Слайд 1 Титульник
Слайд 2 Постановка задачи
В рамках данной работы мне была представлена задача выполнить следующее:

1. Предварительный стат. анализ
2. Вариограммный анализ
3. Исследование стат. свойств оценки вариограммы
4. Прогнозирование методом Кригинг. Исследование точности прогноза в зависимости от оценки вариограммы и модели

Слайд 3 Содержание. Можно опустить, либо быстро.
Для решения поставленной задачи было

1. Мной реализовано программное обеспечение, позволяющее решать класс задача аналогичных поставленной
2. Исследование проведено в реализованном приложении.

Слайд 4 Обзор ПО. Особенности
Для решения поставленной задачи на языке программирования R мной было реализовано клиент-серверное приложение, позволяющее решать класс аналогичных по структуре задач. Для этого написаны несколько модулей, включающих в себя функционал, необходимый для решения конкретной подзадачи. Для удобства работы, каждый модуль имеет отдельные страницы, отвечающие за конкретные инструменты. Таким образом весь процесс работы в приложении разбивается на несколько этапов, на каждом из которых решается конкретная подзадача. В данной работе можно выделить три этапа: первичный анализ данных, анализ остатков и вариограммный анализ.

Приложение доступно по адресу на экране с любого устройства. Имеет простой, в случае необходимости легко расширяемый интерфейс, с широкими графическими возможностями. Любое изменение контрольных параметров сразу отображается в результатах.

Следует отметить, что каждая страница приложения имеет единый дизайн: экран можно условно поделить на панель выбора этапа анализа сверху и область исследования снизу. В свою очередь область исследования можно разделить также на две части: контрольная панель параметров и инструментов слева, и результаты вычислений и анализа справа. Любое изменение параметров контрольной панели сразу же отображается в качестве результата в области исследования.

Слайд 5 Обзор ПО. Модуль предварительного стат. анализа. Первичный анализ

Данный модуль включает в себя возможности по просмотру и анализу непосредственно данных: графически и с помощью таблицы, позволяющей сортировать и производить поиск по определённому признаку. Непосредственно на слайде отображена вкладка первичного анализа. В которой представлены возможности по определению 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. В случае выбора для отображения гистограммы, появляются управляющие элементы, позволяющие выбрать ширину столбца на гистограмме и правило по её вычислению (например, правило Стерджеса), отобразить плотность выборочного распределения и кривую нормального.

Также на данной странице отображается таблица с вычисленными описательными статистиками.

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

Данная страница позволяет оценить зависимость исследуемых данных с помощью диаграммы рассеяния, вычисляет коэффициент корреляции и с помощью критерия Стьюдента проверяет значимость вычисленного коэффициента, а также вычисляет для него доверительный интервал. Среди прочего, данная страница позволяет оценить наличие выбросов с помощью критерия Граббса.

Слайд 7 Обзор ПО. Модуль предварительного стат. анализа. Регрессионный анализ

Вкладка регрессионного анализа позволяет получить регрессионную модель по исследуемым данным. График временного ряда содержит также линию регрессии. Страница демонстрирует возможности по анализу вычисленной модели: определение значимости вычисленных коэффициентов, адекватность модели с помощью критерия Фишера и проверки линейности.

Вывод Инструменты, рассмотренные в рамках данного модуля, позволяют быстро получить информацию по исследуемым данным. А также сделать первые выводы и наметить шаги по дальнейшему исследованию. Заметим, что на каждом из этапов анализа и использования каждого из инструментов реализована возможность изменять объёмы выборки. Как снизу, так и сверху. Другими словами можно отбросить первые или последние наблюдения. Это позволяет быстро оценить, насколько влияют данные на результат в конкретном случае.

Слайд 8 Обзор ПО. Модуль анализа остатков. Автокорреляционная функция

Данный модуль является логическим продолжением рассмотренного ранее. После регрессионного анализа и удаления из исходного временного ряда тренда, основанного на регрессионном уравнении, получаем ряд остатков. Для его анализа реализованы возможности, которые включают в себя некоторые возможности предыдущего. Исключение составляют инструменты регрессионного и корреляционного анализов. Поскольку исследуется ошибка.

Таким образом данный модуль позволяет проверить остатки на нормальность как с помощью графиков квантилей и гистограммы, так и различными критериями: Шапиро-Уилка, χ^2-Пирсона, Колмогорова-Смирнова. В дополнение к этому имеется возможность проанализировать описательные статистики, а также исследовать автокорреляционную функцию. На слайде продемонстрирован график автокорреляционной функции, позволяющий визуально определить наличие автокорреляций в исследуемых данных. Также проверить наличие значимых автокорреляций позволяет проверка реализованного теста Льюнга-Бокса. В свою очередь, расширенный тест Дики-Фуллера, также представленный на рассматриваемой странице, проверяет наличие стационарности в исследуемом случайном процессе.

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 удовлетворила ..., либо в случае, когда не выполняются условия для проведения следующего этапа.

Слайд 9 Обзор ПО. Модуль вариограммного анализа. Семивариограмма

В данном модуле используются современные геостатистические методы и инструменты, которые, в рамках **R**, реализованы пакетом *gstat*. В этом пакете представлены функции для вычисления вариограмм, подбору моделей и параметров, интерполирования методами кригинга и методы валидации конечных результатов. Интерполирование методами кригинга подразумевает наличие подобранной модели вариограммы, поэтому в рассматриваемом модуле акцент сделан именно на подборе и анализе различных моделей вариограмм.

Начальный шаг состоит в подборе модели и её параметров к экспериментальной вариограмме. Для построения экспериментальной вариограммы присутствует возможность использовать две разновидности оценок вариограммы: оценка Матерона и робастная оценка Кресси-Хокинса. Для подбора модели вариограммы, в общем случае, существует два подхода: подбор визуально силами исследователя, и автоматическими методами. В данном модуле в полной мере реализованы оба подхода. В первом случае, изменение любого из параметров модели позволяет незамедлительно оценить эффект как на графике непосредственно вариограммы, так и по конечному прогнозу кригингом.

На слайде изображён скриншот начального этапа вариограммного анализа. Инструменты данной страницы позволяют выбрать модель из следующих: Линейная, Сферическая, Экспоненциальная, Гауссовская, Круговая, Бесселя, Пентасферическая, Волновая, Логарифмическая. А также задать к выбранной модели параметры. Заданные параметры считаются начальными, если выбрать опцию подгона методом наименьших квадратов. На этом шаге также можно воспользоваться реализованным в рамках данной работы алгоритмом автоматического подбора модели. Данная функциональность позволяет сразу перейти к вычислению прогнозных значений и не требует каких-либо прикладных знаний у пользователя. Алгоритм заключается в переборе всех представленных в пакете *gstat* моделей, и подборе параметров с помощью функции *fit.variogram* из того же пакета. Каждая итерация сопровождается оптимальным набором параметров для конкретной модели и невязкой. Выбор наилучшей модели осуществляется по минимальному значению невязки. Представленная страница позволяет оценить по графику вариограммы подобранные либо вручную, либо автоматически модель и параметры. Можно также проследить, как влияет тот или иной параметр на теоретическую вариограмму.

При использовании той или иной модели интерполяции крайне важно правильно подобрать значения модельно-зависимых параметров. Для кригинга такими параметрами являются параметры модели вариограммы. Для проверки качества модели в дальнейшем используется кросс-валидация.

Слайд 10 Обзор ПО. Модуль вариограммного анализа. Подбор параметров модели

На данной странице заключена функциональность по подбору параметров. В большей мере это относится к ручному выбору. В общем случае, подбор осуществляется следующим образом:

- выбирается параметр для подбора, диапазон поиска и шаг итерации
- на каждом шаге кригингом вычисляются прогнозные значения
- на основе полученных значений строится выбранная статистика

В результате такого процесса получается ряд статистик, из которых выбирается оптимальная. Затем процесс повторяется для другого параметра и так далее, пока не найдётся оптимальная модель.

В реализованном приложении имеется два подхода по оценке качества построенной модели. Используя первый подход, модель оценивается с помощью описанного ранее метода кросс-валидации. При втором подходе, адаптивном, в исследуемых данных отдаётся предпочтение последним наблюдениям. Для этого из исходных данных исключается некоторое количество значений и модель строится по оставшимся. Подбор параметров осуществляется по статистикам, основанным на отклонении вычисленных значений от исключенных. Таким образом достигается наилучший прогноз в краткосрочной перспективе.

Таким образом на данной странице можно оценить поведение модели при изменении какого-либо из параметров и для каждого подобрать оптимальное значение.

Слайд 11 Обзор ПО. Модуль вариограммного анализа. Прогнозирование кригинг

Страница кригинга является наглядной демонстрацией применения всего вышеописанного. На ней изображается график с наблюдаемыми значениями и прогнозными значениями, вычисленными кригингом и по регрессионной модели. Это позволяет оценить полученную модель и сделать различные заключения. График также сопровождается вспомогательными таблицами с произведёнными в процессе расчётами. В первую очередь это результаты кригинга с ошибкой для каждого из значений. Также отображается табличный вариант данных, изображённых на графики. И последняя таблица показывает значения статистик после применения кросс-валидации, что сразу позволяет сравнить конкретную модель с другими.

Слайд 12 Исходные данные

Таким образом в описанном приложении, в рамках данной работы, решалась следующая задача. Данные получены от учебно-научного центра «Нарочанская биологическая станция им. Г.Г.Винберга». И представляют собой объёма 38, состоящую из значений средней температуры воды в июле месяце каждый год в период с 1975. Следует отметить, что для непосредственного изучения в данном разделе были использованы наблюдения с 1975 по 2006 год. Наблюдения за 2007-2012 годы были намеренно исключены из исследования в целях дальнейшего оценивания результатов анализа и прогнозирования. ?На результаты предварительного анализа применение этого приема существенно не повлияло.

Слайд 13 Проверка на нормальность

Были вычислены описательные статистики, по ним сделан вывод о небольшой скошенностью вправо (коэффициент асимметрии) и пологостью пика кривой распределения (коэффициент эксцесса) относительно нормального. По графикам гистограммы и квантилей и проверкой тестов показана близость выборочного распределения к нормальному.

Слайд 14 Корреляционный анализ

Проведённым тестом Граббса показано отсутствие выбросов. Показана умеренная положительная зависимость температуры воды от времени.

Слайд 15 Регрессионный анализ

По причине того, что в данном случае мы рассматриваем среднюю температуру июля месяца каждого года на протяжении длительного периода, сделан вывовд об отсутствии циклической и сезонной составляющих временного ряда. Так как не происходит увеличения амплитуды колебаний с течением времени, сделан вывод об аддитивности искомой модели. Таким образом вид исследуемого временного ряда на слайде. И так же уравнение тренда. На графике отображен ряд остатков, после вычитания тренда из исходного временного ряда.

Слайд 16 Качество регрессионной модели

Доказана значимость коэффициентов регрессионной модели, критерием Фишера 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. Это, в свою очередь, может значить, что изменение температуры зависит не только от времени, но ещё и от каких-то других, неучтённых, факторов.

На таблице представлены исходные данные, вычисленный тренд, и ошибка.

Слайд 17 Анализ остатков

Как и в случае исходных данных показана близость распределения к нормальному, с чуть большими отклонениями. По графику (столбцы автокорреляций не выходит за интервал пунктиром) и тестом Льюнга-Бокса сделано заключение об отсутсвии значимых автокорреляций. На графике значения имеют тенденцию к затуханию, что говорит о стационарности в широком смысле, что подтвердил расширенный тест Дики-Фуллера.

Результатам предварительного анализа позволили перейти к геостатистическим методам.

Слайд 18 Оценка вариограммы

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

Слайд 19 Первые два момента

Мной найдены первые два момента введённой оценки. Следует заметить, что в рамках данной теоремы доказана несмещённость оценки.

Слайд 20 Асимптотика

Проведёно исследование асипмтотического поведения оценки.

Слайд 21 Асимптотика

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

Слайд 22 Вариограммный анализ

Прогнозные значения вычисляются как сумма значения по тренду и по кригингу. Для оценки качества используются коэффициент корреляции между известными значениями ошиьки и интерполяционными и среднеквадратическая ошибка.

На рисунке оценка семивариограммы.

Слайд 23 Визуальный подход. Линейная

На практике советуют начинать с простейшей модели семивариограммы: линейной. Ее вид на слайде. Значение параметра подбирается в общем случае по первому значению семивариограммы, чтобы уловить поведение вначале. График с моделью на слайде, под ним график сравнительного прогноза(пояснить обозначения).

КК оказался близким к нулю, МСЕ высока, результат получен не очень хороший.

Слайд 24 Визуальный подход. Наггет

С помощью подгонки параметров средствами R получена модель с чистым эффектом самородков. Она характеризуется разрывом первых значений и началом координат. Объяснить результат подгонки можно тем, что автоматический подбор параметров основан на методе наименьших квадратов. А поскольку значения семивариограммы сразу достигают порогового значения, приблизительно равному дисперсии, 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.

КК оказался -1, МСЕ высока, таким образом результата прогноза данная модель не улучшила, поскольку значения оказались близки к нулю.

Слайд 25 Визуальный подход. Линейная с порогом

По аналогичной схеме рассмотреная линейная модель с порогом. Подобранные с помощью приложения (адаптивным методом) параметры позволили довольно точно предсказать неизвестные значения. Поэтому данная модель хороша в данном случае 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 26 Визуальный подход. Сферическая

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

Слайд 27 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, невысокий МСЕ. А по графику видно, что прогноз не точен. Но и некоторую тенденцию уловил.

Слайд 28 Автоматический подход.

В рамках приложения также реализован функционал по автоматическому подбору моделей семивариограмм. Заключается он в последовательном переборе всех заданных моделей, подгонке параметров каждой и выборе оптимальной по минимальному значению суммы квадратов невязок. Так как в подгонке параметров используется МНК, то на его результат влияет максимальный лаг, до которого вычисляется значения семивариограммы. Также для сравенения введём оценку Кресси-Хокинса. На графике зависимость качества модели от макимального лага. Как видно, для робастной оценки значение 5, для Матерона – 28.

Слайд 29 Автоматический подход. Оценки

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

Слайд 30 Автоматический подход. Волновая модель

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

Слайд 31 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.

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.