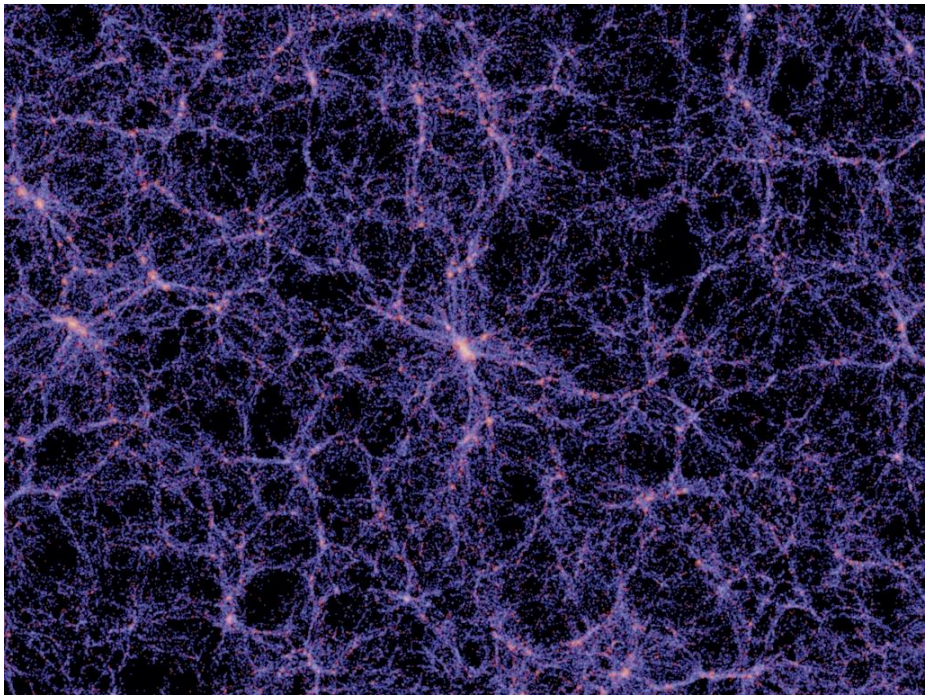
Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

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

Васильев Семён Михайлович, группа 620
Научные руководители: к.ф.-м.н. Мещеряков Александр Валерьевич
Герасимов Сергей Валерьевич

Введение



На больших масштабах темная материя образует связные структуры, объединенные в сеть (крупномасштабная структура Вселенной).

- Гало темной материи.
- Волокна темной материи.
- Стены темной материи.
- Пустоты (войды).

Распределение видимого вещества (галактик) отражает эти структуры.

Построение карты крупномасштабной структуры Вселенной

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

- Определение положения галактик в пространстве. Прогноз расстояния до галактик (красное смещение галактик).
- Поиск волокон галактик.
- Отождествление скоплений галактик – оценка вероятности наличия скопления в точке или направлении (задача классификации).

Прогноз расстояния до астрономических объектов

- 2 основных подхода:
  - Спектрографический прогноз: прямой метод, точный проноз, ~1% объектов.
  - Фотометрическия прогноз: модель регрессии на табличных данных со спектральным пронозом в качестве целевой переменной, прогноз с погрешностью, все наблюдаемые объекты.
- Данные, используемые для фотометрического прогноза, являются мультимодальными. Учет этого свойства позволит увеличить качество прогнозирования.

Актуальность

- Исследование крупномасштабной структуры Вселенной необходимо для понимания процесса эволюции и состава Вселенной.
- Для построения карты крупномасштабной структуры Вселенной необходима общая технология.
- Существующие нейросетевые модели прогноза расстояния до галактик не учитывают многомодальность данных.
- Для построения карты крупномасштабной структуры Вселенной необходимо уметь искать волокна галактик и скопления галактик.

Постановка задачи

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

Подзадачи:

- Исследование и разработка модели прогноза фотометрических красных смещений галактик, учитывающей многомодальность данных.
- Исследование и применение модели DisPerSe для построения и оценки карт волокон галактик.
- Исследование и разработка модели для отождествления скоплений галактик и прогноза их красного смещения по информации о ближайших структурах галактик и локальной плотности галактик.
- Разработка библиотеки для построения и анализа карты крупномасштабной структуры Вселенной.

План доклада

- Обзор:
  - Глубокие ансамбли.
  - Прогноз фотометрического красного смещения галактик. Существующие решения.
  - Данные. Прогноз красного смещения галактик.
  - DisPerSe.
  - Данные. Отождествление скоплений.
  - Метрики.
- Построение решения:
  - Модель прогноза фотометрического красного смещения галактик.
  - Модель построения карт волокон.
  - Модель отождествления скоплений галактик и прогноза их красного смещения.
- Результаты.
  - Модель прогноза фотометрического красного смещения галактик.
  - Модель построения карт волокон.
  - Модель отождествления скоплений галактик и прогноза их красного смещения.
- Практическая реализация.

Обзор

- Глубокие ансамбли.
- Прогноз фотометрического красного смещения галактик. Существующие решения.
- Данные. Прогноз красного смещения галактик.
- DisPerSe.
- Данные. Отождествление скоплений.
- Метрики.

Глубокие ансамбли

Ансамбль нейронных сетей, оценивающий параметры нормального распределения, обученных независимо на одном наборе данных, но с разной инициализацией весов. Применяется для оценки алеаторической и эпистемической компонент неопределенности модели.

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

- $L_i\left(y_j \middle| x_j\right) = \frac{\log(\sigma_i^2(x_j))}{2} + \frac{\left(y_j - \mu_i(x_j)\right)^2}{2\sigma_i^2(x_j)}$
- $\sigma^2(x) = \sigma_a^2(x) + \sigma_e^2(x)$
- $\sigma_a^2(x) = \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2(x)$
- $\sigma_e^2(x) = \frac{1}{M}\sum_{i=1}^{M}[\mu_i(x) - \frac{1}{M}\sum_{k=1}^{M}\mu_k(x)]$

Прогноз фотометрического красного смещения галактик. Существующие решения.

- Модель на основе алгоритма квантильного случайного леса [1]. Наиболее точная модель прогноза красных смещений галактик..
- Модель основанная на аппроксимации шаблонов спектров [2].
- Многослойный перцептрон, обученный на MSE [3].

[1] Borisov, V., Mescheryakov A., et al, 2022, 2107.01891

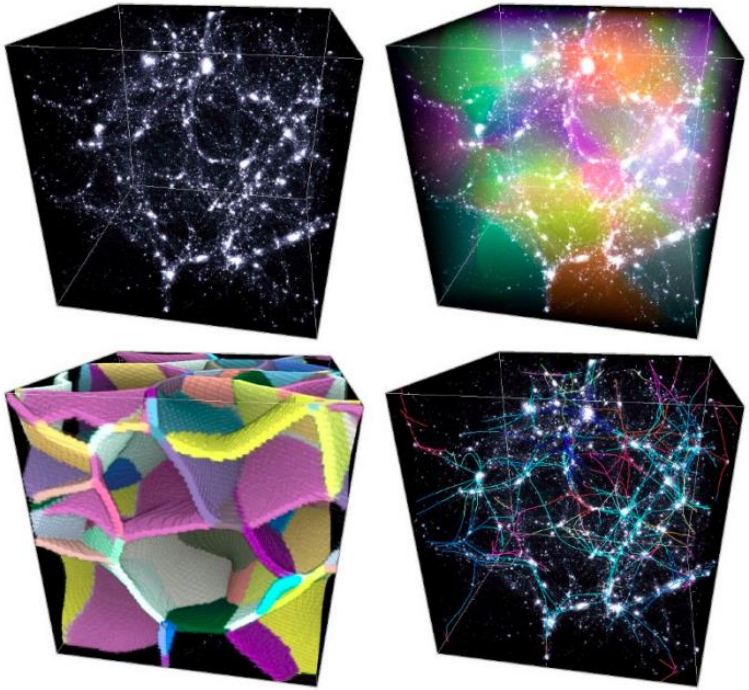
[2] Ananna, T. T., et al. 2017, ApJ, 850, 66. 1710.01296

[3] Brescia, M., Salvato, M., Cavuoti, S., Ananna, T. T., Riccio, G., LaMassa, S. M., Urry, C. M.,& Longo, G. 2019, MNRAS, 489, 663. 1909.00606

Данные. Прогноз фотометрического красного смещения галактик.

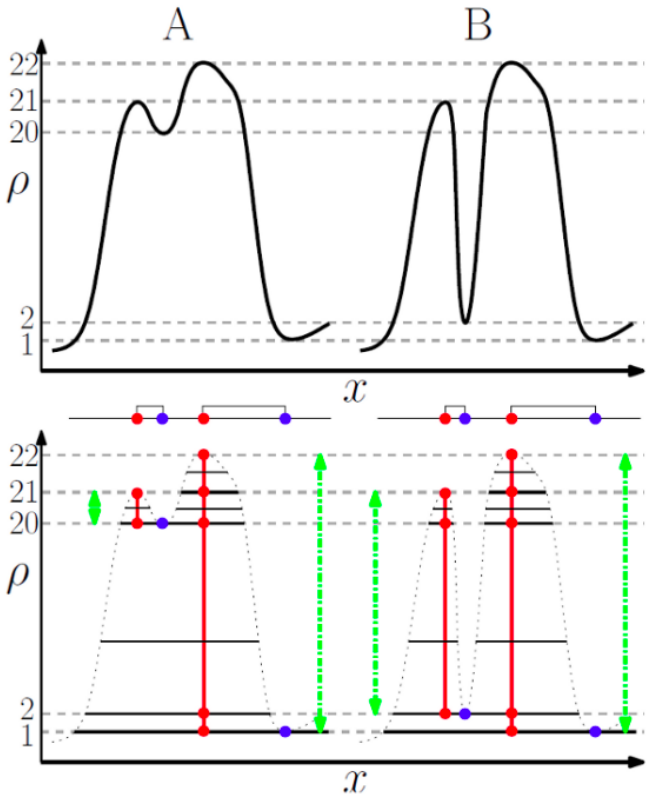
- Выборка составлена из объектов каталогов DR14Q и VHzQ. 576176 объектов.
- Признаки объектов - из объединенных данных 4-х фотометрических обзоров: DESI Legacy Imaging Surveys, PanSTARRS1, SDSS, WISE. Данная выборка была взята из работы [1] для корректного сравнения с моделью на основе случайного леса.
- В качестве тестового набора данных использовалась выборка рентгеновских источников из каталога Stripe 82X. 1164 объекта.

DisPerSe



- Выделяются области возрастания функции плотности распределения галактик.
- Области возрастания функции - войды.
- Грани областей - стены.
- Ребра областей - волокна.
- Локальные максимумы - скопления галактик.

Disperse. Параметры. Удаление шумовых структур



Пары экстремумов функции. Одна из точек порождает топологическую структуру, другая уничтожает.

Разность значений функции в точках - устойчивость пары.

Структуры относящиеся к парам с низкой устойчивостью - шумовые.

Параметр DisPerSe - сигма.

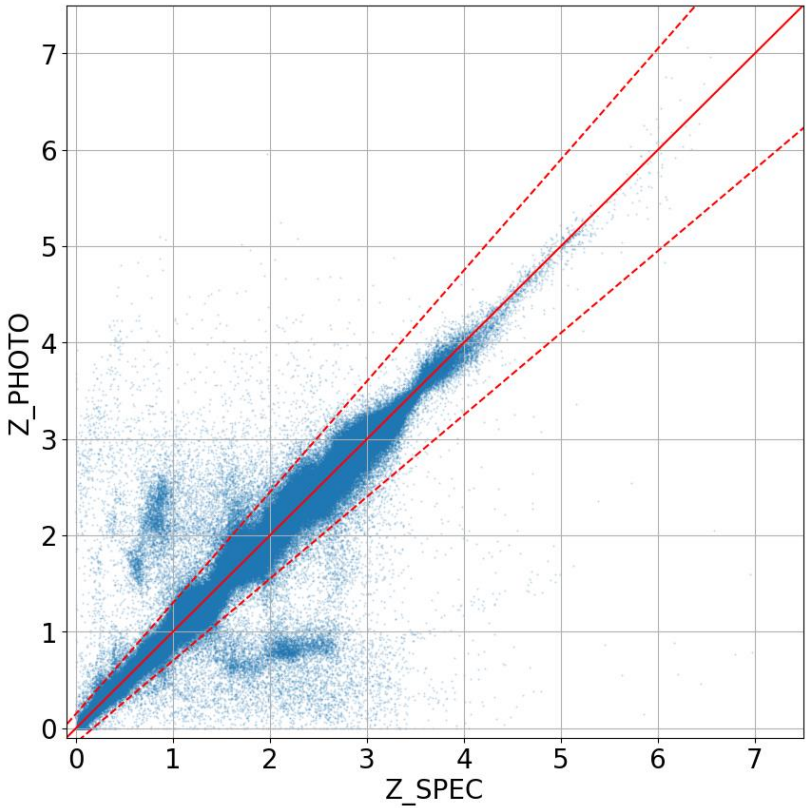
Параметр DisPerSe – число итераций сглаживания оценки плотности.

Данные. Отождествление скоплений

- Каталог галактик SDSS DR16 для построения карт волокон галактик и оценки плотности галактик в областях скоплений.
  Сферические координаты и спектрографические прогнозы красного смещения.
  931860 объектов.
- ACT DR5 Cluster catalog – каталог скоплений галактик.
  Сферические координаты и спектрографические прогнозы красного смещения.
  400 объектов.

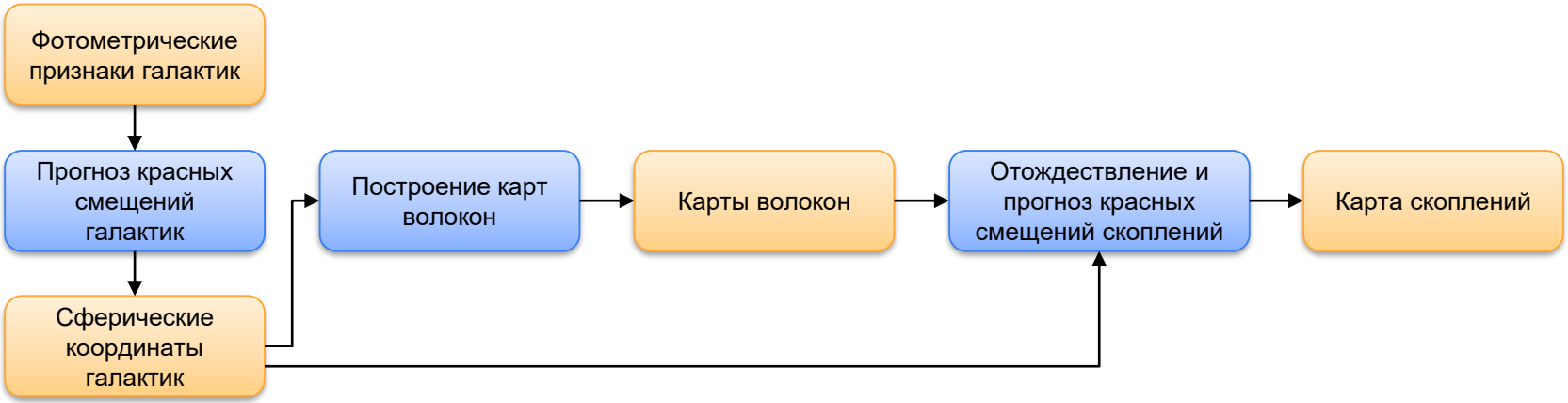
Метрики

- $\sigma_{nmad} = 1.48 * median(|\delta z_i|)$
- $n_{>0.15} = \frac{\#\{i=\overline{1,N} || \delta z_i| > 0.15\}}{N}$
- $\delta z_i = \frac{z_{ph}^i - z_{spec}^i}{1 + z_{spec}^i}$
- ROC-AUC



Построение решения

- Модель прогноза фотометрического красного смещений галактик.
- Модель построения карт волокон.
- Модель отождествления скоплений галактик и прогноза их красного смещения.



Модель прогноза фотометрического красного смещения галактик. Смесь нормальных распределений.

- $p_j(y|x,W) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\left(\frac{\left(x-\mu_j\right)^2}{2\sigma_j^2}\right)}$
- $-\log\big(p(y|x,W)\big) = -\log\big(\sum_{j=1}^{m} \pi_j * p_j(y|x,W)\big)$
- π_j, μ_j, σ_j - оцениваются моделью
- m – число гауссиан в смеси

- MSE: $-\log\big(p(y|x,W)\big) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left(x-\mu_j\right)^2$

Архитектура сети и параметры обучения

- 5 моделей в глубоком ансамбле, смесь из 5 нормальных распределений.
- Полносвязный перцептрон: 6 скрытых слоёв по 400 нейронов.
- ReLU, BatchNorm
- π – SoftMax
- μ – ReLU
- σ – ELU + 1
- Adam, learning rate = $0.001 * 0.96^{\#\text{эпохи}}$, 70 эпох.
- Для увеличения качества модели на далеких объектах (Z > 5) они подавались несколько раз эпоху (до 500).

Модель построения и оценки карт волокон галактик

- Была использована реализация алгоритма DisPerSe на ЯП C++:
  - Каталог галактик с декартовыми или сферическими координатами.
  - Оценка плотности галактик.
  - Построение волокон галактик, как ломаных линий в 2D или 3D пространстве.
- Разработать библиотеку на ЯП Python3.
  - Построение карт.
  - Сериализация.
  - Визуализация.
  - Оценка.

Метрики оценки карт волокон

$$Recall_{true} = \frac{N_{true\ cl\ inter}}{N_{true\ cl\ total}}$$

$$Precision_{true} = \frac{N_{true\ fil\ inter}}{N_{total\ fil}}$$

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

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

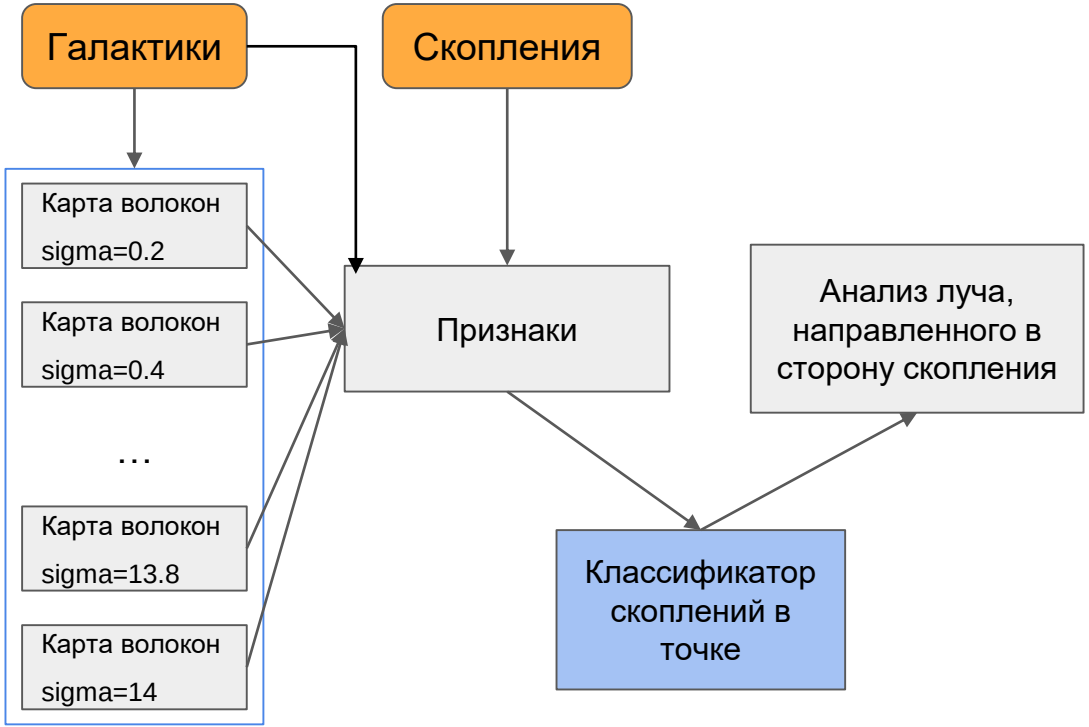
$$Recall_{diff} = Recall_{true} - Recall_{false}$$

$$Precision_{diff} = Precision_{true} - Precision_{false}$$

$$F1_{diff} = 2\frac{Recall_{diff} * Precison_{diff}}{Recall_{diff} + Precision_{diff}}$$

- Для оценки необходим каталог скоплений.
- Ложные скопления генерируются случайно вне окрестностей истинных скоплений.

Модель отождествления скоплений и прогноза их красных смещений. Схема решения



Классификатор скоплений в точке

Объекты:

- Положительные примеры: объекты из каталога ACT DR5.
- Отрицательные примеры: случайные точки.

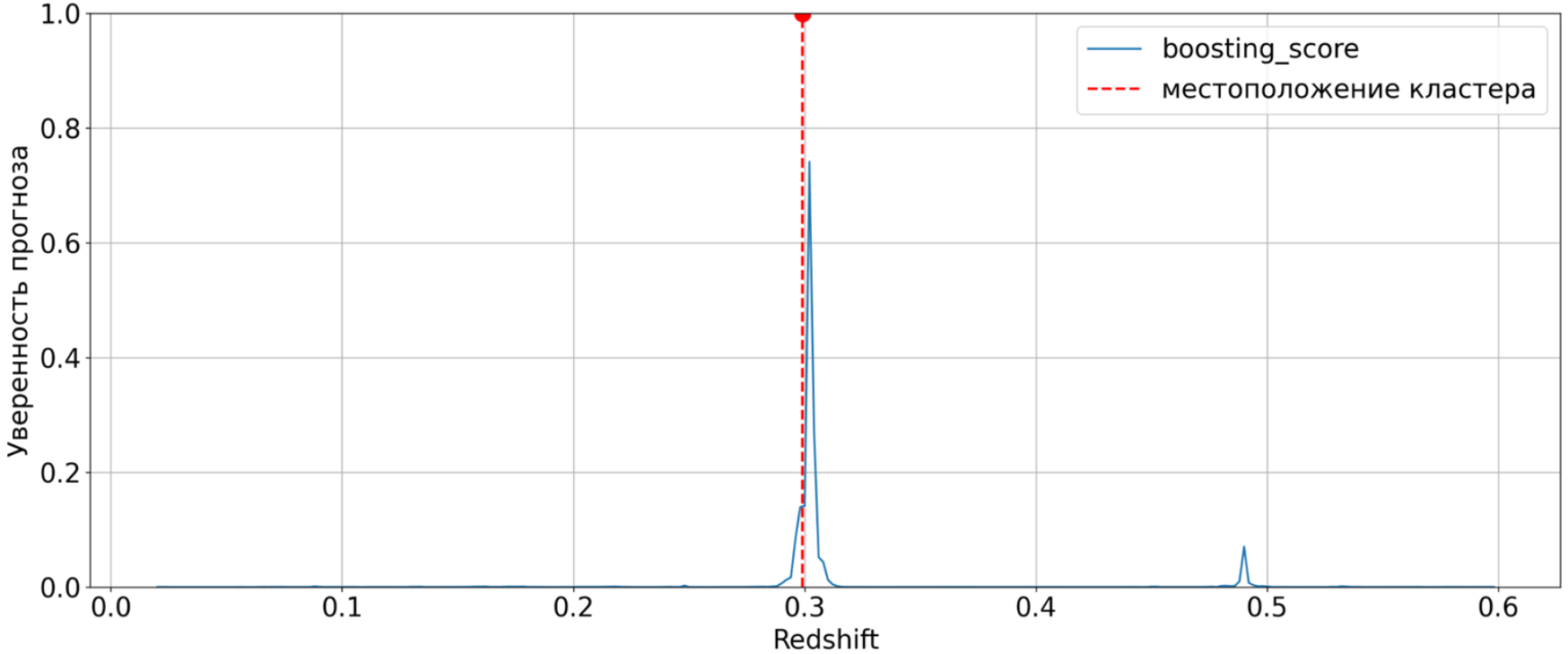
Признаки скоплений:

- Расстояние до ближайшего филамента на картах, соответствующих различным значениям параметров.
- Признаки, описывающие плотность галактик в окрестности точки.
- Расстояние до ближайших соседей на небесной сфере.

Классификатор: градиентный бустинг.

- Число деревьев: 200
- Максимальная глубина: 2
- Скорость обучения: 0.05

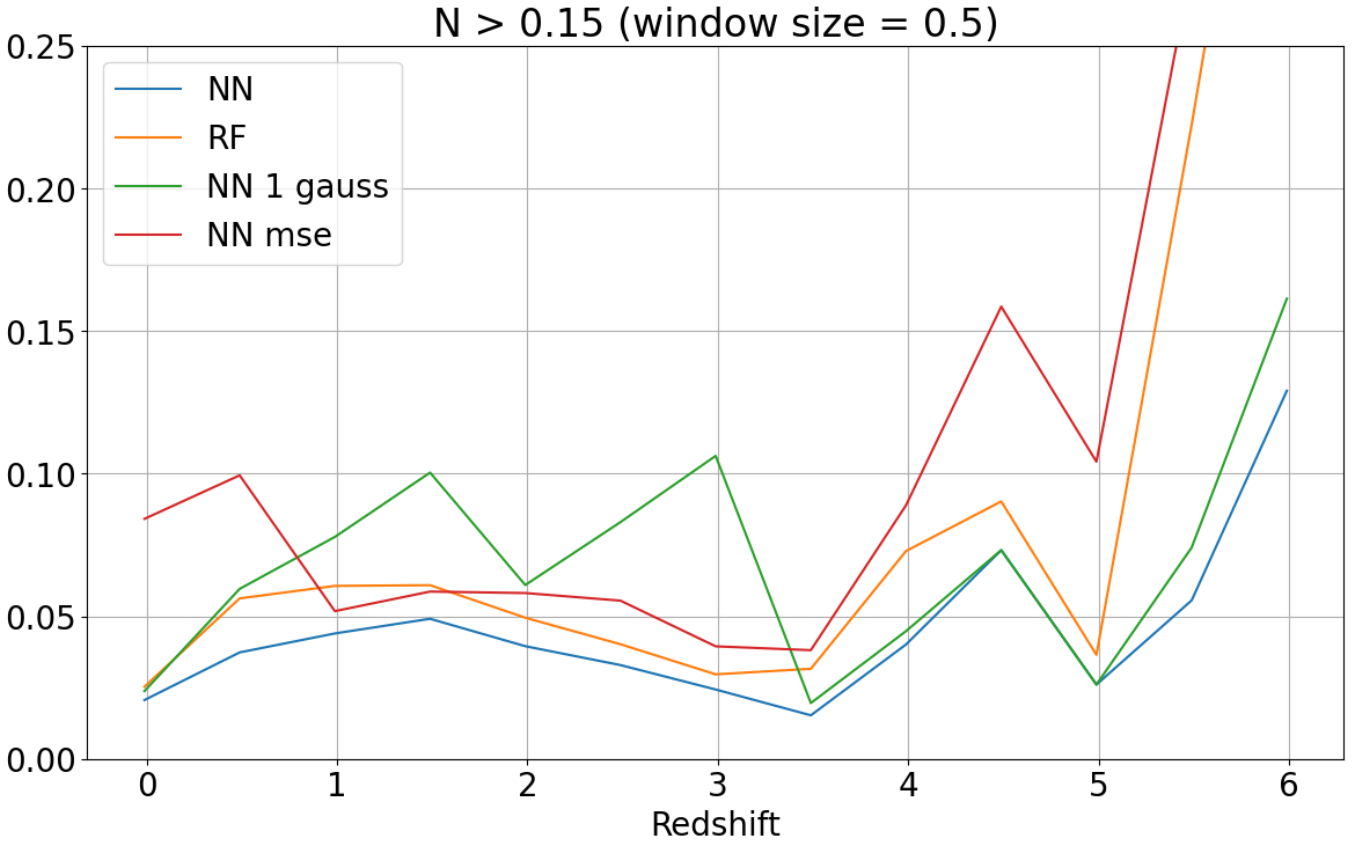
Оценка вероятности скопления и его красного смещения по лучу



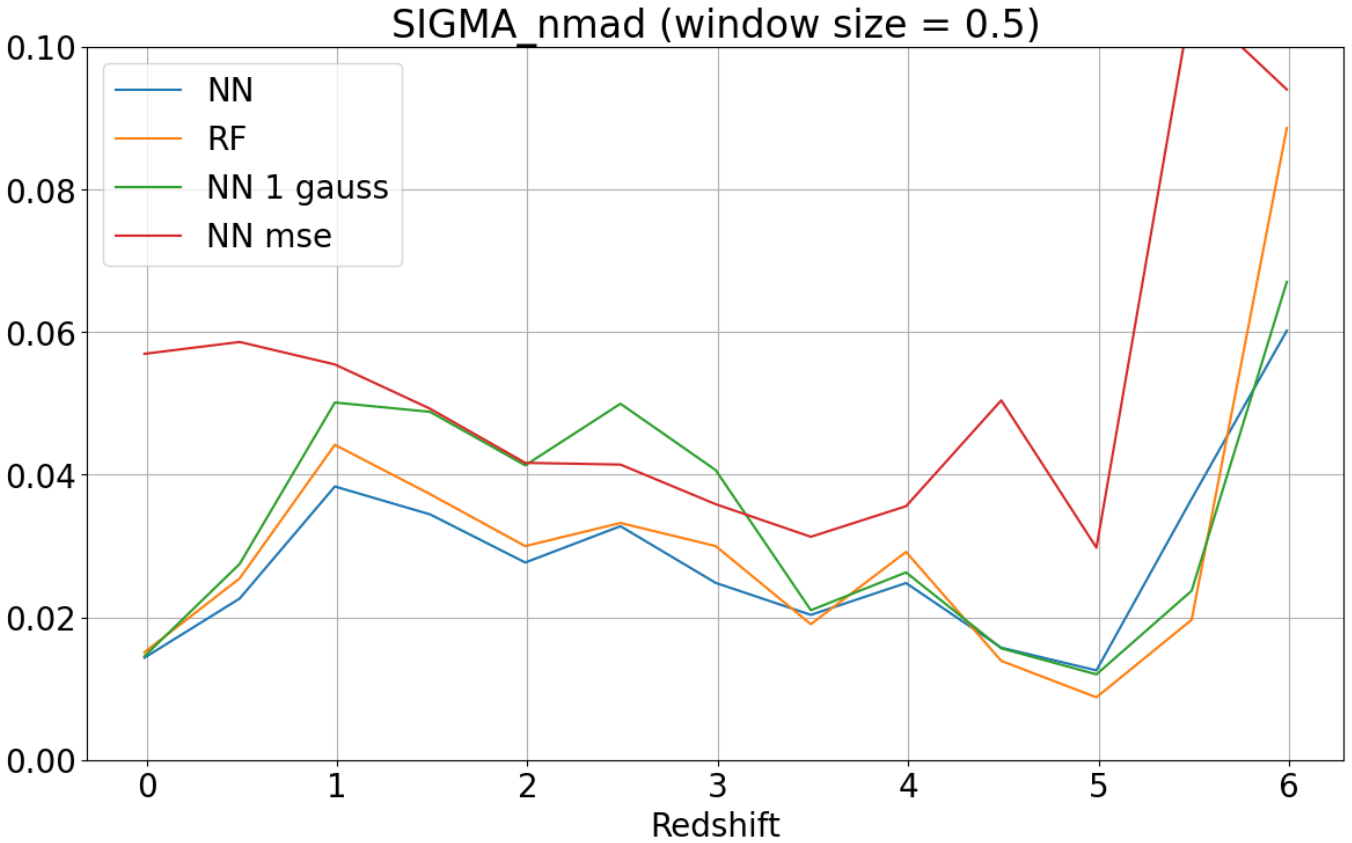
Результаты

- Модель прогноза фотометрического красного смещения галактик.
- Модель построения карт волокон.
- Модель отождествления скоплений галактик и прогноза их красного смещения.

Модель прогноза фотометрического красного смещения галактик. Результаты. Метрики. $n_{>0.15}$

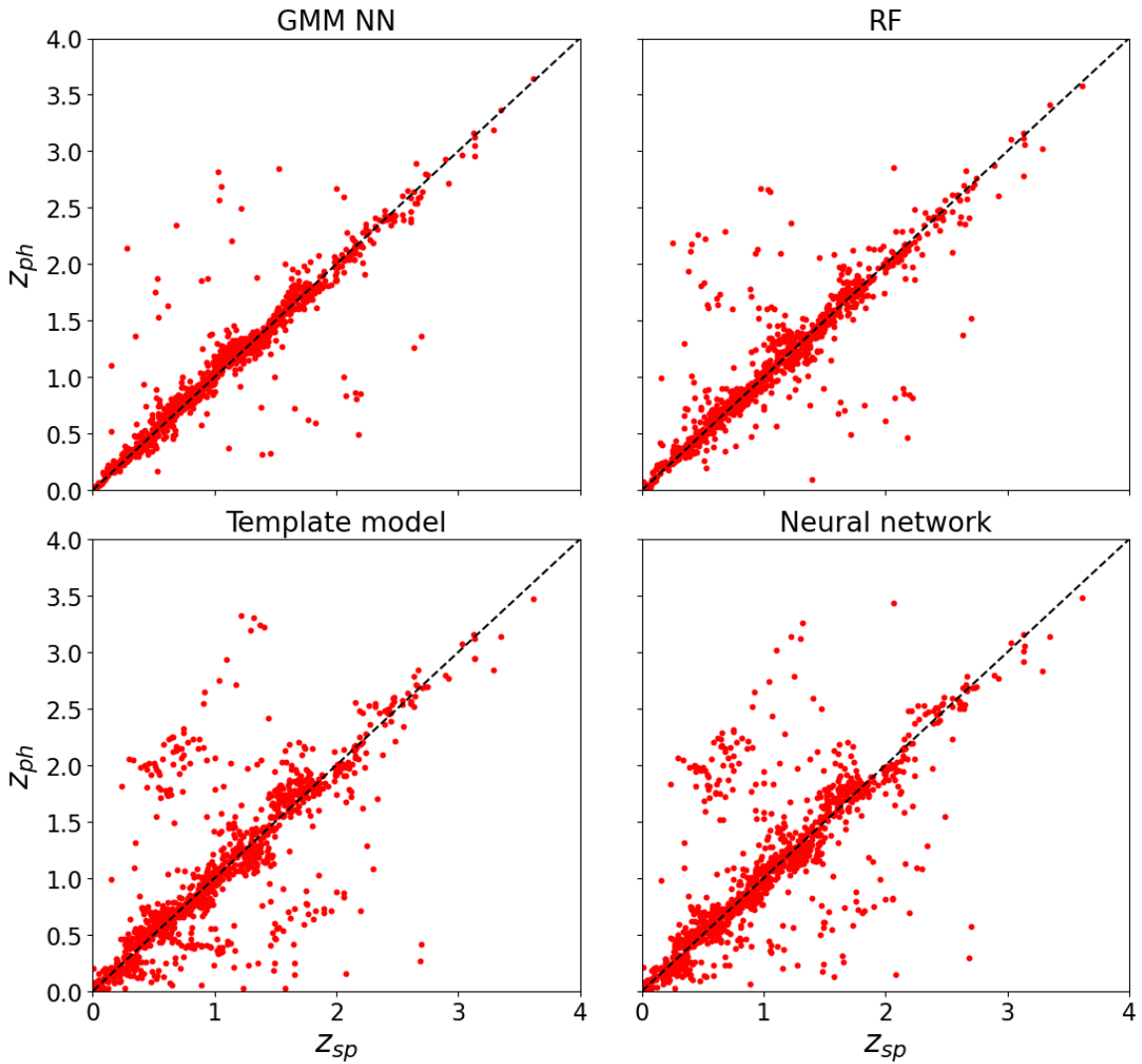


Модель прогноза фотометрического красного смещения галактик. Результаты. Метрики. σ_{nmad}



Модель прогноза фотометрического красного смещения галактик. Результаты. Метрики на каталоге Stripe 82X.

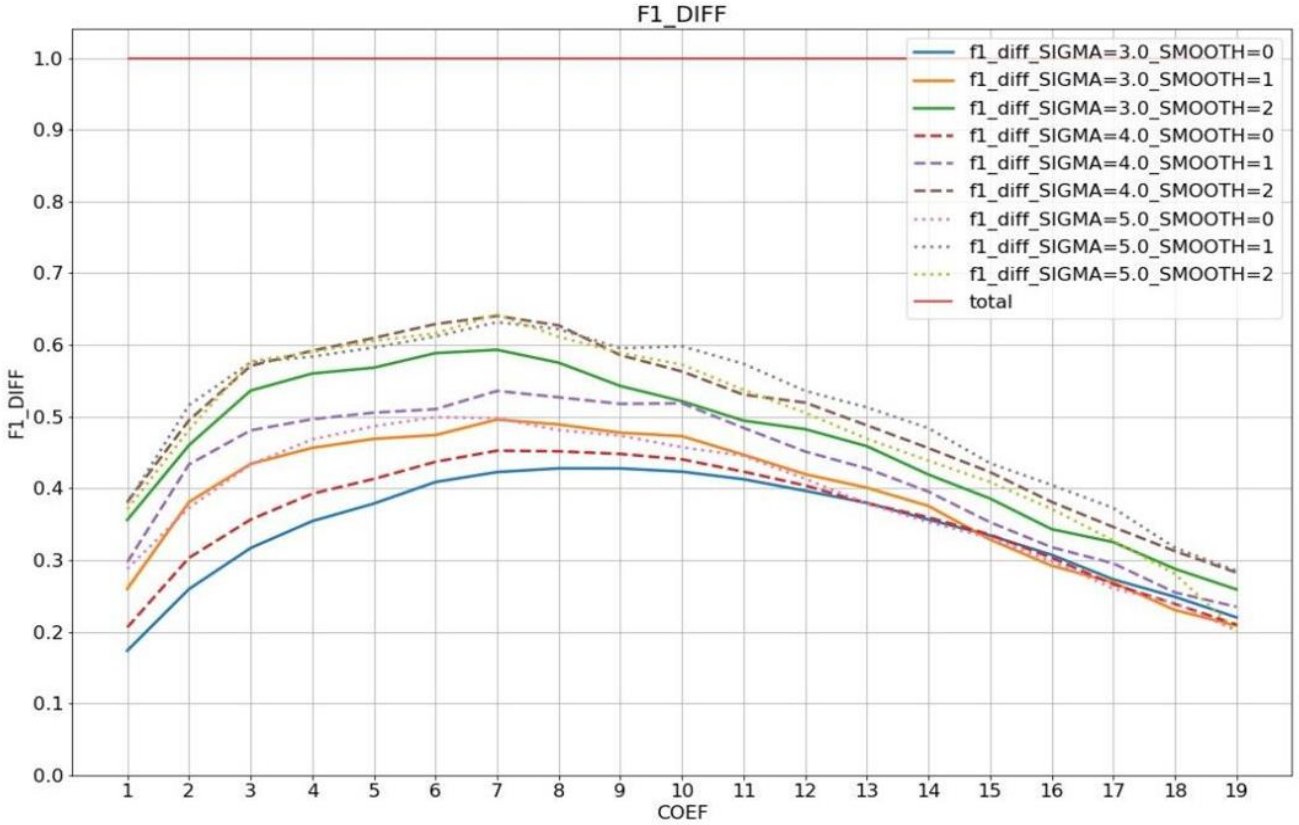
	$N_{>0.15}$	σ_{NMAD}
GMM NN	4.30%	2.95%
RF	7.04%	2.97%
Template model	18.04%	6.79%
Neural network	17.53%	6.58%


GMM NN
RF
Template model
Neural network
z_{ph}
z_{sp}
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
1
2
3
4

Модель прогноза фотометрического красного смещения галактик. Результаты. Выводы.

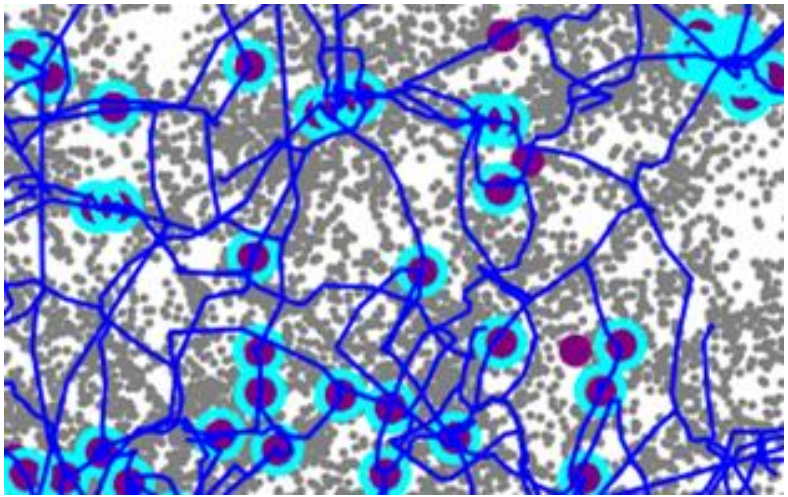
- Исследована и разработана нейросетевая модель, оценивающий фотометрическое красное смещение галактик, как смесь нормальных распределений.
- Качество построенной модели превосходит качество существующих решений на наборе Stripe 82X. В том числе модели на основе алгоритма квантильного случайного леса.
- Построенная модель имеет (относительно случайного леса):
  - Валидационная выборка. σ_{NMAD}: 0.026, $n_{>0.15}$: 0.036 (на 0.82% меньше выбросов).
  - Валидационная выборка (Z>5). σ_{NMAD}: 0.028, $n_{>0.15}$: 0.061 (на 6.78% меньше выбросов).
  - Тест Stripe 82X. σ_{NMAD}: 0.029, $n_{>0.15}$: 0.043 (на 2.74% меньше выбросов).

Модель построения и оценки карт волокон. Результаты. Оценка карт волокон

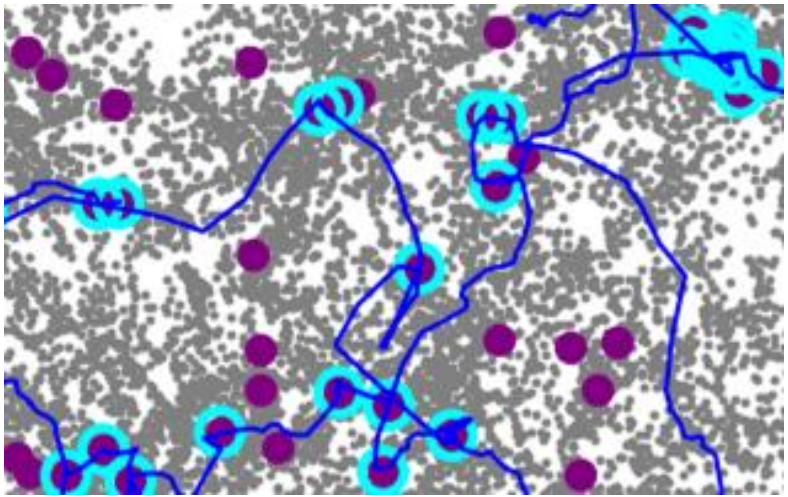


Модель построения и оценки карт волокон. Результаты. Карты волокон для различных сигм

Сигма = 3



Сигма = 7



Модель построения и оценки карт волокон. Результаты. Выводы.

Был разработан программный модуль для построения и анализа карт волокон, позволяющий:

- Строить 2D и 3D карты волокон по набору точечных объектов с декартовыми или сферическими координатами.
- Сериализовывать полученные карты волокон.
- Визуализировать карты волокон.

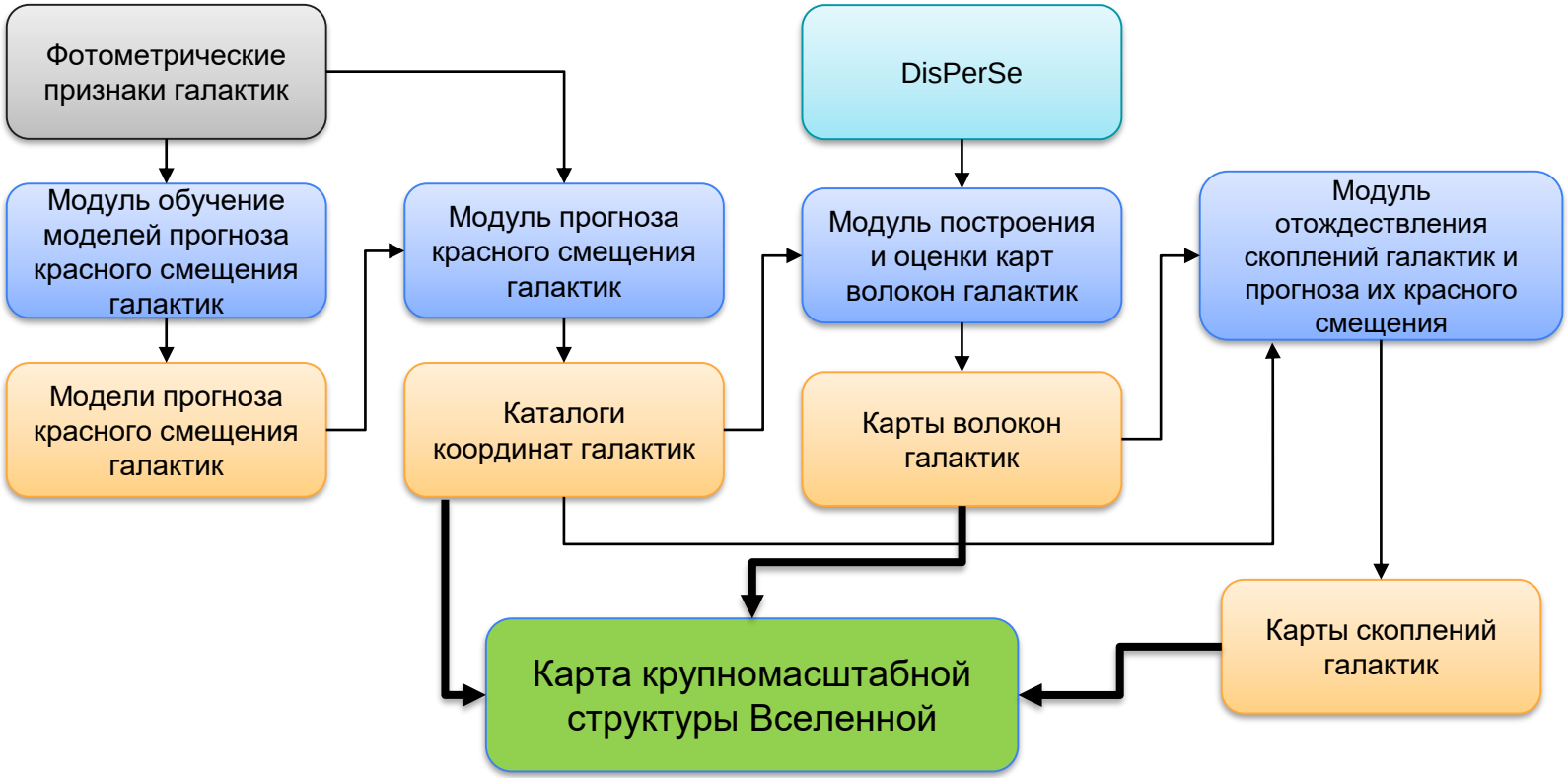
Были предложены метрики для численной оценки качества построенных карт волокон.

Модель отождествления скоплений и прогноза их красных смещений. Результаты

- Базовая модель: красное смещение ближайшей галактики на небесной сфере. Позволяет оценить красное смещение скопления.

- Оценка качества проводилась на каталоге ACT с использованием двойной перекрестной валидации.
- Доля выбросов $n_{>0.15}$: 0.017 (Базовая модель: 0.027)
- σ_{NMAD}: 0.0011 (Базовая модель: 0.0010)
- ROC-AUC: 0.88

Схема библиотека для построения и анализа карты крупномасштабной структуры Вселенной



Практическая реализация.

- Модуль обучение моделей прогноза красного смещения галактик. ~650 строк кода.
- Модуль прогноза красного смещения галактик. ~300 строк кода.
- Модуль построения и оценки карт волокон галактик. ~1000 строк кода.
- Модуль отождествления скоплений галактик и прогноза их красного смещения. ~300 строк кода.

Результаты

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

- Исследована и разработана модель на основе глубоких ансамблей, оценивающая смеси нормальных распределений, для решения задачи прогноза фотометрического красного смещения галактик. Метрики и сравнение с моделью квантильного случайного леса:
  - 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).
  - Тест 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 качества карт волокон галактик.
- Исследована и разработана модель для отождествления скоплений галактик и прогноза их красного смещения, учитывающая крупномасштабную структуру Вселенной. Метрики:
  - $n_{>0.15}$: 0.017
  - σ_{NMAD}: 0.0011
  - ROC-AUC: 0.88
- Разработана библиотека для построения и анализа карты крупномасштабной структуры Вселенной.

Результаты были представлены на конференциях: «Ломоносовские чтения» (2022, 2023) и «Тихоновские чтения» (2023)