Знания-Онтологии-Теории (ЗОНТ-09)

Использование ГИС-технологии и таксономии для визуального анализа данных о субъектах РФ

А. М. Пуртов

Омский филиал Института математики СО РАН, ул. Певцова, д. 13, г.Омск, 644099, Россия

amp@ofim.oscsbras.ru

Аннотация. *В докладе на примере анализа показателей экономического развития субъектов РФ демонстрируется эффективность совместного использования методов геоинформационных систем и таксономии.*

Ключевые слова: Геоинформационные системы, таксономия, анализ данных, экономические показатели субъектов РФ.

1 Введение

Методы визуального интеллектуального анализа данных являются эффективным способом осознания фактов и принятия решений. Основные направления научных исследований в этой области связаны с развитием методов визуализации многомерных данных, автоматизацией интерпретации данных с помощью аналитических зависимостей, созданием интеллектуальных систем выявления закономерностей.

Технология геоинформационных систем (ГИС) в настоящее время является одной из наиболее востребованных и быстро развивающихся. Подсистемы визуализации и анализа больших массивов разнотипных данных о пространственных объектах считаются основными в полнофункциональных ГИС [1]. В ГИС существует даже специальное понятие – геоанализ, суть которого в визуальной оценке взаимного расположения объектов. Поэтому удивляет то, что работы по математическим основам анализа, представления, интерпретации данных редко используются в ГИС.

В докладе приводится пример использования технологии ГИС для визуальной эвристической таксономии (классификации, категоризации) субъектов РФ по экономическим показателям и для отображения результатов на карте. Эта процедура была апробирована при выполнении проекта, поддержанного фондом РГНФ (проект 04-01-12019в) «ГИС-карта археологических памятников юга Западной Сибири»[2], а также при выполнении курсовых проектов по дисциплине ГИС в вузах г. Омска.

2 Визуальная таксономия

По классификации задач анализа данных, предложенной Н.Г.Загоруйко [3], таксономия заключается в разделении «объектов по похожести их свойств». При этом в шкале наименований каждая группа похожих объектов как-то обозначается. Основным научно-практическим направлением в таксономии является автоматизация классификации в многомерном пространстве параметров (признаков). В настоящее время нет универсального алгоритма таксономии, «который мог бы составить реальную конкуренцию человеческой способности к обобщению» [4]. Пока в таксономии не создали «искусственного шахматиста», иногда при решении практических задач помогает собственный интеллект, каким бы он ни был. Без него трудно обойтись и при интерпретации результатов работы интеллектуальных систем.

Визуальная таксономия заключается в объединении объектов с помощью произвольных геометрических фигур по заранее сформулированному критерию. Основным недостатком метода является ограничение размерностью изображений, воспринимаемых человеком.

Достижения в области визуализации многомерных данных могут частично решить эту проблему[5]. Достоинства метода состоят в возможности использования обычных для человека процедур принятия решений. В этом случае можно освободиться от проблемы формализации, быстро менять цели, критерии, алгоритмы. Иногда люди, создающие и реализующие алгоритмы и люди, не знающие этих алгоритмов, могут гордиться друг другом. На рис. 1 представлены два независимо сделанных разбиения на таксоны.

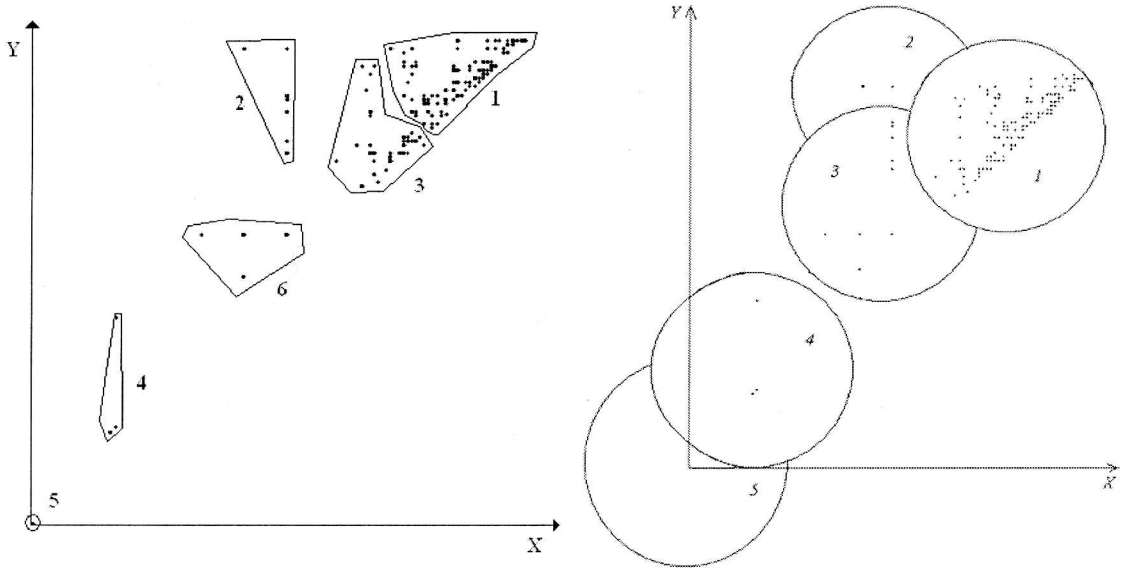

Рис. 1 Визуальная и алгоритмическая таксономия

На рис. 1 представлены нормализованные данные о датах начала (ось X) и окончания (ось У) проживания людей на территории археологических памятников Омской области [2]. В левой части рисунка показано визуальное разбиение на таксоны студенткой Гавриленко Ю. С. (СибАДИ, Омск), в правой части рисунка показано более позднее разбиение на таксоны с помощью программно реализованного алгоритма FOREL [3]. Рисунок демонстрирует похожесть разбиений и право на существования методов визуальной и автоматической (алгоритмической) таксономий.

3 Методика

Предлагаемая методика наиболее эффективна для анализа пространственных объектов, таких объектов, для которых важно, где они расположены. В первую очередь речь идет о географических объектах. Раскраска объектов по категориям средствами ГИС является одним из способов визуализации результатов предварительного анализа. В результате можно делать общие выводы, например, о том, что происходит в Европе, на Севере, за Уралом (с точки зрения Москвы), на Юге страны. Эта методика может быть применима и при предварительном определении логического двумерного пространства. Для выполнения анализа данных методами ГИС-технологии и таксономии можно предложить следующую последовательность действий.

1. Определение цели исследования.
2. Поиск, выбор данных.
3. Создание базы данных.
4. Нормализация данных (например, приведение всех данных в диапазон от 0 до 1000).
5. Выбор пар параметров для анализа.
6. Отображение объектов на плоскости средствами ГИС. Пара параметров является координатами объектов.
7. Выбор критериев для классификации (близость параметров, отличие от среднего, справедливость и др.).
8. Группировка объектов геометрическими фигурами в соответствии с выбранными критериями.
9. Обозначение таксонов символами (классификация в шкале наименований).
10. Создание для объектов столбца таблицы, в котором записываются обозначения таксонов.

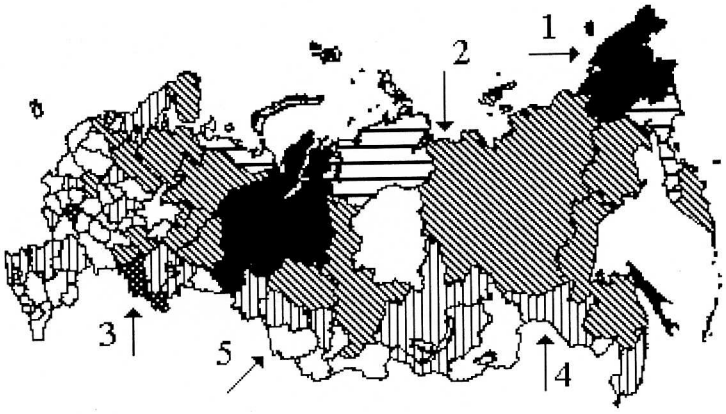

Рис.3 Визуализация классификации 1

На рис. 4 показана классификация субъектов РФ с точки зрения производительности и уровня бедности. Здесь критерием для объединения/разделения объектов выбраны средние значения параметров.

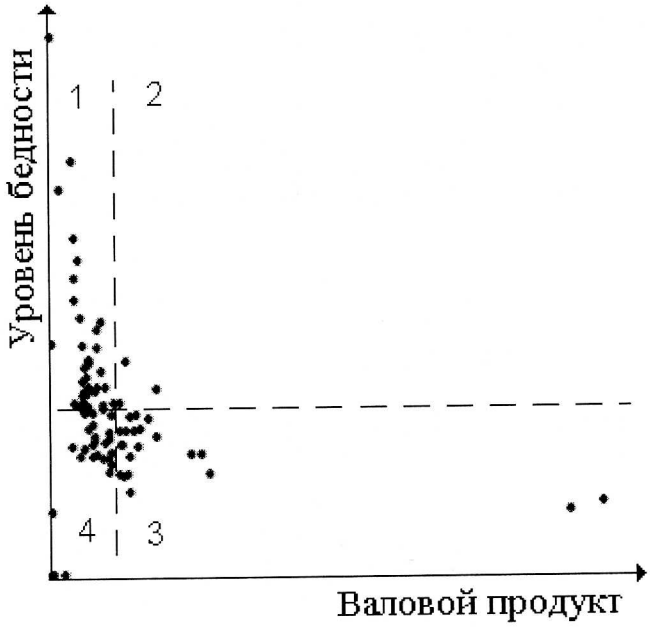

Рис. 4 Классификация 2

В первый таксон попали регионы с высоким уровнем бедности и низким уровнем производительности. Во второй таксон попали регионы с высоким уровнем бедности и высоким уровнем производительности (несправедливость). В третий таксон попали регионы с высоким уровнем производительности и низким уровнем бедности (справедливость). В четвертый таксон попали регионы с относительно низкими уровнями бедности и производительности. Раскраска регионов в соответствии с классификацией 2 приведена на рис. 5. Рисунок показывает, что регионы с высоким уровнем бедности расположены в основном на юге страны, на границе с Китаем и Монголией (таксон 1).

11. Раскраска объектов на карте в зависимости от таксонов, в которые они попали.

12. Проведение дальнейшего анализа на основе классификации и полученной тематической карты.

В современных программах для ГИС есть все необходимые средства для технической поддержки указанных действий.

4 Примеры

Для иллюстрации предложенной методики воспользуемся тремя экономическими показателями субъектов Российской Федерации [6]:

- валовой региональный продукт на душу населения (2006г.);
- среднемесячный доход на душу населения (по регионам, 2007г.);
- численность населения с доходами меньше прожиточного минимума (процент, 2007г.).

Нормализуем параметры (приводим параметры в диапазон от 0 до 1000) и считаем их координатами субъектов РФ на плоскости. Средствами пакета GIS ArcView 3.0 отображаем субъекты на плоскости (рис.2). Классифицируем объекты с помощью графики. Пунктирные линии показывают средние значения параметров. Цифры обозначают номера таксонов.

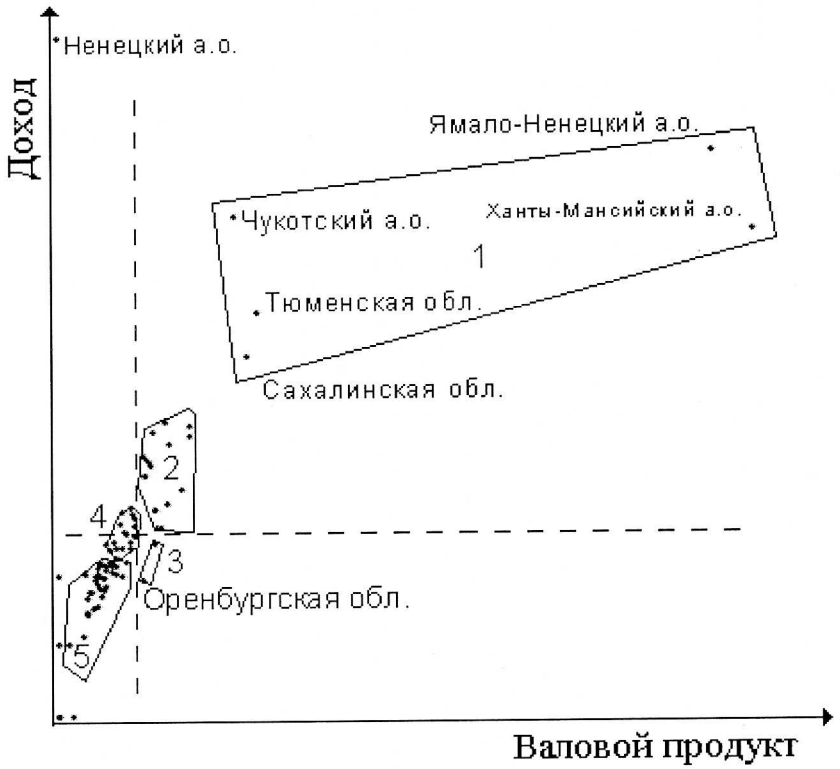

Рис.2 Классификация 1

В первый таксон попали регионы с очень высоким уровнем валового продукта и доходов на душу населения. Во второй таксон попали регионы с производительностью и доходами выше среднего. В третий таксон попали регионы с высокой производительностью и малыми доходами. В четвертый таксон попали регионы со средними показателями. В пятый таксон попали регионы с малыми уровнями производства и доходов. В нулевой таксон попали регионы, по которым отсутствуют данные (нулевой таксон не обозначен).

На рис. 3 показана раскраска регионов (средствами пакета GIS ArcView 3.0) в соответствии с проведенной классификацией. Таксоны 1, 2 показывают регионы с высоким уровнем валового продукта и дохода на душу населения. Это северные и восточные территории РФ, в которых развиты в основном добывающие отрасли. Европейские и южные территории, наиболее привлекательные для жизни, характеризуются относительно низким уровнем производительности и официальных доходов (таксон 5).

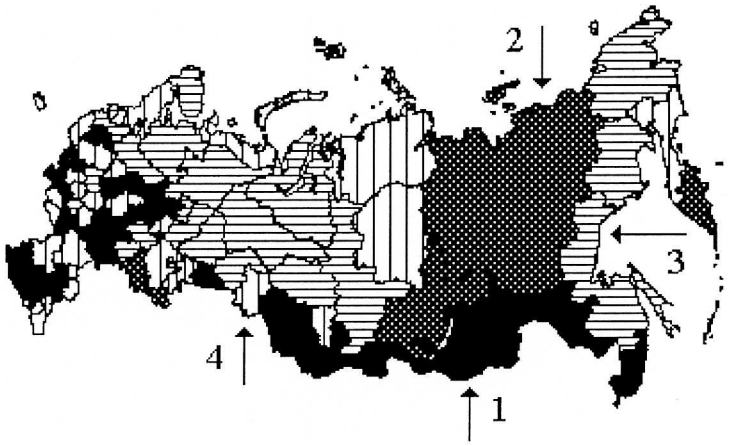

Рис.5 Визуализация классификации 2

Благополучные регионы с точки зрения относительно низкого уровня бедности и высокой производительности (таксон 3) расположены в центре страны и на Дальнем Востоке (ближнем западе для Японии).

5 Заключение

В докладе показана эффективность использования ГИС-технологии для процесса анализа и визуализация данных, результатом которого является получение новых знаний об объектах исследования. Основные недостатки визуальной таксономии связаны с субъективностью исследователя и проблемами работы в многомерных пространствах признаков. Интеграция человеческого интеллекта с методами автоматической таксономии и визуализации многомерных данных значительно повысит качество обработки фактической информации.

Литература

[1] ДеМерс, Майкл Н. Географические Информационные системы. Основы.: Пер. с англ. – М.: Дата+, 1999, 489с.

[2] А.М. Пуртов, С.Ф. Татауров, А.В. Шлюшинский. Разработка ГИС «Археологические памятники юга Западной Сибири». Омский научный вестник.- 2006. N7(43). – с. 136-139.

[3] Н.Г. Загоруйко. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270с.

[4] Борисова И.А., Загоруйко Н.Г. : Функции конкурентного сходства в задаче таксономии. Знания-Онтологии-Теории (ЗОНТ-07). Сб. науч. тр. Т.2 - Новосибирс, 2007. –с 67-76.

[5] Боровиков Н.П. STATISTICA. Искусство анализа данных на компьютере. Для профессионалов. 2-е изд. – ЗАО Издательский дом «Питер», 2003. – 344с.

[6] Социальное положение и уровень жизни населения России. 2008: С69 Стат.сб. / Росстат - М., 2008. – 502 с.