ЭНТРОПИЙНЫЙ АНАЛИЗ ПРОСТРАНСТВЕННО РАСПРЕДЕЛЕННЫХ СИСТЕМ НА ПРИМЕРЕ ГЕОИНФОРМАЦИОННЫХ БАЗ ДАННЫХ КУЗНЕЦКОГО УГОЛЬНОГО БАССЕЙНА

Р. Ю. Замараев, С. Е. Попов, О. Л. Пястунович
Институт угля и углехимии СО РАН, Кемерово, Россия
e-mail: r.zamaraev@kemsc.ru, s.popov@kemsc.ru, skiporol@mail.ru

A new method of estimation and ranging of the complex characteristics of geo information databases (GDB) objects is formulated. Advantages of the method are illustrated in the case of the coal parameters analysis and ranging for a set of the Kuznetsk's coal basin mining lots.

Введение

Развитие ГИС и наполнение ассоциированных баз данных показателями из различных предметных областей позволяют решать ряд интересных задач, имеющих большое практическое значение. В данной работе рассматривается задача анализа данных геологоразведки совместно с технологическими, транспортными и другими факторами для оценки (ранжирования) участков месторождений полезных ископаемых по комплексным характеристикам.

Объекты геоинформационных баз данных (ГБД) представлены обычно широким набором показателей, отличающихся размерностью, типом (натуральные показатели, коэффициенты, баллы) и направленностью. С расширением набора учитываемых факторов при сравнении объектов ГБД очевидность выбора резко снижается. Возникают вопросы о сопоставимости показателей и способах учета их значимости.

Таким образом, в подходе к анализу и ранжированию по широкому набору показателей ключевым становится тезис об уникальности объектов ГБД и их совокупностей и невозможности выработать для них надежные статистические эталоны и правила.

Неоспоримым является факт пространственной разрывности и нерегулярности совокупностей объектов ГБД и, соответственно, их свойств, вследствие чего некорректно рассматривать их априорно как некое поле или однородную выборку для статистического описания и построения многофакторных регрессий. Спорные и неоднозначные результаты получаются также при использовании методов кластеризации. Это вызвано отсутствием в определении кластера меры, по которой можно произвести ранжирование

объектов. Результаты кластеризации сильно разнятся в зависимости от вида кластера и порядка учета показателей. Также возникают непреодолимые трудности в интерпретации результатов в задачах большой размерности и показателей различной природы. Эти же ограничения справедливы и для методов многомерного шкалирования.

Новым и весьма продуктивным инструментом является метод Айтчисона [1], ориентированный на анализ данных закрытого типа (композиций). При всех достоинствах метода в его теоретической части на практике возникают серьезные ограничения. Например, невозможно анализировать совместно показатели из разных композиций и автономные показатели.

Таким образом, для анализа ГБД требуется метод, имеющий в своем определении понятие меры для оценивания и ранжирования объектов, способный работать устойчиво, по единым алгоритмам с большим объемом данных о показателях различной природы и размерности.

Энтропийный метод в полной мере соответствует этим требованиям, но на начальном этапе его использование было обусловлено прежде всего положительным опытом анализа и ранжирования сопоставимых по сложности систем [2–4]:

— периодической системы химических элементов, представленной в тестовом исследовании 56 показателями;

— шахт и разрезов Кемеровской области в составе угольных компаний по широкому набору технологических, экономических и социальных показателей;

— минерального состава вод затопленных шахт и др.

Важно то, что при анализе перечисленных систем не было найдено противоречий с хорошо известными научными фактами и здравым смыслом, что рассматривается как подтверждение работоспособности и корректности метода.

1. Основы энтропийного метода анализа данных

Рассмотрим вектор $\mathbf{x} = \{x_i\}_{i=1,\ldots,m}$, образованный положительными значениями некоторого показателя x для совокупности R мощностью m функционально подобных автономных объектов.

Функциональное подобие здесь понимается как тождественность значимости, размерности и направленности избранных для анализа показателей для всех объектов совокупности. Автономность подразумевает априорную независимость формирования показателей объектов, отсутствие среди избранных показателей очевидного и доказанного аргумента.

Для перехода к вектору безразмерных величин в унифицированном масштабе используем преобразование

$$\mathbf{q} = \left\{\ln x_i^P\right\}, i = 1, ..., m, \tag{1}$$

где $P = 1/\ln \prod\limits_i x_i$.

Важнейшими свойствами этого преобразования являются:

— аддитивность элементов вектора $\mathbf{q}$, так как для любого подмножества W рассматриваемой совокупности R правило вычисления суммы имеет вид

$$\sum_{i\in W}^{q_i} = \ln \prod_{i\in W} x_i^P; \tag{2}$$

— аддитивность набора векторов $\mathbf{q}_1, ..., \mathbf{q}_n$, так как правило вычисления суммы имеет вид

$$\sum_{j=1}^{n} \mathbf{q}_j = \left\{ \ln \prod_j x_{i,j}^{P_j} \right\}, i = 1, ..., m, j = 1, ..., n. \tag{3}$$

В обоих случаях, как для суммы элементов вектора, так и для суммы векторов, получаем мультипликативные функции, математическая корректность которых не вызывает сомнений.

Значение i-го элемента вектора $\mathbf{q}$ численно равно доле i-го объекта в величине производственной функции $\ln \prod_i x_i$ совокупности R по показателю x. Отождествление значения i-го элемента вектора $\mathbf{q}$ с весом или вероятностью выбора i-го объекта по показателю x из совокупности позволяет перейти к отображению энтропийного типа:

$$\mathbf{E} = -\mathbf{q} \otimes \ln \mathbf{q} = \{-q_i \ln q_i\}. \tag{4}$$

Формула (4) подобна определению информационной энтропии по К. Шеннону [5], с этим связано название метода, хотя аналогия с информационной энтропией не является строгой. Сумма элементов вектора $\mathbf{E}$ из его определения — это характеристика упорядоченности совокупности по уровням (логарифмическим) показателя x. Элементы E_i вектора являются мерами неопределенности выбора i-го объекта: чем больше E_i, тем больший вес имеет i-й объект в совокупности и меньше неопределенность (риск) его выбора.

Для вектора $\mathbf{q}$ можно также ввести отображение, производное от энтропийного:

$$\mathbf{L} = \frac{d\mathbf{E}}{dq} = -\ln \mathbf{q} = \{-\ln q_i\}. \tag{5}$$

По аналогии с (2) и (3) для результатов отображений (4) и (5) можно определить алгебраические суммы элементов и векторов.

Отмеченные выше свойства предлагаемых преобразований исходных данных позволяют комбинировать отображения показателей в комплексные характеристики с учетом их направленности:

$$\begin{aligned} \hat{\mathbf{E}} &= \sum_{j \in U} a_j \mathbf{E}_j - \sum_{k \in V} b_k \mathbf{E}_k = \left\{ \sum_{j \in U} a_j E_{i,j} - \sum_{k \in V} b_k E_{i,k} \right\}, \\ \hat{\mathbf{L}} &= \sum_{j \in U} a_j \mathbf{L}_j - \sum_{k \in V} b_k \mathbf{L}_k = \left\{ \sum_{j \in U} a_j L_{i,j} - \sum_{k \in V} b_k L_{i,k} \right\}, \end{aligned} \tag{6}$$

где U и V — подмножества положительно и отрицательно направленных показателей соответственно; a_j и b_k — коэффициенты значимости показателей.

Методическая новизна энтропийного анализа в отличие, например, от подхода, изложенного в [1], заключается в отказе от построения регрессий и симплексов при достигнутой аддитивности операндов. Полагая, что пара векторов ($\hat{\mathbf{E}}$ и $\hat{\mathbf{L}}$) определяет аналог фазовой плоскости $\hat{L}\left(\hat{E}\right)$ дискретно заданной системы, для анализа состояния и ранжирования объектов совокупности предлагается использовать свойства их изображающих точек с координатами $\left(\hat{E}_i, \hat{L}_i\right)$. При этом оказываются доступными для

использования строгие границы и критерии видов состоянии систем и их элементов на фазовой плоскости, известные из динамики.

В энтропийном анализе также реализуется уникальный методический прием — формирование осей фазовой плоскости из независимых и произвольных подмножеств показателей, что является мощным инструментом исследования связи показателей и декомпозиции совокупности на подсистемы.

Однако в этом случае для унификации масштаба требуются дополнительные преобразования фазовых координат. Наиболее удобное для анализа представление дает стандартизация:

$$\overset{\circ}{E} = \frac{\hat{\mathbf{E}} - M\left[\hat{\mathbf{E}}\right]}{\sigma\left[\hat{\mathbf{E}}\right]},$$

$$\overset{\circ}{L} = \frac{\hat{\mathbf{L}} - M\left[\hat{\mathbf{L}}\right]}{\sigma\left[\hat{\mathbf{L}}\right]},$$

где $M[\dots]$ и $\sigma[\dots]$ определяют среднее значение и среднеквадратическое отклонение элементов вектора соответственно.

Для разделения фазовой плоскости на области предлагаются следующие границы, показавшие свою пригодность в ряде тестовых задач [3]:

— главная ось совокупности изображающих точек как предельная линия прямой связи фазовых координат

$$r = \frac{1}{m}\sum_i \overset{\circ}{E}_i \overset{\circ}{L}_i,$$

где r — тангенс угла наклона главной оси, проходящей через начало координат фазовой плоскости $\overset{\circ}{L}\left(\overset{\circ}{E}\right)$;

— эллипс как предельная линия гармонического решения уравнения связи фазовых координат

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} = 1,$$

$$A = \sqrt{2\left(\frac{m-1}{m} + \frac{2r^2}{1+r^2}\right)}, \quad B = \sqrt{2\left(\frac{m-1}{m} - \frac{2r^2}{1+r^2}\right)},$$

где x и y — фазовые координаты; A и B — большая и малая полуоси эллипса соответственно и большая полуось A полагается совпадающей с главной осью совокупности изображающих точек;

— гиперболы, софокусные эллипсу, и гиперболы, сопряженные с ним:

$$\frac{x^2}{A_h^2} - \frac{y^2}{B_h^2} = \pm 1,$$

полученные из условия $A/B = A_h/B_h$, где A_h и B_h — большая и малая полуоси гиперболы соответственно.

Полуоси эллипса получены из условия $\pi AB = mr$, что для фазовой плоскости эквивалентно равенству потенциала производящих динамических систем — совокупности рассматриваемых объектов в выбранных показателях и системы двух математических маятников.

Содержание и оценка границ в значительной мере зависят от показателей, включенных в определение фазовых осей. Неизменным остается качественное разделение фазовой плоскости на области с различными видами состояния и тенденциями поведения объектов.

Решения дифференциального уравнения связи фазовых координат в окрестности эллипса и гипербол места изображающих точек объектов можно по принадлежности к трем видам состояния подразделить:

— на устойчивое — внутри эллипса и над гиперболами;

— неустойчивое — вне эллипса и в створе гипербол;

— автоколебательное — на граничных линиях.

Важно то, что модели границ в своих определениях не зависят от показателей, вовлеченных в построение фазовых осей, т. е. являются универсальными.

2. Анализ показателей углей, извлекаемых на действующих участках угледобывающих предприятий Кузнецкого угольного бассейна

Сегодня в угольной промышленности основной тенденцией стало формирование металлоугольных и энергоугольных холдингов. В холдинге угледобывающее предприятие становится поставщиком одного из компонентов шихты или концентрата, ориентированных на коксование, сжигание и/или химическую переработку. Вследствие широкого разброса свойств углей, условий их залегания и ограниченности выбора задачу комплектования шихты невозможно решить оптимально, ее можно решить только рационально, т. е. выбрать для разработки участок, “лучший из доступных”.

Основные показатели, учитываемые при выборе участка для добычи угля, приведены в таблице. Здесь просматривается не просто многомерность, а принципиальные различия в природе и размерности данных.

Так, пп. 1–6 входят в петрографическую композицию пробы, а пп. 11–14 — в химическую композицию, причем обе из них неполные. Показатели 7–10 относятся к технологической группе и имеют повышенный приоритет при выборе участка. Ряд показателей имеет в прикладном анализе негативное значение — расстояние до железной дороги и ЛЭП. Без умаления общности в таблице не приведены геохимические показатели, которые также составляют неполную композицию.

Надо отметить, что показанные на примере Кузнецкого угольного бассейна многомерность и многоплановость данных свойственны всем ГБД.

Источником данных в настоящей работе послужила комплексная ГИС Кузнецкого угольного бассейна, разработанная в Институте угля и углехимии СО РАН. Она ассоциирована с базой данных по петрографическим, технологическим, химическим и геохимическим свойствам кузнецких углей.

Объем базы данных составляет более 7000 уникальных записей с результатами анализа проб, географически привязанных к разрабатываемым и нераспределенным участкам, шахтным полям, геолого-экономическим районам, физической, административной и транспортной картам Кузбасса.

Для примера рассмотрена совокупность участков, представленная в анализе таблицей значений показателей углей по 2325 пробам. Для каждой пробы известны уникаль-

Перечень показателей, учитываемых при выборе участка

№ п/п	Наименование показателя	Единица измерения
1	Витринит	%
2	Семивитринит	%
3	Инертинит	%
4	Липтинит	%
5	Влага максимальная	%
6	Зольность угольных пачек	%
7	Показатель отражения витринита	%
8	Выход летучих на сухое беззольное состояние	%
9	Толщина пластического слоя	мм
10	Теплота сгорания высшая	МДж/кг
11	Теплота сгорания низшая	МДж/кг
12	Углерод	%
13	Водород	%
14	Азот + кислород	%
15	Влага аналитическая	%
16	Мощность угольных пачек пласта	м
17	Глубина залегания	м
18	Объем запасов	млн т
19	Расстояние до ж/д	км
20	Расстояние до ЛЭП	км

ный идентификатор в базе данных, географическая привязка и марка угля, идентифицированная по стандартным методикам.

Прежде всего была поставлена задача сравнения участков между собой по наиболее значимым технологическим показателям — теплоте сгорания (п. 10, см. таблицу) и толщине пластического слоя (п. 9). Такую постановку можно соотнести с задачей выбора участков с гарантированным сбытом угля — сжигание и/или коксование.

На рис. 1, *a* приведен полный фазовый портрет в координатах $\overset{\circ}{L}_9\left(\overset{\circ}{E}_{10}\right)$, на нем выделена область с концентрацией изображающих точек проб с высшими значениями технологических показателей. На рис. 2–4 изображены отфильтрованные по маркам угля варианты полного портрета. Первое очевидное заключение состоит в том, что заложенная в стандарты модель марок является несостоятельной.

Разброс значений основных технологических показателей внутри марок так велик, что делает марки угля статистически не различимыми. Кроме того, толщина пластического слоя проявляется как грубый ранжированный показатель типа балльной шкалы, а теплота сгорания характеризуется наличием некоторого предельного уровня своего значения в окрестности $0.6\sigma\left[\overset{\circ}{E}_{10}\right]$.

Фазовые координаты совокупности не имеют корреляционной связи ($r \to 0$). Ненулевое значение r можно объяснить только различной мощностью марочных подмножеств.

На фазовых портретах (рис. 3 и 4) привлекает внимание ряд изображающих точек — пробы марок К и Г во втором и третьем квадрантах. Они выходят за граничный эллипс и при этом оторваны от подсистем своих марок значимым смещением по абсциссе. Из этого можно сделать вывод о специфических свойствах углей на данных участках либо об ошибке при определении теплоты сгорания.

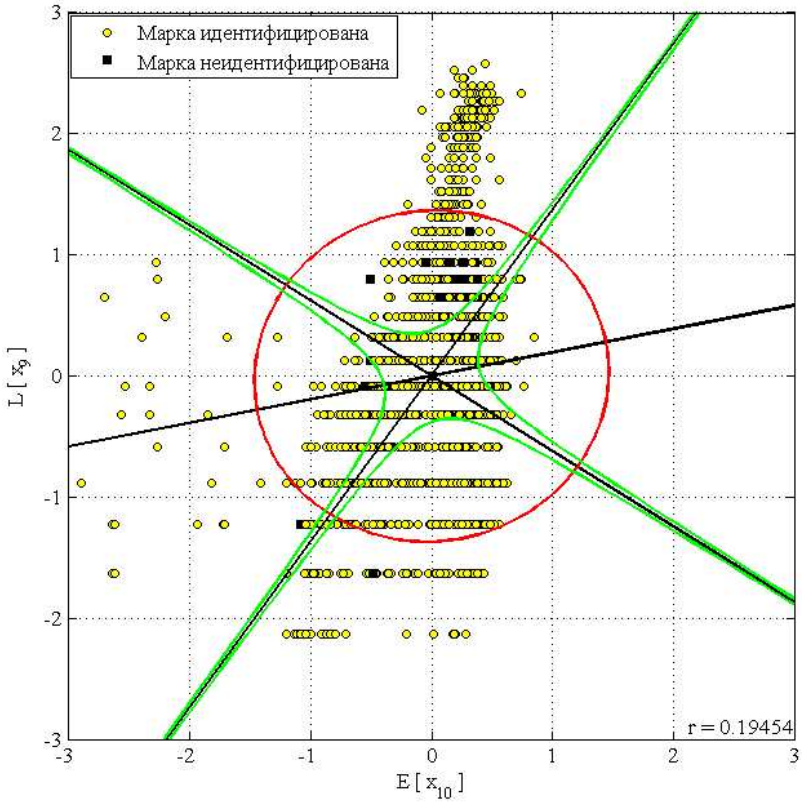
Марка идентифицирована
Марка неидентифицирована
r = 0.19454
E [x10]
L [x9]

Рис. 1.

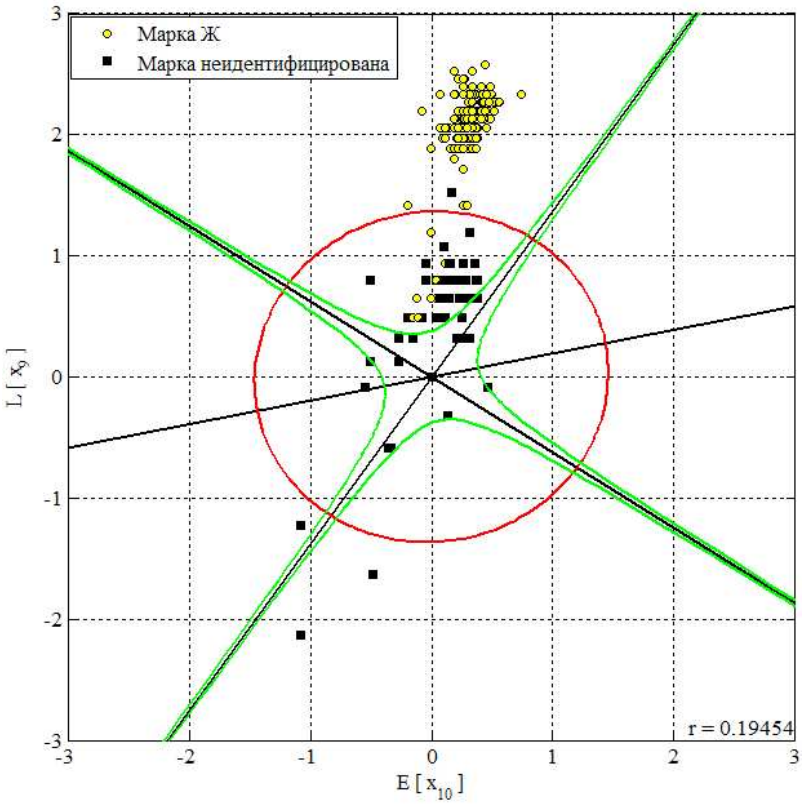
Марка Ж
Марка неидентифицирована
r = 0.19454
E [x10]
L [x9]

Рис. 2.

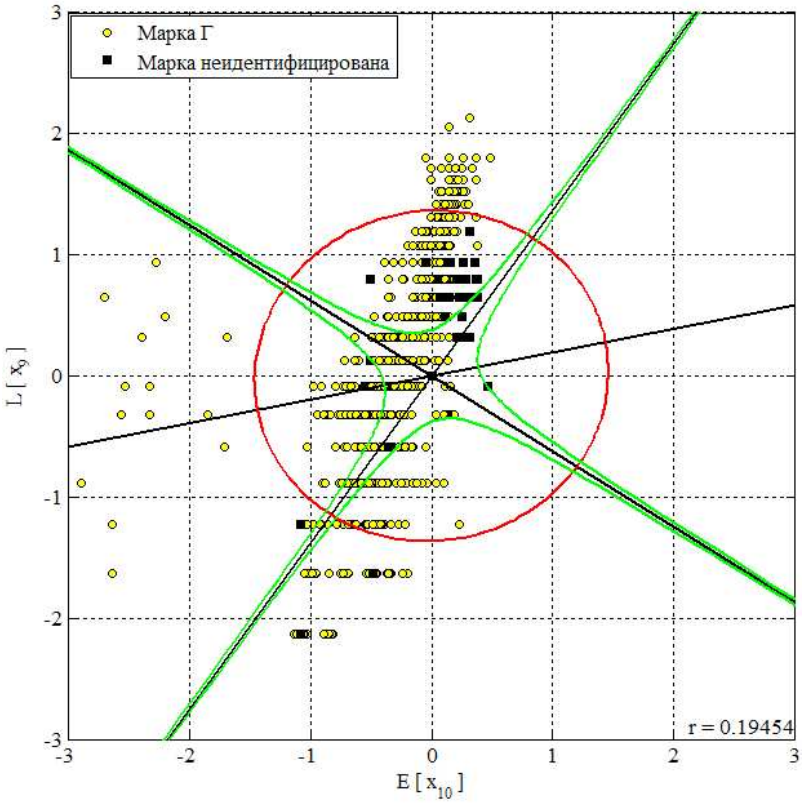
Марка Г
Марка неидентифицирована
r = 0.19454
L [x9]
E [x10]

Рис. 3.

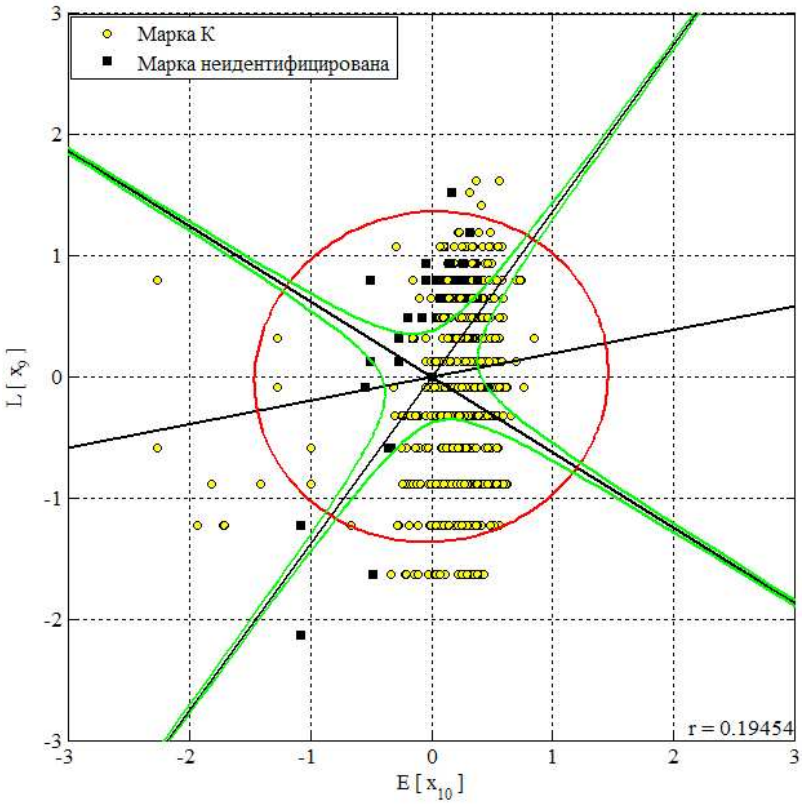
Марка К
Марка неидентифицирована
r = 0.19454
L [x9]
E [x10]

Рис. 4.

Обособленную подсистему на портрете составляют только угли марки Ж. Они выделяются как высокой теплотой сгорания, так и аномально высокой толщиной пластического слоя. Подавляющее число изображающих точек этой марки находится в первом квадранте и выходит за граничный эллипс, что свидетельствует о значимом, системном отличии свойств этих углей от среднего уровня совокупности.

Марки К и Г также представлены в зоне аномально высоких показателей. Граница между марками Ж и Г может быть проведена на уровне 1.8σ ординаты. Марка К по толщине пластического слоя заметно отстает от марки Ж.

Наличие на рынке большого количества углей смешанных или переходных марок (ГЖ, КЖ и т. п.) косвенно подтверждает невозможность статистического различения худших углей марки Ж от лучших углей марок К и Г по основным технологическим показателям. Это в свою очередь открывает возможности для поиска адекватной замены между марками при комплектовании шихты.

Так как алгоритмы энтропийного анализа сохраняют дискретное представление совокупности, имеем возможность сразу отобразить заинтересовавшие нас объекты на карте. На рис. 5 на контуре Кемеровской области с границами геологических районов отмечены центры участков, на которых получены пробы с высшими значениями технологических показателей.

На рис. 6 приведен фазовый портрет в координатах $\overset{\circ}{L}_6\left(\{E_{10}\overset{\circ}{+}E_{16}\}\right)$. Такой вариант иллюстрирует задачу ранжирования и выбора участков из условия максимума теплотворной способности при минимуме затрат на добычу и обогащение угля. Эту комбинацию показателей можно интерпретировать так же, как сравнение инвестиционной привлекательности участков для теплоэнергетики.

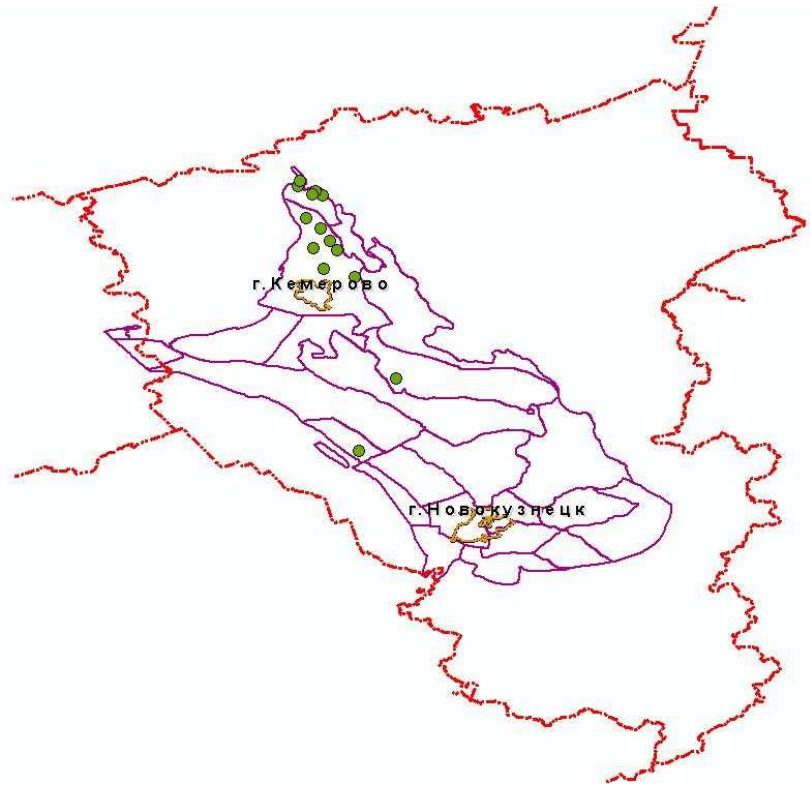

Рис. 5.

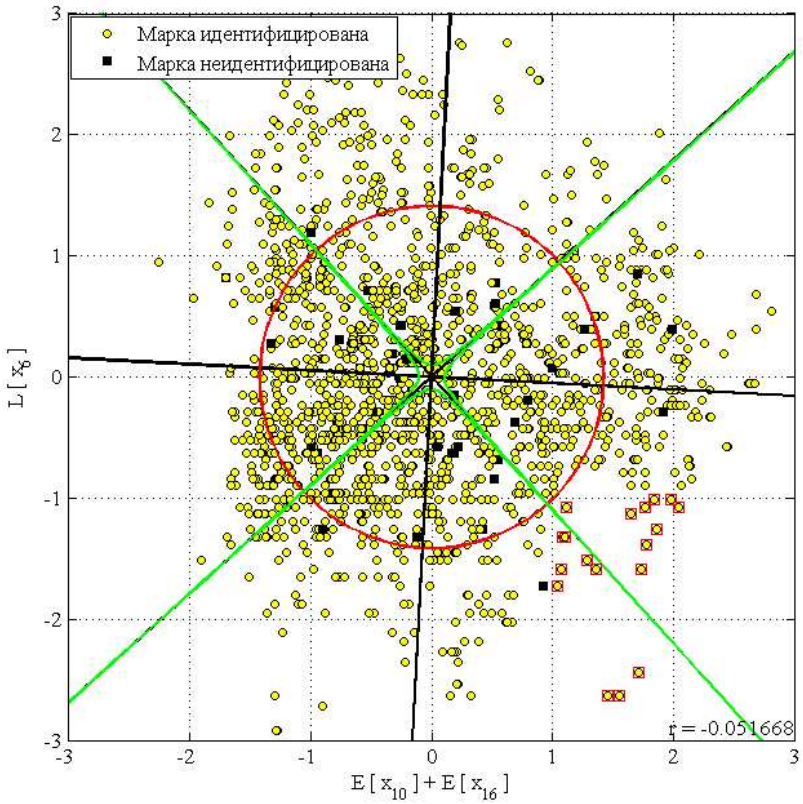

Рис. 6.

Из поставленного условия нас интересует четвертый квадрант плоскости, где сосредоточены изображающие точки с малыми значениями зольности x_6 и большими значениями комплексной характеристики, объединяющей теплоту сгорания x_{10} и мощность пачек пласта x_{16}.

Наибольший интерес должны вызвать прежде всего участки, выходящие за граничный эллипс в створе главных осей совокупности. Эта область дополнительно разделена на две части асимптотой гипербол. Асимптота делит выделенную область на зоны с различными видами состояния изображающих точек и соответствующих объектов относительно совокупности:

— участки с аномально низкой зольностью при высокой комплексной характеристике $E_{10} + E_{16}$ (I);

— участки с аномально высокой комплексной характеристикой $E_{10} + E_{16}$ при низкой зольности (II).

Такое деление предоставляет дополнительные возможности при дифференциации участков в случае выдвижения особых требований к показателям, например минимизации именно зольности.

В качестве абсолютных лидеров на рисунке отмечены (заключены в квадраты) изображающие точки, тяготеющие к асимптоте и выходящие как по абсциссе, так и по ординате за уровень σ. Участки, на которых получены эти пробы, отмечены на карте, приведенной на рис. 6. Кроме проб с идентифицированными марками на рис. 1–4 и 6 вынесены пробы с участков, находящихся в стадии геологической доразведки (черные квадраты). Естественно, по известным для них показателям установить марку несложно. Но на фазовом портрете мы можем увидеть принадлежность полученных проб

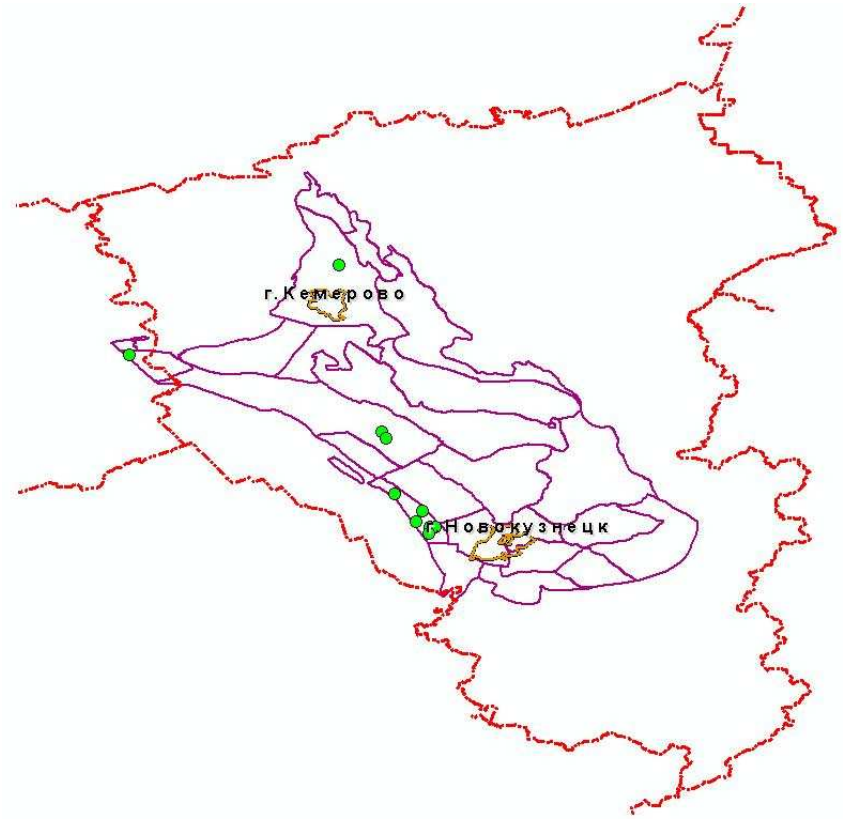

Рис. 7.

к определенной подгруппе углей по их природным свойствам, а не по марке. Например, новый участок, примыкающий к группе абсолютных лидеров (рис. 7), должен вызвать наибольший интерес у энергоугольных холдингов.

Рассмотренные примеры являются наиболее простыми, так как показатели, вошедшие в определение осей фазовой плоскости, не имеют корреляционной связи ($r \to 0$) и главные оси совокупности стремятся к координатным осям. Возможны более сложные варианты, когда определения фазовых осей имеют корреляционную связь. Тогда правила интерпретации положения изображающих точек усложняются.

Заключение

Энтропийный метод анализа в приложении к пространственно распределенным системам является мощным инструментом ранжирования элементов (объектов ГБД) и выделения подсистем.

Возможность комбинирования комплексных характеристик из показателей различной природы и размерности позволяет ставить и решать аналитические задачи любой сложности. Фазовое представление комплексных характеристик и фундаментальные критерии состояния изображающих точек, известные из динамики, обеспечивают строгость заключений и выводов.

Сохранение исходного дискретного представления системы позволяет быстро и просто получать карты для выделенных подсистем и объектов.

Список литературы

[1] Aitchison J., Egozcue J.J. // Mathematical Geology. 2005. Vol. 37, N 7. P. 829–850.

[2] Логов А.Б., Замараев Р.Ю., Логов А.А. Анализ состояния систем уникальных объектов // Вычисл. технологии. 2005. Т. 10, № 5. С. 49–53.

[3] Логов А.Б., Замараев Р.Ю., Логов А.А. Моделирование тенденций поведения элементов систем уникальных объектов // Вычисл. технологии. 2005. Т. 10, № 5. С. 54–56.

[4] Логов А.Б., Замараев Р.Ю., Логов А.А. Алгоритмы энтропийного метода анализа для отображения свойств объекта в фазовом пространстве // Вычисл. технологии. 2005. Т. 10, № 6. С. 75–81.

[5] Шеннон К. Работы по теории информации и кибернетике. М.: Изд. иностр. лит., 1963.

Поступила в редакцию 21 ноября 2007 г.