

**КАЗАНСКИЙ (ПРИВОЛЖСКИЙ)
ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**

Институт экологии и географии

**ГЕОСТАТИСТИЧЕСКИЙ АНАЛИЗ
ДАННЫХ В ЭКОЛОГИИ И
ПРИРОДОПОЛЬЗОВАНИИ
(с применением пакета R)**

Учебное пособие

КАЗАНЬ – 2012

УДК 519.2, 004.43

*Печатается по решению Редакционно-издательского совета
ФГАОУВПО*

«Казанский (Приволжский) федеральный университет»

*Методической комиссии Института экологии и географии
Протокол № 2 от 16 марта 2012 г.*

*Заседания кафедры моделирования экосистем
Протокол № 8 от 7 марта 2012 г.*

Авторы-составители

доктор биологических наук, проф. А.А.Савельев,
канд. биологических наук, доц. С.С. Мухарамова,
ст. преп. А.Г.Пилюгин,

канд. биологических наук, асс. Н.А.Чижикова

Научный редактор

доктор физ.-мат. наук, проф. Э.В. Скворцов

Рецензенты:

доктор физ.-мат. наук, проф. КФУ М.Г. Храмченков
канд. тех. наук, доц. КНИТУ-КАИ И.А. Сабаев

Геостатистический анализ данных в экологии и природопользовании (с применением пакета R): Учебное пособие / А.А. Савельев, С.С. Мухарамова, А.Г. Пилюгин, Н.А. Чижикова. – Казань: Казанский университет, 2012. – 120 с.

Учебное пособие предназначено для студентов естественных факультетов, и может быть использовано при проведении курсов, посвященных изучению методов численного описания природных пространственных явлений. В пособии приводятся основные понятия геостатистики, описываются методы геостатистического оценивания, рассматривается пошаговый подход к анализу, обработке и моделированию пространственных данных. Приводятся примеры реализации изучаемых методов с использованием среды статистического программирования R.

© Казанский университет, 2012

ОГЛАВЛЕНИЕ

Основные понятия геостатистики	5
Случайные величины, пространственные переменные, случайные функции	7
Моменты, используемые в линейной геостатистике	9
Первый момент - математическое ожидание	9
Моменты второго порядка	9
Эргодичность.....	10
Гипотеза стационарности.....	11
Строгая стационарность	12
Стационарность второго порядка	12
Стационарность приращений.....	13
Многомерные гауссовы случайные функции.....	14
Методы геостатистического моделирования	17
Простой кригинг	18
Кригинг с неизвестным матожиданием.....	19
Параметры кригинга.....	21
Оценка качества модели.....	23
Картографическая визуализация модели.....	24
Условия, необходимые для применения геостатистических методов.....	25
Предварительный анализ и обработка пространственных данных	29
1) Визуализация выборки на базовой карте	29
2) Декластеризация	31
Обнаружение кластеров	32
Методы декластеризации	33

3) Проверка ограничений и предположений геостатистики	39
Статистическое описание данных	40
Анализ пространственного тренда	45
Проверка мультинормальности	49
Вариограммный анализ данных.....	58
h-графики	60
Поверхность вариограммы	66
Вариограмма по направлению	70
Ранг, порог, вложенные структуры	71
Поведение около нуля и эффект самородков	72
Анизотропия	74
Проявление присутствия дрифта	75
Пример оценки вариограммы в R	76
Моделирование вариограммы	78
Базисные модели	79
Итеративное построение модели вариограммы	84
Пример построения модели вариограммы в R	89
Пример оценки пространственной переменной методом	
кригинга в R	101
Оценка остатков методом ординарного кригинга	101
Оценка пространственного тренда на основе модели тренда	104
Оценка изучаемой пространственной переменной и построение	
прогнозной карты	106
Список литературы	108
Приложения	111

Основные понятия геостатистики

Многие природные феномены, рассматриваемые в экологии и природопользовании, имеют сильно выраженный пространственный аспект. Для изучения таких явлений недостаточно обычных статистических процедур, необходимо применение специальных методов, основу которых составляет геостатистика. *Целью* геостатистического анализа и моделирования (или просто геостатистики) является численное описание природных явлений, распределенных в пространстве или во времени и пространстве.

Примерами подобных пространственно распределенных переменных могут служить: концентрации загрязняющих веществ в различных средах (воздухе, воде, почве и пр.); свойства геологических пластов (пористость, проницаемость и пр.); содержание руд в минеральных отложениях; свойства почвенного покрова; различные характеристики растительного покрова; количество осадков, температура воздуха и др.

Наблюдение, оценка, моделирование этих явлений проводится для заданной территории. Как правило, для таких переменных характерна исключительная сложность деталей, а в силу различных причин, чаще всего экономических, они измерены на изучаемой территории недостаточно «густо» (в ограниченном числе точек), что приводит к значительной пространственной неопределенности. Все это делает невозможным их описание простыми моделями или сглаживание стандартными математическими функциями.

В рамках геостатистического подхода строится вероятностная модель изучаемого явления. Оценка и

моделирование выполняются с использованием свойств этой модели, а получаемые оценкой или моделированием величины являются некоторым отображением действительности. При этом в высшей степени важно различать, что есть реальность и что относится к вероятностной модели.

На рис.1 представлена общая схема изучения явления в рамках геостатистического подхода.

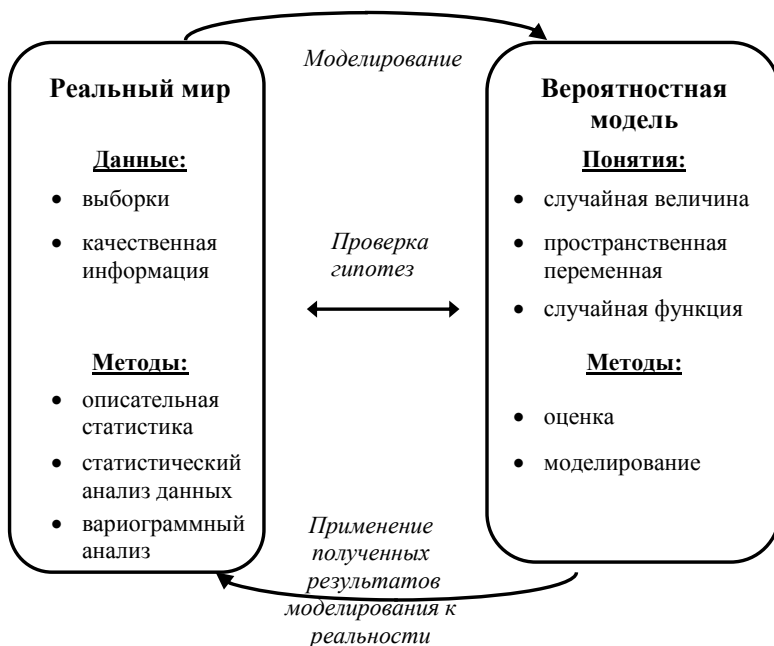


Рис.1

Геостатистический анализ позволяет учитывать при моделировании явлений пространственные законы их распределения, наличие природных барьеров, основные направления материальных, энергетических, информационных потоков и т.п.

Случайные величины, пространственные переменные, случайные функции

Случайная величина Z есть переменная, которая принимает свои значения (образующие множество элементарных событий) случайно, согласно некоторому вероятностному механизму. Например, результат бросания кубика можно рассматривать как случайную величину, принимающую одно из шести равновероятных значений.

Пространственная переменная - это переменная, распределенная в пространстве. То есть, это некоторая числовая функция $z(\mathbf{x})$, где \mathbf{x} – точка в пространстве (на плоскости), задаваемая своими координатами. Пространственная переменная используется для представления поведения изучаемого феномена в заданной области исследования. Например, содержание тяжелых металлов в верхнем слое почвы может быть рассмотрено как пространственная переменная в двумерном пространстве.

Как правило, природный феномен, рассматриваемый в пространстве, совмещает в себе две составляющие: регулярную и случайную. Регулярная составляющая обусловлена неким общим законом распределения этого феномена. Например, на исследуемой территории существуют зоны, которые имеют, в среднем, более высокое содержание тяжелых металлов в почве, чем остальная территория. Случайная составляющая обычно обусловлена локальной изменчивостью природного явления. Например, на ограниченном участке территории содержание тяжелых металлов выглядит случайным. Учесть эту двойственность (одновременно случайность и регулярность

феномена) позволяет вероятностная формализация, обеспечиваемая аппаратом случайных функций.

Случайная функция $Z(\mathbf{x})$ есть множество случайных величин $\{Z(\mathbf{x}) | \mathbf{x} \text{ принадлежит области исследования}\}$.

В рамках геостатистического подхода строится следующая вероятностная модель:

Изучаемое пространственно-распределенное явление рассматривается как случайная функция $Z(\mathbf{x})$, т.е. бесконечное множество случайных величин, представляющих феномен в каждой точке пространства. А та единственная (уникальная) реальность, которая имеет место на исследуемой территории и которой соответствует пространственная переменная $z(\mathbf{x})$ (множество значений феномена, распределенное в пространстве), рассматривается как одна из возможных реализаций случайной функции $Z(\mathbf{x})$.

При этом случайная и регулярная составляющие изучаемого явления выражаются следующим образом:

С одной стороны, в каждой точке \mathbf{x} (локально) значение $z(\mathbf{x})$ рассматривается как реализация случайной величины $Z(\mathbf{x})$, т.е. оно случайно.

С другой стороны, значение $z(\mathbf{x})$ зависит от положения точки \mathbf{x} в пространстве. Эта зависимость проявляется следующим образом: для любой пары точек \mathbf{x} и $\mathbf{x}+\mathbf{h}$ соответствующие случайные величины $Z(\mathbf{x})$ и $Z(\mathbf{x}+\mathbf{h})$ связаны корреляцией. Такая пространственная корреляционная зависимость отражает пространственную структуру явления, его статистическую регулярность.

Моменты, используемые в линейной геостатистике

Рассмотрим случайную функцию $Z(\mathbf{x})$. Любому множеству k точек пространства $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ соответствует множество k случайных величин

$$\{Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_k)\},$$

которые полностью характеризуются k -мерной функцией распределения

$$F_{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k}(z_1, z_2, \dots, z_k) = \text{Prob}\{Z(\mathbf{x}_1) < z_1, \dots, Z(\mathbf{x}_k) < z_k\}.$$

Множество всех таких функций распределения для любого положительного целого k и для любой комбинации точек \mathbf{x}_i образует *закон распределения* случайной функции $Z(\mathbf{x})$.

Заметим, что для решения большинства задач оценки/прогноза не требуется точное знание закона распределения. Более того, объем данных обычно недостаточен, чтобы сделать надежные выводы о виде закона распределения. Поэтому на практике ограничиваются применением *линейной геостатистики*, использующей только первые два момента случайной функции. В ее рамках не может быть обнаружено отличие между двумя случайными функциями, имеющими одинаковые первый и второй моменты.

Первый момент - математическое ожидание

Рассмотрим случайную величину $Z(\mathbf{x})$ в точке \mathbf{x} . Если для величины $Z(\mathbf{x})$ существует математическое ожидание, то оно является функцией от \mathbf{x}

$$E\{Z(\mathbf{x})\} = m(\mathbf{x}).$$

Моменты второго порядка

Приведем три момента второго порядка, рассматриваемые в геостатистике:

1. Дисперсия

$$\begin{aligned} \text{Var}\{Z(\mathbf{x})\} &= E\{[Z(\mathbf{x})-m(\mathbf{x})]^2\} = \\ &= E\{Z^2(\mathbf{x})\}-2E\{Z(\mathbf{x})\}m(\mathbf{x})+m^2(\mathbf{x}) = \\ &= E\{Z^2(\mathbf{x})\}-m^2(\mathbf{x}). \end{aligned}$$

Так же как и математическое ожидание, дисперсия является функцией от \mathbf{x} .

2. Ковариация. Если две случайные величины $Z(\mathbf{x}_1)$ и $Z(\mathbf{x}_2)$ имеют дисперсии, то для них существует ковариация, которая есть функция от \mathbf{x}_1 и \mathbf{x}_2 :

$$\begin{aligned} C(\mathbf{x}_1, \mathbf{x}_2) &= E\{[Z(\mathbf{x}_1)-m(\mathbf{x}_1)][Z(\mathbf{x}_2)-m(\mathbf{x}_2)]\} = \\ &= E\{Z(\mathbf{x}_1)Z(\mathbf{x}_2)\}-E\{Z(\mathbf{x}_1)\}m(\mathbf{x}_2)-E\{Z(\mathbf{x}_2)\}m(\mathbf{x}_1)+m(\mathbf{x}_1)m(\mathbf{x}_2) = \\ &= E\{Z(\mathbf{x}_1)Z(\mathbf{x}_2)\}-m(\mathbf{x}_1)m(\mathbf{x}_2). \end{aligned}$$

3. Вариограмма (или полувариограмма) определяется как половина дисперсии приращения $[Z(\mathbf{x}_1)-Z(\mathbf{x}_2)]$:

$$\begin{aligned} \gamma(\mathbf{x}_1, \mathbf{x}_2) &= 1/2 \text{Var}\{Z(\mathbf{x}_1)-Z(\mathbf{x}_2)\}, \\ 2 \gamma(\mathbf{x}_1, \mathbf{x}_2) &= E\{[Z(\mathbf{x}_1)-Z(\mathbf{x}_2)]^2\}- [E\{Z(\mathbf{x}_1)-Z(\mathbf{x}_2)\}]^2, \\ \gamma(\mathbf{x}_1, \mathbf{x}_2) &= 1/2 E\{[Z(\mathbf{x}_1)-Z(\mathbf{x}_2)]^2\} \quad \text{при} \quad E\{Z(\mathbf{x}_1)-Z(\mathbf{x}_2)\}=0. \end{aligned}$$

Эргодичность

При вероятностном подходе мы имеем дело с ожидаемыми величинами, которые есть не что иное, как обобщение всех возможных реализаций случайного процесса. Эргодичность случайной функции означает, что среднее по всем возможным реализациям равно среднему отдельной безграничной в пространстве реализации. Выполнение этого свойства позволяет оценивать моменты случайной функции по пространственной переменной.

Гипотеза стационарности

Для того чтобы статистически оценить закон распределения или первые два момента случайной функции $Z(\mathbf{x})$, где \mathbf{x} - точка в пространстве, требуется множество реализаций $z_1(\mathbf{x}), \dots, z_k(\mathbf{x})$ случайной функции $Z(\mathbf{x})$. Однако, в отличие от бросания кубика, для природного феномена невозможно получить новые реализации. Действительность уникальна и параметры модели должны быть определены на основе единственной уникальной реализации.

Если в пределах определенного региона исследуемый феномен пространственно однороден, то можно считать, что явление повторяет само себя в пространстве. Это повторение эквивалентно множественной реализации одной и той же случайной функции $Z(\mathbf{x})$ (что схематично показано на рис.2). И, таким образом, две наблюдаемые в разных точках величины $z(\mathbf{x}_0)$ и $z(\mathbf{x}_0 + \mathbf{h})$ могут быть рассмотрены как две различные реализации одной случайной функции $Z(\mathbf{x})$, что позволит (при условии, что случайная функция предполагается еще и эргодичной) получать достоверные статистические выводы на основе наблюдаемых значений пространственной переменной.

Разной степени пространственной однородности изучаемого феномена соответствуют разные гипотезы стационарности случайной функции. Тип стационарности говорит о том, какие статистические выводы допустимы в рамках данной вероятностной модели.

Общий случай

Гипотеза стационарности

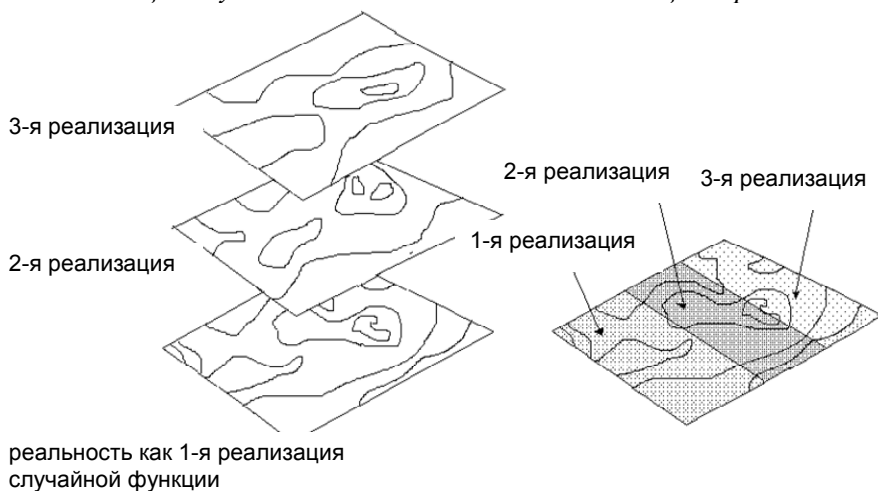


Рис.2. Множество реализаций случайной функции и гипотеза стационарности

Строгая стационарность

Случайная функция стационарна в строгом смысле, если ее закон распределения инвариантен относительно сдвига, то есть

$\{Z(x_1), Z(x_2), \dots, Z(x_k)\}$ и $\{Z(x_1+h), Z(x_2+h), \dots, Z(x_k+h)\}$ имеют одинаковый k -мерный закон распределения при любом h .

На практике это очень сильное предположение редко выполняется, оно обычно заменяется гипотезой о стационарности второго порядка или гипотезой о стационарности приращений.

Стационарность второго порядка

Так как линейная геостатистика рассматривает только первые два момента закона распределения, достаточно предположить, что эти моменты существуют и ограничить дальнейшие предположения ими.

Случайная функция $Z(\mathbf{x})$ обладает стационарностью второго порядка, если:

- математическое ожидание существует и не зависит от \mathbf{x} :

$$E\{Z(\mathbf{x})\} = m, \quad \forall \mathbf{x},$$

- для каждой пары случайных величин $\{Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})\}$ ковариация существует и зависит только от \mathbf{h} :

$$C(\mathbf{h}) = E\{Z(\mathbf{x}+\mathbf{h}) \cdot Z(\mathbf{x})\} - m^2, \quad \forall \mathbf{x}.$$

Стационарность ковариации предполагает стационарность дисперсии:

$$Var\{Z(\mathbf{x})\} = E\{[Z(\mathbf{x})-m]^2\} = C(0), \quad \forall \mathbf{x},$$

и стационарность вариограммы:

$$2C(\mathbf{h}) = 2E\{Z(\mathbf{x}+\mathbf{h}) \cdot Z(\mathbf{x})\} - 2m^2 = [E\{Z^2(\mathbf{x}+\mathbf{h})\} - m^2] + [E\{Z^2(\mathbf{x})\} - m^2] -$$

$$- [E\{Z^2(\mathbf{x}+\mathbf{h})\} - 2E\{Z(\mathbf{x}+\mathbf{h}) \cdot Z(\mathbf{x})\} + E\{Z^2(\mathbf{x})\}] = 2C(0) - 2\gamma(\mathbf{h}),$$

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}).$$

Это соотношение показывает, что, при условии справедливости гипотезы о стационарности второго порядка, ковариация и вариограмма есть два равноценных понятия для характеристики связи между случайными величинами $Z(\mathbf{x})$ и $Z(\mathbf{x}+\mathbf{h})$, разделенными вектором \mathbf{h} .

Стационарность приращений

Многие природные феномены имеют бесконечный разброс значений (броуновское движение). Для такого явления не существует дисперсия или ковариация. Однако для его приращений дисперсия существует, а, значит, существует и может быть определена вариограмма для самого явления. Как следствие, гипотеза стационарности второго порядка может быть ослаблена, если предположить только существование и стационарность вариограммы.

Случайная функция $Z(\mathbf{x})$ обладает стационарностью приращений, если

- математическое ожидание существует и не зависит от \mathbf{x} :

$$E\{Z(\mathbf{x})\} = m, \quad \forall \mathbf{x},$$

- для любого \mathbf{h} приращение $Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})$ имеет конечную дисперсию, не зависящую от \mathbf{x} :

$$Var\{Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})\} = E\{[Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})]^2\} = 2\gamma(\mathbf{h}), \quad \forall \mathbf{x}.$$

Стационарность второго порядка влечет за собой стационарность приращений, обратное неверно.

Стационарность приращений можно также интерпретировать как стационарность второго порядка для приращения случайной функции.

Если принята гипотеза о стационарности приращений без гипотезы о стационарности второго порядка, то для характеристики связи между случайными величинами $Z(\mathbf{x})$ и $Z(\mathbf{x}+\mathbf{h})$ может быть использована только вариограмма.

Многомерные гауссовы случайные функции

Многомерная гауссова модель удобна в силу своей аналитической простоты. Кроме того, согласно *центральной предельной теореме*, это распределение является предельным случаем многих аналитических распределений.

Если пространственный феномен получен *суммированием* нескольких *независимых* источников, имеющих похожие распределения, то пространственное распределение феномена может быть смоделировано с помощью многомерной гауссовой случайной функции:

$$Z(\mathbf{x}) = \sum_{k=1}^K Y_k(\mathbf{x}),$$

где $Z(x)$ - случайная функция, представляющая феномен; $Y_k(x)$ - независимые случайные функции, имеющие схожие законы распределения.

Природные процессы, которые генерируют наблюдаемый феномен, редко независимы один от другого, и способ их взаимодействия редко аддитивен. Однако многомерные гауссовы модели были удачно применены в ряде исследований. Этот практический успех в сочетании с удобством аналитических свойств гауссовых моделей делает их наиболее привлекательными при моделировании непрерывных переменных (если только в силу каких-либо априорных знаний или предварительных статистических исследований не становится известно, что такую модель использовать нельзя).

Многомерная гауссовая модель определяется следующими двумя гипотезами:

1) Возможно преобразование исходной переменной в новую переменную, имеющую одномерное стандартное нормальное распределение:

$$Z(x) = \Phi\{Y(x)\} \text{ или } Y(x) = \Phi^{-1}\{Z(x)\},$$

где $Z(x)$ - исходная случайная величина; $Y(x)$ - случайная величина, имеющая стандартное нормальное распределение; Φ - строго возрастающая функция преобразования. Функция преобразования, известного как Normal Score преобразование, всегда существует. Она, обычно, определяется графически с помощью взаимно однозначного соответствия между функцией распределения случайной величины $Z(x)$ и стандартной нормальной функцией распределения.

2) Многомерное распределение случайной функции $Y(x)$ гауссово и строго стационарно.

Как следствие, случайная функция $Y(\mathbf{x})$ полностью определяется двумя параметрами:

1) ее средним $m(\mathbf{x})$:

$$E\{Y(\mathbf{x})\} = m(\mathbf{x}) = 0, \quad \forall \mathbf{x},$$

2) ее ковариационной функцией $C(\mathbf{h})$:

$$E\{Y(\mathbf{x}) \cdot Y(\mathbf{x} + \mathbf{h})\} = C(\mathbf{h}), \quad \forall \mathbf{x},$$

Заметим, что для гауссового распределения гипотеза строгой стационарности эквивалентна гипотезе стационарности второго порядка.

Свойство, которое делает гауссовую случайную функцию такой полезной моделью для геостатистической оценки и моделирования, заключается в том, что все условные распределения случайной функции $Y(\mathbf{x})$ также многомерно нормальные. Для пространственных данных это означает, что *локальное условное распределение* случайной величины в некоторой точке пространства (то есть функция распределения в некоторой точке при заданном множестве значений в ее окрестности) может быть описано с помощью лишь двух параметров: ее среднего (т.е. условного математического ожидания) и ее дисперсии (условной дисперсии).

Методы геостатистического моделирования

Основная задача геостатистики - реконструировать изучаемое явление в области исследования на базе его значений, измеренных в ограниченном числе точек. С математической точки зрения эта задача может рассматриваться как задача интерполяции. При классическом подходе неизвестная функция аппроксимируется параметрической функцией, чья форма задается заранее или явно (полином) или неявно (условие минимальной кривизны). Параметры выбираются так, чтобы оптимизировать некоторый критерий наилучшего приближения значений в точках, который может быть статистическим (наименьшие квадраты) или детерминированным (точное совпадение в точках измерения).

Геостатистика развивает другой подход, известный как *кригинг* (термин введен Матероном в 1963 г. по имени D.G.Krige). Кригинг строит скорее статистическую модель реальности, чем модель интерполяционной функции. Формальная аналогия здесь может быть проведена с решением задач предсказания в анализе временных рядов: по известным значениям в прошлом (обычно полученным через регулярные временные интервалы) требуется предсказать значение сигнала в некоторый момент в будущем. Сначала сигнал анализируется, как правило, вычисляется и моделируется спектр, а затем конструируется фильтр, т.е. предиктор. Кригинг следует подобному подходу, но в пространстве, где нет понятий прошлое и будущее. К тому же в отличие от временных процессов в пространственных явлениях данные обычно измерены неравномерно.

Теория геостатистического оценивания развивалась в рамках статистических моделей второго порядка, использующих лишь первые два момента случайной функции - среднее $m(\mathbf{x})$ и ковариацию $C(\mathbf{x}, \mathbf{y})$ или вариограмму $\gamma(\mathbf{x}, \mathbf{y})$, предполагая их известными. Поэтому ставшие классическими и реализованные в приложениях методы кригинга решают задачу интерполяции с применением линейных оценок. Это три основные формы кригинга, используемые, соответственно, при указываемых условиях:

- *простой* кригинг - математическое ожидание $m(\mathbf{x})$ известно;
- *ординарный* кригинг - математическое ожидание $m(\mathbf{x})$ неизвестно, но постоянно;
- *универсальный* кригинг - математическое ожидание $m(\mathbf{x})$ неизвестно и непостоянно.

Простой кригинг

Обозначим $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ точки выборки объема n , в которых известны значения изучаемого явления - $z_1=z(\mathbf{x}_1), z_2=z(\mathbf{x}_2), \dots, z_n=z(\mathbf{x}_n)$. Оценка значения пространственной переменной в некоторой точке \mathbf{x}_0 , где она не измерена, при известном матожидании есть:

$$[z(\mathbf{x}_0)]_{SK}^* = z_0^* = m(\mathbf{x}_0) + \sum_{i=1}^n \lambda_i [z_i - m(\mathbf{x}_i)],$$

где n весов λ_i даются *системой простого кригинга*:

$$\sum_{i=1}^n \lambda_i C(\mathbf{x}_i, \mathbf{x}_j) = C(\mathbf{x}_0, \mathbf{x}_j), \quad j = 1, \dots, n.$$

В случае гауссовой случайной функции эта оценка совпадает с условным матожиданием:

$$[z(\mathbf{x}_0)]_{SK}^* = E\{Z(\mathbf{x}_0) | z_1, \dots, z_n\},$$

которое является идеальной оценкой $z(\mathbf{x}_0)$ в смысле наименьших квадратов.

Система простого кригинга имеет единственное решение при условии, что функция ковариации $C(\mathbf{x}, \mathbf{y})$ положительно определенная и все выборочные точки различны; поиск решения – стандартная вычислительная задача.

Дисперсия оценки (или ее квадратный корень - стандартное отклонение) дает меру *ошибки* оценки кригинга:

$$\sigma_{SK}^2 = \text{Var}\{Z(\mathbf{x}_0) | z_1, \dots, z_n\} = E[(Z_0^* - Z(\mathbf{x}_0))^2] = C(0) - \sum_{i=1}^n \lambda_i C(\mathbf{x}_0, \mathbf{x}_i)$$

Следует отметить, что предположения, на которых основан метод простого кригинга, на практике почти никогда не выполняются. В реальности среднее может быть известно лишь тогда, когда либо имеются повторные наблюдения изучаемого явления в случае пространственно-временного процесса, либо среднее постоянно и число данных достаточно для его оценки.

Кригинг с неизвестным матожиданием

В практических ситуациях среднее $m(\mathbf{x})$ неизвестно. В этом случае может быть применен один из двух подходов.

Первый подход состоит в предварительной оценке $m(\mathbf{x})$ и вычитании этой оценки из данных (детрендрование). Такая предварительная оценка $m(\mathbf{x})$ может быть получена построением модели тренда $\hat{m}(\mathbf{x})$, т.е. регрессионной модели зависимости $z(\mathbf{x})$ от координат \mathbf{x} (собственный тренд) и/или от одной или нескольких других пространственных переменных, чье поведение на исследуемой территории известно (внешний тренд) (например, в геологии это может быть опорная поверхность).

Разности между данными и модельными значениями $e_i = z_i - \hat{m}(\mathbf{x}_i)$ называются *остатками* (residuals). Переход к ним сводит задачу к случаю постоянного среднего. Предполагается, что модель тренда $\hat{m}(\mathbf{x})$ достаточно точна для того, чтобы считать матожидание остатков константой (по крайней мере в окрестностях оцениваемого значения). Последующая оценка выполняется методом ординарного кригинга:

$$[e(\mathbf{x}_0)]_{OK}^* = e_0^* = \sum_{i=1}^n \lambda_i e_i,$$

который при поиске параметров λ_i решает *систему ординарного кригинга*, отличающуюся от системы простого кригинга тем, что в систему уравнений включается ограничение - сумма весов должна быть равна 1 (для получения несмещенной оценки отклонения от тренда). Оценка самой пространственной переменной получается как:

$$z_0^* = \hat{m}(\mathbf{x}_0) + e_0^*.$$

Таким образом, при этом подходе имеет место комбинация двух процедур оценки – моделирование тренда и моделирование остатков.

Второй подход: параметрическое оценивание $m(\mathbf{x})$ включается в уравнения кригинга, т.е. имеет место только один шаг оценки. Эта более общая модель, называемая универсальным кригингом, предполагает, что среднее может быть записано как сумма:

$$m(\mathbf{x}) = \sum_{l=0}^L a_l f_l(\mathbf{x}),$$

где $f_l(\mathbf{x})$ - известные базисные функции, a_l - неизвестные коэффициенты. Обычно первая базисная функция ($l=0$) - константа, а другие функции - полиномы невысокой степени от координат \mathbf{x} . Тогда:

$$[z(\mathbf{x}_0)]_{UK}^* = z_0^* = \sum_{l=0}^L a_l f_l(\mathbf{x}_0) + \sum_{i=1}^n \lambda_i [z_i - \sum_{l=0}^L a_l f_l(\mathbf{x}_i)],$$

и для поиска параметров λ_i и a_l решается *система универсального кригинга* (здесь не приводится).

Метод универсального кригинга не требует знания матожидания, а локально оценивает его. Такая оценка, несмотря на свою универсальность, зачастую может приводить к различным артефактам в местах стыковки локальных моделей. Поэтому оптимальным выбором при построении моделей представляется ординарный кригинг с предварительной оценкой тренда, особенно если такой тренд (что часто имеет место на практике) может быть оценен с использованием априорной информации о моделируемых объектах. Еще одним вариантом может быть применение ко-кригинга, когда используются функции не только от координат, как выше, но и от значений других переменных в точках оценки (например, влажность почвы оценивается с использованием отметок рельефа).

Параметры кригинга

Итак, что бы получить z_0^* , необходимо найти значения параметров λ_i , решив систему кригинга. Для этого должны быть известны значения функции ковариации $C(\mathbf{x}_i, \mathbf{x}_j)$, характеризующей связь между случайными величинами $Z(\mathbf{x}_i)$ и $Z(\mathbf{x}_j)$. Здесь $i, j=0, \dots, n$, причем $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ - точки выборки, а

точка x_0 – любая точка области исследования. Поэтому на самом деле необходимо знать $C(x,y)$, где x и y – две любые точки области исследования.

И вот здесь понадобятся свойства эргодичности и стационарности. Если для изучаемого явления выполняются гипотезы эргодичности и стационарности, то на основе выборочных данных можно оценить ковариацию $C(h)$ для некоторых h (повторяющихся в выборке). А затем, на основе полученных оценок $C^*(h)$ может быть построена модель $\hat{C}(h)$ для любого h , т.е. модель зависимости корреляции двух случайных величин $Z(x)$ и $Z(x+h)$, разделенных вектором h , от величины и направления в пространстве этого вектора. Эта модель $\hat{C}(h)$ и подставляется в систему кригинга.

Таким образом, ключевой задачей при геостатистическом моделировании является оценка моментов второго порядка (ковариации или вариограммы), характеризующих пространственную связь. Эта оценка $C^*(h)$ и построение модели пространственной связи $\hat{C}(h)$ выполняются в процессе *вариограммного анализа* данных. После чего одним из методов кригинга с использованием построенной модели ковариации $\hat{C}(h)$ (или модели вариограммы $\hat{\gamma}(h)$) может быть получена искомая оценка z_0^* самой изучаемой переменной.

Заметим, что оценка z_0^* зависит от значений данных $\{z_i\}$ и ковариационной модели $\hat{C}(h)$, а ошибка оценки σ^2 зависит только от пространственной конфигурации данных (т.е. от

взаимного расположения точек $\{x_i\}$) и ковариационной модели $\hat{C}(h)$ и не зависит от значений самих данных $\{z_i\}$.

Дополнительным параметром кригинга является, так называемый, радиус поиска (Search), который задает радиус круга с центром в точке x_0 , за пределом которого при оценке значения z_0 выборочные точки игнорируются; в более общем случае вместо круга используется эллипс. Если радиус поиска недостаточен, то в модели появляются различные «шумовые» артефакты, снижающие качество модели. Если же параметр выбран слишком большим, то, кроме существенного увеличения времени расчетов, модель получается необоснованной за счет влияния на результат расчетов слишком удаленных точек, «смазывающих» локальные детали. Поэтому оптимальным считается выбор значения этого параметра, основанный на вариограммном анализе и составляющий примерно два ранга наиболее генерализованной структуры в модели вариограммы.

Если на исследуемой территории имеют место подтвержденные линии существенных нарушений непрерывности моделируемой переменной, то такая информация при моделировании должна быть учтена. При оценке значения $z(x)$ в точке x_0 выборочные точки, находящиеся «по другую сторону» таких линий, игнорируются. Линии нарушений непрерывности задаются путем указания координат вершин представляющих их ломаных (векторное представление).

Оценка качества модели

Для оценки качества построенной модели принято использовать такой статистический метод, как «перекрестная проверка» (cross-validation). Суть метода состоит в разделении

исходных данных на «обучающую» часть (по которой строится модель), и «проверочную» (на которой оценивается качество модели путем сравнения прогноза с известными данными).

О качестве модели в точке x_0 также можно судить и по величине стандартной ошибки кригинга в этой точке, которая, как отмечалось выше, зависит только от пространственной конфигурации данных и модели пространственной корреляционной зависимости.

Картографическая визуализация модели

Для итогового отображения результатов геостатистического моделирования изучаемого пространственно распределенного явления используется построение на их основе различных карт. С этой целью обычно применяется растровая модель представления пространственных данных. То есть в качестве точки x_0 последовательно рассматриваются узлы (или центры ячеек) регулярной или нерегулярной сетки, которая полностью покрывает территорию исследования. Размеры ячейки задаются существенно меньшими характерного размера представляемого пространственного явления. Наиболее популярным является использование регулярных прямоугольных сеток, ячейки которых в картографической проекции являются квадратами или прямоугольниками.

Оценка значений пространственной переменной в ячейках растровой сетки, проведенная методами геостатистической интерполяции на основе кригинга, дает растровый слой. Для его картографического отображения (визуализации) можно воспользоваться, например, методом изолиний или методом попиксельного изображения в соответствии с заданной легендой,

либо импортировать его в тот или иной ГИС-пакет и использовать стандартные средства картографирования ГИС.

Кроме сеточной модели самой изучаемой переменной полезно построить и рассмотреть соответствующую ей сетку со значениями стандартной ошибки кригинга в каждой ячейке, которая отображает степень доверия к полученной модели или величину пространственной неопределенности.

Условия, необходимые для применения геостатистических методов

Из вышеизложенного следует, что применение геостатистических методов для численного описания (оценки, моделирования) явлений, распределенных в пространстве, предполагает выполнение ряда условий: случайная функция $Z(\mathbf{x})$ должна быть стационарной, эргодичной, а также иметь многомерный нормальный (гауссовый) закон распределения.

Выполнение условий эргодичности и стационарности позволяет получать статистические оценки из единственной реализации случайной функции, то есть на основе выборочных значений пространственной переменной. Мультиформальность позволяет при определении параметров модели ограничиться статистической оценкой только среднего и ковариации.

Может ли изучаемая пространственная переменная $z(\mathbf{x})$ считаться реализацией стационарной эргодичной гауссовой случайной функции? Ответу на этот вопрос должны быть посвящены первые шаги любого геостатистического исследования.

Известно, что свойство эргодичности не может быть проверено на основе выборки из одной единственной уникальной

реализации $z(x)$, какой является действительность. Поэтому в рамках выбранной вероятностной модели случайная функция $Z(x)$ предполагается эргодичной.

Не существует статистической проверки, которая могла бы опровергнуть в целом гипотезу о стационарности случайной функции $Z(x)$; считается, что всегда можно построить стационарную случайную функцию, имеющую реализацию, соответствующую наблюдаемым $z(x)$. На практике отсутствие стационарности обычно проявляется наличием в данных пространственного тренда (дрифта), что указывает на нестационарность математического ожидания - $E[Z(x)] = m(x)$, или тем, что данные показывают наличие в области исследования неких, отличных друг от друга, зон. Пространственная переменная $z(x)$ может считаться реализацией стационарной случайной функции, если это обосновывается пространственной однородностью данных, а также, следует из физического смысла явления и априорных знаний о нем.

Гипотеза о мультинормальности случайной функции $Z(x)$ предполагает, что каждая случайная величина $[Z(x)]$, где x любая точка области исследования/ распределена нормально. Она также предполагает, что каждая пара случайных величин $[Z(u), Z(v)]$, где u и v любые точки/ имеет двумерное нормальное распределение, каждая тройка случайных величин $[Z(u), Z(v), Z(w)]$, u , v и w любые точки/ имеет трехмерное нормальное распределение и так далее. Полностью такая гипотеза, так же как и гипотеза стационарности, не может быть проверена. Однако многомерное нормальное распределение имеет ряд свойств, которые могут быть использованы для проведения некоторых проверок:

1) Так как любая линейная комбинация мультинормальных функций также нормально распределена, проверка на бинормальность заключается в проверке того, что разность $[Y(\mathbf{x}) - Y(\mathbf{x} + \mathbf{h})]$ распределена нормально со средним 0 и дисперсией $2\gamma(\mathbf{h})$.

2) Для многомерной нормальной случайной функции условная дисперсия не зависит от условного среднего. Как следствие, мультинормально распределенные данные не должны показывать такую зависимость, то есть не должны показывать гетероскедастичность (эффект пропорциональности).

Эти две процедуры вполне эффективны и просты в реализации. Заметим, что на практике обычно проверяется только двумерная нормальность случайной функции $Z(\mathbf{x})$. Проверка трех-, четырех-, k -мерной нормальности требует множества k -мерных выборок при одинаковой пространственной конфигураций данных. Это неосуществимо на практике, если только данные не расположены на «густой» регулярной сетке. Поэтому обычно гипотезу о многомерной нормальности случайной функции принимают при согласии эмпирического распределения выборочных данных с одномерным нормальным распределением, и если при проведении, по крайней мере, одной из двух указанных выше проверок не опровергается ее бинормальность.

Противоречие гипотезе стационарности или гипотезе мультинормальности может также быть выявлено и на стадии вариограммного анализа данных. Ниже будет показано, какое поведение оценки ковариации $C^*(\mathbf{h})$ (или оценки вариограммы $\gamma^*(\mathbf{h})$) противоречит этим гипотезам.

Если какая-либо из указанных гипотез в результате проверок была отвергнута, то для возможности проведения дальнейшего анализа (в рамках линейной геостатистики) исходная переменная может быть преобразована к некоторой новой переменной, удовлетворяющей нужным условиям.

В следующих разделах более подробно представлены методы предварительного анализа и обработки пространственных данных, призванные проверить и обеспечить выполнение условий, необходимых для применения геостатистических методов, а также подробно описаны этапы проведения вариограммного анализа данных, в результате которого должна быть получена модель пространственной связи $\hat{C}(h)$ (или $\hat{\gamma}(h)$) для дальнейшей подстановки ее в систему кригинга и получения искомых оценок изучаемой пространственной переменной.

В библиотеках среды статистического программирования **R** присутствуют процедуры, реализующие все необходимые методы предварительного анализа и обработки данных, вариограммного анализа данных, а также геостатистического оценивания на основе кригинга. Кроме этого, методы кригинга реализованы в целом ряде программных продуктов. В частности, это пакет пространственного анализа и моделирования **Surfer**; в котором представлены и средства картографической визуализации результатов моделирования.

В настоящее время методы кригинга активно развиваются, и в первую очередь в направлении расширения допустимых форм нестационарности, многомерного оценивания (кокригинг), нелинейного оценивания.

Предварительный анализ и обработка пространственных данных

В этом разделе рассматривается пошаговый подход к анализу пространственных данных. Предлагается строго придерживаться следующей последовательности действий, предшествующих проведению вариограммного анализа:

- 1) Визуализация выборки на базовой карте.
- 2) Декластеризация данных (при необходимости).
- 3) Проверка ограничений и предположений геостатистики методами одномерной и многомерной статистики, и выполнение (при необходимости) процедур преобразования данных.

1) Визуализация выборки на базовой карте

Одна из наиболее значимых характеристик выборочных данных в науках о Земле - их локализация (положение в пространстве). На начальном этапе анализа пространственных данных элементы выборки должны быть изображены на карте для того, чтобы уловить основные пространственные особенности изучаемого феномена.

Эффективным средством для этого является, так называемая, базовая карта, где символы, представляющие данные, имеют размер или цвет, соответствующий значению данных. Такая карта может быть использована для проведения следующих проверок:

- выявления ошибок локализации данных (проверка координат);

- выявления ошибок в самих данных: невысокие значения, окруженные высокими значениями, и наоборот, могут быть ошибочными (они должны быть исправлены) или могут отражать другой феномен, отличный от изучаемого (таковые не должны браться в рассмотрение);

- обнаружения кластеров (скоплений);

- выявления основных пространственных черт изучаемого явления: трендов, однородных и неоднородных зон, предположительных направлений анизотропии.

Средства для построения базовых карт имеют пакеты Surfer, MapInfo и др. Покажем, как это можно сделать средствами R.

Исходные данные

Табличный набор данных `stud.dat`. Два столбца с именами X и Y содержат координаты выборочных точек x_1, x_2, \dots, x_n (где $n=515$); каждый из столбцов с именами V01, V02, ..., V13 содержит значения в этих точках некоторой пространственной переменной - $z_1=z(x_1), z_2=z(x_2), \dots, z_n=z(x_n)$. Всего таких переменных тринадцать. Пусть нас интересует пространственная переменная $z(x)$, значения которой в выборочных точках даны в столбце V03. Эти данные будем использовать в рассмотренных ниже примерах.

Чтение табличных данных:

```
p=read.table(file="stud.dat", header=TRUE, sep="\t",  
             na.strings="-99999")
```

`read.table()` - чтение табличных данных, `file="stud.dat"` - имя файла, `header=TRUE` - в первой строке заголовок, `sep="\t"` - разделитель табуляция (если данный ключ пропущен, то пробелы и табуляции), `na.strings="-99999"` - какое значение означает отсутствие данных.

Выборка данных:

```
p=p[,c("X","Y","V03")] # выборка только нужных данных
```

Наносим данные на карту, размеры значков делаем пропорциональными значениям в V03:

```
plot(p$X,p$Y,сex=0.2+2*(p$V03-min(p$V03))/(max(p$V03)-  
min(p$V03)),main="Положение на карте",xlab="X",  
ylab="Y",asp=1)
```



Рис.3. Базовая карта выборки

Полученная базовая карта показывает наличие в данных пространственного тренда, а также присутствие кластера.

2) Декластеризация

Образование кластеров (пространственных сгущений) элементами выборки связано с особенностями сбора данных в

исследуемой области пространства. Если данные не образуют пространственных кластеров, то каждый элемент выборки несет то же количество информации, что и любой другой. В противном случае кластеризованные данные (т.е. взятые из одного кластера) являются избыточными, и несут меньше информации, чем некластеризованные. Такая избыточность может быть учтена взвешиванием - использованием меньших весов для кластеризованных данных, чем для некластеризованных.

Наличие кластеров в выборке существенно влияет на оценку параметров генеральной совокупности (среднего, дисперсии, вариограммы), а именно приводит к смещению оценок. Например, если элементы выборки кластеризуются в области более высоких значений, выборочное среднее даст завышенную оценку среднего генеральной совокупности.

Обнаружение кластеров

Наличие кластеров обычно обнаруживается визуально при рассмотрении базовой карты выборки.

Количественный подход к обнаружению кластеров состоит в наложении на базовую карту регулярной сетки с размером ячеек, примерно равным среднему расстоянию между элементами выборки (возможно использование сетки с перекрывающимися ячейками). Затем строится график, где для каждой ячейки по одной оси откладывается среднее значение выборочных данных, попавших в ячейку, по другой - число точек выборки в ячейке. Такой график позволяет обнаружить наличие кластеров в области высоких и/или низких значений.

Методы декластеризации

При декластеризации, необходимой для правильной оценки параметров распределения, применяется несколько методов, из них наиболее распространены два: 1) полигональная декластеризация; 2) клеточная декластеризация. Оба метода обеспечивают определение весов, приписываемых каждому элементу выборки. В дальнейшем одномерные параметры генеральной совокупности (математическое ожидание, дисперсия) оцениваются с учетом этих весов: чем больше вес, тем больше вклад данного элемента выборки в оценку.

Метод *полигональной декластеризации* связывает с каждым элементом выборки вес, пропорциональный площади его многоугольника влияния (каждая точка внутри такого многоугольника ближе к данному элементу, чем к любому другому из той же выборки). Сумма площадей масштабируется единицей, и кластеризованные точки получают меньшие веса, чем некластеризованные (рис.4). Отметим, что при использовании данного метода необходимо решить задачу построения границ многоугольников влияния.

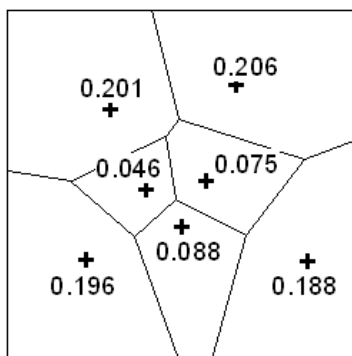


Рис.4. Полигональная декластеризация. Показанные числа есть декластеризующие веса

Метод *клеточной декластеризации* делит всю область выборки на прямоугольные ячейки, и каждая точка выборки получает вес, обратно пропорциональный числу точек, попавших в ту же ячейку (рис.5).

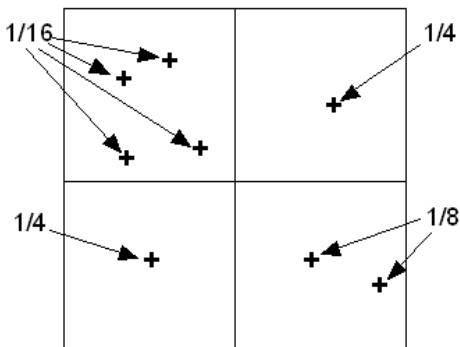


Рис.5. Клеточная декластеризация. Показанные числа есть декластеризующие веса

Вес, приписываемый i -й точке, вычисляется как:

$$w_i = 1/(N \cdot n_i),$$

где N - общее число ячеек, n_i - число точек выборки, попавших в i -ую ячейку. Поскольку сумма всех весов равна единице, а элементы выборки, попавшие в одну ячейку, получают одинаковые веса, далее оценивание получают в два шага:

- 1) в каждой ячейке оценивается локальное математическое ожидание как среднее попавших в нее n_i точек выборки.
- 2) оценка глобального математического ожидания есть среднее из оценок локальных математических ожиданий по N ячейкам.

Если кластеризация наблюдается в области с высокими значениями, размер ячеек выбирают как дающий наименьшее глобальное среднее. Если кластеризация наблюдается в области

низких значений, размер ячеек выбирается как дающий наибольшее значение глобального среднего. Если кластеризация наблюдается в обеих областях, размер ячеек для декластеризации выбирают так, чтобы в областях, где нет кластеризации, в ячейку попадала одна точка данных.

Значительно труднее получить декластеризованные оценки для параметров двумерных распределений (полувариограмма, коррелограмма, ковариация и т.д.). В этом случае должны взвешиваться не кластеризованные точки выборки, а кластеризованные пары точек. Если есть основания полагать, что кластеризация оказывает существенное влияние на оценку вариограммы, то можно «сломать» кластеры, взяв для расчетов небольшое число точек из каждого кластера.

Клеточная декластеризация может быть выполнена в R с помощью пользовательской функции:

```
my.declas(cur.data,cur.X,cur.Y,cur.var,minmax.flag,  
          n.size,w.size0,w.size1,w.nshift)
```

Параметр	Назначение
<code>cur.data</code>	таблица с данными
<code>cur.X</code>	имя столбца с координатой X
<code>cur.Y</code>	имя столбца с координатой Y
<code>cur.var</code>	имя столбца с переменной, для которой вычисляются средние
<code>minmax.flag</code>	0-при кластеризации в области высоких значений; 1-при кластеризации в области низких значений
<code>n.size</code>	число различных размеров ячейки
<code>w.size0</code>	минимальный размер ячейки
<code>w.size1</code>	максимальный размер ячейки
<code>w.nshift</code>	число сдвигов (5 – на 20% ячейки)

Эта функция, текст которой приведен в приложении 1, последовательно рассчитывает средние в локальных окнах

(ячейках), покрывающих территорию, и среднее среди средних (размер окон меняется от минимального до максимального).

Обратите внимание, что минимальный, максимальный размеры ячеек усреднения и число шагов (сдвигов) должны соответствовать размерам области, занятой данными. Визуальный анализ базовой карты (рис.3) показывает, что кластеризация наблюдается в области низких значений, поэтому для нашего примера ищем размер ячеек, дающий наибольшее значение глобального среднего (декластеризованный максимум):

```
p<-my.declus(p,"X","Y","V03",1,26,500,15000,5)
```

Функция выводит в текстовый файл **report.txt** следующую информацию (отчет):

```
minmax flag = 1
ncell min max = 26 500.000000 15000.000000
offsets = 5
There are      515 data with:
mean value      = -840.728932
minimum and maximum = -967.200000 -692.200000
size of data vol in X = 57706.000000
size of data vol in Y = 69729.000000
declustered mean  = -826.977385
min and max weight = 0.071460 2.163866
```

В данном случае наибольшая декластеризованная оценка среднего равна -826.977385, для нее декластеризующие веса лежат в диапазоне от 0.071460 до 2.163866.

Файл результатов **data_V03_w.dat** будет содержать исходные данные, к которым добавлен столбец полученных весов:

X	Y	V03	Weight
37183	19442	-873	0.2754
37978	19649	-872.5	0.31018
39318	17332	-874.7	0.24659
39480	16199	-857.3	0.14325
39489	15475	-857.5	0.22408

...

Кроме этого, функция выводит в текстовый файл **Mean_vs_window_size.txt** информацию, позволяющую проанализировать соответствие размера ячеек и оценки среднего:

Size	Mean
0	-840.729
500	-831.794
1080	-826.977
1660	-827.853
2240	-829.382
2820	-830.989
3400	-832.088
3980	-833.872
4560	-835.361
5140	-836.065
5720	-836.836
6300	-837.339
6880	-838.885
7460	-838.89
8040	-840.229
8620	-841.158
9200	-841.167
9780	-840.94
10360	-842.164
10940	-842.008
11520	-843.706
12100	-842.307
12680	-842.307
13260	-842.667
13840	-843.124
14420	-843.112
15000	-844.018

Средствами R построим график зависимости среднего от размера ячеек, используемых для декластеризации:

```
tmp<-read.table("Mean_vs_window_size.txt",header=TRUE)
plot(tmp$Size,tmp$Mean,xlab="Размер ячеек",
      ylab="Среднее", main="Декластеризованное среднее")
```

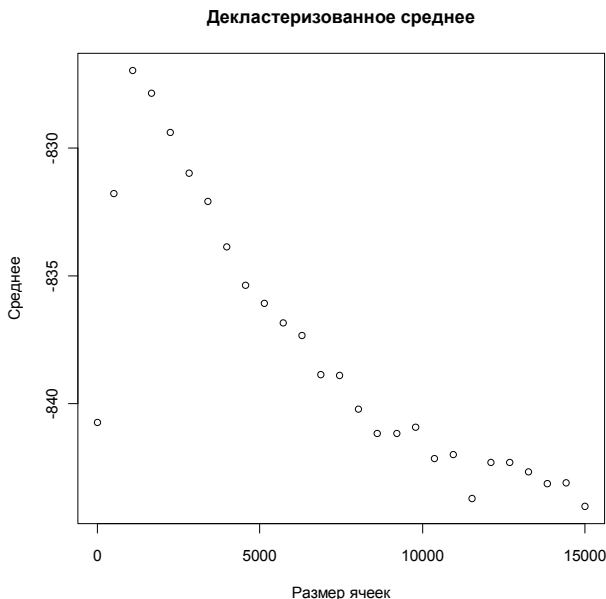


Рис.6. Зависимость оценки среднего от размера ячеек декластеризации

Видно, что исходная оценка среднего (при нулевом размере ячеек) занижена, и, по-видимому, фактическое значение несколько выше. Оценка растёт, достигая максимума -826.977 при размере ячеек 1080. Для более детального анализа полезно рассмотреть распределение весов в пространстве, отобразив их значения на базовой карте с помощью команды:

```
plot(p$X,p$Y,сех=2+ ((p$Weight-mean(p$Weight))/
    sd(p$Weight)),main="Beca",xlab="X",ylab="Y",asp=1)
```

или же используя следующую конструкцию:

```
library(sp)
tmp=p
coordinates(tmp)=~X+Y
bubble(tmp,zcol="Beca",fill=FALSE,
    col=c("black","black"), scales=list(draw=TRUE))
```

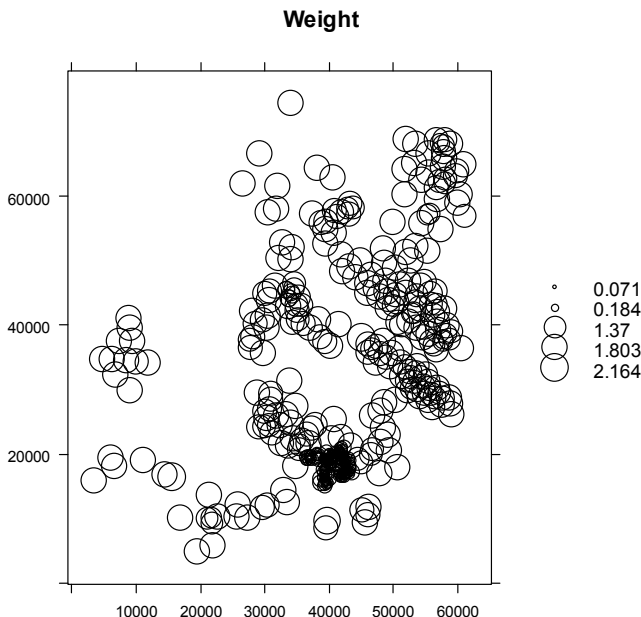


Рис.7. Базовая карта декластеризующих весов

Анализ распределения весов показывает, что в целом по территории веса распределены равномерно, за исключением места сгущения выборочных точек, где веса значительно меньше, и на краях области, где веса несколько увеличены, что соответствует плотности точек. Декластеризацию можно считать успешно выполненной.

3) Проверка ограничений и предположений геостатистики

Здесь остановимся на том, какие методы одномерной и многомерной статистики помогут провести анализ пространственных данных на предмет выполнения условий, необходимых для применения геостатистических методов, а именно:

- данные не противоречат гипотезе стационарности;
- данные не противоречат гипотезе мультиформальности.

Статистическое описание данных

Изложим кратко, что включает в себя описательная одномерная и многомерная статистика. Подробная информация об этом предмете имеется в многочисленных книгах по математической статистике.

Описательная одномерная статистика обеспечивает простой путь для организации и систематизации выборочных данных. Статистическое описание данных в общем случае начинается с анализа *гистограммы*, которая есть график *таблицы частот*, то есть таблицы, регистрирующей, как часто наблюдаемые данные попадают в определенные классы (интервалы значений). Таблица частот и гистограмма дают представление о виде закона распределения изучаемой величины. Кроме этого, на основе анализа гистограммы, в первую очередь, определяются выбросы или ошибки в данных, которые либо корректируются, либо удаляются из анализируемой выборки.

Для числовых данных важнейшие особенности большинства распределений могут быть представлены с помощью нескольких показателей описательной статистики. Эти показатели можно разбить на три группы:

1) характеристики положения – описывают положение данных на числовой оси - выборочные среднее, медиана, мода, минимум, максимум, квартили, квантили;

2) характеристики рассеяния - описывают степень разброса данных относительно своего центра – выборочные дисперсия, среднеквадратическое отклонение, размах выборки;

3) характеристики формы – выборочные коэффициент асимметрии, коэффициент эксцесса, положение выборочной медианы относительно выборочного среднего и относительно выборочных квартилей.

Статистическое описание выборочных данных должно предварять любое фундаментальное статистическое исследование.

Двумерная статистика выявляет взаимосвязи между двумя случайными величинами и их распределениями. Графическое представление элементов выборок, когда по одной из осей откладываются значения из первой выборки, а по другой – соответствующие значения из второй выборки, так называемый *scatterplot*, позволяет напрямую сравнивать две величины. Связь, видимая на такой диаграмме, может быть оценена с помощью коэффициента корреляции или рангового коэффициента корреляции. Графическое изображение квантилей одного распределения против квантилей другого, так называемый *q-q график*, служит для сравнения двух распределений. Если два распределения одинаковы, точки, представляющие квантили, будут изображены как прямая линия под углом 45° . Особенный интерес представляют графики *нормальной* и *логнормальной вероятности*, которые сравнивают квантили экспериментального распределения с квантилями нормального или логнормального распределения.

Многомерная статистика обеспечивает способы систематизации многомерных выборочных данных. Основным инструментом здесь является *ковариационная матрица*, которая отражает особенности изменчивости выборочных данных и взаимосвязи между переменными. Анализ главных компонент,

кластерный анализ, каноническая корреляция и дискриминантный анализ относятся к таким многомерным статистическим методам, чья цель обобщение выборочных данных путем выделения небольшого числа линейных комбинаций исходных переменных, «объясняющих» большую часть изменчивости данных (снижение размерности признакового пространства).

Средствами R построим гистограмму и подсчитаем статистические показатели:

```
hist(p$V03,ylab="Частота",xlab="V03",
      main="Гистограмма V03")
```

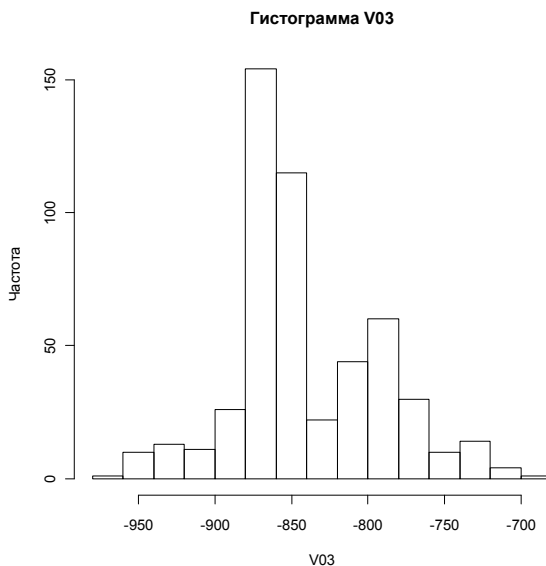


Рис.8. Гистограмма частот

Статистики по V03:

<code>V03.mean</code>	<code><- mean</code>	<code>(p\$V03, na.rm=TRUE)</code>	# среднее
<code>V03.median</code>	<code><- median</code>	<code>(p\$V03, na.rm=TRUE)</code>	# медиана
<code>V03.var</code>	<code><- var</code>	<code>(p\$V03, na.rm=TRUE)</code>	# дисперсия
<code>V03.sd</code>	<code><- sd</code>	<code>(p\$V03, na.rm=TRUE)</code>	# СКО

```
library(e1071)
V03.skewness <- skewness(p$V03,na.rm=TRUE) # асимметрия
V03.kurtosis <- kurtosis(p$V03,na.rm=TRUE) # эксцесс
cat(V03.mean,V03.median,V03.var,V03.sd,V03.skewness,
    V03.kurtosis,"\n")
-840.729 -855.6 2339.672 48.37016 0.4274284 0.1538570
```

Статистики по V03 с учетом декластеризующих весов:

```
V03.wt <- cov.wt(p,p$Weight)
# среднее и дисперсия
cat(V03.wt$center[3],V03.wt$cov[3,3],"\n")
-826.9774 3321.654
```

Статистики по трем столбцам таблицы:

```
summary(p[1:3])
```

X	Y	V03
Min. : 3426	Min. : 4731	Min. : -967.2
1st Qu.: 38918	1st Qu.: 18476	1st Qu.: -869.4
Median : 41232	Median : 20220	Median : -855.6
Mean : 41655	Mean : 29217	Mean : -840.7
3rd Qu.: 47322	3rd Qu.: 39935	3rd Qu.: -803.0
Max. : 61132	Max. : 74460	Max. : -692.2

Квантили по V03:

```
quantile(p$V03,c(0,0.025,0.050,0.95,0.975,1))
0%      2.5%      5%      95%      97.5%      100%
-967.200 -931.855 -917.780 -754.560 -728.435 -692.200
```

Построим график, отражающий соотношение медианы, интерквартильного размаха выборки, наименьшего и наибольшего выборочных значений (график «ящик с усами», см. рис. 9):

```
library(lattice)
bwplot(p$V03,xlab="V03") # ящик с усами
```

На графике левая граница прямоугольника показывает 25% квартиль выборочного распределения, правая - 75% квартиль. Расстояние между ними (ширина прямоугольника) – это интерквартильный размах выборки. Точка внутри прямоугольника показывает положение медианы. Тонкие

вертикальные линии («усы») показывают минимум и максимум выборки в том случае, если выборка не содержит выбросов (аномальных значений). Иначе эти линии указывают расстояние от «ящика», равное 1.5 его ширины (интерквартильного размаха), и точки за ними могут быть интерпретированы как выбросы.

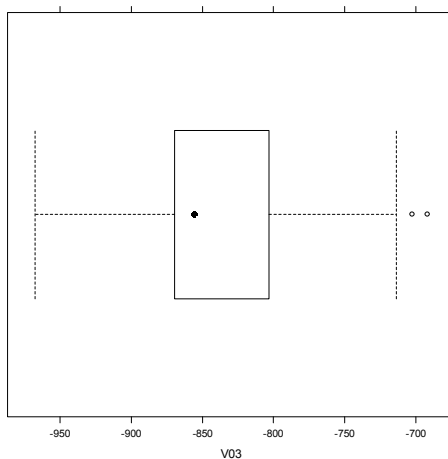


Рис.9. График «ящик с усами»

Судя по гистограмме и графику «ящик с усами», выбросы в данных присутствуют в области высоких значений. Если в пространстве (рис.3) эти аномальные значения расположены равномерно, без образования скоплений, то это может сильно исказить дальнейшую оценку пространственной зависимости. Поэтому их лучше просто исключить из выборки при вычислениях оценок. Например, отсечь по 2.5% двухсторонним квантилям, сохранив результат в колонке с именем V03q:

```
p$V03q<-pmax(quantile(p$V03,0.025),
              pmin(quantile(p$V03,0.975),p$V03))
write.table(p,"data_V03_wq.dat",row.names=FALSE,
            quote=FALSE,sep="\t")
```

Анализ пространственного тренда

Если базовая карта выборки показывает возможное наличие в данных пространственного тренда, то этот факт должен быть проверен статистическими методами, так как в подобном случае данные будут противоречить гипотезе стационарности.

С этой целью сначала выявляется взаимосвязь между значениями исследуемой переменной $z(x)$ и координатами в двухмерном пространстве с использованием различных показателей (мер) корреляционной связи.

Если достигаемый уровень значимости (p-value) оценок коэффициентов корреляции меньше 0.05, то это говорит о наличии статистически значимой ненулевой корреляции между двумя переменными с 95%-й надежностью, т.е. переменные достоверно связаны корреляционной зависимостью. Такое поведение изучаемой переменной противоречит гипотезе стационарности матожидания $m(x)$. Исправить ситуацию может ее преобразование к некоторой новой переменной, удовлетворяющей нужной гипотезе, для чего необходимо построить модель тренда $\hat{m}(x)$, например, методом множественной линейной регрессии.

Средствами R оценим линейную корреляцию переменной V03 с координатами:

```
V03.corX <- cor.test(p$V03q,p$X,  
                    alternative=c("two.sided"),  
                    method=c("pearson"),  
                    conf.level=0.95)  
V03.corY <- cor.test(p$V03q,p$Y,  
                    alternative=c("two.sided"),  
                    method=c("pearson"),  
                    conf.level=0.95)  
cat("corX=", V03.corX$estimate,"t=",
```

```

V03.corX$statistic,"df=", V03.corX$parameter,"
p.value=", V03.corX$p.value,"\n")
corX= 0.7579349 t= 26.31619 df= 513 p.value= 0
cat("conf.int=", V03.corX$conf.int,"\n")
conf.int= 0.7185908 0.7924425
cat("corY=", V03.corY$estimate,"t=",
V03.corY$statistic,"df=", V03.corY$parameter,"
p.value=", V03.corY$p.value,"\n")
corY= 0.7995807 t= 30.15543 df= 513 p.value= 0
cat("conf.int=", V03.corY$conf.int,"\n")
conf.int= 0.7661049 0.8287299

```

Полученные оценки указывают на сильную статистически значимую связь изучаемой переменной с координатами, т.е. на пространственный тренд. Построим модель тренда с использованием линейной регрессии (строим билинейную модель), учитывая декластеризующие веса:

```

p.lm<-lm(V03q~X*Y,weights=Weight,data=p)
summary(p.lm)

```

Call:

```
lm(formula = V03q ~ X * Y, data = p, weights = Weight)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-64.225	-5.715	0.389	6.218	70.706

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.009e+03	5.766e+00	-174.958	<2e-16	***
X	2.667e-03	1.399e-04	19.062	<2e-16	***
Y	2.080e-03	1.649e-04	12.610	<2e-16	***
X:Y	-3.525e-09	3.693e-09	-0.955	0.34	

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

```

Residual standard error: 15.99 on 511 deg.of freedom

Multiple R-squared: 0.9162, Adjusted R-squared: 0.9157

F-statistic: 1862 on 3 and 511 DF, p-value: < 2.2e-16

Доля объясненной дисперсии составляет 91.6%, поэтому считаем тренд статистически значимым (доля объясненной дисперсии велика) и полученную модель тренда адекватной.

В явном виде:

$$\hat{m}(x_i) = -1009 + 0.002667 \cdot x + 0.002080 \cdot y - 3.525 \cdot 10^{-9} \cdot x \cdot y$$

Вычтем модельные значения из данных (элиминация тренда), получаем *остатки*: $e_i = z_i - \hat{m}(x_i)$, которые запишем в новый столбец Res и сохраним в файл:

```
p$Res<-residuals(p.lm) # добавляем столбец с остатками
write.table(p,"data_V03_wqr.dat",sep="\t",
            row.names=FALSE,quote=FALSE)
```

Далее проводится анализ остатков по указанной выше схеме (базовая карта, статистическое описание: гистограмма, и т.д.), и все дальнейшие методы анализа уже будут применяться не к самой переменной, а к остаткам.

Построим гистограмму для остатков и вычислим основные статистики:

```
hist(p$Res,ylab="Частота",xlab="Res",
     main="Гистограмма для остатков")
```

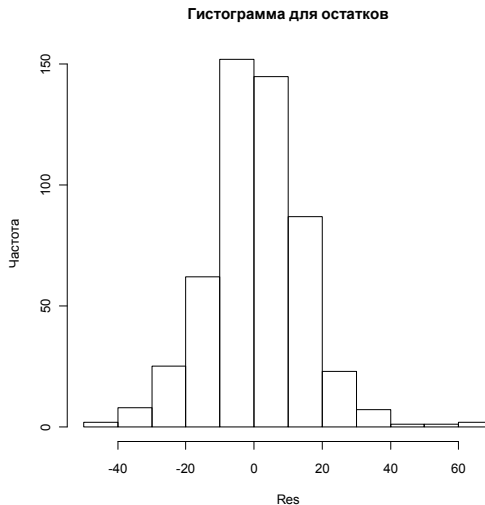


Рис.10. Гистограмма частот для остатков

Статистики по Res:

```
Res.mean      <- mean      (p$Res,na.rm=TRUE) # среднее
Res.median    <- median    (p$Res,na.rm=TRUE) # медиана
Res.var       <- var       (p$Res,na.rm=TRUE) # дисперсия
Res.sd        <- sd        (p$Res,na.rm=TRUE) # СКО
Res.skewness  <- skewness  (p$Res,na.rm=TRUE) # асимметрия
Res.kurtosis  <- kurtosis  (p$Res,na.rm=TRUE) # эксцесс
cat(Res.mean, Res.median, Res.var, Res.sd,Res.skewness,
    Res.kurtosis,"\n")
0.9486555 0.5195577 195.4194 13.97925 0.2361028 1.97978
```

Иногда для исследуемой территории может быть дано пространственное распределение некоторой другой известной пространственной переменной (в геологии - опорная поверхность). В этом случае на выборочных точках проверяется наличие корреляционной зависимости между исследуемой и известной переменной, и если эта связь существует и значима, то известная переменная может быть включена как еще один из регрессоров в модель тренда, что часто улучшает ее качество.

Кроме наличия пространственного тренда, гипотезе стационарности может противоречить поведение данных (в том числе и остатков), обусловленное влиянием на их однородность априори известных структурных нарушений, природных границ и пр. на территории исследования. В этом случае данные кодируются номерами предполагаемых зон или подобластей, которым они принадлежат, и гипотеза об однородности может быть проверена, например, с помощью дисперсионного анализа (где фактором будет номер зоны). В случае выявления неоднородности, определяются правила выделения однородных областей данных, выборка разбивается на подвыборки (однородные подмножества данных) и вся дальнейшая работа проводится только внутри этих областей.

Проверка мультинормальности

Проверка нормальности одномерного распределения

Существует несколько тестов (критериев согласия), которые позволяют проверить гипотезу о нормальности распределения. К ним относятся критерии хи-квадрат, Колмогорова-Смирнова, критерии асимметрии и эксцесса и др.

Для проверки гипотезы необходимо задать уровень значимости - это допустимая для данной задачи вероятность ошибки 1-ого рода при проверке гипотезы по статистическому критерию, т.е. вероятность отклонить нулевую гипотезу (в нашем случае отвергнуть гипотезу о нормальности распределения), когда на самом деле она верна. Стандартные значения уровня значимости: 0.005; 0.01; 0.05; 0.1.

Достижимый уровень значимости (p-value) - это значение функции распределения критерия для его наблюдаемого значения. Чем меньше получаемое значение p-value, тем сильнее совокупность данных свидетельствует против нулевой гипотезы.

Достижимый уровень значимости (p-value) сравнивается с заданным уровнем значимости. Если p-value ниже заданного уровня значимости, то гипотеза отвергается. Если выше, то принимается.

В нашем примере, судя по гистограмме остатков, вид распределения можно считать приблизительно нормальным (хотя асимметрия и эксцесс отличны от 0, их значения не превышают 2).

Построим q-q график (квантильный график) для сравнения квантилей экспериментального распределения с квантилями нормального распределения.

```
qqnorm(p$Res, xlab="Теоретические
```

```
квантили", ylab="Выборочные квантили")
qqline(p$Res, col="black", lwd=2)
```

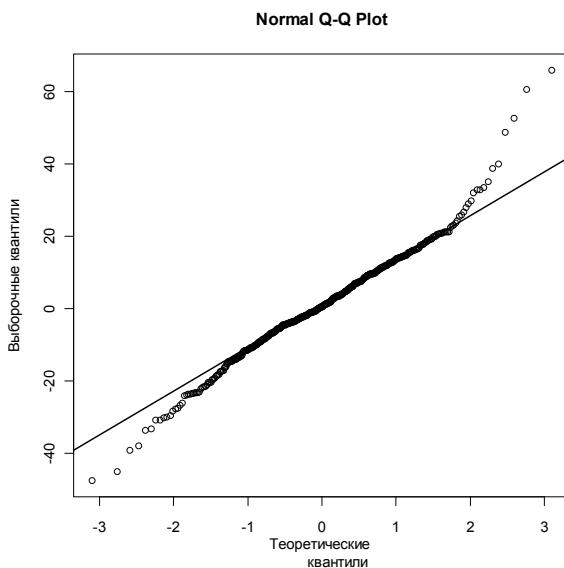


Рис.11. q-q график

Проверим гипотезу одномерной нормальности с помощью критерия Колмогорова-Смирнова:

```
ks.test(p$Res, "pnorm", mean(p$Res), sd(p$Res))
```

One-sample Kolmogorov-Smirnov test

data: p\$Res

D = 0.0464, p-value = 0.2172

alternative hypothesis: two-sided

В данном случае нет оснований отвергнуть гипотезу о нормальном распределении (достигаемый уровень значимости равен 21.7%).

Если данные не подчиняются закону нормального распределения, их можно трансформировать таким образом, что распределение приблизится к нормальному. При правосторонней асимметрии («хвост» вправо) чаще всего применяют следующие

виды «нормализующей трансформации»: извлечение квадратного корня \sqrt{x} , логарифмическое преобразование $\text{Ln}(x)$ или $\text{Log}_{10}(x)$, гармоническое преобразование $(-1/x)$ (знак «минус» используется для сохранения направленности отношений; в противном случае наименьшие и наибольшие значения поменяются местами, что может затруднить интерпретацию результатов). Если переменная может принимать нулевое значение, то для проведения логарифмического или гармонического преобразования следует добавлять к переменной некоторое малое число, например 0.001. При левосторонней асимметрии имеет смысл преобразовывать данные путем их возведения в степень (обычно во вторую или третью). Выбор наиболее подходящего вида трансформации для имеющихся данных определяется методом проб и ошибок, а об успешности преобразования смотрят по графикам, коэффициентам асимметрии и эксцесса и результатам проверки распределения с помощью статистических критериев.

Проверка гомоскедастичности (отсутствие эффекта пропорциональности)

Локальная дисперсия может быть постоянной или изменяться при перемещении по изучаемой области. Рис.12 иллюстрирует два случая, когда локальная дисперсия является константой. В первом случае локальное среднее тоже является константой, а во втором оно подвержено изменению (тренду). Если локальная дисперсия примерно постоянна, то оценка в произвольном месте области будет иметь одинаковую точность.

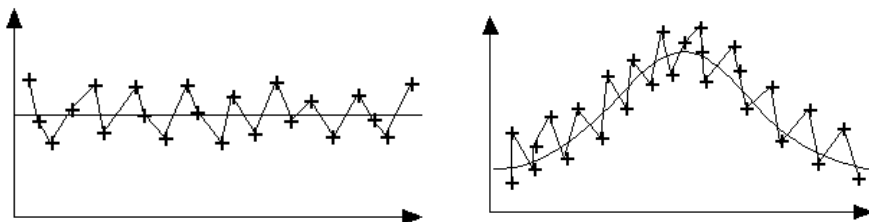


Рис.12. Два примера с постоянной локальной дисперсией

Рис.13 соответствует двум случаям с тем же поведением локальных средних, когда локальная дисперсия непостоянна.

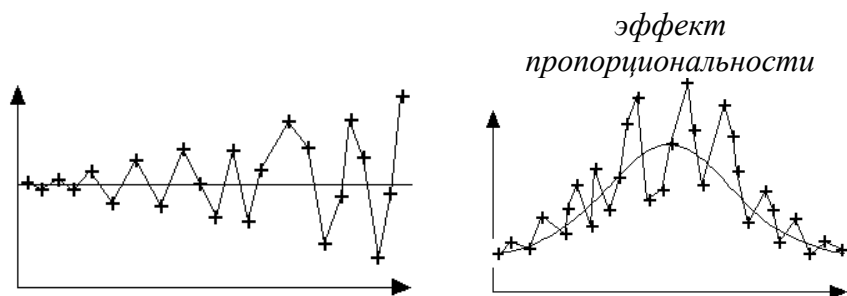


Рис.13. Два примера с непостоянной локальной дисперсией

Оценка в областях с большей локальной дисперсией менее надежна, чем оценка в областях с меньшей локальной дисперсией. Случай, когда локальная дисперсия пропорциональна локальному среднему, представляет особый интерес, поскольку локальная дисперсия может быть предсказана по локальному математическому ожиданию. Наличие такой связи, так называемый *эффект пропорциональности* или *гетероскедастичность*, необходимо устанавливать на начальном этапе анализа пространственных данных, так как это дает некоторую информацию о типе многомерного распределения, описывающего изучаемое явление. Многомерные нормально распределенные данные не показывают эффекта пропорциональности. Напротив, логнормально распределенные

данные дают линейную связь между локальным среднеквадратическим отклонением и локальным средним.

Хорошим средством для проверки данных на эффект пропорциональности является график, показывающий связь локального среднего и локальной дисперсии, оцененных путем вычисления статистик в скользящем локальном окне.

Статистики в скользящем окне используются для изучения аномалий, как в значении среднего, так и дисперсии изучаемого явления. Скользящее окно является не чем иным как локальной окрестностью, в которой вычисляются описательные статистики. Из соображений вычислительной эффективности обычно используются прямоугольные окна. Размер скользящего окна зависит от среднего расстояния между точками данных и от общего размера изучаемой области. Окно должно быть достаточно большим, чтобы в нем находилось достаточное число точек выборки (для надежного вычисления статистик), однако оно не должно быть слишком большим, ибо в этом случае не наберется достаточного числа окон для выявления расположения аномалий. Хорошим компромиссом обычно являются перекрывающиеся окна, когда две соседних окрестности имеют часть общих данных (рис.14).

Поведение локальных средних и дисперсий можно проследить и путем построения карт, показывающих их изменение по территории. При изучении таких карт в первую очередь следует обратить внимание на области с большими значениями локального среднего и локальной дисперсии, которые могут свидетельствовать о наличии эффекта пропорциональности.

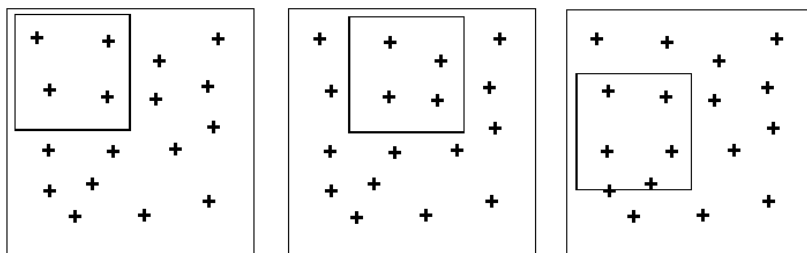


Рис.14. Перекрывающиеся окна для вычисления статистик скользящего окна

Вычисление статистик скользящего окна довольно легко программируется. Оно может быть выполнено в R с помощью пользовательской функции:

```
my.movewin(cur.data,cur.X,cur.Y,cur.var,w.nmin,  
            w.size,w.nshift)
```

Параметр	Назначение
<code>cur.data</code>	таблица с данными
<code>cur.X</code>	имя столбца с координатой X
<code>cur.Y</code>	имя столбца с координатой Y
<code>cur.var</code>	имя столбца с переменной, для которой вычисляются средние
<code>w.nmin</code>	минимальное число точек, при котором вычисляется статистика
<code>w.size</code>	размер окна
<code>w.nshift</code>	число сдвигов (сколько раз окно сдвигается, чтобы сдвинуться на размер окна)

Текст функции приведен в приложении 2. Следует обратить внимание на то, что размер окна должен соответствовать размерам области исследования и густоте точек (для надежного вычисления статистик в окно должно попадать не менее 10 точек).

Запускаем функцию:

```
my.movewin(p, "X", "Y", "Res", 10, 3000, 5)
```

Фрагмент вывода функции в файл результатов `out.dat` приведен ниже:

X	Y	Mean	Var	Num
38526	14631	10.60523	88.7301	13
39126	14631	10.60523	88.7301	13
39726	14631	10.60523	88.7301	13
40326	14631	12.6128	39.63879	12
37926	15231	6.18045	125.60844	13

...

Здесь по столбцам содержатся: координаты центра текущего окна, среднее, дисперсия и число точек выборки в окне.

Оцениваем эффект пропорциональности средствами R:

```
tmp<-read.table("out.dat",header=TRUE)
plot(tmp$Mean,tmp$Var,main="Эффект пропорциональности",
      xlab="Среднее",ylab="Дисперсия")
```



Рис.15. График, показывающий связь локального среднего и локальной дисперсии

Рис. 15 показывает, что никакой явной зависимости дисперсии от среднего не наблюдается.

Чтобы дополнительно убедиться, что остатки не содержат пространственного тренда, построим карты их локальных среднего и дисперсии:

```
coordinates (tmp) =~X+Y  
# отклонение локального среднего от общего  
tmp$Z<- (tmp$Mean-mean (p$Res) ) /sd (p$Res)  
bubble (tmp,zcol="Z",fill=TRUE,col=c ("black","grey") ,  
      scales=list (draw=TRUE) )
```

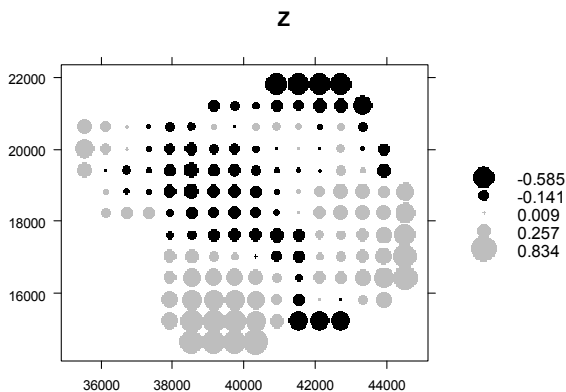


Рис.16. Карта отклонения локального среднего от общего

```
# отклонение локальной дисперсии от общей  
tmp$dF<- (tmp$Var/var (p$Res) -1)  
bubble (tmp,zcol="dF",fill= TRUE,col=c ("grey","black") ,  
      scales=list (draw=TRUE) )
```

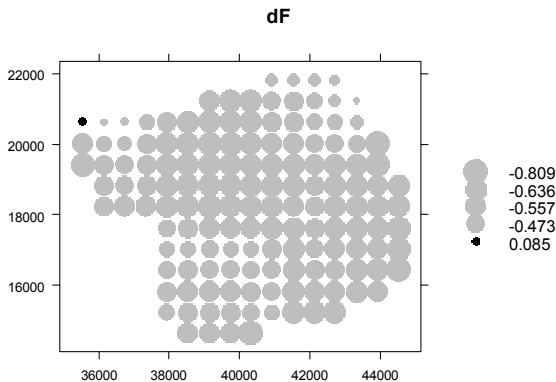


Рис.17. Карта отклонения локальной дисперсии от общей

Из визуального анализа можно заключить, что региональные тренды для среднего отсутствуют. Отношение локальной и глобальной дисперсий не превышает двух, это говорит о том, что их можно считать одинаковыми (критерий Фишера имеет критические значения больше двух, а все полученные значения меньше).

Вариограммный анализ данных

Основным инструментом анализа пространственной связи является вариограмма. По определению вариограмма (или полувариограмма) это один из моментов второго порядка стационарной случайной функции $Z(\mathbf{x})$:

$$\gamma(\mathbf{h}) = 1/2 \text{Var}\{Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})\} = 1/2 E\{[Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})]^2\},$$

описывающий корреляционную связь между двумя случайными величинами $Z(\mathbf{x})$ и $Z(\mathbf{x}+\mathbf{h})$, разделенными в пространстве вектором \mathbf{h} . Сила связи может меняться с изменением расстояния между точками и направления разделяющего вектора.

Если разделяющий вектор \mathbf{h} нулевой, это выражение равно нулю. Вариограмма симметрична относительно \mathbf{h} : $\gamma(\mathbf{h})=\gamma(-\mathbf{h})$. Связь может быть *изотропной*, если зависит только от модуля вектора \mathbf{h} (то есть $\gamma(\mathbf{h})=\gamma(|\mathbf{h}|)$), или *анизотропной*, если зависит как от модуля, так и от направления \mathbf{h} .

Вариограмма и ковариация при условии справедливости гипотезы стационарности второго порядка являются двумя равноценными мерами для характеристики пространственной связи:

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}).$$

Выборочная оценка вариограммы – *экспериментальная вариограмма*:

$$\gamma^*(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [z(\mathbf{x}_i + \mathbf{h}) - z(\mathbf{x}_i)]^2,$$

она вычисляется на основе выборочных значений $\{z_1=z(\mathbf{x}_1), z_2=z(\mathbf{x}_2), \dots, z_n=z(\mathbf{x}_n)\}$. Для вычисления $\gamma^*(\mathbf{h})$ при фиксированном \mathbf{h} используются пары выборочных точек $\{\mathbf{x}_i, \mathbf{x}_{ij}\}$, имеющие

разделяющий вектор, близкий к вектору \mathbf{h} . Эта близость может быть задана двумя способами – рис. 18 и рис.19:

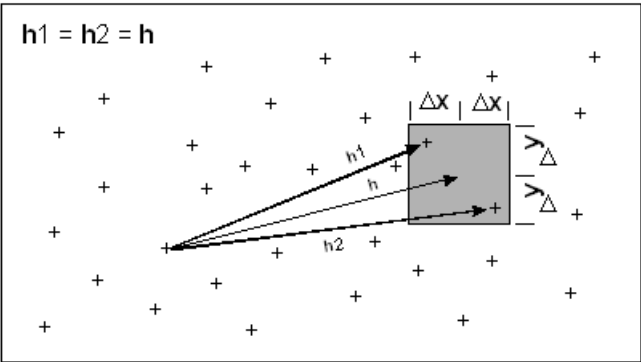


Рис.18. Допустимое отклонение задается для окончания разделяющего вектора \mathbf{h}

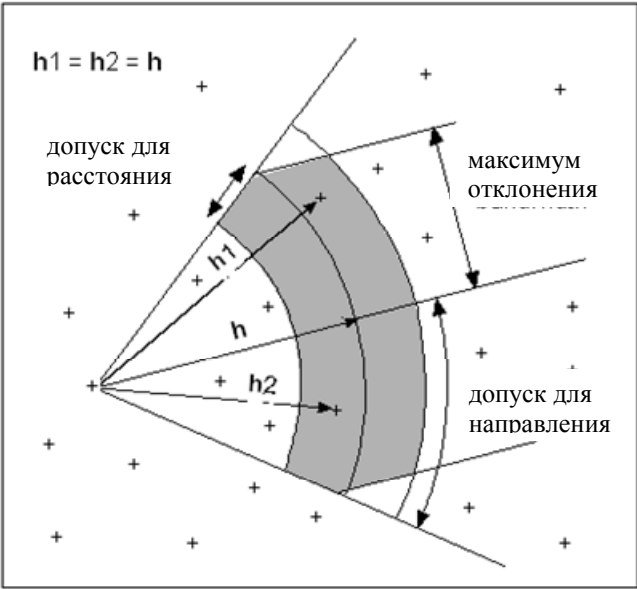


Рис.19. Допустимое отклонение задается для направления и для модуля разделяющего вектора \mathbf{h}

h-графики

Перед тем, как приступить к оценке вариограммы, полезно рассмотреть так называемые *h-графики*, обеспечивающие эффективный способ изучения пространственной связи и позволяющие наглядно выявить проблемы согласия данных со сделанными предположениями о стационарности и мультинормальности.

На *h-графике* значение переменной $z(\mathbf{x})$ в точке \mathbf{x} изображается напротив значения $z(\mathbf{x}+\mathbf{h})$ в точке $\mathbf{x}+\mathbf{h}$.

Для построения *h-графика* при фиксированном \mathbf{h} берутся пары точек $\{\mathbf{x}_i, \mathbf{x}_{ij}\}$, разделенные векторами, близкими к вектору \mathbf{h} , где близость определяется способом, поясненным на рис.19.

Для каких \mathbf{h} следует построить *h-графики*? По каким направлениям, с какими длинами? В общем случае надо «просканировать» плоскость с некоторым шагом по направлению, а на каждом направлении – с некоторым шагом по расстоянию. Обычно берут не менее 4-х направлений (0° , 45° , 90° , 135°) с допуском 30° , и по каждому направлению откладывают несколько лагов (расстояний), обычно кратных некоторому заданному шагу, с допуском для расстояния равным половине шага (то есть допустимые отклонения задают как на рис.19). Как правило угол отсчитывается по часовой стрелке от направления на Север (0°). Задать шаг лагов и число лагов поможет анализ расстояний между выборочными точками.

Рассмотрение *h-графиков* позволяет убедиться, что оценка вариограммы обеспечивается надежным обобщением данных.

1) Если некоторые точки на *h-графике* лежат обособленно («выброс»), нужно определить соответствующую им пару точек данных и, возможно, исключить их из рассмотрения.

2) h -график, имеющий форму «крыльев бабочки» (рис.20 а), противоречит гипотезе о мультинормальности данных - для мультинормальной случайной функции h -графики выглядят как облака эллиптической формы вокруг диагональной прямой (рис.20 б).

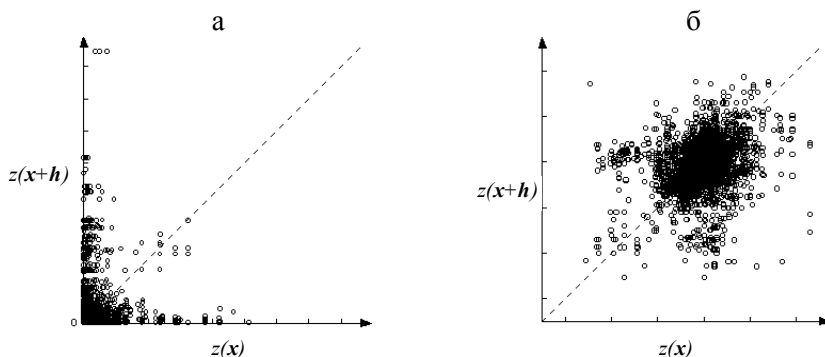


Рис.20. h -график, противоречащий гипотезе мультинормальности (а), и h -график, показывающий согласие с гипотезой мультинормальности (б)

3) Если h -график показывает различные группы точек, это может свидетельствовать о том, что выборка извлечена из смеси нескольких генеральных совокупностей. Облака точек, далеко отстоящих от диагональной прямой, могут соответствовать парам элементов выборки, происходящим из двух различных генеральных совокупностей (рис.21);

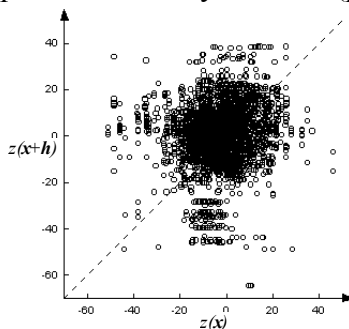


Рис.21. h -график, показывающий наличие неоднородности данных

4) Если последовательность h -графиков, построенных для векторов \mathbf{h} одного и того же направления, но разной длины (при разных значениях модуля $|\mathbf{h}|$ - лагах), показывает отклонение от диагональной прямой, то случайная функция, представляющая изучаемый феномен, не может считаться стационарной (на рис.22 приведены h -графики на 3-х последовательных лагах по заданному направлению).

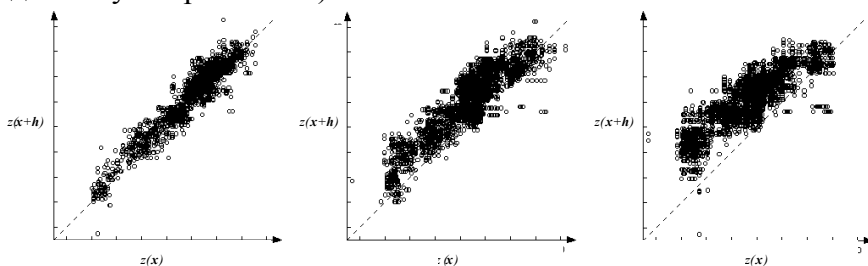


Рис.22. Последовательность h -графиков, противоречащая гипотезе стационарности

Построим и рассмотрим h -графики для данного примера. Сначала проанализируем расстояния между выборочными точками с целью последующего выбора шага лагов и числа лагов. Для этого вычислим расстояния между точками выборки, строим гистограмму расстояний, количественно оцениваем разброс расстояний с помощью квантилей их распределения:

```
p.dist<-as.matrix(dist(p[,c("X", "Y")]))
hist(p.dist)
# --- квантили (для оценки разброса расстояний)
(dist.breaks<-quantile(p.dist,seq(0.1,0.9,by=0.1)))
```

10%	20%	30%	40%	50%	60%	70%
2503.565	4188.440	9406.585	16354.753	20370.213	24770.460	28267.678
33671.850	42253.323					

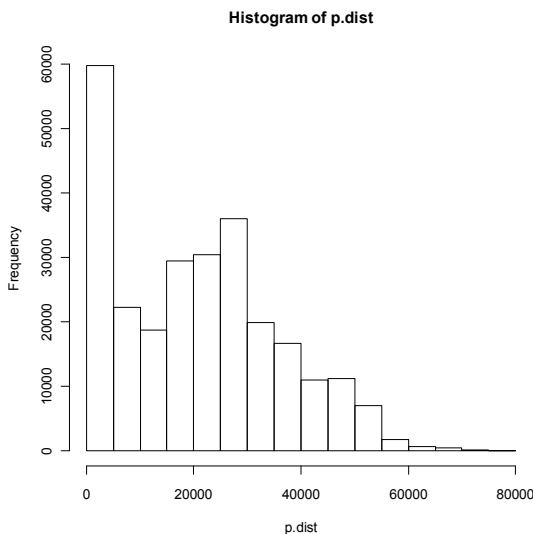


Рис.23. Гистограмма частот расстояний между выборочными точками

Прежде чем начать вариограммный анализ средствами R, следует преобразовать таблицу с данными к новому типу данных – пространственной структуре, с которой работают процедуры и функции пространственного анализа библиотеки **library(sp)** и надстроенных над ней библиотек. При этом табличную структуру предварительно пересохраним с новым именем.

```
p.src<-p
# -- создадим пространственную структуру с координатами
coordinates(p)~X+Y
```

Зададим лаги с шагом равным 1000, и число лагов – 12:

```
p.breaks<-500+(0:12)*1000
```

Построим *h*-графики для векторов с направлениями 0°, 45°, 90°, 135° и длиной, определяемой лагами:

```
# --- h-plot (стандартный)
library(gstat)
hscat(Res~1,p,breaks=p.breaks,col="black",
```

```

variogram.alpha = 0,tol.hor=30)
hscat(Res ~1,p,breaks=p.breaks,col="black",
      variogram.alpha = 45,tol.hor=30)
hscat(Res ~1,p,breaks=p.breaks,col="black",
      variogram.alpha = 90,tol.hor=30)
hscat(Res ~1,p,breaks=p.breaks,col="black",
      variogram.alpha = 135,tol.hor=30)

```

Ниже на рис.24 показаны h -графики с направлением 0° .

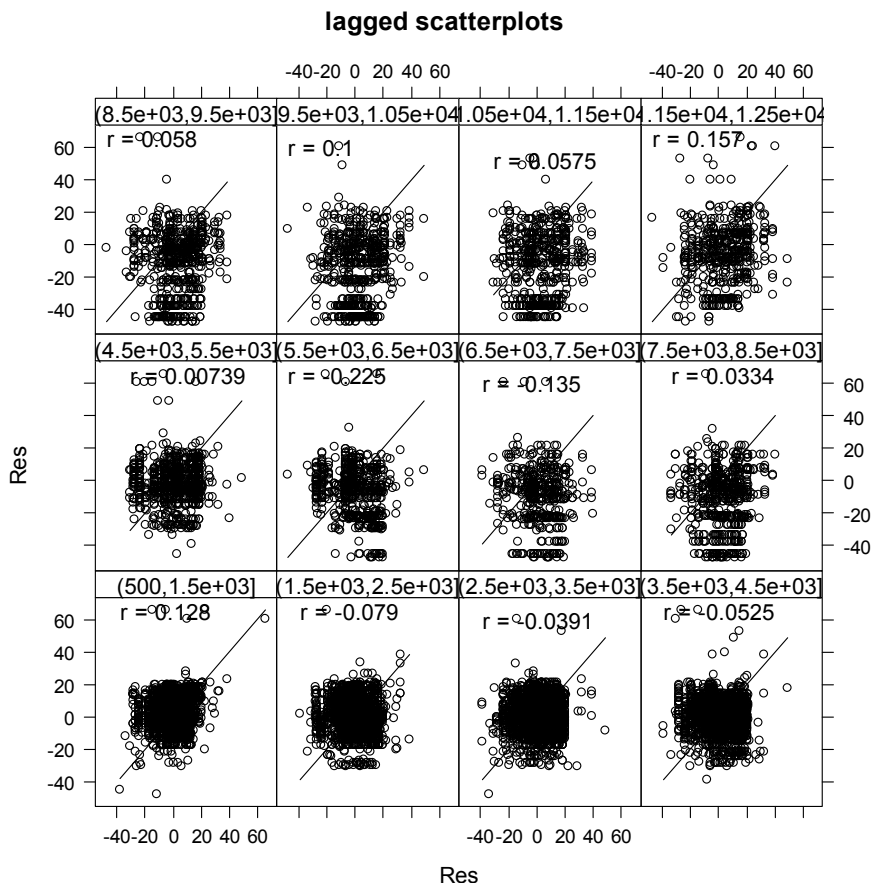


Рис.24. h -графики на 12 лагах по направлению 0°

Для более детального анализа можно построить h -графики для произвольных интервалов длин векторов, используя пользовательскую функцию:

```
my.hscat (cur.var,cur.data,d.lo,d.hi,
          variogram.alpha,tol.hor)
```

Параметр	Назначение
cur.var	Переменная, для значений которой строится график
cur.data	Пространственная структура с данными
d.lo	Нижняя граница длины разделяющего вектора
d.hi	Верхняя граница длины разделяющего вектора
variogram.alpha	Направление разделяющего вектора (угол)
tol.hor	Допуск для направления

Текст этой функции приведен в приложении 3. Функция дополнительно наносит на h -график положение средних значений в «голове» и в «хвосте» вектора h (большая точка; она должна лежать близко к диагонали, если данные не противоречат стационарности) и границу 95% интервала для бинормального распределения (эллипс; 95% точек должны лежать в пределах этого эллипса, если данные не противоречат бинормальности).

Пример для расстояний от 100 до 800 по направлению 0° приведен ниже.

```
# --- h-plot с доверительным интервалом)
my.hscat("Res",p,d.lo=100,d.hi=800,
          variogram.alpha=0, tol.hor=30)
```

Анализ построенных h -графиков для рассматриваемого примера не показал существенного противоречия данных (остатков) предположениям о стационарности и мультинормальности.

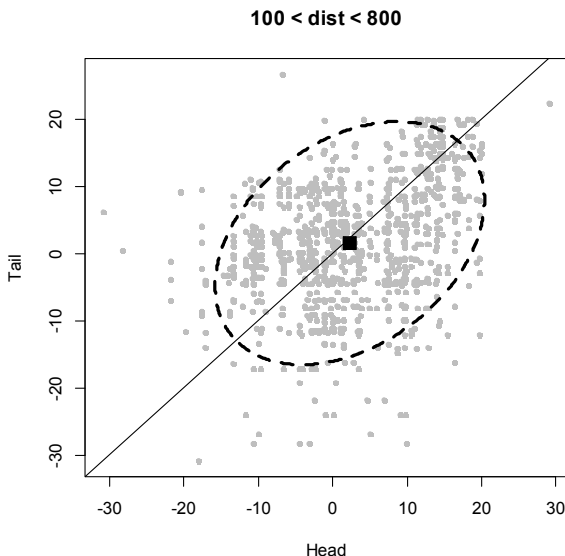


Рис.25. h -график для расстояний от 100 до 800 по направлению 0° с 95%-м доверительным интервалом

Поверхность вариограммы

Исследование вариограммы, как правило, начинают с рассмотрения поверхности экспериментальной вариограммы - изолинейной или пиксельной карты, отображающей значения $\gamma^*(\mathbf{h})$ в зависимости от \mathbf{h} .

Для построения поверхности экспериментальной вариограммы задается растровая сетка с помощью шага лагов по X , шага лагов по Y , числа лагов по X , числа лагов по Y - рис.26, и в каждой ячейке этой сетки вычисляется значение $\gamma^*(\mathbf{h})$ для соответствующего разделяющего вектора \mathbf{h} . Иначе говоря, допустимые отклонения для разделяющего вектора \mathbf{h} задаются способом рис.18.

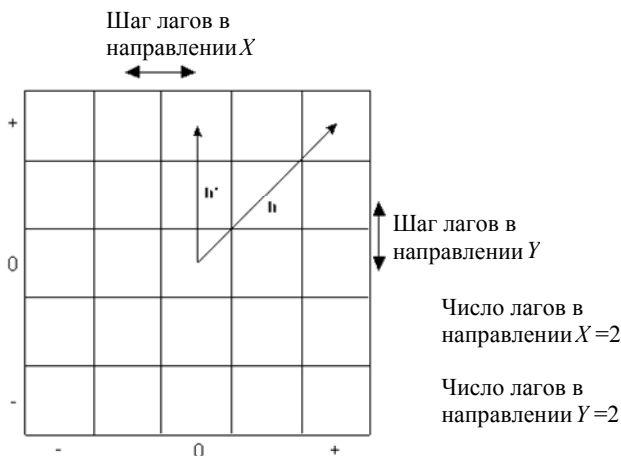


Рис.26. Параметры построения поверхности вариограммы

При этом важно оптимально подобрать параметры построения сетки: шаг лагов и число лагов. Размер шага лагов должен быть не настолько большим, чтобы пропустить характерные структурные особенности вариограммы, и не настолько малым, чтобы какое-либо значение $\gamma^*(h)$ вычислялось по недостаточному числу $N(h)$ пар точек (обычно достаточным считается не менее 50 пар точек). Число лагов должно выбираться таким, чтобы максимальное расстояние, для которого вычисляется $\gamma^*(h)$, было разумным. Обычно оптимальные значения этих параметров выбираются итеративно в процессе исследования вариограммы.

Поверхность вариограммы предоставляет эффективный способ визуализации анизотропии изучаемого явления. Наличие анизотропии пространственной связи и ее направление (направления) хорошо видны на поверхности вариограммы. Эти

направления необходимо определить, чтобы впоследствии вдоль них проанализировать вариограмму более подробно.

Средствами R построим поверхность вариограммы для рассматриваемого примера. Для вычисления экспериментальной вариограммы предназначена команда `variogram(...)`. Чтобы с ее помощью построить поверхность вариограммы, необходимо задать логический параметр `map=TRUE`, и задать параметры сетки - шаг сетки (шаг лагов) `width` и максимальное расстояние `cutoff`.

```
p.gam_map<-variogram(Res ~1,p,cutoff=15000,width=1000,  
                      map=TRUE)
```

```
plot(p.gam_map,col.regions=gray((16:0)/16))
```

Здесь `col.regions` - цвета (чем значение больше, тем темнее).

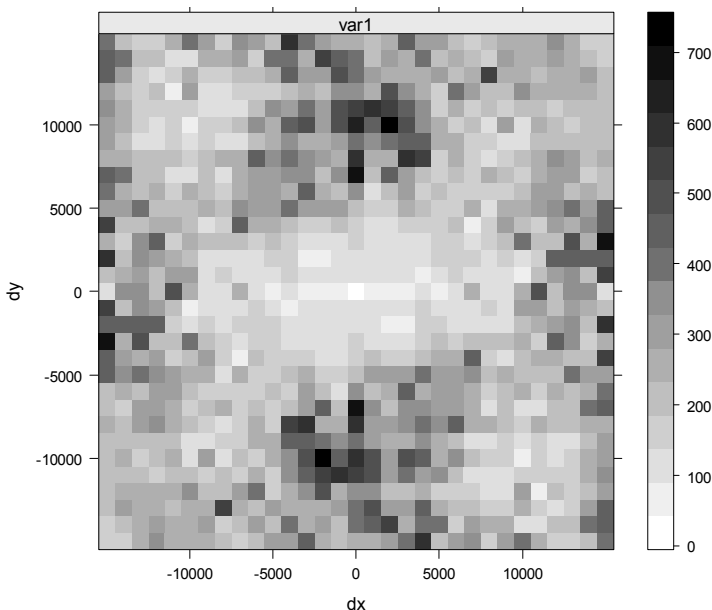


Рис.27. Поверхность вариограммы

В силу того, что в каждой ячейке для надежной оценки необходимо иметь достаточно много значений, поверхность

вариограммы получается грубой, и иногда имеет смысл ее сгладить, чтобы лучше оценить наличие и направление имеющихся структур. Для этого используем пользовательскую функцию, текст которой приведен в приложении 4:

```
my.vario_map(cur.gam_map,no.X,no.Y,h,  
             lev.lo,lev.hi, lev.step)
```

Параметр	Назначение
cur.gam_map	Экспериментальная вариограмма
no.X	Размер ячеек по X
no.Y	Размер ячеек по Y
h	Параметр сглаживания
lev.lo	Минимальный уровень изолиний
lev.hi	Максимальный уровень изолиний
lev.step	Шаг изолиний

```
my.vario_map(p.gam_map, 100, 100,5, 100, 800, 50)
```

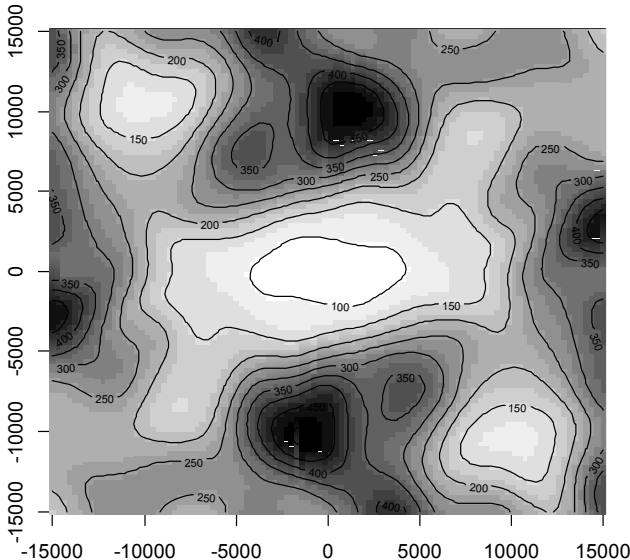


Рис.28. Поверхность вариограммы после сглаживания

Для рассматриваемого примера на поверхности экспериментальной вариограммы хорошо просматривается анизотропия по направлению около 75° (от Севера).

Вариограмма по направлению

Дальнейшее исследование пространственной структуры заключается в подробном анализе поведения экспериментальной вариограммы в некоторых заданных направлениях. Обычно рассматриваются направления максимума и минимума пространственной связи (направление анизотропии и перпендикулярное к нему), определенные по поверхности вариограммы, и два промежуточных направления. Для этих направлений строятся графики изменения значений экспериментальной вариограммы $\gamma^*(h)$ в зависимости от значений модуля $|h|$.

При вычислении $\gamma^*(h)$ по фиксированному направлению задаются такие же параметры (рис.29), как и при построении h -графиков: угол направления, допустимое отклонение от направления, шаг лагов по длине $|h|$, допустимое отклонение по длине $|h|$ (обычно равное половине шага) (то есть близость векторов определяется способом рис.19).

На лаге равном нулю рассматриваются пары точек, имеющих разделяющее расстояние меньшее, чем допустимое отклонение.

Использование параметра для максимальной ширины полосы обеспечивает вычисление вариограммы более точно по направлению, что видно из рис.29.



Рис.29. Параметры, задаваемые при вычислении вариограммы по направлению

Вычисление вариограммы с допустимым отклонением направления вектора \mathbf{h} от заданного равным 90° эквивалентно усреднению вариограмм, вычисленных по всем направлениям. Этот тип вариограммы известен как всенаправленная вариограмма.

В общем случае график вариограммы по направлению показывает следующее поведение:

- начинается в нуле – для $|\mathbf{h}|=0$ $z(\mathbf{x}+\mathbf{h})-z(\mathbf{x})=0$;
- возрастает с ростом $|\mathbf{h}|$;
- стабилизируется на определенном уровне или же неограниченно растет.

Ранг, порог, вложенные структуры

Если вариограмма достигает предельного значения, называемого *порогом*, на некотором расстоянии, называемом *рангом*, то за этим расстоянием $Z(\mathbf{x})$ и $Z(\mathbf{x}+\mathbf{h})$ становятся некоррелированными (рис.30 а). Ранг вариограммы определяет

такой параметр кригинга, как радиус поиска (Search), задающий «площадь влияния выборки» (Search Ellipse).

Если изменчивость феномена не имеет предела, на больших расстояниях вариограмма растет неограниченно (рис.30 б).

Часто вариограмма показывает наличие, так называемых, вложенных структур (иерархических структур), отражающих взаимодействие разномасштабных пространственных зависимостей, каждая из которых характеризуется своим собственным рангом (рис.30 в).

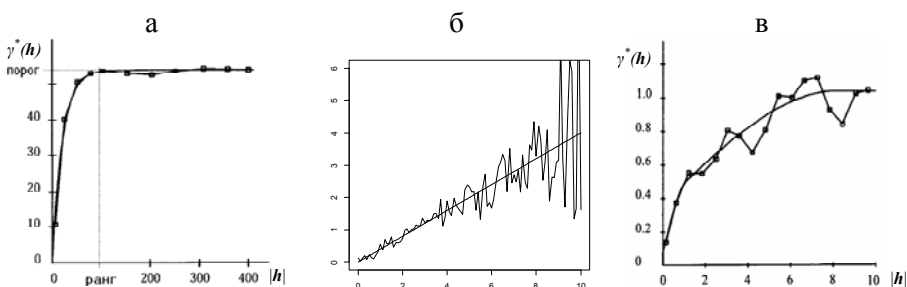


Рис.30. Графики вариограммы по заданному направлению, показывающие: наличие ранга и порога (а); отсутствие ранга и порога (б); наличие вложенных структур (в)

Поведение около нуля и эффект самородков

Поведение вариограммы при малых $|h|$, т.е. для рядом расположенных точек данных, связано с непрерывностью и пространственной гладкостью изучаемой переменной. Можно считать типичными следующие четыре случая (рис.31):

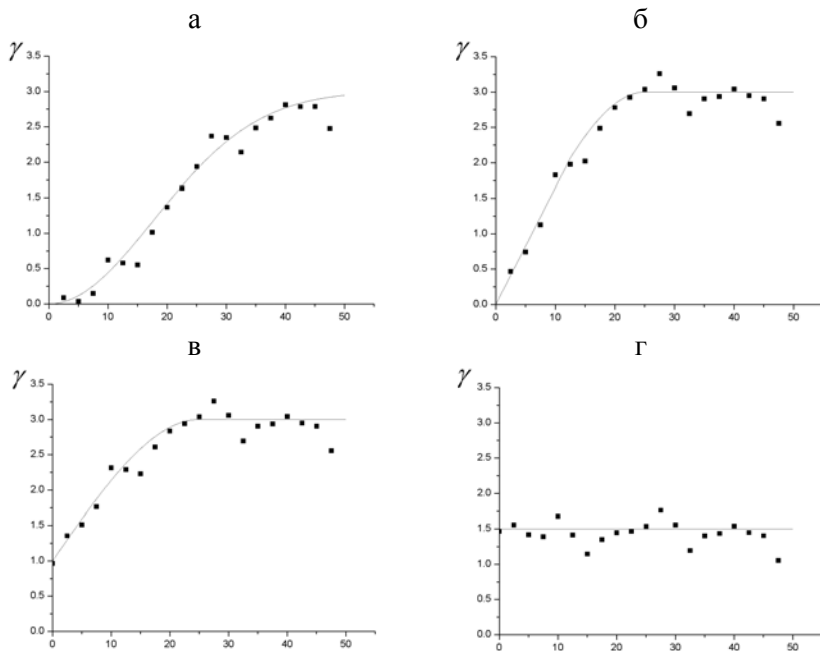


Рис.31. Графики вариограммы по заданному направлению, показывающие: параболическое поведение около нуля (а); линейное поведение около нуля (б); разрыв в нуле - эффект самородков (в); белый шум (г)

(а) Параболическое поведение. Характеризует высокую степень гладкости пространственной переменной. Однако, если оно сохраняется на больших расстояниях, то может означать наличие сильного тренда (дрифта).

(б) Линейное поведение. Говорит о непрерывности пространственной переменной, хотя бы частичной, но с меньшей гладкостью, чем в предыдущем случае.

(в) Разрыв в 0 ($\gamma^*(|h|)$ не стремится к 0 при $|h| \rightarrow 0$) – эффект самородков. Означает разрывность пространственной переменной - наличие микроструктуры, то есть компоненты

феномена с рангом короче чем экспериментальное разрешение (sampling support). Однако, такое поведение вариограммы может быть обусловлено и просто особенностями выборки: слишком большим расстоянием между выборочными точками, что не позволяет обнаружить структуры с рангом короче, чем наименьшее расстояние между точками, а также ошибками измерения или позиционирования. При отсутствии в выборке близко расположенных точек, невозможно выяснить истинные источники эффекта самородков, более того, они могут быть смешаны. Для его объяснения становятся важны априорные знания об изучаемом явлении, его физической сущности.

(г) «Плоская» линия – чистый эффект самородков или белый шум. Означает отсутствие корреляции между двумя точками, как бы близки они не были. Это экстремальный случай общего отсутствия пространственной структуры.

Анизотропия

Вариограмма по направлению может проявлять два типичных вида анизотропии:

1) Порог постоянен, а ранг изменяется с изменением направления – геометрическая анизотропия (рис.32 а, б).

2) Вариограмма показывает низкий порог в некотором специфическом направлении – зональная анизотропия (рис.32 в). Такое поведение часто бывает обусловлено наличием линии нарушения непрерывности, например, разлома (изменчивость сильнее в направлении перпендикулярном разлому), что может свидетельствовать о нестационарности.

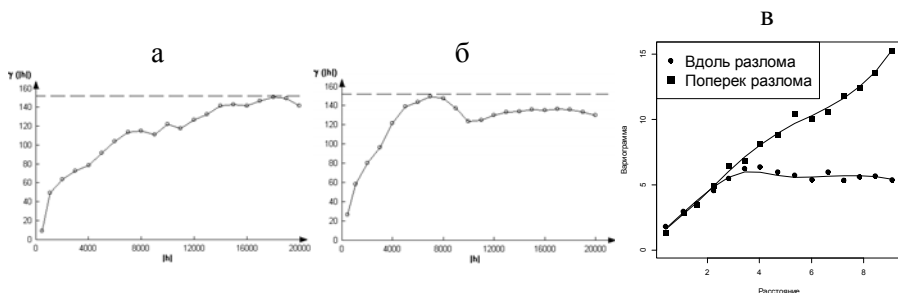


Рис.32. Графики вариограммы по двум различным направлениям, показывающие: наличие геометрической анизотропии (а,б); наличие зональной анизотропии (в)

Проявление присутствия дрифта

Из математических свойств вариограммы следует, что если экспериментальная вариограмма возрастает так же быстро как $|h|^2$ (по параболе) или быстрее, проявляя сильную изменчивость с направлением, то такое поведение указывает на наличие дрифта (пространственного тренда) в данных (рис.33).

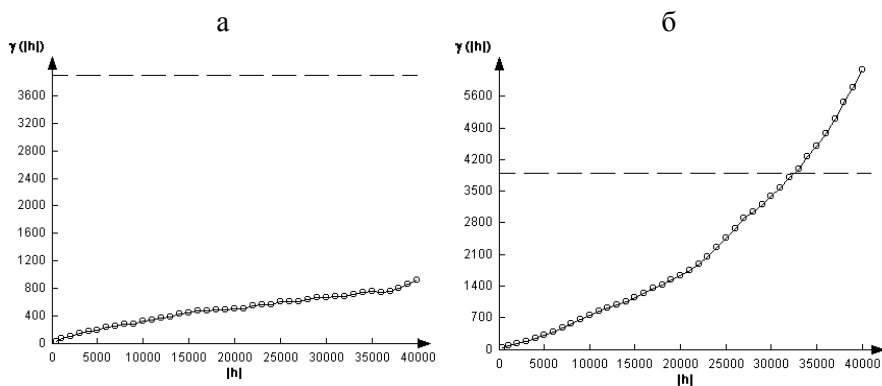


Рис.33. Графики вариограммы по двум заданным направлениям, показывающие наличие пространственного тренда в данных

Пример оценки вариограммы в R

Как отмечалось выше, для вычисления экспериментальной вариограммы в R используется команда `variogram(...)`. Чтобы получить значения $\gamma^*(h)$ на заданных направлениях, необходимо указать следующие параметры:

alpha	Направление, вдоль которого вычисляется $\gamma^*(h)$; угол в градусах; отсчитывается по часовой стрелке от направления на Север (0°); по умолчанию $=0^\circ$
tol.hor	Допустимое отклонение для направления (угол в градусах); по умолчанию $= 90/\text{length}(\text{alpha})$
cutoff	Максимальное расстояние, для которого вычисляется $\gamma^*(h)$
width	Шаг лагов; по умолчанию $= \text{cutoff}/15$
cressie	Если TRUE, то используется робастная оценка вариограммы; если FALSE, то – классическая оценка; по умолчанию $= \text{FALSE}$
...	...

Средствами R оценим вариограмму для данных (остатков) по 4-м направлениям: по направлению анизотропии, по перпендикулярному к нему направлению и по двум промежуточным направлениям - (75° , 120° , 165° , 210°); с допуском равным 30° , шагом лагов равным 500, до максимального расстояния 15000. Если данных много, то для получения более детального результата можно использовать большее число направлений.

```
p.gam<-variogram(Res~1,p,alpha=c(75,120,165,210),
                 width=500,cutoff=15000,tol.hor=30)
```

Нарисуем графики вариограммы:

```
plot(p.gam)
```

Или с помощью собственной функции:

```

my.panel = function(x,y,var=0,...) { # своя рисовалка
  panel.lines(x,y,lty=1,lwd=1,col="black") # линия
  panel.lines(range(x),c(var,var),lty=2,lwd=1,
    col="black") # дисперсия
  vgm.panel.xyplot(x,y,...) # точки }
plot(p.gam,pch=1,col="black",сех=1,var=Res.var,
  panel=my.panel)

```

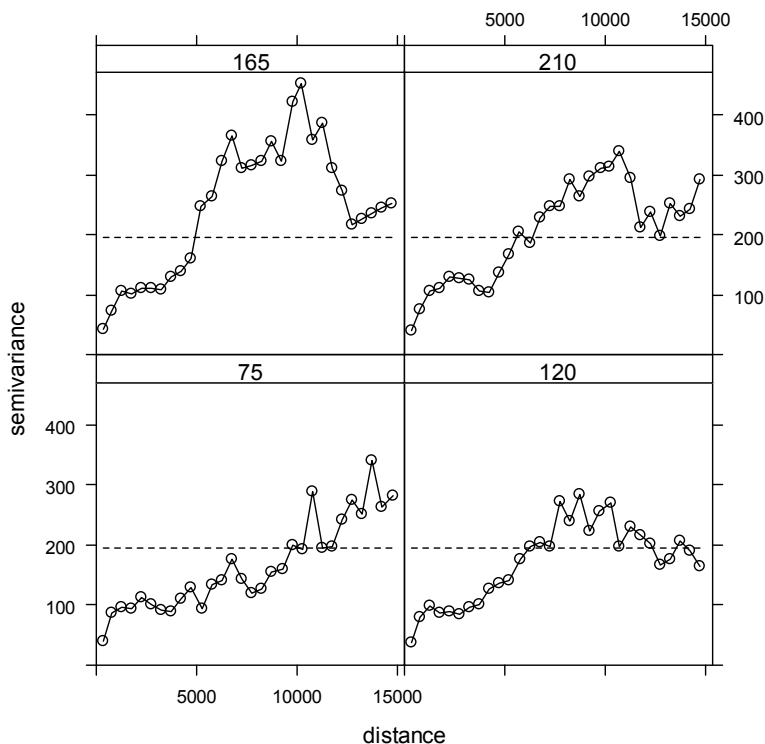


Рис.34. Вариограмма остатков по 4-м направлениям

На рис.34 видно, что поведение вариограммы действительно различно при разных направлениях. Дисперсия переменной равна 195.42, и вариограмма не сильно выходит за ее пределы на небольших расстояниях, которые наиболее важны. Вариограмма показывает наличие двух вложенных структур: первая с порогом около 100 и рангом около 1000 м; порог второй

структуры около 300, а ее ранг зависит от направления, что свидетельствует о геометрической анизотропия пространственной связи. Поведение около нуля линейное, эффект самородков отсутствует.

В качестве примера, иллюстрирующего проявление в вариограмме присутствия в данных пространственного тренда, вычислим вариограмму для недетрендриванной переменной V03q по направлениям 45° и 135°:

```
p.gam_tr<-variogram(V03q~1, p,alpha=c(45,135),
                     width=1000,cutoff=40000,tol.hor=30)
plot(p.gam_tr,pch=1,col="black",cex=1,panel=my.panel)
```

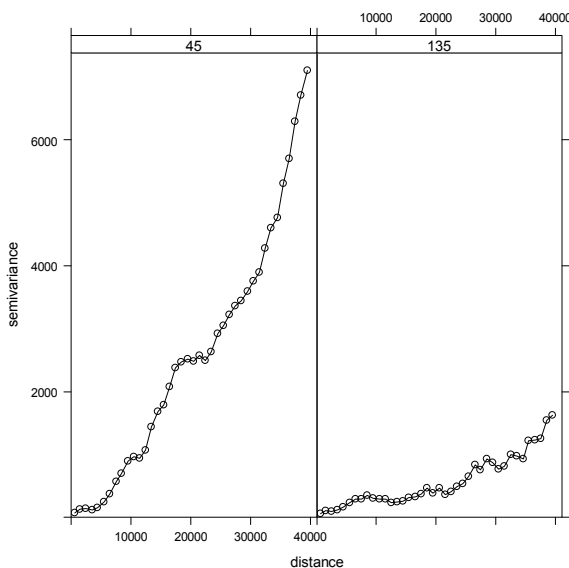


Рис.35. Вариограмма исходной переменной (без вычитания тренда) по 2-м направлениям

Моделирование вариограммы

На завершающем шаге вариограммного анализа на основе экспериментальной вариограммы $\gamma^*(h)$ должна быть построена

модель вариограммы $\hat{\gamma}(\mathbf{h})$ для дальнейшей подстановки ее в систему кригинга и получения искомых оценок изучаемой пространственной переменной.

Базисные модели

Моделью вариограммы может служить не каждая функция, а только та, для которой выполнено условие положительной определенности. Положительная определенность модели вариограммы гарантирует, что уравнения кригинга, построенные с использованием данной модели, имеют единственное устойчивое решение. Поэтому при моделировании используются только те функции, для которых положительная определенность установлена, а также их взвешенные линейные комбинации с неотрицательными весами, которые тоже будут являться положительно определенными.

Эти функции, называемые *базисными моделями*, просты по форме и имеют один или два параметра. Они бывают двух типов:

- 1) модели, достигающие предельного значения (переходящие на плато) с увеличением длины вектора \mathbf{h} ,
- 2) модели, растущие неограниченно.

Модели первого типа используются, если случайная функция обладает стационарностью второго порядка. Плато, которое они достигают, - это порог, а расстояние, на котором порог достигается, - ранг. Некоторые из моделей достигают порога асимптотически; для них ранг - это расстояние, при котором модель достигает 95% от порога.

Модели второго типа не достигают порога, а непрерывно растут с увеличением расстояния. Такие модели используются

для случайных функций, для которых выполнена гипотеза стационарности приращений.

Модель вариограммы $\hat{\gamma}(\mathbf{h})$ строится как линейная комбинация подходящих базисных моделей. На практике чаще всего это одно или два слагаемых, реже больше.

Рассмотрим основные базисные модели, используемые для моделирования вариограммы.

Модель с эффектом самородков

Модель с эффектом самородков (nugget effect) используется для моделирования разрывов вариограммы, имеющих место около нуля (рис.36). Хотя для $|\mathbf{h}|=0$ значение вариограммы есть в точности 0, для весьма малых разделяющих расстояний $|\mathbf{h}|$ оно может существенно отличаться от 0, указывая на наличие разрывности. Такая разрывность может моделироваться следующим уравнением:

$$\gamma(|\mathbf{h}|) = \begin{cases} 0, & \text{если } |\mathbf{h}| = 0 \\ c, & \text{в остальных случаях} \end{cases}$$

где c - положительная величина, характеризующая разрывность .

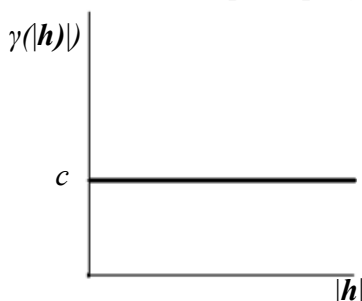


Рис.36. Модель с эффектом самородков

Сферическая модель

Это наиболее известная модель первого типа. Ее уравнение записывается как:

$$\gamma(|\mathbf{h}|) = c \text{ Sph}_a(|\mathbf{h}|) = \begin{cases} c \cdot [1.5|\mathbf{h}|/a - 0.5(|\mathbf{h}|/a)^3], & \text{если } |\mathbf{h}| \leq a \\ c, & \text{в остальных случаях} \end{cases}$$

где a - ранг, c - вклад дисперсии или порог.

Эта модель имеет линейное поведение около начала координат и достигает порога на расстоянии a . Проводя приближение с помощью этой модели, полезно помнить, что касательная из начала координат достигает порога на расстоянии около 2/3 ранга (рис.37).

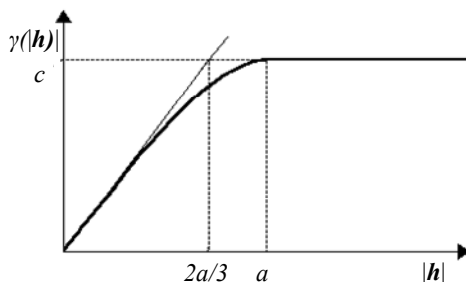


Рис.37. Сферическая модель

Экспоненциальная модель

Это другая общеизвестная модель первого типа. Ее уравнение записывается как:

$$\gamma(|\mathbf{h}|) = c \cdot \text{Exp}_a(|\mathbf{h}|) = c \cdot [1 - e^{-3|\mathbf{h}|/a}],$$

где c - вклад дисперсии или порог, a - фактический ранг, то есть расстояние, на котором значение вариограммы есть 95% от порога. Модель достигает своего порога асимптотически и имеет линейное поведение в начале координат. Используя ее, полезно

помнить, что касательная из начала координат достигает порога на расстоянии около 1/5 ранга (рис.38).

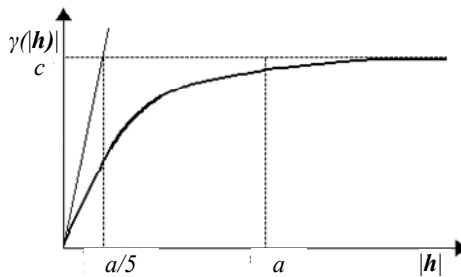


Рис.38. Экспоненциальная модель

Гауссовая модель

Это модель первого типа, часто используемая при моделировании непрерывных феноменов. Ее уравнение записывается как:

$$\gamma(|\mathbf{h}|) = c \cdot \text{Gauss}_a(|\mathbf{h}|) = c \cdot [1 - \text{Exp}(-3|\mathbf{h}|^2/a^2)],$$

где c - вклад дисперсии или порог, a - фактический ранг, то есть расстояние, на котором значение вариограммы есть 95% от порога. Модель достигает порога асимптотически и имеет параболическое поведение около начала координат. Ее график имеет точку перегиба (рис.39).

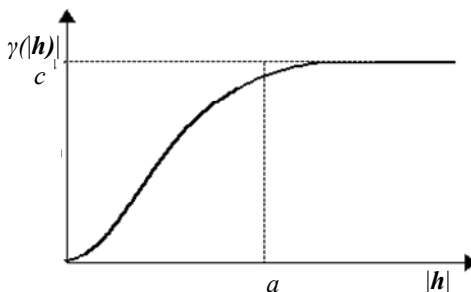


Рис.39. Гауссовая модель

Степенная модель

Степенная модель является моделью второго типа, поскольку она не достигает порога, а растет с увеличением длины разделяющего вектора. Уравнение для нее записывается как:

$$\chi(|\mathbf{h}|) = c \cdot \text{Pow}_a(|\mathbf{h}|) = c \cdot |\mathbf{h}|^a,$$

где c есть вклад дисперсии, степень a находится в интервале от 0 до 2. Данная модель интересна тем, что целый спектр поведений графика модели около нуля определяется значением показателя степени a (рис.40). Ее частный вариант при $a=1$ – линейная модель. Степенная модель является положительно определенной для случайных функций, для которых предполагается только стационарность приращений. Ее нежелательно использовать для случайных функций, имеющих стационарность второго рода.

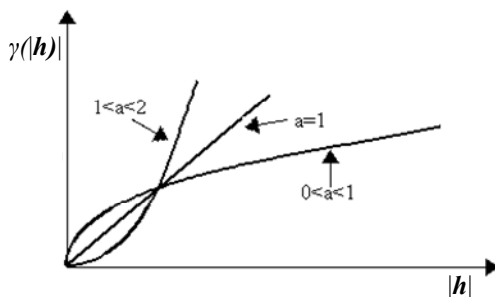


Рис.40. Степенная модель

Модель с эффектом дырок

Эту модель также называют синусоидальной или моделью со скважинным эффектом в трехмерном пространстве; она соответствует непрерывным структурам с периодическим изменением свойств в пространстве. Ее уравнение имеет вид:

$$\gamma(|\mathbf{h}|) = c \cdot \left(1 - \frac{\sin(|\mathbf{h}|/a)}{|\mathbf{h}|/a} \right),$$

где $|\mathbf{h}|/a$ измеряется в радианах. Эта модель имеет порог и характеризуется параболическим поведением в начале координат (рис.41).

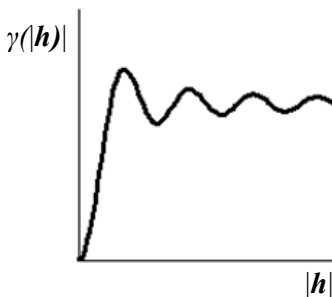


Рис.41. Модель с эффектом дырок

Итеративное построение модели вариограммы

В общем случае модель $\hat{\gamma}(\mathbf{h})$ должна с некоторой точностью воспроизводить поведение экспериментальной вариограммы $\gamma^*(\mathbf{h})$. Построение модели вариограммы представляет собой итеративный процесс, на каждом шаге которого следует наилучшим образом подобрать параметры очередного модельного приближения. Для этого совместно рассматриваются графики экспериментальной и теоретической вариограмм по различным направлениям (обычно, по 4-м направлениям: анизотропии, перпендикулярному к нему направлению и по двум промежуточным направлениям), а также поверхности этих вариограмм.

Приближение поведения около нуля

Рассмотрим график вариограммы по некоторому направлению на рис.42. Для нее возможны следующие три способа приближения поведения около $|h|=0$:

- (a) экстраполируем линейное поведение первых двух точек экспериментальной вариограммы и получаем приближение с эффектом самородков равным 1.2;
- (b) предполагаем наличие структуры с рангом меньше, чем минимальное расстояние между выборочными точками, и строим приближение без эффекта самородков;
- (c) строим приближение, дающее гладкое поведение около нуля в комбинации с сильным эффектом самородков.

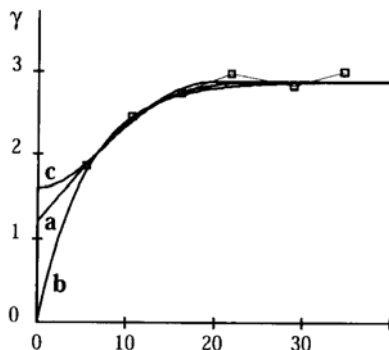


Рис.42. Вариограмма по направлению: три варианта приближения поведения около нуля

Выбор способа моделирования поведения около нуля всегда достаточно уязвим: он может быть как подтвержден, так и опровергнут добавлением новых данных, уплотняющих выборочную сеть. Пока такие данные недоступны, истинное поведение около нуля может только предполагаться с учетом априорных знаний об изучаемом феномене. Например:

1) Если изучается поле потенциалов (гравитационное или магнитное), про которое известно, что оно имеет очень гладкую пространственную структуру, выбирается приближение (с), где эффект самородков отражает ошибки измерения.

2) Если изучается поверхность довольно гладкого геологического пласта, а данные - это замеры на скважинах, не содержащие ошибок, то видимый эффект самородков можно считать структурой с коротким рангом и выбрать приближение (b), моделируемое, например, с использованием сферической базисной модели.

3) Приближение (a) используется для переменных, таких как концентрации или пористость, которые показывают сильную изменчивость и измеряются в небольших ограниченных областях, обнаруживая наличие микроструктур - настоящий эффект самородков.

Необходимо понимать, что выбор способа приближения поведения около нуля связан с риском грубой ошибки: если он не обоснован, кригинг теряет свою оптимальность и величина ошибки для оценки кригинга может значительно увеличиться. Ошибки измерения, микроструктуры и компоненты с коротким рангом часто смешаны. Иногда бывает известна дисперсия ошибок измерения, что облегчает выбор приближения; например, стандартное отклонение (или более часто удвоенное стандартное отклонение) ошибки измерения дает установленная точность оборудования. Однако для того чтобы точно определить компоненты с коротким рангом, существует только один путь - собрать дополнительные экспериментальные данные на более плотной сетке.

Моделирование непрерывной компоненты

Как только тип поведения около нуля (на расстояниях меньших первого лага) выбран, остается смоделировать вариограмму на средних и больших расстояниях.

Для этого можно использовать либо одну из базисных моделей (рис.43), либо линейную комбинацию нескольких базисных моделей, соответствующих разным рангам (вложенные структуры) (рис.44). Для них нужно наилучшим образом подобрать значения параметров, воспроизводящих наклон, ранг (ранги) и порог экспериментальной вариограммы по различным направлениям, а также вид ее поверхности.

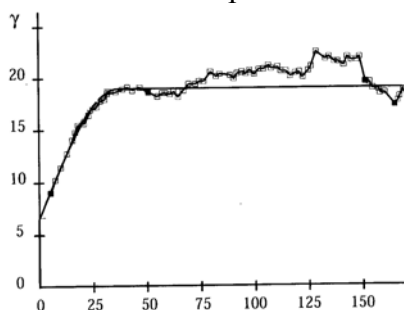


Рис.43. Вариограмма по направлению: моделирование одной структуры

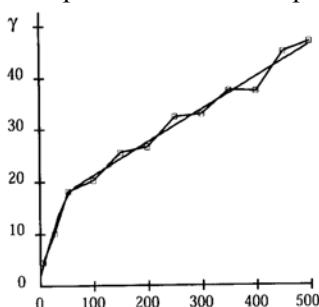


Рис.44. Вариограмма по направлению: моделирование двух вложенных структур

Наклон на расстояниях близких к нулю оценивается по первым точкам графика вариограммы по направлениям. Порог устанавливается на уровне стабилизации вариограммы. Ранг иногда найти не так просто, так как точка, на которой он встречается с порогом, не всегда очевидна. Для сферической модели, часто используемой на практике, ранг можно определить из того свойства графика, что касательная к модели в начале координат пересекает порог на расстоянии $2/3$ ранга.

В отсутствие физических подсказок или внутренних устойчивых ограничений, существует определенная широта в выборе базисных моделей. Как только поведение около нуля установлено, различные возможные приближения вариограммы на других расстояниях будут графически весьма схожи. Другими словами, при заданном поведении около нуля приближение устойчиво относительно аналитической формы выбранной модели.

Моделирование анизотропии

В случае, если поверхность экспериментальной вариограммы показывает наличие анизотропии пространственной связи, эту зависимость от направления необходимо отразить в модели вариограммы.

Все базисные модели позволяют моделировать анизотропное поведение путем учета двух дополнительных параметров: направления анизотропии (*Dir*) и отношения анизотропии (*Anis*), которое есть отношение малой и большой полуосей эллипса, представляющего анизотропию (рис.45).

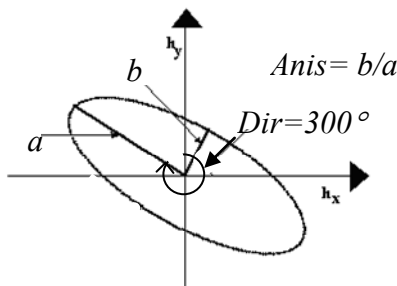


Рис.45. Параметры описания анизотропии

Анизотропная модель создается из изотропной базисной модели с помощью подстановки в нее скорректированного по длине вектора-аргумента $\mathbf{h}_{\text{corrected}}$:

$$|\mathbf{h}_{\text{corrected}}| = \sqrt{x^2 + y^2},$$

где

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{Anis} \end{bmatrix} \begin{bmatrix} \cos(Dir) & \sin(Dir) \\ -\sin(Dir) & \cos(Dir) \end{bmatrix} \begin{bmatrix} h_x \\ h_y \end{bmatrix}$$

Изотропия есть частный случай, когда параметр *Anis* равен 1.

Таким образом, в процессе подгонки модели вариограммы помимо параметров используемой базисной модели (или моделей в случае вложенных структур) нужно определить значения дополнительных параметров, описывающих анизотропию для каждой базисной модели. На параметры самой базисной модели анизотропия не влияет.

Пример построения модели вариограммы в R

В пакете *gstat* среды R доступны все указанные выше базисные модели, а также целый ряд других. Команда **show.vgms()** позволяет показать их поведение при заданных значениях параметров. Например:

```
show.vgms(models = c("Sph", "Exp", "Pow", "Hol"),
  sill=0.2, nugget = 0.05, range=0.3)
```

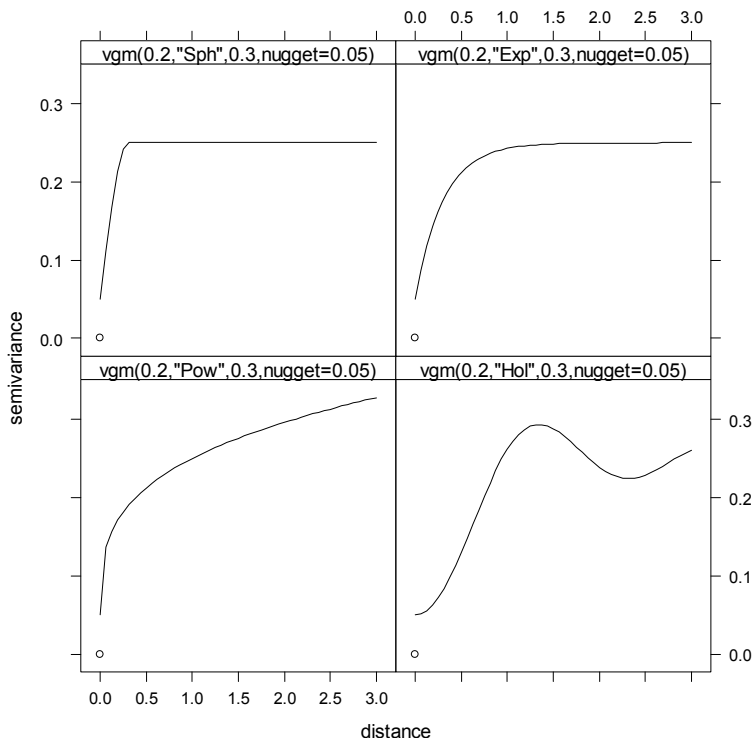


Рис.46. Базисные модели в R: сферическая, экспоненциальная, степенная, с эффектом дырок (ранг = 0.3, эффект самородков = 0.05)

Для создания модели вариограммы используется команда **vgm(...)** с параметрами:

psill	Полная или частная дисперсия (порог) для данной структуры
model	Название базисной модели (“Sph”, “Exp”, и т.д.)
range	Ранг
nugget	«Эффект самородков»
anis	Параметры анизотропии (см. далее)

add.to	Имя переменной, уже содержащей модель вариограммы, к которой нужно прибавить данную модель. Используется при моделировании вложенных структур
...	...

Параметры анизотропии имеют вид: **anis = c(angle,ratio)**, где

angle	Направление анизотропии: угол для направления главной оси эллипса, представляющего анизотропию (вдоль которой вариограмма растет наиболее медленно)
ratio	Отношение анизотропии: отношение ранга второй оси и главной оси эллипса

На основании анализа экспериментальной вариограммы наших данных (остатков) выберем для моделирования первой структуры (с коротким рангом) в качестве базисной сферическую модель и зададим следующие значения параметров: «эффект самородков» = 0.0, суммарная дисперсия = 100=0.0+100 (порог), ранг = 1 500. В силу того, что на небольших расстояниях экспериментальная вариограмма показывала одинаковое поведение для различных направлений, будем моделировать изотропное поведение пространственной связи.

```
gam.sph <- vgm(100,"Sph",1500,nugget=0.0)
```

Построим графики, показывающие совместное поведение экспериментальной и модельной вариограмм по четырем направлениям.

```
plot(p.gam,pch=1,col="black",cex=1,panel=my.panel,
      model= gam.sph)
```

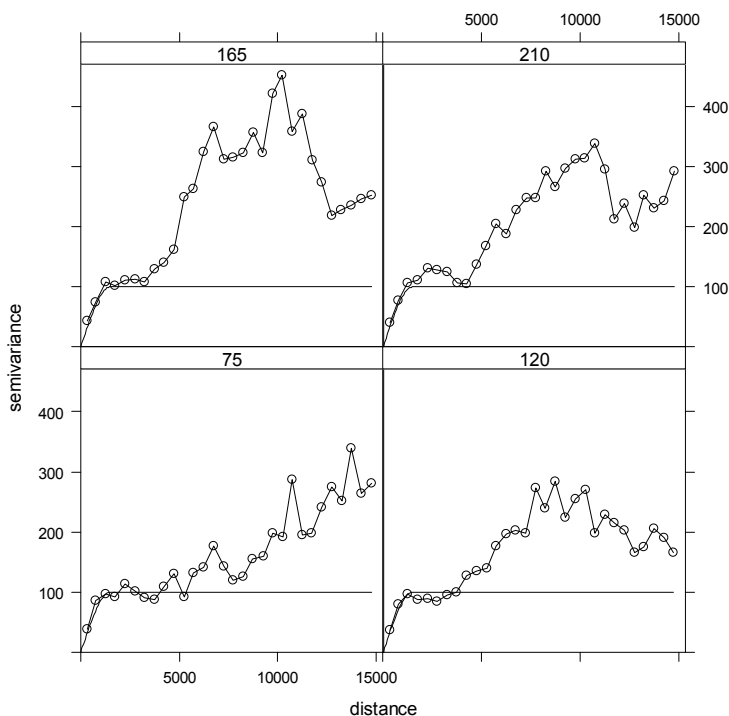


Рис.47. Экспериментальная и модельная (сферическая) вариограммы по 4-м направлениям

Так как экспериментальная вариограмма показывала наличие двух иерархических структур, пробуем добавить в модель вторую структуру, где возьмем как базисную модель с эффектом дырок с параметрами: порог = 180, ранг = 5 000, направление анизотропии = 75°, отношение анизотропии = 0.5.

```
gam.sph <- vgm(80,"Sph",1400,nugget=0.0)
gam.sph_hol <- vgm(180,"Hol",5000,anis=c(75,0.5),
  add.to= gam.sph)
plot(p.gam,pch=1,col="black",cex=1,panel=my.panel,
  model= gam.sph_hol)
```

Получим графики, изображенные на рис. 48:

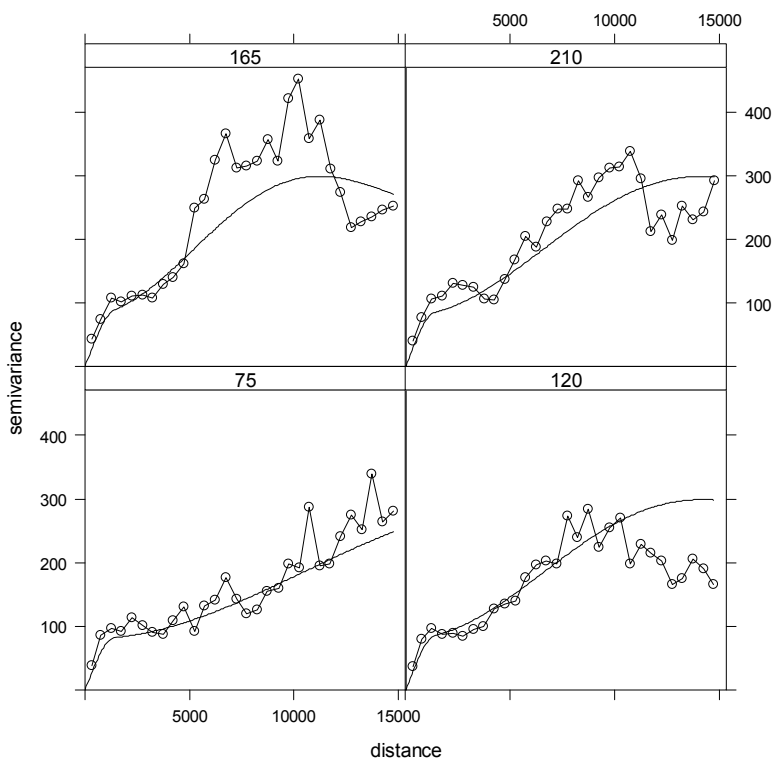


Рис.48. Экспериментальная и модельная (сферическая + модель с эффектом дырок) вариограммы по 4-м направлениям

Проведем другой эксперимент, добавив в модель вторую структуру с помощью Гауссовой базисной модели с параметрами: порог = 180, ранг = 12 000, направление анизотропии = 75°, отношение анизотропии = 0.5.

```
gam.sph <- vgm(80,"Sph",1400,nugget=0.0)
gam.sph_gau <- vgm(180,"Gau",12000,anis=c(75,0.5),
                  add.to= gam.sph)
plot(p.gam,pch=1,col="black",cex=1,panel=my.panel,
     model= gam.sph_gau)
```

Получим такие графики (рис. 49):

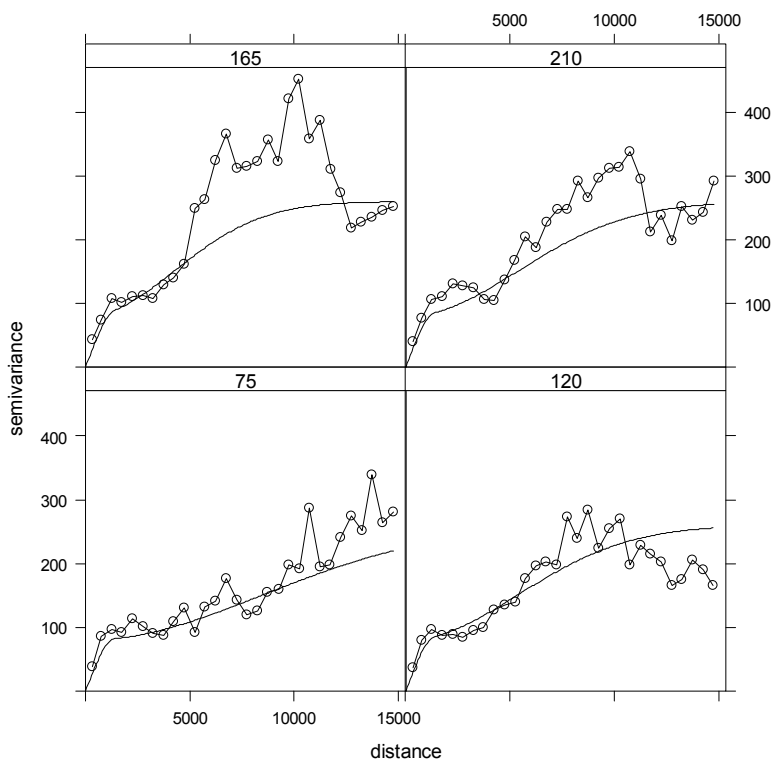


Рис.49. Экспериментальная и модельная (сферическая + Гауссовая) вариограммы по 4-м направлениям

Для того, чтобы проанализировать степень согласия модели $\hat{\gamma}(h)$ с экспериментальной вариограммой $\gamma^*(h)$, также рассмотрим совместно поверхности экспериментальной и теоретической вариограмм. Построим эти поверхности для двух моделей.

С этой целью будем использовать пользовательскую функцию, текст которой приведен в приложении 5:

```
my.vario_map_mod(cur.gam_map, cur.gam_mod, no.X, no.Y,
                 h, lev.lo, lev.hi, lev.step)
```

Параметр	Назначение
cur.gam map	Экспериментальная вариограмма
cur.gam mod	Теоретическая вариограмма (модель)
no.X	Размер ячеек по X
no.Y	Размер ячеек по Y
h	Параметр сглаживания
lev.lo	Минимальный уровень изолиний
lev.hi	Максимальный уровень изолиний
lev.step	Шаг изолиний

Для первой модели:

```
my.vario_map_mod(p.gam_map, gam.sph_hol, 100, 100,5,
                  100, 800, 50)
```

Получаем:

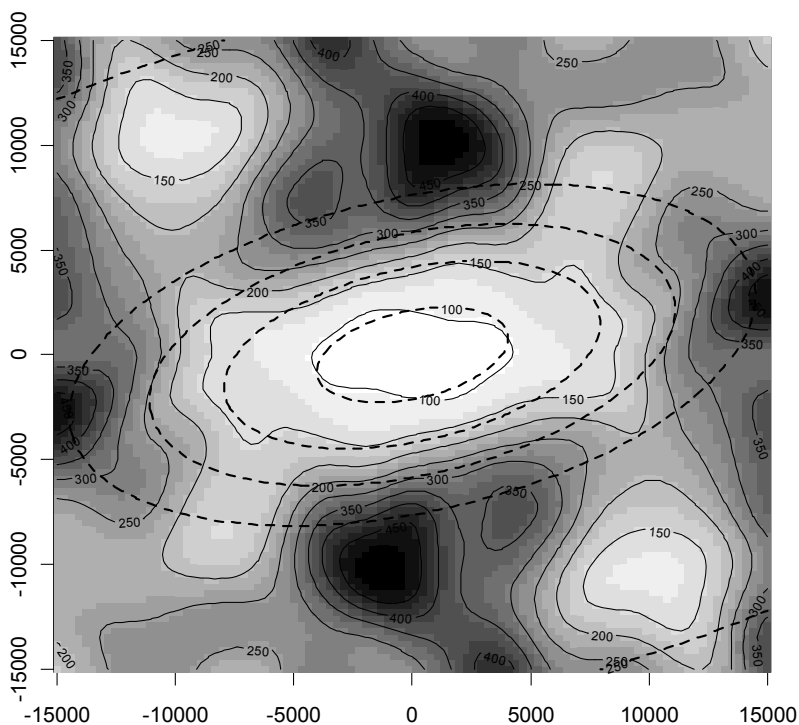


Рис.50. Поверхности экспериментальной и модельной (сферическая + модель с эффектом дырок) вариограмм

Для второй модели:

```
my.vario_map_mod(p.gam_map, gam.sph_gau, 100, 100,5,  
100, 800, 50)
```

Получим следующую картину:

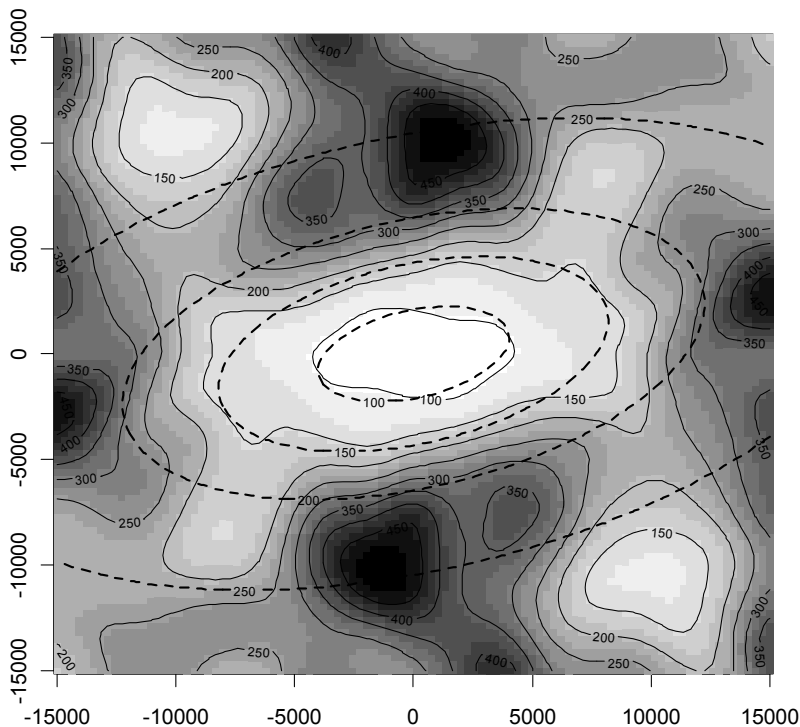


Рис.51. Поверхности экспериментальной и модельной
(сферическая + Гауссовая) вариограмм

Анализ рис.48-51 позволяет сказать, что обе модели достаточно хорошо воспроизводят основные закономерности пространственной структуры (особенно на небольших расстояниях), отражая анизотропное поведение вариограммы. Более предпочтительной на больших расстояниях выглядит первая модель.

Для того, чтобы добиться наилучшего согласия модели с экспериментальными данными можно далее воспользоваться командой `fit.variogram(...)`, которая подгоняет параметры модельной вариограммы специальными методами. Параметры команды:

<code>gam.exp</code>	Экспериментальная вариограмма
<code>gam.model</code>	Модель вариограммы с начальными значениями параметров

Подгоним параметры двух моделей и построим соответствующие графики и поверхности, показывающие совместное поведение экспериментальной и модельных вариограмм.

```
gam.hol.fit<-fit.variogram(p.gam,gam.sph_hol)
gam.hol.fit

  model      psill      range angl anisl
1  Nug    0.00000    0.000   0   1.0
2  Sph   87.35346 1072.468   0   1.0
3  Hol 102.26943 4049.998  75   0.5

plot(p.gam,pch=1,col="black",cex=1,panel=my.panel,
      model= gam.hol.fit)
my.vario_map_mod(p.gam_map, gam.hol.fit, 100, 100,5,
                  100, 800, 50)

#-----
gam.gau.fit<-fit.variogram(p.gam,gam.sph_gau)
gam.gau.fit

  model      psill      range angl anisl
1  Nug    1.840596    0.000   0   1.0
2  Sph   93.340517 1204.215   0   1.0
3  Gau 311.614453 24231.126  75   0.5

plot(p.gam,pch=1,col="black",cex=1,panel=my.panel,
      model= gam.gau.fit)
my.vario_map_mod(p.gam_map, gam.gau.fit, 100, 100,5,
                  100, 800, 50)
```

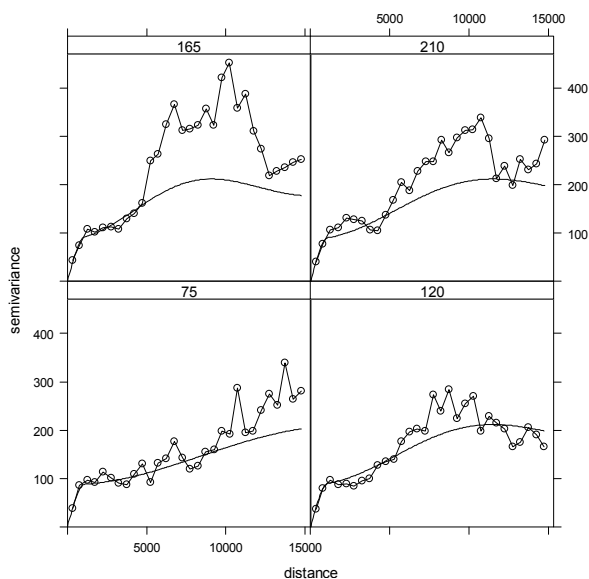


Рис.52. Экспериментальная и модельная после подгонки (сферическая + модель с эффектом дырок) вариограммы по 4-м направлениям

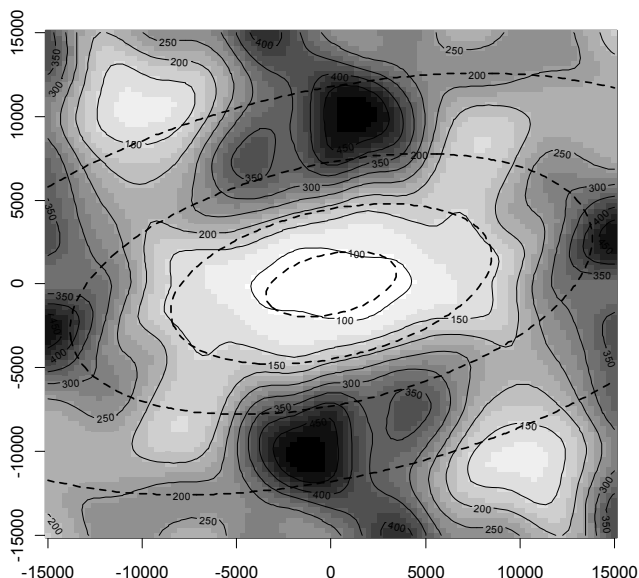


Рис.53. Поверхности экспериментальной и модельной после подгонки (сферическая + модель с эффектом дырок) вариограмм

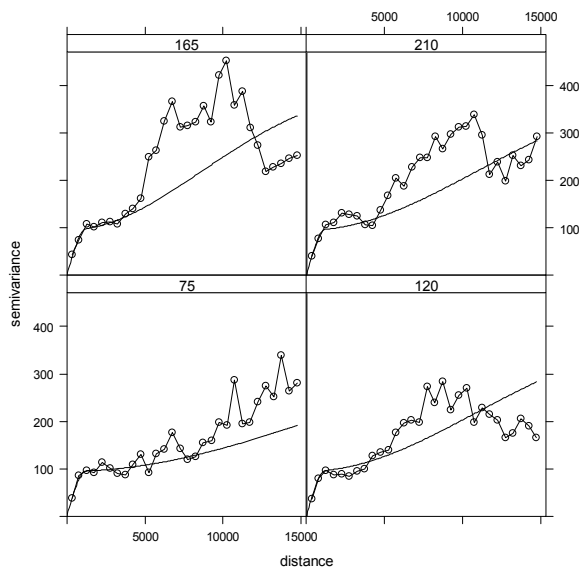


Рис.54. Экспериментальная и модельная после подгонки (сферическая + Гауссовая) вариограммы по 4-м направлениям

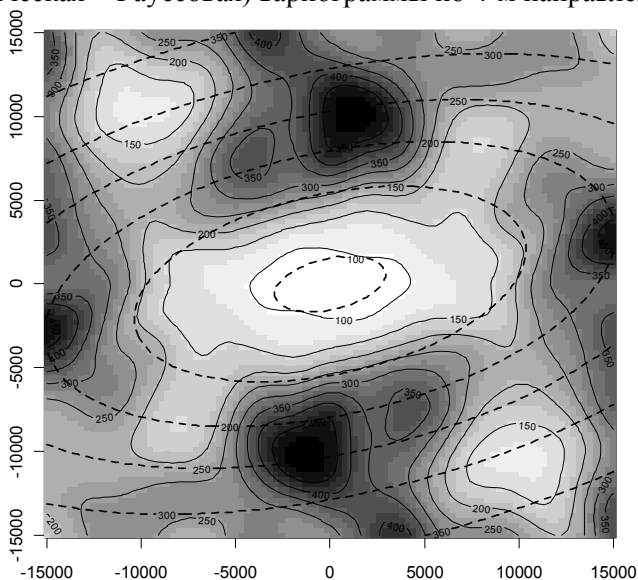


Рис.55. Поверхности экспериментальной и модельной после подгонки (сферическая + Гауссовая) вариограмм

Видно, что после подгонки значения порогов и рангов вложенных структур обеих моделей несколько изменились. В первой модели порог сферической базисной модели изменился с 80 на 87.35346, ранг с 1400 до 1072.468, а порог базисной модели с эффектом дырок изменился со 180 на 102.26943, ранг с 5000 на 4049.998. Во второй модели добавился небольшой эффект самородков (1.840596), порог сферической базисной модели вместо 80 стал равным 93.340517, ранг из 1400 стал равным 1204.215, тогда как порог базисной Гауссовой модели сменился со 180 на 311.614453, ранг с 12000 на 24231.126.

Окончательно, в качестве модели вариограммы исследуемой переменной (остатков) выберем модель `gam.hol.fit`, представляющую собой линейную комбинацию двух базисных моделей - сферической и модели с эффектом дырок, параметры которых уточнялись путем подгонки (рис.52, 53). В явном виде эта модель выписывается следующим образом:

$$\hat{\gamma}(|\mathbf{h}|) = 87 \cdot \left(1 - e^{-3 \cdot |\mathbf{h}|/1072}\right) + 102 \cdot \left(1 - \frac{\sin(|\mathbf{h}_c|/4050)}{|\mathbf{h}_c|/4050}\right),$$

где $|\mathbf{h}_c| = \sqrt{x^2 + y^2}$,

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \cos(75^\circ) & \sin(75^\circ) \\ -\sin(75^\circ) & \cos(75^\circ) \end{bmatrix} \begin{bmatrix} h_x \\ h_y \end{bmatrix}$$

В заключение заметим, что построение модели вариограммы $\hat{\gamma}(\mathbf{h})$ предпочтительнее проводить вручную с учетом знаний эксперта, чем в автоматическом режиме. На практике анализ и интерпретация поведения вариограммы, ее моделирование должны выполняться в контексте знаний об изучаемом явлении.

Пример оценки пространственной переменной методом кригинга в R

Задача геостатистики - оценить значения изучаемой пространственной переменной $z(\mathbf{x})$ в произвольных точках области исследования на основе анализа ее значений, измеренных в ограниченном числе выборочных точек.

По результатам проведенного нами анализа данных (столбец V03 набора данных stud.dat) покажем, как производится такая оценка.

Оценка остатков методом ординарного кригинга

Как было показано выше, в рассмотренном случае в данных присутствовал пространственный тренд. Поэтому на этапе предварительной обработки данных была проведена процедура детрендрования – вычитании модели тренда $\hat{m}(\mathbf{x})$ из данных, и переход к остаткам $e_i = z_i - \hat{m}(\mathbf{x}_i)$. Оценка остатков e^* в произвольной точке (точках) территории может быть получена методом ординарного кригинга с использованием построенной модели вариограммы остатков $\hat{\gamma}(\mathbf{h})$.

Для вычисления оценки ординарным кригингом в R используется команда `krige(...)` с параметрами:

z~1	Переменная, для которой выполняется оценка
locations	Пространственная структура с данными
newdata	Пространственная структура с точками, где должна быть выполнена оценка
model	Модель вариограммы
...	...

Например, оценим значение остатка в точке с координатами $x=40000$, $y=30000$:

```
new<-data.frame(x=40000,y=30000)
coordinates(new) = ~x+y
e0 <- krige(Res~1, p, new, model = gam.hol.fit)
e0["var1.pred"]
      coordinates  var1.pred
1 (40000, 30000) -0.4931573
```

Получим оценку остатков в ячейках регулярной сетки, полностью покрывающей территорию исследования.

```
# границы сетки - по данным
x.lim<- range(coordinates(p) [, "X"])
y.lim<- range(coordinates(p) [, "Y"])
# шаг сетки
x.step<-100
y.step<-100
# позиции линий сетки
x.pos<-seq(x.lim[1],x.lim[2],by=x.step)
y.pos<-seq(y.lim[1],y.lim[2],by=y.step)
# собственно сетка, пока в виде таблицы
p.grid<-as.data.frame(expand.grid(x.pos,y.pos))
names(p.grid)<-c("x","y")
# сетка как специальная структура gstat
gridded(p.grid) = ~x+y
# ординарный кригинг с заданной моделью вариограммы
e <- krige(Res~1, p, p.grid, model = gam.hol.fit)
```

Для визуализации пространственного поведения геостатистической оценки остатков построим картографическое отображение, для чего воспользуемся методом попиксельного изображения.

```
spplot(e["var1.pred"], main = "Модель остатков",
      xlab="X", ylab="Y", scales=list(draw=TRUE),
      col.regions= gray((32:0)/32), asp=1)
```

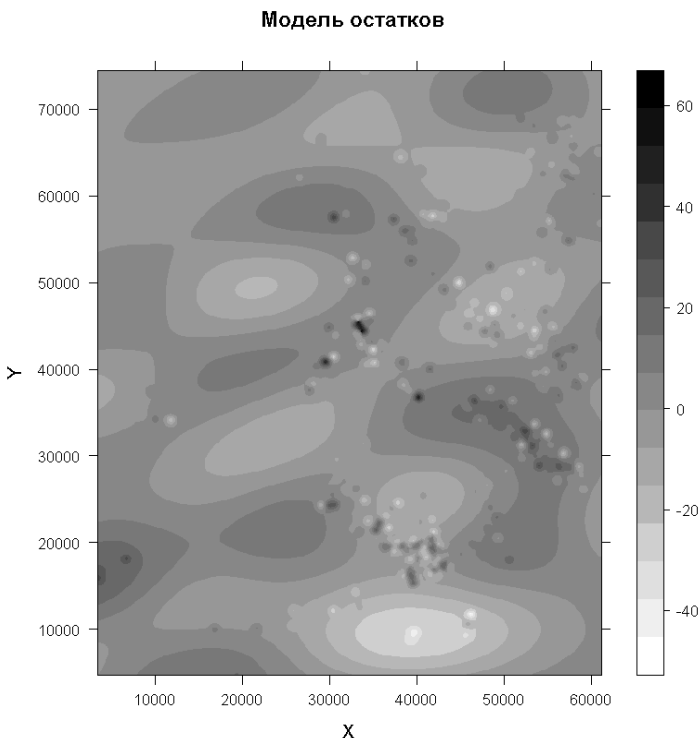


Рис.56. Оценка остатков методом кригинга

Отобразим сетку со значениями стандартной ошибки кригинга в каждой ячейке, показывающей степень доверия к полученным e^* (величину пространственной неопределенности).

```
spplot(e["var1.var"], main = "Дисперсия кригинга",
       xlab="X", ylab="Y", scales=list(draw=TRUE),
       col.regions= gray((0:32)/32), asp=1)
```

Можно видеть (рис.57), что как и следовало ожидать, наименьшие ошибки оценки будут там, где рядом расположены выборочные точки. Чем дальше от точек выборки, тем пространственная неопределенность становится выше.

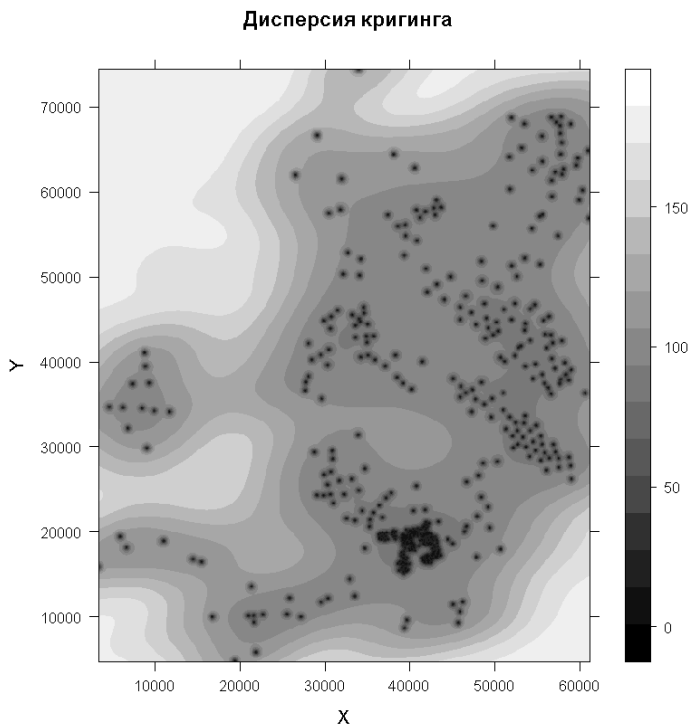


Рис.57. Стандартная ошибка кригинга

Оценка пространственного тренда на основе модели тренда

Найдем оценку пространственного тренда в произвольной точке (точках) исследуемой территории согласно модели тренда $\hat{m}(x)$, полученной на этапе анализа данных. Вычислим модельное значение в точке с координатами $x=40000$, $y=30000$:

x=40000

y=30000

trend=(-1009)+0.002667*x+0.002080*y-0.000000003525*x*y

trend

-844.15

Вычислим модельные значения в ячейках заданной выше сетки и построим картографическое отображение тренда:

```
p.grid_t<-as.data.frame(expand.grid(x.pos,y.pos))
names(p.grid_t)<-c("x","y")
p.grid_t<- data.frame(p.grid_t$x, p.grid_t$y,
  (-1009)+0.002667*p.grid_t$x+0.002080*p.grid_t$y
  -0.000000003525*p.grid_t$x*p.grid_t$y)
names(p.grid_t)<-c("x","y","trend")
gridded(p.grid_t) = ~x+y
spplot(p.grid_t, main = "Модель тренда", xlab="X",
  ylab="Y", scales=list(draw=TRUE),
  col.regions= gray((32:0)/32), asp=1)
```

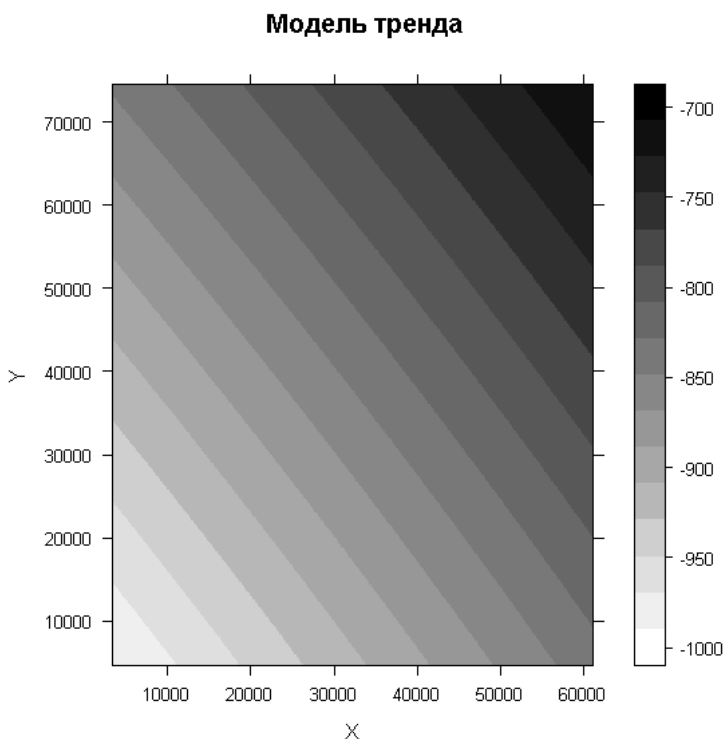


Рис.58. Оценка пространственного тренда

Оценка изучаемой пространственной переменной и построение прогнозной карты

На заключительном шаге получим искомую оценку изучаемой пространственной переменной на исследуемой территории. Например, найдем ее значение в точке с координатами $x=40000$, $y=30000$:

```
z= e0$var1.pred+trend
z
-844.6432
```

Получим оценку пространственной переменной в ячейках регулярной сетки, полностью покрывающей территорию исследования.

```
p.grid_z<-as.data.frame(expand.grid(x.pos,y.pos))
names(p.grid_z)<-c("x","y")
p.grid_z= data.frame(p.grid_z$x, p.grid_z$y,
                    as.data.frame(e)[,"var1.pred"]+
                    as.data.frame(p.grid_t)[,"trend"])
names(p.grid_z)<-c("x","y","z")
gridded(p.grid_z) = ~x+y
```

Покажем ее поведение, построив картографическое отображение.

```
image(p.grid_z, xlab="X", ylab="Y", main = "Модель
изучаемой переменной", col=gray((16:0)/16),asp=1)
tmp.z<-matrix(p.grid_z$z,ncol=length(x.pos),
              nrow=length(y.pos), byrow=TRUE)
contour(x.pos,y.pos,t(tmp.z),
        levels=seq(-2000,-500,by=50),col="black",
        add=TRUE,asp=1)
```

Дополнительно нанесем на карту точки выборки, отобразив их символами, имеющими размер, соответствующий значению переменной в этих точках.

```
points(p, cex=0.2+2*(p$V03-min(p$V03))/(max(p$V03)-
min(p$V03)))
```

Модель изучаемой переменной

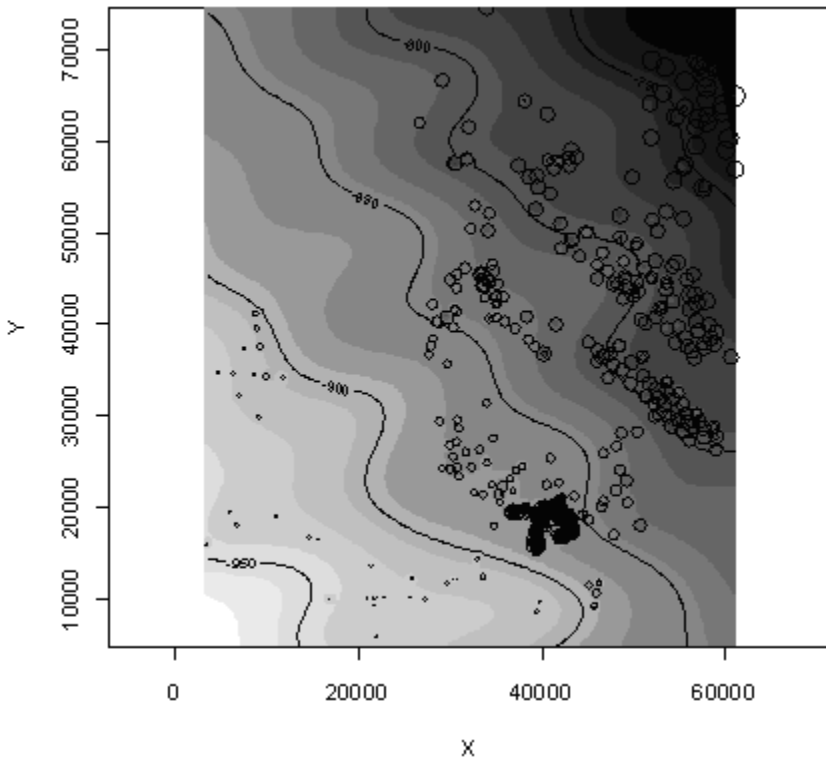


Рис.59. Оценка изучаемой пространственной переменной

Таким образом, поставленная задача - реконструировать изучаемое явление в области исследования на базе его значений, измеренных в ограниченном числе точек, - решена. При этом, кроме модели самой пространственной переменной, получена и величина пространственной неопределенности, отображающая качество (точность) модельных результатов.

Список литературы

1. Давид М. Геоestatистические методы при оценке запасов руд / Л.:Недра, 1980. - 360 с.
2. Девис Дж. Статистика и анализ геологических данных / М.:Мир, 1997. - 572 с.
3. Мухарамова С.С., Савельев А.А., Пилюгин А.Г. Основы геоestatистического анализа и моделирования в экологии. Учебно-методическая разработка / Казань: Лаборатория оперативной полиграфии КГУ, 2002а. – 38 с.
4. Мухарамова С.С., Савельев А.А., Пилюгин А.Г. Геоestatистический анализ данных в экологии. Учебно-методическая разработка / Казань: Лаборатория оперативной полиграфии КГУ, 2002б. – 54 с.
5. Савельев А.А. Моделирование пространственной структуры растительного покрова (геоинформационный подход) / Казань: Казанский гос.университет, 2004. – 244 с.
6. Савельев А.А., Пилюгин А.Г., Мухарамова С.С. Основные понятия языка R. Учебно-методическая разработка / Казань: Изд-во Казан. ун-та, 2007. – 30 с.
7. Савельев А.А., Пилюгин А.Г., Мухарамова С.С. Использование языка R для статистической обработки данных. Учебно-методическая разработка / Казань: Изд-во Казан. ун-та, 2007. - 30 с.
8. Chiles J.-P., Delfiner P. Geostatistics. Modeling Spatial Uncertainty / J.Wiley&Sons,Inc., 1999. - 695 p.
9. Cressie N.A.C. Statistics for Spatial Data / New York, 1991. - 900 p.

10. Deutsch C.V., Journel A.G. GSLIB Geostatistical Software Library / Oxford University Press, 1992. - 340 p.
11. Englund E., Sparks A. Geo-EAS 1.2.1. User's Guide. US-EPA Report, 1991.
12. Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer and and Andreas Weingessel (2011). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-25. <http://CRAN.R-project.org/package=e1071>.
13. Goovaerts P. Geostatistics for Natural Resources Evaluation / Oxford University Press, 1997. - 483 p.
14. Pannatier Y. VARIOWIN 2.1 / Inst.of Mineralogy and Petrography, University of Lausanne, Switzerland, 1994.
15. Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30: 683-691.
16. Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), <http://cran.r-project.org/doc/Rnews/>.
17. R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
18. Ripley B. The R project in statistical computing // MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network. - 2001. - №1(1). - P.23-25.
19. Roger S. Bivand, Edzer J. Pebesma, Virgilio Gomez-Rubio, 2008. Applied spatial data analysis with R. Springer, NY. <http://www.asdar-book.org/>.

20. Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York.
21. Saveliev A.A., Mukharamova S.S., Chizhikova N.A., Budgey R., Zuur A.F. Spatially continuous data analysis and modelling // Analysing ecological data. / A.F. Zuur, E.N. Ieno, G.M. Smith. - Springer-Verlag, 2007. - Chapter 19. - P. 341-372. - 672 p.
22. Surfer for Windows Users Guide. Golden Software Inc. USA, 1998.

Приложения

Приложение 1

```
#=====
#вспомогательная функция (вычисляет все положения окна)
#=====
# генерация массива окон.
# r.x0,r.y0 - левый нижний угол границы области
# r.x1,r.y1 - правый верхний угол границы области
# w.size     - размер (ширина и высота) окна
# w.nshift   - число сдвигов окна (внутри w.size)
#=====
my.win2<-function(r.x0,r.y0,r.x1,r.y1,w.size,w.nshift){
  b.dxy<-w.size/w.nshift          # шаг по XY
  # число шагов по X
  b.nx<-round((r.x1-r.x0)/b.dxy)+w.nshift
  # число шагов по Y
  b.ny<-round((r.y1-r.y0)/b.dxy)+w.nshift
  b.xi<-r.x0+(-w.nshift+1:(b.nx-1))*b.dxy
  b.yi<-r.y0+(-w.nshift+1:(b.ny-1))*b.dxy
  tmp<-expand.grid(b.xi,b.yi)
  tmp<-data.frame(x0=tmp[,1],y0=tmp[,2],
                  x1=tmp[,1]+w.size,y1=tmp[,2]+w.size)
  tmp
}
#=====
# --- декластеризация с использованием локального окна.
#=====
my.declus<-function(p,cur.X,cur.Y,cur.var,minmax.flag,
                    n.size,w.size0,w.size1,w.nshift){
  cur.X=p[,cur.X]
  cur.Y=p[,cur.Y]
  # задаем размеры области
  r.x0=min(cur.X)
  r.y0=min(cur.Y)
  r.x1=max(cur.X)
  r.y1=max(cur.Y)
  # вычисляем шаг изменения размера
  w.dsize<-(w.size1-w.size0)/(n.size-1)
  # матрица для результатов
```

```

# 1-й столбец - среднее
# 2-й столбец - число окон, по которым вычисл.среднее
# 3-й столбец - размер окна
tot.sum<-matrix(0,ncol=3,nrow=n.size)
# цикл по всем размерам окна
for(i.size in 1:n.size) {
  w.size<-w.size0+(i.size-1)*w.dsize
  # получить все возможные положения окна
  # (координаты углов)
  tmp.win<-my.win2(r.x0,r.y0,r.x1,r.y1,
                   w.size,w.nshift)
  # цикл по всем положениям окна
  for(i in 1:nrow(tmp.win)) {
    # точки, попавшие в окно
    cur.ind<-((tmp.win$x0[i] <= cur.X &
               cur.X <= tmp.win$x1[i])&
              (tmp.win$y0[i] <= cur.Y &
               cur.Y <= tmp.win$y1[i]))
    # если попало больше нуля, то ...
    if(sum(cur.ind) > 0) {
      # посчитать среди них среднее,
      # и прибавить к общей сумме средних
      cur.mean<-mean(p[cur.ind,cur.var],na.rm=TRUE)
      tot.sum[i.size,1]<-tot.sum[i.size,1]+cur.mean
      tot.sum[i.size,2]<-tot.sum[i.size,2]+1
    }
  }
  # заполнить строку матрицы для данного размера окна
  if(tot.sum[i.size,2] > 0) {
    tot.sum[i.size,1]<-tot.sum[i.size,1]/
      tot.sum[i.size,2]
  } else {
    tot.sum[i.size,1]<-NA
  }
  tot.sum[i.size,3]<-w.size
}
# вычисляем декластеризованное значение
if(minmax.flag == 1) {
  i.win<-which.max(tot.sum[,1]);
  print(c(i.win,tot.sum[i.win,1]))
} else {
  i.win<-which.min(tot.sum[,1]);

```



```

    print(c(i.win,tot.sum[i.win,1]))
}
# вычисляем веса
p$Weight<-0
w.size<-tot.sum[i.win,3]
tmp.win<-my.win2(r.x0,r.y0,r.x1,r.y1,w.size,w.nshift)
for(i in 1:nrow(tmp.win)) {
    cur.ind<-(tmp.win$x0[i] <= cur.X &
              cur.X <= tmp.win$x1[i] &
              tmp.win$y0[i] <= cur.Y &
              cur.Y <= tmp.win$y1[i])
    if((cur.num<-sum(cur.ind)) > 0) {
        cur.wts<-nrow(p) / (cur.num*tot.sum[i.win,2])
        p$Weight[cur.ind]<-p$Weight[cur.ind]+cur.wts
    }
}
# контроль (должно совпадать с найденным средним)
print(sum(p[,cur.var]*p$Weight)/nrow(p))
p$Weight<-round(p$Weight,5)
# сохранить результаты
sink("report.txt")
cat(sprintf(" minmax flag = %d",minmax.flag),"\\n")
cat(sprintf(" ncell min max =  %d %f %f",
            n.size,w.size0,w.size1),"\\n")
cat(sprintf(" offsets = %d",w.nshift),"\\n")
cat(sprintf(" There are      %d data with:",
            nrow(p)),"\\n")
cat(sprintf("      mean value                = %f",
            mean(p[,cur.var],na.rm=TRUE)),"\\n")
cat(sprintf("      minimum and maximum      = %f %f",
            min(p[,cur.var],na.rm=TRUE),
            max(p[,cur.var],na.rm=TRUE)),"\\n")
cat(sprintf("      size of data vol in X =  %f",
            max(cur.X,na.rm=TRUE)-
            min(cur.X,na.rm=TRUE)),"\\n")
cat(sprintf("      size of data vol in Y =  %f",
            max(cur.Y,na.rm=TRUE)-
            min(cur.Y,na.rm=TRUE)),"\\n")
cat(sprintf("      declustered mean          = %f",
            tot.sum[i.win,1]),"\\n")
cat(sprintf("      min and max weight        = %f %f",
            min(p$Weight,na.rm=TRUE),

```

```

        max(p$Weight, na.rm=TRUE)), "\n")
sink()
tmp<-data.frame(Size =c(0,tot.sum[,3]),
                Mean=c(round(mean(p[,cur.var],
                                na.rm=TRUE), 3), round(tot.sum[,1], 3)))
write.table(tmp, "Mean_vs_window_size.txt",
            row.names=FALSE, quote=FALSE)
cur.fn<-sprintf("data_%s_w.dat", cur.var)
write.table(p, cur.fn, row.names=FALSE, quote=FALSE,
            sep="\t")
p
}

```

Приложение 2

```
#=====
# --- СТАТИСТИКИ В ЛОКАЛЬНОМ ОКНЕ
#=====
my.movewin<-function(p, cur.X, cur.Y, cur.var,
                    w.nmin, w.size, w.nshift) {
  cur.X=p[, cur.X]
  cur.Y=p[, cur.Y]
  # задаем размеры области
  r.x0=min(cur.X)
  r.y0=min(cur.Y)
  r.x1=max(cur.X)
  r.y1=max(cur.Y)
  # задать возможные положения окна
  tmp.win<-my.win2(r.x0, r.y0, r.x1, r.y1, w.size, w.nshift)
  # матрица для результатов
  # 1-й столбец - X
  # 2-й столбец - Y
  # 3-й столбец - среднее
  # 4-й столбец - дисперсия
  # 5-й столбец - число точек
  tot.stat<-matrix(NA, ncol=5, nrow=nrow(tmp.win))
  # цикл по всем положениям окна
  for(i in 1:nrow(tmp.win)) {
    # точки, попавшие в окно
    cur.ind<-(tmp.win$x0[i] <= cur.X &
              cur.X <= tmp.win$x1[i] &
              tmp.win$y0[i] <= cur.Y &
              cur.Y <= tmp.win$y1[i])
    # если попало больше порога, то ...
    if(sum(cur.ind) > w.nmin) {
      # посчитать центр окна, среднее, дисперсию,
      # и заполнить строку
      tot.stat[i,1]<-0.5*(tmp.win$x0[i]+tmp.win$x1[i])
      tot.stat[i,2]<-0.5*(tmp.win$y0[i]+tmp.win$y1[i])
      tot.stat[i,3]<-mean(p[cur.ind, cur.var],
                        na.rm=TRUE)
      tot.stat[i,4]<-var (p[cur.ind, cur.var],
                        na.rm=TRUE)
      tot.stat[i,5]<-sum(cur.ind)
    }
  }
}
```

```

tmp.ind<-!apply(is.na(tot.stat),1,any)
tot.stat<-as.data.frame(tot.stat[tmp.ind,])
names(tot.stat)<-c("X","Y","Mean","Var","Num")
# сохранить результаты
tmp<-tot.stat
write.table(tmp,"out.dat",row.names=FALSE,
            quote=FALSE)
}

```

```
#=====
# --- h-h plot (собственная функция)
#=====
my.hscat<-function(cur.var,cur.data,d.lo,d.hi,
                   variogram,alpha,tol.hor) {
  cur.form<-as.formula(sprintf("%s~1",cur.var))
  # строим "облако вариограммы"
  p.cloud<-variogram(cur.form,cur.data,
                    cloud=TRUE,alpha=variogram.alpha,
                    tol.hor=tol.hor)
  p.BigInt = attr(p.cloud,"BigInt")
  # "вытаскиваем" номер "головы"
  p.cloud$head = p.cloud$np %% p.BigInt + 1
  # "вытаскиваем" номер "хвоста"
  p.cloud$tail = p.cloud$np %/% p.BigInt + 1
  # значение в "голове"
  p.cloud$ValHead = cur.data@data[p.cloud$head,cur.var]
  # значение в "хвосте"
  p.cloud$ValTail = cur.data@data[p.cloud$tail,cur.var]
  # нужный диапазон расстояний
  ind<-(p.cloud$dist >= d.lo & p.cloud$dist < d.hi)
  # рисуем точки h-h графика ...
  sbuf<-sprintf("%g < dist < %g",d.lo,d.hi)
  plot(p.cloud$ValHead[ind],p.cloud$ValTail[ind],
       pch=20,cex=1,xlab="Head",ylab="Tail",
       main=sbuf, col=" gray")
  abline(0,1,lwd=1.5,col="black")
  # --- границы доверительного интервала
  # первая главная компонента (поскольку дисперсия
  # положительная, то это диагональ)
  p.PC1<-c( 1,1)/sqrt(2)
  # вторая главная компонента
  p.PC2<-c(-1,1)/sqrt(2)
  # проекции на главные компоненты
  p.PC1load<-as.matrix(cbind(p.cloud$ValHead[ind],
                             p.cloud$ValTail[ind])) %*% p.PC1
  p.PC2load<-as.matrix(cbind(p.cloud$ValHead[ind],
                             p.cloud$ValTail[ind])) %*% p.PC2
  # собственные числа (дисперсии)
  p.lam1<-var(p.PC1load);
  p.lam2<-var(p.PC2load);
```

```

# координаты точек на единичной окружности
p.sn<-sin(seq(0,2*pi,by=0.01))
p.cs<-cos(seq(0,2*pi,by=0.01))
# координаты центра
p.x0<-mean(p.cloud$ValHead[ind])
p.y0<-mean(p.cloud$ValTail[ind])
# рисуем центр
points(p.x0,p.y0,pch=22,col="black",cex=2,
       bg="black")
# радиусы с учетом разложения дисперсии по главным
# компонентам
p.R1<-1.96*sqrt(drop(p.lam1))
p.R2<-1.96*sqrt(drop(p.lam2))
# эллипс доверительного интервала
p.xy<-t(p.PC1%*%t(p.cs*p.R1))+t(p.PC2%*%t(p.sn*p.R2))
lines(p.xy[,1]+p.x0,p.xy[,2]+p.y0,col="black",
      lty=2.0,lwd=3)
}

```

Приложение 4

```
#=====
# --- сглаженная поверхность вариограммы
#=====
my.vario_map<-function(cur.gam_map, no.X, no.Y, h,
                        lev.lo, lev.hi, lev.step) {
  library(MBA)
# координаты и значение вариограммы
  vario.surf.data<-
  as.matrix(cbind(cur.gam_map$map@coords,
                  cur.gam_map$map@data$var1))
# только те значения, которые реально вычислились
  ind<-!is.na(vario.surf.data[,3])
  vario.surf.mba<-mba.surf(vario.surf.data[ind,],
                           no.X, no.Y, h=h,
                           extend=TRUE)$xyz.est
# --- рисуем поверхность и изолинии
  image(vario.surf.mba,col=gray((16:0)/16))
  contour(vario.surf.mba,levels=seq(lev.lo, lev.hi,
                                   by= lev.step), col=" black",add=TRUE)
}
```

```
#=====
# --- поверхности экспериментальной и
# --- модельной вариограмм
#=====
my.vario_map_mod<-function(cur.gam_map, cur.gam_mod,
                           no.X, no.Y, h,
                           lev.lo, lev.hi, lev.step) {
  library(MBA)
# координаты и значение вариограммы
  vario.surf.data<-
  as.matrix(cbind(cur.gam_map$map@coords,
                  cur.gam_map$map@data$varl))
# только те значения, которые реально вычислились
  ind<-!is.na(vario.surf.data[,3])
  vario.surf.mba<-mba.surf(vario.surf.data[ind,],
                           no.X, no.Y, h=h,
                           extend=TRUE)$xyz.est

  g.x=vario.surf.mba$x
  g.y=vario.surf.mba$y
  t.gam<-list(x=g.x,y=g.y,z=matrix(ncol=length(g.x),
                                   nrow=length(g.y)))
  for(i.x in 1:length(t.gam$x)) {
    for(i.y in 1:length(t.gam$y)) {
      g.dir<-atan2(t.gam$x[i.x],t.gam$y[i.y])
      dir.sn <-sin(g.dir)
      dir.cs <-cos(g.dir)
      cur.dist<-sqrt(t.gam$x[i.x]^2+t.gam$y[i.y]^2)
      cur.gam<-variogramLine(cur.gam_mod,
                            dir=c(dir.cs,dir.sn,0),
                            dist_vector=c(cur.dist))
      t.gam$z[i.y,i.x]<-cur.gam$gamma
    }
  }
# --- рисуем поверхность и изолинии
  image(vario.surf.mba,col=gray((16:0)/16))
  contour(vario.surf.mba,levels=seq(lev.lo, lev.hi,
                                   by= lev.step), col="black",add=TRUE)
  contour(t.gam,levels=seq(lev.lo, lev.hi,
                          by= lev.step),col="black",add=TRUE,
          lty=2,lwd=2)
}
```