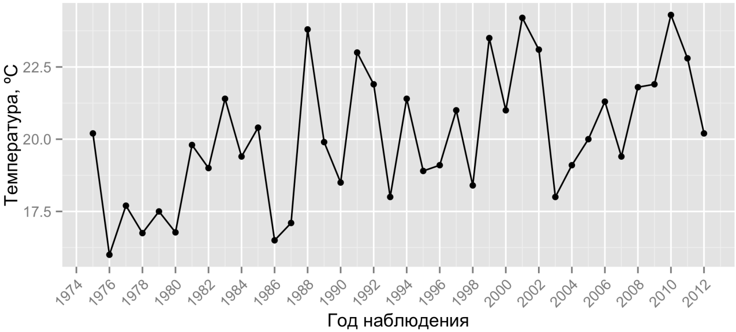

Рис. 1: Исходные данные

- 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

Обзор реализованного программного обеспечения

Модуль предварительного анализа

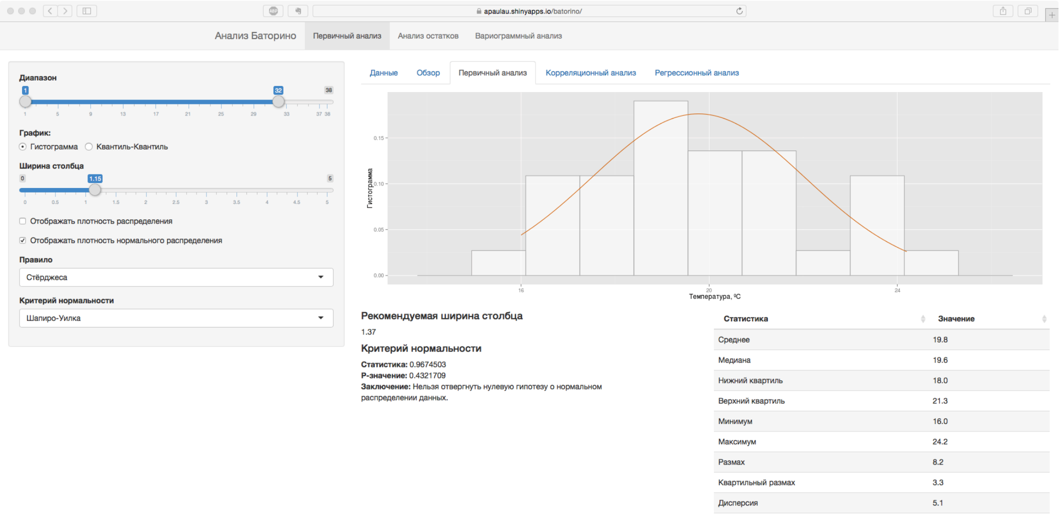

Рис. 2: Первичный анализ и описательные статистики

Проверка на нормальность

Выборочное распределение характеризуется небольшой скошенностью вправо (коэффициент асимметрии 0.30) и пологостью пика кривой распределения (коэффициент эксцесса -0.746) относительно нормального.

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

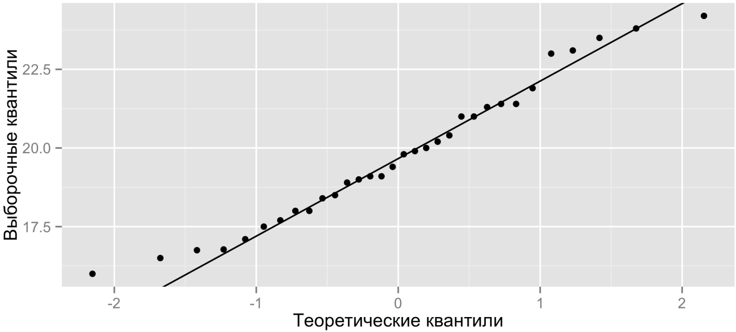

Рис. 3: График квантилей

Обзор реализованного программного обеспечения
Модуль предварительного анализа

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

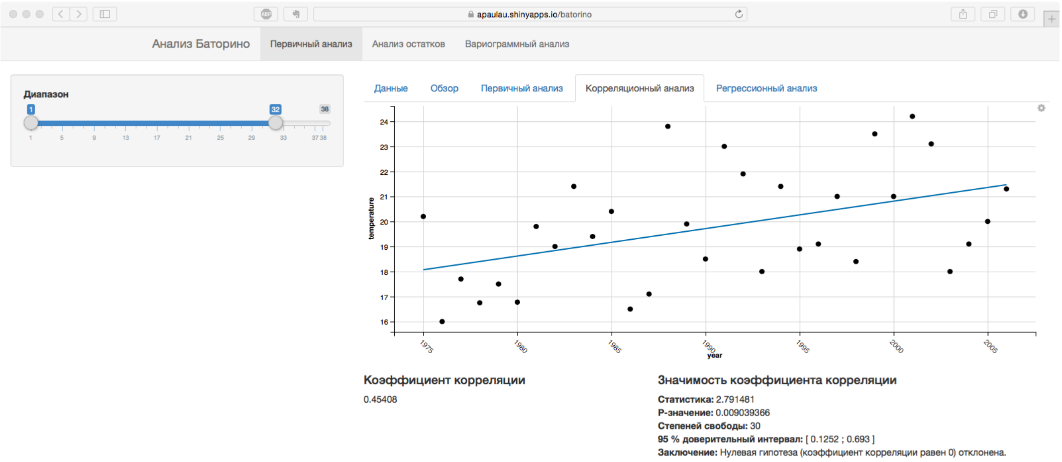

Рис. 4: Корреляционный анализ

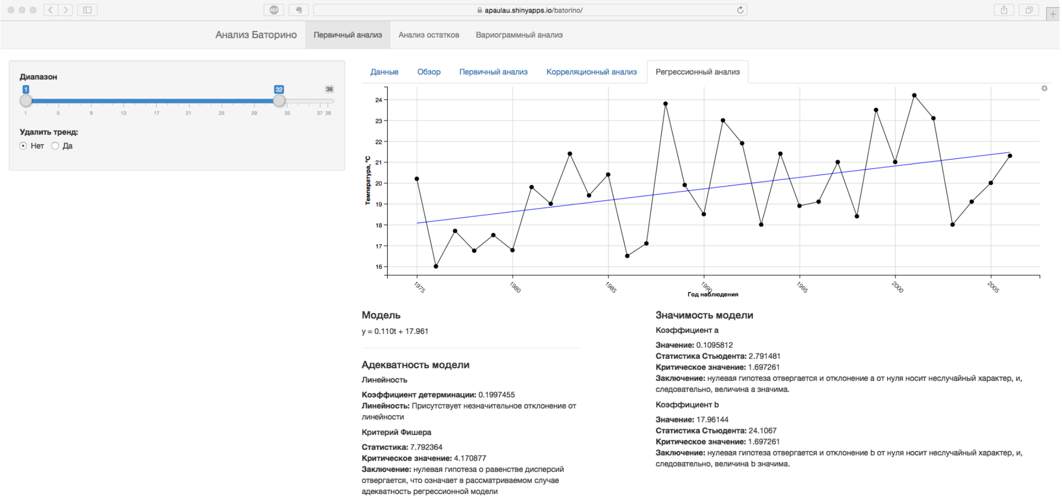

Рис. 5: Регрессионный анализ

Регрессионная модель

Выявлено, что исследуемый временной ряд является аддитивным:

$$X(t) = y(t) + \varepsilon(t), \tag{1}$$

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

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

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

Таблица 1: Сравнение прогнозных значений (модель $y(t)$)

	$X(t)$	$y(t)$	$X(t) - y(t)$
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

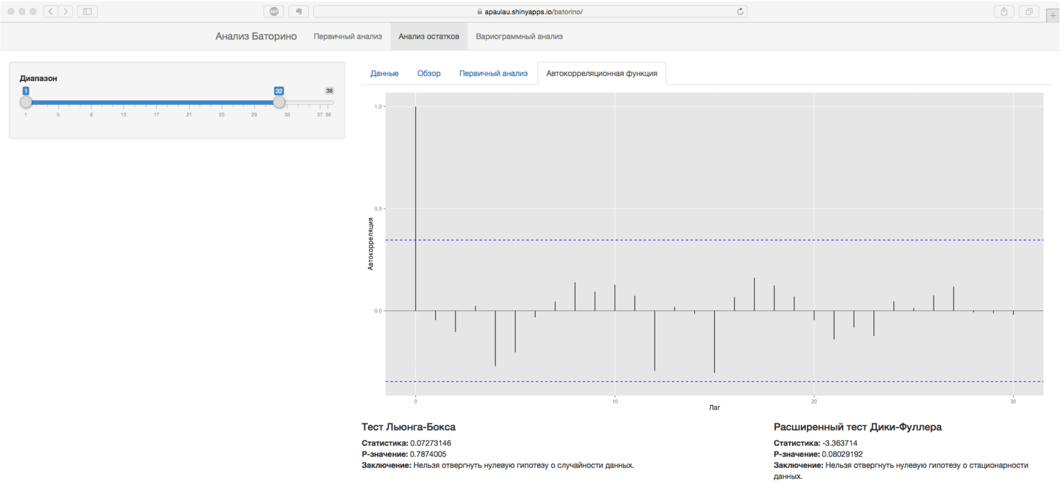

Рис. 6: Автокорреляционная функция

Анализ остатков

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.

Визуальный подход

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

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

где n — объём выборки

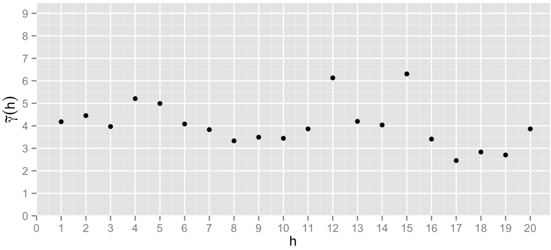

Рис. 7: Оценка семивариограммы Матерона

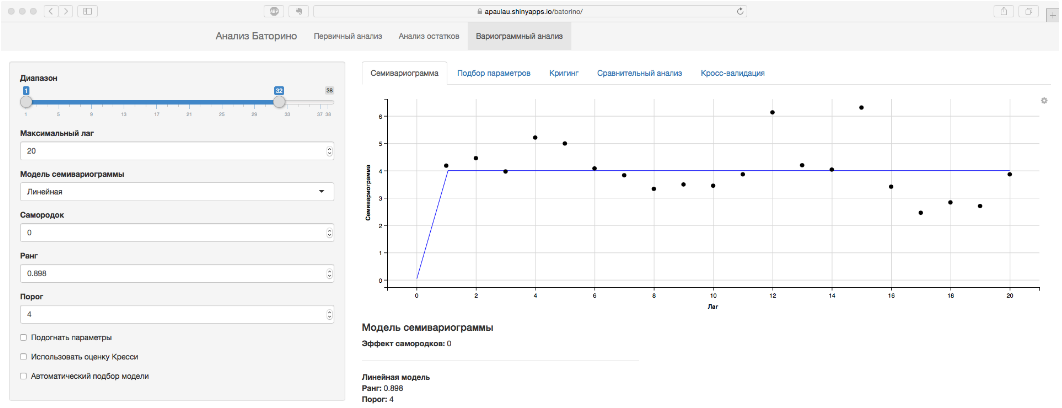

Рис. 8: Возможности по подбору модели семивариограммы

Обзор реализованного программного обеспечения

Модуль вариограммного анализа

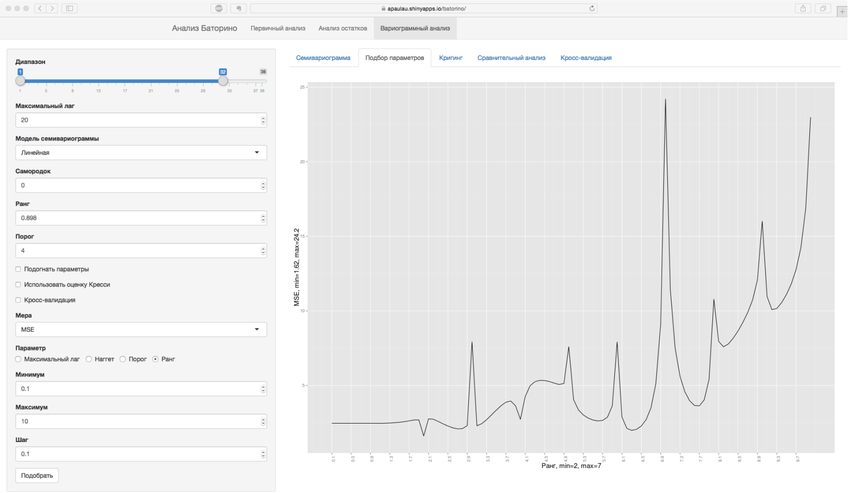

Рис. 9: Подбор параметров модели семивариограммы

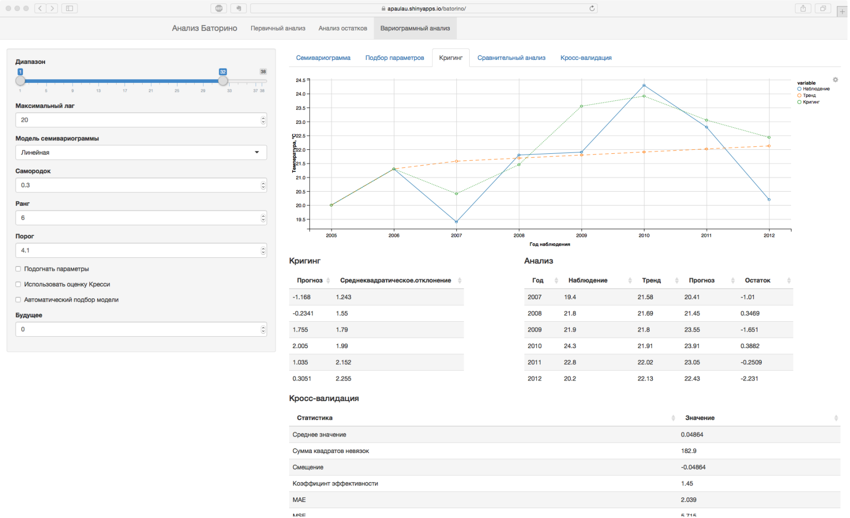

Рис. 10: Сравнение прогнозных значений

$$\widehat{\gamma}(h) = c_0 + Lin(h) = \begin{cases} c_0 + b \cdot h, & h > 0, \\ c_0, & h \leq 0, \end{cases} \qquad (3)$$

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

Подобранная модель:

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

Показатели качества

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

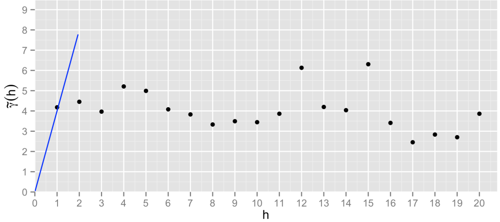

Рис. 11: Модель семивариограммы $\widehat{\gamma}_1(h)$

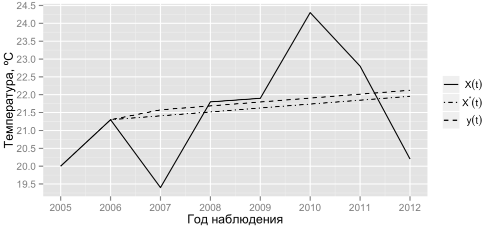

Рис. 12: Прогноз по модели $\widehat{\gamma}_1(h)$

$$\widehat{\gamma}(h) = c_0 + c \cdot Lin(h, a) = \tag{5}$$
$$= \begin{cases} c_0 + c \cdot \frac{h}{a}, & 0 \le h \le a, \\ c_0 + c, & h > a, \end{cases}$$

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

Подобранная модель:

$$\widehat{\gamma}_3(h) = 4 \cdot Lin(h, 2). \tag{6}$$

Показатели качества

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

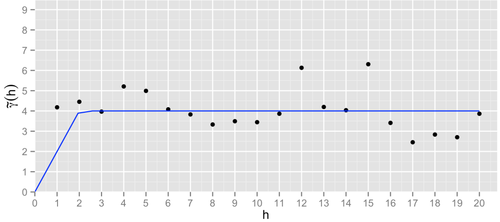

Рис. 13: Модель семивариограммы $\widehat{\gamma}_3(h)$

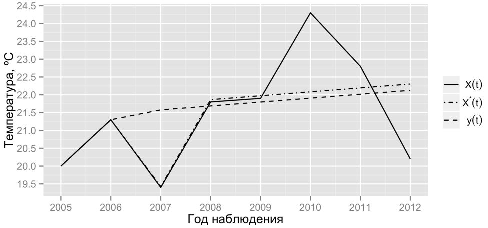

Рис. 14: Прогноз по модели $\widehat{\gamma}_3(h)$

$$\widehat{\gamma}(h) = c_0 + c \cdot Sph(h, a) =$$
$$= \begin{cases} c_0 + c \cdot (\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3), & h \leq a, \\ c_0 + c, & h \geq a, \end{cases} \quad (7)$$

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

Подобранная модель:

$$\widehat{\gamma}_4(h) = 0.9 + 4Sph(h, 6.9), \quad (8)$$

Показатели качества

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

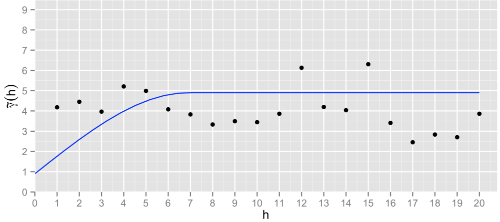

Рис. 15: Модель семивариограммы $\widehat{\gamma}_4(h)$

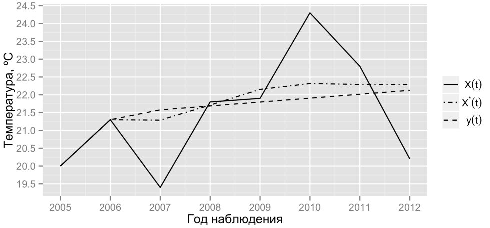

Рис. 16: Прогноз по модели $\widehat{\gamma}_4(h)$

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

$$\widehat{\gamma}(h) = c_0 + c \cdot Per(h, a) = 1 - cos(\frac{2\pi h}{a}), \qquad (9)$$

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

Подобранная модель:

$$\widehat{\gamma}_5(h) = 4 \cdot Per(h, 0.898), \qquad (10)$$

Показатели качества

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

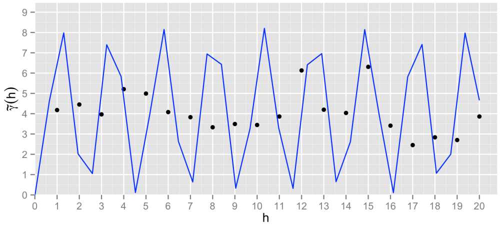

Рис. 17: Модель семивариограммы $\widehat{\gamma}_5(h)$

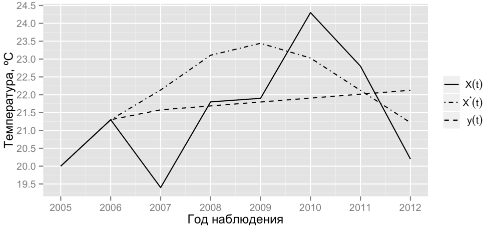

Рис. 18: Прогноз по модели $\widehat{\gamma}_5(h)$

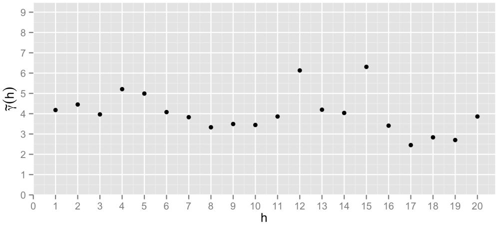

Рис. 19: Оценка Матерона

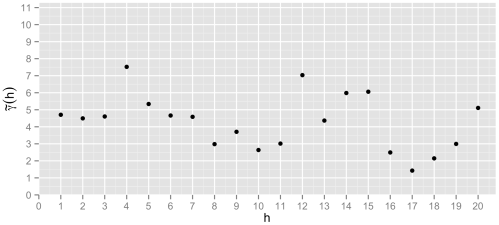

Рис. 20: Оценка Кресси-Хоккинса

Робастная оценка семивариограммы Кресси-Хокинса:

$$2\tilde{\gamma}(h) = \frac{1}{n-h}\Big(\sum_{t=1}^{n-h} |X(t+h) - X(t)|^{\frac{1}{2}}\Big)^4 / \Big(0.457 + \frac{0.494}{n-h} + \frac{0.045}{(n-h)^2}\Big), \quad h = \overline{0, n-1}. \quad (11)$$

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

$$\widehat{\gamma}(h) = c_0 + c \cdot Wav(h, a) = 1 - \frac{a}{h} \cdot sin(\frac{h}{a}), \quad (12)$$

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

Подобранная модель:

$$\widehat{\gamma}_6(h) = 4.11 + 1.65 \cdot Wav(h, 3.59), \quad (13)$$

Показатели качества

$$r_{\varepsilon\varepsilon^*} = -0.03, \quad MSE = 4.20$$

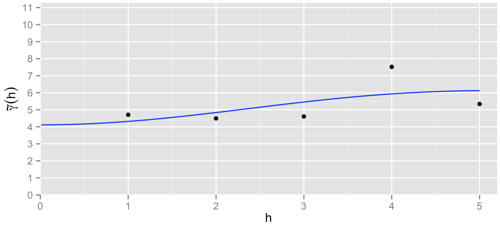

Рис. 21: Модель семивариограммы $\widehat{\gamma}_6(h)$

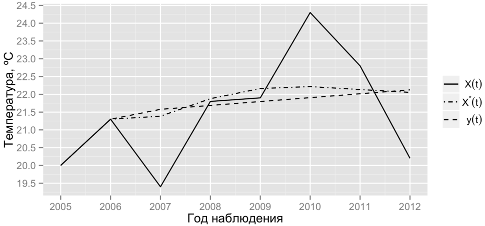

Рис. 22: Прогноз по модели $\widehat{\gamma}_6(h)$

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

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

Подобранная модель:

$$\widehat{\gamma}_7(h) = 3.8 + 0.32 \cdot Per(h, 1.3) \qquad (14)$$

Показатели качества

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

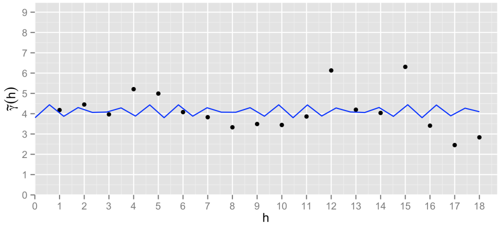

Рис. 23: Модель семивариограммы $\widehat{\gamma}_7(h)$

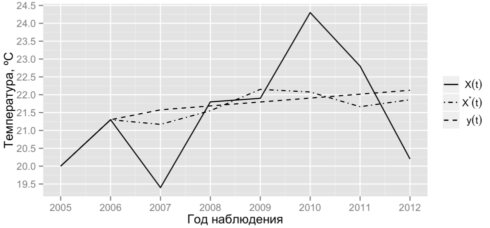

Рис. 24: Прогноз по модели $\widehat{\gamma}_7(h)$

Оценка вариограммы

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

$$2\gamma(h) = V\{X(t+h) - X(t)\},\ t, h \in \mathbb{Z}. \tag{15}$$

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

В качестве оценки вариограммы рассматривается статистика, предложенная Матероном:

$$2\tilde{\gamma}(h) = \frac{1}{n-h} \sum_{t=1}^{n-h} (X(t+h) - X(t))^2, \quad h = \overline{0, n-1}, \tag{16}$$

Первые два момента оценки вариограммы

Теорема 1

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

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Теорема 2

Если имеет место соотношение

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

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

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

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

Асимптотическое поведение оценки вариограммы

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Следствие 1

Из теоремы 2 следует соотношение

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

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

Заключение

1. Проведён предварительный статистический анализ данных
   - показана близость выборочного распределения к нормальному $\mathcal{N}(19.77, 5.12)$
   - выявлена умеренная положительная зависимость температуры от времени
   - построена линейная регрессионная модель
   - вычислен и исследован ряд остатков
2. Выполнен вариограммный анализ
   - Рассмотрены два подхода по подбору моделей семивариограмм: визуальный и автоматический
   - Визуальным подходом показано, что линейная модель с порогом (6) и периодическая (10) являются наилучшими.
   - Автоматическим подходом показано, что волновая (13) и периодическая (14) являются наилучшими.
3. По различным моделям построены прогнозные значения методом кригинг. Исследована зависимость точности прогноза от оценки вариограммы и модели.
4. Исследованы статистические свойства оценки вариограммы гауссовского случайного процесса. Показана несмещённость и состоятельность в среднеквадратическом смысле оценки вариограммы (15)
5. Реализовано программное обеспечение для решения класса задач, аналогичных исходной

Список использованных источников

Cressie N.
Statistics for Spatial Data.
New York. — Wiley, 1993.

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.

Н.Н. Труш
Асимптотические методы статистического анализа временных рядов.
Белгосуниверситет, 1999.

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.

Rudolf Dutter
On Robust Estimation of Variograms in Geostatistics
Robust Statistics, Data Analysis, and Computer Intensive Methods, 109:153-171, 1996.

Mingoti Sueli Aparecida, Rosa Gilmar
A note on robust and non-robust variogram estimators
Rem: Revista Escola de Minas., Vol. 61:87–95, 2008.

Спасибо за внимание!

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