УДК 332.1

КЛАСТЕРИЗАЦИЯ ПЛОТНОСТИ С КОМБИНИРОВАННЫМ РАССТОЯНИЕМ КАК ИНСТРУМЕНТ АНАЛИЗА ГОРОДСКОЙ СРЕДЫ

Митрофанов Алексей Юрьевич,
кандидат экономических наук, доцент кафедры прикладной математики и информатики Саратовского социально-экономического института (филиала) Российского экономического университета им. Г.В. Плеханова; MitrofanovAY0@gmail.com
Файзлиев Алексей Раисович,
преподаватель кафедры информационных систем в экономике Саратовского социально-экономического института (филиала) Российского экономического университета им. Г.В. Плеханова; FaizlievAR1983@mail.ru

Предлагается новый метод кластеризации пространственных данных, основанный на расстоянии, обобщающем различия плотности распределения количественного признака и пространственную удаленность. Предлагаются два метода выбора оптимального числа кластеров, один из которых основывается на непараметрической оценке плотности и корреляционном отношении. В качестве иллюстрации проводится сравнение результатов кластеризации данных по г. Саратову за 2006-2007 гг.; алгоритм реализован в виде набора функций для статистического пакета R, доступного по запросу к авторам.
Ключевые слова: разведочный анализ данных, кластеризация, пространственные данные, плотность, сглаживание, корреляционное отношение, энтропия Шеннона, статистический пакет R.

Введение

Пространственные аспекты экономической деятельности существенно влияют на выбор, осуществляемый экономическими агентами: покупатели учитывают близость расположения и транспортную доступность магазинов; продавцы принимают решения об ассортименте и ценах товаров, учитывая уровень достатка и другие признаки потенциальных покупателей, проживающих недалеко от торговых предприятий. Взаимообусловленность поведения покупателей и продавцов приводит к возникновению территорий с относительно высокой концентрацией предприятий торговли[1]. Вместе с тем,

[1] В работах П. Кругмана и других исследователей описаны механизмы положительной обратной связи, приводящие к пространственному сближению предприятий, реализующих схожие товары,например, букинистических магазинов в Лондоне [7].

действует и обратный процесс – усиление местной конкуренции, насыщение локального рынка ограничивают дальнейшую концентрацию торговли.

Социально-экономические механизмы, приводящие к вариациям плотности населения и/или коммерческой деятельности, очень сложны. Вместе с тем, многие практически важные задачи в сфере бизнеса и государственного управления, требующие учета пространственного расположения объектов, должны быть решены за ограниченное время в условиях недостатка ресурсов и/или невозможности привлечения сторонних исполнителей (например, вследствие необходимости ограничить доступ к конфиденциальной информации). Практическая важность этих задач, с одной стороны, и сложность построения соответствующих моделей, с другой, обуславливают необходимость разработки доступных методов анализа сорасположения городского населения и объектов городской среды.

В качестве примера рассмотрим задачу выбора местоположения будущего магазина. Предположим, что имеются данные о распределении численности населения (потенциальных покупателей), а также о местоположении предприятий-конкурентов. Также предположим, что территория города разбита на небольшие ячейки, каждая из которых дает вариант расположения магазина. Для сравнения этих вариантов используется мера «перспективности» ячейки, в качестве которой можно принять, например, отношение числа потенциальных покупателей в ней к числу соседствующих конкурентов (возможно, учитываемых с весами, зависящими от удаленности, подобно известному закону Рейли). Поскольку число ячеек может достигать десятков и даже сотен, сбор соответствующих данных и сравнение вариантов могут быть затруднены. Возможен альтернативный подход, включающий следующие этапы: 1) выделение территорий города, относительно однородных по плотности населения и «интенсивности» присутствия магазинов-конкурентов (например, отношению суммарной торговой площади к площади ячейки); 2) выделение одной-двух таких территорий по критериям удобства логистики и др. (поскольку общее число таких территорий относительно невелико, их сравнение, скорее всего, не вызовет трудностей); 3) выбор определенной ячейки в пределах ранее выбранных территорий на основе вышеупомянутой меры «перспективности».

Преимущество данного подхода состоит в существенном уменьшении числа сравниваемых вариантов благодаря применению двухступенчатой процедуры сравнения. Разумеется, нельзя гарантировать, что найденная таким образом ячейка будет одной из наиболее «перспективных» при непосредственном сравнении всех ячеек, однако выигрыш в простоте выбора, возможно, скомпенсирует этот недостаток.

Другая задача, требующая анализа пространственного сорасположения, – задача оптимизации маршрутов патрулирования городской территории правоохранительными органами на основе данных о плотности населения, уровне его достатка, а также плотности распределения мест совершения противоправных действий.

Ключевым условием успешности двухступенчатой процедуры сравнения является наличие территорий (зон), однородных по одному или нескольким количественным признакам (таким, как плотность населения, «интенсивность» присутствия конкурентов, «плотность» преступлений определенного вида). С этой целью мы предлагаем новый метод кластеризации значений плотности распределения аддитивного признака.

Кластеризация пространственных данных

Современный крупный город представляет собой сложный конгломерат точечных (нуль-мерных), линейных (одномерных) и территориальных (двумерных) объектов. В случае, когда точечные объекты близки по размерам (жители города без учета их демографических и/или социально-экономических различий, автозаправочные станции и др.), реальная их сеть задается совокупностью точек на плоскости (point pattern[2]): $(x_i, y_i), i = 1,2,\ldots n.$ В случае, когда точечные объекты существенно различаются по размерам (торговые предприятия, объекты недвижимости и др.), реальная сеть задается как совокупность маркированных точек[3]: $(x_i, y_i, z_i), i = 1,2,\ldots,n,$ где «марка» z_i служит характеристикой размера объекта (для однородных по размеру объектов будем считать, что $z_i = 1$). Предположим, что задан маркированный точечный процесс[4], при этом марки будем считать аддитивными.

Ключевую роль в дальнейшем играет плотность (интенсивность) распределения размеров точечных объектов. Для объектов, находящихся на заданной территории A, средняя плотность распределения размеров определяется как $\overline{p} = Z/\#(A)$, где $Z = \sum z_j$, $\#(A)$ – площадь A.

Для определения плотности распределения для каждого точечного объекта в отдельности, построим сеть ячеек Дирихле (Вороного), окружающих каждый объект. Ячейка Дирихле a_i объекта i определяется как совокупность всех таких точек на плоскости, которые лежат к точке (x_i,y_i) ближе (в смысле обычного евклидова расстояния), чем ко всем остальным объектам, и представляет собой выпуклый многоугольник. Обозначим плотность распределения для отдельного объекта (ячейки) $p_i = z_i/\#(a_i)$.

К настоящему времени в зарубежной научной литературе опубликовано множество работ, посвященных методам пространственной статистики и пространственного анализа. Поскольку подобные исследования, чаще всего, имеют тесную связь с приложениями, большое значение имеет доступность программных реализаций имеющихся методов. По нашему мнению, своими возможностями в этом смысле выделяется свободно распространяемый пакет для статистических вычислений «R» [10]. В его (также общедоступных) расширениях реализованы несколько алгоритмов пространственной кластеризации [12]. В алгоритме spatcluster (пакет расширения

[2] В терминологии пакета-расширения spatstat [3] общедоступного статистического пакета R [10].

[3] «Marked point pattern» в пакете spatstat.

[4] В рамках вероятностной парадигмы данное наименование не вполне корректно – следовало бы назвать это реализацией маркированного точечного процесса.

spatgraphs [11]) кластеры рассматриваются как связные компоненты графа удаленности. Алгоритм допускает кластеризацию точечных объектов с использованием расстояний, учитывающих пространственную удаленность и марки (Mass geometric, Mark crossing). Другой алгоритм – dbscan (пакет расширения fpc [8]) ориентирован на поиск кластеров в больших базах пространственных данных, при этом кластеры выделяются на основании плотности распределения точечных объектов, достижимости и связности [6]. Еще один алгоритм [2], реализуемый функцией pdfCluster пакета pdfCluster [1], также основан на плотности точечных объектов, при этом кластер определяется как область высокой плотности. В перечисленных алгоритмах кластеризация основывается на различных мерах близости, сочетающих как соседство объектов (связность), так и сгущения последних. Наличие многих пакетов для пространственного анализа обусловило возникновение трудностей с переносом пространственных данных из одного пакета в другой. Эту трудность частично преодолевает пакет sp [4, 9], предлагающий унифицированные структуры (классы) пространственных данных.

После знакомства со свойствами этих алгоритмов на тестовых примерах, нами сделан вывод о том, что ни один из них в полной мере не удовлетворяет потребностям анализа расположения объектов городской среды – некоторые из перечисленных методов не допускают использования маркированных данных, большинство же реализуют идею связности из теории графов. Последнее, по нашему мнению, непригодно для объектов городской среды, в которой зоны высокой плотности признака могут быть разделены полосами относительно низкой плотности, но это не исключает рассмотрения всех близко расположенных объектов (приблизительно одинаковой) высокой плотности как одного целого.

Алгоритм кластеризации пространственных данных с «комбинированным расстоянием»

Предлагаемый метод кластеризации ячеек Дирихле предназначен для выделения относительно компактных зон городской территории, в пределах которых плотность распределения z_i относительно мало изменяется в сравнении с плотностями соседних зон (заметим, что мы не требуем односвязности зон). Для полного определения алгоритма кластерного анализа необходимо выбрать общую схему алгоритма (агломеративную, дивизимную, К-средних и др.), правило вычисления расстояния между отдельными объектами, правило вычисления расстояния между кластерами, а также метод выбора числа кластеров.

Мы предлагаем использовать агломеративную схему, как одну из наиболее употребительных, и не требующую априорного задания числа кластеров.

Наиболее важным элементом алгоритма кластеризации является используемая формула расстояния между объектами. Особенностью городской среды является значительная вариация плотности (значения которой между ячейками могут различаться в десятки и даже сотни раз). В результате (предполагая, что все значения плотности положительны) мы приходим

к следующей формуле расстояния между тройками[5]:

$$d\left((x_i, y_i, z_i),(x_j, y_j, z_j)\right) = \max\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, c\left|\ln p_i - \ln p_j\right|\right).$$

Мы называем данное расстояние «комбинированным», поскольку оно объединяет как пространственную близость ячеек, так и различие плотностей в них (в логарифмической шкале).

Константа $c > 0$, «уравнивающая» пространственную удаленность и различие логарифмов плотности, определяется как оценка коэффициента линейной регрессии без свободного члена множества пар евклидовых расстояний между объектами на абсолютные расхождения между логарифмами плотностей ячеек Дирихле:

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \approx \hat{c}\left|\ln p_i - \ln p_j\right|, i = 1,2,..., n-1,\ j = 2,..., n, i < j\ ;$$

использование МНК приводит к оценке $\hat{c} = \sum_{1 \le i < j \le n} g_{ij} h_{ij} \Big/ \sum_{1 \le i < j \le n} h_{ij}^2$.

В целях сравнения результатов кластеризации мы рассматриваем три варианта расстояния между кластерами, определяемыми как подмножества номеров ячеек (табл. 1).

Таблица 1

Варианты определения расстояния между кластерами

Вариант	Вид расстояния	Определение
1	ближайшего соседа (single link)	$d^{(1)}(Cl, Cl') = \min_{\alpha \in Cl, \alpha' \in Cl'} d_{\alpha\alpha'}$
2	дальнего соседа (complete link)	$d^{(2)}(Cl, Cl') = \max_{\alpha \in Cl, \alpha' \in Cl'} d_{\alpha\alpha'}$
3	среднее расстояние (pair-group average)	$d^{(3)}(Cl, Cl') = \frac{1}{\#(Cl)\#(Cl')} \sum_{\alpha \in Cl, \alpha' \in Cl'} d_{\alpha\alpha'}$ (# обозначает число элементов)

Для изучения чувствительности результатов кластеризации к значению с проводились эксперименты, в которых значение с уменьшалось и увеличивалось на 20%. Они показали, что измененные значения с не дают преимуществ в смысле качества кластеризации (см. далее) по сравнению со значением, найденным по вышеприведенной формуле.

Выбор числа кластеров

Поскольку город в большинстве случаев представляет собой «непрерывное» целое[6], выбор числа кластеров по данным об объектах городской среды может представлять значительные трудности. Как и при реализации стандартных вариантов кластерного анализа, выбор числа кластеров определяет компромисс между точностью и полнотой представления исходных данных – положений точечных объектов и плотностей p_i по ячейкам Ди-

[5] Данная мера представляет собой настоящее расстояние; рефлексивность и симметричность очевидны, для обоснования неравенства треугольника рассматриваются четыре варианта определения максимума и используются неравенства треугольника для пространственной удаленности и для разности логарифмов плотности.

[6] Четкие кластеры могут быть связаны, например, с микрорайонами-новостройками или особенностями топографии города.

рихле, с одной стороны, и простотой результатов кластеризации, с другой.

С этой целью нами использовались три подхода: а) по максимальному «скачку» расстояния между кластерами, б) по относительной энтропии Шеннона, в) по корреляционному отношению, связанному с непараметрической оценкой плотности. Рассмотрим их более подробно.

По максимальному скачку расстояния. На каждом шаге кластеризации фиксируется минимальное расстояние d_{min} между кластерами перед очередным объединением; эта величина монотонно возрастает с ростом шага алгоритма и монотонно убывает в зависимости от числа кластеров k. Обозначим величины скачков расстояния:

$$\Delta d(k) = d_{\min}(k) - d_{\min}(k+1), 2 \le k \le n-1.$$

Число кластеров k предлагается выбрать соответственно одному из наибольших значений $\Delta d(k)$ (на практике соответственно максимальной длине вертикальных ветвей дендрограммы, изображающих d_{min}).

По относительной энтропии Шеннона. Мы предполагаем, что кластеризация наиболее информативна в том случае, когда кластеры в наибольшей степени «сбалансированы» по числу ячеек, либо по суммарной «массе» аддитивного признака Z. Нами построена количественная мера «сбалансированности» кластеров, основанная на энтропии Шеннона вектора q (в битах):

$$H(q) = -\sum_{i=1}^{n} q_i \log_2 q_i \ ,$$

где $q_i \ge 0, \sum_{i=1}^{n} q_i = 1$ при этом величина $\#(q) = 2^{H(q)}/n$ может рассматриваться как относительная характеристика однородности элементов q:

$$\#((1,0,...,0)) = 1/\mathrm{n} \le \#(\mathrm{q}) \le \#((1/\mathrm{n},1/\mathrm{n},...,1/\mathrm{n})) = 1.$$

Предлагается выбрать число кластеров, максимизирующее $\#(q)$, где q получается нормировкой: 1) численностей объектов в кластерах; 2) итогов аддитивного признака по кластерам. В дальнейшем приводятся результаты вычислений согласно итогам признака по кластерам.

По корреляционному отношению, связанному с непараметрической оценкой плотности – интенсивности точечного процесса с постоянным параметром сглаживания σ и гауссовым ядром[7]: $\phi(z,\sigma) = \exp(-\|z\|^2/2\sigma^2), z \in R^2$. Если наблюдаемые значения $v_1, v_2, ..., v_n$ располагаются в точках $x_1, x_2, ..., x_n$ соответственно, сглаженное значение в точке u определяется как [5]:

$$G_\sigma(u) = \sum_{i=1}^{n} v_i \phi(u - x_i, \sigma) \Big/ \sum_{i=1}^{n} \phi(u - x_i, \sigma), u \in R^2.$$

В нашем случае $v_i = \ln p_i$ (выбор параметра сглаживания будет рассмотрен далее). Пусть в соответствии с одним из вышеперечисленных подходов выполнена пространственная кластеризация с числом кластеров k. Обозначим C_i, $i = 1, 2, ..., k$ – кластеры как подмножества точек на плоскости, полученные объединением ячеек Дирихле. Обозначим площадь

[7] Мы использовали функцию density.ppp пакета-расширения spatstat [3], в котором реализован метод P.J. Diggle [5].

i-го кластера $\#(C_i) = \int_{C_i} dxdy$. Если $g_\sigma(x,y)$ – сглаженное значение плотности, тогда среднее значение плотности по кластеру C_i определяется так: $\overline{g}_i = \int_{C_i} g_\sigma(x,y)dxdy \Big/ \#(C_i)$.

Рассмотрим корреляционное отношение, определяемое по сглаженной плотности соответственно рассматриваемой кластеризации [15]:

$$\eta_k^2(\sigma) = 1 - \frac{\sum_{i=1}^{k} \int_{C_i} (g_\sigma(x,y) - \overline{g}_i)^2 dxdy}{\sum_{i=1}^{k} \int_{C_i} (g_\sigma(x,y) - \overline{g})^2 dxdy},$$

где $\overline{g} = \sum_{i=1}^{k} \int_{C_i} g_\sigma(x,y)dxdy \Big/ \sum_{i=1}^{k} \#(C_i)$ – генеральное среднее значение плотности. Корреляционное отношение представляет собой долю «разброса» значений плотности, объясненную кластеризацией; будем рассматривать величину корреляционного отношения в качестве характеристики «информативности» кластеризации.

Пусть требуется выбрать «наилучший» вариант кластеризации, т.е. число кластеров k в пределах от 2 до некоторого k_{max} (мы использовали $k_{max} = 30$). Как правило, с ростом k корреляционное отношение возрастает, однако следует учесть, что при определении сглаживания требуется задать определенное значение параметра сглаживания σ.

Пусть зафиксирован ряд его значений $\sigma_1, \sigma_2, ..., \sigma_m$ (например, $\sigma = 0.1, 0.2, \ldots, 1$ для $m=10$). Пусть матрица $ETA2$ содержит значения корреляционного отношения для различных значений k и σ:

$$ETA2_{ij} = \eta_{1+i}^2(\sigma_j) \quad i = 1,2,...,k_{\max} - 1,\ j = 1,2,...,m.$$

Также пусть задана нижняя граница корреляционного отношения $\eta_{\min}^2 \le \max_{i,j} ETA2_{ij}$. Обозначим j^* номер первого столбца $ETA2$, для которого имеется хотя бы один элемент, не меньший, чем $\eta_{\min}^2$:

$$j^* = \min_{1 \le j \le m} \left[\max_{1 \le i \le k_{\max} - 1} ETA2_{ij} \ge \eta_{\min}^2 \right].$$

Оптимальное число кластеров предлагается выбрать как соответствующее первому элементу в столбце j^* матрицы $ETA2$, не меньшее, чем $\eta_{\min}^2$:

$$k^* = 1 + \min_{1 \le i \le k_{\max} - 1} \left[ETA2_{ij^*} \ge \eta_{\min}^2 \right].$$

Отметим, что в примерах реализации данного метода принято значение минимального корреляционного отношения $\eta_{\min}^2 = 0.6$, однако для цен на коммерческую недвижимость это значение оказалось слишком большим и было заменено на 0.5.

Кластеризация пространственных данных на примере г. Саратова

Исследования пространственных аспектов социально-экономических явлений в г. Саратове в основном опирались на пространственную статистику и эконометрику (например, [14]), а также на использование нейронных сетей [16].

Для иллюстрации предлагаемых методов воспользуемся данными о размещении населения, предприятий торговли г. Саратова, коммерческой недвижимости по состоянию на конец 2006 г. – начало 2007 г. Данные о населении – численности жителей 18 лет и старше, приписанных к избирательным участкам, были предоставлены городской избирательной комиссией; данные о предприятиях торговли были предоставлены торговыми отделами районных администраций г. Саратова. Выборочная информация об объектах коммерческой недвижимости была получена из различных печатных изданий и Интернет. Для геокодирования информации была использована ГИС Управления по архитектуре г. Саратова.

С целью устранения очень малых по размеру ячеек Дирихле перед реализацией основного алгоритма пространственной кластеризации было выполнено предварительное агрегирование близко расположенных объектов. Для этого использовался обычный кластерный анализ, использующий евклидово расстояние. Кластеризация останавливалась, когда расстояние между кластерами превышало 165 м. Размеры агрегированных объектов были получены как суммы размеров, а их центры – как центры тяжести объединяемых объектов.

Пространственной кластеризации были подвергнуты данные о следующих семи признаках: населении; общей и торговой площади продовольственных и промтоварных магазинов, цене и площади объектов коммерческой недвижимости.

На основе соответствующих картосхем авторами экспертно была оценена «информативность» результирующей кластеризации для каждой комбинации вида расстояния между кластерами и оптимального числа кластеров, выбранного по каждому из трех методов (все результаты доступны по адресу: http://mitrofanovay.ucoz.com/index/klasterizacija/0-12).

Информативность каждого варианта кластеризации оценивалась как «низкая», «средняя» и «высокая», при этом учитывалась потенциальная полезность кластеризации для практических нужд. Варианты кластеризации с очень малым числом кластеров, а также варианты, в которых имеется один кластер, включающий подавляющую часть изучаемой территории города, рассматривались как не информативные. Результаты представлены в табл. 2.

Таблица 2

Оптимальные числа кластеров и экспертные оценки информативности кластеризации[1]

Кластеризуемый признак	Вид расстояния между кластерами	Метод определения числа кластеров			
		По скачку расстояния	По относительной энтропии	По корреляционному отношению	
				$\eta^2_{\min} = 0.6$	$\eta^2_{\min} = 0.5$
Население	Ближн. сосед	22 (1)	2 (1)	20 (1)	20 (1)
	Дальн. сосед	5 (2)	2 (1)	17 (3)	17 (3)
	Ср. расст.	4 (1)	2 (1)	19 (3)	19 (3)

Окончание табл. 2

Кластеризуемый признак	Вид расстояния между кластерами	Метод определения числа кластеров			
		По скачку расстояния	По относительной энтропии	По корреляционному отношению	
				$\eta^2_{\min} = 0.6$	$\eta^2_{\min} = 0.5$
Продовольств. магазины – общая площадь	Ближн. сосед	4 (1)	2 (1)	21 (1)	11 (1)
	Дальн. сосед	2 (1)	8 (3)	4 (2)	12 (3)
	Ср. расст.	2 (1)	4 (1)	6 (3)	14 (2)
Продовольств. магазины – торговая площадь	Ближн. сосед	5 (1)	2 (1)	23 (1)	16 (1)
	Дальн. сосед	2 (1)	9 (3)	5 (2)	12 (3)
	Ср. расст.	7 (2)	3 (1)	8 (3)	16 (3)
Промтоварные магазины – общая площадь	Ближн. сосед	2 (1)	2 (1)	24 (1)	24 (1)
	Дальн. сосед	2 (1)	2 (1)	17 (3)	20 (3)
	Ср. расст.	3 (1)	2 (1)	21 (3)	21 (3)
Промтоварные магазины – торговая площадь	Ближн. сосед	3 (1)	2 (1)	26 (1)	24 (1)
	Дальн. сосед	3 (1)	3 (1)	12 (3)	12 (3)
	Ср. расст.	5 (1)	2 (1)	15 (3)	17 (3)
Коммерческая недвижимость – цена	Ближн. сосед	2 (1)	2 (1)	-	20 (2)
	Дальн. сосед	2 (1)	2 (1)	-	28 (3)
	Ср. расст.	3 (1)	2 (1)	-	16 (3)
Коммерческая недвижимость – площадь	Ближн. сосед	2 (1)	30 (3)	8 (1)	10 (1)
	Дальн. сосед	2 (1)	2 (1)	9 (3)	13 (3)
	Ср. расст.	2 (1)	5 (1)	7 (1)	10 (2)

[1] 1 – низкая, 2 – средняя, 3 – высокая информативность (в скобках).

С помощью алгоритма рекурсивного подразделения rpart пакета R [13] было построено классификационное дерево, предсказывающее экспертную оценку качества кластеризации на основании трех факторов: кластеризуемого признака (первый столбец табл. 2), вида расстояния между кластерами и метода выбора числа кластеров. Общее число проанализированных вариантов составило 81 (3 варианта из 84 были исключены, так как они соответствуют отсутствующим вариантам для цен коммерческой недвижимости). Построенное классификационное дерево включает три терминальных узла («листа»). Первый «лист» (42 варианта) выделяется комбинацией двух методов определения числа кластеров – по относительной энтропии и по скачку расстояния и содержит 37 случаев (88%) «низко информативных» вариантов; в то же время, он содержит также 2 «средне информативных» и 3 «высоко информативных». Второй «лист» (13 вариантов), выделяемый методом определения числа кластеров, основанным на корреляционном отношении и расстоянием между кластерами по ближайшему соседу, также, в целом, дает «низко информативные» результаты (12 вариантов или 92%). Наконец, третий «лист» (26 вариантов),

выделяемый комбинацией метода выделения числа кластеров, основанной на корреляционном отношении и расстояниях между кластерами по дальнему соседу либо среднему расстоянию, чаще всего дает «высоко информативную» (21 вариант или 81%), а также «средне информативную» (4 варианта или 15%) кластеризацию. Это позволяет рекомендовать для приложений именно такую комбинацию вариантов расстояния и выбора числа кластеров, хотя, в отдельных случаях другие методы также могут давать приемлемые результаты.

Отметим, что предлагаемый метод кластеризации может давать как односвязные, так и многосвязные кластеры. По нашему мнению, наличие многосвязных кластеров отражает реальную структуру городской среды, например, наличие нескольких, расположенных по соседству зон высокой плотности торговых площадей может свидетельствовать о формировании нового делового квартала. На раннем этапе такой кластер является многосвязным, однако, возможно, по мере формирования делового района его число компонентов связности будет уменьшаться.

Заключение

Городская среда имеет сложную структуру, которая отражается в наличии значительного числа вариантов, формирующих городскую среду в каждом заданном пункте. При решении практических вопросов возникает необходимость сравнения таких вариантов, причем эта задача часто плохо сформулирована – критерий сравнения задан недостаточно четко, либо имеется несколько критериев, либо практически невозможно собрать весь объем информации, требуемый для ранжирования вариантов.

К подобным данным можно применить предлагаемый алгоритм иерархической пространственной кластеризации, дополненный методом выбора числа кластеров. Найденные кластеры упорядочиваются по убыванию средней плотности «перспективности», далее сопоставляются ячейки, входящие в наиболее «перспективные» кластеры, что позволяет оптимизировать затраты на сопоставление вариантов местоположения.

Отличительная черта предлагаемого метода кластеризации – использование «комбинированного» расстояния, учитывающего как пространственную удаленность объектов, так и относительное отклонение плотностей (интенсивностей). В работе на основе данных за 2006-2007 гг. по г. Саратову проанализированы различные комбинации расстояний между кластерами и методов выбора числа кластеров (по максимальному скачку расстояния, относительной энтропии и корреляционному отношению). Сравнение «информативности» результатов кластеризации позволяет рекомендовать только две комбинации: расстояние между кластерами по дальнему соседу, либо среднее расстояние; выбор числа кластеров на основе корреляционного отношения (мы рекомендуем граничное значение 0.6 или 0.5 как «запасной» вариант). Такой выбор для данных по г. Саратову в 81% случаев дает хорошие, а в 15% случаев удовлетворительные результаты. Исполь-

зование методов выбора числа кластеров по максимальному скачку расстояния, а также по относительной энтропии и/или расстояния по ближайшему соседу, в большинстве случаев, дает малоинформативную кластеризацию. Таким образом, метод выбора числа кластеров по максимальному скачку расстояния, часто используемый в «непространственной» статистике, для пространственных данных не дает удовлетворительных результатов (по крайней мере, применительно к используемому нами определению расстояния).

Предлагаемый метод часто приводит к многосвязным кластерам; мы не считаем эту особенность недостатком алгоритма. Напротив, мы выдвигаем гипотезу о том, что число односвязных компонентов кластера можно рассматривать как характеристику «зрелости» соответствующей городской структуры: с течением времени кластер должен приближаться к односвязному (имеющиеся в нашем распоряжении данные не позволили проверить эту гипотезу).

Все предлагаемые алгоритмы реализованы в виде набора функций для свободно распространяемого статистического пакета R и предоставляются авторами по запросу (например, через сайт http://mitrofanovay.ucoz.com/).

Возможным продолжением данного исследования может быть анализ метода выбора числа кластеров, основанный на верхней границе коэффициента вариации плотности по кластерам; данный подход не противоречит интуиции, и предварительное изучение показало его перспективность.

Список источников

1. Azzalini, A. R package 'pdfCluster': Cluster analysis via nonparametric density estimation (version 0.1-9) [электронный ресурс] / A. Azzalini, G. Menardi, T. Rosolin. [2004]. URL: http://cran.r-project.org/package=pdfCluster.

2. Azzalini, A. Clustering via nonparametric density estimation [текст] / A. Azzalini, N. Torelli // Statistics and Computing. – 2007. – № 17. – С. 71 – 80.

3. Baddeley, A. Spatstat: an R package for analyzing spatial point patterns [электронный ресурс] / A. Baddeley, R. Turner // Journal of Statistical Software. – 2005. – № 12(6). – С. 1 – 42. – ISSN: 1548-7660. URL: www.jstatsoft.org.

4. Bivand, R.S. Applied spatial data analysis with R [текст] / R.S. Bivand, E.J. Pebesma, V. Gomez-Rubio. – NY: Springer. – 2008. URL: http://www.asdar-book.org/.

5. Diggle, P.J. A kernel method for smoothing point process data [текст] / P.J. Diggle // Applied Statistics (Journal of the Royal Statistical Society, Series C). – 1985. – № 34. – С. 138 – 147.

6. Ester, M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [текст] / M. Ester, H.-P. Kriegel, J. Sander, Xiaowei Xu // Institute for Computer Science, University of Munich. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). – 1996.

7. Fujita, M. The Spatial Economy: Cities, Regions, and International Trade [текст] / M. Fujita, P. Krugman, A.J. Venables. – Cambridge, Massachusetts: The

MIT Press. – 2001.

8. Hennig, C. [электронный ресурс] fpc: Flexible procedures for clustering. R package version 2.0-3 / C. Hennig. [2010]. URL: http://CRAN.R-project.org/package=fpc.

9. Pebesma, E.J. Classes and methods for spatial data in R [электронный ресурс] / E.J. Pebesma, R.S. Bivand // R News. – 2005. – № 5(2). URL: http://cran.r-project.org/doc/Rnews/.

10. R Development Core Team. R: A language and environment for statistical computing [электронный ресурс] // R Foundation for Statistical Computing, Vienna, Austria. – ISBN 3-900051-07-0. [2010]. URL: http://www.R-project.org/.

11. Rajala, T. spatgraphs: Graphs for spatial point patterns. R package version 2.44 [электронный ресурс] / T. Rajala. [2011]. URL: http://CRAN.R-project.org/package=spatgraphs.

12. Struyf, A. Clustering in an Object-Oriented Environment [электронный ресурс] / A. Struyf, M. Hubert, P.J. Rousseeuw // Journal of Statistical Software. – 1996. – № 1. URL: http://www.stat.ucla.edu/journals/jss/.

13. Therneau, T. rpart: Recursive Partitioning. R package version 3.1-48 [электронный ресурс] / T. M. Therneau, B. Atkinson, B. Ripley. [2010]. URL: http://CRAN.R-project.org/package=rpart.

14. Балаш, В.А. Пространственная корреляция в статистических исследованиях [текст] / В.А. Балаш, А.Р. Файзлиев // Вестн. Сарат. гос. соц.-экон. ун-та. – 2008. – № 4(23). – С. 122 – 125.

15. Корреляционное отношение [электронный ресурс]. URL: http://slovari.yandex.ru/~книги/Энциклопедия социологии/Корреляционное отношение/.

16. Салахутдинов, Р.З. Применение карты Кохонена для анализа цен объектов недвижимости [текст] / Р.З. Салахутдинов, М.Г. Тиндова // Вестн. Сарат. гос. соц.-экон. ун-та. – 2006. – № 13(2). – С. 117 – 119.

DENSITY CLUSTERING WITH A COMBINED DISTANCE AS A TOOL FOR THE ANALYSIS OF URBAN ENVIRONMENT

Mitrofanov Aleksey Yuryevich,
Ph. D. of Economy, Associate Professor of the Chair of Applied Mathemaics and Informatics of Saratov Social and Economic Institute (filial-branch of) Russian Economic University named by G.V. Plekhanov; MitrofanovAY0@gmail.com
Phayzliyev Aleksey Raisovich,
Lecturer of the Chair of Information Systems in Economy of Saratov Social and Economic Institute (filial-branch of) Russian Economic University named by G.V. Plekhanov; FaizlievAR1983@mail.ru

A new method for clustering spatial data, based on distance, generalizing the difference density of the distribution of the quantitative trait and spatial distance is offered. Two methods to select the best number of clusters are considered, one of which is based on the non-parametric density estimation and correlation ratio. As an illustration a comparison of the results of cluster polarization of data for Saratov in 2006-2007; algorithm is realized as a set of functions for the statistical package R, available on request to the authors.
Keywords: exploratory data analysis, clustering, spatial data, density, smoothing, correlation ratio, Shannon entropy, statistical package R.