УДК 519.863

М. О. Солнцева[1], *Б. Г. Кухаренко*[1,2]

[1]Московский физико-технический институт (государственный университет)
[2]Институт машиноведения РАН

Применение методов кластеризации узлов на графах с разреженными матрицами смежности в задачах логистики

Рассматривается решение задачи оптимального расположения логистических центров в сети поставок на графе транспортной сети с помощью различных алгоритмов кластеризации. Разбиение узлов транспортной сети на кластеры выполняется после предварительного анализа данных, представляемых сильно разреженной матрицей смежности графа этой сети. В качестве альтернативного подхода к кластеризации узлов на графе рассматриваются процедуры построения и усечения минимального дерева Штейнера.

Ключевые слова: анализ данных, анализ графов, разреженные матрицы смежности, алгоритмы кластеризации, деревья Штейнера, транспортные сети, логистика.

Введение

При проектировании логистических сетей решается задача оптимального расположения логистических центров [1]. Принято выделять три основные модели расположения логистических центров: создание единого центра по обслуживанию потребителей (single sink network design problem); расположение нескольких новых центров (multi sink network design problem); разбиение всех потребительских центров на подмножества и создание в каждом подмножестве своей сети (selective network design problem) [2]. Точное решение этих задач на неориентированном графе принадлежит к классу NP-сложных задач.

Настоящая работа посвящена решению задачи оптимального расположения логистических центров при проектировании цепи поставок (facility location problem) с помощью алгоритма кластеризации K-means [3] и его модификаций [4] и также с помощью построения и усечения дерева Штейнера [5]. При решении этой задачи существует ограничение, налагаемое топологией транспортной сети, которая задаётся матрицей смежности (affinity matrix) графа этой сети. Для построения графов транспортной сети используются данные проекта OpenStreetMap [6]. На рис. 1 приведён пример графа транспортной сети Омского региона. До настоящего времени ограничения, налагаемые реальной транспортной сетью, при решении задачи оптимизации не принимались во внимание. Введение такого ограничения необходимо для регионов с транспортной сетью, представимой разреженной матрицей смежности. При решении поставленной задачи разбиение графа транспортной сети на произвольное число кластеров нецелесообразно. Необходимо предварительно выявить скрытую структуру графа, которая отражена в его матрице смежности. Анализ матрицы смежности позволяет оценить количество кластеров в структуре графа. Такую оценку можно использовать в качестве входного параметра для алгоритмов кластеризации узлов, заданных на графе.

Алгоритм ALA (Alternate Location Allocation) [7], обычно используемый для решения этой задачи, не учитывает топологию транспортной сети, поэтому его результаты применимы только для областей с хорошо развитой транспортной инфраструктурой [8]. Существенным недостатком этого алгоритма является то, что его целевая функция не достигает глобального минимума и зависит от начальных условий, задаваемых случайным образом. Наиболее близким к алгоритму ALA является алгоритм кластеризации K-means [3].

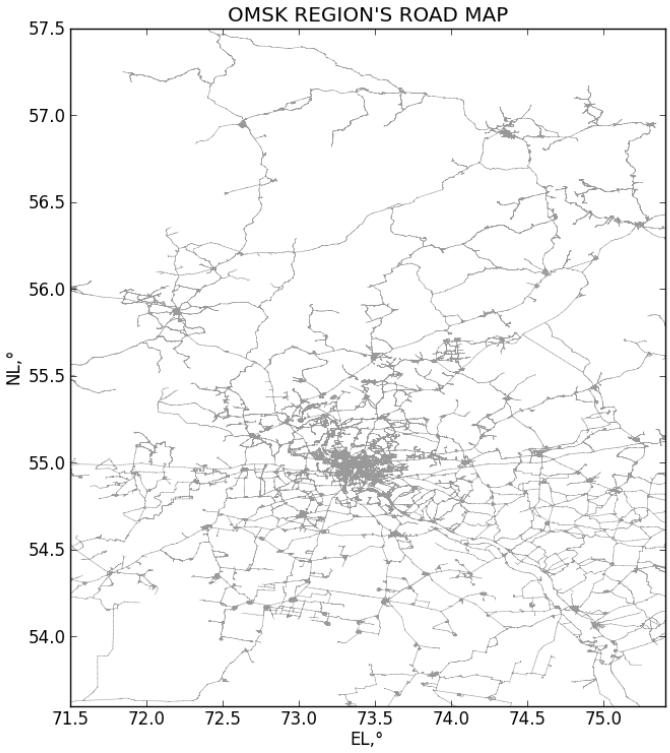

Рис. 1. Граф транспортной сети Омского региона, построенный по данным OpenStreetMap

Для того чтобы оценить максимально допустимое число логистических центров, в настоящей работе проводится предварительный анализ скрытой структуры графа транспортной сети, представленного разреженной матрицей смежности. Такой анализ выполняется с помощью алгоритма кросс-ассоциаций [9], выделяющего в матрице смежности кластеры с однородной внутренней структурой, в результате чего определяется число кластеров в структуре графа, которое является входным параметром для применяемых далее алгоритмов кластеризации. Алгоритмы K-means [3] и K-means++ [4] используются для решения задачи выделения нескольких логистических центров в сети поставок (multi sink network design problem). Процедуры построения и усечения дерева Штейнера [5] применяются для выделения как нескольких логистических центров (multi sink network design problem), так и единого центра (single sink network design problem).

В настоящей работе разработан комплекс программ на языке C++ и других объектно-ориентированных языках, реализующих указанные выше алгоритмы кластеризации [3, 4, 5, 9], который позволяет производить численные эксперименты на графах реальных транспортных сетей.

Материалы и методы

Данные OSM

Для построения графа транспортной сети используются массивы данных проекта OpenStreetMap [6]. В настоящей работе исследуется граф транспортной сети Омского региона (см. рис. 1). Исходный граф включает 140 000 узлов. Это число сокращается до 38 777 за счёт исключения несущественной информации. В описании графа сохраняются узлы, соответствующие исходному пункту каждой дороги, пункту её окончания и пересечениям с другими дорогами. Объём данных, представляющий эту разреженную матрицу смежности с 100 378 ненулевых элементов, занимает на диске 3 Гб и 1,5Гб в памяти компьютера.

Предварительный анализ матрицы смежности графа транспортной сети. Алгоритм кросс-ассоциаций

В настоящей работе для предварительного анализа, выявляющего скрытую структуру графа, используется метод кросс-ассоциаций [9]. Этот метод предлагает общую модель, основанную на сжатии данных согласно принципу MDL (Minimum description length) [10].

Структура графа отображается в его матрице смежности – бинарной матрице из нулей и единиц, поэтому проблема её анализа аналогична задачам анализа ассоциативных правил, информационного поиска и выделения групп в сетях и т.п. [9]. Во всех этих случаях маркировка столбцов и строк не имеет значения. Группировка независимых строк и столбцов матрицы основывается на их сходстве. Эти перекрёстные ассоциации выявляют скрытую структуру графа. Соответствующие прямоугольные области различной плотности используются для быстрого продвижения по структуре матрицы.

Если исходные элементы матрицы составляют $m \times n$ «прямоугольников» с «плотностью» либо 0, либо 1, то по завершении процедуры выявляются прямоугольные блоки с плотностью от 0 до 1, число которых подлежит определению. Отбрасывание части этих прямоугольников уменьшает сложность описания данных. Для оценки сложности описания матрицы смежности в используемом алгоритме кросс-ассоциаций [9] применяется MDL-принцип [10], в котором стоимость структуры данных, выделяемой в матрице, оценивается числом бит, необходимых для передачи всей структуры вместе с данными о каждой выделенной прямоугольной области.

Пусть $D = [d_{ij}]$ — матрица бинарных данных размерности $m \times n$, где m и n – количество строк и столбцов матрицы соответственно. Припишем строки к группам строк и столбцы к группам столбцов:

$$\Psi : \{1, 2, ..., m\} \to \{1, 2, ..., k\}, \Phi : \{1, 2, ..., n\} \to \{1, 2, ..., l\},$$

где k и l обозначают число независимых групп среди строк и столбцов соответственно. Согласно этому приписанию (или «перекрёстной ассоциации» $\{\Psi,\ \Phi\}$) элементы матрицы D перестраиваются таким образом, чтобы строки, соответствующие группам 1, 2 и т.д., перечислялись в соответствии с этим порядком и аналогично – для столбцов. Такая перестройка разделяет исходную матрицу D на прямоугольные блоки меньшего размера – подматрицы D_{ij} размерности (a_i, b_j), $i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, l$.

Введём обозначения: A – бинарная матрица, размерности $a \times b$, $n_1(A)$ – число ненулевых элементов матрицы A, $n_0(A)$ – число нулевых элементов матрицы A, $n(A) = n_1(A) + n_0(A) = a \times b$, $P_A(i) = \frac{n_i(A)}{n(A)}$, $i = 1, 0$.
Тогда полная длина кода в битах

$$C(A) = \sum_{i=0}^{1} n_i(A) \log \left(\frac{n(A)}{n_i(A)} \right) = n(A) H(P_A(0)),$$

где H – бинарная энтропия Шэннона. Полная длина кода для матрицы D с учётом заданной кросс-ассоциации имеет вид

$$T(D; k; , l; \Psi; \Phi) =$$

$$= \log^* k + \log^* l + \sum_{j=1}^{l-1} \lceil \log \bar{a}_i \rceil + \sum_{j=1}^{l-1} \lceil \log \bar{b}_j \rceil + \sum_{i=1}^{k} \sum_{j=1}^{l} \lceil \log(a_i b_j + 1) \rceil + \sum_{i=1}^{k} \sum_{j=1}^{l} C(D_{i,j}),$$

где $\bar{a}_i = \left(\sum_{t=i}^{k} a_t \right) - k + i, \quad i = 1, ..., k-1; \bar{b}_j = \left(\sum_{t=j}^{l} b_t \right) - l + j, \quad j = 1, ..., l-1.$

Оптимальная кросс-ассоциация соответствует числу групп строк k^*, числу групп столбцов l^* и кросс-ассоциации $\{\Psi^*,\ \Phi^*\}$ таких, что полная результирующая длина кода $T(D; k^*; l^*; \Psi^*; \Phi^*)$ минимизируется.

Задача определения оптимальной кросс-ассоциации является вычислительно сложной, поэтому для её решения применяется эвристика [9], состоящая из двух шагов:

1) Внутренний цикл: для заданных чисел k и l найти подходящую перегруппировку (т.е. кросс-ассоциацию), соответствующую достижению локального минимума функции

$$\sum_{i=1}^{k}\sum_{j=1}^{l} C(D_{i,j}). \tag{1}$$

2) Внешний цикл: поиск наилучших k и l среди тех, которые рассматриваются во внутреннем цикле. Этот этап использует резкое падение функции (1) при малых значениях k и l (см. рис. 2, где $k = l = N$).

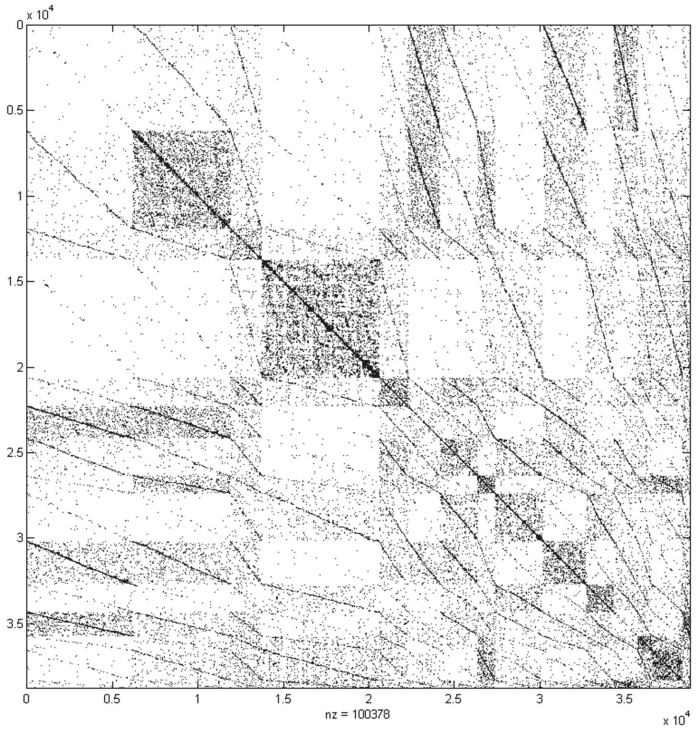

Рис. 2. Значения функции (1) в зависимости от выделенного числа однородных кластеров в матрице смежности графа транспортной сети Омской области (см. рис. 3)

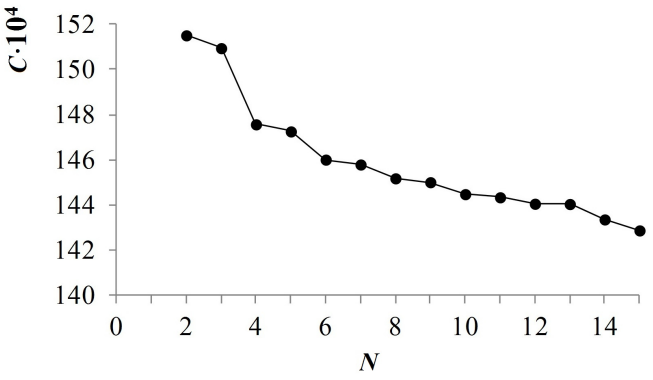

Рис. 3. Результат применения алгоритма кросс-ассоциаций к матрице смежности графа транспортной сети Омской области, представленного на рис. 1

Оценка сложности алгоритма: на каждом шаге алгоритма возрастает либо k, либо l, поэтому сумма $k + l$ всегда возрастает на 1. Следовательно, общая сложность алгоритма

оценивается как $O(n_i(D)(k^* + l^*)^2$. На практике для нахождения кросс-ассоциации достаточно 20 итераций.

На рис. 3 показаны результаты применения алгоритма поиска кросс-ассоциаций [9] для матрицы смежности графа транспортной сети Омской области.

Алгоритм кластеризации K-means

Пусть $\{\mathbf{X}_1, \ldots, \mathrm{X_N}\}$ — векторное представление кластеризуемых узлов, $\{\mu_k\}$ — набор векторов, соответствующих центрам кластеров, $r_{nk} \in \{0, 1\}$ – индикаторные переменные, указывающие на приписание узлов кластерам. Алгоритм K-means [3] минимизирует суммарное среднеквадратичное отклонение узлов кластера от его центра:

$$\min \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{X}_n - \mu_k \right\|^2 \right\},$$

откуда $\mu_k = \frac{\sum_n r_{nk} \mathbf{X}_n}{\sum_n r_{nk}}$.

Сложность алгоритма определяется как $O(N^{\,2K+1} \log\, N)$, N – число узлов, K – число кластеров. Этот алгоритм гарантирует нахождение локальной оптимальной конфигурации кластеров, но не применим для нахождения глобального оптимума. Алгоритм K-means [4] имеет по меньшей мере два основных недостатка. Первый из них – время работы алгоритма описывается сверхполиномиальной зависимостью от числа кластеров K. Второй – произвольная аппроксимация недостаточно хорошо оптимизирует целевую функцию. Чтобы избежать главным образом второго недостатка, используется алгоритм K-means++ [4]. Сначала этот алгоритм осуществляет процедуру оптимальной инициализации центров кластеров, затем выполняются стандартные итерации алгоритма K-means [3]. Благодаря процедуре инициализации сложность алгоритма K-means++ определяется как $O(\log\, K)$.

Алгоритм кластеризации K-means++

Рассмотрим n-мерное евклидово пространство (хотя подход может быть представлен и в более общем виде). Пусть имеется множество точек X и требуется найти множество C, состоящее из K точек, называемых центрами кластеризации, которые минимизируют ошибку

$$Err(C, X) = \frac{1}{|X|} \sum_{x \in X} d(x, C(x)),$$

где точка $C(x)$ из множества C является ближайшим центром к точке x.

$$C(x) = \arg\min_{c \in C} d(x, c).$$

На первом шаге алгоритма происходит выбор первого центра c [1] кластеризации из однородного распределения точек множества X. Далее выполняется выбор k-го центра $c[k]$ $(k = 2, ..., K)$ из мультиномиального распределения на множестве X, где точка x имеет вероятность

$$\theta_x = \frac{D(x)^2}{\sum_{x' \in X} D(x')^2} \propto D(x)^2.$$

$D(x)$ определяется как расстояние до ближайшего существующего центра.

После того как начальные центры выбраны, переходят к стандартной процедуре кластеризации K-means [3].

Процедуры построения и усечения дерева Штейнера

Алгоритм построения и усечения дерева Штейнера также можно использовать для выделения центров кластеризации. Полное описание алгоритмов построения дерева Штейнера (Steiner tree) можно найти в работе [5].

Пусть $G = (V, E)$ — неориентированный граф, в котором каждое ребро $ij \in E$ имеет неотрицательную реальную стоимость, и пусть N — множество концевых узлов (вершин) графа, $N \subseteq V$. Дерево T, выделяемое в графе G, называют деревом Штейнера, если оно содержит все концевые узлы графа $V(T) \supseteq N$. Требуется найти такое дерево Штейнера T, стоимость которого $c(T) = \sum_{ij \in E(T)} c(ij)$ минимальна среди всех деревьев Штейнера для графа G. Оптимальное решение T является минимальным деревом Штейнера.

Для усечения дерева Штейнера строят сеть расстояний $D_G(N)$, где N — выделенное указанное множество концевых узлов, и затем находят минимальное остовное дерево M в $D_G(N)$. Заменяя каждое ребро $v\ w$ в M кратчайшим путём $v\ -w$ в G, получают новый подграф $T\,'$ в G. В результате находят минимальное остовное дерево $T\,'$, последовательно удаляя любые узлы Штейнера первой степени.

Результаты и обсуждения

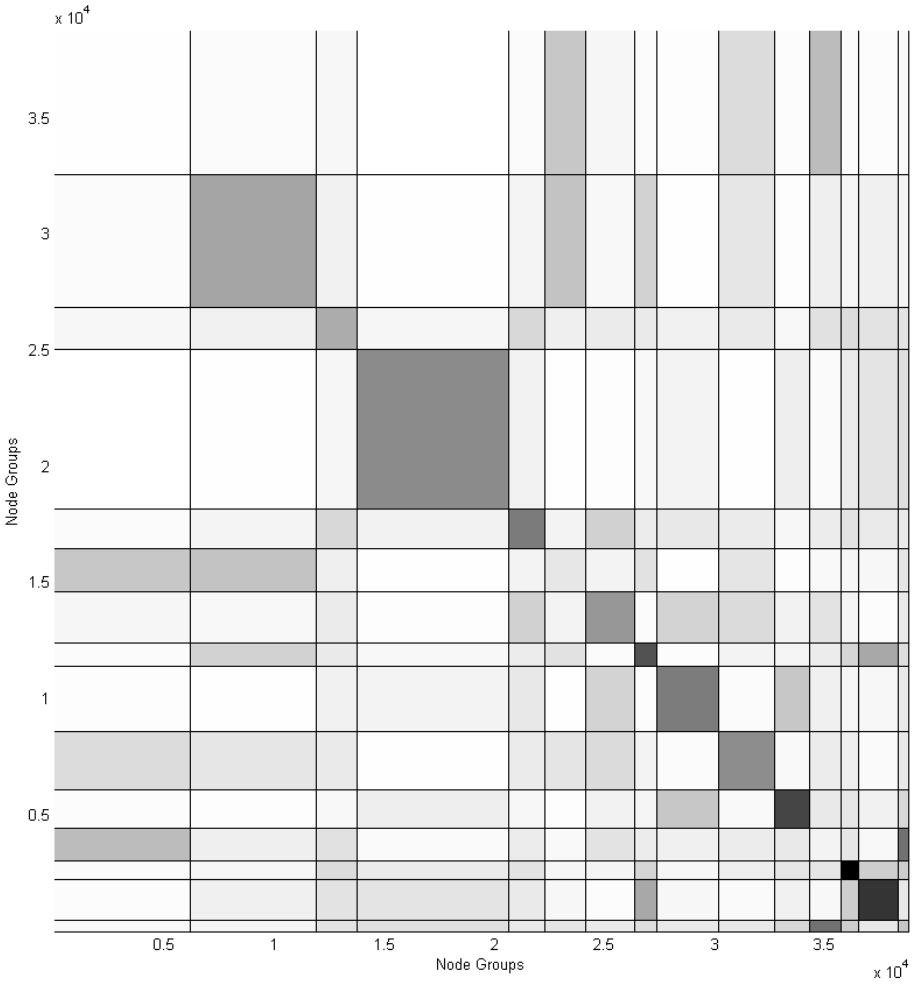

Рис. 4. Схематичное представление структуры матрицы смежности графа транспортной сети Омской области (см. рис. 3). Чем темнее прямоугольник, тем выше плотность связей между узлами в нём

В настоящей работе задача оптимизации расположения логистических центров в условиях ограничений решается с помощью алгоритмов кластеризации после того, как выявлена структура разреженной матрицы смежности графа рассматриваемой транспортной сети. Следует отметить, что ранее проблема оптимизации расположения логистических

центров при ограничениях транспортной сетью в литературе не рассматривалась.

Результаты предварительного анализа графа транспортной сети Омской области (см. рис. 3), выполненные с помощью метода кросс-ассоциаций [9], показали, что в структуре этого графа выделяются 15 кластеров, расположенных на главной диагонали матрицы смежности графа. На рис. 4 показано идеализированное изображение структуры матрицы смежности, представленной на рис. 3. Насыщенность цвета прямоугольников на рис. 4 зависит от величины плотности связей в каждом выделенном кластере. Чем темнее прямоугольник, тем больше плотность связей в кластере.

Отметим, что полученное на этом этапе число кластеров отражает всего лишь общую структуру матрицы смежности, в то время как для размещения логистических центров имеет смысл рассматривать только те кластеры, плотность связей в которых выше некоторого среднего уровня. На рис. 5 показано распределение значений плотности связей в выделенных 15 кластерах. Как видно из рис. 5, только 8 кластеров имеют плотность связей выше среднего значения. Это число кластеров следует рассматривать как предельное количество логистических центров, к которым можно приписать выделенные узлы графа. В зависимости от числа рассматриваемых узлов можно варьировать число размещаемых логистических центров.

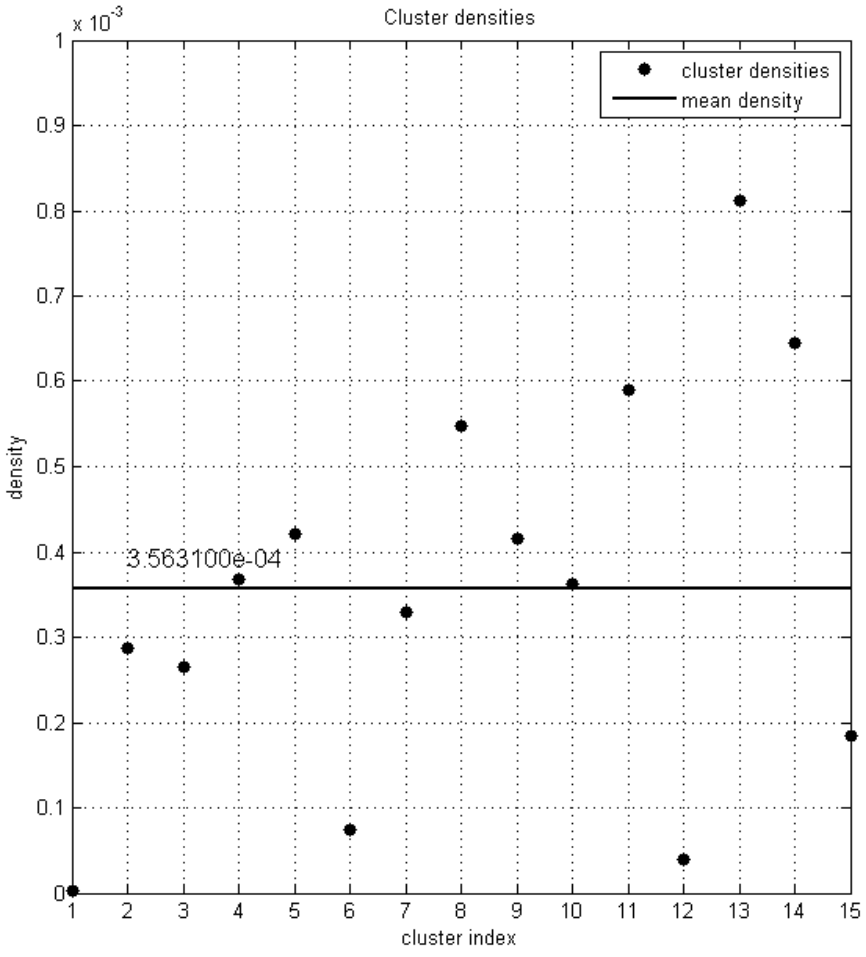

Рис. 5. Точки графика показывают значения плотности связей в каждом из 15 выделенных кластеров. Сплошная линия соответствует средней плотности связей в выделенных кластерах

Для решения задачи оптимального расположения новых логистических центров в сети поставок в настоящей работе рассматривались наиболее популярные алгоритмы класте-

ризации, такие как K-means и его модификации [3, 4], и также процедуры построения и усечения дерева Штейнера [5]. Исходные целевые функции этих алгоритмов не учитывают естественные ограничения, возникающие в конкретных приложениях, где наиболее значительным ограничением является транспортная сеть. В настоящей работе такие ограничения были учтены. Для построения графа транспортной сети были использованы массивы данных проекта OpenStreetMap [6]. Работа программ демонстрируется на примере региона Омской области.

На рис. 6 приведены примеры кластеризации с помощью алгоритмов K-means [3] и K-means++ [4]. При этом искомые центры кластеризации соединяются с каждым рассматриваемым узлом по маршруту, вычисляемому алгоритмом Дейкстры [11]. С помощью алгоритма K-means++ [4] можно не только существенно уменьшить время работы алгоритма K-means [3], но и получить более рациональное разбиение на кластеры, соответствующее минимальному расстоянию между узлами в кластере, как это видно из рис. 6 (б).

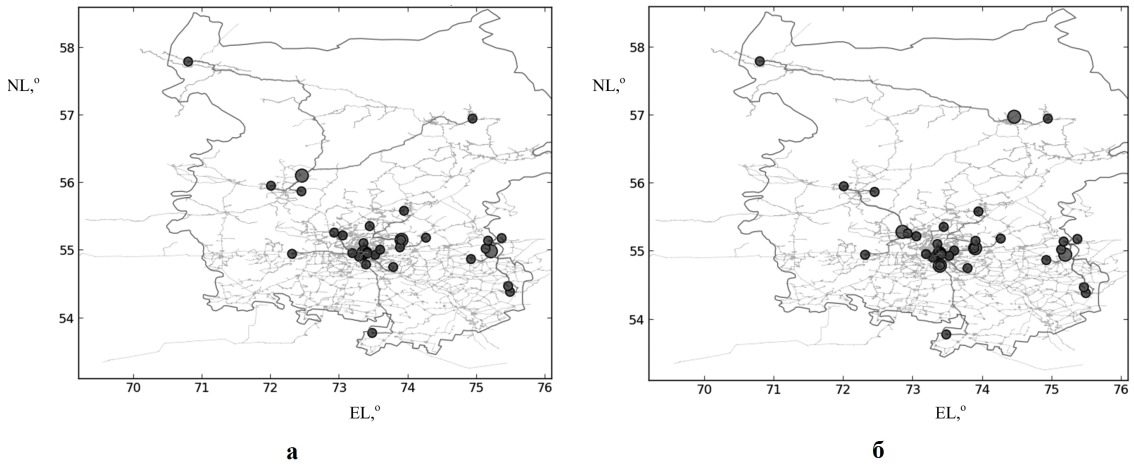

Рис. 6. Сравнение результатов кластеризации алгоритма K-means (а) и K-means++ (б) в условиях ограничений, налагаемых транспортной сетью, на карте Омской области. Центры кластеров обозначены кружками большего диаметра

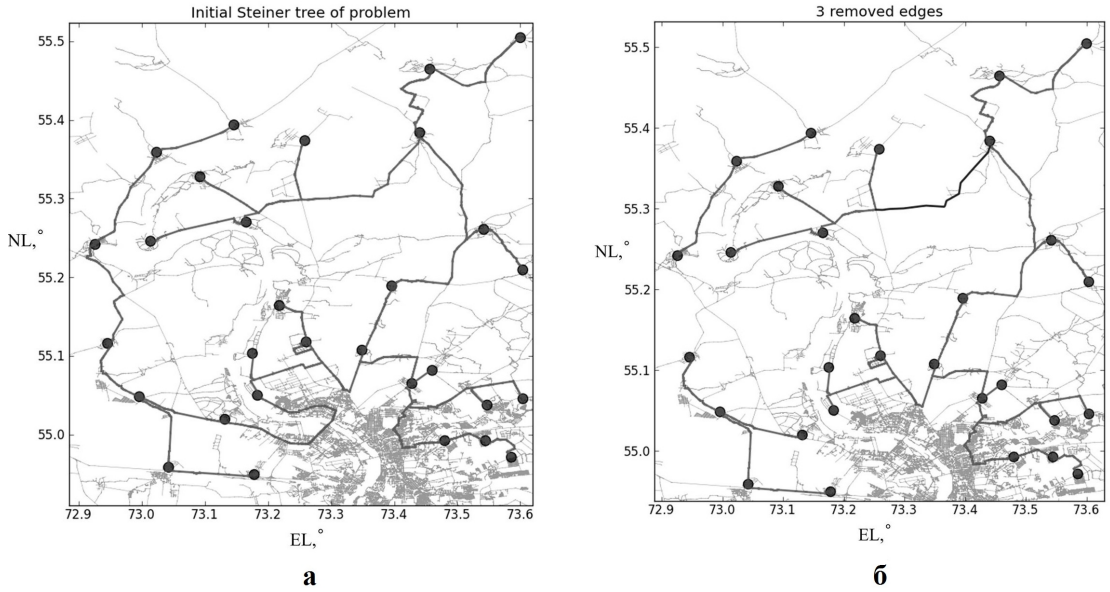

Рис. 7. Результаты работы алгоритма построения минимального дерева Штейнера (а) и процедуры усечения этого дерева для выделения трёх кластеров (б)

Построение дерева Штейнера на неориентированном графе с заданными n концевыми узлами начинается с построения сети расстояний (distance network) [5]. Положение кон-

цевых узлов задаётся пользователем на карте дорог. В соответствии с алгоритмом среди множества деревьев Штейнера для заданных концевых узлов находится минимальное дерево. Результаты алгоритма, реализующего построение минимального дерева Штейнера, представлены на рис. 7а.

Полученное дерево представляет собой минимальный маршрут, соединяющий все заданные узлы. Любой узел может быть выделен в качестве логистического центра, если решается задача выбора расположения единого центра в сети (single sink network design). Если ставится задача выбора нескольких логистических центров, то к полученному минимальному дереву Штейнера (рис. 7а) сначала применяется процедура его усечения. На рис. 7б показано усечение минимального дерева Штейнера с целью выделения трёх кластеров. Усечение производится последовательным удалением наиболее длинных рёбер полученного дерева.

Заключение

Предварительный анализ структуры транспортной сети, выполненный на матрице смежности её графа, показал, что число кластеров, выделенное в структуре матрицы смежности, может быть использовано как оценка предельного значения числа логистических центров.

Литература

1. Facility Location: Applications And Theory / Ed. by Drezner Z., Hamacher H.W. — Berlin-Heidelberg: Springer-Verlag, 2004.
2. *Awerbuch B., Azar Y.* Buy-at-bulk network design // In IEEE Symposium on Foundations of Computer Science (FOCS). — 1997. — P. 542–547.
3. *MacQueen J.* Some Methods for Classification and Analysis of Multivariate Observations // Proc. Fifth Berkeley Symp. Math. Statistics and Probability. — 1967. — P. 281–296.
4. *Arthur D., Vassilvitskii S.* K-means++: the advantages of careful seeding / SODA'07 Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. — CityPhiladelphia, StatePA: placecountry-regionSIAM Press. — 2007. — P. 1027–1035.
5. *Cheriyan J., Ravi R.* Lecture Notes on Approximation Algorithms for Network Problems — Canada: University of Waterloo, 1998.
6. *Ramm F., Topf J., Chilton S.* OpenStreetMap: Using and Enhancing the Free Map of the World // Cambridge, United Kingdom: UIT Cambridge Ltd. — 2010. — P. 386.
7. *Cooper L.* Location-allocation problems // Operations Research. — 1963. — V. 11. — P. 331–343.
8. *Michael G.K., Ajithkumar N.P.* Material Flow Analysis of Public Logistics Networks // Progress in Material Handling Research. — 2002. — P. 205–218.
9. *Chakrabarti D., Papadimitriou S., Modha D., Faloutsos C.* Fully automatic cross-associations // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2004. — P. 79–88.
10. *Rissanen J.* Information and Complexity in Statistical Modeling // Springer. — 2007. — P. 97–103.
11. *Dijkstra E.W.* A note on two problems in connection with graphs // Numerische Mathematik. — 1959. – I. 1. — P. 269–271.

Поступила в редакцию 14.01.2012.