УДК 519.25

А. Р. Файзлиев

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОПРЕДЕЛЕНИЯ ЧИСЛА ЛОКАЛЬНЫХ ЦЕНТРОВ НА ТЕРРИТОРИИ ГОРОДА

В статье предлагается метод кластеризации, учитывающий пространственное местоположение и интенсивность. Также предлагается новый критерий для выбора числа кластеров.

Цель работы: разработка метода выделения «сгущений» населения, магазинов и других объектов городской среды на основе комбинированного критерия пространственной близости и близости по плотности распределения объектов.

Задачи:

1. Подразделить территорию города Саратова на неперекрывающиеся пространственные ячейки.

2. Разбить множество таких ячеек $\{I_1, I_2, \ldots, I_n\}$ на непересекающиеся множества $\{C_1, C_2, \ldots, C_k\}$ (называемые кластерами), чтобы в пределах каждого кластера ячейки были как можно ближе в смысле:

а) евклидова расстояния между их центрами;

б) близости значений плотностей распределения объектов.

Такая постановка задачи предполагает собой задачу кластерного анализа [1].

Исходные данные:

Пространственная ячейка, заранее построенная с помощью алгоритма мозаики Воронового [2], характеризующаяся:

1. местоположением x_i, y_i;

2) площадью S_i;

3) средней интенсивностью P_i (плотностью населения, плотностью коммерческой недвижимости).

Основные параметры алгоритма кластеризации:

1. Форма меры «удаленности» для двух отдельно взятых ячеек.

2. Метод вычисления удаленности между двумя заданными кластерами.

3. Общая схема алгоритма (агломерация, подразделение и другие).

Мера удаленности между ячейками

$$d(K_i, K_j) = max(d_{ij}, c|lnp_i - lnp_j|),$$

где K_i, K_j – кластеры, состоящие из одной ячейки, d_{ij} – евклидово расстояние между центрами ячеек, $c > 0$ – постоянный коэффициент, позволяющий привести логарифмическую величину плотности к шкале евклидовых расстояний (рассчитывается с помощью вспомогательной линейной регрессии методом МНК-оценок.)

Методы вычисления удаленности между кластерами:

1) минимальное (single link);

2) максимальное (complete link);

3) среднее пар расстояний (pair-group average).

Общая схема алгоритма

Агломеративная, т.е., начиная от одноэлементных кластеров, определяются два самых близких, которые объединяются и т.д. Выбор числа кластеров:

1. По скачкам расстояния (классический метод);

2. По энтропии Шеннона (предложенный нами) [3]:

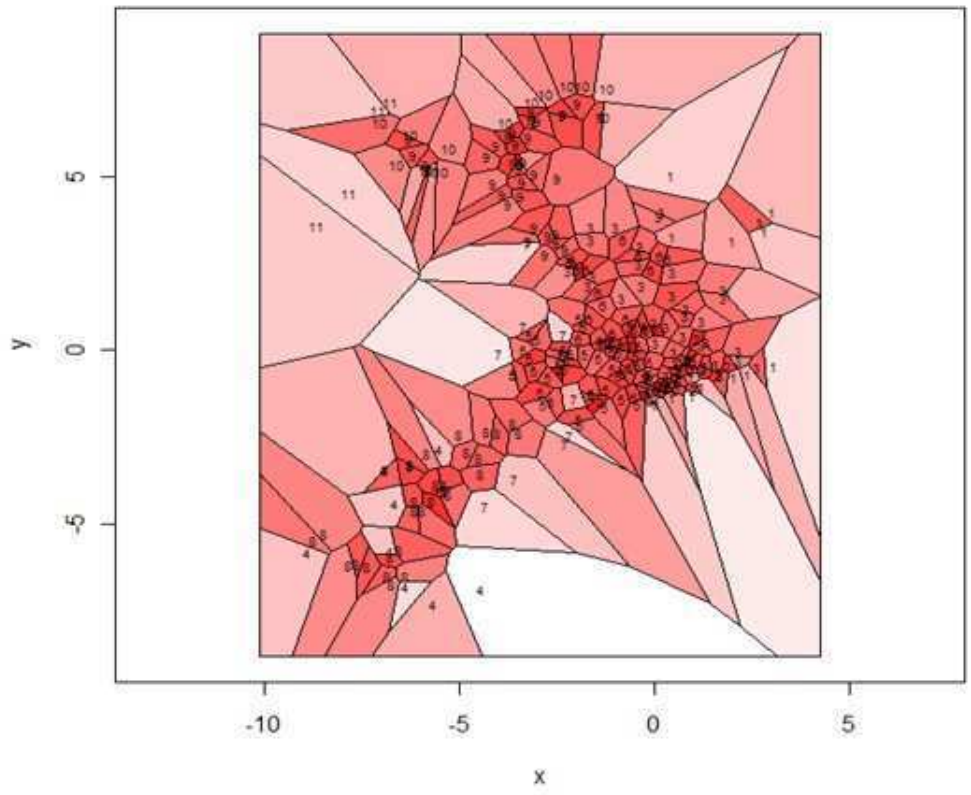
$$H(p) = -\sum_{i=1}^{n} p_i ln p_i, \text{ где } p_i \geq 0, \sum_{i=1}^{n} p_i = 1.$$

Для удобства мы рассматривали величину $\frac{e^{H(p)}}{n}$, так как она будет изменяться в пределах от 0 до 1.

В таблице представлены результаты кластеризации по населению (максимальное расстояние).

Скачок расстояния	Число кластеров	Энтропия	Число кластеров
1.8524755	4	0.9597912	3
1.7774426	2	0.8604969	2
1.2235018	6	0.7996000	4
1.0698915	3	0.7631253	11
1.0538180	11	0.7152539	19
0.8566795	7	0.7074681	12
0.4954933	12	0.7050943	5

Как видно из таблицы, по скачкам расстояния хорошо выделяются 4 и 2 кластера, которым соответствуют высокие показатели энтропии. Для большего числа кластеров хороший вариант с 11 кластерами, который подтверждается и скачком расстояния и энтропией. Рисунок иллюстрирует кластеризацию по населению (максимальное расстояние, 11 кластеров).



Выводы:

1. Предложен метод кластеризации, учитывающий пространственное местоположение и интенсивность.

2. Для выбора числа кластеров предлагается использовать не только скачки расстояния, но и новый критерий, основанный на энтропии Шеннона.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ким Дж.-О., Мюллер Ч. У., Клекка У. Р.* Факторный, дискриминантный и кластерный анализ / пер. с англ. И. С. Унюкова. М. : Финансы и статистика, 1989. 215 с.

2. *Препарата Ф., Шеймос М.* Вычислительная геометрия : Введение / пер. с англ. М. : Мир, 1989. С. 478.

3. *Волькенштейн М. В.* Энтропия и информация. М.: Наука, 1986. С. 192.