Записка Ласкина МБ в СПб НМСО 24 августа 2022г.

Слайд 1.

Можно ли построить кластеры для наших задач. Технические проблемы, достоинства, недостатки.

Слайд 2.

Несколько предварительных ремарок, по поводу прочитанного и услышанного.

1. Кластерный анализ – обширный набор алгоритмов, позволяющих разбивать множества объектов на подмножества (кластеры), обладающие признаками сходства, попадающих в подмножества (кластеры) объектов.

Кластерный анализ относится к методам машинного обучения, называемых обучением без учителя.

Корреляционная кластеризация – это один из методов кластеризации, при котором в качестве метрики, определяющей «сходство» элементов множества выбирается корреляция между координатами различных элементов исходного множества. В наших задачах «в лоб» работать не будет.

2. У нас имеется набор, включающий в себя цену и все ЦОФы

$$V, X_1, X_2, \dots, X_n$$

Понятие ценовой кластер интуитивно понятно, но не очень корректно – при разбиении на кластеры мы определяем кластеры «сходных» объектов по координатам $X_1, X_2, \dots, X_n$.

Цена – это уже свойство кластера.

3. Мы хотим построить кластеры так, чтобы в одном кластере оказывались объекты в некотором смысле «сходные» (по ЦОФ), да еще и так, чтобы цена для каждого элемента кластера не слишком «плавала» внутри кластера.

Цену можно включить в какой либо алгоритм кластеризации, но тогда надо понимать, что в кластере могут оказаться объекты заметно разной стоимости, т.к. задача многомерная. Все зависит от метода кластеризации и выбора метрик.

Слайд 3.

ПРОБЛЕМЫ:

- выбор метрики, определяющей сходство элементов исходного множества, позволяющей построить кластеры. В нашем случае – задача не тривиальная, т.к. ЦОФов много, часть из них непрерывные (вещественные факторы), часть ранговые, часть бинарные.

Как установить меру близости между векторами $X_1, X_2, \dots, X_n$ $Y_1, Y_2, \dots, Y_n$

или между такими векторами $V, X_1, X_2, \dots, X_n$ $W, Y_1, Y_2, \dots, Y_n$?

- визуализация в многомерным случае (когда координат – в нашем случае это ЦОФы - более трех).

ДОСТОИНСТВО кластерного анализа:

- любое множество можно разбить на какие либо кластеры

НЕДОСТАТОК:

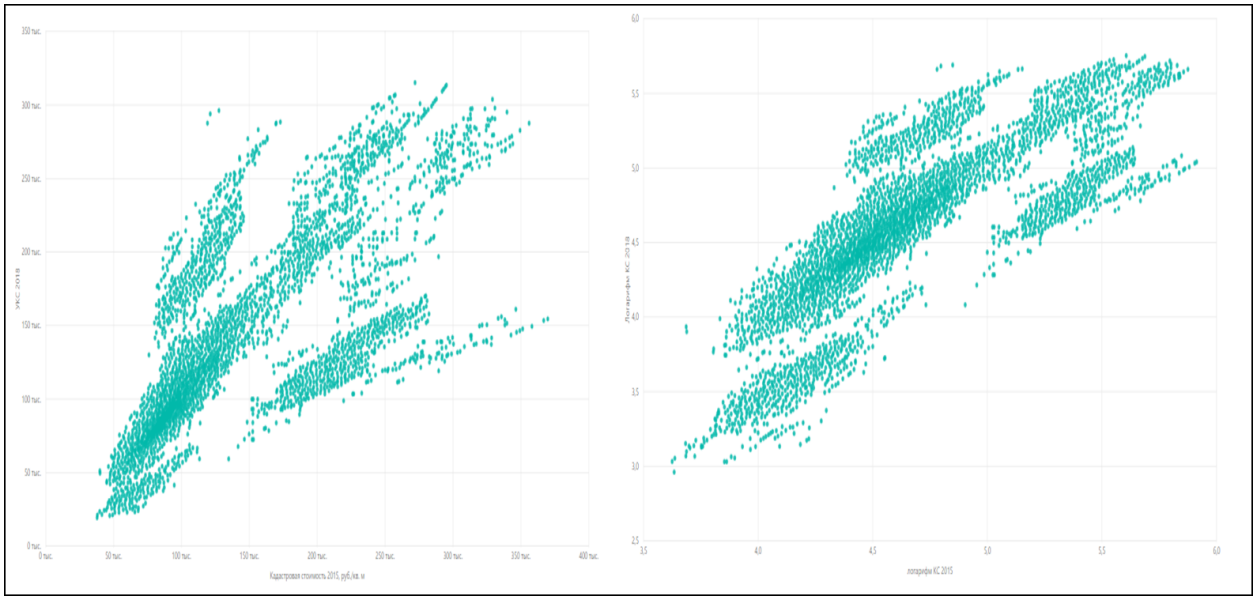
- в результате работы алгоритмов кластерного анализа НЕ остается никаких записей, позволяющих описать кластер в терминах предметной области. Единственный выход – просматривать каждый кластер и искать в нем объяснение «общности» попавших в него объектов с точки зрения предметной области.

Слайд 4

Пример, который уже рассматривался ранее.

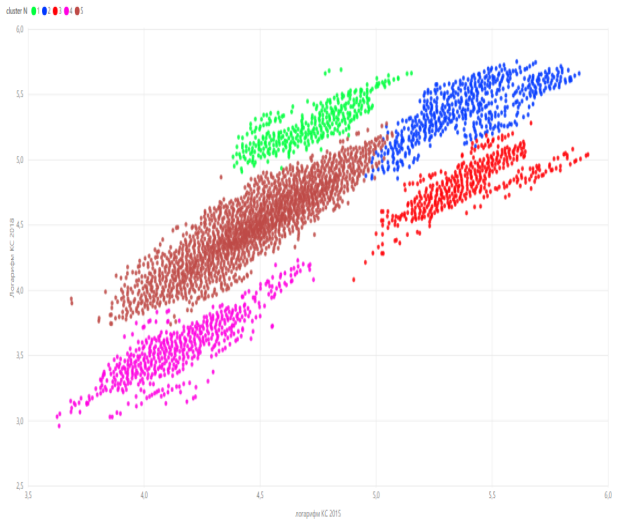
Кадастровая база в разных периодах. Две цены позволяют создать визуальный образ.

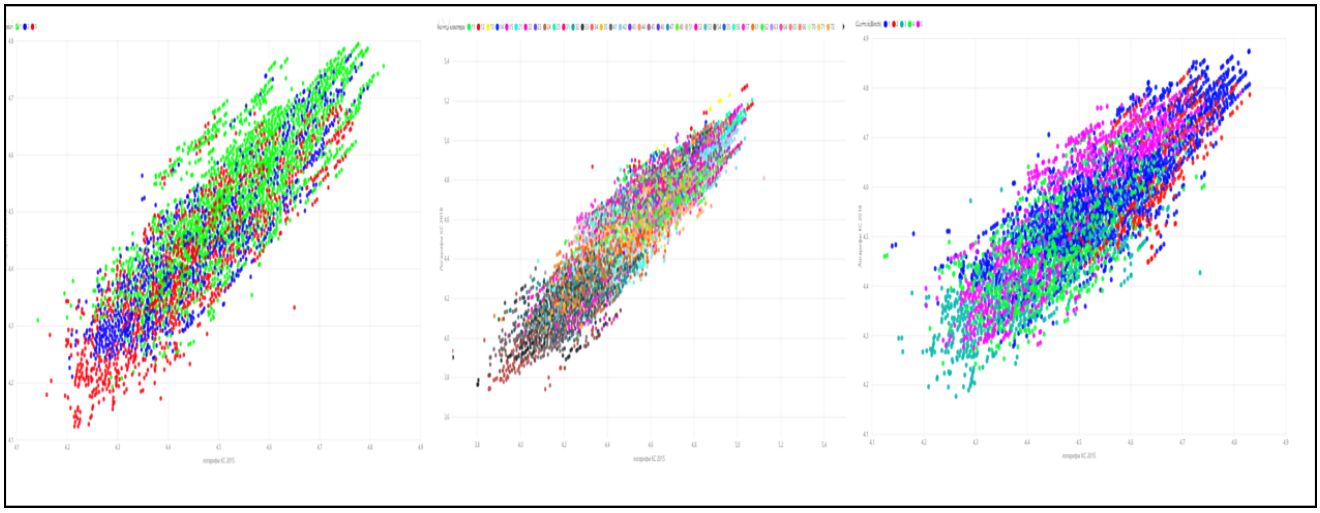
Жилая недвижимость Санкт-Петербурга.



Его не удалось разделить на кластеры большинством библиотечных алгоритмов кластерного анализа: K-means, K-medoids, CLARA (Clustering Large Application), CURE (Clustering Using Representatives), DBSCAN (Density-Based Spatial Clustering Application with Noise), OPTICS, STING (Statistical Information Grid Approach), Wave Cluster, Fuzzy C-means.

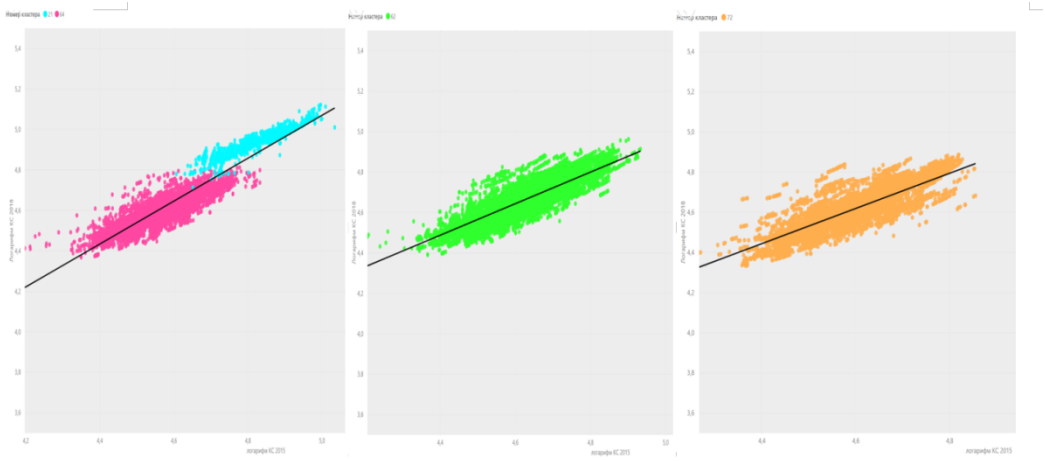
Он развалился на кластеры только применением комбинации методов BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)и WARD.





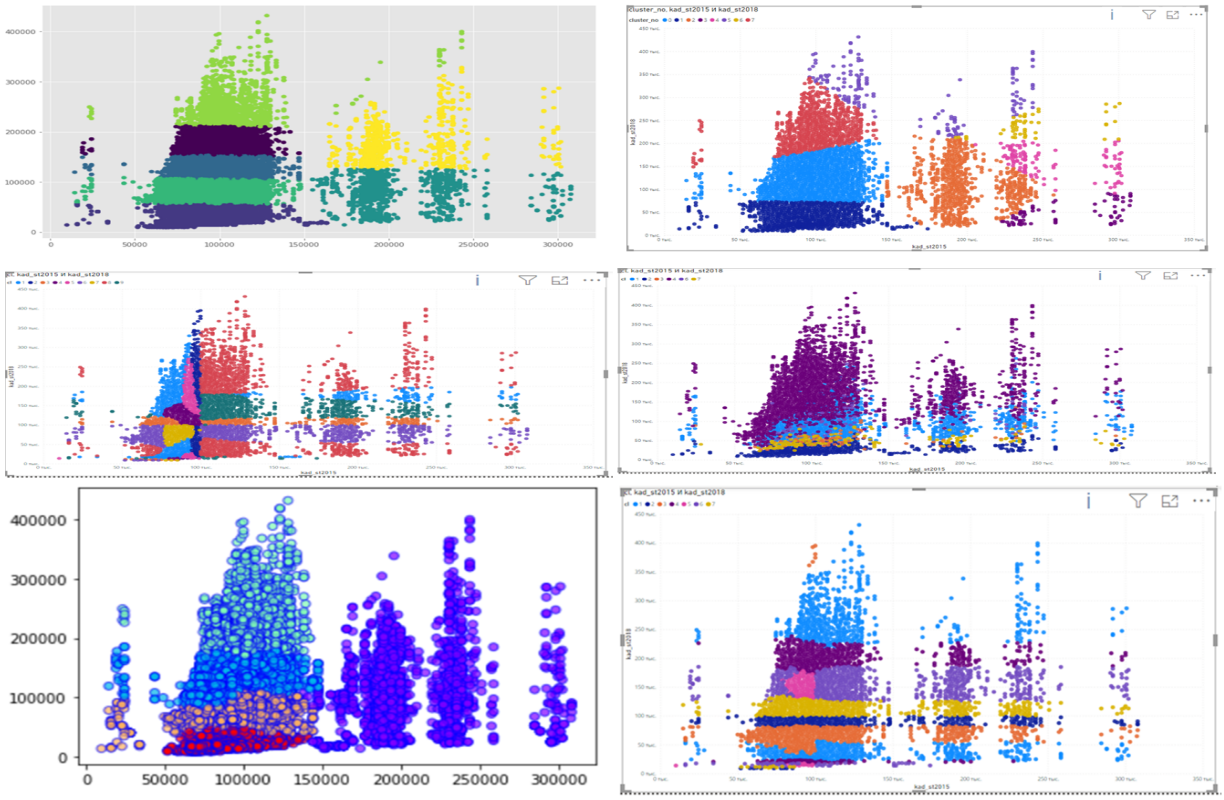
Кластеры получились хорошо интерпретируемыми, но все равно содержащими некоторую долю «примесей» - объектов имеющими с точки зрения формальной метрики, достаточные основания для отнесения объекта к определенному кластеру, но не объяснимыми с точки зрения предметной области.

Что получим в итоге? Кластеры приблизительно «одинаково» развивающихся во времени объектов.

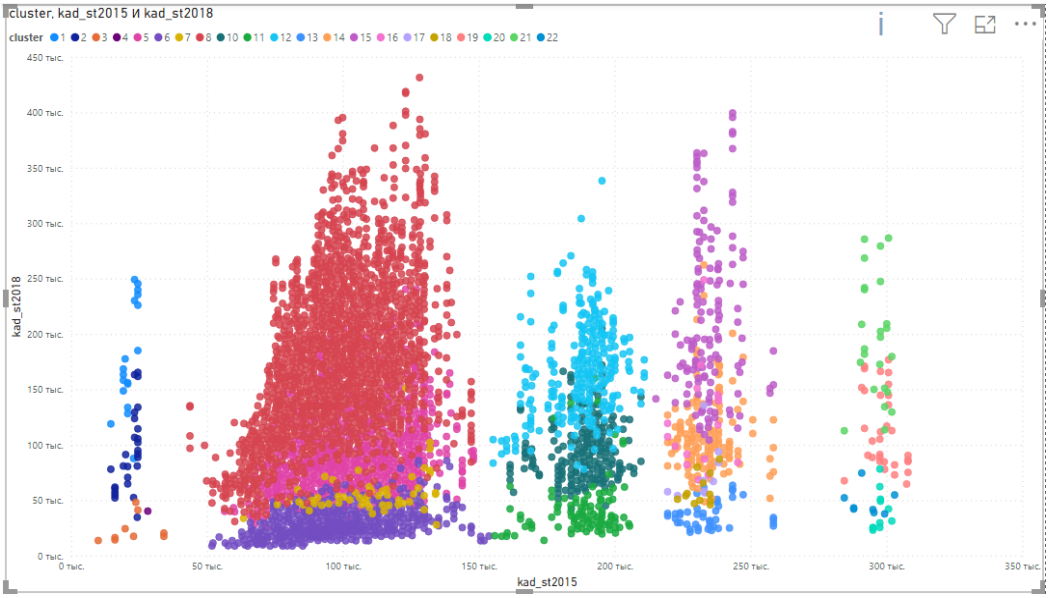


Слайд 5.

Пример некорректного разбиения на кластеры методами «K-means», «MeanShif», EM-алгоритм и даже BIRCH без предварительной подготовки по качественным и количественным переменным.



И корректное разбиение комбинацией методов BIRCH и WARD с предварительной подготовкой по качественным и количественным переменным.



Слайд 6.

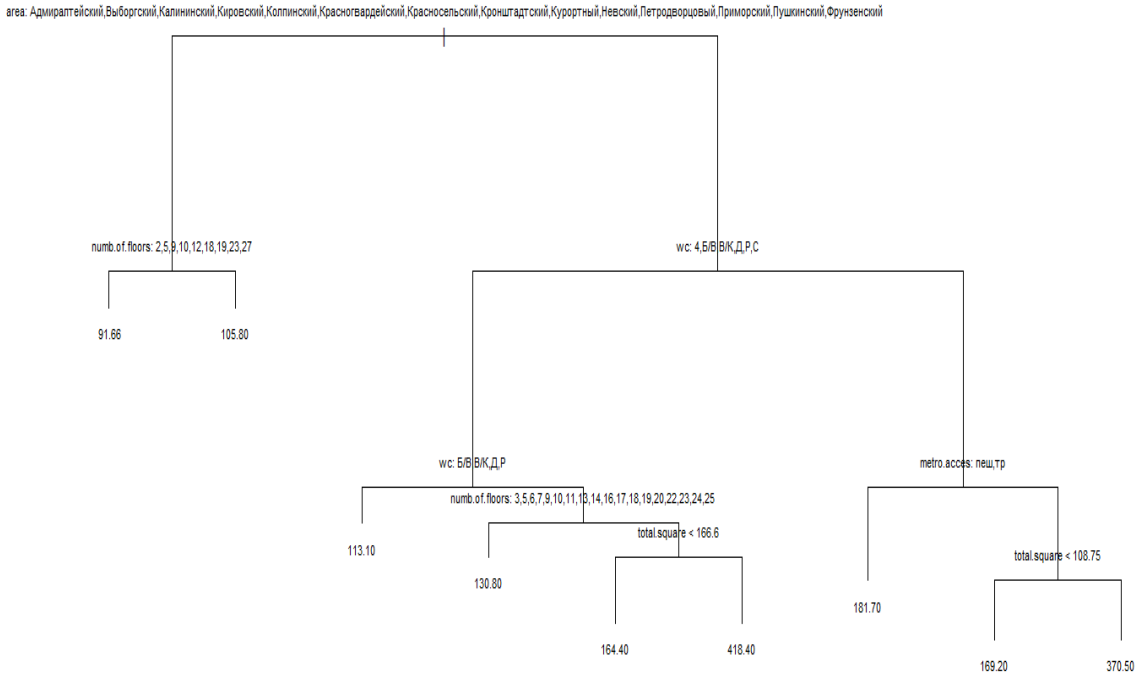
Еще один класс алгоритмов, позволяющих разбить наше множество на «ценовые» подмножества:

- решающее дерево.

Недостаток – недостаточная точность предсказания, т.к. всем элемента подмножества присваивается значение, равное среднему арифметическому всех его элементов.

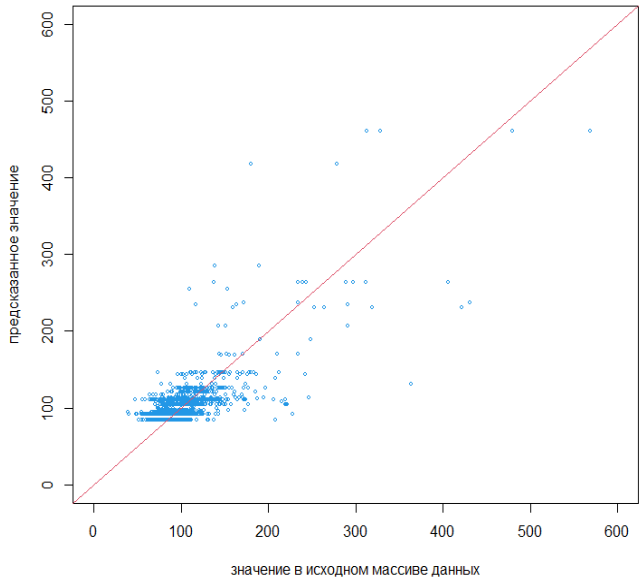
Однако, построив достаточно глубокое разбиение, можно из этого алгоритма взять только нужную нам часть – разбиение на подмножества. Останется только исследовать насколько велико стандартное отклонение на каждом множестве.

Пример.



НЕДОСТАТОК решающего дерева:

- как правило слишком большой разброс в конечных множествах (кластерах)



ДОСТОИНСТВО решающего дерева:

- в библиотечных алгоритмах сохраняется правила формирования дерева (т.е. конечных множеств), любой новый объект легко может быть интерпретирован и отнесен к нужному подмножеству (чего нет у алгоритмов кластерного анализа).

Имеется класс алгоритмов, основанных на решающих деревьях: случайный лес, бэггинг, бустинг. Но это – предсказательные алгоритмы, часто дающие неплохую точность в наших задачах.

Их недостаток – отсутствие визуализации и итогового разбиения на подмножества:
у каждого дерева – свое разбиение.

Поэтому использовать их для разбиения наших множеств на кластеры не удастся.

Слайд 7.

Вывод:

Задача своевременная, важная, интересная и разрешимая.

Для её реализации потребуются:

- важные административные решения,
- данные кадастровых баз, т.к. они содержат все зарегистрированные на дату оценки объекты
- подготовленные программисты.