Каневский М.Ф., Демьянов В.В., Савельева Е.А., Чернов С.Ю., Тимонин В.А.
Элементарное введение в геостатистику
серия Проблемы окружающей среды и природных ресурсов,
№ 11, ВИНИТИ, Москва, 1999

Введение в методы анализа данных по окружающей среде

М.Ф. Каневский, В.В. Демьянов

Проблема анализа пространственно-распределенной информации по окружающей среде является чрезвычайно важной в настоящее время. Глубокий анализ и моделирование пространственных данных требует применения комплексного подхода и различных методов, характеризующих ту или иную особенность явления. Сложность такого анализа обусловлена несколькими факторами: наличием большого количества количественной и качественной информации по исследуемому явлению, многомасштабностью и многопеременностью, наличием различных факторов влияния. Это прежде всего относится к задачам анализа загрязнения окружающей среды.

В настоящей работе изложены элементы адаптивной методологии для анализа пространственно-распределенных данных, включающей статистическое описание, анализ сети мониторинга, анализ пространственных корреляций, существующие геостатистические методы пространственного оценивания в сравнении с широко используемыми традиционными детерминистическими интерполяциями. Существенно, что выбор (адаптация) метода/модели в рамках предложенной схемы зависит как от количества и качества исходных данных, так и от целей и задач анализа: глобальные оценки, подготовка карт для принятия решений, вероятностное картирование, сравнение результатов моделирования и измерений и т.п.

Применение предлагаемой методологии проиллюстрировано на примере анализа пространственного радиоактивного загрязнения почвы в результате Чернобыльских выпадений.

Постановка задачи анализа пространственных данных

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. Наиболее распространенной проблемой при работе с пространственно распределенными данными является получение пространственной оценки. Так, например, было подготовлено много различных карт по радиоактивному загрязнению почвы в результате Чернобыльской аварии. При этом остается открытым вопрос о качестве и точности этих карт, неопределенности оценки, чувствительности использованных методов интерполяции, и т.п. В рамках этой

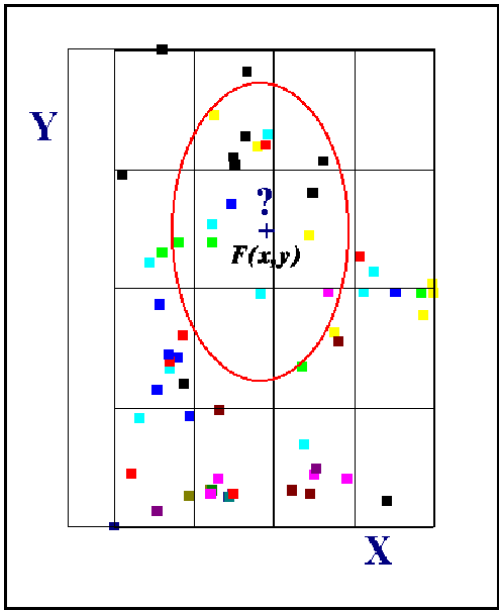

Рис.1. Постановка задачи пространственного анализа данных.

проблемы можно выделить ряд конкретных задач, для решения которых необходимо применение комплекса исследований с помощью методов геостатистики – статистики пространственно распределенной (региональной) информации:

- Оценить значение в точке, где измерения не проводились.
- Нарисовать карту, построить изолинии (определить значения на плотной сетке).
- Оценить ошибку интерполяции.
- Учесть при интерполировании ошибки измерений.
- Определить вероятность превышения заданного уровня.
- Провести совместный пространственный анализ коррелированных переменных.
- Получить набор равновероятных пространственных реализаций распределения.
- Описать пространственную вариабельность и неопределенность.

Традиционные детерминистические методы, широко используемые в задачах пространственной интерполяции, позволяют практически ответить только на первые два из выше поставленных вопросов.

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

Подходы к анализу пространственно распределенных данных

Существует несколько подходов к анализу и обработке пространственно распределенных данных, которые можно условно разделить на 4 группы [1-13].

1. Детерминистические модели (интерполяторы): триангуляция, метод обратных расстояний в степени, мультиквадратичные уравнения, и т.п.
2. Геостатистика – модели, базирующиеся на статистической интерпретации данных.
3. Алгоритмы искусственного интеллекта (искусственные нейронные сети различной архитектуры, генетические алгоритмы).
4. Модели, базирующиеся на статистической теории обучения (теория Вапника-Червоненкиса): машины векторов поддержки (Support Vector Machines).

Конечно, такое деление является крайне условным. Так, геостатистические модели можно изложить в детерминистической формулировке и наоборот, ряд детерминистических моделей 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.

Современная геостатистика – это широкий спектр статистических моделей и инструментов для анализа, обработки и представления пространственно-распределеннойинформации. В настоящей работе мы подробно опишем наиболее часто используемые модели и инструменты, из которых можно составить замкнутый цикл исследования и решить поставленные выше задачи.

При работе с пространственными данными важно прежде всего понять насколько эффективна имеющаяся *сеть мониторинга*. Для оценки этого используются различные характеристики, описывающие топологию сети, включая фрактальную размерность. Это позволяет оценить чувствительность сети и понять, какие явления она может детектировать.

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. Следует отметить, что пространственные данные могут быть различного рода: непрерывные (загрязнение), категориальные (типы почв), интерполируемые и неинтерполируемые.

Если данные собраны на нерегулярной кластерной сети мониторинга, то необходимо проведение пространственной декластеризации для получения репрезентативной глобальной статистики – средних, вариаций, гистограмм. Оценить пространственные особенности данных позволяет статистика с движущимся окном, когда 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 – его оценка обладает минимальной дисперсией. Важным свойством кригинга является точное воспроизведение значений измерений в имеющихся точках (интерполяционные свойства). В отличие от многочисленных детерминистических методов оценка кригинга сопровождается оценкой ошибки интеполяции в каждой точке. Полученная ошибка позволяет охарактеризовать неопределенность интерполяционной оценки данных при помощи доверительных интервалов и “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 при помощи методов геостатистики и фракталов. Эффективными инструментами подбора модельных параметров являются методы кросс-валидации (cross-validation), складного ножа (jack-knife), бутстрэп (bootstrap).

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 имеется 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 не превышению 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 и 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 резервуара, "длины" береговой линии, и т.п.

Еще одной проблемой, часто осложняющей проведение анализа пространственных данных, является пространственная нестационарность . Поведение данных в природе обычно зависит от множества различных факторов . Это приводит к появлению разномасштабных пространственных структур. Проблема моделирования и удаления крупномасштабного тренда из данных решается различными способами. 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 свою высокую эффективность по сравнению с другими существующими методами на различных данных, имеющих сложный пространственный характер (нестационарность, периодичность, пятнистость). Перечисленные методы успешно применялись авторами в процессе анализа данных по радиоактивному загрязнению почвы в результате Чернобыльской аварии, данных по химическому загрязнению почвы тяжелыми металлами в Швейцарии и Японии, климатических данных.

Приведем ряд типичных проблем, с которыми обычно сталкиваются при анализе и моделировании пространственной информации и которые можно решить в рамках предлагаемой методологии:

Проблема	**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.
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.
Имеет ли смысл задача интерполяции?	При отсутствии пространственной корреляции между данными получение оценки в точке путем взвешивания соседних измерений ***не имеет смысла***.
Как выявить и промоделировать пространственную непрерывность данных на различных масштабах?	Исследовать и промоделировать пространственную корреляцию данных с учетом возможной нестационарности и анизотропии при помощи стандартных приемов вариографии, анализа трендов.
Получить наилучшую в статистическом смысле оценку значения пространственной переменной в точке, где измерения отсутствуют. Оценить ошибку полученной оценки. Построить карты оценок и ошибок оценки.	Применить спектр моделей кригинга – наилучших несмещенных линейных оценивателей.
Как учесть при интерполяции ошибки измерений?	Геостатистическое оценивание позволяет учесть ошибку измерений и ее пространственное распределение при интерполяции.
Как подобрать оптимальные параметры модели интерполяции?	Методы кросс-валидации, jack-knife, bootstrap позволяют эффективно подобрать оптимальные параметры и не зависят от выбранной модели интерполяции.

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).
Построить оценку вероятности превышения заданного уровня значений (провести оценку риска). Получить не единственную оценку функции в точке, построить равновероятные реализации пространственного распределения	Методы вероятностного картирования, включающие индикаторный кригинг и стохастическое моделирование, позволяют получать множество равновероятных реализаций функции и оценивать на их основе различные статистические характеристики, описывать пространственную вариабельность и неопределенность данных.
Как избежать "сглаженной оценки" и воспроизвести изначальную вариабельность данных?	Стохастическое моделирование дает не сглаженную картину и воспроизводит исходные данные наряду с параметрами распределения (статистические моменты 1-го и 2-го порядков). Оно позволяют описать неопределенность и пространственную вариабельность данных
Как учесть дополнительную априорную информацию о данных и/или соседних областях?	Применить модели Байесовского кригинга.
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.

Истоки геостатистической методологии

Геостатистика возникла в начале 60-х годов, как теория региональных переменных, сформулированная Ж. Матероном (Matheron) для анализа данных о природных ископаемых (горнорудное дело) (Matheron, 1963; Матерон, 1968). Матероном был организован Центр геостатистики в Фонтенбло, который внес заметный вклад в теоретические исследования и практические применения.

Практически в это же время, и даже несколько ранее, Л.С. Гандиным была сформулирована теория оптимальной интерполяции для объективного анализа метеополей (Гандин, 1976). К сожалению работы российских ученых в этой области не были широко известны на западе (Вистелиус, 1984, 1986).

В настоящее время одними из наиболее активных центров являются Стэнфордский университет, руководителем геостатистического направления в котором является А.

Жорнель (Journel A.G., Huijbregts Ch.J., 1978) и центр геостатистики в Фонтенбло (Франция).

Современная геостатистика имеет различные области применения: природные ископаемые и ресурсы, картография, экология, гидрогеология, финансовая активность и анализ рынков, эпидемиология, лесное и рыбное хозяйство, анализ общественного мнения, криминогенная ситуация, восстановление загрязненных территорий, обработка изображений и многое другое (http://cg.ensmp.fr/anglais/LesDomaines.html).

Международная ассоциация по математической геологии (International Association for Mathematical Geology- IAMG) (www.iamg.org) является организацией, среди прочего курирующей и вопросы геостатистики. Под ее эгидой издаются основные журналы по этой тематике: Mathematical Geology, Computers & Geosciences. IAMG ежегодно проводит научные конференции, раз в четыре года проводятся геостатистические конгрессы (Кейптаун 2000). Раз в два года проводятся достаточно популярные Европейские конференции по применению геостатистики для окружающей среды – geoENV.

Подробная информация о геостатистическом мире и его связи с географическими информационными системами может быть получена на сервере AI-GEOSTAT: http://curie.ei.jrc.it/ai-geostats.htm. На этом сервере находится большое количество информации, касающейся часто задаваемых вопросов, математического обеспечения, конференций, публикаций и т.п. Этот сервер может стать хорошей отправной точкой для начинающих геостатистиков.

Методология анализа пространственной информации

В рамках настоящего сборника статей описан и применен широкий спектр современных методов комплексного исследования распределенной в пространстве и зависящей от времени информации. Основные использованные модели базируются на геостатистическом подходе, теории фракталов, стохастическом моделировании, искусственных нейронных сетях. Такое рассмотрение позволяет значительно улучшить качество анализа, обработки и представление радиоэкологических данных. На основе перечисленных подходов была выработана передовая методика анализа радиоэкологической информации. Методология анализа состоит из нескольких крупных этапов. В пределах каждого этапа используются специфические методы, позволяющие оценить те или иные характеристики изучаемого явления. Алгоритмическая схема методологии представлена на Рис. 1. Методика анализа пространственно распределенных данных состоит из следующих блоков:

- Подготовка баз данных, конвертация координат из географической системы отсчета в различные метрические системы и обратно. Работа с категориальными данными и графическими базами данных, кодирование категориальной информации.
- Визуализация/отображение пространственных измерений и другой информации: пост плот, полигоны Вороного/ячейки Дирихле, триангуляция Делоне.
- Анализ сети мониторинга: фрактальная размерность сети, индекс Моришита, генерация новых точек.

- Декластеризация: ячейковая, случайная, по гистограмме, комбинированная, по площадям полигонов.
- Одномерное описание: общая экспресс-статистика, нормальная бумага, гистограммы.
- Многопеременное распределение: диаграмма взаимного разброса, корреляция, линейная регрессия.
- Статистика движущегося окна, эффект пропорциональности, вариабельность статистических параметров.
- Структурный анализ (вариография): экспериментальные вариограммы, вариограмные поверхности, диаграммы разброса пар, вариограмные облака; кросс-вариограмма; анализ тренда; вариография невязок; автоматический подбор модели вариограмм, вариограмная модель.
- Кросс-валидация: анализ невязок, карты ошибок и невязок.
- Подготовка плотной пространственной сетки для проведения интерполяций, генерация адекватной сети мониторинга.
- Пространственные интерполяции: детерминистические (обратные квадраты расстояний, мультиквадратичные уравнения), геостатистические (простой кригинг, обычный кригинг, логнормальный кригинг, кокригинг).
- Непараметрические методы – индикаторный кригинг (оценки локальных функций распределения).
- Стохастическое моделирование – альтернативные равновероятные пространственные реализации.
- Визуализация результатов: карты оценок и ошибок, контурные и мозаичные карты, “толстые” изолинии неопределенности; карты вероятностей и риска, квантильные карты.
- Географические информационные системы (ГИС) – картография, карты для принятия решений.

Приведенная выше методология основана на многолетнем мировом опыте анализа пространственно распределенных данных (в том числе и собственном опыте авторов) Cressie N. 1991; Isaaks E.H., Shrivastava R.M. 1989; Goovaerts P. 1997, Wackernagel H. 1995; Armstrong M. 1997; Каневский, 1999; Kanevsky et. al., 1996, 1997).

Безусловно, геостатистический анализ требует не только глубоких экспертных знаний и опыта работы с данными, но и наличия современного и эффективного программного обеспечения. В настоящее время в мире существуют разнообразные программные реализации тех или иных геостатистических моделей, а так же полного цикла анализа. Современные прикладные пакеты геостатистических программ делятся на дорогостоящие коммерческие (Isatis, Lynx) с ценой порядка десятков тысяч долларов, и более доступные по цене, но обладающие либо ограниченным набором средств (Variowin, GS+), либо несовершенным пользовательским интерфейсом (GSLIB, Geo-EAS). В рамках настоящей работы авторами будет продемонстрирован пакет прикладных программ “Геостат офис”, который реализован как рабочее место эксперта для анализа пространственно распределенных данных [12]. “Геостат офис” предоставляет модели и инструменты для проведения полного цикла исследования и картирования данных в среде MS Windows™.

Данные для примера исследования

В качестве примера для демонстрации всех излагаемых ниже моделей и методов были выбраны данные по радиоактивному загрязнению почвы в результате Чернобыльской аварии. Авария на Чернобыльской АЭС произошла в 1986 году и сопровождалась выбросом радиоактивных материалов из активной зоны реактора. Это повлекло за собой загрязнение огромных территорий в Европе, на Украине, в Белоруссии и России. Радиоактивное загрязнение, образовавшееся в результате выпадения радионуклидов из облака, имеет крайне сложный пятнистый характер с присутствием анизотропии и большого количества экстремальных значений. Такая структура выпадений обусловлена многочисленными факторами влияния: атмосферные осадки, метеорологические условия, сухое и влажное осаждение, орографические эффекты, характеристики подстилающей поверхности, типы почв, и т.д. Наиболее существенное влияние на долговременное заражение почвы оказали радионуклиды цезий 137 (^{137}Cs) и стронций 90 (^{90}Sr), существенное наличие которых в выбросе обусловлено особенностями источника (типа реактора). Период полураспада этих элементов составляет около 30 лет. К настоящему времени составлено много карт Чернобыльских выпадений, как в России так и за рубежом (De Cort M. и др.,1996). При их составлении использовались различные методы – в основном детерминистические.

Для проведения полного анализа таких сложных данных необходимо применять широкий спектр методов пространственной статистики. В качестве конкретных данных для примера исследования были выбраны измерения радиоактивного загрязнения ^{137}Cs и ^{90}Sr в западной части Брянской области – наиболее загрязненной части России.

В данной работе предполагается описать и продемонстрировать практическое применение основных элементарных подходов из предложенной методологии, не углубляясь в более продвинутые методы анализа сети мониторинга, вероятностного оценивания, стохастического моделирования и искусственных нейронных сетей.

Все результаты, полученные в настоящем сборнике, носят научный и методический характер и не могут использоваться для принятия конкретных решений.

В основу настоящего сборника легли работы авторов, выполненные при поддержке грантов ИНТАС 94 2361, 96 1957 и 98 1726.

Литература

1. Матерон Ж. Основы прикладной геостатистики. М.: Мир, 1968, 407 с.
2. Matheron G. Principles of Geostatistics. Economic Geology, V.58, 1963, pp.1246-1266.
3. Вистелиус А.Б. Математическая геология и ее вклад в фундаментальные геологические разработки. Препринт ЛОМИ Р-5-86, 1986, 27 с.
4. Вистелиус А.Б. Математическая геология: история, состояние, перспективы. Препринт ЛОМИ, Р-10-84, 53 с.
5. Гандин Л.С., Каган Р.Л. Статистические методы интерполяции метеорологических данных. Гидрометеоиздат, Ленинград, 1976. 359 с.
6. Journel A.G., Huijbregts Ch.J. Mining Geostatistics. London: Academic Press. 1978, 600 p.
7. Cressie N. Statistics for spatial data. John Wiley & Sons, New-York, 900p., 1991.
8. Isaaks E.H., Shrivastava R.M. An Introduction to Applied Geostatistics. Oxford University press, Oxford, 1989.
9. Goovaerts P. Geostatistics for Natural Resources Evaluation. Oxford University Press, 1997.
10. Wackernagel H. Multivariate Geostatistics. Springler-Verlag, Berlin, 1995.
11. Armstrong M. Basic Linear Geostatistics. Springer Verlag, 1997.
12. De Cort M., Tsaturov Yu.S. Atlas on caesium contamination of Europe after the Chernobyl nuclear plant accident. European Commission, report EUR 16542 EN, 1996, 39 p.
13. Vapnik V.N. Statistical Learning Theory. John Wiley & Sons, Inc. N.Y. 736 p.
14. М. Каневский, В. Демьянов, С. Чернов, Е. Савельева, В. Тимонин, Геостатистика и искусственные нейронные сети для анализа и моделирования пространственно распределенных данных. Известия РАН. Энергетика, том. 1, 1999.
15. M. Kanevsky, R. Arutyunyan, L. Bolshov, S. Chernov, V. Demyanov, I. Linge, N. Koptelova, E. Savelieva, T. Haas, M. Maignan. Chernobyl Fallouts: Review of Advanced Spatial Data Analysis, geoENV I – Geostatistics for Environmental Applications, ed. A. Soares, J. Gomez-Hernandes, R. Froidvaux, Kluwer Academic Publishers, 1997, pp. 389-400.
16. Kanevsky M., Arutyunyan R., Bolshov L., Demyanov V., Linge I., Savelieva E.,. Shershakov V., Haas T., Maignan M. Geostatistical Portrayal of the Chernobyl Fallout. Geostatistics Wollongong '96, ed. E.Y. Baafi, N.A. Schofield, Kluwer Academic Publishers, 1996, volume 2, pp.1043-1054.

Рисунок 1. Схема методологии анализа, обработки и представления пространственно-распределенных данных.

Каневский М.Ф., Демьянов В.В., Савельева Е.А., Чернов С.Ю., Тимонин В.А.
Элементарное введение в геостатистику
серия Проблемы окружающей среды и природных ресурсов,
№ 11, ВИНИТИ, Москва, 1999