Международный государственный экологический
университет им. А. Сахарова
Кафедра экологических информационных систем

Г. Куканков

Анализ пространственно распределенных данных

Методические указания по выполнению лабораторных работ

Минск 2005

Автор:
Г. Куканков

Анализ пространственно распределенных данных. Методические указания по выполнению лабораторных работ / Г. Куканков -- Минск, 2005

Приведены сведения из теории и методические указания по выполнению заданий по курсу с использованием пакета Matlab. Для выполнении заданий используются данные о загрязнении Полесского радиационного заповедника ИРБ АНРБ.

1. Моделирование связи поверхностной активности цезия-137 и мощности экспозиционной дозы с использованием пакета Matlab

Данные: В нашем распоряжении имеются реальные данные, которые представляют собой результаты измерения поверхностной активности цезия-137 и мощности экспозиционной дозы. Измерения проводились в 1986 году на территории, которая в настоящее время принадлежит Полесскому радиационному заповеднику. Всего измерения, включающие отбор проб для определения поверхностной активности цезия-137 и замер мощности экспозиционной дозы гамма-излучения были произведены почти в 3-х тысячах точек местности.

В файле `inidata.mat` содержится матрица `drcs` состоящая из двух колонок. В первой содержатся значения мощности экспозиционной дозы в микрорентгенах в час, а во второй – поверхностная активность цезия-137 в Кюри на квадратный километр.

Задача: Предлагается средствами `Matlab` исследовать связь между мощностью экспозиционной дозы и оценить возможность получения модельных оценок поверхностной активности цезия по результатам измерения мощности экспозиционной дозы гамма-излучения.

Поставленную задачу будем решать в несколько этапов.

1. Преобразование данных о поверхностной активности цезия-137 к международным единицам kBq/m.sq.
2. Знакомство с данными – описательная статистика, графическое представление исходных данных, построение гистограмм, оценка коэффициента корреляции.
3. Построение регрессионной модели по исходным данным и оценка качества моделирования.
4. Построение регрессионной модели по логарифмам исходных данных.

Рекомендуется в случае неочевидности применения той или иной команды перед ее использованием ознакомиться с ее описанием. Например

```
help load
```

выведет на экран описание команды `load`.

Загрузка данных

Произведем очистку памяти (команда `clear`) и загрузим исходные данные командой `load inidata.`

Ознакомимся с содержимым памяти (команда `whos` покажет, что в памяти содержится одна переменная-массив `drcs`), обратим внимание на размер массива данных – количество строк по числу точек отбора проб и количество столбцов – 2 – по числу наблюдаемых величин. Каждый столбец переменной-массива `drcs` содержит данные только об одной переменной – мощности экспозиционной дозы (1-й столбец) и поверхностной активности цезия (2-й столбец).

Преобразование данных о поверхностной активности цезия-137 из Ки/км.кв. к международным единицам Бк/м.кв.

для удобства дальнейшей работы переведем результаты измерения поверхностной активности из Ci/km^2 в kBq/m^2:

```
dc(:,2)=drcs(:,2)*37;
```

Здесь “37” – коэффициент перехода от размерности Ci/km^2 к размерности kBq/m^2.

Размерность мощности экспозиционной дозы менять не будем:

```
dc(:,1)=drcs(:,1);
```

Знакомство с данными – описательная статистика, графическое представление вариационного ряда исходных данных, построение гистограмм, оценка коэффициента корреляции.

Определим (a) максимальное, (b) минимальное, (c) среднее значения измеренных величин, а также (d) медиану и (e) среднеквадратичное (стандартное) отклонение для каждой из переменных:

```
a)max(dc) b)min(dc) c)mean(dc) d)median(dc)
e)std(dc)
```

Обратим внимание на то, что указанные статистические характеристики подсчитываются по столбцом переменной-массива.

Построим вариационный ряд измеренных величин и отобразим его графически, т.е. расположим выборочные значения в порядке возрастания. Для этого воспользуемся командой `sort:`

```
[y,i]=sort(dc(:,1));
```

Затем построим вариограмму – график зависимости выборочного значения от его номера в вариационном ряду (ранг – rank):

```
plot(dc(i,1),'r'), grid
ylabel('Dose rate, \muRe per hour)
```

Аналогично строится вариационный ряд для второй переменной (постройте его самостоятельно).

Рекомендация: Для того, чтобы окно с гистограммой мощности экспозиционной дозы не “пропадало”, перед построением очередного графика набирайте команду

```
figure
```

которая создает новое окно.

Некоторое представление о распределении случайной величины дает гистограмма – ступенчатая кривая, представляющая зависимость частоты попадания в некоторый интервал от значения величины в центре интервала.

```
hist(dc(:,1)), title('Histogram of Dose Rate')
```

Эти команды строят в графическом окне гистограмму распределения значений в выборке мощности экспозиционной дозы.

Самостоятельно постройте гистограмму распределения значений поверхностной активности. *Напоминаем*: Перед построением очередного графика набирайте команду `figure` которая создает новое окно.

Подпишите заголовки и оси.

Рекомендуется использовать повторный нызов ранее использованных команд. Вызывается нажатием клавиши “стрелка вверх”, после чего команду можно подредактировать обычным образом. Выполнение производится по нажатию клавиши “Enter.”

Сравните гистограммы.

Построим скатерограмму – графическое изображение связи между двумя переменными:

```
plot(dc(:,1), dc(:,2),'.')
```

Напоминаем: Перед построением очередного графика набирайте команду `figure` которая создает новое окно.

Обратим внимание на характер скатерограммы. Связь двух выбранных переменных очевидна – возрастание одной величины влечет за собой возрастание второй (связь прямая). Характер выборки таков, что большинство выборочных значений сосредоточено в области небольших значений измеренных величин. Рассматривая полученную скатерограмму можно предположить близкую к линейной связь между переменными.

Степень линейной связи между переменными характеризуется коэффициентом корреляции. Чем ближе модуль коэффициента корреляции к единице – тем ближе связь между величинами к линейному виду. Если коэффициент корреляции близок к 0, говорят, что связь отсутствует (не обнаружена). Связь считается тесной если коэффициент корреляции находится в пределах 0.75-1.

Рассчитаем коэффициент корреляции между признаками:

```
corrcoef(dc)
```

Оцените степень линейной связи между признаками.

Скатерограммы и оценка коэффициента корреляции для логарифмов исходных данных,

Во многих практически важных случаях исходные данные подвергаются предварительному преобразованию – трансформации. Наиболее часто на практике применяют логарифмирование исходных данных.

Прологарифмируем все значения исходной матрицы данных и запишем ее в новую переменную `logdc`

```
logdc=log(dc);
```

Самостоятельно проведите анализ переменной `logdc` аналогичный тому, который проводился для переменной `dc`

Построение регрессионной модели по исходным данным и оценка качества моделирования

В случае, когда выявлена связь между переменными, полезно построить модель выявленной зависимости. Наиболее просто построить линейную модель. Метод, когда подгонка линейной зависимости к данным осуществляется методом наименьших квадратов, традиционно называется методом *линейной регрессии.*

В пакете Matlab подгонка модели к данным может быть осуществлена с использованием команд `polyfit` и `polyval`

Как видно из названий эти команды работают с полиномами (произвольной степени). Полином вида $a_1*x^0+a_2*x^1$ даст нам искомую линейную модель.

Самостоятельно постройте модель линейной регрессии для переменных из массива `dc`. Постройте зависимость мощности экспозиционной дозы (зависимая переменная) от поверхностной активности (независимая переменная) [Заметим, что использование в качестве независимой переменной поверхностной активности не совсем корректно -- независимая переменная в методе линейной регрессии не должна быть случайной величиной].

Напоминание: Формат используемых команд `polyval` и `polyfit` определите с помощью команды `help`

```
help polyfit
help polyval
```

Оценка качества моделирования может быть проведена с использованием различных тестов примененных к *остаткам* – значениям, которые определяются как разность между измеренными значениями зависимой переменной и ее значениями предсказанными регрессионной моделью.

Мы ограничимся визуальной оценкой структуры распределения остатков вдоль линии независимой переменной. Для этого вычислите остатки и постройте график зависимости остатков от поверхностной активности.

Самостоятельно проведите анализ переменной `logdc` аналогичный тому, который проводился для переменной `dc`

Сравните полученные графики.

Приложение

Выдержки из файла документации к пакету Matlab – “Getting started with Matlab” в переводе на русский язык.

2. Анализ расположения объектов в пространстве

Постановка задачи: Провести анализ расположения точек используя представленную выборку. Проверить гипотезы о кластерном или равномерном распределении а) методом ближайших соседей, б) методом квадратов.

Данные:

Выборка значений положений точек измерения поверхностной активности цезия-137 и мощности экспозиционной дозы. Измерения проводились в 1986 году на территории, которая в настоящее время принадлежит Полесскому радиационному заповеднику. Всего измерения, включающие отбор проб для определения поверхностной активности цезия-137 и замер мощности экспозиционной дозы гамма-излучения были произведены почти в 3-х тысячах точек местности. Представленная выборка содержит около 30 точек измерений. из всего набора точек.

3. Геостатистика: основные сведения о полувариограмме

Полудисперсия есть мера степени пространственной зависимости между пробами вдоль заданной базы. Для простоты мы предположим, что пробы являются точечными измерениями некоторого свойства, такого, как глубина подповерхностного горизонта. Для облегчения вычислений мы будем далее предполагать, что база регулярная, т. е. пробы равномерно расположены в пространстве вдоль прямых линий. Если расстояние между пробами по прямой линии равно некоторой величине Δ, то полудисперсия может быть вычислена для расстояний, кратных Δ:

$$\gamma_h = \frac{1}{2n} \sum_{i=1}^{n-h} (X_i - X_{i+h})^2 \ .$$

В этих обозначениях X_i – значение регионализованной переменной, взятой в точке i; $X_i + h$ – другое значение, взятое через h интервалов. Мы поэтому нашли сумму квадратов разностей между значениями регионализованной переменной в паре точек, разделенных расстоянием Δh. Число точек равно n, так что число сравнений между парами точек есть $n - h$.

В случае расположения точек на плоскости бывает затруднительно выбирать расстояния кратные Δ. Поэтому приведенную формулу можно заменить используя все возможные расстояния между точками близкие к Δh, и формула примет вид

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n_h}(Z(x,y) - Z(x,y|h))^2 ,$$

где $Z(x,y)$ – значение регионализованной переменной, взятой в точке (x,y); $Z(x,y|h)$ – другое значение, взятое на расстоянии h от первой точки; n_h – число пар точек на расстоянии h (число сравнений между парами точек будет разным для выбранного расстояния h). Заметим, что данная формула записана в предположении изотропности свойств регионализованной переменной.

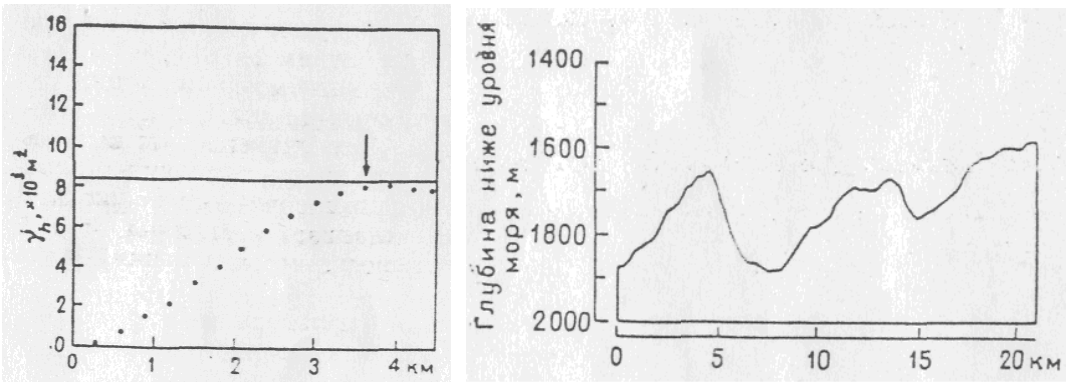

Рис. Полувариограмма абсолютных отметок кровли меловой формации, измеренной вдоль морского сейсмического разреза в Магеллановом проливе, Чили.
Линия, изображенная точками, представляет порог, или дисперсию, возвышений и равна 8380 м^2. Ранг, указанный стрелкой, – расстояние, ниже которого разность между дисперсиями и порогом считается пренебрежимо малой (3.5 км)

Рис. Подпочвенные структурные абсолютные отметки кровли меловой формации, оцененные по отражению сейсмических волн вдоль 21-километрового морского траверса в Магеллановом проливе. Сейсмические измерения взяты в 300-метровом интервале

Вернемся к одномерному случаю. Если мы вычислим полудисперсии для различных значений h, то мы можем нанести результаты на график в виде полувариограммы, являющейся аналогом коррелограммы. На следующих рис. представлена полувариограмма, соответствующая глубине сейсмически отражающего горизонта и построенная по измерениям вдоль приведенного сейсмического профиля. Заметим, что когда расстояние между точками опробования равно нулю, то значение в каждой точке сравнивается с самим собой. Следовательно, все разности равны нулю, и полудисперсия для γ_0 есть нуль. Если Δh – малое расстояние, точки при сравнении оказываются очень похожими, и полудисперсия будет мала. По мере увеличения расстояния Δh

сравниваемые точки становятся слабее связанными друг с другом и расстояния между ними увеличиваются, что приводит к большим значениям Δh. Предположим, что на некотором расстоянии сравниваемые точки находятся так далеко, что они не связаны друг с другом, и их квадраты разностей будут равны по величине дисперсии относительного среднего значения. Полудисперсия более не растет и полувариограмма переходит в плоскую область, называемую порогом. Расстояние, на котором полудисперсия приближается к дисперсии, называется рангом, или размахом регионализованной переменной, оно определяет окрестность, в пределах которой все положения связаны друг с другом.

Для некоторой произвольной точки в пространстве мы можем представить себе окрестность как симметричный интервал (или площадь, или объем, в зависимости от размерности) вокруг точки. Если регионализованная переменная стационарна или всюду имеет одно и то же среднее значение, то любое положение вне этого интервала совершенно независимо от центральной точки и не может давать информацию вокруг значения регионализированной переменной в этой точке. В пределах этой окрестности, однако, регионализированная переменная во всех наблюдаемых точках связана с регионализованной переменной в центральной точке и, следовательно, может быть использована для оценки ее значения. Если мы используем множество измерений, сделанных в точках внутри этой окрестности для оценки значения регионализованной переменной в центральной точке, то полувариограмма обеспечит собственные веса, которые должны быть приписаны каждому измерению.

Полудисперсия равна не только среднему значению квадратов разностей для пар точек, расположенных на расстоянии Δh друг от друга, но и дисперсии этих разностей, т. е. полудисперсия может быть определена по формуле

$$\gamma_h = \frac{1}{2n} \sum \left\{ (X_i - X_{i+h}) - \frac{1}{n} \sum_i (X_i - X_{i+h}) \right\}^2 .$$

Заметим, что среднее значение регионализованной переменной X_i, есть также среднее регионализованной переменной X_{i+h}, так как это – те же самые наблюдения, только взятые в другом порядке, т. е.

$$\frac{\sum X_i}{n} = \frac{\sum X_{i+h}}{n} .$$

Поэтому их разность должна быть равна нулю

$$\frac{\sum X_i}{n} - \frac{\sum X_{i+h}}{n} = 0 .$$

Комбинируя суммы, получаем

$$\frac{(\sum X_i - \sum X_{i+h})}{n} = \frac{\sum (X_i - X_{i+h})}{n} = 0 .$$

Заметим, что это соотношение строго справедливо только тогда, когда регионализованная переменная стационарна. Если данные не стационарны, то среднее значение последовательности изменяется вместе с h, и должно быть модифицировано.

Полувариограмма отражает пространственное поведение регионализованных переменных или их остатков. Некоторые идеализированные формы полувариограмм даны на следующем рис.

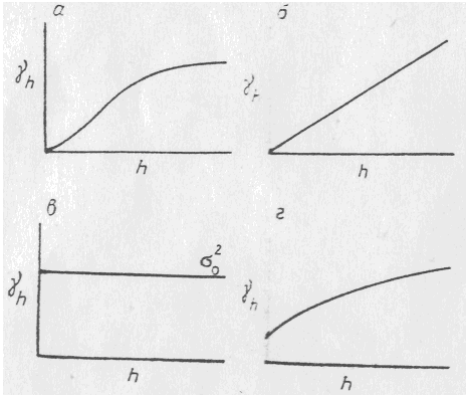

Рис. Идеализированные полувариограммы:
а – параболическая форма, показывающая отличную непрерывность регионализованной переменной; б–линейная форма, показывающая умеренную непрерывность; в – горизонтальная форма уровня σ_0^2, соответствующая случайной переменной, не имеющей пространственной автокорреляции: г–эффект самородков, или явное отклонение полувариограммы от начала координат, показывающее, что регионализованная переменная сильно изменчива при расстояниях, меньших чем интервал опробования

В принципе экспериментальная полувариограмма может быть прямо использована для получения оценок, которые мы рассмотрим позже. Однако полувариограмма известна только в дискретном наборе точек, расположенных на расстояниях Δh; на практике, однако, полувариограммы могут потребоваться для любых расстояний независимо от того, является ли оно кратным Δ или нет. По этой причине дискретная экспериментальная полувариограмма должна быть представлена некоторой непрерывной функцией, которая может быть вычислена для любого желаемого расстояния.

Рассмотрим только линейную модель. Линейная модель проще других, так как она имеет только один параметр, наклон. Модель имеет вид

$$\gamma_h = \alpha h$$

и представляет собой прямую, проходящую через начало координат. Очевидно, эта модель не может иметь пика, так как она растет неограниченно. Иногда линейная модель произвольно модифицируется с помощью вставки внезапного излома в точке пика, как, например,

$$\gamma_h = \alpha h \qquad \text{для} \qquad h < a\,,$$
$$\gamma_h = \sigma_0^2 \qquad \text{для} \qquad h \geq a\,.$$

Постановка задачи: Используя набор измерений регионализованной переменной – поверхностной активности цезия-**137** – проанализировать методами геостатистики ее статистические свойства.

Данные:

Результаты измерения поверхностной активности цезия-137 и мощности экспозиционной дозы. Измерения проводились в 1986 году на территории, которая в настоящее время принадлежит Полесскому радиационному заповеднику. Всего измерения, включающие отбор проб для определения поверхностной активности цезия-137 и замер мощности экспозиционной дозы гамма-излучения были произведены почти в 3-х тысячах точек местности.

Ориентировочная последовательность действий:

1. Провести обычный статистический анализ переменной: вычислить показатели описательной статистики, построить необходимые распределения, визуализировать пространственное распределение точек измерений и т. д.

2. Вычислить на основе представленных данных полувариограмму и провести структурный анализ: определить пространственный лаг корреляции переменной, оценить на основе данных ошибку прибора (дисперсия на лаге 0). Для начала пользоваться упрощенными предположениями, в частности считать, что свойства переменной изотропны, предположить линейность модели полувариограммы.

3. Геостатистика: кригинг

Точечный кригинг – простейшая форма кригинга, в котором наблюдения состоят из измерений, взятых в безразмерных точках и оценки проводятся в других местах, которые сами также являются безразмерными точками. Точечный кригинг используется, например, в построении карты в изолиниях для наблюдений, являющихся абсолютными отметками кровли формации,

измеренными в ряде разведочных скважин. Построение структурной карты в изолиниях требует, чтобы оценки абсолютных отметок кровли формации были сделаны в близко расположенных точках на картируемой площади. Проделав это, можно провести изолинии через эти оценки так, как описано в предыдущем разделе.

Для упрощения задачи можно допустить, что картируемая переменная статистически стационарна или свободна от дрифта. Значение в точке, не принадлежащей выборке, может быть оценено как взвешенное среднее известных наблюдений, т. е. значение в точке p основано на ограниченном множестве близлежащих контрольных точек:

$$\hat{Y}_p = \sum W_i Y_i \ .$$

Следует ожидать, что оценка $\hat{Y}_p$ будет отличаться от истинного (но неизвестного) значения Y_p на величину, которую можно назвать ошибкой оценки:

$$\varepsilon_p = \left(\hat{Y}_p - Y_p\right) \ .$$

Если сумма весов, используемых в оценке, равна единице, то полученная оценка называется несмещенной при условии, что дрифта нет. Это значит, что для большого множества оценок средняя ошибка будет равна нулю, так как положительные и отрицательные отклонения взаимно компенсируют друг друга. Однако даже если средняя ошибка оценки оказывается нулевой, оценки могут быть широко рассеянными относительно истинных значений. Это рассеяние можно охарактеризовать дисперсией ошибки:

$$s_\varepsilon^2 = \frac{1}{n} \sum \left(\hat{Y}_p - Y_p\right)^2$$

или после извлечения квадратного корня стандартной ошибкой оценки:

$$s_\varepsilon = \sqrt{s_\varepsilon^2} \ .$$

Интуитивно представляется правдоподобным, что ближайшие контрольные точки оказываются наиболее влияющими на оценку значения в точке поверхности, не являющейся точкой опробования, и что более удаленные контрольные точки оказывают меньшее влияние. Мы также вправе ожидать, что используемые веса в процессе оценивания и ошибки оценок некоторым образом должны быть связаны с полувариограммой поверхности.

Предположим, что требуется оценить значение Y в точке p по трем близким точкам, используя в качестве оцениваемого параметра взвешенное среднее трех известных значений:

$$\hat{Y}_p = W_1 Y_1 + W_2 Y_2 + W_3 Y_3 \ .$$

Веса подчинены условию, что сумма их равна единице, поэтому в отсутствие тренда оценка является несмещенной. Предположим, что вес W_1 выбран равным 1.0. Тогда веса W_2 и W_3 должны быть равны нулю, и оценка в точке p есть

$$\hat{Y}_p = 1Y_1 + 0Y_2 + 0Y_3 \ .$$

или

$$\hat{Y}_p = Y_1 \ .$$

Очевидно, ошибка оценки есть просто $\varepsilon = Y_p - Y_1$, так как Y_1 есть оценка Y_p. Если многие другие значения, подобные Y_p, оцениваются по точкам, размещенным в пространстве подобно Y_1, то оценка дисперсии может быть вычислена как средняя квадратичная разность между этими парами точек. Для удобства можно обозначить эти оцененные значения через Y_{pi} и другие оценки через Y_{1i}. Тогда

$$s_\varepsilon^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_{pi} - Y_{1i})^2 \ .$$

Ясно, что оценка дисперсии равна удвоенной полувариограмме для расстояния, равного интервалу, разделяющему точки Y_{pi} и Y_{1i}.

В данном случае выбрана одна частная комбинация весов для получения оценки Y_p и для определения ошибки оценки. Имеется бесконечное множество способов комбинирования весов, которые должны быть выбраны, и каждый из них дает различную оценку и различную ошибку оценки. Однако имеется только одна комбинация, которая будет давать минимум ошибки оценивания. Именно эту комбинацию весов и позволяет найти кригинг.

Получение уравнении кригинга требует вычислений и здесь не рассматривается. Оптимальные значения для весов можно найти решением системы совместных уравнений, которые включают значения из вариограммы оцениваемой переменной. Эти веса оптимальны в том смысле, что окончательные оценки являются несмещенными и имеют минимальную оценку дисперсии. Никакая другая линейная комбинация наблюдений не может дать оценки, которые имеют меньшее рассеяние относительно их истинных значений.

В простейших случаях задача кригинга состоит в оценке значения Y в точке p по трем известным наблюдениям. Для ее нахождения требуется решить систему трех уравнений:

$$W_1\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) = \gamma(h_{1p}) ,$$
$$W_1\gamma(h_{12}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) = \gamma(h_{2p}) ,$$
$$W_1\gamma(h_{13}) + W_2\gamma(h_{23}) + W_3\gamma(h_{33}) = \gamma(h_{3p}) .$$

Здесь $\gamma(h_{ij})$ – полувариограмма на расстоянии h, соответствующем интервалу между контрольными точками i и j. Например, $\gamma(h_{13})$ – полувариограмма для расстояния между известными точками 1 и 3; $\gamma(h_{1p})$ – полувариограмма для расстояния между известной точкой 1 и точкой p, в которой производится оценка. Матрица в левой части системы симметрична, так как $h_{ij} = h_{ji}$. Диагональные элементы этой матрицы равны нулю, так как h_{ii}, представляет расстояние точки от себя самой, которое равно нулю. В предположении, что полувариограмма проходит через начало координат, полувариограмма для нулевого расстояния равна нулю. Значения полудисперсии взяты из полувариограммы, которая должна быть известна или оценена до кригинга.

Однако, для того чтобы обеспечить несмещенность решения, необходимо наложить ограничение на веса: их сумма должна быть равна единице. Четвертое уравнение

$$W_1 + W_2 + W_3 = 1 .$$

В итоге получается набор четырех уравнений для трех неизвестных. Так как уравнений больше, чем неизвестных, то для того чтобы обеспечить минимально возможную ошибку оценки, нужно использовать дополнительные степени свободы. Это делается добавлением в систему уравнений немой переменной λ, называемой множителем Лагранжа. Полная система уравнений имеет следующий вид

$$W_1\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) + \lambda = \gamma(h_{1p}) ,$$
$$W_1\gamma(h_{12}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) + \lambda = \gamma(h_{2p}) ,$$
$$W_1\gamma(h_{13}) + W_2\gamma(h_{23}) + W_3\gamma(h_{33}) + \lambda = \gamma(h_{3p}) ,$$
$$W_1 + W_2 + W_3 + 0 = 1$$

или в матричной форме

$$\begin{bmatrix} \gamma(h_{11}) & \gamma(h_{12}) & \gamma(h_{13}) & 1 \\ \gamma(h_{12}) & \gamma(h_{22}) & \gamma(h_{23}) & 1 \\ \gamma(h_{13}) & \gamma(h_{23}) & \gamma(h_{33}) & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma(h_{1p}) \\ \gamma(h_{2p}) \\ \gamma(h_{3p}) \\ 1 \end{bmatrix}$$

В общем виде нужно решить матричное уравнение

$$[A][W] = [B]$$

для вектора неизвестных коэффициентов $[W]$. Члены матрицы $[A]$ и вектора $[B]$ берутся непосредственно из полувариограммы или из математических функций, описывающих ее вид.

Определив неизвестные веса, значение оцениваемой переменной в точке p представим в виде

$$\hat{Y}_p = W_1 Y_1 + W_2 Y_2 + W_3 Y_3 \ .$$

Оценка дисперсии имеет вид

$$s_\varepsilon^2 = W_1 \gamma(h_{1p}) + W_2 \gamma(h_{2p}) + W_3 \gamma(h_{3p}) + \lambda \ .$$

Иными словами, дисперсия оценки есть в сущности взвешенная сумма полудисперсий для расстояний до точек, использованных в оценивании, плюс вклад от коэффициента λ, который эквивалентен постоянному члену. Крайгинг имеет два больших преимущества перед обычными процедурами оценивания, которые используются при построении карт в изолиниях. Оценки процедур кригинга в среднем имеют наименьшую возможную ошибку, и также обеспечивают явное выражение величины этой ошибки.

Для иллюстрации точечного кригинга мы приведем оценку уровня воды в точке p на карте, представленной на рис. Оценка будет проведена по известным уровням, измеренным в трех наблюдательных скважинах. Координаты карты скважин и расстояния между ними приведены в табл. Предварительный структурный анализ позволил получить линейную полувариограмму. Значения полудисперсии, соответствующие расстоянию между скважинами, даны в табл.; они могут быть получены прямо из полувариограммы или вычислены из наклона.

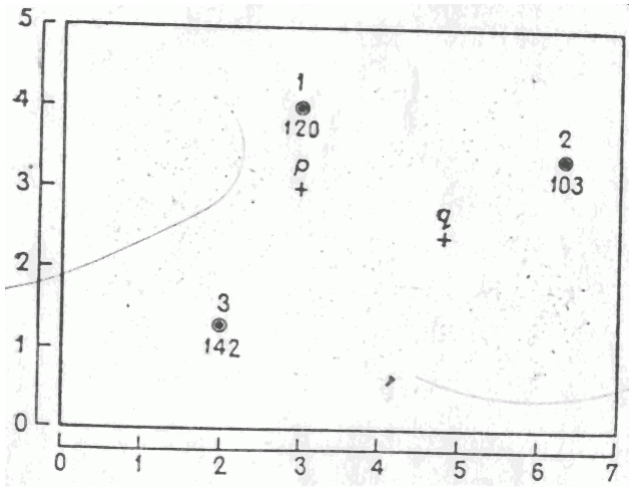

Рис. Карта уровней воды (в м) в трех наблюдательных скважинах.
Оценки уровня воды сделаны в точках *p* и *q*. Координаты указаны (в км) для произвольного начала

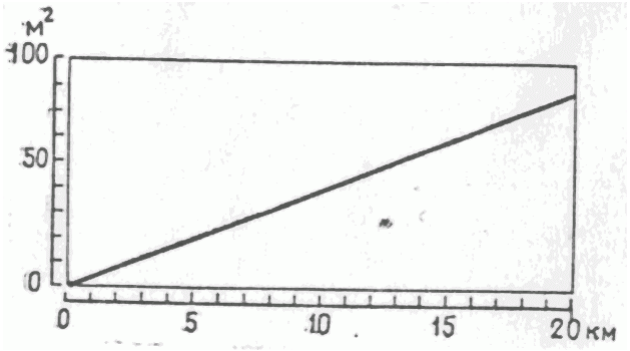

Рис. Линейная полувариограмма уровней воды на площади, включающей карту на предыдущем рис. Полувариограмма имеет наклон 4.0 м^2/км внутри 20-километровой зоны

Таблица

Наблюдения, проведенные в скважинах,
используемых для оценки уровня воды в точке *p*

Скважина	Координата X_1	Координата X_2	Уровень воды
1	3.0	4.0	120
2	6.3	3.4	103
3	2.0	1.3	142
Точка *p*	3.0	3.0	

Расстояния между скважинами и точкой *p*

Скважина	1	2	3	*p*
1	0	3.35	2.88	1.00
2		0	4.79	3.32
3			0	1.97

Полудисперсии для расстояний между скважинами и точкой *p*

Скважина	1	2	3	p
1	0	13.42	11.52	4.00
2		0	19.14	13.30
3			0	7.89

Уравнения, которые должны быть решены для определения весов, в этом примере имеют вид

$$W_1(0) + W_2(12.2) + W_3(11.5) + \lambda = 4.0 ,$$
$$W_1(12.2) + W_2(0) + W_3(18.1) + \lambda = 12.1 ,$$
$$W_1(11.5) + W_2(18.1) + W_3(0) + \lambda = 7.9 ,$$
$$W_1 + W_2 + W_3 + \lambda = 1$$

или в матричной форме

$$\begin{bmatrix} 0 & 12.2 & 11.5 & 1 \\ 12.2 & 0 & 18.1 & 1 \\ 11.5 & 18.1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 4.0 \\ 12.1 \\ 7.9 \\ 1.0 \end{bmatrix} .$$

В целях упрощения вычислений целесообразно поменять порядок уравнений для того, чтобы избежать нулей на диагонали. Обратная матрица есть

$$\begin{bmatrix} -0.0680 & 0.0326 & 0.0354 & 0.1932 \\ 0.0326 & -0.0433 & 0.0106 & 0.4072 \\ 0.0354 & 0.0106 & -0.0461 & 0.3995 \\ 0.1932 & 0.4072 & 0.3995 & -9.5851 \end{bmatrix} .$$

Теперь можно найти неизвестные веса умножением справа на транспонированную матрицу вектора правой части, состоящего из полудисперсий. В результате получим

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0.5954 \\ 0.0975 \\ 0.3071 \\ -0.7298 \end{bmatrix}.$$

Оценка уровня воды в точке p находится подстановкой подходящих весов в линейное уравнение:

$$\hat{Y}_p = 0.5954(120) + 0.975(103) + 0.3071(142) = 125.1\text{м}\ .$$

Аналогично находится и ошибка дисперсии во взвешенной сумме полувариограмм для расстояний от контрольных точек до оцениваемой точки. В матричных обозначениях $s_\varepsilon^2 = [W]^T[B]$:

$$s_\varepsilon^2 = 0.5954(4) + 0.0975(12.1) + 0.3071(7.9) - 0.7298(1) = 5.25\text{м}^2\ .$$

Стандартная ошибка оценки есть просто квадратный корень из оценки дисперсии или $s_\varepsilon = \sqrt{5.25} = 2.3$ м. Если мы предположим, что ошибки оценивания распределены нормально относительно истинного среднего значения, то мы можем использовать стандартную ошибку для определения доверительного интервала этой оценки. Вероятность того, что истинный уровень воды в точке p находится в пределах одной стандартной ошибки выше или ниже оцениваемого значения, равна 68%, а вероятность того, что истинный уровень лежит в пределах двух стандартных ошибок, равна 95%. Иными словами, уровень воды в точке должен быть $Y_p = 125.1 \pm 4.6$ м с вероятностью 95%. В каждой точке этой карты мы можем оценить уровень воды и можем также определить стандартные ошибки этих оценок. Из них мы можем построить две карты; первая основана на самих оценках и является наилучшим образом предсказанной конфигурацией картируемых переменных, вторая – это ошибка карты, показывающая доверительную обертывающую поверхность, которая окружает оцениваемую поверхность; она выражает относительную надежность первого отображения. На площадях слабого контроля ошибка карты может принимать большие значения, показывая, что оцениваемый параметр подвергается большой изменчивости. На площадях слабого контроля ошибка карты будет показывать низкие значения, и в самих контрольных точках ошибка оценки будет равна нулю.

Система уравнений, используемая для нахождения весов кригинга, должна решаться для каждой оцениваемой точки до тех пор, пока пробы расположены по регулярной схеме так, что расстояния между точками остаются одинаковыми. Если мы пожелаем оценить уровень воды в точке q,

то необходимо рассмотреть расстояния между q и тремя наблюдаемыми скважинами. Эти расстояния и соответствующие полудисперсии следующие:

$$\begin{matrix} 1 \\ 2 \\ 3 \end{matrix}\begin{bmatrix} 2.4 \\ 1.6 \\ 3.0 \end{bmatrix}; \qquad \begin{bmatrix} 9.6 \\ 6.2 \\ 12.0 \end{bmatrix}.$$

Так как расположение наблюдаемых скважин остается тем же самым, то все расстояния между ними одинаковы и левая часть системы совместных уравнений неизменна. Обратная матрица тоже не изменяется. Поэтому, умножив ее на новый вектор полудисперсий, мы получим веса для оценки уровня воды в точке q. Новое множество весов таково:

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0.1676 \\ 0.5796 \\ 0.2528 \\ -0.3711 \end{bmatrix}.$$

Оценим уровень воды $\hat{Y}_p$ и дисперсию s_ε^2:

$$\hat{Y}_p = 0.1676(120) + 0.5796(103) + 0.2528(142) = 115.7\text{м};$$

$$s_\varepsilon^2 = 0.1676(9.6) + 0.5796(6.3) + 0.2528(12.0) - 0.3711(1) = 7.91\text{м}^2.$$

Стандартная ошибка оценки в точке $s_\varepsilon = \sqrt{7.91} = 2.8$ м, поэтому уровень воды в этой новой точке может быть выражен в виде $Y_p = 115.7 \pm 5.6$ м с вероятностью 95%. Поверхность подземных вод в точке q ниже, чем в точке p, и стандартная ошибка, больше, что отражает большое общее расстояние до контрольных скважин.

Постановка задачи: с использованием результатов предыдущего раздела продолжить анализ регионализованной переменной

3. На основе нескольких значений из набора данных провести интерполяцию данных используя точечный кригинг. Помимо произвольных точек интересно проинтерполировать данные в точках для которых измеренное значение переменной известно, и сравнить вычисленное и измеренное значение. В предположении нормальности распределения ошибки построить 95% доверительный интервал для интерполированного значения.

4. Рассмотреть варианты применения других методов не использующих упрощающих предположений о поведении регионализованной переменной, в частности универсальный кригинг. Предположить анизотропность

распределения переменной.

Дополнительно:

5. Рассмотреть варианты применения к данным методов работающих с перешкалированием исходных данных: индикаторный кригинг.