Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

Васильев Семён Михайлович

Применение машинного обучения для построения карты крупномасштабной структуры Вселенной по данным многоволновых обзоров неба

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:
к.ф-м.н. А.В.Мещеряков
С.В.Герасимов

Москва, 2023

Оглавление.

1. Аннотация.

Целью данной работы является исследование и разработка технологии построения карты крупномасштабной структуры Вселенной, включающий в себе, модель фотометрического прогноза красного смещения галактик, модель построения и оценки карт волокон галактик и модель отождествления и прогноза красного смещения скоплений галактик. Для прогноза фотометрических красных смещений галактик была исследована и разработана модель на основе глубокого ансамбля, оценивающая вероятностное распределение красного смещения галактики, как смесь нормальных распределений. Для построения карт волокон галактик была применена и исследована модель DisPerSe. Для оценки карт волокон галактик используется свойство связности крупномасштабной структуры Вселенной, рассматривались пересечения построенных волокон с реальными скоплениями из тестового каталога. Для оценки вероятности нахождения скопления галактик в точке трехмерного пространства используется информация о ближайших волокнах галактик, информация о локальной плотности галактик вблизи скопления и классические алгоритмы машинного обучения: случайный лес и градиентный бустинг. Модель оценки вероятности нахождения скопления галактик в заданном направлении и прогноза его красного смещения использует построенные модели оценки вероятности скопления в точке для анализа луча, проходящего в исследуемом направлении.

2. Введение.

2.1. Крупномасштабная структура Вселенной.

Темная материя во Вселенной на больших масштабах распределена неравномерно и формирует связные структуры, образующие систему, похожую на паутину. Эта система и называется крупномасштабной структурой Вселенной [1]. Темная материя не взаимодействует с электромагнитным излучением, однако распределение видимого вещества (газа и галактик) отражает распределение тёмной материи.

Структуры, которые образует темная материя, можно разделить на несколько видов в зависимости от их формы:

- Гало темной материи – компактные области повышенной плотности вещества. Скопления (кластеры) галактик являются маркерами гало темной материи. Они могут содержать до 1000 галактик. Радиус скоплений галактик варьируется от 1 до 5 мегапарсек. Парсек – единица измерения расстояния, широко используемая в астрономии, равная расстоянию до объекта, у которого изменение координат на небесной сфере, связанное с движением Земли вокруг Солнца, составляет одну угловую секунду. Один парсек приблизительно равен 3,26 световым годам или $3 \cdot 10^{13}$ км.
- Волокна (филаменты) темной материи – крупнейшие из наблюдаемых космических структур во Вселенной. Волокна галактик являются маркерами волокон темной материи. Они имеют продолговатую нитевидную структуру. Средняя длина волокон составляет от 50 до 80 мегапарсек.
- Стены темной материи. Стены галактик являются их маркерами. Стены галактик – плоские области повышенной плотности галактик, обрамляющие пустоты (войды).
- Войды – обширные области низкой плотности темной материи и, соответственно, галактик между стенами и волокнами. Размеры войдов могут варьироваться от 10 до 100 мега парсек, а плотность вещества в них может быть на порядок меньше, чем средняя во Вселенной.

Описанные выше структуры тесно связаны между собой в сеть, простирающуюся через всю наблюдаемую Вселенную.

На рис. 1 можно видеть результат выполнения N-body симуляции, иллюстрирующий строение крупномасштабной структуры Вселенной.

N-body симуляция – симуляция динамической системы частиц под действием физических сил, таких как гравитация. Широко применяется в астрономии для исследования динамики систем нескольких тел. Применяется в космологии для симуляции формирования структур галактик и процессов эволюции крупномасштабной структуры Вселенной.

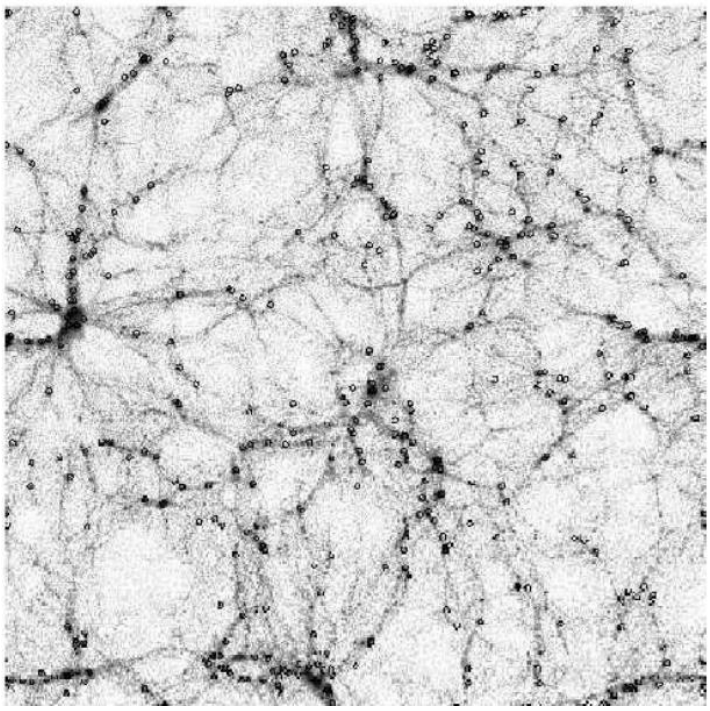

Рисунок 1. Результат выполнения N-body симуляции [1].

2.2. Построение карты крупномасштабной структуры Вселенной.

Для построения карты крупномасштабной структуры Вселенной необходимо иметь общую технологию, решающую задачи, возникающие на разных этапах построения карты:

- Галактики – основные объекты, формирующие крупномасштабную структуру Вселенной. Для построения карты крупномасштабной структуры Вселенной необходимо иметь точные координаты галактик, для этого нужно осуществлять прогноз расстояний до галактик по их табличным описаниям.
- Для построения карты крупномасштабной структуры Вселенной необходимо уметь выделять формируемые галактиками структуры (волокна, скопления), составляющие крупномасштабную структуру Вселенной.
- Необходимо уметь проводить анализ связности этих структур.

2.3. Прогноз расстояний до астрономических источников.

В астрономии существует 2 основных подхода для прогноза расстояний до объектов, являющихся источниками электромагнитного излучения.

2.3.1. Спектрографический прогноз расстояния.

У источника снимается спектр его электромагнитного излучения рис. 2. В спектре выделяются линии, соответствующие излучению различных химических элементов, входящих в состав данного источника. Выделенные линии будут иметь большую длину волны (смещены в сторону красного диапазона, красное смещение – мера расстояния до астрономических объектов, она обозначается, как Z) по сравнению с линиями, полученными в лабораторных условиях. Данный эффект связан с расширением Вселенной и удалением астрономических объектов друг от друга. По величине смещения можно определить скорость удаления объекта. Закон Хаббла дает зависимость скорости удаления объекта от расстояния до него.

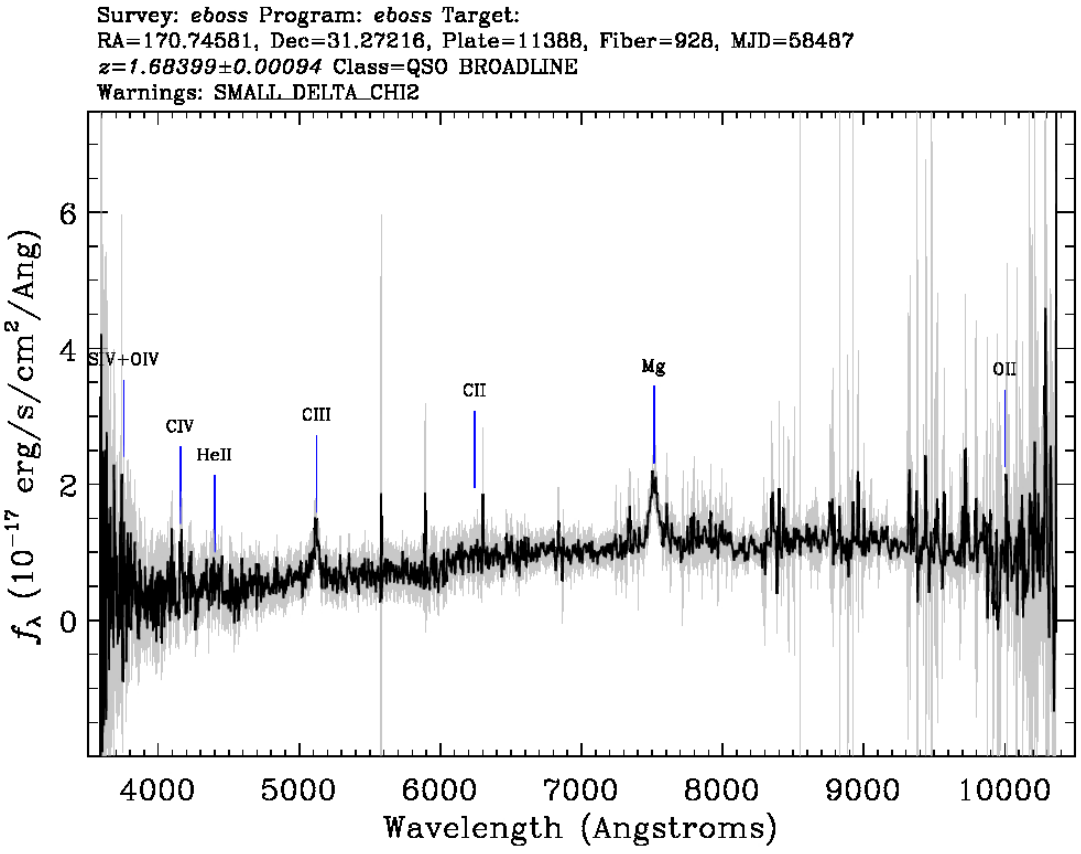

Рисунок 2. Спектр активного ядра галактики (квазара).

Спектрографический прогноз является прямым методом прогноза расстояния, дающим точным результат. Однако, снятие спектра – дорогостоящая операция, поэтому только ~1% всех наблюдаемых объектов имеют спектрографические прогнозы красного смещения.

2.3.2. Фотометрический прогноз расстояния.

Для фотометрического прогноза расстояния необходимо строить модели регрессии на табличных данных. В качестве признаков используются значения яркости объекта в различных цветовых фильтрах. В качестве целевой переменной для настройки параметров модели регрессии используют спектрографические прогнозы красного смещения.

Данный метод не является прямым и обладает большей погрешностью, чем спектрографический, но даёт более полные прогнозы, так как все наблюдаемые объекты имеют фотометрические признаки.

Также стоит отметить, что прогнозы фотометрического красного смещения некоторых типов галактик (квазары – активные ядра галактик) имеют многомодальное распределение. Учет этого факта при построении модели может привести к значительному улучшению качества предсказания.

2.4. Методы поиска скоплений галактик.

На данный момент существует целый ряд методов поиска скоплений галактик. Некоторые из них используют наблюдения в определенных диапазонах электромагнитного излучения, другие ищут области повышенной плотности галактик или группы галактик, обладающих определенными свойствами:

- Поиск скоплений галактик в рентгеновском диапазоне электромагнитного излучения. Наблюдения рентгеновских телескопов показали, что подавляющее большинство крупных скоплений галактик являются мощными источниками рентгеновского излучения. Это обуславливает наличием горячего разреженного межгалактического газа во внутрикластерной среде [2]. Массовая доля газа может превышать массовую долю галактик скопления на порядок. eRosita [3] (extended Roentgen Survey with an Imaging Telescope Array) – рентгеновский телескоп, установленный на базе космической обсерватории СРГ. Целью данного проекта

является поиск рентгеновских источников за счет сканирования всего неба в мягком рентгеновском диапазоне. Стоит отметить, что не все далекие рентгеновские источники являются скоплениями галактик. Существуют точечные рентгеновские источники, например, активные ядра галактик (квазары). Поэтому встает задача отождествления скоплений галактик, то есть задача оценки вероятности наличия скопления в точке пространства или в некотором направлении (задача классификации). Также важно подчеркнуть, что при использовании данного метода обнаружения скоплений галактик возникают проблемы с определением их точного положения, так как устанавливается только направление излучения, а не расстояние до источника

- Поиск скоплений галактик в микроволновом диапазоне электромагнитного излучения. Интенсивность радиоизлучения реликтового фона изменяется при прохождении через межгалактический газ во внутрикластерной среде. Этот эффект называется эффектом Сюняева – Зельдовича [4]. Благодаря ему есть возможность использовать профиль реликтового излучения для поиска скоплений галактик. Телескоп ACT (Atacama Cosmology Telescope) [5] – один из инструментов осуществления, описанного метода.
- Поиск красных последовательностей. Большинство богатых скоплений галактик содержит галактики, между яркостью и цветом которых наблюдается линейная зависимость. Группы таких галактик называются красными последовательностями. На рис. 3 представлен пример красной последовательности скопления Abell 1084. Метод предложенный в работе [6] использует описанный выше факт. Производится поиск галактик, близких на небесной сфере и составляющих красную последовательность. Такие группы галактик считаются кандидатами в скопления галактик.

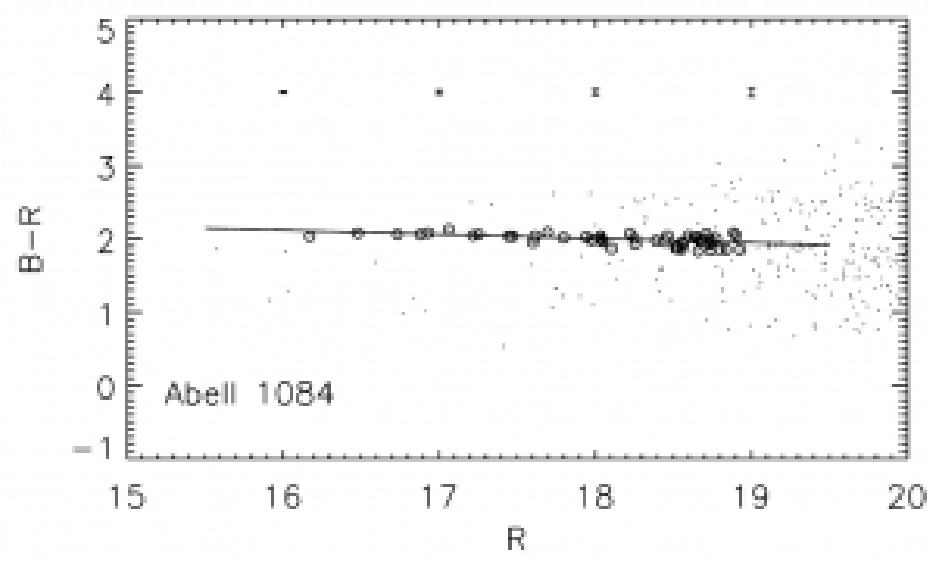

Рисунок 3. Красная последовательность скопления Abell 1084.

- Поиск локальных максимумов плотности распределения галактик. Для поиска скоплений часто применяется алгоритм Friend-of-friends [7]. Имеется набор точек в трехмерном пространстве. Точки, расстояние между которыми меньше заданного, называются друзьями. Каждая точка напрямую связана со своим друзьями и опосредовано с друзьями друзей. Результатом работы алгоритмы является набор групп связных объектов. Для работы алгоритма необходимо настроить свободный параметр – максимальная длина связи. Кандидатов в скопления галактик можно отбирать по численности групп.

2.5. Актуальность.

Исследование крупномасштабной структуры Вселенной необходимо для понимания процесса эволюции и состава Вселенной.

Для построения карты крупномасштабной структуры Вселенной необходима общая технология, решающая задачи, возникающие на разных этапах построения карты

Для понимания распределения вещества во Вселенной и построения карты крупномасштабной структуры Вселенной необходимо уметь точно оценивать расстояния до астрономических источников.

Текущие нейросетевые модели прогноза фотометрического красного смещения галактик не учитывают многомодальную природу данных.

Для построения карты крупномасштабной структуры Вселенной необходимо уметь искать волокна галактик и скопления галактик.

Связность крупномасштабной структуры Вселенной позволяет предположить, что для поиска скоплений галактик можно использовать глобальные карты волокон галактик, однако такой подход до сих пор слабо изучен.

Некоторые инструменты поиска скоплений галактик могут принимать за скопления объекты другой природа, например, квазары. Необходимо решать задачу отождествления кандидатов в скопления галактик с помощью независимых данных.

3. Постановка задачи.

Целью данной работы является исследование и разработка технологии для построения карты крупномасштабной структуры Вселенной по данным многоволновых обзоров неба.

Постановка задачи:

- Исследовать и разработать модель прогноза фотометрических красных смещений галактик, учитывающую многомодальность данных.
- Исследовать и применить модель DisPErSe для построения и оценки карт волокон галактик.
- Исследовать и разработать модель для отождествления скоплений галактик и прогноза их красного смещения по информации о ближайших структурах галактик и локальной плотности галактик.
- Разработать библиотеку для построения и анализа карты крупномасштабной структуры Вселенной.

4. Обзор существующих решений.

4.1. Цели обзора.

Целями данного обзора являются:

- Обзор существующих решений для оценки уверенности предсказаний нейросетевых моделей машинного обучения.
- Метрики прогноза красного смещения и классификации.
- Обзор существующих алгоритмов поиска волокон галактик.
- Поиск исследований, проводящих сравнение алгоритмов поиска волокон галактик или предлагающих метод их сравнения.
- Разбор существующих методов для поиска скоплений галактик, использующих глобальные карты волокон галактик.

4.2. Глубокие ансамбли.

Глубокие ансамбли – это ансамбли глубоких нейронных сетей, обученных независимо на одних и тех же наборах данных, но с разной инициализацией весов [8].

Глубокие ансамбли применяются для оценки уверенности прогнозирования нейронной сети. Ошибку можно разложить на две составляющие:

- Aleatoric error σ_a^2 – часть ошибки, вызванная шумом в данных.
- Epistemic error σ_e^2 – часть ошибки, вызванная несовершенством модели и неполнотой данных.

В случае, когда нейронные сети, составляющие глубокий ансамбль моделируют нормальное распределение, они генерируют для каждого объекта два параметра нормального распределения: μ_i – математическое ожидание и σ_i – дисперсия.

Тогда:

- $\sigma^2(x) = \sigma_a^2(x) + \sigma_e^2(x)$
- $\sigma_a^2(x) = \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2(x)$
- $\sigma_e^2(x) = \frac{1}{M}\sum_{i=1}^{M}[\mu_i(x) - \frac{1}{M}\sum_{j=1}^{M}\mu_j(x)]$

4.3. Метрики прогноза красного смещения и классификации.

В астрономии существует 2 основные метрики для оценки качества решения задачи прогноза красного смещения:

- Доля катастрофических выбросов: $n_{>0.15} = \frac{\#\{i=\overline{1,N} || \delta z_i| > 0.15\}}{N}$
- Нормированное медианное абсолютное отклонение: $\sigma_{nmad} = 1.48 * median(|\delta z_i|)$,

где $\delta z_i = \frac{z_{ph}^i - z_{spec}^i}{1+z_{spec}^i}$, z_{ph}^i – предсказанное красное смещение, z_{spec}^i – спектрографическое красное смещение.

В задачах классификации используется метрика ROC-AUC – площадь под FPR-TPR кривой.

4.4. Алгоритмы поиска волокон галактик.

4.4.1. DisPerSe.

В качестве входных данных алгоритм DisPerSe (Discrete Persistent Structures Extractor) [9] принимает дискретное множество точек (галактики) в двумерном или трехмерном пространстве. На первом шаге работы алгоритма необходимо провести оценку плотности их распределения. В алгоритме DisPerSe оценка плотности производится с помощью алгоритма DTFE (The Delaunay Tessellation Field Estimator) [10].

4.4.1.1. DTFE.

Алгоритм оценки плотности DTFE принимает дискретное множество точек в двумерном или трехмерном пространстве. На первом шаге работы алгоритма производится построение триангуляции Делоне. Пространство разбивается на треугольники (или тетраэдры в случае трех измерений) таким образом, чтобы для любого треугольника все точки из заданного множества за исключение точек, являющихся его вершинами, лежали вне окружности, описанной вокруг данного треугольника.

Триангуляция Делоне однозначно определена. Триангуляция Делоне максимизирует минимальный угол из всех углов всех построенных треугольников. Таким образом, метод стремиться строить равносторонние треугольники.

Как видно из рис. 4, в областях повышенной плотности точек строятся треугольники малой площади, а в областях пониженной плотности большой. Следовательно, площадь треугольников может служит мерой локальной плотности точек.

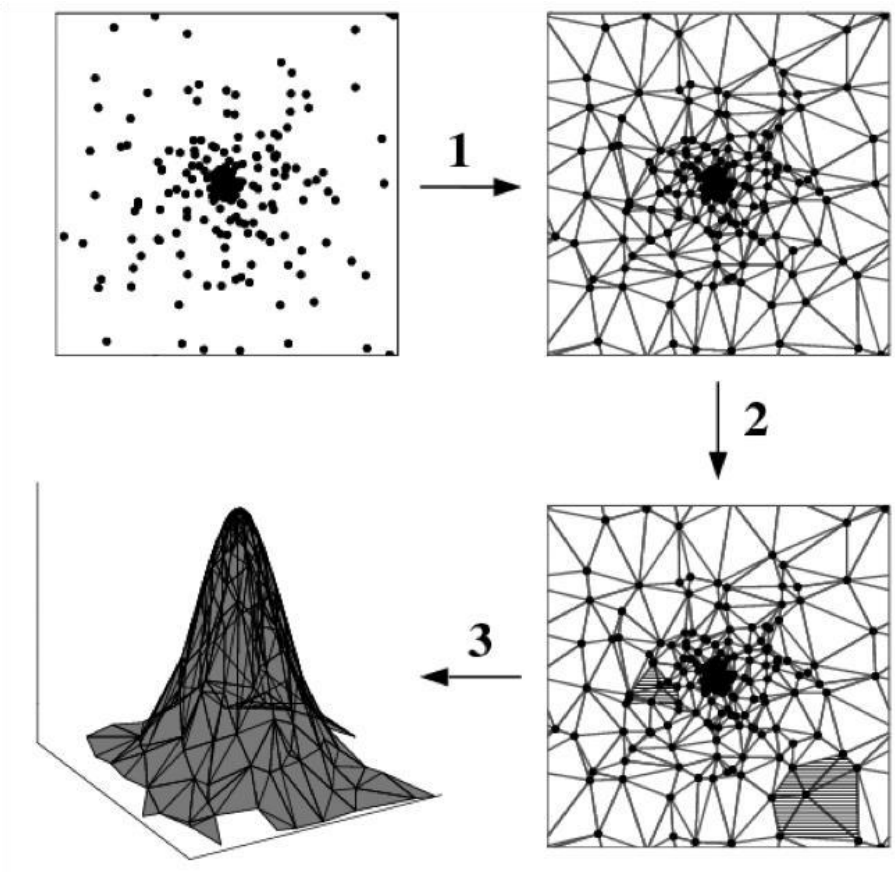

Рисунок 4. Этапы работы алгоритмы DTFE.

На втором этапе работы алгоритма производится оценки плотности в точках заданного множества. Плотность в этих точках определяется, как $\frac{M_i}{S_i}$, где M_i – масса точки, а S_i – сумма площадей смежных треугольников.

На третьем шаге работы алгоритма функция плотности распределения приближается во внутренних точках треугольников линейной функцией.

Свойства алгоритма DTFE:

- Не требует настройки параметров.
- Учитывает не только плотность точек, но и их взаимное расположение.
- Адекватно оценивает разреженные регионы.
- Чувствителен к шуму.
- Линейная оценка плотности не гладкая.

- Ошибки на границах области оценки плотности. Треугольники на границах имеют бесконечную площадь.
- Ошибки в пустых регионах, так как площадь треугольников всегда конечна (кроме треугольников на границах).

Алгоритм DTFE позволяет проводить итеративное сглаживание оценки плотности за счет усреднения оценки в вершинах треугольников по значениям оценки в вершинах смежных треугольников.

4.4.1.2. DisPerSe. Основные этапы.

На втором этапе работы алгоритма DisPerSe производится оценка градиента функции плотности распределения и строится множество критических точек.

Критическая точка – точка, в которой градиент равен нулю. У двумерной функции существует 3 типа критических точек (минимум, седловая, максимум). У трехмерной 4 типа (есть 2 типа седловых точек).

Вводится понятие интегральной линии. Интегральная линия – это кривая, сонаправленная полю градиента. Через каждую неинтегральную точку проходит ровно одна интегральная линия. Отсюда следует, что интегральные линии заполняют собой всю область определения функции. Каждая интегральная линия начинается и заканчивается в критических точках.

В следствии этих замечаний, интегральные линии порождают разбиение пространства на области возрастания (убывания) функции. В каждую подобласть разбиения входят только точки, через которые проходят интегральные линии, начинающиеся в одном и том же минимуме (заканчивающиеся в одном и том же максимуме). Иллюстрация этапов работы алгоритма представлена на рис. 5.

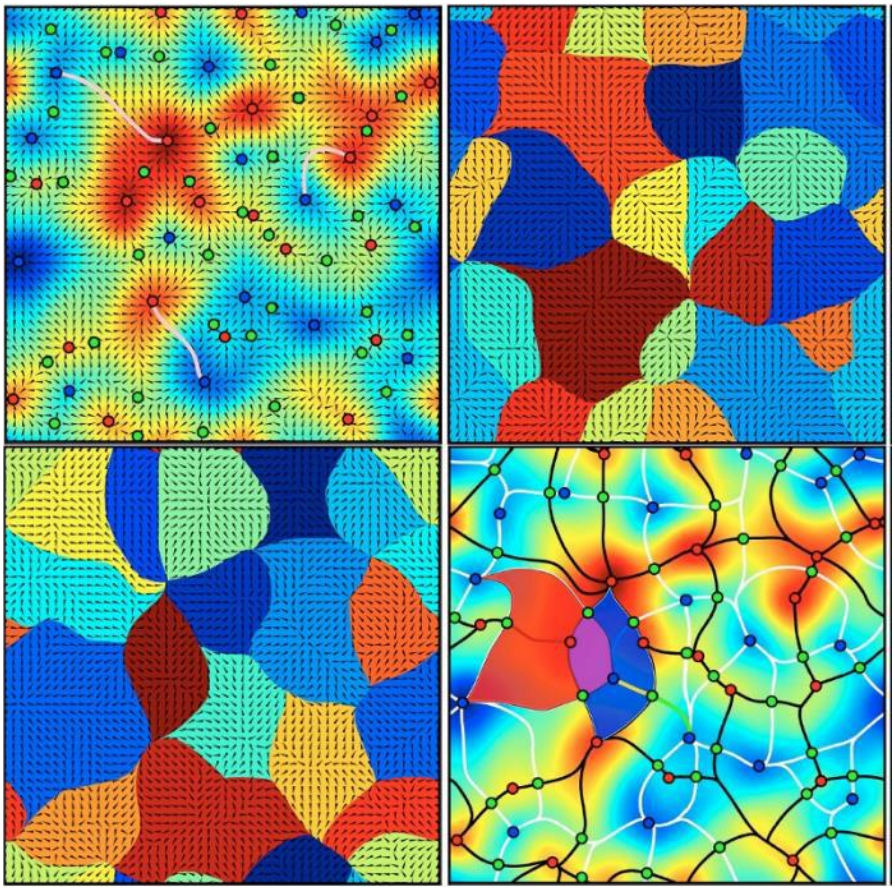

Рисунок 5. Этапы работы алгоритма DisPerSe. Оценка градиента. Выделение критических точек. Выделение интегральных линий. Построение разбиения.

При решении задачи выделения компонентов крупномасштабной структуры Вселенной нас интересует разбиение на области возрастания функции. В этом случае топологические структуры, найденные алгоритмом DisPerSe, имеют следующий физический смысл (рис. 6):

- Точки максимума функции распределения – скопления галактик.
- Области возрастания функции – войды.
- Грани областей возрастания функции – стены галактик.
- Ребра областей возрастания функции – волокна галактик.

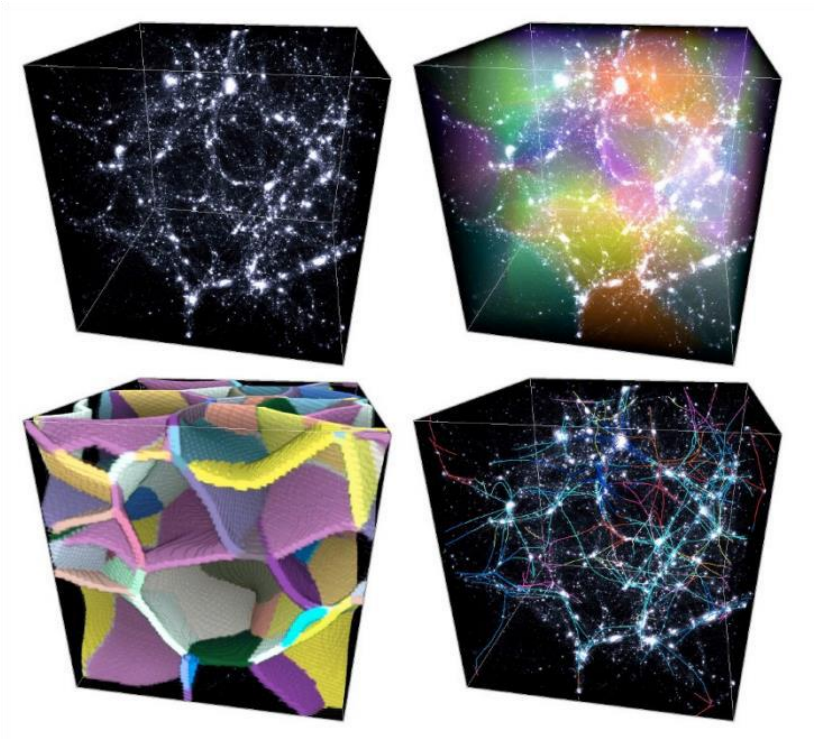

Рисунок 6. Физические смысл найденных топологических структур. Верх-лево: максимумы плотности, скопления. Верх-право: области возрастания плотности, войды. Низ-лево: грани областей возрастания, стены. Низ-право: ребра областей возрастания, волокна.

4.4.1.3. Устойчивость найденных топологических структур.

Важной особенностью алгоритма DisPerSe является возможность отфильтровать найденные топологические структуры по их устойчивости. Если просматривать профиль одномерной функции от больших значений к меньшим (рис. 7), то видно, что топологические структуры порождаются точками локальных максимумов и уничтожаются точками локальных минимумов. Разность между порождающей и уничтожающей критической точкой называется устойчивостью топологической структуры.

Алгоритм DisPerSe позволяет исключать из результата выполнения топологические структуры, обладающие низкой устойчивостью, либо по порогу абсолютного значения устойчивости структур, либо по порогу вероятности появления структуры с данным значением устойчивости в случайно сгенерированных данных. Вероятность задается в терминах «сигма» (как у нормального распределения).

Удаление структуры производиться путем локального сглаживания функции в области структуры.

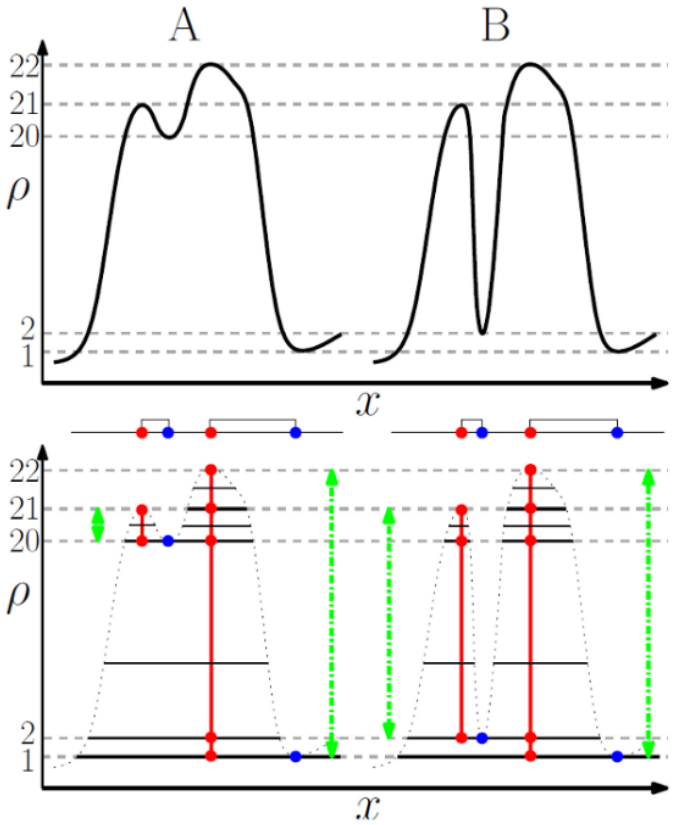

Рисунок 7. Порождение и уничтожение топологических структур в профиле одномерной функции.

4.4.2. Случайный лес, обученный на легковесной симуляции.

Авторы работы [11] предлагают метод генерации обучающих данных, совмещающий эвристический подход и физические законы.

Этапы генерации обучающих данных:

- Генерация скоплений галактик. Массы скоплений M сэмплируются из функции распределения масс скоплений (HMF – halo mass function), описанной в работе [12] с минимальным и максимальным числом галактик $N_{min} = 8$ и $N_{max} = 13245$ соответственно. Радиус скопления сэмплируется из логнормального распределения со средним $R(M) = R_0 \left(\frac{M}{M_{gal}}\right)^{\alpha}$ и стандартным отклонением $\sigma(M) = \sigma_0 \left(\frac{M}{M_{gal}}\right)^{\beta}$, $M_{gal} = 7.55 * 10^{10} M_{\odot}$, $R_0 = 0.12\ Mpc$, $\alpha = 0.38$, $\sigma_0 = 0.12\ Mpc$, $\beta = 0.16$. После определения радиуса и массы скопления он наполняется галактиками согласно сферическому нормальному распределению со стандартным отклонением равным $\sigma_H(R_H) = \frac{R_H}{5}$ обрезанному по радиусу скопления. Центр скопления размещается в исследуемом объеме случайно, согласно равномерному распределению.

- Генерация волокон. Случайным образом выбираются центры двух сгенерированных скоплений (за исключением тех пар, между которыми уже сгенерировано волокно). Между выбранными точками строится кривая Безье степени 2 [13]. Радиус волокна R_F выбирается из равномерного распределения на отрезке $[R_{F,min}, R_{F,max}]$, где $R_{F,min} = 0.3\ Mpc$, а $R_{F,max} = 0.6\ Mpc$. Радиальная плотность волокна выбирается из равномерного распределения на отрезке $[\lambda_{min}(R_F), \lambda_{max}(R_F)]$, где $\lambda_{min}(R_F) = B_{min}\left(\frac{R_{F,max}-R_F}{R_{F,max}-R_{F,min}}\right)^3 + \lambda_0$, $\lambda_{max}(R_F) = B_{max}\left(\frac{R_{F,max}-R_F}{R_{F,max}-R_{F,min}}\right)^3 + \lambda_0$, $B_{min} = 0.65\ Mpc^{-4}$, $B_{max} = 1.15\ Mpc^{-4}$, $\lambda_0 = 2.85\ Mpc^{-1}$. Частицы генерируются вокруг волокна согласно нормальному распределению со стандартным отклонением равным $\sigma_F(R_F) = \frac{5R_F}{4}$ обрезанному по радиусу волокна.
- Генерация фона. Частицы генерируются по всему исследуемому объему согласно равномерному распределению до тех пор, пока не будет достигнуто суммарное количество частиц, задаваемое при запуске генерации обучающих данных.
- Разметка сгенерированных галактик. Галактики, сгенерированные на этапе построения скоплений и волокон, размечаются, как галактики, принадлежащие скоплениям и волокнам, соответственно. Галактики, сгенерированные на этапе генерации волокон или фона, попавшие в радиус скоплений, размечаются, как принадлежащие скоплениям. Галактик, сгенерированные на этапе генерации фона, попавшие в радиус волокон, но не попавшие в радиус скоплений, размечаются, как принадлежащие волокнам. Остальные фоновые галактики размечаются, как принадлежащие войдам.

Далее для каждой сгенерированной галактики вычисляется набор признаков, несущих информацию о локальной плотности распределения галактик.

Признаки:

- Объем ячейки разбиения Вороного соответствующей данной галактике.
- Число частиц в некотором радиусе.
- Расстояние до центра масс частиц в некотором радиусе.
- Момент инерции частиц в некотором радиусе.
- Расстояние до k-го ближайшего соседа.
- Разность между долями объясненной дисперсии компонент PCA (анализ главных компонент), построенного по галактикам в некотором радиусе.

Имея для каждой сгенерированной галактики признаки и разметку классов (скопление, волокно, войд) авторы обучают на этих данных классификатор. В качестве алгоритма классификации авторы выбрали случайный лес [14].

4.5. Выводы.

Был проведен анализ 30 научных статей за последние 10 лет. По результатам проведения обзора:

- Были найдены решения для оценки уверенности предсказаний нейросетевых моделей – глубокие ансамбли.
- Были рассмотрены классическое метрики оценки качества решения поставленных задач.
- Были рассмотрены существующие алгоритмы поиска волокон галактик различных типов: топологические, графовые, с использованием машинного обучения и эвристических симуляций для генерации данных.
- Выяснилось, что не существует исследований, которые проводили бы численную оценку и сравнение различных алгоритмов поиска волокон галактик. Также не существует универсальных методов численной оценки карт волокон галактик.
- Не было найдено исследований, предлагающих подход поиска скоплений галактик, использующий карты волокон галактик.

5. Построение решения задачи.

5.1. Модель прогноза фотометрического красного смещения галактик.

5.1.1. Описание наборов данных DR14Q + VHzQ и Stripe 82X.

Модель для прогноза красного смещения галактик обучалась и валидировалась на объектах из объединения каталогов DR14Q [15] и Very High Z Quasars (VHzQ) [16]. Этот набор данных содержит 580456 объектов на $Z \in [0, 6.97]$. Для каждого объекта было получено 65 фотометрических признаков из объединения четырёх фотометрических обзоров: DESI Legacy Imaging Surveys, PanSTARRS1, SDSS, WISE. Объекты из описанного набора данных содержат точные спектрографические прогнозы красного смещения, которые могут быть использованы для обучения в качестве целевой переменной.

Стоит отметить, что описанный выше набор данных также использовался при обучении модели квантильного случайного леса [17] – наиболее точного решения задачи прогноза фотометрического красного смещения галактик. Таким образом, можно провести корректное сравнение построенной модели с моделью квантильного случайного леса.

Используемый набор данных содержит малое количество объектов на больших красных смещениях, что приводит к ухудшению качества работы построенной модели для далеких объектов. Для решения этой проблемы было принято решение продублировать объекты на Z > 4.5 500 раз.

В качестве тестового набора данных использовался астрономический каталог Stripe 82X [18], содержащий 1164 рентгеновских квазара.

5.1.2. Оценка смеси нормальных распределений с помощью нейронных сетей.

Для многих задач машинного обучения можно сделать предположение, что распределение $p(y|x)$ – нормально. Однако, существуют примеры, когда истинное распределение целевой переменной значительно отличается от нормального, оно может быть многомодальным. В таких случаях предположение о нормальности распределения целевой переменной влечет низкую точность предсказаний. Решить эту проблему может помочь моделирование распределения целевой переменной смесью нормальных распределений [19].

При моделировании смеси нормальных распределений нейронной сетью для каждой из гауссиан предсказывается три параметра: π_i – вероятность, того что объект порожден данной гауссианой, μ_i – математическое ожидание данной гауссианы и σ_i^2 – дисперсия данной гауссианы.

Итоговое распределение выглядит следующим образом:

$$p(y|x) = \sum_{i=1}^{K} \pi_i(x,w)\mathcal{N}(y|\mu_i(x,w), \sigma_i^2(x,w))$$

Ошибка обучения:

$$E(w) = -\sum_{n=1}^{N} \ln\big(p(y_n|x_n)\big),$$

где N – число объектов обучающей выборки, w – веса нейронной сети, K – число гауссиан.

5.1.3. Архитектура модели прогноза фотометрического красного смещения галактик.

При решении задачи прогноза фотометрического красного смещения необходимо оценивать ошибку модели. С этой задачей справляются глубокие ансамбли, описанные в обзоре. Также необходимо отметить, что задача прогноза фотометрического красного смещения галактик является многомодальной.

Исходя из описанных выше замечаний было принято решение использовать глубокие ансамбли на основе глубоких полносвязных нейронных сетей, моделирующих смеси нормальных распределений.

Подбор параметров осуществлялся за счет использования двойной кросс-валидации.

В результате проведенных экспериментов подбирались следующие параметры модели:

- Число сетей в ансамбле: 5.
- Архитектура нейронной сети:
  - Число слоев: 6.
  - Число нейронов в слое: 400.
  - Число гауссиан в смеси: 5.
  - Использование batch-norm: да.
  - Использование dropout: нет.
- Процесс обучения:
  - Скорость обучения: $0.001 * 0.96^{\#\text{эпохи}}$.

- Batch size: 2^{13}.
- Weight decay: 0.001.

- Дублирование далеких объектов:
  - Порог по Z: 4.5.
  - Степень дублирования: 500.

5.2. Построение и оценка карт волокон галактик.

5.2.1. Описание наборов данных.

5.2.1.1. Астрономические координаты.

Для описания положения объекта на небесной сфере используются две угловые координаты:

- Прямое восхождение (right ascension; RA) - координата объекта на небесной сфере, равная дуге небесного экватора от точки весеннего равноденствия до круга склонения объекта. Измеряется в градусах [0, 360) или угловых часах [0, 24).
- Склонение (declination; DEC) - координата объекта на небесной сфере, равная угловому расстоянию от плоскости земного экватора до объекта. Измеряется в градусах [-90, 90]. Принимает положительные значения для объектов в северном полушарии и отрицательные для объектов в южном полушарии.

Совместно эти координаты образуют вторую экваториальную систему координат, широко используемую в астрономии.

5.2.1.2. Main Galaxy Sample SDSS DR12.

Слоуновский цифровой небесный обзор (Sloan Digital Sky Survey; SDSS) [20] - крупномасштабный проект по исследованию многоспектральных изображений и красных смещений звезд и галактик. Двенадцатая версия каталога (SDSS DR12) содержит значения координат RA, DEC и спектроскопического прогноза красного смещения для 2599191 галактик.

Для проведения экспериментов при решении задачи разработки метода построения карт волокон галактик использовалась подвыборка спектрального каталога SDSS DR12, называемая Main Galaxy Sample (MGS) [21] и содержащая 584449 объектов (рис. 8).

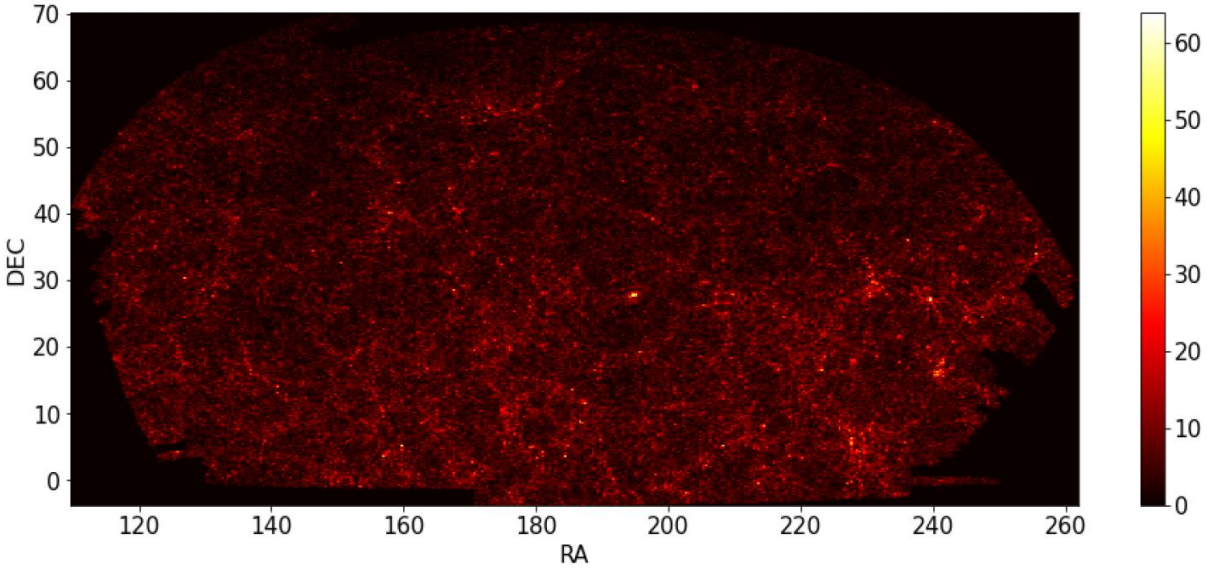

Рисунок 8. Тепловая карта галактик из каталога SDSS DR12 Main Galaxy Sample.

Эксперименты проводились в области сверхскопления Волос Вероника (Coma cluster). Также отбирались только галактики с точными значениями красного смещения. Эти детали постановки эксперимента брались из работ [22, 23].

- RA ограничено отрезком [130, 260]
- DEC ограничено отрезком [-10, 70]
- Z ограничено отрезком [0.005, 0.04]
- ZWARNING = 0
- ZCONFFINAL > 0.35

5.2.1.3. SDSS12 Optical Groups.

Данный каталог [24] содержит скопления галактик, построенные по галактикам из спектрального каталога SDSS DR12 в интервале Z [0; 0.2] с помощью алгоритма Friends-of-friends. Для проведения эксперимента были отобраны скопления, расположенные в области сверхскопления Волос Вероники. Также для обеспечения работы только с массивными (значимыми скоплениями) применялся фильтр по числу галактик в скоплении. Были отобраны только скопления, содержащие не менее 15 галактик, это соответствует третьему квартилю распределения числа галактик в скоплениях сверхскопления волос Вероники.

5.2.2. Метрики качества карт волокон галактик.

Для оценки качества волокон галактик, были разработаны метрики, основанные факте связности крупномасштабной структуры Вселенной. С одной стороны, необходимо, чтобы как можно большая доля построенных волокон соединяла скопления галактик. С другой, необходимо, чтобы как можно большая часть скоплений были соединены волокнами. Основываясь на этих соображениях, были разработаны аналоги метрик: полнота (recall) и точность (precision).

Сначала вычисляются метрики полнота и точность для реальных скоплений из выбранного каталога, $Recall_{true}$ и $Precision_{true}$ по формулам (1) и (2) соответственно.

$$Recall_{true} = \frac{N_{true\ cl\ inter}}{N_{true\ cl\ total}} \text{ (1)},$$

где $N_{true\ cl\ total}$ – число реальных скоплений, пересеченных волокнами, а $N_{true\ cl\ total}$ – общее число реальных скоплений.

$$Precision_{true} = \frac{N_{true\ fil\ inter}}{N_{total\ fil}} \text{ (2)},$$

где $N_{true\ fil\ inter}$ – число найденных волокон, пересеченных реальными скоплениями, $N_{total\ fil}$ – общее число найденных волокон.

Далее генерируются несколько выборок случайных (ложных) скоплений внутри выпуклой оболочки галактик такого же размера, что и каталог реальных скоплений. Для них считаются метрики $Recal_{false}$ и $Precision_{false}$ по формулам (3) и (4) соответственно. Далее полученные метрики усредняются по выборкам ложных скоплений.

$$Recall_{false} = \frac{1}{M}\sum_{i=1}^{M}\frac{N_{i,false\ cl\ inter}}{N_{i,false\ total\ cl}} \text{ (3)},$$

где $N_{i,false\ total\ cl}$ – число найденных ложных скоплений i -ой случайной выборки, пересеченных волокнам.

$$Precision_{false} = \frac{1}{M}\sum_{i=1}^{M}\frac{N_{i,false\ fil\ inter}}{N_{total\ fil}} \text{ (4)},$$

где $N_{i,false\ fil\ inter}$ – число найденных волокон, пересеченных ложными скоплениями из i-ой случайной выборки.

После этого происходит вычисление окончательных метрик $Recall_{diff}$ и $Precision_{diff}$ по формулам (5)и (6)соответственно.

$$Recall_{diff} = Recall_{true} - Recall_{false} \text{ (5)}.$$

$$Precision_{diff} = Precision_{true} - Precision_{false} \text{ (6)}.$$

Вычисление метрик для ложных скоплений штрафует карту волокон за наличие в ней большого количества ложных скоплений.

Для агрегации точности и полноты можно использовать F1-меру (7).

$$F1_{diff} = 2\frac{Recall_{diff} * Precison_{diff}}{Recall_{diff} + Precision_{diff}} (7).$$

5.3. Модели оценки вероятности нахождения скопления галактик с известным положением в трехмерном пространстве.

5.3.1. Описание наборов данных.

5.3.1.1. ACT DR5 Cluster catalog.

Для построения моделей оценки вероятности нахождения скоплений галактик в некоторой точке трехмерного пространства использовался каталог скоплений ACT DR5 SZ Cluster Catalog [25]. Поиск скоплений, содержащихся в данном каталоге, производился радиотелескопом в микроволновом диапазоне на основании эффекта Сюняева – Зельдовича. Инструмент - космологический телескоп Атакама (Atacama Cosmology Telescope) [5].

Каталог содержится содержит координаты на небесной сфере (прямое восхождение и склонение) и расстояние (красное смещение) для 1648 объектов. Выбор участка небесной 22 сферы для исследования связан с чувствительностью телескопа (рис. 9) и пересечением с областью используемого каталога галактик. А выбор промежутка по красному смещению обусловлен неравномерным распределением галактик по координате прямого восхождения в выбранном каталоге галактик на больших значениях красного смещения.

- RA ограничено отрезком [160, 240]
- DEC ограничено отрезком [0, 20]
- Z ограничено отрезком [0.0, 0.6]

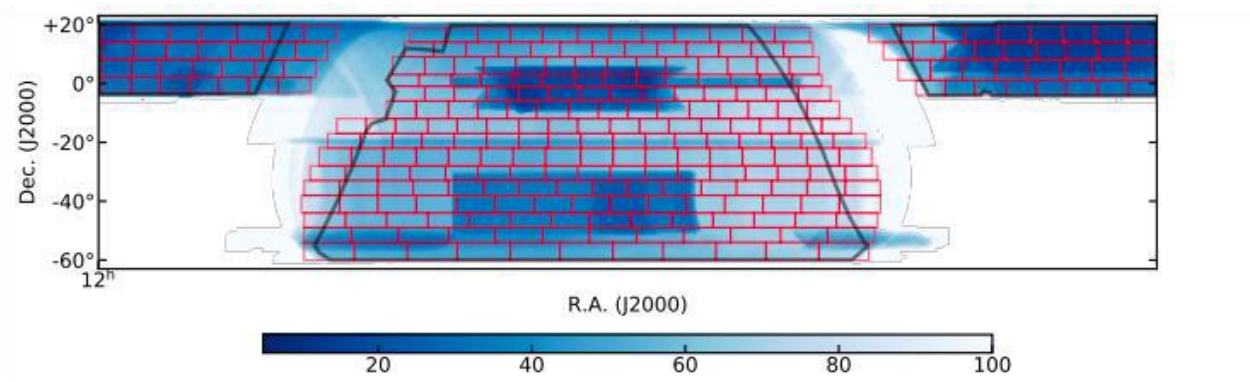

Рисунок 9. Область покрытия и чувствительность телескопа ACT. Градация синего – величина шума.

5.3.1.2. SDSS DR16.

В качестве каталога галактик был выбран слоуновский цифровой небесный обзор шестнадцатой версии (Sloan Digital Sky Survey; SDSS DR16). Каталог содержится содержит координаты на небесной сфере (прямое восхождение и склонение) и расстояние (красное смещение) для 3234563 галактик. Интервалы по координатам RA и DEC выбраны таким образом, чтобы обеспечить обрамление выбранной подобласти каталога скоплений галактик на 20 градусов по координате RA и на 10 градусов по координате DEC. Верхнее ограничение на красное смещение выбрано из-за неравномерного распределения галактик по координате прямое восхождение для галактик на большом красном смещении (более 0.7) (рис. 10 и рис. 11).

- RA ограничено отрезком [140, 260]
- DEC ограничено отрезком [-10, 30]
- Z ограничено отрезком [0.0, 0.6]

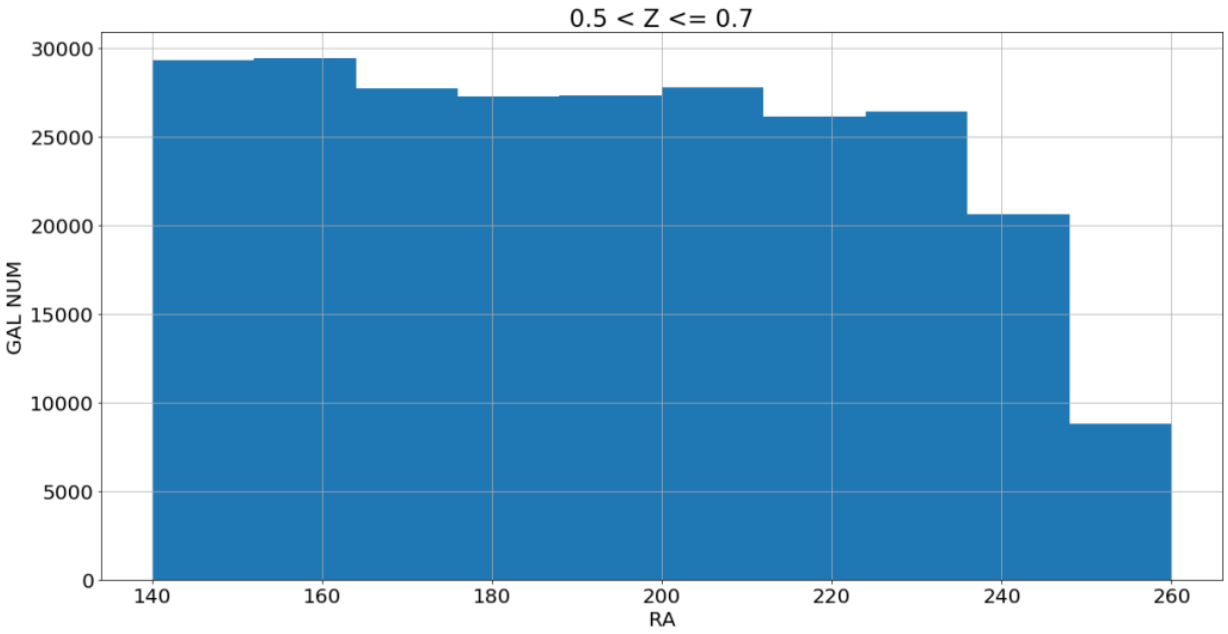

Рисунок 10. Распределение галактик по координате RA для интервала Z (0.5; 0.7].

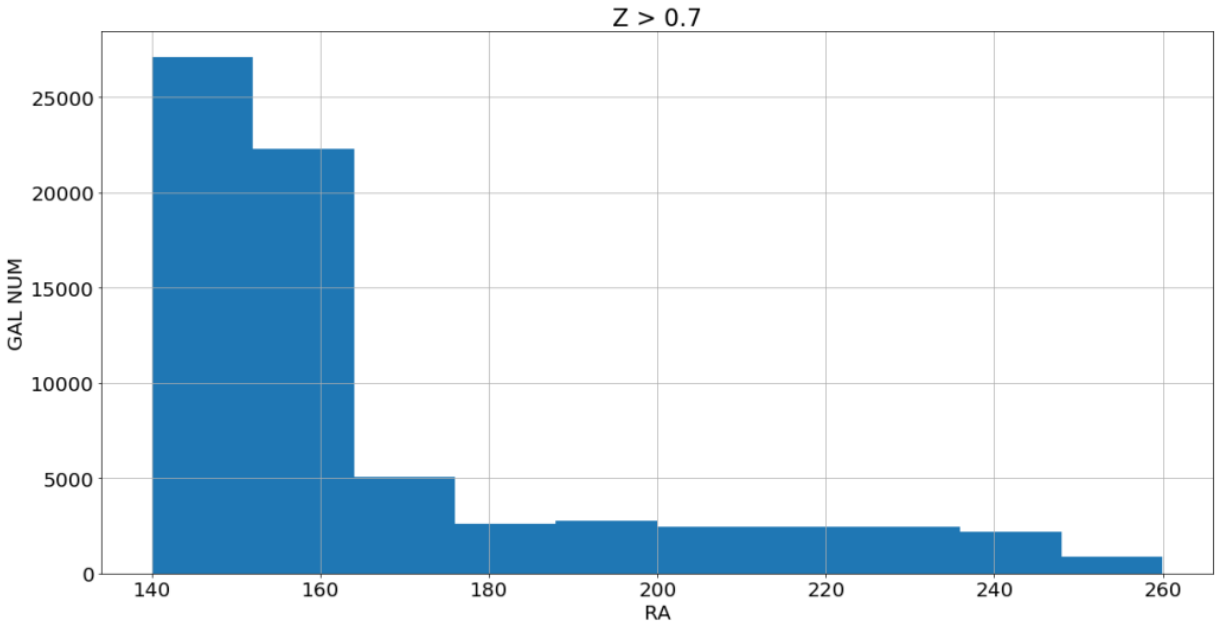

Рисунок 11. Распределение галактик по координате Z для интервала Z (0.7; 1.0].

5.3.2. Базовая модель оценки вероятности скоплений галактик с известным положением в трехмерном пространстве.

Основная идея оценки вероятности нахождения скопления галактик в некоторой точке трехмерного пространства на основании глобальной карты волокон галактик крупномасштабной структуры Вселенной заключается в использовании информации близлежащих к рассматриваемому скоплению волокнах, расстоянии до них и их значимости.

Для построения карт волокон используется алгоритм DisPerSe. Сигма, один из параметров алгоритма, отвечает за значимость волокон, возвращаемых алгоритмом. На рис. 12 можно видеть примеры карт волокон, построенных на небольшом участке небесной сферы алгоритмом DisPerSe со значениями параметра сигма 3 (слева) и 5 (справа).

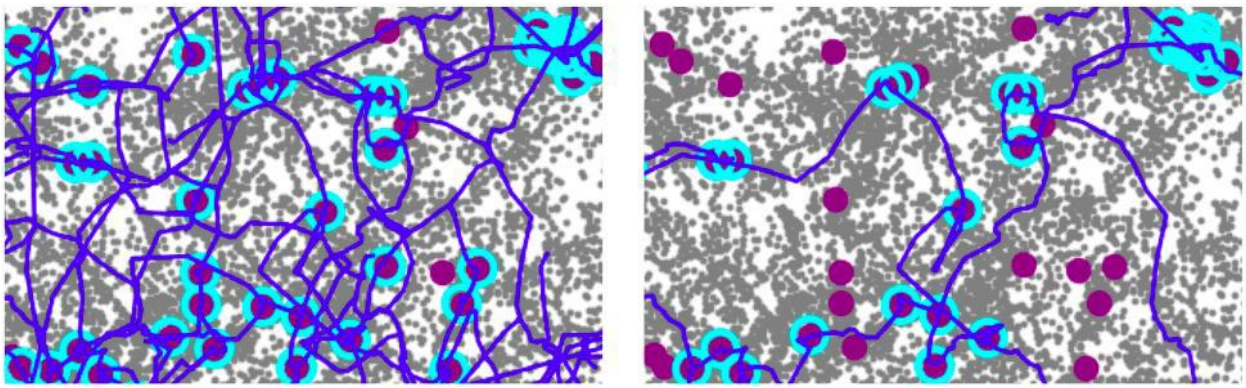

Рисунок 12. Карты волокон, соответствующие значениям сигма: 3 (слева) и 5 (справа).

Как видно на карте, соответствующей значению параметра сигма 5, выделено меньше волокон галактик, но они обладают большей значимостью (плотностью галактик).

Для построения базовой модели оценки вероятности скопления с заданной точке пространства на исследуемой области Вселенной с помощью алгоритма DisPerSe строится набор карт волокон для различных значений параметра сигма (от 0.2 до 14.0 с шагом 0.2).

Далее для исследуемой точки пространства рассматривается окрестность некоторого радиуса. В качестве меры уверенности нахождения скопления галактик в рассматриваемой точке берется максимальное значение параметра сигма, такое что соответствующая карта волокон содержит волокно, пересекающее рассматриваемую окрестность данной точки.

Радиус окрестности является свободным параметром модели и нуждается в оптимизации.

5.3.3. Машинно-обученные модели оценки вероятности скоплений галактик с известным положением в трехмерном пространстве.

5.3.3.1. Генерация примеров обучения.

Примерами для обучения служат некоторые точки в исследуемой области Вселенной, для которых необходимо оценить вероятность нахождения там скопления галактик.

В качестве положительных примеров были взяты точки расположения истинных скоплений галактик из каталога ACT DR5 SZ Cluster Catalog (400 объектов).

В качестве отрицательных примеров была сгенерирована выборка случайных (ложных) скоплений галактик в исследуемой области (4119 объектов). Было исследовано 2 способа генерации ложных скоплений:

- Вводится только ограничение на распределение ложных скоплений по красному смещению. Оно должно совпадать с распределением истинных скоплений.
- Помимо требования к распределению ложных скоплений по красному смещению вводится требование к их расположению относительно волокон. Ложные скопления привязываются к галактикам из используемого каталога. Такой подход обеспечивает большую вероятность близости ложных скоплений к волокнам. Предполагается, что такая генерация породит более сложные для классификации примеры, что формально уменьшит метрики качества модели, но заставит модель находит большее сложные зависимости и уменьшит переобучение. Также для обоих методов

генерации вводится требование на расстояние между истинными и ложными скоплениями. Они должно превышать 10 Мрс для любой пары скоплений различных классов.

5.3.3.2. Генерация признаков.

Набор признаков состоит из четырех групп:

- Признаки, основанные на построенных картах волокон. Так же, как и в случае базовой модели, строится с помощью алгоритма DisPerSe строится набор карт волокон для различных значений параметра сигма (от 0.2 до 14.0 с шагом 0.2). Далее для рассматриваемой точки пространства рассчитываются расстояния до ближайших волокон. Получается группа из 70 признаков. На рис. 13 представлено распределение расстояний до волокон различной значимости (соответствующих различным значениям параметра сигма) для положительных (синие точки) и отрицательных (красные кресты) примеров.

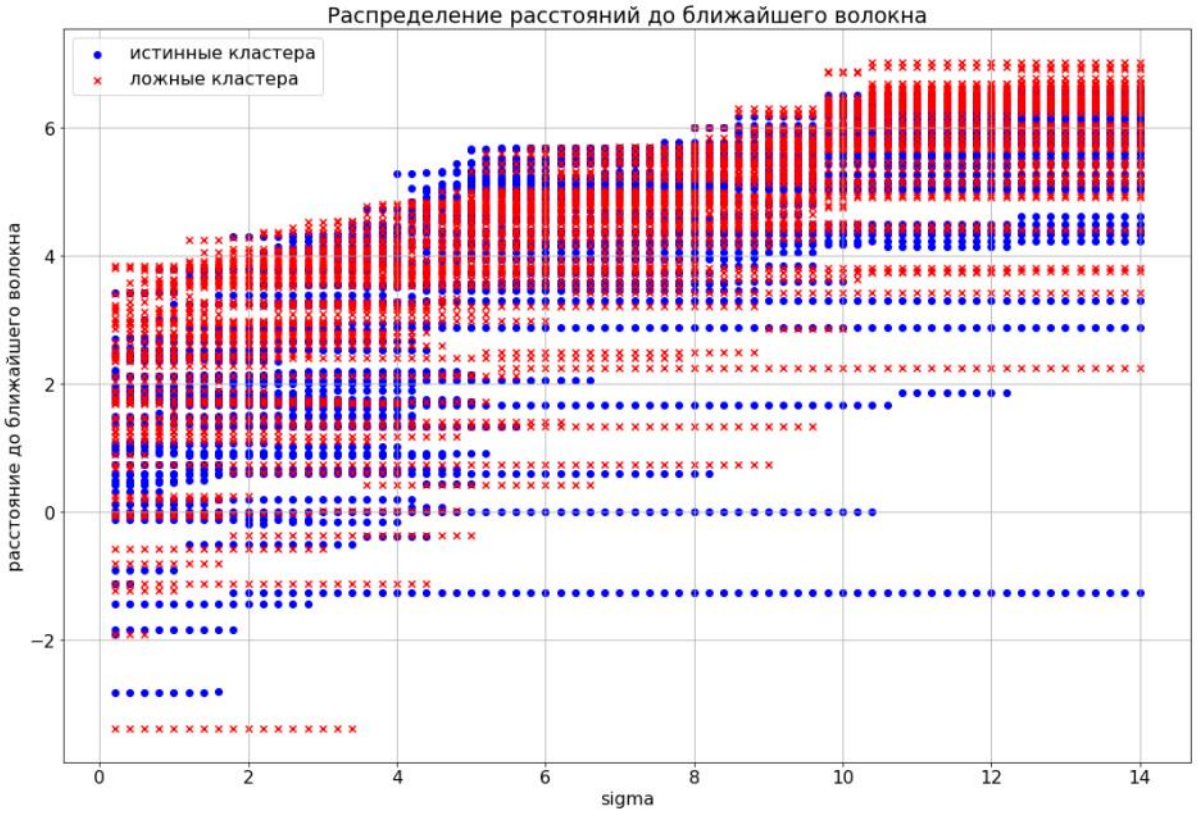

Рисунок 13. Распределение до волокон различной значимости для положительных (синие точки) и отрицательных (красные кресты) примеров. По горизонтальной оси отложены значения сигма (значимость), с которым были построены волокна.

- Признаки, основанные на локальном распределении галактик вокруг скопления. Данный набор признаков был взят из работы [11].
    - Расстояние до центра масс галактик в радиусе R_{cme}.
    - Момент инерции частиц в радиусе R_{cme}.
    - Число частиц в радиусе R_{cme}.
    - $R_{cme} \in [0.2, 0.5,\ 0.8, 1.0, 1.5, 2.0, 3.0, 5.0, 7.5, 10.0, 12.0]\ Mpc$.
    - Расстояние до k−го ближайшего (в трехмерном пространстве) соседа.
    - $k \in [1, 2, 3, 5, 8, 10, 15, 20, 30, 35, 40]$.
    - Разность максимальной и минимальной долей объясненной дисперсии, полученных в результате применения метода главных компонент [24] на точках в радиусе R_{pca}.
    - $R_{pca} \in [1.5, 1.6, 1.7, 1.8, 1.9, 2.0]\ Mpc$.
- Красное смещение рассматриваемой точки. Необходимо для учета моделью различной плотности галактик в каталоге на различном красном смещении.
- Разность в красном смещении с k-м ближайшим соседом на небесной сфере. $k \in [1, 2, 3, 4, 5, 6, 7]$.

5.3.3.3. Модели классификации.

В качестве моделей классификации использовались:

- Случайный лес. Алгоритм машинного обучения, решающий задачи классификации, регрессии и кластеризации. Представляет из себя ансамбль решающих деревьев. Сочетает в себе две идеи: метод бэггинга [26] и метод случайных подпространств [27]. Гиперпараметры используемой модели: число деревьев – 600, максимальная глубина деревьев – 13, доля используемых признаков – корень квадратный из общего числа.
- Градиентный бустинг на деревьях [28]. Алгоритм машинного обучения, решающий задачи классификации и регрессии. Представляет из себя ансамбль решающих деревьев, построенных методом бустинга [29]. Гиперпараметры используемой модели: число деревьев – 200, максимальная глубина деревьев – 2, доля используемых признаков – корень квадратный из общего числа, скорость обучения – 0.05.

5.3.4. Модель оценки вероятности нахождения скопления галактик в заданном направлении и прогноза его красного смещения.

Дается не точка в трехмерном пространстве, а только угловые координаты на небесной сфере. Необходимо оценить вероятность того, что в этом направлении находится скопление галактик, и спрогнозировать его красное смещение.

Далее модели оценки вероятности нахождения скопления с заданной точке будут называться моделями первого уровня, а в заданном направлении моделями второго уровня.

5.3.4.1. Генерация признаков.

Вдоль заданного направления проводится луч. Далее по заданному лучу с некоторым шагом по красному смещению выставляются точки. В этих точках с помощью одной из моделей первого уровня оценивается вероятность того, что там находится скопление. Получает график зависимости вероятности нахождения скопления галактик от удаленности точки на луче от Земли (рис. 14).

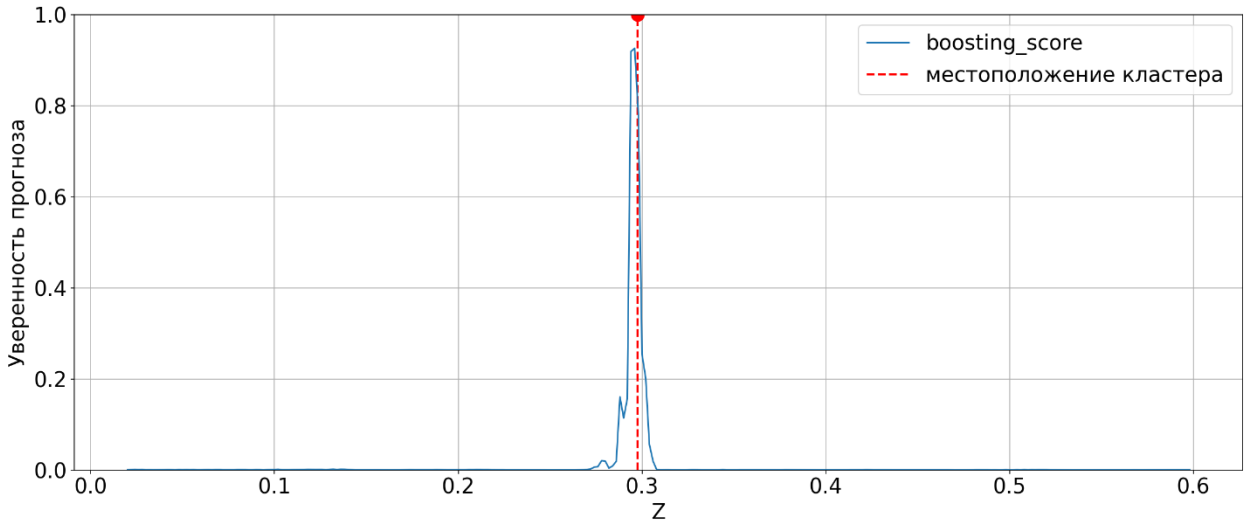

Рисунок 14. Кривая, показывающая, вероятность нахождения скопления в точках некоторого луча на разной удаленности от Земли. График построен для реального скопления. Красная вертикальная линия показывает истинной положение скопления.

5.3.4.2. Базовая модель оценки вероятности скопления галактик в заданном направлении и прогноза его красного смещения.

Была предложена базовая модель второго уровня. В качестве уверенности прогноза для объекта берется величина максимального пика на кривой, описывающей данный объект. В качестве прогноза красного смещения берётся положение максимального пика.

5.3.5. Общая схема решения задачи отождествления скоплений галактик и прогноза их красных смещений.

На рис. 15 представлена общая схема построенного решения задачи отождествления скоплений галактик и прогноза их красных смещений.

- На основании каталога SDSS DR17 с помощью алгоритма DisPerSe строится набор карт волокон галактик для различных значений параметра сигма.
- Генерация отрицательных примеров обучения.
- Далее на основании построенных карт волокон и каталога SDSS DR17 генерируются признаки для модели первого уровня.
- Обучение модели первого уровня.
- Построение кривых, описывающих примеры, для использования модели второго уровня с помощью модели первого уровня.
- Оценка вероятности нахождения скопления и прогноза его красного смещения с помощью модели второго уровня.

Рисунок 15. Общая схема решения задачи отождествления скоплений галактик и прогноза их красных смещений.

5.4. Общая схема решения задачи построения карты крупномасштабной структуры Вселенной.

На рис. 16 представлена общая схема решения задачи построения карты крупномасштабной структуры Вселенной.

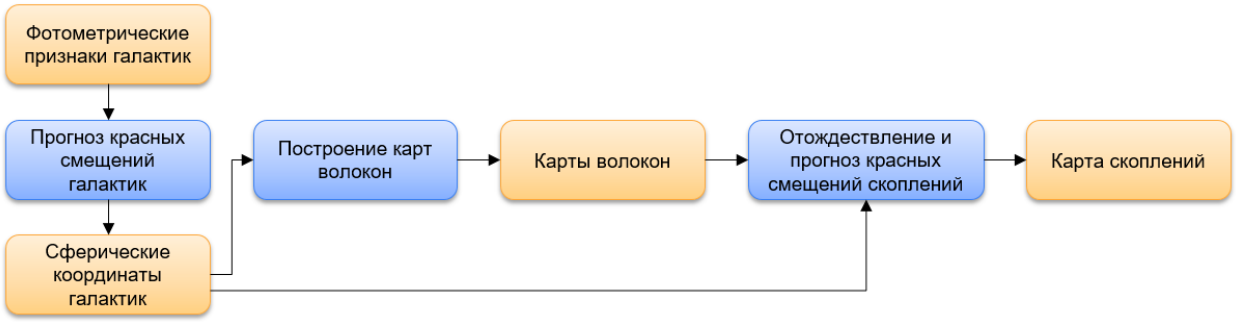

Рисунок 16. Общая схема решения задачи построения карты крупномасштабной структуры Вселенной.

- На основании фотометрических признаков галактик вычисляются прогнозы их красных смещений. Имея прогнозы, получаем положения галактик в пространстве.
- По координатам галактик строятся карты волокон галактик.
- По картам волокон галактик и координатам галактик вычисляются признаки для модели отождествления скоплений галактик и прогноза их красных смещений.
- Имея признаки, можем оценить вероятности нахождения скопления и сделать прогноз его красного смещения.

6. Описание практической части.

6.1. Программная реализация.

6.1.1. Библиотека для решения задачи построения и анализа крупномасштабной структуры Вселенной.

В результате проделанной работы была реализована библиотека для решения задачи построения и анализа крупномасштабной структуры Вселенной. Реализация выполнялась средствами языка программирования Python3, ряда open source библиотек и реализации алгоритма DisPerSe на языке программирования C++ [30].

Набор используемых open source библиотек:

- Pandas v1.2.4: работа с табличными данными.
- NumPy v1.20.1: векторные вычисления.
- AstroPy v4.2.1: работа с астрономическими данными.
- Matplotlib v3.3.4: визуализация.
- Scikit-learn v0.24.1: алгоритмы поиска ближайших соседей, работа с обучающими выборками, алгоритмы машинного обучения.
- Lightgbm v3.2.1: реализация градиентного бустинга [31].
- PyTorch v1.13.1: проектирование и обучение нейронных сетей.

На рис. 17 представлена общая схема разработанной библиотеки.

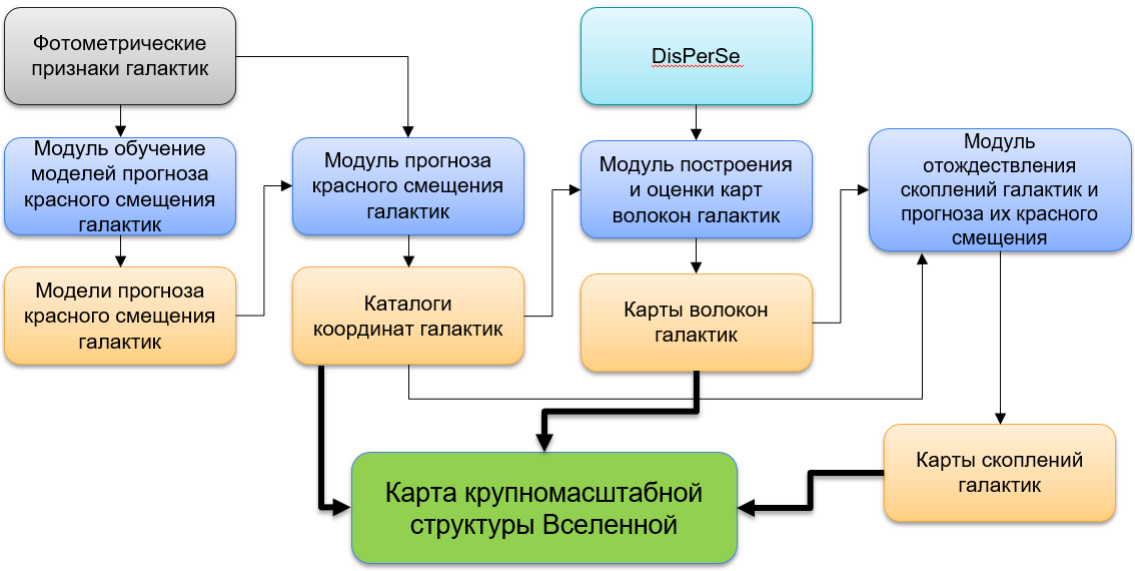

Рисунок 17. Общая схема разработанной библиотеки.

Программный код, обученные модели и необходимые для запуска данные доступны по ссылке: https://disk.yandex.ru/d/iEExSYWFpc3yCg. Общий объём кода составил порядка 2000 строк.

6.1.2. Модуль обучения моделей прогноза красного смещения галактик.

Модуль обучения моделей прогноза красного смещения галактик содержит классы, реализующие модели глубокого ансамбля и нейросетевой модели для оценки смеси нормальных распределений на основе архитектуры полносвязного перцептрона.

Разработанные классы позволяют настраивать архитектуру обучаемой модели:

- Число сетей в ансамбле.
- Число слоев и нейронов в сети.
- Параметры обучения.
- Параметры дублирования далеких объектов.

Модуль предоставляет возможности для визуализации процесса обучения и расчета метрик.

В качестве входных данных модуль принимает фотометрические признаки галактик. Модуль возвращает обученные модели.

6.1.3. Модуль прогноза красного смещения галактик.

Модуль прогноза красного смещения галактик содержит методы для выполнения прогноза. Модуль позволяет осуществлять прогноз несколькими моделями и отбирать для каждого объекта лучший из имеющихся прогнозов.

Также модуль позволяет оценить для каждого объекта меры неопределенности прогноза и эмпирическое распределение прогноза.

В качестве входных данных модель принимает фотометрические признаки галактик и обученные модели. Модуль возвращает прогнозы красных смещений галактик.

6.1.4. Модуль построения и оценки волокон галактик.

Модуль построения и оценки карт волокон галактик содержит класс, использующий реализацию алгоритма DisPerSe [30], позволяющий:

- Строить карты волокон галактик в двумерном и трехмерном пространствах.
- Вычислять значения предложенных метрик качества карт волокон галактик.
- Визуализировать построенные карты волокон.
- Сериализовывать/десериализовывать построенные карты волокон.

В качестве входных данных модель принимает каталоги галактик с известными декартовыми или сферическими координатами. Модуль возвращает карты волокон галактик в формате JSON.

6.1.5. Модуль отождествления скоплений галактик и прогноза их красных смещений.

Модуль отождествления скоплений галактик и прогноза их красного смещений содержит методы:

- Генерации обучающих примеров.
- Генерации признаков.
- Построения моделей первого уровня. Реализованы функции разделения обучающих данных для выполнения процедуры скользящего контроля, обучения моделей первого уровня, оценки построенных моделей за счет вычисления метрики ROC-AUC.
- Генерации признаков для модели второго уровня.
- Применения и оценки базовой модели второго уровня. Выполняется применение базовой модели второго уровня и ее оценка за счет вычисления метрик: $n_{>0.15}$, σ_{NMAD}, ROC-AUC.

6.2. Экспериментальные исследования.

6.2.1. Модель прогноза фотометрического красного смещения галактик.

В табл. 1 приведены значения метрик обученной нейросетевой модели прогноза фотометрических красных смещений галактик, оценивающей смесь нормальных распределений, и модели на основе квантильного случайного леса из работы [17], полученные на обучающей выборке с помощью двойной перекрестной валидации.

- Вся выборка.
- Далекие объекты на красном смещении Z > 5.

Таблица 1. Метрики нейросетевой модели и модели на основе квантильного случайного, полученные с помощью двойной перекрестной валидации.

Модель	$n_{>0.15}$	σ_{NMAD}	$n_{>0.15}\ (Z > 5)$	$\sigma_{NMAD}\ (Z > 5)$
Нейронная сеть	0.036	0.026	0.061	0.021
Случайный лес	0.048	0.028	0.129	0.017

На рис. 18 и рис. 19 представлены значения метрик $n_{>0.15}$ и σ_{NMAD}, соответственно, для моделей:

- Оценка смеси нормальных распределений (5 гауссиан).
- Квантильный случайный лес.
- Оценка смеси нормальных распределений (1 гауссиана).
- Нейросетевая модель, обученная на MSE.

При использовании нейронной сети удается добиться значительного уменьшения числа выбросов. Особенно данный эффект заметен для далеких объектов. Средняя же точность предсказаний примерно равна у двух моделей. Это объясняется тем, что обученный случайный лес сильно привязывается к примерам обучающей выборки.

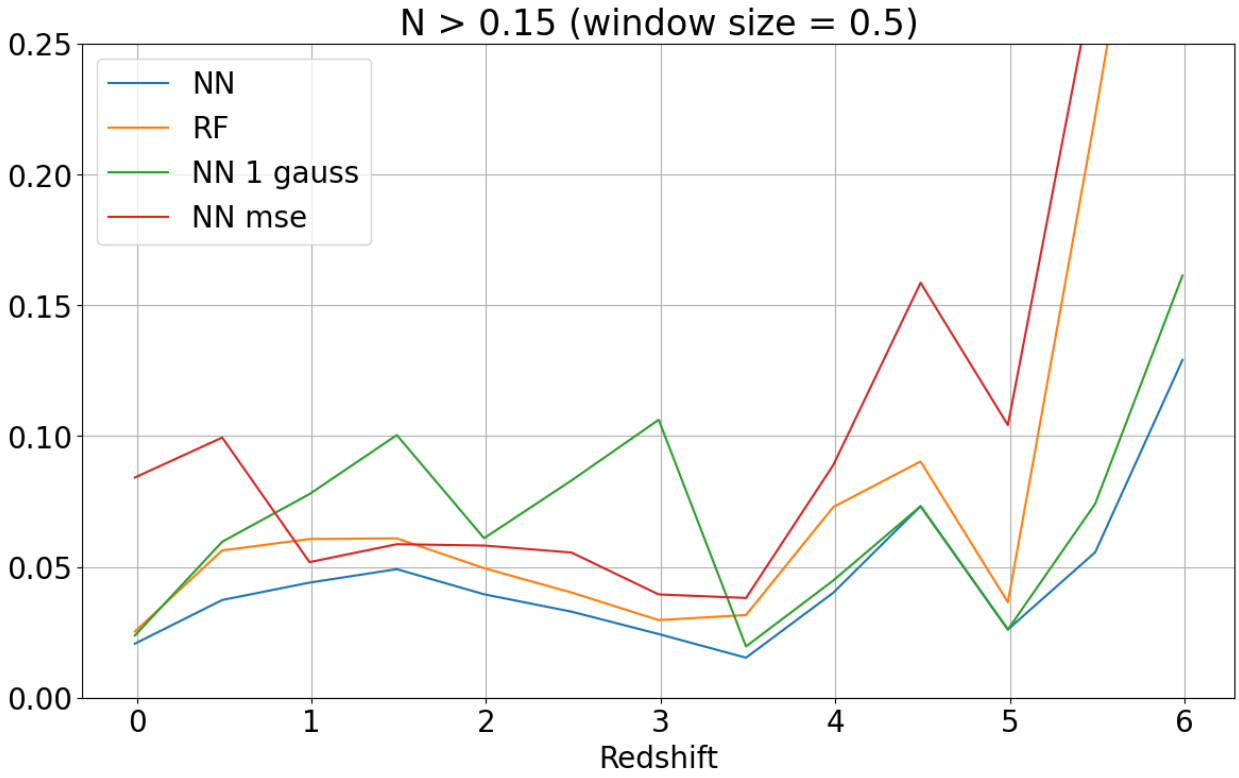

Рисунок 18. $n_{>0.15}$ в разрезах по красному смещению.

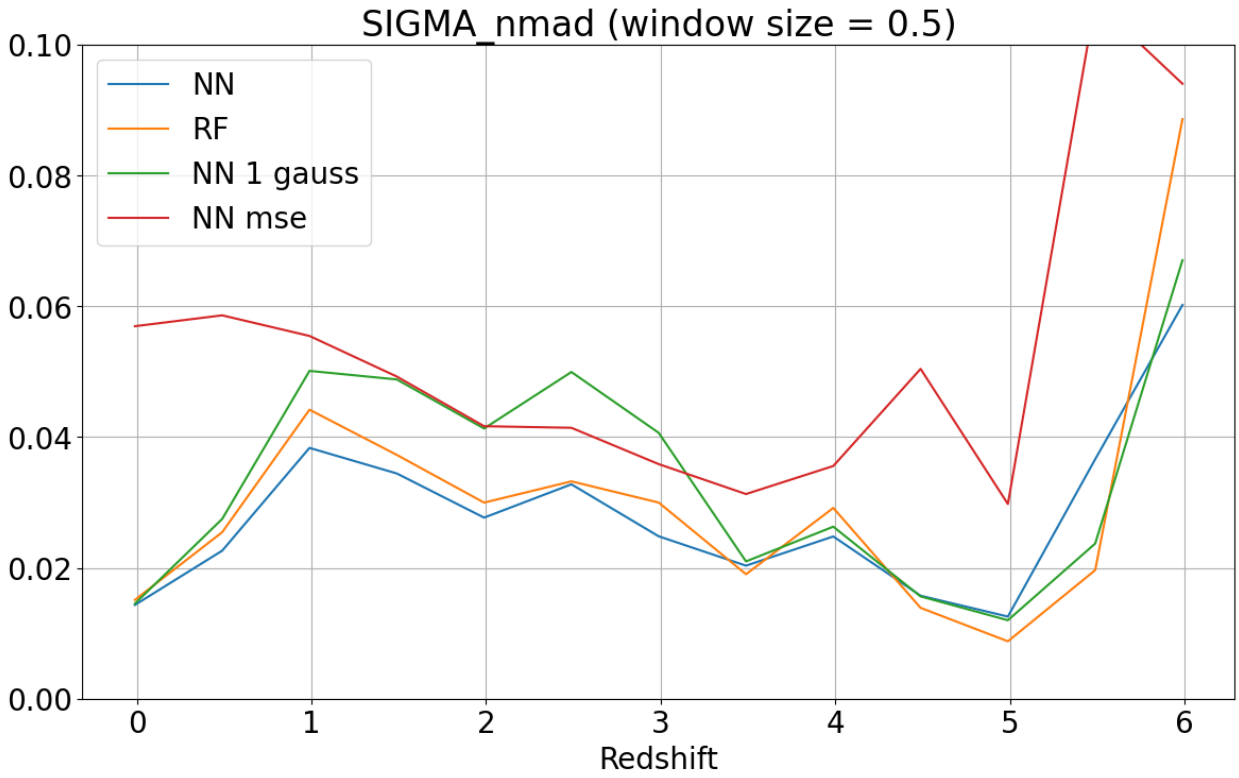

Рисунок 19. σ_{NMAD} в разрезах по красному смещению.

При оценке на наборе данных Stripe 82X обученная модель сравнивалась с:

- Моделью на основе алгоритма случайного леса [17].
- Модель основанной на аппроксимации шаблонов спектров галактик [32].
- Многослойным перцептроном, обученным на MSE [33].

В табл. 2 представлены значения метрик описанных моделей.

Таблица 2. Метрики на наборе данных Stripe 82X.

	$n_{>0.15}$	σ_{nmad}
Нейронная сеть (GMM)	**0.043**	**0.029**
Случайный лес	0.070	0.029
Аппроксимация шаблонов	0.180	0.069
Нейронная сеть (MSE)	0.173	0.065

На рис. 20 представлены диаграммы рассеивания предсказаний описанных моделей.

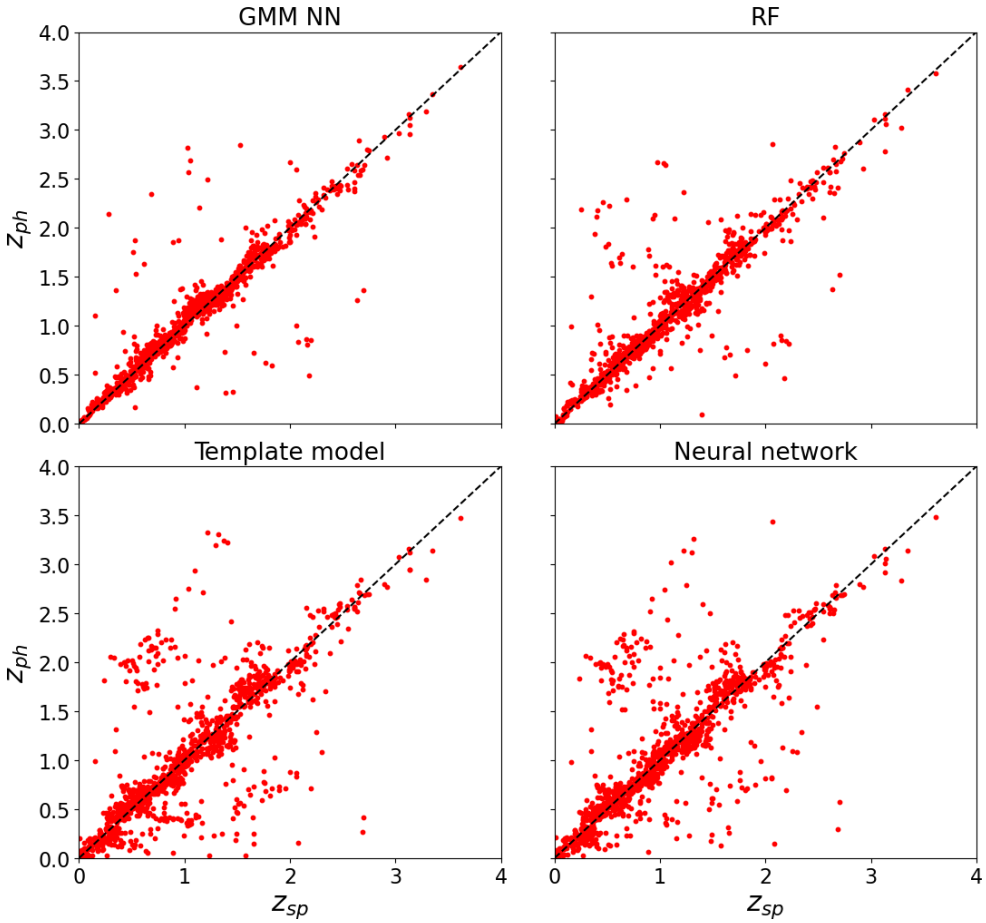

Рисунок 20. Диаграммы рассеивания на наборе данных Stripe 82X.

Обученная модель имеет на 0.8% меньше выбросов на всей валидационной выборке галактик и на 6.5% меньше выбросов среди далеких объектов (Z > 5), чем модель на основе квантильного случайного леса.

Обученная модель показывает наилучшую точность среди имеющихся на текущий момент решений на тестовом наборе данных Stripe 82X, в том числе представленные в литературе нейросетевых решений, и имеет на 2.74% меньше выбросов, чем модель на основе квантильного случайного леса – основного подхода для прогноза расстояния до рентгеновских квазаров из каталогов eRosita.

6.2.2. Метрики качества карт волокон галактик.

При проведении эксперимента оценивали качество карт волокон галактик из области сверхскопления Волос Вероники, построенных с использованием алгоритма DisPerSe с различными значениями параметров.

Настраивались 2 параметра алгоритма DisPerSe:

- smooth - число сглаживания оценки плотности распределения галактик с помощью алгоритма DTFE.
- sigma - порог фильтрации волокон с низкой устойчивостью.

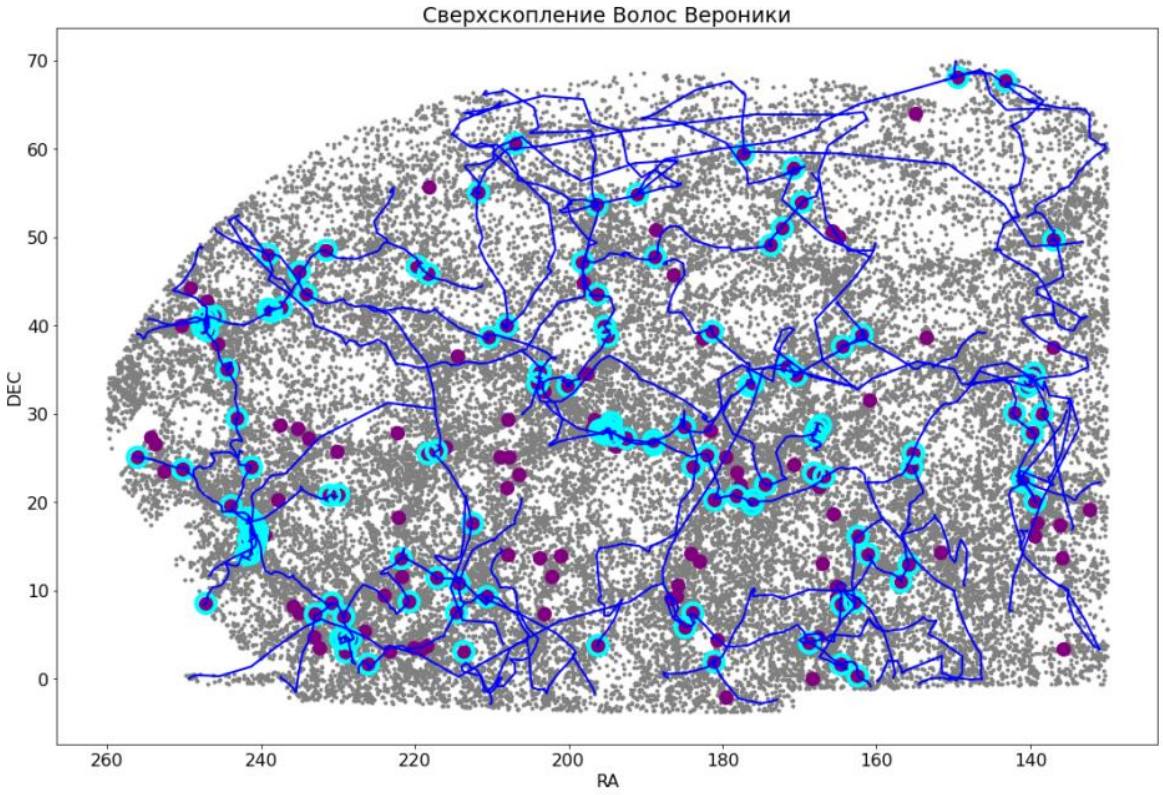

Рисунок 21. Карта волокон для области сверхскопления Волос Вероники.

На рис. 21 представлен пример карты волокон, построенный алгоритмом DisPerSe с параметрами: sigma = 5 и smooth = 2. Синими кривыми отображены найденные волокна галактик. Серые точки - галактики. Фиолетовые точки - скопления из каталога скоплений SDSS12 Optical Groups. Если фиолетовый круг обрамлен голубой окружность, то соответствующее скопление пересечено волокнами галактик из построенной карты волокон.

Увеличение значений обоих параметров приводит к тому, что из карты удаляются волокна, обладающие низкой значимостью, то есть тех, плотность галактик в которых слабо отличается от средней плотности Вселенной.

Эти параметры нуждаются в настройке, ведь при их низких значениях карты будут содержать шумовые волокна, а при высоких из карт будут удалены некоторые реальные волокна, обладающие относительно низкой значимостью.

На рис. 22 и рис. 23 показаны графики зависимости метрик $F1_{diff}$ от заданного радиуса волокон галактик.

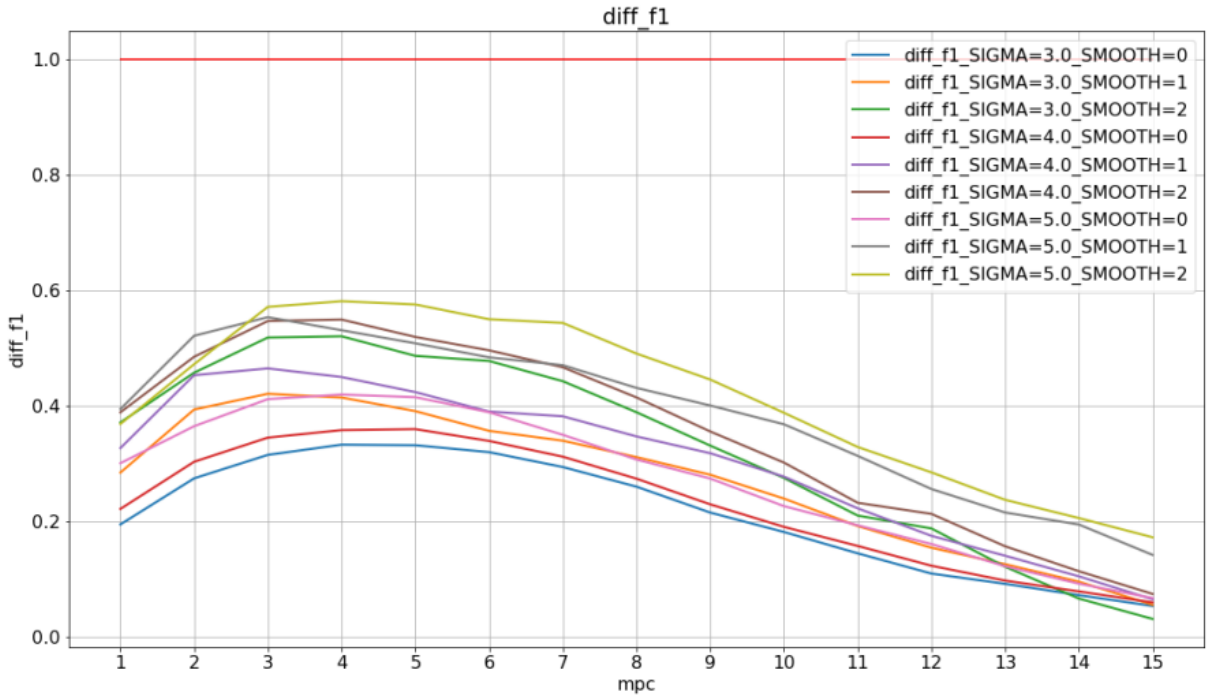

Рисунок 22. Значения метрики F1 для карт волокон, построенных в области сверхскопления Волос Вероники с помощью алгоритма DisPerSe. Sigma: 3.0, 4.0, 5.0. Smooth: 1, 2, 3.

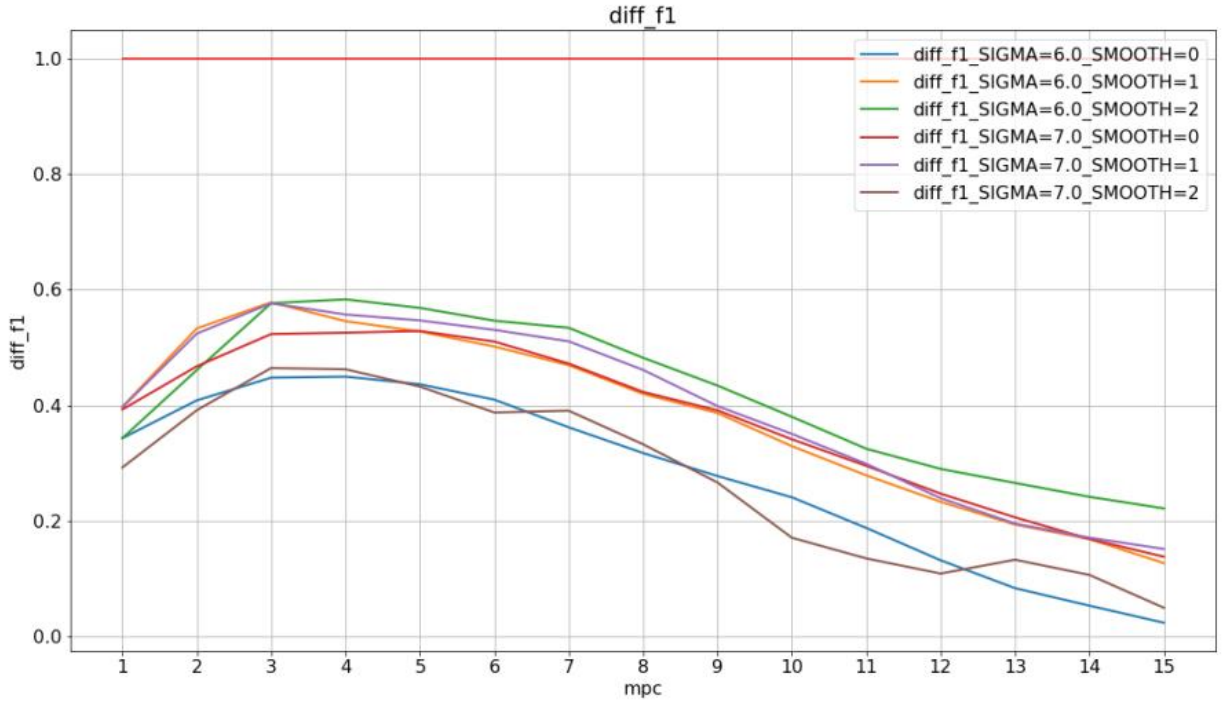

Рисунок 23. Значения метрики F1 для карт волокон, построенных в области сверхскопления Волос Вероники с помощью алгоритма DisPerSe. Sigma: 6.0, 7.0. Smooth: 1, 2, 3.

Алгоритм DisPerSe оценивает волокна, как одномерные кривые, не вычисляя их радиус. Но для определения пересечений скоплений с волокнами необходимо знать значение радиуса. Поэтому производится расчет метрик для различных значений радиуса построенных волокон галактик. По пику значений метрик на графиках можно оценить средний радиус скоплений. В нашем случае он равен 3-4 Мрс.

В табл. 1 приведены значения метрики $F1diff$ для радиуса волокон галактик равного 4 Мрс.

Таблица 3. Значения исследуемых метрик для карт волокон галактик, построенных с помощью алгоритма DisPerSe в области сверхскопления Волос Вероники. Sigma: 3.0, 4.0, 5.0, 6.0, 7.0. Smooth: 0, 1, 2.

Sigma	Smooth	$Recall_{true}$	$Precision_{true}$	$Recall_{diff}$	$Precision_{diff}$	$F1_{diff}$
3	0	**0.9864**	0.2548	0.7360	0.2149	0.3327
	1	0.9099	0.3601	0.7432	0.2872	0.4143
	2	0.7612	0.5210	0.6621	0.4286	0.5204
4	0	**0.9649**	0.2828	0.7333	0.2367	0.3579
	1	0.8783	0.4077	0.7405	0.3230	0.4498
	2	0.7387	0.5723	0.6550	0.4729	0.5492
5	0	0.9189	0.3509	**0.7666**	0.2887	0.4195

	1	0.7837	0.5437	0.6829	0.4341	0.5307
	2	0.6351	0.6746	0.5837	0.5783	**0.5810**
6	0	0.8378	0.4107	0.7207	0.3269	0.4498
	1	0.6891	0.6415	0.6180	0.4887	0.5458
	2	0.5360	**0.7954**	0.5099	**0.6818**	**0.5835**
7	0	0.7567	0.5478	0.6594	0.4372	0.5258
	1	0.5585	0.7758	0.5126	0.6103	0.5572
	2	0.4054	**0.8127**	0.3838	0.5818	0.4624

Из табл. 1 можно следует, что:

- Как и предполагалось, увеличение значений обоих параметров ведет к уменьшению значения метрики $Recall_{true}$ и увеличению значения метрики $Precision_{true}$.
- Метрики $Recall_{diff}$ и $Precison_{diff}$ ведут себя менее предсказуемо из-за штрафов за пересечение с ложными скоплениями, однако тренды все равно заметны.
- Лучшими в смысле метрики $F1_{diff}$ оказались значения параметров sigma=5, smooth=2 и sigma=6, smooth=2.

6.2.3. Метрики качества модели отождествления скоплений галактик и прогноза их красного смещения.

В табл. 4 приведены значения метрик: $n_{>0.15}$, σ_{NMAD} и ROC-AUC для моделей отождествления скоплений галактик и прогноза их красного смещения.

Таблица 4. Метрики моделей отождествления скоплений галактик и прогноза их красного смещения.

Модель первого уровня	ROC-AUC модели первого уровня	ROC-AUC модели второго уровня	$n_{>0.15}$	σ_{NMAD}
Случайный лес	0.978	0.842	0.065	0.0012
Град. бустинг	0.986	0.884	0.017	0.0011

Из табл. 4 видно, что модель градиентного бустинга превосходит модель случайного леса и даёт качество, достаточное для применения модели к реальным данным.

7. Заключение.

В результате проделанной работы была исследована и разработана технология для построения и анализа крупномасштабной структуры Вселенной. Разработанная технология позволяет решать задачи, возникающие на разных этапах построения карты крупномасштабной структуры Вселенной: прогноз красных смещений галактик, построение карт волокон галактик, отождествление и прогноз красных смещений скоплений галактик.

- Была исследована и разработана модель прогноза фотометрических красных смещений галактик. В основе модели лежат глубокие ансамбли на базе нейронных сетей, оценивающих смеси нормальных распределений. Метрики качества и сравнение с моделью на основе квантильного случайного леса – основного алгоритма прогноза фотометрических красных смещений рентгеновских источников каталогов eRosita:
    - Валидационная выборка. σ_{NMAD} : 0.026, $n_{>0.15}$: 0.036 (на 0.82% меньше выбросов).
    - Валидационная выборка. σ_{NMAD} : 0.028, $n_{>0.15}$: 0.061 (на 6.78% меньше выбросов).
    - Тестовая выборка Stripe 82X. σ_{NMAD}: 0.029, $n_{>0.15}$: 0.043 (на 2.74% меньше выбросов).
- Применена и исследована модель для построения и оценки карт волокон галактик.
    - Впервые предложены метрики для численной оценки карт волокон галактик.
- Исследована и разработана модель отождествления и прогноза красного смещения скоплений галактик на основе информации о локальной плотности галактик и расположении относительно крупномасштабной структуры Вселенной. Метрики разработанной модели:
    - $n_{>0.15}$: 0.017
    - σ_{NMAD}: 0.0011
    - ROC-AUC: 0.88
- Разработана библиотека для построения и анализа карты крупномасштабной структуры Вселенной.

8. Планы дальнейших исследований.

Модель прогноза фотометрического красного смещения галактик может быть улучшена за счёт использования более сложных нейросетевых архитектур для обработки табличных данных. Например, трансформерной архитектуры SAINT [34].

Модель отождествления и прогноза красных смещений скоплений галактик может быть улучшена за счёт использования фотометрических оценок красного смещения галактик.

Модель прогноза фотометрического красного смещения галактик может быть применена для поиска ошибочных спектрографических прогнозов красного смещения далеких квазаров.

9. Список источников.

1. Coil A. L. Large scale structure of the universe //arXiv preprint arXiv:1202.6633. – 2012.
2. Sarazin C. L. X-ray emission from clusters of galaxies //Reviews of Modern Physics. – 1986. – Т. 58. – №. 1. – С. 1.
3. Predehl P. et al. The eROSITA X-ray telescope on SRG //Astronomy & Astrophysics. – 2021. – Т. 647. – С. A1.
4. Novaes C. P., Wuensche C. A. Identification of galaxy clusters in cosmic microwave background maps using the Sunyaev-Zel'dovich effect //Astronomy & Astrophysics. – 2012. – Т. 545. – С. A34.
5. Thornton R. J. et al. The Atacama cosmology telescope: the polarization-sensitive ACTPol instrument //The Astrophysical Journal Supplement Series. – 2016. – Т. 227. – №. 2. – С. 21.
6. Gladders M. D., Yee H. K. C. A new method for galaxy cluster detection. I. The algorithm //The Astronomical Journal. – 2000. – Т. 120. – №. 4. – С. 2148.
7. Farrens S. et al. Friends-of-friends groups and clusters in the 2SLAQ catalogue //Monthly Notices of the Royal Astronomical Society. – 2011. – Т. 417. – №. 2. – С. 1402-1416.
8. Peirson A. L. Neural network analysis of X-ray polarimeter data //arXiv preprint arXiv:2206.10537. – 2022.
9. Sousbie T. DisPerSE: robust structure identification in 2D and 3D //arXiv preprint arXiv:1302.6221. – 2013.
10. Schaap W. E. The Delaunay tessellation field estimator //Ph. D. Thesis. – 2007.
11. Buncher B., Carrasco Kind M. Probabilistic cosmic web classification using fast-generated training data //Monthly Notices of the Royal Astronomical Society. – 2020. – Т. 497. – №. 4. – С. 5041-5060.
12. Buncher B., Carrasco Kind M. Probabilistic cosmic web classification using fast-generated training data //Monthly Notices of the Royal Astronomical Society. – 2020. – Т. 497. – №. 4. – С. 5041-5060.
13. Hermes D. Helper for Bézier curves, triangles, and higher order objects //Journal of Open Source Software. – 2017. – Т. 2. – №. 16. – С. 267.
14. Breiman L. Random forests //Machine learning. – 2001. – Т. 45. – №. 1. – С. 5-32.
15. Pâris I. et al. The Sloan Digital Sky Survey quasar catalog: fourteenth data release //Astronomy & Astrophysics. – 2018. – Т. 613. – С. A51.

16. Ross N. P., Cross N. J. G. The near and mid-infrared photometric properties of known redshift z≥ 5 quasars //Monthly Notices of the Royal Astronomical Society. – 2020. – Т. 494. – №. 1. – С. 789-803.

17. Borisov V. et al. Probabilistic photo-z machine learning models for X-ray sky surveys //arXiv preprint arXiv:2107.01891. – 2021.

18. LaMassa S. M. et al. SDSS-IV eBOSS Spectroscopy of X-Ray and WISE AGNs in Stripe 82X: Overview of the Demographics of X-Ray-and Mid-infrared-selected Active Galactic Nuclei //The Astrophysical Journal. – 2019. – Т. 876. – №. 1. – С. 50.

19. Bishop C. M., Nasrabadi N. M. Pattern recognition and machine learning. – New York : springer, 2006. – Т. 4. – №. 4. – С. 738.

20. SDSS [Электронный ресурс]. – Электрон. дан. – [Б. м.] : 2019. – Режим доступа: https://www.sdss.org. – 26.04.2022.

21. Strauss M. A. et al. Spectroscopic target selection in the Sloan Digital Sky Survey: the main galaxy sample //The Astronomical Journal. – 2002. – Т. 124. – №. 3. – С. 1810.

22. Malavasi N. et al. Like a spider in its web: a study of the large-scale structure around the Coma cluster //Astronomy & Astrophysics. – 2020. – Т. 634. – С. A30.

23. Malavasi N. et al. Characterising filaments in the SDSS volume from the galaxy distribution //Astronomy & Astrophysics. – 2020. – Т. 642. – С. A19.

24. Tempel E. et al. Merging groups and clusters of galaxies from the SDSS data-The catalogue of groups and potentially merging systems //Astronomy & Astrophysics. – 2017. – Т. 602. – С. A100.

25. Hilton M. et al. The atacama cosmology telescope: A catalog of> 4000 Sunyaev–Zel'dovich galaxy clusters //The Astrophysical Journal Supplement Series. – 2021. – Т. 253. – №. 1. – С. 3.

26. Breiman L. Bagging predictors //Machine learning. – 1996. – Т. 24. – №. 2. – С. 123-140.

27. Ho T. K. The random subspace method for constructing decision forests //IEEE transactions on pattern analysis and machine intelligence. – 1998. – Т. 20. – №. 8. – С. 832-844.

28. Chen T. Introduction to boosted trees //University of Washington Computer Science. – 2014. – Т. 22. – №. 115. – С. 14-40.

29. Friedman J. H. Stochastic gradient boosting //Computational statistics & data analysis. – 2002. – Т. 38. – №. 4. – С. 367-378.

30. DisPerSe Overview [Электронный ресурс]. – Электрон. дан. – [Б. м.] : 2019. – Режим доступа: http://www2.iap.fr/users/sousbie/web/html/index3c4a.html?category/Overview - 26.04.2022

31. Ke G. et al. Lightgbm: A highly efficient gradient boosting decision tree //Advances in neural information processing systems. – 2017. – Т. 30.
32. Ananna T. T. et al. AGN populations in large-volume X-ray surveys: photometric redshifts and population types found in the stripe 82X survey //The Astrophysical Journal. – 2017. – Т. 850. – №. 1. – С. 66.
33. Brescia M. et al. Photometric redshifts for X-ray-selected active galactic nuclei in the eROSITA era //Monthly Notices of the Royal Astronomical Society. – 2019. – Т. 489. – №. 1. – С. 663-680.
34. Somepalli G. et al. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training //arXiv preprint arXiv:2106.01342. – 2021.