Организация Объединенных Наций ECE/CES/2010/12

Экономический и Социальный Совет

Distr.: General
30 March 2010
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Пятьдесят восьмая пленарная сессия
Париж, 8−10 июня 2010 года
Пункт 6 предварительной повестки дня
Пространственная статистика

Комбинирование переменных пространственных данных с сетками с целью улучшения визуализации данных

Записка Бюро переписей Соединенных Штатов Америки

Резюме

В настоящей записке дается общий обзор юридической, статистической и административной географии, используемой Бюро переписей Соединенных Штатов. В ней анализируются схожие черты и различия между пространственными данными и статистическими сетками и обсуждаются преимущества и недостатки их использования. В некоторых случаях оба подхода комбинируются в виде интегрированного решения. В качестве примеров различных подходов к использованию пространственной статистики описываются два тематических исследования: анализ населения Гаити и сельскохозяйственная статистика Соединенных Штатов.

GE.10-21548 (R) 070510 100510

I. Введение

1. Все более широкая доступность пространственных данных в комбинации со статистикой низких уровней географии привела к повышению в последние годы интереса к использованию пространственной статистики. Статистический образ того или иного района представляет собой, как правило, многоугольник, формируемый административными единицами и геометрическими сетками. Решение о выборе формы зависит от цели представления данных, вида анализа, характеристик статистических и пространственных данных и характеристик графического материала, который будет представлять пользователь. Использование статистических сеток и административных многоугольников обладает как преимуществами, так и недостатками. В некоторых случаях в качестве интегрированного решения используется комбинация обоих подходов.

2. Характер элементов и характеристик пространственных данных, который лежит в основе переписного районирования, оказывает влияние на их использование. Пространственная статистика концентрируется на схемах и кластерах деятельности. В настоящем документе дается общий обзор юридической, статистической и административной географии, используемый Бюро переписей Соединенных Штатов Америки, которая сопоставляется с характеристиками геометрических сеток. В качестве примеров различных подходов к использованию пространственной статистики в нем приводятся два тематических исследования. В заключение излагаются вопросы для стимулирования дальнейшей дискуссии и будущих разработок.

II. Пространственные данные

3. Пространственные данные могут быть классифицированы по трем группам: геостатистические данные, данные точечных образов и сеточные данные (Cressie, 1993). Геостатистические данные представляют собой данные, собираемые по непрерывной пространственной области, которая привязывается к земле. Геостатистические данные характеризуются "Результатами наблюдений, связанных с непрерывной вариацией в пространстве, как правило, в функции рассеяния" (Anselin, 1992). Примером геостатистических данных могут являться образцы почв, собираемых в определенном регионе.

4. Когда интерес представляет то, в каком месте происходят события, речь идет о данных точечного образа. Данные точечного образа концентрируются на местоположениях индивидуальных точек данных и конкретно созданном пространственном образе (Cressie, 1993). Объекты пространственных данных точечного образа нерегулярно распределяются в пространстве. Данный тип пространственных данных не позволяет с уверенностью предсказать место появления события. Примером данных точечного образа могут служить местоположения жилищных единиц.

5. Сеточные данные собираются по регулярной или нерегулярной сетке с некоторой определяющей соседней структурой (Cressie, 1993). Район, в котором собираются точечные данные, обладает конечным числом мест сбора. Размер пространства ограничен. Значения присваиваются точкам данных, а местоположения точек данных известны. Примером сеточных данных может служить число жителей в каждом округе штата.

III. Пространственная статистика

6. Пространственная статистика, также называемая геостатистикой, является одной из форм статистики, анализирующей пространственные временные наборы данных. Пространственная статистика отличается от других форм статистики тем, что она касается местоположения значений данных. Все данные имеют пространственные и временные признаки. Близость этих данных зачастую является показателем схожести данных. Как указывал Уолдо Тоблер в своем "Первом законе географии", "все связано со всем, но более близко расположенные объекты связаны более тесно" (1970 год). Таким образом, данные, являющиеся более близкими друг другу с пространственной или временной точек зрения, по всей видимости, будут являться более схожими, чем более удаленные от них данные. Пространственная статистика использует различные методы изучения данных и их топологических, геометрических и географических признаков. Цель пространственной статистики заключается в определении величины пространственной изменчивости между точечными данными, которые изменяются в пространстве и/или времени. Пространственная статистка может использоваться для описания пространственных характеристик набора данных или для интерполяции какого-то заданного набора данных на районы, по которым существует мало либо никакой информации. В пространственной статистике каждое местоположение описывает пространственную схему, будь то в форме окружающей среды, климата, загрязнения, урбанизации или состояния здоровья населения.

7. Люди всегда пытались находить схемы в окружающем их мире. Таким образом, истоки пространственного анализа могут уходить глубоко в историю к началу географии, составления карт и геодезической съемки. Однако формализованное исследование пространственной статистки началось только во второй половине XX века. Сегодня пространственный анализ опирается на компьютерные методы благодаря наличию огромного объема географических данных, сложных программ статистического и географического анализа и передовых методов пространственного моделирования. Эти богатые данными среды являются результатом новейших технологий. Сегодня данные могут собираться с помощью методов дистанционного зондирования, интеллектуальных транспортных систем и мобильных устройств, оборудованных глобальными системами позиционирования (GPS), которые могут указывать местоположение практически в режиме реального времени. Благодаря развитию географических информационных систем (ГИС) управление большими объемами данных стало повседневной практикой. В результате этого пространственный анализ стал инструментом, доступным широкой аудитории. Это позволяет большему числу людей становиться аналитиками и рассчитывать и анализировать взаимосвязи и закономерности в данных и между ними.

8. Ввиду широкого изобилия имеющихся данных сегодня ведутся новые разработки в области хранения, представления, поиска, передачи и, что более важно, обобщения данных. Исследование пространственной статистики требует автоматизированных методов обобщения, классификации и прогнозирования или моделирования. Хотя пространственная статистика используется во многих дисциплинах, унификация найдена в использовании схем данных. Данные, имеющие значение в пространстве и во времени, зачастую связаны взаимодействиями и очевидны в пространственных схемах, которые служат ориентиром для исследований пространственной статистики. Общей целью пространственной статистики является выявление и изучение этих взаимосвязей и регули-

рующих схем, их классификация, а затем моделирование взаимосвязей и схем будущих данных.

9. Существует четыре фундаментальных вопроса, на которые пытается дать ответы пространственная статистика:

a) Каким образом распределяются данные?

b) Какой образ создают данные?

c) Где расположены кластеры?

d) Каковы взаимосвязи между наборами данных или значений?

IV. Статистические сетки

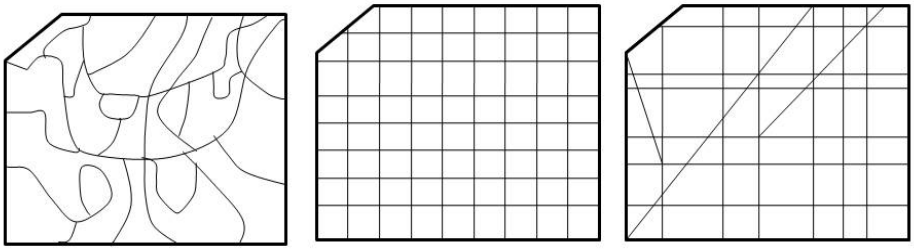
10. Статистические сетки представляют собой прямоугольные вместилища данных, которые обычно имеют одинаковые параметры и согласованный размер для конкретного вида использования. Создание сетчатого плана начинается с регулярного откладывания интервалов по осям x и y (горизонтальной и вертикальной). Сетки обеспечивают реляционные перспективы в рамках сетчатого плана от одной клетки ячейки к другой. В силу своей геометрической конфигурации ячейки являются масштабируемыми с точки зрения повышения или снижения разрешающей способности данных. Ячейка является держателем места, т.е. пространством для хранения единичных значений данных. Само по себе пространство клетки не обладает ни определением, ни значением.

11. Статистические данные применяются к ячейке сетки. Точечные данные рассеиваются по регулярной схеме или же по нерегулярной случайной схеме. Цель заключается в обеспечении того, чтобы одно место занималось только одной точкой данных. Для создания реальной ассоциации классифицированные данные применяются ко всей поверхности ячеек сетки и соотносятся с прилегающими ячейками, которые имеют либо тот же класс, либо различные классы. Схожести ведут к появлению кластеров и схем, в то время как различия свидетельствуют об уникальных или резко отклоняющихся событиях.

12. Наложение сетки на нерегулярную пространственную сеть (например, схемы мобильности) может позволить пользователю выявить взаимосвязи между двумя системами географических координат. Как правило, крупные сетки в сопоставлении с пространственными характеристиками меньшего масштаба (большие районы) улучшают сопоставимость взаимосвязей между данными. Антропоморфные характеристики, такие как транспортные сети (автомобильные и железные дороги), имеют нерегулярные и неупорядоченные формы. Природные характеристики, такие как горные гряды и реки, имеют схожие свойства нерегулярной ориентации.

13. Границы административных районов зачастую являются нерегулярными. В некоторых случаях прямоугольные административные районы, например исторические участки административного межевания земель и пастбищ в Соединенных Штатах, могут иметь одинаковую протяженность в пространстве. Вероятность одинаковой протяженности в пространстве зависит от многочисленных факторов, таких как происхождение, размер и цель сетки.

Рисунок 1
Сопоставление различных географических районов



14. Сетки служат подвижным окном переменного размера для изучения данных. Они также служат механизмом интеграции данных из других источников. В настоящее время интеграция данных является сложной задачей при использовании нерегулярных пространственных данных. Хотя сетки обладают хорошо определенными свойствами, они не соотносятся с нерегулярным характером явлений реального мира, такими как пространственные данные.

V. Схожести и различия между пространственными данными и статистическими сетками

15. Пространственные данные описывают явления реального мира. Некоторые данные являются естественными и преобладающе непредсказуемыми. Антропоморфные данные, как правило, являются непредсказуемыми. Однако некоторые характеристики создаются на основе схем, которые отвечают требованиям, таким как спецификации. Примерами могут служить схема посадки деревьев в саду, открытое пространство между противоположными направлениями дороги ограниченного доступа и расположение дорожных указателей расстояния через регулярные интервалы.

16. Пространственные данные меняются во времени и пространстве. Согласно Гриффиту и Пелинку (Griffith and Paelinck), данные реального мира страдают помехами, загрязненностью и неупорядоченностью (2007 год). Неопределенность (или непредсказуемость) данных усиливает помехи в данных. На месте можно обнаружить относительно засушливый ландшафт с неожиданным водоемом, питаемым природными ключами.

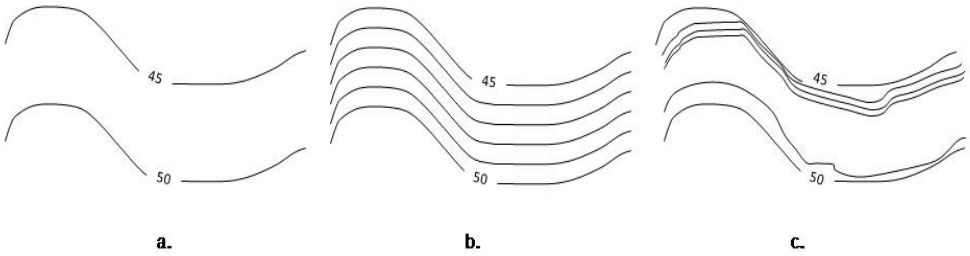
17. Загрязненные данные содержат различные противоречия, которые могут включать в себя неполные данные или резко отклоняющиеся значения и аномалии. Ведение полных и точных пространственных данных в реальном времени по мере развертывания событий и/или изменений невозможно в случае набора данных национального уровня. Даже в случае непрерывного зондирования весьма малое количество пространственных данных собирается в режиме реального времени. Неполнота данных обусловлена такими факторами, как ограничения источников, качество данных и наличие ресурсов. Резко отклоняющиеся значения являются крайними значениями. Так, например, температурные данные на любом конце температурной шкалы, опирающиеся на местные и необычные микроклиматические условия, могут указывать крайние значения. Под аномалиями понимаются исключения. Термин "округ" имеет единое определение для административно-территориальной единицы второго уровня. Термин "приход" в Луизиане эквивалентен округу и является исключением в отношении стандартного термина.

18. Неупорядоченные данные, как правило, зависят от условий наблюдения. Индивидуальные дома поддаются выявлению на местном уровне обычно с помощью спутникового изображения. Однако гараж, преобразованный в жилищную единицу, как правило, невозможно выявить в качестве таковой без дополнительной информации или посещения.

19. Пустое пространство на карте имеет пространственные характеристики, которые не указываются. Это явление усиливается при увеличении масштаба и размера пустого пространства на карте. Интерполяция осложняется характером пространственных данных. Условный расчет пространственных данных для заполнения пробелов или пустот на базовой карте схож с гаданием вслепую. Возьмем широко распространенный пример интерполяции контуров. Известные интервалы между изолиниями, значение возвышения вдоль изолинии определяются относительно точными фотограмметрическими и землемерными или геодезическими процессами. Значение возвышения между одной горизонталью и следующей горизонталью неизвестно (рис. 2a)). Можно интерполировать ориентацию линии, которая указывает на постепенное, регулярное изменение в возвышении (рис. 2 b)). Суть заключается в том, что крутизна или относительная стабильность возвышения неизвестны без точных расчетов (рис. 2 c)).

Рисунок 2

Сопоставление различных географических районов



20. В пространственных данных метаданные являются единственной разработкой, которая позволяет усовершенствовать использование данных и максимально облегчить усилия по интеграции геопространственных данных. Стандарты метаданных обеспечивают процесс для документирования информации, такой как качество, период, источник данных и другой необходимой информации по каждому элементу данных. Хотя они могут быть объемными, метаданные обеспечивают возможность принятия точных информированных решений в отношении использования и интеграции дискретных данных.

21. Сами по себе географические координаты являются недостаточными для определения пространственных данных. Характеристики придают дополнительное значение и обусловливают цель геоданных. Примеры характеристик, такие как схема классификации, географическое название и многочисленные другие дескрипторы обеспечивают полное характеризующее описание и дополнительную полезность географического явления.

22. Управление геопространственными данными является сложной задачей. Существуют многочисленные факторы ввода погрешности в пространственные данные. Поскольку географические данные привязаны к земле, одним из первых источников погрешности является ошибочная локализация данных. Другими факторами погрешности могут являться многочисленность атрибутов, опре-

деляющих элемент данных. Кроме того, геопространственные процессы сами по себе создают благоприятные условия для распространения погрешностей.

23. Точность пространственных данных является одним из искомых результатов. Обеспечение правильных географических взаимосвязей в контексте пространственных данных имеет первостепенное значение. Применение концепции топологии в геопространственных процессах обеспечивает соблюдение требования о поддержании корректных взаимосвязей между простыми геометрическими элементами геопространственной точки, линии и области. Регулярность формы или состояния географического района не создает проблемы для обеспечения корректных географических взаимосвязей. Так, например, топологические принципы обеспечивают взаимосвязь между точкой, описывающей жилищную единицу, и переписным участком независимо от формы и географического размера района. Жилищная единица расположена в своем переписном участке и указана довольно точно, даже если границы переписного участка не точно позиционированы.

24. Существуют различия между применением пространственной статистики и использованием статистических сеток. Визуализация этих подходов позволяет глубже понять их характеристики, различия и схожести. Были разработаны инструменты, помогающие определять наилучший подход в зависимости от целей анализа и использования данных.

VI. Географическая основа Бюро переписей Соединенных Штатов Америки

25. Бюро переписей Соединенных Штатов ведет географию на уровне переписных участков, которые являются наименьшей неделимой географической единицей, по которой собираются и составляются в табличной форме данные в рамках проводящиеся раз в 10 лет переписи. Дороги и другие видимые характеристики служат границами многогранников переписных участков. Другие характеристики, такие как границы городов (которые могут быть измеренной границей, не видимой на земле) также используются для делимитации переписных участков. Соединенные Штаты и их территории разбиты на миллионы переписных участков. Все земли распределены по переписным участкам.

26. В сопоставлении с согласованной формой статистических сеток переписные участки имеют нерегулярную форму. Выбор характеристик для использования в качестве границ переписных участков производится на основе четко определенных критериев. Хотя их форма носит нерегулярный характер, особенно вне зон городских улиц, размер многогранников участков находится в рамках общего допуска. Переписные участки, которые являются слишком большими, создают трудности при проведении полевых операций. Если учитывать все пространственные данные, то можно сказать, что существуют "загрязненные" участки, которые имеют аномальные характеристики. Так например, автомобильная и/или железная дорога может идти вдоль берега извивающейся реки в долине. Узкие многоугольники, сформированные такими непрерывными характеристиками, зачастую ведут к формированию вытянутых переписных участков.

27. Переписные участки также являются составными элементами всех других единиц переписного районирования, что объединяет их всех между собой. Многие другие уровни районирования имеют определенную общую форму вложенной взаимосвязи, т.е. более высокий уровень включает в себя часть бо-

лее низких уровней географии, находящихся в рамках его границ. Общим примером служит уровень округа. Округа состоят из переписных районов, которые содержат группы переписных участков, и в конечном итоге переписные участки, как это показано на рисунке 3 стандартной иерархии единиц переписного районирования.

Рисунок 3

Стандартная иерархия географических единиц переписного районирования

28. Существуют три типа переписных районов: юридические, статистические и административные. Юридические районы определяются другими уровнями управления. Так, например, границы самоуправляющегося города утверждаются избранными должностными лицами. Статистические районы определяются Бюро переписей Соединенных Штатов, зачастую по согласованию с партнерами из плановых организаций и схожих учреждений с целью создания значимых районов для табулирования данных. Примером административного района может служить школьный округ в рамках города.

29. Уровень точности применительно к конкретному географическому району зависит от различных факторов. Для делимитации юридических районов, таких как города, проводится съемка границ включенных земель для определения изменений в границах, а также новых включений. Качество информации зависит от источника поставщика данных и качества его процессов. Делимитация статистических районов опирается на определенные критерии. Поскольку толкование и местные интересы характеризуются различиями, результаты также являются различными.

30. Малый размер и практичность переписных участков делают их хорошими кандидатами для использования в качестве рабочей единицы в целях сбора, об-

работки и использования пространственной статистики. Использование сеток взамен данного конечно уровня географии, как правило, сужает возможности и добавляет сложность по сравнению с использованием переписных участков. В целом, чем меньшим является уровень переписного районирования, тем большие надежды могут возлагаться на повышенную точность.

VII. Совершенствование визуализации и анализа - тематические исследования

31. В одном из документов, который скоро выйдет в свет, Отдел народонаселения Бюро переписей Соединенных Штатов предпринял демографический анализ Гаити (Azar et al.). Целью этого проекта являлось картирование населения в масштабе 100-метровых ячеек сетки с использованием переписных данных и анализа спутниковых изображений. Итоги численности населения согласно переписи были распределены по ячейкам сетки исходя из площади антропоморфной непроницаемой поверхности (застроенные площади, такие как здания и дороги) в каждой ячейке. Сетчатая карта была затем усовершенствована с помощью онлайновых инструментов картирования (Бюро переписей Соединенных Штатов, 2010 год), позволяющих составление специфических разрезов данных и создание единой основы для общенационального анализа.

32. В одной предыдущей публикации "Сельскохозяйственный атлас Соединенных Штатов 1992 года" для более точного указания местоположения элементов данных в серии карт точечного распределения использовалась комбинация сеток и административной географии (Министерство сельского хозяйства США, 2010 год). Публикация содержала примерно 190 карт точечного распределения, а также более 120 тематических карт. Точечные карты Соединенных Штатов были подготовлены с использованием данных уровня округов.

33. Поскольку сельскохозяйственные характеристики подвержены изменениям, было очевидно, что в границах многих округов случайное размещение точек приведет к появлению сельскохозяйственной деятельности в тех зонах, в которых она невозможна или в значительной степени невероятна, например пастбищ в городских районах или зерновых в тундре. Источники включали в себя отдельные обобщенные файлы границ округов, береговых линий и городских районов. Для составления набора данных о землепользовании/почвенном покрове используется файл растровой сетки. На более низких уровнях в виде файлов многогранников покрова были созданы тематические данные о землепользовании/почвенном покрове, интегрированные в границы административного округа.

34. Сельскохозяйственные данные классифицировались по одной из пяти общих группировок: многоотраслевое сельское хозяйство; растениеводство; луга и пастбища; животноводство и сады. В рамках каждой категории землепользования (например, леса и лесопокрытые пастбища) каждой сельскохозяйственной группе присваивалось значение оценочной вероятности наличия (высокое в отношении пастбищ и лугов). Результирующему весу присваивался процент числа точек, размещенных в рамках части округа, относящийся к специфицированному виду землепользования. Каждая точка на карте представляла определенное количество элементов данных (значение точки).

35. Объединение двух различных типов данных позволило повысить точность представления картированных данных. Аналитики получили распределения данных, отражающие влияние землепользования на различные виды сель-

скохозяйственной деятельности. Аналогичные усилия могут принести более благоприятные результаты в тех случаях, когда сетчатые данные комбинируются с административной географией.

VIII. Выводы

36. Относительно недавно начавшаяся конвергенция данных, технологии, инструментов программного обеспечения и вычислительных мощностей открыла новые возможности для аналитиков с точки зрения изучения статистических данных применительно к их местоположению. Хотя базовые функции предлагались уже в рамках предыдущих географических информационных систем, в настоящее время наблюдается рост интереса к оценке распределений и кластеров, а также к прогнозированию будущей деятельности.

37. В тех случаях, когда в наличии имеются данные о малых географических районах, таких как переписные участки, рекомендуется рассматривать использование данного типа данных в качестве альтернативы статистической сетке, поскольку природа пространственных данных влияет на элементы статистических данных. Существуют также возможности объединения административных единиц со статистическими сетками, что может открывать новые возможности использования комплексных данных в целях анализа. В последние три десятилетия усилия были сосредоточены на построении наборов пространственных данных и использовании их характеристик традиционным образом. Визуализация эффекта пространственной статистики в различных формах открывает новые возможности для аналитиков с точки зрения углубления и расширения анализа.

IX. Справочные материалы

Anselin, L. (1992), Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences, Technical Report 92-10, National Center for Geographic Information and Analysis

Azar D., et al. (Forthcoming), Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti, *International Journal of Remote Sensing*

Cressie, N.A.C., (1993), *Statistics for Spatial Data*, Wiley: New York

Griffith, D.A. and Paelinck, J.H.P. (2007), An equation by any other name is still the same:

Spatial econometrics and spatial statistics, *Annals of Regional Science,* Vol. 41

Tobler W. (1970) A computer movie simulating urban growth in the Detroit region, Economic Geography, 46(2)

United States Census Bureau. 2010. Haiti Earthquake: United States Census Bureau Population Data. Online: https://www.geoint-online.net/community/haitiearthquake/default.aspx.

USDA. 2010. Agricultural Atlas of the United States. Online: http://www.agcensus.usda.gov/Publications/1992/Agricultural_Atlas/textfile/introduc.asc.