Слайд 1 Титульный лист

Слайд 2 Постановка задачи

В рамках данной работы мне была представлена задача выполнить на основе реальных данных: 1) предварительный стат. анализ; 2) вариограммный анализ; 3) исследование стат. свойств оценки вариограммы 4) прогнозирование методом кригинг

Слайд 3 Содержание. Можно опустить, либо быстро.

Слайд 4 Исходные данные

Таким образом в рассмотренном приложении мной решалась следующая реальная задача. Данные получены от учебно-научного центра «Нарочанская биологическая станция им. Г.Г.Винберга». И представляют собой выборку объёма 38, состоящую из значений средней температуры воды в июле месяце каждый год в период с 1975 по 2012. В целях дальнейшего оценивания результатов прогнозирования были использованы наблюдения только с 1975 по 2006 год.

Слайд 5 Обзор ПО. Особенности

Для решения поставленной задачи на языке программирования R мной было реализовано клиент-серверное приложение, позволяющее решать класс аналогичных по структуре задач. Приложение разбито на 3 модуля в соответствии с поставленной задачей: модуль предварительного анализа данных, модуль анализа остатков и модуль вариограммного анализа. Приложение доступно по адресу на экране с любого устройства. Имеет простой, в случае необходимости легко расширяемый интерфейс, с широкими графическими возможностями.

Слайд 6 Обзор ПО. Модуль предварительного стат. анализа. Первичный анализ

Страница первичного анализа представляет возможности по подбору закона распределения исследуемых данных с помощью как проверки различными тестами, так и визуально: на гистограмме и графике квантилей. Контрольная панель позволяет изменять отображаемый в данный момент график, а также позволяет выбрать критерий нормальности. Также на данной странице отображается таблица с вычисленными описательными статистиками.

Слайд 7 Проверка на нормальность

По описательным статистикам, графикам гистограммы и квантилей и проверкой тестов показана близость выборочного распределения к нормальному.

Слайд 8 Обзор ПО. Модуль предварительного стат. анализа. Корреляционный анализ

Страница корреляционного анализа позволяет оценить зависимость исследуемых данных с помощью диаграммы рассеяния, вычисляет коэффициент корреляции, проверяет его значимость и позволяет оценить наличие выбросов в данных. Тестом Граббса показано отсутствие выбросов. Показана умеренная положительная зависимость температуры воды от времени.

Слайд 9 Обзор ПО. Модуль предварительного стат. анализа. Регрессионный анализ

Вкладка регрессионного анализа позволяет получить регрессионную модель по исследуемым данным, а также провести анализ модели: определить значимость вычисленных коэффициентов, адекватность модели и проверка линейности.

Слайд 10 Регрессионный анализ

Найден вид исследуемого временного ряда, а так же уравнение тренда. На графике отображен ряд остатков, после его удаления из исходного временного ряда. Доказана значимость коэффициентов регрессионной модели, показана адекватность, отклонение от линейности незначительно. Но при этом коэффициент детерминации оказался не высоким(< 0.7)//, это говорит о том, что построенная регрессионная модель не описывает в достаточной мере поведение временного ряда//.

Вывод Инструменты, рассмотренные в рамках данного модуля, позволяют быстро получить информацию по исследуемым данным и сделать первые выводы и наметить шаги по дальнейшему исследованию.

Слайд 11 Обзор ПО. Модуль анализа остатков. Автокорреляционная функция

После регрессионного анализа и удаления из исходного временного ряда тренда получаем ряд остатков. Для его исследования реализован модуль анализа остатков включающий в себя первичный анализ (аналогичный рассмотренному) и анализ автокорреляционной функции. На слайде продемонстрирована страница анализа автокорреляционной функции, позволяющая как визуально, так и с помощью ряда тестов (Льюнга-Бокса, расширенный тест Дики-Фуллера) определить наличие автокорреляций и стационарности в исследуемом временном ряду.

Слайд 12 Анализ остатков

Показана близость распределения ряда остатков к нормальному. По графику и тестами Льюнга-Бокса и Дики-Фуллера сделано заключение об отсутсвии значимых автокорреляций и стационарности в широком смысле.

Вывод В зависимости от результатов, полученных на рассмотренном этапе, можно либо закончить исследование, либо продолжить в модуле вариограммного анализа. Закончить исследование стоит в том случае, если модель удовлетворительного качества, либо в случае, когда не выполняются условия для проведения следующего этапа.

Слайд 13 Вариограммный анализ

Прогнозные значения вычисляются как сумма значения вычисленных по тренду и по кригингу. Для оценки качества используются коэффициент корреляции между известными значениями ошибки и интерполяционными, и среднеквадратическая ошибка. На рисунке вычисленная оценка семивариограммы.

Слайд 14 Обзор ПО. Модуль вариограммного анализа. Семивариограмма

Начальный шаг состоит в подборе модели и её параметров к экспериментальной семивариограмме. Для построения экспериментальной семивариограммы присутствует возможность использовать оценку Матерона и робастную оценку Кресси-Хокинса. А также реализовано два подхода по подбору: визуально силами исследователя, и автоматическими методами. Инструменты данной страницы позволяют выбрать модель семивариограммы из списка и задать параметры. По графику можно оценить подобранные параметры.

Слайд 15 Обзор ПО. Модуль вариограммного анализа. Подбор параметров модели

Инструмент подбора параметров позволяет оценить как определённый параметр влияет на качество модели и конечный результат. В реализованном приложении имеется два подхода по оценке качества построенной модели. Используя первый подход, перекрёстный, модель оценивается с помощью метода кросс-валидации. При втором подходе, адаптивном, в исследуемых данных отдаётся предпочтение последним наблюдениям. На данной странице при выбранном подходе можно оценить поведение модели при изменении какого-либо из параметров и для каждого подобрать оптимальное значение.

Слайд 16 Обзор ПО. Модуль вариограммного анализа. Прогнозирование кригинг

Страница кригинга является наглядной демонстрацией результатов подбора моделей. На ней изображается график с наблюдаемыми значениями и прогнозными значениями. Это позволяет оценить полученную модель и сделать различные заключения. График также сопровождается вспомогательными таблицами с произведёнными в процессе расчётами.

Слайд 17 Визуальный подход. Линейная

Были рассмотрены различные модели семивариограмм, из них линейная является простейшей. График с этой моделью на слайде, под ним график сравнительного прогноза(пояснить обозначения). КК оказался близким к нулю, МСЕ высока, результат получен не очень хороший.

Слайд 18 Визуальный подход. Наггет

С помощью подгонки параметров получена модель с чистым эффектом самородков. Объяснить результат подгонки можно тем, что автоматический подбор параметров основан на методе наименьших квадратов. А поскольку значения семивариограммы сразу достигают порогового значения, то эффект самородков $\widehat{\gamma}_2(h)$, изображённый на рисунке, оказывается наилучшей моделью. Но при этом данная модель не учитывает особенностей исследуемых данных, поэтому результатов прогнозирования она не улучшила. Другими словам, данный подход не учитывает поведение оценки семивариограммы около нуля, поскольку в исходных данных нет информации о ближайших к исследуемому месяцах.

Слайд 19 Визуальный подход. Линейная с порогом

По аналогичной схеме рассмотрена линейная модель с порогом. Подобранные с помощью приложения (адаптивным методом) параметры позволили довольно точно предсказать неизвестные значения. Поэтому данная модель хороша для краткосрочных прогнозов. При этом КК не высок, но значительно выше предыдущих, а МСЕ наоборот. Что говорит о том, что данная модель не очень справляется с описанием всего ряда остатков.

Слайд 20 Визуальный подход. Сферическая

Данная модель похожа на предыдущую своим видом. Как результат, прогноз уловил поведение исходных данных. Но значения не очень точны.

Слайд 21 Визуальный подход. Периодическая

По графику оценки семивариограммы можно заметить некоторую периодичность. Поэтому была использована периодическая модель. В результате подбора, получился высокий КК, по сравнению со всеми подобранными. А по графику видно, что прогноз не точен, но и некоторую тенденцию он уловил.

Слайд 22 Автоматический подход.

Как было сказано ранее также реализован функционал по автоматическому подбору моделей семивариограмм. Так как подгонка параметров основывается на МНК, то на его результат влияет максимальный лаг, до которого вычисляются значения семивариограммы. Также для сравенения введём оценку Кресси-Хокинса.

На графике зависимость качества модели от макимального лага при автоматическом подборе. Как видно, для робастной оценки наилучшее значение 5, для Матерона – 28.

Слайд 23 Автоматический подход. Оценки

Для сравнения графики оценок семивариограмм. Они не сильно отличаются, в Кресси-Хоккинса наблюдается более четкие периоды.

Слайд 24 Автоматический подход. Волновая модель

По результатам построена такая модель с таким результатом. Плохо описывает ошибку в целом, но улавливает поведение неизвестных значений.

Слайд 25 Автоматический подход. Периодическая модель

Как и в случае ручного подбора, данная модель лучше себя показывает для описания ошибки. И в целом дает неплохой результат по прогнозу. Но не точен.

Слайд 26 Оценка вариограммы

Для непосредственно перехода к вариограммному анализу, введем следующие понятия: вариограмма, оценка вариограммы.

Слайд 27 Первые два момента

Мной найдены первые два момента введённой оценки. В теореме доказана несмещённость оценки.

Слайд 28 Асимптотика

Проведёно исследование асипмтотического поведения оценки.

Слайд 29 Асимптотика

Как следствие из доказанных теорем, показано что рассмотренная оценка вариограммы является несмещённой и состоятельной в среднеквадратическом смысле оценкой неизвестной вариограммы.

Вывод Визуальные методы точнее и надёжнее так как исследователь знает специфику данных, но требует определённый знаний у пользователя, как статистических так и опыт по вариограммному анализу. В свою очередь автоматический подбор может использоваться любыми пользователями, так как он сразу выдает результат. Ну и следует отметить так же, что температура воды является специфическим показателем(ОСОБЕННО В НАШЕМ СЛУЧАЕ), поскольку она может изменяться в широком диапазоне. В случае с данными, которые имеют более плавный характер изменений (например глубина), автоматический подбор будет показывать себя лучше.