3-я Международная конференция
«Субмиллиметровая и миллиметровая
астрономия: цели и инструменты»"

Поиск областей звездообразования в Галактике с помощью методов машинного обучения

Плакитина К.В.[1], Кирсанова М.С.[1], Островский А.Б.[3], Салий С.В.[2], Гималиева А.Д.[3]

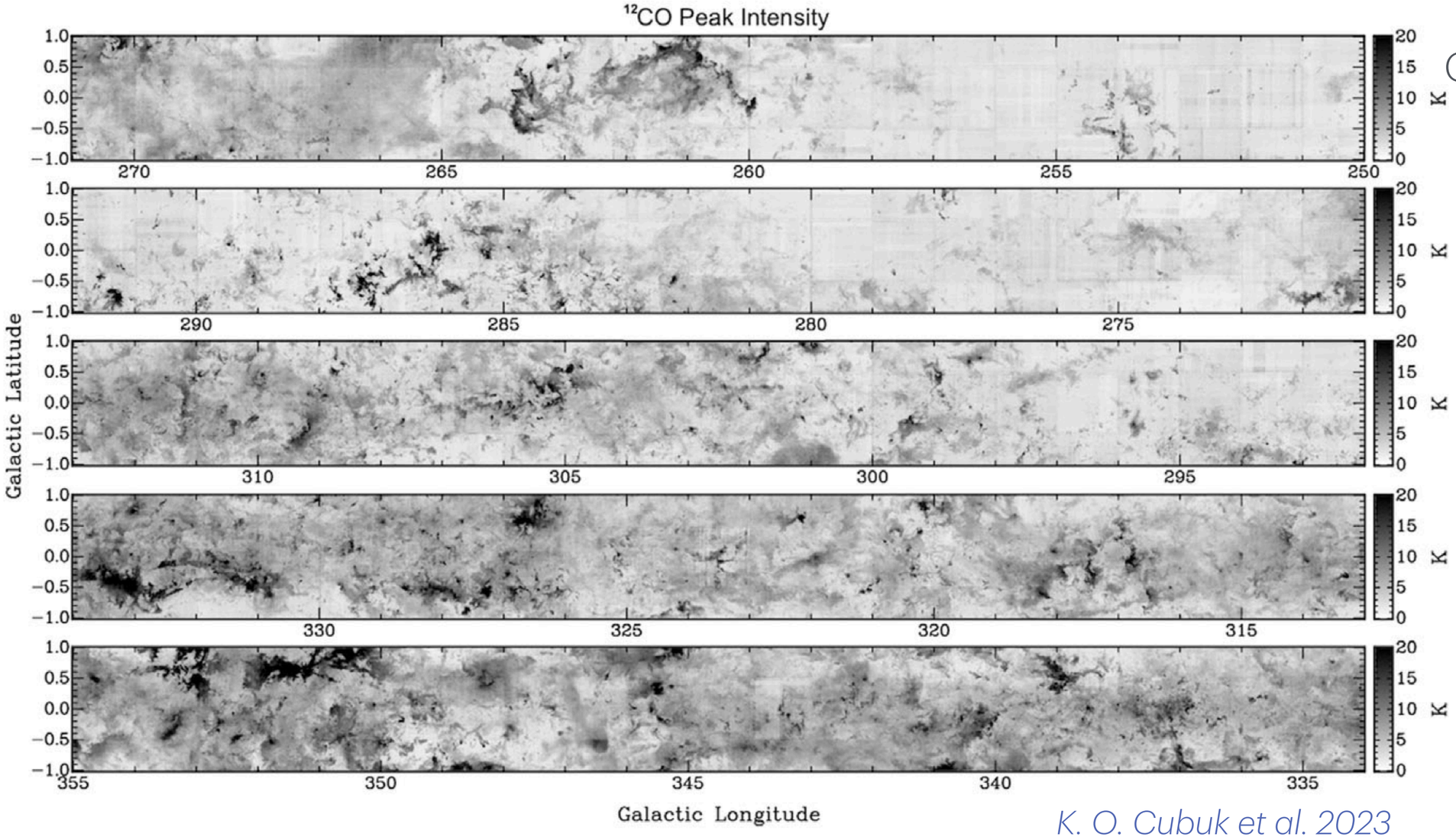
[1]Институт астрономии РАН

[2]Коуровская астрономическая обсерватория

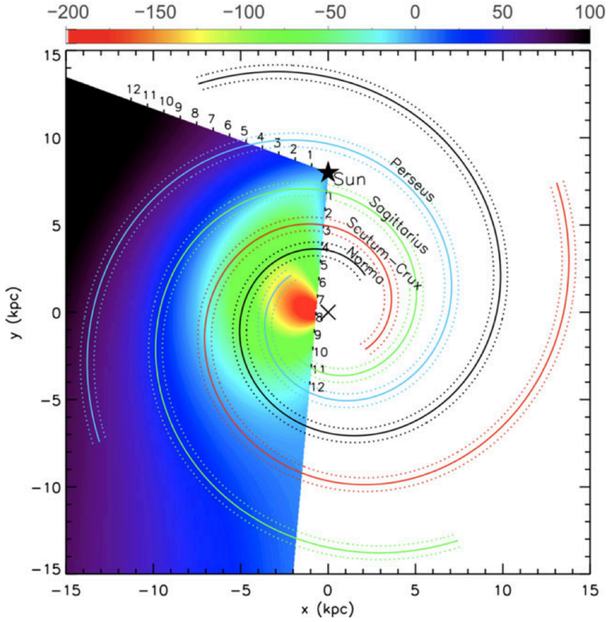
[3]Уральский федеральный университет

Работа выполнена при поддержке гранта РНФ 24-22-00097

Актуальность и цель работы

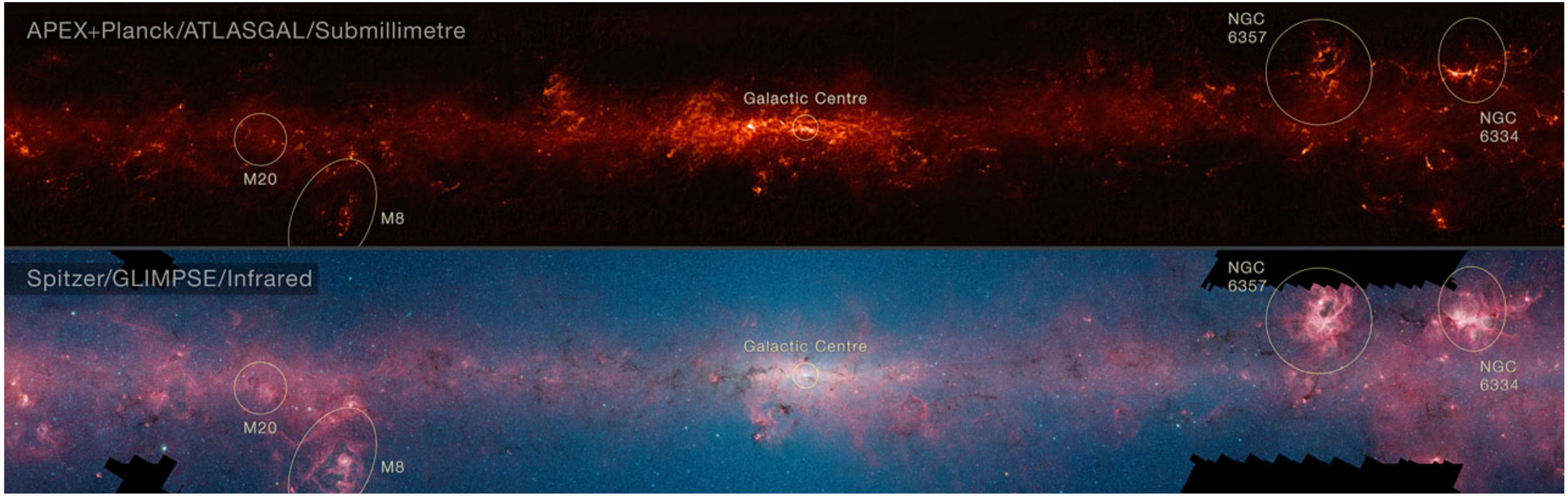


The Mopra Southern Galactic Plane CO Survey



K. O. Cubuk et al. 2023

Актуальность и цель работы



https://www.eso.org/public/images/eso1606e/

ЦЕЛЬ: Показать, как методы машинного обучения помогают в анализе и идентификации областей звездообразования

Каталог MALT90

The Millimetre Astronomy Legacy Team 90 GHz

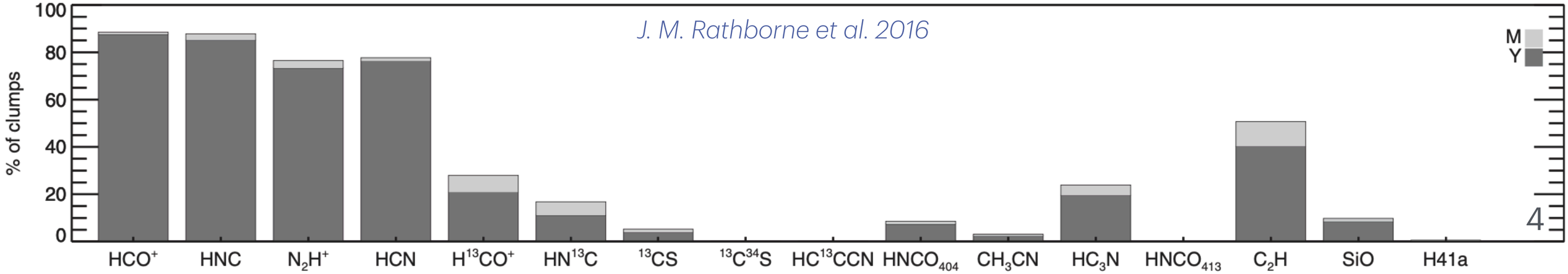
Цель — изучение физической и химической эволюций областей звездообразования методом картирования плотных сгустков из обзора ATLASGAL 870мкм

Апертура Mopra — 22 м

Угловое разрешение — 38″

Размер карт — 3′ x 3′

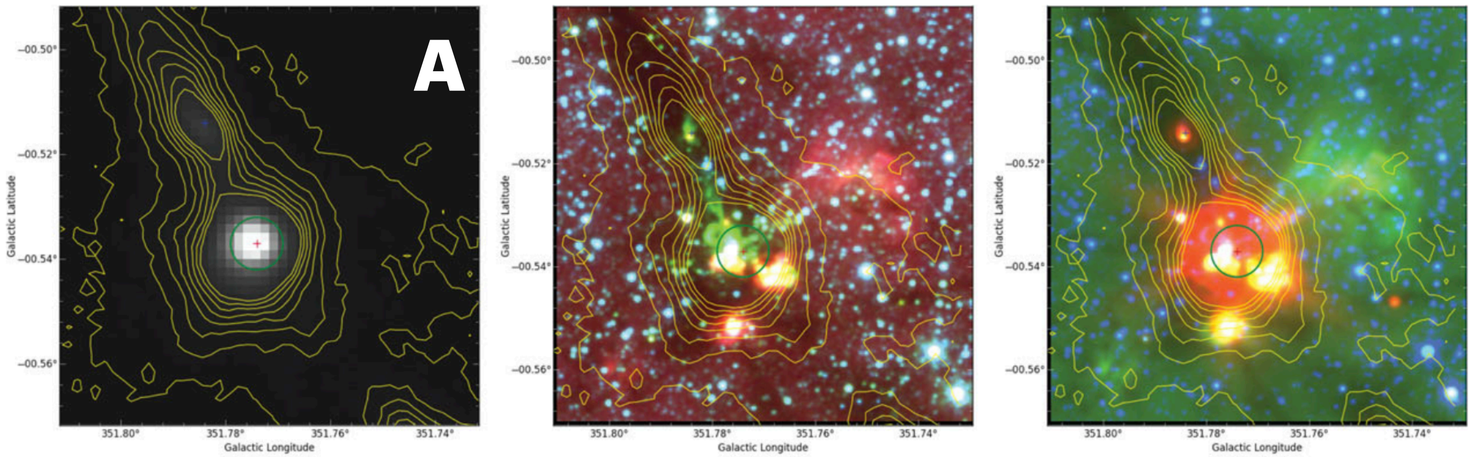
Спектральное разрешение — 0.1 км/с



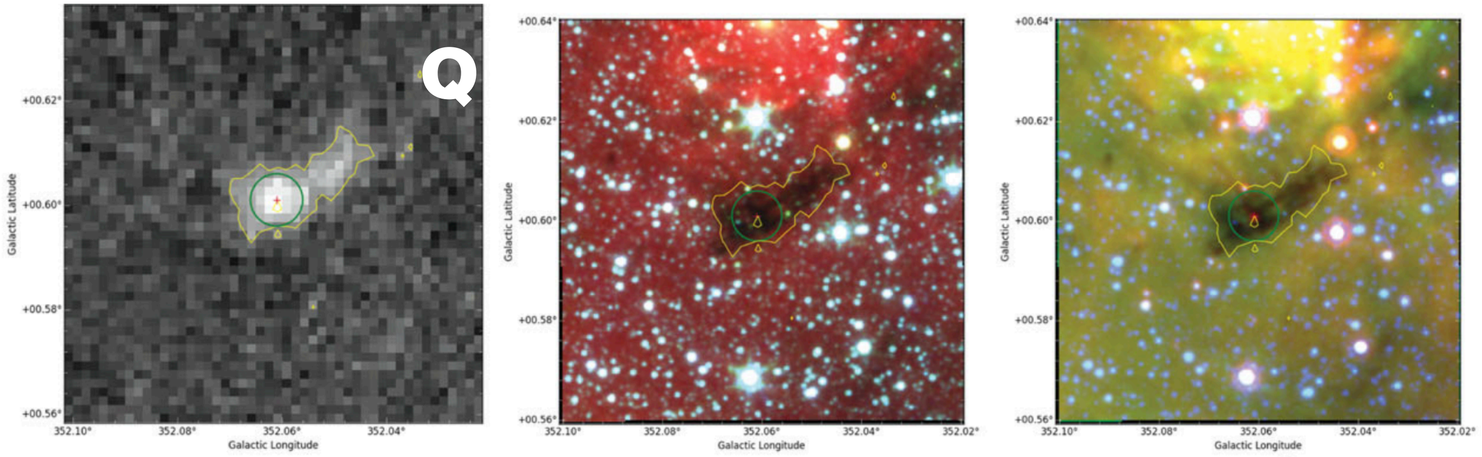
Каталог MALT90

Spitzer

ATLASGAL 870мкм | 3.6 мкм 4.5 мкм 8 мкм | 3.6 мкм 8 мкм 24мкм



AGAL351.774−00.537_S : V_c= −2.1 km s^{-1}



AGAL352.061+00.601_S : V_c= 1.3 km s^{-1}

Всего картировано 3 246 сгустков

- **A** — протозвездные области (753)
  - Протяженное излучение на 4.5 мкм
  - Точечные источники на 24 мкм
- **H** — протяженные области HII (688)
- **U** — неопределенные (673)
- **Q** — темные ИК области (616)
- **P** — области ФДО (345)
- **C** — компактные области HII (171)

Методы машинного обучения

Предобработка данных

Анализ исходных данных → **Масштабирование данных** → **Уменьшение размерности**

- Удаление признаков с большим количеством пропусков
- Удаление нерелевантных и избыточных признаков

Для чего:

- Обеспечение эффективности и надежности результатов кластеризации

- Приведение признаков к единому масштабу

Для чего:

- Признаки с большими значениями могут доминировать, искажая результаты кластеризации
- Масштабирование обеспечивает равный вклад всех признаков в процесс кластеризации

- Создание линейной или нелинейной комбинации признаков

Для чего:

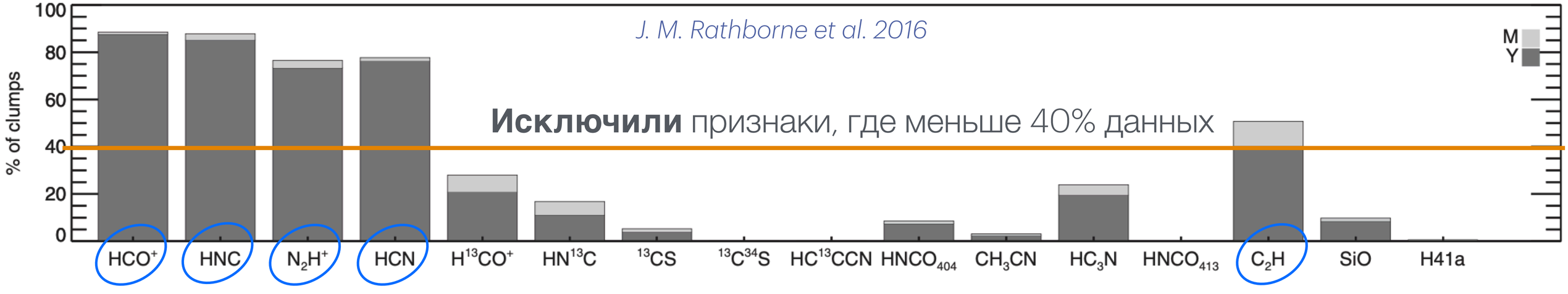
- Облегчает визуализацию и интерпретацию кластеров
- Повышение качества кластеризации

Применение методов машинного обучения на данных из каталога MALT90

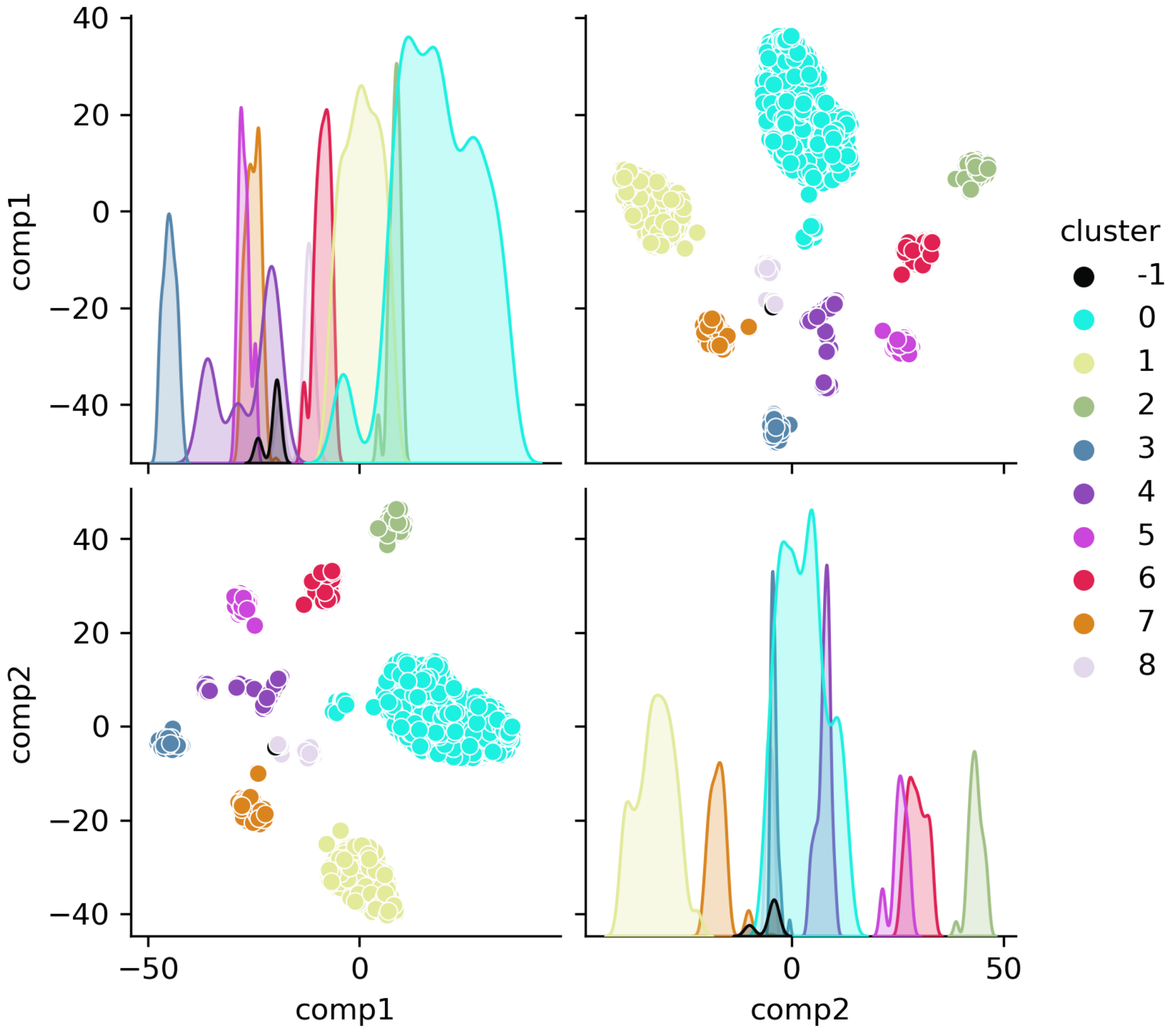
Предобработка данных

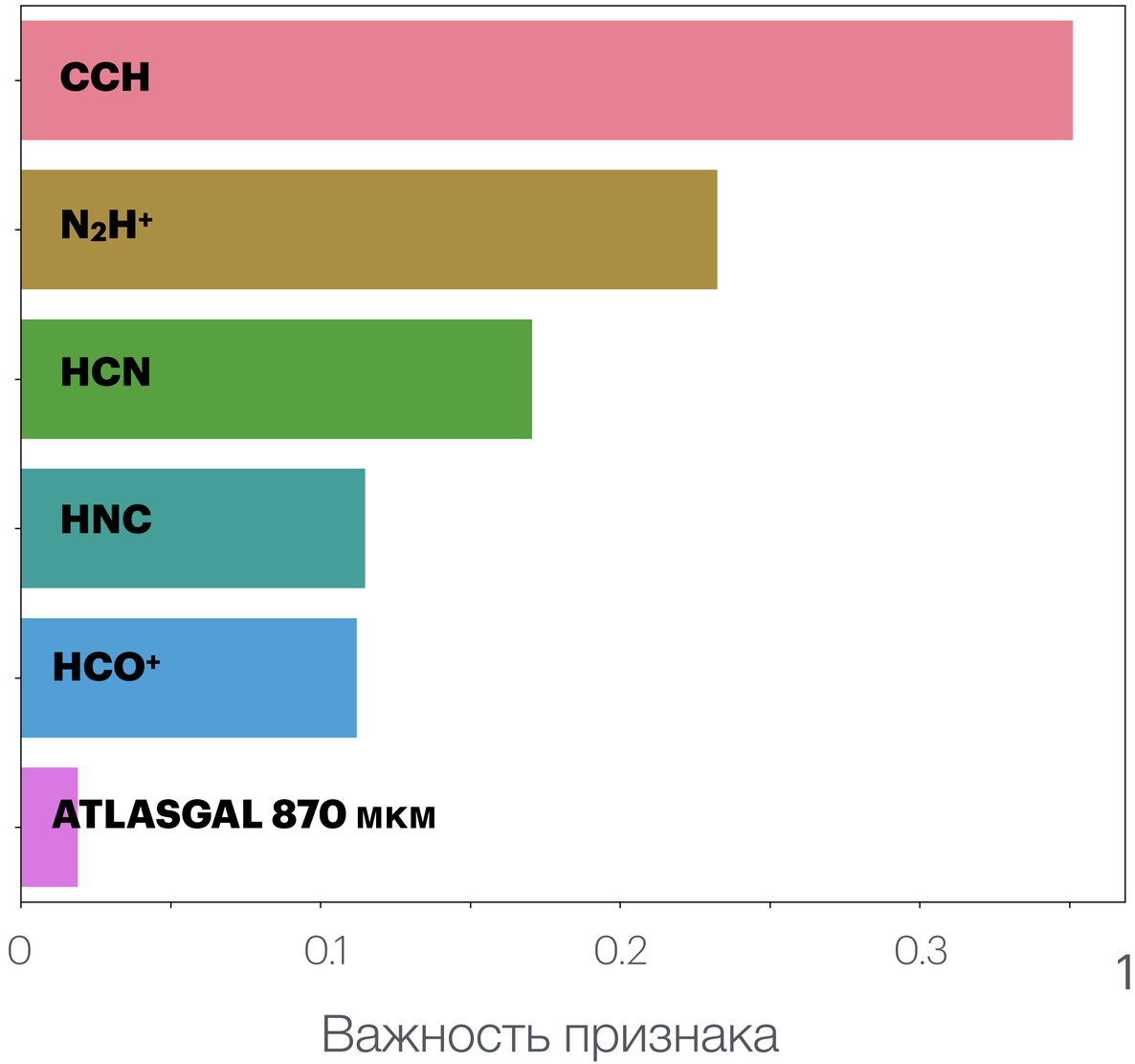
Исключили нерелевантные и избыточные признаки:

- Названия объектов
- Координаты
- Маркеры обнаружения линий
- Маркеры классификаций объектов
- И др.

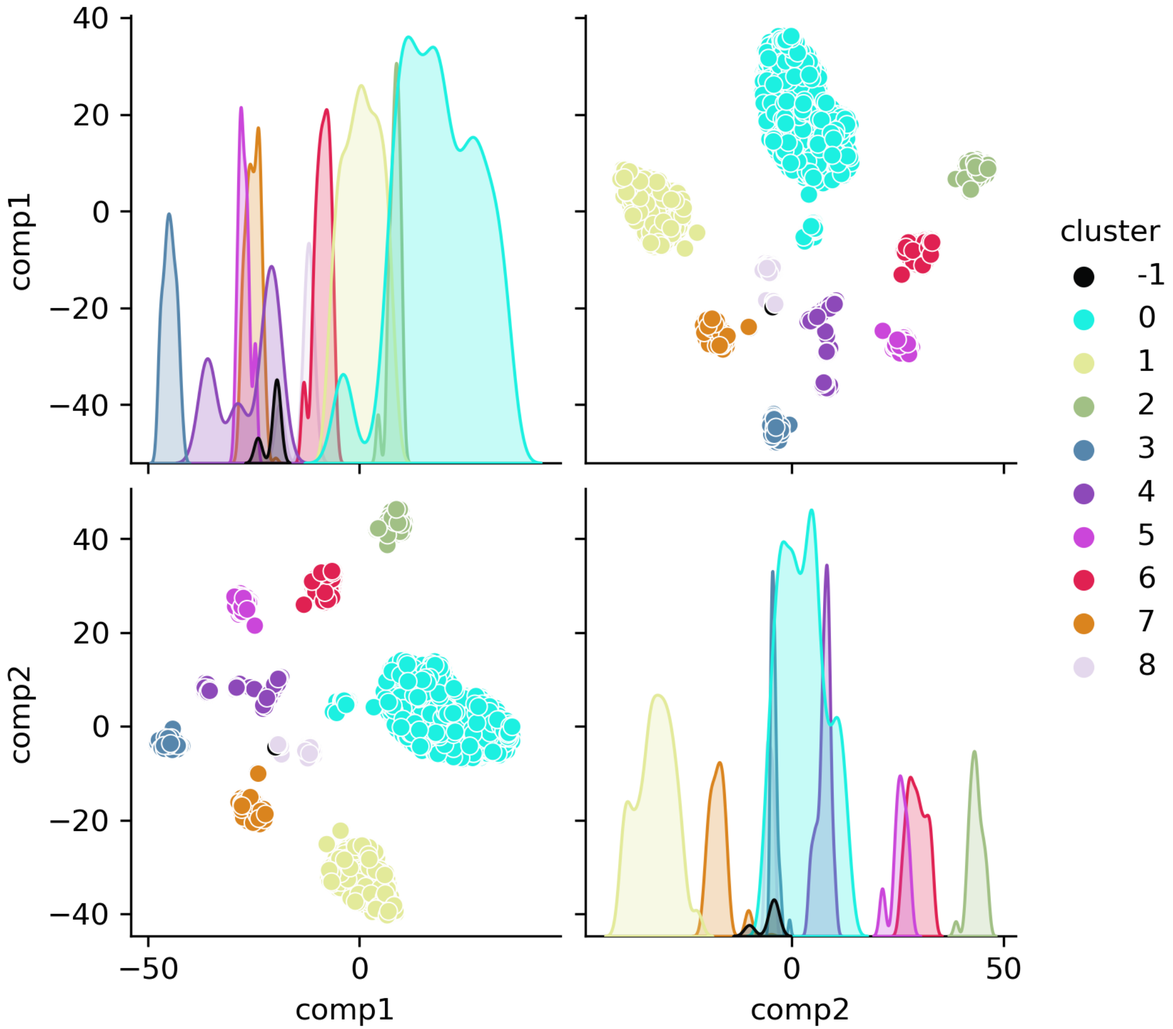


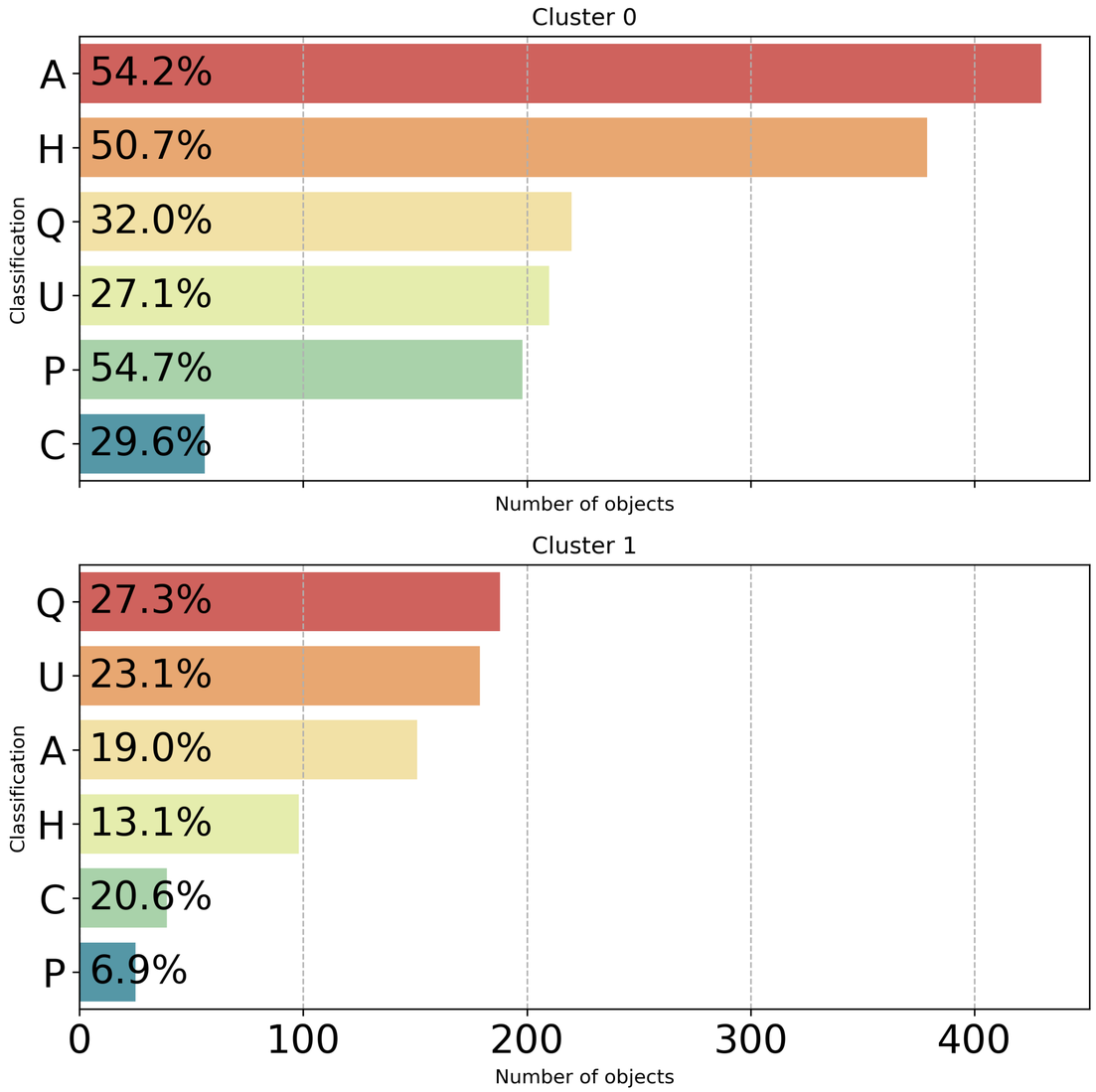
Результаты кластеризации MALT90





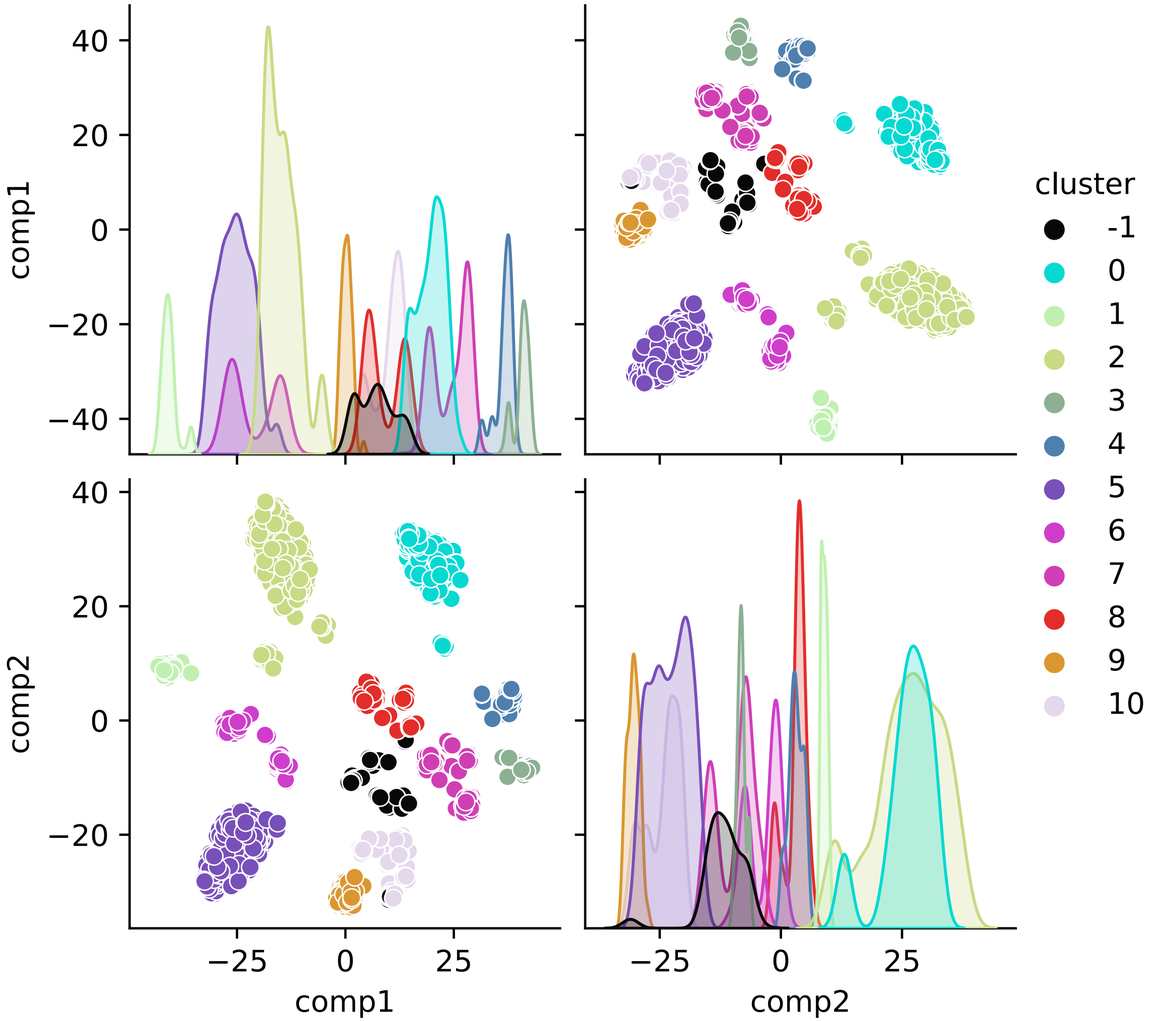
Результаты кластеризации MALT90

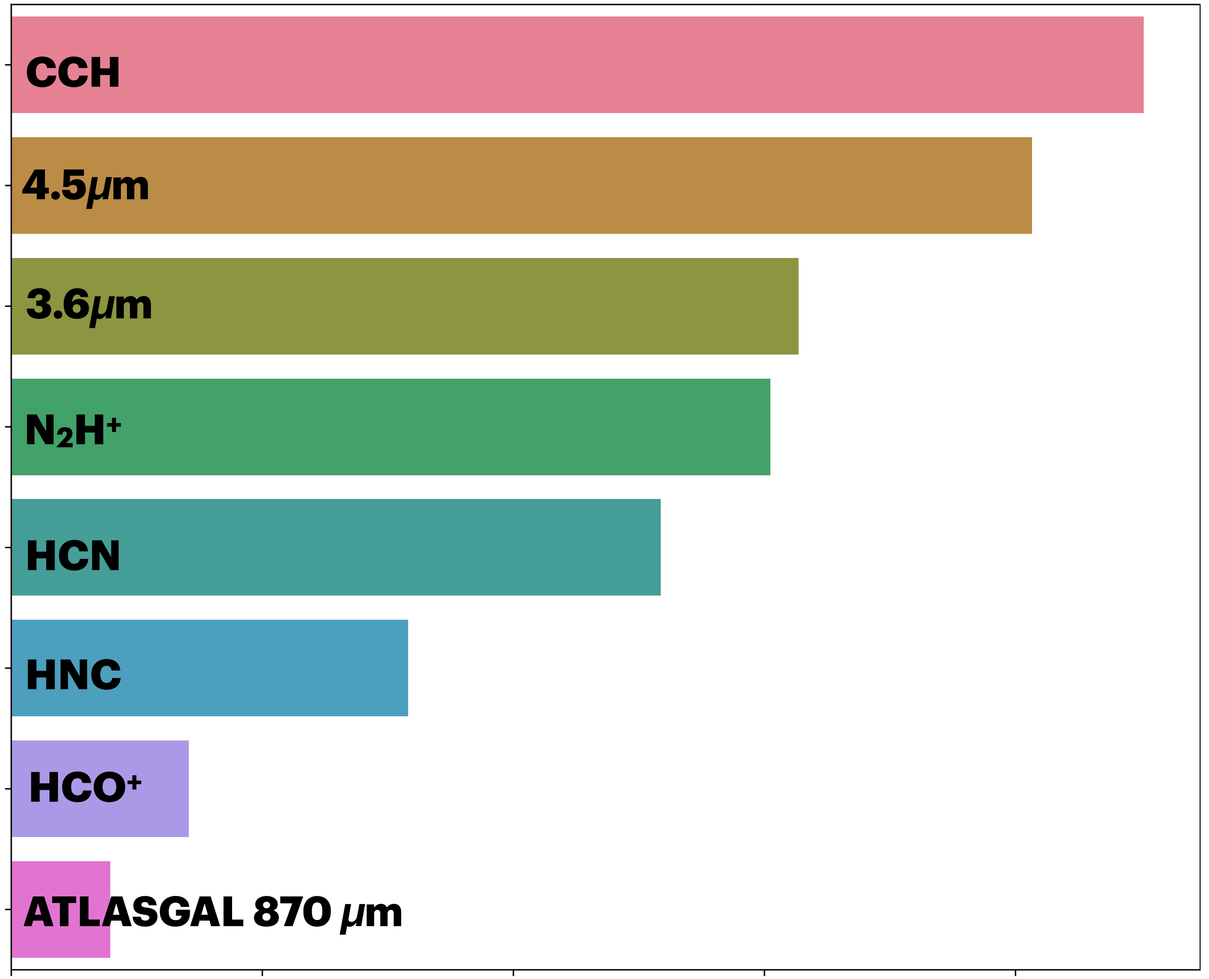




Результаты кластеризации

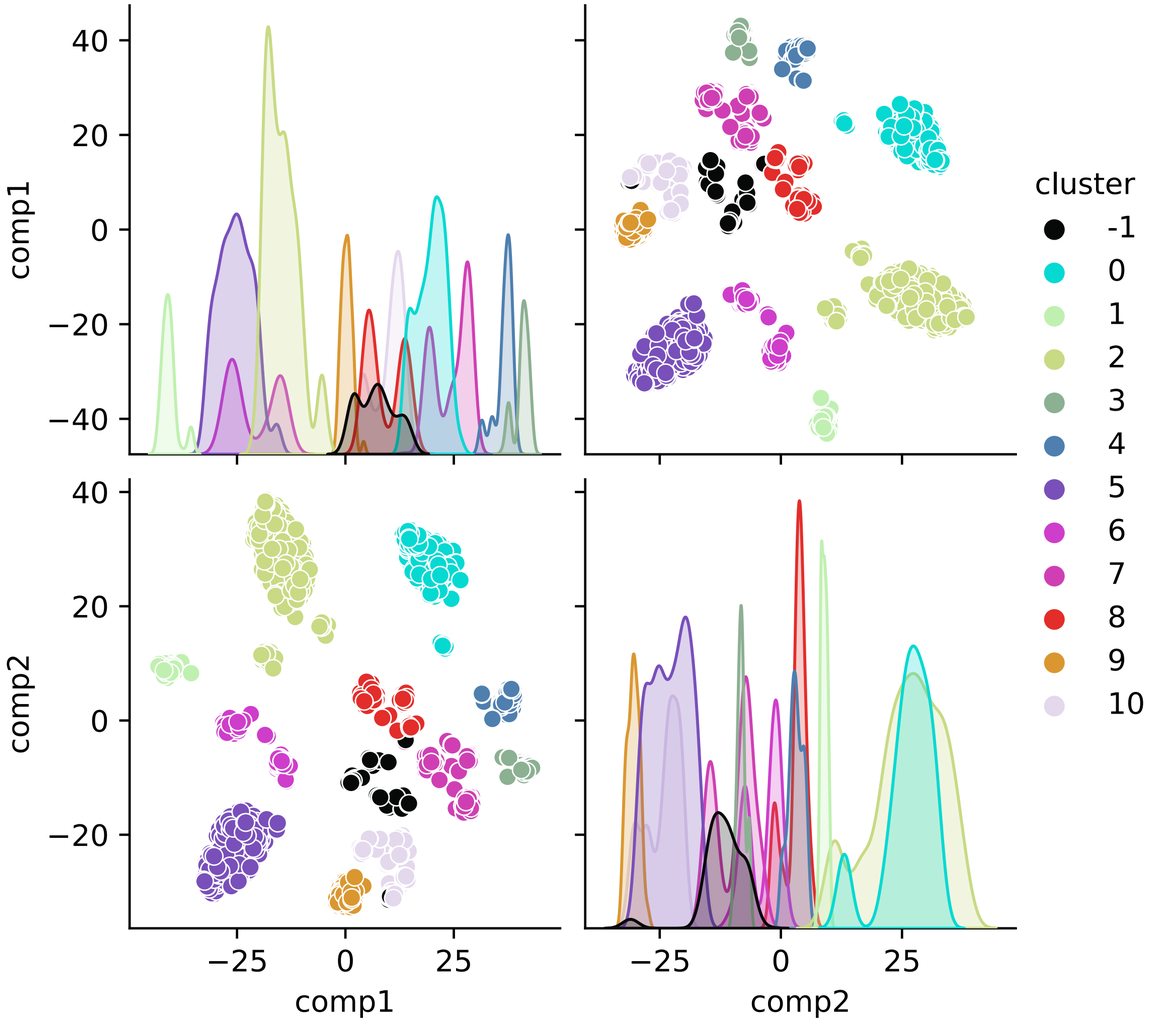
MALT90+Spitzer

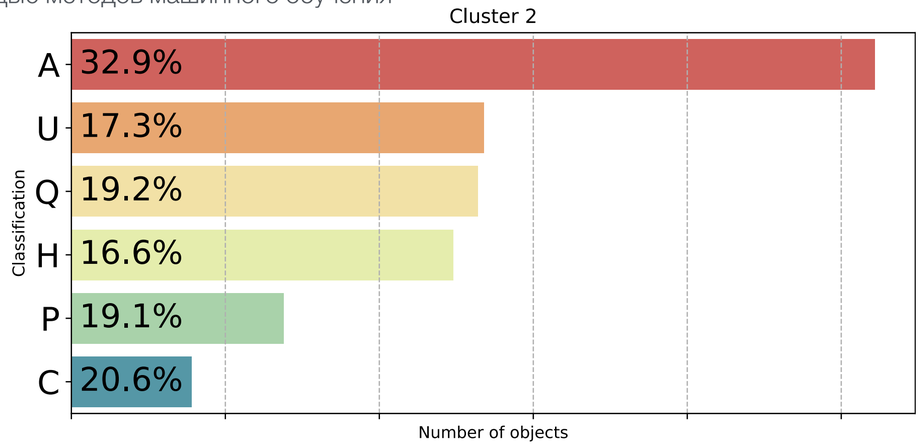


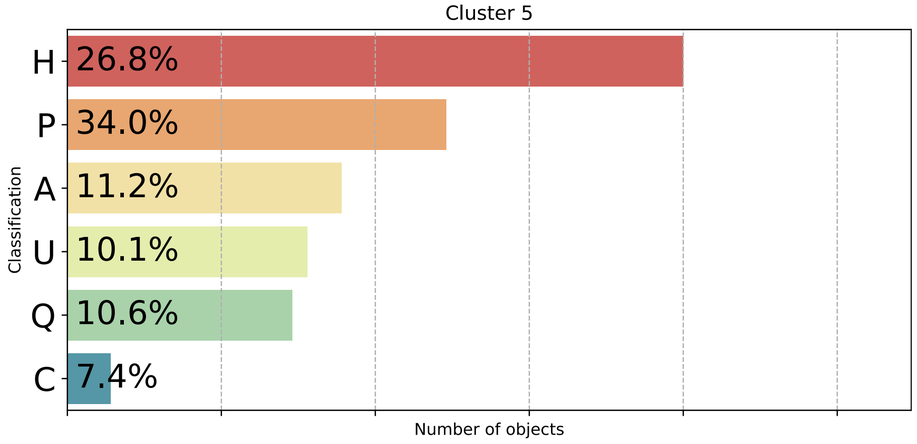


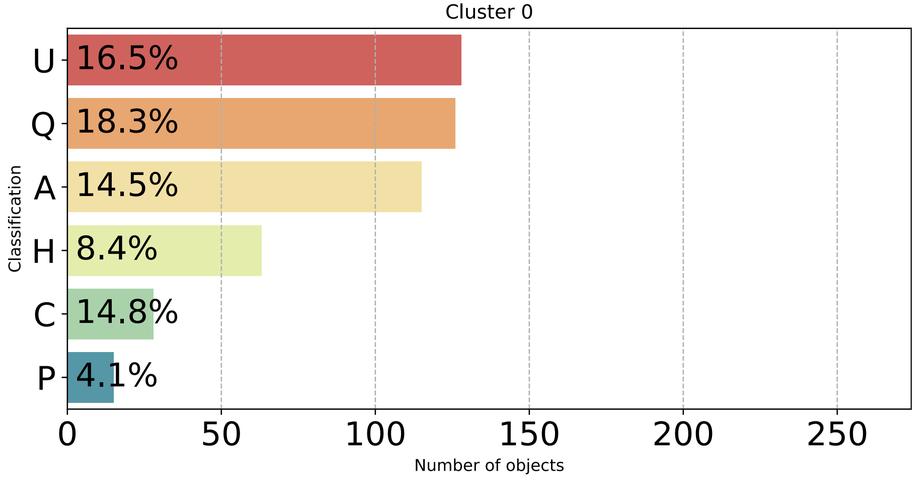
Результаты кластеризации

MALT90+Spitzer



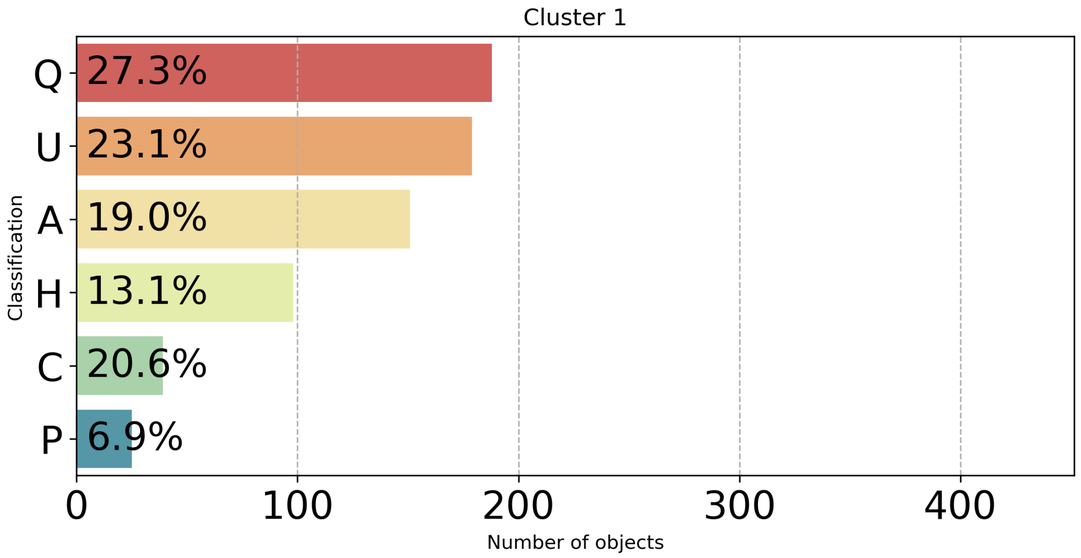




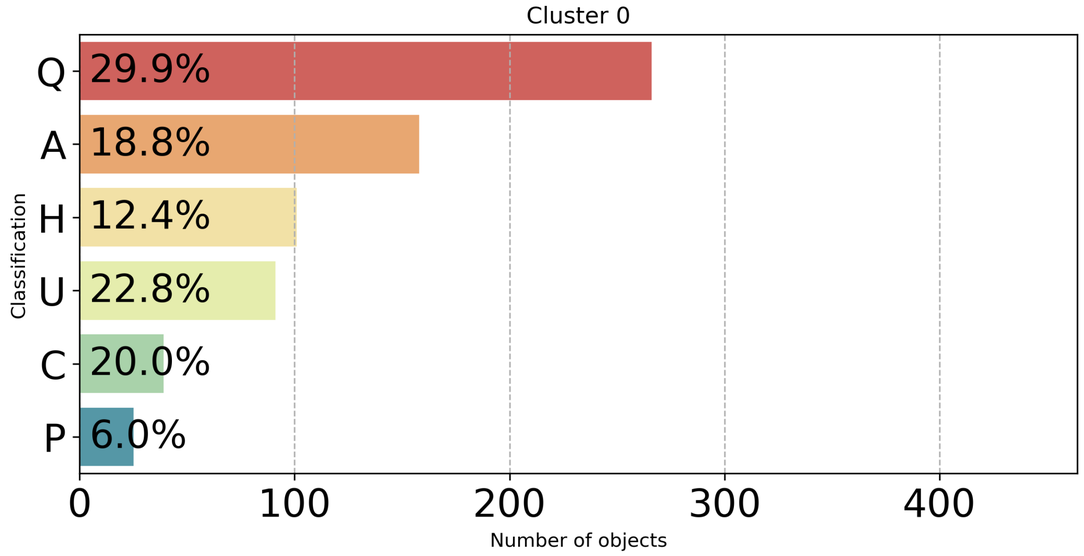


Результаты кластеризации с обучением

U — неклассифицированы



U — определены



Заключение

- Ничего не зная о природе объектов, алгоритм выделил три кластера объектов на основе интегральной интенсивностей линий молекул и потоков в ИК-диапазоне. Мы ассоциируем эти кластеры с холодными облаками, облаками с протозвездами и ФДО.

- С помощью машинного обучения нам удалось классифицировать объекты, которые не смог классифицировать человек на основе своих знаний о свойствах объектов.

- Типы объектов в кластерах, выделенные алгоритмом, отличаются от типов объектов, которые выделяет человек.

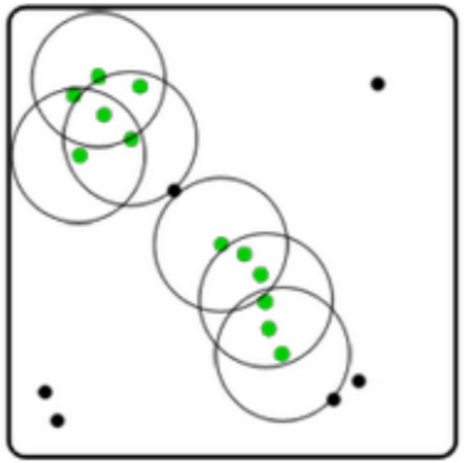
Работа выполнена при поддержке гранта РНФ 24-22-00097

Параметры HDBSCAN

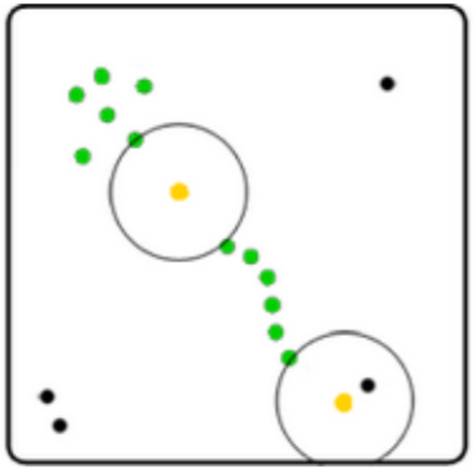
min_cluster_size — определяет минимальное количество точек для кластера

min_samples — определяет минимальное количество “соседей”, для того, чтобы точка считалась ядром

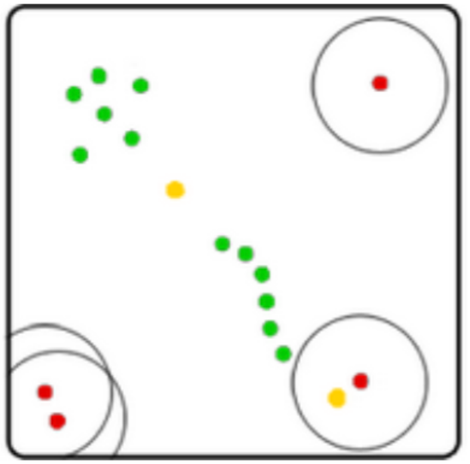
Поиск ядер кластера



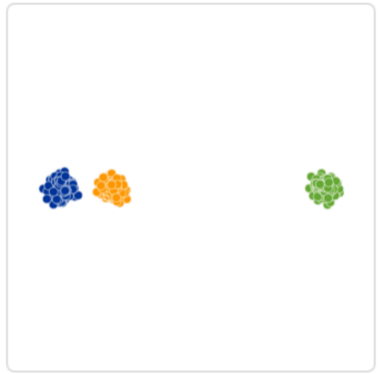
Поиск граничных точек



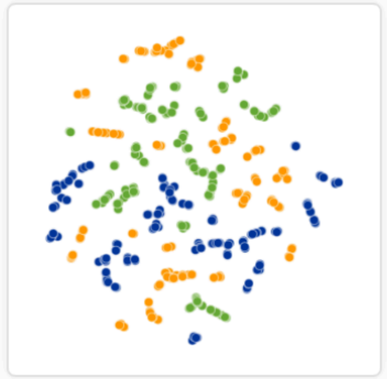
Определение выбросов



Влияние параметра perplexity в t-SNE

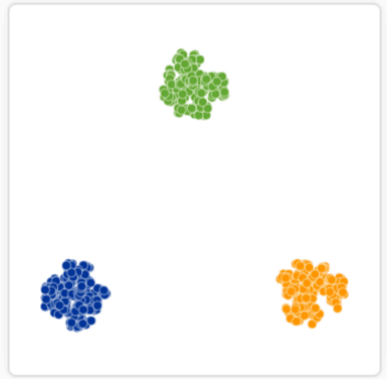


Original

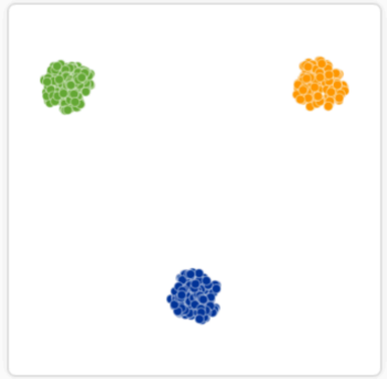


Perplexity: 2
Step: 5,000

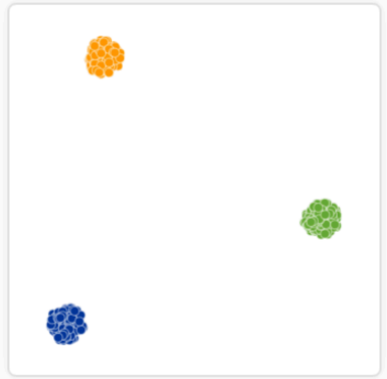
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000

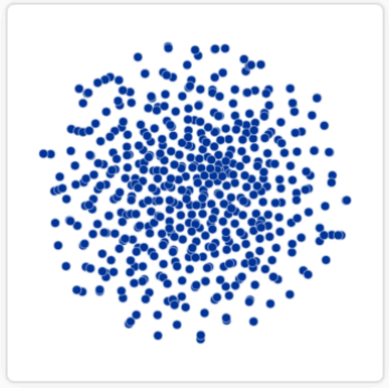


Perplexity: 100
Step: 5,000

Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

Масштабирование данных

StandardScaler

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Данные центрируются относительно среднего и масштабируются по стандартному отклонению.

Выбросы могут сильно влиять на среднее и стандартное отклонение, что может привести к искажению масштабирования.

MinMaxScaler

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \times$$

$$\times (max - min) + min$$

Не искажает распределение признака

Выбросы могут существенно влиять на X min и X max, сжимая остальные данные в узкий диапазон

RobustScaler

$$X_{\text{scaled}} = \frac{X - \text{Median}(X)}{\text{IQR}(X)}$$

Устойчивость к выбросам

Не центрирует данные вокруг нуля, если распределение асимметрично

Масштабирование данных (нелинейное)

QuantileTransformer

1: Вычисление эмпирических квантилей исходного признака

2: Сопоставление этих квантилей с квантилями целевого распределения.

3: Преобразование исходных значений на основе этого соответствия.

Преобразует данные так, чтобы их распределение стало **равномерным** в диапазоне от 0 до 1

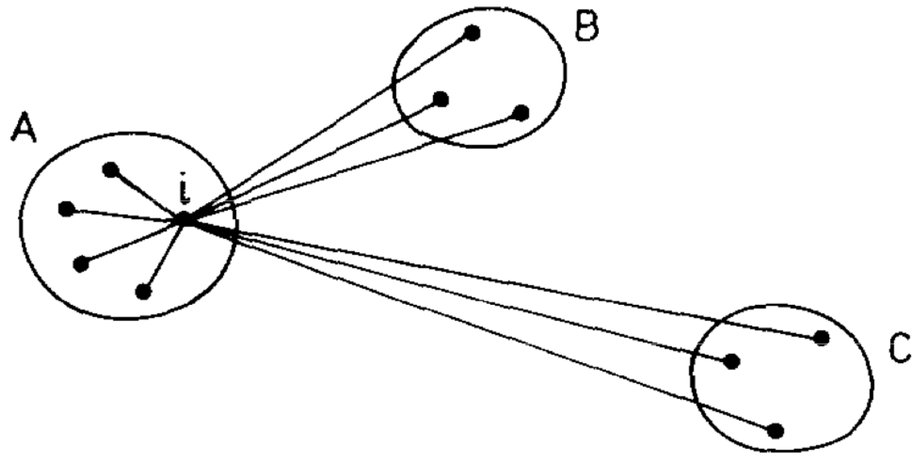
Преобразует данные так, чтобы их распределение стало **нормальным** (гауссовым) со средним 0 и стандартным отклонением 1

Оценка силуэта

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

$$a_i = \frac{1}{N_A - 1} \sum_{j \neq i} d(x_i, x_j) \qquad d(x_i, C) = \frac{1}{N_C} \sum_{k=1}^{N_C} d(x_i, x_k)$$

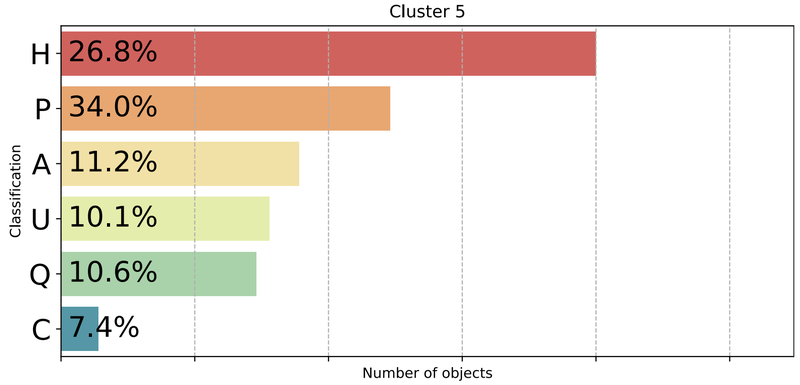
$$b_i = \min d(x_i, C)$$



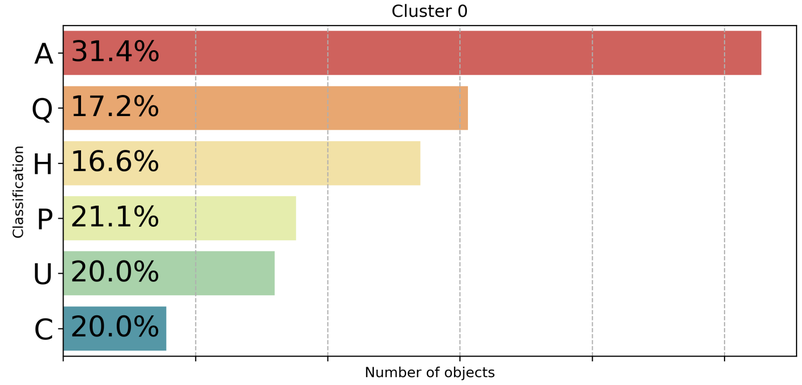
ROUSSEEUW 1986

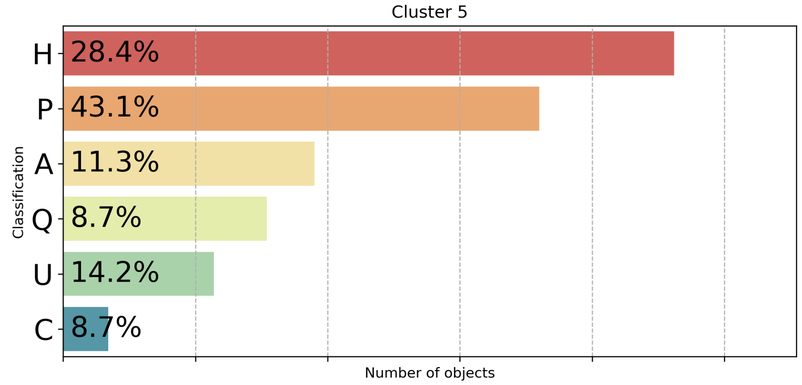
+Spitzer 3.6 & 4.5μm

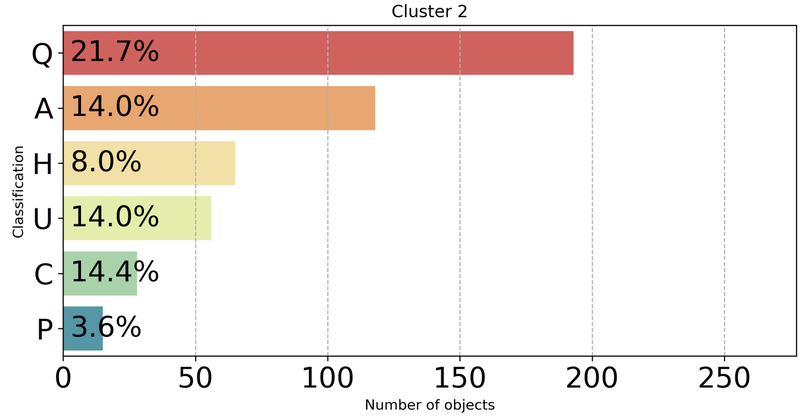
BEFORE



Identified U







Масштабирование данных (нелинейное)

QuantileTransformer

1: Вычисление эмпирических квантилей исходного признака

2: Сопоставление этих квантилей с квантилями целевого распределения.

3: Преобразование исходных значений на основе этого соответствия.

Преобразует данные так, чтобы их распределение стало **нормальным** (гауссовым) со средним значением 0 и стандартным отклонением 1

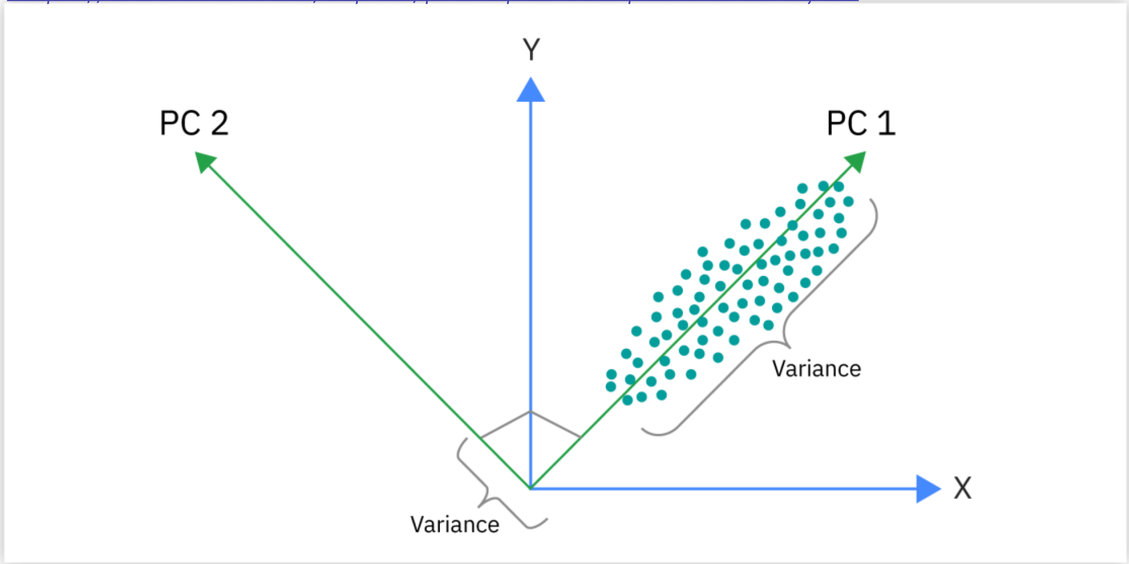
Уменьшение размерности данных (линейное)

PCA (Principal Component Analysis)

Особенности:

- Уменьшение размерности: сокращает количество признаков в данных, сохраняя при этом как можно больше информации
- Упрощение структуры данных: устраняет корреляции между признаками, создавая новый набор независимых переменных
- Визуализация: упрощает визуализацию высокоразмерных данных путем проекции их на пространство меньшей размерности

https://www.ibm.com/topics/principal-component-analysis



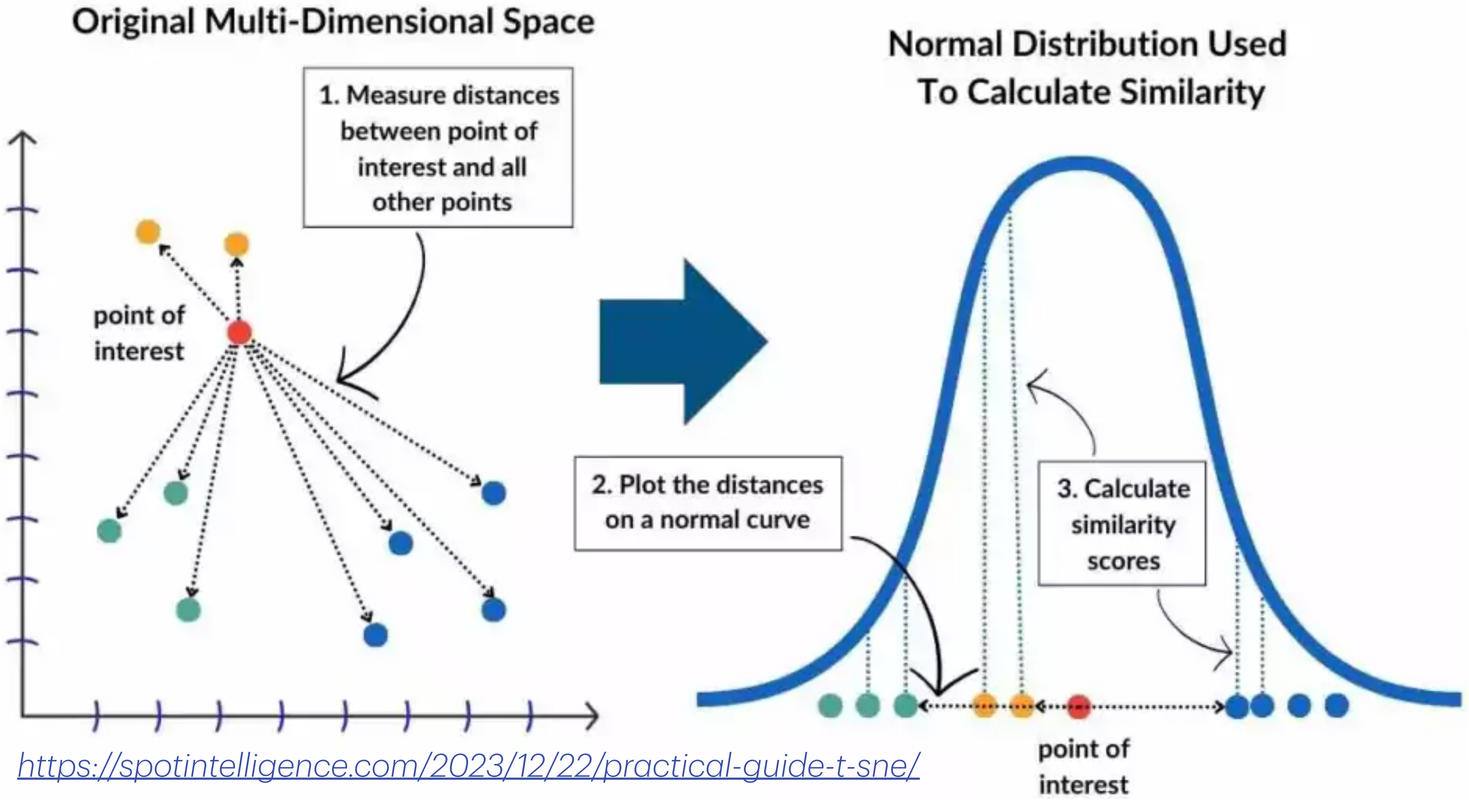
PCA предполагает линейные зависимости между признаками. Он не справляется с нелинейными структурами данных

Уменьшение размерности данных (нелинейное)

t-SNE (t-distributed Stochastic Neighbor Embedding)

Особенности:

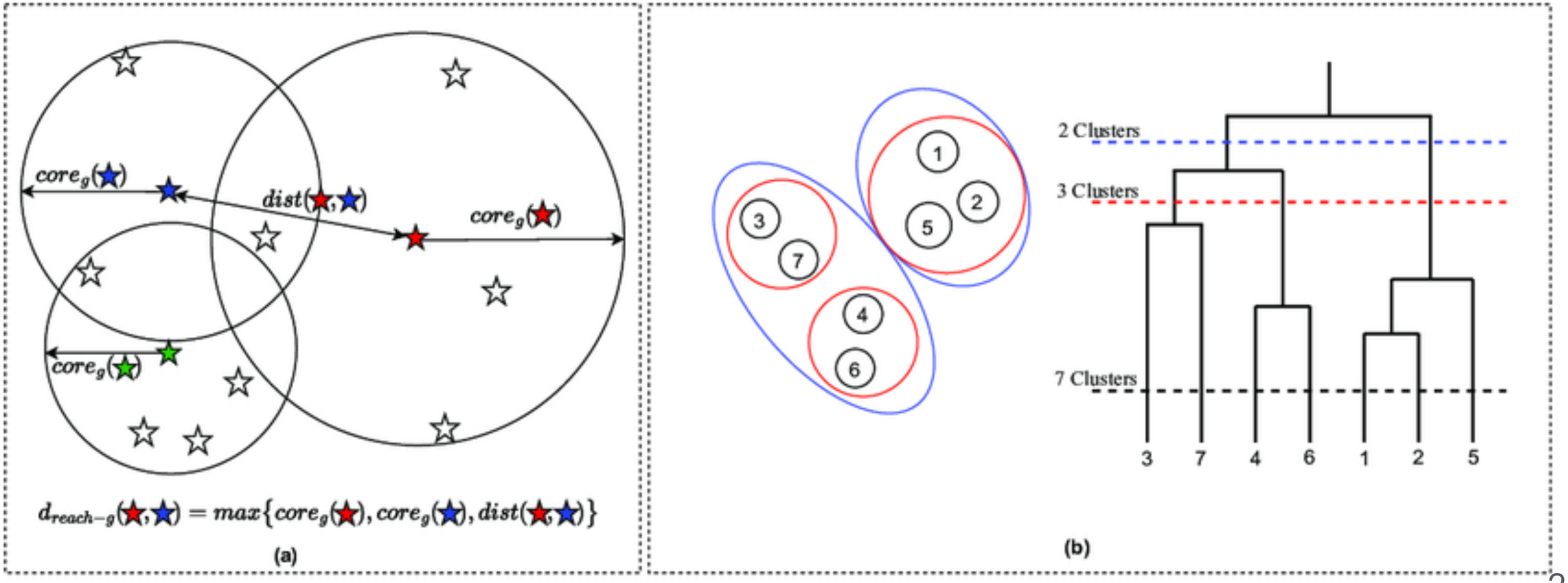
- Сохранение локальных отношений: Если две точки близки в исходном пространстве, алгоритм стремится расположить их близко и в низкоразмерном пространстве.
- Полезен для обнаружения скрытых структур и кластеров в данных



https://spotintelligence.com/2023/12/22/practical-guide-t-sne/

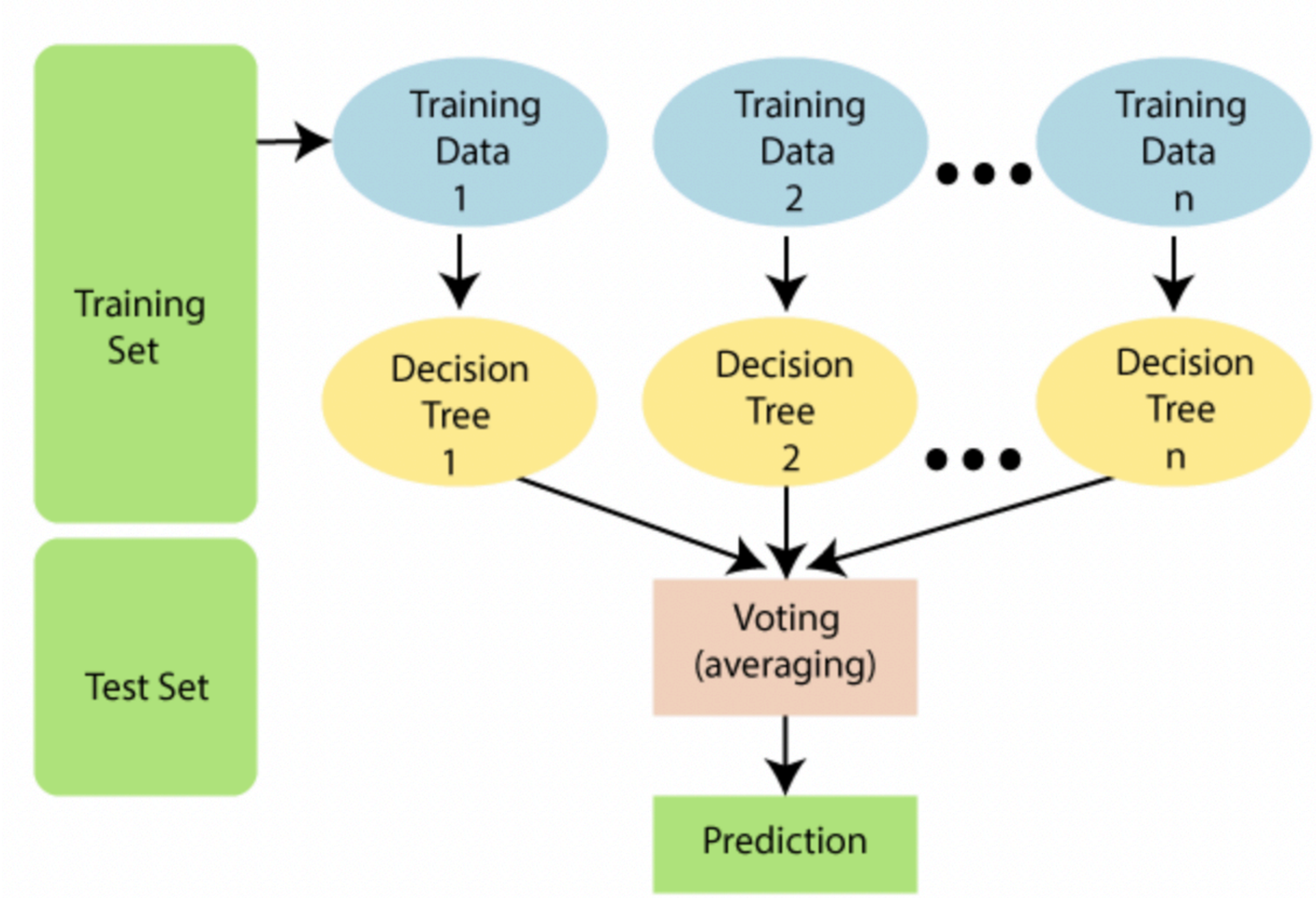
Алгоритм кластеризации HDBSCAN

Предназначен для выявления кластеров произвольной формы в наборах данных с неравномерной плотностью



Owiredu et al 2022. A Quasi-Intelligent Maritime Route Extraction from AIS Data. Sensors.

Метод случайного леса (Random Forest)



https://www.tpointtech.com/machine-learning-random-forest-algorithm

Spitzer+MALT90+ U

	870μm		HCO+		HNC		N2H+		HCN		CCH		3.6μm		4.5μm	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
cluster 2	3.05	1.21	7.65	3.96	5.67	2.85	6.40	3.72	9.03	4.09	2.59	1.55	28.87	2.20	65.69	3.10
cluster 5	2.64	1.31	6.78	4.85	4.49	3.22	5.38	3.78	7.74	5.29	2.66	1.98	0.00	0.00	0.00	0.00
cluster 0	1.25	0.84	6.64	2.85	4.10	2.14	3.58	2.43	8.80	2.71	0.00	0.00	15.19	1.44	16.48	1.86

Spitzer+MALT90+identified U

	870μm		HCO+		HNC		N2H+		HCN		CCH		3.6μm		4.5μm	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
cluster 0	3.05	1.21	7.64	3.95	5.65	2.85	6.38	3.71	9.00	4.07	2.58	1.55	28.80	2.20	65.47	3.10
cluster 5	2.57	1.28	6.36	4.61	4.16	2.94	4.73	3.09	7.31	5.03	2.54	1.88	0.00	0.00	0.00	0.00
cluster 2	1.25	0.84	6.64	2.85	4.10	2.14	3.58	2.43	8.80	2.71	0.00	0.00	15.19	1.44	16.48	1.86

MALT90+ U

	870μm		HCO+		HNC		N2H+		HCN		CCH	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
cluster 1	1.27	0.83	5.73	2.83	3.60	2.07	3.30	2.31	7.27	2.63	0.00	0.00
cluster 0	2.96	1.34	7.26	4.54	5.14	3.15	6.19	4.10	8.36	4.72	2.59	1.73

MALT90+identified U

	870μm		HCO+		HNC		N2H+		HCN		CCH	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
cluster 0	1.27	0.83	5.73	2.83	3.60	2.07	3.30	2.31	7.27	2.63	0.00	0.00
cluster 5	2.96	1.34	7.26	4.54	5.14	3.15	6.19	4.10	8.36	4.72	2.59	1.73