Федеральное агентство по образованию

Санкт-Петербургский государственный горный институт им. Г.В.Плеханова (технический университет)

Кафедра минералогии, кристаллографии и петрографии

МАТЕМАТИЧЕСКИЕ МЕТОДЫ МОДЕЛИРОВАНИЯ В ГЕОЛОГИИ

Методические указания к лабораторным работам

САНКТ-ПЕТЕРБУРГ
2005

УДК 519.2 (075.83)

МАТЕМАТИЧЕСКИЕ МЕТОДЫ МОДЕЛИРОВАНИЯ В ГЕОЛОГИИ: Методические указания к лабораторным работам / Санкт-Петербургский горный ин-т. Сост. *Ю.Л.Гульбин*. СПб, 2005. 46 с.

Изложены методы вариационной статистики и геостатистики, предназначенные для моделирования сложных геологических объектов. Приведено краткое руководство по работе с программами Statistica и Surfer.

Предназначены для студентов специальности 130306 «Прикладная геохимия, петрология, минералогия».

Ил.11. Библиогр.: 8 назв.

Научный редактор проф. *Ю.Б.Марин*

© Санкт-Петербургский горный институт им. Г.В.Плеханова, 2005 г.

ВВЕДЕНИЕ

Важной составной частью любого геологического исследования является вывод количественных закономерностей, характеризующих изучаемый объект земной коры. Математические модели, в рамках которых анализируют эти закономерности, можно условно разделить на две группы: детерминированные и статистические. Детерминированные модели в подавляющем большинстве являются физико-химическими. В их основе лежат аналитические выражения, описывающие сравнительно простые явления, доступные для исследования в лаборатории. Многие природные явления настолько сложны, что не поддаются физико-химическому моделированию. Для их описания используют статистические модели, позволяющие оценивать численные значения геологических свойств, выявлять корреляции между признаками, сравнивать объекты друг с другом и т.д. Создание подобных моделей осуществляется с помощью методов вариационной статистики и геостатистики.

Знакомству с этими методами посвящены лабораторные занятия по курсу «Математические методы моделирования в геологии». В ходе выполнения лабораторных работ студенты закрепляют теоретические знания, полученные на лекциях, учатся самостоятельно выполнять расчеты, необходимые для построения моделей, получают практические навыки использования компьютерных программ для решения поставленных задач.

Лабораторные занятия проводятся в компьютерных классах. Необходимым условием выполнения лабораторных работ является установка на персональных компьютерах пакетов современных прикладных программ в области статистического анализа данных. К их числу относятся популярные программы Excel компании Microsoft (www.microsoft.com), Statistica компании StatSoft (www.ststsoft.com) и Surfer компании Golden Software (www.goldensoftware.com). В методических указаниях содержится краткое руководство по работе с последними версиями этих программ (версией 2002 программы Excel, версией 6.0 программы Statistica и версией 8.0 программы Surfer).

1. МОДЕЛИРОВАНИЕ СЛУЧАЙНОЙ ПЕРЕМЕННОЙ

1.1. НОРМАЛЬНАЯ МОДЕЛЬ

Базовым понятием вариационной статистики выступает понятие «случайной переменной». Одномерной случайной переменной X называют физическую величину (мощность осадочного слоя, содержание химического элемента, количество знаков золота в шлиховой пробе), принимающую различные (заранее не известные) значения в серии независимых испытаний.

Случайная переменная считается заданной, если известны все ее возможные значения $x_1, x_2, ..., x_N$ и соответствующие им вероятности $p_1, p_2, ..., p_N$ ($0 \le p_i \le 1$, $\sum_{i=1}^{N} p_i = 1$). Математическое выражение, связывающее x и p, носит название закона или функции распределения вероятностей случайной переменной. По определению в этом качестве рассматривают функцию $F(x)$, определяющую вероятность того, что случайная переменная X в результате испытания примет значение, меньшее x

$$F(x) = P(X < x).$$

Данная функция является неубывающей. Ее значения принадлежат отрезку [0,1]. Графиком служит ступенчатая (дискретная переменная) или гладкая (непрерывная переменная) кривая. Первая производная функции распределения непрерывной случайной переменной носит название плотности распределения (плотности вероятностей) $f(x) = F'(x)$. Графическим выражением этой зависимости является кривая плотности.

На практике при описании случайной переменной вероятности заменяют абсолютными $n_1, n_2, ..., n_N$ или относительными $w_1, w_2, ..., w_N$ ($w_i = n_i / N$) частотами наблюденных значений признака (совокупность которых образует выборку объема N). Для непрерывной случайной переменной подобное распределение задается в виде последовательности равных

интервалов числовой оси и частот попадания наблюденных значений переменной в эти интервалы. Математическое выражение, связывающее x и w, называется функцией распределения выборки. Для каждого значения x эта функция определяет относительную частоту события $X < x$

$$F^*(x) = W(X < x).$$

В отличие от эмпирической функции распределения выборки функцию распределения генеральной совокупности называют теоретической функцией распределения. Как правило, вид ее неизвестен, но может быть оценен по выборочным данным. Процедура оценивания сводится к проверке статистических гипотез о соответствии теоретического распределения одному из модельных распределений, задаваемых априори. В этом качестве могут фигурировать различные распределения (биномиальное, Пуассона, Вейбулла, геометрическое), но чаще всего используется так называемое нормальное или Гауссово распределение, плотность вероятности которого задается формулой

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Графиком записанного выражения служит колоколообразная кривая, положение и форма которой всецело определяются параметрами μ (математическим ожиданием) и σ (стандартным отклонением).

Наиболее строгий подход к проверке соответствия теоретического распределения нормальной модели основан на использовании критериев согласия (χ^2 Пирсона, Колмогорова, Смирнова и др.). В его основе лежит сравнение эмпирических (наблюденных) частот и тех частот, которые бы характеризовали распределение значений переменной при условии его подчинения нормальному закону. В случае критерия Пирсона расчет ведется по формуле $\chi^2 = \sum \frac{(n_i^* - n_i)^2}{n_i}$, где n_i^* – эмпирическая частота, n_i – теоретическая частота. Вывод о соответствии изучаемого

распределения нормальной модели делается при условии, что рассчитанное значение статистики не превышает критического значения распределения хи-квадрат (для заданного уровня значимости и числа степеней свободы)

$$\chi^2 < \chi^2(\alpha; k),\ k = m - 3,$$

где m – число интервалов группирования выборки.

Альтернативой критерию χ^2 выступает критерий Колмогорова. Статистикой критерия служит максимум модуля разности между значениями теоретической и эмпирической функций распределения $d = \max\left|F(x_i) - F^*(x_i)\right|$. Если рассчитанное значение статистики меньше критического

$$d < d(\alpha; N) = \frac{\lambda}{\sqrt{N}},$$

где λ – процентная точка распределения Колмогорова, изучаемое распределение может быть описано в рамках нормальной модели.

Для использования критериев согласия объем выборки быть достаточно велик, во всяком случае не менее 50. Каждый частичный интервал должен содержать не менее 5-8 вариант, малочисленные группы следует объединять, суммируя частоты. В случае малых выборок (30-50 наблюдений) для проверки гипотезы о нормальном распределении генеральной совокупности вместо критериев согласия удобно использовать выборочные оценки асимметрии и эксцесса

$$A^* = \frac{\sum (x_i - \bar{x})^3}{N} \cdot \frac{1}{s^3},$$

$$E^* = \frac{\sum (x_i - \bar{x})^4}{N} \cdot \frac{1}{s^4} - 3.$$

где s – выборочная оценка стандартного отклонения. Будучи статистическими параметрами, асимметрия и эксцесс (в свою очередь) обладают свойствами случайных величин. При условии

соответствия теоретического распределения нормальному закону, их значения распределены (асимптотически) нормально с математическими ожиданиями, равными 0

$$E(A) = 0\,,\ E(E) = 0$$

и средними квадратичными отклонениями, приближенно равными

$$\sigma_A \cong \sqrt{\frac{6}{N}}\ ,\ \ \sigma_E \cong \sqrt{\frac{24}{N}}\ .$$

Если выборочные оценки A и E выходят за пределы 3σ

$$\left|A^*\right| > 3\sigma_A\,,\ \ \left|E^*\right| > 3\sigma_E\,,$$

нормальная модель отвергается.

Третий способ проверки гипотезы о нормальном распределении состоит в использовании так называемых графиков на вероятностной бумаге. Для построения этих графиков: 1) значения случайной переменной записывают в виде (ранжированного по возрастанию) вариационного ряда $x_1, x_2, ..., x_N$, 2) для каждого значения переменной рассчитывают значения эмпирической функции распределения по формуле $F_i = \frac{i-1}{N}$, 3) пользуясь рассчитанными значениями F_i, находят значения квантиля нормального распределения $y_i = Z_i^{-1}(Z_i = F_i)$. График строят в координатах $x - y$ (исходные значения переменной – ожидаемые значения переменной). Если распределение изучаемой переменной приближенно описывается нормальной моделью, все фигуративные точки такого графика располагаются на одной прямой (рис.1). Если распределение характеризуется положительной асимметрией – точки располагаются вдоль дуги, обращенной выпуклостью вверх (рис.2).

Задание. *Проверить гипотезу о соответствии теоретического распределения нормальной модели* и *рассчитать числовые характеристики случайной переменной.*

Порядок решения. Запускаем программу Statistica и редактируем таблицу ввода данных. С помощью диалогового окна спецификации переменных (которое отрывается, если дважды щелкнуть левой кнопкой мыши на любом из заголовков столбцов Var1, Var2…) указываем имена переменных. С помощью кнопок панели инструментов **Vars** (**Переменные**) и **Cases** (**Случаи**) выполняем различные операции со столбцами и строками таблицы.

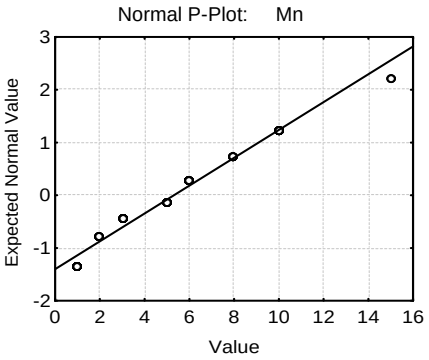

Рис.1

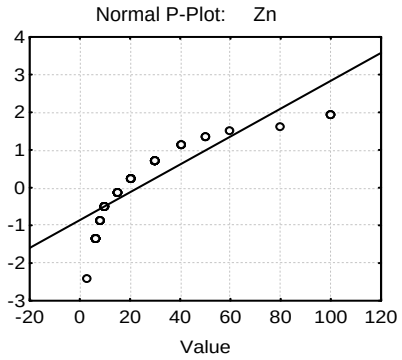

Рис.2

Копируем исходные данные (например, таблицу содержаний химических элементов в гранитах, подготовленную в формате электронных таблиц Exel) в таблицу ввода данных через буфер Windows (в ходе копирования число строк и столбцов задается автоматически).

Приступаем к расчетам. С помощью выпадающего меню **Statistics/Basic Statistics** (**Статистики/Основные статистики**) открываем одноименный модуль. Выбираем строку **Descriptive Statistics** (**Описательные статистики**). В одноименном диалоговом окне с помощью кнопки **Variables** отмечаем имена интересующих нас переменных.

Открываем закладку **Normality** и с помощью кнопки **Histograms** строим гистограммы частот статистических распределений признаков. По виду гистограмм судим о степени их скошенности (асимметрии). В зарамочном оформлении гистограмм

считываем значение статистики Колмогорова-Смирнова d, достигнутые уровни значимости p (при проверке простой гипотезы) и Lillieforce p (при проверке сложной гипотезы)[1]. Если значение статистики велико ($p < 0.05$), отвергаем нулевую гипотезу и считаем нормальную модель непригодной для описания изучаемого распределения.

Открываем закладку **Prob&Scatterplot** и с помощью кнопки **Normal Probability Plot** строим графики на вероятностной бумаге. В ходе просмотра графиков отмечаем случаи умеренного отклонения от нормального закона и случаи с резко выраженной положительной асимметрией. В отдельную группу выделяем графики с аномальными (резко выделяющимися) значениями переменных.

Открываем закладку **Advanced** и отмечаем требуемые для вывода описательные статистики: **Valid N** (**Объем выборки**), **Mean** (**Среднее**), **Median** (**Медиана**), **Mode** (**Мода**), **Standard Deviation** (**Стандартное отклонение**), **Variance** (**Дисперсия**), **Std. err. of mean** (**Стандартная ошибка среднего**), **Conf. limits for mean** (**Доверительный интервал среднего**), **Scewness** (**Асимметрия**), **Std. err. Scewness** (**Стандартное отклонение асимметрии**), **Kurtosis** (**Эксцесс**), **Std. err. Kurtosis** (**Стандартное отклонение эксцесса**). С помощью кнопки **Summary** выводим на экран таблицу числовых характеристик исследуемых переменных. Используем выборочные оценки E^*, A^*, σ_A, σ_E для проверки нулевой гипотезы о соответствии изучаемого распределения нормальной модели.

С учетом результатов проведенного статистического анализа выполняем преобразования исходных данных (удаляем аномальные значения, логарифмируем переменные с положительной асимметрией). Для логарифмирования значений признака: создаем новую переменную, открываем диалоговое окно ее спецификации, указываем имя переменной, затем в нижней части окна, в поле **Long**

[1] Простой гипотезой называют предположение о виде распределения, сделанное при условии, что известны параметры этого распределения. Сложной гипотезой – предположение о виде распределения, параметры которого оцениваются по той же самой выборке, по которой проверяют гипотезу о согласии.

name записываем формулу преобразования. Синтаксис записи: знак равенства, символ используемой функции (например, Log10), имя переменной в круглых скобках. Нажимаем кнопку **OK**. На вопрос программы **Expression OK. Recalculate the variable now?** Отвечаем: **да**.

Еще раз проверяем соответствие теоретических функций распределения преобразованных переменных нормальной модели (например, с помощью графиков на вероятностной бумаге) и рассчитываем их числовые характеристики. Сохраняем результаты расчетов.

1.2. РЕГРЕССИОННЫЕ МОДЕЛИ

1.2.1. Двумерная линейная регрессия

При решении разнообразных геологических задач часто возникает необходимость совместного рассмотрения двух случайных переменных. При этом исследователей в первую очередь интересуют статистические зависимости между изучаемыми признаками. Подобные зависимости возникают, когда фиксированному значению переменной $X = x_j$ соответствует не одно, но несколько значений переменной $Y = y_1, y_2, ..., y_k$, каждое из которых характеризуется условной вероятностью $p'(y_i \left| X = x_j \right.)$. Последний ряд задает условное распределение вероятностей, центр которого соответствует условному математическому ожиданию случайной переменной Y при фиксированном значении переменной X

$$E(Y \left| X = x_i \right.) = \sum y_i p'(y_i \left| X = x_i \right.).$$

Условное математическое ожидание является функцией от X

$$E(Y \left| X \right.) = f(X).$$

В общем случае эта функция называется функцией регрессии Y на X. Если мы имеем дело с выборочными данными, вид ее заранее

неизвестен, но может быть определен в ходе статистического исследования. Традиционный метод решения поставленной задачи состоит в моделировании подобных функций линейными или полиномиальными уравнениями.

Построение линейной регрессионной модели осуществляется в два этапа. На первом этапе проверяется гипотеза о наличии линейной связи между переменными и оценивается сила этой связи. На втором этапе находится выборочное уравнение регрессии (конкретизирующее вид искомой зависимости).

Первая задача на качественном уровне решается путем визуального анализа диаграмм рассеяния и выявления трендов в расположении точек. Количественным критерием для оценки наличия и силы линейной связи служит статистика

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{N-2}\,,$$

предназначенная для проверки нулевой гипотезы о равенстве нулю выборочного коэффициента корреляции r при условии нормального распределения двумерной генеральной совокупности (X,Y). Если рассчитанное значение статистики превышает критическое значение распределения Стьюдента для заданных уровня значимости и числа степеней свободы

$$t > t(\alpha;k)\,, \quad k = N-2\,,$$

нулевая гипотеза отвергается и линейная связь признается существующей.

В предположении, что линейная модель не противоречит наблюдаемым закономерностям, она может быть описана с помощью выборочного уравнения среднеквадратической регрессии

$$Y^* = b_0 + b_1 X\,,$$

где Y^* – регрессионная оценка зависимой переменной, b_0, b_1 – выборочные коэффициенты регрессии. По определению данное уравнение учитывает только часть изменчивости зависимой переменной. Для характеристики неучтенной изменчивости находят

остаточную сумму квадратов $SS_{ERROR} = \sum \left(y_i^* - y_i \right)^2$, которая вычисляется как разность $SS_E = SS_T - SS_R$, где $SS_{TOTAL} = \sum \left(y_i - \bar{y} \right)^2$ – общая сумма квадратов, $SS_{REGR} = \sum \left(y_i^* - \bar{y} \right)^2$ – регрессионная сумма квадратов.

С помощью остаточной суммы квадратов:

1) проверяют гипотезу о значимости линейной модели. С этой целью вычисляют статистику

$$F = \frac{SS_R}{SS_E} \cdot \frac{k_2}{k_1}.$$

Если рассчитанное значение статистики превышает критическое значение распределения Фишера-Снедекора

$$F > F(\alpha; k_1; k_2), \quad k_1 = 1, \quad k_2 = N - 2,$$

нулевую гипотезу отвергают и линейную модель считают значимой;

2) рассчитывают среднеквадратическую ошибку регрессионной модели, равную корню квадратному из остаточной дисперсии

$$s_E^2 = \frac{SS_E}{N - 2};$$

3) рассчитывают коэффициент детерминации

$$R^2 = \frac{SS_R}{SS_T}.$$

Коэффициент детерминации служит показателем качества регрессионной модели. Чем ближе R^2 к единице, тем лучше линейная модель объясняет экспериментальные данные.

Задание. *Рассчитать уравнение линейной регрессии Y на X, оценить ее значимость и качество.*

Порядок решения. Открываем файл исходных данных (например, результаты химического анализа рудных проб на свинец и золото) в программе Exel.

Проверяем гипотезу о соответствии выборочного распределения переменных нормальной модели (например, с помощью графиков на вероятностной бумаге в программе Statistica). При необходимости логарифмируем данные.

Строим диаграмму рассеяния. В программе Exel открываем **Мастер диаграмм** и выбираем **тип диаграммы: Точечная**. Щелкаем правой кнопкой мыши на любой точке графика, в открывшемся контекстном меню выбираем строчку **Добавить линию тренда**. В диалоговом окне **Линия тренда** на закладке **Тип** выбираем окно **Линейная**, на закладке **Параметры** помечаем опцию **показывать уравнение на диаграмме** (рис.3).

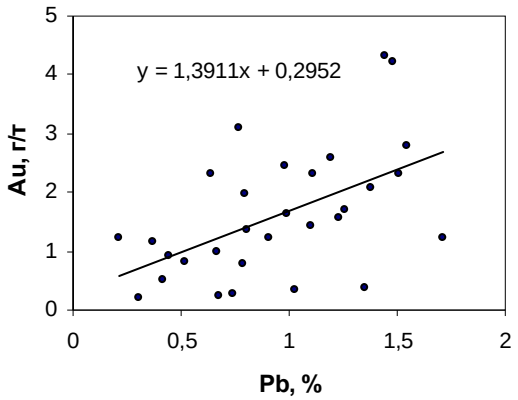

Рис.3

Рассчитываем уравнение регрессии. С помощью выпадающего меню **Сервис/Анализ данных** открываем одноименный модуль. Выбираем строку **Регрессия**. В диалоговом окне задаем входные интервалы переменных и нажимаем на кнопку **ОК**. Знакомимся с выходными данными. В таблице **Регрессионная статистика** считываем значение коэффициента корреляции (строка *Множественный R*), коэффициента детерминации (*R-квадрат*), среднеквадратической ошибки регрессии (*Стандартная ошибка*). Из таблицы **Дисперсионный анализ** берем значения регрессионной, остаточной и общей суммы квадратов (столбец *SS*),

остаточной дисперсии (столбец *MS*, строка *Остаток*), F - статистики (столбец *F*), рассчитанного уровня значимости (столбец *Значимость F*). Если рассчитанный уровень значимости меньше 0.05, делаем вывод, что регрессионная модель значима. Ниже считываем значения выборочных коэффициентов регрессии (столбец *Коэффициенты*), их среднеквадратических ошибок (*Стандартная ошибка*), нижней и верхней границ доверительных интервалов (*Нижние и Верхние* 95%).

1.2.2. Полиномиальная и множественная регрессия

Иногда визуальное изучение диаграмм рассеяния показывает наличие криволинейной связи между переменными (или подобная связь вытекает из содержательного анализа задачи). Для ее количественной характеристики используют регрессионные модели, основанные на нелинейных аппроксимирующих функциях. Среди подобных функций (степенных, логарифмических, показательных и др.) важную роль играют полиномиальные функции вида

$$Y^* = b_0 + b_1 X + b_2 X^2 + ... + b_k X^k = b_0 + \sum b_i X^i ,$$

где k – порядок полинома.

Расширением двумерных регрессионных моделей являются модели множественной регрессии, в которых переменная Y рассматривается как линейная функция не одной, но нескольких независимых переменных $X_1, X_2, ..., X_m$

$$Y^* = b_0 + b_1 X_1 + b_2 X_2 + ... + b_m X_m = b_0 + \sum b_j X_j ,$$

где Y^* – регрессионная оценка зависимой переменной (результирующий признак), $b_0, b_1, b_2, ..., b_m$ – частные коэффициенты или веса регрессионной модели. Они носят название частных, поскольку каждый их них характеризует влияние одной независимой переменной на результирующий признак при условии, что значения всех остальных переменных фиксированы на среднем уровне. Для вычисления весов исходные данные предварительно

стандартизируют (т.е. подвергают преобразованию вида $\hat{X} = (X - \bar{x}) / s_X$ с целью получения вероятностных распределений с единичной дисперсией и нулевым средним). В результате стандартизации регрессионное уравнение принимает следующий вид

$$Y^* = \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2 + ... + \beta_m \hat{X}_m = \sum \beta_j \hat{X}_j ,$$

где $\beta_1, \beta_2, .., \beta_m$ – стандартизированные частные коэффициенты регрессии. Соотношение между b и β:

$$b_j = \beta_j \frac{s_Y}{s_{X_j}} ,$$

где s_Y – оценка среднеквадратичного отклонения переменной Y, s_{X_j} – оценка среднеквадратичного отклонения переменной X_j. По аналогии с двумерной регрессией , «беты» можно рассматривать как частные коэффициенты корреляции между независимыми и зависимой переменной[2].

Среди независимых переменных, учитываемых уравнением множественной регрессии, могут встречаться признаки, влиянием которых можно пренебречь. Для их выявления проверяется статистическая гипотеза о равенстве нулю соответствующего регрессионного коэффициента. В ходе проверки гипотезы рассчитывается статистика

$$t_{b_j} = b_j / s_{b_j} ,$$

где s_{b_j} – среднеквадратичная ошибка коэффициента b_j, имеющая распределение Стьюдента с $N - 1$ степенями свободы. Если рассчитанное значение статистики превышает критическое

[2] Соотношение, связывающее угловой коэффициент двумерной регрессии и коэффициент корреляции: $b_1 = r(s_Y / s_X)$.

$t_{b_j} > t(\alpha; k)$, нулевая гипотеза отвергается и коэффициент регрессии считается значимым.

Если выясняется, что изучаемый частный коэффициент регрессионной модели незначим, то соответствующая ему переменная исключается из рассмотрения, после чего рассчитывается новая регрессионная модель.

Задание. *Построить модели полиномиальной и множественной регрессии, описывающие статистические связи между переменными*, *оценить их значимость и качество.*

Порядок решения. Открываем файл исходных данных (например, результаты измерений содержаний железа и микротвердости сфалерита) в программе Exel.

Строим диаграмму рассеяния. С помощью правой кнопки мыши открываем диалоговое окно **Линия тренда**, на закладке **Тип** выбираем окно **Полиномиальная**, на закладке **Параметры** помечаем опцию **показывать уравнение на диаграмме**.

Копируем данные в программу Statistica. Редактируем таблицу ввода данных. С помощью выпадающего меню **Statistics/Advanced Linear-Nonlinear Models /General Regression Models** (**Статистики/Линейные и нелинейные модели/Главные регрессионные модели**) открываем одноименный модуль. Выбираем строку **Polynomial Regression**. В открывшемся диалоговом окне с помощью кнопки **Variables** указываем имена переменных: в ответ на приглашение **Select dependent variables and continuous predictors** в левом поле отмечаем зависимую переменную, в правом – независимую переменную. Нажимаем на кнопку **ОК**. Знакомимся с результатами расчетов. С помощью кнопки **Coefficients** выводим на экран таблицу выборочных коэффициентов регрессии, с помощью кнопки **All Effects** – таблицу дисперсионного анализа. В обеих таблицах значимые статистики подсвечиваются красным.

Открываем следующий файл исходных данных (например, результаты опробования нескольких горизонтов пегматитовой жилы на редкие щелочи) в программе Exel. Копируем данные в программу Statistica. Редактируем таблицу ввода данных. Проверяем гипотезу о соответствии выборочного распределения переменных нормальной

модели. Открываем модуль **General Regression Models**, выбираем строку **Multiple Regression** (**Множественная регрессия**). С помощью кнопки **Variables** указываем имя зависимой переменной (dependent variable) и имена независимых переменных (continuous predictors). Нажимаем на кнопку **ОК** и знакомимся с результатами расчетов (с помощью кнопок **Coefficients** и **All Effects**).

Выбираем значимые регрессионные коэффициенты (коэффициенты, для которых значение $t-$статистики велико, а рассчитанный уровень значимости $p<0.05$). Сокращаем число признаков (за счет переменных с незначимыми коэффициентами) и строим новую регрессионную модель.

1.3. КЛАССИФИКАЦИОННЫЕ МОДЕЛИ

1.3.1. Метод линейной дискриминантной функции

Большое значение при геологических исследованиях имеют задачи классификации. Можно выделить два рода таких задач.

Первый род включает задачи, связанные с поиском решающего правила для отнесения произвольного объекта совокупности к одной из групп, которые выделяются заранее. Примером могут служить две группы сланцев, о которых доподлинно известно, что они сформировались в условиях морского и пресноводного бассейнов. Подобные задачи получили название задач разделения или дискриминации. Для их решения был предложен метод линейной дискриминантной функции. Подобная функция преобразует исходное множество измерений, характеризующих изучаемый объект (например, данные химического анализа единичного образца), в дискриминантное число (индекс), величина которого позволяет судить о принадлежности объекта к той или иной (заданной) группе. Функция подбирается таким образом, чтобы при разделении многомерной нормально распределенной совокупности на группы достигались максимальная однородность внутри групп и минимальная между группами.

Поставленным условиям удовлетворяет функция вида

$$D = a_1 X_1 + a_2 X_2 + ...a_m X_m = \sum a_j X_j \ ,$$

коэффициенты которой a_j подбираются таким образом, чтобы при делении совокупности на группы отношение межгрупповой дисперсии к внутригрупповой дисперсии было бы максимальным

$$F = \frac{s_R^2}{s_E^2} = \max \ .$$

Можно показать, что для двух групп многомерных данных $F-$статистика равна отношению

$$F = \frac{\overline{D}_U - \overline{D}_V}{s_D^2} \ ,$$

где $\overline{D}_U , \overline{D}_V -$ выборочные средние значения дискриминантной функции для каждой группы, $s_D^2 -$ средневзвешенная дисперсия дискриминанта объединенной выборки. На дискриминантной оси точки $\overline{D}_U$ и $\overline{D}_V$ соответствуют центрам двух групп, вокруг которых рассеяны отдельные наблюдения. Расстояние между центрами

$$d^2 = \left| \overline{D}_U - \overline{D}_V \right|$$

получило название расстояния Махалонобиса. Можно показать, что оно численно равно квадрату евклидового расстояния между многомерными средними 2-х групп в $m-$мерном пространстве признаков, стандартизированному относительно дисперсии объединенной выборки. Чем больше d^2, тем увереннее можно провести разделение объектов, тем более эффективна дискриминационная модель. Степень эффективности оценивают по априорной вероятности ошибочной классификации

$$p = 1 - Z(d^2) \ ,$$

где $Z(d^2)$ – функция нормального распределения. Помимо априорной ошибки вычисляется эмпирическая ошибка модели. Она определяется как выраженная в % доля наблюдений обучающей выборки, ошибочно классифицированных с помощью дискриминантной функции.

Задание. *Рассчитать уравнение линейной дискриминантной функции для разделения двух групп объектов, оценить априорную и эмпирическую ошибки классификации*.

Порядок решения. Открываем файл исходных данных (например, таблицу содержаний петрогенных компонентов в рудоносных и безрудных аляскитах) в программе Exel. Копируем данные в программу Statistica. Проверяем гипотезу о соответствии выборочного распределения переменных нормальной модели.

Задаем группирующую переменную (значения которой «сообщают» программе о принадлежности индивидуальной пробы к первой или второй группе). С помощью модуля **Basic statistics/t-test, independent, by group** (**Статистики/t-тест для сравнения двух независимых выборок**) проверяем гипотезу о значимости различий одномерных средних.

Приступаем к расчету линейной дискриминантной функции. С помощью выпадающего меню **Statistics/Multivariate Exploratory Techniques/Discriminant Analysis** (**Статистики/Методы многомерной статистики/Дискриминантный анализ**) открываем одноименное диалоговое окно. С помощью кнопки **Variables** указываем имена группирующей переменной (grouping var.) и независимых переменных (independent vars.). С помощью кнопки **Codes for grouping variable** специфицируем две группы объектов. Нажимаем на кнопку **OK** и знакомимся с результатами расчетов.

В верхней части открывшегося диалогового окна считываем значение F – статистики и рассчитанного уровня значимости p. Если p<0.05, считаем модель значимой. Открываем таблицу **Classification Functions** и копируем ее в программу Exel. Вычитаем из левого столбца (A) правый столбец (B), результирующий столбец используем для построения дискриминантной функции (первые строки) и вычисления дискриминантного индекса (последняя строка). Пользуясь рассчитанной функцией, находим значение

дискриминанта для каждой пробы, затем рассчитываем среднее значение дискриминанта для каждой из двух групп и вычисляем расстояние Махалонобиса. Для наглядной демонстрации полученного результата копируем столбец рассчитанных значений функции D обратно в программу Statistica и строим гистограмму частот распределения (рис.4).

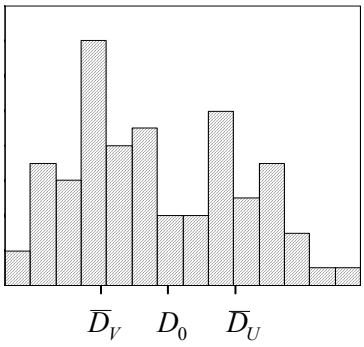

Рис.4

Оцениваем качество дискриминантной модели. Открываем **Вероятностный калькулятор** (**Statistics/Probability Calculator /Distribution**). Помечаем строку **Z(Normal)**; в окно **X** вписываем рассчитанное значение d^2, нажимаем кнопку **Compute** (**Вычислить**), в окне **p** считываем значение $Z(d^2)$. Вычисляем априорную ошибку классификации. Открываем таблицу **Classification Matrix**, подсчитываем общее число неправильно классифицированных объектов и вычисляем эмпирическую ошибку классификации.

1.3.2. Кластерный анализ

Второй род включает задачи, связанные с разделением совокупности геологических объектов на однородные группы, число которых заранее не определено. Классическим примером подобных задач является задача классификации ископаемых останков по

набору морфометрических признаков. В результате подобной классификации виды организмов объединяются в роды, роды в семейства, семейства в отряды, отряды в классы и т.д. Объединение объектов в группы на каждом уровне описанной иерархии производится либо путем субъективной оценки, либо с помощью количественных критериев. Последний подход получил название «численной таксономии».

Среди формальных процедур численной таксономии ведущее место занимает кластерный анализ. Эти термином называют процедуру оптимального разбиения исходного множества на несколько соподчиненных подмножеств или кластеров. В основу процедуры кладется вычисление степени сходства (близости) двух произвольных объектов матрицы наблюдений. В качестве меры сходства выступает евклидово расстояние между объектами в m – мерном пространстве признаков

$$d_{ik} = \sqrt{\sum^{m} (x_{ij} - x_{kj})^2} ,$$

где x_{ij} , x_{kj} – значения j – й переменной в i – м и k – м объектах.

Множество мер сходства представляют в виде симметричной матрицы порядка n, где n – число объектов. Из этой матрицы выбирают недиагональный элемент, обладающий наименьшим значением. Пару объектов, соответствующую этому элементу, называют минимальным кластером. К этому двухчленному кластеру добавляют новый объект, расположенный к любому из объектов кластера ближе, чем все другие объекты. В результате получают трехчленный кластер. К нему добавляют следующий ближайший объект и т.д. Описанная процедура носит название кластеризации по методу ближайшего соседа (одиночной связи). Уровни сходства, при которых объединяются наблюдения, используются для построения дендрограммы.

Дендрограмма представляет собой древовидный граф, по оси абсцисс которого располагаются изучаемые объекты, по оси ординат – значения меры сходства, которая используется для кластеризации. Кроме евклидова расстояния в этом качестве могут применяться

расстояние Махалонобиса, манхэттенское расстояние и др. Влияет на вид дендрограммы и выбор процедуры кластеризации. Помимо метода ближайшего соседа для построения кластера используют метод невзвешенного попарного центроидного усреднения (в котором для присоединения нового объекта к кластеру ищется минимальное расстояние между объектом-кандидатом и центром тяжести кластера), невзвешенного попарного арифметического среднего (в котором рассчитывается среднее арифметическое всех расстояний между объектом-кандидатом и объектами, образующими кластер) и т.д.

Существенное влияние на вид дендрограммы может оказать и процедура стандартизации данных. В результате ее использования удается избежать зависимости рассчитанного евклидова расстояния от переменных, имеющих резко повышенные значения или измеренных в других единицах.

Задание. *Используя кластерный анализ, осуществить иерархическую классификацию объектов.*

Порядок решения. Открываем файл исходных данных (например, таблицу средних содержаний петрогенных и редких элементов в различных геохимических типах гранитов) в программе Exel. Копируем данные в программу Statistica.

Приступаем к построению дендрограммы. С помощью выпадающего меню **Statistics/Multivariate Exploratory Techniques/Cluster Analysis** открываем модуль **Clustering Method**. Выбираем строку **Joining** (**Tree clustering**) и нажимаем кнопку **OK**.

В открывшемся диалоговом окне с помощью кнопки **Variables** указываем переменные, с помощью выпадающего меню **Input file** (**Входной файл**) – тип входных данных (**Raw data**), с помощью выпадающего меню **Cluster** – объекты кластеризации (**Cases**). С помощью выпадающего меню **Amalgamation (linkage) rule** выбираем процедуру кластеризации: **Single linkage** (**Одиночной связи**), **Unweighted pair-group centroid average** (**Невзвешенного попарного центроидного усреднения**), **Unweighted pair-group arithmetic averages** (**Невзвешенного попарного арифметического среднего**) и т.д. С помощью выпадающего меню **Distance measure** выбираем меру сходства:

Euclidian distance (**Евклидово расстояние**), **Squared Euclidian distance** (**Квадратичное евклидово расстояние**) и т.д. Нажимаем на кнопку **OK** и знакомимся с результатами расчетов.

С помощью кнопки **Vertical icicle plot** выводим на экран дендрограмму. С помощью кнопки **Cancel** возвращаемся назад, к предыдущему диалоговому окну, задаем другие процедуру кластеризации и меру сходства, строим новую дендрограмму.

Переключаемся на таблицу исходных данных и выполняем процедуру стандартизации. С этой целью: выделяем мышью блок данных, с помощью правой клавиши мыши открываем контекстное меню, выделяем строки **Fill Standardize Block/Standardize Columns** (**Стандартизировать колонки**). Строим новую дендрограмму. Сравниваем полученные результаты.

1.4. ФАКТОРНЫЕ МОДЕЛИ

1.4.1. Метод главных компонент

Общей особенностью геологических объектов является сложность их состава и строения. Большинство из них характеризуются не одним, но множеством признаков, доступных измерению (примером могут служить горные породы, для характеристики которых необходимо измерить содержания ряда петрогенных и редких элементов). Увеличение размерности признакового пространства неизбежно создает трудности для обработки информации и понимания изучаемых явлений.

Одним из наиболее эффективных методов многомерной статистики выступает метод главных компонент. С его помощью строятся факторные модели многомерных совокупностей, позволяющие решать целый комплекс исследовательских задач, включая сокращение размерности признакового пространства, классификацию признаков и наблюдений (объектов), нахождение причин (скрытых переменных), ответственных за геологическую изменчивость.

Основу метода составляет вычисление собственных векторов и собственных значений ковариационной матрицы. Обычно перед

расчетом этой матрицы исходные данные стандартизируются, в результате чего ковариационная матрица превращается в корреляционную. С помощью собственных векторов корреляционной матрицы осуществляется переход от исходных переменных к новым некоррелированным признакам, дисперсии которых максимизированы, т.е. принимают наибольшее значение среди всех линейных комбинаций исходных переменных.

Допустим, что в нашем распоряжении имеется матрица наблюдений или случайный вектор $[X]$ размерности $n \times m$, где n – число объектов, m – число признаков. Будем считать, что составляющие вектор переменные характеризуются нормальным распределением с нулевым математическим ожиданием и единичной дисперсией. Для нахождения корреляционной матрицы $[R]$ умножим матрицу наблюдений на ее транспонированный аналог слева

$$[R] = \frac{1}{n-1}[X]^T \cdot [X].$$

Данную матрицу можно рассматривать как набор координат точек в m – мерном пространстве. Каждому ее столбцу (или строке) соответствует вектор, исходящий из начала координат. Координатными осями служат случайные переменные $X_1, X_2, ..., X_m$, для которых рассчитана корреляционная матрица.

Вычислим диагональную матрицу собственных значений $[\Lambda]^2$ и матрицу собственных векторов $[Y]$ корреляционной матрицы. Последняя определит направление новых осей координат (главных компонент). Для задания масштаба измерения новых переменных найдем произведение матриц собственных векторов и сингулярных значений $[\Lambda]$

$$[W] = [Y] \cdot [\Lambda].$$

Результатом произведения выступит матрица факторных нагрузок $[W]$, с помощью которых исходные переменные проектируются на факторные оси. Произведение матриц наблюдений и факторных

нагрузок позволит найти результирующую матрицу значений статистических факторов

$$[Z] = [X] \cdot [W].$$

Полученная матрица характеризуется той же размерностью, что и матрица наблюдений. Ее столбцы представляют собой стандартизированные переменные, распределение которых соответствует нормальной модели. В отличие от исходных переменных они ортогональны, т.е. не коррелируют между собой. Их вклад в общую изменчивость системы может быть оценен с помощью собственных значений корреляционной матрицы. Каждое из собственных значений имеет смысл дисперсии фактора. Сумма собственных значений равна суммарной дисперсии корреляционной матрицы

$$\sum \lambda_i = \sum \sigma_j^2 ,$$

где λ_i – дисперсия i – го фактора, σ_j^2 – дисперсия j – й переменной.

Дисперсия фактора является мерой для проверки гипотез о значимости главных компонентов с помощью специальных статистических критериев. На практике вместо них используют эмпирические критерии. Согласно одному их них (критерию Кайзера) значимыми признаются факторы с дисперсией, превышающей 1 (т.е. дисперсию исходных стандартизированных переменных).

Дисперсии 2-3 первых факторов (факторов с максимальными весами $D_i = \frac{\lambda_i}{m} \cdot 100$, $\sum D_i = 100\%$), как правило, заметно превышают единицу. Для хорошо обусловленных корреляционных матриц (матриц с высокими коэффициентами корреляции) относительный вес первых факторов достигает 70-90%. На практике это означает, что первые 2-3 фактора несут основную информацию о системе, тогда как влиянием прочих переменных можно пренебречь. Сверхзадача факторного анализа состоит в интерпретации главных статистических компонент и их отождествлении с теми

физическими причинами, которые обусловили общность в поведении исходных переменных.

Интерпретация факторов основывается прежде всего на анализе матрицы факторных нагрузок. Последним часто придается смысл парных коэффициентов корреляции между исходными переменными и факторами. Для оценки значимости факторных нагрузок используют тесты, разработанные для оценки значимости коэффициентов корреляции.

Для демонстрации результатов факторного анализа строят диаграммы в координатах факторных нагрузок (фигуративными точками здесь являются переменные) и координатах значений факторов (фигуративные точки здесь соответствуют наблюдениям). При построении диаграмм следует учитывать, что факторные нагрузки изменяются в пределах от –1 до +1, значения факторов – от –3 до +3. Если значения факторов выходят за эти пределы, построенная факторная модель некорректна.

С помощью факторных диаграмм выявляются комбинации переменных (коррелирующих между собой) и группы наблюдений (близких по свойствам). Такие переменные (наблюдения) характеризуются высокими положительными или отрицательными нагрузками (значениями факторов).

Задание. *Используя метод главных компонент*, *построить факторную модель многомерной совокупности.*

Порядок решения. Открываем файл исходных данных (например, результаты опробования рудного тела на главные и примесные компоненты) в программе Exel. Копируем данные в программу Statistica. Проверяем гипотезу о соответствии выборочного распределения переменных нормальной модели.

Приступаем к построению факторной модели. С помощью выпадающего меню **Statistics/Multivariate Exploratory Techniques/Factor Analysis** открываем одноименное диалоговое окно. С помощью кнопки **Variables** указываем имена переменных. Нажимаем на кнопку **OK.** В следующем диалоговом окне на закладке **Advanced** выбираем метод факторизации (по умолчанию **Principal Component – Метод главных компонент**). В окне **Max. no of factors** указываем максимальное число выводимых факторов

(по умолчанию 2), в окне **Mini. eigenvalue** – минимальное собственное значение фактора (по умолчанию 1.00). Нажимаем на кнопку **OK** и знакомимся с результатами расчетов.

Открываем закладку **Loadings** (Факторные нагрузки) и с помощью кнопки **Summary: Factor loadings** выводим на экран таблицу факторных нагрузок (рис.5). В нижней части таблицы считываем собственные значения (строка **Expl.Var**) и относительные веса (строка **Prp.Totl.**) факторов. Оцениваем значимость факторных нагрузок. Для этого открываем вероятностный калькулятор (**Statistics/Probability Calculator /Correlations**), в окно **N** вписываем число проб (объем выборки), в окно **p** – уровень значимости (0.05), нажимаем кнопку **Compute**, в окне **r** считываем пороговое значение факторной нагрузки.

С помощью кнопки **Plot of loadings, 2D** строим диаграммы факторных нагрузок.

Открываем закладку **Scores** (**Значения факторов**). С помощью кнопки **Factor scores** выводим на экран таблицу значений факторов (рис.6). Копируем таблицу в программу Exel и строим диаграммы значений факторов.

Variable	Factor 1	Factor 2	Factor 3
K	0.862	-0.380	-0.040
Rb	0.811	-0.357	0.083
Sr	0.371	-0.220	-0.870
Ba	0.703	-0.374	-0.548
V	-0.535	0.442	-0.411
Cr	-0.674	-0.654	0.205
Co	-0.753	-0.281	-0.138
Ni	-0.698	-0.620	0.088
Sn	-0.849	-0.465	-0.034
W	-0.871	0.178	-0.209
Be	0.941	-0.089	0.176
Expl.Var	6.776	1.805	1.432
Prp.Totl	0.565	0.150	0.119

Рис.5

Case	Factor 1	Factor 2	Factor 3
1	-0.548	0.677	0.479
2	-0.559	0.502	0.556
3	-0.585	0.531	0.458
4	-0.654	0.496	0.383
6	-0.616	0.600	0.386
7	-0.402	0.167	0.206
8	-0.464	1.104	0.067
9	-0.797	0.481	0.346
10	-0.745	0.330	0.523
11	-1.860	-2.750	1.123
12	-0.655	0.638	0.344
13	-1.001	-0.192	0.491
14	-0.484	0.459	0.228

Рис.6

2. МОДЕЛИРОВАНИЕ ПРОСТРАНСТВЕННОЙ ПЕРЕМЕННОЙ

2.1. ДЕТЕРМИНИСТИЧЕСКИЕ МОДЕЛИ

2.1.1. Анализ поверхности тренда

К числу главных задач геологического исследования относится изучение закономерностей пространственного строения объектов земной коры. Эти закономерности отражаются на геологических картах и разрезах, учитывающих данные количественного анализа горных пород (данные геохимического опробования, шлихоминералогической съемки, результаты бурения и т.п.). Вопросами картирования количественных признаков занимается специальный раздел математической геологии, получивший название геостатистики. Ключевым понятием геостатистики выступает понятие «пространственной переменной».

Пространственной переменной называют переменную, обладающую географическим распространением и являющуюся функцией пространственных координат. К числу подобных переменных относятся, например, высота на уровнем моря, глубина залегания кровли осадочного пласта или содержание металла в рудном теле.

Пространственная переменная всегда определена в конкретной области пространства, называемой ее геометрическим полем. Это понятие может быть заменено более близким нам понятием «рудного поля». В пределах рудного поля пространственную переменную $Z(\mathbf{x})$ можно рассматривать как функцию точки $\mathbf{x}$ с координатами (x, y, z). Пространственная переменная считается заданной, если известна ее функция распределения

$$F(\mathbf{x}, z) = P\{Z(\mathbf{x}) < z\}.$$

Можно показать, что в общем случае пространственная переменная не удовлетворяет условиям, которые используются для определения случайной переменной. Это связано, во-первых, с

невозможностью точного повторения испытания, заключающегося в отборе пробы в точке **x** рудного поля, во-вторых, с зависимостью испытаний, заключающихся в отборе проб из обогащенной (или обедненной) металлом части рудного поля.

Сказанное свидетельствует о том, что пространственная перемененная обладает свойствами, промежуточными между свойствами полностью детерминированных и полностью случайных величин. Одним из возможных подходов к ее описанию (получившим название тренд-анализа) служит представление функции $Z(\mathbf{x})$ в виде суммы неслучайной и случайной компонент

$$Z(\mathbf{x}) = U(\mathbf{x}) + \varepsilon(\mathbf{x}).$$

Неслучайная компонента $U(\mathbf{x})$ задается аналитически (т.е. аппроксимируется с помощью элементарной функции) и описывает некоторую функциональную поверхность (поверхность тренда), вокруг которой рассеяны значения пространственной переменной.

Одним из распространенных методов тренд-анализа служит метод полиномиальной регрессии. В рамках этого метода искомая функция $U(\mathbf{x})$ заменяется алгебраическим полином от координат $P(\mathbf{x})$ порядка k. Для двумерного случая

$$U(x,y) = P^k(x,y) = b_0 + b_1 x + b_2 y + b_3 xy + b_4 x^2 + b_5 y^2 + \ldots,$$

где $b_0, b_1, \ldots -$ регрессионные коэффициенты, определяемые методом наименьших квадратов из условия $\sum (P^k - Z)^2 = \min$.

В зависимости от величины k поверхность тренда представляет собой плоскость (линейный полином), поверхность 2-го порядка – параболоид, гиперболоид или эллипсоид (квадратичный полином), поверхность третьего порядка (кубический полином) и т.д (рис.7). Как правило, для построения карт используют тренды относительно низких порядков (во всяком случае не выше 5-го), дающие представление о главных закономерностях изменчивости пространственной переменной.

Выбор той или иной поверхности тренда остается за исследователем и может быть осуществлен с учетом остаточной

суммы квадратов отклонений $SS_E = \sum \left(P^k - Z\right)^2$. С помощью этой величины рассчитывается среднеквадратичная ошибка регрессии, проверяется гипотеза о значимости тренда, оценивается величина коэффициента детерминации и степень приближения тренда к исходным данным.

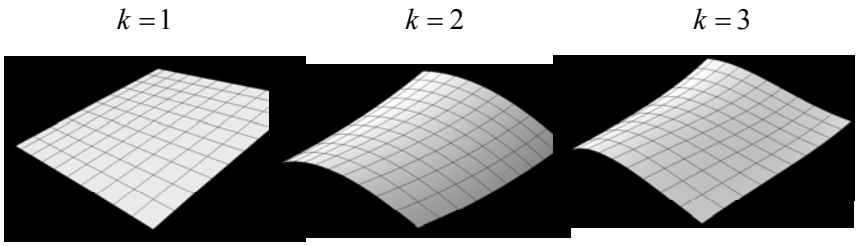

Рис.7

Корректное применение статистических критериев требует соблюдения определенных условий. В первую очередь это касается особенностей распределения регрессионных остатков (отклонений от поверхности тренда)

$$\varepsilon = P^k - Z .$$

Как правило, постулируется нормальное распределение остатков с нулевым средним и конечной (постоянной) дисперсией, некоррелированность остатков и их независимость от значений регрессионной оценки.

Обычно в ходе тренд-анализа строят несколько моделей, последовательно повышая степень аппроксимирующего полинома. При интерпретации получаемых результатов систематическую компоненту пространственной переменной (фон или региональный тренд) связывают с действием регионального геологического фактора. Случайные вариации объясняют влиянием локальных структур (аномалий). Обнаружению этих структур помогает устранение региональной компоненты путем вычитания тренда из значений пространственной переменной и построения карты остатков ε. С формальной точки зрения подобные аномалии

являются областями автокоррелированных (неслучайных) величин, поэтому обычные статистические критерии для оценки их значимости неприменимы.

Задание. *Используя метод полиномиальной регрессии от координат, построить поверхности тренда, моделирующие поведение пространственной переменной. Построить карту остатков и выделить локальные аномалии рудного поля*.

Порядок решения. Открываем программу Surfer. Строим карту фактического материала. С помощью выпадающего меню **Map/Post Map/New Classed Post Map** (**Карта/Карта фактического материала**) обращаемся к файлу данных (например, результатам структурного картирования кровли нефтеносного пласта по данным разведочного бурения), подготовленному в формате Exel. Структура файла данных: первая строка – имена переменных (координата x, координата y, пространственные переменные Z_1, Z_2,…), последующие строки – значения переменных.

Редактируем карту фактического материала. Щелкаем на ней два раза левой кнопкой мыши, открываем диалоговое окно **Map Properties** (**Свойства карты**), на закладке **General** (**Общая**) с помощью выпадающих меню **X Coord**, **Y Coord**, **Z Value** проверяем правильность выбора переменных. Переходим на закладку **Classes** (**Классы**). С помощью выпадающего меню **Binning Method** указываем способ группирования данных **Equal Intervals** (**Равные интервалы**), в поле **Number of cases** вписываем число интервалов группирования (например, 15). Нажимаем на кнопку **Apply** (**Применить**). При необходимости редактируем символы точек наблюдений (путем двойного щелчка левой кнопкой мыши на выбранном символе в пределах окна группирования). Изменяя число интервалов группирования, выявляем аномальные точки наблюдений (например, единичные точки с резко повышенными значениями картируемого признака, расположенные в окружении точек с фоновыми значениями). Закрываем диалоговое окно.

Приступаем к вычислению поверхности тренда. С помощью выпадающего меню **Grid/Data** (**Пространственная интерполяция/Данные**) открываем одноименное диалоговое окно. С помощью выпадающих меню **X**, **Y**, **Z** указываем имена

переменных. Из выпадающего меню **Gridding Method** выбираем метод гриддинга **Polynomial Regression**. (В пакете Surfer двумерная аппроксимация поверхностями названа «гриддинг» (от grid –сетка, англ.), Суть метода - интерполяция данных с целью определения значений пространственной переменной в узлах регулярной сети, плотность которой многократно превышает плотность исходной сети опробования. Результаты гриддинга сохраняются в виде специального файла с расширением *.grd. В дальнейшем подобные грид-файлы используются для построения карт. В геодезии подобный метод называется сгущением сети). Кнопкой **Advanced Options** следует выбрать метод аппроксимации поверхностями, образованными двумерными полиномами регрессии: линейными (**Simple planar surface**), квадратичными (**Quadratic surface**), гиперболическими (**Cubic surface**) или полиномами более высоких степеней, заданных пользователем (**User Define Polynomial**). В случае выбора последней опции пользователю предоставляется возможность задания поверхности тренда 4-го или 5-го порядков (с помощью окон **Max X Order**, **Max Y Order**, **Max Total Order**, в которых указывается максимальная степень, в которую возводятся члены регрессионного уравнения, содержащие координату x, координату y и их смешанные произведения). В окне **Output Grid File** указываем имя файла вывода данных. Нажимаем кнопку **OK**.

При помеченной опции **Grid Report** создание грид-файла сопровождается отчетом, в разделе **Gridding Rules** которого можно найти основные параметры регрессионной модели (для более детальной характеристики регрессионного уравнения лучше обратиться к программе Statistica).

Строим карту поверхности тренда в изолиниях. С помощью выпадающего меню **Map/Contour Map/New Contour Map** открываем созданный грид-файл. Редактируем карту: с помощью двойного щелчка левой кнопкой мыши на карте открываем диалоговое окно, на закладке **General** которого отмечаем опции **Fill Contour** (**Заполнить контуры**) и **Color Scale** (**Цветовая шкала**), нажимаем кнопку **Apply**. Задаем цвет карты: переходим на закладку **Levels** (**Уровни**), щелкаем левой кнопкой мыши на заголовке

таблицы **Fill**, еще раз – на палитре **Foreground Color** и выбираем нужный цвет. Нажимаем кнопку **OK**.

Строим аналогичные карты в режимах **Image Map** (**Растровая карта**) и **Surface** (**Объемная карта**). Редактируем карты: два раза щелкаем левой кнопкой мыши на карте, открываем диалоговое окно, на закладке **General** отмечаем опцию **Show Color Scale**, щелкаем левой кнопкой мыши на палитре **Colors**, затем с помощью кнопки **Load** загружаем цветовую схему Rainbow (файл с расширением *.clr, расположенный в директории Program Files/Golden Software/ Surfer8/Samples).

Строим карту остатков. С помощью выпадающего меню **Grid/Residuals** (**Пространственная интерполяция/Остатки**) последовательно открываем грид-файл и файл исходных данных (в формате Exel), в диалоговом окне **Grid Residuals** в поле **Store Residuals in column** указываем столбец, в котором будут сохранены отклонения от поверхности тренда. Сохраняем результирующий файл. На его основе создаем грид-файл (используя в качестве пространственной переменной остатки, в качестве процедуры гриддинга метод крайгинга) и строим карту.

2.1.2. Метод обратных расстояний

По способу построения поверхность тренда непрерывна, сильно сглаживает исходные данные и не воспроизводит ее значений на исходном множестве точек наблюдений. Поэтому поверхности тренда используются для изучения только наиболее общих (генерализованных) закономерностей поведения пространственной переменной. Если наша задача состоит в том, чтобы воспроизвести функцию $Z(\mathbf{x})$ во всех подробностях (так, чтобы она максимально точно отображала исходные данные), следует поискать другой алгоритм пространственной интерполяции.

Этот алгоритм известен и состоит в том, чтобы представить искомую функцию в виде линейной комбинации соседних наблюдений

$$Z_0^* = \sum_{i=1}^{N} \lambda_i z_i ,$$

где Z_0^* – регрессионная оценка значения функции $Z(\mathbf{x})$ в произвольной точке рудного поля (пункте оценки), z_i – известные значения пространственной переменной в соседних точках наблюдений $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, λ_i – весовые коэффициенты регрессионной модели. Записанное выражение носит название линейного интерполятора. Формально оно представляет собой уравнение регрессии, независимыми переменными которого выступают значения пространственной переменной, измеренные в точках наблюдений. В зависимости от способа вычисления весовых коэффициентов (вида весовой функции) различают: детерминистические и геостатистические интерполяторы.

Один из простейших методов пространственной интерполяции носит название интерполяции по ближайшему соседу. Этот метод состоит в том, что значение функции $Z(\mathbf{x})$ в произвольной точке рудного поля полагается равным значению пространственной переменной в ближайшей точке наблюдения. Для нахождения ближайшего соседа рудное поле разбивается на области влияния (полигоны Вороного), всем точкам которых присваивается значение пространственной переменной в точке наблюдения, принадлежащей этой области. Для произвольной области влияния A вид весовой функции принимает вид

$$\lambda_i = \begin{cases} 1, \mathbf{x}_i \in A \\ 0, \mathbf{x}_i \notin A \end{cases}.$$

Поверхность, построенная с помощью метода полигонов, является ступенчатой и обладает многочисленными точками разрыва на границах областей влияния. Для сглаживания этой поверхности используется метод триангуляции, согласно которому точки наблюдений, являющиеся центрами соседних полигонов, соединяются отрезками прямых, после чего изучаемая поверхность представляется в виде набора плоских треугольных пластинок, вершины которых характеризуются известными значениями картируемого признака. Подобная поверхность обладает многочисленными изломами и характеризуется искусственной треугольной формой изолиний (для построения карты в изолиниях

поверхность триангуляции рассекается серией горизонтальных плоскостей). Поэтому часто вместо плоских треугольников используют изогнутые треугольные пластинки, плавно переходящие друг в друга в местах стыков. Подобного результата удается достигнуть при помощи сплайнов[3].

Несмотря на использование сплайнов, метод триангуляции является весьма приблизительным методом аппроксимации искомой функции $Z(\mathbf{x})$. Это связано с тем, что в ходе пространственной интерполяции для оценки Z_0^* в расчет берется только одна (ближайшая) точка наблюдения. Данного недостатка лишены интерполяторы, построенные с помощью метода обратных расстояний. В основе этого метода лежит очевидное допущение: чем дальше находится точка наблюдения от пункта оценивания, тем меньшее влияние она оказывает на значение функции $Z(\mathbf{x})$ в этом пункте. Считая влияние обратно пропорциональным расстоянию, можно присвоить соответствующему весовому коэффициенту значение обратного нормированного расстояния

$$\lambda_i = \frac{1/d_i}{\sum 1/d_i}, \qquad \sum \lambda_i = 1$$

(нормировка здесь является обязательной операцией, поскольку для любой весовой функции сумма лямбд должна равняться единице). Вычислив обратные расстояния до всех ближайших точек наблюдения, можно рассчитать значение оценки Z_0^*

$$Z_0^* = \frac{\sum z_i (1/d_i)}{\sum (1/d_i)}.$$

В рассмотренном примере использован лишь один из возможных видов весовой функции. С не меньшим успехом лямбды можно вычислять в предположении, что влияние точек наблюдения на пункт оценивания обратно пропорционально квадрату расстояний

[3] Сплайнами называют семейство кусочно-определенных функций, которые описывают гладкие кривые, соединяющие заданные последовательности точек.

$$\lambda_i = \frac{(1/d_i)^2}{\sum(1/d_i)^2}.$$

По сравнению с методом обратных расстояний, в последнем случае усиливается влияние точек, приближенных к пункту оценивания, и ослабляется влияние удаленных точек.

Задание. *Построить карты площадного распределения значений пространственной переменной с помощью детерминистических интерполяторов.*

Порядок решения. Открываем программу Surfer. С помощью выпадающего меню **Map/Post Map/New Classed Post Map** обращаемся к файлу данных (например, результатам геохимического картирования рудного поля) и строим карту фактического материала.

С помощью выпадающего меню **Grid/Data** указываем имена переменных и последовательно специфицируем перечисленные методы гриддинга: **Nearest Neighbor** (**Метод ближайшего соседа**), **Triangulation with Linear Interpolation** (**Метод триангуляции**), **Inverse Distance to a Power** (**Метод обратных расстояний**). При использовании метода обратных расстояний дополнительно с помощью кнопки **Advanced Options** указываем степень (**Power**) в которую возводится обратное расстояние при расчете весового коэффициента (например, 1 и 2). Результаты гриддинга сохраняем в виде отдельных файлов.

На основе созданных грид-файлов строим карты (контурные, растровые, объемные) и сравниваем полученные результаты.

2.2. ГЕОСТАТИСТИЧЕСКИЕ МОДЕЛИ

2.2.1. Моделирование вариограммы

Рассмотренные детерминистические интерполяторы требуют для своего задания априорных суждений о виде весовой функции, с помощью которой рассчитывается регрессионная оценка Z_0^*. Они не

учитывают наблюдающуюся изменчивость пространственной переменной и не позволяют оценивать ошибку пространственной интерполяции. Этих недостатков лишены геостатистические интерполяторы. Ключевую роль в их построении играет вариограмма.

По определению, вариограммой (полувариограммой) называют функцию, значения которой равны половине среднего квадрата разности значений пространственной переменной в точках $\mathbf{x}$ и $\mathbf{x}+\mathbf{h}$

$$\gamma(\mathbf{h}) = 0.5E\left[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right]^2 .$$

Данная формула справедлива при условии стационарности 2-го порядка (стационарности приращений). Это означает, что: 1) аргументом вариограммы выступает вектор $\mathbf{h}$, соединяющий две произвольные точки рудного поля, 2) вариограмма показывает, как в среднем различаются значения пространственной переменной (например, содержания металла в горной породе) в зависимости от расстояния и направления вектора в изучаемой области пространства. В случае изотропных рудных полей аргументом вариограммы является скалярная величина $h = |\mathbf{h}|$ (расстояние между точками наблюдений). Для стационарных переменных (т.е. переменных, обладающих конечной дисперсией) выполняется следующее соотношение между изотропной вариограммой, дисперсией $C(0)$ и пространственной ковариацией $C(h)$

$$\gamma(h) = C(0) - C(h) .$$

Графиком идеализированной вариограммы служит кривая, исходящая из начала координат (рис.8). По мере увеличения расстояния h значения функции $\gamma(h)$ монотонно растут; угол наклона кривой при этом показывает, насколько быстро убывает влияние контрольной точки на ее окружение. Достигнув некоторого предела (порога вариограммы), кривая выполаживается (данное свойство характеризует только стационарные переменные). Взаимное влияние точек рудного поля становится здесь равным нулю. Расстояние, на котором вариограмма достигает порога,

называется радиусом корреляции (рангом вариограммы). Вариограммы, не обладающие порогом, свойственны нестационарным переменным.

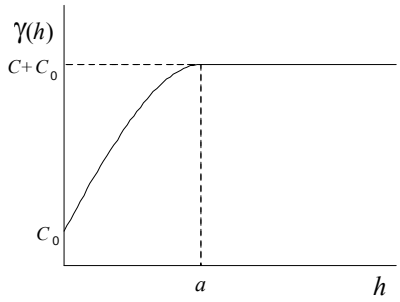

Рис.8

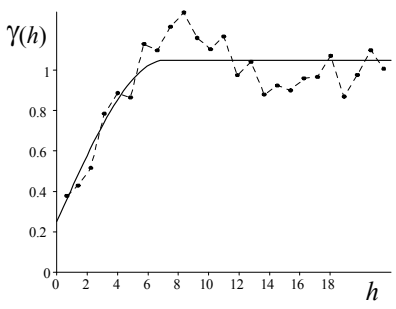

Рис.9

Порог и радиус корреляции характеризуют особенности вариограммы на дальних расстояниях. Не менее важно рассмотреть поведение функции $\gamma(h)$ при малых h. Особую роль тут играют случаи пересечения вариограммой вертикальной оси в точке с ординатой, отличной от нуля. Подобная ситуация, как правило, связана с тем, что пространственная переменная не обладает свойством непрерывности (если речь идет о содержании металла, отсутствие непрерывности проявляется как скачкообразное изменение содержаний в соседних точках опробования). Поскольку особенно часто данный эффект проявляется на месторождениях золота, он получил название эффекта самородков.

Вариограмма используется для изучения пространственной корреляционной структуры рудного поля. Ее значения служат входными параметрами для реализации уравнений крайгинга (см. ниже). На практике вид функции $\gamma(h)$ неизвестен и должен быть оценен при помощи выборочной вариограммы

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[z(x_i + h) - z(x_i) \right]^2 .$$

При этом важно, чтобы модельная вариограмма имела вид непрерывной гладкой кривой (рис.9). Обычно она выбирается из числа аналитических функций, удовлетворяющих базовым свойствам вариограммы (положительной определенности, поведению в нуле и на бесконечности и т.д.), и затем подгоняется к опытным данным. Примером подобных функций может служить сферическая модель

$$\gamma(h)=\begin{cases} C_0 + C\left(\dfrac{3}{2}\cdot\dfrac{h}{a}-\dfrac{1}{2}\cdot\dfrac{h^3}{a^3}\right), & h \le a \\ C_0 + C, & h > a \end{cases}$$

где a – радиус корреляции, C_0 – эффект самородков, $C_0 + C$ – порог (рис.10). Эта модель ведет себя линейно вблизи нуля и резко выполаживается на расстоянии, равном радиусу корреляции. Другой пример – экспоненциальная модель

$$\gamma(h) = (C_0 + C)\left[1 - \exp\left(-\frac{h}{a}\right)\right].$$

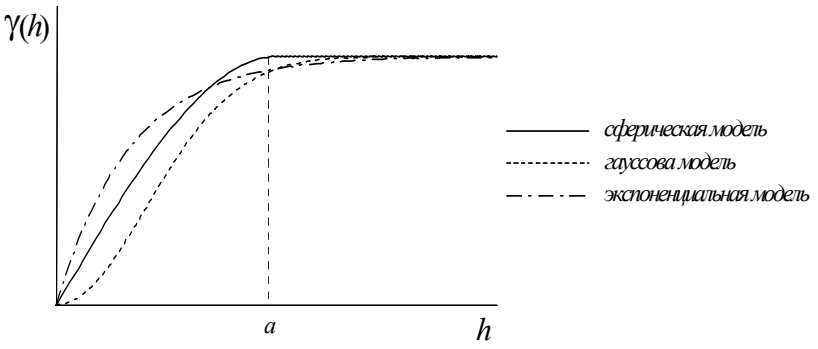

Рис.10

По сравнению со сферической эта модель растет быстрее в начале координат, но медленнее (точнее асимптотически) достигает порога. Для ее характеристики используется понятие эффективного радиуса

корреляции, т.е. расстояния, на котором вариограмма достигает 95% пороговых значений. В случае экспоненциальной модели он равен $3a$. Еще один пример – гауссова модель

$$\gamma(h) = (C_0 + C)\left[1 - \exp\left(-\frac{h^2}{a^2}\right)\right].$$

Отличительной чертой этой модели является параболическое поведение вариограммы вблизи нуля. Она достигает порога асимптотически с эффективным радиусом корреляции равным $\sqrt{3}a$.

Для всех рассмотренных моделей критическое значение имеет поведение вариограммы вблизи нуля. Именно это поведение определяет степень непрерывности пространственной переменной (степень сходства соседних точек наблюдения). В данной связи особенно сильны различия между гауссовой и экспоненциальной моделью. Первая характеризует непрерывную пространственную переменную, вторая – пространственную переменную, отличающуюся повышенной вариабельностью даже на малых расстояниях. Промежуточными свойствами обладает сферическая модель. Все три модели могут испытывать скачок (разрыв) в нуле при ненулевом эффекте самородков, что резко повышает вариабельность пространственной переменной на малых расстояниях и сказывается на виде картируемой поверхности.

Описанные модели используются для аппроксимации изотропных вариограмм. Вместе с тем довольно часто рудные поля обладают отчетливо выраженной анизотропией (например, в случае тектонического контроля оруденения). Для ее характеристики рассчитывают директивные вариограммы. В простейшем случае на карте задают четыре сектора с углом раствора 90^o, ориентированные по направлениям 0, 45, 90 и 135^o. Для каждого сектора в расчет принимают только те пары наблюдений, направления между которыми попадают внутрь сектора. Построенные вариограммы демонстрируют особенности поведения пространственной переменной в различных направлениях. При так называемой геометрической анизотропии директивные вариограммы обладают одинаковыми порогами, но разными рангами (радиусами

корреляции). Меньший радиус корреляции соответствует направлению наибольшей изменчивости картируемого признака (направлению вкрест простирания осадочных слоев, зон разломов, рудных жил и т.п.), больший радиус – наименьшей изменчивости (рис.11). Эти радиусы берутся за основу при построении эллипса анизотропии. Он характеризуется коэффициентом анизотропии

$$k = \frac{a_1}{a_2}, \quad a_1 > a_2 .$$

С помощью коэффициента анизотропии вариограмма, соответствующая произвольному вектору $\mathbf{h}$, рассчитывается как

$$\gamma(\mathbf{h}) = \gamma_1 \sqrt{h_1^2 + k^2 h_2^2} ,$$

где γ_1 – вариограмма с б***о***льшим рангом, h_1 – компонента вектора вдоль направления с меньшей изменчивостью (б***о***льшим рангом), h_2 – компонента вектора вдоль направления с большей изменчивостью (меньшим рангом).

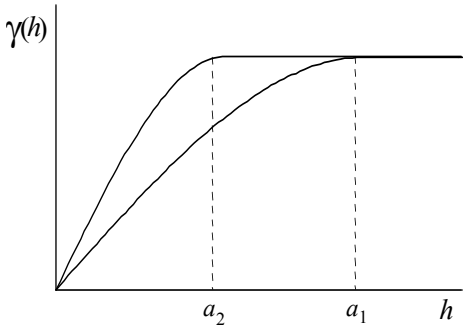

Рис.11

Задание. *Построить модели изотропной и директивной вариограмм пространственной переменной*.

Порядок решения. Открываем программу Surfer. С помощью выпадающего меню **Grid/Variogram/New Variogram** обращаемся к файлу данных. В поле **Data Columns** диалогового

окна **New Variogram** указываем имена переменных и нажимаем кнопку **OK**.

Анализируем график. Если построенная вариограмма не имеет порога, снова открываем окно **New Variogram**, на закладке **General** в поле **Detrend** отмечаем опции **Linear** или **Quadratic** (**Снять линейный или квадратичный тренд**) и строим новую вариограмму. Если на экспериментальной вариаграмме отмечаются выбросы в области малых h, открываем карту фактического материала, удаляем аномальные точки наблюдений и строим новую вариограмму.

Редактируем изотропную вариограмму. С помощью двойного щелчка левой кнопкой мыши на графике, открываем диалоговое окно **Variogram Properties** (**Свойства вариограммы**). Сглаживаем график, уменьшая в поле **Number of Lags** закладки **Experimental** число лагов (лагом длины h называют целое число интервалов, разделяющих две любые точки рудного поля).

Приступаем к моделированию изотропной вариограммы. Переходим на закладку **Model** и с помощью кнопки **Remove** удаляем строку **Linear** поля **Variogram Components**. Нажимаем на кнопку **Add** и добавляем с помощью открывшегося меню одну из моделей вариограмм: **Exponential** (**Экспоненциальная**), **Gaussian** (**Гауссова**), **Linear** (**Линейная**), **Power** (**Степенная**), **Spherical** (**Сферическая**), **Hole Effect** (**Эффекта включений**) и т.д. Подгоняем модель к экспериментальным данным. Для этого указываем параметры модели: в поле **Scale** – порог, в поле **Length** – радиус корреляции, в поле **Error Variance** (активизируется при подсвечивании строки **Nugget Effect**) – эффект самородков и нажимаем кнопку **Apply** (**Применить**). Действуя методом проб и ошибок, добиваемся примерного совпадения экспериментальной и модельной вариограмм (черной и синей кривой).

Строим директивную вариограмму. Для этого переходим на закладку **Experimental** и в поле **Lag Direction** (Направление лагов) вписываем следующие значения параметров: **Tolerance** (**Допуск**) 45, **Step Amount** (**Шаг**) 45. С помощью кнопок **Step CW** и **Step CCW** вращаем построенный сектор и наблюдаем, как изменяется график. Выбираем направление с наибольшим рангом. Переходим на

закладку **Model**. Подсвечиваем строку с названием выбранной вариограммы и, изменяя ранг (**Length**), подгоняем модель к экспериментальной кривой. Возвращаемся на закладку **Experimental** и, вращая сектор, выбираем направление с наименьшим рангом. Снова переходим на закладку **Model**. Подсвечиваем строку с названием выбранной вариограммы, в поле **Anisotropy** указываем коэффициент анизотропии (**Ratio**) и направление длинной оси эллипса анизотропии (**Angle**). Нажимаем кнопку **Apply**. Действуя методом проб и ошибок, добиваемся примерного совпадения экспериментальной и модельной кривой.

2.2.2. Геостатистическая интерполяция (крайгинг)

Семейство геостатистических интерполяторов объединяется под общим названием «крайгинга». В сравнении с детерминистическими интерполяторами крайгинг обладает минимальной дисперсией и позволяет рассчитывать ошибку интерполяции в любом пункте оценивания.

В статистическом смысле крайгинг является наилучшим несмещенным линейным оценивателем. Его использование оптимизирует процедуру пространственной интерполяции и дает возможность получить наилучшую регрессионную оценку Z_0^*. Последняя считается таковой, если она подчиняется требованию несмещенности $E(Z_0^* - Z_0) = 0$ и обладает минимальной дисперсией $\mathrm{var}(Z_0^* - Z_0) = \min$, где Z_0 – истинное значение функции $Z(\mathbf{x})$ в пункте оценивания.

Легко показать, что первому условию в случае стационарных переменных, которые характеризуются постоянным (хотя и неизвестным) математическим ожиданием, удовлетворяет равенство суммы весов регрессионной модели единице. Отсюда для нахождения наилучшей регрессионной оценки требуется найти такие лямбды, которые бы минимизировали дисперсию оценивания при условии $\sum \lambda_i = 1$.

Решение задачи сводится к минимизации выражения

$$F = \sigma_E^2 - 2\mu(\sum \lambda_i - 1)$$

или более подробно

$$F = 2\sum_i \lambda_i \gamma_{0i} - \sum_i \sum_j \lambda_i \lambda_j \gamma_{ij} - 2\mu(\sum \lambda_i - 1)$$

где σ_E^2 – дисперсия оценивания, γ_{0i} – значения вариограммы для векторов, соединяющих пункт оценивания с точками наблюдений, γ_{ij} – значения вариограммы для векторов, соединяющих точки наблюдений друг с другом, μ – неизвестный параметр (множитель Лагранжа). После преобразований искомые регрессионные коэффициенты λ_i находятся из матричного уравнения

$$[\lambda] = [\gamma_{ij}]^{-1} \cdot [\gamma_{0i}].$$

Дисперсия регрессионной оценки (дисперсия крайгинга) рассчитывается по формуле

$$\sigma_E^2 = \sum_i \lambda_i \gamma_{0i} + \mu.$$

Она оказывается наименьшей из всех возможных.

Задание. *Построить карту площадного распределения значений пространственной переменной методом крайгинга.*

Порядок решения. Открываем программу Surfer. Обращаемся к файлу данных и строим вариограмму, характеризующую поведение пространственной переменной (например, изменчивость содержаний металла) по различным направлениям.

С помощью выпадающего меню **Grid/Data** указываем имена переменных и указываем выбранный метод гриддинга **Kriging**. С помощью кнопки **Advanced Options** открываем диалоговое окно и нажимаем кнопку **Get Variogram**, затем – кнопку **OK**. Сохраняем результаты гриддинга.

На основе созданного грид-файла строим карты площадного распределения значений пространственной переменной.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. *Гмурман В.Е*. Теория вероятностей и математическая статистика. М.: Высшая школа, 1997.

2. *Боровиков В.П.* Популярное введение в программу STATISTICA. М.: Компьютер Пресс, 1998.

3. *Давид М*. Геостатистические методы при оценке запасов руд. Л.: Недра, 1980.

4. *Дэвис Дж.С*. Статистический анализ данных в геологии. М.: Недра, 1990.

5. *Каждан А.Б., Гуськов О.И*. Математические методы в геологии. М.: Недра, 1990.

6. Проблемы окружающей среды и природных ресурсов. 1999. N 11.

7. *Смоленский В.В*. Статистические методы обработки экспериментальных данных. СПб.: СПГГИ, 2003.

8. *Goovaertz P*. Geostatistics for natural resources evaluation. N.Y.: Oxford University Press, 1997.

СОДЕРЖАНИЕ