

4. Технология картирования произвольных данных

4.1. Упругие карты

Описывается технология визуализации произвольных данных. Рассматривается представление данных и алгоритм построения упругой карты, моделирующей данные. Предложены различные способы проектирования многомерных данных на двумерную сетку. Продемонстрировано применение методов визуализации произвольных данных на примере картографирования экономических таблиц.

4.1.1. Введение

В самых разных областях человеческой деятельности (в медицине, биологии, экономике и т.д.) исследователи сталкиваются с необходимостью осмысления больших таблиц данных, собранных в результате наблюдения за свойствами объектов той или иной природы. Как правило, такие таблицы содержат информацию о состоянии нескольких сотен или тысяч объектов, по каждому из которых известны значения определенного набора интересующих исследователя свойств. Число таких свойств (*признаков*) может также достигать нескольких сотен. Естественно, среди исследуемых объектов могут найтись такие, некоторые свойства которых неизвестны или недоступны для измерения. В таких случаях данные называются неполными или *данными с пробелами*.

Традиционным приемом при анализе таблиц является их представление, когда каждому исследуемому объекту сопоставляется точка в некотором абстрактном *многомерном пространстве данных*. Размерность этого пространства равна числу свойств-признаков в наборе, который характеризует состояние каждого из объектов. Значение каждой из координат точки в пространстве данных равно значению соответствующего признака, так что близким (в той или иной метрике) точкам в пространстве данных соответствуют объекты со сходными свойствами.

Таким образом, таблица данных представлена как облако точек в многомерном пространстве. Следует заметить, что такая картина не совсем верна в случае неполных данных. Так, например, если для объекта неизвестно значение только одной из координат, то его правильнее представлять как прямую, параллельную соответствующей координатной оси. Если на значение отсутствующего признака наложены априорные ограничения, то прямая превращается в отрезок. В случае, когда число неизвестных признаков более одного, то объект представляется в виде m -мерной плоскости (m – число неизвестных признаков), параллельной m координатным осям, или, соответственно, многомерным прямоугольным параллелепипедом.

Необходимый элемент анализа (осмысления) набора данных – это их красивое, наглядное и компактное описание. Более того, для многих

случаев построение такого описания уже решает задачу анализа данных. Если же в ходе такого анализа удастся решать вопросы о восстановлении пропущенных данных, то, как правило, этого достаточно для большинства приложений. На сегодняшний день известно два основных способа такого описания. Первый из них заключается в том, что точки данных разбиваются на несколько больших классов (кластеров), затем в классах ищутся более мелкие подклассы и т.д. В конечном итоге таблица данных описывается как иерархическая система кластеров в облаке точек [35, 74, 129, 130].

Второй способ – сокращение размерности описания данных. Дело в том, что человеческий мозг неспособен к эффективному анализу объектов размерности более трех, поэтому возникает необходимость в построении различных проекций многомерных данных, причем эти проекции малой размерности, как правило, выбираются так, чтобы максимально полно сохранить информацию об имеющихся в наборе данных закономерностях. Такой подход, например, характерен для традиционного линейного факторного анализа, когда в качестве такой проекции выбирается ортогональная проекция на подпространство, натянутое на несколько главных собственных векторов корреляционной матрицы (*главных компонент*) [129-133].

Под *визуализацией данных* можно понимать такой способ описания данных, когда размерность их описания сокращается до *двух* измерений. В этом случае данные можно изображать, например, в виде точек на экране монитора компьютера. Если в процессе сокращения размерности будет сохранена существенная часть закономерностей, присущих данным (разбиение на классы, отношения соседства), то исследователь получает возможность наглядно представить себе исходный набор многомерных данных, сделать выводы об их распределении.

Для визуализации многомерных данных разработан метод, заключающийся в построении вложенного в многомерное пространство данных двумерного многообразия, называемого *картой*, которое определенным способом *моделирует* или аппроксимирует данные (то есть большая часть точек данных лежит в окрестности карты). После этого точки данных с помощью определенной процедуры переносятся или *проецируются* на карту (см. рис. 4.1.). После того, как найдены положения всех образов точек данных на карте, каждому объекту из набора данных можно поставить в соответствие пару координат, характеризующих положение образа на двумерной карте.

На левом рисунке (рис. 4.1.) показано облако точек данных в пятимерном пространстве (показана проекция на первые две координатные оси), и карта, моделирующая эти данные. Белый кружок с линией отображает процесс проецирования многомерных данных на двумерное многообразие. На правом рисунке (рис. 4.1.) двумерная карта с нанесенными на нее образами данных «развернута» и каждому объекту соответствует теперь только две координаты образа на карте.

Сама идеология построения моделирующих карт была предложена

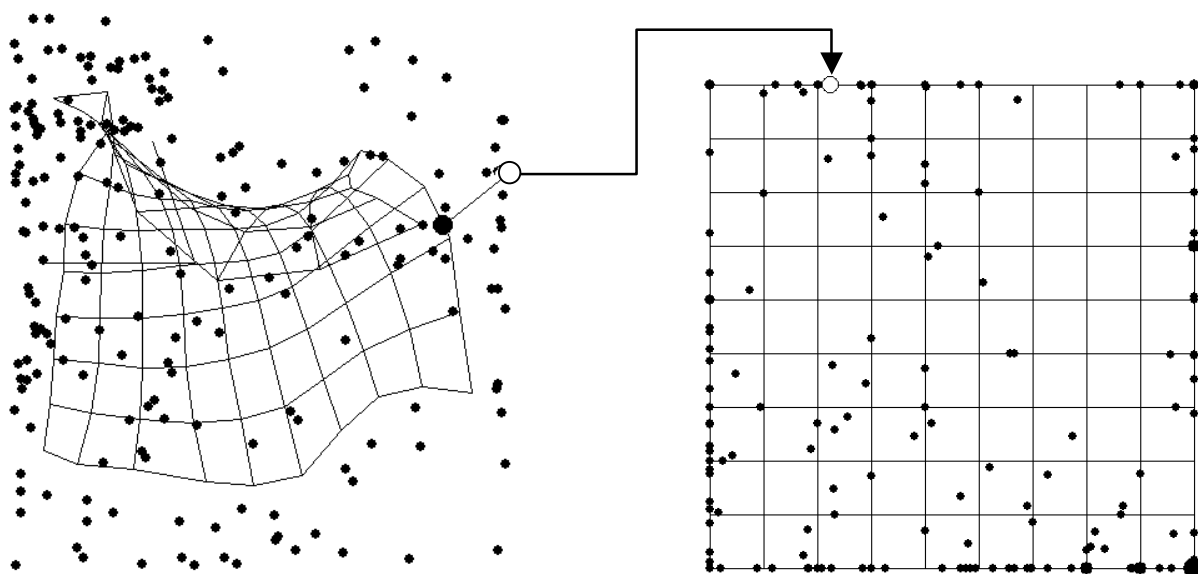


Рис. 4.1. Карта и ее развертка

Кохоненом [77]. Им же был разработан алгоритм построения самоорганизующихся карт (SOM – Self-Organizing-Maps), который с успехом применяется на практике. В дальнейшем алгоритм SOM совершенствовался в нескольких направлениях и на данный момент известно большое количество его модификаций. В данной работе предлагается иной алгоритм построения карт, названных *упругими* [68-71, 134, 135]. Кроме этого, описываются различные способы непрерывного проектирования данных на карту (а не только в узлы, как это было предложено Кохоненом), что позволяет посмотреть на принципы визуализации данных в ином ракурсе. Приведен пример картографирования экономических показателей двухсот крупнейших российских предприятий по данным журнала «Эксперт».

4.1.2. Постановка задачи

Проблема визуализации многомерных данных может быть разбита на несколько задач. Сформулируем эти задачи на математическом языке.

Рассматривается пространство R^n , в котором задано конечное множество точек X , которое интерпретируется как набор произвольных данных, полученных в результате наблюдений за состоянием исследуемой системы. Результат одного измерения (или набор признаков для одного объекта) изображаются точкой x^i ($i=1..N$, N – количество точек в X) в R^n . Расстояние между точками – обычное евклидово.

1) Построение карты

В целях исследования и визуализации множества X оно аппроксимируется двумерным многообразием. Можно указать два основных способа построения такого моделирующего многообразия. Во-первых, можно построить какое-либо стандартное линейное многообразие (например, плоскость двух первых главных компонент) и искать малые

отклонения от него. Такой подход можно назвать *квазилинейным*.

В работе предложен и используется принципиально нелинейный способ. В R^n размещается двумерная сетка G , расположение узлов которой r^{ij} , удовлетворяет определенным требованиям. Для упругой карты этими требованиями являются а) близость узлов сетки к данным; б) не слишком сильная «растянутость» сетки; в) не слишком сильная изогнутость сетки.

Будем считать, что на сетке изначально задана «внутренняя» ортонормированная система координат u, v так, что целым положительным значениям координат соответствуют вершины сетки (в этих координатах сетка является "прямоугольной"). Эти целые значения и используются в качестве значений индексов i, j для r^{ij} ($1 \leq i \leq n_1, 1 \leq j \leq n_2$, где n_1, n_2 – число узлов сетки по горизонтали и вертикали).

После того, как сетка G размещена в многомерном пространстве, она *доопределяется* до многообразия, которое обозначим через M . Это производится с помощью той или иной процедуры интерполяции между узлами. Самой простой такой процедурой является какой-либо вариант триангуляции, в результате чего получается кусочно-линейное многообразие M , составленное из треугольников.

2) Проектирование данных

Для того, чтобы можно было бы представлять X с помощью двумерного многообразия M , необходимо построить отображение $P: x \in X \rightarrow r \in M$, которое будем называть правилом проектирования или *проектором*. Для рассматриваемых задач желательно, чтобы проектор обладал следующими качествами:

а) проектор должен сохранять отношения соседства, то есть желательно, чтобы близким точкам в R^n соответствовали близкие точки на карте (по крайней мере, в некоторой окрестности сетки);

б) проектор, по крайней мере, в некоторой окрестности сетки должен быть однозначным;

в) проектор должен быть по возможности непрерывным, чтобы плавным изменениям состояния системы в X соответствовали непрерывные изменения положения образа в M .

г) проекция должна не слишком сильно отличаться от ближайшей вершины сетки G или, если доопределена карта, то от ближайшей точки карты.

Создателем SOM Кохоненом был применен самый очевидный и в некотором смысле естественный вариант проектирования – кусочно-постоянный. При этом каждой точке из X сопоставляется та вершина сетки, которая является ближайшей к этой точке:

$$P: x^k \rightarrow r^{ij}, x^k \in X, r^{ij} \in G, (x^k - r^{ij})^2 \rightarrow \min$$

Достоинством такого проектирования являются его логическая прозрачность и простота, очевидный его недостаток – разрывность, что не позволяет подробно изобразить картой структуру X . Тем не менее, такой вид проектирования успешно применяется на практике для раскраски карт

Кохонена и может изображать разбиение данных по кластерам, размеры кластеров и их взаимное расположение.

Для предложенной технологии построения упругих карт кусочно-постоянный способ проектирования малоприменим, поскольку, в отличие от SOM, каждый из узлов сетки G , вообще говоря, не располагается в центре локального сгущения точек данных. Напротив, упругая карта представляет собой более-менее равномерно натянутую на данные сетку, и поэтому существенная часть данных может быть расположена в промежутках между узлами. В этой работе рассмотрены два варианта кусочно-линейного непрерывного проектирования многомерных данных на двумерную карту.

4.1.3. Построение упругой карты

Основой для построения упругой карты является двумерная прямоугольная сетка G , вложенная в многомерное пространство, аппроксимирующая данные и обладающая регулируемыми свойствами упругости по отношению к растяжению и изгибу.

Расположение узлов сетки ищется в результате решения вариационной задачи на нахождение минимума следующего функционала:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{m} + \mu \frac{D_3}{m} \rightarrow \min, \quad (4.1)$$

где $|X|$ - число точек в X ; m - число узлов сетки (в случае прямоугольной сетки $m = n_1 \cdot n_2$); λ, μ - коэффициенты упругости, отвечающие за растяжение и изогнутость сетки соответственно; D_1, D_2, D_3 - слагаемые, отвечающие за свойства сетки, именно:

$$D_1 = \sum_{ij} \sum_{x \in K_{ij}} \|x - r^{ij}\|^2 \quad - \text{является мерой близости расположения}$$

узлов сетки к данным. Здесь K_{ij} - подмножества точек из X , для которых узел сетки r^{ij} является ближайшим (*таксоны*):

$$x \rightarrow r^{ij}, \|r^{ij} - x\|^2 \rightarrow \min, K_{ij} = \left\{ x \in X \mid x \rightarrow r^{ij} \right\};$$

$$D_2 = \sum_{ij} \|r^{ij} - r^{i,j+1}\|^2 + \sum_{ij} \|r^{ij} - r^{i+1,j}\|^2 \quad - \text{мера растянутости сетки};$$

$$D_3 = \sum_{ij} \|2r^{ij} - r^{i,j-1} - r^{i,j+1}\|^2 + \sum_{ij} \|2r^{ij} - r^{i-1,j} - r^{i+1,j}\|^2 \quad - \text{мера}$$

изогнутости (кривизны) сетки.

Алгоритм решения вариационной задачи напоминает метод динамических ядер. На каждом шаге алгоритма:

1) Квадратичный функционал D минимизируется при заданном разбиении множества X на таксоны - подмножества точек K_{ij} , для которых данный узел сетки является ближайшим. При фиксированном разбиении

на таксоны функционал D квадратичен по переменным r^{ij} , поэтому задача нахождения его минимума сводится к решению системы линейных уравнений с матрицей размерами $m \times m$, причем коэффициенты самой матрицы системы зависят только от положения узлов r^{ij} , а сама матрица имеет очень простую – трехдиагональную структуру заполнения.

2) Производится новое разбиение множества X на таксоны, соответствующее новым положениям узлов сетки.

Общий критерий D уменьшается при каждой итерации алгоритма, он естественно ограничен снизу нулем, а число разбиений конечного множества X по таксонам конечно. Следовательно, алгоритм сходится.

При построении карты возникает общая для всех оптимизационных задач проблема нахождения глобального минимума функционала D , связанная с выбором начальной карты. В работе не предлагается ее абсолютного решения, однако, достаточно разумных экстремальных значений можно достичь следующим образом:

1) Первоначально карта располагается в плоскости первых двух главных компонент. Ее размеры выбираются равными порядка $1/2 - 1/3$ от длины двух главных полуосей эллипсоида рассеяния.

2) К положениям узлов добавляется малый случайный шум для того, чтобы вывести расположение узлов из одной плоскости и избежать возникновения различных вырожденных случаев.

3) При нахождении минимума функционала D параметры λ , μ меняются от больших значений (порядка 10) к малым (порядка 0,01). В результате карта в начале расчета очень «жесткая», ее кривизна невелика, в результате чего моделируются самые общие особенности распределения точек данных. При уменьшении коэффициентов упругости карта становится «мягкой» и ее узлы располагаются к точкам данных гораздо ближе. Такой метод можно назвать методом *отжига*.

В результате в пространстве данных располагается относительно равномерная сетка G , аппроксимирующая данные. После проектирования данных на сетку расстояния между образами и их скоплениями на карте отражают соответствующие особенности распределения данных в многомерном пространстве.

4.1.4. Проектирование многомерных данных на двумерную сетку

После построения сетки G , которая является точечной аппроксимацией множества X , она должна быть доопределена до многообразия. Самым простым способом является выбор какого-либо способа триангуляции, один из вариантов которой показан на рисунке 4.2.

В результате в многомерном пространстве данных располагается двумерное кусочно-линейное многообразие, на котором соседние вершины соединяются ребрами, а ребра ограничивают плоские грани сетки.

В качестве одного из самых очевидных способов проектирования многомерных данных на двумерную сетку можно предложить нахождение ближайшей точки карты (а не только ближайшей вершины, как в случае SOM). Задача состоит в нахождении ближайшей к данной точке данных x

точки кусочно-линейного многообразия M . Заметим, что ближайшая точка карты может оказаться на грани сетки, на ребре или в вершине сетки. Таким образом, пространство вокруг карты разбивается на области, где ближайшей точкой карты является точка грани, ребра или вершина.

Прежде всего, определим ближайшую к x вершину сетки и обозначим ее радиус-вектор как r^0 , а окружающие ее вершины перенумеруем изображенным на рисунке 2 образом.

Помимо ближайшей вершины r^0 , можно выбрать еще две наиболее близкие к x вершины r^k и r^l из ближайшего окружения r^0 , причем таким образом, чтобы вершины r^0 , r^k , r^l являлись вершинами грани карты при выбранном способе триангуляции. Таким образом, определяем *ближайшее ребро и ближайший треугольник* (грань) карты.

Алгоритм нахождения ближайшей точки карты заключается в следующем:

1. Выполняется ортогональное проектирование x на плоскость, натянутую на вектора $m=r^k-r^0$ и $n=r^l-r^0$. Образом x окажется точка $r^* = r^0 + um + vn$, где

$$u = \frac{(x-r^0, m) \cdot n^2 - (n, m)(x-r^0, m)}{m^2 n^2 - (m, n)^2}, \quad v = \frac{(x-r^0, n) \cdot m^2 - (n, m)(x-r^0, n)}{m^2 n^2 - (m, n)^2}.$$

В случае, если $u \geq 0, v \geq 0, u+v \leq 1$, то ближайшей точкой карты является точка r^* . Иначе:

2. Выполняется ортогональное проектирование на ближайшее ребро, натянутое на вектор $m=r^k-r^0$. При этом

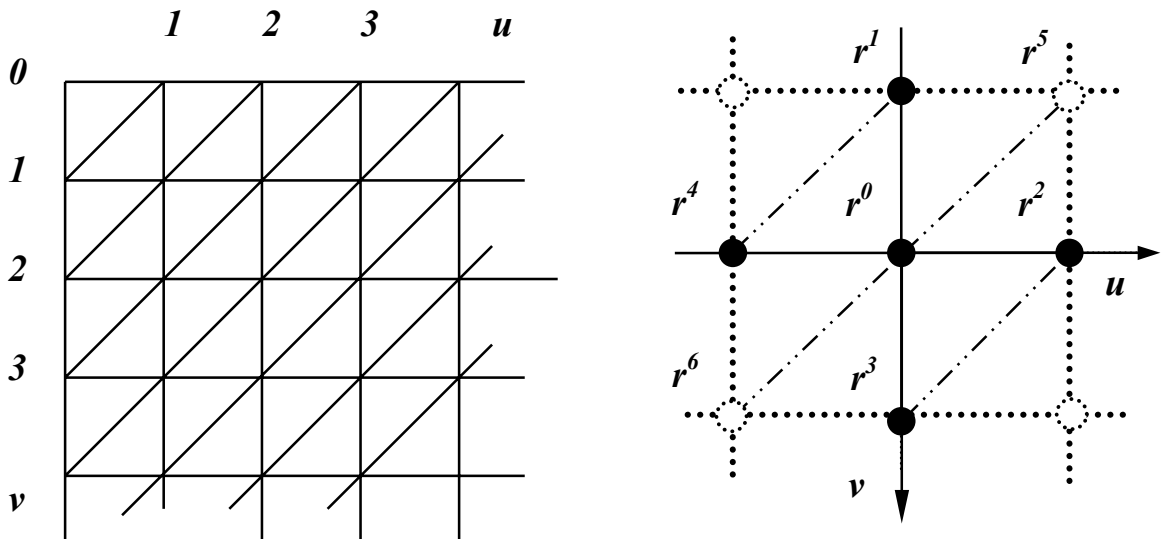


Рис. 4.2. На левом рисунке показан один из вариантов триангуляции сетки. На правом – выбранный способ нумерации узлов, окружающих ближайший к данной точке данных узел.

$$x \rightarrow r^{**} = r^0 + um, \quad u = \frac{(x - r^0, m)}{m^2},$$

и если $0 \leq u \leq 1$ - ближайшей точкой карты является r^{**} . Иначе ближайшей точкой карты является вершина r^0 .

Другой вариант кусочно-линейного проектора. Для случая моделирования конечного множества точек данных с помощью самоорганизующейся кривой (SOC) [131-133] был предложен вариант центральной проекции на одномерную сетку. Точки данных сначала ортогонально проектируются в плоскость трех вершин на сетке – ближайшей к заданной точке данных и двух ее соседей, после этого на плоскости выполняется центральное проектирование, в качестве центра проекции O выбирается точка пересечения серединных перпендикуляров к ребрам ломанной, непосредственно прилегающих к ближайшей к точке данных x вершине (см. рис. 4.3).

Можно перенести этот вариант проектирования на случай двумерной карты, если считать, что проектирование поочередно выполняется сначала для одной, а потом для другой координатной линии двумерной сетки. Таким образом, можно определить значения внутренних координат u, v для точки проекции.

Следует заметить, что для описанных выше алгоритмов (как при

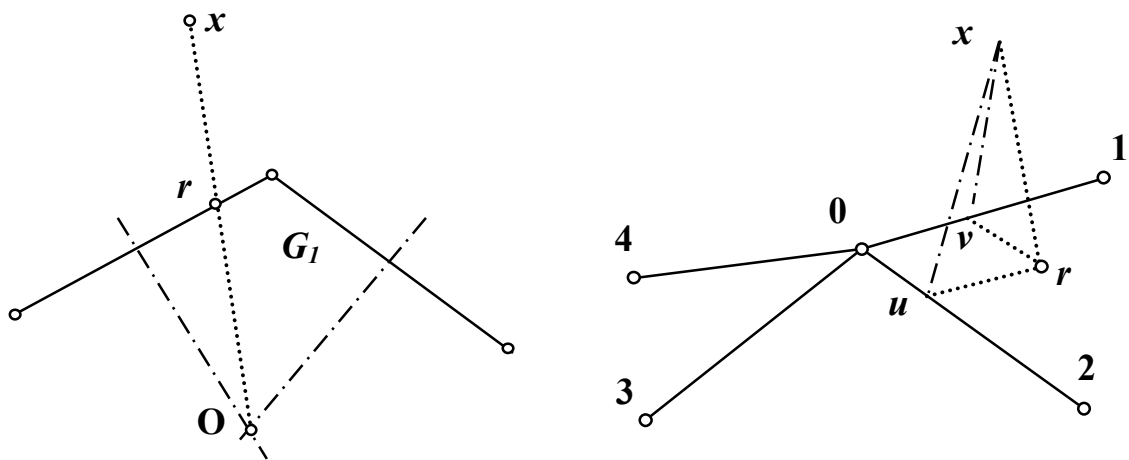


Рис. 4.3. На правом рисунке показан способ нахождения центральной проекции точки x на ломаную, которая является одной из координатных линий u, v двумерного многообразия M . На левом рисунке изображен способ нахождения координат на карте образа точки x .

построении карты, так и при проектировании) не является принципиальным наличие в данных пробелов. В случае, если по какой-либо точке данных известен не полный набор признаков, то вычисление всех необходимых расстояний от точки данных до карты выполняется в пространстве меньшей размерности, где для выбранной точки данных известны все координаты. Таким образом, на карте находят свое

положение также и неполные данные, что позволяет использовать изложенные алгоритмы в качестве средства восстановления пробелов в данных. Хотя практика показывает, что точность восстановления данных при таком способе невысока, тем не менее, пользователь может получить представление о месте данных с пробелами среди всей совокупности данных. Кроме того, ниже рассматривается способ существенно увеличить точность восстановления данных.

4.1.5. Использование карты для анализа распределения данных

После того как карта построена, ее можно «развернуть» (см. рис. 4.1.) и наносить разного рода слои информации [68-71, 134, 135]. В результате карта приобретает разные раскраски, дающие представление об исходном распределении данных. Такие раскраски могут быть построены с использованием богатого арсенала средств и методов ГИС. Таким образом, открывается новое широкое поле деятельности для использования ГИС-технологий для картирования информации самого разного происхождения.

Во-первых, на карте можно изобразить сами данные. При этом можно отображать различные разбиения на подмножества данных, в соответствии со значением того или иного признака.

Во-вторых, на карте можно изобразить произвольные функции координат, поскольку каждой точке с координатами u, v на двумерной карте соответствует точка в многомерном пространстве данных. Самыми простыми раскрасками являются раскраски, отражающие значение той или иной координаты в точках размещения карты. Кроме этого, на карте можно отображать такие координатные функции, как плотность распределения данных в пространстве или плотность того или иного подмножества данных. Сама плотность может быть рассчитана с помощью какой-либо непараметрической оценки. Кроме собственно плотностей подмножеств интерес могут представлять раскраски, отвечающие значению относительных плотностей подмножеств на фоне общего распределения.

В-третьих, на карте можно изображать различные свойства самого моделирующего многообразия. Например, можно получить раскраски, отражающие значения метрических коэффициентов карты или ее кривизны, свидетельствующие о мере ее растянутости и изогнутости на разных участках. Также интерес представляет такая раскраска карты, на которой цветом отражены расстояния от точки карты до ближайшей точки данных в многомерном пространстве. Такая раскраска позволяет оценить степень близости моделирующей карты к точкам данных в разных областях пространства.

Наконец, весьма плодотворной оказалась идея составления по множеству X нескольких карт, которые можно по аналогии с начертательной геометрией образно назвать «главными проекциями множества». Первая из таких карт визуализирует сами данные, вторая «главная проекция» визуализирует погрешности аппроксимации данных (то есть разности между положениями точек данных и их образами на

карте) и позволяет оценить точность описания данных первой картой, и так далее. Последовательность таких «главных проекций» позволяет моделировать данные с высокой точностью, что в случае неполных данных позволяет правдоподобно восстанавливать пропущенные или ремонтировать недостоверные данные.

4.2. Применение методов визуализации произвольных данных к картографированию экономических таблиц

В качестве примера применения описанной технологии метод визуализации произвольных данных был применен для картографирования таблицы крупнейших российских предприятий, взятой из журнала «Эксперт-200» [136]. Файлы исходных данных были получены с официального сайта журнала <http://www.expert.com>.

Исходная таблица содержала информацию об экономическом положении двухсот крупнейших российских предприятий, ранжированную в порядке убывания валового объема производства продукции. Изначально таблица содержала следующие поля-признаки (часть из них является независимыми признаками, часть рассчитывается по явным формулам):

- Название предприятия;
- Регион местонахождения предприятия;
- Отрасль, к которой относится предприятие;
- Валовой объем производства продукции в 1998 году;
- Валовой объем производства продукции в 1997 году;
- Темпы роста предприятия
- Валовой объем производства в 1998 году, выраженный в долларовом эквиваленте;
- Балансовая прибыль предприятия;
- Прибыль предприятия после налогообложения;
- Прибыльность предприятия;
- Число работающих на предприятии;
- Производительность труда.

В работе [137] уже была предпринята попытка визуализации таблицы предприятий, взятой из журнала «Эксперт» за 1997 год. В этой работе были использованы традиционные самоорганизующиеся карты Кохонена и диаграммы Хинтона. Там же было предложено использовать в качестве координат пространства данных отношения некоторых независимых признаков из таблицы. Было предложено четыре таких координаты.

В данной работе было решено расширить пространство исходных данных еще одним измерением, в результате чего был получен набор независимых признаков.

Таблица 1

Описание полей обрабатываемой таблицы		
Обозначение признака	Значение	Условное Название

LG_VO1998	Логарифм валового объема производства продукции в 1998 году	Размер
TEMP	Валовой объем производства продукции в 1998 году / Валовой объем производства продукции в 1997 году	Темп роста
PROFIT_BAL	Балансовая прибыль предприятия / Валовой объем производства продукции в 1998 году	Балансовая прибыль
PROFIT_NAL	Прибыль предприятия после налогообложения / Валовой объем производства продукции в 1998 году	Чистая Прибыль
PRODUCTIV	Прибыль предприятия после налогообложения / Число работающих на предприятии	Производительность

В результате была составлена таблица из двухсот записей с пятью полями. Часть записей содержала неполную информацию (по отдельным признакам информация отсутствовала).

Данные были предварительно нормированы по формуле $\tilde{x}_i = th\left(\frac{x_i - M}{\sqrt{D}}\right)$, где $\tilde{x}_i, x_i, M, D$ – новое, старое значения признака, среднее значение и дисперсия признака соответственно, th – гиперболический тангенс.

Карта, с помощью которой осуществлялась визуализация множества данных, была построена по описанному выше алгоритму построения упругих карт. Первоначальная сетка содержала 10 узлов по вертикали и 10 по горизонтали. Для нахождения локального минимума функционала применялся описанный выше метод отжига. Параметры μ и λ медленно (так чтобы при каждом изменении карта успевала перейти в близлежащий локальный минимум) менялись от значений $\mu = 5, \lambda = 5$ до $\mu = 0.01, \lambda = 0.01$.

После построения упругой карты, данные из пространства признаков, были спроектированы на карту, с помощью описанной выше процедуры нахождения ближайшей точки карты в случае кусочно-линейной интерполяции между узлами.

В качестве иллюстрации анализа экономических данных ниже приведены раскраски полученной карты по координатным полям, а также слой рассчитанной плотности данных в точках карты. На раскрасках большими точками с номерами выделена группа предприятий, принадлежащих нефтегазовой промышленности. Такое выделение позволяет проанализировать место той или иной отрасли промышленности среди других предприятий.

4.2.1. Раскраска по признакам

На рис.4.4а изображено значение признака LG_VO1998 в точках карты. При этом более светлым участкам соответствуют более высокие показатели признака. Самый яркий цвет соответствует первым 10%

предприятий с самым большим валовым объемом производства. Для примера кружками с цифрами выделены предприятия нефтегазовой промышленности. Цифрам соответствуют следующие названия предприятий:

1 – ОАО «Газпром»; 2 – Нефтяная компания «ЛУКОЙЛ»; 3 – Башкирская топливная компания; 4 – Нефтяная компания «Сургутнефтегаз»; 5 – Тюменская нефтяная компания; 6 – «Татнефть»; 7 – Нефтяная компания «Славнефть»; 8 – Нефтяная компания «Роснефть»; 9 – Оренбургская нефтяная компания «Онако»; 10 – Центральная топливная компания; 11 – Нефтяная компания «КомиТЭК».

Рис.4.4б изображает раскраску по показателю TEMP. Как видно из рис.4.4б, область крупнейших предприятий не пересекается с областью наиболее высоких темпов роста. В правом нижнем углу карты, например, располагаются предприятия пищевой промышленности, цветной металлургии и другие, быстро развивающиеся отрасли.

На рис. 4.4в, 4.4г, 4.4д показаны раскраски по признакам PROFIT_BAL, PROFIT_NAL, PRODUCTIV. Раскраски этих признаков схожи, что указывает на корреляцию последних трех признаков. Вместе с этим различия в раскраске позволяют выделить предприятия, которые

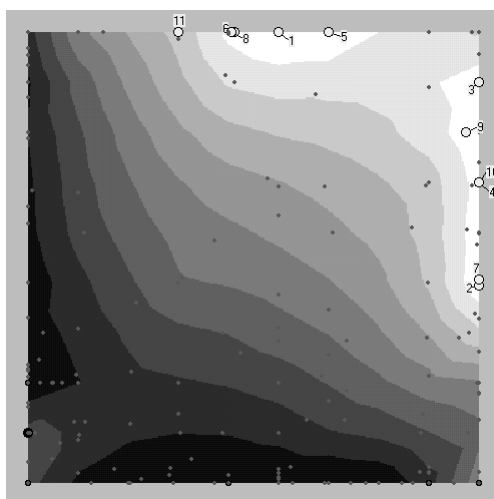


Рис. 4.4а

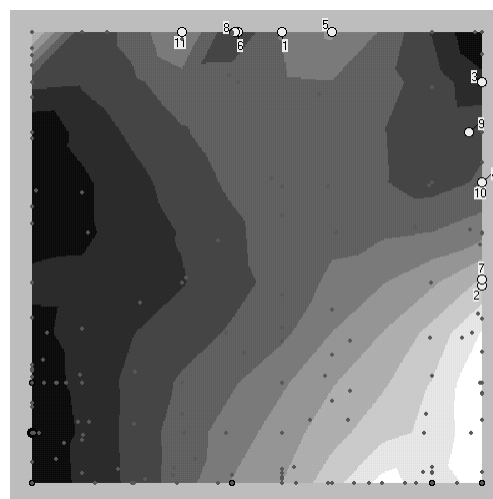


Рис. 4.4б

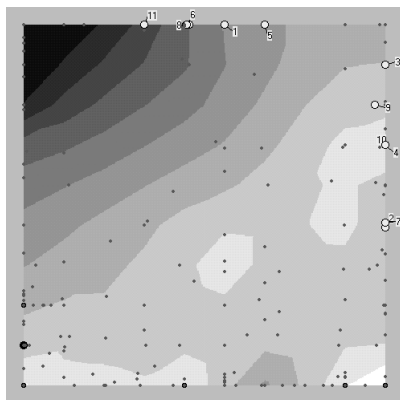


Рис. 4.4в

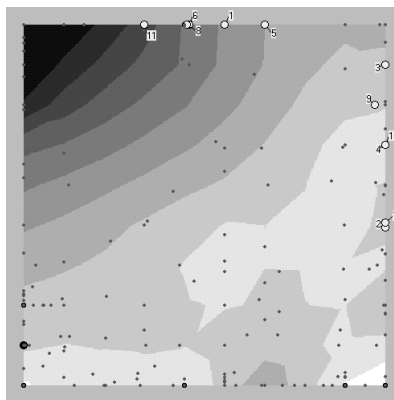


Рис. 4.4г

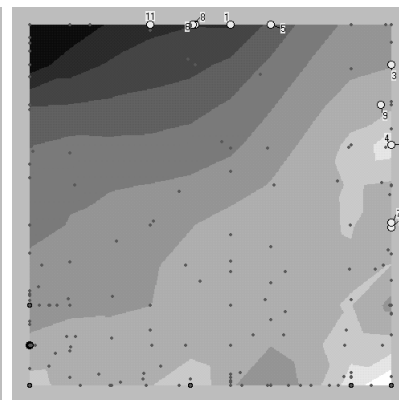


Рис. 4.4д

выпадают из корреляционной зависимости.

4.2.2. Раскраска по плотности данных

На рис. 4.5а, 4.5б, 4.5в показана раскраска карты по плотности

данных, оцененной с помощью какой-либо непараметрической оценки. Существует два способа оценить плотность данных. Во-первых, можно рассматривать двумерное распределение точек на карте. Во-вторых, можно рассчитать плотность точек в исходном n -мерном пространстве, и изображать на карте значения этой плотности в точках расположения карты. На рисунках изображено применение первого способа. Более темным участкам соответствуют более высокие значения плотности.

Рис.4.5а изображает двумерное распределение общей плотности данных. На рис.4.5б – распределение плотности предприятий нефтегазовой промышленности. Рис.4.5в отражает удобную для оценок относительную плотность предприятий нефтегазовой промышленности (то есть отношение первых двух плотностей).

На рис.4.6 отражено расстояние от каждой из точек карты до ближайшей точки данных. Более темным участкам соответствуют большие

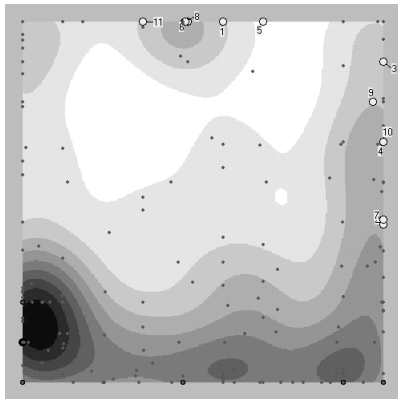


Рис. 4.5а

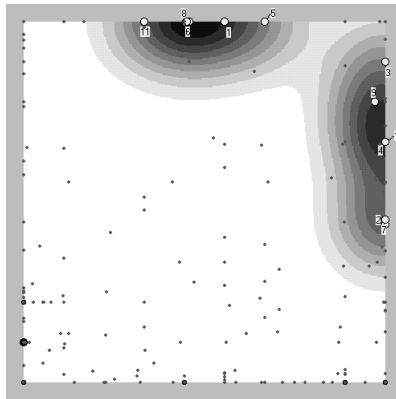


Рис. 4.5б

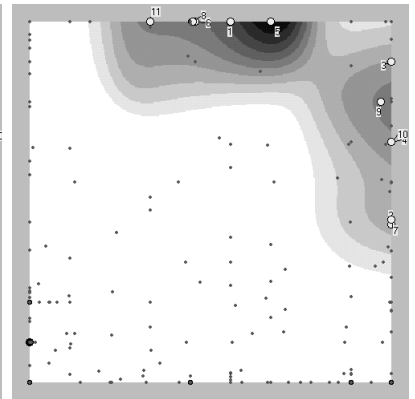


Рис. 4.5в

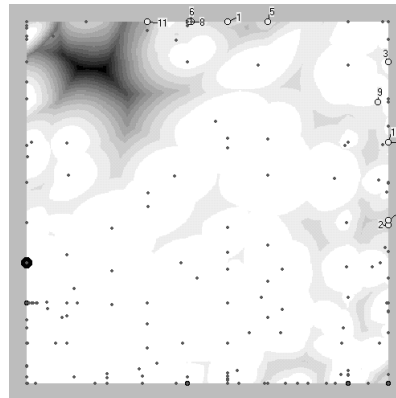


Рис. 4.6

расстояния. Видно, что в целом данные достаточно плотно прилегают к карте, за исключением участка в левом верхнем углу (впрочем, точки данных там отсутствуют и темный цвет указывает на то, что точки в левом верхнем углу карты расположены в n -мерном пространстве достаточно далеко от основного массива данных).

Беглый взгляд на рисунки позволяет сделать, например, такие выводы. Предприятия нефтегазовой промышленности являются лидерами по объему валового производства, но темпы роста этой области промышленности невелики по сравнению, например, с пищевой промышленностью. Предприятия нефтегазовой промышленности распадаются на две группы, которые существенно отличаются по прибыльности производства. В целом, набор таких рисунков могут служить удобным средством анализа для специалистов в макроэкономике.

На приведенных рисунках карты изображены во внутренних координатах сетки, в которых сетка является прямоугольной. Можно также изображать карту в проекции на координатные плоскости: данные проецируются на карту, а карта, в свою очередь, на одну из координатных

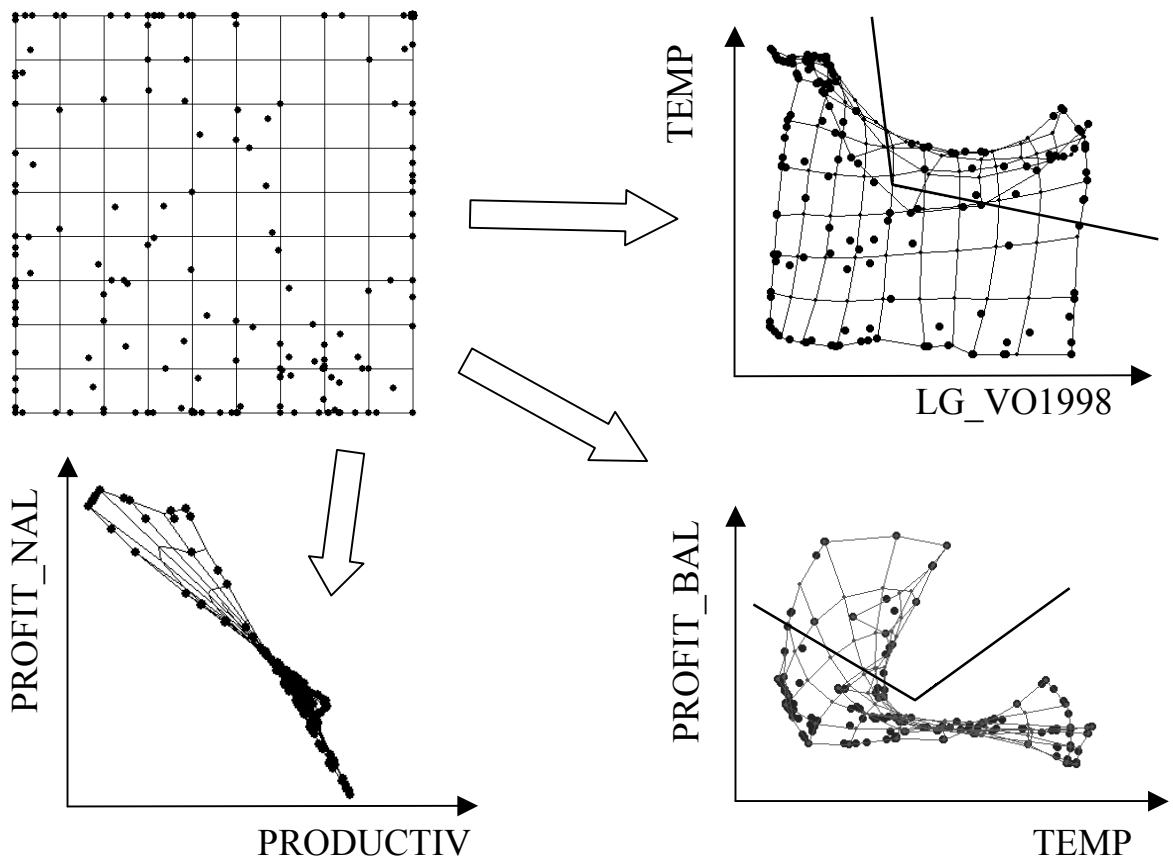


Рис. 4.7. Слева сверху изображена развернутая карта во внутренних координатах. Справа и ниже изображены три проекции карты на три разные пары координатных плоскостей. Две линии, расположенные вне карты, показывают направления первых двух главных компонент.

плоскостей (см. рис.4.7). Еще один из многих вариантов проекций – проекции на плоскости главных компонент.

4.2.3. Восстановление данных, регрессия и прогноз

Как уже упоминалось, данные, предоставляемые для визуализации, могут содержать недостающие значения. Тем не менее, наличие пробелов

в данных нисколько не препятствует построению упругих карт. Все описанные алгоритмы претерпевают минимальную модификацию, которая заключается в следующем: все расстояния между точками при наличии пробелов рассчитываются в пространстве меньшей размерности, где значения признаков известны полностью. На практике это означает, что в

формуле расчета расстояния $r(x, x^0) = \sqrt{\sum_{i=1}^N (x_i - x_i^0)^2}$, где x – точка, чьи

координаты известны не полностью, в сумме исключаются те слагаемые, для которых значение координаты x_i неизвестно. Это замечание относится к вычислению расстояния от точки до узла сетки и к вычислению сумм в функционале D_I .

Таким образом, по неполным данным можно построить упругую карту, положения узлов которой в пространстве известны *полностью*, что позволяет использовать такую карту в качестве средства для восстановления данных с пробелами.

Задачу восстановления данных можно сформулировать следующим образом:

Пусть дано конечное множество точек в многомерном пространстве, причем значения отдельных координат некоторых точек неизвестны. Требуется построить математическую модель множества, с помощью которой можно было бы правдоподобно восстановить недостающие значения координат. К сожалению, не так просто точно определить понятия *правдоподобия* [35]. Поскольку при построении модели данных не используется никаких гипотез о подчинении распределения данных какому-либо статистическому закону (не существует никакой генеральной совокупности), применение статистических оценок не является уместным. Единственным способом оценить правдоподобность процедуры восстановления значений координат является эмпирическое тестирование модели данных, когда на время делается предположение о том, что значение той или иной координаты неизвестно, затем координата восстанавливается с помощью модели и сравнивается с ее известным «истинным» значением.

С практической точки зрения задача восстановления данных может быть сформулирована как задача *регрессии*, то есть построения эмпирических зависимостей одних данных от других. С другой стороны, методы заполнения пробелов в данных могут быть использованы с целью *прогноза*, то есть предоставления правдоподобной оценки каких-либо параметров изучаемой системы при определенных предположениях о других параметрах.

Естественно, что значения восстановленных координат зависят от метода построения моделирующего многообразия и от способа проектирования данных на него. Так при использовании в качестве моделирующего многообразия SOM точка данных, отдельные координаты которой подлежат восстановлению, сопоставляется с ближайшим *узлом*

сетки, что означает замену точки на центр локального сгущения точек данных в пространстве. При использовании упругих карт также возможно кусочно-постоянное проектирование, однако более естественным являются кусочно-линейные способы проектирования (например, алгоритм нахождения ближайшей *точки карты*), описанные выше. Можно ожидать, что в случае кусочно-линейных проекторов точность восстановления будет выше.

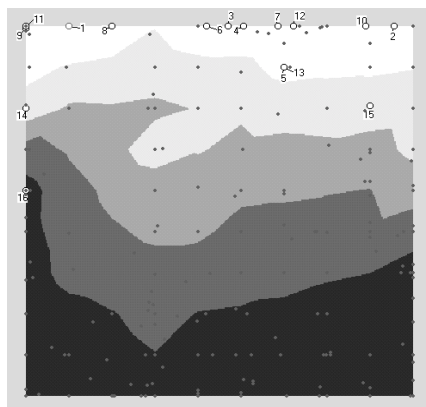
Точность восстановления пропущенных данных можно увеличить, делая упругую карту все более «мягкой». Однако возможности такого подхода ограничены, прежде всего тем, что количество узлов карты ограничено. Кроме того, совершенно «мягкая» карта фактически начинает моделировать случайный шум. Более перспективным является уже упомянутое «мультикартирование», когда после построения основной карты данных для каждой точки вычисляются «остатки» – разности между положением точки и ее проекцией на карту. Полученные остатки также картографируются, в результате чего возникает карта погрешностей. Процедура может быть продолжена до получения необходимой точности.

Рассмотрим процедуру восстановления данных с помощью нескольких карт более формально. Обозначим через M первую карту (то есть карту, построенную по самим данным), а через $P(x)$ – проектор на M , то есть функцию, сопоставляющую произвольной точке данных $x \in X$ точку $r \in M$. Обозначим через X^n , M_n , P_n – множество точек $\{x^n\}$, представляющих собой n -ые остатки (погрешности), карту построенную на основе этих остатков и процедуру проектирования на M_n соответственно.

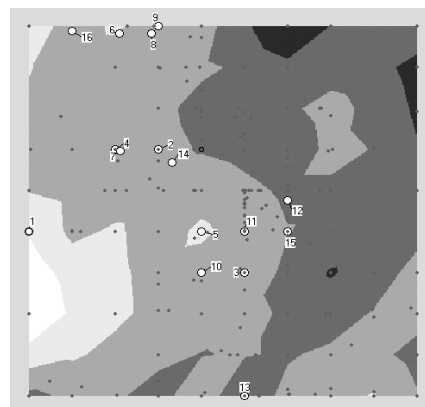
Остатки рассчитываются по формулам $x^1 = x - P(x)$, $x^2 = x^1 - P_1(x^1) \dots x^n = x^{n-1} - P_{n-1}(x^{n-1})$. Заменяя в тождестве $x = P(x) + P_1(x^1) + \dots + P_{n-1}(x^{n-1}) + x^n$ значение x^n на его смоделированное значение $P_n(x^n)$, получаем формулу для восстановления пропущенных значений в данных при помощи картирования n остатков:

$\tilde{x} = P(x) + \sum_{i=1}^n P_i(x^i)$, где $\tilde{x}$ – восстановленный радиус-вектор (в правой

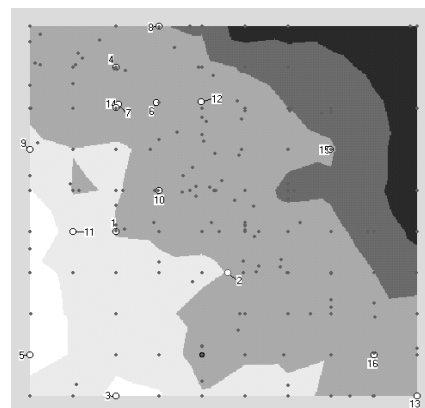
части выражения все координаты известны, несмотря на то, что координаты точки x и остатков x^i могут содержать пропущенные значения). Проиллюстрируем применение изложенных идей на примере картографирования экономических таблиц, приведенном в предыдущем разделе. На рис. 4.8 показаны первые три карты (карта данных, карты первых и вторых остатков). Все три карты приведены с нанесенной раскраской по признаку объема производства. При раскраске применена пятицветная раскраска, средний полутон отвечает значениям координат, близким к нулю. Это означает, что на карте первых остатков точки, попавшие в область среднего оттенка, наилучшим образом описываются моделью данных.



Карта данных



Карта первых остатков



Карта вторых остатков

Рис. 4.8. Карты данных и погрешностей модели.

Раскраска по признаку LG VO1998.

4.2.4. Проблема экстраполяции

Опыт визуализации данных показывает, что в случае не слишком «мягкой» карты проекции на карте имеют тенденцию к группировке вблизи краев карты, поскольку крайние точки карты оказываются ближайшими для большей части точек данных. Вследствие этого возникает идея экстраполяции карты, по крайней мере, на некоторую ее окрестность для того, чтобы не возникало подобных эффектов. В качестве условий, налагаемых на процедуру экстраполяции естественно взять требование того, чтобы вдали от данных карта становилась бы плоской (например, совпадала бы с плоскостью первых двух главных компонент).

При аппроксимации данных одномерными моделями [131-133] приемлемые результаты дает применение формул Карлемана [138].

4.3. Применение технологии для визуализации и анализа таблицы экологических измерений

Для практического применения описанной технологии была создана база данных по результатам измерений экологических показателей (примесей в слое приземного воздуха) за период с 1994-го по 1997 годы на территории города Красноярска.

Данные любезно предоставлены сотрудниками ООО "Экология" А.А. Дубковым, В.Н. Шевниным и А.В. Кучеренко.

По таблице 2 было построено два набора данных:

Набор №1 (Полный). Полный набор данных, содержащий все числовые поля и все измерения.

Набор №2 (Средние значения). Набор данных, в котором были усреднены значения условий и примесей по всем измерениям для каждой из точек, в которой проводились измерения. В результате каждая точка отображает средние многолетние значения измеряемых величин.

Оба набора данных состояли из одинакового набора количественных признаков.

Таблица 2

Описание полей исходной таблицы данных

Названия групп полей	Название поля	Описание
Географическое положение точек на карте	X Y ZONE REGION POINT	географические координаты на карте города Красноярска условные зоны города части зон условные названия точек
Даты и времена замеров	DATE TIME	день, месяц, год доля суток от полуночи
Метеорологические характеристики условия замеров	WS WD TEMP RH BP SR	скорость ветра, м/с направление ветра, градусы температура воздуха, °C влажность, % атмосферное давление, мм.рт.ст. солнечная радиация, МДж/м ²
Концентрации примесей в приземном слое атмосферы, в %	CO NO NO2 CO2 H2S SO2 O3 HCL	окись углерода окись азота диоксид азота диоксид углерода сероводород диоксид серы озон хлористый водород

Таблица 3

Способы нормировки признаков

Признак	Название	Нормировка
XX	Координата X	На среднеквадратичное отклонение
YY	Координата Y	На среднеквадратичное отклонение
CO	Концентрация CO	На среднеквадратичное отклонение
NO	Концентрация NO	На среднеквадратичное отклонение
NO2	Концентрация NO ₂	На среднеквадратичное отклонение
CO2	Концентрация CO ₂	На среднеквадратичное отклонение
H2S	Концентрация H ₂ S	Логарифм. преобразование и на ср.кв.отк.
SO2	Концентрация SO ₂	Логарифм. преобразование и на ср.кв.отк.
O3	Концентрация O ₃	Логарифм. преобразование и на ср.кв.отк.
HCL	Концентрация HCl	Логарифм. преобразование и на ср.кв.отк.
WS	Скорость ветра	Логарифм. преобразование и на ср.кв.отк.
WD	Направление ветра	На среднеквадратичное отклонение
TEMP	Температура	На среднеквадратичное отклонение
RH	Влажность	На среднеквадратичное отклонение
BP	Давление	На среднеквадратичное отклонение
SR	Солнечная радиация	Логарифм. преобразование и на ср.кв.отк.

Как видно из таблицы 3, к некоторым признакам предварительно

применялось логарифмическое преобразование в связи с тем, что их значения отличались на порядки.

По *Набору №1* была построена информационная модель полных данных, на основе которой был получен *Атлас №1* информационных раскрасок, представленный в стандартном ГИС-формате ГеоГраф 1.5. Атлас описан в разделе в разделе "*Атлас №1, №2 информационных раскрасок*". Отдельные слои из этого Атласа приведены в приложении.

По *Набору №2* была построена информационная модель усредненных данных, на основе которой был получен *Атлас №2* информационных раскрасок в формате ГеоГраф 1.5. Отдельные слои из *Атласа №2* приведены в приложении.

Проведенные эксперименты показали высокую правдоподобность данных в таблице. В частности, проводились эксперименты по предсказанию содержания примесей в атмосфере в зависимости от места, времени и атмосферных условий. При этом было выявлено, что более чем в 50% случаях относительная ошибка предсказания не превышает 10%. На долю же сильных ошибок (более 100% ошибки) выпало максимум 10% всех примеров, что может говорить о том, что в данных примерах имеют место ошибки в сборе и представлении информации. Причем по разным показателям эта доля варьируется – к примеру, содержание примесей серы предсказывается с ошибкой 5-10% более чем в 80% примерах и на долю больших ошибок приходится не более 1-2% всех примеров. Такая же картина и по содержанию озона и хлористого водорода. Несколько другая картина с соединениями углерода и азота – здесь процент ошибок выше, но все равно позволяет сделать вывод о правдоподобности этих данных.

При использовании моделей для предсказания содержания определенных примесей в атмосфере по всем остальным параметрам и соединениям ошибка достаточно мала и лишь в небольшом проценте примеров есть серьезные отклонения предсказанных значений от исходных. Что говорит о высокой правдоподобности данных и наличии небольшого числа, возможно, ошибочных значений.

Следует заметить, что наличие искажений не обязательно влечет за собой вывод об ошибочной их природе. Искажения могут быть вызваны влиянием каких-нибудь неучтенных фактора, поэтому большое наличие неправдоподобных значений может потребовать пересмотра всей модели данных и используемых при этом признаков.

Атлас №1, №2 информационных раскрасок

Основа визуализации данных – *карта данных*. Точки, близкие в исходном многомерном пространстве, оказываются близкими на карте данных. В нашем случае точка данных представляет собой проведенное измерение экологической обстановки. Это означает, что близкие на карте данных близкие точки изображают измерения, показавшие «схожесть» экологической обстановки по совокупности измеряемых величин. На рис. 1-16 приложения показана карта данных с нанесенными на нее точками данных.

Можно сказать, что карта данных подобна географической карте с той разницей, что на географической карте рядом оказываются объекты с близкими географическими координатами, а на карте данных – *объекты с близкими свойствами*. На географической карте рядом могут оказаться объекты с совершенно различными характеристиками, кроме того, некоторые объекты просто не имеют никакой привязки к географии. На карте данных учитывается сходство объектов «по совокупности свойств» в смысле расстояния в многомерном пространстве.

Простое изображение данных дает для исследователя не слишком много информации. Для того, чтобы представить себе какие особенности измерений соответствуют различным областям карты, исследователь получает представленный на рис. 1–16 (см. приложение) ГИС-атлас раскрасок. Сравнивая различные раскраски карты, исследователь получает представление о наборе данных. Схожесть раскрасок указывает на скореллированность соответствующих признаков.

Важной особенностью предлагаемой технологии является представление конечного продукта – атласа, в формате ГИС ГеоГраф 1.5. С помощью стандартной ГИС исследователь получает возможность генерации собственных раскрасок, подчеркивая те или иные интересующие его особенности набора данных, видоизменяя представление точек данных, накладывая различные слои друг на друга и находя их пересечения.

Атлас №2 имеет структуру, аналогичную *Атласу №1*. Разница состоит лишь в том, что он построен по *Набору данных №2* (усредненные значения). Соответственно, если на первом атласе показано 560 точек измерений, то на втором – 113 точек, соответствующих разным точкам города, в которых проводились измерения.

Итак, разработана и опробована на практике технология визуализации или картографирования многомерных данных (в которых могут содержаться пробелы) с помощью вложенных в пространство данных двумерных многообразий, названных упругими картами. И алгоритм построения этих многообразий, и общая идеология визуализации данных с их помощью существенно отличается от общепринятой на сегодняшний день технологии SOM. Особенностью технологии также является возможность непрерывного проектирования данных на карту, что существенно повышает точность представления данных.

Еще раз стоит отметить, что описанная технология открывает перспективы для использования всего арсенала методов и средств, накопленных в ГИС-технологиях для картирования данных самой различной природы, без привязки к географическим координатам. Можно сказать, что вместо географической карты в описанной технологии используется подложка, образованная структурой самих данных.

Атлас информационных раскрасок №1

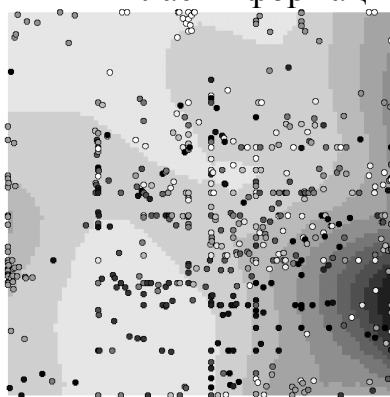


Рис. 1. Концентрация CO

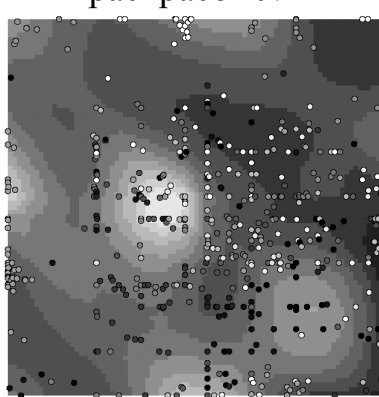


Рис. 2. Концентрация CO2

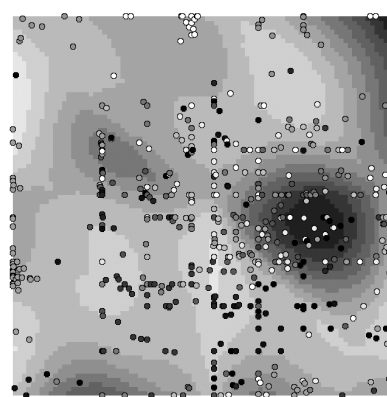


Рис. 3. Концентрация NO

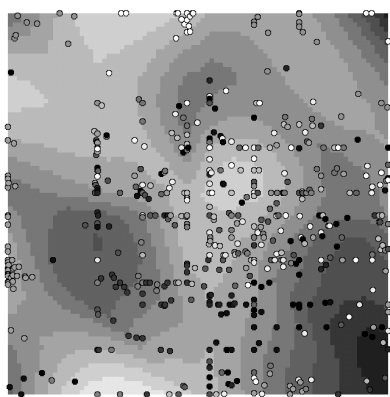


Рис. 4. Концентрация NO2

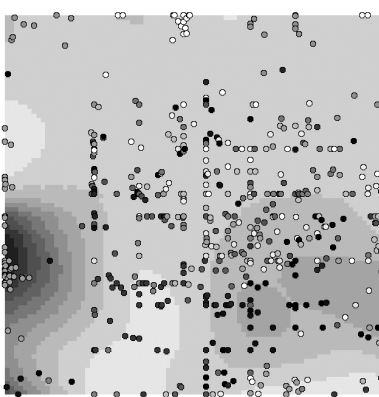


Рис. 5. Концентрация H2S

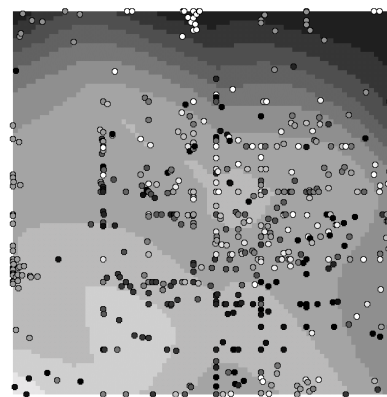


Рис. 6. Концентрация SO2

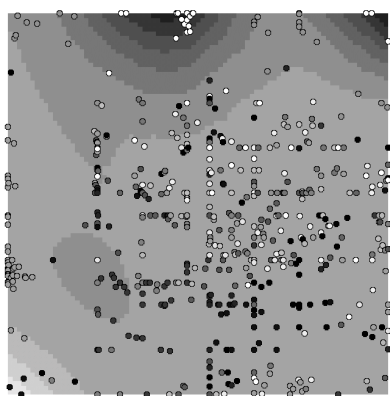


Рис. 7. Концентрация O3

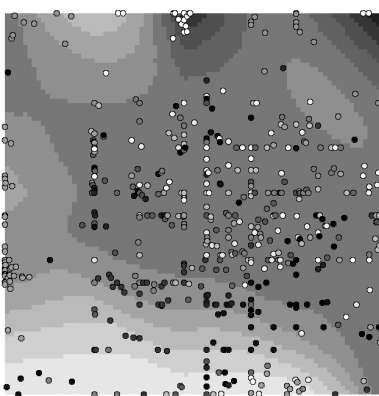
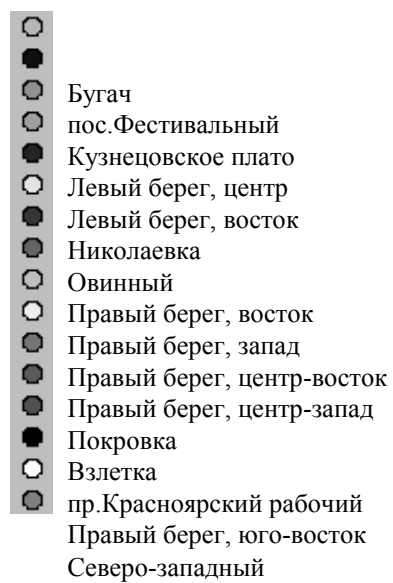


Рис. 8. Концентрация HCL



Атлас информационных раскрасок №2

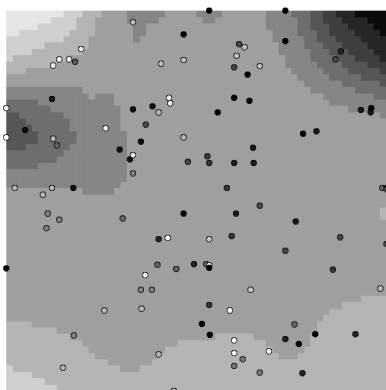


Рис. 9. Концентрация CO

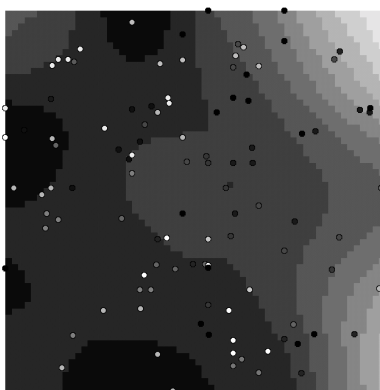


Рис. 10. Концентрация CO2

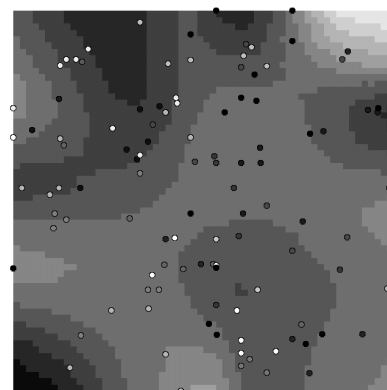


Рис. 11. Концентрация NO

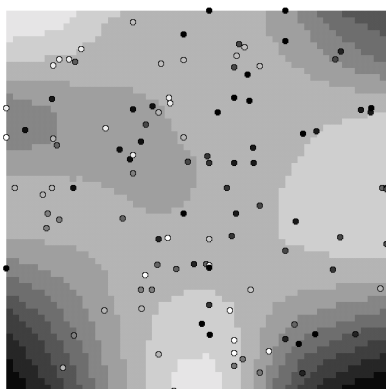


Рис. 12. Концентрация NO2

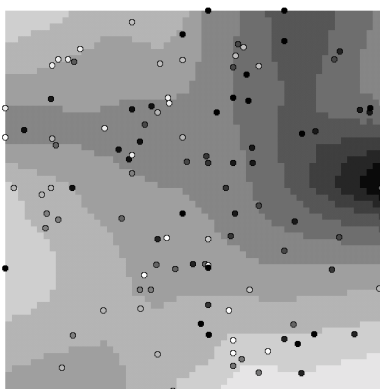


Рис. 13. Концентрация H2S

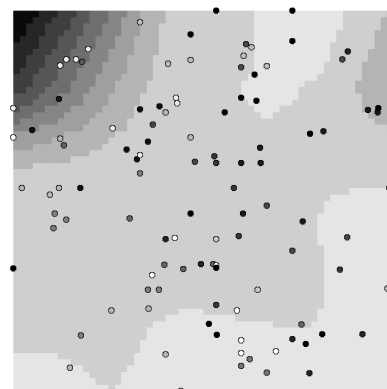


Рис. 14. Концентрация SO2

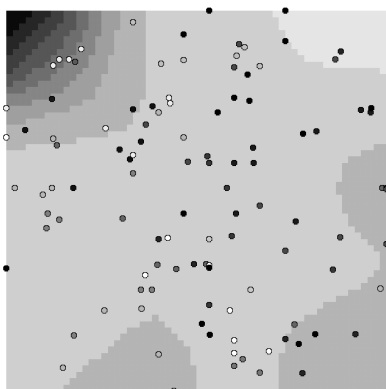


Рис. 15. Концентрация O3

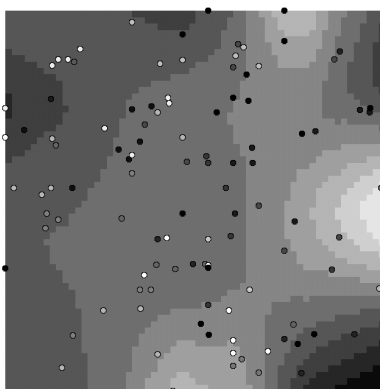


Рис. 16. Концентрация HCL

