## § 1.5. Обзор программных продуктов для кластеризации многомерных данных.

Нами изучались интернет-материалы на тему «Кластерный анализ, программы кластерного анализа». При обилии ссылок на справочный материал и использование элементов кластерного анализа в статистических исследованиях в различных областях, где требуется обработка статистических данных, наблюдается дефицит программных продуктов, реализующих методы кластерного анализа. В большинстве ссылок упоминаются следующие пакеты программ:​ «ClusterDelta»,  «Сегментация рынка -​​​​​​​ позиционирование товара»,​ «Statistical Package for the Social Sciences»(SPSS Statistics)», MatLab.

В MathCad и Maple стандартных методов кластеризации не найдено, но в некоторых статьях предлагается реализовать их вручную.

​В пакете программ «ClusterDelta – кластерный график как инструмент анализа рынка» (http://tempofox.com/klasternyj-grafik-kak-instrument-analiza-rynka/) кластерный анализ используется для вычисления точечного объема заданных сделок в ограниченном диапазоне времени и финансов (ценовой бар). Кластер представляет собой «ценовой бар», отображаемый на графике, то есть бар, который разбит на ценовые уровни, на каждом из которых отображается проторгованный объем [26].

Пакет программ  «Сегментация рынка – ​​​​​​​ позиционирование товара» (<http://www.segmentation-market.ru/index.html>) позволяет производить:

1) выделение сегментов рынка методами экспертного сегментирования;

2) выделение целевых сегментов рынка методами кластерного анализа;

3) при оценке сходства потребителей учитывается важность отдельных переменных;

4) анализ перекрытия выделяемых сегментов рынка, сегментация рынка;

5) сегментация рынка - выделение и интерпретация семантики выделенных сегментов;

6) исследование выделенных сегментов с помощью функций полезности, главных факторов, профилей сегментов при сегментации рынка;

7) позиционирование товаров и бренда на выделенных сегментах, построение карты восприятия товара (карта позиционирования);

8) исследование стратегического потенциала целевых сегментов рынка;

9) представление результатов анализа доходных сегментов рынка в виде таблиц и графиков.

Используемый в этом программном пакете метод K-means – это итерационный метод кластеризации. Метод разбивает множество объектов на заданное аналитиком число кластеров (K). Объединение объектов в кластеры происходит на основе вычисляемой меры сходства. После первого назначения объектов кластерам (первой итерации) происходит повторное вычисление мер сходства и перенос объектов из одного кластера в другой. Итерационный процесс продолжается до тех пор, пока не будет получено наилучшее разбиение объектов на кластеры или достигнуто максимальное число итераций [27].

Программный пакет​ «StatisticalPackagefortheSocialSciences» (SPSSStatistics) http://www.datuapstrade.lv/rus/spss/section\_20/2/ – «статистический пакет для социальных наук» –  компьютерная программа для [статистической обработки](https://docviewer.yandex.ru/r.xml?sk=y7641a37c0c52a9f4a3be81098b2d23ee&url=http%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%25A1%25D1%2582%25D0%25B0%25D1%2582%25D0%25B8%25D1%2581%25D1%2582%25D0%25B8%25D1%2587%25D0%25B5%25D1%2581%25D0%25BA%25D0%25B8%25D0%25B9_%25D0%25B0%25D0%25BD%25D0%25B0%25D0%25BB%25D0%25B8%25D0%25B7%22+%5Co+%22%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9+%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7) данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в [социальных науках](https://docviewer.yandex.ru/r.xml?sk=y7641a37c0c52a9f4a3be81098b2d23ee&url=http%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%25A1%25D0%25BE%25D1%2586%25D0%25B8%25D0%25B0%25D0%25BB%25D1%258C%25D0%25BD%25D1%258B%25D0%25B5_%25D0%25BD%25D0%25B0%25D1%2583%25D0%25BA%25D0%25B8%22+%5Co+%22%D0%A1%D0%BE%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5+%D0%BD%D0%B0%D1%83%D0%BA%D0%B8) [7], [28].

Он использует 7 методов кластеризации: межгрупповые связи (Between-groups linkage); связь внутри групп (Within-groups linkage); близлежащий сосед (Nearest neighbor); дальний сосед (Furthest neighbor); центроидная кластеризация (Centroid clustering); медианная кластеризация (Median clustering); метод Варда (Ward's Method). Этот пакет не ориентирован на решение задач кластеризации в условиях неоднородных показателей, на обработку нечетких данных.

В Маtlab [14] реализуются методы четкой кластеризации: метод ближнего соседа, метод дальнего соседа, метод средней связи, центроидный метод, метод медианной связи. Возможна реализация нечеткого fcm-метода с использованием командной строки или графического интерфейса пользователя. Для нахождения центров кластеров в Маtlab имеется встроенная функция fcm: [*center*, *U*, *obj*\_*fcm*] = *fcm*(*data*, *cluster*\_*n*).

Аргументами данной функции являются: data – множество данных, подлежащих кластеризации, каждая строка описывает точку в многомерном пространстве характеристик; cluster\_n – количество кластеров (более одного).

Функцией возвращаются следующие параметры: center – матрица центров кластеров, каждая строка ко­торой содержит координаты центра отдельного кластера; U – результирующая матрица ФП; obj\_fcm – значение целевой функции на каждой итерации.

В «MatLab» [14] имеются встроенные функции, позволяющие реализовать алгоритмы четкой и нечеткой кластеризации. Этот пакет рассчитан на использование профессиональными математиками и программистами и в этом их ограничение.

Таблица со сравнительным анализом программных пакетов для кластеризации приведена в приложении 1.

## Выводы

Кластеризация – это разбиение множества объектов на некоторые однородные подмножества (кластеры), параметры которых изначально известны. Для кластера характерны внутренняя однородность (объекты одного класса схожи между собой по определенным признакам) и внешняя изолированность (объекты разных классов существенно отличаются).

Методы кластеризации делятся на иерархические и неиерархические. В иерархических – происходит последовательное объединение маленьких кластеров в большие или наоборот разделении больших кластеров на маленькие. Неиерархические методы основаны на разделении набора данных на определенное количество кластеров и выполнении итеративного процесса оптимизации некоторой целевой функции, определяющей оптимальность (обусловленную особенностями алгоритма) данного разбиения множества объектов на кластеры, на итеративный процесс накладывается условие остановки.

При решении задач структуризации сложных систем большинство формируемых классов объектов размыты по своей природе. Это приводит к необходимости использования «нечеткой кластеризации». Задачей нечеткой кластеризации является нахождение нечеткого разбиения или нечеткого покрытия множества элементов исследуемой совокупности, которые образуют структуру нечетких кластеров, присутствующих в рассматриваемых данных.

Кластеризация применяется в статистике, экономике, финансовой математике, в социологии, в информатике для «интеллектуального» анализа данных и др.

Изучение возможностей программных пакетов показало, что, с одной стороны, существуют узконаправленные пакеты программ для конкретных предметных областей, например, экономики: для сегментации рынка товаров, построения ценового бара и т.д. С другой стороны профессиональные программные пакеты с обширным набором кластерных методов, но без привязки к терминологии предметных областей. Эти программные пакеты не ориентированы на решение задач кластеризации в условиях неоднородных показателей, на обработку нечетких данных. Рассчитаны на использование профессиональными математиками и в этом их ограничение.