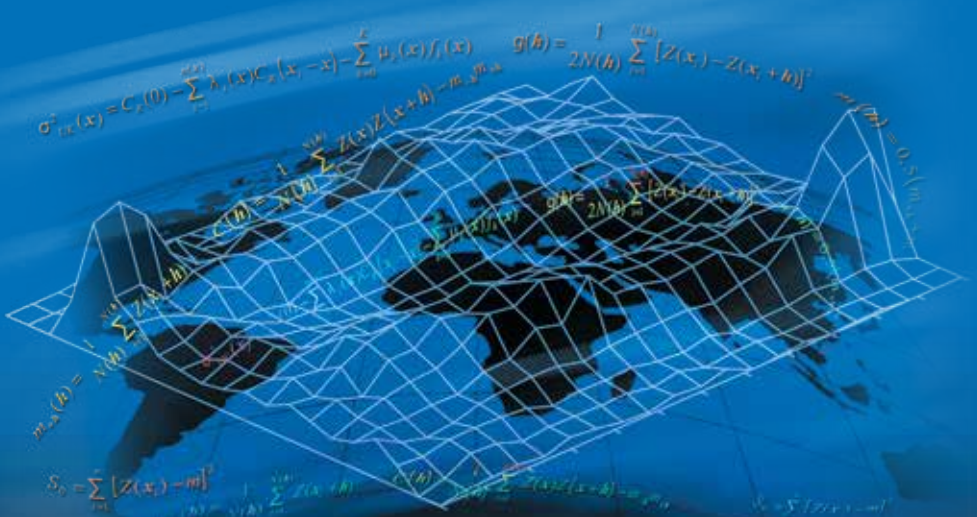
РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИБРАЭ Институт проблем безопасного развития атомной энергетики

В. В. Демьянов, Е. А. Савельева

ГЕОСТАТИСТИКА
ТЕОРИЯ И ПРАКТИКА



НАУКА

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Институт проблем безопасного развития атомной энергетики

В. В. Демьянов, Е. А. Савельева

ГЕОСТАТИСТИКА
теория и практика

Под редакцией
профессора, доктора физико-математических наук
Р. В. Арутюняна

Москва Наука 2010

УДК 91:519.8
ББК 26.8в6
Г35

Рецензенты:
доктор технических наук *Б. И. Яцало,*
доктор физико-математических наук *В. М. Головизнин*

Геостатистика: теория и практика / В. В. Демьянов, Е. А. Савельева ; под ред. Р. В. Арутюняна; Ин-т проблем безопасного развития атомной энергетики РАН. — М. : Наука, 2010. — 327 с. — ISBN 978-5-02-037478-2 (в пер.).

В монографии подробно изложены методы геостатистики и смежных разделов пространственного моделирования. Изложение теории сопровождается примерами использования моделей в различных областях: экологии, геологии, гидрогеологии, нефтедобыче, энергетике, оценке рыбных запасов и т. п. В заключительном разделе очерчены основные направления развития современной геостатистической теории. Издание может быть использовано в качестве учебного пособия. Материал излагается с постепенным усложнением. Для закрепления полученных знаний даны вопросы и упражнения. В книгу включены приложения, позволяющие использовать ее как справочник по геостатистике.

Для ученых, инженеров и практиков, интересующихся проблемами анализа пространственных данных, студентов (геологов, географов, почвоведов, геофизиков, биологов, нефтяников, социологов и др.).

ISBN 978-5-02-037478-2

Содержание

Нашим родным и близким

Введение

История создания книги

Авторы этой книги познакомились с геостатистикой в начале 1990-х гг. В это время в Институте проблем безопасного развития атомной энергетики РАН по инициативе проф. М. Ф. Каневского геостатистика начала активно применяться для анализа и моделирования радиоактивного загрязнения почвы, образовавшегося в результате Чернобыльской аварии. В течение более 10-ти лет лаборатория под руководством М. Ф. Каневского развивала геостатистические приложения для картирования пространственного загрязнения с применением методов геостатистики и искусственного интеллекта. Работы лаборатории в этом направлении поддерживались пятью грантами европейской программы Международной ассоциации содействия сотрудничеству с учеными независимых государств б. СССР (ИНТАС), грантами Civilian Research and Development Foundation (CRDF), Российского фонда фундаментальных исследований, РАН, контрактами с Министерством РФ по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий, совместными европейскими и американскими проектами. Достижения лаборатории в области геостатистики были признаны на ведущих международных форумах (в частности, на Геостатистическом конгрессе, Конференции по математической геологии и Конференции по применению геостатистики для окружающей среды). Сотрудники лаборатории опубликовали более 100 статей и тезисов докладов, защитили одну докторскую и три кандидатские диссертации, в ИБРАЭ РАН по этой тематике были выполнены десятки дипломных работ.

В 1999 г. Всероссийский институт научной и технической информации (ВИНИТИ) выпустил первую книгу по геостатистике на русском языке после ранней работы Ж. Матерона [1968]. Сборник ВИНИТИ, в работе над которым авторы принимали самое активное участие, представлял собой краткое изложение известных моделей геостатистики и описание их применения к картированию радиоактивного загрязнения [Каневский и др., 1999]. Несмотря на ограниченный тираж, сборник оказался очень популяр-

ным — первый и два дополнительных тиража разошлись, даже не поступив в открытую продажу. К нам приходили оклики на него от исследователей, работающих в самых различных сферах — от добычи нефти и газа до рыбного хозяйства.

За 10 лет, прошедших с момента публикации сборника, методы геостатистики нашли широкое применение в России. За это время в нашей стране было издано несколько хороших монографий и статей по этой теме на русском языке, но они ориентированы на специалистов геологов и почвоведов. Наша книга призвана привлечь к геостатистике внимание всех, кто заинтересован в проведении анализа пространственных данных. По сравнению с первым сборником авторы систематизировали описываемые методы, усилили доходчивость изложения, подобрали разнообразные примеры из различных сфер приложений, исправили опечатки и доработали материал.

Мы надеемся, что книга вызовет широкий интерес и будет хорошим подспорьем для многих российских исследователей, практиков, студентов и аспирантов.

Цель и структура издания

Книга — наиболее полное изложение современной геостатистики на русском языке. Содержащийся в ней материал не предполагает специальных знаний по статистике. Теоретические положения сопровождаются большим количеством примеров. Книга может быть использована в качестве учебного пособия: в нее включен ряд упражнений и вопросов.

Издание будет интересно тем, кто сталкивается с пространственной информацией и нуждается в ее анализе, мониторинге и моделировании. Список приложений геостатистики огромен: география и геофизика, окружающая среда и экология, геология и геологоразведка включая добычу нефти и газа, эпидемиология и социология, рыбное и лесное хозяйство и т. п.

Книга состоит из Введения, 11-ти глав, 7-ми приложений и Указателя. Уровень изложения материала постепенно усложняется. Последовательное чтение книги знакомит с пошаговым исследованием пространственных данных. На каждом шаге ставятся задачи и описываются методы их решения. В конце глав приведены списки литературы.

Глава 1 посвящена общим проблемам, связанным с пространственными данными и постановкой различных задач. Она дает общее представление о широком спектре вопросов, которые затрагивает геостатистика.

В Главе 2 введены основные понятия геостатистики и обсуждены ключевые предположения, т. е. закладывается фундамент для понимания методов, изложенных в последующих главах. В эту главу включены также понятия из смежных с геостатистикой областей, таких как анализ сети мониторинга, визуализация данных, пространственное разрешение и пр.

Детерминистические модели интерполяции, изложенные в Главе 3, не являются частью геостатистической теории, однако авторы сочли необходимым включить их в книгу, поскольку эти методы, известные задолго до разработки геостатистической теории, нашли широкое применение в практических исследованиях. Они популярны и в настоящее время, в том числе благодаря своей доступности. В то же время их простота и одновременно ограниченность являются хорошей мотивацией для использования моделей геостатистики.

Глава 4 посвящена ключевой теме геостатистики — исследованию и моделированию пространственной корреляции. Здесь подробно изложено понятие вариограммы — одно из ключевых в классической геостатистике, которое будет использоваться во всех последующих главах.

Геостатистические модели пространственного оценивания семейства кригинга подробно описаны в Главе 5, где рассмотрены различные типы кригинга и приведены примеры моделирования.

Глава 6 посвящена методам многопеременного анализа и моделирования. В ней обсуждены проблемы совместного оценивания нескольких переменных, преимущества и недостатки многопеременных геостатистических моделей.

Вероятностное картирование и моделирование категориальных данных при помощи методов индикаторного кригинга изложены в Главе 7.

В Главе 8 излагаются методы стохастического моделирования пространственных данных. Это наиболее современные методы, находящие все большее применение в различных приложениях. В этой главе представлен весь спектр существующих подходов к стохастическому геостатистическому моделированию (некоторые модели, разработанные совсем недавно, приведены в Главе 11).

В Главу 9 включено несколько примеров исследования реальных данных при помощи геостатистических моделей, которые описаны в предыдущих

главах. В качестве примеров использованы данные по радиоактивному загрязнению почвы и зонированию гидрогеологических слоев. Здесь же приведен сравнительный анализ геостатистических методов на примере картирования риска превышения пороговых значений загрязнения почвы.

Глава 10 посвящена комбинированным методам геостатистики и искусственных нейронных сетей (ИНС), которые были разработаны для решения проблемы анализа и моделирования данных в присутствии нелинейного крупномасштабного тренда.

Глава 11 содержит описание некоторых наиболее перспективных, на наш взгляд, направлений развития современной геостатистики: пространственно-временно́го моделирования, многоточечной статистики, теории байесовской максимальной энтропии.

В приложениях собрана дополнительная информация для облегчения работы с книгой и дальнейшего знакомства с геостатистикой. Математические символы, использованные в формулах, сведены в нотацию в Приложении 1. Приложение 2 содержит определения базовых статистических величин, которые часто используются в книге. Таким образом, книгу можно использовать и как справочник по геостатистике. Для дальнейшего углубленного изучения геостатистики служат краткие обзоры геостатистических монографий, изданных на английском языке (Приложение 3), существующего программного обеспечения (Приложение 4), список избранных геостатистических ресурсов в Интернете (Приложение 5). В Приложении 6 собраны ответы к упражнениям из различных глав книги. Приложение 7 содержит глоссарий ключевых понятий геостатистики.

За рамками данной книги осталось достаточно много смежных тем, которые, однако, не относятся напрямую к геостатистике. Например, географические информационные системы используются в качестве инструмента для получения пространственных данных и отображения результатов моделирования. Также в книге нет описания моделей машинного обучения (искусственных нейронных сетей, машин поддерживающих векторов и др.), которые в настоящее время активно используются наряду и совместно с геостатистикой. Описание методов, основанных на обучении, и их применение для пространственного моделирования можно найти в [Kanevski, Maignan, 2004; Advanced…, 2008].

Данные и примеры исследований, использованные в книге

Для иллюстрации возможностей и особенностей геостатистики помимо синтетических примеров использовались реальные данные из различных областей исследования. Мы специально старались расширить их разнообразие, чтобы показать широту возможных приложений геостатистики. Ниже описаны основные из них.

Климатические данные. Рассматривались два набора климатических данных. Первый — данные по усредненным за 10 дней выпадениям осадков в Швейцарии в 1986 г. Эти данные распространялись в рамках международного конкурса сравнения методов пространственной интерполяции (Spatial Interpolation Comparison — SIC'97) [SIC'97]. Описание данных и полученные результаты опубликованы в [Kanevski et al., 1998; SIC'97]. Второй набор — мгновенный срез поля температуры (результат разового измерения на метеостанциях) в Приаралье. Эти данные распространялись среди участников гранта ИНТАС по Аральскому морю 1072 «Prospect for the development of natural-economic resources in the Kazakh Priaralie». Некоторые результаты их анализа представлены в [Kanevski et al., 2005].

Чернобыльское загрязнение почвы. Данные по загрязнению почвы ^{137}Cs и ^{90}Sr в Брянской области были первыми, на которых авторы использовали геостатистические методы и отрабатывали геостатистическую методологию в приложении к анализу пространственного загрязнения. Эти данные использовались во многих их работах [Kanevsky et al., 1996; Savelieva et al., 1998; Savelieva et al., 2005]. Авторы благодарны сотрудникам ИБРАЭ РАН С. В. Панченко, О. А. Павловскому и И. И. Линге за предоставленные данные и помощь в их обработке и интерпретации. Работы по анализу этих данных были поддержаны международными грантами CRDF RG2-2236, INTAS 94-2361 и ИНТАС INTAS 97-31726.

Загрязнение почвы и донных отложений. Кроме данных по загрязнению радиоактивными изотопами почвы в результате Чернобыльской аварии, для иллюстрации использовались данные по пространственному загрязнению радиоактивными изотопами и тяжелыми металлами. Анализ данных по загрязнению ^{241}Am проводился в рамках совместных исследований ИБРАЭ РАН и Sandia National Laboratory по программе РАН и Министерства энергетики США [Kanevski et al., 2002; Kanevski et al., 2006]. Данные по загрязнению тяжелыми металлами донных отложений Женевского озера были получе-

ны в рамках сотрудничества по программе ИНТАС (гранты INTAS 96-1957и INTAS 99-00099) [Parkin et al., 2001].

Гидрогеологические данные. Приведен пример моделирования гидрогеологического осадочного слоя в рамках гидрогеологической системы из 10-ти слоев, а также зонирования гидрогеологического слоя. Анализ этих данных проводился в рамках совместных исследований ИБРАЭ РАН и Pacific Northwest National Laboratory по программе РАН и Министерства энергетики США [Savelieva et al., 2002].

Электропотребление. В Главе 10 рассмотрен пример использования геостатистики для описания неопределенности прогноза временно́го ряда по электропотреблению в Московском регионе. Данные по электропотреблению были предоставлены «Энергосбытом» «Мосэнерго» [Арутюнян и др., 1999]. Работа проводилась в рамках соглашения о научно-техническом сотрудничестве между ОАО «Энергосбыт» «Мосэнерго» и ИБРАЭ РАН.

Распределение популяции крабов. В качестве иллюстрации применения нелинейных методов геостатистики использовались данные траловых съемок пространственного распределения различных видов крабов (краб опилио, краб Берди и камчатский краб). Данные получены от Всероссийского НИИ рыбного хозяйства и океанографии (ВНИРО) для проведения совместных исследований [Savelieva et al., 2007]. Авторы благодарны С. М. Гончарову и В. А. Бизикову за предоставленные данные и продуктивное обсуждение полученных результатов.

Издание этой книги было бы невозможно без поддержки и помощи широкого круга людей в России и за рубежом. В первую очередь авторы глубоко признательны проф. М. Ф. Каневскому — нашему бывшему научному руководителю и другу — за приобщение нас к геостатистике и бесценный опыт многолетней совместной работы, а также за глубокие обсуждения и идеи, многие из которых нашли место в этой книге. Мы рады возможности поддерживать постоянные научные контакты и вести совместные исследования с М. Ф. Каневским, который руководит Институтом геоматики и анализа риска в Университете Лозанны (IGAR, University of Lausanne), Швейцария. При написании книги мы также использовали материалы книги, изданной М. Ф. Каневским на английском языке [Kanevski, Maignan, 2004], и сборника под его редакцией [Advanced…, 2008].

Издание нашей книги было поддержано ИБРАЭ РАН. Авторы благодарны чл.-кор. РАН проф. Л. А. Большову и проф. Р. В. Арутюняну за поддержку и помощь.

Авторы благодарны сотрудникам лаборатории моделирования окружающей среды и системных исследований С. Ю. Чернову и В. А. Тимонину за разработку пакета программ «Геостат Офис», который был незаменим в нашей научной деятельности и активно использовался для работы над настоящей книгой [Kanevski, Maignan, 2004]. Также авторы признательны коллегам и студентам ИБРАЭ РАН за участие в обсуждениях различных аспектов геостатистики и их приложений.

В. В. Демьянов благодарен проф. М. Кристи (M. Christie) из Университета Хериот-Ватт (Heriot-Watt University), Великобритания, за поддержку при написании книги, советы и помощь в научных исследованиях. Также В. В. Демьянов признателен проф. П. Корбетту (P. Corbett), который ведет курс геомоделирования в Университете Хериот-Ватт, за полезные обсуждения и идеи.

В работе над книгой авторам помогали курсы лекций, которые они читают студентам. Курс «Методы анализа данных» для студентов III курса МФТИ читает Е. А. Савельева в ИБРАЭ РАН. В. В. Демьянов читает курс прикладной геостатистики для студентов-магистров в Институте нефтяного инжиниринга (Institute of Petroleum Engineering) Университета Хериот-Ватт.

Авторы благодарят коллегу и старого друга проф. М. Майгнана (M. Maignan) из Университета Лозанны за многолетнее сотрудничество, поддержку и обсуждение проблем геостатистики. Авторы признательны проф. Д. Кристакосу (G. Christakos) из Университета Сан-Диего, США, за многолетнее сотрудничество, помощь в освоении теории байесовской максимальной энтропии и предоставление пакета программ BMElib для исследований, результаты которых приведены в настоящей книге [Christakos, 2000; Christakos et al., 2002]. Авторы также благодарны проф. Дж. Каерсу (J. Caers) и Стэнфордскому центру прогнозирования месторождений (SCRF, Stanford University, USA) за возможность использования моделей многоточечной статистики [SGeMS] и помощь в их освоении.

Литература

Арутюнян Р. В., Богданов В. И., Большов Л. А. и др. Прогноз электропотребления: Анализ временных рядов, геостатистика, искусственные нейронные сети. — М., 1999. — 45 с. — (Препринт ИБРАЭ; IBRAE-99-05).

Каневский М., Демьянов В., Савельева Е. и др. Элементарное введение в геостатистику. — М., 1999. — 136 с. — (Проблемы окружающей среды и природных ресурсов / ВИНИТИ; № 11).

Матерон Ж. Основы прикладной геостатистики. — М.: Мир, 1968. — 407 с.

Advanced Mapping of Environmental Data: Geostatistics, Machine Learning and Bayesian Maximum Entropy / Ed. M. Kanevski; ISTE Ltd. — [S. l.], 2008. — 313 p.

Christakos G. Modern Spatiotemporal Geostatistics. — New York: Oxford Univ. Press, 2000.

Christakos G., Bogaert P., Serre M. Temporal GIS: Advanced Functions for Field-Based Applications. — [S. l.]: Springer, 2002. — 250 p.

Kanevski M., Arutyunyan R., Bolshov L. et al. Geostatistical Portrayal of the Chernobyl Fallout // Geostatistics Wollongong ’96 / Ed. E. Y. Baafi, N. A. Schofield. — [S. l.]: Kluwer Academic Publ., 1996. — Vol. 2. — P. 1043—1054.

Kanevski M., Demyanov V., Chernov S. et al. Neural Network Residual Kriging Application For Climatic Data // The J. of Geographic Information and Decision Analysis (GIDA). — 1998. — Vol. 2, N 2.

Kanevski M., Maignan M. Analysis and modelling of spatial environmental data. — Lausanne: EPFL Press, 2004. — 288 p. — (With a CD and educational/research MS Windows software tools) (http://www.ppur.org/auteurs/1000772.html).

Kanevski M., Pozdnukhov A., McKenna S. et al. (Transductive decision-oriented mapping of environmental data // Proceedings of IAMG2002 conference, September 2002, Berlin, Germany. — [S. l.], 2002. — P. 519—524.

Kanevski M., Pozdnukhov A., Tonini M. et al. Statistical Learning Theory for Geospatial Data. Case study: Aral Sea // 14th European colloquium on Theoretical and Quantitative Geography. Portugal, September 2005. — [S. l.], 2005.

Kanevski M., Demyanov V., Savelieva E. et al. Validation of Geostatistical and Machine Learning Models for Spatial Decision-Oriented Mapping // Proceeding of StatGIS 99 / Ed. J. Piltz, J. Heyn. — Klagenfurt, 2006.

Parkin R., Kanevski M., Maignan M. et al. Multivariate Geostatistical Mapping of Contamination in Geneva Lake Sediments: Case Study with Multigeo. — Moscow: Nuclear Safety Inst. RAS, 2001. — (Препринт / ИБРАЭ; IBRAE-01-4).

Savelieva E., Bizikov V., Goncharov S. et al. Stochastic Simulations for Assessment of Uncertainty of Spatial Distribution and Biomass of Marine Living Resources // Proceedings of the Sixth European Conference on Ecological Modelling, Triest, Italy, 27—30 November 2007. — [S. l.], 2007.

Savelieva E., Demyanov V., Kanevski M. et al. BME Based Uncertainty Assessment of the Chernobyl Fallout // Geoderma. — 2005. — Vol. 128. — P. 312—324.

Savelieva E., Kanevski M., Demyanov V. et al. Conditional Stochastic Cosimulations of the Chernobyl Fallout // geoENV II — Geostatistics for Environmental Applications / Ed. J. Gomez-Hernandez, A. Soares, R. Froidevaux. — [S. l.]: Kluwer Academic Publishers, 1998. — P. 453—464.

Savelieva E., Kanevski M., Timonin V. et al. Uncertainty in the hydrogeologic structure modeling // Proceedings of IAMG2002 conference, September 2002, Berlin, Germany. — [S. l.], 2002. — P. 481—486.

S-GeMS The Stanford Geostatistical Modeling Software (S-GeMS) // http://sgems.sourceforge.net.

SIC’97 Spatial Interpolation Comparison Exercise 1997 // http://www.ai-geostats.org/index.php?id=45.

Глава 1

Основные задачи анализа пространственных данных

В этой главе мы начнем с постановки задачи при анализе и моделировании пространственных данных и приведем примеры типовых задач. В разделе 1.3 приведен обзор общих подходов к пространственному моделированию, кратко описана история создания и развития геостатистики. В разделе 1.4 представлена методология последовательного анализа и моделирования пространственных данных. В разделе 1.5 приведен список типовых вопросов и ответов по проблемам пространственных данных, которые будут подробно освещены в последующих главах книги.

1.1. Проблемы пространственного моделирования

В 1986 г. произошел выброс радиоактивных веществ из реактора на Чернобыльской АЭС. Радиоактивное загрязнение распространилось по воздуху на сотни километров и затронуло многие европейские страны [De Cort, Tsaturov, 1996]. Измерения радиоактивного загрязнения почвы проводились во многих местах. Встали вопросы: Как построить карту загрязнения? Можно ли обойтись простыми методами интерполяции? Можно ли дать однозначный ответ о том, где проходит граница повышенного уровня загрязнения? На эти и многие другие вопросы могут дать ответ анализ и моделирование пространственных данных с использованием статистических методов [Kanevski et al., 1996; Kanevski et al., 1997; Каневский и др., 1999б].

Существует огромное количество пространственно распределенной информации, собранной в базы и банки данных по окружающей среде. Задача ее интерпретации, анализа и дальнейшего использования представляется чрезвычайно важной и требует комплексного системного подхода. Статистическое моделирование пространственных явлений позволяет обобщить имеющиеся измерения и получить модель их распределения в пространстве.

Наиболее распространенной проблемой при работе с пространственно распределенными данными является получение пространственной оценки. Так, было подготовлено много различных карт по радиоактивному загрязнению почвы в результате Чернобыльской аварии [De Cort, Tsaturov, 1996]. При этом оставался открытым вопрос о качестве и точности этих карт, неопределенности оценки, чувствительности использованных методов интерполяции и т. п.

Пространственное моделирование применяется во многих сферах человеческой деятельности. Так, при климатическом моделировании анализируются измерения температуры, осадков, скорости ветра и т. д. в различных точках пространства. При моделировании загрязнения окружающей среды используются измерения (пробы грунта, воды, воздуха, дистанционное зондирование) в различных местах. В задачах геологии моделируются свойства пород в промежутке между скважинами, где делаются измерения. В медицинской географии анализируются факторы, влияющие на уровень заболеваний, и моделируется распространение эпидемий. Пространственно распределенные данные используются при моделировании запасов полезных ископаемых и рыбных ресурсов, криминогенной ситуации и природных катастроф (оползней, лавин и пр.).

Глубокий анализ и моделирование пространственных данных требуют применения комплексного подхода и различных методов, характеризующих ту или иную особенность явления. Сложность такого анализа обусловлена несколькими факторами: наличием больших объемов количественной и качественной информации по исследуемому явлению, многомасштабностью и многопеременностью, наличием различных факторов влияния.

Мы опишем элементы методологии геостатистического анализа пространственно распределенных данных и приведем примеры исследования с применением этих методов для реальных данных, связанных с загрязнением окружающей среды, климатическими условиями, геомоделированием свойств пород, гидрогеологией, моделированием рыбных ресурсов.

1.2. Постановка задачи

При работе с пространственными данными обычно имеется некоторое количество измерений изучаемой переменной в различных точках, число которых ограниченно. Итак, есть область, на которой проведен ряд измерений некоторой величины Z. Эти измерения проведены в произвольно распреде-

ленном по области наборе точек (x, y), которые мы будем называть сетью мониторинга (рис. 1.1). Но есть и участки области, не покрытые измерениями, о значениях величины Z в которых хотелось бы получить информацию. Наиболее часто требуется оценить значение наблюдаемой величины в непромеренной точке X на основе имеющихся данных, т. е. решить задачу интерполяции.

Данные измерений, как правило, дискретны и пространственно неоднородно распределены. Анализ данных и его результаты зависят от качества и количества исходных данных, от методов и моделей обработки данных.

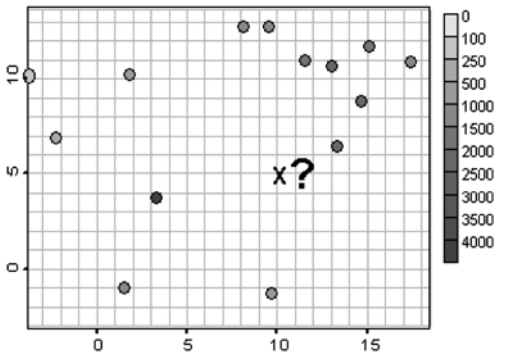


Рис. 1.1. Постановка задачи пространственного оценивания

Приведем здесь ряд конкретных задач, для решения которых необходимо применение комплекса исследований с помощью методов геостатистики — статистики пространственно распределенной (региональной) информации:

- оценить значение в точке, где измерение не проводилось;
- нарисовать карту, построить изолинии (определить значения на плотной сетке);
- оценить ошибку интерполяционной оценки;
- оценить значение переменной, по которой мало измерений, используя значения другой коррелированной с ней переменной, по которой проведено много измерений;
- определить вероятность того, что значения наблюдаемой переменной превысят заданный уровень в интересующей нас области;
- получить набор равновероятных стохастических пространственных реализаций распределения наблюдаемой переменной.

Первые три задачи — примеры задач регрессии или классификации (в зависимости от типа исходных значений). Две последние задачи относятся к вероятностному анализу и связаны с оценками риска. Отдельные главы данной книги будут посвящены решению этих задач.

1.3. Подходы к анализу пространственно распределенных данных

Существует несколько подходов к анализу и обработке пространственно распределенных данных, которые можно условно разделить на три группы:

- детерминистические модели (интерполяторы) — линейная интерполяция на основе триангуляции, метод обратных расстояний в степени, мульти-квадратичные уравнения и т. п. [Каневский и др., 1999б];
- геостатистика — модели, базирующиеся на статистической интерпретации данных [Journel, Huijbregts, 1978];
- алгоритмы, основанные на обучении — искусственные нейронные сети, генетические алгоритмы, статистическая теория обучения машин векторов поддержки (Support Vector Machines) [Vapnik, 1998].

Конечно, это деление до известной степени условно. Так, геостатистические модели можно изложить в детерминистической формулировке, и наоборот, некоторые детерминистические модели имеют близкие статистические аналоги. В свою очередь, статистический подход, на котором базируется геостатистика, включает регрессионные модели пространственных интерполяций (предсказаний) и методы стохастического моделирования, цели и задачи которых различны. Алгоритмы, основанные на обучении (или искусственный интеллект), также имеют статистическую интерпретацию.

Современная геостатистика — это широкий спектр статистических моделей и инструментов для анализа, обработки и представления пространственно распределенной информации [Cressie, 1991]. Ниже мы подробно опишем наиболее часто используемые модели и инструменты, из которых можно составить замкнутый цикл исследования и решить поставленные выше задачи.

Традиционные детерминистические методы, широко используемые для пространственной интерполяции, позволяют решать только первую и вторую задачи из приведенного выше списка. Геостатистическая теория позволяет решать весь набор задач, в том числе оценить неопределенность оценки и описать ее вариабельность.

Геостатистика возникла в начале 1960-х гг. как теория региональных переменных, сформулированная Ж. Матероном (Matheron) для анализа данных о природных ископаемых (горнорудное дело) [Matheron, 1963; Матерон, 1968]. Он организовал Центр геостатистики в Фонтенбло. Этот центр внес заметный вклад в теоретические исследования и их практические применения.

Независимо от Ж. Матерона и практически в то же время Л. С. Гандин сформулировал теорию оптимальной интерполяции для объективного анализа метеополей [Гандин, Каган, 1976]. В этой теории также приведены основы теории геостатистической. К сожалению, последующие работы российских ученых в этой области не нашли в то время широкой поддержки [Вистелиус, 1984, 1986].

Современная геостатистика — это быстро развивающаяся область прикладной статистики с огромным набором методов, линейных и нелинейных, параметрических и непараметрических моделей для анализа, обработки и представления пространственной информации. Спектр ее применения весьма широк — от традиционного использования в области добычи ископаемых до современных приложений в экономике, финансах, окружающей среде, эпидемиологии [Goovaerts, 1997; Wackernagel, 1995]. В Приложении 3 приведен краткий обзор книг по геостатистике на английском языке.

Геостатистический анализ позволяет значительно повысить уровень надежности и качество решений, принимаемых на основе использования пространственно распределенной информации. Современные тенденции геостатистики связаны с развитием методов стохастического моделирования (пространственных аналогов методов Монте-Карло), методов, основанных на многоточечной статистике, гибридных моделей с использованием алгоритмов искусственного интеллекта, с использованием дополнительной информации различного вида и приложениями в области обработки и передачи изображений, с расширением на временной и пространственно-временной анализы и многими направлениями [Kanevski et al., 2007]. Некоторые из продвинутых методов, разработанных в последние годы, описаны в Главе 11.

Одним из важных составляющих традиционной геостатистики является пространственный корреляционный анализ, или *вариография*. Несмотря на кажущуюся простоту исходных формул, вариография позволяет сделать глубокие выводы о статистической природе данных и структуре адекватных моделей. В принципе экспериментальная вариография, основанная

на исходных данных, может быть использована в большинстве задач пространственного оценивания независимо от метода интерполяции наравне с традиционным статистическим анализом.

1.4. Основные этапы анализа и моделирования пространственных данных

Первым и весьма важным этапом исследования является современный статистический анализ данных, позволяющий определить наличие ошибок и выбросов (outliers) в данных, оценить базовые статистические закономерности, провести корреляционный анализ при наличии нескольких переменных и т. п.

Если данные собраны на нерегулярной кластерной сети мониторинга, может потребоваться пространственная декластеризация для получения репрезентативной глобальной статистики — средних, вариаций, гистограмм. Если сеть мониторинга имеет зоны с заметно более высокой плотностью измерений, чем остальная область, то сеть мониторинга кластерная. Если при этом зоны повышенной плотности измерений характеризуются более высокими (или, наоборот, низкими) значениями измерений, возникает необходимость в декластеризации. Иначе оценки всех статистических характеристик будут искажены, например оценка среднего будет завышена (или, наоборот, занижена). Процедура декластеризации ориентирована на устранение такого рода искажений. Можно рассматривать два основных типа декластеризации — выборочную и весовую. Выборочная декластеризация связана с выбором части данных из кластеров, весовая предполагает задание весов, с которыми используются измерения. Подробнее кластерность и декластеризация рассмотрены в Главе 2.

Оценить некоторые пространственные особенности данных позволяет статистика с движущимся окном: область разбивается на подобласти, в каждой из которых проводится независимый статистический анализ.

Дальнейший пространственный анализ предполагает исследование и моделирование *пространственной корреляции* между данными по одной или нескольким переменным. Мерой пространственной корреляции является *вариограмма* — статистический момент второго порядка.

Для получения наилучшей в статистическом смысле пространственной оценки используются модели из семейства *кригинга* (kriging) — наилучшего линейного несмещенного оценивателя (best linear unbiased estimator — BLUE).

Кригинг является «наилучшим» оценивателем в статистическом смысле в классе линейных интерполяторов — его оценка обладает минимальной вариацией ошибки. Важное свойство кригинга — точное воспроизведение значений измерений в имеющихся точках (точный оцениватель). В отличие от многочисленных детерминистических методов, оценка кригинга сопровождается оценкой ошибки интерполяции в каждой точке. Полученная ошибка позволяет охарактеризовать неопределенность полученной оценки данных при помощи доверительных интервалов или «толстых» изолиний.

При применении любой модели интерполяции встает вопрос о подборе оптимальных модельно-зависимых параметров. Легко показать, что даже в случае использования одного и того же метода интерполяции можно получить качественно разные результаты в зависимости от выбора модельных параметров. Выбор оптимальных параметров опирается на пошаговое исследование характера и структуры данных. Эффективными инструментами подбора модельных параметров являются методы кросс-валидации (cross-validation), складного ножа (jack-knife), бутстреп (bootstrap) [Armstrong, 1997]. Все они основаны на проведении оценки для части точек измерений, выбранных из основного набора по остальным данным с последующим вычислением ошибки оценки. После оценок по всем точкам, наборам или выборкам оценивается среднеквадратичная ошибка полученных оценок. По ней сравниваются различные методы или выбираются наилучшие параметры метода. В геостатистике традиционно более широко используется кросс-валидация.

При проведении анализа реальных данных эксперты часто сталкиваются с проблемой малого количества измерений по интересующей переменной, например вследствие их дороговизны или небезопасности взятия проб. При этом в наличии может оказаться большое (избыточное) количество «дешевых» измерений переменной, которая достаточно сильно коррелированна с основной. Встает вопрос, как можно использовать «дешевую» информацию для улучшения оценки переменной, информация по которой «дорога». В рамках многопеременной геостатистики существует модель совместной пространственной интерполяции нескольких коррелированных переменных — *кокригинг*. Кокригинг позволяет значительно улучшить качество оценки, перейти из области экстраполяции в область интерполяции, уменьшить ошибку оценки за счет использования дополнительной «дешевой» информации по коррелированным переменным.

Часто результатом пространственного анализа данных в рамках квалифицированной поддержки принятия решений являются вероятностные карты. Вероятностное картирование дает возможность оценить уровень риска по превышению или непревышению заданного уровня значения пространственной переменной. Оно также используется при оптимизации решений, когда пространственный анализ данных является только промежуточным этапом. В рамках геостатистики для вероятностного картирования используются *нелинейные модели кригинга,* в частности индикаторный кригинг. Он позволяет рассчитать локальную функцию распределения в точке оценивания. В качестве результатов составляются карты вероятности, карты средних оценок, карты оценок с заданной вероятностью превышения, которые используются в процессе принятия решений.

Применение различных детерминистических или геостатистических моделей интерполяции/оценивания всегда дает единственное и сглаженное, не воспроизводящее изначальную вариабельность данных значение оценки в интересующей точке при выбранных модельных параметрах. *Стохастическое моделирование* является альтернативным подходом, дающим возможность воспроизвести исходную вариабельность и получить сколь угодно много равновероятных реализаций пространственной функции в области. Равновероятные реализации позволяют описать пространственную вариабельность (изменчивость) и неопределенность пространственной функции, оценить вероятности и риск. При использовании стохастического моделирования удается избежать «сглаженной» картины оценки, которая присуща большинству моделей интерполяции. Это позволяет получать корректные результаты в таких задачах, как, например, расчет объема нефтяного резервуара, «длины» береговой линии и т. п.

На основе описанных этапов анализа и моделирования пространственных данных можно сформулировать блок-схему пошагового анализа (рис. 1.2). В ее основе лежит методология, опробованная в различных исследованиях, в том числе и на данных радиоактивного Чернобыльского загрязнения [Каневский и др., 1999а, б]. На основе аналогичной блок-схемы был создан пакет программ «Геостат Офис», включающий набор моделей для пространственного анализа и картирования данных [Kanevski, Maignan, 2004]. Мы будем следовать этой методологии и подробно опишем спектр алгоритмов, которые можно применить на каждом этапе.

Обучаемые статистические модели, такие как искусственные нейронные сети и машины поддерживающих векторов (support vector machines), мож-

но использовать наряду с геостатистическими моделями для решения задач пространственной регрессии и классификации [Kanevski, Maignan, 2004; Advanced..., 2008]. Подробное описание этих моделей выходит за рамки настоящей книги. Однако некоторые примеры совместного использования геостатистики и ИНС разобраны в Главе 10.

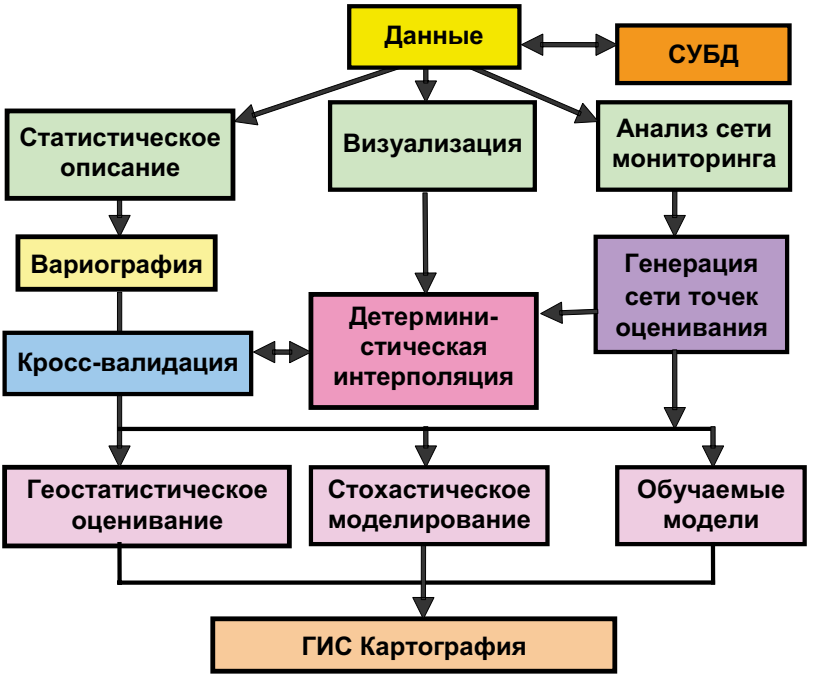


Рис. 1.2. Блок-схема методологии последовательного анализа и моделирования пространственно-распределенных данных

1.5. Вопросы, возникающие при пространственном моделировании

Таблица 1.1. Вопросы и методы решения

Вопрос	Метод решения
Какое разрешение имеет сеть мониторинга и какие явления она может обнаружить?	Анализ сети мониторинга проводится с привлечением фрактальных моделей, геометрических характеристик, статистических индексов и зависимостей (см. Главу 2)
Как описать количество и качество имеющейся информации и составить репрезентативное корректное статистическое описание данных?	Наряду со средствами традиционной статистики используется пространственная статистика движущегося окна и методы декластеризации (см. Главу 2)
Имеет ли смысл задача интерполяции?	При отсутствии пространственной корреляции между данными получение оценки в точке путем взвешивания соседних измерений не имеет смысла (см. Главу 4)
Как выявить и смоделировать пространственную непрерывность данных на различных масштабах?	Исследовать и моделировать пространственную корреляцию данных с учетом возможной нестационарности и анизотропии при помощи стандартных приемов вариографии, анализа трендов (см. Главу 4)
Как получить наилучшую в статистическом смысле оценку значения пространственной переменной в точке, где измерения отсутствуют? Как оценить ошибку полученной оценки? Как построить карты оценок и ошибок оценки?	Применить модель из семейства кригинга — наилучших несмещенных линейных оценивателей (см. Главу 5)
Как учесть при интерполяции ошибки измерений?	Геостатистическое оценивание позволяет учесть ошибку измерений и ее пространственное распределение при интерполяции (см. Главу 5)
Как подобрать оптимальные параметры модели интерполяции?	Методы кросс-валидации, складного ножа, бутстрепа позволяют эффективно подобрать оптимальные параметры и не зависят от выбранной модели интерполяции (см. Главу 2)
Как использовать избыточную «дешевую» информацию для улучшения оценки переменной, измерения которой «дороги»?	Провести совместный анализ и интерполяцию нескольких коррелированных переменных при помощи многомерных геостатистических моделей (кокригинг) (см. Главу 6)
Как получить оценку вероятности превышения заданного уровня значений (провести оценку риска)?	Метод вероятностного картирования — индикаторный кригинг (см. Главу 7)

Таблица 1.1 (окончание)

Вопрос	Метод решения
Как получить не единственную оценку функции в точке, построить равновероятные реализации пространственного распределения?	Стохастическое моделирование позволяет получить множество равновероятных реализаций функции и оценивать на их основе различные статистические характеристики, описывать пространственную вариабельность и неопределенность данных (см. Главу 8)
Как избежать «сглаженной» оценки и воспроизвести изначальную вариабельность данных?	Стохастическое моделирование дает несглаженную картину и воспроизводит исходные данные наряду с параметрами распределения (статистические моменты первого и второго порядков), позволяют описать неопределенность и пространственную вариабельность данных (см. Главу 8)
Как оптимизировать сеть мониторинга?	Эта задача решается путем геостатистического анализа существующей сети и оптимизации функции стоимости для получения наименьшей ошибки оценки с учетом затрат на дополнительные измерения
Какие модели можно использовать, если в данных измерений присутствуют крупномасштабный тренд, периодичность, пятнистость?	Одним из эффективных подходов представляется применение искусственных нейронных сетей (ИНС). В процессе обучения ИНС адаптируются к исходным данным и хорошо моделируют крупномасштабные нелинейные эффекты. Смешанные модели ИНС в сочетании с геостатистикой продемонстрировали высокую эффективность по сравнению с другими методами на различных данных, имеющих сложный характер (см. Главу 10)
Пространственно-временной прогноз — как одновременно смоделировать данные по пространству и времени?	Геостатистические модели оценивания могут применяться и в пространственно-временно́м континууме с использованием пространственной и временно́й компонент модели пространственной корреляции (см. Главу 11)
Как учесть дополнительную априорную информацию о наблюдаемой переменной и/или о подобных явлениях?	Применить байесовские модели или модели интеграции данных (см. Главу 11)

Перечисленные проблемы успешно решались авторами в процессе анализа данных по радиоактивному загрязнению почвы, данных по химическому загрязнению донных отложений Женевского озера, распределению популяции рыбы в море, климатических данных (температуры, осадков), данных по моделированию гидрогеологической структуры, данных по электропотреблению и др. Перечисленные данные используются в книге для иллюстрации использования геостатистических методов.

Литература

Вистелиус А. Б. Математическая геология: история, состояние, перспективы. — Л., 1984. — 53 с. — (Препринт / ЛОМИ; Р-10-84).

Вистелиус А. Б. Математическая геология и ее вклад в фундаментальные геологические разработки. — Л., 1986. — 27 с. — (Препринт / ЛОМИ; Р-5-86).

Гандин Л. С., Каган Р. Л. Статистические методы интерполяции метеорологических данных. — Л.: Гидрометеоиздат, 1976. — 359 с.

Каневский М., Демьянов В., Савельева Е. и др. Элементарное введение в геостатистику. — М., 1999а. — 136 с. — (Проблемы окружающей среды и природных ресурсов / ВИНИТИ; № 11).

Каневский М., Демьянов В., Чернов С. и др. Геостатистика и искусственные нейронные сети для анализа и моделирования пространственно распределенных данных // Изв. РАН. Энергетика. — 1999б. — № 1.

Матерон Ж. Основы прикладной геостатистики. — М.: Мир, 1968. — 407 с.

Advanced Mapping of Environmental Data: Geostatistics, Machine Learning and Bayesian Maximum Entropy / Ed. M. Kanevski; ISTE Ltd. — [S. l.], 2008. — 313 p.

Armstrong M. Basic Linear Geostatistics. — [S. l.]: Springer Verl., 1997.

Cressie N. Statistics for spatial data. — New York: John Wiley & Sons, 1991. — 900 p.

De Cort M., Tsaturov Yu. S. Atlas on caesium contamination of Europe after the Chernobyl nuclear plant accident / European Commission. — [S. l.], 1996. — 39 p. — (Report EUR 16542 EN).

Goovaerts P. Geostatistics for Natural Resources Evaluation. — [S. l.]: Oxford Univ. Press, 1997.

Isaaks E. H., Srivastava R. M. An Introduction to Applied Geostatistics. — Oxford: Oxford Univ. Press, 1989.

Journel A. G., Huijbregts Ch. J. Mining Geostatistics. — London: Academic Press, 1978. — 600 p.

Kanevsky M., Arutyunyan R., Bolshov L. et al. Geostatistical Portrayal of the Chernobyl Fallout // Geostatistics Wollongong ’96 / Ed. E. Y. Baafi, N. A. Schofield. — [S. l.]: Kluwer Academic Publ., 1996. — Vol. 2. — P. 1043—1054.

Kanevsky M., Arutyunyan R., Bolshov L. et al. Chernobyl Fallouts: Review of Advanced Spatial Data Analysis // geoENV I — Geostatistics for Environmental Applications / Ed. A. Soares, J. Gomez-Hernandes, R. Froidvaux. — [S. l.]: Kluwer Academic Publ., 1997. — P. 389—400.

Kanevski M., Maignan M. Analysis and modelling of spatial environmental data. — Lausanne: EPFL Press, 2004. — 288 p. — (With a CD and educational/research MS Windows software tools) (http://www.ppur.org/auteurs/1000772.html).

Matheron G. Principles of Geostatistics // Economic Geology. — 1963. — Vol. 58. — P. 1246—1266.

Vapnik V. N. Statistical Learning Theory. — New York: John Wiley & Sons, Inc., 1998. — 736 p.

Wackernagel H. Multivariate Geostatistics. — Berlin: Springler-Verl., 1995.

Глава 2
Основные понятия и элементы геостатистики

Эта глава посвящена базовым понятиям и предположениям геостатистики, а также смежных областей. В разделе 2.1 даны определения пространственно распределенных данных, с которыми работает геостатистика. В Разделах 2.2—2.5 сделан экскурс в смежные области, связанные с пространственными данными: метрику пространства, пространственное разрешение, описание сети мониторинга, декластеризацию. Раздел 2.6 посвящен одному из важнейших понятий геостатистики — пространственной непрерывности. Различные виды стационарности и связанные с ними предположения описаны в Разделе 2.7. В Разделе 2.8 речь идет об основной модели геостатистического оценивания — кригинге. Раздел 2.9 посвящен кроссвалидации и другим методам проверки качества моделей.

2.1. Пространственно распределенные данные

При анализе данных различных измерений часто крайне трудоемко или вовсе невозможно получить формульный закон распределения данных на основе физических процессов, обуславливающих соответствующие явления. Альтернативный подход — статистическое (а не детерминистическое) описание пространственного распределения. В отличие от детерминистических методов геостатистические оценки опираются на информацию о внутренней структуре данных, зависят от самих данных, т. е. являются адаптивными. Геостатистика базируется на статистической интерпретации данных. Предполагается, что данные измерений $z(\boldsymbol{x}_i)$ являются реализациями случайных переменных $Z(\boldsymbol{x}_i)$, которые описываются некоторыми функциями распределения. Это, однако, не означает, что природа самого процесса является случайной. Чтобы использовать геостатистику, необходимо определить пространственную корреляционную структуру поля $Z(\boldsymbol{x})$, задаваемую всеми случайными переменными в области исследования. Геостатистический подход позволяет исходить при анализе из строгих критериев.

Предметом анализа геостатистики являются *пространственные переменные* (или регионализованные переменные — regionalised variables), что аналогично переменным с координатной привязкой. Примеры пространственных переменных: количество осадков, плотность населения в некоторой географической области, мощность геологической формации, плотность загрязнения почвы, среднее потребление электроэнергии в определенный час и т. п. Пространственные переменные не следует путать со случайными величинами, изучаемыми методами обычной статистики.

Случайная функция определяется как набор обычно зависимых между собой случайных переменных $Z(\boldsymbol{x}_i)$, по одной для каждого местоположения $\boldsymbol{x}_i$ в рассматриваемой области. Любому набору из N местоположений $\{\boldsymbol{x}_k, k = 1, ..., N\}$ можно поставить в соответствие N случайных переменных $\{Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_N)\}$, которые характеризуются N-мерной условной функцией распределения:

$$F(\boldsymbol{x}_1, ..., \boldsymbol{x}_N; z_1, ..., z_N) = \text{Prob}\{Z(\boldsymbol{x}_1) \le z_1, ..., Z(\boldsymbol{x}_N) \le z_N\}. \quad (2.1)$$

Понятие *случайной величины* в классической статистике имеет конкретный смысл только при соблюдении следующих условий:

1) должна быть хотя бы теоретическая возможность бесконечного повторения испытаний (реализаций), в результате которых случайная величина приобретает численные значения;
2) результат каждого из испытаний должен быть независим от результатов всех предыдущих испытаний.

Пространственная переменная не удовлетворяет ни одному из этих условий. Если, например, испытание состоит в отборе пробы в точке $\boldsymbol{x}$, то содержание искомого вещества в такой пробе будет единственным, физически определенным и ни в коей мере не случайным. Нет никакой возможности повторить такое испытание, поскольку проба в конкретной точке уже взята, что влечет невыполнение условия 1. Однако есть возможность отобрать новую пробу в непосредственной близости от точки $\boldsymbol{x}$, что можно в приближении принять за выполнение условия 1. Но тогда нарушается условие 2: если первая проба отобрана в обогащенной зоне, то вторая проба, взятая в непосредственной близости от первой, как правило, будет иметь высокое содержание. Таким образом, испытания оказываются зависимыми.

В дальнейшем мы будем использовать для удобства привычный в статистике термин случайной величины, понимая под ней пространственную регионализованную переменную и учитывая вышеописанные особенности.

Наблюдаемая переменная может быть *непрерывной* (например, любая физическая величина — плотность, давление, концентрация и т. п.) или кате-

гориальной (например, временной бинарный сигнал или тип почвы либо геологической породы). Для анализа переменных разного типа естественно использовать различные подходы.

2.2. Метрика в пространстве

Мы будем рассматривать так называемые регионализованные данные, а именно измерения, обладающие координатной привязкой. Координатная привязка может быть:

- пространственной, определяющей географическое положение измерения (географические координаты) в пространстве или его относительное положение по отношению к другим объектам (специальная координатная система для определенной местности);
- временно́й, определяющей время проведения измерения (абсолютное или относительное);
- пространственно-временно́й, т. е. и пространственной, и временно́й одновременно.

Основное требование к координатной системе — ее метричность, т. е. координаты должны сопровождаться метрикой, возможностью вычислять расстояния между точками. В большей части книги, если иное не оговорено, для простоты будем предполагать, что мы работаем в двумерном евклидовом пространстве R^2, где метрика такова, что расстояние между точками пространства $X_1 = (x_1, y_1)$ и $X_2 = (x_2, y_2)$ определяется евклидовой нормой:

$$\|(x_1, y_1)(x_2, y_2)\| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \qquad (2.2)$$

Введение третьей пространственной координаты идеологически не добавляет ничего, кроме усложнения выкладок, связанных с введением дополнительных направлений в пространстве и различием масштабов вертикальной координаты по сравнению с горизонтальными. Евклидово расстояние между точками $X_1 = (x_{11}, ..., x_{1n})$ и $X_2 = (x_{21}, ..., x_{2n})$ в n-мерном пространстве вычисляется аналогично двумерному случаю:

$$\|(\bar{X}_1)(\bar{X}_2)\| = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}. \qquad (2.3)$$

Введение дополнительной временно́й координаты и проблемы построения пространственно-временно́го континуума будут рассмотрены в Главе 10, посвященной развитию пространственно-временно́й геостатистической теории.

Различие масштабов связано с измерениями: например, если рассматривать распространение загрязнения в почве, то горизонтальные пространственные области распространяются на десятки километров (10^4—10^5 м), в то время как рассматриваемая глубина при анализе почвы не превышает 0,3 м, а геологические слои могут иметь толщину от нескольких сантиметров до сотен метров. При работе с такими различными масштабами обычно производится нормировка координат — переход к другой системе, где размерности соизмеримы, например линейное преобразование на отрезок (0, 1).

Пространственная переменная всегда определена в конкретной области пространства — в *геометрическом поле*. Пространственную переменную V можно рассматривать как функцию точки пространства $\boldsymbol{x}$: $Z = Z(\boldsymbol{x})$. Однако чаще интерес представляют не точечные, а средние значения величины $Z(\boldsymbol{x})$ в пределах малой области пространства — *геометрической базы* (support). Например, для такого признака, как содержание чего-либо в грунте, геометрической базой является объем пробы. База должна быть определена весьма точно. Необходимо знать ее объем, форму и ориентацию в пространстве. Если изменяется геометрическая база, то возникает новая пространственная переменная, близкая к предыдущей, но не совпадающая с ней:

$$z^*(x_0) = \frac{1}{S} \int_{S(x_0)} Z(x) dx. \qquad (2.4)$$

Теория пространственных переменных, которая называется *геостатистикой*, позволяет предсказывать характеристики *переменной Z^*, связанной с геометрической базой* в поле S, по известным характеристикам другой *точечной переменной V*, заданной в поле X, отличном от поля S. Эта возможность составляет одно из важнейших преимуществ названной теории.

2.3. Пространственное разрешение

Одним из ключевых свойств пространственно распределенных данных является их пространственное разрешение. При исследовании того или иного пространственного явления очень важно, чтобы имеющиеся данные могли адекватно отразить его. Обычно под пространственным разрешением понимается наименьший размер особенности, которую могут отражать данные и пространственные оценки.

Разрешение интерполяционной пространственной оценки на регулярной сетке характеризуется размером ячейки. Если сетка оценивания нерегу-

лярная, то ее разрешение можно охарактеризовать распределением расстояний между узлами сетки (см. ниже).

Эффект разрешения сетки оценивания может быть значительным, особенно при решении динамических задач с граничными условиями на сетке. В статических задачах пространственного картирования разрешение сетки также имеет большое значение. Существуют характеристики, связывающие разрешение сетки с картографическим масштабом [Hengl, 2006], более подробное описание которых выходит за рамки настоящей работы.

Опора (support) *данных измерений* является одним из основных свойств при анализе пространственно распределенных данных. Опору не следует путать с пространственным разрешением модели (карты интерполяционной оценки). Опора характеризуется процессом измерения и обработки данных, а не моделирования. Под опорой измерения понимается физический объем, подвергнутый измерению. Например, при измерении радиоактивности образца опора измерения характеризуется размером пробы. Однако не всегда удается однозначно оценить опорный размер: так, при аэрогаммасъемке загрязненных территорий опорный размер может варьироваться от десятков до сотен метров.

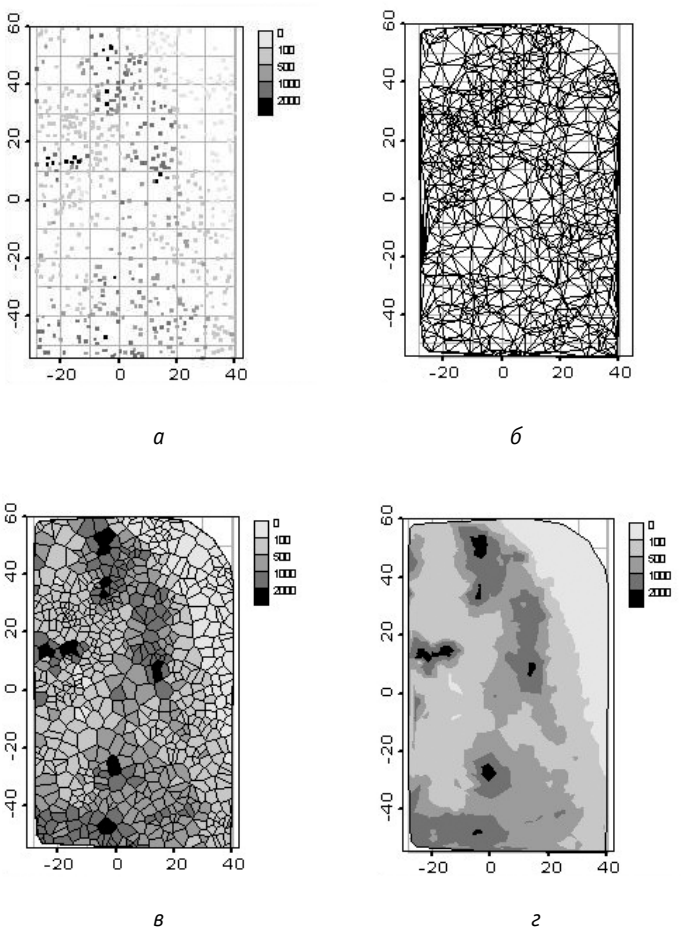
Определение опорного размера данных измерений, использующихся в моделях пространственного оценивания, чрезвычайно важно для адекватного моделирования вариабельности данных. Так, если при интерполировании на сетку с разрешением 1 км используются данные с опорой 10 см, надо понимать, что такие данные обладают вариабельностью на подсеточном масштабе. Другими словами, величина наблюдаемой переменной в ячейке сетки оценивания не может быть однозначно определена на основе данных с опорой более мелкого масштаба.

При моделировании свойств пористости и проницаемости пород в подземных месторождениях размером несколько километров используются данные с различной опорой. Так, пористость и проницаемость, измеренные на основе кернов из скважин, имеют высокую точность и опору порядка нескольких сантиметров. Данные же сейсмического зондирования обладают зашумленностью, и размер их опоры не всегда удается однозначно определить (от единиц до сотен метров). Динамические измерения давления в скважине имеют опору порядка нескольких километров, поскольку отражают поведение сред в связанной пористой системе месторождения. Все это необходимо учитывать при моделировании неопределенности и вариабельности пространственных распределений на основе данных различных типов.

В геостатистике можно учесть изменения размера опоры при блочном кригинге (см. Главу 5).

2.4. Сеть мониторинга и кластерность

Простейшим общепринятым видом визуализации данных является нанесение точек на плоскость пространственных координат, причем цвет нанесенной точки может соответствовать измеренной в них величине (рис. 2.1*а*).



а *б*

в *г*

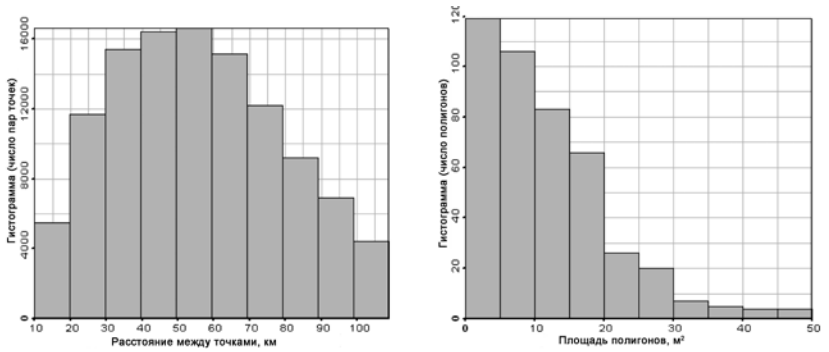
Рис. 2.1. Диаграмма расположения точек измерений (*а*), триангуляция сети мониторинга (*б*), полигоны Вороного (*с*) и контуры данных измерений по триангуляции (*г*)

Для визуализации сети мониторинга и ее кластерной структуры часто используется *триангуляция Делоне* [Preparata, Shamos, 1985] — система треугольников с вершинами в точках измерений, непересекающимися ребрами и минимальным количеством тупоугольных треугольников (рис. 2.1*б*). Такая визуализация позволяет качественно обособить области с повышенной плотностью измерений — с кластерами. Кроме того, триангуляция Делоне строит *систему соседства*: точки, которые соединены друг с другом ребрами треугольников, являются ближайшими соседями по отношению друг к другу.

Триангуляция также является основой для построения простейшего метода линейной интерполяции: три точки в пространстве (вершины треугольников) однозначно определяют плоскость, в пределах которой значения функции вычисляются согласно геометрическим принципам (рис. 2.1*г*).

Другим видом визуализации данных являются *полигоны Вороного*, или, как их еще называют, *разбиение Тиссена, ячейки Дирихле* и *области влияния*. Полигон Вороного P_i, построенный для точки измерений $\boldsymbol{x}_i$, характеризуется тем, что содержит те и только те точки, расстояние от которых до точки $\boldsymbol{x}_i$ меньше или равно расстоянию до любой другой точки измерений $\boldsymbol{x}_j$ (рис. 2.1*в*). При построении полигонов Вороного используется система соседства, полученная в процессе триангуляции Делоне. Границы полигона Вороного P_i состоят из отрезков серединных перпендикуляров, проведенных к сторонам треугольников Делоне. Полигоны Вороного можно использовать как разрывную интерполяционную оценку (оценка по ближайшему соседу). Для этого каждой точке, попавшей в полигон, присваивается значение, соответствующее его материнской точке. Эти полигоны также используются в задачах пространственной классификации — классификация по ближайшему соседу.

Для выявления особенностей, а именно наличия кластерных структур или разреженностей в сети мониторинга (наборе точек измерений), проводят анализ сети мониторинга. Простейшими методами такого анализа можно считать описание топологии сети с помощью гистограммы расстояний между точками (рис. 2.2*а*) и гистограммы площадей полигонов Вороного (рис. 2.2*б*). Гистограмма в данном случае — это график числа каких-либо событий (числа пар или числа полигонов), попавших в какой-либо интервал значений.



а *б*

Рис. 2.2. Гистограмма расстояний между точками (*а*) и гистограмма площадей полигонов Вороного (*б*)

При равномерном распределении точек в пространстве число пар должно быть одинаково для всех расстояний (или уменьшаться при увеличении расстояния за счет граничного эффекта). Рост числа пар с ростом расстояния между точками свидетельствует о наличии кластеров. Гистограмма площадей полигонов для регулярной сетки должна представлять собой дельта-функцию (один пик), так как все полигоны одного размера. Любые искажения (широкий пик, длинный хвост, несколько пиков) означают присутствие каких-либо особенностей в сети.

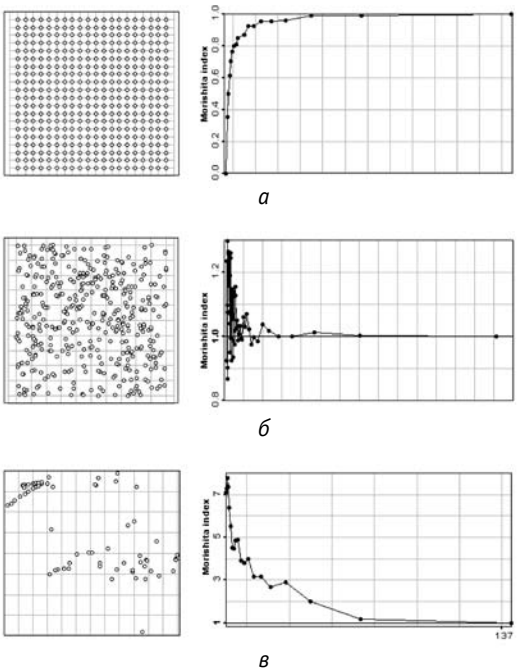
Другим методом анализа сети мониторинга является статистический подход [Cressie, 1991], рассматривающий точки измерений как случайный точечный процесс. Характеризовать распределение точек можно с использованием статистических индексов. Примером такого подхода является *диаграмма Моришита*. Индекс Моришита вычисляется для области, разбитой на прямоугольные ячейки равного размера, по формуле [Morishita, 1959]

$$I_\delta = Q \frac{\sum_{i=1}^{Q} n_i (n_i - 1)}{N(N-1)}, \qquad (2.5)$$

где N — полное число точек сети мониторинга; Q — число ячеек разбиения; n_i (i = 1, 2, ..., Q) — число точек сети мониторинга, попавших в i-ю ячейку. Этот индекс характеризует вероятность того, что при выборе двух случайных точек они окажутся в одной ячейке. Диаграмма Моришита представляет собой зависимость индекса Моришита от размера ячейки разбиения. Существуют три типа характерного поведения диаграммы Моришита, комбинации которых позволяют судить о характеристиках сети мониторинга:

- величина индекса Моришита с ростом размера ячейки растет и стремится к 1; тогда распределение точек можно считать равномерным;
- величина индекса Моришита не зависит от размера ячейки и примерно равна ≈ 1 (колеблется около 1); это означает, что распределение точек случайно и не имеет кластерных структур.
- величина индекса Моришита с ростом размера ячейки уменьшается или растет выше 1 — распределение точек сети кластерное.

На рис. 2.3 приведены примеры диаграмм Моришита для различных типов сетей мониторинга. Так, в случае мониторинга на регулярной равномерной сетке диаграмма имеет вид гладкой кривой логарифмического типа, стремящейся к единице (рис. 2.3*а*). При наличии многочисленных кластеров в плотной сети мониторинга кривая Моришита изобилует точками перегиба, которые характеризуют размеры различных кластеров (рис. 2.3*б*). В случае произвольного мониторинга с несколькими четко выраженными кластерами кривая Моришита имеет более гладкий вид и уменьшается, стремясь к единице (рис. 2.3*в*). Размер кластеров характеризуют в этом случае точки изменения кривизны.



а

б

в

Рис. 2.3. Примеры диаграммы Моришита для различных сетей мониторинга: регулярная равномерная сеть (*а*), произвольная сеть со слабой кластерной структурой (*б*), произвольная слабо связанная кластеризованная сеть (*в*)

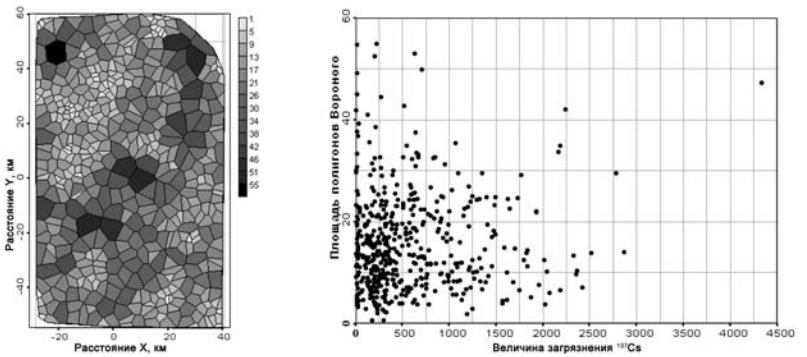
Для анализа сети мониторинга на неоднородность можно также использовать теорию фракталов и фрактальную размерность [Mandelbrot, 1982] (характеристику степени самоподобия объекта). Фрактальная размерность характеризует размерностное (dimensional resolution) разрешение сети мониторинга. Методы вычисления и использования фрактальной размерности подробно рассмотрены в [Raes et al., 1991].

2.5. Декластеризация

Бóльшая часть пространственно распределенных данных, которые анализируются в геостатистике, имеет кластерную структуру. Кластер образуется, если в одной области было проведено значительно большее число измерений, чем в другой. В этом случае могут возникнуть существенные искажения при вычислении, например среднего значения. Это влечет невозможность получить репрезентативную гистограмму распределения. Пусть, например в области высоких значений измеряемой величины, находится в двое больше точек, чем в области низких значений. Если при оценке среднего и других статистических параметров все значения будут иметь одинаковый вес, то область высоких значений будет слишком сильно влиять на такую оценку. В этом случае точки из зоны с бóльшими значениями нужно было бы учитывать с весом, в двое меньшим, чем все остальные. Проблема вычисления статистического веса каждой точки в параметрах распределения решается путем проведения процедуры *декластеризации* (declustering) данных.

Декластеризация не требуется, если измерения были выполнены на регулярной сетке. В этом случае наилучшее описание распределения получится при работе с равными весами. Тем не менее во многих случаях невозможно или нежелательно получить данные на равномерной сетке.

При анализе измерений, проведенных на нерегулярной сетке, предполагается существование такого набора весов, при котором может быть получено репрезентативное распределение данных. Здравый смысл подсказывает, что данные из области с большей плотностью измерений нужно брать с меньшим весом (для уменьшения их влияния на распределение в целом), чем данные из области с меньшей плотностью измерений. Для вычисления весов могут быть использованы разные подходы: метод ячейковой декластеризации, метод ячеек Дирихле (полигонов Вороного, рис. 2.4), кригинг.



а *б*

Рис. 2.4. Площади полигонов Вороного (*а*), корреляция площади полигона и величины пространственной переменной ^{137}Cs (*б*)

Метод ячейковой декластеризации (cell-declustering) был предложен в [Journel, 1983]. Его идея заключается в разбиении рассматриваемой области на подобласти кластеризованных данных и в определении равных весов для всех точек внутри каждой подобласти в соответствии с их количеством.

Так, если в ячейку a_k попало n_k точек, то каждое измерение будет учтено с весом $1/n_k$. Область a_k пространства обычно имеет размерность 3 (время может стать четвертым измерением). Для ячейки, не содержащей опытных точек, веса не рассчитываются, т. е. *область декластеризованных данных* состоит из ячеек, содержащих по крайней мере по одному измерению. Это ограничивает влияние граничных данных весом 1,0. На рис. 2.5 показан пример разбиения области на ячейки. Расчет соответствующих весовых коэффициентов приведен в табл. 2.1. После вычисления весов в такой форме они должны быть отнормированы так, чтобы их сумма была равна 1.

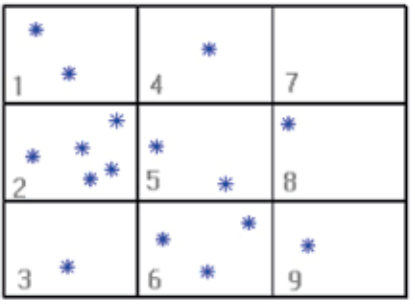


Рис. 2.5. Пример расчета весов ячейковой декластеризации

Таблица 2.1. Расчет весов ячейковой декластеризации к рис. 2.5

N ячейки	1	2	3	4	5	6	7	8	9
Количество точек	2	5	1	1	2	3	0	1	1
Вес	1/2	1/5	1	1	1/2	1/3	0	1	1

Для вычисления весов декластеризации нужно знать два параметра: размер ячейки (в каждом направлении) и начальную точку сетки (левый нижний угол).

Возможны два предельных случая. Если размер ячейки слишком мал, то каждая ячейка будет содержать не более одной точки, что приведет к присвоению всем точкам равных весов, и возникнет исходная ситуация недекластеризованных данных. В противоположном случае, когда размер ячейки слишком велик, все данные попадут в одну единственную ячейку и результат будет тот же — все точки получат равные веса.

Метод выбора размера ячейки зависит от типа кластеризации. Если данные кластеризованы случайным образом (есть области скопления точек, никак не связанных с их значениями), размер ячейки выбирается так, чтобы в областях с низкой плотностью измерений на одну ячейку приходилось приблизительно по одной точке измерений. Если же известно, что есть области высоких или низких значений с большим количеством измерений, то размер ячейки может быть выбран так, чтобы оптимально получить максимальное или минимальное взвешенное среднее. При декластеризации областей высоких или низких значений нужно пробовать наборы ячеек разного размера. В этом случае строится график зависимости взвешенного среднего значения от размера ячейки и в соответствии с ним выбирается подходящий размер [Deutsch, 1989].

Ячейки не обязательно должны быть квадратными. С помощью параметра анизотропии (отношение размеров ячейки) можно построить описанные выше зависимости и на их основе также выбрать параметры ячейки, соответствующие минимуму или максимуму взвешенного среднего. Результаты можно представить, например, в виде контурной карты с размерами ячеек в каждом из направлений в качестве координат.

Если при фиксированном размере ячейки перемещать начало декластеризующей сетки, то веса декластеризации могут существенно меняться. Чтобы исключить влияние этого фактора, проводят несколько шагов декластеризации, вводя систематическое смещение начала сетки. Веса, полученные

после каждого шага смещения, нормируются на единицу, и результаты суммируются. Обычно бывает достаточно пяти смещений. По окончании манипуляций веса всех точек снова должны быть отнормированы так, чтобы их сумма была равна 1.

Таким образом, формулу для вычисления декластеризованного среднего можно записать следующим образом:

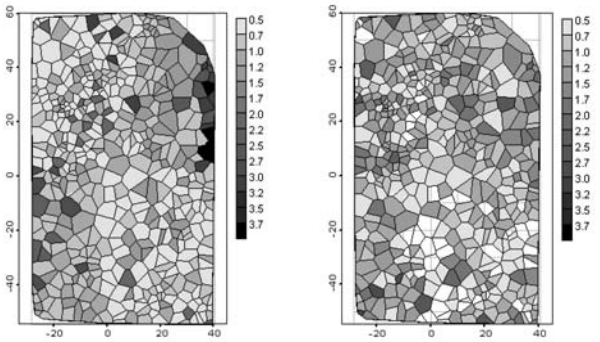
$$m = \frac{1}{nN_{\text{of}}} \sum_{i=1}^{N_{\text{of}}} \sum_{k=1}^{n} w_{ik} Z(\boldsymbol{x}_k), \tag{2.6}$$

где n — общее число исходных данных; N_{of} — число использующихся при вычислении смещений; w_{ik} — веса декластеризации для k-й ячейки при i-м смещении начала ячеек. Но в алгоритме декластеризации, реализованном в популярном пакете геостатистических программ GSLib [Deutsch, Journel, 1998], используется нормализация весов не к 1, а к числу измерений. При этом формула для вычисления декластеризованного среднего (2.6) несколько изменяется:

$$m = \frac{1}{n^2 N_{\text{of}}} \sum_{i=1}^{N_{\text{of}}} \sum_{k=1}^{n} w_{ik}^{*} Z(\boldsymbol{x}_k), \tag{2.7}$$

где w_{ik}^{*} — веса декластеризации, связанные с весами из (2.6) соотношением $w_{ik}^{*} = n w_{ik}$.

На рис. 2.6 приведены значения весов ячейковой декластеризации по формуле (2.7) для данных по радиоактивному загрязнению изотопом ^{137}Cs почвы. Можно сравнить эти значения с исходными данными, приведенными на рис. 2.8. На рис. 2.7 для тех же данных приведен график зависимости декластеризованного среднего от размера декластеризующей ячейки. Чтобы компенсировать влияние кластеров высоких значений, следует, видимо, выбрать ячейку размером 75 км.



а *б*

Рис. 2.6. Веса ячейковой декластеризации для декластеризации кластеров низких значений (*а*) и кластеров высоких значений (*б*)

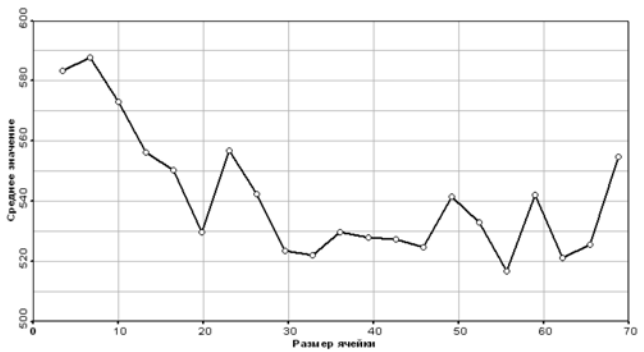


Рис. 2.7. Зависимость декластеризованного среднего значения от размера ячейки, метод ячейковой декластеризации

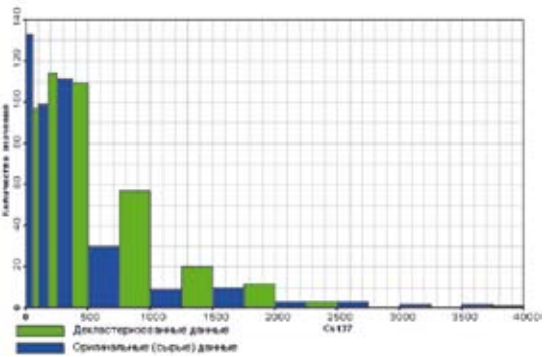


Рис. 2.8. Гистограммы декластеризованных и исходных данных ^{137}Cs

Упражнение 2.1. *При расчете декластеризованного среднего значения значение каждого данного учитывается с определенным весом. Почему при ячейковой декластеризации можно получить различные наборы весов для кластеров низких и высоких значений? Как при этом будут различаться средние значения?*

2.6. Пространственная непрерывность

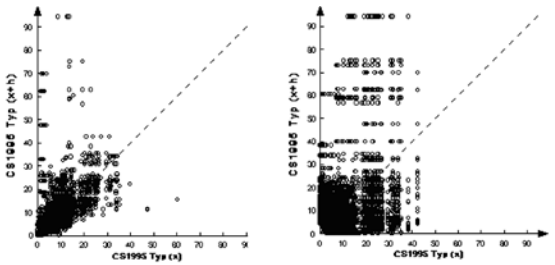
Пространственная непрерывность присутствует в большинстве геофизических явлений и выражает простое свойство исследуемой функции $Z(\boldsymbol{x})$: в двух точках, находящихся ближе друг к другу, *скорее будут* близкие значения, чем в более удаленных друг от друга точках. Подчеркнем вероятностный, статистический характер этого понятия.

Пространственную непрерывность в данных можно наглядно продемонстрировать, если построить зависимость значений, удаленных друг от друга, от расстояния между ними. Такая диаграмма называется диаграммой взаимного разброса пар точек (h-scatterplot), разделенных расстоянием h (рис. 2.9). Диаграмма взаимного разброса пар позволяет увидеть пространственную непрерывность и проверить наличие корреляции в данных как качественно, так и количественно.

На плоскости отмечают все возможные пары измерений, разделенные вектором $\boldsymbol{h}$. Если значения в паре, разделенной вектором $h = \boldsymbol{x}_i - \boldsymbol{x}_j$, обозначить $Z(\boldsymbol{x})$ и $Z(\boldsymbol{x} + \boldsymbol{h})$, то по оси абсцисс откладывается значение переменной $Z(\boldsymbol{x})$, а по оси ординат — $Z(\boldsymbol{x} + \boldsymbol{h})$. Диаграмма характеризует коррелированность значений в точках, разделенных данным расстоянием, и в определенном направлении. Если значения в точках, разделенных вектором (либо расстоянием) $\boldsymbol{h}$, близки, то точки диаграммы сгруппируются вдоль прямой $y = x$. При большей разнице между значениями в парах облако на диаграмме будет расплываться. Это обычно происходит при увеличении расстояния h. Часто на итоговую статистику диаграммы влияют отдельные отклонения. Такие пары точек лежат в отдалении от прямой $y = x$. В этом случае стоит попробовать посчитать статистику, исключив эти точки из рассмотрения.

На рис. 2.9 изображены диаграммы разброса пар для данных по загрязнению почвы в западной части Брянской области изотопом ^{137}Cs для расстояний 10 (слева) и 70 км (справа). На расстоянии 10 км пространственная корреляция существенна: точки на диаграмме пар сгруппированы вдоль

прямой $y = x$. На расстоянии 70 км пространственная корреляция уже очень слаба — диаграмма принимает форму прямоугольника.



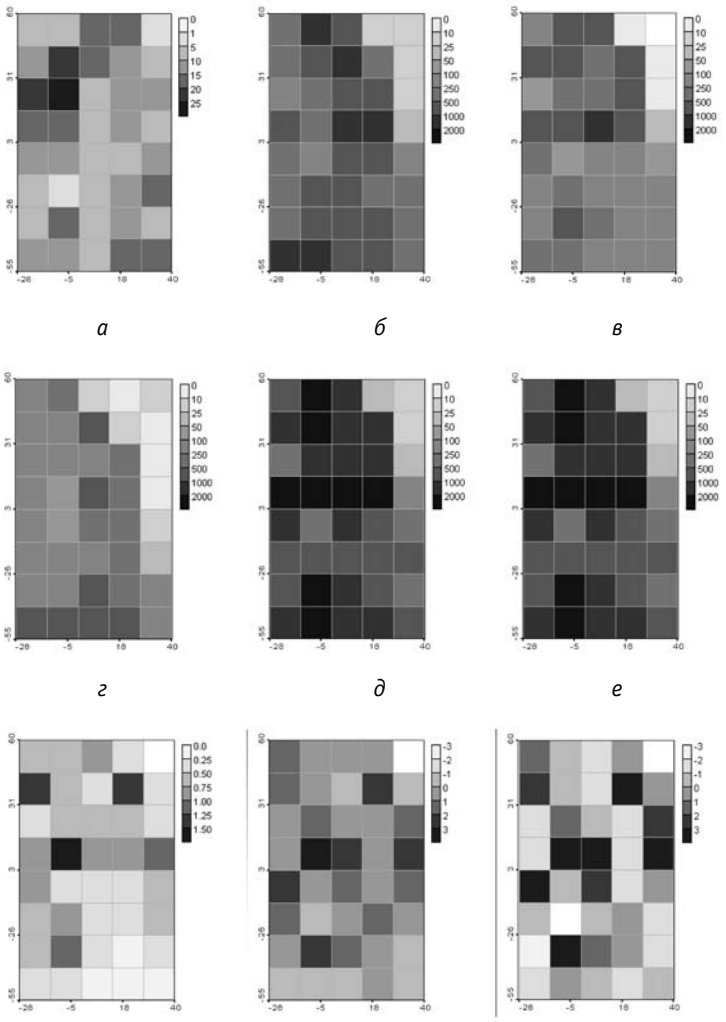
а *б*

Рис. 2.9. Диаграммы разброса пар точек h = 10 км демонстрирует корреляцию между данными (*а*); на расстоянии h = 70 км между точками отсутствует корреляция (*б*) для данных по загрязнению западной части Брянской области изотопом ^{137}Cs

Пространственная непрерывность может быть исследована простым методом вычисления локальных статистических характеристик: среднего, вариации и т. п.

Статистика движущегося окна (moving window statistics) — это подсчет описанной выше статистики, но не для всей области данных в целом, а в ее подобластях (окнах). Такой метод очень полезен для поиска зон аномальных средних значений и при наличии зон различной вариации значений (heteroscedasticity) [Isaaks, Srivastava, 1989]. Метод состоит в разбиении области данных на несколько одинаковых, обычно прямоугольных окрестностей — окон. Размер окна зависит от среднего расстояния между точками. Хорошим компромиссом между большими и маленькими окнами являются перекрывающиеся окна. При этом два соседних окна имеют несколько общих точек. Это повышает количество окон при достаточно большом их размере, дающем достоверную статистику. Таким образом, мы как бы берем в руки окно-лупу и рассматриваем всю область, передвигая по ней окно. Статистические характеристики вычисляются для каждого поднабора данных, попавших в отдельное окно.

Можно построить карту локальных средних значений и стандартных отклонений в окнах. При сравнении с образами данных, приведенными выше, можно увидеть те же области, где локальное среднее велико. Но в дополнение к этому можно выделить области локального изменения вариабельности, которые не детектировались предыдущими методами (рис. 2.10).



а *б* *в*

г *д* *е*

ж *з* *и*

Рис. 2.10. Локальные значения статистики с движущимся окном:
а — количество точек в окне; *б* — среднее значение; *в* — стандартное отклонение; *г* — минимальное значение; *д* — максимальное значение; *е* — размах значений; *ж* — коэффициент вариации, *з* — коэффициент симметрии, *и* — эксцесс

Эффект пропорциональности (proportional effect) состоит в наличии явной зависимости между локальными средними значениями и локальной вариабельностью, описываемой локальным стандартным отклонением, т. е. когда коэффициент вариации $CV=\sigma/m$ демонстрирует явное детерминированное поведение. Можно выделить четыре самых общих случая этой зависимости [Isaaks, Srivastava, 1989]:

- среднее и вариабельность постоянны;
- среднее имеет локальный тренд, в то время как вариабельность остается постоянной;
- среднее постоянно, но изменяется вариабельность;
- и среднее, и вариабельность изменяются вместе пропорционально.

Для определения эффекта пропорциональности можно построить диаграмму разброса (scatterplot) локального стандартного отклонения в зависимости от локального среднего (рис. 2.11). При нормальном распределении данных эффект пропорциональности не наблюдается, и стандартное отклонение обычно постоянно. При логнормальном распределении зависимость между локальным средним и локальным стандартным отклонением линейная. В исследуемых данных корреляция между локальным средним и локальным стандартным отклонениями достаточно высока и равна 0,69 (см. рис. 2.11). Это свидетельствует о наличии в данных эффекта пропорциональности.

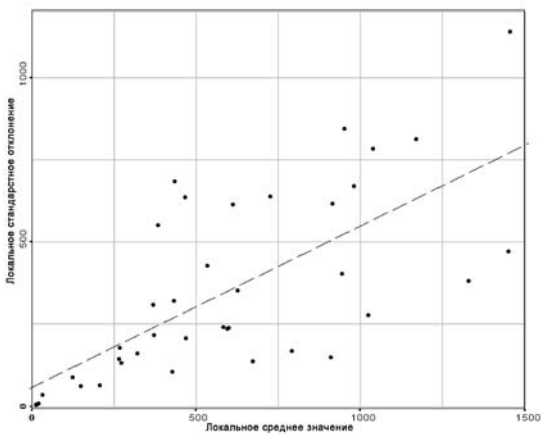


Рис. 2.11. Корреляция локального среднего значения с локальным стандартным отклонением по результатам статистики с движущимся окном

2.7. Стационарность в строгом и мягком смыслах

Пространственная непрерывность связана с другим краеугольным понятием — стационарностью. Стационарность в строгом теоретическом смысле определяется следующим образом.

Если совместная функция распределения (2.1) инвариантна относительно положения начала координат, то в этом случае говорят о стационарности случайной функции $Z(\boldsymbol{x})$ в области S. Это означает, что любые два вектора случайных переменных $\{Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_N)\}$ и $\{Z(\boldsymbol{x}_1 + \boldsymbol{h}), ..., Z(\boldsymbol{x}_N + \boldsymbol{h})\}$ имеют одинаковые условные многомерные функции распределения независимо от вектора сдвига $\boldsymbol{h}$:

$$F\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N; z_1, ..., \boldsymbol{z}_N\right) = F\left(\boldsymbol{x}_1 + \boldsymbol{h}, ..., \boldsymbol{x}_N + \boldsymbol{h}; z_1, ..., z_N\right), \quad \forall\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right) \text{ и } \boldsymbol{h}, \tag{2.8}$$

т. е. функция распределения является трансляционно инвариантной.

Пространственная стационарность в строгом смысле означает, что распределения случайной величины в двух различных зонах области распределения являются идентичными. Таким образом, полная стационарность является скорее теоретическим, чем реально применимым для моделирования природных явлений понятием.

Пространственная нестационарность заключается в меняющемся характере функции распределения в зависимости от местоположения точек измерения.

Гипотеза о пространственной стационарности функции распределения часто необходима при решении задач пространственной интерполяции. Условие стационарности является весьма строгим, поэтому на практике используются более мягкие условия стационарности второго порядка (стационарность в широком смысле) или внутренняя гипотеза. В рамках предположения о стационарности второго порядка, в частности, работает базовый метод геостатистики — кригинг.

Случайная функция $Z(\boldsymbol{x})$ обладает стационарностью второго порядка, если [Journel, Huijbregts, 1978]:

- математическое ожидание $m(\boldsymbol{x})$ существует и не зависит от местоположения $\boldsymbol{x}$:

$$m\left(\boldsymbol{x}\right) = E\left\{Z\left(\boldsymbol{x}\right)\right\} = \text{const}, \ \forall \boldsymbol{x};$$

- для каждой пары значений случайной переменной $\{Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\}$ ковариация существует и зависит только от разности координат $\boldsymbol{h}$:

$$C(\boldsymbol{h}) = E\{Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})\} - m^2, \ \forall \boldsymbol{x}. \quad (2.9)$$

Таким образом, стационарность второго порядка — это стационарность только для моментов первого и второго порядка.

Случайная функция $Z(\boldsymbol{x})$ удовлетворяет *внутренней гипотезе*, если:

- математическое ожидание $m(\boldsymbol{x})$ существует и не зависит от местоположения $\boldsymbol{x}$:

$$m(\boldsymbol{x}) = E\{Z(\boldsymbol{x})\} = \text{const}, \ \forall \boldsymbol{x};$$

- для любого вектора $\boldsymbol{h}$ разность $Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})$ имеет конечную вариацию, не зависящую от $\boldsymbol{x}$ (стационарность приращений):

$$\text{Var}\{Z(\boldsymbol{x}+\boldsymbol{h}) - Z(\boldsymbol{x})\} = E\left\{Z(\boldsymbol{x}+\boldsymbol{h}) - \left[Z(\boldsymbol{x})\right]^2\right\} = 2\gamma(\boldsymbol{h}), \ \forall \boldsymbol{x}. \quad (2.10)$$

Упражнение 2.2. *Если среднее значение и ковариация стационарны, что можно сказать о поведении вариации разности значений функции с расстоянием?*

Из внутренней гипотезы следует определение одного из ключевых понятий геостатистики — *вариограммы*. Функция $\gamma(\boldsymbol{h})$ носит название *полувариограммы* (или вариограммы) и является статистическим моментом второго порядка. Внутренняя гипотеза (intrinsic hypothesis) соответствует стационарности второго порядка для приращений функции.

Центральная идея геостатистики состоит в использовании знаний о пространственной корреляции экспериментальных данных для построения пространственных оценок и интерполяций. Вариограмма — ключевой инструмент для оценки степени пространственной корреляции, имеющейся в данных, и для ее моделирования. Модель вариограммы является функцией, определяющей зависимость изменения исследуемой величины в пространстве от расстояния. Следовательно, интерполяционная модель, основанная на такой корреляционной функции, будет отражать реальные явления, которые лежат в основе данных измерений.

В условиях стационарности второго порядка корреляция между измерениями в двух точках, как уже указывалось, предполагается зависящей только от разности местоположений этих точек. С точки зрения пространственных корреляций это означает, что различные регионы статистически подобны, что, кстати, позволяет интерпретировать различные регионы как различные реализации стохастической региональной функции и делать статистические выводы. Таким образом, значения измерений, проведенных в некотором конечном множестве точек, могут быть исследованы с точки зрения поведения разности между ними. Всевозможные пары точек могут быть рассортированы по классам в соответствии с разностью их координат $\boldsymbol{h} = \boldsymbol{x}_i - \boldsymbol{x}_j$, называемой *лагом* (или лэгом — lag). Для близких точек разность значений функции в них обычно меньше и растет с увеличением расстояния между точками. Вычислив среднее значение квадратов разностей для каждого значения лага $\boldsymbol{h}$ (для каждого собранного класса пар измерений), можно получить дискретную функцию, называемую *экспериментальной вариограммой* (sample variogram, или raw variogram — вариограммой сырых данных). Более подробно построение вариограммы рассмотрено в Главе 4.

Теоретически поведение экспериментальной вариограммы должно иметь отношение к пространственной корреляции между образцами и может содержать количественную информацию о пространственном процессе. Но чтобы использовать эту информацию в теоретических исследованиях и практических оценках, необходимо построить непрерывную гладкую функцию, которая будет представлять собой *теоретическую модель* экспериментальной вариограммы. После такой подгонки (fitting) модельной вариограммы к экспериментальному образцу первая может быть использована для вычисления весов при интерполяции кригингом.

Вариограмма, вообще говоря, — это функция векторного аргумента $\boldsymbol{h}$. Часто случается, что пространственная корреляция зависит не только от расстояния между точками измерений, но и от направления, т. е. данные могут обладать пространственной анизотропией. В этом случае оцениваются вариограммы по направлениям (directional variograms) и строится общая анизотропная модель вариограммы.

Свойство *эргодичности* по отношению к пространственным данным означает, что при вычислении различных статистических моментов можно переходить от усреднения по реализациям к усреднению по пространству, а также делать при этом статистические выводы.

2.8. Геостатистическое оценивание

Основной геостатистической моделью, которая в том или ином виде используется во всех методах геостатистики, является *кригинг* (kriging) — линейный интерполятор, использующий для получения оценки значения функции в некоторой точке пространства $\boldsymbol{x}_0$ экспериментально измеренные значения этой функции в других точках:

$$Z^*(\boldsymbol{x}_0) = \sum_{i=1}^{N(\boldsymbol{x})} w_i(\boldsymbol{x}_0) Z(\boldsymbol{x}_i). \qquad (2.11)$$

Для определения весов $w_i(\boldsymbol{x}_0)$ могут быть использованы различные детерминистические методы, например веса могут браться обратно пропорциональными расстоянию от измеренной точки до оцениваемой или в соответствии с каким-либо другим предположением о природе связей в данных. Однако все эти методы пренебрегают использованием информации о структуре внутренней корреляции пространственных данных.

Следующим критерием при построении модели является условие *несмещенности оценки*, что эквивалентно условию

$$E\left\{Z^*(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right\} = 0, \qquad (2.12)$$

где $Z(\boldsymbol{x}_0)$ — истинное (неизвестное) значение оцениваемой функции в точке $\boldsymbol{x}_0$. Иными словами, ошибки интерполяции должны иметь в каждой точке среднее, равное нулю. Это условие может быть реализовано и в рамках детерминистических подходов.

Еще одно условие, которое мы хотим наложить, — оптимальность интерполяции в смысле *минимизации вариации ошибки оценки,* т. е. веса w_i линейной регрессии в уравнении (2.11) должны быть выбраны так, чтобы минимизировать значение вариации ошибки оценки:

$$\mathrm{Var}\left[Z^*(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right] = E\left\{\left[Z^*(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right]^2\right\}. \qquad (2.13)$$

Таким образом, кригинг является наилучшим (в смысле минимума вариации оценки) линейным и несмещенным оценивателем (the best linear unbiased estimator — BLUE). В процессе поиска минимума вариации (2.13) ключевую роль играет использование модели вариограммы исходных данных. В результате поиска весовых коэффициентов для получения оценки, удовлетворяющей всем перечисленным условиям, удается оценить и значение

вариации (2.13), которое может интерпретироваться как описание точности кригинговой оценки. Более подробно теория кригинга изложена в Главе 5.

2.9. Проверка качества модели — кросс-валидация

При использовании той или иной модели интерполяции крайне важно правильно подобрать значения модельно-зависимых параметров. Для кригинга такими параметрами являются параметры модели вариограммы. При работе с реальными данными не всегда удается сразу выбрать теоретическую модель экспериментальной вариограммы. Для проверки качества выбранной модели используют различные количественные методы: кросс-валидацию (cross-validation), метод складного ножа (jack-knife), бутстреп (bootstrap).

Кросс-валидация — наиболее простой и часто использующийся не только в геостатистике подход при сравнении результатов, получаемых различными методами или одним и тем же методом, но с различными параметрами. Выполняется кросс-валидация следующим образом:

- из базы данных временно изымается одна точка, и для нее проводится оценка значения;
- полученное значение сравнивается с известным, и вычисляется *невязка* — разница между измеренными и оцененными значениями:

$$\Delta Z(\boldsymbol{x}) = Z(\boldsymbol{x}) - Z^*(\boldsymbol{x});$$

- первые два шага проводятся для всех точек базы данных.

Полученные невязки $\Delta Z(\boldsymbol{x})$ могут быть графически представлены в виде карты (карты невязок), по которой можно посмотреть, в каких зонах метод срабатывает лучше, а в каких хуже. Вместо невязок можно визуализировать *относительные ошибки:*

$$\text{relative error}(\boldsymbol{x}_i) = \frac{Z(\boldsymbol{x}_i) - Z^*(\boldsymbol{x}_i)}{Z(\boldsymbol{x}_i)}.$$

Полезно также представить результаты кросс-валидации в виде графика $Y\left(Z(\boldsymbol{x})\right) = Z^*(\boldsymbol{x})$ или аналогичного ему — $Y\left(Z(\boldsymbol{x})\right) = \Delta(\boldsymbol{x})$. Проведение на таком графике биссектрисы (или соответственно прямой $Y = 0$), соответствующей равенству оценки и исходного значения, позволяет видеть характер отклонения: большее отклонение для высоких или для низких значений Z, какие-либо тренды в поведении оценки и т. п. Вместе с тем

на графиках невязок можно проследить эффект сглаживания — область низких значений в среднем переоценивается, а область высоких значений недооценивается.

Кроме локальных характеристик кросс-валидация позволяет оценить и глобальные характеристики оценки для сравнения:

1. Смещение $\Delta m = m - m^*$, где m — среднее, оцененное по исходным данным; m^* — среднее, оцененное по полученным результатам.
2. Сумму квадратов невязок:

$$S = \sum_{i=1}^{n} \left[Z(\boldsymbol{x}_i) - Z^*(\boldsymbol{x}_i) \right]^2 + R,$$

   где R — штрафной член, вводящийся для контроля количества неоцененных точек.

3. Среднюю квадратичную ошибку (root mean square error — RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{n} \left[Z(\boldsymbol{x}_i) - Z^*(\boldsymbol{x}_i) \right]^2}.$$

4. Коэффициент эффективности:

$$E = \frac{S}{S_0},$$

   где $S_0 = \sum_{i=1}^{n} \left[Z(\boldsymbol{x}_i) - m \right]^2.$

5. Коэффициент корреляции ρ, угол наклона регрессионной прямой на графике $Y\left(Z(\boldsymbol{x})\right) = Z^*(\boldsymbol{x})$.

Вообще говоря, кросс-валидация — это частный случай метода складного ножа, когда выбираемый набор состоит из одной точки (leave-one-out).

Метод складного ножа (jack-knife) является общим случаем кросс-валидации, когда оценивание проводится не в одной, а в нескольких точках измерений, данные о которых предварительно изымаются из рассмотрения. Полученные в результате невязки анализируется методом, аналогичным описанному выше. Поскольку при джек-найфе изымается произвольный набор данных, комбинации этого набора могут варьироваться, что делает этот метод стохастическим.

Бутстреп (bootstrap) состоит в оценке на основе случайных выборок из набора данных. Выборки делаются из исходного набора случайным образом. Выбранная точка не изымается, она может попасть в выборку несколько раз. Оценка проводится по оставшимся не выбранными точкам. Обычно процедура выборки и оценки повторяется много раз.

Литература

Cressie N. Statistics for spatial data. — New York: John Wiley & Sons, 1991. — 900 p.

Deutsch C. DECLUS: a FORTRAN 77 program for determining optimal declustering weights // Computers and Geosciences. — 1989. — Vol. 15. — P. 325—332.

Deutsch C. V., Journel A. G. GSLIB: Geostatistical Software Library and User's Guide. —New York; Oxford: Oxford Univ. Press, 1998. — 369 p.

Engineering and Design: Practical aspects of applying geostatistics at hazardous, toxic and radioactive waste sites: Technical Letter ETL 1110-1-175 / Department of the US Army. — Washington, 30 June 1997. — 93 p.

Goovaerts P. Geostatistics for Natural Resources Evaluation. — [S. l.]: Oxford Univ. Press, 1997.

Hengl T. Finding the right pixel size // Computers and Geosciences. — 2006. — Vol. 32. — P. 1283—1298.

Isaaks E. H., Srivastava R. M. An Introduction to Applied Geostatistics. — Oxford: Oxford Univ. Press, 1989.

Journel A. G. Nonparametric estimation of spatial distributions // Mathematical Geology. — 1983. — Vol. 15. — P. 445—468.

Journel A. G., Huijbregts Ch. J. Mining Geostatistics. — London: Academic Press, 1978. — 600 p.

Mandelbrot B. B. The fractal theory of nature. — New York: Freeman, 1982.

Morishita M. Measuring of the dispersion and analysis of distribution patterns // Memoires of the Faculty of Science, Kyushu University. Series E. Biology. — 1959. — Vol. 2. — P. 215—235.

Preparata F. P., Shamos M. I. Computational Geometry. — New York: Springer-Verl., 1985. — P. 198—218.

Raes F., Graziani G., Girardi F. A simple and fractal analysis of the European on-line network for airborne radioactivity monitoring // Environmental Monitoring and Assessment. — 1991. — Vol. 18. — P. 221—234.

Глава 3

Детерминистические методы пространственной интерполяции

Детерминистические методы традиционно широко используются в различных областях прикладной и научной деятельности. Например, широко известный пакет SURFER содержит достаточно большую коллекцию таких методов [SURFER..., 2002]. Приведем некоторые наиболее часто встречающиеся алгоритмы и отметим особенности их использования. В трех разделах этой главы описаны три основных подхода к детерминистической интерполяции: линейные модели, полиномиальные модели и модели базисных функций.

Детерминистические методы интерполяции предполагают наличие заданной аналитической зависимости между значениями функции в пространстве. Эти методы популярны из-за простоты вычисления оценки по заданной параметрической формуле. Наиболее широко применяемые «формульные» зависимости: обратная пропорциональность расстоянию (или его степени), сплайны, полиномы различных степеней и пр. Однако детерминистические интерполяции имеют ряд серьезных недостатков: они не дают возможности характеризовать качество оценки, настройка параметров часто не предполагается или производится скрыто от пользователя, многие методы пренебрегают пространственной корреляцией и т. п. Тем не менее рассмотрим наиболее распространенные детерминистические подходы для пространственной интерполяции.

При использовании детерминистических методов предполагается, что анализируемые данные описываются некоторой детерминистической функцией $Z(\boldsymbol{x}, \lambda)$, определенной на исследуемой области S, где $\boldsymbol{x} \in S$ — координаты точки; λ — набор внутренних параметров модели. Задача состоит в том, чтобы, базируясь на известных данных ($Z_i = Z(\boldsymbol{x}_i)$ — значения, измеренные в точках $\boldsymbol{x}_i \in S$) и на другой контекстной информации об исследуемом явлении, подобрать набор параметров λ и построить функцию $Z(\boldsymbol{x}, \lambda)$ для всей исследуемой области S. После этого значение в любой точке просто вычисляется по формуле.

Детерминистические интерполяторы могут быть глобальными (все точки с известными значениями используются при интерполяции) или локальными

(только часть значений в точках, ближайших к оцениваемой, используются для интерполяции). Глобальные интерполяторы делают искомую функцию более сглаженной. При использовании локальных методов окрестность, используемая для оценки, может задаваться различными способами. Может быть фиксировано число ближайших к оцениваемой точке соседей, использующихся при интерполяции: $N(\boldsymbol{x}) = N = \text{const}$. Тогда размер зоны, влияющей на значение в точке $\boldsymbol{x}$, зависит от локальной плотности точек измерения. Возможно, наоборот, задание размера области (например, область поиска D), точки из которой используются при оценивании значения в $\boldsymbol{x}$. В этом случае $N(\boldsymbol{x})$ будет меняться и в области с редкими измерениями при малом значении D могут возникнуть неоцененные зоны (в D-окрестности точки $\boldsymbol{x}$ недостаточно измерений).

В этой главе все методы проиллюстрированы на данных по радиоактивному загрязнению почвы ^{137}Cs в западной части Брянской области, которые уже использовались в Главе 2.

3.1. Линейные интерполяторы

Линейные интерполяторы представляют искомую функцию в виде линейной комбинации известных значений:

$$Z^*(\boldsymbol{x}_0) = \sum_{i=1}^{N(\boldsymbol{x}_0)} w_i(\boldsymbol{x}_0) Z(\boldsymbol{x}_i), \tag{3.1}$$

где $Z^*(\boldsymbol{x}_0)$ — оцениваемое значение в точке $\boldsymbol{x}_0$; $Z(\boldsymbol{x}_i)$ — известные значения в точках измерений $\boldsymbol{x}_i$; $N(\boldsymbol{x}_0)$ — количество исходных точек, принимающих участие в оценке для координаты $\boldsymbol{x}_0$; $w_i(\boldsymbol{x}_0)$ — весовые коэффициенты. В данном случае набор параметров состоит из весовых коэффициентов и количества точек для оценки. Эти параметры определяются отдельно для каждой точки, подлежащей оценке.

Линейные интерполяторы различаются формой весовых коэффициентов, которые задают различные особенности функции. Например, в форме линейного интерполятора можно задать полигонный *метод Тиссена (метод ближайшего соседа)*. В этом случае весовые коэффициенты задаются формулой

$$w_i(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in A(\boldsymbol{x}_i), \\ 0, & \boldsymbol{x} \notin A(\boldsymbol{x}_i), \end{cases}$$

где $A(\boldsymbol{x}_i)$ — область влияния точки x_i.

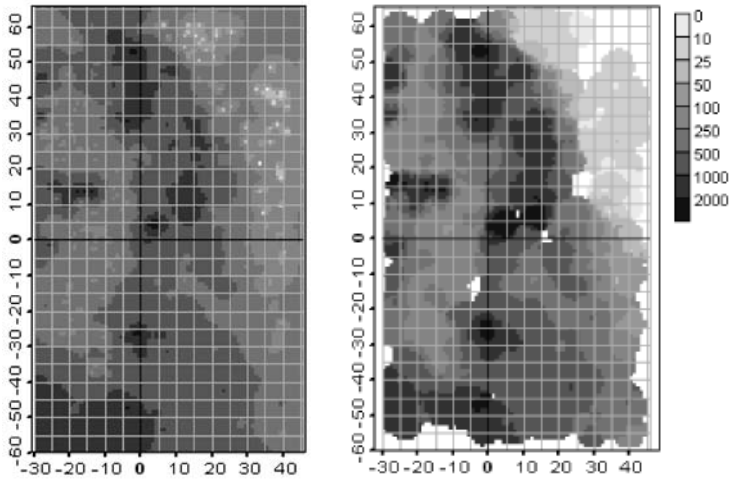
Широко используется линейный интерполятор с весовыми коэффициентами, обратно пропорциональными расстоянию до оцениваемой точки в степени. Весовые коэффициенты, определяются по формуле

$$w_i(\boldsymbol{x}) = \frac{\frac{1}{\boldsymbol{h}_i^{\beta}}}{\sum_{j=1}^{n} \frac{1}{\boldsymbol{h}_j^{\beta}}},$$

где β — степень; $\boldsymbol{h}_i = \sqrt{d_i^2 + \delta^2}$; d_i — расстояние между точками $\boldsymbol{x}_i$ и $\boldsymbol{x}_0$; δ — сглаживающий параметр. Нижняя часть дроби в весовом коэффициенте вводится для выполнения условия несмещенности оценки $E\{Z^*(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\} = 0,$ которое соответствует условию на весовые коэффициенты $\sum_{i=1}^{n} w_i = 1.$

Этот метод может быть точным (точное воспроизведение значений в исходных точках) и сглаженным, что характеризуется сглаживающим параметром. Точным метод будет при $\delta = 0$. В этом случае возникает артефакт в виде «бычьих глаз» (выгибание к точным значениям) вокруг точек измерений. Сглаживающий параметр способствует удалению этого артефакта.

В качестве значения степени чаще всего используется значение 2. Такой вариант метода известен как *метод обратных квадратов* [Grimm, Lynch, 1991].



а *б*

Рис. 3.1. Результат интерполяции методом обратных квадратов:
а — радиус поиска 50, *б* — радиус поиска 5

Примеры использования метода обратных квадратов в глобальном и локальном вариантах приведены на рис. 3.1 для различных значений параметра модели — радиуса области поиска $N(\boldsymbol{x}_0)$ соседних данных для оценки (3.1). Чем больше радиус поиска, тем более размыта интерполяционная оценка (см. рис. 3.1*а*). При малом радиусе поиска оценка становится более контрастной, при этом высокие значения не сглаживаются (см. рис. 3.1*б*).

Проблема сильной зависимости простейшей интерполяционной оценки метода обратных квадратов от единственного параметра — размера области поиска — является на самом деле более глубокой. Линейная регрессионная оценка предполагает наличие некоторой зависимости между данными, участвующими в интерполяции. Это может быть обратная пропорциональность квадрату расстояния, как в описанном выше методе, либо более сложная зависимость. Зависимости между данными в пространстве могут распространяться на ограниченное расстояние. Так, при использовании всех данных для оценки в любой точке предполагается, что данные на любых расстоянии имеют влияние на значения оценки. В противоположном предельном случае зависимости между данными не существует даже на минимальном расстоянии между точками. Это означает, что данные распределены абсолютно случайно и ни один из методов интерполяции не имеет смысла, поскольку при их использовании предполагается та или иная зависимость.

Таким образом, выбор подходящего радиуса поиска тесно связан с понятиями пространственной непрерывности и пространственной корреляции, которые подробно рассмотрены в Главе 4.

Для подбора оптимального значения радиуса поиска можно использовать один из алгоритмов проверки качества оценки — кросс-валидацию или метод складного ножа, которые были описаны в главе 2. При использовании кросс-валидации оптимальный радиус поиска определяется путем минимизации кросс-валидационной ошибки. На этом принципе основаны некоторые алгоритмы пространственного *автокартирования* [Kanevski et al., 1999].

На рис. 3.2 изображены интерполяционные оценки методом обратных расстояний в степени на основе трех точек: (2, 2), (5, 8) и (8, 5).

Упражнение 3.1. *Какие степени — 1, 2 или 3 — использованы для получения интерполяционных оценок A, B и C?*

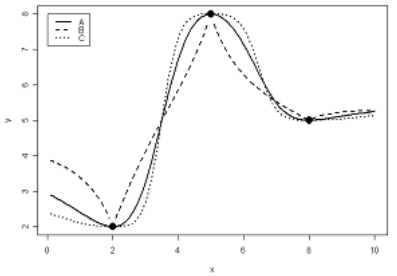


Рис. 3.2. Оценки методом обратных расстояний в степени для степеней 1, 2 и 3

Если область поиска определяется радиусом, то он может быть в явном виде введен в весовые коэффициенты, что реализовано, например, в *весовых коэффициентах Крессмана* [Grimm, Lynch, 1991]. Этот метод по сути аналогичен методу обратных квадратов:

$$w_i(\boldsymbol{x}) = \frac{D_i^2 - R_i^2}{D_i^2 + R_i^2}, \tag{3.2}$$

где D_i — радиус влияния i-й точки. Радиус влияния может быть задан постоянным по всей исследуемой области, а может меняться в зависимости от локальной плотности точек измерения. При использовании весовых коэффициентов Крессмана их рекомендуется нормализовать так, чтобы выполнялось условие

$$\sum_{i=1}^{N(\boldsymbol{x})} w_i(\boldsymbol{x}) = 1.$$

Весовые коэффициенты можно определять и совсем иначе, например основываясь на системе полигонов Вороного (подробнее о полигонах Вороного см. в Главе 2). Такой подход называется *методом естественного соседа*.

В этом случае определению весовых коэффициентов предшествует построение системы полигонов Вороного (областей влияния) для исходного набора точек. Чтобы вычислить весовые коэффициенты для некоторой точки оценивания, систему полигонов модифицируют для включения и этой точки. При этом часть полигонов уменьшается в размере, так как полигон новой точки отнимает зоны от полигонов соседних с ней точек из исходного набора. Эти зоны называют «арендованными». Весовые коэффициенты точек x_i определяются пропорционально арендованной у них зоне:

$$w_i = \frac{p_{io}}{\sum_{j=1}^{n} p_{j0}},$$

где p_{i0} — площадь зоны, арендованной точкой x_0 у точки x_i.

Основной проблемой этого метода является неоцененная зона за пределами выпуклого многоугольника, окружающего исходный набор точек. Пример работы этого метода приведен на рис. 3.3.

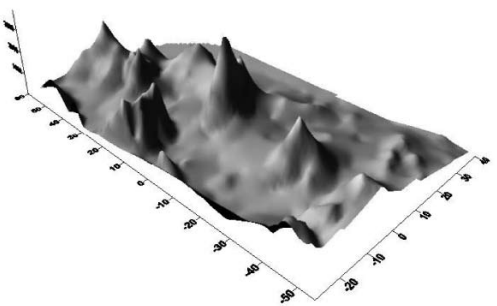


Рис. 3.3. Результат интерполяции методом естественного соседа

В зависимости от знаний или предположений о природе процесса, задаваемого исходными данными, могут вводиться и другие способы описания весовых коэффициентов.

3.2. Полиномиальные методы

Полиномиальные интерполяторы [Goodin et al., 1979] представляют значение в точке в виде полинома от координат. В двумерном случае — для точки $\boldsymbol{x}$ с координатами (x, y) $Z^*(x, y) = P_n(x, y)$, где P_n — полином n-й степени.

Обычно на практике для двумерного случая используют один из четырех типов полиномов:

- плоскость: $P_1(x, y) = a + bx + cy;$
- билинейно-седловой: $P_{1.5}(x, y) = a + bx + cy + dxy; \frac{n!}{r!(n-r)!}$
- квадратичный: $P_2(x, y) = a + bx + cy + dxy + ex^2 + fy^2;$
- кубический: $P_3(x, y) = a + bx + cy + dxy + ex^2 + fy^2 + gx^2y + hxy^2 + ix^3 + jy^3.$

Теоретически можно использовать и полиномы более высокого порядка. Они определяются максимальной степенью при x, максимальной степенью при y и совместной максимальной степенью. Все промежуточные степени в полиноме будут присутствовать.

Задача полиномиальной интерполяции сводится к тому, чтобы определить неизвестные коэффициенты a_i так, чтобы полиномы максимально хорошо соответствовали данным в заданных точках. Для этого находят минимум по всем коэффициентам (a, b, c, d и т. д.) функции χ^2, задающей интегральную ошибку интерполяции и определяемой следующим образом:

$$\chi^2 = \sum_{i=1}^{N} \left[Z(x_i, y_i) - P_n(x_i, y_i) \right]^2.$$

Минимизация состоит в решении системы линейных уравнений с числом неизвестных, равным числу уравнений. Число уравнений (неизвестных) зависит от выбранного полинома.

Любой глобальный полиномиальный метод, вообще говоря, не является интерполятором в строгом смысле, скорее он относится к аппроксиматорам. Его можно использовать, например, для выделения крупномасштабного тренда.

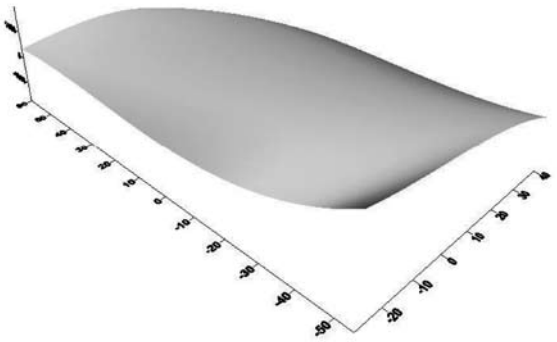
Можно воспользоваться и локальным вариантом полиномиального метода, когда поиск коэффициентов производится только на основе данных, попавших в зону поиска. Примеры применения глобального полигона третьего порядка и локального варианта с полигоном второго порядка приведены на рис. 3.4.

3.3. Метод базисных функций

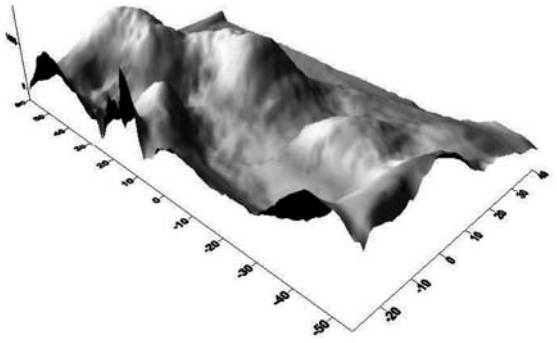
Оценка методом базисных функций строится как линейная комбинация из базисных функций:

$$Z^*(\boldsymbol{x}_0) = \sum_{i=1}^{n} c_i B(h_{0i}),$$

где h_{0i} — расстояние от точки x_0 до точки x_i; $B(h_{0i})$ — базисная функция, определяемая от расстояния; c_i — весовые коэффициенты. Коэффициент c_i определяет алгебраический знак вхождения соответствующего члена и степень его влияния. Классический вариант метода является точным, но возможно введение сглаживающего параметра δ.



а



б

Рис. 3.4. Результат глобальной интерполяции полигоном третьей степени (*а*) и локальной интерполяции полигоном второй степени (*б*)

Традиционно используются следующие типы базисных ядерных функций:

- обратный мультиквадрик: $B(h) = \frac{1}{\sqrt{h^2 + \delta^2}}$;
- мультилогарифмический: $B(h) = \lg\left(h^2 + \delta^2\right)$;
- мультиквадратичный: $B(h) = \sqrt{h^2 + \delta^2}$;
- естественный кубический сплайн: $B(h) = \left(h^2 + \delta^2\right)^{\frac{1}{2}}$;
- тонкий сплайн: $B(h) = (h^2 + \delta^2)\lg(h^2 + \delta^2)$.

Весовые коэффициенты c_i определяются из условия точности оценки в известных точках, т. е. во всех заданных точках (x_i, y_i) модель интерполяции должна давать оценку, равную заданным значениям $V(x_i, y_i)$. Определение весов производится при $\delta = 0$, при использовании сглаженного варианта δ используется при оценке. Таким образом, чтобы найти коэффициенты c_i, необходимо решить систему из N линейных уравнений с N неизвестными:

$$\begin{cases} \sum_{i=1}^{N} c_i \left[q\left(x_i, y_i, x_1, y_1\right) \right] = V\left(x_1, y_1\right), \\ \ldots \\ \sum_{1}^{N} c_i \left[q\left(x_i, y_i, x_N, y_N\right) \right] = V\left(x_N, y_N\right). \end{cases}$$

Метод базисных функций обладает универсальностью и эффективностью. Пример применения метода базисных функций (мультиквадратичные ядра) приведен на рис. 3.5.

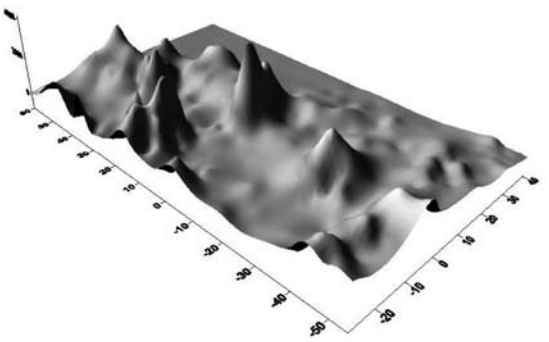


Рис. 3.5. Результат интерполяции методом базисных функций (мультиквадратичные ядра)

Литература

Bartier P. M., Keller C. P. Multivariate interpolation to incorporate thetic surface data using inverse distance weighting // Computers and Geosciences. — 1996. — Vol. 22, N 7. — P. 795—799.

Franke R. Scattered Data Interpolation: Test of Some Methods // Mathematics of Computation. — 1982. — Vol. 38, N 157. — P. 181—200.

Goodin W. R., McRae G. J., Seinfeld J. H. A comparison of interpolation methods for sparse data: application to wind and concentration fields // J. of Applied Meteorology. — 1979. — Vol. 18. — P. 761—771.

Grimm J. W., Lynch J. A. Statistical analysis of error in estimating wet deposition using five surface estimation algorithms // Atmospheric Environment. — 1991. — Vol. 25A. — P. 317—127.

Kanevski M., Demyanov V., Chernov S. et al. Geostat Office for Environmental and Pollution Spatial Data Analysis // Mathematische Geologie. — 1999. — Vol. 3, N 4. — P. 73—83.

Lattuada R., Raper J. Applications of 3D Delaunay triangulation algorithms in geoscientific modelling // http://www.iah.bbscr.ac.uk/phd/gisruk95.html.

Macidonio G., Pareschi M. T. An algorithm for the triangulation of arbitrarily distributed points: applications to the volume estimate and terrain fitting // Computers and Geosciences. — 1991. — Vol. 17, N 7. — P. 859—874.

Okabe A., Boot B., Sugihara K. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. — New York: J. Wiley & Sons, 1992. — 532 p.

Saunderson H. C. Multiquadric surfaces in C // Computers and Geosciences. — 1994. — Vol. 20, N 7/8. — P. 1103—1122.

SURFER (R) Version 8.0: Surface Mapping System / Golden Software, Inc. — [S. l.], 2002.

Tabois G. Q., Salas J. D. A Comparative Analysis, for Spatial Interpolation of Precipitation // Water Resources Bul. — 1985. — Vol. 21, N 3. — P. 365 —380.

Weber D., Englund E. Evaluation and Comparison of Spatial Interpolators // Mathematical Geology. —1992. — Vol. 24, N 4. — P. 381—391.

Глава 4

Анализ и моделирование пространственной корреляции. Вариография

В этой главе подробно рассмотрена основная тема геостатистики — вариография. Под вариографией понимают анализ и моделирование пространственной корреляционной структуры данных. В Разделе 4.1 мы еще раз вернемся к пространственной непрерывности, описанной в Разделе 2.6. В Разделе 4.2 собраны различные меры пространственной корреляции, количественно характеризующие пространственную непрерывность. Раздел 4.3 посвящен построению экспериментальной вариограммы для набора пространственных данных. Моделирование построенной экспериментальной вариограммы с помощью аналитических функций описано в Разделе 4.4. В Разделах 4.5, 4.6 изложены свойства вариограммы на больших расстояниях и вблизи нуля. Различные типы анизотропной пространственной корреляции описаны в Разделе 4.7. В Раздел 4.8 вынесено практическое упражнение на определение вариограмм для различных пространственных образов. Проблемы, связанные с использованием вариограммы, рассмотрены в Разделах 4.8 и 4.9. В Разделе 4.10 приведен пример анализа и моделирования пространственной корреляционной структуры для реальных данных.

4.1. Пространственная непрерывность

Важным свойством пространственно распределенных данных является *пространственная непрерывность*, которая означает, что близко расположенные в пространстве измерения скорее всего будут иметь близкие значения. Пространственная непрерывность данных обычно описывается с помощью корреляционных и ковариационных функций (статистических моментов), выражающих меру этой непрерывности. В геостатистике корреляция может быть представлена статистическими моментами. Одной из наиболее популярных функций является *вариограмма* — статистический двухточечный момент второго порядка. Использование вариограммы обусловлено про-

стотой ее применения в интерполяционных моделях кригинга. По этой причине этап анализа и описания пространственной корреляционной структуры данных в геостатистике принято называть *вариографией*.

Анализ пространственной корреляционной структуры данных можно разбить на два этапа:

- построение и интерпретация мер пространственной непрерывности на основе данных;
- моделирование пространственной корреляционной структуры; построение теоретической функции, аппроксимирующей экспериментальные значения мер корреляции аналитической формулой.

Сущность вариографии состоит в выявлении наличия корреляционной структуры в данных и ее описании. Подробнее это означает, например, проверку данных на наличие или отсутствие крупномасштабного пространственного тренда (видимой связи значения в точке с ее реальным местоположением). Тренд может быть описан некоторой математической функцией. Проверяется также зависимость корреляционной структуры от взаимной пространственной ориентации точек, т. е. наличие или отсутствие пространственной анизотропии. Определяется эффективный радиус корреляции данных (если он существует) — максимальное расстояние, на котором еще наблюдается зависимость между значениями в точках.

Конечной целью этапа вариографии является построение аналитической функции, описывающей пространственную корреляционную структуру данных для использования в геостатистических моделях интерполяции (в кригинге). Иными словами, конечной целью этапа вариографии является построение модели вариограммы. Качество этой модели определяет и качество последующей геостатистической оценки, и величину ее ошибки.

4.2. Меры пространственной корреляции

Для описания пространственной корреляции данных можно использовать различные моменты второго порядка. Все они характеризуют похожесть (или непохожесть) данных в зависимости от их взаимного расположения (расстояния и направления), тем самым описывая пространственную непрерывность. Чтобы делать статистические выводы о характере распределения при наличии только одной реализации случайной величины (данных

измерений), требуется принять дополнительные предположения о пространственной стационарности. Понятия стационарности, стационарности второго порядка и внутренняя гипотеза подробно рассмотрены в Главе 2.

Приняв предположение о *стационарности второго порядка* или *внутреннюю гипотезу*, считаем, что функции корреляции между данными зависят только от взаимного расположения точек измерений, а не от их конкретного местоположения в пространстве. Это означает, что корреляционные функции определяются вектором $\boldsymbol{h}$. Для изотропного случая, когда корреляция не зависит от направления, а только от расстояния, — вектор $\boldsymbol{h}$ переходит в скаляр (расстояние): $h = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$.

Для проведения пространственного корреляционного анализа можно использовать следующие моменты второго порядка, обеспечивающие различное описание пространственной непрерывности на основе двухточечной статистики (пар точек).

Ковариация (covariance) — статистическая мера корреляции между двумя значениями $Z(\boldsymbol{x})$ и $Z(\boldsymbol{x} + \boldsymbol{h})$ в точках, разделенных вектором $\boldsymbol{h}$:

$$C(\boldsymbol{h}) = E\left\{\left[Z(\boldsymbol{x}) - m(\boldsymbol{x})\right]\left[Z(\boldsymbol{x}+\boldsymbol{h}) - m(\boldsymbol{x}+\boldsymbol{h})\right]\right\}.$$

Для $N(\boldsymbol{h})$ экспериментальных точек, разделенных вектором $\boldsymbol{h}$,

$$C(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})} Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h}) - m_{-\boldsymbol{h}}m_{+\boldsymbol{h}},$$

где $m_{-\boldsymbol{h}}$ — среднее значение для данных, находящихся в началах вектора $\boldsymbol{h}$:

$$m_{-\boldsymbol{h}}(\boldsymbol{h}) = \frac{1}{N(h)}\sum_{i=1}^{N(\boldsymbol{h})} Z(\boldsymbol{x}_i);$$

$m_{+\boldsymbol{h}}$ — среднее значение для данных, находящихся в концах вектора $\boldsymbol{h}$:

$$m_{+\boldsymbol{h}}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})} Z(\boldsymbol{x}_i + \boldsymbol{h}).$$

Таким образом, при условии локальной стационарности

$$m(\boldsymbol{h}) = 0,5\left(m_{+\boldsymbol{h}} + m_{-\boldsymbol{h}}\right).$$

Ковариацией можно пользоваться только в рамках предположения о стационарности второго порядка. Ковариация характеризует степень похожести данных — чем более похожи данные (ближе значения), тем больше значение ковариации.

Полувариограмма (semivariogram), или просто вариограмма — вариация разницы значений переменной в двух точках как функция расстояния и направления между ними:

$$\gamma(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) = 0,5\mathrm{Var}\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right] = 0,5E\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right]^2.$$

Для $N(\boldsymbol{h})$ экспериментальных точек, разделенных вектором $\boldsymbol{h}$,

$$\gamma(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}\left[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{h})\right]^2.$$

Для существования вариограммы не требуется стационарности второго порядка, достаточно выполнения внутренней гипотезы. Вариограмма характеризует степень различия данных в зависимости от расстояния между ними. Чем ближе значения данных (меньше разница между ними), тем больше значение вариограммы.

Вариограмма обладает рядом полезных свойств.

- Вариограмма, как функция приращений значений переменных, не подвержена влиянию постоянных компонент переменной, которые отфильтровываются в предположении о стационарности второго порядка.
- Квадрат разницы делает вариограмму очень чувствительной к влиянию предельных значений (крайних высоких и низких). Ниже описаны менее чувствительные корреляционные функции с более стабильным поведением в присутствии предельных значений.
- Вариограмма точечно симметрична относительна нуля: $\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$.
- Теоретически в соответствии с определением в нуле вариограмма должна быть равна нулю ($\gamma(0) = 0$), но на практике бывают случаи, когда это условие приходится игнорировать. Они рассмотрены ниже при обсуждении проблем моделирования вариограммы.
- При выполнении условия стационарности второго порядка выполняются следующие соотношения между вариограммой и ковариацией:

$$\gamma(\infty) = C(0),$$
$$\gamma(\boldsymbol{h}) = C(0) - C(\boldsymbol{h}),$$

где $C(0)$ — априорная ковариация или экспериментальная вариация $\mathrm{Var}[Z(\cdot)]$ функции Z.

Упражнение 4.1. Вывод выражения для ковариации

Показать, что $C(\boldsymbol{x},\boldsymbol{h}) = E\{Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})\} - m^2$.
При условии стационарности среднего значения

$$m(\boldsymbol{x}) = m(\boldsymbol{x}+\boldsymbol{h}) = m = E\{X(\cdot)\} = \text{const}.$$

В соответствии со свойством аддитивности $E\{A+B\} = E\{A\} + E\{B\}$.

Упражнение 4.2. Связь между ковариацией и вариограммой

Показать, что при условии стационарности второго порядка

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

Упражнение 4.3. Симметрия вариограммы

Показать, что $\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$.

Существуют статистические моменты, аналогичные вариограмме, но отличающиеся степенью, в которую возводится разница значений пар точек [Goovaerts, 1997].

Мадограмма (madogram) — модуль разницы — позволяет уменьшить влияние больших разбросов значений по сравнению с вариограммой:

$$M(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})}\sum_{i=1}^{N(h)} |Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})|.$$

Родограмма (rodogram) — квадратный корень — еще более понижает влияние значений с большим разбросом:

$$R(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})}\sum_{i=1}^{N(h)} |Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})|^{\frac{1}{2}}.$$

Дрейф (drift) — очень важная функция при анализе пространственной корреляции:

$$D(\boldsymbol{h}) = E\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right].$$

Для экспериментальных данных он вычисляется по формуле

$$D(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right].$$

В отличие от вариограммы дрейф не обладает точечной симметрией.

Дрейф может служить указателем правомочности предположения о внутренней гипотезе для данных. Такой вывод можно сделать, если значение $D(\boldsymbol{h})$

колеблется вблизи нуля. Если же $D(\boldsymbol{h})$ растет (или убывает) с ростом $|\boldsymbol{h}|$, то данные не подчиняются даже внутренней гипотезе, не говоря уже о более строгом условии стационарности второго порядка. Это может, в частности, означать, что у данных имеется систематический тренд, т. е. определенная зависимость значения исследуемой функции от пространственного местоположения (координаты). Для таких данных моделирование вариограммы и использование обычных геостатистических оценивателей может привести к необоснованным результатам. В этом случае необходимо использовать специальные методы. О некоторых из них рассказано в последующих главах.

Пример графического представления описанных мер пространственной корреляции приведен на рис. 4.1, иллюстрирующем связь между различными мерами пространственной корреляции для одного направления при выполнении гипотезы о стационарности второго порядка.

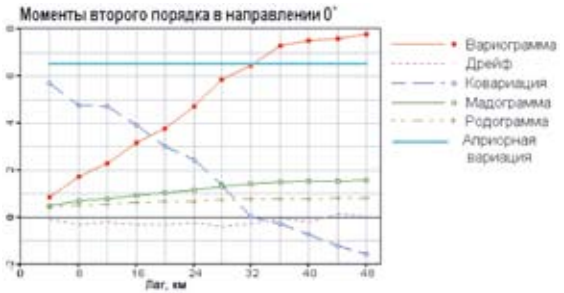


Рис. 4.1. Пример поведения характеристик пространственной корреляционной структуры данных

И для ковариации, и для вариограммы существуют стандартизованные варианты: *коррелограмма* и *стандартизованная вариограмма*. Эти моменты второго порядка являются более робастными, т. е. более устойчивыми к зашумленным данным и присутствию выбросов (outliers). Они вычисляются по следующим формулам:

коррелограмма (correlogram): $\rho(\boldsymbol{h}) = \dfrac{C(\boldsymbol{h})}{\sigma_{-\boldsymbol{h}}\sigma_{+\boldsymbol{h}}}$;

стандартизованная вариограмма (standardised variogram):

$$\gamma_{\text{st}}(\boldsymbol{h}) = \frac{\gamma(\boldsymbol{h})}{\sigma_{-\boldsymbol{h}}\sigma_{+\boldsymbol{h}}},$$

где σ_{-h} и σ_{+h} — стандартные отклонения для точек, находящихся соответственно в начале и в конце вектора $\boldsymbol{h}$:

$$\sigma^2_{-\boldsymbol{h}} = \frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}\left[Z(\boldsymbol{x}_i) - m_{-\boldsymbol{h}}\right]^2,$$

$$\sigma^2_{+\boldsymbol{h}} = \frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}\left[Z(\boldsymbol{x}_i + \boldsymbol{h}) - m_{+\boldsymbol{h}}\right]^2.$$

Упражнение 4.4. Вариограмма и коррелограмма

Определить, при каком предположении верно следующее соотношение:
$\gamma(\boldsymbol{h}) = \rho(0) - \rho(\boldsymbol{h})$.

Очень часто в реальных данных наблюдается *эффект пропорциональности*. Он уже рассматривался в разделе 2.6. Здесь мы рассмотрим проявление эффекта пропорциональности в появлении зависимости между средним $m(\boldsymbol{h})$ и вариограммой $\gamma(\boldsymbol{h})$. Эффект пропорциональности значительно усложняет понимание результата анализа пространственного корреляционного анализа, внося в вариограмму дополнительную зависимость. Как было описано в Главе 2, обнаружить этот эффект можно с помощью вычисления локальных статистических характеристик — среднего и вариации. Для этого используется метод движущегося окна (см. раздел 2.6). При наличии эффекта пропорциональности предлагается использовать *относительные вариограммы*, которые нивелируют эффект пропорциональности [Isaaks, Srivastava, 1989].

Можно выделить два вида *относительных вариограмм* (relative variogram):

- *общую относительную вариограмму* (general relative variogram):

$$\gamma_{\mathrm{GR}}(\boldsymbol{h}) = \frac{\gamma(\boldsymbol{h})}{m^2(\boldsymbol{h})},$$

  когда вариограмма просто нормируется на квадрат среднего значения данных, разделенных вектором $\boldsymbol{h}$;

- *парную относительную вариограмму* (pairwise relative variogram):

$$\gamma_{\mathrm{PR}}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}\frac{\left[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{h})\right]^2}{\left[\dfrac{Z(\boldsymbol{x}_i) + Z(\boldsymbol{x}_i + \boldsymbol{h})}{2}\right]^2},$$

  когда нормировка на квадрат среднего производится для каждой конкретной пары данных отдельно.

4.3. Построение вариограммы

Описанные выше моменты первого и второго порядков служат для анализа пространственной корреляции данных. Для выявления пространственной структуры используется несколько различных инструментов: вариограммы по направлениям, вариограммные поверхности, вариограммные облака. Как уже указывалось, значения мер пространственной корреляции вычисляются с использованием формулы для статистической несмещенной оценки математического ожидания (среднего). В данном случае главный вопрос заключается в выборе набора пространственных ориентаций и образовании пар для вычисления, чтобы для каждой выбранной ориентации было количество пар, достаточное для получения статистически достоверной оценки среднего.

Пространственная ориентация задается вектором $\boldsymbol{h}$, определяемым длиной (лагом) и направлением. Количество и размер лагов определяются конкретными данными — важно, чтобы было несколько лагов на росте вариограммы и несколько лагов, когда она достигает некоторого уровня, сравнимого со значением априорной вариации, и перестает расти. Если значение вариограммы не перестает расти, это может означать, что для данных не выполнена гипотеза о стационарности второго порядка. Проблема нестационарности более подробно обсуждается в Разделе 4.9.

При подозрении о наличии различий в пространственной структуре в зависимости от направления φ рассчитываются экспериментальные вариограммы по направлениям (directional variogram). Число направлений обычно определяется количеством данных. Для получения общего представления о наличии анизотропии достаточно двух взаимно перпендикулярных направлений. Для более точного моделирования анизотропной вариограммы удобно иметь 6-8 направлений (см. рис. 4.6). Выбор направления расчета вариограммы чрезвычайно важен для четкого выявления корреляционной структуры.

Исходные данные для анализа обычно произвольным нерегулярным образом распределены по области, поэтому трудно предположить, что удастся набрать достаточное количество пар точек измерений, разделенных точно зафиксированными расстояниями в заданном направлении. Чтобы преодолеть эту проблему, используют допущение по разбросу значения расстояния лага (lag tolerance) и угла раствора вокруг направления (direction tolerance). Допуск расстояния лага Δh (lag tolerance) определяет отклонение расстояния в парах от значения расстояния лага h. При $\Delta h = h/2$ все

данные будут учитываться при расчете вариограммы и каждая точка попадет хотя бы в один лаг. Если $\Delta h < h/2$, то часть точек может не попасть ни в один лаг из-за ограничения размера допуска расстояния в лаге. Если $\Delta h > h/2$, то некоторые данные могут учитываться при расчете значения вариограммы для нескольких лагов. Такое перекрытие лагов бывает полезно при малом количестве данных.

Угол раствора $\Delta\varphi$ (direction tolerance) вокруг направления φ позволяет выявлять узконаправленные анизотропные корреляции. Ширина полосы b_w (bandwidth) сужает область поиска на больших расстояниях, ограничивая угол раствора $\Delta\varphi$. Когда угол раствора равен 90°, вариограмма становится обобщенной по всем направлениям (omnidirectional). Такая вариограмма используется, если никакая анизотропия в данных не обнаружена или если анизотропией решено пренебречь, например из-за малого количества данных измерений.

На рис. 4.2 представлена схема параметров для вычисления вариограммы в рамках одного направления φ: h, Δh, $\Delta\varphi$, b_w. Можно представить, что такие сектора (как на рис. 4.2) перемещаются по области данных от одной точки к другой для учета всех пар точек, которые ранжируются по расстоянию между ними и попадают в тот или иной лаг для рассматриваемого направления.

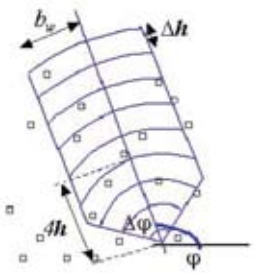
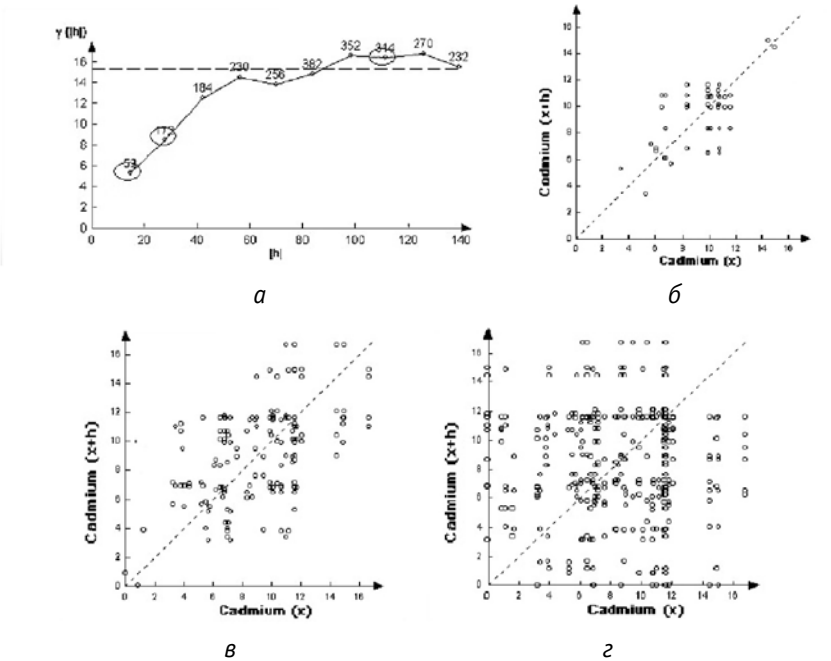


Рис. 4.2. Параметры для расчета вариограммы:
φ — угол направления вариограммы; $\Delta\varphi$ — угол раствора; h — лаг; Δh — разброс лага; b_w — (полу)ширина полосы; квадратиками помечены точки измерений, которые используются при вычислении вариограммы

Вариограмма, рассчитанная по схеме, изображенной на рис. 4.2, приведена на рис. 4.3*а*. На графике вариограммы указано число пар для каждого лага. Возрастание вариограммы с расстоянием лага указывает на наличие корреляции между значениями в парах. Скорость роста вариограммы с рас-

стоянием лага характеризует величины пространственной корреляции. Постоянное значение вариограммы для больших расстояний лага показывает отсутствие корреляции между значениями в парах. Это можно проиллюстрировать диаграммами разброса значений в парах лага (lag scatterplot) для различных лагов, они уже приводились в разделе 2.6. Выберем три лага вариограммы (на рис. 4.3*а* — 1-й, 2-й и 8-й). Из рис. 4.3*б* для 1-го лага видно, что значения в парах сгруппированы вдоль диагонали графика — это означает хорошую корреляцию. Однако из-за малого расстояния между точками 1-го лага количество точек для расчета значения вариограммы мало — 52 пары. Для 2-го лага с вдвое бóльшим расстоянием количество пар точек возрастает до 172, что значительно повышает статистическую репрезентативность значения вариограммы. Значения в парах для 2-го лага также имеют высокую корреляцию (рис. 4.3*в*) в среднем, хотя видно, что разброс значений в парах возрастает. В 8-й лаг вошли пары точек на большом расстоянии (115). На рис. 4.3*г* видно, что значения в парах для 8-го лага не коррелированны — точки разбросаны по всему графику и не группируются вдоль диагонали, как для 1-го и 2-го лагов.

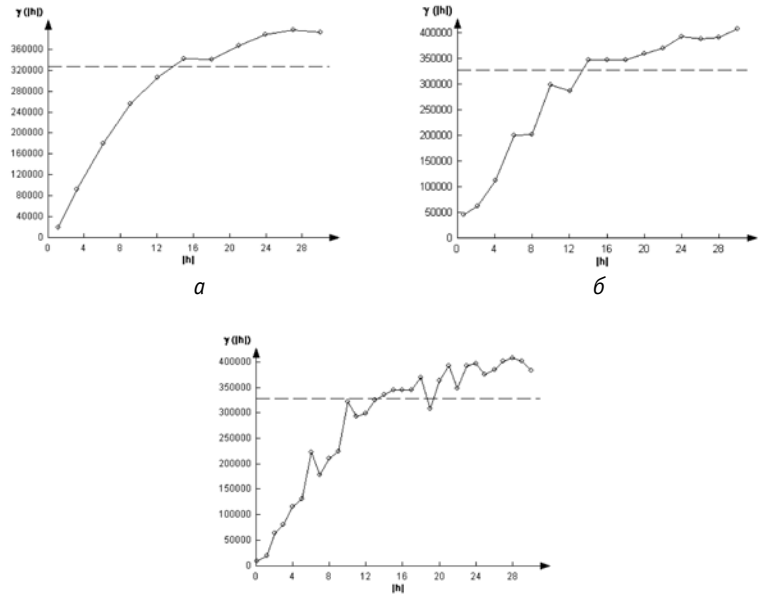


а *б*

в *г*

Рис. 4.3. Вариограмма, количество пар в лагах (*а*) и диаграммы разброса пар для 1-го лага (*б*), 2-го лага (*в*) и 3-го лага (*г*)

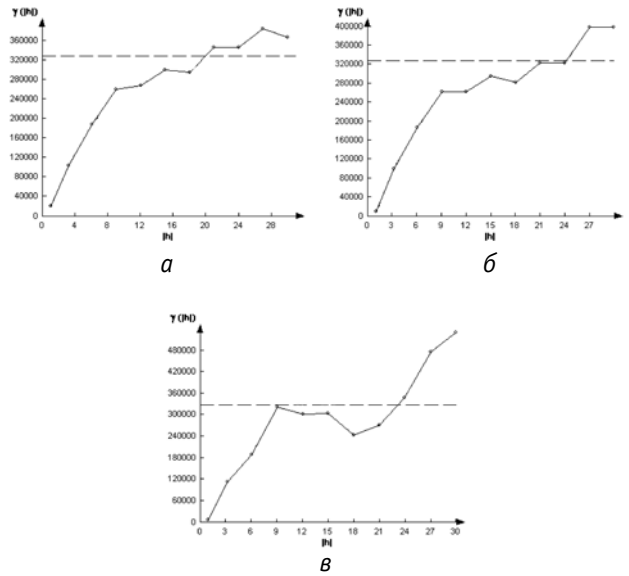
Выбор размера отклонений лага и угла направления зависит от количества данных. Если данных много и они распределены плотно, разбросы могут быть небольшими. Важно только следить, чтобы число пар, попавших в каждый сектор, было достаточным. При малом количестве данных допустимо использование перекрывающихся секторов. Они не нарушают общую структуру вариограммы, а делают ее более гладкой, удобной для последующего моделирования. Пример экспериментальных вариограмм, построенных на основе одних и тех же данных, для различных лагов представлен на рис. 4.4. Можно видеть, что с уменьшением значения лага экспериментальная вариограмма становится менее гладкой. Обычно используют равные по длине лаги, что ведет к равномерному их распределению. Однако в особых случаях можно использовать и неравномерно распределенные лаги [Flamm et al., 1994]. Поскольку расчет вариограммы весьма чувствителен к выбору длины лага, их значение является принципиальным для дальнейшего моделирования пространственной корреляции. На практике используют интерактивные программы подбора, такие как описанные в [Pannatier, 1996; Kanevski, Maignan, 2004].



а *б* *в*

Рис. 4.4. Вариограммы, рассчитанные с лагом различной длины: 3 (*а*), 2 (*б*), 1 (*в*)

Раствор угла направления также сильно влияет на поведение вариограммы. С его помощью можно ограничить разброс направлений пар точек и таким образом выявить узконаправленную корреляцию. На рис. 4.5 приведены вариограммы, рассчитанные для различных значений раствора угла: 45°, 30° и 15°. Можно видеть, что с уменьшением раствора поведение вариограммы становится менее гладким, если направление не полностью соответствует направлению доминирующей пространственной непрерывности. На рис. 4.5 также видно, что с уменьшением угла раствора число пар, указанное для каждого лага, уменьшается.



а *б* *в*

Рис. 4.5. Вариограммы по одному направлению для различных углов раствора $\Delta\varphi$: 45° (*а*), 30° (*б*), 15° (*в*)

Для визуализации вариограмм можно использовать графики, как, например, на рис. 4.6, но для изображения и исследования пространственной анизотропии более удобны двумерные изображения. Одним из них является *вариограммная роза* [Chernov et al., 1998] (рис. 4.7). Роза имеет вид лепестков, представляющих вариограммы по направлениям. Вариограммная роза симметрична относительно центра в силу свойств симметрии моментов второго порядка. По вариограммной розе можно построить изолинии значений вариограммы методом линейной интерполяции на основе триангуляции (см. рис. 4.8). Изолинии вариограмм хорошо визуализируют анизотропную корреляционную структуру.

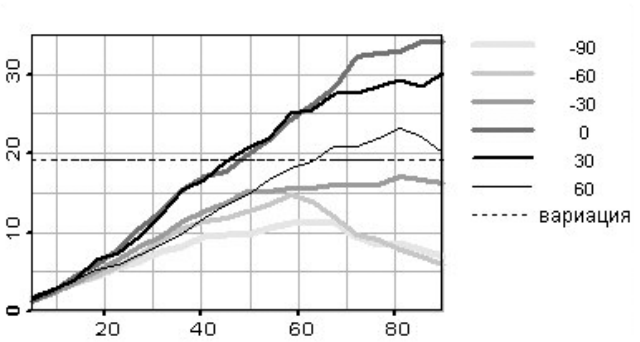


Рис. 4.6. Вариограммы по шести направлениям и уровень априорной вариации $C(0)$

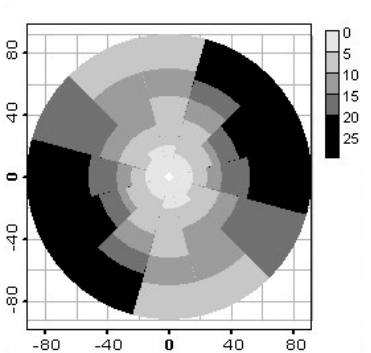


Рис. 4.7. Вариограммная роза

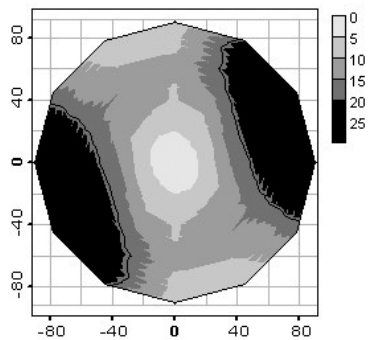


Рис. 4.8. Сглаженные изолинии вариограммной розы (светлым контуром отмечен уровень априорной вариации)

Другим инструментом, который дает представление о поведении пространственной структуры в целом, является *вариограммная поверхность* (variogram surface) (рис. 4.9) [Pannatier, 1996]. Для построения вариограммной поверхности вектор $\boldsymbol{h}$ представляется не в полярном виде, как для вариограммной розы (расстояние и направление), а в виде проекций лага на оси координат Δx и Δy. Вариограммная поверхность представляет собой поверхность значений, вычисленных на регулярной сетке в пространстве лагов по формуле вариограммы. При вычислении (наборе пар) вариограммные координаты, естественно, тоже берутся с разбросом (lag tolerance). Вариограммная поверхность позволяет сразу увидеть анизотропию и определить приоритетные направления для построения вариограмм. Следует отметить, что вариограммная поверхность обладает центральной симметрией относительно точки (0, 0).

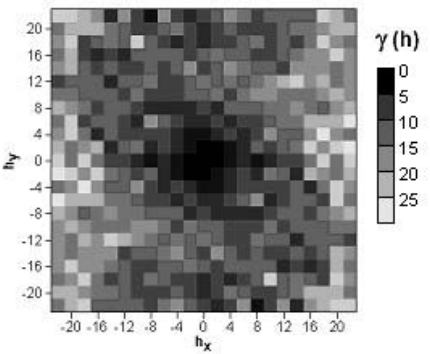


Рис. 4.9. Вариограммная поверхность для ^{137}Cs

Еще одним инструментом пространственного корреляционного анализа является *вариограммное облако* (variogram cloud) (рис. 4.10). Это диаграмма разброса квадратов разности значений для всех пар в зависимости от расстояния между точками в паре. Такая диаграмма помогает распознать пары с большим значением квадрата разности значений, поскольку они вносят существенный вклад в значение экспериментальной вариограммы. Присутствие пар, дающих необоснованно большие значения для малых лагов, помогает выявлять крайние экстремальные значения — выбросы (outliers). Вариограммное облако также помогает определить оптимальный лаг для вычисления вариограммы. Вариограммное облако может быть построено для любого направления и раствора угла. Следует отметить, что вариограммное облако также обладает центральной симметрией относительно точки (0, 0).

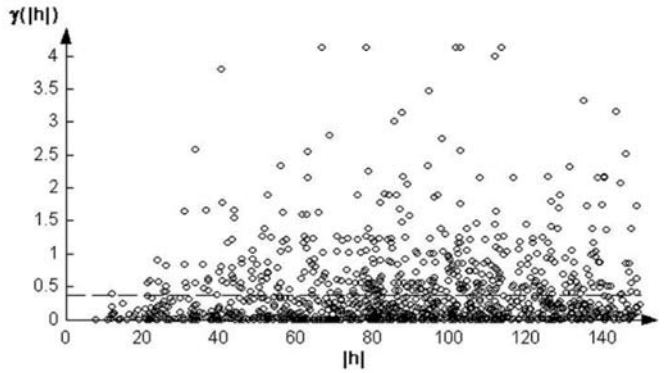
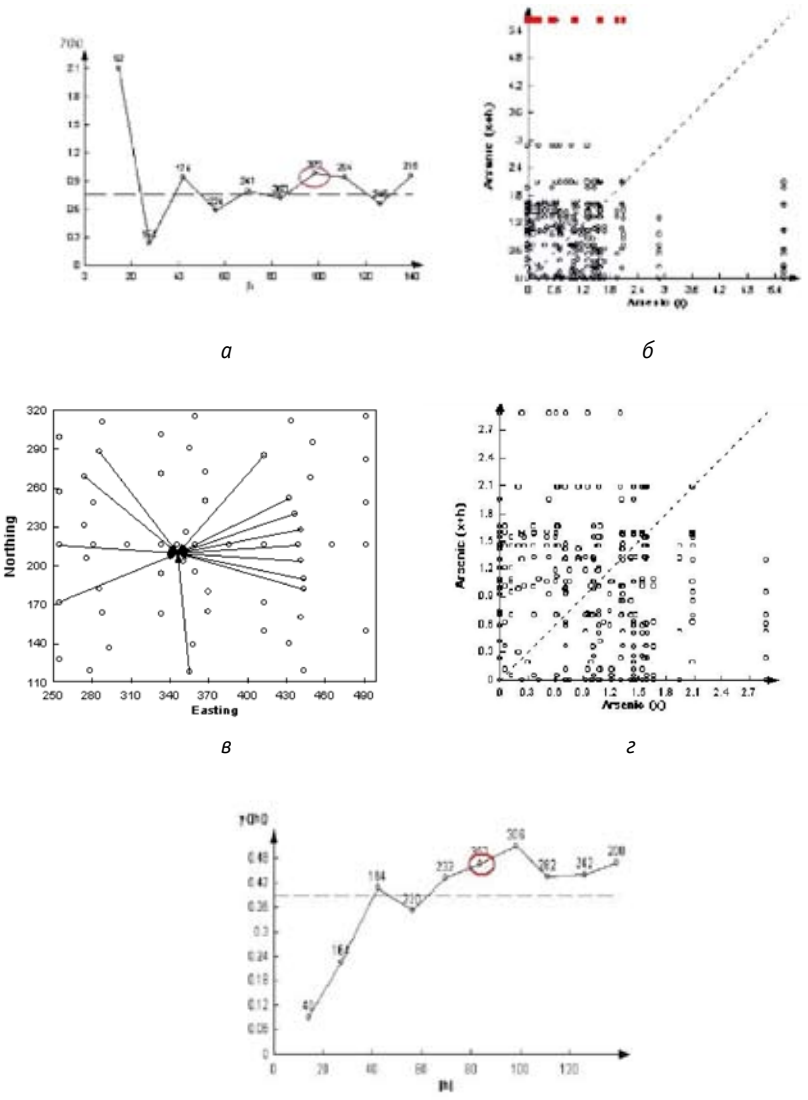


Рис. 4.10. Вариограммное облако

На практике поиск пространственной корреляции при помощи вариограммы является трудоемким процессом. Зашумленность данных или присутствие множества предельных значений затрудняет выявление пространственной структуры.

Приведем пример влияния на вариограмму единичного предельного значения — экстремально высокого измерения. Вариограмма, построенная с учетом всех данных (рис. 4.11*а*), казалось бы, демонстрирует отсутствие структуры в данных. Рассмотрим более пристально один из лагов (7-й лаг с расстоянием в парах 100). На диаграмме разброса пар (рис. 4.11*б*) видно экстремально высокое значение (больше 5), выделенное в кружке, которое значительно отличается от основной массы данных. Очевидно, что это предельное значение дает наибольший вклад в разницу между парами точек лага. На рис. 4.11*в* нанесены пары точек, включающие крайне высокое значение, выделенные на рис 4.11*б*. Если исключить точку с предельным значением из рассмотрения и не учитывать ее при вычислении вариограммы, то вклады всех остальных данных в вариограмму будут сопоставимы (рис. 4.11*г*). Таким образом, без учета предельного значения можно выявить пространственную корреляцию в данных (за исключением выколотого максимума) до расстояния 100 (рис. 4.11*д*).



а *б* *в* *г* *д*

Рис. 4.11. Влияние экстремального значения на вариограмму:

а — вариограммы с учетом крайне высокого значения; *б* — диаграммы разброса пар для 7-го лага вместе с крайним значением; *в* — местоположение пар точек с крайним значением; *г* — диаграммы разброса пар для 6-го лага вместе с крайним значением; *д* — вариограммы без учета крайнего значения

Упражнение 4.5. *Для построения анизотропной модели вариограмм производится расчет вариограммы по нескольким направлениям (углам φ). Чему равен раствор угла $\Delta\varphi$, если вариограммы построены в шести направлениях без перекрытия одинаковых секторов (т. е. без повторного учета одних и тех же точек в разных направлениях)?*

4.4. Моделирование вариограммы

Как будет показано в Главе 5, значения вариограммы напрямую входят в систему уравнений, решаемую для получения оценки методом кригинга. Чтобы составить эту систему, требуются значения вариограммы для любых пространственных ориентаций. Для этого используют теоретическую модель вариограммы, специальным образом построенную на основе экспериментальной вариограммы.

С другой стороны, система уравнений кригинга имеет единственное решение при несингулярности матрицы системы, что эквивалентно положительной определенности ковариации, что, в свою очередь, эквивалентно отрицательной определенности вариограммы [Armstrong, 1984; Christakos, 1984]:

$$-\sum_j \sum_i \boldsymbol{b}_i \boldsymbol{b}_h \gamma\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) \geq 0 \ \text{ при } \ \sum_i b_i = 0,$$

где $\boldsymbol{x}_i$ — конечное число точек в пространстве ($\boldsymbol{x}_i$: i = 1, 2, 3, …, m); b_i — действительные числа (b_i, b_j: i, j = 1, 2, 3, …, m).

Чтобы избежать трудоемкой процедуры доказательства отрицательной определенности функции, используют специальные модели, для которых это свойство уже доказано. Остается только выбрать модель (или линейную комбинацию моделей) и подобрать параметры, делающие ее подходящей для экспериментальной вариограммы, рассчитанной по данным.

Ниже приведены наиболее известные типы моделей вариограмм, удовлетворяющие требованию отрицательной определенности [Goovaerts, 1997].

Модель наггет:

$$\gamma(h) = \begin{cases} 0, & h = 0, \\ c_0, & h \neq 0, \end{cases}$$

Константа $c_0 = C(0)$ носит название *наггет* (nugget), что означает самородок. Это понятие было заимствовано из золотодобычи и означает некоррелированный случайный характер. Наличие у данных вариограммы только типа наггет означает отсутствие пространственной корреляции. Данные в этом случае распределены абсолютно случайно (pure nugget). Отсутствие корреляции в данных может иметь следующие причины: мелкомасштабная вариабельность (меньше, чем расстояние между измерениями), ошибки измерений, ошибки в определении местоположений точек.

Сферическая модель:

$$\gamma(h) = \begin{cases} c_0 + c\left[\dfrac{3h}{2a} - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3\right] \text{ для } h \le a, \\ c_0 + c \text{ для } h > a, \end{cases}$$

где a — действительный радиус корреляции (range), на рис. 4.12 $a = 40$. Радиус корреляции означает, что данные, находящиеся на расстояния a и ближе, коррелированны, а находящиеся на расстоянии больше a — не коррелированны.

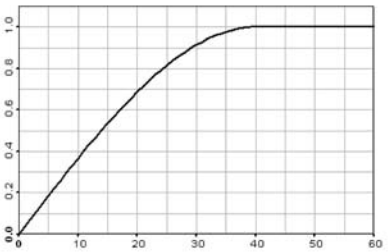


Рис. 4.12. Сферическая модель

Для сферической модели $\gamma(a) = C(0) = c_0 + c$ — плато (sill). Эта модель ведет себя линейно вблизи нуля.

Экспоненциальная модель:

$$\gamma(h) = \begin{cases} 0, h = 0, \\ c_0 + (c - c_0)\left[1 - \exp\left(\dfrac{-3h}{a}\right)\right], h \ne 0, \end{cases}$$

где a — эффективный радиус корреляции (range), на рис. 4.13 $a = 40$. На этом расстоянии значение вариограммы достигает 95% плато. Данная модель достигает плато асимптотически.

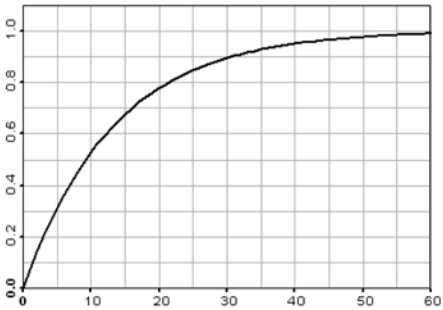


Рис. 4.13. Экспоненциальная модель

Гауссова модель:

$$\gamma(h) = c_0 + c\left[1 - \exp\left(\frac{-3h^2}{a^2}\right)\right],$$

где a — эффективный радиус корреляции (range), на рис. 4.14 a = 40. На этом расстоянии значение вариограммы достигает 95% плато.

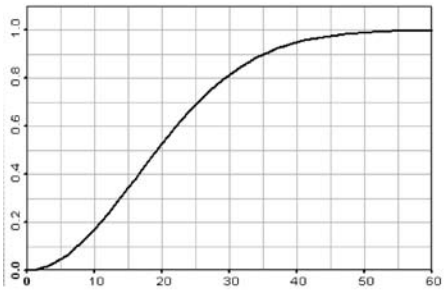


Рис. 4.14. Гауссова модель

Отличительной чертой этой модели является ее гладкость: параболическое поведение вблизи нуля и асимптотическое приближение к плато. Случайная компонента в корреляции данных при гладком гауссовом поведении вариограммы обычно обусловлена ошибками измерений.

Степенная модель:

$$\gamma(h) = \begin{cases} 0, h = 0, \\ ch^{\alpha}, h \neq 0, \end{cases}$$

где α — степень. Эта модель (рис. 4.15) при $\alpha = 1$ иногда выделяется в отдельный тип и называется линейной.

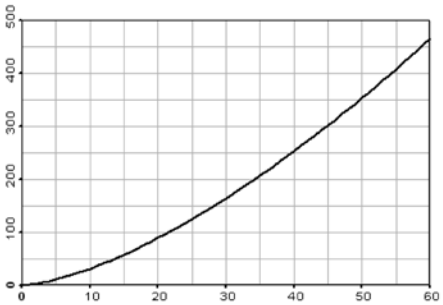


Рис. 4.15. Степенная модель

Данная модель отражает корреляцию на всех расстояниях, поэтому для нее радиус корреляции стремится к бесконечности. В случае степенной модели вариограммы не выполняется предположение о стационарности второго порядка, это соответствует модели статистического самоподобия данных.

Степенная и линейная модели могут использоваться и в обрезанном виде в предположении о стационарности второго порядка. В этом случае можно говорить о существовании радиуса корреляции $\alpha < \infty$ и модель принимает вид

$$\gamma(h) = \begin{cases} 0, h = 0, \\ ch^{\alpha}, 0 < h \le a, \\ ca^{\alpha}, h > a. \end{cases}$$

Степенная модель кусочно-интегрируема, с особой точкой при выходе на плато (обрезанная).

Периодическая модель (hole effect):

$$\gamma(h) = 1 - \cos\left(\frac{2\pi h}{a}\right),$$

где a — период периодической структуры, эквивалентный радиусу корреляции.

Эта модель (рис. 4.16) используется для периодических структур. Периодическая модель работает только в одном направлении.

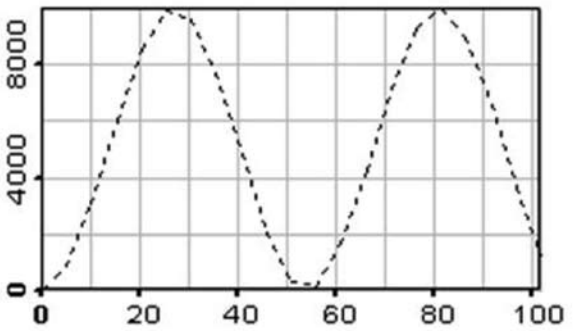


Рис. 4.16. Периодическая модель

Затухающая периодическая модель (dampened Hole effect model):

$$\gamma(h) = 1 - \exp\left(-\frac{h}{a}\right)\cos\left(\frac{2\pi h}{a}\right).$$

Эта модель (рис. 4.17) представляет собой произведение экспоненциальной модели ковариации и периодической функции.

Затухающая периодическая структура встречается чаще, чем чисто периодическая.

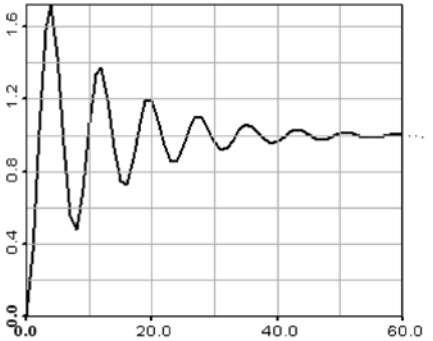


Рис. 4.17. Затухающая периодическая модель

Кубическая модель (Cubic mode):

$$\gamma(h) = \begin{cases} c_0 + (c - c_0)\left[7\left(\frac{h}{a}\right)^2 - \frac{35}{4}\left(\frac{h}{a}\right)^3 + \frac{7}{2}\left(\frac{h}{a}\right)^5 - \frac{3}{4}\left(\frac{h}{a}\right)^7\right], 0 \le h < a, \\ c, a \le h. \end{cases}$$

Эта модель (рис. 4.18) состоит из линейной комбинации степенных моделей и ограничена постоянным значением плато c за пределами радиуса корреляции a.

Данная модель может быть использована в одно-, дву- и трехмерных случаях.

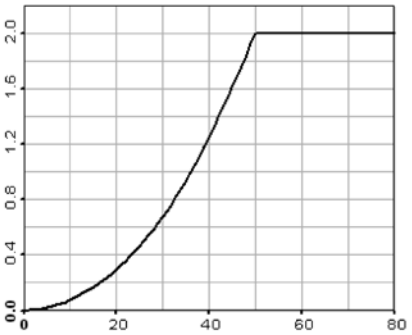


Рис. 4.18. Кубическая модель

Пентасферическая модель (Pentaspherical model):

$$\gamma(h)=\begin{cases} c_0+(c-c_0)\left[\frac{15}{8}\frac{h}{a}-\frac{5}{4}\left(\frac{h}{a}\right)^3+\frac{3}{8}\left(\frac{h}{a}\right)^5\right], 0\le h<a, \\ c, a\le h. \end{cases}$$

Эта модель является вариантом сферической модели более высокого порядка. Она также ограничена значением плато c за пределами радиуса корреляции a, на рис. 4.19 a = 50. В отличие от классической сферической модели данная обладает большим градиентом на участке роста.

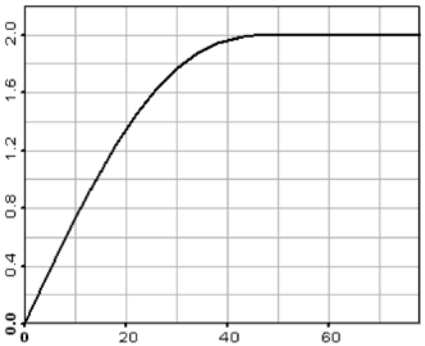


Рис. 4.19. Пентасферическая модель

Кубическая и пентасферическая модели используются достаточно редко и не входят во многие распространенные геостатистические программы.

Во всех моделях вариограмм плато c может быть только положительным.

Упражнение 4.6. *Перечислить стационарные и нестационарные модели вариограмм.*

На практике наиболее часто применяются сферическая, экспоненциальная и гауссова модели (или их комбинации). При использовании кригинга важно, чтобы вариограммные модели были стационарными, поскольку кригинг предполагает стационарность среднего (см. Главу 5).

Графическое изображение наиболее часто использующихся в геостатистике типов моделей вариограмм приведены на рис. 4.20. Изображенные модели имеют одинаковые параметры: $c_0 = 0$, $c = 1$, $a = 10$.

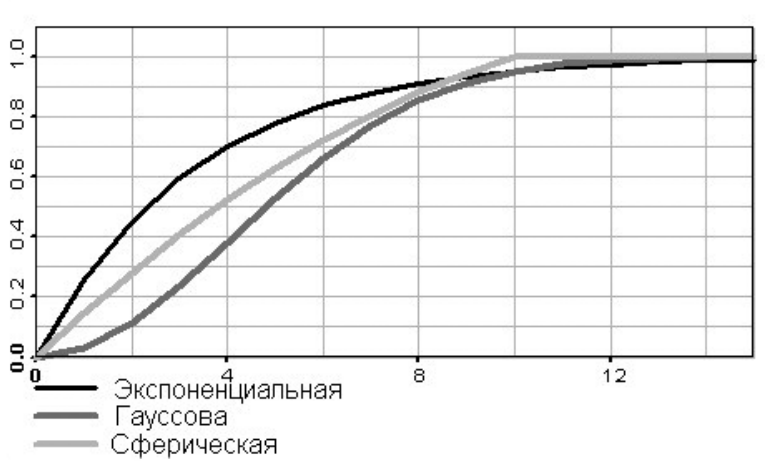


Рис. 4.20. Основные типы стационарных моделей вариограмм при $a = 10$, $c = 1$, $c_0 = 0$

Во многих случаях используются линейные комбинации моделей различных типов. Их плато и радиусы могут быть различными. Так, наличие суммы моделей с разными радиусами корреляции свидетельствует о присутствии *гнездовой структуры* (nested structure) данных:

$$\gamma_{\text{nested}}(h) = \sum |\lambda_l| \gamma_l(h).$$

Как уже указывалось, построение вариограммы и подбор для нее теоретической модели — весьма трудоемкий процесс, требующий некоторых навыков и опыта.

При непосредственном подборе формы и параметров модели необходимо оценить ее качество, т. е. близость к экспериментальной вариограмме. Одним из доступных подходов при этом является визуальное сходство. В этом случае многое зависит от опыта эксперта, проводящего моделирование. Хорошим подспорьем может оказаться набор специальных критериев. Существует целый спектр критериев качества соответствия как общего назначения, так и специально разработанных для подбора параметров и модели вариограмм. На основе некоторых из них создаются программы автоматического подбора модели вариограмм. Однако практика показывает, что автоматическая процедура не всегда дает корректные результаты, особенно в случаях негладкой вариограммы при сильном влиянии зашумленных и экстремальных значений. Тогда приходится полагаться на мнение эксперта и ручной подбор параметров модели.

Индексы качества аппроксимации являются функциями разницы между значением экспериментальной вариограммы $\gamma(h_i)$ для i-го лага h и значением модели вариограммы $\gamma^*(h_i, \lambda)$ для лага h_i и набора параметров модели λ (c_0, c, a). Приведем несколько критериев, которые применяются в геостатистике.

Индекс взвешенных наименьших квадратов:

$$I_{wls} = \sum_{i=1}^{k} W(i)\left[\gamma^*(h_i,\lambda) - \gamma(h_i)\right]^2.$$

Суммирование проводится по количеству лагов экспериментальной вариограммы k. Если веса $W(i) = 1$, то этот метод вырождается в обычный метод наименьших квадратов, который предполагает, что невязки между экспериментальной вариограммой и модельными значениями независимы, нормально распределены и имеют одну и ту же вариацию. Это достаточно сильное предположение, оно не всегда соблюдается на практике.

Индекс Кресси [Cressie, 1985]:

$$I_C = \sum_{i=1}^{k} \frac{N(h_i)}{\left[\gamma^*(h_i,\lambda)\right]^2}\left[\gamma^*(h_i,\lambda) - \gamma(h_i)\right]^2,$$

где $N(h_i)$ — число пар точек, по которым вычислялось значение для лага.

Веса для индекса Кресси зависят от количества пар в лаге вариограммы, которое определяется при построении экспериментальной вариограммы (см. раздел 4.3).

Индикатор Кресси может иметь модифицированный вариант [Zhang et al., 1995]:

$$I_C = \sum_{i=1}^{k} \frac{N(h_i)}{h_i^2} \left[\gamma^*(h_i, \lambda) - \gamma(h_i) \right]^2.$$

Веса для модифицированного индекса зависят напрямую от расстояния между точками в лаге. Модифицированный индекс дает близкие индексу Кресси результаты на малых расстояниях лага. Для больших расстояний модифицированный индекс более устойчив по сравнению с оригинальным индексом Кресси.

Индекс качества подбора (indicative goodness of fit), используемый в программе VARIOWIN [Pannatier, 1996]:

$$I = \frac{1}{P} \sum_{k=1}^{P} \sum_{i=0}^{n(k)} \frac{\dfrac{N(h_i)}{\sum_{i=0}^{n(k)} N(h_i)}}{\dfrac{h_i}{h_{\max}(k)}} \left[\frac{\gamma(h_i) - \gamma^*(h_i, \lambda)}{\sigma^2} \right]^2,$$

где $h_{\max}(k)$ — максимальная длина лага для k-го направления; $N(h_i)$ — число пар точек, по которым вычислялось значение для лага; n — число лагов; σ^2 — вариация оценки вариограммы; P — число направлений, которые участвуют в подборе параметров модели.

Значение этого критерия стремится к нулю при улучшении качества подбора. Поэтому критерий может использоваться как для вариограммы в одном направлении, так и для одновременного подбора параметров вариограмм по нескольким направлениям.

Информационный критерий Акайк [Xiaodong et al., 1996]:

$$I_{\text{AIK}} = k \ln \left(\frac{\sum_{i=1}^{k} \left[\gamma(h_i) - \gamma^*(h_i) \right]^2}{k} \right) + 2p,$$

где p — число параметров в модели.

Информационный критерий Акайк учитывает также сложность модели вариограммы — модели с большим числом параметров (гнездовых структур) при том же качестве соответствия получают более высокое значение критерия в соответствии с принципом Оккама.

4.5. Поведение вариограмм на больших расстояниях

Вариограммы приведенных выше типов можно классифицировать по поведению на больших расстояниях. При наличии стационарности второго порядка значение вариограммы на бесконечности равно значению ковариации в нуле (ковариации исходных данных) [Barnes, 1991]. Это характерно для вариограмм сферического, экспоненциального и гауссового типов. Модель вариограммы сферического типа достигает плато на расстоянии a, экспоненциального типа — на расстоянии $3a$, гауссова модель достигает 95%-ного значения плато на расстоянии $a\sqrt{3}$.

Если вариограмма не имеет предела роста на бесконечности, это означает, что ковариации не существует. В этом случае стационарность второго порядка заменяется более слабой внутренней (intrinsic) гипотезой. Бесконечной вариации данных соответствует степенная и линейная модели вариограмм без ограничений на радиус роста, т. е. данные, даже очень удаленные друг от друга, все еще продолжают оказывать взаимное влияние.

Упражнение 4.7. *Каково соотношение между ковариацией на бесконечности и вариограммой в предположении стационарности второго порядка?*
Упражнение 4.8. *Каково соотношение между вариограммой на бесконечности и ковариацией в предположении стационарности второго порядка?*

4.6. Поведение вариограмм вблизи нуля

Вариограммы также различаются по характеру поведения в нуле. Теоретически $\gamma(0) = 0$ независимо от типа вариограммы. Однако очень часто вариограмма имеет скачок в нуле, что и называется *наггет-эффектом*. Такой разрыв вариограммы вблизи нуля моделируется включением соответствующей наггет-составляющей (константы). Эффект можно объяснить, например, присутствием ошибок измерений или вариабельностью данных на более

мелких масштабах. Поскольку структура этих микровариабельностей имеет меньший масштаб, чем масштаб анализируемых данных, они (микровариабельности) проявляются как белый шум.

Как уже указывалось, колеблющееся вокруг некоторой константы значение вариограммы представляет собой *чистый наггет-эффект* (pure nugget effect). При этом $\gamma(0) = 0$ в некоторой окрестности нуля ε, а при $h > \varepsilon\ \gamma(h) = C(0)$, т. е. чистый наггет-эффект соответствует полному отсутствию корреляций.

При *параболическом поведении* вблизи нуля ($\gamma(h) \sim A|h|^2$) вариограмма дважды дифференцируема в нуле. Такое поведение характеризует сильно регулярную структуру, которая соответствует гауссовой модели вариограммы.

При линейном поведении вблизи нуля ($\gamma(h) \sim A|h|$) вариограмма не дифференцируема в нуле, но остается непрерывной при $h = 0$. Этот случай представлен линейной моделью.

4.7. Анизотропия вариограмм

До сих пор мы рассматривали модель вариограммы для одного направления или для *изотропной* вариограммы, которая зависит только от расстояния между точками. При изотропии изолинии вариограммы на вариограммной поверхности или вариограммной розе будут иметь форму окружности. Если вариограмма зависит и от ориентации пары точек в пространстве, то можно говорить о наличии *анизотропии*. Это означает существование структур данных с различными пространственными характеристиками в различных направлениях. В такой ситуации есть два выхода. Первый выход — построить одну модель изотропной вариограммы и при ее использовании всякий раз производить с вектором $\boldsymbol{h}$ преобразования пространства и только после этого подставлять в качестве аргумента величину $|\boldsymbol{h}|$. Второй выход состоит в полномасштабном моделировании анизотропной структуры вариограммы и использовании ее в вычислениях. Для моделирования сложной анизотропии используют гнездовую структуру. Для выбора подхода к моделированию анизотропии следует проверить, к какому типу она относится.

В традиционной геостатистике анизотропию делят на два класса: *геометрическую* и *зонную* (все остальные варианты анизотропии, кроме геометрической). Но с точки зрения подхода к моделированию удобнее подразделять анизотропию вариограмм по основным параметрам, использующимся в моделях: радиусу и плато. Третий параметр — наггет — определяет значение вариограммы в малой окрестности нуля. Если рассматривать такую

анизотропию, то это — различное поведение вариограмм вблизи нуля для разных направлений. Оно может быть вызвано только коррелированностью ошибок измерений. Поэтому на практике такая анизотропия не рассматривается и не моделируется, лишь подбирается одинаковое значение наггета для всех направлений, наиболее подходящее по индикаторам.

В случае анизотропии радиуса вариограммы (ковариации) по различным направлениям имеют одинаковые форму и значения плато, но разные эффективные радиусы корреляции, другими словами, значения вариограммы достигают значения плато на различных расстояниях в зависимости от направления. При этом возможны два случая: геометрическая анизотропия и зонная анизотропия радиуса.

Геометрическая (geometric) *анизотропия*. В этом случае изолинии вариограммы на вариограммной поверхности или вариограммной розе имеют форму эллипса (рис. 4.21). Это означает, что существует положительно определенная матрица $\boldsymbol{B}$ — такая, что $\gamma(\boldsymbol{h}') = \gamma\left(\boldsymbol{h}^T \mathbf{B} \boldsymbol{h}\right)$ — изотропная вариограмма. Преобразование пространства, после которого геометрически анизотропная вариограмма становится изотропной, определяется следующим образом:

$$\boldsymbol{h}' = \begin{bmatrix} 1 & 0 \\ 0 & a_{\min} / a_{\max} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \boldsymbol{h},$$

где $a_{\min}$ — длина малой оси эллипса; $a_{\max}$ — длина основной оси эллипса; θ — угол, определяющий направление основной оси эллипса. Вариограммы по четырем направлениям для случая геометрической анизотропии в горизонтальном направлении (90°) представлены на рис. 4.22. Можно заметить, что графики вариограмм по направлениям следуют в последовательности углов направлений.

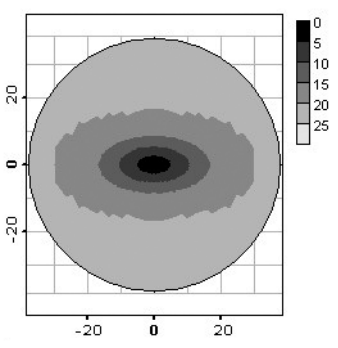


Рис. 4.21. Геометрическая анизотропия: вариограммная роза

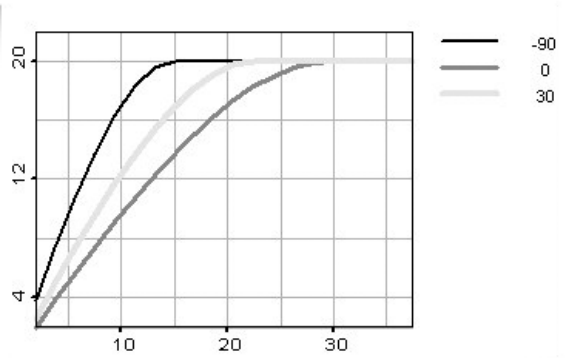


Рис. 4.22. Геометрическая анизотропия: вариограммы по направлениям

Негеометрическая анизотропия радиуса (non-geometric range anisotropy). В этом случае изолинии на вариограммной поверхности или вариограммной розе образуют форму, отличную от эллипса (пример на рис. 4.23). При моделировании такого рода вариограмм удобнее пользоваться гнездовыми структурами, хотя при желании можно построить и преобразование пространства [Zimmerman, 1993]. Для такого преобразования сначала нужно построить и запомнить функцию зависимости радиуса корреляции от угла $R(\varphi)$, потом необходимо построить изотропную модель вариограммы, взяв за основу одно из направлений. Например, это может быть направление θ, в котором радиус корреляции максимален и равен R_θ. Тогда при вычислении вариограммы достаточно воспользоваться преобразованием пространства:

$$\boldsymbol{h}' = \begin{bmatrix} R_\theta / R(\varphi) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \boldsymbol{h},$$

где θ — угол, определяющий направление, в котором достигается максимум значения модуля $|\boldsymbol{h}'|$ (иначе говоря, это угол, для которого вектор $\boldsymbol{h}$ имеет проекцию максимальной длины, т. е. коллинеарный вектору $\boldsymbol{h}$). Пример, иллюстрирующий негеометрическую анизотропию, представлен на рис. 4.24: вариограммы по четырем направлениям. Заметим, что в отличие от геометрической анизотропии при негеометрической анизотропии радиуса вариограммы по направлениям не следуют в порядке углов направлений. Негеометрическая анизотропия радиуса является одним из вариантов зонной анизотропии.

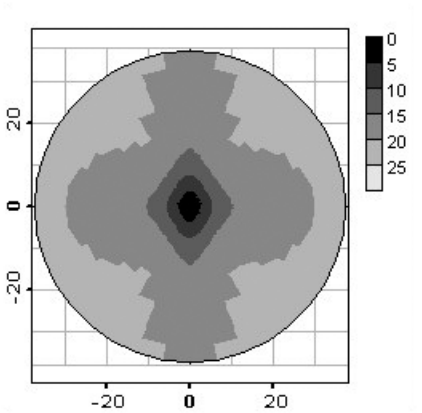


Рис. 4.23. Негеометрическая (зонная) анизотропия радиуса: вариограммная роза

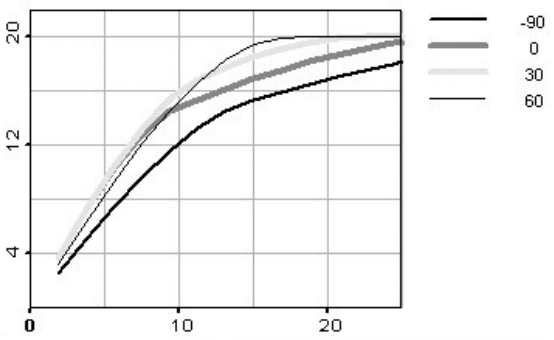
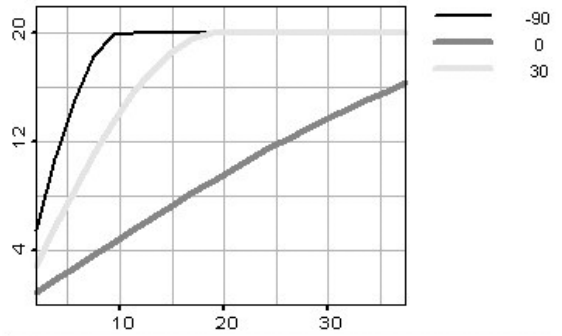


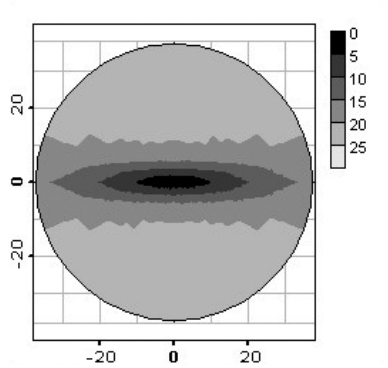
Рис. 4.24. Негеометрическая (зонная) анизотропия радиуса: вариограммы по направлениям

Другой случай зонной анизотропии — *анизотропия плато* (sill anisotropy). Здесь для различных направлений различаются значения плато (рис. 4.25). Наличие плато у вариограммы означает, что процесс не только удовлетворяет внутренней гипотезе, но и обладает стационарностью второго порядка. Тогда, учитывая поведение вариограммы на больших расстояниях, для любого фиксированного $\boldsymbol{h}$ можно написать

$$\lim_{\alpha\to\infty}\gamma(\alpha\boldsymbol{h}) = C(0) - \lim_{\alpha\to\infty}C(\alpha\boldsymbol{h}).$$



а



б

Рис. 4.25. Пример зонной анизотропии плато:
а — вариограммы по направлениям; *б* — вариограммная роза

Очевидно, что такое равенство справедливо для любого направления, где у вариограммы обнаружено плато. А так как мы имеем различные значения плато для различных направлений, существуют хотя бы два направления $\boldsymbol{h}_1$ и $\boldsymbol{h}_2$ такие, что $\lim_{\alpha\to\infty} C(\alpha\boldsymbol{h}_1) \neq \lim_{\alpha\to\infty} C(\alpha\boldsymbol{h}_2)$, т. е. по крайней мере в одном направлении $\lim_{\alpha\to\infty} C(\alpha\boldsymbol{h}) \neq 0.$ Таким образом, если значение плато меняется в зависимости от направления, то, следовательно, существует хотя бы одно направление, на котором корреляция между значениями не пропадает ни при каких расстояниях.

При обнаружении анизотропии плато можно сделать два предположения о характере процесса: либо стационарность второго порядка присутствует,

но радиус корреляции в одном из направлений больше области исследования и настоящее значение плато еще не достигнуто, либо стационарности нет из-за присутствия тренда. Хотя второй вариант более правдоподобен, на практике чаще делается первое предположение. В этом случае для получения изотропной вариограммы можно попробовать следующее преобразование:

$$\boldsymbol{h}' = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \boldsymbol{h},$$

где θ — угол, перпендикулярный направлению, в котором вариограмма имеет наибольшее значение плато. Можно использовать вариограммную модель гнездовой структуры.

Упражнение 4.9. *На рис. 4.26 изображены вариограммы по четырем направлениям: 0°, 30°, 60° и 90°. Найти радиус корреляции в направлениях 270°, 240°, 210° и 180°.*

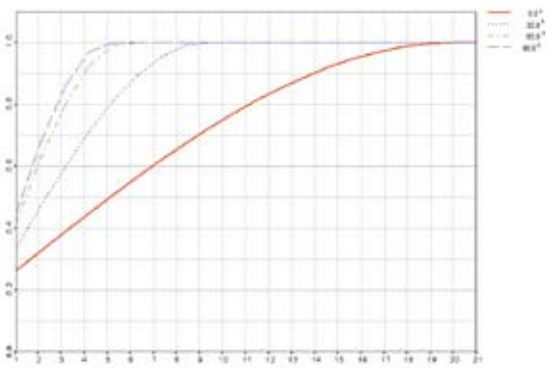
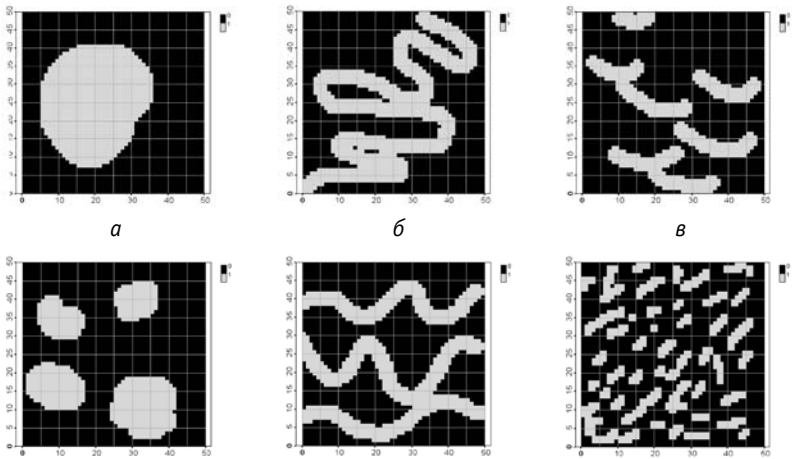


Рис. 4.26. К упражнению 4.9

В случае использования гнездовой структуры для моделирования анизотропной вариограммы необходимо строить ее так, чтобы все *i*-е элементы (для всех направлений) имели одинаковую модель (сферическую, гауссову и т. п.) и одинаковое значение плато, но радиусы могут быть любые, в том числе и такие большие, чтобы скрывать анизотропию плато. В гнездовых структурах по всем направлениям должно быть одинаковое число элементов. Преобразования пространства делаются отдельно для каждого элемента, а в конечном счете опять получается линейная комбинация моделей.

Упражнение 4.10. Вариограммы для различных геологических структур

Поставьте пространственным (геологическим) образам на рис. 4.27 в соответствие вариограммы на рис. 4.28 и вариограммные розы на рис. 4.29. Что можно сказать о корреляционных структурах приведенных образов?

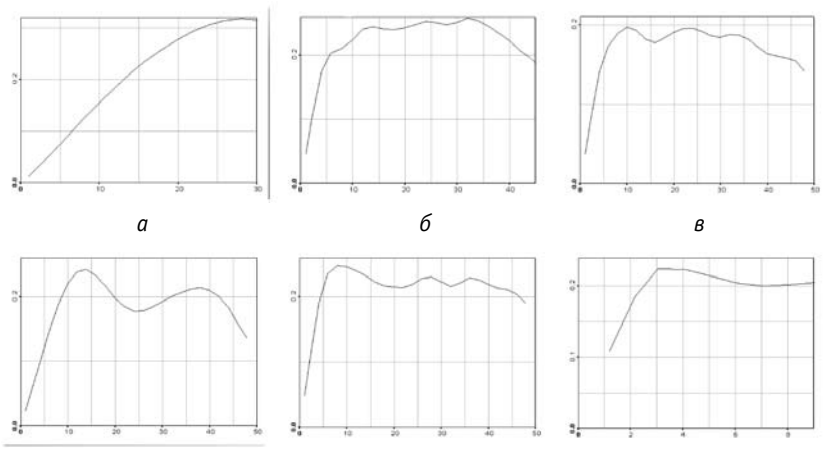


а *б* *в*

г *д* *е*

Рис. 4.27. Образы наблюдаемой величины:

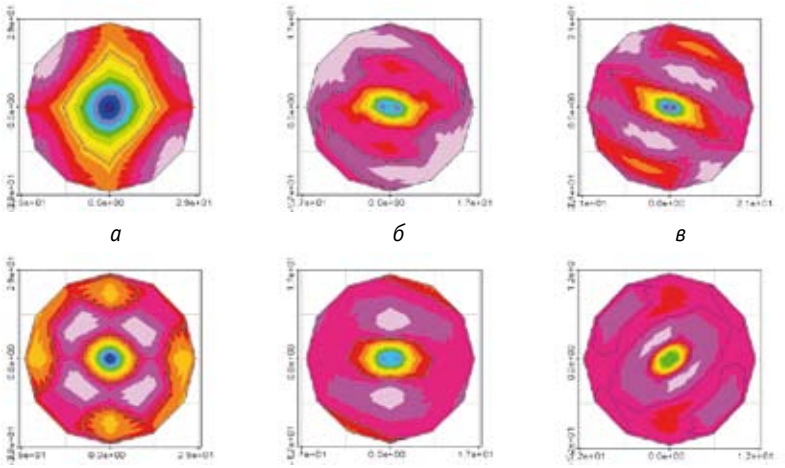
а — одиночное тело; *б* — извилистое русло; *в* — эоловые дюны; *г* — множественные тела; *д* — параллельные русла; *е* — вкрапления



а *б* *в*

г *д* *е*

Рис. 4.28. Вариограммы, соответствующие образам на рис. 4.27



а *б* *в*

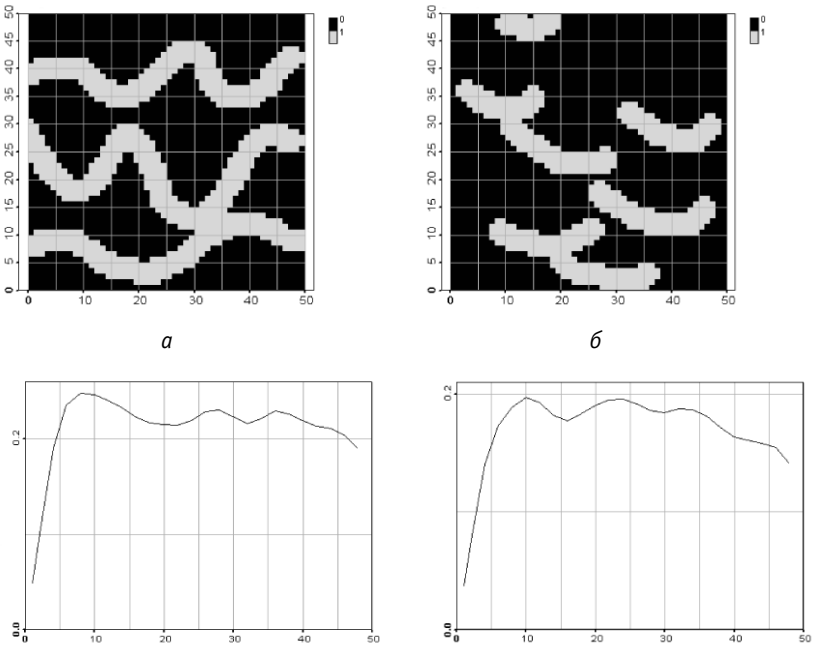
г *д* *е*

Рис. 4.29. Вариограммные розы, соответствующие образам на рис. 4.27

4.8. Неоднозначность при моделировании пространственных структур при помощи вариограммы

На рис. 4.30*а,б* приведены два пространственных образа, которые представляют собой совершенно различные геологические структуры: параллельные русла и смыкающиеся эоловые структуры дюн. Эти геологические структуры обладают противоположной связностью — русла связаны горизонтально и способны поддерживать течение потока в себе, в то время как смыкающиеся дюны не имеют сквозной связанности и поэтому не могут поддерживать течение потока. Свойство связности крайне важно при моделировании течения в подземных месторождениях нефти, газа и задачах гидрогеологии.

Теперь обратимся к соответствующим вариограммам по всем направлениям для приведенных образов (рис. 4.30*в,г*). Эти вариограммы очень похожи и не отражают коренного отличия в связанности приведенных образов. Таким образом, моделирование пространственной структуры на основе только вариограммы является ограниченным и в определенных случаях может привести к некорректным результатам.

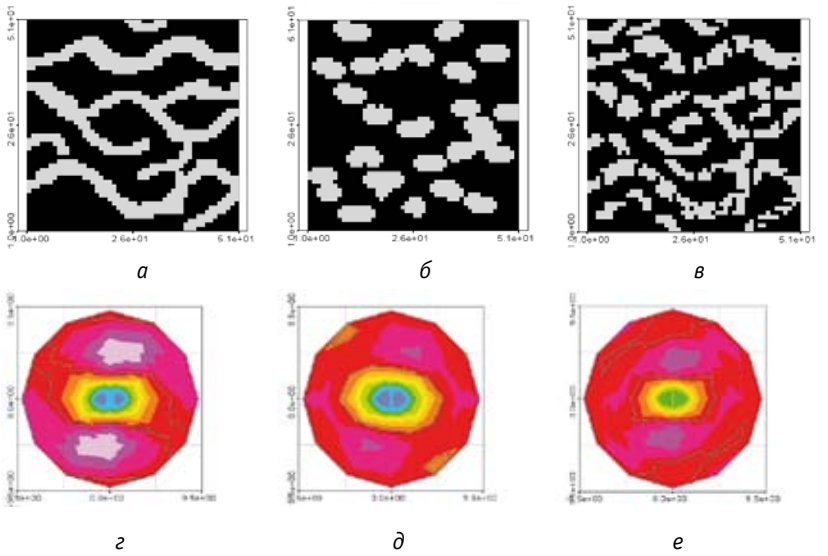


а *б*

в *г*

Рис. 4.30. Пространственные образы параллельных русел (*а*), смыкающихся структур (*б*) и соответствующие им вариограммы *(в, г*)

Проблемы с репрезентативностью и однозначностью вариограммы связаны с тем, что вариограмма является двухточечным моментом, т. е. зависит от поведения только пар точек. Если же корреляция определяется более чем парами точек (например, тройками или паттернами из десятка точек), то вариограмма не может охарактеризовать такие структуры. В рассмотренном случае с параллельными руслами структура определяется более чем двумя парами связанных соседних точек. Аналогичные выводы можно сделать на основе вариограммных роз для анизотропных структур (рис. 4.31).



а *б* *в*

г *д* *е*

Рис. 4.31. Пространственные образы параллельных русел (*а*), эллиптических объектов (*б*), прерывистых русел (*в*) и соответствующие им вариограммные розы (*г, д, е*)

Для решения подобных задач были разработаны методы многоточечной статистики, базирующиеся на многоточечных структурных функциях, которые можно рассматривать как более общий случай модели пространственной корреляции [Caers, 2005]. Более подробно методы многоточечной статистики изложены в главе 10.

4.9. Пространственный тренд и нестационарность

Тренд — систематическое изменение наблюдаемой величины с изменением координаты. В случае присутствия тренда в данных измерений предположение о стационарности наблюдаемой величины неправомерно. Так, значение температуры в горной местности зависит от высоты над уровнем моря, поэтому в такой местности локальное среднее значение температуры уменьшается с высотой, что нарушает предположение о стационарности среднего. Существование трендов обычно связано с тем или иным крупномасштабным явлением, которое оказывает систематическое влияние на наблюдаемую величину. Например, высота и орография влияют на количество осадков,

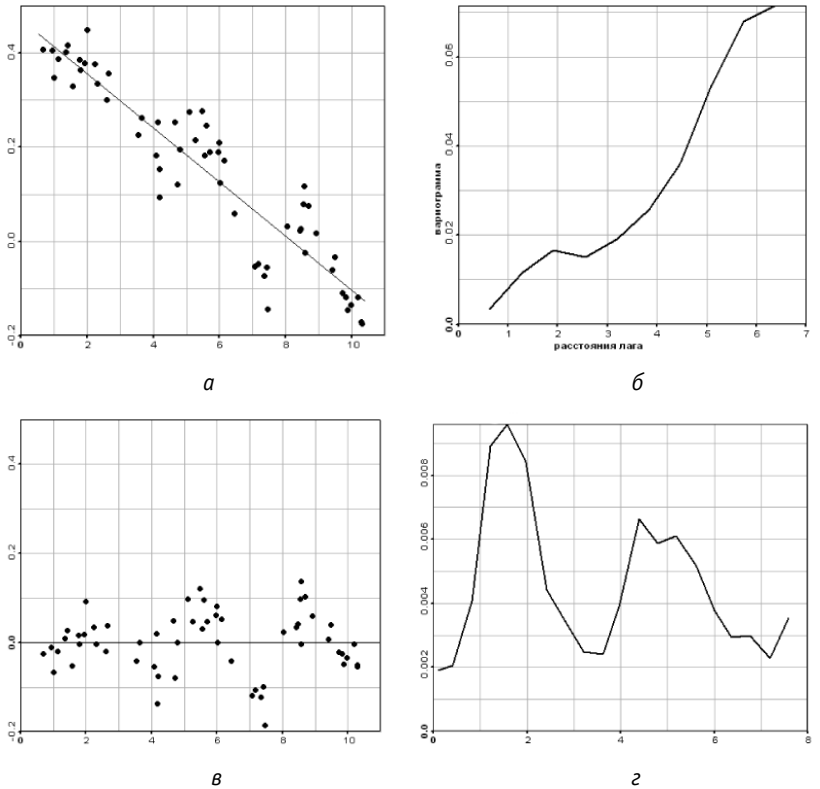
изменение влажности влияет на загрязнение атмосферы и процесс осаждения загразнения на поверхность. Пространственный тренд может иметь как простой линейный характер (в одном направлении), так и очень сложную нелинейную пространственную зависимость на различных масштабах.

Систематический пространственный тренд должен быть промоделирован и удален из данных измерений до построения вариограмм. Иначе вариограммы будут воспроизводить крупномасштабный тренд, что приведет к потере собственной корреляции наблюдаемой величины на более мелком масштабе. Более того, вариограмма данных, имеющих систематический тренд, будет нестационарной и потому не может быть использована в геостатистических моделях кригинга.

После анализа, моделирования корреляционной структуры невязок и получения интерполяционной оценки невязок пространственный тренд добавляется к оценке для получения итогового значения переменной.

Влияние линейного тренда на вариограммы представлено на рис. 4.32. Видно (см. рис. 4.32*а*), что данные имеют сильный линейный тренд (отрицательную корреляцию), в результате которого вариограмма ведет себя нестационарно на больших расстояниях, демонстрируя постоянный рост (см. рис. 4.32*б*). Если вычесть из данных компоненту линейного тренда ($y = -0{,}575x - 0{,}4712$), то оставшиеся невязки (см. рис. 4.32*в*) демонстрируют стационарную периодическую корреляционную структуру (см. рис. 4.32*г*), которую не отражала вариограмма данных с трендом. Таким образом, предварительное моделирование тренда и его вычитание из данных позволяют найти локальную корреляцию. Однако на практике модель тренда зачастую является более сложной и нелинейной.

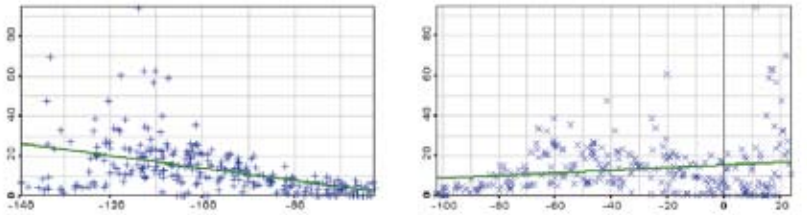
Нелинейный пространственный тренд можно увидеть в форме непрерывной крупномасштабной зависимости наблюдаемой величины от направления (рис. 4.33). Тренд в горизонтальном направлении X можно грубо представить при помощи линейной модели, в то время как тренд в направлении Y имеет ярко выраженный нелинейный характер. При наличии тренда в одном из направлений вариограмма обычно превышает уровень априорной вариации (рис. 4.34). Вариограмма в направлении 45° указывает на присутствие тренда. Вариограмма в направлении 0° обладает стационарностью только на расстоянии до 30. Вариограмма в направлении –45° также указывает на присутствие крайних значений либо тренда.



а *б*

в *г*

Рис. 4.32. Удаление тренда из данных и вариограмма:
а — данные с линейным трендом; *б* — вариограммы данных с трендом;
в — невязки после удаления линейного тренда; *г* — вариограммы невязок



а *б*

Рис. 4.33. Зависимости значений переменной в горизонтальном X (*а*)
и вертикальном Y (*б*) направлениях

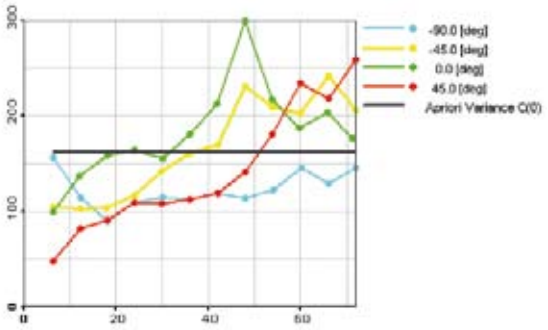


Рис. 4.34. Вариограммы по направлениям при наличии нелинейного пространственного тренда

Присутствие пространственного тренда приводит к невозможности использовать гипотезу о стационарности в каком-либо виде (второго порядка и др.). Следовательно, традиционные геостатистические модели кригинга не могут быть использованы напрямую для интерполяции данных.

Однако существуют методы учета пространственного тренда, которые позволяют адаптировать и все-таки получать оценку кригингом. Для этого следует выделить составляющую пространственного тренда при помощи отдельной модели. В качестве моделей тренда широко используются полиномы, сплайны либо более сложные нелинейные модели.

Здесь мы только приведем список геостатистических моделей, позволяющих оценивать данные в присутствии пространственного тренда:

1. *Кригинг с трендом* (или универсальный — universal — кригинг) использует модель тренда как линейную комбинацию набора базисных функций (см. Раздел 5.4). Универсальный кригинг прост в применении, не требует дополнительных настроек параметров, если правильно выбраны базисные функции. Их выбор и представляет наибольшую трудность. Чаще всего используется полиномиальная модель (линейная комбинация полиномов). Но такая жесткая модель не всегда может адекватно описать сложную многомасштабную пространственную структуру тренда.
2. *Кригинг с внешним дрейфом* (external drift) использует дополнительные данные измерений коррелированной переменной в качестве модели тренда. Он позволяет достаточно точно оценить тренд при наличии данных дополнительной тренд-переменной во всех точках оценивания. Кригинг с внешним дрейфом часто используется в климатических

приложениях, где пространственный тренд наблюдаемой переменной (например, температуры) часто связан с высотой. В этом случае в качестве модели тренда используется модель высот (см. Раздел 6.1).

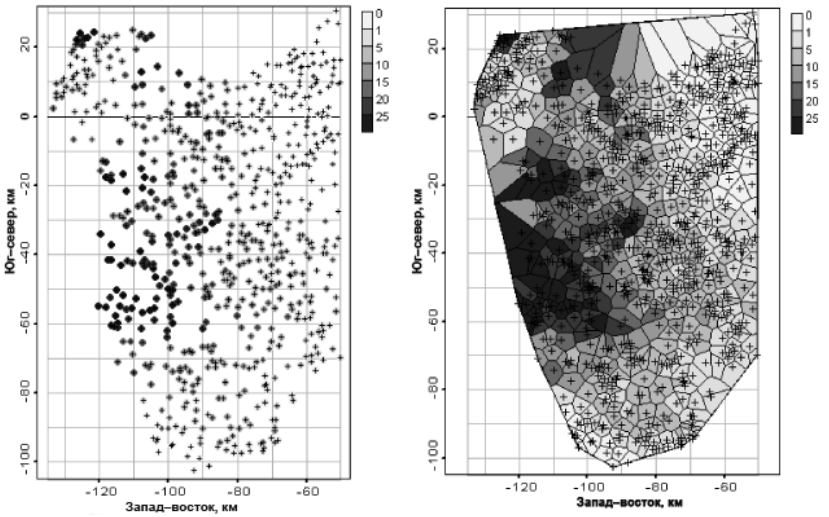
3. *Локально меняющееся среднее* (locally varying mean) использует в качестве модели тренда локальное среднее значение, которое может быть получено при помощи метода движущегося окна (см. Главу 2).
4. *Кригинг невязок с движущимся окном* (moving window residual kriging) похож на модель с локально меняющимся средним. Но он вычислительно гораздо более сложен, поскольку предполагает подбор модели тренда и модели вариограммы в каждой локальной окрестности (окне) [Haas, 1990].
5. *Внутренняя случайная функция порядка k* (intrinsic random function of order k — IRFk) использует моменты более высокого порядка, чем второй, вместе с вариограммой для моделирования трендов [Marcotte, David, 1988].
6. Моделирование нелинейного тренда на разных масштабах при помощи искусственной нейронной сети (ИНС). *Кригинг невязок искусственной нейронной сети* (neural netwrok residual kriging — NNRK) был предложен в 1996 г. [Kanevski et al., 1996]. На его основе было разработано целое семейство методов с применением различных типов ИНС (более подробно см. раздел 10.2).

4.10. Пример анализа пространственной корреляционной структуры

Для иллюстрации приведем пространственный корреляционный анализ данных по загрязнению поверхности западной части Брянской области изотопом ^{137}Cs (рис. 4.37). Загрязнение произошло после аварии на Чернобыльской АЭС. Оно было принесено облаком, которое частично выпало на поверхность в виде сухого осаждения, а где-то было вымыто локальными дождями.

На рис. 4.38 представлено вариограммное облако. Оно показывает, что в данных нет ярко выраженных выбросов, т. е. точек с необоснованно высокими (или низкими) значениями. Этот вывод следует из того, что ширина вариограммного облака растет с расстоянием, а значит, для близких точек разница значений меньше, чем для более удаленных. Присутствие на вариограммном облаке точек с большим значением квадрата разности значений

должно было бы вызывать опасения при построении функций пространственной корреляции, так как они могут вызвать серьезные искажения структуры.



а *б*

Рис. 4.37. Пример данных для вариографии — загрязнения в Брянской области ^{137}Cs:
а — диаграмма расположения точек; *б* — полигоны Вороного

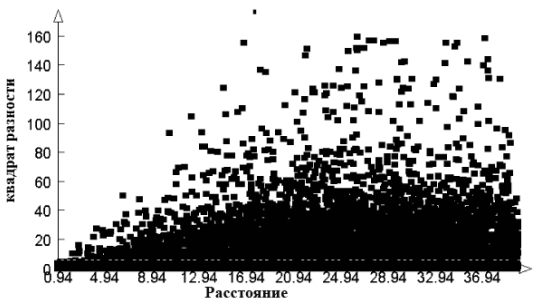


Рис. 4.38. Вариограммное облако для данных по загрязнению ^{137}Cs

Следующее, на что следует обратить внимание, — возможность присутствия тренда и анизотропии. Для прояснения этих вопросов рассмотрим функцию дрейфа. На рис. 4.39 представлен дрейф, рассчитанный для различных направлений (углы указаны в градусах от направления запад-восток против часовой стрелки). По этому рисунку, а также по розе дрейфа (рис. 4.40) видно, что поведение дрейфа — не колебание вокруг нуля и что оно раз-

личается в направлениях с северо-востока на юго-запад и с северо-запада на юго-восток. Значит, в данных присутствует тренд, но, судя по значениям модуля дрейфа, незначительный. Для таких данных предпочтительно использовать интерполяционные модели, учитывающие тренд. Но в данном случае традиционные геостатистические методы вполне могут дать приемлемый результат (нужно моделировать направления, в которых дрейф несущественен).

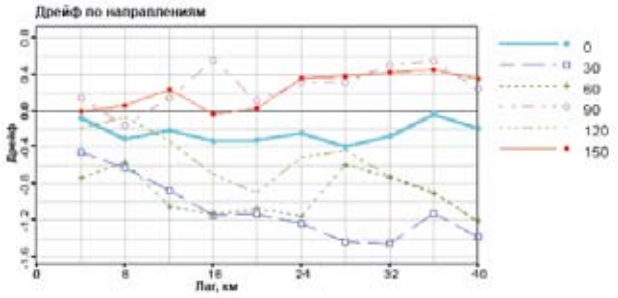


Рис. 4.39. Дрейф по направлениям для данных по загрязнению ^{137}Cs

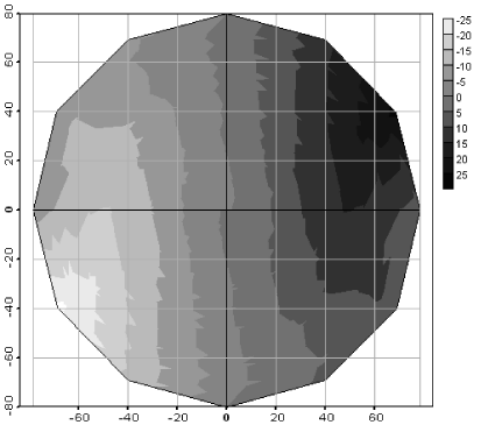
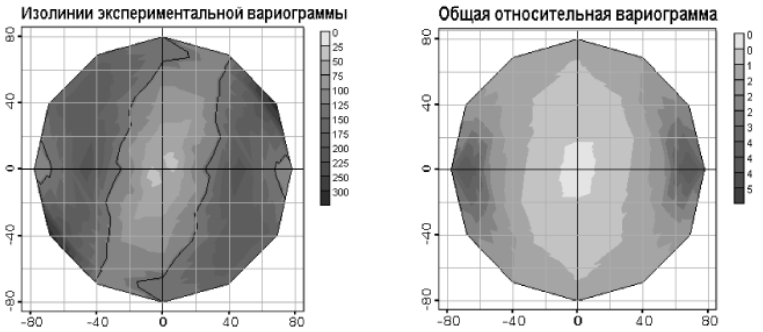


Рис. 4.40. Изолинии розы дрейфа для данных по загрязнению ^{137}Cs



а *б*

Рис. 4.41. Изолинии розы вариограммы (*а*) и розы общей относительной вариограммы (*б*) для данных по загрязнению ^{137}Cs

Вариограммная роза (рис. 4.41*а*) и вариограммная поверхность (рис. 4.42) демонстрируют анизотропную пространственную структуру. В данном случае на вариограммной поверхности и вариограммной розе довольно четко прочертился эллипс. Это указывает на геометрическую анизотропию. Основная ось расположена вдоль направления 60° от горизонтали, а малая — соответственно –30° от горизонтали.

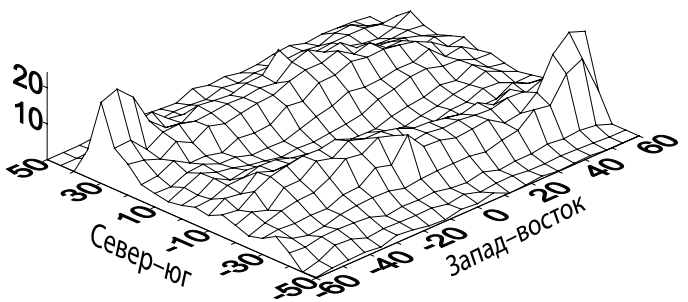


Рис. 4.42. Вариограммная поверхность для данных по загрязнению ^{137}Cs

Теперь можно выполнить моделирование вариограммы, т. е. осуществить подбор математической модели, удовлетворяющей всем свойствам вариограммы и позволяющей описать ее для любого лага и направления (напомним, что экспериментальную вариограмму мы оценили лишь для конечного набора лагов и направлений). Из вариограмм по направлениям наиболее стационарными и подходящими для моделирования являются вариограммы в направлениях 0° и 150°. Они отражают устойчивые корреляционные

структуры данных в этих направлениях. Как уже отмечалось, дрейф в этих направлениях несущественен (см. рис. 4.39), что позволит применить модель обычного кригинга.

Рисунки 4.43, 4.44 иллюстрируют процесс подбора параметров теоретической модели вариограммы при помощи интерактивной программы «Геостат Офис» [Kanevski, Maignan, 2004]. Там же приведены значения индексов (см. Раздел 4.4). Окончательные параметры теоретической модели вариограммы приведены в табл. 4.1. Она имеет гнездовую структуру, состоящую из наггет-модели и двух сферических анизотропных моделей и хорошо аппроксимирует точки экспериментальных вариограмм (рис. 4.45). Для анизотропной модели с геометрической анизотропией был получен радиус корреляции около 60 км вдоль направления 50° (СВ-Ю3), что соответствует примерно половине размера области. Зонная анизотропия моделируется второй сферической структурой в направлении 165° (С33-ЮВВ). Рисунок 4.46 показывает хорошее соответствие анизотропной структуры экспериментальной и модельной вариограмм.

Таблица 4.1. Параметры моделей вариограмм

Наггет c_0	Модель	Направление φ	Плато c	Продольный радиус $a\|\|$, км	Поперечный радиус $a\perp$, км
9,17	Сферическая	50°	55,39	57,37	38,41
	Сферическая	–15°	108,70	68,40	243,00

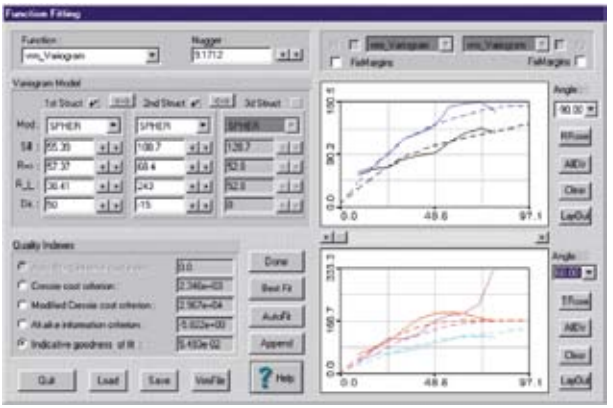


Рис. 4.43. Подбор параметров теоретической вариограммной модели для соответствия экспериментальным вариограммам по направлениям

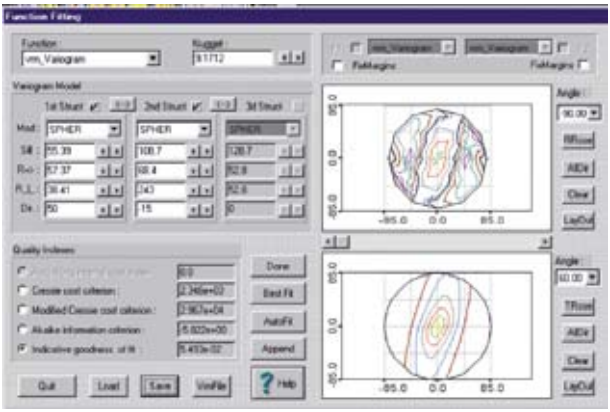


Рис. 4.44. Подбор параметров теоретической вариограммной модели для соответствия изолиний теоретической и экспериментальной вариограммных роз

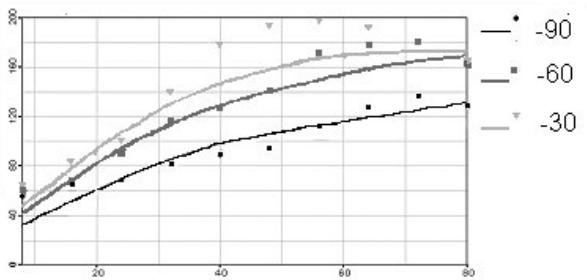
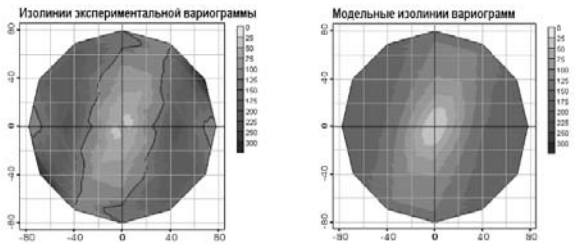


Рис. 4.45. Анизотропная модель вариограмм и точки значений экспериментальных вариограмм для данных по загрязнению ^{137}Cs в различных направлениях



а *б*

Рис. 4.46. Изолинии экспериментальной (*а*) и модельной (*б*) вариограммных роз для данных по загрязнению ^{137}Cs

В заключение отметим, что анализ и моделирование пространственных корреляционных структур — очень важный момент геостатистического анализа данных. По-видимому, теоретические знания, опыт и искусство вариографии при этом одинаково важны. Как правило, опытные геостатистики избегают автоматического подбора параметров вариограмм, больше опираясь на опыт, а также интерпретируемость анализа и моделирования. Дальнейшее использование промоделированных вариограмм — решение систем линейных уравнений, что является тривиальным при современном развитии численных методов и вычислительной техники. Основные результаты анализа и их отображение получены с помощью программного обеспечения «Геостат Офис» [Kanevski, Maignan, 1996].

Литература

Armstrong M. Common Problems Seen in Variograms // Mathematical Geology. — 1984. — Vol. 16, N 3. — P. 305—313.

Barnes R. J. The Variogram Sill and the Sample Variance // Mathematical Geology. — 1991. — Vol. 23, N 4. — P. 673—678.

Caers J. Petroleum Geostatistics / SPE. — [S. l.], 2005. — 98 p.

Chernov S., Demyanov V., Kanevski M., Savelieva E. VarRose — a Way of Variogram Analysis. — Moscow, 1998. — 27 p. — (Препринт / ИБРАЭ; IBRAE-98-03).

Christakos G. On the problem of permissible covariance and semivariogram models // Water Resources Research. — 1984. — Vol. 20, N 2. — P. 251—265.

Clark I. Practical Geostatistics. — London; New York: Elsevier Applied Science Publ., 1984.

Cressie N. Fitting models by weighted least squares // Mathematical Geology. — 1985. — Vol. 17, N 5. — P. 563—586.

David M. Handbook of Applied Advanced Geostatistical Ore Reserve Estimation. — Amsterdam B.V.: Elsevier Applied Science Publ., 1988.

Flamm C., Kanevsky M., Savelieva E. Non-regular variography and multi-method mapping to determination of origin of heavy metals // International Association for Mathematical geology Annual Conference: Papers and Extended Abstracts. — [S. l.], 1994.

Goovaerts P. Geostatistics for Natural Resources Evaluation. — [S. l.]: Oxford Univ. Press, 1997.

Haas T. C. Kriging and automated variogram modeling within a moving window // Atmospheric Environment. — 1990. — Vol. 24A. — P. 1759—1769.

Isaaks E. H., Srivastava R. M. An Introduction to Applied Geostatistics. — Oxford: Oxford Univ. Press, 1989.

Kanevsky M., Arutyunyan R., Bolshov L. et al. Artificial neural networks and spatial estimations of Chernobyl fallout // Geoinformatics. — 1996. — Vol. 7, N 1—2. — P. 5—11.

Kanevski M., Maignan M. Analysis and modelling of spatial environmental data. — Lausanne: EPFL Press, 2004. — 288 p. — (With a CD and educational/research MS Windows software tools).

Marcotte D., David M. Trend surface analysis as a special case of IRF-k kriging // Mathematical Geology. — 1988. — Vol. 20, N 7. — P. 821—824.

Pannatier Y. VARIOWIN Software for Spatial Data Analysis. — New York: Springer-Verl., 1996.

Xiaodong J., Olea R. A., Yu Y.-S. Semivariogram modeling by weighted least squares // Computers and Geosciences. — 1996. — Vol. 22, N 4. — P. 387—397.

Zhang X. F., Van Eijkeren J. C. H., Heemink A. W. On the weighted least-square method for fitting a semivariogram model // Computers and Geosciences. — 1995. — Vol. 21, N 4. — P. 605—608.

Zimmerman D. L. Another Look at Anisotropy in Geostatistics // Mathematical Geology. 1993. — Vol. 25, N 4.

Глава 5

Геостатистические интерполяции для одной переменной

Данная глава посвящена семейству моделей кригинга для анализа одной пространственной переменной. В Разделе 5.1 формулируются основные постулаты кригинга. Основные типы кригинга (простой и обычный) подробно описаны в Разделах 5.2 и 5.3. В Разделах 5.4 и 5.5 рассмотрены некоторые другие типы кригинга — универсальный, логнормальный. Раздел 5.6 посвящен дополнительным аспектам теории кригинга, в частности некоторым свойствам весовых коэффициентов и вариации кригинга.

Кригинг — базовая интерполяционная модель геостатистики. Он является основой всех методов, связанных с геостатистикой, — интерполяции, вероятностного картирования, стохастического моделирования. Термин «кригинг» служит для обозначения семейства алгоритмов линейной пространственной регрессии. Он происходит от фамилии инженера Д. Крига, который первым применил интерполятор на основе модели пространственной корреляции данных для анализа золотых месторождений Южной Африки [Krige, 1951]. Л. С. Гандин независимо от Д. Крига применил аналогичный метод для объективного анализа метеополей [Гандин, Каган, 1976].

Выделяют несколько вариантов моделей кригинга (простой, обычный, универсальный, логнормальный, невязок и др.), которые различаются принятыми предположениями и используемой информацией о моделируемой переменной.

5.1. Основные постулаты кригинга

Рассмотрим проблему оценивания значения непрерывной переменной Z в произвольной точке $\boldsymbol{x}$, принадлежащей области пространства S. Исходная информация о переменной представлена в виде набора $\{z(\boldsymbol{x}_i), i = 1, \ldots, n\}$ из n измерений, сделанных в точках $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ пространства.

Все интерполяторы семейства кригинга являются различного рода модификациями базового линейного регрессионного оценивателя $Z^*(\boldsymbol{x})$, определяемого следующим образом:

$$Z^*(\boldsymbol{x}) - m(\boldsymbol{x}) = \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x})\left[Z(\boldsymbol{x}_i) - m(\boldsymbol{x}_i)\right], \tag{5.1}$$

где $\lambda_i(\boldsymbol{x})$ — весовые коэффициенты, относящиеся к данным $z(\boldsymbol{x}_i)$. Значения $z(\boldsymbol{x}_i)$ интерпретируются как реализации случайной переменной $Z(\boldsymbol{x}_i)$. Величины $m(\boldsymbol{x})$ и $m(\boldsymbol{x}_i)$ являются математическими ожиданиями случайных переменных $Z(\boldsymbol{x})$ и $Z(\boldsymbol{x}_i)$. Число данных, использующихся при оценке, и значения весовых коэффициентов могут меняться в зависимости от местоположения оцениваемой точки $\boldsymbol{x}$.

Тип оценивателя зависит от модели случайной функции $Z(\boldsymbol{x})$. Ее всегда можно разложить на две компоненты — детерминистический тренд $m(\boldsymbol{x})$ и случайную невязку $R(\boldsymbol{x})$:

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + R(\boldsymbol{x}). \tag{5.2}$$

Компонента невязки $R(\boldsymbol{x})$ моделируется как стационарная случайная функция с нулевым математическим ожиданием $m_R(\boldsymbol{x})$ и ковариацией $C_R(\boldsymbol{h})$:

$$E\{R(\boldsymbol{x})\} = 0,$$
$$\mathrm{Cov}\{R(\boldsymbol{x}), R(\boldsymbol{x}+\boldsymbol{h})\} = E\{R(\boldsymbol{x})R(\boldsymbol{x}+\boldsymbol{h})\} = C_R(\boldsymbol{h}).$$

Математическое ожидание пространственной переменной Z в точке $\boldsymbol{x}$, таким образом, будет равно значению тренда:

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

Далее мы рассмотрим разновидности кригинга для моделирования одной переменной, которые определяются предположением о виде тренда.

Все методы семейства кригинга используют одну и ту же целевую функцию для минимизации, а именно вариацию ошибки оценки $\sigma_E^2(\boldsymbol{x})$ при дополнительном условии несмещенности оценки, иными словами, вариация

$$\sigma_E^2(\boldsymbol{x}) = \mathrm{Var}\{Z^*(\boldsymbol{x}) - Z(\boldsymbol{x})\} \tag{5.3}$$

минимизируется при ограничении

$$E\{Z^*(\boldsymbol{x}) - Z(\boldsymbol{x})\} = 0. \tag{5.4}$$

Изначально все кригинги рассматривались как глобальные оцениватели, т. е. для оценки значения в точке x_0 из области S предполагалось использовать все имеющиеся измерения $\{z(\boldsymbol{x}_i), i = 1, \ldots, n\}$. Тогда предположение, например о постоянстве среднего распространяется на всю (возможно, достаточно большую) область, что, вообще говоря, редко встречается в природе. Чтобы не делать такого сильного предположения, на практике обычно используют локальную оценку на основе $n(\boldsymbol{x})$ ближайших к точке оценивания данных. Можно сказать, что используемые при оценке данные выбираются из некоторой окрестности $W(\boldsymbol{x})$ точки оценивания $\boldsymbol{x}$. Размер и форма этой окрестности зависят от исходных данных: предлагается использовать зону, ориентированную в соответствии с эллипсом корреляции, но возможно и меньшего или большего размера. Уменьшение окрестности позволяет получать более вариабельную (менее сглаженную) оценку.

5.2. Простой кригинг

Простой кригинг (simple kriging — SK) работает в предположении о стационарности второго порядка случайной переменной $Z(\boldsymbol{x})$ (см. Раздел 2.6). Кроме того, предполагается, что детерминистическая компонента $m(\boldsymbol{x})$ в (5.2) постоянна и известна на всей области исследования S:

$$m(\boldsymbol{x}) = m, \ \forall \in S.$$

Знание среднего значения m дает возможность сделать простое преобразование путем вычета постоянного тренда

$$Y(\boldsymbol{x}) = Z(\boldsymbol{x}) - m \tag{5.5}$$

и далее строить линейный оцениватель для случайной функции $Y(\boldsymbol{x})$ на всей области S

$$Y^*(\boldsymbol{x}) = \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) Y(\boldsymbol{x}_i), \tag{5.6}$$

автоматически получая несмещенность оценки (сохранение глобального среднего). Так как $E\{Y(\boldsymbol{x})\} = 0, \quad \forall \boldsymbol{x} \in S$, то

$$E\left\{Z^*(\boldsymbol{x}) - Z(\boldsymbol{x})\right\} = E\left\{Y^*(\boldsymbol{x}) - Y(\boldsymbol{x})\right\} = E\left\{\sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) Y(\boldsymbol{x}_i) - Y(\boldsymbol{x})\right\} = 0.$$

Окончательная оценка простого кригинга из (5.5) и (5.6) имеет вид

$$Z^*(\boldsymbol{x}) = m + \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) Y(\boldsymbol{x}_i). \tag{5.7}$$

Теперь рассмотрим вариацию ошибки σ_R^2 для оценки функции $Y(\boldsymbol{x})$.

$$\begin{aligned} \sigma_R^2 &= \mathrm{Var}\left\{Y^*(\boldsymbol{x}) - Y(\boldsymbol{x})\right\} = E\left\{\left(Y^*(\boldsymbol{x}) - Y(\boldsymbol{x})\right)^2\right\} = \\ &= \mathrm{Var}\left\{Y^*(\boldsymbol{x})\right\} - 2\mathrm{Cov}\left\{Y^*(\boldsymbol{x})Y(\boldsymbol{x})\right\} + \mathrm{Var}\left\{Y(\boldsymbol{x})\right\}. \end{aligned} \tag{5.8}$$

Так как функция $Z(\boldsymbol{x})$ удовлетворяет стационарности второго порядка, этому условию удовлетворяет и функция $Y(\boldsymbol{x})$. Тогда все ковариации и вариации, входящие в (5.8), существуют. Чтобы получить вариацию оценки, подставим в первое и второе слагаемые суммы (5.8) формулу оценки (5.6):

$$\mathrm{Var}\left\{Y^*(\boldsymbol{x})\right\} = \sum_{i=1}^{n(\boldsymbol{x})} \sum_{j=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) \lambda_j(\boldsymbol{x}) C_{ij}, \tag{5.9}$$

$$2\mathrm{Cov}\left\{Y^*(\boldsymbol{x})Y(\boldsymbol{x})\right\} = 2\sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) C_{i0}, \tag{5.10}$$

где $C_{ij} = \mathrm{Cov}\left\{Y(\boldsymbol{x}_i)Y(\boldsymbol{x}_j)\right\}$, $C_{i0} = \mathrm{Cov}\left\{Y(\boldsymbol{x}_i)Y(\boldsymbol{x})\right\}$.

Вариация неизвестной случайной переменной $Y(\boldsymbol{x})$ также существует и связана с априорной вариацией исходных данных σ_z^2:

$$\mathrm{Var}\left\{Y(\boldsymbol{x})\right\} = \mathrm{Var}\left\{Z(\boldsymbol{x})\right\} - m^2 = \sigma^2{}_Z - m^2 = \sigma_Y^2. \tag{5.11}$$

В итоге получим значение вариации ошибки оценки переменной Y как сумму (5.9), (5.10) и (5.11):

$$\sigma_R^2(\boldsymbol{x}) = \sum_{i=1}^{n(\boldsymbol{x})} \sum_{j=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) \lambda_j(\boldsymbol{x}) C_{ij} - 2\sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) C_{i0} + \sigma_Y^1. \tag{5.12}$$

Кригинг как наилучший оцениватель из класса линейных должен иметь минимальную вариацию ошибки. Весовые коэффициенты $\lambda_i(\boldsymbol{x})$ в (5.6) подбираются так, чтобы они минимизировали вариацию ошибки (5.12), т. е. чтобы производная от вариации по всем весам равнялась нулю. В результате дифференцирования получается система уравнений простого кригинга — линейная система из $n(\boldsymbol{x})$ уравнений с $n(\boldsymbol{x})$ неизвестными:

$$\sum_{j=1}^{n(\boldsymbol{x})} \lambda_j(\boldsymbol{x}) C_{ij} = C_{i0}, \forall i = 1, \ldots, n(\boldsymbol{x}). \tag{5.13}$$

Система уравнений простого кригинга (5.13) имеет единственное решение, если матрица ковариаций несингулярна. Это условие выполнено при положительной определенности функции ковариации и отсутствии среди набора исходных точек $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n(x)}$, пространственно совпадающих или очень близко расположенных. Совпадающие или близкие точки формируют линейно зависимые строки матрицы C_{ij}.

Оценка функции $Z(\boldsymbol{x})$ получается подстановкой полученных весовых коэффициентов в формулу (5.7).

Ошибка оценки простого кригинга (вариация простого кригинга) получается из формулы (5.12) подстановкой в нее (5.13). Вариацию простого кригинга можно вычислить по формуле

$$\sigma^2_{SK}(\boldsymbol{x}) = \sigma^2_Z - \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) C_{i0}. \qquad (5.14)$$

Простой кригинг обладает рядом свойств.

- Оценка простого кригинга является точной. Это означает, что если координата оцениваемой точки x_0 совпадает с какой-то координатой из исходного набора данных ($\boldsymbol{x}_0 = \boldsymbol{x}_i$, $i = 1, \ldots, n$), то полученная оценка будет также совпадать с исходным значением: $Z^*(\boldsymbol{x}_0) = Z(\boldsymbol{x}_k)$. Это легко доказать, пользуясь единственностью решения системы уравнений простого кригинга.
- Веса простого кригинга не зависят от значений исходного набора данных, а зависят только от пространственной корреляции поля, построенного на основе данных. Таким образом, если есть несколько наборов исходных данных, измеренных в одних и тех же точках и описывающихся одинаковыми (или мультипликативно связанными) функциями ковариации, то для вычисления оценки простого кригинга в общей точке x_0 систему уравнений простого кригинга достаточно решить один раз, а потом использовать полученные веса для всех переменных.
- Оценка простого кригинга является сглаженной по сравнению с распределением исходных данных. Как видно из (5.14), вариация оценки простого кригинга σ^2_{SK} меньше значения вариации исходных данных σ_2.
- Ошибка простого кригинга ортогональна оценке простого кригинга в гильбертовом пространстве, построенном из всех возможных линейных комбинаций исходных данных и имеющем в качестве метрики ковариацию. Это свойство лишний раз подтверждает, что простой кригинг является лучшей оценкой в классе линейных оценивателей.

Основным недостатком простого кригинга является предположение о знании среднего. Использование в качестве среднего его статистической оценки (математического ожидания) делает веса зависимыми от значений исходного набора данных. Кроме того, оценка математического ожидания может оказаться искаженной, смещенной и т. п., например при высокой кластерности исходной сети мониторинга (о кластерности и декластеризации см. Разделы 2.5 и 2.6). Поэтому простой кригинг редко применяется как самостоятельный метод оценивания, обычно его использование связано с искусственными комбинациями, где среднее известно вследствие предварительных манипуляций с исходными данными.

5.3. Обычный кригинг

Обычный кригинг (ordinary kriging — OK) отличается от простого кригинга тем, что не предполагает знание среднего значения. В обычном кригинге среднее значение считается постоянным, но оно неизвестно. Кроме того, обычный кригинг при использовании локальной оценки не требует постоянства среднего по всей зоне оценивания; предполагается, что среднее постоянно только в окрестности точки оценивания $W(\boldsymbol{x})$. Предположение о постоянстве среднего в рамках малой окрестности более реалистично, тем более что данные обладают пространственной непрерывностью.

Оценка обычного кригинга строится, как линейная комбинация исходных данных:

$$Z^*(x) = \sum_{i=1}^{n(x)} \lambda_i(x) Z(x_i). \qquad (5.15)$$

Рассмотрим условие несмещенности (5.4) в случае неизвестного среднего:

$$E\left\{Z^*(x) - Z(x)\right\} = E\left\{\sum_{i=1}^{n(x)} \lambda_i(x) Z(x_i) - Z(x)\right\} = \left[\sum_{i=1}^{n(x)} \lambda_i(x) - 1\right] m \equiv 0,$$

т.е. условие несмещенности будет выполнено, если сумма весов, использующихся при оценке, равна единице:

$$\sum_{i=1}^{n(x)} \lambda_i(x) = 1. \qquad (5.16)$$

Таким образом, отсутствие знания о значении среднего накладывает на веса $\lambda_i(\boldsymbol{x})$ дополнительное требование. Теперь, чтобы выполнялось свой-

ство наилучшего оценивателя, нужно находить веса, которые минимизируют вариацию при дополнительном ограничении (5.16).

Решение этой задачи осуществляется с использованием минимизации лагранжиана $L(\boldsymbol{x})$, куда помимо вариации (5.8) включается условие (5.16) с весовым коэффициентом $\mu(\boldsymbol{x})$ (множителем Лагранжа):

$$L(x) = \sum_{i=1}^{n(x)} \sum_{j=1}^{n(x)} \lambda_i(x)\lambda_j(x)C_{ij} - 2\sum_{i=1}^{n(x)} \lambda_i(x)C_{i0} + \sigma_z^2 + 2\mu(x)\left(\sum_{i=1}^{n(x)} \lambda_i(x) - 1 \right),$$

где C_{ij} — ковариации случайных переменных:

$$C_{ij} = \mathrm{Cov}\{Z(x_i)Z(x_j)\}, i, j = 1, ..., n,$$
$$C_{i0} = \mathrm{Cov}\{Z(x_i)Z(x)\}, \quad i = 1, ..., n.$$

Для минимизации лагранжиана $L(x)$ необходимо его продифференцировать по всем весам $\lambda_i(x)$ и коэффициенту $\mu(x)$, а потом приравнять эти производные нулю. В результате получается линейная система из $n(x)$ + 1 уравнений с $n(x)$ + 1 неизвестными — система уравнений обычного кригинга:

$$\begin{cases} \sum_{j=1}^{n(x)} \lambda_j(x)C_{ij} + \mu(x) = C_{i0}, i = 1, \ldots, n(x), \\ \sum_{i=1}^{n(x)} \lambda_i(x) = 1. \end{cases} \tag{5.17}$$

Система уравнений (5.17) аналогично с системой уравнений простого кригинга (5.13) имеет единственное решение при положительной определенности функции ковариации C и отсутствии пространственно совпадающих или очень близких точек.

Для вычисления оценки найденные веса $\lambda_i(x)$ подставляются в линейную комбинацию (5.15). Вариация обычного кригинга вычисляется из формулы (5.12) с использованием первой части системы (5.17):

$$\sigma_{OK}^2(x) = \sigma_Z^2 - \sum_{i=1}^{n(x)} \lambda_i(x)C_{i0} + \mu(x). \tag{5.18}$$

На практике чаще вместо предположения о стационарности второго порядка и функций ковариации пользуются менее слабым предположением о внутренней гипотезе и связанной с ней вариограммой (о внутренней гипотезе см. Раздел 2.6, о вариограмме — Раздел 4.2). Система уравнений обычного кригинга (5.17) легко может быть переписана в терминах вариограммы:

$$\begin{cases} \sum_{j=1}^{n(x)} \lambda_j(x)\gamma_{ij} - \mu(x) = \gamma_{i0}, i = 1, ..., n(x), \\ \sum_{i=1}^{n(x)} \lambda_i(x) = 1. \end{cases}$$

Вариация обычного кригинга (5.18) также может быть переписана в терминах вариограммы:

$$\sigma^2_{OK}(x) = \sum_{i=1}^{n(x)} \lambda_i(x)\gamma_{i0} + \mu(x). \quad (5.19)$$

Все свойства, описанные в разделе 5.2 для простого кригинга, относятся в той же мере и к обычному кригингу. Но сравнение формул вариаций простого (5.14) и обычного (5.18), (5.19) кригингов показывает, что платой за неизвестное значение среднего является увеличение вариации, что ведет к росту неопределенности оценки.

Для того чтобы понять, как влияет на оценку и вариацию кригинга отсутствие знания среднего, рассмотрим искусственный пример. Пусть известны одно значение $Z(x_1)$ функции $Z(x)$ и модель пространственной корреляции, заданная ковариацией $C(h)$ или вариограммой $\gamma(h)$. Используя эти данные, построим оценку функции в точке x_0. Если известно среднее (пусть для простоты это будет нуль), модель простого кригинга строится следующим образом:

1. Уравнение простого кригинга:
   $\lambda C_{11} = C_{10}$, где C_{11} — значение априорной вариации; C_{10} — значение ковариационной функции для вектора, разделяющего точки x_1 и x_0.
2. Весовой коэффициент кригинга
   $$\lambda = \frac{C_{10}}{C_{11}}.$$
3. Определяется оценка:
   $$Z(x_0) = \frac{C_{10}}{C_{11}} Z(x_1).$$
   Значение оценки зависит от взаимной пространственной ориентации точек через пространственную корреляцию.

4. Вариация кригинга
   $$\sigma^2 = C_{00} - \frac{C_{10}}{C_{11}} C_{10}$$
   также определяется пространственной корреляцией.

Пример, иллюстрирующий зависимость оценки от расстояния до известного значения, приведен на рис. 5.1. Использована сферическая модель пространственной корреляции с нулевым наггетом, единичным плато и радиусом корреляции 1. Известное значение помечено кружком. Видно, что оценка по мере удаления от исходной точки все сильнее отличается от исходного значения, а при достижении расстояния, равного радиусу корреляции, влияние исходной точки пропадает, и определить значение дальше невозможно. При этом возраставшая вариация кригинга достигает значения априорной вариации (1).

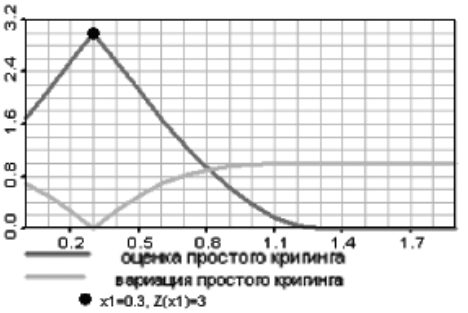


Рис. 5.1. Зависимость оценки и вариации простого кригинга от расстояния при одном известном значении функции

Теперь рассмотрим обычный кригинг, когда среднее неизвестно.

1. Система уравнений обычного кригинга имеет вид

$$\begin{cases} \lambda\gamma_{11} + \mu = \gamma_{10}, \\ \lambda = 1. \end{cases}$$

   Коэффициент известен сразу из введенного для обычного кригинга дополнительного условия.
2. Оценка обычного кригинга $Z(x_0) = Z(x_1)$ не зависит от пространственного расположения точек. Она постоянна везде и равна известному (т. е. среднему по набору данных) значению.
3. Вариация кригинга $\sigma^2 = 2\gamma_{10}$ зависит от расположения точек, растет по мере удаления, а потом выходит на удвоенное значение плато.

Иллюстрация оценки простого и обычного кригинга в одномерном случае

Проиллюстрируем оценку кригинга в одномерном случае на основе шести точек измерений. В реальных задачах данных может быть намного больше, однако в случае такого небольшого числа точек построить адекватную вариограммную модель бывает невозможно. Если не удается промоделировать пространственную корреляцию на основе имеющихся данных, то можно сделать предположения о величине радиуса корреляции и направлении, исходя из другой косвенной информации. Например, в геологии часто используют геологическую качественную информацию — экспертное описание слоев пород.

Оценка кригинга для одномерного случая, изображенная на рис. 5.2, демонстрирует точное воспроизведение данных простым и обычным кригингом (по построению оценки). Оценки простого и обычного кригинга весьма близки в приведенном одномерном примере при одинаковых параметрах модели (радиус корреляции $r = 50$, плато $c = 5$, наггет $c_0 = 0$). Увеличение наггета до $c_0 = 2$ с соответствующим уменьшением плато до $c = 3$ ведет к росту случайной компоненты в оценке. Это приводит к более высокой вариации оценки в окрестностях точек измерений вместе с более гладкой оценкой вне этих окрестностей. Таким образом, оценка вне точек измерений стремится к среднему значению, в то время как сами точки измерения воспроизводятся точно в виде «выколотых». Это означает, что данные в модели с высоким наггетом предполагаются зашумленными и менее репрезентативными, т. е. не характерными и поэтому оказывающими слабое влияние вне своей непосредственной окрестности. Вариация соответствующей оценки кригинга (рис. 5.3) демонстрирует, что значения обычного кригинга выше, чем значения простого кригинга. Вариация кригинга с высоким наггетом значительно выше вне точек измерений, чем для модели с нулевым наггетом. Отметим, что вариация оценки кригинга в точках измерений равна нулю, что не отражает существующую вариабельность, например при повторных измерениях или обрубку измерений. Ошибка измерений может быть учтена в кригинге (см. подраздел 5.6.3).

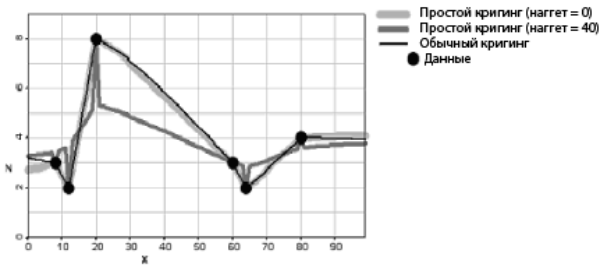


Рис. 5.2. Оценка простого и обычного кригинга в одномерном случае (для сферической модели вариограммы с радиусом корреляции r = 50 и различных значений наггета c_0 = 0 и 2, соответственно плато c равно 5 и 3)

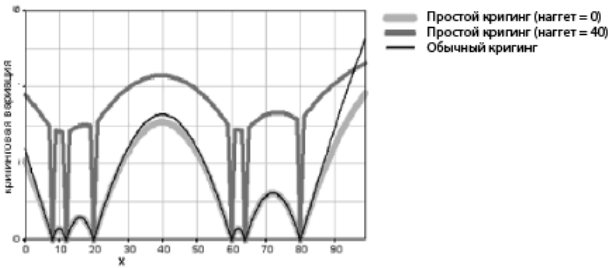


Рис. 5.3. Вариация оценки простого и обычного кригинга в одномерном случае (для различных значений наггета c_0 равно 0 и 2)

Упражнение 5.1. Несмещенность оценки кригинга

Показать несмещенность оценки обычного кригинга при условии стационарности переменной Z.

Упражнение 5.2. Оценка кригинга и нестационарность вариограммы

Одним из условий кригинга является стационарность (внутренняя гипотеза). Что происходит с оценками кригинга, если нет стационарности среднего значения?

Упражнение 5.3. Точное воспроизведение данных кригингом

Кригинг является точным оценивателем — воспроизводит значения исходных данных. Показать, что для заданного значения $Z(x_0)$ оценка кригинга $Z^*(x_0) = Z(x_0)$.

Упражнение 5.4. Вариация оценки кригингом

Показать, что для заданного значения $Z(x_0)$ в отсутствие ошибки измерения вариация оценки кригинга $\sigma_{SK}(x_0) = 0$.

Упражнение 5.5. Заниженная вариация оценки кригинга

Вариация оценки кригинга не зависит от значений данных, а только от их взаимного местоположения.

А. Что больше — вариация оценки кригинга или вариации исходных данных?

Б. В каком случае оценка кригинга будет обладать той же вариацией, что и исходные данные?

Упражнение 5.6. Сравнение оценок обычного кригинга и метода обратных квадратов расстояния

На рис. 5.4. приведены три интерполяционные оценки на основе пяти данных. Две из них получены обычным кригингом с большим и маленьким радиусами корреляции, третья — методом обратных квадратов расстояния (см. Раздел 3.1). Поставить в соответствие оценкам A, B, C приведенные выше интерполяционные модели.

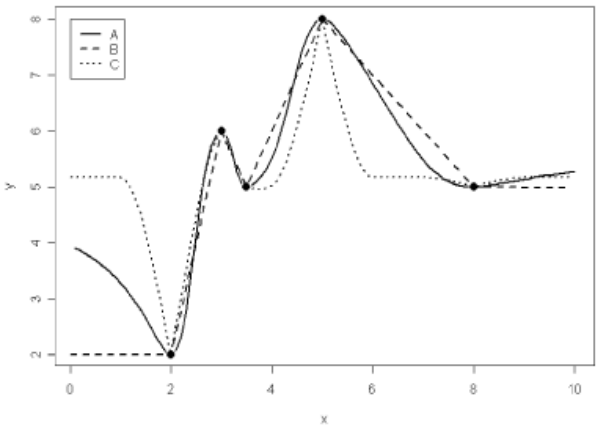


Рис. 5.4. Интерполяционные оценки A, B, C на основе пяти данных: обычный кригинг с большим и маленьким радиусами корреляции и метод обратных квадратов расстояния

Пример результата обычного кригинга

Для более полного понимания функционирования обычного кригинга рассмотрим пример его применения на искусственных данных, моделирующих пористость. Пространственное распределение исходных данных приведено на рис. 5.5 — шесть точек, имеющих значения в диапазоне от 0,05 до 0,30. Цель примера — показать влияние на результат обычного кригинга модели пространственной корреляционной структуры (варио-

граммы). При этом будем рассматривать оценку обычного кригинга и вариацию обычного кригинга.

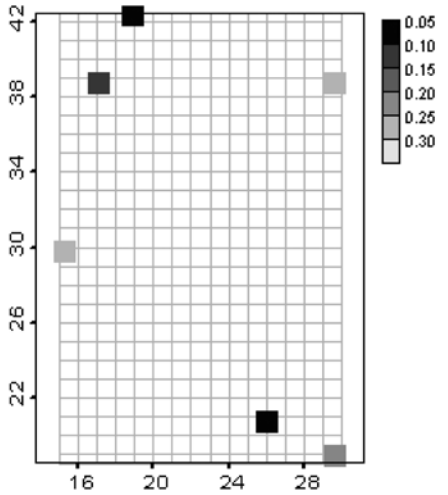


Рис. 5.5. Пространственное расположение исходных данных для обычного кригинга

Будем использовать сферическую модель вариограммы с различными параметрами (см. главу 4). Постоянным будет только значение плато, которое характеризуется априорной вариацией исходных данных. Варьировать будем значение радиуса корреляции и направление главной оси эллипса при геометрической анизотропии, а также значение наггета. На приведенных ниже рисунках шкала значений оценки обычного кригинга одинакова для всех случаев — диапазон значений оценки примерно одинаков для всех рассмотренных вариантов модели пространственной корреляции. Вариация кригинга представлена в различных шкалах (четырех типов), так как диапазон ее значений зависит от параметров модели пространственной корреляции.

На рисунках 5.6, 5.7 приведены соответственно оценка и вариация кригинга для случая изотропной модели вариограммы (т. е. радиус корреляции не зависит от направления). Увеличение радиуса корреляции приводит к использованию большего числа точек и вследствие этого к их заметному влиянию. У оценки кригинга (см. рис. 5.6) увеличиваются зоны и больших, и малых значений, уменьшается зона, соответствующая среднему. Для вариации кригинга (см. рис. 5.7) заметно уменьшение значения и рост области с более низким значением.

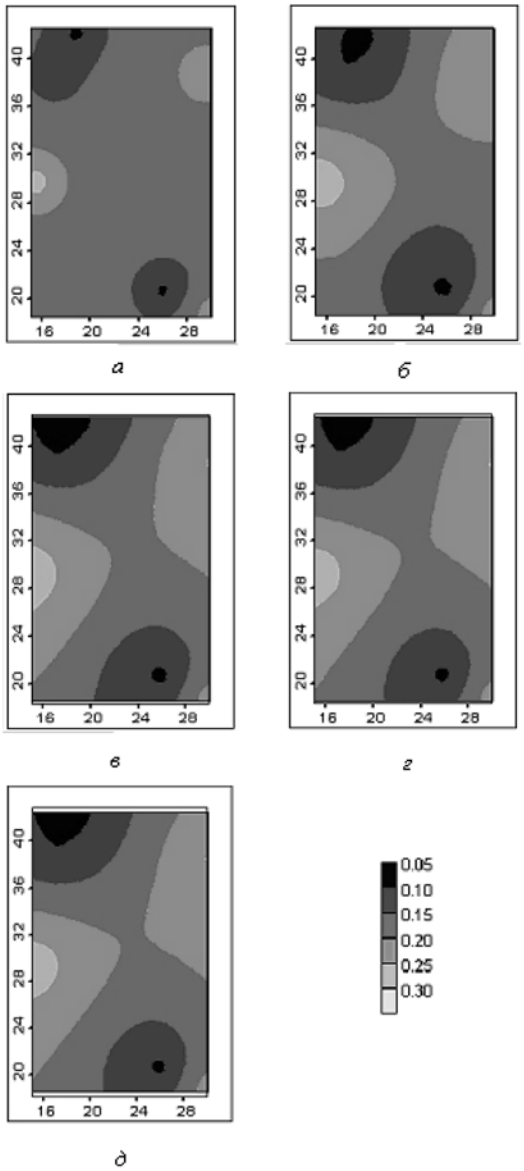


Рис. 5.6. Оценка обычного кригинга при изотропной модели пространственной корреляционной структуры (вариограмме) и различных значениях радиуса корреляции: *а* — 5 м; *б* — 10 м; *в* — 20 м; *г* —30 м; *д* — 40 м

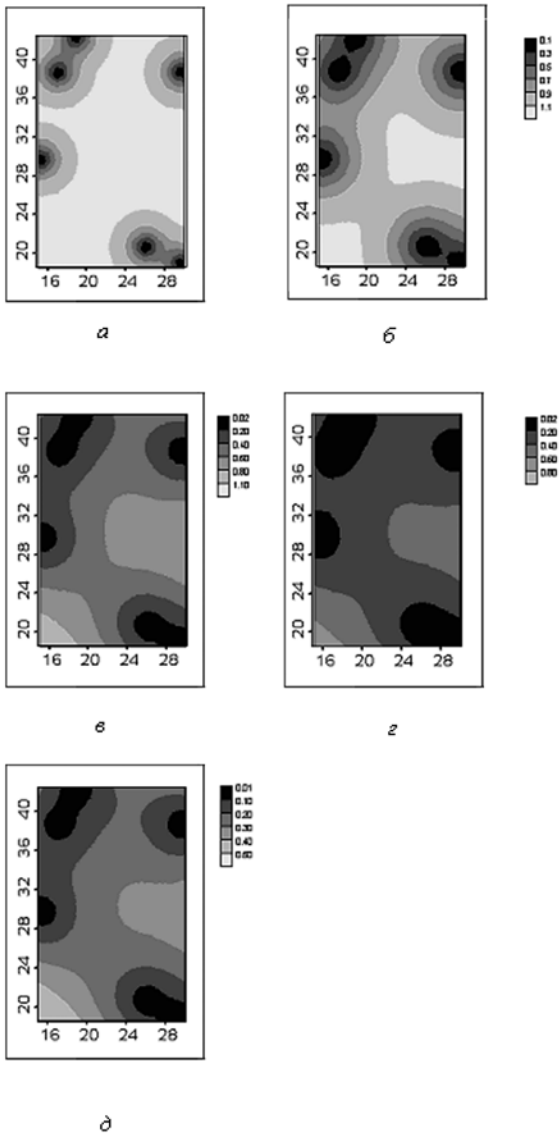


Рис. 5.7. Вариация обычного кригинга при изотропной модели пространственной корреляционной структуры (вариограмме) и различных значениях радиуса корреляции:
а — 5 м; *б* — 10 м; *в* — 20 м; *г* — 30 м; *д* — 40 м

На рис. 5.8—5.13 представлены оценка и вариация кригинга для случаев с моделью вариограммы с геометрической анизотропией: рис. 5.8 и 5.9 соответствуют направлению главной оси вдоль оси X, рис. 5.10 и 5.11 — вдоль оси Y, рис. 5.12 и 5.13 — направлениям 45°, 30° и 60° от оси X против часовой стрелки. Радиус корреляции вдоль главной оси равен 40 м. Значение радиуса корреляции вдоль малой оси варьируется. Везде заметно влияние направления главной оси, а изменение значения второго радиуса корреляции приводит к результатам, аналогичным описанным выше для изотропного случая.

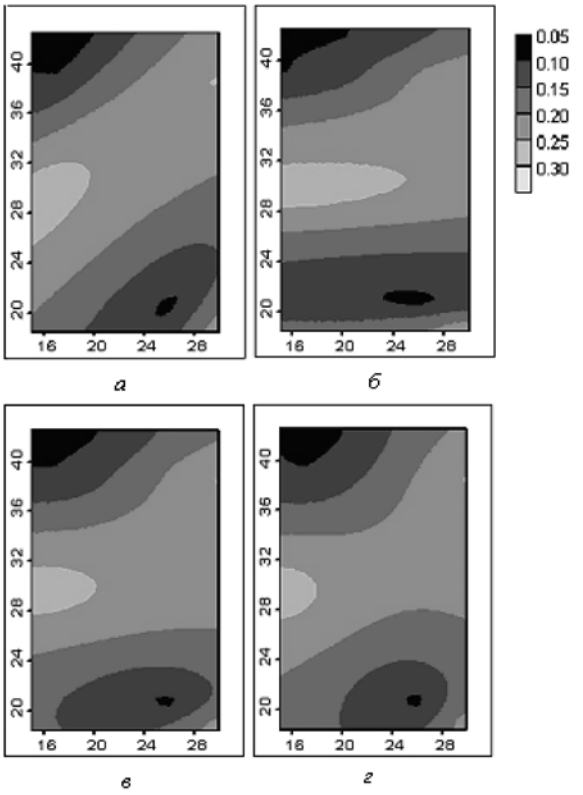


Рис. 5.8. Оценка обычного кригинга при анизотропной модели пространственной корреляционной структуры (вариограмме), направление главной оси — вдоль оси X, радиус корреляции вдоль главной оси — 40 м, значения радиуса корреляции вдоль малой оси: *а* — 5 м; *б* — 10 м; *в* — 20 м; *г* — 30 м

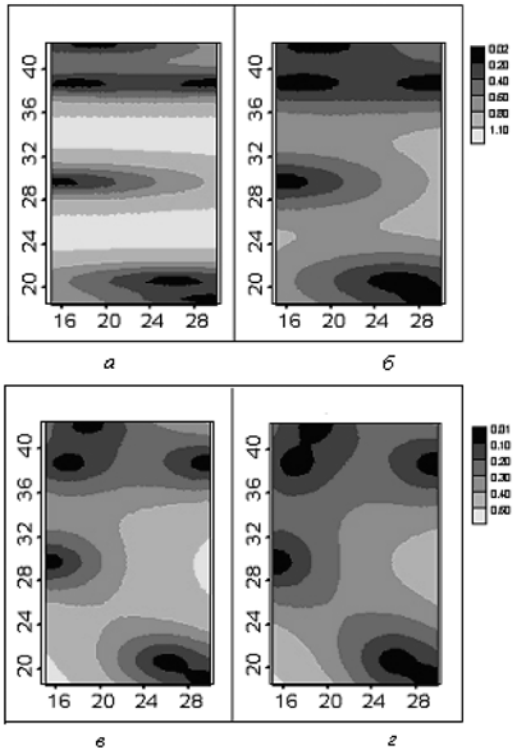


Рис. 5.9. Вариация обычного кригинга при анизотропной модели пространственной корреляционной структуры (вариограмме), направление главной оси — вдоль оси X, радиус корреляции вдоль главной оси — 40 м, значения радиуса корреляции вдоль малой оси: *а* — 5 м; *б* — 10 м; *в* — 20 м; *г* — 30 м

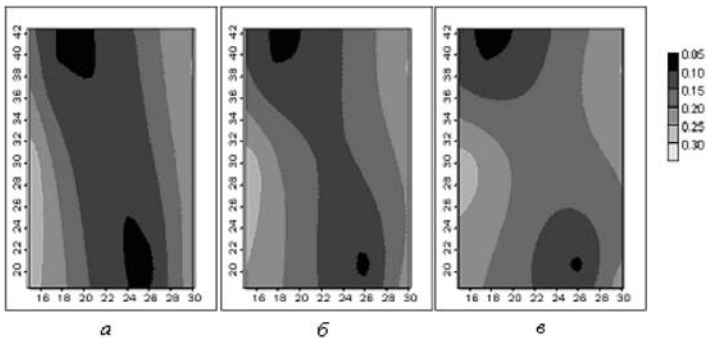


Рис. 5.10. Оценка обычного кригинга при анизотропной модели пространственной корреляционной структуры (вариограмме), направление главной оси — вдоль оси Y, радиус корреляции вдоль главной оси — 40 м, значения радиуса корреляции вдоль малой оси: *а* — 10 м; *б* — 20 м; *в* — 30 м

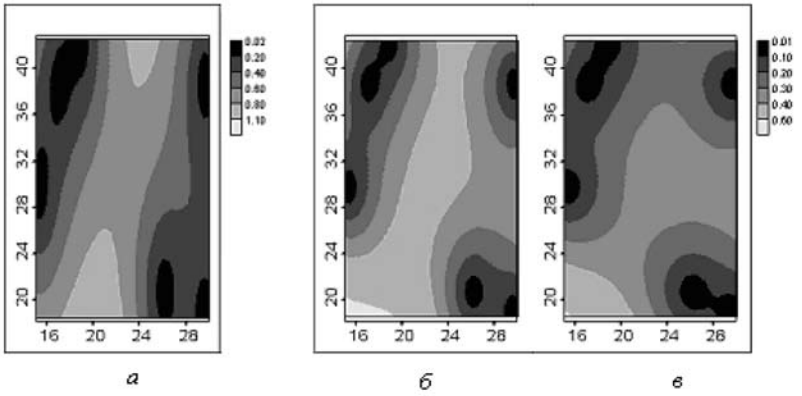


Рис. 5.11. Вариация обычного кригинга при анизотропной модели пространственной корреляционной структуры (вариограмме), направление главной оси — вдоль оси Y, радиус корреляции вдоль главной оси — 40 м, значения радиуса корреляции вдоль малой оси: *а* — 10 м; *б* — 20 м; *в* — 30 м

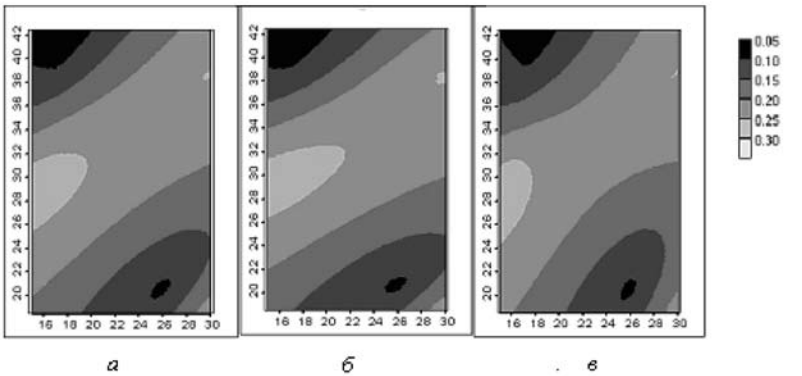


Рис. 5.12. Оценка обычного кригинга при анизотропной модели пространственной корреляционной структуры (вариограмме), радиус корреляции вдоль главной оси — 40 м, радиус корреляции вдоль малой оси — 20 м, направление главной оси: *а* — 45°; *б* — 30°; *в* — 60° (от оси X против часовой стрелки)

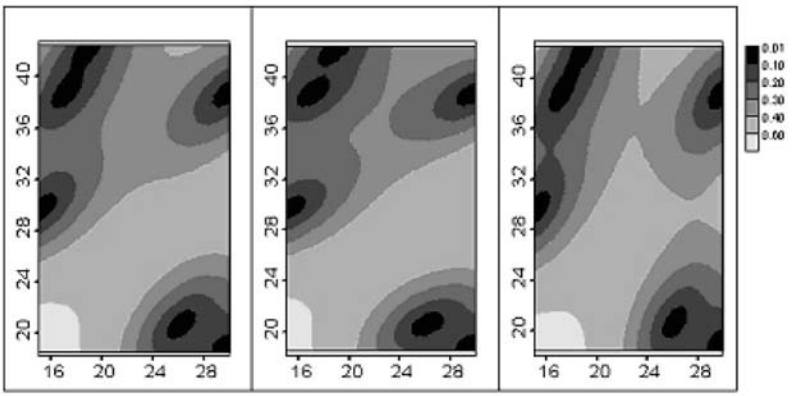


Рис. 5.13. Вариация обычного кригинга при анизотропной модели пространственной корреляционной структуры (вариограмме), радиус корреляции вдоль главной оси — 40 м, радиус корреляции вдоль малой оси — 20 м, направление главной оси: *а* — 45°, *б* — 30°; *в* — 60° (от оси X против часовой стрелки)

На рис. 5.14 и 5.15 показано влияние на оценку и вариацию кригинга вариации наггета. Во всех рассмотренных до этого случаях значение данного параметра было равно нулю. Здесь рассматривается случай с изотропной моделью пространственной корреляции, радиус корреляции равен 5. При таком радиусе корреляции большая часть оцениваемой области соответствует среднему значению оценки. Это усредненное значение больше всего подвержено влиянию наггета. Рассмотрены случаи со значением наггета 0,5 и 0,9. Картина оценки при этом меняется не очень существенно, но можно заметить, что зоны низких и высоких значений (вокруг точек) уменьшаются, как при уменьшении радиуса корреляции. А значение вариации кригинга существенно возрастает при увеличении этого параметра (можно сравнить шкалы на рис. 5.15*а* и 5.15*б*). Наггет характеризует уровень неопределенности модели пространственной корреляции.

Таким образом, очевидно, что результат применения обычного кригинга во многом определяется качеством подобранной модели пространственной корреляционной структуры, поэтому ее оценке и моделированию должно быть уделено существенное внимание.

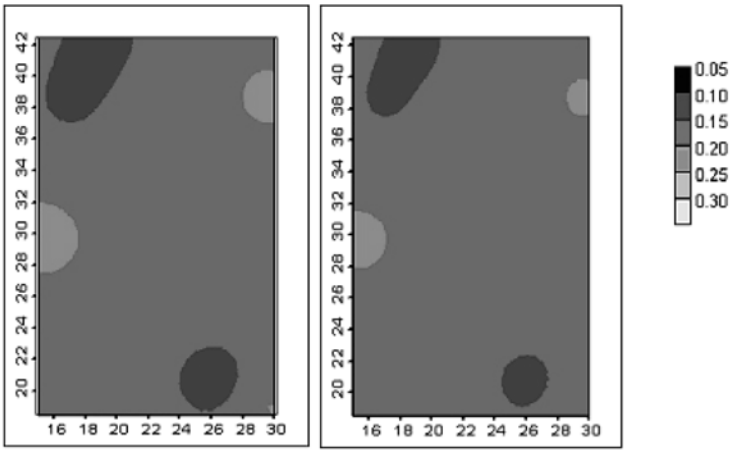


Рис. 5.14. Оценка обычного кригинга при изотропной модели пространственной корреляционной структуры (вариограмме), радиус корреляции 5 м, наггет: *а* — 0,5; *б* — 0,9

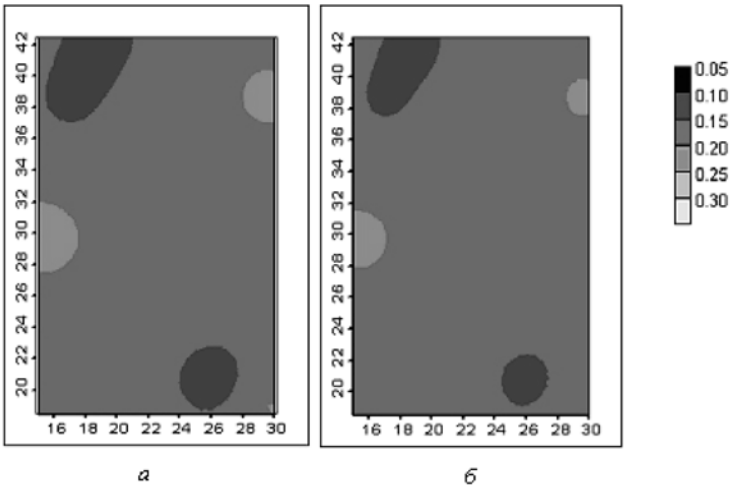


Рис. 5.15. Вариация обычного кригинга при изотропной модели пространственной корреляционной структуры (вариограмме), радиус корреляции 5 м, наггет: *а* — 0,5; *б* — 0,9

5.4. Универсальный кригинг

Универсальный кригинг, или кригинг с трендом (universal kriging, UK), предполагает, что неизвестное среднее значение $m(\boldsymbol{x})$ плавно меняется во всей области исследования S. В некоторых случаях невозможно предположить локальное постоянство среднего даже в окрестности оцениваемой точки $W(\boldsymbol{x})$. Одним из возможных в таком случае подходов является именно универсальный кригинг. Предполагается, что детерминистическая компонента случайной переменной (тренд) моделируется как линейная комбинация $K+1$ базисных (известных) функций $f_k(\boldsymbol{x})$ (по принятому соглашению $f_0(\boldsymbol{x})=1$) с коэффициентами $a_k(\boldsymbol{x})$, неизвестными и постоянными внутри окрестности оцениваемой точки x $W(\boldsymbol{x})$:

$$m(\boldsymbol{x}')=\sum_{k=0}^{K} a_k(\boldsymbol{x}')f_k(\boldsymbol{x}'),\quad a_k(\boldsymbol{x}')=a_k,\quad \forall x'\in W(\boldsymbol{x}). \tag{5.20}$$

Рассмотрим, как выполнить в таком случае условие несмещенности оценки:

$$E\left\{Z^*_{UK}(\boldsymbol{x})-Z(\boldsymbol{x})\right\}=\sum_{i=1}^{n(\boldsymbol{x})}\lambda_i(\boldsymbol{x})m(\boldsymbol{x}_i)-m(\boldsymbol{x})=$$
$$=\sum_{i=1}^{n(\boldsymbol{x})}\lambda_i(\boldsymbol{x})\sum_{k=0}^{K}a_k(x)f_k(x_i)-\sum_{k=0}^{K}a_k(\boldsymbol{x})f_k(\boldsymbol{x})=$$
$$=\sum_{k=0}^{K}a_k(\boldsymbol{x})\left[\sum_{i=1}^{n(\boldsymbol{x})}\lambda_i(\boldsymbol{x})f_k(\boldsymbol{x}_i)-f_k(\boldsymbol{x})\right].$$

Получаем набор из $K+1$ дополнительных ограничений, но избавляемся от необходимости оценивать коэффициенты $a_k(\boldsymbol{x})$:

$$\sum_{i=1}^{n(\boldsymbol{x})}\lambda_i(\boldsymbol{x})=1,$$
$$\sum_{i=1}^{n(\boldsymbol{x})}\lambda_i(\boldsymbol{x})f_k(\boldsymbol{x}_i)=f_k(\boldsymbol{x}),\quad k=1,\ldots,K.$$

Построив соответствующий лагранжиан, продифференцировав его по всем неизвестным переменным и приравняв к нулю соответствующие производные, получаем систему уравнений универсального кригинга:

$$\begin{cases} \sum_{j=1}^{n(\boldsymbol{x})} \lambda_j(\boldsymbol{x}) C_R\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) + \sum_{k=0}^{K} \mu_k(\boldsymbol{x}) f_k(\boldsymbol{x}_i) = C_R\left(\boldsymbol{x}_i - \boldsymbol{x}\right), \ i = 1, ..., n(\boldsymbol{x}), \\ \sum_{j=1}^{n(\boldsymbol{x})} \lambda_j(\boldsymbol{x}) = 1, \\ \sum_{j=1}^{n(\boldsymbol{x})} \lambda_j(\boldsymbol{x}) f_k(\boldsymbol{x}_j) = f_k(\boldsymbol{x}), \ k = 1, ..., K. \end{cases} \quad (5.21)$$

Вариация универсального кригинга также может быть записана следующим образом:

$$\sigma^2_{UK}(\boldsymbol{x}) = C_R(0) - \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) C_R\left(\boldsymbol{x}_i - \boldsymbol{x}\right) - \sum_{k=0}^{K} \mu_k(\boldsymbol{x}) f_k(\boldsymbol{x}).$$

Здесь следует обратить внимание, что в системе уравнений универсального кригинга используются ковариации $C_R(\cdot)$ для случайной компоненты $R(\boldsymbol{x})$ функции $Z(\boldsymbol{x})$, априорное знание которых предполагается. Кроме того, требуется априорное знание набора базисных функций. Подробнее ознакомиться с тем, как на практике подходят к решению задач подготовки модели тренда и ковариационной функции остатков, можно, например, в [Goovaerts, 1997; Armstrong, 1984]. Но, вообще говоря, универсальный кригинг не получил широкого распространения, так как задача подбора функций для моделирования тренда не является прозрачной.

5.5. Логнормальный кригинг

Логнормальным случайным процессом $\left\{Z(\boldsymbol{x}) : \boldsymbol{x} \in S\right\}$ называется такой положительно-определенный процесс, когда $Y(\boldsymbol{x})$ является гауссовым процессом:

$$Y(\boldsymbol{x}) \equiv \log Z(\boldsymbol{x}) \sim N\left(m(\boldsymbol{x}), \sigma^2\right), \ \boldsymbol{x} \in S.$$

Пусть $Y^*(\boldsymbol{x}_0)$ — оценка функции $Y(\boldsymbol{x})$ в точке $\boldsymbol{x}_0$, где нет измерения, полученная с помощью кригинга на основании известных данных $Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n)$ в точках измерений $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$. Теперь требуется получить оценку функции $Z(\boldsymbol{x})$ в этой точке. Если получать такую оценку, просто делая обратное логарифму преобразование

$$Z^*(\boldsymbol{x}_0) = \exp\left\{Y^*(\boldsymbol{x}_0)\right\},$$

то она будет смещенной [Dowd, 1982], т. е.

$$E\left\{Z^*(\boldsymbol{x}_0)\right\} \neq E\left\{Z(\boldsymbol{x}_0)\right\}.$$

А цель любого кригинга, в том числе и логнормального, — наилучшая несмещенная оценка. Поэтому, чтобы учесть и исправить возникающее при обратном преобразовании смещение, используется формула

$$Z^*(\boldsymbol{x}_0) = \exp\left\{Y^*(\boldsymbol{x}_0) + \frac{1}{2}\left[\mathrm{Var}\{Y(\boldsymbol{x}_0)\} - \mathrm{Var}\{Y^*(\boldsymbol{x}_0)\}\right]\right\}.$$

Здесь вариация $\mathrm{Var}\left\{Y^*(\boldsymbol{x}_0)\right\}$ представляет собой не вариацию обычного кригинга $\sigma^2_{OK}(\boldsymbol{x}_0)$, а значение, определяемое как

$$\mathrm{Var}\left\{Y^*(\boldsymbol{x}_0)\right\} = \lambda^{\mathrm{T}}\Sigma\lambda,$$

где Σ — вариационно-ковариационная матрица значений $Y = \left[Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), \ldots, Y(\boldsymbol{x}_n)\right]^{\mathrm{T}}$. Таким образом, несмещенная оценка $Z(\boldsymbol{x}_0)$ может быть получена на основании следующей формулы:

$$\begin{aligned} Z^*(\boldsymbol{x}_0) &= \exp\left\{Y^*(\boldsymbol{x}_0) + \frac{1}{2}\left[\mathrm{Var}\{Y(\boldsymbol{x}_0)\} - \mathrm{Var}\{Y^*(\boldsymbol{x}_0)\}\right]\right\} = \\ &= \exp\left\{Y^*(\boldsymbol{x}_0) + \frac{1}{2}\sigma^2_{OK}(\boldsymbol{x}_0) - \mu\right\}, \end{aligned} \qquad (5.22)$$

где μ — множитель Лагранжа, значение которого находится при решении системы уравнений обычного кригинга для оценки $Y^*(\boldsymbol{x}_0)$.

На практике логнормальный кригинг обычно используется для данных, где значения различаются на порядки. Такое сильное различие не дает возможности получить модель пространственной корреляции. Нелинейное логарифмическое преобразование делает данные пригодными для геостатистики.

Пример данных для логнормального кригинга

Примером могут служить данные по пространственному распределению крабов. Они получены от Всероссийского научно-исследовательского института рыбного хозяйства и океанографии (ВНИРО) и представляют собой результат траловой съемки краба Берди в Беринговом море в 2003 г. Для крабов характерны огромные скопления на фоне практически нулевых значений. Разброс значений на пять порядков не дает возможности оценивать пространственную корреляцию с помощью вариограммы (рис. 5.16).

Рис. 5.16. Экспериментальная вариограмма для данных траловой съемки пространственного распределения краба Берди

После логарифмического преобразования данные вполне пригодны для геостатистического анализа и моделирования — вариограмма на рис. 5.17 легко моделируется сферической моделью.

Рис. 5.17. Экспериментальная вариограмма для данных траловой съемки пространственного распределения краба Берди после логарифмического преобразования

Логарифмически преобразованные данные не всегда соответствуют требованию нормальности. Но невыполнение этого условия нарушает корректность обратного преобразования (5.22), что может привести к сомнительной по качеству оценке. В случае с крабами корректность соблюдается, их распределение отлично соответствует логнормальному (рис. 5.18).

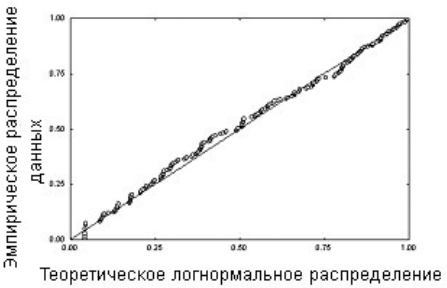


Рис. 5.18. График соответствия эмпирического распределения теоретическому (probability-probability plot) для данных траловой съемки пространственного распределения краба Берди

5.6. Некоторые дополнительные аспекты кригинга

5.6.1. Веса кригинга и эффект экранирования

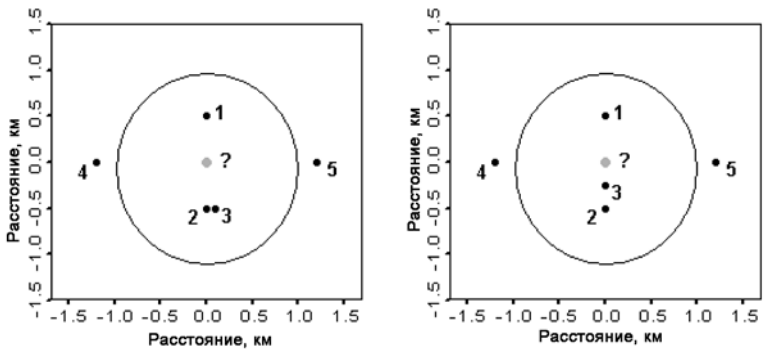
Выше (при обсуждении основных свойств простого кригинга — см. раздел 5.2) упоминалось, что веса кригинга не зависят от известных значений функции, а определяются моделью пространственной корреляции. Отметим еще несколько свойств весов кригинга, которые позволят лучше понять кригинг как метод.

Веса кригинга зависят от формы функции пространственной корреляции (относительный наггет-эффект, анизотропия, радиус корреляции), но не от глобального значения плато или множителя при значении функции ковариации или вариограммы. Изменение глобального плато или умножение функции ковариации на некоторый множитель для уравнения кригинга — (5.13) или (5.17) — это все равно что умножение обеих сторон системы линейных уравнений на одно число — решение системы при этом не изменится.

Как легко понять (например, из свойств функции ковариации), значения весов кригинга уменьшаются при удалении точки с данными от оцениваемой. Но веса кригинга зависят и от кластерности сети мониторинга. При примерно одинаковом пространственном положении относительно оцениваемой точки (одинаковое расстояние, одинаковое направление в терминах ковариационной функции, но в разные стороны от оцениваемой точки, как, например, точки *1* и *2* на рис. 5.19) две точки будут иметь одинаковые веса, только если картина абсолютно симметрична. Если около одной из них больше соседей, использующихся при оценке (например, около точки *2* на рис. 5.19*а* есть еще точка *3*), то такая избыточность информации уменьшит вес этой точки. Более того, точка может оказаться экранированной, если между нею и оцениваемой находится еще одна (например, точка *2* на рис. 5.19*б* экранирована точкой *3*). В таком случае вес кригинга может стать отрицательным.

При использовании кригинга нужно внимательно относиться к наличию эффекта экранирования и отрицательных весов. Такие веса могут приводить к выпадению оценки из области допустимых значений (например, отрицательные значения концентрации или значение пропорции больше 1).

Можно ставить дополнительное ограничение на положительность весов, но на практике обычно контролируют и вводят ограничения на значения оценки.



а *б*

Рис. 5.19. Два примера пространственной конфигурации точек

Важно отметить, что отсутствие корреляции между оцениваемой точкой и точкой с измерением (нулевое значение функции ковариации) не означает априорного равенства нулю веса кригинга для этого значения. Если точка изолированная (не имеет рядом других) и при этом не экранирована, то она с точки зрения кригинга является существенной (например, точки *4* и *5* на рис. 5.19). Это важно помнить при выборе зоны поиска для оценки — она должна быть больше области корреляции, особенно в случае малой плотности исходных данных.

Все описанные выше свойства можно проверить. Для этого достаточно рассмотреть точки *1—5* на рис. 5.19 как пространственную модель точек для оценки в центре (большой серой точке, помеченной «?»). В качестве ковариационной функции можно, например, использовать изотропную сферическую модель с нулевым наггетом, единичным плато и радиусом корреляции 1 км.

Также следует отметить очень существенное влияние относительного наггет-эффекта на веса кригинга. Увеличение относительного наггет-эффекта ведет к уменьшению влияния расстояния от точки оценивания, он также уменьшает влияние эффекта экранирования. Например, в случае вариограммы типа «чистый наггет» веса при всех точках равны, а оценка кригинга равна арифметическому среднему всех данных из области оценивания.

5.6.2. Вариация кригинга и неопределенность оценки

Модели кригинга позволяют получить оценки локального среднего (оценка кригинга) и локальной вариации (вариация кригинга). Полученную вариацию кригинга можно использовать для описания неопределенности оценки. Если принять гипотезу о мультинормальности случайных переменных, то 95%-ные доверительные интервалы, в которых лежит истинное значение функции $Z(\boldsymbol{x}_0)$, будут определяться как

$$Z(x_0) \in Z^*(x_0) \pm 2\sigma,$$

где σ — стандартное отклонение, полученное из кригинговой вариации.

Если не делать предположения о мультинормальности, то размер 95%-ного доверительного интервала увеличивается [Chiles, Delfiner, 1999] (отсутствие знания увеличивает неопределенность) до 6σ. Это прямое следствие неравенства Высочанского — Петунина, полученного в 1980 г. Единственное предположение является достаточно слабым, предполагается, что распределение ошибки является непрерывным и унимодальным.

Неравенство Высочанского — Петунина формулируется следующим образом: если X — случайная переменная с плотностью распределения f, неубывающей до моды v, а потом невозрастающей, и если α — ожидаемое стандартное отклонение от произвольного значения, то

$$\Pr\left(|X-\alpha| \geq td\right) \leq \begin{cases} \dfrac{4}{9t^2}, & \forall t \leq \sqrt{\dfrac{8}{3}}, \\ \dfrac{4}{3t^2} - \dfrac{1}{3}, & \forall t > \sqrt{\dfrac{8}{3}}. \end{cases}$$

Если X — ошибка кригинга и $\alpha = 0$, то $d^2 = \sigma_{OK}^2$ и $\Pr\left(\left|Z^* - Z\right| \geq 2\sigma_{OK}\right) \leq \dfrac{1}{9}$.

При сформулированных предположениях доверительный интервал является примерно 90%-ным. Чтобы получить 95%-ный интервал, требуется положить $t = 3$, так как $\dfrac{4}{81} = 0,049$.

В любом случае в каждой точке оценивания при использовании моделей кригинга определены три величины: оценка и две границы доверительного интервала. Это позволяет провести изолинии для всех трех величин. Изолиния оценки будет с заданной вероятностью (например, 0,9) лежать между контурами доверительных интервалов, которые образуют «кори-

дор». Ширина «коридора» характеризует неопределенность изолинии оценки. Эта так называемая «толстая» изолиния содержит в себе с определенной вероятностью изолинию действительного распределения, истинное положение которой в точности неизвестно. Пример «толстой» изолинии, представляющей 90%-ный доверительный интервал ($\pm 2\sigma_{ОК}$), приведен на рис. 5.20. Этот пример относится к анализу данных по загрязнению поверхности в результате Чернобыльского выброса в мае 1986 г. [Kanevski et al., 1999]. Результаты измерений картированы с использованием обычного кригинга. Изолиния относится к одному из критических уровней загрязнения — 1000 Ки/км^2.

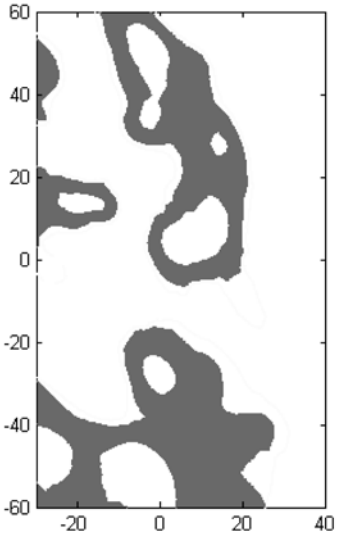
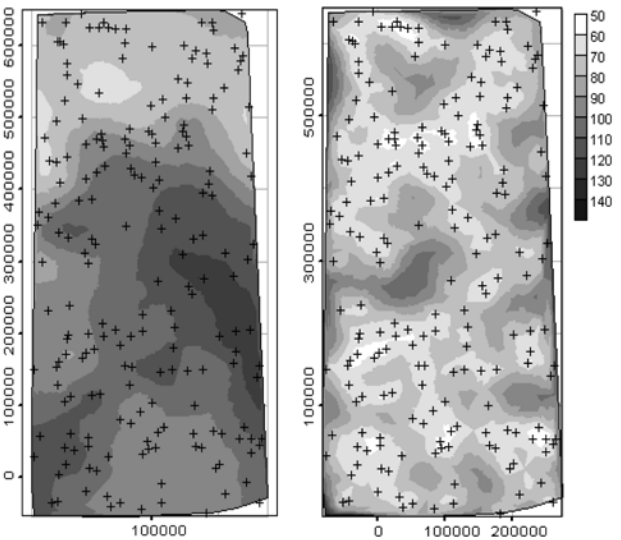


Рис. 5.20. «Толстые» изолинии, представляющие 90%-ный доверительный интервал для изолинии 1000 Ки/км^2 по результатам картирования обычным кригингом загрязнения поверхности ^{137}Cs в районе Чернобыльского выброса

С другой стороны, в соответствии с (5.14) для простого кригинга и (5.18) или (5.19) для обычного кригинга, вариация кригинга не зависит от значений исходных данных и зависит от ковариационной функции и конфигурации (взаимного расположения) данных. Это также было отмечено и в искусственном примере на обычный кригинг. Таким образом, кригинг дает одинаковую вариацию для всех точек с аналогичной конфигурацией исходных данных (при использовании одной и той же модели пространственной корреляционной структуры) и для случаев, когда исходные данные близки по значениям и когда они сильно различаются, т.е. фактически вариация

кригинга является характеристикой плотности исходных данных. Это можно наблюдать на рис. 5.21, где изображена вариация кригинга вместе с исходными данными, полученная для оценки кригинга. Видно, что вариация оценки не зависит от значения самой оценки. Зависимость вариации кригинга только от плотности исходных данных позволяет использовать ее для оптимизации сети мониторинга.



а *б*

Рис. 5.21. Пример картирования:
а — оценка кригинга; *б* — вариации кригинга
(крестиками отмечены точки расположения исходных данных)

5.6.3. Учет собственных ошибок измерений в уравнениях кригинга

Бывают случаи, когда измеренные значения сопровождаются ошибкой измерения, полученной, например, как сведения о систематической ошибке прибора или методологии измерений, т.е. вместе со значением функции $Z(x_i)$ имеется еще и ошибка измерения, представленная как локальная вариация $\sigma(x_i) = \sigma_i$. Если сами ошибки не являются пространственно коррелированными, то система уравнений обычного кригинга может быть видоизменена так, чтобы учитывать ошибки при оценке [Kanevski et al., 1993]:

$$\begin{cases} \sum_{j=1}^{n(x)} \lambda_j \gamma_{ij} + \mu - \lambda_i \sigma_i^2 = \gamma_{i0}, \\ \sum_{i=1}^{N} \lambda_i = 1. \end{cases}$$

При этом оценка кригинга будет определяться по традиционной формуле — сумма взвешенных значений с найденными весами, а вариация кригинга будет больше в сравнении с обычным кригингом:

$$\sigma_{OK}^2(x) = \sum_{i=1}^{n(x)} \lambda_i(x)\gamma_{i0} + \sum_{i=1}^{n(x)} \lambda_i^2 \sigma_i^2 + \mu(x).$$

Увеличение неопределенности (кригинговой вариации) вызвано введением дополнительной неопределенности в данные. Обычный кригинг предполагает все данные абсолютно точными.

5.6.4. Блочный и точечный кригинг

Любое измерение $Z(x_\alpha)$ всегда соотносится с некоторым ненулевым конечным объемом. Это может быть кусок породы или почвы, где берется проба. Обычно размер измерения позволяет приписать его к точке с координатой x_α. Размер оцениваемого значения такой же, как и измеряемого, и оценка также приписывается к координате. Но иногда целью оценки является среднее по определенному объему, если именно такой размер соответствует определенным действиям (например, очистке и т. п.). *Блочный кригинг* — обобщенное название метода для определения среднего значения функции *z* по какому-либо измеримому сегменту (длине, площади, объему) любого размера или формы в противоположность *точечному кригингу*, относящемуся к нулевому размеру оценки (пробы).

Блочная оценка может выполняться, например, как усреднение точечных оценок кригинга, попавших в требуемый объем. Другой подход состоит в использовании непосредственно системы уравнений блочного кригинга.

Система уравнений блочного кригинга выглядит точно так, как система уравнений обычного кригинга (5.17). Единственное различие состоит в том, что в правой части системы стоит ковариация не для двух точек, а для блока (объема) и точки $\bar{C}(x_\alpha, V(x))$. Определяется такая ковариация в соответствии с формулой [Journel, Huijbregts, 1978]

$$\bar{C}\left(x_\alpha, V(x)\right) = \mathrm{Cov}\left\{Z(x_\alpha), Z_V(x)\right\} = \frac{1}{|V|}\int_{V(x)} C\left(x_\alpha - x'\right)dx'.$$

На практике такая ковариация определяется как среднее точечных ковариаций точки x_α и N точек x_i, дискретно описывающих объем $V(x)$:

$$\bar{C}\left(x_\alpha, V(x)\right) \cong \frac{1}{N}\sum_{i=1}^{N} C(x_\alpha - x_i).$$

Ковариации могут быть обобщены на случай, когда сами измерения тоже представлены объемом. Это существенно, если размер измерения сопоставим с размером области исследования и точечный кригинг является некорректным.

Литература

Гандин Л. С., Каган Р. Л. Статистические методы интерполяции метеорологических данных. — Л.: Гидрометеоиздат, 1976. — 359 с.

Armstrong M. Problems with universal kriging // Mathematical Geology. — 1984. — Vol. 16. — P. 101—108.

Chiles J. P., Delfiner P. Geostatistics. Modeling Spatial Uncertainty. — New York: A Wiley-Interscience Publication, 1999.

Dowd P. A. Lognormal kriging — the general case // Mathematical Geology. — 1982. — Vol. 14. — P. 475—499.

Goovaerts P. Geostatistics for Natural Resources Evaluation. — [S. l.]: Oxford Univ. Press, 1997. — 483 p.

Journel A. G., Huijbregts C. J. Mining Geostatistics. — London: Academic Press, 1978. — 600 p.

Kanevski M. F., Arutyunyan R. V., Bolshov L. A. et al. Spatial data analysis of Chernobyl fallout data. — 1. Preliminary results / Nuclear Safety Inst. — Moscow, 1993. — 91 p. — (Preprint NSI-23-93).

Kanevski M., Arutyunyan R., Bolshov L. et al. Mapping of Radioactively Contaminated Territories with Geostatistics and Artificial Neural Networks // Contaminated Forests / Eds. I. Linkov and W. R. Schell. — [S. l.]: Kluwer Academic Publ., 1999. — P. 249—256.

Krige D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand // J. of the Chem., Metal. and Mining Soc. of South Africa. — 1951. — Vol. 52. — P. 119—139.

Глава 6

Многопеременное пространственное моделирование

Данная глава посвящена проблемам использования дополнительной информации в рамках классической геостатистики. В Разделе 6.1 рассмотрены особенности использования информации, известной на всей области (кригинг с внешним дрейфом). В Разделах 6.2, 6.3 описаны взаимная пространственная корреляция нескольких переменных и их совместное оценивание (кокригинг). Уменьшение размерности пространства переменных с помощью метода принципиальных компонент и переход к факторному кригингу рассмотрены в Разделе 6.4.

Во многих практических задачах пространственного оценивания измерения основной переменной могут сопровождаться дополнительной информацией, представленной в виде измерений других переменных или внешнего параметра (свойства), заданного на всем поле наблюдений. При определенных условиях дополнительная информация может способствовать оценке основной переменной. Например, если измерений дополнительной переменной больше (скажем, из-за того, что их дешевле проводить), то их использование может позволить проводить оценку в областях, которые для основной переменной были зоной экстраполяции, а при использовании измерений дополнительной переменной становятся зоной интерполяции. Основным условием возможности и полезности использования дополнительной информации является ее коррелированность с основной оцениваемой переменной. Однако избыточное количество дополнительной информации может необоснованно усложнить модель и привести к увеличению ошибки оценки.

6.1. Кригинг с внешним дрейфом

Кригинг с внешним дрейфом (kriging with external drift) можно рассматривать как модификацию универсального кригинга (см. Раздел 5.4). В универсальном кригинге тренд моделируется линейной комбинацией базис-

ных функций (5.20), в то время как в кригинге с внешним дрейфом тренд моделируется как линейная функция гладкой дополнительной переменной $y(\boldsymbol{x})$, внешней по отношению к оцениваемой $Z(\boldsymbol{x})$:

$$m(\boldsymbol{x}) = a_0(\boldsymbol{x}) + a_1(\boldsymbol{x})y(\boldsymbol{x}).$$

Неизвестные коэффициенты предполагаются постоянными в окрестности оцениваемой точки и вычисляются при решении кригинговой системы уравнений. В оценку $Z^*(\boldsymbol{x})$ внешний дрейф входит опосредованно через влияние на весовые коэффициенты $\lambda_j(\boldsymbol{x}), j = 1, ..., n(u)$:

$$Z^*(\boldsymbol{x}) = \sum_{j=1}^{n(\boldsymbol{x})} \lambda_j(\boldsymbol{x}) Z(\boldsymbol{x}_j).$$

Система уравнений кригинга с внешним дрейфом при этом имеет вид

$$\begin{cases} \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) C_R(\boldsymbol{x}_j - \boldsymbol{x}_i) + \mu_0(\boldsymbol{x}) + \mu_1(\boldsymbol{x}) y(\boldsymbol{x}_j) = C_R(\boldsymbol{x}_j - \boldsymbol{x}), \; j = 1, ..., n(\boldsymbol{x}), \\ \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) = 1, \\ \sum_{i=1}^{n(\boldsymbol{x})} \lambda_i(\boldsymbol{x}) y(\boldsymbol{x}_i) = y(\boldsymbol{x}), \end{cases}$$

где $C_R(\boldsymbol{x}_j - \boldsymbol{x})$ — ковариационная функция невязки $R(\boldsymbol{x}) = Z(\boldsymbol{x}) - m(\boldsymbol{x})$. Система уравнений кригинга с внешним дрейфом является частным случаем системы уравнений универсального кригинга (5.21), если $K = 1$ и $f_l(\boldsymbol{x})$ в любой точке совпадает со значением вторичной переменной $y(\boldsymbol{x})$.

Для использования кригинга с внешним дрейфом требуется выполнение следующих условий.

- Между трендом оцениваемой переменной и вторичной переменной должна быть линейная зависимость. При наличии другого типа зависимости можно провести некоторое преобразование, чтобы сделать ее линейной.
- Значение вторичной переменной должно быть доступно в любой точке исследуемой области: в точках измерения и в точках оценивания основной переменной.
- Значение вторичной переменной должно достаточно гладко изменяться на исследуемой области, чтобы не вызывать нестабильности системы уравнений.

- Должна быть возможность оценки и моделирования ковариации $C_R(\boldsymbol{x}_j - \boldsymbol{x})$ или вариограммы $\gamma_R(\boldsymbol{x}_j - \boldsymbol{x})$ остатков по значениям реальных измерений, так как сами остатки становятся известны только после решения кригинговой системы. Это требование непосредственно связано с предыдущим о плавности изменений вторичной переменной — $y(\boldsymbol{x}_j) \approx y(\boldsymbol{x}_j + h)$.

Примером использования кригинга с внешним дрейфом может служить моделирование поля температуры при наличии дополнительной информации о высоте над уровнем моря на подробной сетке (практически в любой точке). Такая информация доступна в виде цифровой карты уровней (digital elevation map). На рис. 6.1 приведены карта пространственного распределения данных о температуре и цифровая карта уровней для этого места. Рисунок 6.2 показывает практически линейную зависимость между температурой и высотой. Такая ярко выраженная линейная зависимость позволяет использовать данные о высоте как внешний дрейф для температуры. Применение обычного кригинга на специально выбранном из исходных данных валидационном наборе дало среднеквадратичную ошибку 3,13. Использование дополнительной информации позволило уменьшить среднеквадратичную ошибку на валидационном наборе до 1,42.

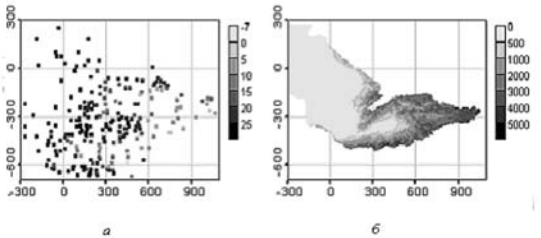


Рис. 6.1. Точки измерений температуры на поверхности (*а*), цифровая карта уровней для этого места (*б*)

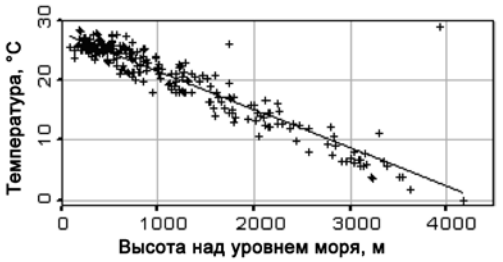


Рис. 6.2. Зависимость между температурой воздуха у поверхности и высотой над уровнем моря

6.2. Меры корреляции и пространственной корреляции нескольких переменных

Теперь рассмотрим общий случай измерений нескольких параметров, интерпретируя их как многопеременную случайную функцию. Итак, пусть $Z_\alpha(x_i)$ — многопеременная функция, заданная на области S ($i = 1, ..., n$; $\alpha = 1, ..., K$), где i — индекс, означающий номер измерения (пространственной точки); α — индекс, означающий номер переменной. Все измерения всех переменных можно представить в виде матрицы $\mathbf{Z}$ размерностью $K \times n$.

Корреляция переменных описывается ковариационной матрицей:

$$C_{\mathbf{Z}} = \left[\sigma_{ij}\right] = \frac{1}{K} E\left\{\left[\mathbf{Z} - \boldsymbol{m}\right]^{\mathrm{T}} \left[\mathbf{Z} - \boldsymbol{m}\right]\right\}, \tag{6.1}$$

где $\boldsymbol{m}$ — вектор средних значений отдельных переменных.

Диагональ ковариационной матрицы соответствует собственным вариациям переменных, остальные элементы характеризуют ковариации пары переменных. Ковариационная матрица статистически описывает взаимную связь различных пар переменных многопеременной функции, а также используется для выделения главных компонент многопеременных векторов. Все такого рода типы многопеременного анализа обладают недостатком с точки зрения геостатистики (пространственной статистики) — они никак не учитывают и не используют пространственное расположение точек измерения, т. е. не дают возможности описывать, оценивать и использовать пространственные связи между различными переменными.

Для описания пространственной корреляции пар переменных используют кросс-ковариацию или кросс-вариограмму.

Кросс-ковариация (cross-covariance). Для $N(\boldsymbol{h})$ экспериментальных точек, разделенных вектором $\boldsymbol{h}$, в которых есть измерения обеих переменных $Z_\alpha(\boldsymbol{x})$ и $Z_\beta(\boldsymbol{x})$, кросс-ковариация определяется как [Dowd, 1989]

$$C_{\alpha\beta}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} Z_\alpha(\boldsymbol{x}) Z_\beta(\boldsymbol{x} + \boldsymbol{h}) - m_{Z^{\alpha}-\boldsymbol{h}} m_{Z^{\beta}+\boldsymbol{h}}, \tag{6.2}$$

где $\boldsymbol{m}_{Z^{\alpha-h}}$ — среднее значение переменной Z_α по началу вектора $\boldsymbol{h}$, а $\boldsymbol{m}_{Z^{\beta+h}}$ — среднее значение переменной Z_β по концам вектора $\boldsymbol{h}$.

Кросс-ковариационная функция не является априори ни симметричной, ни инвариантной относительно перестановки переменных:

$$C_{\alpha\beta}(\boldsymbol{h}) \neq C_{\beta\alpha}(\boldsymbol{h}) \text{ или } C_{\alpha\beta}(\boldsymbol{h}) \neq C_{\alpha\beta}(-\boldsymbol{h}).$$

Но она сохраняется при одновременной перестановке переменных и замене разделяющего вектора на симметричный:

$$C_{\alpha\beta}(\boldsymbol{h}) = C_{\alpha\beta}(-\boldsymbol{h}).$$

В общем случае кросс-ковариационная функция не является положительно определенной: ее максимум может быть смещен относительно нуля на некоторое расстояние r. Такое смещение максимума корреляции очень распространено в многопеременных временны́х рядах, когда одна из переменных воздействует на другую, но не мгновенно. Время, которое требуется второй переменной на реакцию на изменение первой, называется временем задержки. Оно и вызывает сдвиг максимума кросс-ковариации из нуля.

Однако при использовании многопеременных геостатистических оценивателей на кросс-ковариацию накладывается требование положительной определенности (для несингулярности кокригинговой матрицы). В случае пространственных переменных эффект задержки встречается крайне редко, поэтому обычно проблемы такого рода не возникают. При работе с временны́ми рядами и пространственно-временны́ми функциями нужно внимательно следить за корректностью использования для кросс-ковариационной функции положительно определенной модели.

Кросс-вариограмма (cross-variogram). Кросс-вариограмма для $N(\boldsymbol{h})$ экспериментальных точек, разделенных вектором $\boldsymbol{h}$, в которых есть измерения обеих переменных $Z_\alpha(\boldsymbol{x})$ и $Z_\beta(\boldsymbol{x})$, определяется по формуле

$$\gamma_{\alpha\beta}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}\left[Z_\alpha(\boldsymbol{x}_i) - Z_\alpha(\boldsymbol{x}_i+\boldsymbol{h})\right]\left[Z_\beta(\boldsymbol{x}_i) - Z_\beta(\boldsymbol{x}_i+\boldsymbol{h})\right]. \quad (6.3)$$

Как и обычная вариограмма, кросс-вариограмма обладает симметрией, кроме того, она инвариантна относительно перестановки переменных:

$$\gamma_{\alpha\beta}(\boldsymbol{h}) = \gamma_{\alpha\beta}(-\boldsymbol{h}) = \gamma_{\beta\alpha}(\boldsymbol{h}) = \gamma_{\beta\alpha}(-\boldsymbol{h}).$$

Если построить матрицу $\Gamma(\boldsymbol{h}_0)$ из обычных вариограмм и кросс-вариограмм для фиксированного вектора аналогично ковариационной матрице (6.1), то она обычно является отрицательно определенной, так как это вариационно-ковариационная матрица приращений.

На практике измерения, относящиеся к различным компонентам вектора функций $Z_\alpha(\boldsymbol{x})$, могут быть проведены в разных точках пространства (в разное время). Различают следующие возможные случаи пространственного взаимного расположения двух переменных:

- полная гетеротопия (complete heterotopy) — измерения переменных находятся в различных точках и не имеют ни одной общей точки;
- частичная гетеротопия (partial heterotopy) — переменные имеют и общие, и различные точки измерений;
- изотопия (isotopy) — в каждой точке сети мониторинга есть данные измерений по всем переменным.

В случае полной гетеротопии возникает проблема с корреляционной моделью, поскольку экспериментальные кросс-вариограммы невозможно вычислить при отсутствии общих точек измерений [Myers, 1991]. Однако можно воспользоваться экспериментальными кросс-ковариациями, хотя они не относятся к тем же точкам, что и соответствующие значения собственной ковариации.

Другим путем при полной гетеротопии может быть использование *псевдокросс-вариограммы*, которая учитывает значения дополнительной переменной в точках, в которых отсутствуют данные по основной переменной [Papritz et al., 1993]. Для учета всех имеющихся данных в совпадающих и не совпадающих точках измерений переменных даже в случае полной гетеротопии псевдокросс-вариограмму можно вычислить следующим образом:

$$g_{\alpha\beta}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \left[Z_{\alpha}(\boldsymbol{x}_i) - Z_{\beta}(\boldsymbol{y}_i) \right]^2, \qquad (6.4)$$

где $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ — векторы координат данных, разделенных вектором $\boldsymbol{h}$ (см. Раздел 4.3).

Псевдокросс-вариограмма может быть выражена через обыкновенную кросс-вариограмму:

$$g_{\alpha\beta}(\boldsymbol{h}) = \gamma_{\alpha\beta}(\boldsymbol{h}) + 0,5\left[C_{\alpha\alpha}(0) - 2C_{\alpha\beta}(0) + C_{\beta\beta}(0) \right].$$

Значения псевдокросс-вариограммы $g_{\alpha\beta}$ могут быть использованы в уравнениях кокригинга (см. Раздел 5.3), как и обыкновенная кросс-вариограмма $\gamma_{\alpha\beta}$.

В случае частичной гетеротопии рекомендуется строить кросс-вариограмму или кросс-ковариацию на основе изотопного поднабора измерений, вычлененного из полного объема исходных данных, если размер этого поднабора позволяет провести статистически достоверную оценку. Если возможности выделить такой поднабор нет, можно воспользоваться описанной выше псевдокросс-вариограммой.

Вычисление экспериментальных кросс-ковариаций и кросс-вариограмм производится, как и соответствующих функций, для одной переменной (для набора лагов и направлений, с заданием отклонений по лагу и направлению). По экспериментальной кросс-вариограмме подбирается теоретическая модель — параметры к соответствующей формуле. Лаги, направления, отклонения и типы теоретических моделей подробно обсуждались в Главе 4.

6.3. Линейная модель корегионализации

Есть два подхода к упрощению анализа данных по многим переменным и замене его анализом однопеременного набора данных. В первом случае проводится анализ одной или более линейных комбинаций компонент и требуется определить подходящие линейные комбинации. При другом подходе представляют различные компоненты как линейную комбинацию некоррелированных частей, так что каждая из них может быть проанализирована отдельно. Оба эти не связанных между собой подхода используются в современной геостатистике.

Линейная модель корегионализации предполагает, что каждая компонента вектора случайной функции может быть представлена как линейная комбинация некоррелированных компонент. Они обычно представляются в виде моделей ковариации или вариограмм одного типа, но с разными радиусами корреляции. Преимущество линейной модели корегионализации состоит в том, что условие положительной определенности сводится к проверке положительной определенности постоянной матрицы. Она наиболее полезна при малом количестве измерений. Ее недостаток — ограниченный выбор моделей кросс-вариограмм и кросс-ковариаций. Другими словами, если каждая компонента представлена как линейная комбинация некоррелированных компонент, то это соответствует диагонализации матрицы структурной функции [Myers, 1995].

Рассмотрим вектор значений случайной функции $Z(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), \ldots, Z_m(\boldsymbol{x})]$. Пусть $Y_1(\boldsymbol{x}), \ldots, Y_p(\boldsymbol{x})$ — некоррелированные случайные функции, где p может быть больше или меньше m. Стационарность компонент Y следует из стационарности компонент Z и наоборот. Предположим, что

$$Z_i(\boldsymbol{x}) = \sum_k Y_k(\boldsymbol{x}) a_{kj}(\boldsymbol{x}),\ j = 1, \ldots, m. \qquad (6.5)$$

Выражение (6.5) можно представить в матричной форме:

$$\left[Z_1(\boldsymbol{x}), ..., Z_m(\boldsymbol{x})\right] = \left[Y_1(\boldsymbol{x}), ..., Y_p(\boldsymbol{x})\right] \cdot \mathbf{A}. \quad (6.6)$$

Пространственные функции Z соотносятся с пространственными функциями Y следующим образом:

$$C_Z(\boldsymbol{h}) = \boldsymbol{A}^{\mathrm{T}} C_Y(\boldsymbol{h}) A,$$

$$\gamma_Z(\boldsymbol{h}) = \boldsymbol{A}^{\mathrm{T}} \gamma_Y(\boldsymbol{h}) A,$$

где $C_Z(\boldsymbol{h})$, $C_\gamma(\boldsymbol{h})$ — ковариационные матрицы с компонентами:

$$C_{st,Z}(\boldsymbol{h}) = \mathrm{Cov}\{Z_s(\boldsymbol{x}+\boldsymbol{h}), Z_t(\boldsymbol{x})\}, \quad (6.7)$$

$$C_{uv,Y}(\boldsymbol{h}) = \mathrm{Cov}\{Y_u(\boldsymbol{x}+\boldsymbol{h}), Y_v(\boldsymbol{x})\} = 0, \text{ при } u \neq v, \quad (6.8)$$

а $\gamma_Z(\boldsymbol{h})$, $\gamma_Y(\boldsymbol{h})$ — матрицы вариограмм с компонентами:

$$\gamma_{st,Z}(\boldsymbol{h}) = \mathrm{Cov}\{Z_s(\boldsymbol{x}+\boldsymbol{h}) - Z_s(\boldsymbol{x}), Z_t(\boldsymbol{x}+\boldsymbol{h}) - Z_t(\boldsymbol{x})\}, \quad (6.9)$$

$$\gamma_{uv,Y}(\boldsymbol{h}) = \mathrm{Cov}\{Y_u(\boldsymbol{x}+\boldsymbol{h}) - Y_u(\boldsymbol{x}), Y_v(\boldsymbol{x}+\boldsymbol{h}) - Y_v(\boldsymbol{x})\} = 0, \text{ при } u \neq v. \quad (6.10)$$

Заметим, что в уравнениях для вариограмм (6.9) и (6.10) $\gamma_Z(h)$ и $\gamma_Y(h)$ являются стандартными кросс-вариограммами, а не приведенными выше псевдокросс-вариограммами — см. (6.4) [Myers, 1991].

Из (6.7) и (6.8) для ковариации получим

$$C_{st,Z}(\boldsymbol{h}) = \sum a_{su} C_{uu,Y}(\boldsymbol{h}) a_{ut} = \sum b_{st}^u C_{uu,Y}(\boldsymbol{h}), \quad (6.11)$$

или

$$C_Z(\boldsymbol{h}) = \sum B^u C_{uu,Y}(\boldsymbol{h}). \quad (6.12)$$

Линейная модель корегионализации обычно записывается в виде (6.12). По этому построению коэффициенты B^u будут автоматически удовлетворять требуемому условию положительной определенности.

Для случая двух переменных U и V модели авто- и кросс-вариограмм строятся следующим образом:

$$\gamma_U(\boldsymbol{h}) = u_0\gamma_0(\boldsymbol{h}) + u_1\gamma_1(\boldsymbol{h}) + \ldots + u_m\gamma_m(\boldsymbol{h}),$$

$$\gamma_v(\boldsymbol{h}) = v_0\gamma_0(\boldsymbol{h}) + v_1\gamma_1(\boldsymbol{h}) + \ldots + v_m\gamma_m(\boldsymbol{h}), \qquad (6.13)$$

$$\gamma_{UV}(\boldsymbol{h}) = w_0\gamma_0(\boldsymbol{h}) + w_1\gamma_1(\boldsymbol{h}) + \ldots + w_m\gamma_m(\boldsymbol{h}),$$

где $\gamma_U(\boldsymbol{h})$, $\gamma_v(\boldsymbol{h})$, $\gamma_{UV}(\boldsymbol{h})$ — авто- и кросс-вариограммные модели для U и V соответственно. Базисные модели задаются $\gamma_0(\boldsymbol{h})$, $\gamma_1(\boldsymbol{h})$, …, $\gamma_m(\boldsymbol{h})$; u, v, w — коэффициенты, возможно отрицательные.

Уравнения (6.13) можно записать в матричной форме:

комбинация первых базисных моделей $\gamma_0(\boldsymbol{h})$:

$$\begin{bmatrix} \gamma_{U,0}(\boldsymbol{h}) & \gamma_{UV,0}(\boldsymbol{h}) \\ \gamma_{VU,0}(\boldsymbol{h}) & \gamma_{V,0}(\boldsymbol{h}) \end{bmatrix} = \begin{bmatrix} u_0 & w_0 \\ w_0 & v_0 \end{bmatrix} \cdot \begin{bmatrix} \gamma_0(\boldsymbol{h}) & 0 \\ 0 & \gamma_0(\boldsymbol{h}) \end{bmatrix}; \qquad (6.14)$$

комбинация вторых базисных моделей $\gamma_1(\boldsymbol{h})$:

$$\begin{bmatrix} \gamma_{U,1}(\boldsymbol{h}) & \gamma_{UV,1}(\boldsymbol{h}) \\ \gamma_{VU,1}(\boldsymbol{h}) & \gamma_{V,1}(\boldsymbol{h}) \end{bmatrix} = \begin{bmatrix} u_1 & w_1 \\ w_1 & v_1 \end{bmatrix} \cdot \begin{bmatrix} \gamma_1(\boldsymbol{h}) & 0 \\ 0 & \gamma_1(\boldsymbol{h}) \end{bmatrix}; \qquad (6.15)$$

комбинация m-х базисных моделей $\gamma_m(\boldsymbol{h})$:

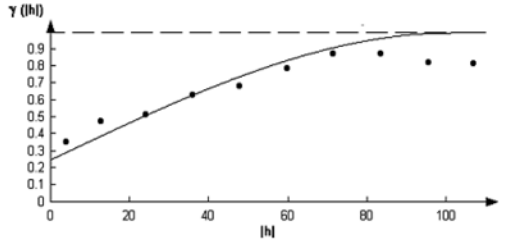
$$\begin{bmatrix} \gamma_{U,m}(\boldsymbol{h}) & \gamma_{UV,m}(\boldsymbol{h}) \\ \gamma_{VU,m}(\boldsymbol{h}) & \gamma_{V,m}(\boldsymbol{h}) \end{bmatrix} = \begin{bmatrix} u_m & w_m \\ w_m & v_m \end{bmatrix} \cdot \begin{bmatrix} \gamma_m(\boldsymbol{h}) & 0 \\ 0 & \gamma_m(\boldsymbol{h}) \end{bmatrix}. \qquad (6.16)$$

Для удовлетворения условия положительной определенности линейной модели (6.13) достаточно положительности коэффициентов u, v, w в уравнениях (6.14)—(6.16). Это достигается наложением на коэффициенты следующих условий:

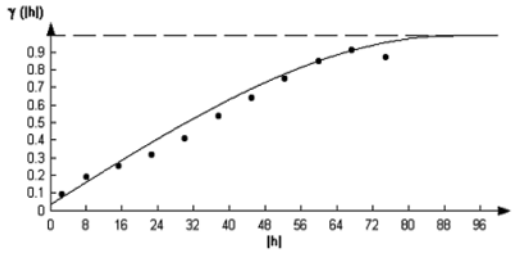
$$\begin{aligned} &u_i > 0, v_i > 0, \text{ для всех } i = 0, ..., m, \\ &u_i v_i > w_i w_i \text{ для всех } i = 0, ..., m. \end{aligned} \qquad (6.17)$$

Ограничение, накладываемое условиями (6.17), может значительно усложнить моделирование. Часто одна из авто- или кросс-вариограммных моделей не подгоняется под соответствующую экспериментальную вариограмму, в то время как другие модели подходят хорошо. В таком случае следует рассматривать каждую индивидуальную модель как часть общей модели и

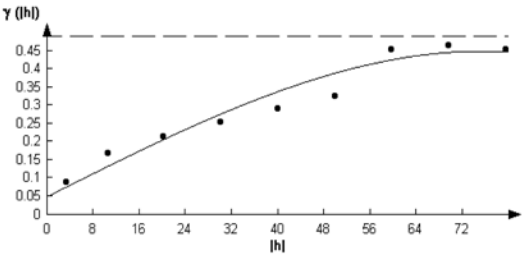
судить о качестве подгонки в соответствии с этим. Из уравнения (6.17) следуют два полезных замечания для моделирования корегионализации. Во-первых, базисная модель, содержащаяся в любой из автовариограммных (собственных) моделей, не обязательно должна быть включена в кросс-вариограммную модель. Во-вторых, любая базисная модель, содержащаяся в модели кросс-вариограммы, должна быть включена во все модели собственных вариограмм.



а



б



в

Рис. 6.3. Экспериментальная и модельные вариограммы ^{90}Sr (*а*), ^{137}Cs (*б*) и кросс-вариограмма ^{90}Sr-^{137}Cs (*в*) с применением линейной модели корегионализации

Рисунок 6.3 и табл. 6.1 иллюстрируют разрешение проблемы корегионализации при совместном моделировании пространственной структуры ^{137}Cs и ^{90}Sr. Для всех моделей выполняется условие положительной определенности:

для наггета: $\begin{vmatrix} 0,25 & 0,05 \\ 0,05 & 0,04 \end{vmatrix} = 0,0075 > 0;$

для параметров сферической модели: $\begin{vmatrix} 0,75 & 0,4 \\ 0,4 & 0,96 \end{vmatrix} = 0,56 > 0.$

Таблица 6.1. Сферические модели авто- и кросс-вариограмм для ^{90}Sr и ^{137}Cs

Переменная	Наггет	Плато	Радиус
^{137}Cs	0,04	0,96	95
^{90}Sr	0,25	0,4	100
		0,3	75
^{90}Sr-^{137}Cs	0,05	0,4	75

6.4. Кокригинг

Кокригинг (cokriging) — естественное обобщение кригинга на случай многопеременных данных, когда между переменными имеется пространственная корреляция. Основная переменная оценивается на основе ее собственных измерений и данных по другим (дополнительным) переменным. Знание всех переменных во всех точках не требуется. Для обычного кокригинга обязательно по крайней мере одно измерение основной переменной, для простого же достаточно знания ее среднего значения, остальная информация вносится за счет дополнительных переменных. С другой стороны, случай полной изотопии данных даже при взаимной коррелированности переменных эквивалентен кригингу и не дает дополнительного улучшения оценки.

Случай частичной гетеротопии, когда есть достаточный поднабор изотопных данных для построения пространственных кросс-корреляционных моделей, является наиболее интересным в плане применения кокригинга. В этом случае использование дополнительных переменных позволяет увеличить область интерполяции и/или уменьшить неопределенность оценки.

Оценка обычного кокригинга функции $Z_{\alpha_0}(x_0)$ — линейная комбинация значений различных переменных из окрестности точки x_0. Количество

участников оценивания среди различных переменных n_α может быть различно:

$$Z^*_{\alpha_0}(\boldsymbol{x}_0) = \sum_{\alpha=1}^{K}\sum_{i=1}^{n_\alpha}\lambda_i^\alpha Z_\alpha(\boldsymbol{x}_i). \tag{6.18}$$

Весовые коэффициенты линейной комбинации (6.4) определяются с использованием традиционных для геостатистики условий: несмещенности и минимизации вариации ошибки. Эти условия являются базовыми для любого геостатистического оценивателя.

Рассмотрим условие несмещенности для так построенной оценки:

$$\begin{aligned} E\left\{Z^*_{\alpha_0}(\boldsymbol{x}_0) - Z_{\alpha_0}(\boldsymbol{x}_0)\right\} &= E\left\{\sum_{\substack{\alpha=1\\ \alpha\neq\alpha_0}}^{K}\sum_{i=1}^{n_\alpha}\lambda_i^\alpha Z_\alpha(x_i) + \sum_{i=1}^{n_{\alpha_0}}\lambda_i^{\alpha_0} Z_{\alpha_0}(\boldsymbol{x}_i) - Z_{\alpha_0}(\boldsymbol{x}_0)\right\} = \\ &= \sum_{\substack{\alpha=1\\ \alpha\neq\alpha_0}}^{K}\left(m_\alpha\sum_{i=1}^{n_\alpha}\lambda_i^\alpha\right) + m_{\alpha_0}\left(\sum_{i=1}^{n_{\alpha_0}}\lambda_i^{\alpha_0} - 1\right) = 0. \end{aligned} \tag{6.19}$$

Наиболее логичным из возможных вариантов, когда условие (6.19) выполнено, будет равенство нулю всех членов суммы, т. е.

$$\sum_{i=1}^{n_\alpha}\lambda_i^\alpha = \delta_{\alpha\alpha_0} = \begin{cases} 1, & \alpha = \alpha_0, \\ 0, & \alpha \neq \alpha_0, \end{cases} \tag{6.20}$$

или, другими словами, сумма весов при основной переменной равна 1, а сумма весов при каждой из дополнительных переменных равна нулю. Этот вариант выполнения условия несмещенности является традиционным и наиболее распространенным в многопеременной геостатистике. Такое предположение приводит к традиционному обычному кокригингу.

После минимизации вариации ошибки с дополнительными условиями (6.20) получается система уравнений традиционного обычного кокригинга. Она может быть выражена в терминах кросс-ковариации и ковариации или в терминах кросс-вариограммы и вариограммы. Здесь приведен вариант выражения вариограммы, который более характерен при использовании обычного кокригинга (когда неизвестно среднее) [Wackernagel, 1995]:

$$\begin{cases} \sum_{\beta=1}^{K}\sum_{j=1}^{n_{\beta i}}\lambda_j^\beta\gamma_{\alpha\beta}(\boldsymbol{h}_{ij}) + \mu_i = \gamma_{\alpha\alpha_0}(\boldsymbol{h}_{ii0}),\ \alpha = 1, ..., K;\ i = 1, ..., n_\alpha, \\ \sum_{j=1}^{n_\alpha}\lambda_j^\alpha = \delta_{\alpha\alpha_0},\ \alpha = 1, ..., K. \end{cases}$$

Кокригинг, как и интерполяторы семейства кригингов, позволяет оценить и вариацию ошибки:

$$\sigma^2_{CK} = \sum_{\alpha=1}^{K}\sum_{i=1}^{n_\alpha}\lambda_i^\alpha\gamma_{\alpha\alpha_0}(\boldsymbol{h}_{ii_0}) + \mu_{i_0} - \gamma_{\alpha_0\alpha_0}(0).$$

Кроме традиционного обычного кокригинга, рассмотренного выше, различают несколько типов кокригинга в зависимости от способа удовлетворения условия несмещенности (6.20).

Стандартизованный обычный кокригинг. Часто улучшение состоит во введении новых дополнительных переменных с таким же средним значением, как и у основной переменной. Тогда сумма весов для всех переменных приравнивается к единице. В этом случае оценка (6.18) может быть записана так:

$$Z^*_{\alpha_0}(\boldsymbol{x}_0) = \sum_i \lambda_i^{\alpha_0} Z_{\alpha_0}(\boldsymbol{x}_i) + \sum_{\beta\neq\alpha_0}\sum_j\left[\lambda_j^\beta Z_\beta(\boldsymbol{x}_j) + m_\beta - m_{\alpha_0}\right],$$

где m_β — стационарные средние значения $Z_\beta(x)$. Для реализации условия несмещенности (2.12) в этом случае используется другое условие:

$$\sum_{i=1}^{n}\sum_{\alpha=1}^{K}\lambda_i^\alpha = 1.$$

Простой кокригинг. Как и в случае простого кригинга, средние значения m_α для всех K переменных предполагаются известными, и, следовательно, условие (6.19) для весов выполняется автоматически. Оценка простого кокригинга в точке x_0 вычисляется следующим образом:

$$Z^*_{\alpha_0}(\boldsymbol{x}_0) = m_{\alpha_0} + \sum_{\alpha=1}^{K}\sum_{i=1}^{n_i}\lambda_i^\alpha\left[Z_\alpha(\boldsymbol{x}_i) - m_\alpha\right].$$

Во всех типах кокригинга, кроме традиционного обычного кокригинга, используются не вариограммы и кросс-вариограммы, а ковариации и кросс-ковариации, поскольку предполагаются выполненными условия стационарности второго порядка.

Как и в случае кригинга, кокригинг может быть *точечным* и *блочным*. Его использование ничем не отличается от случая блочного кригинга (см. Подраздел 5.6.4), только еще определяются и кросс-функции, усредненные по некоторой зоне.

Пример использования кокригинга для Чернобыльских данных. Полезность кокригинга удобно проиллюстрировать на примере анализа пространственных данных по загрязнению окружающей среды в результате Черно-

быльской аварии изотопами ^{137}Cs и ^{90}Sr, которые сильно коррелированны между собой (рис. 6.4). Если рассматривать их корреляцию как линейную, то коэффициент корреляции равен 0,74. Количество измерений ^{90}Sr было меньше, чем ^{137}Cs, из-за их дороговизны (рис. 6.5). Методика кокригинга позволила использовать измерения ^{137}Cs для уточнения оценки загрязнения ^{90}Sr. В частности, зоны, где отсутствуют измерения по ^{90}Sr, но присутствуют измерения по ^{137}Cs (северо-восточная часть данной территории), перестают быть зонами экстраполяции. Именно для них использование кокригинга и представляет особый интерес, так как там, где есть измерения по ^{90}Sr, достаточно оценки обычного кригинга.

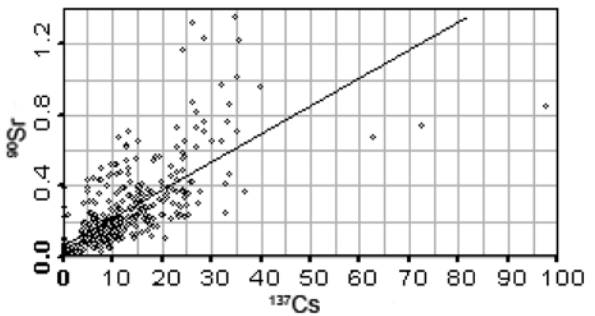


Рис. 6.4. Корреляция между значениями выпадений ^{137}Cs и ^{90}Sr в западной части Брянской области

При использовании обычного кригинга оценивание в областях экстраполяции связано с выбором области поиска используемых при оценке данных. Размер анизотропной корреляционной структуры для ^{90}Sr (до 40 км, см. вариограммную поверхность на рис. 6.6*б*) не позволяет получить корректную оценку в этих областях экстраполяции. Кокригинг же использует для оценивания дополнительные измерения ^{137}Cs, присутствующие в этих областях. Кроме того, более протяженная корреляционная структура ^{137}Cs (до 70 км, см. вариограммную поверхность на рис. 6.6*а*) и соответствующая ей область поиска дают возможность захватить более обширную территорию. Таким образом, кокригинг позволяет избежать появления не оцененных (или оцененных некорректно) областей по сравнению с обычным кригингом.

На рис. 6.7 представлена пространственная кросс-корреляционная структура ^{137}Cs и ^{90}Sr. Пространственная корреляция и пространственная кросс-корреляция представлены в виде вариограммных (кросс-вариограммных)

поверхностей. На рисунках 6.6, 6.7 приведены и экспериментальные, и соответствующие им модельные структуры.

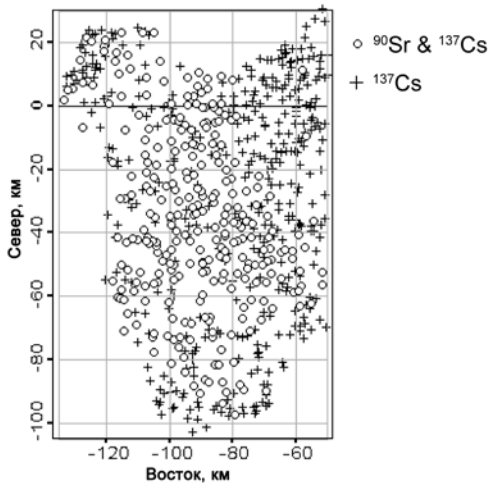


Рис. 6.5. Схема точек снятия проб по загрязнению поверхности ^{137}Cs и ^{90}Sr в западной части Брянской области

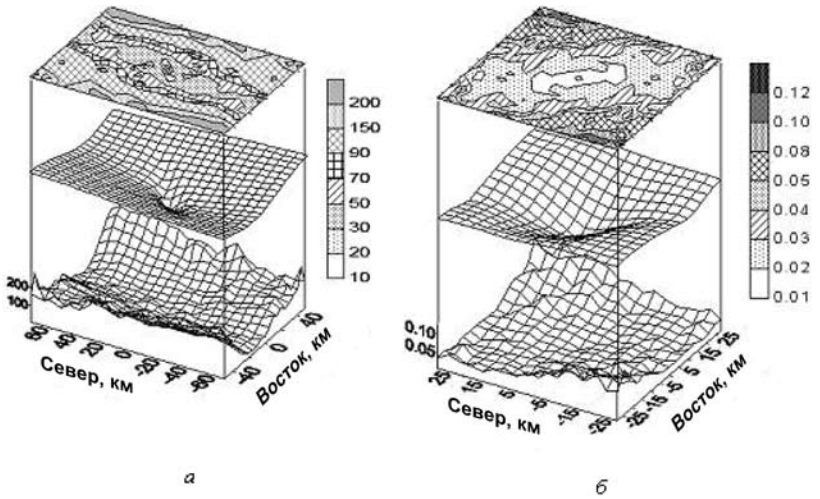


Рис. 6.6. Экспериментальные (верхняя и нижняя)
и модельные (средняя) вариограммные поверхности
для данных по загрязнению поверхности западной части Брянской области:
а — для ^{137}Cs; *б* — для ^{90}Sr

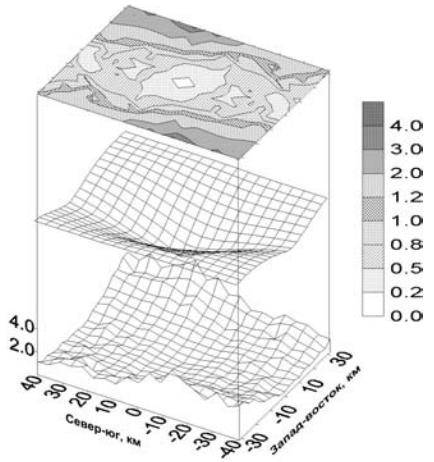


Рис. 6.7. Экспериментальные (верхняя и нижняя) и модельная (средняя) кросс-вариограммные поверхности для данных по загрязнению ^{137}Cs и ^{90}Sr поверхности западной части Брянской области

Кокригинг выполнялся на прямоугольной сетке 45×70 ячеек с размером ячейки 2×2 км. Полученная карта оценок ^{90}Sr представлена на рис. 6.8. Для сравнения оценка загрязнения ^{90}Sr почвы в этой же области с использованием обычного кригинга представлена на рис. 6.9. При сравнении карт оценок видно, что обычный кригинг сделал размазывающее усреднение к границам и оставил не оцененными значительные области по краям. Это связано с отсутствием в этих областях измерений ^{90}Sr. Кокригинг же использует для оценивания дополнительные измерения ^{137}Cs в областях, где измерения ^{90}Sr отсутствуют.

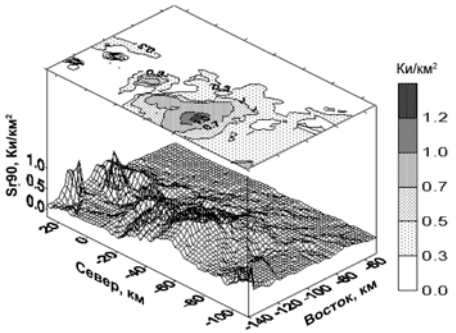


Рис. 6.8. Карта результатов оценки загрязнения ^{90}Sr западной части Брянской области с помощью обычного кокригинга

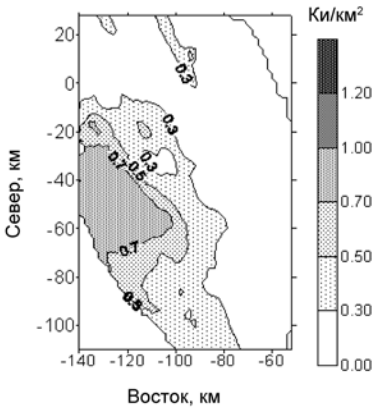


Рис. 6.9. Карта результатов оценки загрязнения ^{90}Sr западной части Брянской области с помощью обычного кригинга

Оценка обычного кригинга более размазана, менее вариабельна. Это означает занижение оценки в точках с высоким загрязнением и ее завышение в точках с низким загрязнением. В отличие от обычного кригинга кокригинг дал значительно менее сглаженную оценку. На карте четко видны характерные пятна большого загрязнения вблизи точек выбросов высоких значений. Сравнение итоговых статистик оценки кокригинга и исходных измерений (с учетом и без учета декластеризации, см. Раздел 2.5) также подтверждает ее превосходство над оценкой обычного кригинга (табл. 6.2). Отметим, что кокригинг позволяет улучшить оценку высоких значений.

Теоретически кокригинг не имеет ограничений на число переменных, и добавление новой информации должно вести к улучшению оценки. На практике это не совсем так.

В случае K переменных для кокригинга требуется K^2 моделей вариограмм. Проверка всех гипотез для такого количества данных и последующее совместное моделирование становится достаточно трудоемким. Кроме того, оценка экспериментальных вариограмм, кросс-вариограмм и их моделирование на практике выполняется с некоторой ошибкой. Большое количество моделей вариограмм может настолько усложнить вычисление окончательной оценки, что результат даже ухудшится. Поэтому важно подбирать правильное количество переменных и выбирать те, использование которых действительно приводит к улучшению оценки.

Таблица 6.2. Статистика распределений измерений ^{90}Sr и оценок кокригинга и обычного кригинга

Статистика	Кластеризованные измерения	Декластеризованные измерения	Оценки кокригинга	Оценки обычного кригинга
Количество данных	286	286	3150	2779
Среднее значение	0,292	0,251	0,309	0,323
Вариация	0,052	0,049	0,050	0,051
Стандартное отклонение	0,227	0,222	0,224	0,228
Коэффициент вариации, %	77,9	88,46	72,6	70,2
Минимум	0,018	0,018	0,180	0,034
Нижний квартиль (25%)	0,144	0,109	0,149	0,162
Медиана	0,226	0,183	0,258	0,252
Верхний квартиль (75%)	0,372	0,310	0,403	0,411
Максимум	1,361	1,361	1,303	1,043

Пример использования кокригинга для исследования загрязнения Женевского озера. Можно сравнить результаты применения кригинга и кокригинга с различным количеством переменных для случая девятипеременной функции (загрязнение донных отложений Женевского озера металлами). Одна из переменных является основной (Pb), остальные (Zn, Cu, Mn, Cd, Ni, Be, B, Cr) — дополнительной информацией. На рис. 6.10 представлены кросс-вариограммы основной переменной со всеми дополнительными переменными. Видно, что пространственная корреляция для всех, кроме одной (правая верхняя с B), достаточно хорошо моделируется.

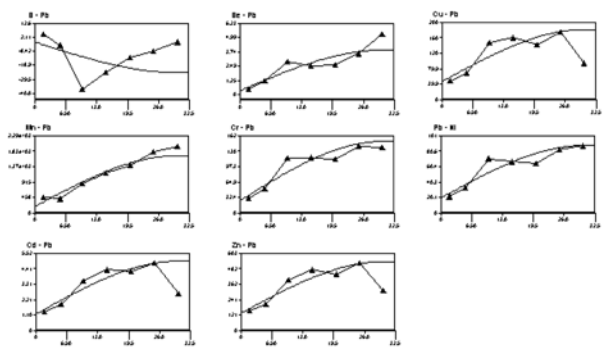


Рис. 6.10. Кросс-вариограммы основной переменной (Pb) с дополнительными переменными. Линии, соединяющие значки, — экспериментальные кросс-вариограммы, кривые — модели кросс-вариограмм

Сравнение результатов кокригинга проводилось по среднеквадратичной ошибке кросс-валидации. Результаты собраны в табл. 6.3. Все кокригинги дают меньшую ошибку кросс-валидации, чем кригинг (86,37). Использование третьей переменной улучшает оценку по сравнению с двумя. Но введение четвертой переменной (при использовании различных наборов) уже ведет к ухудшению. Использование всех переменных дает худший для кокригинга результат.

Таблица 6.3. Среднеквадратичные ошибки кросс-валидации (RMSE) при оценке кокригинга основной переменной (Pb) с использованием различных наборов дополнительных переменных (Zn, Cu, B, Cd, Cr, Mn, Be, Ni)

Количество дополнительных переменных	Дополнительные переменные	RMSE
0 (обычный кригинг)	—	86,37
1	Zn	59,10
2	Zn, Cu	58,72
2	Zn, B	58,69
3	Zn, Cu, B	58,25
4	Zn, Cu, B, Cd	58,66
4	Zn, Cu, B, Mn	58,74
4	Zn, Cu, B, Cr	61,37
4	Zn, Cu, B, Be	60,63
4	Zn, Cu, B, Ni	58,81
8	Все	63,35

Другой причиной того, что кокригинг не очень популярен, является эффект экранирования более коррелированными данными (измерениями основной переменной) менее коррелированных (измерений дополнительной переменной). Влияние эффекта экранирования на значения весов и их отрицательный эффект на оценку уже обсуждались при рассмотрении обычного кригинга (см. Подраздел 5.6.1). В случае кокригинга эффект экранирования проявляется чаще за счет дополнительной информации в дополнительных точках измерения.

6.5. Колокационный кокригинг

Один из способов борьбы с избыточной информацией по дополнительным переменным — колокационный (collocated) кокригинг. В этом случае для оценки используются только значения дополнительных переменных, находящихся в ближайшей окрестности точки оценивания, и они приписываются к пространственному положению точки оценивания, т.е. оценка колокационного кокригинга может быть записана так:

$$Z^*_{\alpha_0}(x_0) = \sum_{i=1}^{n} \lambda_i^{\alpha_0} Z_{\alpha_0}(x_i) + \sum_{\beta \neq \alpha_0} \lambda_{i_0}^{\beta} Z_{\beta}(x_0).$$

Кроме того, колокационный кокригинг предполагает линейную связь между ковариацией основной переменной и кросс-ковариацией:

$$\rho = \frac{C_{12}(0)}{C_{11}(0)}, \quad C_{12}(h) = \rho C_{11}(h).$$

Это снимает необходимость моделирования кросс-ковариаций, ограничивая моделирование только ковариациями (или вариограммами). Колокационный кокригинг вычислительно проще и быстрее полного кокригинга.

В Разделе 6.1 уже упоминался пример моделирования поля температур с использованием дополнительной информации о высоте над уровнем моря. Применение полного кокригинга в этом примере привело к среднеквадратичной ошибке на валидационном наборе в размере 3,97. Это даже хуже, чем результат обычного кригинга, где среднеквадратичная ошибка на валидационном наборе составляла 3,13. Использование колокационного кокригинга позволило уменьшить среднеквадратичную ошибку на валидационном наборе до 3,05.

6.6. Анализ принципиальных компонент в геостатистике

Выше уже упоминалась проблема выбора оптимальной комбинации дополнительных переменных из большого набора переменных для оценки основной переменной. Эту задачу можно переформулировать как проблему сжатия (понижения размерности) пространства при условии сохранения максимальной информативности.

Одним из способов уменьшения вычислительной сложности многопеременного кокригинга является использование кригинга принципиальных компо-

нент. Этот подход предполагает переход от K коррелированных переменных к d ($d \leq K$) некоррелированным факторам. Эти факторы представляют собой линейную комбинацию исходных переменных и могут быть использованы для описания данных при меньшей размерности. Для перехода к факторам используется известный в многопеременной статистике анализ принципиальных компонент [Вентцель, 1964], где эти принципиальные компоненты и являются искомыми факторами.

Анализ принципиальных компонент — это линейное ортогональное преобразование данных в новую систему координат, построенную так, что первая координата направлена вдоль проекции данных, где они имеют максимальную вариацию, а вторая координата — вдоль направления проекции данных, представляющего следующую по значению вариацию (рис. 6.11), и т. д. Уменьшение размерности осуществляется за счет пренебрежения координатами, соответствующими малым значениям вариации данных. Это дает возможность сохранить наиболее значимую информацию.

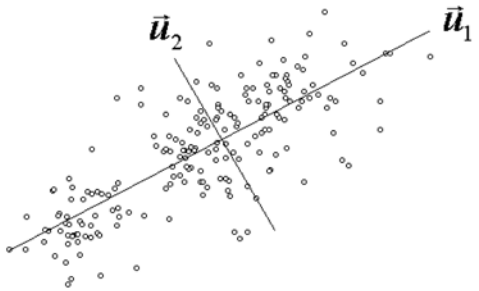


Рис. 6.11. Схема построения новой системы координат при анализе принципиальных компонент

Рассмотрим анализ принципиальных компонент в математическом плане. Пусть имеется матрица данных, из которых вычтены значения средних по каждой переменной. Обозначим ее через $\mathbf{Z}$, она имеет размерность $K \times N$ — N точек измерений по K переменных в каждой. Для нее можно построить ковариационную матрицу $\mathbf{V}_Z$ по формуле (6.1) (среднее равно нулю). Требуется найти такую ортогональную матрицу $\mathbf{A}$ размерности $N \times N$, что в результате ее умножения на исходную матрицу $\mathbf{Z}$ получим новую матрицу $\mathbf{Y}$, обладающую диагональной ковариационной матрицей $\mathbf{V}_Y$, т. е.

$$\mathbf{AZ} = \mathbf{Y},$$

где $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{E}$ и

$$\mathbf{V}_Y = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \begin{vmatrix} \lambda_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_{NN} \end{vmatrix}.$$

Сделав подстановку и простейшие преобразования

$$\mathbf{V}_Y = \frac{1}{n}(\mathbf{ZA})^{\mathrm{T}}(\mathbf{ZA}) = \frac{1}{n}\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{ZA} = \mathbf{A}^{\mathrm{T}}\frac{1}{n}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\mathbf{A} = \mathbf{A}^{T}\mathbf{V}_Z\mathbf{A},$$

получаем, что решаемая нами задача свелась к поиску собственных значений и собственных векторов ковариационной матрицы исходных данных.

Собственные значения λ_{ii} (будем для простоты обозначать их просто λ_i), характеризующие вариацию факторов, могут быть отсортированы в порядке убывания. Собственные вектора выстраиваются в порядке соответствующих им собственных значений. Таким образом, получается последовательность из N некоррелированных факторов. Они обеспечивают оптимальное (в смысле аппроксимации методом конечных квадратов) разложение полной вариации:

$$\mathrm{tr}\left(\mathbf{V}_Z\right) = \sum_{i=1}^{N}\sigma_{ii} = \sum_{p=1}^{N}\lambda_p.$$

Собственные значения характеризуют вклад вариации фактора в полную вариацию, а отношение

$$\frac{\lambda_i}{\mathrm{tr}\left(\mathbf{V}_Z\right)}100\% \qquad (6.21)$$

дает численное значение (обычно выражаемое в процентах) значимости соответствующего фактора. Значимыми считаются факторы, для которых соотношение (6.21) дает больше 90%.

На рис. 6.12 показаны значения принципиальных компонент для примера, уже использовавшегося в этой главе (содержание металлов в донных отложениях Женевского озера). Пунктирная линия проводит границу значимости (90%). Три компоненты вносят основной вклад. Этот результат согласуется с тем, что был получен при использовании кокригинга, — три переменные (если они были правильно выбраны) давали лучший результат.

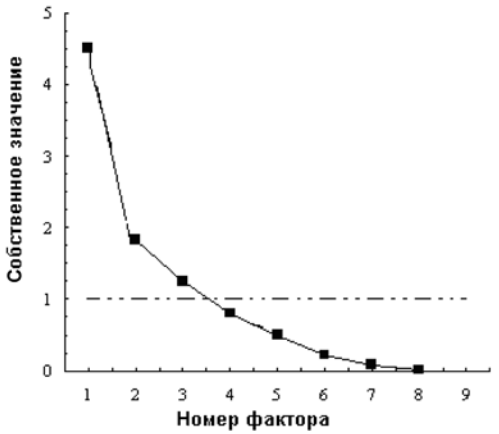


Рис. 6.12. Значения вариаций принципиальных компонент для набора из девяти переменных

Соотнесение исходных переменных с факторами (корреляцию фактора и исходных данных) можно увидеть на кругах корреляции. Они изображают проекции положения переменных на поверхность гиперсферы по отношению к плоскости, определяемой парой осей-факторов. Два примера кругов корреляции (две плоскости факторов) для трех значимых факторов представлены на рис. 6.13. По ним видно, что на фактор *1* проецируется много переменных, а фактор *3* связан практически исключительно с бором.

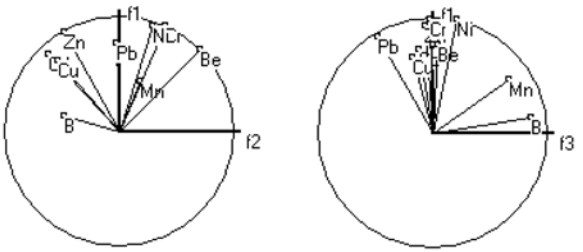


Рис. 6.13. Круги корреляции девяти переменных для плоскостей, определяемых факторами *1* и *2* (слева) и факторами *1* и *3* (справа)

Для каждого из значимых факторов, полученных в результате анализа принципиальных компонент, может быть проведен кригинг. Так как факторы по определению ортогональны (и в соответствии с нашим желанием не коррелированны), их совместную оценку проводить бессмысленно. По

оцененным факторам можно вернуться к исходным данным, используя обратное преобразование $\mathbf{Z} = \mathbf{A}^{-1}\mathbf{Y}$. Таким образом, вместо моделирования K^2 моделей пространственной корреляции и кросс-корреляции мы обходимся моделированием $d \leq K$ пространственных корреляций значимых факторов, но в дополнение решаем задачу поиска собственных векторов и собственных значений ковариационной матрицы.

Литература

Вентцель Е. С. Теория вероятностей. — М., 1964.

Каневский М. Ф., Арутюнян Р. В., Большов Л. А. и др. Геостатистический подход к анализу чернобыльских выпадений // Изв. Рос. акад. наук. Энергетика. — 1995. — № 3. — С. 34—46.

Cressie N. Statistics for Spatial Data. — New York: John Wiley & Sons, 1991. — P. 141.

Deutsch C., Journel A. G. GSLIB: Geostatistical Software Library and User's Guide. — [S. l.]: Oxford Univ. Press, 1998.

Dowd P. A. Generalised cross-covariances // Geostatistics. — 1989. — Vol. 1. — P. 579—590.

Isaaks Ed. H., Srivastava R. M. An Introduction to Applied Geostatistics. — Oxford, Oxford Univ. Press, 1989.

Myers D. E. Pseudo-Cross Variograms, Positive-Definiteness and cokriging // Mathematical Geology. — 1991. — Vol. 23, N 6. — P. 805—816.

Myers D. E. The Linear coregionalization and simultaneous diagonalization of the variogram matrix function // Sciences de la Terre. — 1995. — Vol. 32. — P. 125—139.

Pan G., Gaard D., Moss K., Heiner T. A Comparison Between Cokriging and Ordinary Kriging: Case Study with a Pollymetallic Deposit // Mathematical Geology. — 1993. — Vol. 25, N 3. — P. 377—398.

Papritz A., Kunsch H. R., Webster R. On the Pseudo Cross Variogram // Mathematical Geology. — 1993. — Vol. 25, N 8. — P. 1015—1026.

Wackernagel H. Multivariate Geostatistics. — Berlin: Springer-Verl., 1995.

Глава 7

Вероятностное моделирование локальной неопределенности

Эта глава включает описание индикаторного преобразования (Раздел 7.1), применения индикаторного подхода для анализа непрерывной и категориальной переменных (Раздел 7.2). Рассмотрены примеры использования индикаторного подхода для различных типов данных (Раздел 7.3).

Как уже было показано (в частности, в Главе 5), каждая оценка обладает неопределенностью, т. е. дает значение лишь с некоторой долей точности. Наиболее общим подходом при такой интерпретации результата является не сама оценка значения в точке, а описание локальной кумулятивной функции распределения (кумулятивной функции распределения в точке). Знание функции распределения дает возможность строить различные типы оценки значения функции: максимально вероятную (максимум плотности функции распределения), среднюю (минимизирующую вариацию ошибки), медианную (минимизирующую среднее абсолютное значение ошибки), с учетом штрафов (минимум специально построенной функции от ошибок) и т. п. Функция распределения позволяет получать различные вероятностные и статистические оценки: вероятность превышения некоторого уровня, вероятность попадания значения в интервал, доверительные интервалы, среднее значение для определенного вероятностного интервала и т. п.

Одним из возможных решений поставленной задачи в рамках классической геостатистики является индикаторное рассмотрение. Оно основано на предположении о пространственной непрерывности анализируемой функции, т. е. о том, что поведение функции в окрестности точки некоторым образом аналогично поведению в точке. Это предположение компенсирует наличие только одной реализации случайной функции.

Индикаторный подход был изложен в работах [Journel, 1983, 1985]. Было предложено использовать при анализе и моделировании переход от исходных значений переменных к специальным индикаторам. Индикаторные переменные — бинарные, т. е. принимают значения либо 0, либо 1. Для категориальной переменной такое преобразование дает индикаторную

переменную для каждой из категорий, характеризуя присутствие или отсутствие данной категории. В случае непрерывной переменной переход представляет собой нелинейное преобразование, моделирующее кумулятивную функцию распределения. Переход к индикаторным переменным позволяет получить оценку условной локальной функции распределения (но не ее явный вид) во всем пространстве. Преобразованные данные более устойчивы к выбросам (outliers). Индикаторное преобразование может быть применено и при анализе категориальных данных, а также при совместном анализе данных различных типов, что, в свою очередь, добавляет такому подходу дополнительную привлекательность.

7.1. Индикаторное преобразование

Для индикаторного преобразования непрерывной случайной функции $Z(\boldsymbol{x})$ сначала выбирается набор пороговых значений z_k, $k = 1, \ldots, K$. Затем производится переход к индикаторам, для каждого порогового значения z_k определяется индикатор

$$I(\boldsymbol{x}, z_k) = \begin{cases} 1, & \text{если } Z(\boldsymbol{x}) \le z_k, \\ 0, & \text{если } Z(\boldsymbol{x}) > z_k. \end{cases} \tag{7.1}$$

В результате для каждой точки $\boldsymbol{x}$ пространства получается вектор индикаторов (размерности K):

$$I(\boldsymbol{x}) = \left(I(\boldsymbol{x}, z_1), \ldots, I(\boldsymbol{x}, z_K) \right)$$

где K — число пороговых отсечений.

Пример проведения индикаторного преобразования, определяющего тип фации (породы) для одного порогового значения, представлен на рис. 7.1 на примере профиля каротажа Z в скважине. Разбиение на категории осуществляется в соответствии с пороговым значением z_k — его превышением и непревышением. Значения функции больше порогового значения трансформируются в значения 0, значения функции меньше порогового значения преобразуются в 1. Использование набора пороговых значений в случае непрерывной переменной дает возможность построить подробную локальную функцию распределения с требуемым разрешением.

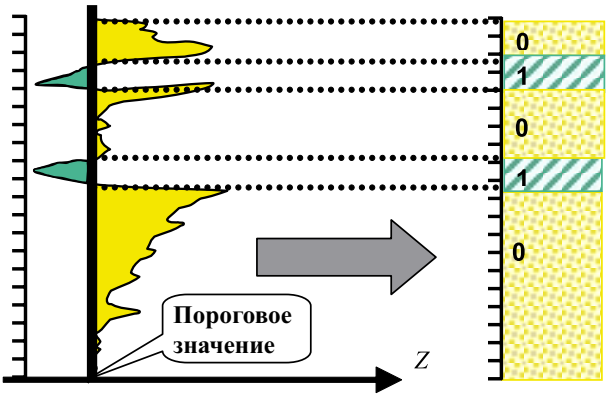


Рис. 7.1. Пример индикаторного преобразования функции

Следует иметь в виду, что вектор индикаторов не воспроизводит точно исходные значения функции $Z(\boldsymbol{x})$. Для каждой точки измерений $\boldsymbol{x}_i$ он представляет модель ступенчатой функции $\Phi(z)$ из 0 и 1, где переход с 0 на 1 зависит от значения исходной функции $Z(\boldsymbol{x}_i)$. Глобальное по области усреднение индикаторов примерно воспроизводит одномерную функцию распределение исходных данных:

$$E\{I(\boldsymbol{x}, z_k)\} = \Pr\{Z(\boldsymbol{x}) \le z_k\} = F(z_k).$$

Важно правильно задать количество порогов и их значения. Количество порогов должно быть достаточно большим, чтобы обеспечивать допустимое дискретное представление локальной функции распределения. С другой стороны, оно не должно быть чрезмерным, чтобы не приводить к излишним вычислительным сложностям и уменьшить влияние искажений, вызванных процедурой преобразования. На практике число порогов всегда больше 5 и редко больше 15.

В качестве пороговых обычно используются значения, разделяющие примерно равномерно полный диапазон значений функции на $K + 1$ класс. Кроме того, полезно использовать в качестве пороговых критические значения, связанные с конкретной задачей, например значения концентрации, требующие проведения профилактических или защитных мероприятий. Важно также обращать внимание на то, чтобы для каждого порога не было слишком сильного преобладания нулевых или единичных значений. Такой случай приведет к проблемам при моделировании пространственной корреляции для этого порога.

Теперь рассмотрим функцию $Q(x)$, определенную на области S и принимающую конечный набор значений (c_1, …, c_C). Такую функцию называют пространственной (региональной) категориальной. На практике это могут быть типы почв, типы геологического формирования, предупредительный сигнал прибора и т. п. Для пространственной категориальной переменной $Q(x)$ индикаторное преобразование делается для каждого возможного значения (класса) c_i, $i = 1, \ldots, C$:

$$I(x, c_i) = \begin{cases} 1, & \text{если} \quad Q(x) = c_i, \\ 0, & \text{если} \quad Q(x) \neq c_i. \end{cases} \tag{7.2}$$

Для категориальной переменной вектор индикаторов

$$I(x) = \left[I(x, c_1), \ldots, I(x, c_C) \right]$$

состоит из 0 и одной 1 в соответствии со значением $Q(x)$. Глобальное усреднение индикаторов даст относительное распределение исходного набора данных по классам (значениям категориальной функции):

$$E\left\{ I(x, c_i) \right\} = \Pr\left\{ Q(\cdot) = c_i \right\} = P(c_i).$$

Индикаторы и непрерывной, и категориальной переменных по сути относятся к одному типу — это категориальные индикаторы, принимающие два значения — 0 и 1. Поэтому их можно использовать вместе.

Отдельный индикатор — $I(\boldsymbol{x}, z_k)$ или $I(\boldsymbol{x}, c_i)$ — можно рассматривать как пространственную функцию от $\boldsymbol{x}$. Соответственно можно ввести и пространственные корреляционные функции для этих переменных.

Нецентральная индикаторная ковариация:

$$K_I(\boldsymbol{h}, z_k) = E\left\{ I(\boldsymbol{x}, z_k) I(\boldsymbol{x} + \boldsymbol{h}; z_k) \right\} = \Pr\left\{ Z(\boldsymbol{x}) \leq z_k, Z(\boldsymbol{x} + \boldsymbol{h}) \leq z_k \right\};$$

$$\begin{aligned} K_1(\boldsymbol{h}, z_k) &= E\left\{ I(\boldsymbol{x}, z_k) I(\boldsymbol{x} + \boldsymbol{h}, z_k) \right\} = \\ &= \Pr\left\{ Z(\boldsymbol{x}) \leq z_k, Z(\boldsymbol{x} + \boldsymbol{h}) \leq z_k \right\}, k = 1, \ldots, K; \end{aligned} \tag{7.3}$$

$$\begin{aligned} K_1(\boldsymbol{h}, c_{ik}) &= E\left\{ I(\boldsymbol{x}, c_i) I(\boldsymbol{x} + \boldsymbol{h}, c_i) \right\} = \\ &= \Pr\left\{ Q(\boldsymbol{x}) = c_i, Q(\boldsymbol{x} + \boldsymbol{h}) = c_i \right\}, i = 1, \ldots, C. \end{aligned}$$

Центральная индикаторная ковариация:

- для индикатора непрерывной переменной:

$$C_I(\boldsymbol{h}, z_k) = K_I(\boldsymbol{h}, z_k) - F^2(z_k);$$

- для индикатора категориальной переменной:

$$C_I(\boldsymbol{h}, c_i) = K_I(\boldsymbol{h}, c_i) - P^2(c_i).$$

Индикаторная полувариограмма:

- для индикатора непрерывной переменной:

$$2\gamma(\boldsymbol{h}, z_k) = E\left\{\left[I(\boldsymbol{x}, z_k)\right] - I(\boldsymbol{x}+\boldsymbol{h}, z_k)^2\right\} = 2K_I(0, z_k) - 2K_I(\boldsymbol{h}, z_k);$$

- для индикатора категориальной переменной:

$$2\gamma(\boldsymbol{h}, c_i) = E\left\{\left[I(\boldsymbol{x}, c_i)\right] - I(\boldsymbol{x}+\boldsymbol{h}, c_i)^2\right\} = 2K_I(0, c_i) - 2K_I(\boldsymbol{h}, c_i).$$

На практике при использовании индикаторного подхода для анализа пространственных данных используется индикаторная полувариограмма. Анализ и моделирование индикаторной полувариограммы проводится по методике, описанной в Главе 4.

Индикаторное преобразование позволяет также обойти проблему наличия крайних нехарактерных значений (высоких и низких крайних экстремальных значений, которые характеризуют длинные хвосты распределения) и проблему с широким разбросом значений. Индикаторное преобразование является нелинейным, что позволяет уменьшить влияние на вариограмму крайних высоких и низких значений (аналогично логнормальному преобразованию, см. Раздел 5.5).

Упражнение 7.1. *В предположении внутренней гипотезы вариограмма стремится к постоянному значению (плато) на расстоянии радиуса корреляции, которое соответствует априорной вариации. К чему стремится априорная вариация для медианной индикаторной переменной при большом количестве данных?*

В Главе 5 рассматривался пример вариограммы для траловой съемки краба Берди в Беринговом море в 2003 г. Там разброс значений составлял от 0 до 170000, (т. е. пять порядков), что не позволяло оценивать пространственную корреляцию с помощью вариограммы (см. рис. 5.16). Индика-

торные вариограммы для четырех порогов, соответствующих квантилям 0,5 (321 краб), 0,625 (870,5 краба), 0,75 (2186 крабов) и 0,875 (7421,13 краба), представлены на рис. 7.2. Каждая из них имеет структуру, которую вполне можно моделировать.

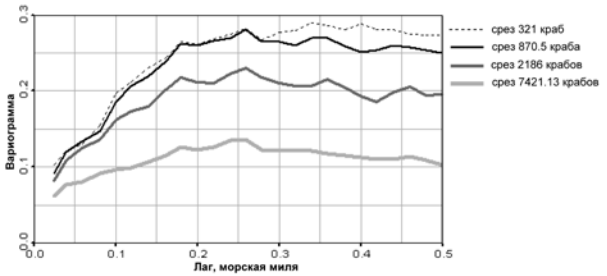


Рис. 7.2. Индикаторные вариограммы для данных траловой съемки пространственного распределения краба Берди

7.2. Индикаторный кригинг

Индикаторным кригингом (*indicator kriging*) называется обычный кригинг, выполненный для индикаторов, т. е. это линейный оцениватель, построенный по аналогии с обычным кригингом, но не для значений анализируемой переменной, а для индикатора:

$$I^*(\boldsymbol{x_0}, z_k) = \sum_{i=1}^{n} \lambda_{ki} I(\boldsymbol{x_i}, z_k). \tag{7.4}$$

Доказано, что если в (7.4) использовать весовые коэффициенты, которые найдены, исходя из предположения о минимизации вариации ошибки, то полученная оценка индикатора является оценкой вероятности, в случае непрерывной переменной — оценкой кумулятивной функции распределения:

$$I^*\left(x_0, z_k\right) = F^*\left(x_0, z_k \,|\, (n)\right).$$

Веса λ_{ki}, полученные при решении системы уравнений обычного кригинга для индикаторов,

$$\begin{cases} \sum_{i=1}^{n} \lambda_{ki} C_I\left(h_{ij}, z_k\right) - \mu_k = C_I\left(h_{0j}, z_k\right), \; j = 1, \ldots, n, \\ \sum_{i=1}^{n} \lambda_{ki} = 1, \end{cases} \tag{7.5}$$

удовлетворяют требованию минимизации вариации ошибки. Здесь n — число точек измерений с заданными значениями оцениваемой функции $Z(\boldsymbol{x}_i)$; $\boldsymbol{x}_0$ — точка, в которой производится оценивание; $\boldsymbol{h}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$ — вектор разности координат между соответствующими точками, который является аргументом функций ковариаций и вариограмм (по условию стационарности для $Z(\boldsymbol{x})$).

Решение системы уравнений (7.5) выполняется в каждой точке, где проводится оценка для всех индикаторов. Существует ситуация, при которой систему уравнений нужно решать всего один раз. Это случай, когда модели пространственных корреляций для различных порогов связаны множителями с некоторой базовой моделью (рис. 7.3):

$$C_I\left(h, z_k\right) = \phi_k C_I\left(h, z_M\right). \tag{7.6}$$

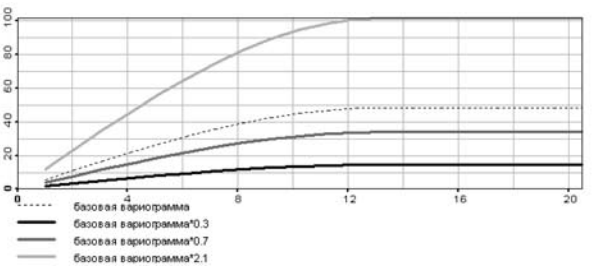


Рис. 7.3. Пример вариограмм, связанных множителями с базовой вариограммой

Здесь выделен базовый порог z_M, через модель пространственной корреляции которого описываются пространственные корреляции всех остальных отсечений. Как уже упоминалось (см. Главу 5), веса кригинга λ_i не изменяются при мультипликативном преобразовании модели пространственной корреляции. Поэтому, будучи вычислены один раз, они могут быть использованы и для получения оценки индикатора, соответствующего другому порогу. Этот упрощенный вариант индикаторного кригинга принято называть *медианным кригингом*, так как обычно при выполнении условия (7.6) в качестве базового порога выбирается значение медианы исходного набора данных.

Получив оценки для индикаторов, соответствующих всем порогам, мы должны получить модель локальной условной функции распределения. Условность функции распределения обусловлена использованием исходных данных. Однако последовательное использование индикаторного кригинга для набора порогов не гарантирует согласованности между

оценками для отдельных порогов, т. е. может быть не выполнено одно из математических свойств функции распределения: ограниченность снизу нулем, ограниченность сверху единицей и монотонное неубывание. Такого рода ситуации могут возникать, например, из-за эффекта экранирования, рассмотренного в Главе 5, который приводит к отрицательным весам кригинга. Математически эти требования записываются следующим образом:

$$\begin{gathered}\left[F\left(u, z_k \mid (n)\right)\right]^* = i^*(u, z_k) \in [0,1], \\ \left[F\left(u, z_k \mid (n)\right)\right]^* = i^*(u, z_k) \le \left[F\left(u; z_{k'} \mid (n)\right)\right]^* = i^*(u, z_{k'}) \quad \forall z_{k'} > z_k .\end{gathered} \tag{7.7}$$

Поскольку эффект экранирования является искусственной проблемой, не зависящей от свойств самого индикаторного кригинга и значений функции в точках оценивания, для удовлетворения требований (7.7) проводят коррекцию оценки условной функции распределения. Простейший способ коррекции состоит в использовании среднего значения от поправок при проходе вверх и при проходе вниз. Подробнее эту процедуру можно описать следующим образом:

- сначала делается коррекция на попадание в отрезок [0, 1]:

$$\begin{aligned} F^*(z_1) &= \begin{cases} F^*(z_1) & F^*(z_1) \ge 0, \\ 1 & F^*(z_1) < 0, \end{cases} \\ F^*(z_K) &= \begin{cases} F^*(z_K) & F^*(z_K) \le 1, \\ 1 & F^*(z_K) > 1; \end{cases} \end{aligned} \tag{7.8}$$

- проход вверх производит коррекцию в сторону увеличения при убывании оцененного значения при возрастании порогового значения:

$$\forall i = 1 : K-1 \Rightarrow F^*_U(z_{i+1}) = \begin{cases} F^*(z_{i+1}) & F^*(z_{i+1}) \ge F^*(z_i), \\ F^*(z_i) & F^*(z_{i+1}) < F^*(z_i); \end{cases}$$

- при проходе вниз проводится коррекция в сторону уменьшения, если наблюдается рост оцененного значения при уменьшении порогового значения:

$$\forall i = K : 2 \Rightarrow F^*_D(z_{i-1}) = \begin{cases} F^*(z_{i-1}) & F^*(z_{i-1}) \le F^*(z_i), \\ F^*(z_i) & F^*(z_{i-1}) > F^*(z_i); \end{cases}$$

- на последнем шаге в качестве окончательной оценки значения функции распределения используется среднее от двух проведенных поправок:

$$\forall i = 1:K \quad F^{**}(z_i) = \frac{F_U^*(z_i) + F_D^*(z_i)}{2}.$$

Очевидно, что если проблем с оценкой не было, она в результате этих манипуляций не изменится.

После проведения оценки индикаторным кригингом и необходимой коррекции получены значения оценки локальной функции распределения для пороговых значений. Для получения оценок кумулятивной локальной функции распределения для других значений используется простейший способ степенной интерполяции. Для значения из интервала между двумя порогами (случай интерполяции) это выполняется так:

$$F^{**}(z) = F^{**}(z_{k-1}) + \left[\frac{z - z_{k-1}}{z_k - z_{k-1}}\right]^{\omega} \left[F^{**}(z_k) - F^{**}(z_{k-1})\right] \quad z \in (z_{k-1}, z_k),$$

где ω — положительная величина. При $\omega = 1$ получается линейная интерполяция. При $\omega < 1$ более быстрый, чем линейный, рост оценки функции распределения в начале интервала сменяется более медленным на его второй половине. Чем меньше значение ω, тем дальше оценка функции распределения от линейной. При $\omega > 1$ все происходит наоборот. Примеры такой степенной оценки для одного отрезка интерполяции приведены на рис. 7.4.

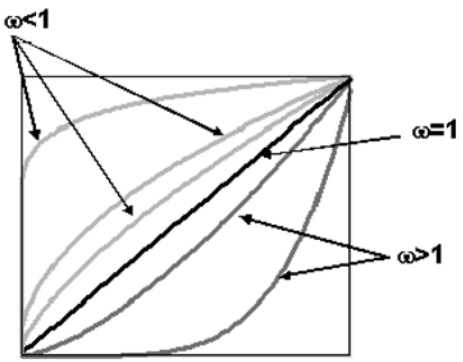


Рис. 7.4. Примеры степенной интерполяции, применяемой для интерполяции функции распределения

При необходимости оценить значение функции распределения за пределами минимального или максимального порогов (случай экстраполяции) можно пользоваться тем же методом, но требуется задать локальные значения $z_{\min}$ и $z_{\max}$. Для случая оценки меньше минимального порога (нижний хвост) используется формула

$$F^{**}(z) = \left[\frac{z - z_{\min}}{z_1 - z_{\min}}\right]^{\omega} F^{**}(z_1),$$

а при z больше максимального порога (верхний хвост) —

$$F^{**}(z) = F^{**}(z_K) + \left[\frac{z - z_K}{z_{\max} - z_K}\right]^{\omega} \left[1 - F^{**}(z_K)\right].$$

К этим случаям относится все, что было сказано выше относительно параметра степенной экстраполяции ω (см. рис. 7.4).

Для оценки верхнего хвоста можно использовать гиперболическую модель экстраполяции. Это позволяет строить функцию распределения до бесконечности, асимптотически приближая ее к единице. Гиперболическая модель имеет вид

$$F^{**}(z) = 1 - \frac{\lambda}{z^{\omega}},$$

где параметр ω (> 1) контролирует скорость приближения к граничному значению функции распределения (т. е. к единице). Чем больше ω, тем короче хвост. Параметр λ определяется оценкой функции распределения, полученной для максимального порога:

$$\lambda = z_K^{\omega}\left[1 - F^{**}(z_K)\right].$$

После выполнения индикаторного кригинга, корекции и интерполяции (экстраполяции) получены оценки локальных кумулятивных функций распределения в наборе точек пространства. Примеры таких локальных кумулятивных функций распределения (данные по загрязнению поверхности Брянской области ^{137}Cs) представлены на рис. 7.5. Для моделирования этих локальных функций распределения было сделано 19 пороговых значений.

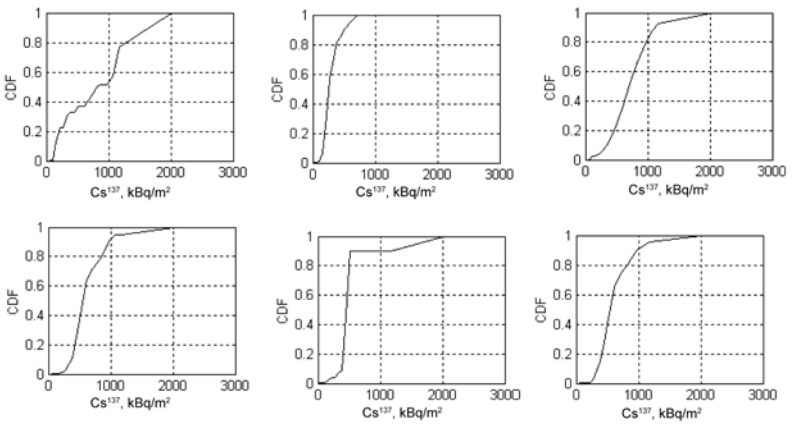


Рис. 7.5. Примеры локальных кумулятивных функций распределения, полученных с помощью индикаторного кригинга

По оцененным локальным кумулятивным функциям распределения можно получать значения и оценки, полезные для различного рода принятия решений.

Оценки E-типа (E-type) — среднее значение в точке. Название происходит от английской транскрипции математического ожидания (expectation). Они вычисляются по формуле

$$Z_E^*(x) = \int z dF(x,z) \approx \sum_{i=1}^{K-1} z_i \left(F^{**}(x, z_{i+1}) - F^{**}(x, z_i) \right).$$

Эти оценки можно сравнивать с интерполяциями другими методами, например обычным кригингом. Средние оценки E-типа можно представить на карте и гистограмме. Как всякое усреднение, оценка E-типа дает сглаженное значение.

P-квантиль — вероятность превышения (или непревышения $1 - p$) некоторого уровня значений z_c, что соответствует оцененной функции распределения. Понять это можно следующим образом: квантиль $p = 0{,}1$ значит, что настоящее (но неизвестное) значение функции с вероятностью 90% превысит этот уровень. Одновременно можно рассчитать средние значения оценок выше и ниже порога отсечения. Картирование таких вероятностей может быть полезно при анализе последствий выбросов для поддержки принятия квалифицированных решений, например о проведении контрмер.

Оценки, которые могут быть превышены с заданной вероятностью, — это оценки, соответствующие значению условной кумулятивной функции распределения (вероятности). Частым случаем такой оценки является *M-оценка* (медианная), которая может быть равновероятно превышена или не превышена. M-оценка соответствует значению p-квантиля 0,5 условной кумулятивной функции распределения.

При работе с категориальной переменной индикаторный кригинг интерпретируется несколько иным образом: оценка индикаторного кригинга дает вероятность некоторого значения. Эти вероятности позволяют решать задачу классификации по правилу выбора наиболее вероятного значения категориальной переменной в точке (т. е. классу с максимальной вероятностью). В этом случае также имеются математические ограничения на набор оценок в точке: значение вероятности должно быть положительным, значение вероятности не должно превышать 1, сумма вероятностей по всем возможным значениям категориальной переменной должна быть равна 1. В силу уже обсуждавшихся выше причин индикаторный кригинг не гарантирует выполнение этих ограничений. Таким образом, в случае категориальной переменной также проводится коррекция. Она состоит в ограничении значений, эквивалентном тому, которое делалось для непрерывной функции (7.8):

$$\forall i = 1 : K \; P^*(i) = \begin{cases} 1 & P^*(i) \geq 1, \\ P^*(i) & P^*(i) \in [0,1], \\ 0 & P^*(i) \leq 0. \end{cases}$$

Коррекция для удовлетворения равенства суммы единице состоит в традиционной нормировке:

$$\forall i = 1 : K \quad P^{**}(i) = \frac{P^*(i)}{\sum_{j=1}^{K} P^*(j)}.$$

Можно предположить, что при оценке индикаторов (каждого по отдельности) информация о данных используется не полностью. Более полным было бы построение многопеременной модели с учетом всех индикаторов сразу (индикаторный кокригинг, по аналогии с кокригингом при многопеременном анализе, рассмотренным в Главе 6). Но такой подход имеет ряд настолько существенных недостатков, что не применяется на практике.

- Кокригинг требует оценки и моделирования пространственных корреляционных структур для каждого индикатора и, кроме того, оценки и моделирования пространственной кросс-корреляционной структуры для всех пар порогов. В случае с 10 порогами требуется промоделировать 55 вариограмм и кросс-вариограмм (при условии, что они симметричны при перестановке номеров индикаторов). Моделирование такого их количества вносит ошибки, возникающие из-за неточности оценки и при самом моделировании.
- Системы уравнений кокригинга также имеют большую размерность.
- Использование ограничений кокригинга (возникающих из-за требования несмещенности оценки) еще чаще вызывает появление отрицательных весовых коэффициентов, а следовательно, и искаженной оценки локальной кумулятивной функции распределения.

Хотя в теории индикаторный кокригинг должен быть лучше, на практике это не подтверждается [Goovaerts, 1997].

7.3. Примеры использования индикаторного подхода

7.3.1. Зонирование гидрогеологического слоя

В этом примере рассматривается применение индикаторного кригинга к категориальным данным — зонирование гидрогеологического слоя. Зонирование — это подход, альтернативный анализу пространственной вариабельности гидрологических параметров, таких как пористость, проводимость и т. п. Он состоит в том, что область разбивается на зоны, гидрологические параметры внутри которых считаются постоянными. Обычно использование такого подхода обусловлено недостатком измерений для моделирования непосредственно гидрологических параметров.

В рассматриваемом примере проводится зонирование хорошо проводящего гидрогеологического слоя [Savelieva et al., 2003], который представлен пятью типами (зонами): тремя типами гравия, песка и ила. Исходные данные представляют собой 225 измерений, которые изображены на рис. 7.6 с помощью полигонов Вороного, которые здесь использованы исключительно для улучшения визуализации, так как при рисовании пространственного распределения точек измерений часть из них перекрывается.

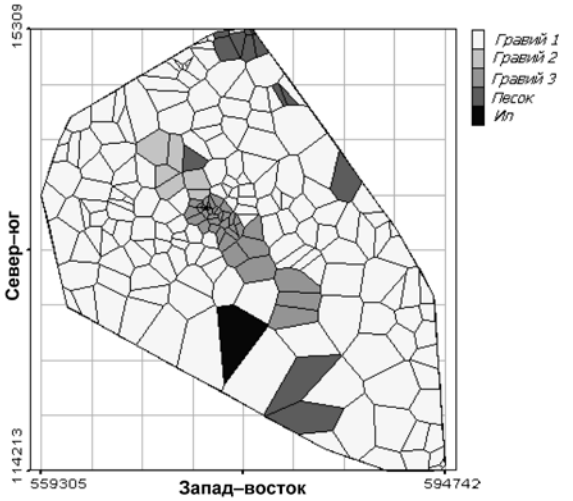


Рис. 7.6. Исходные данные по гидрогеологическим классам в виде полигонов Вороного

Первый шаг состоит в индикаторном преобразовании исходных данных: каждой точке измерений ставится в соответствие набор из пяти значений, четыре из которых являются 0, а один — 1. Их также можно рассматривать как пять пространственно распределенных индикаторных функций, принимающих значения 0 или 1. Для каждой индикаторной функции оценивается и моделируется пространственная корреляция. Индикаторный кригинг дает оценки вероятности принадлежности точки к соответствующему классу. Как указывалось выше, эти оценки могут требовать коррекции. В данном случае такой необходимости не было. На рис. 7.7 приведены вероятностные карты для гравия первого типа и песка.

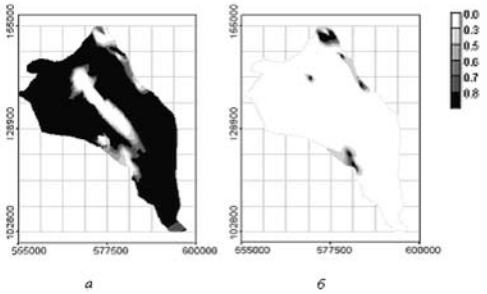
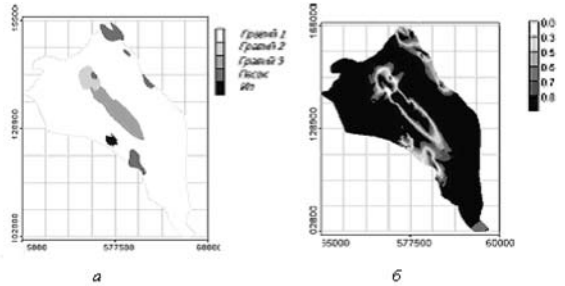


Рис. 7.7. Вероятность классов:
а — гравий; *б* — песок

Окончательное решение (классификация) принимается в пользу класса с максимальной вероятностью. На рис. 7.8 приведены результат классификации и вероятность класса-победителя, т. е. максимальная среди классов в этой точке.



а *б*

Рис. 7.8. Результат классификации (*а*) и вероятность класса победителя (*б*)

Правило принятия решения может быть изменено, если нужно классифицировать только те области, где вероятность одного из классов действительно высока (например, больше 0,7). Если в задаче присутствует больше двух классов, то могут быть области, где ни один класс не преодолевает этот барьер, а такие области остаются неклассифицированными. Они характеризуют зону неопределенности данной задачи.

На рис. 7.9 представлен результат классификации для задачи зонирования гидрогеологического слоя с использованием такого правила классификации и границей допустимой вероятности классификации 0,7. На рисунке видны белые зоны — это неклассифицированные зоны. Видно, что все зоны неопределенности расположены вдоль границы смены классов, следовательно, они являются аналогом «толстых изолиний» (подробнее о «толстых изолиниях» см. в Главе 5); где-то в пределах этих зон происходит смена одного класса на другой.

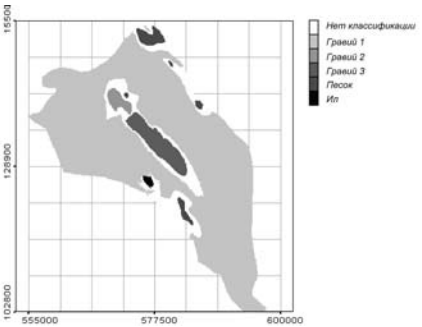


Рис. 7.9. Результат классификации с использованием барьера вероятности 0,7

7.3.2. Позиционирование скоплений крабов

Этот пример показывает использование индикаторного кригинга при работе с переменной, которую можно рассматривать как непрерывную — она имеет бесконечное множество значений (целые числа). Рассматриваем пространственное распределение краба опилио в Беринговом море. Измерения проводились траловой съемкой. Результат измерения представлен числом крабов в определенном месте. Диапазон значений от 0 до 821 442 характерен для краба. Как уже было показано, такой диапазон значений существенно усложняет задачу интерполяции.

Но на самом деле интерполяционные значения не представляют особого интереса, гораздо важнее обнаружить места скоплений крабов. Таким образом, задачу можно свести к определению вероятности обнаружения скопления крабов в данном месте. Количество крабов больше 5000 будем считать скоплением. На рис. 7.10 представлено пространственное распределение индикатора для порогового значения 5000.

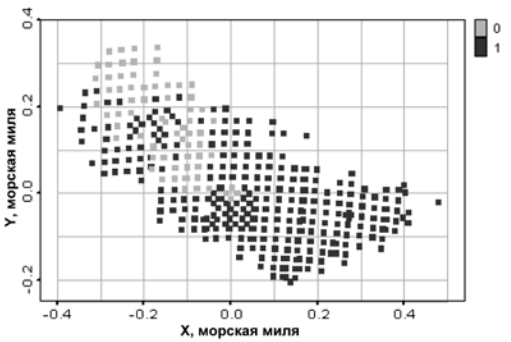


Рис. 7.10. Пространственное распределение индикатора для порогового значения 5000 крабов опилио

Индикаторный кригинг позволяет оценить вероятность найти скопление крабов, т. е. вероятность, что число крабов в данном месте будет больше или равно 5000. Оценка индикаторного кригинга дает вероятность того, что значение функции не превышает порогового значения. Но если вычесть оценку индикаторного кригинга из единицы, то получится искомая величина. На рис. 7.11 изображена вероятность обнаружения скопления краба опилио. Белые плюсы показывают места, где результаты измерений указывали на скопления. Соответствие оценки и реальных измерений представляется вполне хорошим.

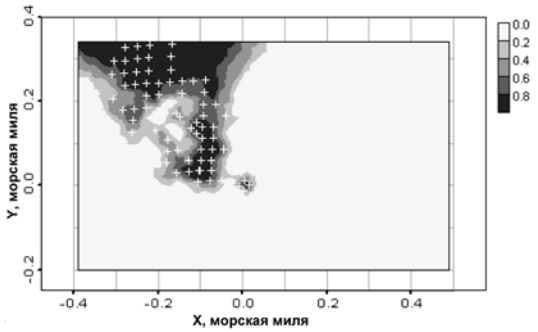


Рис. 7.11. Карта вероятности обнаружить скопление крабов опилио

Литература

Goovaerts P. Geostatistics for Natural Resources Evaluation. — [S. l.]: Oxford Univ. Press, 1997. — 483 p.

Journel A. G. Nonparametristic Estimation of Spatial Distribution // Mathematical Geology. — 1983. — Vol. 15, N 3.

Journel A. G. The Deterministic Side of Geostatistics // Mathematical Geology. — 1985. — Vol. 17, N 1.

Savelieva E., Bolshov L., Pozdnukhov A. et al. Zonation of Hanford Formation at the Hanford Site / Nuclear Safety Inst. RAS. — Moscow, 2003. — 55 p. — (Preprint IBRAE-2003-07).

Глава 8

Стохастическое моделирование пространственной неопределенности

В предыдущих главах были рассмотрены регрессионные геостатистические модели — кригинг. Кригинг, как и другие регрессионные оцениватели, позволяет получить единственное значение оцениваемой функции в точке для заданного набора данных и выбранных параметров модели. Единственность оценки наряду с известными преимуществами несет в себе ряд ограничений. В этой главе мы рассмотрим альтернативный метод — стохастическое моделирование.

Раздел 8.1 посвящен основам стохастического моделирования и его принципиальному отличию от регрессионного оценивания. Здесь же представлены основные подходы к стохастическому моделированию, его виды и кратко перечислены существующие алгоритмы. В разделе 8.2 описан ключевой для большинства стохастических моделей геостатистики принцип последовательного моделирования. Следующие разделы посвящены конкретным алгоритмам, которые базируются на последовательном принципе моделирования. В разделе 8.3 подробно описано последовательное гауссово моделирование, в разделе 8.4 — обрезанное гауссово моделирование, в разделе 8.5 — последовательное индикаторное моделирование, в разделе 8.6 — последовательное прямое моделирование. Принципиально другой алгоритм — моделирование отжига — представлен в разделе 8.7. Раздел 8.8 посвящен объектному подходу к стохастическому моделированию. В разделе 8.9 собраны практические упражнения и вопросы по нескольким тонким моментам стохастического моделирования.

8.1. Основы стохастического моделирования

Рассмотрим проблему мониторинга радиоактивного загрязнения почвы. Измерения активности пробы, взятой на участке 10 км^2, могут различаться и подвержены ошибке измерительного прибора. Далее, измерения, собранные на площади 1 м^2, могут иметь еще больший разброс значений, обусловленный пространственной вариабельностью загрязнения на микромасштабе (1 м). Если использовать регрессионную модель для интерполяции загрязнения на сетке с разрешеньем 1 км (шаг сетки), то полученная

оценка кригинга будет отражать некое среднее значение загрязнения в каждой ячейке сетки. Принимая во внимание, что ошибка кригинга имеет безусловный характер (она зависит не от данных измерений, а только от их плотности), можно заключить, что модель кригинга не позволяет адекватно оценить неопределенность и вариабельность пространственного распределения в точке оценивания.

Для оценки вариабельности пространственной функции используют методы стохастического моделирования, которое в отличие от кригинга позволяет получить множество реализаций значений функции в точке оценивания для заданного набора данных и выбранных параметров модели [Journel, Huijbregts, 1978].

Проиллюстрируем сказанное на простом одномерном примере (рис. 8.1). Через имеющиеся шесть данных измерений проведены кривые оценок кригинга (жирные линии). Они проходят через все шесть точек измерений, поскольку кригинг является точным оценивателем. Кривые оценок плавно соединяют точки измерений, при этом оценка не может иметь значения выше максимального и ниже минимального измеренного. Разница между обычным (ОК) и простым (ПК) кригингом в данном случае несущественна. Стохастические реализации нанесены пятью более тонкими линиями. Они тоже проходят через точки измерений, поскольку удовлетворяют данным точно. Принципиальное различие между реализациями состоит в разнообразии возможных значений вне точек измерений. Вследствие стохастической природы модели функция может принимать значения выше и ниже измеренного и воспроизводить разнообразие стохастической динамики изменения значений — вариабельности. Отметим, что вариабельность меняется в зависимости от близости данных измерений. Так, в интервале абсцисс {8, 20} и {59, 76} вариабельность значительно ниже, чем там, где отсутствуют измерения.

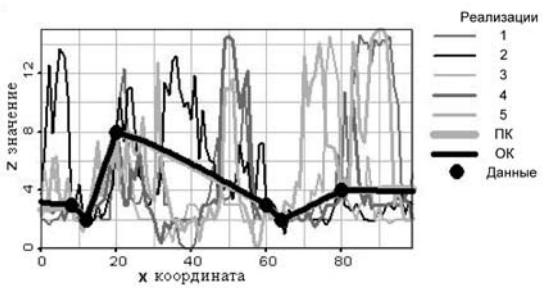


Рис. 8.1. Стохастические реализации и оценки кригинга в одномерном случае

Методы стохастического моделирования строят набор оценок значения (реализаций) функции на всей области. При этом каждая реализация обладает определенными свойствами (в том числе и вариабельностью) исходного процесса, т. е. любая реализация является возможной моделью исходных данных. Таким образом, стохастическое моделирование позволяет оценивать пространственную неопределенность процесса.

При стохастическом моделировании оценивается совместная условная функция распределения всего процесса, поэтому каждая сгенерированная пространственная реализация стремится воспроизвести следующие свойства исходного распределения:

- плотность распределения;
- статистические характеристики исходных данных;
- пространственную корреляционную структуру.

Задача оценки совместной условной функции распределения решается путем построения набора стохастических равновероятных пространственных реализаций. Таким образом, разброс значений реализаций в каждой локальной точке определяет вариабельность модельной оценки. Совместное пространственное распределение вариабельности позволяет воспроизвести неопределенность оценки реальных распределений, а также локальные флуктуации значений неизвестного пространственного распределения.

Стохастическое моделирование может быть *условным* — зависимым от данных (см. рис. 8.1) — или *безусловным*, когда нет данных измерений. При условном моделировании данные измерений воспроизводятся точно, как и при оценке кригинга, и влияют на остальные значения реализации. При безусловном моделировании воспроизводятся только заданные априори функционалы — статистические моменты первого и второго порядков (среднее, вариация, пространственная корреляционная структура).

Существуют методы безусловного моделирования как для категориальных, так и для непрерывных переменных.

Самый известный метод безусловных симуляций — метод «вращающихся лент» (turning bands) [Mantoglou, Wilson, 1981]. Он позволяет построить безусловные реализации гауссова поля $Y(x)$ с известной ковариационной функцией $C_Y(\boldsymbol{h})$. Основная идея метода состоит в построении одномерных реализаций вдоль 15 линий, различным образом ориентированных и разбивающих трехмерную область на примерно равные части. Каждый узел 3D-пространства, где производится моделирование, проектируется в некоторую точку на каждой линии. Значение в узле задается суммой значений этих проекций.

Моделирование одномерной реализации (вдоль линии) обычно выполняется с использованием спектрального подхода на основе быстрого преобразования Фурье [Christakos, 1992]. Функция ковариации $C_Y(\boldsymbol{h})$, описывающая пространственную корреляцию моделируемого поля, при этом подходе заменяется соответствующей ей функцией спектральной плотности $S_Y(\omega)$:

$$S_Y(\omega) = \frac{1}{(2\pi)^n} \int_{R^n} C_Y(\boldsymbol{h}) e^{-i\omega \boldsymbol{h}} d\boldsymbol{h}.$$

Подробности безусловного моделирования выходят за рамки данной книги.

Существуют различные подходы к стохастическому пространственному моделированию [Chiles, Delfiner, 1999]. Один из них основан на последовательном принципе моделирования. Другой подход использует методы глобальной минимизации целевой функции.

Основой *последовательного подхода* является возможность перейти от совместной условной функции распределения к произведению локальных условных функций распределения. Последовательный принцип может использоваться как в условном, так и в безусловном моделировании. На принципе последовательного моделирования основано большинство геостатистических стохастических методов, строящих значения функции последовательно в каждой ячейке сетки. Такие модели называют *ячейковыми* (или *пиксельными*). Они отличаются от *объектных* моделей, которые моделируют значения в локальной окрестности, определяющейся геометрической формой (объектом).

При использовании *минимизации целевой функции* все характеристики исходных данных, которые требуется воспроизвести, формализуются в целевую функцию. Для ее минимизации применяются стохастические методы глобальной минимизации. Такие подходы могут использоваться и для условного, и для безусловного моделирования. К ним относятся объектные модели и некоторые пиксельные, например моделирование отжига. Объектные модели основаны на случайном пространственном распределении объектов заданных геометрических форм и размеров. Это предполагает априорное задание этих форм в качестве альтернативы модели пространственной корреляции. Методы глобальной минимизации могут использоваться как для условных, так и для безусловных симуляций. Это определяется включением дополнительного компонента, отвечающего за воспроизведение исходных данных в целевую функцию. Условное моделирование при помощи объектных моделей может быть сопряжено с рядом трудностей, связанных с

сохранением точного воспроизведения данных при итерационной оптимизации. Обычно итерационная подгонка объектных моделей требует значительных вычислительных затрат.

В результате стохастического моделирования получаются равновероятные пространственные реализации переменной. Они характеризуют пространственную вариабельность и локальную неопределенность пространственной функции. Анализируя пространственные реализации, можно получить вероятности превышения значений функции и пр.

Пространственные реализации, полученные в результате применения любого метода стохастического моделирования, могут использоваться как входные данные модели оценки функции распределения пространственной переменной. Это дает возможность оценить локальную вариабельность пространственного распределения. Возможные реализации геологической среды, полученные в результате стохастического моделирования, могут использоваться для оценки вероятности попадания загрязнения через грунтовые воды. Реализации стохастического моделирования для ошибок при прогнозировании потребления электроэнергии могут использоваться при оценке неопределенности (доверительного интервала) прогноза и т. п. По результату стохастического моделирования можно оценивать вероятность превышения или непревышения определенных значений как в отдельной точке (задача моделирования локальной функции распределения), так и в нескольких точках одновременно (рассматривается совместная функция распределения). На практике чаще моделируются и используются двухточечные совместные функции распределения. Стохастическое моделирование позволяет также получить оценки вероятности превышения заданного уровня (p-value) и оценки с заданным уровнем вероятности превышения. Такие оценки, а также анализ неопределенности пространственной оценки, крайне важны для поддержки принятия квалифицированных решений. При усреднении стохастических реализаций можно получить среднюю оценку (E-type), сравнимую с аналогичной оценкой для индикаторного кригинга (см. Раздел 7.2). Разница между стохастическими реализациями позволяет оценить вариацию и разброс значений локальных функций распределения.

Стохастическое моделирование может проводиться и в случае многопеременной функции, тогда реализации строятся для основной переменной с использованием дополнительных аналогично тому, как делается кокригинг (см. Главу 6). В данной главе для простоты будем рассматривать одномерную функцию.

Пиксельные модели

- моделируют значение в каждой отдельно взятой ячейке;
- позволяют легко интегрировать дополнительную локальную информацию;
- используют модели пространственной корреляции.

Объектные модели используют реалистичные структуры (объекты различной геометрической формы), но имеют следующие недостатки:

- предположение о формах объектов требует априорных знаний о структуре системы;
- возникают неопределенности, связанные с выбором форм, размеров и местоположения объектов;
- требуются высокие вычислительные затраты на итерационные алгоритмы выбора оптимального местоположения объектов.

В заключение Раздела кратко перечислим алгоритмы, использующиеся для условного стохастического моделирования в задачах пространственного и пространственно-временно́го оценивания, большинство которых будет подробно разобрано ниже.

1. *Последовательное гауссово моделирование* (sequential Gaussian simulation) — моделирует непрерывные переменные (например, значение загрязнения, пористости породы, биомассу рыбы).
2. *Обрезанное гауссово моделирование* (truncated Gaussian simulation) — моделирует категориальные переменные (например, типы почв, геологических пород).
3. *Последовательное индикаторное моделирование* (sequential indicator simulations) — моделирует как категориальные переменные, так и непрерывные (типы пород, уровень загрязнения).
4. *Прямое моделирование* (direct simulations) — моделирует непрерывные переменные.
5. *Моделирование отжига* (simulated annealing) — моделирует категориальные и непрерывные переменные.
6. *Объектное моделирование* (object modelling) — моделирует категориальные переменные, использует заданный набор геометрических объектов.
7. *Многоточечное моделирование* (multipoint statistics simulation) — первоначально реализовано для категориальных данных, но недавно были разработаны алгоритмы многоточечной статистики для моделирования и непрерывных переменных. Этот подход подробно рассмотрен в Главе 11.

8.2. Последовательный принцип моделирования

Геостатистика интерпретирует измерения пространственной переменной $Z(x_i)$ в точках x_i ($i = 1, ..., n$) как реализацию значений случайной функции $Z(x)$, которая определенна в области S и характеризуется совместной условной функцией распределения $F(x|z)$. В рамках такой системы проблема заключается в генерации K реализаций этой случайной функции, детально покрывающих область S набором из N точек [x_1, ..., x_N]. Каждая реализация соответствует совместной условной функции распределения, в которую включены все статистические характеристики процесса:

$$F\left(x_1, ..., x_N; z_1, ..., z_N \mid (n)\right) = \Pr\left\{Z\left(x_1\right) \le z_1, ..., Z\left(x_N\right) \le z_N \mid (n)\right\}. \quad (8.1)$$

Принцип последовательного моделирования основан на использовании правила Байеса:

$$\Pr\{A, B \mid C\} = \Pr\{A \mid B, C\}\Pr\{B \mid C\}. \quad (8.2)$$

Применяя последовательно (N раз) формулу (8.2) к правой части формулы (8.1) и переписав ее в виде функции распределения, получаем:

$$\begin{aligned} &F\left(x_1, ..., x_N \mid (n)\right) = \\ &= F\left(x_N; z_N \mid (n+N-1)\right) F\left(x_{N-1}; z_{N-1} \mid (n+N-2)\right) ... F\left(x_1; z_1 \mid (n)\right), \end{aligned} \quad (8.3)$$

где $F\left(x_{N-1}; z_{N-1} \mid (n+N-2)\right)$ — условная функция распределения $Z(x)$, определяемая n исходными данными и $N-1$ значениями реализации $z(x_j)$, $j = 1, ..., N-1$. Полученное выражение и является теоретической основой последовательных методов стохастического моделирования.

На практике последовательный принцип реализуется через последовательное включение в процесс моделирования уже сгенерированных значений. Подробнее практическую схему построения одной реализации можно записать в виде выполнения последовательности шагов (рис. 8.2).

1. Построение случайной последовательности из набора N [x_1, ..., x_N] точек для оценки $\left[x_1^k, ..., x_N^k\right]$, где k — номер реализации.
2. Для точки x_1^k проводится оценка локальной функции распределения по набору исходных данных. В соответствии с оцененной функцией рас-

пределения разыгрывается значение функции $Z(x)$ в точке x_1^k $\left(z\left(x_1^k\right)\right)$. Это значение добавляется к исходным данным.

3. Для точки x_m^k проводится оценка локальной функции распределения по набору исходных данных и полученным на предыдущих $m-1$ шагах значениям $z\left(x_1^k\right), \ldots, z\left(x_{m-1}^k\right)$. В соответствии с оцененной функцией распределения разыгрывается значение функции $Z(x)$ в точке x_m^k $\left(z\left(x_m^k\right)\right)$. Это значение также добавляется к исходным данным.
4. Шаг 3 повторяется для всех последующих $N-1$ точек в соответствии с последовательностью, полученной на шаге 1. Каждый раз оценка локальной функции распределения производится по исходным данным и значениям, сгенерированным на предыдущих шагах.

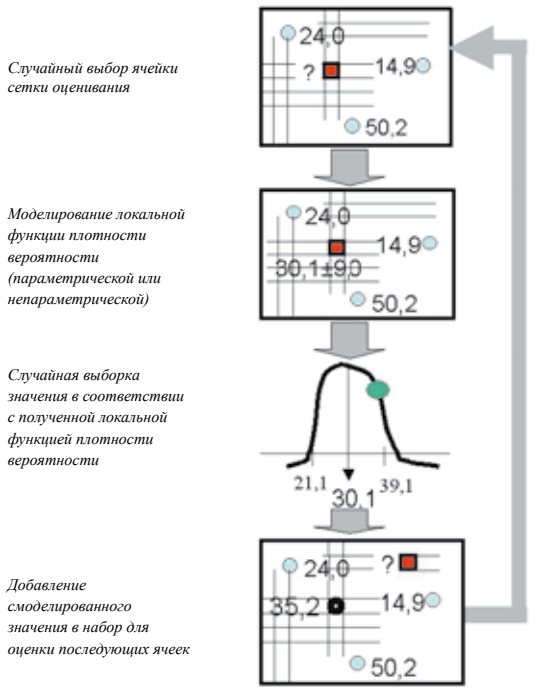


Рис. 8.2. Пошаговый алгоритм последовательного моделирования

Вся процедура повторяется для всей сетки оценивания столько раз, сколько предполагается получить различных реализаций. При этом стохастическая природа реализаций обусловливается случайностью разыгрывания (выборки) значения в каждой точке по оцененной локальной функции рас-

пределения. Случайная последовательность обхода всех ячеек сетки позволяет воспроизвести стохастическое разнообразие во всем пространстве и избежать ненужных артефактов (необоснованно большого количества каких-либо значений).

Ключевым моментом в последовательном моделировании является построение локальной функции распределения в точке оценивания на основе данных в ее окрестности и модели пространственной корреляции. Для построения этой локальной функции распределения требуется принять некоторые предположения относительно ее формы или аналитического вида. В зависимости от этих предположений можно выделить два типа алгоритмов — параметрический и непараметрический. Параметрические алгоритмы предполагают аналитический вид локальной функции распределения, которая зависит от набора параметров. Так, предположение о локальной нормальности распределения позволяет использовать известную параметрическую функцию, заданную двумя параметрами — средним значением и вариацией. Если вид и форма локального распределения не могут быть заданы аналитически, то можно использовать непараметрические методы. Локальную функцию распределения можно задать в табулированном виде, основываясь на предположении о форме распределения. Индикаторное моделирование и прямое моделирование относятся к непараметрическим методам. В них локальная функция задается непосредственными значениями плотности вероятности, которые получены на основе имеющихся данных. Непараметрическое задание локальной функции распределения использует интерполяцию между табулированными значениями плотности вероятности. При этом вид интерполяционной функции выбирается в зависимости от априорных предположений (см. рис. 7.4).

Остановимся на первом шаге описанной выше процедуры. Теоретически используется любой путь от одной точки оценивания к другой. На практике случайный путь, примерно равномерно заполняющий все зоны области, предпочтительнее регулярного, стартующего в одной зоне области и заполняющего сначала ее. Это позволяет избежать возможного распространения артефактов, вызванных сильным ростом количества похожих соседей в результатах.

Если моделирование проводится на регулярной сетке, то имеет преимущества концепция *промежуточных сеток*, в частности при воспроизведении вариограммных структур с очень большими радиусами корреляции. При таком подходе моделирование начинается с очень грубой сетки, потом

она дополняется до менее грубой, и это продолжается до получения сети, на которой проводится моделирование. В каждой из подсеток выбирается случайный путь следования от узла к узлу. Количество промежуточных сеток зависит от радиусов корреляции вариограмм и конечного размера ячеек сетки. Преимуществом такой схемы является большая стабильность моделирования и получаемых симуляций — после моделирования одной подсетки условные значения (по крайней мере, промоделированные) располагаются регулярно, что предотвращает возможную сильную кластеризацию и развитие артефактов.

Многие модели геостатистического последовательного стохастического моделирования используют кригинг при моделировании локальной функции распределения: либо для моделирования параметров распределения, либо для моделирования табулированных вероятностей для непараметрических методов. При использовании кригинга может возникнуть *эффект экранирования* [Isaaks, Srivastava, 1989]. Он состоит в уменьшении веса точек, попадающих между одной из точек измерения и точкой оценивания, что может привести к появлению отрицательных весов. Таким образом, эффект экранирования будет усиливаться за счет добавления вновь смоделированных значений к данным после каждого шага оценивания. На практике не обязательно использовать все существующие значения измерений для построения условного распределения в точке оценивания. Необходимо ограничить используемые условные измерения окрестностью точки оценивания, при этом можно ограничивать отдельно количество исходных данных и количество уже смоделированных значений. Более подробно эффект экранирования описан в Главе 5.

Среди наиболее широко используемых алгоритмов последовательного моделирования можно выделить:

- гауссово;
- обрезанное гауссово;
- индикаторное;
- прямое;
- многоточечное и др.

Первые два алгоритма являются параметрическими, остальные — непараметрическими. Эти алгоритмы отличаются различными предположениями о локальном законе распределения.

Далее мы рассмотрим некоторые наиболее часто используемые алгоритмы.

8.3. Последовательное гауссово моделирование

Метод последовательного стохастического гауссового моделирования предполагает совместное нормальное распределение моделируемой случайной величины в исследуемой области. Совместное нормальное распределение называется мультинормальным и предполагает распределение всех компонент (во всех локальных точках) по стандартному нормальному закону. В этом случае для любой точки области локальная функция распределения будет распределена по нормальному закону и будет определяться двумя параметрами — средним и вариацией.

Реальные данные, как правило, не являются нормально распределенными, поэтому для применения гауссового моделирования требуется предварительная подготовка. Она заключается в преобразовании данных в нормальное распределение и проверке обоснованности гипотезы о мультинормальности.

На первом этапе моделирования предполагается стационарность случайной функции $Z(x)$ и существование такой случайной функции $Y(x)$

$$Y(x) = \varphi[Z(x)] \text{ и } Y(x) \sim N(0,1),$$

где φ — однозначная функция нормализующего (normal score) преобразования (рис. 8.3).

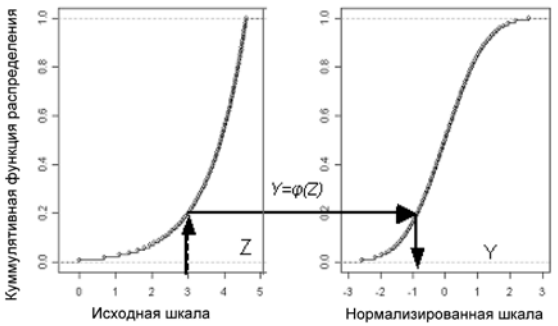


Рис. 8.3. Нормализация исходных данных

Гауссово преобразование φ производится путем постановки в соответствие реальной кумулятивной функции распределения исходных данных кумулятивного стандартного нормального распределения:

$$G(y) = F(z) \Leftrightarrow y = \varphi(z) = G^{-1}[F(z)] \Leftrightarrow z = \varphi^{-1}(y) = F^{-1}[G(y)],$$

где $F(\cdot)$ — кумулятивное распределение частоты исходных данных; $G(\cdot)$ — стандартное кумулятивное нормальное распределение. Такое соответствие проиллюстрировано на рис. 8.3. На практике при проведении прямого преобразования строится специальная таблица соответствия значений для обратного преобразования.

После преобразования распределение данных становится нормальным, т. е. функция $Y(x)$ является стационарной в строгом смысле и подчиняется стандартному нормальному закону распределения. Как следствие этого полный вероятностный закон распределения $Y(x)$ известен, если известны среднее значение и ковариационная функция. Среднее значение равно нулю в силу стандартности нормального распределения. Ковариация выражается через вариограммы следующим образом:

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

где $C(h)$ — ковариация; $\gamma(h)$ — вариограмма функции $Y(x)$.

Для корректного применения гауссова стохастического моделирования требуется, чтобы случайная функция $Y(x)$ была распределена мультинормально. Нормализующее преобразование, вообще говоря, не гарантирует мультинормальность. Полученная в результате нормализующего преобразования переменная распределена одномерно нормально по построению. Это, однако, необходимое, но не достаточное условие мультинормальности ее пространственного распределения. Для корректного использования алгоритма требуется проверка мультинормальности. Вследствие отсутствия простого теста на мультинормальность обычно ограничиваются проверкой бинормальности (совместная функция распределения для любых пар точек нормальна), что может считаться достаточным в общих предположениях классической геостатистики, где стационарность также ограничивается стационарностью второго порядка. В геостатистике достаточность проверки на бинормальность определяется использованием двухточечных моментов второго порядка — вариограмм.

Для проверки на бинормальность условной функции распределения любого набора пар данных $\{y(\boldsymbol{x}_i), y(\boldsymbol{x}_i + \boldsymbol{h}), i = 1, ..., N(\boldsymbol{h})\}$ используется ковариация $C_Y(\boldsymbol{h})$. Существуют аналитическое и табулированное соотношения между ковариацией $C_Y(\boldsymbol{h})$ и значением стандартной нормальной функции распределения [Deutsch, Journel, 1998]:

$$\Pr\{Y(\boldsymbol{x}) \le y_p, Y(\boldsymbol{x}+\boldsymbol{h}) \le y_p\} = p^2 + \frac{1}{2\pi}\int_{0-}^{\arcsin C_Y(\boldsymbol{h})} \exp\left(-\frac{y_p^2}{1+\sin\theta}\right) d\theta,$$

где $y_p = G^{-1}(p)$ — стандартный нормальный p-квантиль; $C_Y(\boldsymbol{h})$ — ковариация стандартной нормальной случайной функции $Y(\boldsymbol{x})$.

Бинормальная вероятность является нецентральной индикаторной ковариацией (см. (7.3)) для порога y_p:

$$\Pr\{Y(\boldsymbol{x}) \le y_p, Y(\boldsymbol{x}+\boldsymbol{h}) \le y_p\} = E\{I(\boldsymbol{x}, p) I(\boldsymbol{x}+\boldsymbol{h}, p)\} = p - \gamma(\boldsymbol{h}, p),$$

где $I(\boldsymbol{x}, p)$ — индикаторное преобразование для функции $Y(\boldsymbol{x})$ (см. (7.1)); $\gamma_I(\boldsymbol{h}, p)$ — индикаторная полувариограмма для p-квантиля отсечения y_p.

Таким образом, на практике тест на бинормальность сводится к сравнению табулированных значений с индикаторной вариограммой для набора квантилей.

Существует и более грубый тест на бинормальность [Emerly, 2005]. Он заключается в проверке соотношения вариограммы γ и мадограммы M:

$$\frac{\gamma}{M^2} \approx \pi.$$

Эта проверка является приблизительной, однако с ее помощью можно быстро определить близость распределения к бинормальному закону (рис. 8.4) — чем ближе значения частного к π, тем ближе распределение данных к бинормальному.

Для моделирования требуется также провести последовательный анализ нормализованных пространственных данных. Важно проанализировать условия принятия гипотезы о пространственной стационарности (второго порядка или внутренней). Существенная нестационарность данных должна быть предметом особого внимания. Так, области с различными статистическими характеристиками должны рассматриваться отдельно. Присутствующие тренды должны выделяться и рассматриваться отдельно с последующим добавлением по окончании моделирования. На этом же этапе проводится декластеризация, если этого требует сеть мониторинга.

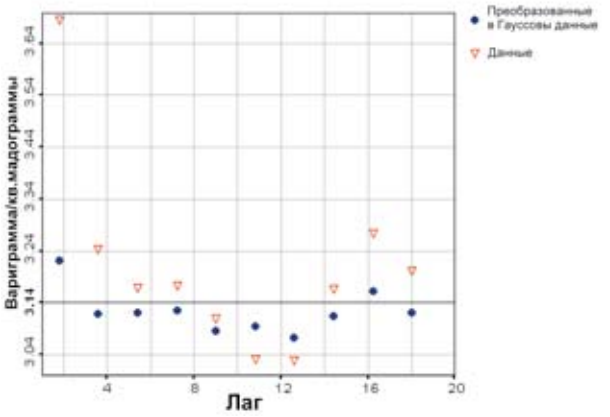


Рис. 8.4. Проверка на бинормальность исходных данных и нормализованных данных

Следующим шагом является построение ковариации для нормализованных значений $Y(\boldsymbol{x})$ и ее моделирование. Эта задача подробно описана в Главе 4. Напомним, что ковариация переменной y для вектора $\boldsymbol{h} = \boldsymbol{x}_i - \boldsymbol{x}_j$ обозначается как $C_{ij} = C(\boldsymbol{x}_i - \boldsymbol{x}_j)$.

Вариограмма нормализованных значений, как правило, стабильнее и устойчивее, чем вариограмма исходных данных. Это объясняется сглаживанием влияния крайних предельных значений при нелинейном нормализующем преобразовании. Плато вариограммы нормализованных значений должно быть равно единице, поскольку она является априорной вариацией стандартного нормального распределения нормализованных данных.

На следующем этапе проводится собственно последовательное моделирование нормализованных значений по алгоритму, соответствующему реализации последовательного принципа моделирования (рис. 8.5).

Для определения очередности, в которой оцениваются точки (ячейки сетки), строится случайная последовательность из всех точек оценивания. Далее для каждой точки из этой последовательности оценка производится по следующему алгоритму.

1. Преобразование исходных данных в нормальное распределение, проверка на бинормальность и моделирование пространственной корреляционной структуры нормализованных данных.
2. Выбор случайной последовательности из всех точек оценивания.
3. В каждой точке оценивания производятся:

- оценка параметров локального распределения плотности вероятности (среднего и вариации) с помощью простого кригинга на основе преобразованных исходных данных и уже сгенерированных значений в других точках последовательности;
- выборка случайного значения в соответствии с нормальным распределением и полученными параметрами;
- добавление сгенерированного значения в общий набор для последующего использования при оценке простого кригинга.

4. Обратное преобразование промоделированных реализаций из нормализованных значений.

Для получения следующей равновероятной реализации повторяются шаги 2—4.

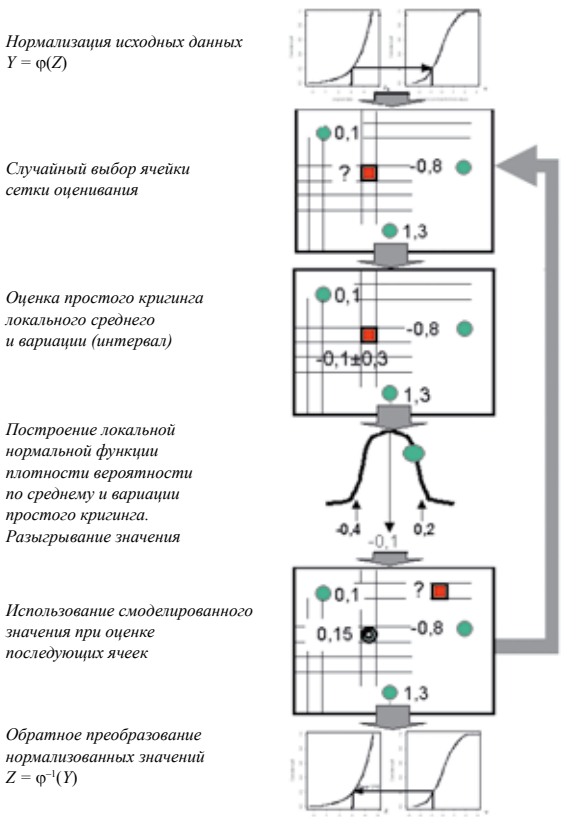


Рис. 8.5. Схема алгоритма последовательного гауссова моделирования

Ключевой момент последовательного гауссового моделирования — построение в каждой точке оценивания нормальной функции плотности вероятности. Локальная гауссова функция распределения определяется двумя параметрами — средним значением и вариацией. Для их получения используется простой кригинг с известным средним значением. Обычный кригинг в движущемся окне может быть использован для оценки среднего значения в нестационарном случае, но при этом вариация оценивается простым кригингом.

Оценка простого кригинга $Y^*_{SK}(x)$ в точке x вычисляется так (см. Раздел 5.2):

$$Y^*_{SK}(x) = \sum_{i=1}^{n(x)} \lambda_i^{SK}(x)\left[Y(x_i) - m\right] + m,$$

где среднее значение нормализованных данных m по области постоянно (в предположении о стационарности второго порядка) и равно нулю.

Веса простого кригинга λ_α^{SK} определяются путем решения системы $n(\boldsymbol{x})$ уравнений простого кригинга:

$$\sum_{j=1}^{n(x)} \lambda_j^{SK} C_{ij} = C_{i0}, \ \forall i = 1, \ldots, n(x),$$

где $n(x)$ — общее количество использующихся соседних точек x_i при оценке точки x.

Вариация оценки простого кригинга

$$\sigma^2_{SK}(x) = C(0) - \sum_{\alpha=1}^{n(x)} \lambda_\alpha^{SK}(x) C\left(x_\alpha - x\right).$$

В результате получаем параметры локальной нормальной функции распределения $N\left(Y^*_{SK}(x), \sigma^2_{SK}(x)\right)$ в точке оценивания x.

Далее проводится случайная выборка (по методу Монте-Карло) из полученного нормального распределения. Разыгранное значение является равновероятной стохастической реализацией значения функции $Y(x)$ в данной точке.

Полученное значение $Y(x)$ добавляется к набору данных и других уже промоделированных значений для использования в последующих оценках.

После прохода через все точки оценивания для получения окончательного результата моделирования полученные нормальные значения оценок $\left\{y(x), x \in A\right\}$ преобразуются обратно в абсолютные значения исходной

функции $\{z(x) = \varphi^{-1}(y(x)), x \in A\}$ с использованием обратного гауссова преобразования.

Если модель хорошо соответствует исходным данным, то реализации последовательного гауссова моделирования воспроизводят:

- стандартное нормальное распределение у преобразованной переменой;
- вариограммы преобразованных переменных;
- значения преобразованных переменных в точках измерений.

При обратном преобразовании промоделированных значений воспроизводятся распределения и измерения исходной переменной. Это также предполагает воспроизводство вариограмм исходной переменной.

Приведенная выше выборка из нормального распределения со средним, равным оценке простого кригинга $Y^*_{SK}(\boldsymbol{x})$, эквивалентна случайной выборке из нормального распределения с нулевым средним $N(0, \sigma_{SK}(\boldsymbol{x}))$. В этом случае значение стохастической реализации будет определяться по формуле

$$Y(\boldsymbol{x}) = Y^*_{SK}(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}),$$

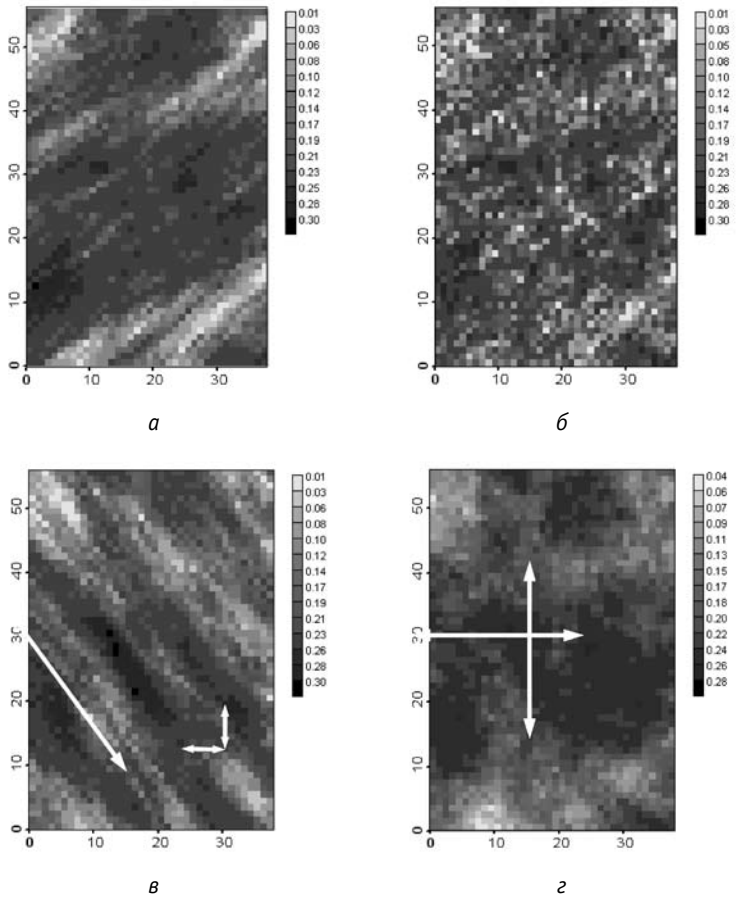
где компонента ошибки $\varepsilon(x)$ разыгрывается случайным генератором, моделирующим нормальное распределение с параметрами

$$E\{\varepsilon(\boldsymbol{x})\} = 0 \text{ и } \mathrm{Var}\{\varepsilon(\boldsymbol{x})\} = \sigma^2_{SK}.$$

Основным преимуществом метода последовательного гауссова моделирования является его простота, основанная на хорошо известном и понятном поведении нормального распределения.

Базовое предположение о мультинормальности совместных функций распределения дает ряд важных преимуществ. Выборка из локальных нормальных распределений гарантирует, что моделируемые стохастические распределения сохранят форму гауссова распределения наряду с другими параметрами (средним, вариацией, вариограммой). Сохранение последних может выполняться и при выборке по негауссовым распределениям, но при этом форма полученного в результате распределения будет изменена. Эта проблема не возникает, когда все локальные распределения имеют одну и ту же форму, что гарантируется предположением о мультинормальности. Также можно отметить, что в соответствии с центральной предельной теоремой последовательное добавление случайно выбранных значений дает в совокупности гауссово распределение.

Примеры реализаций последовательного гауссова моделирования с различными параметрами вариограммы приведены на рис. 8.6 для уровня нагета 0 и 40% от априорной вариации (*а, б*); изотропии (*г, е*), геометрической анизотропии (*в*) и зонной анизотропии (*д*) радиуса корреляции; угла направления длинной корреляционной структуры от вертикали 0° (*ж*), 30° (*в*) и 90° (*з*); соотношения радиусов корреляции в ортогональных направлениях 8 и 40 (*в*), 80 и 4 (*д*), 40 и 8 (*ж, з*) в единицах расстояния.



а *б*

в *г*

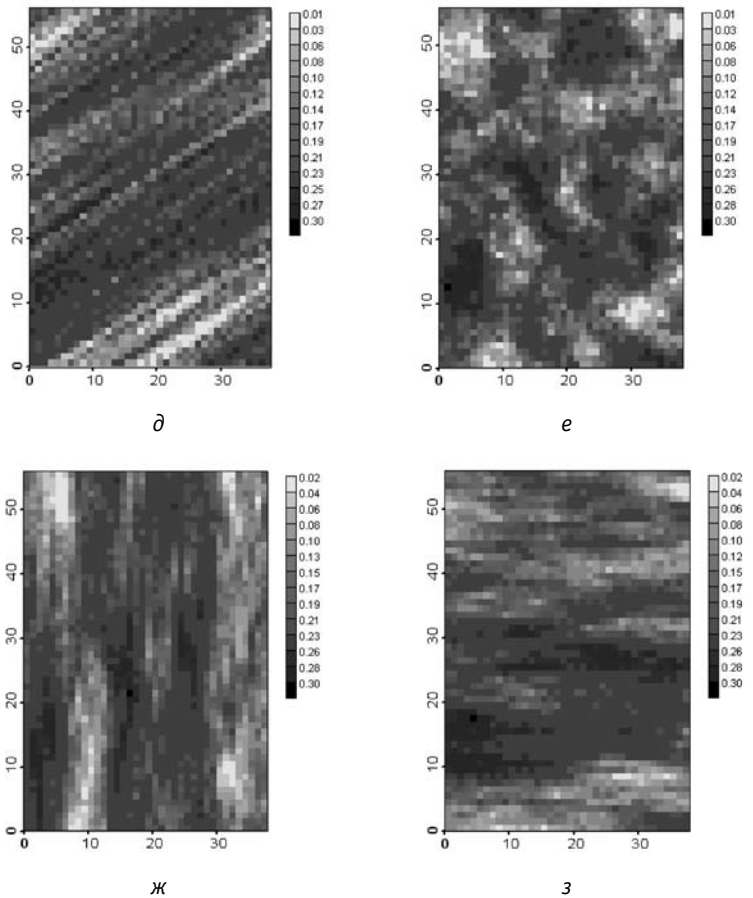


Рис. 8.6. Примеры стохастических реализаций последовательного гауссова моделирования для различных значений параметров модели вариограммы: *а* — R = 40, r = 8, наггет = 0.0, угол = 60; *б* — R = 40, r = 8, наггет = 0.4, угол = 60; *в* — R = 8, r = 40, наггет = 0.0, угол = 60; *г* — R = 40, r = 40, наггет = 0.0, угол = 60; *д* — R = 80, r = 4, наггет = 0.0, угол = 60; *е* — R = 8, r = 8, наггет = 0.0, угол = 60; *ж* — R = 40, r = 8, наггет = 0.0, угол = 0; *з* — R = 40, r = 8, наггет = 0.0, угол = 90

Предположение о локальной нормальности имеет и негативные стороны. Гауссово моделирование является алгоритмом *максимальной энтропии* — максимального беспорядка в стохастической реализации. Это означает слабую связанность предельных значений, т. е. точки с максимальными значениями переменной не будут иметь связи друг с другом по соседним ячейкам с высокими значениями переменной. Такое поведение не характерно, например для

геологических приложений, где пласты высокой проницаемости образуют связные структуры. Дело в том, что вариограмма характеризует корреляцию, основываясь на связи пар точек, в то время как связная структура, образованная несколькими точками, не сохраняется и может быть разрушена при стохастическом моделировании, что и происходит в случае максимальной энтропии (беспорядка), свойственной гауссову моделированию.

Последовательное гауссово моделирование позволяет получить реализации переменной, принимающие непрерывные значения, например концентрации загрязнения или пористости породы. Размерность пространства, в котором используется метод, также не имеет значения.

Однако существуют категориальные (или разрывные) переменные, которые могут принимать только определенные значения, например типы почв или породы. Для моделирования таких переменных можно использовать другие методы, основанные на последовательном принципе.

8.4. Обрезанное гауссово моделирование

Обрезанное гауссово моделирование является модификацией последовательного гауссова моделирования для разрывных и категориальных переменных. Алгоритм обрезанного гауссова моделирования отличается лишь пред- и постобработкой результатов моделирования.

Локальное нормальное распределение непрерывной переменной, полученной в результате последовательного гауссова моделирования, можно разбить по категориям на основе выбранных пороговых значений (рис. 8.7). Так определяются значения искомой категориальной (или разрывной) переменной. Значение стохастической реализации получается при попадании случайно выбранного значения из локального нормального распределения в тот или иной интервал.

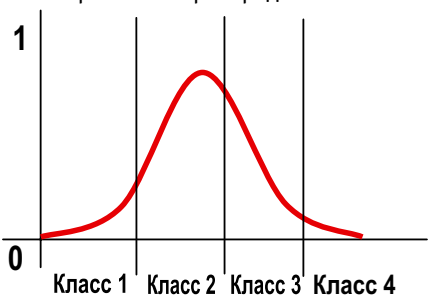


Рис. 8.7. Разбиение гауссова распределения на классы категориальной переменной

Обрезанное гауссово моделирование позволяет быстро получить стохастические реализации категориальных переменных, значения которых могут быть соотнесены с пороговыми значениями непрерывной переменной. Например, типы пород (фаций) можно определить по измерениям пористости или гамма-нейтронного каротажа. Соответственно для моделирования достаточно иметь единственную вариограммную модель нормализованных значений, что облегчает процесс подгонки вариограммы. Однако это же может быть причиной неточности моделирования, связанной с тем, что различные категории могут иметь разную пространственную корреляцию. В этом случае нужно моделировать вариограмму для каждой категории значений отдельно (индикаторный подход). Одной из особенностей обрезанного гауссова моделирования является сохранение последовательности промоделированных категорий. Для фиксированной последовательности пороговых значений категории всегда будут располагаться в той же последовательности, что и соответствующие им пороговые значения. Это важно, когда последовательность категорий имеет под собой физический смысл, как, например, последовательность пластов пород в геологии.

8.5. Последовательное индикаторное моделирование

Стохастическое индикаторное моделирование также базируется на последовательном принципе, но в отличие от гауссова моделирования не предполагает существования определенной аналитической формы локального распределения. Вместо этого локальная функция распределения плотности вероятности оценивается при помощи индикаторного кригинга, который был подробно описан в Главе 7.

Индикаторный подход заключается в моделировании бинарных индикаторных переменных, которые принимают значения либо 1, либо 0 в зависимости от присутствия или отсутствия свойства в данной точке. Индикаторный подход может быть использован для моделирования как категориальных, так и непрерывных переменных. Для получения значений бинарных индикаторных переменных проводится индикаторное преобразование исходных данных для выбранного набора срезов в случае непрерывной функции (7.1) В случае категориальной переменной значения индикаторных переменных (1 и 0) соответствуют присутствию или отсутствию каждой из категорий в

точке измерения — см. (7.2). Индикаторные преобразования, выбор числа и значения срезов обсуждались в Разделе 7.1.

Локальная функция распределения в случае индикаторного моделирования строится на основе вероятностей, полученных индикаторным кригингом для каждой индикаторной переменной. В результате получается вероятность значений категориальной переменной либо вероятность превышения порогового значения (что аналогично категории) в точке оценивания. Оценка индикаторного кригинга в этом случае выглядит так:

$$\Pr^{*}\left\{I(x;z_k)=1\,|\,(n)\right\}=p_k+\sum_{i=1}^{n}\lambda_i\left[I(x_i;z_k)-p_k\right],$$

где $p_k=E\left\{I\left(x;sk\right)\right\}\in\left[0,1\right]$ определяет долю категории z_k в глобальном распределении и находится из исходных данных с учетом декластеризации. Веса λ_i определяются индикаторным кригингом с использованием модели ковариации для соответствующих категорий z_k (см. Главу 7). Если средние значения доли категории варьируются по области, то можно использовать простой индикаторный кригинг с гладко меняющимся локальным средним значением.

В случае непрерывной переменной для подробного оценивания локальной функции плотности вероятности может понадобиться достаточно большое количество пороговых значений. В случае категориальной переменной число индикаторных переменных соответствует числу категорий.

Алгоритм последовательного индикаторного моделирования заключается в следующих этапах [Goovaerts, 1997] (рис. 8.8).

1. Индикаторное преобразование исходных данных по заданному набору порогов отсечений (или дискретному набору категорий) и моделирование пространственной корреляционной структуры для каждой индикаторной переменной.
2. Выбор случайной последовательности через все точки оценивания.
3. В каждой точке оценивания моделируется стохастическая реализация по следующей последовательности операций (см. рис. 8.7):
   - оценка K вероятностей $P_k(x|(z_k))$ $k=1,\ldots,K$ при помощи индикаторного кригинга в выбранной точке x последовательности;
   - построение локальной условной функции плотности вероятности на основе K вероятностей P_k (коррекция, интерполяция, экстраполяция, как описано в Главе 7);
   - выборка случайного значения по построенной локальной функции распределения плотности вероятности (или по набору вероятностей

для категорий), которое определяет смоделированное значение переменной в точке для данной реализации;

- добавление сгенерированного случайного значения к набору данных и других сгенерированных значений для использования в последующих оценках кригинга.

4. Переход к следующей точке оценивания и выбранной последовательности и повторение шагов 2 и 3.

Шаги 2—4 повторяются для получения нескольких равновероятных реализаций в точках оценивания.

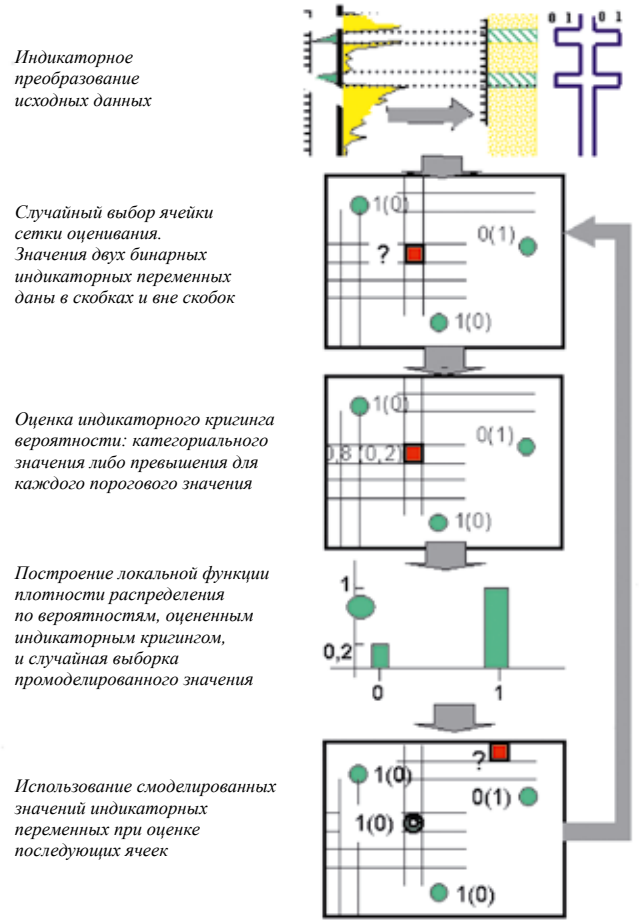


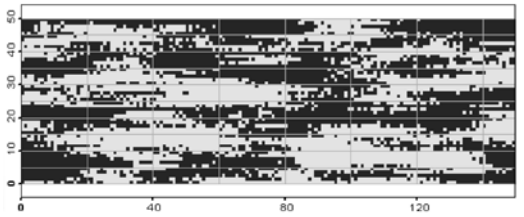
Рис. 8.8. Схема алгоритма последовательного индикаторного моделирования

Ни рис. 8.8 изображена пошаговая схема индикаторного моделирования на примере двух категориальных переменных, которым соответствуют индикаторные переменные. Значения переменных в точках даны вне скобок и в скобках.

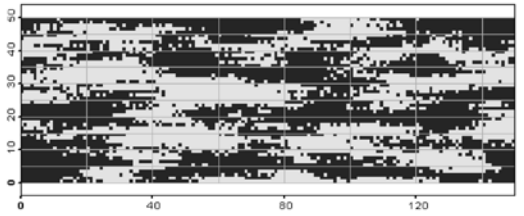
При построении локальной функции плотности вероятности непрерывной переменной важно соблюдать последовательность суммирования составляющих вероятностей индикаторных переменных, которые с увеличением порогового значения образуют кумулятивную функцию распределения вероятности. В случае категориальной переменной соблюдать последовательность не обязательно, поскольку нет последовательности значений отсечений.

Индикаторное моделирование гарантирует приблизительное воспроизводство средней доли каждой категории, исходя из заданного глобального распределения и вариограммы, соответствующей данной категории. Таким образом, аппроксимация статистических моментов первого (среднее) и второго (вариация и вариограмма) порядков зависит от следующих факторов: количества порогов отсечений, условной информации, учитывающейся в индикаторном кригинге (долей категорий, данных), функций интер- и экстраполяции, использующихся для аппроксимации между пороговыми значениями и определения хвостовых значений распределения.

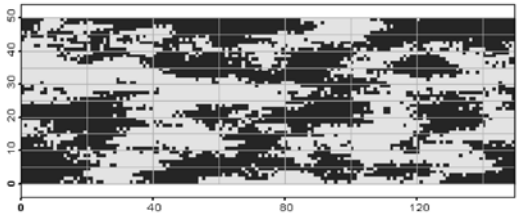
Приведем примеры реализаций последовательного индикаторного моделирования для различных радиусов корреляций для двух индикаторных переменных, соответствующих двум типам пород в задачах моделирования проницаемости пористой среды. На рис. 8.9 приведены реализации для различных значений горизонтального радиуса корреляции в моделях индикаторных вариограмм. На рис. 8.10 аналогичным образом варьируется радиус корреляции вариограммной модели по вертикали. Видно, что при больших значениях радиуса корреляции возникают протяженные связанные структуры индикаторной переменной, и в этом случае можно предполагать хорошее протекание при высокой проницаемости породы. Связанность структур с высокой проницаемостью определяет потоки жидкости и газа в задачах по моделированию добычи углеводородов.



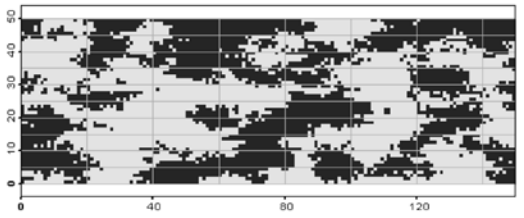
а



б



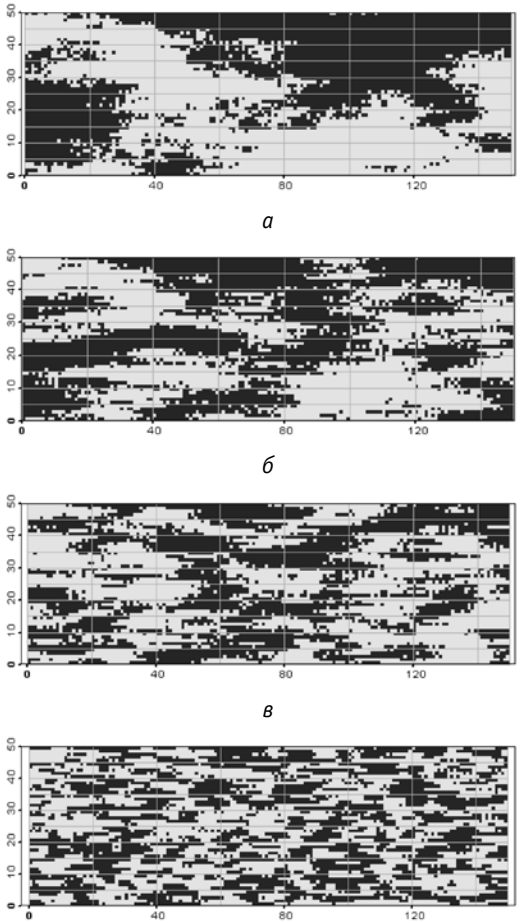
в



г

Рис. 8.9. Реализации последовательного индикаторного моделирования для различных горизонтальных радиусов корреляции:

а — $r = 160$; *б* — $r = 80$; *в* — $r = 40$; *г* — $r = 20$
(вертикальный радиус корреляции $R = 8$)



а

б

в

г

Рис. 8.10. Реализации последовательного индикаторного моделирования для различных вертикальных радиусов корреляции:

а — $r = 20$; *б* — $r = 10$; *в* — $r = 5$; *г* — $r = 2$
(горизонтальный радиус корреляции $R = 80$)

Индикаторное моделирование отличается следующими особенностями.

1. Использование индикаторных вариограмм для каждой категории позволяет учесть индивидуальную корреляционную структуру каждого класса значений (различие в анизотропии, корреляционном расстоянии и т. д.), которая может не проявляться при построении глобальной вариограммы для всего интервала значений переменной.

2. Индикаторная вариограмма более стабильна и устойчива к крайним значениям, чем вариограмма исходных данных, поскольку индикаторное преобразование позволяет избавиться от влияния крайних предельных значений путем нелинейного преобразования.
3. Построение моделей вариограмм для каждой категории достаточно трудоемко по сравнению с подготовкой к обрезанному гауссову моделированию, для которого требуется единственная модель вариограммы для нормализованных значений.
4. Реализации индикаторного моделирования (как и гауссова) воспроизводят статистические моменты (среднее, вариацию и вариограмму) исходного распределения. Причем вариограмма обычно воспроизводится с большей точностью (меньшей вариабельностью), поскольку моделирование пространственной корреляции для набора индикаторных переменных (соответствующих порогам отсечений) точнее, чем для единственной общей переменной.
5. В полученных пространственных реализациях категориальной переменной может не сохраняться последовательность категорий, т. е. соседние точки могут чередовать категории в любой последовательности в отличие от результатов обрезанного гауссова моделирования, где последовательность категорий фиксирована. Это свойство может привести к нереалистичности результатов моделирования, например геологических пластов, где известна последовательность слоев, исходя из физической модели.

В случае моделирования категориальных переменных бинарные пространственные реализации для различных индикаторных переменных можно объединить в единую реализацию, которая будет отображать совместное стохастическое распределение всех категорий. Следующим шагом при моделировании геологических пород является моделирование свойств пористости и проницаемости для каждого типа породы, которое проводится отдельно для области распределения каждого типа породы в соответствии с их пространственной реализацией.

При моделировании непрерывной переменной стохастические индикаторные реализации являются непосредственными значениями выборки из локальных распределений вероятности.

На основе набора стохастических реализаций можно вычислить среднюю оценку (E-type) и разброс локальных значений функции.

8.6. Последовательное прямое моделирование

При прямом моделировании отсутствует предварительное преобразование исходной переменной с последующим поиском функции распределения в преобразованном пространстве, что характерно для двух описанных выше методов. Первый шаг в этом направлении для моделирования непрерывной переменной был сделан в [Journel, 1994], где было показано, что реализации, полученные при использовании в качестве параметров локального распределения оценки и вариации простого кригинга для непрерывной переменной без предварительного преобразования, воспроизводят пространственную корреляцию исходных данных. Но эти реализации не могут воспроизводить гистограммы исходных данных, что считается важным при стохастическом моделировании. В [Soares, 2001] предложен алгоритм, позволяющий воспроизводить различные (даже сложные) гистограммы исходных данных.

Итак, мы рассматриваем непрерывную функцию $Z(x)$ с глобальной условной функцией распределения $F_Z(z) = \Pr\{Z(x) < z\}$ и стационарной вариограммой $\gamma(h)$, которые мы заинтересованы воспроизводить в наших реализациях. При этом для воспроизведения вариограммы нам достаточно использовать локальные условные кумулятивные функции распределения плотности вероятности, центрированные в оценке простого кригинга:

$$z^*(x_u) = m + \sum_i w_i(x_u)\left[z(x_i) - m\right],$$

где x_i — местоположения данных (исходных и уже смоделированных) с условными вариациями, полученными из простого кригинга $\sigma^2_{SK}(x_u)$. При этом неважно, какая именно функция распределения используется.

Основная идея предложенного алгоритма состоит в том, чтобы использовать локальные среднее и вариацию не для оценки локальной условной функции распределения, а для того, чтобы разыгрывать значение в соответствии с глобальной функцией распределения исходных данных. При этом глобальная гистограмма (постоянная для всех шагов) представляется набором классов, а локальные данные определяют, какой класс выбрать для розыгрыша значения. Например, задавать классы можно, используя часть исходных данных, среднее и вариация которых соответствуют локальной оценке и вариации простого кригинга. Разыгранное значение выбирается в соответствии с функцией распределения этих данных. Такой способ требует каждый раз строить функцию распределения некоторого поднабора данных.

Более удобный подход состоит в использовании разбиения на классы, аналогичного обрезанному гауссову распределению [Deutsch, 2002]. Преобразуем исходные данные $z(x)$ в нормальное распределение при помощи функции φ:

$$y(x) = \varphi\big(z(x)\big),$$

где $G\big(y(x)\big) = F_z\big(z(x)\big)$.

Локальная оценка простого кригинга $z^*(x_u)$ имеет эквивалентное гауссово значение $y^*(x_u) = \varphi(z^*(x_u))$, которое совместно со стандартизованной вариацией простого кригинга $\sigma^2_{SK}(x_u)$ позволяет определить гауссову функцию распределения $G\big(y^*(x_u), \sigma^2_{SK}(x_u)\big)$.

Эта гауссова условная функция распределения позволяет определить интервал условной функции распределения $z(x)$, в котором нужно разыгрывать новое значение:

- сгенерировать значение p из распределения $U(1, 0)$;
- сгенерировать значение y^s из распределения $G\big(y^*(x_u), \sigma^2_{SK}(x_u)\big)$:

$$y^s = G^{-1}\big(y*(x_u), \sigma^2_{SK}(x_u), p\big).$$

Разыгрываемое значение получается обратным преобразованием:

$$z^s(x_u) = \varphi^{-1}(y^s).$$

Эта схема включается в стандартную последовательность действий, характерную для последовательного моделирования (см. рис. 8.4) на шагах определения локальной функции распределения и розыгрыша значения.

Главными достоинствами прямого моделирования являются отсутствие предварительного преобразования данных и способность качественно воспроизводить глобальную функцию распределения исходных данных (гистограмму).

Примеры, представленные на рисунках 8.11, 8.12, иллюстрируют это качество метода. Рассмотрены исходные данные с двумя типами глобальной функции распределения: характеризующейся двумя пиками (см. рис. 8.11) и примерно однородным (см. рис. 8.12). Реализации, различающиеся между собой, достаточно четко воспроизводят исходные гистограммы (рис. 8.13 — 2 пика, рис. 8.14 — однородное распределение).

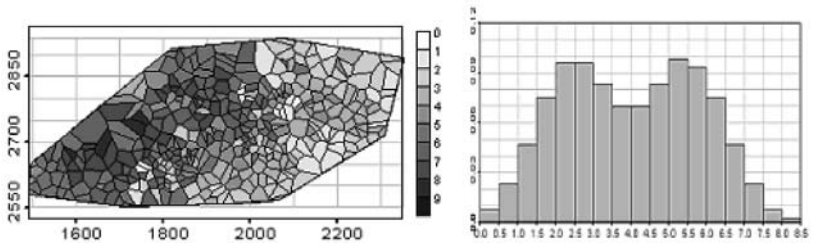


Рис. 8.11. Исходные данные и их глобальная функция распределения

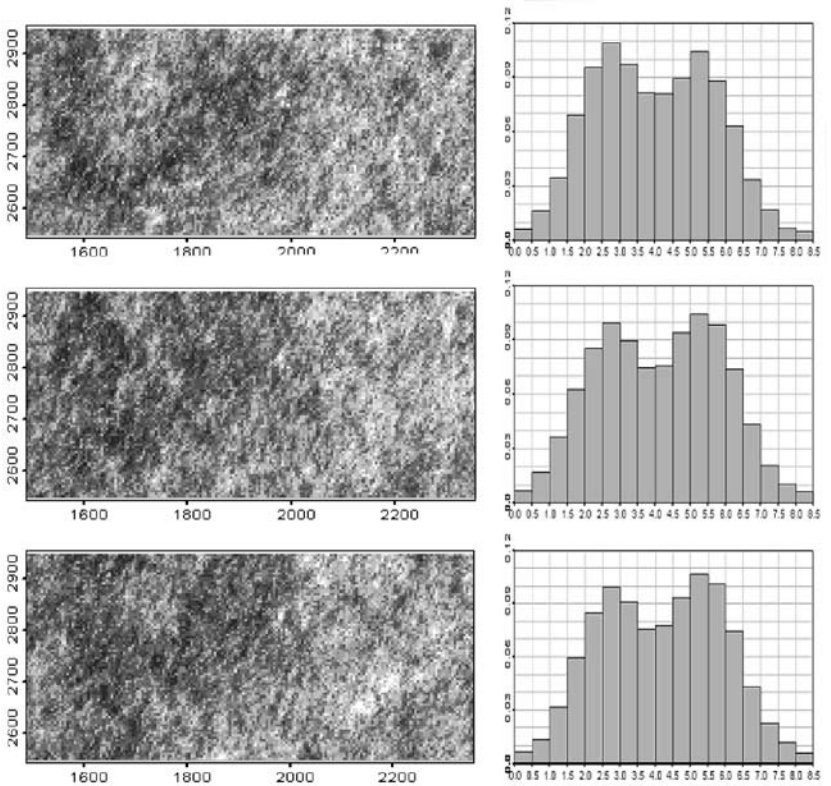


Рис. 8.12. Примеры реализации прямого условного моделирования и их глобальные функции распределения

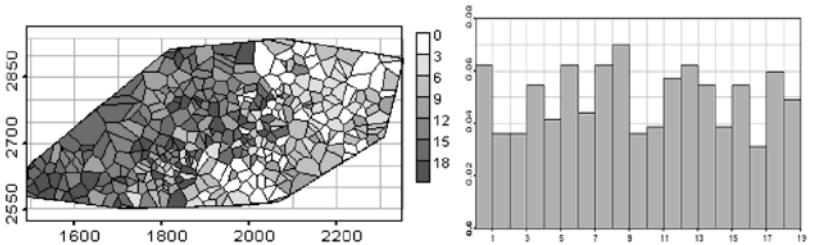


Рис. 8.13. Исходные данные и их глобальная функция распределения

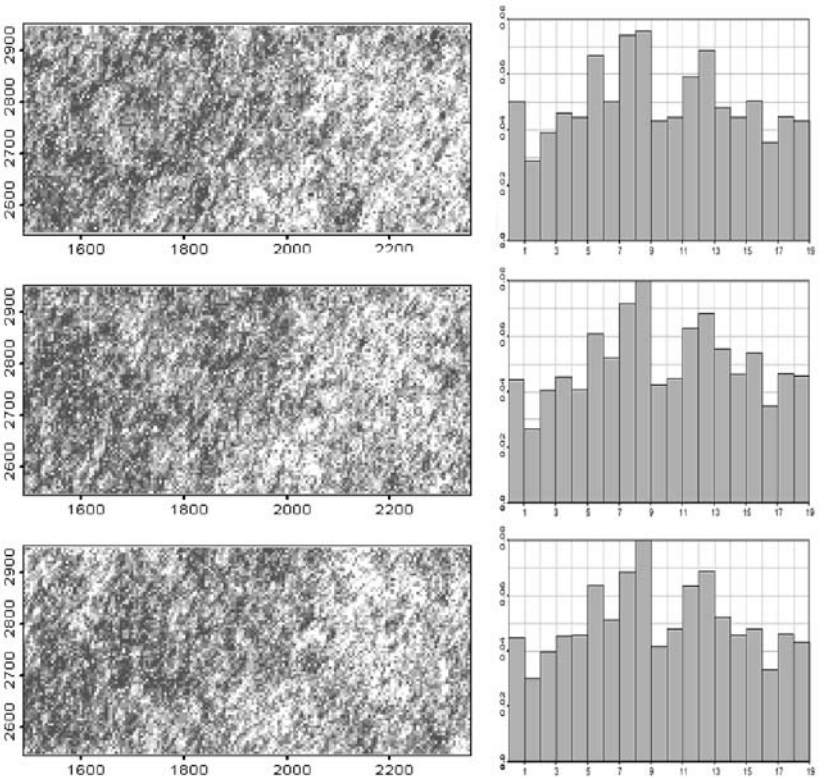


Рис. 8.14. Примеры реализации прямого условного моделирования и их глобальные функции распределения

8.7. Моделирование отжига

Моделирование отжига (simulated annealing) [Metropolis et al., 1985] — это общее название для семейства оптимизационных алгоритмов, основанных на принципах стохастической релаксации. Моделирования отжига аналогично процессу остывания металла в термодинамике и основано на соотношении Больцмана между температурой и энергией [Aarts, Korst, 1989].

Термин «отжиг» (annealing) пришел в математику из металлургии. При высокой температуре металл легко деформируется и меняет форму. Чем выше температура, тем больше скорость колебаний атомов и тем легче металл поддается деформации. Если резко снизить температуру, то атомы «замрут» и мы получим твердый, но очень хрупкий металл. Если стоит задача получить максимально организованную структуру, необходимо сначала сильно разогреть металл, а затем очень медленно охлаждать его. Таким образом, металл будет проходить через множество квазиравновесных состояний и у каждого его атома будет достаточно времени, чтобы найти «лучшее» место среди других атомов в смысле минимума полной энергии системы, что соответствует идеальной кристаллической решетке. С точки зрения математики задачу можно сформулировать так: минимизировать среднюю квадратичную ошибку отклонения уровня поверхности металла от некоторого постоянного значения. Оптимизируемыми параметрами в этом случае будут координаты положения атомов.

При использовании моделирования отжига как метода стохастической минимизации заданный образ постепенно возмущается так, чтобы подогнать его под воспроизводство каких-либо целевых структур (гистограмма, вариограмма, ковариация и т. п.), оставляя исходные данные неизменными.

Рассмотрим моделирование непрерывной величины z в N узлах сетки x_j при заданных условиях $z(x_\alpha)$, $\alpha = 1, \ldots, n$ таким образом, чтобы вариограмма данных воспроизводилась для первых S лагов. Аннилинг требует задания целевой функции (являющейся аналогом энергии), которая измеряет разницу между значениями целевых и текущих статистических параметров на каждом i-м возмущении. Если цель — воспроизвести вариограммную модель, то целевая функция может выглядеть так:

$$O(i) = \sum_{s=1}^{S} \left[\gamma(h_s) - \gamma_{(i)}(h_s) \right]^2,$$

где $\gamma(h_s)$ — значение требуемой вариограммной модели для лага h_s и $\gamma_{(i}(h_s)$ — соответствующее значение вариограммы реализации на i-м возмущении.

Если целевая функция установлена, то процесс моделирования (точнее, оптимизации) включает в себя систематическое модифицирование начальной реализации так, чтобы уменьшить значение целевой функции, делая реализацию приемлемо близкой к целевой статистике.

Общий алгоритм моделирования отжига состоит из следующих этапов.

1. Создаем начальный образ $\left\{z_{(0)}\left(x_j\right), j=1,...,N\right\}$, который сохраняет исходные данные и может сразу аппроксимировать какую-нибудь целевую статистику (дисперсию распределения, плато вариограммы или гистограмму).
2. Считаем начальное значение целевой функции, соответствующее этой начальной реализации.
3. Возмущаем реализацию каким-либо механизмом, например отражением пар z-значений: $z_{(0)}(x_j)$ становится $z_{(0)}(x_i)$, и наоборот. По аналогии с физическими процессами в остывающем металле механизм возмущения зависит от температуры: чем она ниже, т. е. чем больше шагов сделано, тем меньше меняется значение в точке при возмущении. Например, в случае отражения пар уменьшается расстояние между точками, значения которых мы меняем местами.
4. Оцениваем эффект возмущения на воспроизведение целевой функции, снова вычисляя ее значение O_{new}, учитывая модификацию начальной реализации.
5. Принимаем или не принимаем возмущение на основе какого-либо правила. Обычно вероятность принятия задается с помощью распределения Больцмана [Aarts, Korst, 1989]:

$$\Pr\{\text{принять}\} = \begin{cases} 1, \text{ если } O_{\text{new}} \le O_{\text{old}}, \\ \exp\left(\dfrac{O_{\text{new}} - O_{\text{old}}}{T}\right) \text{ в других случаях,} \end{cases}$$

   где T — температура. Чем выше температура, тем больше вероятность принять неблагоприятное (т. е. не уменьшающее целевую функцию) возмущение. Делается это для того, чтобы была возможность избежать локальных минимумов и найти глобальный.
6. Если возмущение принимается, заменяем начальную реализацию на новую $\left\{z^l_{(1)}\left(x_j\right), j=1,...,N\right\}$ с соответствующей целевой функцией $O_{\text{old}} = O_{\text{new}}$.

7. Повторяем шаги с 3-го по 6-й, пока целевая структура не будет приемлемо достигнута или пока возмущения не перестанут уменьшать целевую функцию. Потом снижаем температуру и проделываем указанную процедуру (шаги 3—6) для новой температуры и так поступаем до тех пор, пока не достигнем приемлемого результата (см. ниже критерий остановки).

Последующие равновероятные реализации $\left\{z^{(l')}\left(x_j\right), j=1,...,N\right\}, l' \neq l$ производятся повторением шагов 1—8, начиная с другого начального образа.

Обычно число узлов N так велико, а вариограмма накладывает так мало связей, что существует очень много решений оптимизационной задачи. Конечная реализация выбирается из этого набора приблизительных решений.

Существует много способов осуществления алгоритма моделирования отжига. Варианты отличаются тем, как создавать начальный образ, как его возмущать, компонентами целевой функции и типом критерия, принимать или не принимать эффект после возмущения.

Требования к начальному образу.

- Он должен быть легко производим.
- Значения точек начального образа должны сразу подходить какой-нибудь части целевой структуры (например, воспроизводить гистограмму данных), чтобы ускорить последующий процесс оптимизации.
- Все начальные образы должны быть «равновероятны». Нужно остерегаться использовать один начальный образ в качестве стартовой точки для нескольких различных реализаций, потому что даже разные дальнейшие пути могут привести к слишком похожим конечным реализациям и вызвать, таким образом, недооценку неопределенности.

Обычно начальный образ производится так: исходные данные «замораживаются» на своих местах, а приписываемое каждому узлу значение z-величины выбирается случайным образом в соответствии с глобальной функцией распределения $F(z)$. Такой подход достаточно быстр и дает набор начальных образов, уже удовлетворяющих целевой гистограмме.

Начальный образ также может быть результатом применения какого-либо другого алгоритма, например кригинга или реализации последовательного моделирования. В таком случае аннилинг выступает в качестве постпроцессора. Его цель состоит в улучшении воспроизведения целевой статистики или наложении дополнительной структуры, которая не может быть введена другими алгоритмами. Например, требование присутствия канальных структур в пространственной модели проницаемости.

Чаще всего используются два механизма возмущения.

Первый — отражение z-величин для случайным образом выбранных пар точек x_j и x_k, находящихся на расстоянии $d \le D(T)$, $D(T)$, уменьшается понижением T:

$$\begin{cases} z_{(i)}(x_i) = z_{(i-1)}(x_k), \\ z_{(i)}(x_k) = z_{(i-1)}(x_i). \end{cases}$$

Такой механизм возмущения позволяет сохранить гистограмму начального образа. Значит, нет необходимости включать воспроизведение гистограммы в целевую функцию, если начальный образ уже удовлетворяет ей.

Другой механизм называют возмущением. Случайным образом выбирается одна точка x_j, и модифицируется соответствующее значение $z_{(i-1)}(x_j)$ согласно какому-нибудь механизму, например

$$z_{(i)}(x_j) = F^{-1}(p_j),$$

где p_j — случайное значение из {0, 1}; а $F(z)$ — целевая гистограмма.

В отличие от механизма отражения, сохраняющего начальную гистограмму, здесь требуется включать соответствующую гистограмме компоненту в целевую функцию.

В обоих случаях условные (т. е. исходные) данные никогда не возмущаются, чтобы конечная реализация сохранила эти значения.

Моделирование отжига позволяет принимать в расчет различные типы информации, вводя ее количественные характеристики в глобальную целевую функцию. Эта функция представляет собой взвешенную сумму из C компонент O_c, измеряющих разницу между статистикой текущей реализации (на i-м возмущении) и целевой статистикой:

$$O(i) = \sum_{c=1}^{C} \omega_c O_c(i),$$

где веса ω_c контролируют относительную важность c-й компоненты целевой функции. Приведем примеры наиболее часто используемых в рамках геостатистического моделирования компонент целевой функции.

Кумулятивная функция распределения (учет гистограммы данных).

Типичная целевая статистика — это по возможности декластеризованная однопеременная функция распределения z-данных $F(z)$. Если диапазон из-

менения z описывается серией из K порогов z_k, то разница между целевой и текущей функцией распределения может быть измерена так:

$$O_c(i) = \sum_{k=1}^{K} \left[F(z_k) - F'_{(i)}(z_k) \right]^2,$$

где $F'_{(i)}(z_k)$ — значение на пороге z_k, вычисленное для реализации на i-м возмущении.

Модель полувариограммы. Воспроизведение вариограмной модели $\gamma(h)$ обычно ограничивается определенным количеством S лагов. Разница между целью и текущим значением в этом случае измеряется так:

$$O_c(i) = \sum_{s=1}^{S} \frac{\left[\gamma(h_s) - \widehat{\gamma}_{(i)}(h_s) \right]^2}{\left[\gamma(h_s) \right]^2},$$

где $\widehat{\gamma}_{(i)}(h_s)$ — значение полувариограммы реализации для лага h_s на i-м возмущении.

Модели индикаторных полувариограмм. Моделирование отжига позволяет учитывать специальные пространственные структуры, моделируемые индикаторными полувариограммами $\gamma_I(h, z_k)$, посчитанными для K различных порогов z_k. Так получаем еще одну компоненту целевой функции:

$$O_c(i) = \sum_{k=1}^{K} \sum_{s=1}^{S} \frac{\left[\gamma(h_s; z_k) - \widehat{\gamma}_I^{(i)}(h_s; z_k) \right]^2}{\left[\gamma(h_s; z_k) \right]^2},$$

где $\widehat{\gamma}_I^{(i)}(h_s; z_k)$ — значения индикаторной полувариограммы на лаге h_s по порогу z_k для реализации на i-м возмущении.

Коэффициент корреляции. Пусть Y — лучшая выборка или ранее промоделированная величина, вторичная по отношению к величине Z. Если взаимосвязь Z-Y корректно описывается линейным коэффициентом корреляции $\rho_{ZY}(0)$, то его можно ввести в целевую функцию в виде компоненты

$$O_c(i) = \left[\rho_{ZY}(0) - \widehat{\rho}_{ZY}^{(i)}(0) \right]^2,$$

где $\widehat{\rho}_{ZY}^{(i)}(0)$ — коэффициент корреляции, вычисленный по соответствующим парам y-данных и z-значений на i-м возмущении.

Кросс-вариограмма. Пространственная кросс-корреляция между величинами Z и Y, моделируемая с помощью кросс-вариограммы $\gamma_{ZY}(\boldsymbol{h})$, может быть воспроизведена включением в целевую функцию компоненты типа

$$O_c(i) = \sum_{s=1}^{S} \frac{\left[\gamma_{ZY}(\boldsymbol{h}_s) - \widehat{\gamma}_{ZY}^{(1)}(\boldsymbol{h}_s)\right]^2}{\left[\gamma_{ZY}(\boldsymbol{h}_s)\right]^2},$$

где $\widehat{\gamma}_{ZY}^{(1)}(\boldsymbol{h}_s)$ — значение кросс-вариограммы между z-значениями и y-данными на i-м лаге.

Как и для любого итерационного алгоритма, для этого оптимизационного процесса должен быть определен *критерий остановки*. Возможные критерии таковы:

- целевая функция достигла достаточно малого значения $O_{\min}$;
- количество возмущений при одной температуре превысило допустимое максимальное число;
- доля приемлемых возмущений меньше, чем заданное пороговое значение.

Примеры реализаций, полученных с использованием моделирования отжига, приведены на рис. 8.15. Это реализации пространственного распределения краба Берди в 2006 г. в Беринговом море. Использование моделирования отжига в данном случае обусловлено необходимостью использования индикаторных вариограмм как более робастных (это показало в Главе 7). Но, с другой стороны, небольшое количество ненулевых данных не дает возможность использовать напрямую индикаторный подход (не удается сделать достаточное количество срезов).

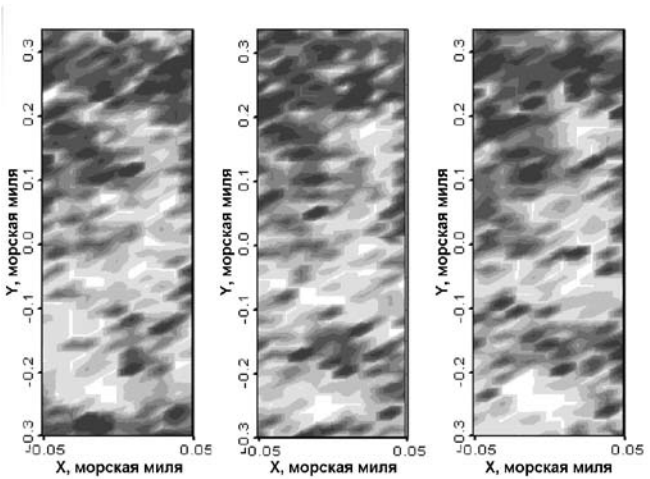


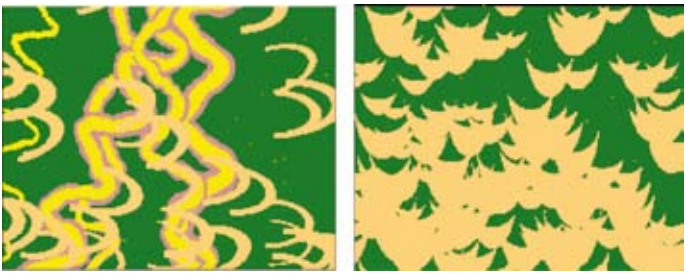
Рис. 8.15. Примеры реализаций с использованием моделирования отжига для данных траловой съемки по пространственному распределению краба Берди в Беринговом море

8.8. Объектное моделирование

Объектное моделирование является альтернативой пиксельному моделированию. В отличие от многих геостатистических моделей оно не основано на двухточечной статистике (вариографии). Однако объектный подход позволяет промоделировать пространственную корреляцию без помощи вариограммы на основе набора пространственных структур, которые имеют определенную заданную форму, — объектов. Для распределения объектов используются булево моделирование и алгоритмы оптимизации. Как правило, объектное моделирование применяется для категориальных переменных. При этом распределение рассматриваемой категории моделируется как совокупность геометрических объектов, которые покрывают области, в которых преобладают значения данной категории.

Примеры объектного моделирования

Для проведения объектного моделирования требуется определить набор форм объектов. Обычно при этом руководствуются экспертным анализом на основе физических представлений об исследуемой системе. Так, в геологических приложениях формы объектов могут быть получены в результате анализа выходов пород, разрезов, шурфов, информации из буровых скважин, а также геологических представлений. Примеры геологических объектов — флювиальные синусоидные и меандрированные русла, песочные линзы, эоловые дюны нетривиальной трехмерной геометрии, диски или эллипсоиды, сланцевые останцы, конусы прорыва, лагуны и т. п. (рис. 8.16).



а *б*

Рис. 8.16. Примеры объектного моделирования:
а — система речных русел с двумя типами объектов: связанные кривые русла с прирусловым валом и дугообразные формы старых русел;
б — эоловая система, характерная для дюнных отложений

Размеры объектов выбранных форм могут варьироваться для моделирования природного разнообразия и более качественной подгонки под имеющиеся данные.

Объекты могут размещаться в пространстве при помощи различных алгоритмов [Deutsch, 2002].

Наиболее простым методом является случайное размещение объектов. В этом случае можно просто и быстро достичь необходимой доли форм в области моделирования, однако этот метод не гарантирует воспроизведения данных измерений.

Для воспроизведения данных измерений необходимо переместить объекты либо изменить их размеры, чтобы они удовлетворяли исходным данным. При этом следует иметь в виду, что если точки измерения выбирались с предпочтением к определенным значениям переменных, то доля рассматриваемой категории в распределении исходных данных измерений может быть завышена (рис. 8.17)

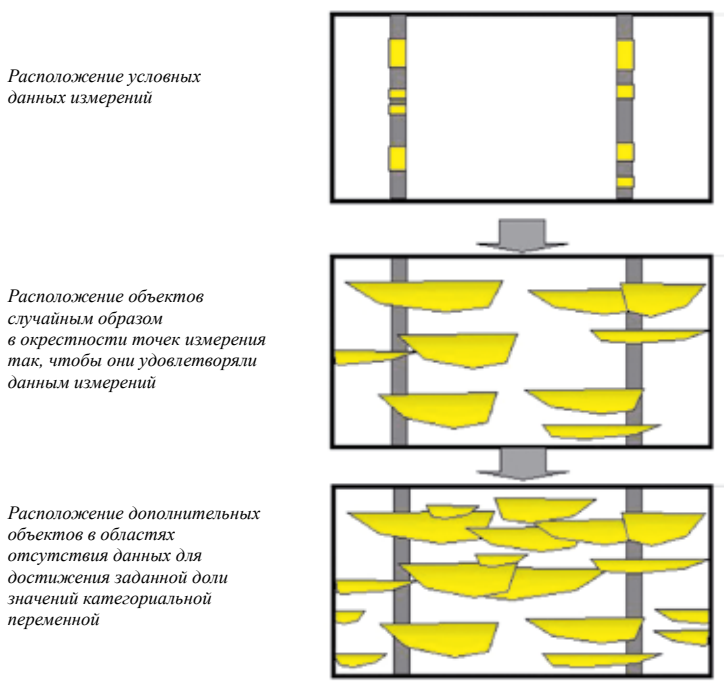


Рис. 8.17. Алгоритм объектного булева моделирования

Оптимальным и наиболее трудоемким методом является итерирование перемещений и изменение размеров объектов для наилучшего удовлетворения данных измерений и заданной доли категории. Для этого осуществляют минимизацию целевой функции при помощи стандартных алгоритмов (градиентных и стохастических), используя случайные возмущения в размещении объектов.

Объектное моделирование отличает хорошая связь с физическим смыслом моделируемых объектов (например, геологических). Однако при этом требуются трудоемкий и корректный экспертный анализ и хорошее понимание исследуемой системы, чтобы выбрать набор форм объектов. При отсутствии достаточной информации о формах исследуемых объектов результаты моделирования могут оказаться далекими от действительности.

При использовании слишком большого разнообразия форм и степеней свободы объектов объектная модель может оказаться слишком громоздкой. В этом случае возможны проблемы со сходимостью алгоритма оптимизации многопараметрической модели и оптимального результата добиться не удастся.

Еще одним недостатком объектного подхода является сложность включения в него дополнительных вероятностных «мягких» данных [Caers, 2005]. «Мягкие» (soft) данные обычно имеют вид вероятности значения переменной, которая определяется на регулярной сетке на основе данных более низкого разрешения, чем сетка оценивания. Эти данные коррелированны с моделируемой переменной, но требуют калибровки и менее точны (например, сейсмическое зондирование, аэрогаммасъемка). В случае пиксельных моделей такие данные приводятся на сетку оценивания более высокого разрешения с учетом изменения вероятности. При моделировании объектами, имеющими достаточно низкое разрешение, использование таких «мягких» данных напрямую затруднительно, однако они могут быть использованы при выборе формы и размеров объектов.

Упражнение 8.1. *Почему оценку кригинга называют сглаженной, чем от нее отличается реализация стохастического моделирования?*

Упражнение 8.2. *При оценивании часто точно неизвестен верхний предел оцениваемой переменной. Чему равна максимальная/минимальная оценка кригинга? Что больше — максимальная оценка кригинга или максимальное значение стохастической реализации?*

Упражнение 8.3. *Вариограмма характеризует пространственную корреляцию и уровень пространственной вариабельности. На рис. 8.18 приведены две вариограммы, построенные для оценки кригинга и стохастической реализации на основе одних и тех же данных и модели вариогрммы. Какая из вариограмм соответствует оценке кригинга, а какая — реализации стохастического моделирования? На чем основан выбор?*

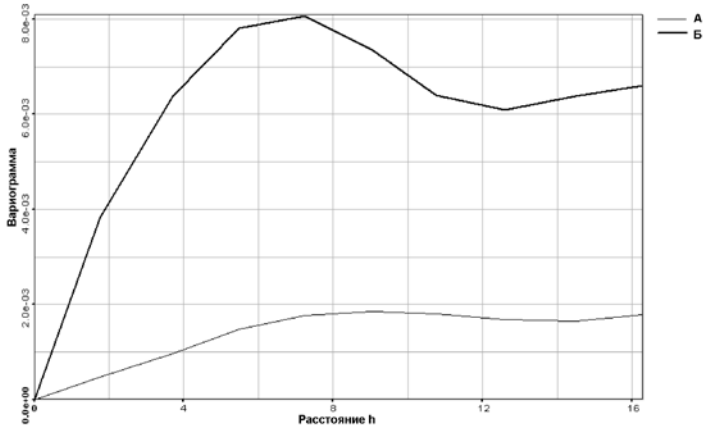


Рис. 8.18. Вариограммы для оценки кригинга и стохастической реализации

Упражнение 8.4. *В последовательном гауссовом моделировании используется модель вариограммы. Какие типы моделей вариограмм могут быть использованы? Почему суммарное плато модели вариограмм для гауссова моделирования должно быть равно единице?*

Литература

Каневский М. Ф., Демьянов В. В., Савельева Е. А. и др. Элементарное введение в геостатистику. — М., 1999. — (Пробемы окружающей среды и природных ресурсов / ВИНИТИ; № 11).

Aarts E., Korst J. Simulated Annealing and Boltzmann Machines. — New York: John Wiley & Sons, 1989.

Caers J. Petroleum Geostatistics / Society of Petroleum Engineers. — Richardson, TX, 2005.

Chiles J.-P., Delfiner P. Geostatistics: Modeling Spatial Uncertainty. — New York: John Wiley & Sons, 1999.

Christakos G. Random Field Models in Earth Sciences. — San Diego, CA: Academic Press, 1992.

Deutsch C. Geostatistical Reservoir Modelling. — [S. l.]: Oxford Univ. Press, 2002.

Deutsch C., Journel A. G. GSLIB: Geostatistical Software Library and User's Guide. — [S. l.]: Oxford Univ. Press, 1998.

Emerly X. Variogram of order ω: A tool to validate a bivariate distribution model // Mathematical Geology. — 2005. — Vol. 37, N 2. — P. 163—181.

Goovaerts P. Geostatistics for Natural Resources Evaluation. — [S. l.]: Oxford Univ. Press, 1997. — 483 p.

Isaaks E. H., Srivastava R. M. An Introduction to Applied Geostatistics. — Oxford, Oxford Univ. Press, 1989.

Journel A. G. Modeling uncertainty: some conceptual thoughts // Geostatistics for the Next Century / Ed. R. Dimitrakopoulos. — Dordrecht: Kluwer Academic Pub., 1994. — P. 30—43.

Journel A. G., Huijbregts C. J. Mining Geostatistics. — London: Academic Press, 1978. — 600 p.

Mantoglou A., Wilson J. Simulation of random fields with the turning band method / Department of Civil Engineering, M.I.T. — [S. l.], 1981. — (Technical Report N 264).

Metropolis N., Rosenbluth A., Teller A., Teller E. Equations of state calculations by fast computing machines // J. of Chem. Physics. — 1985. — Vol. 21, N 6. — P. 1087—1092.

Perrin O., Iovleff S. Estimation a non-stationary spatial structure using simulated annealing // GeoComputation 99 // http://www.geovista.psu.edu/geocomp/geocomp99/Gc99/028/gc_028.htm.

Soares A. Direct Sequential Simulation and Cosimulation // Mathematical Geology. — 2001. — Vol. 33, N 8. — P. 911—926.

Tran T. Improving variogram reproduction on dense simulation grids // Computers and Geosiences. — 1994. — Vol. 20. — P. 1161—1168.

Глава 9

Последовательный геостатистический анализ данных: примеры исследования

9.1. Использование обычного кригинга для мониторинга радиационного загрязнения в режиме реального времени

В данном примере описаны результаты участия обычного кригинга в международном конкурсе сравнения методов пространственной интерполяции (Spatial Interpolation Comparison — SIC2004), организованном Исследовательским центром в Испре, Италия (Joint Research Centre, Ispra, Italy). Подробное описание данных, условий и результатов конкурса опубликовано в специальном выпуске журнала «Applied GIS» в 2005 г. Одним из условий конкурса было использование метода в автоматическом режиме, т. е. настраиваемые параметры метода должны были быть высланы организаторам за сутки до предоставления исходных данных. Для настройки параметров можно было использовать данные по измерению той же величины на той же сети мониторинга, но в другое время.

В конкурсе использовались данные радиационного мониторинга воздуха в районе действующей АЭС, расположенной в Европе [Dubois, Galmarini, 2005]. Таким образом, можно считать, что этот пример демонстрирует возможность применения обычного кригинга для анализа данных мониторинга радиационной обстановки вокруг радиационно опасного объекта в режиме реального времени (on-line) [Savelieva, 2005].

Реальный мониторинг представлен 1008 датчиками. Для проверки качества работы методов в рамках конкурса были выделены тренировочный и валидационный наборы. Тренировочный набор представлен 200 точками, 808 использовались для валидации. Пространственное распределение тренировочных и валидационных точек приведено на рис. 9.1. Сам объект находился в точке с координатами (0, 0).

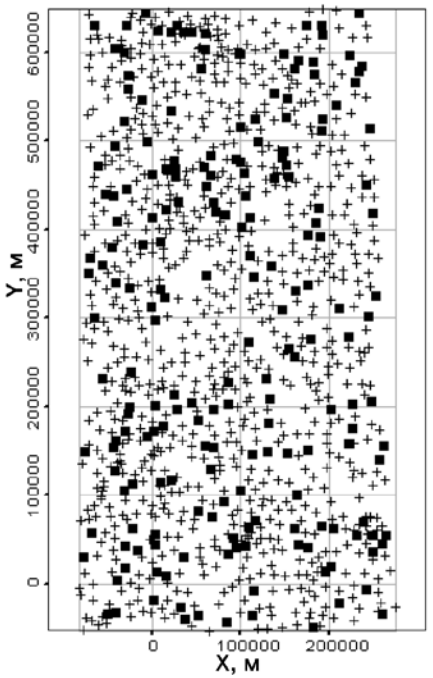


Рис. 9.1. Пространственное распределение тренировочных (■) и валидационных (+) точек

Использование в режиме реального времени предполагает полностью автоматическое функционирование, т. е. такие параметры метода, как модель пространственной корреляционной структуры и область поиска (область оценки), считаются заданными априори. В данном случае для настройки параметров использовалась историческая информация, 10 наборов измерений того же параметра в 200 тренировочных точках.

Для предоставленных 10 наборов был проведен полный статистический анализ. Было обнаружено, что все эти наборы данных являются схожими по таким статистическим характеристикам, как среднее, медиана, минимум, максимум, вариация, диапазон значений и др.

Разница между максимальным и минимальным значениями в каждой точке не превышала 40, больше 30 разница была только в 6 точках. Визуальную схожесть можно наблюдать на рис. 9.2, где визуализированы данные двух наборов (рис. 9.2*а*,*б*) и среднего по 10 наборам (рис. 9.2*в*).

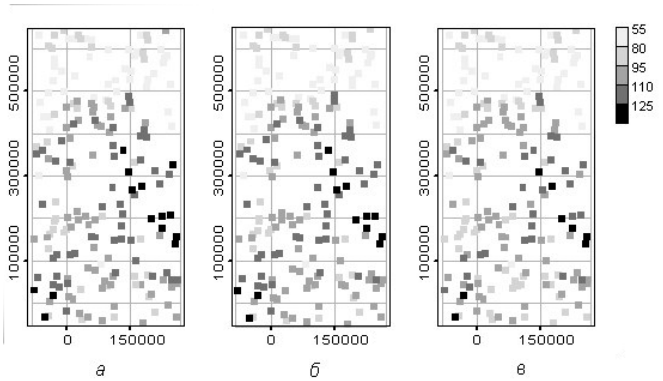


Рис. 9.2. Исторические данные для настройки параметров:
а, б — данные измерений; *в* — среднее по 10 измерениям

Исследование пространственной корреляции также проводилось для всех 10 наборов данных. Результаты этих исследований показали следующее.

- Экспериментальные вариограммы для различных наборов измерений также схожи (пример четырех вариограммных изолиний представлен на рис. 9.3): они не обладают анизотропией до расстояния 60 км и демонстрируют анизотропию (больший радиус корреляции в направлении запад—восток) на расстояниях до 200 км. Значение плато и размер области корреляции одинаковы.
- Вариограмма, усредненная по 10 экспериментальным вариограммам, отражает все свойства отдельных вариограмм (рис. 9.4*а*). Можно предположить, что она будет отражать пространственную корреляционную структуру и других наборов измерений этой переменной.

По усредненной вариограмме построена модель, которая и используется в рамках обычного кригинга (рис. 9.4*б*). Выбрана сферическая модель со значением наггета 33,59, плато — 267,0 и эллипсом корреляции с радиусами 306,3 и 230,4 м, главная ось — в направлении запад—восток.

Зона поиска представлена эллипсом с радиусами 310 и 235 м, главная ось — в направлении запад—восток.

Кросс-валидация выбранных параметров на имеющихся данных подтвердила их адекватность. Коэффициент корреляции между измерениями и оценкой обычного кригинга был в диапазоне от 0,74 до 0,78.

Для картирования были предоставлены два набора данных: обычный выброс (данные, статистически аналогичные априорным данным) и аварийный выброс (искусственно смоделированный выброс, наложенный на обычные

данные) [Dubois, Galmarini, 2005]. Обычный кригинг с описанными выше параметрами был использован для обоих наборов данных при прогнозировании значений в 808 валидационных точках. Полученные обычным кригингом оценки представлены на рис. 9.5. Результаты интерполяции различаются, так как зависят не только от параметров модели, но и от исходных данных.

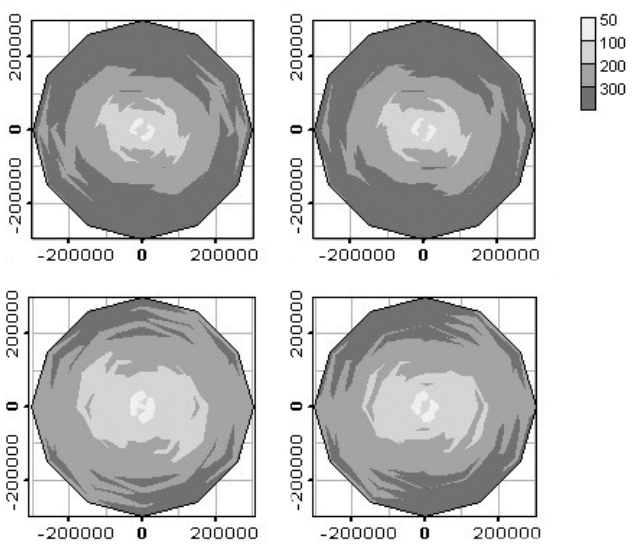


Рис. 9.3. Примеры экспериментальных вариограмм для четырех наборов измерений

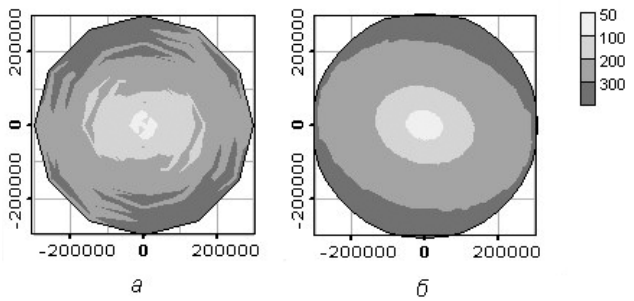


Рис. 9.4. Усредненная по 10 экспериментальным вариограммам вариограмма (*а*) и модель (*б*)

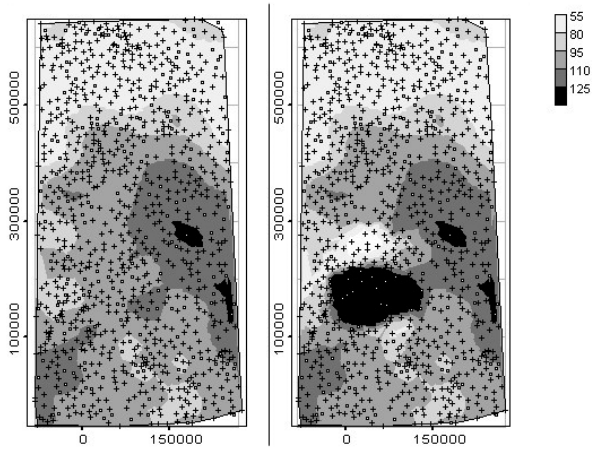


Рис. 9.5. Результат интерполяции обычным кригингом для обычных данных (*а*) и данных со смоделированным выбросом (*б*)

Как и ожидалось, обычный набор статистически похож на исторические наборы, поэтому результат валидации соответствует результатам кросс-валидации (коэффициент корреляции — 0,78).

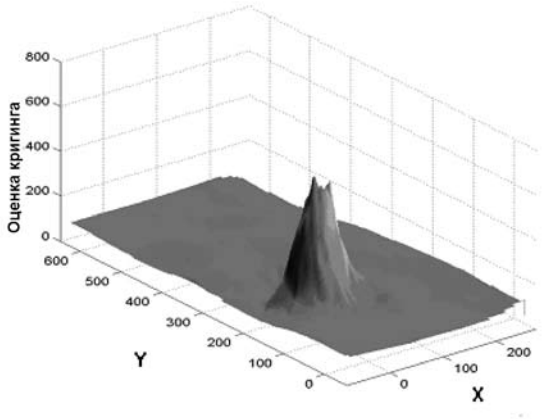


Рис. 9.6. Моделирование обычным кригингом экстремальных значений

Пятно, характеризующее выброс, обнаружено и хорошо видно на рисунке. Но другие области (особенно северная часть) кажутся не подверженными влиянию выброса и выглядят одинаково для обоих исходных наборов данных. Пятно растянуто и сглажено (рис. 9.6), но обнаружено. Основная проблема состоит в том, что в окрестности вокруг выброса (зоне максимально-

го градиента) нарушено предположение обычного кригинга о постоянстве среднего. Это вызывает искажение оценки (светлые пятна вокруг выброса на рис. 9.5*б*). Таким образом, оценка кригинга вокруг выброса в момент выброса не может считаться корректной.

На всех рисунках плюсами отмечены точки с исходными измерениями.

На рис. 9.7 представлена вариация кригинга. Она одинакова для обоих наборов, так как зависит только от модели вариограммы и пространственного распределения точек измерения. К сожалению, она не отражает того, что максимальная неопределенность оценки наличествует при выбросе в его окрестности.

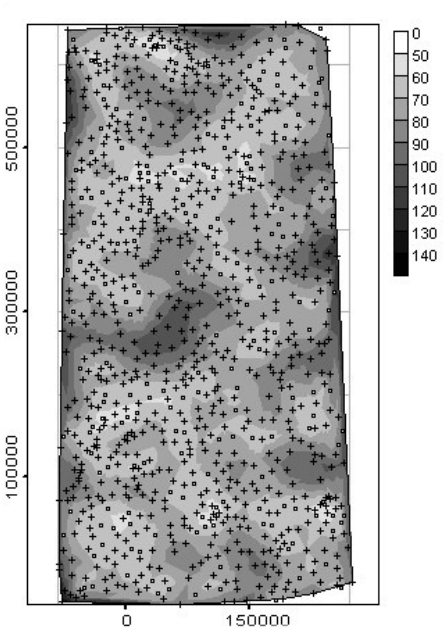


Рис. 9.7. Вариация кригинга одинаковая для обычных данных и данных со смоделированным выбросом

Воспроизведение статистических характеристик на валидационном наборе приведено в табл. 9.1. Обычный набор характеризуется хорошим совпадением. Для набора с выбросом хорошо воспроизведены среднее и медиана. Недооценка максимума и стандартного отклонения соответствует сглаженности оценки. Значение минимума в кригинговой оценке меньше минимума исходных данных, что связано с искажением оценки в области максимального градиента переменной. Там обнаружено 6 таких точек.

Таблица 9.1. Сравнение статистических характеристик, оцененных обычным кригингом и реальных значений

Данные	Минимум	Максимум	Среднее	Медиана	Стандартное отклонение
Реальные обычные	57,00	180,00	98,02	98,80	20,02
Оцененные обычные	66,80	130,35	96,63	98,50	15,19
Реальные с выбросом	57,00	1528,2	105,42	98,95	83,71
Оцененные с выбросом	37,53	651,18	103,23	97,89	53,17

Валидационные ошибки анализировались с помощью таких характеристик, как среднее от абсолютных значений ошибок (САО), среднее ошибок (СО), коэффициент корреляции между оцененными и реальными значениями и корень из среднеквадратичной ошибки (СКО). Эти характеристики представлены в табл. 9.2. Результаты прогноза для обычных данных лучше, чем для данных с выбросом, что вполне очевидно, так как обычные данные по характеристикам совпадают с теми, по которым настраивались параметры модели. Тем не менее результаты по данным с выбросом не являются бессмысленными.

Таблица 9.2. Анализ валидационных ошибок

Данные	САО	СО	Коэффициент корреляции	СКО
Обычные	9,11	–1,39	0,78	12,49
С выбросом	19,68	–2,18	0,56	69,08

Вывод. Обычный кригинг проявил себя вполне пригодным методом для анализа данных мониторинга в районе радиационно опасного объекта в обычных условиях и способным выявить аномальное поведение данных в случае выброса. В соревновании методов обычный кригинг проявил себя лучше многих более сложных методов.

Таким образом, обычный кригинг можно рекомендовать для включения в системы автоматического мониторинга, тем более что он не требует ничего, кроме вычисления линейной комбинации с уже подготовленными весами.

9.2. Анализ неопределенности в моделировании гидрогеологической структуры

Этот пример описывает моделирование одного гидрогеологического осадочного слоя в рамках гидрогеологической системы из 10 слоев. Анализ данных проводился в рамках совместных исследований ИБРАЭ РАН и Pacific Northwest National Laboratory по программе РАН и Министерства энергетики США. Результаты исследований представлены в [Savelieva et al., 2002].

Задача возникла в связи с анализом возможности переноса грунтовыми водами радиоактивного загрязнения из бункеров хранилищ в реку, являющуюся источником питьевой воды. Для моделирования была применена гидродинамическая модель, использующая параметры среды — проницаемость и пористость. Настройка параметров обычно осуществляется с использованием обратной задачи по результатам замеров в скважинах.

Использование настроенной единожды модели дает всегда один и тот же результат и не позволяет оценить его неопределенность. Оценка неопределенности результата и была основной побудительной причиной построения набора альтернативных моделей геологической среды.

Геологическая среда описывается как структура из 10 гидрогеологических слоев, расположенных в определенном порядке, но допускающих пропуски. Моделирование проводилось последовательно для каждого слоя. Здесь рассмотрен один из слабо проводящих слоев (U4), являющийся очень важным в данной гидрогеологической системе.

Моделирование проводилось в два этапа:

- получение зоны присутствия данного слоя (задача бинарной классификации);
- оценка толщины гидрогеологического слоя в областях его присутствия.

Так как изначально речь шла об оценке неопределенности и альтернативных моделях, моделирование толщины производилось с помощью стохастического метода.

Исходный набор данных содержал 401 скважину, где измерялись толщины гидрогеологических слоев. На рис. 9.8 приведено пространственное расположение скважин, каждая помечена в соответствии с присутствием (отсутствием) в ней слоя U4. Несколько скважин обозначены как неопределенные, т. е. эксперт-геолог не смог окончательно решить, присутствует

ли в керне слой U4. Толщина слоя U4 измерена в 117 скважинах, где он определенно обнаружен.

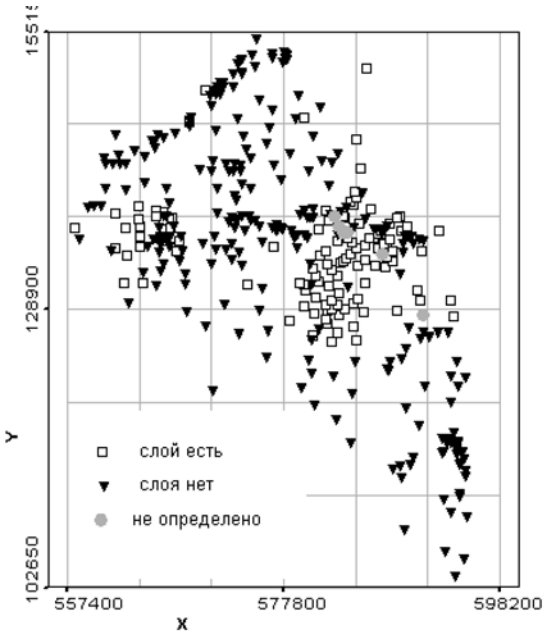


Рис. 9.8. Пространственное распределение присутствия и отсутствия слоя U4 в скважинах

Первая часть задачи — бинарная классификация — может решаться с помощью индикаторного кригинга. Индикаторное преобразование:

$$I(x) = \begin{cases} 1, \text{U4 есть в } x, \\ 0 \text{ в других случаях.} \end{cases}$$

Индикаторный кригинг даст вероятность присутствия слоя U4.

Но сначала требуется провести построение и моделирование индикаторной вариограммы. На рис. 9.9 приведены экспериментальная индикаторная вариограмма и ее модель. Они визуализированы с помощью изолиний вариограммы. Параметры модели вариограммы: наггет — 0,1, плато — 0,16, радиусы корреляции — 29 030 и 10 480 м, главная ось эллипса — в направлении 30° по часовой стрелке от оси север—юг.

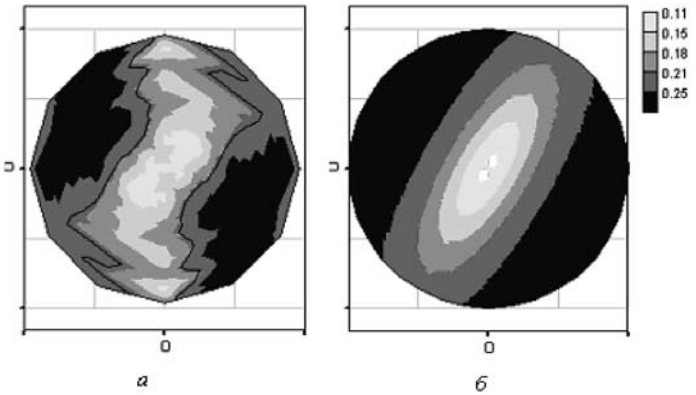


Рис. 9.9. Экспериментальная индикаторная вариограмма присутствия слоя U4 (*а*) и модель вариограммы (*б*)

Результат кросс-валидации дает ошибку классификации 18%.

Расчет проводился на ячейках размером 150×150 м и на области, ограниченной рассматриваемой гидрогеологической моделью. Результат картирования вероятности присутствия слоя U4 представлен на рис. 9.10.

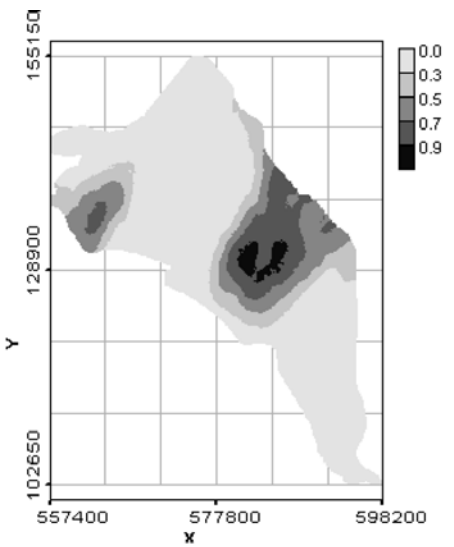


Рис. 9.10. Вероятность присутствия гидрогеологического слоя

Второй шаг — стохастическое моделирование толщины гидрогеологического слоя U4 с учетом его присутствия. Присутствие определяется по вероятности, полученной на предыдущем шаге. Эту вероятность можно учитывать по-разному. Можно сразу провести классификацию, считая, что слой присутствует, если вероятность его присутствия больше пороговой вероятности (например, 0,5). Другой подход состоит в розыгрыше присутствия каждый раз при движении по сетке оценивания. Мы использовали оба варианта.

В качестве метода стохастического моделирования использовалось последовательное гауссово моделирование. Данные подвергались нормализующему преобразованию. Бинормальность проверялась эмпирическим тестом — с использованием мадограммы:

$$\frac{\sqrt{\gamma(h)}}{M(h)} - \sqrt{\pi} \approx 0 .$$

Результат теста приведен на рис. 9.11.

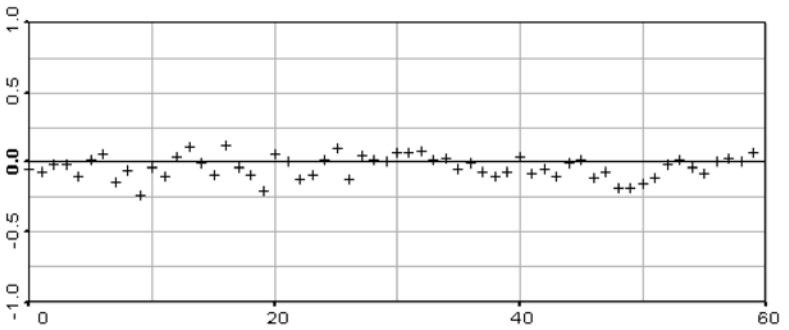


Рис. 9.11. Эмпирический тест на бинормальность

Данные по толщине слоя U4 обладают изотропной пространственной корреляцией. Для моделирования в нормализованных переменных использовалась сферическая модель с нулевым наггетом, единичным плато и радиусом корреляции 2707 м.

Примеры полученных реализаций приведены на рис. 9.12 (на рис. 9.12*б* — без розыгрыша присутствия слоя U4). В табл. 9.3 собраны некоторые статистические характеристики для пяти произвольно выбранных реализаций. Видно, что статистические характеристики реализации достаточно хорошо воспроизводят статистические характеристики. Примеры воспроизведения вариограмм несколькими реализациями приведены на рис. 9.13.

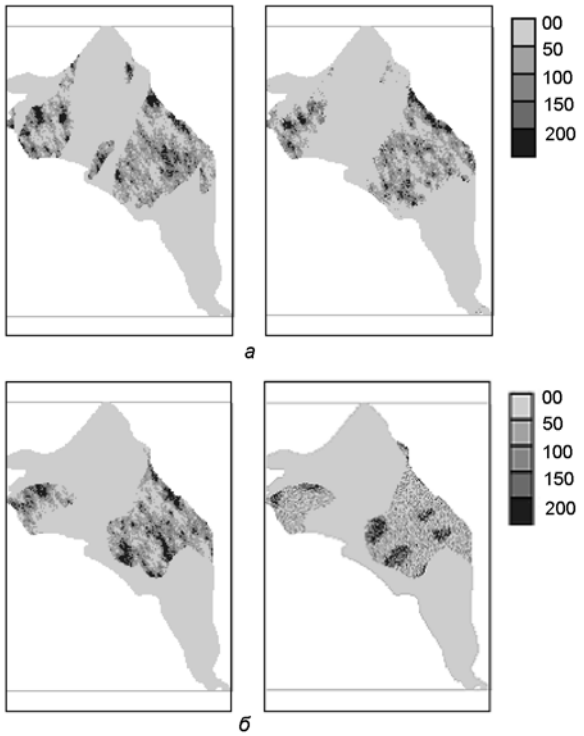


Рис. 9.12. Примеры реализаций толщины гидрогеологического слоя

Таблица 9.3. Глобальные статистические характеристики нескольких реализаций толщины слоя U4

Характеристика	Исходные данные	P1	P2	P3	P4	P5	R6	R7
1/4Q	4,5	4,1	4,5	4,6	4,2	4,5	4,5	4,5
Медиана	7,6	7,6	7,6	7,6	7,6	7,6	7,6	7,6
3/4Q	12,2	12,2	12,9	12,2	12,2	12,2	12,2	12,2
Nscore среднее	0,0	–0,04	0,07	0,02	–0,01	0,04	0,03	–0,02
Среднее	8,84	8,8	9,2	8,9	8,7	9,0	8,95	8,6
Вариация	40,8	40,7	43,3	39,8	39,2	39,7	39,7	35,8

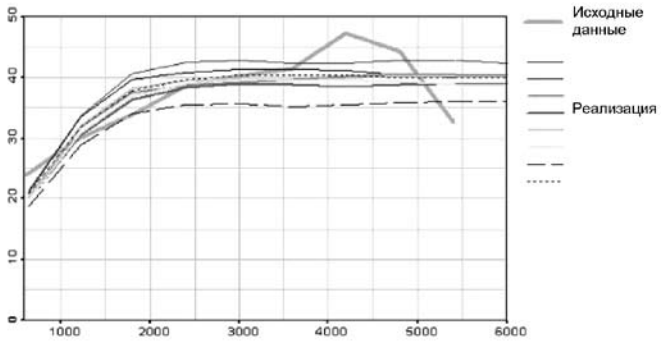


Рис. 9.13. Экспериментальные вариограммы исходных данных и нескольких реализаций

Аналогично можно проводить моделирование для других гидрогеологических слоев.

9.3. Сравнительный валидационный анализ геостатистических методов пространственного моделирования

В этом разделе мы приведем пример количественного сравнения различных геостатистических моделей — кригинга, стохастического моделирования — на примере реальных данных экологического мониторинга.

Воспользуемся данными по радиоактивному загрязнению почвы ^{241}Am, которые использовались в рамках совместных исследований ИБРАЭ РАН и Sandia National Laboratories по программе РАН и Министерства энергетики США. Результаты исследований опубликованы в [Kanevski et al., 2006].

Данные представляют собой набор измерений гамма-детектором в ряде точек, покрывающих большую площадь. Исходные 193 измерения были использованы для моделирования пространственного поля загрязнения и оценки в 917 валидационных точках. Значения в валидационных точках были изначально скрыты от исследователей для обеспечения чистоты эксперимента и приведены лишь после получения оценок для сравнения качества моделирования каждым методом. Исходные измерения приведены на рис. 9.14. Валидационные значения представлены на рис. 9.15. Целью исследования был вероятностный прогноз превышения уровней загрязнений 17, 27 и 38 пКи/г.

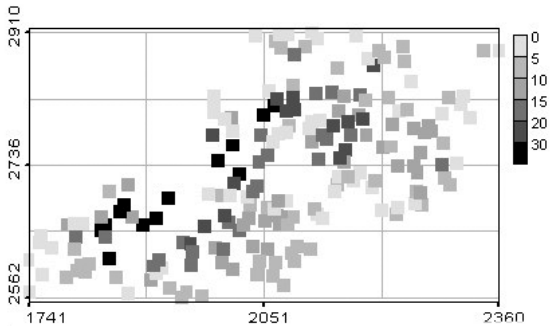


Рис. 9.14. Исходные данные по загрязнению ^{241}Am в 193 точках

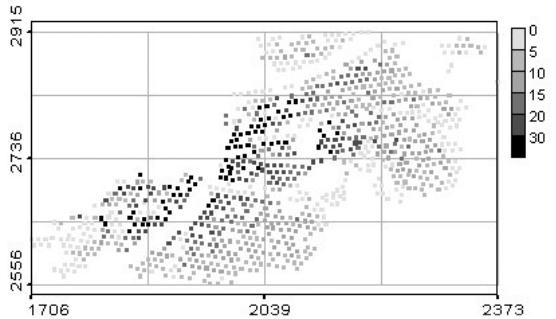


Рис. 9.15. Валидационные данные по загрязнению ^{241}Am в 917 точках

Одним из ключевых факторов успеха при решении проблемы валидации является репрезентативность исходных данных. Так, если исходные данные сильно отличаются от валидационных, трудно качественно оценить значения в валидационных точках на основе исходных данных. Исходный и валидационный наборы данных достаточно однородно распределены в пространстве. Глобальные статистики для обоих наборов, приведенные в табл. 9.4, имеют близкие значения, из чего следует хорошая репрезентативность исходных данных. Заметим, однако, что длинный хвост высоких значений для валидационных данных вдвое превышает максимальное значение исходных данных.

Таблица 9.4. Итоговая статистика для данных по загрязнению Am^{214} (пКи/г)

Статистика	Исходные данные	Валидационные данные
Количество	193	917
Минимум	0,822	0,38
Нижний квартиль 25%	4,67	4,48
Медиана	8,87	9,74
Верхний квартиль 75%	15,56	16,31
Максимум	77,20	115,74
Среднее значение	12,19	12,69
Стандартное отклонение	12,34	12,74
Вариация	152,26	162,4
Коэффициент симметрии	2,51	2,9
Эксцесс	8,03	14,0

Дальнейшее сравнение исходных и валидационных наборов заключается в рассмотрении их пространственной корреляции. Сравнение вариограмм по всем направлениям для исходных и валидационных данных показывает их близость за исключением более высокой вариабельности и меньшей стационарности валидационных данных (рис. 9.16). При рассмотрении вариограмм по различным направлениям можно выявить более значительные различия между корреляционными структурами исходных и валидационных данных. Так, валидационные данные имеют четкую геометрическую анизотропию в горизонтальном направлении восток—запад, в то время как исходные данные демонстрируют слабую анизотропию в вертикальном направлении север—юг (рис. 9.17*а*). Вариограммная модель, построенная на основе исходных данных, приведена на рис. 9.17*б*. Вариограммные модели были также построены для преобразованных значений — нормализованных и индикаторных переменных для 9 пороговых значений (для использования в соответствующих методах). Качество всех моделей было проверено при помощи кросс-валидации и тестирования на 30 данных, предварительно отделенных от исходного набора.

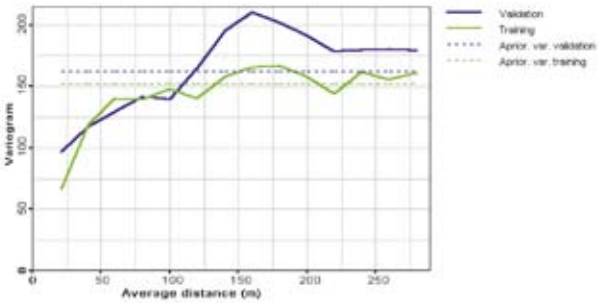
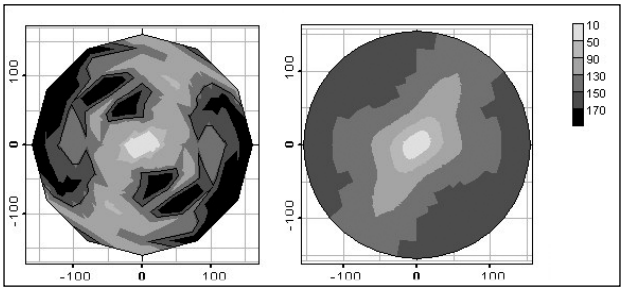


Рис. 9.16. Вариограммы по всем направлениям для исходных (нижняя линия) и валидационных (верхняя линия) данных



а *б*

Рис. 9.17. Изолинии вариограммной розы для исходных сырых данных (*а*) и вариограммной модели (*б*)

Разнообразные геостатистические модели были применены для решения задачи валидации. Использовались геостатистические оцениватели: простой (SK) и обычный (OK) кригинг для получения точечных оценок загрязнения, а также индикаторный кригинг (IK) для получения вероятностных оценок. Три стохастические модели — последовательное гауссово моделирование (SGS), последовательное индикаторное моделирование (SIS) и моделирование отжига (SA). Таким образом, сравнительный анализ методов был проведен для широкого спектра геостатистических моделей. Расчеты простого и обычного кригинга проводились при помощи пакета программ «Геостат Офис» [Kanevski, Maignan, 2004], результаты стохастического моделирования и индикаторного кригинга были получены при помощи программ GSLIB [Deutsch, Journel, 1998].

Таблица 9.5. Сравнение итоговой статистики для валидационных данных с оценками различными геостатическими

Статистика	Данные	SK	OK	IK (E-type)	SGS	SIS	SA
Минимум	**0,38**	0,0	0,0	2,2	**0,65**	0,001	0,82
Нижний квартиль (25%)	**4,4**	6,8	6,6	7,3	5,6	5,4	6,5
Медиана	**9,74**	10,1	**9,98**	10,3	**9,98**	9,12	10,88
Верхний квартиль (75%)	**16,3**	15,9	15,9	15,9	17,02	**16,0**	17,91
Максимум	**115,7**	67,9	69,0	60,3	**79,3**	68,5	64,1
Среднее значение	**12,7**	13,5	13,4	14,2	13,9	**12,5**	14,4
Стандартное отклонение	**12,8**	11,1	11,4	11,6	13,92	11,3	**12,4**
Симметрия	**2,9**	2,2	2,2	2,0	6,08	4,96	3,6
Эксцесс	**14,0**	5,2	**5,3**	4,1	2,32	1,98	1,79

Примечание. SK — простой кригинг, OK — обычный кригинг, IK — усредненная оценка индикаторного кригинга, SGS — среднее значение гауссова моделирования, SIS — среднее значение индикаторного моделирования, SA — среднее значение моделирования отжигом. Полужирным шрифтом выделены значения статистик оценок, наиболее близкие к статистикам валидационного набора.

Оценка кригинга рассчитывалась на основе построенной вариограммной модели. Вариация оценок кригинга в валидационных точках не зависит от значения оценки и отражает плотность сети мониторинга.

Индикаторным кригингом с использованием девяти индикаторных переменных были получены локальные функции распределения вероятности в валидационных точках. Для сравнения с оценками кригинга были использованы усредненные значения E-типа (E-type) (см. Раздел 7.2).

Стохастическое моделирование было проведено на регулярной сетке с шагом 5×5 м. Далее значения в валидационных точках были получены методом ближайшего соседа. Такая последовательность обусловлена ограничениями программ пакета GSLIB, что может внести смещение в окончательные результаты валидации. Однако выбор достаточно высокого разрешения сетки моделирования по сравнению с разрешением сети валидационных данных позволяет считать это хорошей аппроксимацией. Ошибка аппроксимации при использовании в данном случае метода ближайшего соседа

будет значительно меньше, чем ошибка измерения и локальная неопределенность в валидационных точках. При анализе результатов стохастического моделирования вместо использования усредненных оценок E-типа (см. Главу 8) были рассчитаны статистические параметры распределений отдельных реализаций. Далее эти статистики были усреднены для сравнения с валидационным распределением. Как видно из табл. 9.5, усредненная оценка индикаторного кригинга (IK) сильно сглажена по сравнению с валидационными данными. Это обусловлено выбором среднего значения промоделированных распределений. Оценка кригинга дала хорошее совпадение медианы распределения и, более того, эксцесса, который близок к экцессу валидационного распределения. Стохастическое моделирование позволяет лучше воспроизвести валидационное распределение, чем кригинг. Можно видеть, что различные статистические параметры воспроизводятся лучше разными методами. Так, результаты гауссова моделирования (SGS) имеют значения минимума, максимума и медианы, наиболее близкие к соответствующим параметрам валидационного набора. Индикаторное моделирование (SIS) дало наиболее близкие средние значения и значения квартилей. Реализации моделирования отжигом (SA) имеет наилучшие стандартное отклонение и коэффициент симметрии. Еще раз подчеркнем, что статистические параметры были получены путем усреднения статистик каждой из 100 реализаций для каждого алгоритма.

Пространственные корреляционные структуры оценок кригинга и реализаций стохастического моделирования представлены соответствующими вариограммами, которые сравнивались с вариограммой валидационных данных. Вариограмма оценки кригинга ожидаемо недооценивает уровень пространственной вариации, хотя размеры корреляции представлены достаточно хорошо (рис. 9.18*д*). Вариограмма оценки кригинга не имеет наггета, в то время как наггет по вариограмме валидационных данных составляет 25—30% априорной вариации.

Стохастическое моделирование позволяет лучше промоделировать вариабельность и неопределенность пространственной корреляции, которую можно представить доверительными интервалами вариограммы на основе стохастических реализаций. Распределение вариограмм для реализаций представлено средней вариограммой с доверительным интервалом $\pm 2\sigma$ (рис. 9.19). Как было отмечено выше, пространственная корреляция валидационных данных существенно отличается от корреляции исходных данных. Все алгоритмы кроме индикаторного моделирования (SIS) демонстрируют хорошее совпадение пространственной корреляционной структу-

ры. Усредненная вариограмма реализаций индикаторного моделирования (SIS) недооценивает уровень вариабельности даже с учетом интервала неопределенности (см. рис. 9.19*б*). Широкие доверительные интервалы вариограмм реализаций говорят о значительной неопределенности, которую воспроизводят стохастические реализации. Усредненные вариограммные розы реализаций демонстрируют хорошее совпадение со структурой валидационных данных (см. рис. 9.18). Результаты моделирования отжига (SA) дают наиболее близкое совпадение пространственной корреляции с валидационным распределением (см. рис. 9.18*а*). Усредненные вариограммные розы для реализаций гауссова (SGS) и индикаторного (SIS) моделирования более близки к пространственной корреляции исходного распределения (см. рис. 9.18*б,г*).

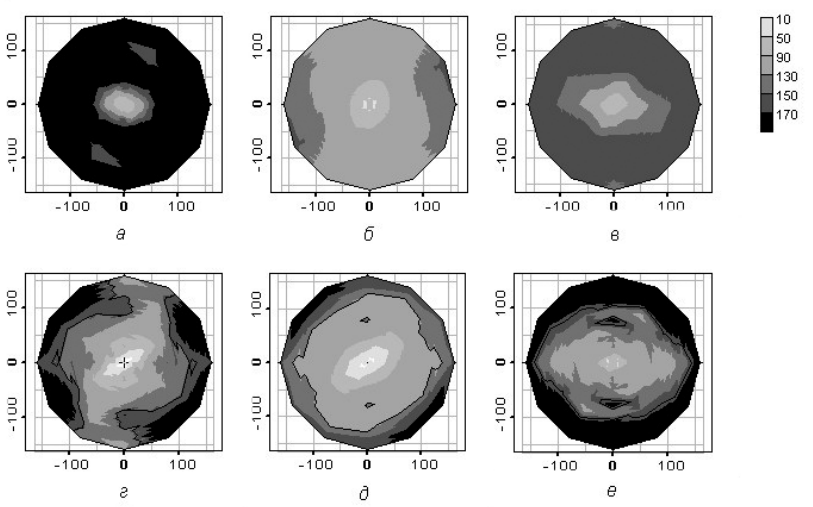


Рис. 9.18. Сравнение контуров вариограммной розы валидационных данных с усредненными вариограммами по 100 стохастическим реализациям для гауссова моделирования SGS (*а*), индикаторного моделирования SIS (*б*), моделирования отжигом (*в*), гауссова моделирования невязок нейронной сети (*г*), вариограммы оценок обычного кригинга (*д*) и валидационных данных (*е*)

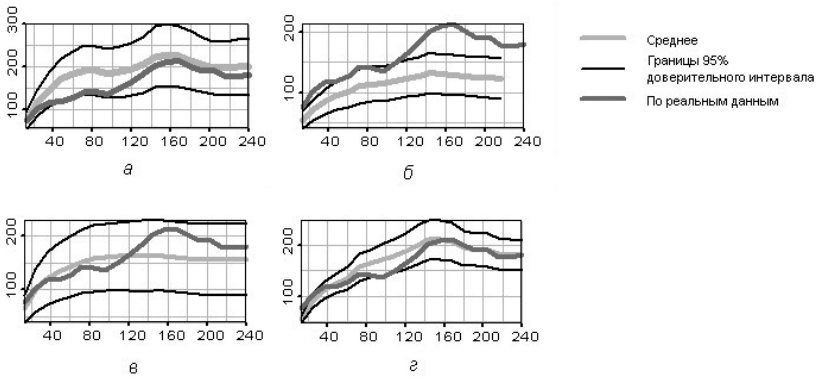


Рис. 9.19. Средняя вариограмма по всем направлениям с доверительным интервалом $\pm 2\sigma$ для 100 реализаций:
гауссово моделирование (SGS) (*а*),
индикаторное моделирование (SIS) (*б*),
моделирование отжигом (SA) (*в*),
гауссово моделирование невязок нейронной сети (*г*)
в сравнении с вариограммой для валидационных данных

Анализ невязок, оставшихся после оценок кригинга, обнаружил существование пространственной корреляции в них на малых расстояниях. Возможно, это связано с присутствием нестационарности, которая не учитывается кригингом. Так, крупномасштабный пространственный тренд, выявленный в валидационных данных, не проявлялся в исходных данных и поэтому не был учтен. Оценки обычного кригинга (OK) вместе с соответствующей вариацией приведены на рис. 9.20.

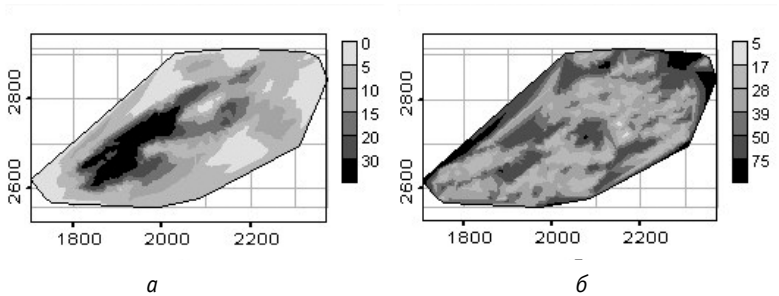


Рис. 9.20. Оценки обычного кригинга (*а*) и соответствующая вариация оценки (*б*)

Вероятностные оценки превышения уровня загрязнения ^{241}Am 27 пКи/г, полученные индикаторным кригингом (IK) и различными алгоритмами стохастического моделирования, приведены на рис. 9.21.

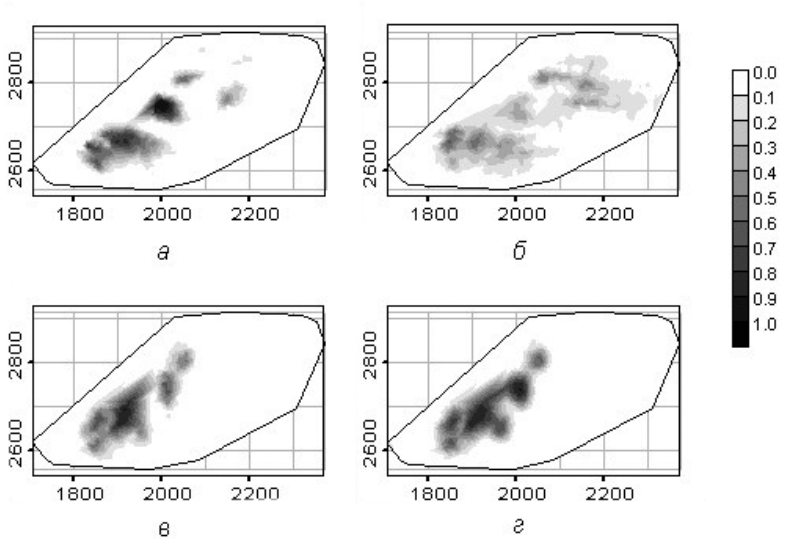


Рис. 9.21. Оценки вероятности превышения уровня концентрации Am^{241} 27 пКи/г, полученные гауссовым моделированием (SGS) (*а*), индикаторным моделированием (SIS) (*б*), моделированием отжига (SA) (*в*), индикаторным кригингом (IK) (*г*)

Качество валидационной оценки можно определить путем сравнения промоделированных локальных функций распределения с валидационными данными. На рис. 9.22 представлены локальные функции распределения, полученные различными методами в четырех валидационных точках. На те же графики нанесены оценки обычного кригинга (OK) с доверительным интервалом $\pm\sigma$ ошибки оценки кригинга.

Количественный анализ удовлетворения доверительных интервалов, построенных на основе локальных функций распределения во всех 917 валидационных точках, приведен в табл. 9.6. Были выбраны четыре доверительных интервала: размах (между минимумом и максимумом оценки), 90% распределения между 5 и 95%; 80% распределения между 10 и 90% распределения; межквартильный интервал между 25 и 75% распределения. Из табл. 9.6 видно, что все три метода стохастического моделирования дают доверительные интервалы, хорошо удовлет-

воряющие валидационным данным. Процент валидационных данных, попавших в соответствующий интервал, хорошо согласуется с размерами интервала. Наилучшие результаты показало моделирование отжига (SA).

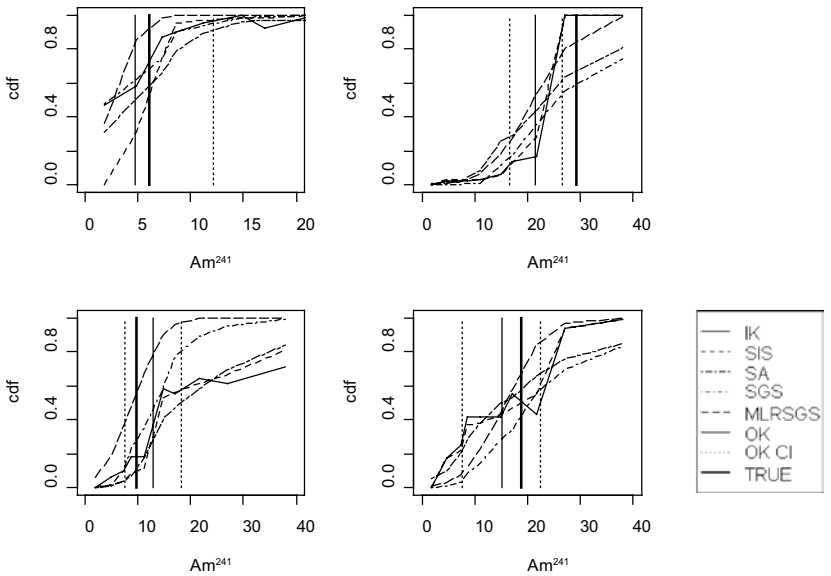


Рис. 9.22. Локальные кумулятивные функции распределения вероятности в четырех валидационных точках, полученные различными методами: индикаторным кригингом (IK), гауссовым моделированием (SGS), индикаторным моделированием (SIS), моделированием отжига (SA), моделированием невязок нейронной сети (MLRSGS), в сравнении с соответствующими валидационными измерениями (жирная вертикальная линия) и значениями оценки обычного кригинга (OK — тонкая вертикальная линия) с доверительным интервалом $\pm\sigma$ (OK CI — вертикальный пунктир)

Таблица 9.6. Доля валидационных данных, попавших в интервалы неопределенности, полученные различными методами стохастического моделирования, %

Интервал	SGS	SIS	SA
От минимума до максимума	94	97	97
Квантиль 5% — квантиль 95%	82	82	86
Квантиль 10% — квантиль 90%	75	70	77
Квантиль 25% — квантиль 75%	49	46	52

Пример исследования показал, что методы стохастического моделирования способны хорошо воспроизводить пространственную вариабельность. Вероятностные оценки на основе стохастического моделирования предпочтительнее вероятностных оценок индикаторного кригинга. Различные алгоритмы стохастического моделирования дают близкие результаты, которые достаточно хорошо согласуются с распределением валидационных данных. Различные методы лучше воспроизводят разные статистические параметры. Количественный анализ неопределенности оценки позволяет сравнить процент валидационных данных, попавших внутрь доверительного интервала.

Литература

Deutsch C. V., Journel A. G. GSLIB: Geostatistical Software Library and User's Guide. — New York: Oxford Univ. Press, 1998. — 369 p.

Dubois G., Galmarini S. Introduction to the Spatial Interpolation Comparison (SIC) 2004 exercise and presentation of the data sets // Applied GIS. — 2005. — Vol. 1, N 2. — P. 9.1—9.10 (http://publications.epress.monash.edu/doi/pdf/10.2104/ag050009).

Kanevski M., Demyanov V., Savelieva E. et al. Validation Of Geostatistical And Machine Learning Models For Spatial Decision-Oriented Mapping // Proceeding of StatGIS 99 / Ed. J. Piltz, J. Heyn. — Klagenfurt, 2006.

Kanevski M., Maignan M. Analysis and modelling of spatial environmental data. — Lausanne: EPFL Press, 2004. — 288 p. — (With a CD and educational/research MS Windows software tools).

Savelieva E. Using Ordinary Kriging to Model Radioactive Contamination Data // Applied GIS. — 2005. — Vol. 1, N 2. — P. 10.1—10.10.

Savelieva E., Kanevski M., Timonin V. et al. Uncertainty in the hydrogeologic structure modeling // Proceedings of IAMG2002 conference. — [S. l.], 2002. — P. 481—486.

Глава 10

Комбинированные модели ИНС и геостатистики

Раздел 10.1 настоящей главы посвящен постановке задачи комбинированного моделирования на основе искусственных нейронных сетей (ИНС) и геостатистики. В Разделах 10.2, 10.3 приведены примеры использования предложенного комбинированного метода для моделирования пространственных и временны́х данных: в Разделе 10.2 рассмотрено картирование атмосферных осадков при помощи кригинга невязок ИНС [Kanevsky et al., 1998], а в Разделе 10.3 — прогнозирование электропотребления при помощи стохастического моделирования невязок ИНС.

Проблема существования пространственной корреляции на различных масштабах обычно связана с различными источниками, процессами образования данных и влияющими эффектами. Так, радиоактивное загрязнение поверхности почвы обусловлено крупномасштабными процессами динамики атмосферы, однако локальные изменения погодных условий, орографические эффекты и свойства подстилающей поверхности также вносят свой вклад. Таким образом, различные физические процессы на разных масштабах сильно влияют на пространственную структуру данных. На практике часто трудно воспроизвести такую структуру на различных масштабах при помощи математической модели. Предположение о стационарности, которое обычно используется в геостатистических моделях, тесно связано с многомасштабностью данных (см. Раздел 2.7). В Разделе 4.10 были кратко перечислены некоторые подходы к решению проблемы присутствия тренда в данных. В этой главе мы подробно опишем один из них — комбинированное моделирование на основе ИНС и геостатистики.

10.1. Геостатистический анализ невязок

Идея метода заключается в моделировании нелинейного крупномасштабного тренда при помощи ИНС и последующего моделирования невязок геостатистическими методами. Этот подход был впервые предложил М. Ф. Каневский

[Kanevsky et al., 1996a] и в дальнейшем развивался в последующих работах [Kanevsky et al., 1997, 1998; Demyanov et al., 2000; 2001; Savelieva et al., 2000]. Изначально невязки, оставшиеся после применения ИНС для пространственного оценивания данных, были оценены обычным кригингом. Таким образом, итоговая оценка была получена как сумма оценки ИНС и кригинга невязок. Последующее развитие подхода касалось обобщения на случай нескольких переменных и применения стохастического моделирования невязок. В случае нескольких переменных используются ИНС с несколькими выходными нейронами и для моделирования невязок применяется кокригинг (см. Главу 6) [Kanevski et al., 1997]. Методы стохастического моделирования (см. Главу 8) привлекаются для оценивания невязок аналогичным образом [Demyanov et al., 2000]. В Разделе 9.3 был приведен один из таких методов с использованием последовательного гауссового моделирования в качестве сравнения (MLRSGS). Комбинированный подход на основе ИНС и геостатистики нашел применение и в других работах, например в [Cortez et al., 1998; Bryan, Adams, 2002; Zhang et al., 2004].

Преимущество использования ИНС для моделирования тренда перед другими моделями тренда (полиномы, сплайн и пр.) заключается в том, что ИНС является универсальным оценивателем и хорошо моделирует нелинейные структуры. ИНС не предполагает фиксированной аналитической зависимости, а, наоборот, способна получить эту зависимость на основе имеющихся данных в процессе обучения. Более подробно теория ИНС и алгоритмов обучения изложена, например, в [Haykin, 1998].

В качестве ИНС может использоваться как наиболее популярный многослойный перцептрон, так и более сложные нейронные сети (обобщенной регрессии, радиальных базисных функций). Ключевым моментом является анализ и моделирование корреляционной структуры невязок, оставшихся после вычета из данных оценки ИНС.

В случае отдельного применения ИНС анализ невязок также представляется важным. Он помогает проинтерпретировать результаты и оценить их качество. Если не обнаружено корреляции между невязками и исходными данными, значит, вся информация из данных успешно моделируется при помощи только ИНС. Таким образом, ИНС применяется для интерполяции напрямую.

Устойчивость (робастность) подхода показывает, насколько он чувствителен к выбору архитектуры ИНС и алгоритма обучения. Итоговая статистика и моменты второго порядка (вариограммы) невязок устойчивы по отношению

к изменению количества скрытых слоев и числа нейронов в них. Таким образом, рекомендуется выбирать наиболее простую по архитектуре ИНС, которая в то же время в состоянии обучиться и воспроизвести нелинейные тренды. Обычно выбор подходящей сети осуществляется на основе теста на аккуратность. Тем не менее могут быть использованы более сложные тесты.

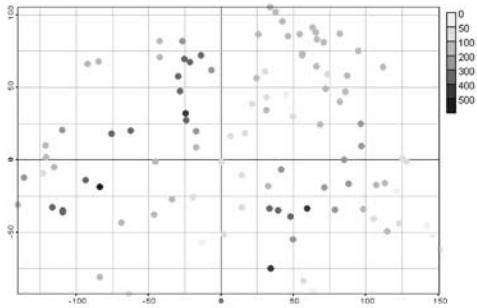
При анализе невязок — разницы между данными и оценками ИНС — возможно несколько вариантов. Если невязки не обладают пространственной корреляцией, а распределены совершенно случайно, это может означать, что ИНС полностью промоделировала структуру данных. В этом случае оценку ИНС можно принять как окончательную. Если невязки имеют пространственную структуру, а также коррелированы с исходными данными, необходимо проводить дальнейшее моделирование невязок. Обычно корреляция невязок и исходных данных слабее, чем корреляция данных с оценками ИНС (в случае корректного обучения и использования ИНС). Можно видеть, что невязки обладают пространственной корреляцией на меньших расстояниях, чем данные. Это обусловлено тем, что ИНС уже промоделировала корреляцию на более крупных масштабах. Это свойство невязок часто позволяет предположить их стационарность на всей области исследования, чего нельзя было предположить для исходных данных с трендом. Таким образом, вариограммная модель для невязок отличается коротким радиусом и стабилизированным плато. Использование такой модели в кригинге дает корректные и точные результаты.

Кригинг невязок (residual kriging), как и универсальный кригинг, предполагает, что неизвестное среднее значение $m(x)$ меняется во всей области исследования S так, что нельзя допустить постоянство даже локального среднего. В этом случае компонента тренда моделируется отдельно, используя другие математические или физические подходы. В качестве модели тренда можно использовать прогноз, выполненный с использованием физической модели процесса, являвшегося причиной формирования поля $Z(x)$. Компонента тренда, как уже указывалось, может моделироваться с помощью нелинейных адаптивных методов (например, искусственных нейронных сетей, вейвлетов [Demyanov et al., 2001], метода регрессии на опорных векторах [Kanevski et al., 2003, 2004] и др.), использующих набор измерений как информацию для настройки своих параметров. После выделения тренда простой или обычный кригинг используется на невязках модели к измеренным значениям поля.

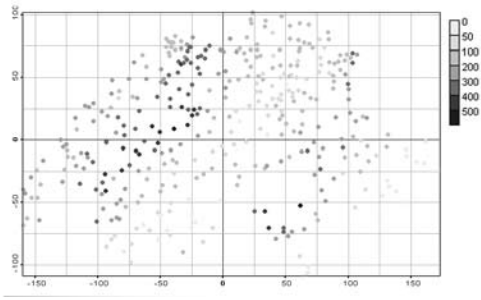
Кригинг невязок можно также интерпретировать как гибридную модель, объединяющую различные по математической или физической базе методы.

10.2. Пример использования кригинга невязок

В данном примере использованы метеорологические данные по усредненным за 10 дней выпадениям осадков в Швейцарии в 1986 г. Задача состояла в оценке значений в 367 точках по 100 измерениям (рис. 10.1). Эти данные распространялись в рамках международного конкурса сравнения методов пространственной интерполяции (Spatial Interpolation Comparison — SIC'97), организованного геостатистическим порталом AI-GEOSTAT [SIC'97] и Группой мониторинга радиоактивности в окружающей среде Института окружающей среды в Объединенном исследовательском центре (Radioactivity Environmental Monitoring group of the Environment Institute at the Joint Research Centre), Испра, Италия.



а



б

Рис. 10.1. Исходные данные (100 точек), использованные для моделирования (*а*), и валидационные данные для проверки качества оценки (367 точек) (*б*)

Данные обладают корреляционной структурой с трендом — дрейфом (определение дрейфа см. в Разделе 4.2), который почти во всех направлениях практически монотонно убывает (рис. 10.2). Очевидно, что для картирования таких данных требуется моделирование тренда. Было предложено построить нелинейную модель тренда с помощью искусственной нейронной сети [Kanevsky et al., 1998] или с использованием вейвлет-метода [Demyanov et al., 2001].

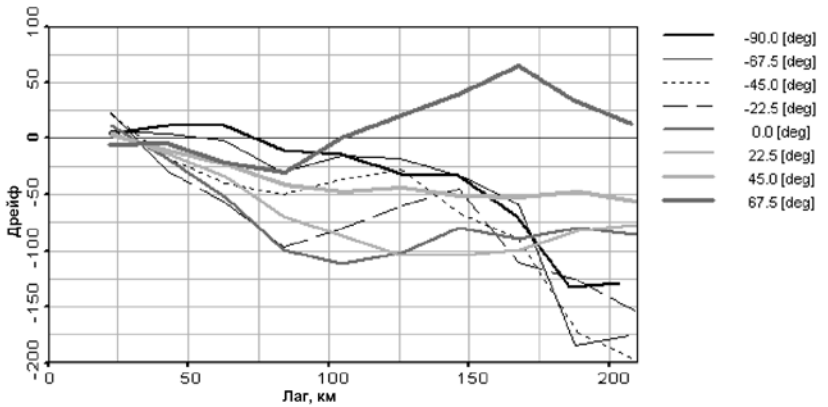


Рис. 10.2. Дрейф данных по выпадению осадков, рассчитанный в различных направлениях

После применения нелинейной модели к исходному набору (100 тренировочных точек) были получены оценки ИНС. Была использована ИНС типа многослойный перцептрон с двум входными нейронами (по количеству пространственных координат) и одним выходным нейроном — оцениваемой переменной. Количество нейронов в единственном скрытом слое варьировалось. Приведенные на рис. 10.3 вариограмма для оценок ИНС и вариограмма исходных данных демонстрируют совпадение, что свидетельствует о хорошем качестве модели ИНС. Вариограммы отражают сложную периодическую корреляционную структуру на нескольких масштабах. Однако если посмотреть на невязки — разницу между данными и оценками ИНС, можно видеть, что они коррелированы со значениями данных (рис. 10.4*б*). Это означает необходимость дальнейшего моделирования невязок.

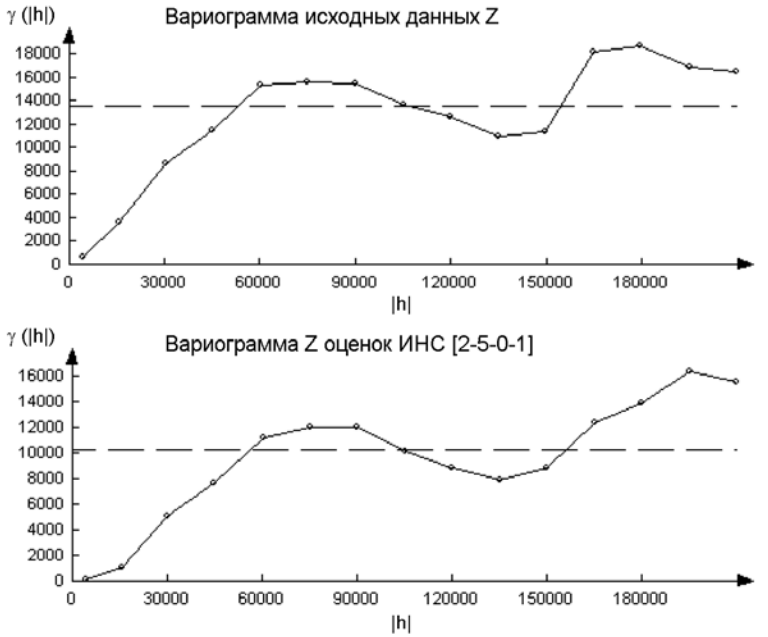
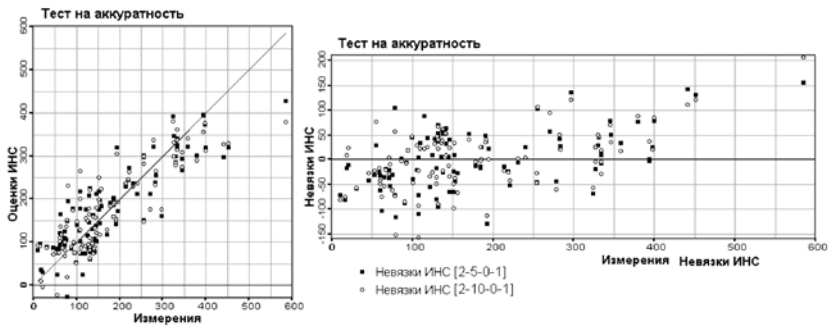


Рис. 10.3. Экспериментальные вариограммы для исходных данных (измерений) (вверху) и оценок ИНС (внизу)



а *б*

Рис. 10.4. Тест на аккуратность: зависимость оценок ИНС от измерений для ИНС [2-5-1] и [2-10-1] (*а*), зависимость невязок ИНС от измерений (*б*)

Невязки ИНС в отличие от исходных данных демонстрируют отсутствие пространственного тренда во всех направлениях (см. рис. 10.5), который полностью оценен ИНС. Пространственная корреляция невязок имеет ко-

роткий радиус — 30—80·10^3 м (по сравнению с радиусом корреляции исходных данных — 80—200·10^3 м) и обладает стационарностью (рис. 10.6). Таким образом, невязки можно эффективно промоделировать обычным кригингом. Пространственная корреляция (вариограмма) хорошо моделируется сферической моделью с учетом анизотропии (см. рис. 10.7*б*): радиусы корреляции 73,01 км и 54,53 км, больший под углом 15° по часовой стрелке от направления с северо-запада на юго-восток.

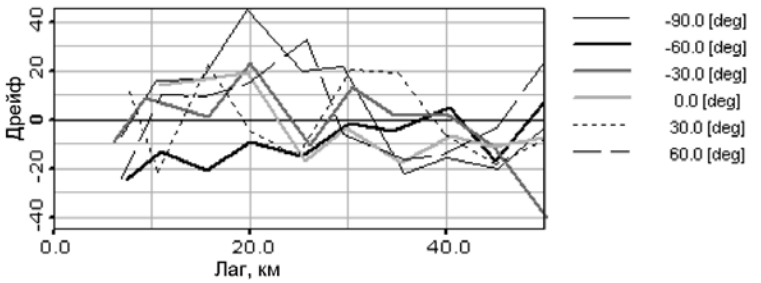


Рис. 10.5. Дрейф невязок модели нелинейного тренда

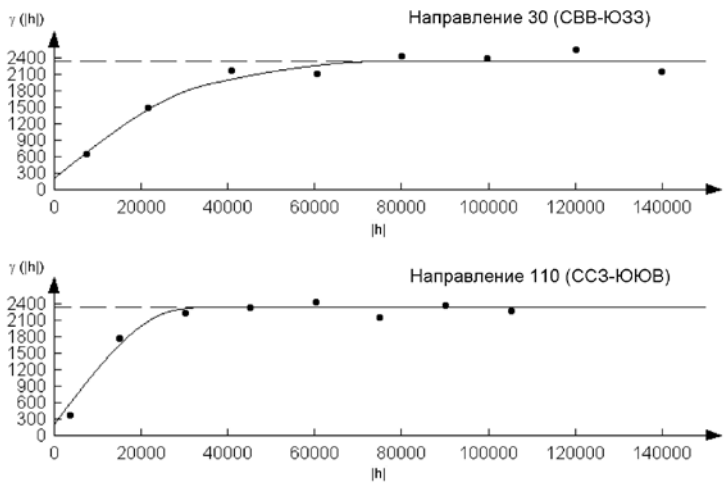
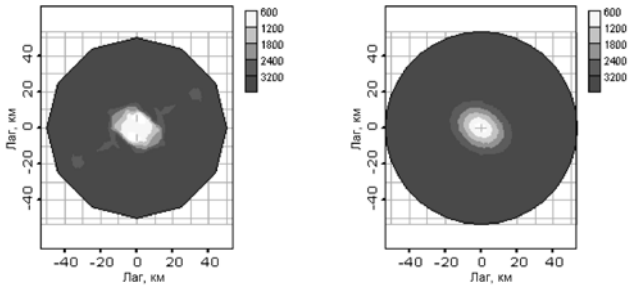


Рис. 10.6. Экспериментальные вариограммы и их анизотропная модель для невязок ИНС [2-5-1]



а *б*

Рис. 10.7. Контуры вариограммной розы: экспериментальная (*а*) и модель ИНС (*б*)

Результат, полученный после применения обычного кригинга к невязкам и сложения результата нелинейной модели и кригинга невязок, представлен на рис. 10.8. Он имеет естественный пятнистый вид, воспроизводящий корреляционную структуру на различных масштабах. Проверка качества оценки модели была проведена на валидационном наборе. Статистические характеристики валидационной оценки близки к валидационным данным (табл. 10.1). При сравнении с обычным кригингом [Atkinson, Lloyd, 1998] выяснилось, что среднеквадратичная ошибка обычного кригинга (5,97) несколько выше, чем таковая кригинга невязок ИНС (5,6). Стандартное отклонение валидационной ошибки также выше (на 6%) у оценки обычного кригинга (59,69), чем оценка кригинга невязок ИНС (56,28), что означает более широкий разброс ошибок. На рис. 10.9 приведен график зависимости валидационной ошибки от исходных данных, который показывает отсутствие корреляции между ними — кригинг невязок ИНС промоделировал всю пространственную структуру.

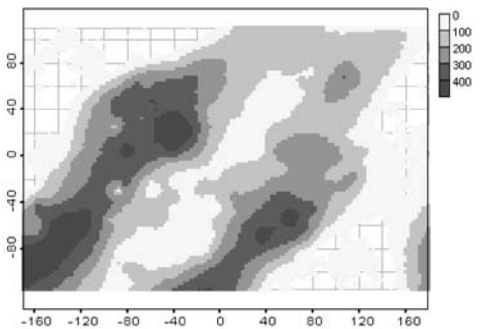


Рис. 10.8. Результат картирования с использованием нелинейной модели и кригинга невязок

Таблица 10.1. Статистические характеристики для валидационного набора

Характеристика	Данные	ИНС + кригинг невязок
Минимум	0	0
Нижний квартиль	100	109
Медиана	162	167
Верхний квартиль	264	267
Максимум	517	513
Среднее	185	187
Стандартное	111	112

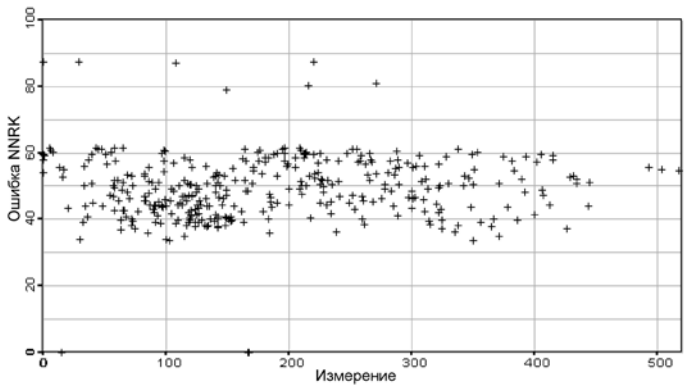


Рис. 10.9. Валидация: зависимость ошибки оценки кригинга невязок ИНС от измеренных значений

10.3. Пример использования стохастического моделирования невязок

В данном примере рассмотрено применение ИНС и геостатистики для краткосрочного (на неделю вперед) прогнозирования электропотребления в Московском регионе. Электропотребление обладает периодической структурой на различных временны́х масштабах (сутки, недели, годы), а также связано сложной нелинейной зависимостью с погодными параметрами (температурой, облачностью, осадками и т. д.). Все погодные параметры, использующиеся при прогнозировании, также являются прогнозными, по-

этому даже самый лучший метод не может дать идеальный прогноз. Таким образом, здесь встает задача анализа неопределенности прогноза.

Для прогнозирования используется искусственная нейронная сеть как универсальный нелинейный аппроксиматор. Прогноз делается на две недели вперед: первая неделя — текущая, т. е. с уже известными значениями электропотребления, следующая — непосредственно прогнозируемая.

Относительные ошибки прогноза ИНС представлены на рис. 10.10. Видно, что в большей части прогноз имеет ошибку меньше 10%. Возможно, прогноз может быть улучшен за счет изменения набора входных параметров. В данной работе количество исходной информации было очень ограничено. На рис. 10.11 приведена вариограмма невязок прогноза ИНС, на которой хорошо наблюдается временна́я корреляция.

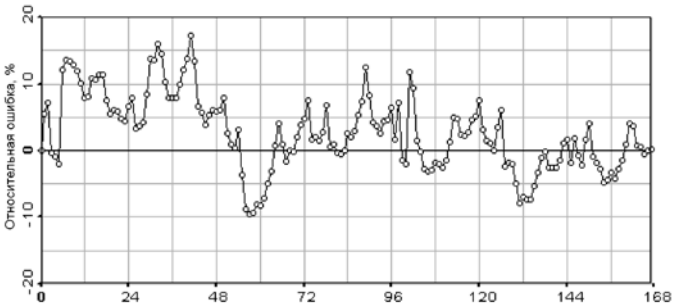


Рис. 10.10. Невязки после прогнозирования электропотребления с помощью ИНС

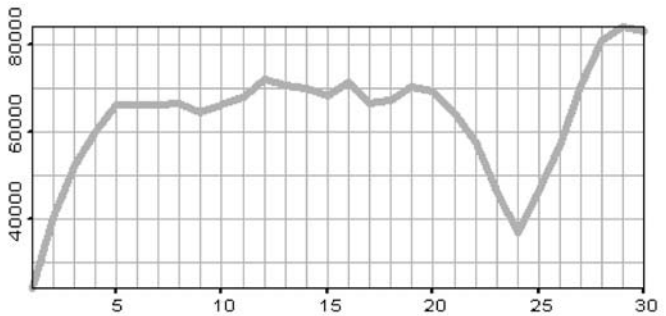


Рис. 10.11. Вариограмма невязок прогноза ИНС

Для прогнозирования неопределенности прогноза используется стохастическое моделирование невязок. Оно выполняется с помощью моделирования отжига. Делаются безусловные симуляции, воспроизводящие вариограмму и гистограмму исходных невязок. В данном случае нет необходимости строить модель вариограммы, так как исходные данные заданы на такой же сетке, как и строящиеся симуляции, т. е. для любого лага значение вариограммы известно.

Несколько полученных реализаций представлено на рис. 10.12, а на рис. 10.13 показано качество воспроизведения вариограммы, где толстой серой линией изображена исходная вариограмма, более тонкой и темно-серой — средняя по набору из 30 реализаций, тонкими линиями — границы разброса значений вариограмм для реализаций.

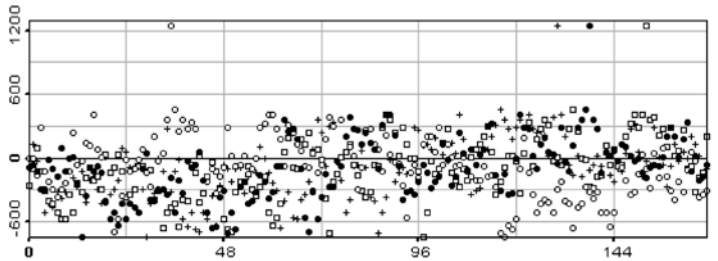


Рис. 10.12. Примеры реализаций невязок, полученных с использованием моделирования отжига

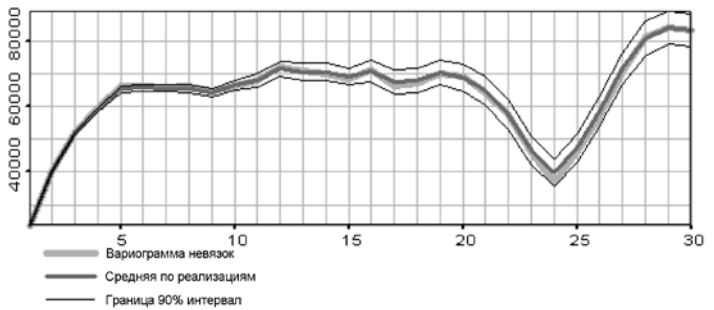


Рис. 10.13. Разброс вариограмм реализаций невязок

Окончательный результат прогноза с доверительными интервалами приведен на рис. 10.13. Это сумма прогноза, полученного ИНС, и среднего по набору реализаций для каждого момента времени. Доверительные 90%-ные интервалы получены как 2σ, где σ — корень из вариации по набору реализаций в каждый момент.

Качество прогнозной оценки и доверительных интервалов видно на рис. 10.14, где приведено и реальное значение электропотребления.

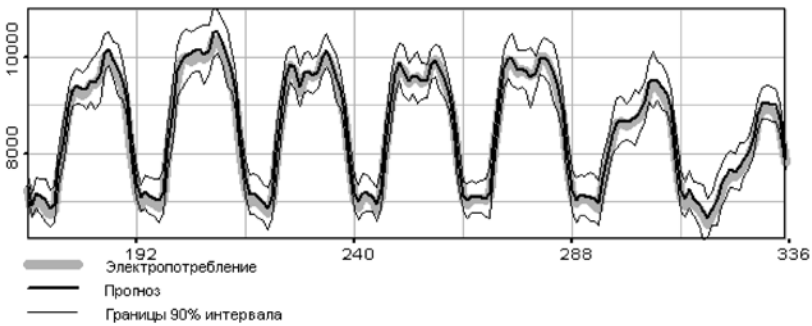


Рис. 10.14. Прогноз электропотребления гибридной моделью с доверительными интервалами

Литература

Atkinson P. M., Lloyd C. D. Mapping precipitation in Switzerland with ordinary and indicator kriging // The J. of Geographic Information and Decision Analysis. — 1998. — Vol. 2, N 2.

Bryan B. A., Adams J. M. Three-Dimensional Neurointerpolation of Annual Mean Precipitation and Temperature Surfaces for China // Geographical Analysis. — 2002. — Vol. 34, N 2. — P. 93—111.

Cortez L. P., Sousa A. J., Durao F. O. Mineral resources estimation using neural networks and geostatistical techniques // APCOM'98 Computer applications in the minerals industries: International symposium N 27, London, ROYAUME-UNI / Centro de Valorização de Recursos Minerais (CVRM), Portugal. — [S. l.], 1998. — P. 305—314.

Demyanov V., Kanevski M., Savelieva E. et al. Neural Network Residual Stochastic Cosimulation for Environmental Data Analysis // Proceedings

of the Second ICSC Symposium on Neural Computation (NC'2000), May 2000, Berlin, Germany. — [S. l.], 2000a. — P. 647—653.

Demyanov V., Serre M, Christakos G. et al. Neural Network residual BME analysis of Chernobyl fallout // Proc. GeoEnv III — 3rd European Conference on Geostatistics for Environmental Applications, Avignon, France. — [S. l.], 2000б.

Demyanov V., Soltani S., Kanevski M. et al. Wavelet analysis residual kriging vs. neural network residual kriging // Stochastic Environmental Research and Risk Assessment. — 2001. — Vol. 15, Iss. 1. — P. 18—32.

Haykin S. Neural Networks: A Comprehensive Foundation Prentice Hall. — [S. l.], 1998. — 842 p.

Kanevsky M., Arutyunyan R., Bolshov L. et al. Artificial neural networks and spatial estimations of Chernobyl fallout // Geoinformatics. — 1996a. — Vol. 7, N 1—2. — P. 5—11.

Kanevsky M., Arutyunyan R., Bolshov L. et al. Chernobyl Fallouts: Review of Advanced Spatial Data Analysis // geoENV I — Geostatistics for Environmental Applications / Ed. A. Soares, J. Gomez-Hernandes, R. Froidvaux. — [S. l.]: Kluwer Academic Publ., 1997a. — P. 389—400.

Kanevski M., Demyanov V., Maignan M. Mapping of Soil Contamination by Using Artificial Neural Networks and Multivariate Geostatistics // Artificial Neural Networks ICANN'97. 7th International Conference, Lausanne, Switzerland, October 1997: Proceedings / W. Gerstner, A. Germond, M. Hasler, J.-D. Nicould (eds.). — [S. l.]: Springer, 1997б. — P. 1125. — (Lecture Notes in Computer Science).

Kanevski M., Demyanov V., Chernov S. et al. Neural Network Residual Kriging Application For Climatic Data // The J. of Geographic Information and Decision Analysis. — 1998. — Vol. 2, N 2.

Kanevski M., Demyanov V., Pozdnukhov A. et al. Advanced Geostatistical and Machine-Learning Models for Spatial Data Analysis of Radioactively Contaminated Regions // Special Iss. of J. of Environmental Science and Pollution Research. — 2003. — Vol. 1. — P. 137—149.

Savelieva E., Kravetskiy A., Chernov S. et al. Application of MLP and stochastic simulations for electricity load forecasting in Russia // Proceeding of 8th European Symposium on Artificial Neural Networks ESANN'2000, Belgium. — [S. l.], 2000. — P. 413—418.

SIC’97 Spatial Interpolation Comparison exercise 1997 // http://www.ai-geostats.org/index.php?id=45.

Kanevski M., Parkin R., Pozdnukhov A. et al. Environmental Data Mining and Modelling Based on Machine Learning Algorithms and Geostatistics // Environmental Modelling & Software. — 2004. — Vol. 19, Iss. 9. — P. 845—855.

Zhang Quan Shen, Shi JieBin, Wang Ke et al. Neural network ensemble residual kriging application for spatial variability of soil properties / Inst. of Remote Sensing and Information System Application, Zhejiang University, Hangzhou, China // Pedosphere. — 2004. — Vol. 14, N 3. — P. 289—296.

Глава 11

Современные направления развития пространственной статистики

11.1. Пространственно-временна́я геостатистика

При анализе пространственно-временны́х явлений часто крайне трудно или вовсе невозможно получить закон распределения данных на основе физических процессов, обусловливающих эти явления. Простые физические методы дают хорошую модель общего тренда, усложнение и детализация физического описания ведет к увеличению числа параметров, бо́льшая часть которых неизвестна. Таким образом, детализация физической модели не уменьшает неопределенность, а может даже увеличивать ее. Альтернативным подходом является статистическое описание пространственно-временного распределения, базирующееся на данных измерений, которые несут в себе информацию о процессе и внешних параметрах. Геостатистические оценки опираются на информацию о внутренней структуре данных, зависят от самих данных, т. е. являются адаптивными. Как уже неоднократно упоминалось в этой книге, геостатистика базируется на статистической интерпретации данных. Это, однако, не означает, что природа самого процесса является случайной.

Пространственно-временны́е данные являются реализацией случайного поля $Z(\boldsymbol{x},t)$ $\left(Z(\boldsymbol{x},t);\boldsymbol{x}\in D,t\in T\right)$, где D — пространственная область; T — временной интервал. Иногда они могут быть представлены в виде пространственно распределенных временных рядов, но могут быть неравномерно распределены и в пространственно-временном континууме $D\times T$. Для того, чтобы использовать геостатистические методы, необходимо определить пространственно-временную корреляционную структуру поля $Z(\boldsymbol{x},t)$, задаваемую всеми случайными переменными в области исследования ($D\times T$).

Для описания пространственно-временно́й корреляции значений используются те же моменты первого и второго порядков, что были описаны в Главе 4. Приведем здесь основные из них (ковариацию и вариограмму) в пространственно-временно́м виде.

Ковариация, которая зависит в случае стационарности второго порядка только от пространственного и временно́го лагов $\boldsymbol{h}$ и τ, определяется так:

$$C_z(\boldsymbol{h},\tau) = E\Big[\big\{Z(\boldsymbol{x}+\boldsymbol{h},t+\tau)-m(\boldsymbol{x}+\boldsymbol{h},t+\tau)\big\}\big\{Z(\boldsymbol{x},t)-m(\boldsymbol{x},t)\big\}\Big], \quad (11.1)$$

где $m(\boldsymbol{x}, t)$ — среднее значение случайного поля Z в пространственно-временно́й точке $(\boldsymbol{x}, t)$. Когда среднее $m = E\big[Z(\boldsymbol{x},t)\big]$ постоянно по пространству и во времени, формула (11.1) преобразуется в

$$C_Z(\boldsymbol{h},\tau) = E\big[Z(\boldsymbol{x},t)Z(\boldsymbol{x}+\boldsymbol{h},t+\tau)\big]-m^2, \quad (11.2)$$

где $C_Z(\mathbf{0}, 0)$ равняется вариации σ_Z^2 по определению. Пространственно-временна́я ковариационная функция C_Z должна обладать теми же свойствами, что и чисто пространственная (см. Главу 4). Только некоторые можно немного переписать для пространственной и временно́й составляющих:

$$\lim_{\|\boldsymbol{h}\|\to\infty} C_Z(\boldsymbol{h},\tau) = \lim_{\|\tau\|\to\infty} C_Z(\boldsymbol{h},\tau) = \lim_{\|\boldsymbol{h}\|,|\tau|\to\infty} C_Z(\boldsymbol{h},\tau) = 0. \quad (11.3)$$

Если среднее предполагается постоянным, то для $N(\boldsymbol{h}, \tau)$ экспериментальных точек, разделенных вектором $\boldsymbol{h}$ и временны́м интервалом τ, пространственно-временна́я ковариационная функция определяется по формуле

$$\widehat{C}_Z(\boldsymbol{h},\tau) = \frac{1}{N(\boldsymbol{h},\tau)}\sum\big[(Z(\boldsymbol{x}+\boldsymbol{h},t+\tau)-\widehat{m})(Z(\boldsymbol{x},t)-\widehat{m})\big], \quad (11.4)$$

где $\widehat{m}$ — классическая оценка среднего по N известным значениям пространственно-временно́й функции $Z(\boldsymbol{x}, t)$:

$$\widehat{m} = \frac{1}{N(\boldsymbol{h},\tau)}\sum_{i=1}^{N} Z(\boldsymbol{x},t).$$

Как видно из (11.4), пространственно-временна́я ковариационная функция может быть вычислена для данных, расположенных на нерегулярной пространственно-временно́й сетке. Поэтому нет необходимости, например, иметь измерения в одной и той же пространственной точке в различные

моменты времени. Однако оценка ковариационной функции, определяемая (11.4), может оказаться смещенной вследствие того факта, что мы используем оценку неизвестного нам среднего $\hat{m}$ вместо неизвестного истинного значения.

Как и в пространственном случае, вариограмма дает возможность избежать оценки среднего, перейдя к приращениям:

$$\gamma_Z(\boldsymbol{h},\tau)=\frac{1}{2}\mathrm{Var}\left[Z(\boldsymbol{x}+\boldsymbol{h},\ t+\tau)-Z(\boldsymbol{x},t)\right]. \qquad (11.5)$$

В предположении о стационарности второго порядка для приращений $Z(\boldsymbol{x}+\boldsymbol{h},\ t+\tau)-Z(\boldsymbol{x},t)$ случайного поля Z (внутренняя гипотеза) пространственно-временна́я вариограмма (11.5) преобразуется к виду

$$\gamma_Z(\boldsymbol{h},\tau)=\frac{1}{2}E\left[\left(Z\left(\boldsymbol{x}+\boldsymbol{h},t+\tau\right)-Z(\boldsymbol{x},t)\right)^2\right] \qquad (11.6)$$

с условием $E\left[Z(\mathbf{x}+\mathbf{h},t+\tau)-Z(\mathbf{x},t)\right]=0.$ Свойства пространственно-временной вариограммы не отличаются от свойств вариограммы пространственной, которые подробно описаны в Главе 4.

Оценивается вариограмма по формуле для оценки математического ожидания:

$$\hat{\gamma}_Z(\boldsymbol{h},\tau)=\frac{1}{2N(\boldsymbol{h},\tau)}\sum\left[\left(Z(\boldsymbol{x}+\boldsymbol{h},t+\tau)-Z(\boldsymbol{x},t)\right)^2\right]. \qquad (11.7)$$

Как и в случае ковариационной функции, пространственно-временну́ю вариограмму можно оценить, даже если данные расположены на нерегулярной пространственно-временно́й сетке.

Основной проблемой при моделировании пространственно-временно́й корреляции является необходимость определения метрики на пространственно-временно́м континууме.

В различное время были предложены разнообразные теоретические модели пространственно-временны́х ковариационных функций и вариограмм, позволяющие объединять пространственные и временны́е координаты. Одним из наиболее подробных обзоров соответствующих геостатистических моделей был обзор П. Кириакидиса и А. Жорнеля [Kyriakidis, Journel, 1999]. Согласно этому обзору модели пространственно-временно́й корреляционной структуры можно подразделить на два вида: предусматривающие разделение на пространственную и временну́ю компоненты и такого

разделения не предусматривающие. Ниже рассмотрены модели, имеющие в настоящее время наибольшее распространение.

Метрическая модель. Одним из подходов является использование «обобщенной» переменной, моделирующей евклидову пространственно-временну́ю метрику [Dimitrakopoulos, Luo, 1994], для ковариационной функции

$$C_Z(\boldsymbol{h}, \tau) = C\left(a^2 |\boldsymbol{h}|^2 + b^2\tau^2\right), \tag{11.8}$$

где a, b — действительные коэффициенты. Следует отметить, что модель (11.8) предполагает одинаковый тип модели для пространственной и временной ковариационных функций с возможными различиями только в радиусе корреляции. На практике эта модель, несмотря на кажущуюся простоту, используется редко.

Линейная модель. Предполагает разделение пространственно-временно́й ковариации на пространственную и временну́ю компоненты. Общая модель пространственно-временно́й ковариации представляет сумму пространственной и временно́й компонент:

$$C_Z(\boldsymbol{h}, \tau) = C_x(\boldsymbol{h}) + C_t(\tau). \tag{11.9}$$

Эта модель обладает существенным недостатком: при некоторых ее конфигурациях матрица ковариаций пространственно-временны́х данных может оказаться сингулярной [Rouhani, Myers, 1990]. В таком случае ковариационная функция является только положительно полуопределенной и, следовательно, не удовлетворяет требуемому условию для использования в кригинге. Это ограничивает сферу применения данной модели.

Модель произведения. Эта модель пространственно-временно́й корреляции также основана на разделении зависимости по пространству и времени [De Cesare et al., 2001, 2002]. Но в отличие от предыдущего случая (11.9) пространственно-временна́я ковариационная модель строится как произведение этих компонент:

$$C_Z(\boldsymbol{h}, \tau) = kC_x(\boldsymbol{h})C_t(\tau). \tag{11.10}$$

Пространственно-временна́я модель ковариации (11.10) может быть переписана в терминах пространственно-временно́й вариограммы:

$$\gamma_Z(\boldsymbol{h}, \tau) = k\left(C_t(0)\gamma_x(\boldsymbol{h}) + C_x(0)\gamma_t(\tau) - \gamma_x(\boldsymbol{h})\gamma_t(\tau)\right), \tag{11.11}$$

где γ_Z — пространственно-временна́я вариограмма; γ_t — временна́я компонента вариограммы; γ_x — пространственная компонента вариограммы; C_t — временна́я компонента ковариационной функции; $C_{\mathbf{x}}$ — пространственная компонента ковариационной функции; $C_Z(0,0)$ — плато (sill) пространственно-временно́й вариограммы γ_Z; $C_{\mathbf{x}}(0)$ — плато пространственной компоненты вариограммы γ_x; $C_t(0)$ — плато временной компоненты вариограммы γ_t.

Параметр k логично определяется из уравнения (11.11):

$$k = \frac{C_Z(0,0)}{C_{\mathbf{x}}(0)C_t(0)}, \tag{11.12}$$

чтобы при нулевых расстояниях по пространству ($|h| = 0$) и/или времени ($\tau = 0$) оставалась только нужная компонента.

Если в выражении (11.10) C_x является положительно-определенной функцией в пространстве действительных чисел размерностью d $\Re^d$, а C_t — положительно-определенной в $\Re^1$, то и модель произведения (11.10) также является положительно-определенной функцией [Cressie, 1993].

Однако класс функций (11.10) сильно ограничен, так как для любой пары пространственных точек кросс-ковариационная функция двух временны́х рядов всегда должна иметь «похожую» форму. Фактически для любых двух фиксированных пространственных векторов $\boldsymbol{h}_1$ и $\boldsymbol{h}_2$

$$C(\boldsymbol{h}_1, \tau) \propto C(\boldsymbol{h}_2, \tau). \tag{11.13}$$

Такой же результат должен быть и для любой пары временны́х точек кросс-ковариационной функции двух пространственных процессов [De Cesare et al., 2001].

Модель произведения-суммы. Линейную модель и модель произведения можно легко свести вместе:

$$C_Z(\boldsymbol{h}, \tau) = k_1 C_{\mathbf{x}}(\boldsymbol{h}) C_t(\tau) + k_2 C_{\mathbf{x}}(\boldsymbol{h}) + k_3 C_t(\tau). \tag{11.14}$$

Чтобы модель произведения-суммы (11.14) была применима, C_x и C_t должны быть положительно-определенными функциями. Кроме того, коэффициенты k_2 и k_3 должны быть неотрицательны ($k_2 \geq 0$, $k_3 \geq 0$), в то время как k_1 должен быть строго положительным ($k_1 > 0$) [De Cesare et al., 2001, 2002].

Модель произведения-суммы (11.14) может быть легко переписана в терминах пространственно-временно́й вариограммы:

$$\gamma_Z(\boldsymbol{h},\tau)=\left[k_1C_{\boldsymbol{x}}(0)+k_3\right]\gamma_t(\tau)+\left[k_1C_t(0)+k_2\right]\gamma_{\boldsymbol{x}}(\boldsymbol{h})-k_1\gamma_{\boldsymbol{x}}(\boldsymbol{h})\gamma_t(\tau). \quad (11.15)$$

При переходе от ковариационной формы (11.14) к вариограммной (11.15) неявно получается следующее условие:

$$k_1C_{\boldsymbol{x}}(0)C_t(0)+k_2C_{\boldsymbol{x}}(0)+k_3C_t(0)=C_Z(0,0). \quad (11.16)$$

Кроме того, из (11.15) получаются условия для пространственной и временно́й компонент вариограммы:

$$\begin{aligned}\gamma_Z(\boldsymbol{h},0)&=\left[k_2+k_1C_t(0)\right]\gamma_{\boldsymbol{x}}(\boldsymbol{h}),\\ \gamma_Z(0,\tau)&=\left[k_3+k_1C_{\boldsymbol{x}}(0)\right]\gamma_t(\tau).\end{aligned} \quad (11.17)$$

Чтобы определить коэффициенты k_1, k_2, k_3, необходимы три уравнения. Два их них получаются из условий (11.17):

$$\begin{aligned}k_2+k_1C_t(0)&=1,\\ k_3+k_1C_{\boldsymbol{x}}(0)&=1.\end{aligned} \quad (11.18)$$

Третье получаем, используя условие (11.16). Таким образом, получены формулы для вычисления всех параметров k_1, k_2 и k_3:

$$\begin{aligned}k_1&=\frac{C_{\boldsymbol{x}}(0)+C_t(0)-C_Z(0,0)}{C_{\boldsymbol{x}}(0)C_t(0)},\\ k_2&=\frac{C_Z(0,0)-C_t(0)}{C_{\boldsymbol{x}}(0)},\\ k_3&=\frac{C_Z(0,0)-C_{\boldsymbol{x}}(0)}{C_t(0)}.\end{aligned} \quad (11.19)$$

При моделировании чисто пространственной и чисто временно́й вариограмм необходимо следить, чтобы значения плато $C_Z(0,0)$, $C_{\boldsymbol{x}}(0)$, $C_t(0)$ были выбраны таким образом, что коэффициенты k_1, k_2, k_3 в (11.15) оставались положительными.

Основное удобство использования моделей произведения (11.10) и произведения-суммы (11.14) заключается в том, что они полностью опре-

деляются чисто временно́й γ_t и чисто пространственной γ_x компонентами вариограммы.

С другой стороны, ограничения (11.18) на ковариационную модель произведения-суммы (11.14) налагают на нее форму симметрии, т. е. симметрии между влиянием пространственной и временно́й корреляционных компонент.

Неразделимая модель. Другой подход к моделированию пространственно-временной корреляции позволяет получить классы неразделимых пространственно-временных стационарных ковариационных функций. Он был предложен. Кресси и Х. Хуаном [Cressie, Huang, 1999]. Этот подход основан на использовании частотного представления ковариационной функции:

$$H(\omega, \tau) = (2\pi)^{-d} \int e^{-ih^T\omega} C_Z(\boldsymbol{h}, \tau)\, d\boldsymbol{h}, \qquad (11.20)$$

где проводится частичное разделение на компоненты. Частотное представление ковариационной функции $H(\omega, \tau)$ имеет вид произведения

$$H(\omega, \tau) = \rho(\omega, \tau) K(\omega).$$

На компоненты произведения наложены два условия:

- для любого $\omega \in \Re^d$, $\rho(\omega, \cdot)$ является непрерывной автокорреляционной функцией;
- $K(\omega)$ — положительная функция с ограниченным интегралом $\int K(\omega) d\omega < \infty$.

Неразделимую модель можно модифицировать так, чтобы учитывать также анизотропию данных, в частности анизотропию в пространственных координатах [Fernández-Casal et al., 2001]. Отсутствие достаточной проработки в плане практического применения этого подхода сильно ограничивает его привлекательность.

Пространственно-временно́й кригинг. Когда модель пространственно-временно́й корреляции построена, проблем по обобщению кригинга (или любого другого геостатистического метода) на пространственно-временной случай нет. Для оценки, например обычным кригингом, используется линейная комбинация исходных измерений:

$$Z^*(\boldsymbol{x}, t) = \sum_{i=1}^{n(x)} \sum_{j=1}^{n(t)} \lambda_{ij}(\boldsymbol{x}, t) Z(\boldsymbol{x}_i, t_j),$$

где $\lambda_{ij}(\boldsymbol{x}, t)$ — веса, присваиваемые данным $Z(\mathbf{x}_i, t_j)$, которые, в свою очередь, являются реализациями пространственно-временно́й переменной Z. Количество данных $n(x)$ и $n(t)$, используемых для оценивания, как и их веса, могут меняться в зависимости от точки оценивания $(\mathbf{x}, t)$.

Для пространственно-временно́го случая может использоваться любая из описанных в Главе 5 моделей кригинга. Все условия и выводы формул без проблем переносятся в пространственно-временно́й континуум. Таким образом, основной сложностью при введении временно́й компоненты является моделирование пространственно-временно́й корреляции данных, а именно понимания связи между пространственной и временно́й зависимостями.

Пример использования пространственно-временно́го кригинга

В этом примере рассмотрено моделирование пространственно-временно́й динамики уровня грунтовых вод. Для моделирования использовалась информация из 31 скважины за период с 1972 г. Более подробно данные описаны в [Нужный и др., 2007].

При моделировании пространственно-временно́й корреляции использовался подход, разделяющий пространственную и временну́ю компоненты. Для каждой из компонент были проведены оценка и моделирование. Пространственная компонента рассматривалась без учета анизотропии. Результаты моделирования отдельных компонент представлены на рис. 11.1 (пространственная) и 11.2 (временна́я). Параметры моделей компонент собраны в табл. 11.1.

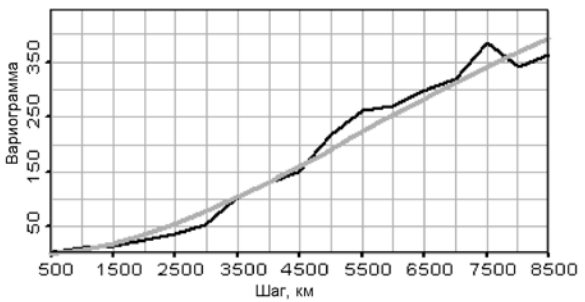


Рис. 11.1. Экспериментальная вариограмма (черная) и ее модель (серая) для пространственной компоненты пространственно-временны́х данных по уровням грунтовых вод

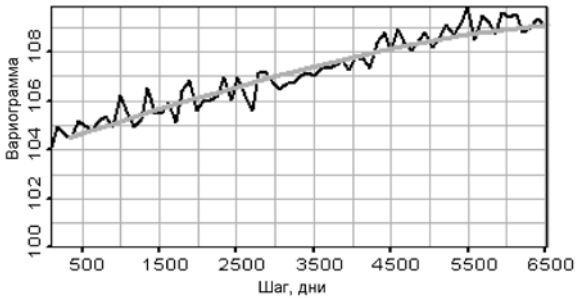


Рис. 11.2. Экспериментальная вариограмма (черная) и ее модель (серая) для временно́й компоненты пространственно-временны́х данных по уровням грунтовых вод

Таблица 11.1. Параметры моделей вариограмм пространственной и временно́й компонент

Компонента	Значение в нуле	Плато	Радиус a
Пространственная	0	556	23 100
Временна́я	104,2	5,09	7 000

Для построения пространственно-временно́й корреляционной структуры из отдельно промоделированных компонент использовалась модель произведения-суммы (11.15). Коэффициенты k_1, k_2, k_3 определялись по формулам (11.19) с использованием параметров моделей отдельных компонент (см. табл. 11.1). Они получились равными $2{,}5 \cdot 10^{-4}$, 0,975 и 0,86 соответственно. Графическое изображение общей модели представлено на рис. 11.3.

С использованием такой модели пространственно-временно́й кригинг был применен к некоторому набору отдельных измерений (в разных скважинах и в разное время). Этот набор не использовался в анализе из-за слабой представительности скважин — не более 10 измерений за весь период. Коэффициент корреляции получился очень высоким — 0,97.

Примеры пространственной оценки уровня грунтовых вод для отдельных временны́х срезов представлены на рис. 11.4.

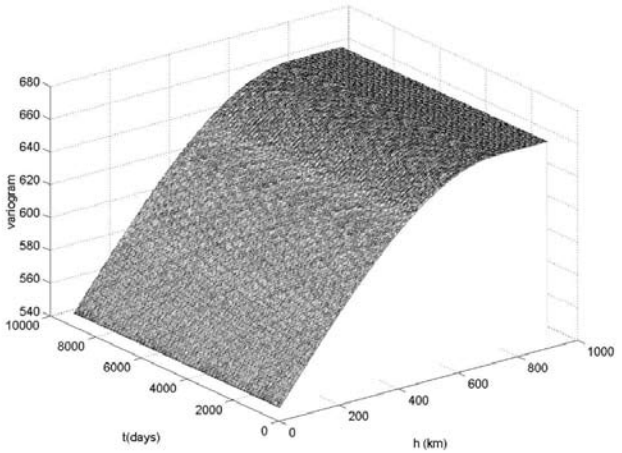


Рис. 11.3. Модель пространственно-временно́й вариограммы данных по уровням грунтовых вод

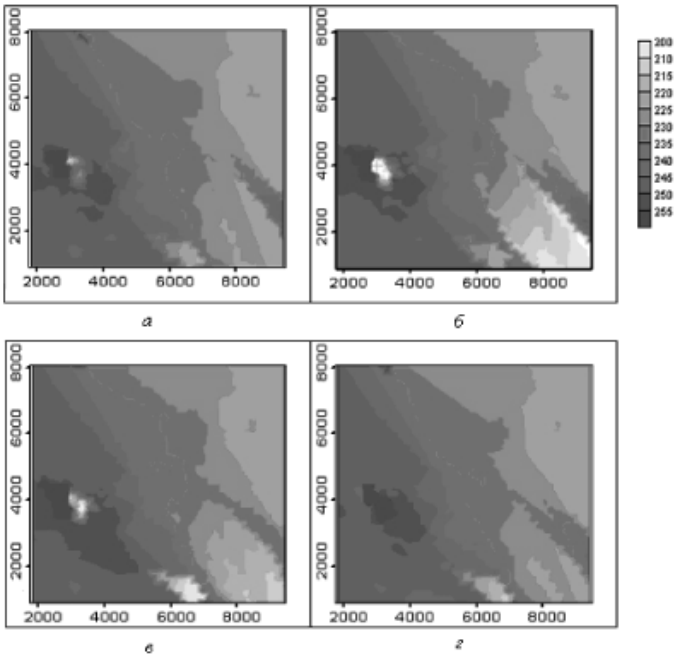


Рис. 11.4. Несколько временны́х срезов результата моделирования уровня грунтовых вод в 2005 г.:

а — январь; *б* — апрель; *в* — июль; *г* — октябрь

11.2. Стохастическое моделирование многоточечной статистики

Ряд геостатистических алгоритмов стохастического моделирования, описанных в Главе 8, базируется на вариограмме, которая отражает пространственную корреляцию данных. Вариограмма рассчитывается как вариация разницы пар значений. Пара измерений, расположенных на близком расстоянии, имеет более близкие значения, чем пара измерений, более удаленных друг от друга. В результате использования такой двухточечной статистики (вариограммы) отсутствует возможность моделировать сложные связные структуры, например протяженные флювиальные пласты породы, речные структуры. Ограниченные возможности моделирования на основе вариограммы кратко обсуждались в Разделе 4.8. Объектный подход к стохастическому моделированию позволяет преодолеть эти ограничения и моделировать сложные связные структуры на основе объектов определенной геометрической формы. Таким образом, в объектном подходе пространственная корреляция жестко привязана к выбору размера и формы объектов. Однако разнообразие форм и размеров объектов не является такой унифицированной мерой пространственной корреляции, как вариограмма. Объектное моделирование также сопряжено с рядом сложностей, которые уже обсуждались в Главе 8.

В начале 1990-х гг. был предложен новый подход к моделированию на основе тренировочного образа [Guardiano, Srivastava, 1993]. Однако в то время вычислительные возможности не позволили его реализовать на практике, и только в начале 2000-х гг. был предложен первый действующий алгоритм стохастического моделирования на основе многоточечной статистики [Strebelle, 2000, 2002].

Тренировочный образ является основой многоточечной статистики, он характеризует совместную связь множества точек, а не только пар с определенной пространственной ориентацией. Тренировочный образ представляет собой концепцию глобальной структуры данных (по аналогии с гистограммой или вариограммой), которая адаптируется к имеющимся локальным данным. При моделировании на основе многоточечной статистики удается воспроизводить глобальную структуру тренировочного образа, которая в то же время удовлетворяет локальной информации, имеющейся в точках измерений.

Многоточечное стохастическое моделирование совмещает в себе свойства объектного и пиксельного моделирования. Так, тренировочный образ может точно описывать достаточно сложные структуры различных геометрических форм, как и объектный подход. В то же время значение каждой ячейки моделируется индивидуально, как в других пиксельных алгоритмах (последовательном гауссовом, индикаторном и пр.). Как и в упомянутых методах, стохастическая природа моделирования проявляется в выборке значения из локальной функции плотности вероятности, построенной в каждой точке оценивания. Функция плотности вероятности строится на основе информации, полученной при обработке тренировочного образа, в отличие от других методов, основанных на вариограммном оценивании. При построении локальной плотности вероятности производится поиск конфигурации данных в локальной окрестности точки оценивания (data event) в тренировочном образе. На основе полученных вариантов значений строится функция для выборки.

Принцип последовательного моделирования используется здесь аналогично другим алгоритмам (см. Раздел 8.2), а именно каждая вновь смоделированная точка добавляется к набору данных для использования при моделировании последующих точек. Обработка тренировочного образа позволяет получить условную плотность распределения вероятности для каждой конфигурации пиксельные данных (data event).

Для иллюстрации рассмотрим примитивный пример тренировочного образа — вертикальные линии в квадрате 6×6 (рис. 11.5). Белые и черные ячейки распределены в равной пропорции (50% на 50%). Пошаговый алгоритм моделирования на сетке 2×2 приведен на рис. 11.6.

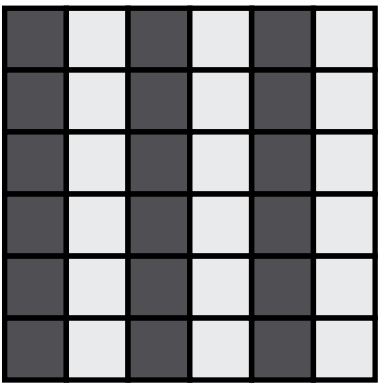


Рис. 11.5. Тренировочный образ

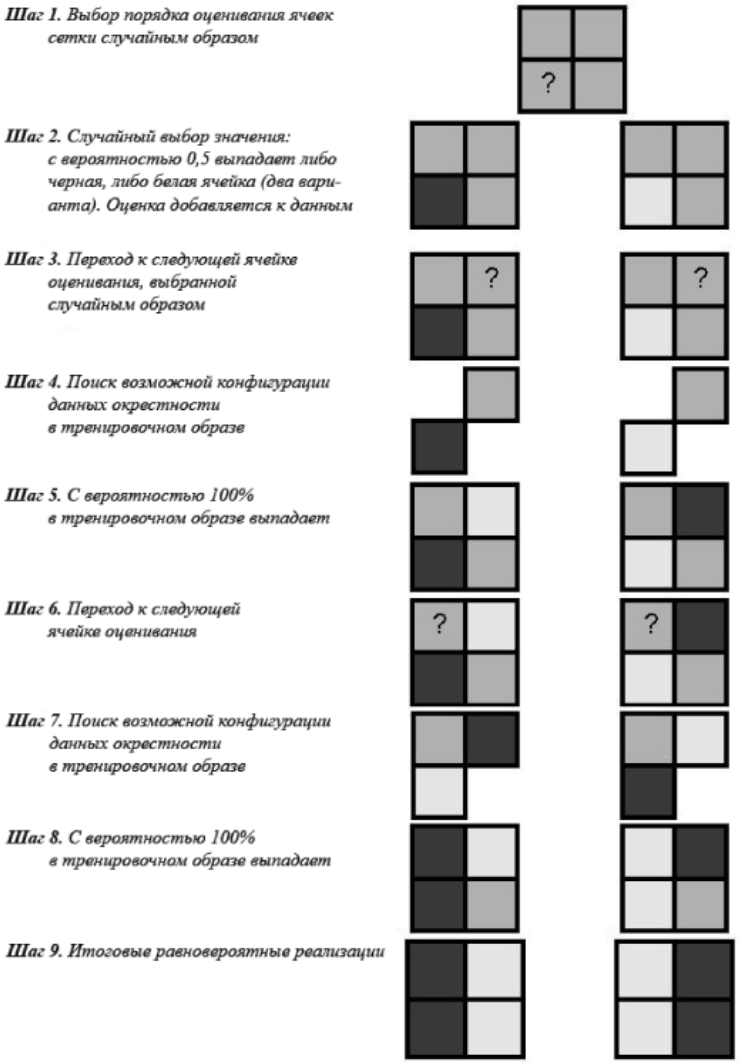


Рис. 11.6. Иллюстрация алгоритма стохастического моделирования с использованием многоточечной статистики на основе тренировочного образа на рис. 11.5

Если рассматривать менее примитивную и более реалистичную конфигурацию тренировочного образа, то на его основе получаются неоднородные условные функции плотности вероятности для конфигурации окрестности данных (рис. 11.7) [Caers, 2005].

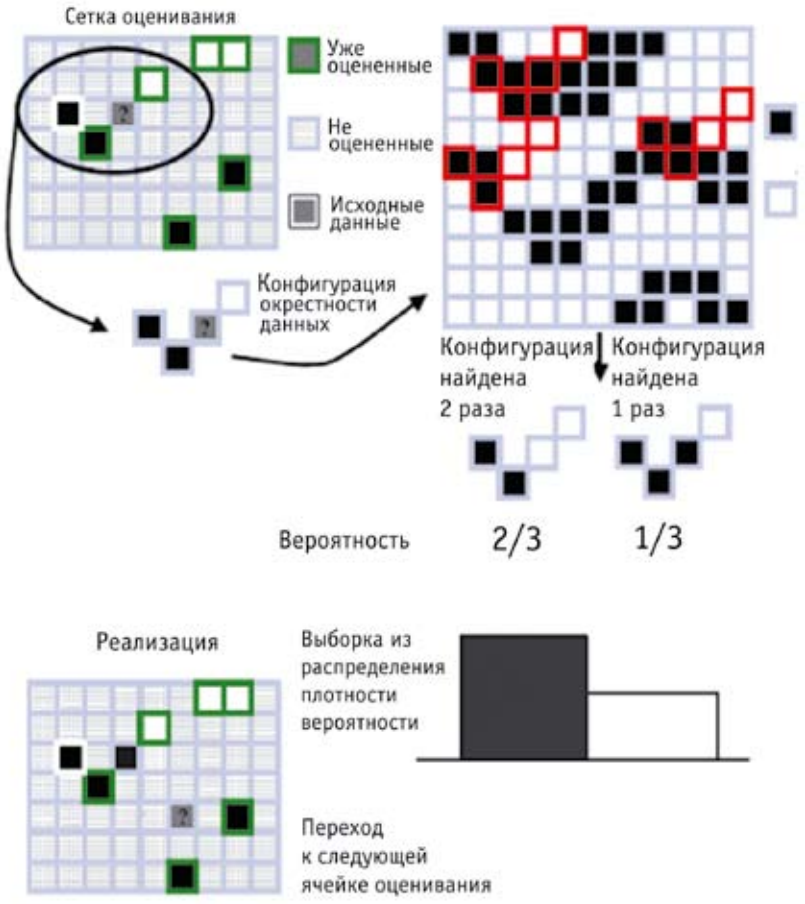
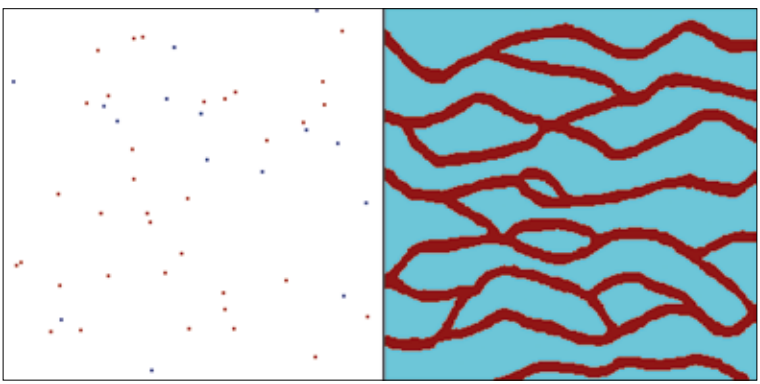


Рис. 11.7. Схема получения и выборки из условной функции плотности вероятности конфигурации окрестности данных [Caers, 2005]

Одним из ключевых вопросов моделирования на основе многоточечной статистики остается источник получения тренировочного образа. В геологии источниками тренировочных образов могут быть физические модели процессов отложений и образования речных систем, подробные описания обнажений пород, сейсмическое зондирование высокого разрешения. При использовании нескольких тренировочных образов можно получить альтернативные сценарии.

Алгоритм моделирования одного нормального уравнения (Single Normal Equation simulation — SNESIM) был предложен в [Strebelle, 2000, 2002]. Он позволяет моделировать категориальные данные. В качестве примера рассмотрим моделирование геологической структуры русел [S-GeMS]. Исходная информация — набор данных в точках измерений (рис. 11.8*а*) и тренировочный образ, описывающий характерную структуру русел, но не привязанный к конкретным данным (рис. 11.8*б*). На рис. 11.9 приведены равновероятные реализации, полученные на основе исходых данных и тренировочного образа при помощи пакета программ S-GeMS [S-GeMS].



а *б*

Рис. 11.8. Данные измерений (*а*) и тренировочный образ (*б*) для задачи моделирования залегания геологических пород

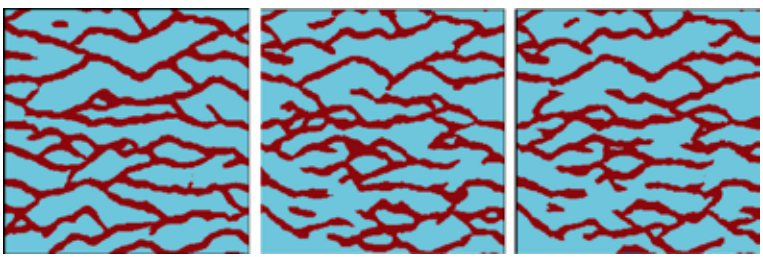


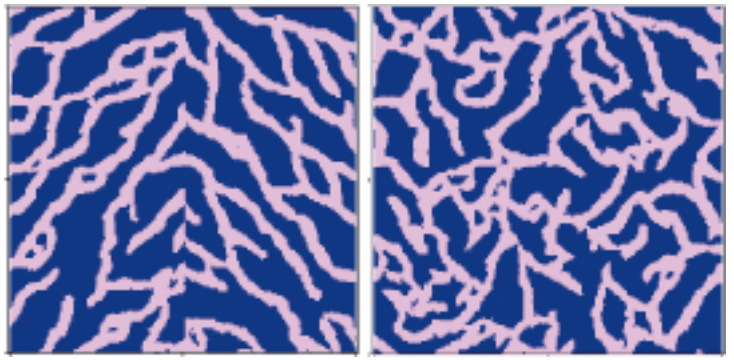
Рис. 11.9. Равновероятные реализации алгоритма SNESIM моделирования на основе многоточечной статистики

За последние годы было разработано несколько алгоритмов на основе многоточечной статистики, которые позволяют моделировать и непрерывные данные. Один из них использует фильтрацию при обработке тренировочного образа [Zhang et al., 2006]. В другом алгоритме выборка производится из

набора самих конфигураций окрестностей данных, полученных из тренировочного образа [Arpat, Caers, 2004].

Одним из ограничений подхода к моделированию на основе многоточечной статистики является проблема стационарности, которая подробно была рассмотрена в Разделе 4.10 и Главе 10. При моделировании в различных точках области оценивания используется один и тот же тренировочный образ, что предполагает стационарность пространственной корреляционной структуры. Решение проблемы учета нестационарности в многоточечном моделировании было предложено в [Strebelle, 2005].

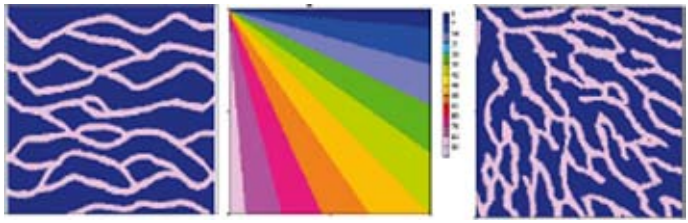
В результате использования нестационарного тренировочного образа, в котором ориентация русел зависит от местоположения (рис. 11.10), полученная реализация не отражает структуры тренировочного образа — ориентации русел перемешаны в пространстве. Во избежание такого эффекта было предложено использовать поле фактора нестационарности в качестве дополнительной локальной информации.



а *б*

Рис. 11.10. Нестационарный тренировочный образ (*а*) и стохастическая реализация на его основе (*б*)

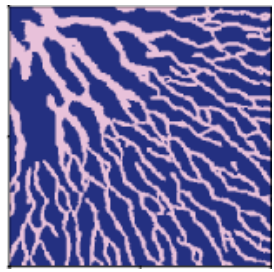
Если построить функцию изменения нестационарного фактора, например угла направления русел, на сетке оценивания (рис. 11.11*б*), то в результате учета этой информации при моделировании получается стохастическая реализация, которая отражает изменение параметров тренировочного образа в пространстве (рис. 11.11*в*). Так же можно учитывать комбинацию факторов — направление русел и их толщину. В результате в реализации можно воспроизвести структуру дельты (рис. 11.12*в*) на основе стационарного образа параллельных русел (рис. 11.12*а*).



а *б* *в*

Рис. 11.11. Учет изменения угла направления русел в пространстве: тренировочный образ (*а*), фактор изменения угла направления русла (*б*), стохастическая реализация (*в*)

а *б* *в*



г

Рис. 11.12. Тренировочный образ (*а*), фактор масштабирования толщины русла (*б*), фактор изменения угла направления русла (*в*), стохастическая реализация (*г*)

11.3. Байесовская геостатистика

Байесовский подход позволяет использовать в качестве дополнительной информации предварительные знания, сформулированные в вероятностном виде как приорные распределения. Приорные распределения совместно с данными позволяют оценивать зоны неопределенности (границы значений) исследуемой переменной. В случае полного байесовского подхода неопределенность представляется как постериорная локальная (или глобальная) функция распределения.

Здесь не представляется возможным подробно изложить все эти теории. Мы приводим только базовые понятия, а желающие могут более подробно изучить материал по англоязычным ссылкам.

Если предварительная (приорная) информация относится к знаниям о пространственном тренде, то формулируется *байесовский кригинг* [Omre, 1987]. В некотором смысле его можно считать модификацией универсального кригинга, рассмотренного в Главе 5.

Напомним, что в универсальном кригинге тренд моделируется линейной комбинацией базисных функций

$$Z(x) = f(x)^T \theta + \varepsilon(x), \quad E\{\varepsilon(x)\} = 0.$$

Оценку универсального кригинга (в векторно-матричном виде), полученную из условий несмещенности и минимизации вариации ошибки, можно записать так:

$$Z^*(x_0) = c_0^T C^{-1}(Z - F\mu) + f_0^T \hat{\theta}, \tag{11.21}$$

где $f_0 = f(x_0), \quad F = (f(x_1), f(x_2), ..., f(x_n))^T$ — вектор и матрица из базисных функций; $(C)_{ij} = C(x_i - x_j)$ и $(c_0)_i = C(x_i - x_0) \quad i, j = 1, ..., n$ — ковариационные функции. Значение $\hat{\theta}$ является оценкой неизвестного параметра θ.

Предположим теперь, что известна дополнительная информация о функции распределения неизвестного параметра θ. Как и в любом другом кригинге, ограничиваемся моментами первого и второго порядка, т. е.

$$E\{\theta\} = \mu, \quad \mathrm{Cov}\{\theta\} = \Phi.$$

В отличие от универсального кригинга в данном подходе отбрасывается условие несмещенности. Вместо него рассматривается компонента смещенности λ_0, т.е. оценка байесовского кригинга записывается как

$$Z^*(x_0) = \lambda^T Z + \lambda_0. \qquad (11.22)$$

Веса же, как и всегда, находятся минимизацией вариации ошибки оценки. Используя решение соответствующей системы уравнений, оценку (11.22) можно записать в виде

$$Z^*(x_0) = \tilde{c}_0^T \tilde{C}^{-1}\left(Z - F\mu\right) + f_0^T \mu,$$

где ковариационные члены видоизменились по сравнению с (11.21):

$$\tilde{c}_0 = c_0 + F\Phi f_0, \quad \tilde{C} = C + F\Phi F^T.$$

При использовании геостатистики неопределенность присутствует не только при моделировании тренда. Тренд вообще можно моделировать отдельно, например, как было описано в Главе 10. Важным аспектом геостатистического анализа является моделирование пространственной корреляционной структуры — вариограммы. Модель вариограммы задается набором параметров $\theta = (c_0, c, a, \nu)$, где ν относится к типу модели (см. Раздел 4.4). Рассмотрение проблем с неопределенностью параметров вариограммы можно найти в [Piltz et al., 1997].

$$\gamma(\boldsymbol{h}; \theta) = c_0 + c\left[1 - (1-\nu)e^{-a|\boldsymbol{h}|} - \nu e^{-a|\boldsymbol{h}|^2}\right]. \qquad (11.23)$$

Использование модели вариограммы (11.23) и наличие предварительной информации о функции распределения исходных данных и параметров тренда позволили провести полное байесовское моделирование [Piltz et al., 2005]. При таком рассмотрении плотность постериорной условной функции распределения выражается следующим образом:

$$p(Z_0 / Z) = \int_{\Theta}\int_{B} p(Z_0 / \beta, \theta, Z)\, p(\beta, \theta / Z)\, d\beta\, d\theta,$$

где Θ — область значений параметра θ; B — область значений параметра β.

Метод байесовской максимизации энтропии

Наиболее общим в рамках пространственной статистики является подход, разработанный Дж. Кристакосом, — метод байесовской максимизации энтропии (БМЭ). Классические геостатистические оцениватели являются част-

ным случаем этого метода. Его теория и применение для анализа различных пространственных (и пространственно-временны́х) данных изложены в книгах [Christakos, 2000, 2002] и серии статей [Christakos, 1990; Christakos, 1998; Christakos, Li, 1998; Serre, Christakos, 1999; D'Or et al., 2001; Bogaert, 2002; Serre et al., 2003; D'Or, Bogaert, 2003; Savelieva et al., 2005].

Метод БМЭ базируется на трех фундаментальных основах:

- стохастическом описании информации, выполненном в вероятностной формализации;
- теории информации Шеннона [Шеннон, 1963];
- привязке к данным измерений.

Использование этих компонентов дает возможность объединять междисциплинарные исходные данные, так как стохастическое описание приводит их к общей формализации. Теория информации дает общую формулу максимизации информации при определенных ограничениях. Учет конкретных измерений позволяет подстроить общую формулу для описания и моделирования конкретного случая.

Стохастическое описание связано с введением набора возможных реализаций и их вероятностями. Набор возможных реализаций, удовлетворяющих заданным условиям, определяет уровень знаний. Если, например, возможна только единственная реализация, то это детерминистический случай, соответствующий полному знанию.

Таким образом, при использовании стохастического описания происходит смещение от некоторого единственного состояния системы к набору возможных реализаций. Изучение единственного состояния заменяется изучением вероятностей различных возможных состояний. А выводы о дальнейшем поведении системы делаются на основе всех возможных предыдущих и последующих состояний.

При работе в рамках неполных знаний (стохастический подход) выводы должны делаться на основе функции распределения, максимизирующей информацию (энтропию) при имеющемся наборе ограничений (исходной информации). Например, если известны статистические моменты данных, то максимизирует энтропию распределение, построенное как экспонента от линейной комбинации этих моментов.

Так как стохастическое описание дается через моменты, мы в общем случае получаем искомое распределение как экспоненту, параметризованную в зависимости от набора исходной информации.

После получения общей формы функции распределения остается выбрать ее вид, удовлетворяющий данным конкретных измерений, т. е. получить условную функцию распределения.

Не вдаваясь в математическую формализацию (ее можно найти в [Christakos, 2000]), рассмотрим процедуру проведения оценки в рамках данного подхода (грубо она приведена на рис. 11.16).

- Первый шаг состоит в сборе информации. Она делится на общие знания о процессе (фундаментальные законы природы, эмпирические формулы, моменты и корреляции и т. д.), которые собирают в общую базу знаний (G-KB), и конкретные проявления процесса (это точные и неточные данные измерений — интервалы, распределения и т. п.), которые собирают в специальную базу данных (S-KB).
- Второй шаг состоит в стохастической формализации всей собранной информации.
- Далее на основе общих знаний строится функция распределения, максимизирующая энтропию (f_G).
- Этап интеграции заключается в построении условной функции распределения на основе общей функции распределения и специальной базы знаний:

$$\begin{cases} f_G(\chi_k) \\ \Xi(X_{\text{data}}) \end{cases} \Rightarrow f_X\left(\chi_k \mid X_{\text{data}}\right).$$

Рис. 11.16. Схема проведения оценки по методу БМЭ

В общем случае нет никаких ограничений на вид полученной условной функции распределения. Оцененная функция распределения дает возможность строить оценки любого типа. Это рассматривалось в Главе 5.

Пример использования БМЭ для моделирования пространственного распределения в рамках неточных данных

В приведенном выше списке работ описано много разнообразных примеров использования БМЭ. Самый интересный из них посвящен моделированию распространения эпидемии чумы в Европе в XIII в. [Christakos et al., 2003]. Здесь мы приведем анализ данных, выполненный самими авторами [Savelieva et al., 2005].

Рассматривались данные по загрязнению почвы радиоактивными изотопами ^{137}Cs, выпавшими в результате Чернобыльской аварии. Исходные данные были двух типов: точные χ_{hard} (единственное измерение) и неточные χ_{soft} (в одном населенном пункте было проведено несколько измерений). Поскольку измерения были приписаны к центру населенного пункта, они не давали возможности строить пространственные зависимости внутри него. Так как все измерения, проведенные в разное время, были пересчитаны на момент аварии, они не описывали временны́х тенденций. Такие данные можно было только использовать для описания неопределенности. Все измерения были представлены в вероятностном виде: единственные рассматривались с вероятностью 1, неточные описывались локальными функциями распределения треугольной формы (рис. 11.17). Диапазон определялся локальными максимумом и минимумом, а в качестве наиболее вероятного значения (максимума плотности вероятности) использовалась экспертная оценка, так называемые официальные данные по загрязнению.

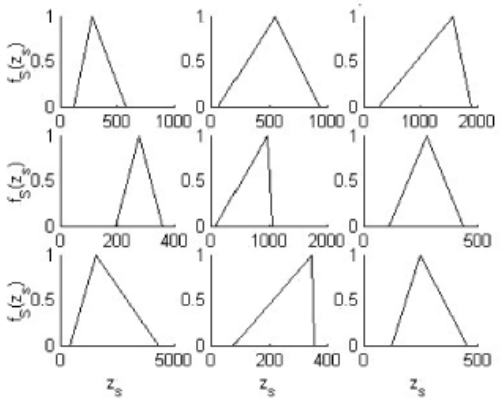


Рис. 11.17. Примеры локальных функций распределения в пунктах с неточными данными

В данном случае общие знания включали в себя тренд (локальное среднее) и модель ковариации. Вообще говоря, они были получены на основе конкретных данных, но по построению метода моменты относятся к общим знаниям. Специальная база включала набор данных, описанный выше, и набор точек, где предполагается провести оценку χ_k, — специально отобранные из исходного набора валидационные точки и точки на сетке размером 2×2 км. Таким образом, полный набор пространственных точек можно описать как $\chi_{map} = (\chi_{hard}, \chi_{soft}, \chi_k)$.

Условные функции распределения в точках оценивания можно формализовать:

$$f_K(\chi_k) = A^{-1} \int d\chi_{soft} f_S(\chi_{soft}) f_G(\chi_{map}).$$

Валидационный набор включал как точки с единственным измерением, так и точки с большим (более 20) количеством измерений. Для точек с единственным измерением это значение интерпретировалось как наиболее вероятное и сравнивалось с наиболее вероятным, оцененным в соответствии с постериорным локальным распределением БМЭ. Коэффициент корреляции для этой части валидационного набора был равен 0,92.

Для точек с большим количеством измерений можно оценить функцию распределения и сравнить ее с предсказанной БМЭ. Несколько примеров сравнения функций распределения с использованием специальных графиков (QQ-plot) представлено на рис. 11.18. График представляет собой оценки значений квантилей по набору измерений (ось X) и по оцененной локальной функции распределения (ось Y). Графики демонстрируют хорошее соответствие.

Для визуализации результатов удобно использовать какую-нибудь оценку. Самой распространенной оценкой при наличии локальной функции распределения является наиболее вероятное значение. Ее выбор обусловлен, в частности, тем, что она соответствует максимуму плотности вероятности, а в рамках методологии БМЭ оценка ориентируется на максимизацию энтропии.

Результаты картирования на регулярную сетку и несколько примеров полученных локальных функций распределения представлены на рис. 11.19.

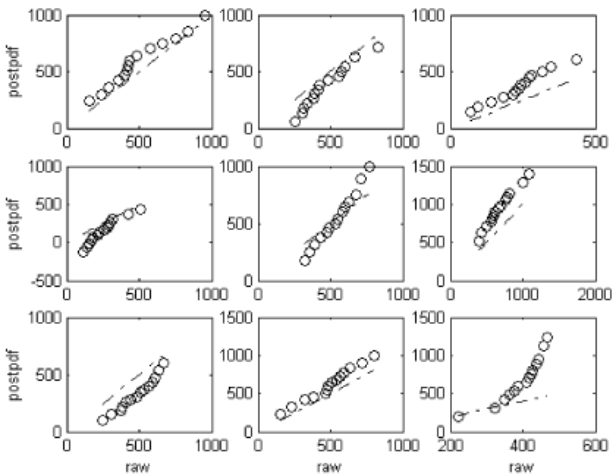


Рис. 11.18. QQ-plot для сравнения локальных функций распределения измерений и постериорных функций распределения БМЭ

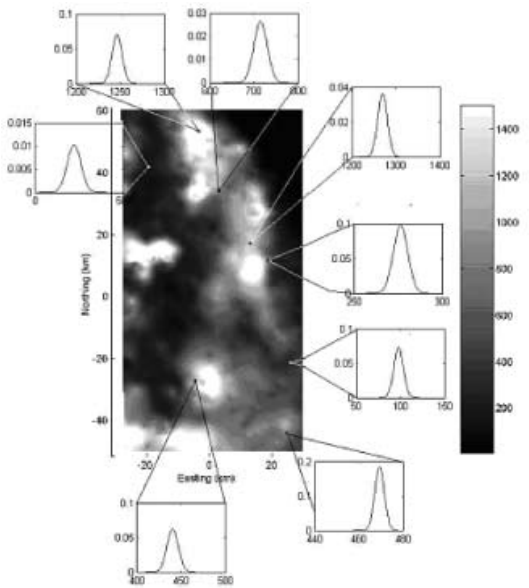


Рис. 11.19. Результат интерполяции (максимально вероятные значения) и примеры локальных функций распределения

Литература

Нужный А. С., Савельева Е. А., Линге И. И., Ястребков А. Ю. Статистический анализ изменения уровней грунтовых вод в районе ПО «Маяк» // Изв. Рос. акад. наук. Энергетика. — 2007. — № 6. — С. 73—79.

Шеннон К. Работы по теории информации и кибернетике. — М.: Изд-во иностр. лит., 1963. — 830 с.

Arpat B. G., Caers J. A. Multiple-scale, Pattern-based Approach to Sequential Simulation // Geostatistics Banff 2004 / O. Leuangthong and C. V. Deutsch (eds). — Dordrecht: Kluwer Academic Publ., 2004. — P. 225—264.

Bogaert P. Spatial prediction of categorical variables: the Bayesian Maximum Entropy approach // Stochastic Environmental Research and Risk Assessment. — 2002. — Vol. 16. — P. 425—448.

Christakos G. A Bayesian/maximum-entropy view to the spatial estimation problem // Mathematical Geology. — 1990. — Vol. 22. — P. 763—776.

Christakos G. Modern Spatiotemporal Geostatistics. — New York: Oxford Univ. Press, 2000.

Christakos G. Spatiotemporal information systems in soil and environmental sciences // Geoderma. — 1998. — Vol. 85. — P. 141— 179.

Christakos G., Bogaert P., Serre M. L. Temporal GIS. — New York: Springer-Verl., 2002.

Christakos G., Li X. Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators? // Mathematical Geology. — 1998. — Vol. 30, N 4. — P. 435—462.

Christakos G., Olea R. A., Yu H.-L., Wang L. L. Interdisciplinary Public Health Reasoning and Epidemic Modeling: Black Death Case. — [S. l.]: Springer, 2003.

Cressie N. A. C. Statistics for Spatial Data. — New York: Wiley, 1993. — 900 p.

Cressie N. A. C., Huang H. Classes of nonseparable, spatio-temporal stationary covariance functions // J. of the American Statistical Association. — 1999. — Vol. 94. — P. 1330—1340.

De Cesare L., Myers D., Posa D. Estimating and modeling space-time correlation structures // Statistics and Probability Letters. — 2001. — Vol. 51. — P. 9—14.

De Cesare L., Myers D., Posa D. FORTRAN 77 programs for space-time modeling // Computers and Geosciences. — 2002. — Vol. 28. — P. 205 —212.

Dimitrakopoulos R., Luo X. Spatiotemporal modeling: covariances and ordinary kriging systems // Geostatistics for the Next Century / R. Dimitrakopoulos (ed.). — Dordrecht: Kluwer Academic Publ., 1994. — P. 88—93.

D'Or D., Bogaert P., Christakos G. Applications of BME to soil texture mapping // Stochastic Environmental Research and Risk Assessment. — 2001. — Vol. 15. — P. 87—100.

D'Or D., Bogaert P. Continuous-valued map reconstruction with the Bayesian Maximum Entropy // Geoderma. — 2003. — Vol. 112. — P. 169—178.

Fernández-Casal R., González-Manteiga W., Febrero-Bande M. General Classes of Flexible Spatio-Temporal Stationary Variogram Models. Spatio-temporal modelling of environmental processes // Proceedings of the 1st Spanish Workshop on Spatio-temporal Modelling of Environmental Processes, Benicasim (Castellón), Spain, 28—31 October 2001 / Ed. by J. Mateu & F. Montes. — [S. l.], 2001.

Gaudard M., Karson M., Sinha E. L. D. Bayesian spatial prediction // Environment and Ecological Statistics. — 1999. — Vol. 6. — P. 147— 171.

Guardiano F., Srivastava R. M. Multivariate geostatistics: Beyond bivariate moments // Geostatistics-Troia / A. Soares, ed. — Vol. 1. — Dordrecht: Kluwer Academic, 1993. — P. 133—144.

Kyriakidis P. C., Journel A. G. Geostatistical space-time models: a review // Mathematical Geology. — 1999. — Vol. 31. — P. 651—684.

Omre Y. Bayesian kriging — merging observations and qualified guesses in kriging // Mathematical Geology. — 1987. — Vol. 19. — P. 25—39.

Piltz J., Pluch P., Spock G. Bayesian Kriging with lognormal data and uncertain variogram parameters // Geostatistics for Environmental Applications / P. Renard, H. Demougeot-Renard, R. Fridevaux (eds). — [S. l.]: Springer, 2005. — P. 51—62.

Piltz J., Schimek M. J., Spock G. Taking into account of uncertainty in spatial covariance estimation // Geostatiatica Wolongong / E. Baafi and N. Schofield (eds). — Vol. 1. — Dordrecht: Kluwer, 1997. — P. 402—413.

Rouhani S., Myers D. E. Problems in Space-Time Kriging of Hydrogeological Data // Mathematical Geology. — 1990. — Vol. 22. — P. 611—623.

Savelieva E., Demyanov V., Kanevski M. et al. BME Based Uncertainty Assessment of the Chernobyl Fallout // Geoderma. — 2005. — Vol. 128. — P. 312—324.

Serre M. L., Christakos G. Modern Geostatistics: Computational BME in the light of uncertain physical knowledge — The Equus Beds Study // Stochastic Environmental Research and Risk Assessment. — 1999. — Vol. 13. — P. 1—26.

Serre M. L., Kolovos A., Christakos G., Modis K. An application of the holistochastic human exposure methodology to naturally occurring Arsenic in Bangladesh drinking water // Risk Analysis. — 2003. — Vol. 23. — P. 515—528.

S-GeMS: The Stanford Geostatistical Modeling Software // http://sgems.sourceforge.net.

Strebelle S. Sequential simulation drawing structures from training images / Stanford Univ. — [S. l.], 2000. — 200 p. — Unpublished doctoral dissertation.

Strebelle S. Conditional simulation of complex geological structure using multiple-point statistics // Mathematical Geology. — 2002. — Vol. 34. — P. 1—22.

Strebelle S. Geostatistical Modeling Using Multiple Sources of Information: The MPS-FDM Workflow // Stanford-Heriot-Watt Forum on Reservoir Description and Modeling, Tiburon, California. — [S. l.], 2005.

Zhang T., Switzer P., Journel A. Filter-Based Classification of Training Image Patterns for Spatial Simulation // Mathematical Geology. — 2006. — Vol. 38, N 1.

Приложения

1. Математические обозначения

В этот раздел вынесены только основные обозначения, использованные в данной книге. Некоторые обозначения, используемые локально, вводятся непосредственно в тексте.

Операторы

Pr — оператор вычисления вероятности

E — оператор математического ожидания

Var — оператор вариации

$|x|$ — метрика в многомерном пространстве

Координаты

R^n, R^2 — пространство действительных чисел размерности n, 2

$x, x_t, x_j, \ldots$ — вектор координат в пространстве R^n

ξ_i — i-я координата в пространстве размерности $n \geq i$

ξ_i^j — i-я координата точки x_j в пространстве размерности $n \geq i$

t — координата времени в пространственно-временно́м континууме

Функции, реализации, оценки

$Z(x)$ — анализируемая непрерывная функция, случайная непрерывная функция

$Z(x, t)$ — случайная пространственно-временна́я функция

$Q(x)$ — случайная категориальная переменная

$Z(x_i), Z_i$ — случайная (анализируемая) функция в точке x_i

U, V — случайные функции при рассмотрении многомерного анализа (случай с двумя переменными)

$Z_\alpha(x), Z_\beta(x)$ — случайные переменные в случае многопеременной функции Z

$z_t, z(x_i)$ — реализация случайной функции в точке x_i

N — число точек, где известна функция $Z(x)$

$n, n(x)$ — число точек, использующееся при оценке функции $Z(x)$ в точке x

K — число переменных при многопеременном анализе, число отсечений (срезов) при индикаторном подходе

$Z^*, Z^*(x_0)$ — оценка исследуемой функции, оценка в точке x_0

$x_1^k, ..., x_i^k, ..., x_N^k$ — случайная последовательность точек для проведения стохастической реализации номер k в рамках последовательного принципа

$z(x_i^k)$ — значение реализации k стохастического моделирования в точке x_i

$Y(x)$ — функция, полученная из исследуемой после проведения операции (например, нормализации)

$F(x; z), F(x_1, ..., x_m; z_1, ..., z_m)$ — кумулятивная условная функция распределения (одноперменная или совместная)

S — поле, где определена случайная (анализируемая) функция

$S(x_0)$ — геометрическая поддержка измерения x_0

P_i — полигон Вороного (область влияния) точки x_i

p_i — площадь полигона Вороного (области влияния) точки x_i

p_{i0} — площадь полигона Вороного точки x_i, арендованная новой точкой x_0 при введении ее в систему полигонов Вороного

Статистические моменты

σ^2 — вариация

m — глобальное среднее или локальное среднее, постоянное по всей области

$\widehat{m}$ — классическая несмещенная оценка математического ожидания (среднего)

$m(x)$ — локальное среднее, функция пространственного тренда

$R(x), \varepsilon(x)$ — функция невязки, случайная функция после удаления детерминистического тренда

$m(\boldsymbol{h})$ — среднее по точкам, разделенным вектором $\boldsymbol{h}$

$m_{+\boldsymbol{h}}$ — среднее по точкам, являющимся концом вектора $\boldsymbol{h}$

$m_{-\boldsymbol{h}}$ — среднее по точкам, являющимся началом вектора $\boldsymbol{h}$

$\sigma^2(\boldsymbol{h})$ — вариация по точкам, разделенным вектором $\boldsymbol{h}$

m^* — среднее оценки исследуемой функции

$\sigma^2_{+\boldsymbol{h}}$ — вариация по точкам, являющимся концом вектора $\boldsymbol{h}$

$\sigma^2_{-\boldsymbol{h}}$ — вариация по точкам, являющимся началом вектора $\boldsymbol{h}$

σ^2_E — вариация ошибки

V_Z, V_Y — вариационно-ковариационные матрицы многомерных случайных переменных Z и Y

Меры пространственной корреляции и связанные с ними параметры

$\boldsymbol{h}$ — вектор, задающий пространственную ориентацию при вычислении и моделировании пространственной корреляции

h — длина вектора $\boldsymbol{h}$

τ — шаг по времени при вычислении вариограммы

$C(x, \boldsymbol{h})$ — нестационарная ковариационная функция

$C(\boldsymbol{h})$ — стационарная ковариационная функция

$C_Z(\boldsymbol{h}, \tau)$ — пространственно-временна́я ковариация

$C(0)$ — глобальная вариация исследуемой функции

$C_Z(\cdot)$ — ковариационная функция функции Z (при необходимости конкретизации в тексте)

C_{ij} — ковариация для вектора, соответствующего вектору, разделяющему точки x_i и x_j

C_{i0} — ковариация для вектора, соответствующего вектору, разделяющему точки x_i и x_0 (точка оценки)

$C_{\alpha\beta}$ — кросс-ковариация переменных Z_α и Z_β

$\ddot{\dot{C}}$ — функция блочной ковариации

$\gamma(x, \boldsymbol{h})$ — нестационарная вариограмма

$\gamma(\boldsymbol{h})$ — стационарная вариограмма

γ_{ij} — вариограмма для вектора, соответствующего вектору, разделяющему точки x_i и x_j

γ_{i0} — вариограмма для вектора, соответствующего вектору, разделяющему точки x_i и x_0 (точка оценки)

$\gamma_{\alpha\beta}$ — кросс-вариограмма переменных Z_α и Z_β

$g_{\alpha\beta}$ — псевдокросс-вариограмма переменных Z_α и Z_β

$C_x(\boldsymbol{h})$ — пространственная компонента пространственно-временно́й ковариации

$C_t(\tau)$ — временная компонента пространственно-временно́й ковариации

$\gamma_x(\boldsymbol{h})$ — пространственная компонента пространственно-временно́й вариограммы

$\gamma_t(\tau)$ — временна́я компонента пространственно-временно́й вариограммы

$M(\boldsymbol{h})$ — мадограмма

$R(\boldsymbol{h})$ — родограмма

$\rho(\boldsymbol{h})$ — дрейф

Δh — допуск по расстоянию

$\Delta\varphi$ — допуск по направлению

b_w — ширина полосы по направлению при больших расстояниях

c_0 — параметр модели вариограммы «самородок»

c — параметр модели вариограммы «плато»

$a\left(a_{\|}, a_{\perp}\right)$ — параметр модели вариограммы (эффективный радиус корреляции)

$w(i)$ — весовые коэффициенты при гнездовом моделировании вариограммы

γ^* — оценка вариограммы

λ — набор параметров модели (оценивателя)

$\gamma(\boldsymbol{h}, \lambda)$ — значение теоретической модели вариограммы с набором параметров λ

Интерполяции

λ_I — весовой коэффициент *i-ŭ* точки, использующейся в линейном интерполяторе

$L(x)$ — лагранжиан

$\mu(x)$ — множитель Лагранжа при минимизации с ограничением

R — штрафной член при вычислении ошибки интерполяции

RMSE — корень из среднеквадратичной ошибки интерполяции

χ^2 — интегральная ошибка интерполяции

Eff — коэффициент эффективности

ρ — коэффициент корреляции между оценкой и известными реальными значениями

D — область поиска в детерминистических интерполяторах

ν — параметр степени

δ — сглаживающий параметр в детерминистических методах

P_n — полином степени n

σ^2_{SK}, σ^2_{OK}, σ^2_{UK} — вариация простого, обычного, универсального кригинга

$f_k(x)$ — базисная функция при моделировании пространственного тренда

$F = \left(f(x_1), ..., f(x_N)\right)$ — матрица базисных функций тренда, определенная для точек измерений

$B(h)$ — ядерные базисные функции

Индикаторный подход и стохастическое моделирование

z_k — значение отсечения (среза) при индикаторном преобразовании непрерывной функции

$I(x; z_k)$ — индикаторное преобразование непрерывной функции $Z(x)$ со срезом (отсечением) z_k

$I(x; c)$ — индикаторное преобразование категориальной функции $Q(x)$ для возможного значения c

$K_I(\boldsymbol{h}, z_k)$ — нецентральная индикаторная ковариация для среза z_k

$C_I(\boldsymbol{h}, z_k)$ — центральная индикаторная ковариация для среза z_k

$\gamma_I(\boldsymbol{h}, z_k)$ — индикаторная вариограмма для среза z_k

λ_{ki} — весовые коэффициенты индикаторного кригинга (k — номер среза, i — номер точки измерений)

$i^*(\cdot)$ — оценка для индикатора, полученная кригингом

F^*, F^{**} — оценки локальной кумулятивной функции распределения

F_U^* — оценки локальной кумулятивной функции распределения при коррекции проходом вверх

F_D^* — оценки локальной кумулятивной функции распределения при коррекции проходом вниз

p_k — среднее для индикатора по срезу z_k

P^*, P^{**} — оценка вероятности класса

$S_g(\omega)$ — функция спектральной плотности ковариационной функции

Φ — оператор преобразования случайной функции к случайной функции с нормальным распределением

$N(m, \sigma^2)$ — нормальное распределение с параметрами m, характеризующим среднее (параметр места), и σ^2, характеризующим вариацию (параметр разброса)

Y^*_{SK} — оценка функции Y с помощью простого кригинга

O — целевая функция при моделировании отжига

$O(i)$ — составная часть целевой функции при моделировании отжига, относящаяся к воспроизводимому статистическому параметру i

$O_{\text{old}}, O_{\text{new}}$ — значения целевой функции непосредственно до и после возмущения

T — температура при моделировании отжига, область значений времени при пространственно-временно́м моделировании

$\omega(i)$ — весовые коэффициенты составных частей целевой функции при моделировании отжига

$z^{(i)}(x_j)$ — значение в точке x_j на i-м шаге итераций при определенной температуре

2. Некоторые определения статистических понятий

Понятие случайной величины активно используется в геостатистике. Приведем наиболее часто встречающиеся статистические формулы, связанные со случайными переменными.

Функция распределения вероятности (кумулятивная функция распределения) определяется для непрерывной случайной величины Z как функция непрерывного переменного, она означает вероятность того, что значение переменной меньше или равно z:

$$F_Z(z) = \Pr\{Z \le z\}.$$

Функция распределения вероятности — монотонно-неубывающая от z, кроме того, она является ограниченной: $0 \le F_Z(z) \le 1$. Функцию распределения вероятности можно представить как интеграл плотности вероятности:

$$F_Z(z) = \int_{-\infty}^{z} f(x)dx.$$

В случае пространственно распределенных данных случайная величина является функцией от координаты. При этом может рассматриваться кумулятивная функция распределения вероятности по выборке (ее называют *глобальной*) или функция распределения вероятности для конкретной точки (*локальная*).

Условная функция распределения вероятности означает условную вероятность.

Нормальное (гауссово) распределение вероятности используется наиболее часто. Оно удобно простотой и множеством доказанных теорем, относящихся к переменным, удовлетворяющим этому распределению. Нормальное распределение определяется формулой

$$F(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(\frac{(x-m)^2}{\sigma^2} \right).$$

Данные, соответствующие *логнормальному распределению вероятности*, характеризуются тем, что после нелинейного логарифмического преобразования $y = \log(x)$ удовлетворяют нормальному (гауссову) распределению.

Меры, характеризующие функцию распределения. *Медиана* определяется как серединное значение, т. е. вероятность быть больше или меньше него для значений из некоторого набора одинакова. Иными словами, это такое x, что $F(x) = 1/2$, или $x = F^{-1}(1/2)$.

Верхний и *нижний квартили* (upper and lower quartile) — значения, соответствующие четверти наибольших Q_1 и четверти наименьших Q_3 значений. Вместе с медианой они делят все множество данных на четыре части с равными вероятностями попадания в них: $Q_1 = F^{-1}(1/4)$, $Q_3 = F^{-1}(3/4)$.

Разность между верхним и нижним квартилями может характеризовать разброс значений в наборе. Основное преимущество такой характеристики в том, что она не подвержена влиянию беспорядочных высоких значений.

Перцентиль (*процент*) — значение переменной, соответствующее процентной доле ранжированного распределения (сотыми долями). Перцентиль p ($0 < p \le 1$) — это значение x, вероятность быть ниже которого равна $p/x = F^{-1}(p)$.

Разбиение на *квантили* — деление множества данных на части с равными вероятностями попасть в каждую из них.

Меры неопределенности, определяемые по функции распределения. Вероятность попадания в интервал $[A, B]$ определяется через разность значений функции распределения:

$$\Pr\{x \in [A, B]\} = F(B) - F(A).$$

Доверительный интервал — интервал значений вокруг наиболее вероятного значения x_{mp}, попадание в который лимитируется определенным процентом (чаще всего рассматривается 95%-ный доверительный интервал). В общем случае 95%-ный доверительный интервал (несимметричный) $[z_{mp} - a, z_{mp} + b]$ задается так:

$$\Pr\{x \in [z_{mp} - a, z_{mp} + b]\} = 0,95.$$

В частном случае гауссова распределения 95%-ный доверительный интервал симметричен и определяется параметром σ ($a = b = 2\sigma$).

Наиболее вероятное значение соответствует максимуму плотности функции распределения (производной от функции распределения).

Для вычисления медианы по набору данных не обязательно оценивать функцию распределения, можно воспользоваться формулой

$$M = \begin{cases} Z\left(\boldsymbol{x}_{\frac{n+1}{2}}\right), n \text{ нечетное}, \\ 0,5\left[Z\left(\boldsymbol{x}_{\frac{n}{2}}\right) + Z\left(\boldsymbol{x}_{\frac{n}{2}+1}\right)\right], n \text{ четное}. \end{cases}$$

По аналогичной схеме можно оценивать и квантили, последовательно разбивая число данных.

В теории вероятностей вводятся *статистические моменты* порядка k (m_k):

$$m_k = E\{\xi^k\} = \int \boldsymbol{x}^k f(\boldsymbol{x}) d\boldsymbol{x}$$

и *центральные моменты* порядка k ($m_k^{(0)}$):

$$m_k^{(0)} = E\{(\xi - m_1)^k\} = \int (\boldsymbol{x} - m_1)^k f(\boldsymbol{x}) d\boldsymbol{x}.$$

Наиболее часто используемые моменты и функции от них:

- *Среднее* — момент первого порядка:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} Z(\boldsymbol{x}_i).$$

- *Вариация* — центральный момент второго порядка:

$$m_2^{(0)} = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(Z(\boldsymbol{x}_i) - m_1\right)^2.$$

- *Стандартное отклонение* — корень из вариации:

$$\sigma = \sqrt{m_2^{(0)}}.$$

Используется для характеристики разброса значений.

- *Коэффициент симметрии* — центральный момент третьего порядка:

$$m_3^{(0)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i) - m_1\right)^3}{\sigma^3}.$$

Характеризует степень перекошенности распределения.

- *Коэффициент вариации* — отношение стандартного отклонения к среднему значению:

$$CV = \frac{\sigma}{m_1}.$$

Используется для описания асимметрии распределения аналогично коэффициенту симметрии. В основном этот показатель используется для описания распределений положительных значений и с положительным коэффициентом симметрии. Если коэффициент вариации больше единицы, это означает наличие беспорядочных больших значений.

- *Эксцесс* — центральный момент четвертого порядка:

$$m_4^{(0)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i) - m_1\right)^4}{\sigma^4} - 3.$$

Характеризует крутизну плотности функции распределения рядом с максимумом.

Меры ошибки. Для сравнения оценок с реальными значениями используются характеристики ошибок оценки. Кроме вероятностных, получаемых по функции распределения, существуют еще детерминистические характеристики.

Невязка — разность между правильным значением и оценкой:

$$Z(x_i) - Z^*(x_i).$$

Абсолютная ошибка — абсолютное значение от невязки:

$$|Z(x_i) - Z^*(x_i)|.$$

Относительная ошибка — невязка, нормированная на реальное значение,

$$\text{relative error}\,(x_i) = \frac{Z(x_i) - Z^*(x_i)}{Z(x_i)}.$$

Часто представляется в процентах. Может также использоваться абсолютный аналог.

Среднеквадратичная ошибка — глобальная характеристика по всем ошибкам на оцениваемом наборе:

$$\text{RMSE} = \sqrt{E\left\{(Z(x) - Z^*(x)\right\}}.$$

Коэффициент корреляции Пирсона — коэффициент корреляции между оценками и реальными значениями:

$$\rho = E\left\{Z(x)Z^*(x)\right\}.$$

3. Краткий обзор книг по геостатистике

Теория

Сложность 1

1. *Isaaks E. H., Srivastava R. M.* An Introduction to Applied Geostatistics. — Oxford: Oxford Univ. Press, 1989. — 592 p.

   Книгу можно рекомендовать в качестве исходного знакомства с геостатистикой для неспециалистов. Содержит основные понятия и модели геостатистики и изобилует примерами. Материал изложен доступно и не требует специальной подготовки.

2. *Clark H. W. A.* Practical Geostatistics 2000 / Publ. by Geostokos Ecosse. — [S. l.], 2004. — 440 p.

   Книга посвящена основам статистики и теоретическим основам линейной геостатистики, включая вариографию и различные модели кригинга. Включены упражнения и программное обеспечение по геостатистике PG2000.

Сложность 2

3. *Wackernagel H.* Multivariate Geostatistics. — [S. l.]: Springer, 2003. — 403 p.

   Книга посвящена многопеременной геостатистике. Подробно изложены основные модели многопеременной геостатистики. Имеются разделы о нелинейных моделях, нестационарных случаях, многопеременных пространственно-временны́х приложениях. Изложение теории сопровождается примерами из различных областей.

Сложность 3

4. *Chile J.-P. Delfiner P.* Geostatistics: Modeling Spatial Uncertainty. — New York: John Wiley & Sons Inc., 1999. — 695 p. — (Wiley Series in Probability and Statistics).

   В книге теоретическое математически насыщенное изложение всех основ геостатистики. Рассмотрены модели с использованием корреляции более высокого порядка.

5. *Cressie N.* Statistics for spatial data. — New York: John Wiley & Sons, 1991. — 900 p.

   Книга содержит наиболее полное изложение различных методов пространственной статистики, в том числе геостатистики, разработанных до 1990-х гг.

6. *Матерон Ж.* Основы прикладной геостатистики. — М.: Мир, 1968. — 407 с.

   Книга написана основателем геостатистической теории Ж. Изложены основы классической геостатистики. Хотя изложение носит чисто математический характер, издание представляет интерес не только для математиков, работающих в области теории вероятностей и функционального анализа, но и для специалистов по прикладным наукам, занимающихся статистическим анализом образцов и структур.

Справочники

Сложность 1

7. *Deutsch C. V., Journel A. G.* GSLIB: Geostatistical Software Library and User's Guide. — New York: Oxford Univ. Press, 1998. — 369 p.

   В книге описаны алгоритмы классической геостатистики, реализованные в Стэнфордском университете в виде библиотеки программ GSLIB на Фортране. Приведено краткое, но исчерпывающее описание основных алгоритмов и рассказано об их практическом применении. Приложен диск с исходными кодами программ и исполняемыми файлами, а также примерами данных.

Приложения к тематическим задачам

Сложность 1

8. *Дюбрюль О.* Геостатистика в нефтяной геологии / Издательство Института компьютерных исследований, НИЦ «Регулярная и хаотическая динамика», 2009. — 256 с.
   Переводная книга по геостатистике. Показано, не прибегая к языку математики, что геостатистика — простой и гибкий формальный подход для количественного представления геологических данных. Рассмотрены все основные аспекты классической геостатистики и способы адаптации геостатистических моделей для решения конкретных геологических задач.
9. *Красильников П. В.* Геостатистика и география почв / Наука, 2007 — 175 с.
   В книге представлено использование простейших методов геостатистики для анализа пространственных особенностей почв.
10. *Caers J.* Petroleum Geostatistics / Society of Petroleum Engineers. — [S. l.], 2005. — 88 p.

    В книге сжато изложена геостатистическая методология в приложении к моделированию нефтяных месторождений. Включены многие современные алгоритмы геостатистики, применяемые в практиче-

ских исследованиях. Материал изложен доступно, без подробных статистических выкладок и сопровождается иллюстративными схемами алгоритмов. Издание ориентировано на широкий круг читателей и не требует специальной математической подготовки.

11. *Deutsch C. V.* Geostatistical Reservoir modelling. — [S. l.]: Oxford Univ. Press, 2002. — 400 p.

    В книге изложена геостатистическая теория и приведены алгоритмы, используемые для моделирования пористых геологических сред нефтесодержащих пластов. Книга ориентирована на широкую аудиторию инженеров без специальной статистической подготовки. Материал сопровождается примерами и блок-схемами алгоритмов.

12. *Webster R., Oliver M. O.* Geostatistics for Environmental Scientists. — [S. l.]: John Wiley & Sons, 2000. — 286 p. — (Statistics in Practice).

    Книга — популярное изложение линейных методов геостатистики для задач окружающей среды. Включена глава о дизъюнктивном кригинге (disjunctive kriging). В приложении приведено описание программы Genstat.

Сложность 2

13. *Kanevski M., Maignan M.* Analysis and modelling of spatial environmental data. — Lausanne: EPFL Press, 2004. — 288 p. — (With a educational/research Geostat Office for Windows software package) (http://www.ppur.org/auteurs/1000772.html).

    Книга посвящена практическому анализу и моделированию пространственных данных. Изложены методы геостатистики и искусственного интеллекта (искусственных нейронных сетей и машин векторов поддержки). Приложен диск с учебной версией пакета программ «Геостат Офис», в котором реализованы описанные модели геостатистики и ИНС (учебная версия ограничена количеством загружаемых данных).

14. *Goovaerts P.* Geostatistics for Natural Resources Evaluation. — New York: Oxford Univ. Press, 1997. — 376 p.

    Книга содержит подробное описание основных методов геостатистики и их применения к пространственному анализу данных экологического мониторинга. Набор основных алгоритмов в целом совпадает с пакетом GSLIB, но сопровождается более разнообразными примера-

ми исследования. Материал изложен подробно и на хорошом математическом уровне.

15. Advanced Mapping of Environmental Data: Geostatitistics, Machine Learning and Bayesian Maximum Entropy / Ed. by M. Kanevski. — [S. l.]: iSTE, Dec. 2007. — 352 p.

    Книга посвящена применению статистических методов моделирования к разнообразным пространственным данным по окружающей среде, геологии, географии, климатическому моделированию, экологии и пр. Изложены модели классической геостатистики, а также современные разработки, методы машинного обучения (ИНС, машины поддерживающих векторов) и теория байесовской максимальной энтропии.

Сложность 3

16. *Kanevski M., Pozdnukhov A., Timonin V.* Machine learning algorithms for analysis and modelling of spatial data: Theory and case studies. — [S. l.]: EPFL Press, 2008. — 300 p.

    Книга — дальнейшее развитие более раннего издания. Наряду с кратким изложением моделей геостатистики рассмотрено применение моделей машинного обучения (искусственных нейронных сетей, машин поддерживающих векторов) к задачам пространственной классификации и регрессии. Изложены последние достижения в статистической теории обучения и представлен огромный спектр различных моделей основанных на обучении. Описание методов сопровождается примерами на реальных данных по окружающей среде. Приложен диск с пакетом программ «Machine Learning Office», дающий возможность применить модели на практике.

17. *Christakos G., Bogaert P., Serre M.* Temporal GIS: Advanced Functions for Field-Based Applications. — [S. l.]: Springer, 2002. — 250 p.

    В книге изложена теория метода байесовской максимальной энтропии (BME) и его приложение к задачам пространственно-временного картирования. Разработанная теория позволяет интегрировать в модели оценивания различные типы информации: интервальную, качественную, экспертную, эмпирическую. Теория метода проиллюстрирована практическими примерами. Включен пакет прикладных программ TGIS для Матлаба, в котором реализован BME.

4. Краткий обзор программного обеспечения по геостатистике

В этом приложении приведен список избранных геостатистических компьютерных программ. Список не претендует на полноту и содержит наиболее популярные и доступные программы, которые в совокупности отражают весь спектр геостатистических моделей. Выбор автора является субъективным и основывается на личном опыте.

GSLIB — набор программ на языке программирования FORTRAN (с открытыми кодами), написанных студентами и аспирантами Стэнфордского университета. Набор программ покрывает практически полный спектр методов классической геостатистики и может работать под различными операционными системами (Windows, UNIX, DOS). Распространяется на диске как приложение к книге *Deutsch C. V., Journel A. G.* GSLIB: Geostatistical Software Library and User's Guide. — New York: Oxford Univ. Press, 1998. — 369 p. (http://www.gslib.com). В этом наборе программ не предусмотрено средство для подбора параметров модели вариограммы. Программы могут запускаться как отдельные модули с использованием больших файлов с параметрами или через специальную интерактивную программу WinGSLIB (http://www.statios.com/WinGslib).

SGEMS — оболочка с набором прикладных геостатистических моделей и библиотека для разработчика, изданная и поддерживаемая Центром прогноза нефтяных месторождений (SCRF, http://ekofisk.stanford.edu/SCRF.html) Стэнфордского университета. Пакет программ включает наиболее современные алгоритмы многоточечной статистики наряду с моделями классической геостатистики (http://sgems.sourceforge.net).

VarioWin — интерактивная программа под Windows для анализа и моделирования пространственной корреляционной структуры данных включая построение модели вариограммы. Распространяется как приложение к книге *Pannatier Y.* VARIOWIN Software for Spatial Data Analysis. — New York: Springer Verl., 1996 (http://www-sst.unil.ch/research/variowin).

«Геостат Офис» (GSOffice) — набор интерактивных программ под Windows для полного анализа и визуализации (2D) пространственных данных. Помимо геостатистических моделей GSOffice содержит другие методы пространственного анализа (искусственные нейронные сети, машины на опорных векторах и пр.), есть возможность экспорта результатов в геоинформационные системы (ГИС) — ArcView, MapInfo. Учебная версия GSOffice (с

ограничением на количество входных данных). Распространяется как приложение к книге *Kanevski M., Maignan M.* Analysis and modelling of spatial environmental data. — Lausanne: EPFL Press, 2004. — 288 p. (http://www.ibrae.ac.ru/~mkanev/eng/gsoffice/HELP/Introduction.html.

Gstat — пакет геостатистических программ под различные платформы (Windows, UNIX, R), разрабатываемый E. J. Pebesma с 1996 г. в Утрехтском университете. Пакет включает различные типы кригинга, стохастическое гауссово и индикаторное моделирование, а также вариографию. Есть возможности обмена данными с ГИС (http://www.gstat.org).

На платформе статистического языка R (http://www.r-project.org) существуют и другие бесплатные дополнительные геостатистические модули (sgeostat, geoR, Rasp, geoRglm, VR и т. д. — http://cran.r-project.org/src/contrib/Views/Spatial.html).

GeoEAS — один из старейших программных пакетов по геостатистике, содержащий набор программ для выполнения геостатистической интерполяции (кригинга) с требующейся для этого предобработкой (вариография) данных и визуализацией. Пакет создавался при участии Агентства по охране окружающей среды США. Находится в свободном доступе (http://www.epa.gov/ada/csmos/models/geoeas.html).

Коммерческие программные продукты. Существуют многичисленные коммерческие геостатистические программные продукты (GS+, Geovariances Isatis, Lynx Geosystmes, SAGE 2001). В один из наиболее распространенных коммерческих пакетов для пространственной интерполяции SURFER Golden Software включены простой и обычный кригинг. Геостатистические алгоритмы находят применение в различных специализированных программных продуктах, таких как геоинформационные системы (ArcView Spatial Analyst™). Также, например, для нефтяной отрасли были разработаны специализированные программы, включающие геостатистические алгоритмы (Schlumberger Petrel™, IRAP™ RMS).

Более обширный список компьютерных программ по геостатистике можно найти на основном сервере по геостатистике AI-GEOSTAT (GIS, geostatistics, spatial analysis) (http://www.ai-geostats.org).

5. Краткий обзор информационных ресурсов по геостатистике в Интернете

В настоящее время в Интернете собрано огромное количество информации по анализу пространственно распределенных данных и по смежным темам. Ниже приведены ссылки на некоторые сайты, связанные с геостатистикой. Этот список неполный, он в основном представляет организации, использующие геостатистику для различных приложений. Геостатистика — динамично развивающаяся область, поэтому число новых ресурсов постоянно растет, многие группы, использующие геостатистику, имеют свои сайты.

1. **AI-GEOSTAT (GIS, geostatistics, spatial analysis)** (http://www.ai-geostats.org). Основной обзорный сервер по геостатистике. На нем можно подписаться на список рассылки электронной конференции AI-GEOSTAT. Здесь находится большое количество ссылок на различные ресурсы в сфере пространственного моделирования: программное обеспечение, публикации, конференции, вакансии и др.
2. **International Association for Mathematical Geology, IAMG** (http://www.iamg.org). Сервер Международной ассоциации математической геологии. Содержит ссылки на основные издания и конференции ассоциации, а также архив кодов компьютерных программ, опубликованных в журнале «Computers and Geosciences».
3. **GEOENVia** (http://www.geoENVia.org). Сервер международной ассоциации, пропагандирующей использование геостатистики для анализа окружающей среды. Здесь приведена информация о планирующихся конференциях, курсах, школах, связанных с геостатистикой. Конференции ассоциации проводятся раз в два года.
4. **PEDOMETRICS** (http://www.pedometrics.org). Сервер международной рабочей группы по применению математических методов для анализа почвы в рамках ассоциации по почвоведению. Геостатистика является основным, но не единственным аппаратом, использующимся при анализе почв. На этом сайте можно найти хорошие примеры практического использования геостатистики.
5. **Environmental Modelling and System Analysis Lab** (http://www.ibrae.ac.ru/~mkanev/). Веб-сайт Лаборатории моделирования окружающей среды и системных исследований ИБРАЭ РАН, который поддерживают авторы этой книги. На нем находится информация об исследованиях лаборатории (геостатистика, искусственные нейронные сети, фракталы,

временны́е ряды, радиоэкологическое моделирование), публикациях, научных проектах, а также о разрабатываемом математическом обеспечении. Приведены различные примеры исследования данных по окружающей среде с помощью геостатистики и ГИС. Также на сайте можно скачать программу 3Plot — визуализационный модуль пакета программы «Геостат Офис».

6. Ответы к упражнениям

Упражнение 2.1

При ячейковой декластеризации веса данных рассчитываются на основе количества попавших в ячейку данных. Размер ячейки может варьироваться и влиять на значения весов. Если в ячейку попадают все точки кластера (характерный размер кластера соответствует размеру ячейки декластеризации), то значения в этих точках учитываются с меньшими весами, что уменьшает их влияние на декластеризованное среднее значение данных. Размеры кластеров высоких и низких значений могут быть различны. При декластеризации кластеров высоких значений декластеризованное среднее значение меньше исходного, поскольку бо́льшие значения данных учитываются с меньшим весом. При декластеризации кластеров низких значений декластеризованное среднее выше исходного. Таким образом, варьируя размер ячейки декластеризации, можно построить кривую зависимости декластеризованного среднего значения от размера ячейки и на ее основе найти минимальное и максимальное декластеризованные средние значения. Максимальное среднее значение соответствует декластеризации кластеров низких значений, а минимальное среднее значение — декластеризации кластеров высоких значений.

Упражнение 2.2

Стационарность второго порядка включает в себя внутреннюю гипотезу, так как существование ковариации означает и существование полувариограммы. Это легко получить, расписав формулу вариограммы. Обратное — неверно.

Упражнение 3.1

A — степень 2; B — степень 1; C — степень 3.

Упражнение 4.1

Доказательство:

$$C(x,h) = E\left\{\left[E(x)-m\right]\left[Z(x+h)-m\right]\right\} =$$
$$= E\left\{\left[Z(x)Z(x+h)-mZ(x+h)-mZ(x)+m^2\right]\right\} =$$
$$= E\left\{Z(x)Z(x+h)\right\}-m^2-m^2+m^2.$$

Упражнение 4.2

Доказательство:

$$2\gamma(\boldsymbol{h}) = E\left\{\left[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})\right]^2\right\} =$$
$$= E\left\{\left[\left(Z(\boldsymbol{x})-m\right)-\left(Z(\boldsymbol{x}+\boldsymbol{h})-m\right)\right]^2\right\} =$$
$$= E\left\{\left[\left(Z(\boldsymbol{x})-m\right)\right]^2\right\}-2E\left\{\left(Z(\boldsymbol{x})-m\right)-\left(Z(\boldsymbol{x}+\boldsymbol{h})-m\right)\right\}+$$
$$+E\left\{\left(Z(\boldsymbol{x}+\boldsymbol{h})-m\right)^2\right\} = C(0)-2C(\boldsymbol{h})+C(0).$$

Упражнение 4.3

Доказательство:

$$\gamma(-\boldsymbol{h}) = 0{,}5\,\mathrm{Var}\left\{Z(\boldsymbol{x})-Z(\boldsymbol{x}-\boldsymbol{h})\right\} = E\left\{\left[Z(\boldsymbol{x})-Z(\boldsymbol{x}-\boldsymbol{h})\right]^2\right\} =$$
$$= E\left\{\left[Z(\boldsymbol{x}+\boldsymbol{h})-Z(\boldsymbol{x})\right]^2\right\} = \gamma(\boldsymbol{h}).$$

Упражнение 4.4

Если вариация функции распределения равна единице (например, в случае стандартного нормального распределения), то $\rho(h) = C(h)$, что приводит к искомому соотношению вариограммы и корелограммы.

Упражнение 4.5

Половина угла раствора $\Delta\varphi$ равна 15°. Для получения шести направлений для расчета вариограммы 180° делится на шесть с учетом свойства симметрии вариограммы. Таким образом, каждый сектор равен 30°, что дает половину раствора 15° в обе стороны от угла направления каждой вариограммы.

Упражнение 4.6

Стационарные модели: наггет, сферическая, гауссова (асимптотически), экспоненциальная (асимптотически), периодическая, затухающая периодическая, а также кубическая и пентасферическая.

Нестационарная модель: степенная. Но при желании ее тоже можно ограничить.

Упражнение 4.7

$C(\infty) = \gamma(0)$.

Воспользуемся результатом упражнения 4.2: $\gamma(0) = C(0) - C(0) = 0 = C(\infty)$.

Упражнение 4.8

$\gamma(\infty) = C(0) = \sigma^2$.

Воспользуемся результатом упражнения 4.2: $\gamma(\infty) = C(0) - C(\infty) = C(0) = \sigma^2$.

Упражнение 4.9

Вариограмма является симметричной функцией: $\gamma(h) = \gamma(-h)$. Для произвольного угла направления α справедливо $\gamma(\alpha) = \gamma(\alpha + 180)$.

Радиус корреляции для 270°~ 5, для 240°~ 6, для 210 ~ 9, для 180 ~ 20.

Упражнение 4.10

Ответы:

а) одиночное тело, радиус корреляции 26—28;

б) набор выпуклых тел, радиус корреляции 10—30, нижняя граница соответствует корреляции внутри каждого объекта, в то время как верхняя

граница характеризует корреляцию между объектами, периодическая структура указывает на повторяющиеся схожие объекты;

в) извилистое русло, радиус корреляции 25—30, присутствует анизотропия на мелких масштабах (повороты русла);

г) параллельные русла, радиус корреляции 5—6 по вертикали соответствует толщине русел и до 10 по горизонтали характеризует горизонтальные участки русел, присутствует геометрическая анизотропия;

д) смыкающиеся объекты, радиус корреляции 10—20, сильная геометрическая анизотропия характеризует ориентацию структур;

е) пикселизированная мозаика, радиус корреляции 3 соответствует размеру мелких объектов различной формы, геометрическая анизотропия характеризует диагональную ориентацию объектов, выход вариограммы на плато указывает на отсутствие корреляции между мелкими объектами случайной формы.

Упражнение 5.1

Доказательство.

Оценка кригинга: $Z^*(\boldsymbol{x}_0) = \sum_{j=1}^{n} w_j Z(\boldsymbol{x}_j)$;

математическое ожидание невязки оценки $Z(x_0)$ в точке x_0:

$$E\left\{\sum_{j=1}^{n} w_j Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_0)\right\} = \sum\nolimits_{i=1}^{n} w_i E\left\{Z(\boldsymbol{x}_i)\right\} - E\left\{Z(\boldsymbol{x}_0)\right\} =$$

$$= E\{Z(\boldsymbol{x}_i)\}E\left\{\sum_{i=1}^{n} w_i\right\} - E\{Z(\boldsymbol{x}_0)\} = 0,5\text{A}; 8\ E\left\{\sum_{i=1}^{n} w_i\right\} = 1,$$

что означает несмещенность оценки при стационарности случайной функции:

$$E\left\{Z(\boldsymbol{x}_i)\right\} = E\left\{Z(\boldsymbol{x}_0)\right\} = E\left\{Z\right\}.$$

Упражнение 5.2

Доказательство.

Если $Z(\boldsymbol{x})$ не обладает стационарностью, то

$$E\left\{Z(\boldsymbol{x}_i)\right\} \neq E\left\{Z(\boldsymbol{x}_0)\right\}.$$

Таким образом,

$$E\{Z(\boldsymbol{x}_i)\}E\left\{\sum_{i=1}^{n} w_i\right\} - E\{Z(\boldsymbol{x}_0)\} \neq 0,$$

что означает смещенность оценки кригинга.

Упражнение 5.3

Доказательство.

Оценка простого кригинга: $Z^*(\boldsymbol{x}_0) = \sum_{j=1}^{n} w_j Z(\boldsymbol{x}_j), \;\; \sum_{j=1}^{n} w_j C_{ij} = C_{i0}.$

Если $\boldsymbol{x}_0 = \boldsymbol{x}_i$, т. е. это точка из набора данных, то $C_{i0} = C_{00} = \sigma^2$, и каждое уравнение системы будет иметь вид

$$\sum_{\substack{j=1,\\ j\neq i}}^{n} w_j C_{ij} = \sigma^2(1 - w_i).$$

Очевидно, что $w_i = 1, w_j = 0, \; j = 1, ..., n, \; j \neq i$ является решением этой системы. И если матрица ковариаций несингулярна, то решение системы единственно.

Тогда получим: $Z^*(x_0) = 1 \cdot Z(x_0) + 0 \cdot Z(x_1) + ... + 0 \cdot Z(x_n) = Z(x_0).$

Упражнение 5.4

Доказательство.

Ошибка простого кригинга при нулевой ошибке измерений

$$\sigma_{SK}^2 = \sigma^2 - \sum_{i=1}^{n} w_i C_{i0},$$

где σ^2 — вариация данных.

Как было показано в упражнении 5.3, $w_i = 0$, $w_0 = 1$, $C_{00} = \sigma^2$, тогда

$$\sigma_{SK}^2 = \sigma^2 - 1 \cdot \sigma^2 - 0 \cdot C_{10} - \ldots - 0 \cdot C_{1n} = 0.$$

Упражнение 5.5

А. Вариация оценки кригинга определяется по формуле (5.14) как разность вариации исходных данных и взвешенной суммы ковариаций. Последняя

положительна в силу неотрицательной ковариации в предположении о стационарности. Таким образом, вариация кригинга не больше вариации исходных данных.

Б. Гладкость оценки кригинга определяется ее глобальной вариацией. Вариация оценки кригинга равна вариации исходных данных в случае нулевой ковариации в формуле (5.14). Это может быть достигнуто в случае полного отсутствия пространственной корреляции — вариограмма с чистым наггетом. В этом случае $C_{ij} = C_{i0} = 0$.

Упражнение 5.6

А — обратные квадраты расстояния, В — обычный кригинг с большим радиусом корреляции ($r = 10$ — все данные), C — обычный кригинг с маленьким радиусом корреляции (r = 1).

Упражнение 7.1

Доказательство.

$$C(0) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[Z(\boldsymbol{x}_i) - m_1\right]^2 = \frac{1}{n-1}\sum_{i=1}^{n} n_1(1-0,5)^2 + n_2(0,5)^2 =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(n_1 + n_2\right)0,25,$$

где $n_1 + n_2 = n$, где n_1 соответствует индикаторным значениям 0, а n_2 — 1. Среднее значение для отсечения по медиане $m_1 = 0,5$.

Упражнение 8.1

Оценка кригинга гладко интерполирует значения оцениваемой переменной в промежутках между данными. Стохастическое моделирование не дает гладкой зависимости в промежутках между данными.

Упражнение 8.2

Оценка кригинга всегда ограничена минимальным и максимальным значениями данных. Максимальное значение стохастической реализации может

быть выше максимального значения исходных данных, оно зависит от задаваемого уровня.

Упражнение 8.3

А — стохастическое моделирование, Б — кригинг. Уровень плато вариограммы характеризует вариация глобального распределения, которая всегда меньше у оценки кригинга, чем у стохастической реализации.

Упражнение 8.4

А. Могут использоваться любые стационарные типы моделей (сферическая, экспоненциальная, гауссова и пр.), степенна́я модель не может быть использована.

Б. Плато вариограммы равно значению априорной вариации, которое для стандартного гауссова распределения нормализованных значений, используемых в гауссовом моделировании, равно 1.

7. Глоссарий

Анизотропия — зависимость некоторого свойства функции от ориентации аргумента.

Анизотропия геометрическая (geometric) — анизотропия, при которой полувариограммы (ковариации) по различным направлениям имеют одинаковую форму и плато, но разные радиусы, изолинии вариограммы на диаграмме имеют форму концентрических эллипсов.

Анизотропия зонная (zonal) — анизотропия, которая не является геометрической.

Валидация — проверка качества работы модели при помощи данных, не использованных для ее настройки.

Вариабельность (пространственная) — свойство пространственно распределенной функции иметь неоднородное поле значений.

Вариограмма (variogram) или полувариограмма, структурная функция — статистический момент второго порядка для разности значений в точках, разделенных некоторым вектором, т. е. зависимость квадрата разности значений функции в точках от вектора расстояния между точками.

Вариограмма анизотропная — зависимость значений вариограмм от направления вектора, разделяющего пары точек.

Вариограмма по всем направлениям (omnidirectional variogram) — вариограмма, не моделирующая анизотропию, т. е. зависящая только от модуля вектора, разделяющего точки пары.

Вариограмма экспериментальная — значение вариограммы, вычисленное на основе данных.

Модель вариограммы — теоретическая функция, аппроксимирующая значения вариограммы между точками экспериментальной вариограммы (лагов).

Параметры модели вариограммы — см. наггет, плато, радиус корреляции.

Поверхность вариограммная (variogram surface) — диаграмма значений вариограммы, полученных на регулярной сетке.

Роза вариограммная (variogram rose) — лепестковая диаграмма типа розы ветров, где вдоль каждого лепестка отложено значение вариограммы в соответствующем направлении.

Вариография (variography) — анализ и моделирование пространственной корреляции (вариограмм).

Внутренняя гипотеза (intrinsic hypothesis) — свойство случайной функции со стационарными приращениями, т. е. функции, у которой существуют математическое ожидание, не зависящее от местоположения, и конечная вариация разницы значений функции в точках независимо от местоположения.

Геостатистика (geostatistics, пространственная статистика) — развитие статистики для анализа пространственно распределенных данных.

Декластеризация (declustering) — приписывание весов значениям пространственной функции в точках в зависимости от характера сети мониторинга.

Дрейф (пространственный, drift) — зависимость средней разности значений функции точек от вектора расстояния между точками.

Индикаторный подход — непараметрический метод для моделирования локальной функции распределения пространственной случайной переменной, основан на нелинейном преобразовании данных, моделирующем функцию распределения в исходных точках.

Интерполяция — оценивание значений функции в точках, где значение неизвестно, но окруженных точками с известными значениям аргумента.

Детерминистические методы пространственной интерполяции — методы, основанные на предположении об аналитической (формуль-

ной) зависимости между данными в пространстве (обратные расстояния в степени, полиномы, сплайны и т. д.).

Статистические (геостатистические) методы пространственной интерполяции — методы, основанные на предположении о существовании случайной пространственной функции, реализациями которой являются значения измерений.

Кластер (cluster) — область повышенной плотности точек измерений пространственной функции.

Ковариационная функция (covariance function) — зависимость разницы среднего значения произведения значений функции в парах точек и квадрата математического ожидания функции от вектора, разделяющего точки пары.

Ковариация пары значений функции — разница среднего значения произведения значений функции в двух точках и квадрата математического ожидания функции.

Кокригинг (cokriging) — геостатистический метод совместной пространственной интерполяции нескольких переменных, основанный на линейной регрессии, обладает несмещенностью и минимальной дисперсией оценки.

Кригинг (kriging) — геостатистический метод пространственной интерполяции, основанный на линейной регрессии, обладает несмещенностью и минимальной дисперсией оценки.

Индикаторный кригинг (indicator kriging) — геостатистический непараметрический метод пространственной интерполяции, основанный на линейной регрессии нелинейно преобразованных переменных (индикаторов), обладает несмещенностью и минимальной дисперсией оценки.

Кригинг с внешним дрейфом — кригинг с трендом, который задается значениями функции тренда в точках оценивания.

Логнормальный кригинг — геостатистический метод пространственной интерполяции для функции, реализации которой обладают свойством логнормальности, т. е. логарифмическое преобразование приводит к нормальному распределению.

Обычный кригинг (ordinary kriging) — кригинг с неизвестным математическим ожиданием случайной функции.

Простой кригинг (simple kriging) — кригинг с известным математическим ожиданием случайной функции.

Универсальный кригинг (с трендом) — кригинг с полиномиальной моделью тренда.

Кросс-валидация (cross-validation) — метод подбора оптимальных параметров модели интерполяции при помощи оценки значения в точке измерения без учета самого измерения в этой точке.

Лаг, лэг (lag) — расстояние, которое выбирается для поиска пар точек при расчете моментов второго порядка (вариограммы, ковариации, мадограммы и т. д.).

Мадограмма (madogram) — зависимость среднего модуля разности значений функции от вектора расстояния между точками.

Наггет (nugget — самородок) — параметр теоретической модели вариограммы, характеризующий значение вариограммы вблизи нуля.

Непрерывность — свойство данных, при котором пара точек, находящихся ближе друг к другу, скорее будет иметь близкие значения, чем пара удаленных друг от друга точек.

Нестационарность — изменяющийся характер распределения в зависимости от области рассмотрения.

Нормальная бумага (normal probability plot) — график зависимости значений функции распределения случайной переменной от значений, соответствующих нормальному распределению.

Плато (sill) — параметр теоретической модели вариограммы, характеризующий значение вариограммы на больших расстояниях (при условии ее стационарности).

Полигоны Вороного (ячейки Дирихле, Тиссена) — область влияния точки X_i, т. е. совокупность всех точек исследуемой области ($Z(X_i)$) таких, что $\forall x \in Z(X_i)$, $\forall j \neq i$: $|X_i, x| < |X_j, x|$.

Последовательный принцип при стохастическом моделировании — использование уже промоделированных значений при моделировании в следующих точках.

Пост плот (post plot) — диаграмма местоположения точек (графическое представление данных).

Пространственная корреляция — зависимость между значениями пространственно распределенной функции от взаимного расположения точек.

Радиус корреляции (range) — параметр теоретической модели вариограммы, характеризующий расстояние достижения вариограммой постоянного значения (плато).

Случайная переменная — переменная, которая может принимать набор значений в соответствии с функцией распределения вероятности.

Стационарность — отсутствие зависимости в поведении случайной функции от местоположения.

Стационарность в строгом смысле — инвариантность функции распределения относительно вектора сдвига.

Стационарность в широком смысле (second order stationarity) — свойство случайной функции: математическое ожидание не зависит от местоположения, существует ковариация, зависящая только от вектора, разделяющего точки (стационарность ковариации).

Стохастическое моделирование (симуляции, stochastic simulations) — метод генерации равновероятных реализаций в соответствии с функцией распределения случайной функции.

Гауссово стохастическое моделирование — алгоритмы стохастического моделирования в предположении о мультинормальности моделируемой случайной функции.

Гауссово обрезанное моделирование (truncated Gaussian) — специальная модификация алгоритма гауссова стохастического моделирования для случая категориальной переменной.

Индикаторное моделирование — алгоритм последовательного стохастического моделирования, использующий индикаторный подход, который требует предварительного индикаторного преобразования данных.

Объектное моделирование — алгоритм стохастического моделирования, основанный на использовании объектов характерной формы.

Отжига моделирование (simulated annealing) — алгоритм генерации равновероятных реализаций распределения случайной функции, основанный на принципе стохастической релаксации и имитирующий металлургический процесс медленного охлаждения раскаленного металла.

Прямое моделирование — алгоритм последовательного стохастического моделирования, не требующий предварительного преобразования данных, так как не делается никаких предположений о характере функции распределения данных.

Структурный анализ (пространственный), вариография (variography) — анализ и моделирование пространственной корреляции (вариограмм).

Тренд пространственный (trend) — крупномасштабная зависимость значений пространственной функции от местоположения.

Триангуляция — разбиение области исследования на треугольники с вершинами в точках измерений так, что их ребра не пересекаются.

Указатель

Научное издание

Демьянов Василий Валерьевич,
Савельева Елена Александровна

ГЕОСТАТИСТИКА
теория и практика

Утверждено к печати Ученым советом
Института проблем безопасного развития атомной энергетики
Российской академии наук

Редактор А. И. Иоффе

Издательство «Наука»
117997, Москва, Профсоюзная ул., 90
Зав. редакцией М. В. Грачева
Редактор изд-ва И. С. Власов

Оригинал-макет подготовлен ООО «Комтехпринт»
Иллюстрации приведены в авторской редакции

Формат 60×90 $^{1}/_{16}$. Бумага офсетная 80 г/м^2
Печать офсетная. Гарнитура «Оффицина»
Уч.-изд. л. 20,4. Заказ № 20138

Заказное

Отпечатано с готовых диапозитивов типографией ООО «Инфолио-Принт»