Национальный исследовательский институт
Высшая Школа Экономики

На правах рукописи

Авдеева Александра Сергеевна

Исследование параметров звёзд и определение межзвёздного поглощения по данным больших современных обзоров неба

Специальность 1.3.1. Физика космоса, астрономия

ДИССЕРТАЦИЯ
на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук,
доцент
Олег Юрьевич Малков

Москва — 2024

Оглавление

Стр.

Стр.

Введение

Современная астрономия сталкивается с вызовом обработки и анализа огромных объемов данных, получаемых благодаря большим астрономическим обзорам и миссиям. Большие проекты, такие как Gaia, WISE, 2MASS, LAMOST и другие предоставляют обширные наборы данных, включающие фотометрические наблюдения, спектроскопические измерения, астрометрические параметры и другие характеристики звезд и галактик. Эти данные являются бесценным ресурсом для ученых, но их анализ требует разработки новых методов обработки и интерпретации.

Одной из конкретных задач является обнаружение и классификация коричневых карликов среди множества небесных объектов. Вследствие низкой светимости коричневых карликов, вместо спектроскопических методов для их обнаружения зачастую используется метод отбора по показателям цвета. Методы машинного обучения, основанные на анализе фотометрических данных из больших обзоров неба, позволяют автоматически выделять и идентифицировать коричневые карлики на основе их цветовых характеристик и блесков.

Другая важная задача – это разработка и использование методов для получения надежных оценок эффективных температур звезд. Эта обширная проблема включает в себя создание флагов качества для температур, полученных в различных обзорах. Зачастую эффективные температуры, полученные даже в рамках одного подхода, могут иметь различную степень надежности. Флаги качества помогают выбрать только наиболее качественные данные для дальнейшего изучения.

Изучение межзвездного поглощения в различных областях Галактики также является крайне важной задачей. Данные спектроскопии, фотометрии и астрометрии позволяют определять характеристики межзвездного поглощения в зависимости от направления на небе. Использование больших астрометрических обзоров, таких как Gaia, в сочетании с данными спектроскопических наблюдений, позволяет оценить параметры поглощения на луче зрения, а также полное галактическое поглощение.

Актуальность темы исследования и степень ее разработанности.

Коричневые карлики – это субзвездные объекты, которые были теоретически предсказаны [1; 2], а затем спустя 30 лет обнаружены [3; 4]. С тех пор

поиск [5—7] и систематическое изучение известных коричневых карликов [8—10] не прекращались. Масса этих объектов недостаточна для начала и поддержания стабильного термоядерного синтеза гелия из водорода, что приводит к их постепенному остыванию. Пик интенсивности излучения приходится на инфракрасный диапазон, объекты относительно слабо излучают в видимом спектре. В спектральной классификации для коричневых карликов выделены спектральные типы L, T и Y.

Согласно работе [11], количество коричневых карликов в Галактике находится в пределах от 25 до 100 миллиардов. Для различных видов исследований необходимы однородные и полные выборки коричневых карликов. Кинематические исследования [12], исследования двойных звезд с коричневыми карликами [13] и исследования параметров Галактики требуют статистических параметров коричневых карликов. Коричневые карлики занимают промежуточное положение между звездами и планетами, таким образом, изучение свойств коричневых карликов помогает уточнить наше понимание их различий. Полные и однородные каталоги позволяют идентифицировать и характеризовать коричневые карлики с большей точностью, что позволяет лучше определить нижний предел массы для формирования звезд и верхний предел массы для формирования планет. Более того, коричневые карлики имеют сходства с экзопланетами-гигантами, что делает их ценными образцами для изучения атмосфер экзопланет. Изучая атмосферы коричневых карликов, аналогичные атмосферам экзопланет, мы можем получить представление о процессах и условиях, определяющих атмосферы экзопланет, включая наличие облаков, состав атмосферы и тепловые профили.

Возможно, самой актуальной проблемой, связанной с коричневыми карликами, является L/T-переход [14—16]. L/T-переход у коричневых карликов – явление, характеризующееся резким изменением показателя цвета $(J - K_s)$ и блеска H коричневых карликов, которые происходят для объектов находящихся на границе L и T классов. Рассматривается несколько механизмов, которые могут быть ответственны за этот эффект. Модели облаков связывают резкий переход с опусканием пылевых облаков под фотосферу [17; 18]. Нестабильность в углеродной химии в атмосферах коричневых карликов была предложена как еще один механизм [19]. Адиабатическая конвекция, вызванная этой нестабильностью, может привести к изменчивости цвета в спектральной последовательности L/T. Рассеивание облаков выдвигается в качестве альтер-

нативного механизма L/T-перехода [20]. Предполагается, что облака, состоящие из более крупных частиц, рассеиваются легче, чем облака, состоящие из меньших частиц. Переход от спектральных типов L к типам T может сопровождаться переходом от мелких частиц к крупным, что приводит к фрагментации облаков и переходу к атмосферам, лишенным облаков [21; 22]. Подробный обзор проблемы можно найти в [23].

Большая часть потока, излучаемого L и T карликами, находится в ближнем инфракрасном диапазоне от 1 до 2.5 микрометра. Низкие температуры карликов поздних M, L и T типов приводят к появлению богатого ближнего инфракрасного спектра, содержащего многие особенности: от относительно узких линий нейтральных атомов до широких молекулярных полос, каждая из которых имеет различные зависимости от температуры, силы тяжести и металличности. Типичные атмосферы известных коричневых карликов имеют температуру от 2200 до 750 K.

Наиболее надежным методом идентификации коричневых карликов является спектроскопия, которая позволяет непосредственно анализировать состав атмосферы этих объектов. Спектроскопические данные позволяют идентифицировать характерные признаки, такие как спектральные линии лития или атмосферного метана, которые служат важными индикаторами коричневых карликов. Кроме того, особенности спектров коричневых карликов классов L, T и Y позволяют отличить их друг от друга. Например, L карлики характеризуются полосами эмиссии и заметными атомными линиями щелочных металлов, таких как натрий и калий. T карлики, в свою очередь, показывают сильные полосы поглощения H_2O и монооксида углерода (CO) в инфракрасном диапазоне. Y карлики имеют спектры с пиками поглощения около 1.55 микрометра, что, вероятно, связано с поглощением аммиака. Эти различия в спектрах позволяют классифицировать коричневые карлики и более точно определять их физические свойства.

В то время как спектроскопия необходима для подтверждения природы коричневого карлика и изучения его детальных свойств, проведение спектроскопических наблюдений для большого количества объектов на всем небе является затратным по времени и ресурсам. Например, самый крупный современный спектроскопический обзор LAMOST не содержит спектров коричневых карликов, поскольку они являются слишком слабыми для наблюдения его инстурментами. С другой стороны, фотометрические обзоры могут охватить

намного большую область неба и получать данные о многочисленных небесных объектах одновременно.

Применяя техники выбора по цвету в фотометрических обследованиях, можно выявить объекты, проявляющие цвета, характерные для коричневых карликов. Преимущество использования фотометрических обследований заключается в том, что они позволяют систематически и в широком масштабе искать потенциальных кандидатов в коричневые карлики, помогая выявлять многообещающие объекты для последующих спектроскопических наблюдений.

Например, авторы [9] провели поиск в обзорах SDSS, UKIDSS и WISE. Они применили критерий выбора по цвету: $(Y - J) > 0.8$ и $J < 17.5$, как правило отбора коричневых карликов из всего массива данных. С таким подходом им удалось обнаружить около 1300 коричневых карликов в области 3000 квадратных градусов, что составляет примерно 7.5% небесной сферы. Еще одним примером поиска коричневых карликов с помощью цветового отбора является работа [7]. В ней использованы данные обзоров DES, VHS и WISE. Были применены следующие критерии по цвету для поиска коричневых карликов: $(i - z) > 1.2$, $(z - Y) > 0.15$, $(Y - J) > 1.6$, и $z < 22$. Наложение ограничения на звездную величину в полосе z необходимо для обеспечения полноты набора данных, чтобы исключить тем самым пропущенные значения. В пределах области, охватывающей 2400 квадратных градусов, что составляет примерно 5.8% небесной сферы, благодаря такому подходу было идентифицировано примерно 12 тысяч коричневых карликов. В работе [7] также представлен обзор других работ по отбору коричневых карликов по цвету.

Методы машинного обучения все чаще применяются для классификации астрономических объектов из-за огромного объема данных, собранных за последние десятилетия. Например, в работе [24] были использованы методы Опорных Векторов (SVM), Случайного Леса (RF) и многослойного персептрона для классификации массивных звезд в близлежащих галактиках. Точность классификации на тестовом наборе данных составила 83%. Применение этих методов к другим галактикам (не включенным в тренировочный набор данных), таким как IC1613, WLM и Sextans A, показало точность на уровне 70%. Снижение точности классификации относительно тестового набора данных авторы связывают с различными значениями металличности и поглощения в других галактиках. В работе исследовались два метода заполнения пропущенных данных в каталогах, а именно, заполнение простыми средними значениями и методом

Iterative Imputer из библиотеки Scikit-learn [25]. Метод Iterative Imputer вычисляет отсутствующие значения на основе имеющихся значений признаков так же, как это делают модели регрессии. Этот метод, будучи в то же время более устойчивым, показал лучшие результаты и на классификации.

Интерпретируемые методы машинного обучения (Localized General Matrix LVQ и RF) были использованы в работе [26] для обнаружения ультракомпактных карликовых галактик (UCDs) и шаровых скоплений. В то время как зачастую модели машинного обучения представляют собой черные ящики, что снижает доверие к результатам, для некоторых методов можно вычислить важность отдельных признаков. Если то, насколько важным является тот или иной признак, можно объяснить с помощью физических механизмов или особенностью исследуемых объектов, метод приобретает особую ценность. Авторы работы [26] проанализировали важность признаков и сравнивали их с признаками, несущими физическую информацию об объектах.

Точное определение эффективных температур звезд имеет первостепенное значение в различных областях астрономии и астрофизики. Знание точных и надежных температур небесных объектов позволяет судить об их внутренних свойствах, изучить стадии их эволюции и получить представление о фундаментальных астрофизических процессах.

Эффективные температуры, полученные из данных различных спектроскопических обзоров, использовались для характеристики асимптотической ветви гигантов в широком диапазоне масс и металличности [27] и необходимы для получения детальной информации о звездах диска и гало [28]. Как упоминалось выше, эффективные температуры используются для характеристики межзвездной среды в работе [29]. Аналогичным образом, в работе [30] были получены межзвездные поглощения для более миллиона звезд с использованием спектроскопических параметров обзоров LAMOST и GALAH [31].

На данный момент существует несколько спектроскопических обзоров с высоким разрешением, которые можно считать надежным источником фундаментальных параметров звезд. Это обзоры APOGEE [32; 33], GALAH, Gaia-ESO [34]. Последняя на данный момент версия каталога APOGEE – DR17 ($R \sim 22500$) включает информацию об объектах северного полушария, а также расположенных на выделенных участках южного неба. APOGEE в первую очередь направлен на изучение проэволюционировавших звезд в галактическом диске, в галактическом центре и во внешнем гало. Самая свежая версия каталога

GALAH, DR3 (R $\sim$ 28000), содержит звезды в диапазоне звездных величин $12 < V < 14$ и с галактической широтой $|b| > 10$ градусов. Обзор Gaia-ESO (R $\sim 16000 - 25000$) – это спектроскопический обзор, целью которого является получение высококачественной спектроскопии 100 000 звезд Млечного Пути, принадлежащих всем основным населениям Галактики. Хотя эти обзоры имеют достаточное разрешение и отношение сигнала к шуму, они ограничены по охвату неба и яркости звезд. Мы сравниваем температуры Gaia GSP-Phot с эффективными температурами обзоров APOGEE на северном небе и GALAH на южном небе.

Третий выпуск данных Gaia (DR3) [35; 36] представляет беспрецедентный набор точных астрометрических, фотометрических данных и данных об астрофизических параметрах, что является по настоящему революционным. Он включает в себя информацию об атмосферных характеристиках сотен миллионов звезд, полученную с помощью нескольких независимых модулей, которые имеют разные наборы входных данных из Gaia [37]. Среди данных Gaia DR3 одними из самых ожидаемых являются спектры низкого разрешения Синего (BP) и Красного (RP) фотометров (описание и внутренняя калибровка [38; 39], внешняя калибровка [40]), позволяющие определять астрофизические параметры для сотен миллионов звезд.

Один из основных модулей, General Stellar Parametrizer from Photometry (GSP-Phot), дает оценки эффективной температуры $T_{\rm eff}$, ускорения свободного падения на поверхности звезды $\log g$, металличности, абсолютной звездной величины M_G, радиуса, расстояния, поглощения на луче зрения A_0, A_G, A_{BP}, A_{RP} и покраснения $E(BP - RP)$ путем прямого моделирования спектров BP/RP с низким разрешением, видимой звездной величины G и параллакса с использованием методов Монте-Карло с марковскими цепями (MCMC). С этой целью GSP-Phot использует модели звездной эволюции для получения самосогласованных температур, ускорения свободного падения на поверхности звезды, металличности, радиусов и абсолютных звездных величин [41]. Модуль GSP-Phot предоставляет ряд астрофизических параметров, полученных с помощью различных кодов расчета звездной атмосферы, а именно, MARCS [42], PHOENIX [43], A и OB [44; 45] модели. Gaia DR3 приводит лучшие значения из полученных моделями в модуле GSP-Phot для 471 миллиона источников.

Эти параметры позволяют изучать свойства звезд, звездную эволюцию и состав различных звездных популяций в Галактике. Однако самосогласованное

определение этих параметров может внести дополнительные систематические ошибки, особенно при оценке эффективной температуры $T_{\rm eff}$. В работе [46] данные Gaia DR3 использовались в качестве литературных значений для сравнения калибровки спектров. Однако, им пришлось исключить все объекты с $T_{\rm eff}^{\rm Gaia} > 7000$ K из-за наблюдаемых систематических различий. В работе [47] также отмечается несоответствие между оценками $T_{\rm eff}$ Gaia DR3 GSP-Phot и собственными оценками авторов для звезд в рассеянных скоплениях Гиады и Плеяды. К сожалению, модуль GSP-Phot, несмотря на то, что он является богатейшим источником астрофизических параметров в Gaia DR3, не предоставляет флагов качества или других индикаторов, на которые можно положиться при использовании эффективных температур и других атмосферных параметров.

Предпринимаются попытки произвести переоценку эффективных температур, предоставляемых Gaia DR3 в модуле GSP-Phot. В исследовании [48] авторы использовали методы машинного обучения, в частности XGBoost, для создания новых оценок эффективных температур, а также металличности и ускорения свободного падения на поверхности звезды. Модель была обучена на атмосферных параметрах из обзора APOGEE для вычисления параметров по спектрам BP/RP, звездным величинам Gaia DR3 и CatWISE. Результаты представляют собой 175 миллионов звезд с пересмотренными атмосферными параметрами, показывающими хорошее согласие с параметрами из обзора APOGEE. В работе [49] предлагаются пересмотренные параметры звездных атмосфер для 220 миллионов звезд из Gaia DR3. В этом подходе используется модель спектров Gaia BP/RP, обученная с использованием данных LAMOST и дополненная фотометрией 2MASS и WISE. Этот метод повышает точность и уменьшает вырождение параметров, что приводит к лучшим оценкам параметров звезд.

Учет поглощения света в межзвездном пространстве является важным этапом в каждом астрономическом и астрофизическом исследовании. По этой причине трехмерные карты и модели межзвездного покраснения или поглощения являются очень важными инструментами. Более того, распределение межзвездной пыли, которая в основном ответственна за поглощение, само по себе представляет интерес в контексте изучения эволюции и структуры Галактики.

Одним из основных источников для оценки покраснения и поглощения любого внегалактического объекта до сих пор является двумерная карта,

представленная в работе [50]. Эта карта основана на данных космических миссий NASA COBE/DIRBE и IRAS (ISSA) по излучению пыли в дальнем инфракрасном диапазоне, в частности, на длине волны 100 мкм. Излучение было рассчитано от наблюдателя до бесконечности, включая весь слой пыли вдоль средней плоскости Галактики. Поскольку диффузное излучение в дальнем инфракрасном диапазоне напрямую связано с поверхностной плотностью межзвездной пыли, такую карту можно использовать как меру поглощения внегалактических объектов.

Здесь мы представляем обзор работ, в которых представлены трехмерные карты поглощения. Работа [51] основана на популяционном синтезе Галактики с использованием фотометрии 2MASS. Полученная трехмерная карта доступна в виде таблицы в базе данных Vizier. В работах [52; 53] используются байесовские методы для создания трехмерной карты. Авторы работы [52] используют фотометрические данные обзора INT Photometric H-Alpha Survey (IPHAS). Однако, поглощение изучается только для низких галактических широт, главным образом в галактической плоскости. В работе [53] используются данные Gaia, Pan-STARRS DR1 и 2MASS. Карта охватывает примерно три четверти небесной сферы и имеет ограничение по минимальному расстоянию из-за перенасыщения близких звезд в обзоре Pan-STARRS. Типичное значение этого предела составляет около 300 парсек, то есть за основной массой пыли в областях высоких галактических широт.

Авторы работы [54] использовали метод регуляризованного байесовского подхода к данным избытка цвета, ранее полученным в работах [55; 56] а также на основе Женевско-Копенгагенского обзора [57; 58]. Этот метод получил развитие в работе [59], где была использована фотометрия обзора 2MASS и астрометрия Gaia DR2. Результатом работы стала трехмерная карта межзвездной пыли в пределах 3 кпк. Обе карты доступны в виде таблиц. У первой упомянутой работы есть онлайн-сервис для извлечения значения покраснения для выделенного направления (или направлений). Результаты второй работы представлены в виде таблицы, построенной в евклидовых координатах, без возможности извлечения данных для конкретного направления.

В работе [60] используется фотометрия 2MASS и WISE, а также оптические наблюдения с телескопа Шмидта Xuyi 1.04/1.20-m (XSTPS-GAC) для создания трехмерной карты поглощения в полосе r. Карта охватывает площадь более 6000 квадратных градусов вокруг антицентра Галактики ($140 < l <$

$240, -60 < b < 40$) с пространственным угловым разрешением (в зависимости от широты) от 3 до 9 угловых минут. Для решения этой задачи в работе сделана выборка из 132316 эталонных звезд с нулевым значением поглощения. Далее эта выборка используется для создания стандартной библиотеки распределений энергий в спектре. Поскольку покраснение или поглощение влияет на видимую и ультрафиолетовую часть спектра, но слабо влияет на инфракрасную область спектра, то сравнение наблюдаемого распределения энергии в спектре с эталоном позволяет определить из этих данных значение покраснения. Для выборки, состоящей из примерно 13 миллионов звезд, обладающих надежной фотометрией, этим методом было определено поглощение в полосе r. Карта охватывает диапазон расстояний от 0 до 4 кпк. Однако на расстояниях более 3 кпк она становится менее надежной из-за меньшего количества звезд с высококачественной фотометрией, доступной для оценки поглощения.

В исследовании [61] проанализировано распределение 70 миллионов звезд из 2MASS с высокоточной фотометрией (лучше 0,05 зв.величины) на диаграмме (J - Ks) – Ks. Один из пиков этого распределения состоит из карликов, субкарликов и субгигантов типа F со средней абсолютной величиной M_{Ks} в 3 звездные величины. Сдвиг этого пика в сторону больших значений (J - Ks) с увеличением Ks отражает покраснение этих звезд с увеличением расстояния или поглощения на луче зрения. В результате в каждой пространственной ячейке среднее расстояние и средняя звездная величина Ks оказываются взаимосвязаны друг с другом, что позволяет оценить поглощение на заданном расстоянии.

В отличие от карт, модели поглощения определяются некоторыми формулами, что делает их более удобными в использовании. Тем не менее, различные параметры формул также могут быть представлены в табличной форме, поскольку их различия в разных частях небесной сферы могут быть значительными. Здесь мы также кратко рассмотрим существующие на сегодняшний день модели.

Первая модель, разработанная [62], представляет собой барометрическую (экспоненциальную) функцию:

$$A_V(b,d) = \frac{a_0 \cdot \beta}{\sin|b|}\left(1 - \exp\left(\frac{-d \cdot \sin|b|}{\beta}\right)\right), \quad (1)$$

Это классическая модель однородного, полубесконечного поглощающего слоя с плотностью, экспоненциально распределенной по высоте. Параметр β

представляет собой высоту шкалы, а a_0 – поглощение на единицу длины в галактической плоскости. Хотя формула 1 по-прежнему актуальна для отдельных направлений, в [62] предполагались постоянные параметры на сфере, что, очевидно, не соответствует действительности.

В [63] была предложена альтернативная трехмерная аналитическая модель пространственной вариации поглощения A_V в Галактике, которая описывает поглощение в каждой точке в пределах околоземного пространства, учитывая пылевые слои вдоль галактической плоскости и пояса Гулда. Авторы работы обработали все доступные точные спектральные и фотометрические данные для более чем 42 000 звезд и разработали аналитическую трехмерную модель межзвездного поглощения A_V на расстоянии около 1 кпк от Солнца. Небо было разделено на 199 ячеек по галактическим координатам. В каждой ячейке величина поглощения задавалась формулой:

$$A_V = k1 \cdot R + k2 \cdot R^2, \tag{2}$$

где R – расстояние до объекта, а k1 и k2 – эмпирически найденные коэффициенты для каждой ячейки в пределах слоя пыли. Вне этого слоя A_V оставалось постоянным и идентичным A_V на краю слоя (в более высоких широтах) или медленно увеличивалось по закону линейной регрессии (в более низких широтах). Были установлены следующие ограничения на A_V: $(A_V < 0.1)$ зв.величины для $(|b| \geqslant 60)$, $(A_V < 1.2)$ зв.величины для $(45 \leqslant |b| < 60)$, и $(A_V < 3)$ зв.величины для $(|b| < 45)$.

Однако эта модель не предоставляет физического объяснения для наблюдаемых систематических пространственных изменений A_V. Внутри каждой области неба использовалась только одна функциональная зависимость A_V от расстояния, что делает зависимости в близких областях неба несовместимыми между собой. Это означает, что A_V может существенно различаться в соседних направлениях. Большинство проанализированных звезд относятся к O-F звездам главной последовательности в пределах 600 пк от Солнца и близко к плоскости Галактики, что приводит к более низкой точности A_V на больших расстояниях и высоких широтах.

Карта [50] и другие источники показывают, что пыль распределена не только вдоль плоскости Галактики, но и вдоль пояса Гулда, который содержит пылевые облака, области звездообразования, молодые звезды, их ассоциации и скопления, оказывая влияние на глобальную структуру поглощения.

В модели [61] два пылевых слоя пересекаются под углом γ, и их ось пересечения повернута относительно оси Y на угол λ_0. Слой пояса Гулда имеет конечный радиус R_{limit} и центрирован относительно Солнца. Этот радиус может быть рассчитан как свободный параметр модели или принят как фиксированный (например, $R_{limit} < 600$ пк). Расчет поглощения в поясе ограничен этим радиусом. Внутри каждого слоя поглощение описывается барометрическим законом, зависящим от цилиндрических координат. Свободные параметры модели определяются для разных направлений путем подгонки карт поглощения, как например описано в работах [64; 65].

В работе [29] предлагается альтернативный метод построения эмпирической модели распределения поглощения в Галактике. Авторы работы вычисляют поглощение A_V, основываясь на данных фотометрии и астрометрии Gaia DR2 и EDR3 и используя данных спектроскопического обзора LAMOST для определения собственных показателей цвета для звезд в выделенных площадках северной части неба. Для определения собственных показателей цвета используется зависимость эффективная температура – собственный показатель цвета. Полученные значения A_V затем аппроксимируются в каждой площадке согласно закону косеканса (1) и для каждой площадки находится пара параметров a_0 и β. Затем значения параметров a_0 и β аппроксимируются сферическими функциями по всей небесной сфере, что обеспечивает согласованность параметров в соседних регионах.

Целью данной работы было определение параметров поглощения межзвездной среды на основе анализа спектроскопических, фотометрических и астрометрических данных различных астрономических обзоров и миссий, а также исследование параметров звезд, включая независимую оценку надежности эффективных температур Gaia DR3 и разработку и апробацию новых методов для поиска и классификации астрономических объектов, в частности, коричневых карликов.

В процессе достижения данной цели в рамках данной работы были решены следующие **задачи**:

– Определение межзвездного поглощения с помощью данных об эффективной температуре из спектроскопических обзоров. Получение значений параметров закона косеканса в различных направлениях по полученным значениям поглощения.

– Разработка и применение цветовых правил для поиска и классификации коричневых карликов в данных фотометрических обзоров, как классическими методами, так и с использованием машинного обучения.
– Создание флагов качества для температур Gaia DR3 модуля GSP-Phot с использованием спектроскопических обзоров высокого разрешения, а также моделей машинного обучения.

Научная новизна: Научная новизна диссертационной работы заключается в использовании новых данных спектроскопических обзоров для вычисления собственных показателей цвета, использовании самой актуальной переписи ближайших коричневых карликов для составления цветовых правил отбора и первом применении алгоритмов машинного обучения к задаче об идентификации коричневых карликов среди объектов других классов, что делает все полученные в ходе работы результаты новыми. Впервые был разработан метод оценки качества эффективных температур, представленных в каталоге Gaia DR3 модуля GSP-Phot, с помощью машинного обучения и получены флаги качества для этих эффективных температур. Подход определения межзвездного поглощения с помощью данных спектроскопических обзоров был распространен на южное небо с использованием данных обзора RAVE.

Теоретическая и практическая значимость.

На данных обзоров RAVE и LAMOST был исследован метод определения поглощения с использованием данных спектроскопических обзоров. Показано, что метод определения собственных показателей цвета с использованием эффективной температуры, предоставленной спектроскопическими обзорами, хорошо согласуется с известными значениями поглощения из литературы. Была показана возможность и перспективы применения моделей машинного обучения к выделению коричневых карликов в фотометрических обзорах WISE, 2MASS и Pan-STARRS. Кроме того, показано, что глубина обзора Gaia недостаточна для полного обнаружения всех коричневых карликов. Показана возможность получения флагов качества для эффективных температур модуля GSP-Phot каталога Gaia с помощью данных спектроскопических обзоров высокого разрешения, а также методов машинного обучения.

В ходе выполнения работы были получены результаты, которые также имеют важную практическую значимость. Были разработаны точные фотометрические правила для отбора коричневых карликов из данных обзоров DES, 2MASS и WISE. Фотометрические правила разработаны для трех се-

мейств коричневых карликов, что позволяет сразу проводить предварительную классификацию между ранними и поздними спектральными классами коричневых карликов. При анализе разработанных моделей машинного обучения было найдено, что коричневые карлики показывают сильные отличия от других спектральных классов в показателе цвета $(i-y)$, а именно: $(i-y) > 1.88$ является более эффективным цветовым правилом отбора для коричневых карликов в сравнении с аналогичными правилами из литературы. В рамках изучения эффективных температур каталога Gaia DR3 GSP-Phot были созданы флаги качества для эффективных температур с помощью моделей машинного обучения, которые позволяют эффективно выделять наиболее качественные значения температур каталога Gaia DR3 GSP-Phot.

Методология и методы исследования. Для решения поставленных в работе задач были использованы статистические методы и методы машинного обучения для классификации и анализа данных. Результаты анализировались с помощью авторского программного обеспечения. Исследование основано на данных современных фотометрических, астрометрических и спектроскопических обзоров, таких как WISE, 2MASS, DES, Gaia DR3, APOGEE, GALAH, LAMOST. Созданные модели прошли верификацию и сравнение с внешними наборами данных.

Основные положения, выносимые на защиту:

1. На основе данных спектроскопического обзора RAVE, фотометрических и астрометрических данных каталога Gaia EDR3, а также соотношения эффективной температуры и собственного показателя цвета, для 40 площадок на южном небе была определена зависимость поглощения от расстояния. С помощью закона косеканса были получены оценки полного галактического поглощения в 36 из 40 площадок.
2. Были разработаны и апробированы фотометрические правила для поиска коричневых карликов в обзорах WISE, 2MASS и DES. По разработанным правилам обнаружено 49 объектов – кандидатов в коричневые карлики.
3. Показана возможность и перспективы применения моделей машинного обучения к выделению коричневых карликов в фотометрических обзорах WISE, 2MASS и Pan-STARRS. Впервые продемонстрировано, что использование показателя цвета, разницы видимого блеска в фильтрах,

$(i-y)$ каталога Pan-STARRS перспективно для эффективного выделения коричневых карликов среди других объектов в указанных обзорах.

4. С помощью моделей классического машинного обучения, а именно моделей, реализующих алгоритм бустинга (XGBoost, CatBoost, LightGBM), из полной выборки эффективных температур Gaia GSP-Phot в 471 миллионов объектов были выбраны только объекты с температурами, отклоняющимися от эталонных в пределах 250K. Согласно сделанным оценкам, в каталоге Gaia GSP-Phot около 66% (313 миллионов) объектов обладают надежными оценками эффективной температуры.

Достоверность

Достоверность полученных результатов обеспечивается использованием актуальных астрометрических, спектроскопических и фотометрических данных, а также применением хорошо разработанных методов статистического анализа и машинного обучения. Достоверность полученных выводов подтверждается через сопоставление с результатами предыдущих исследований, опубликованных другими авторами.

Апробация работы. Основные положения и результаты, вошедшие в диссертацию, докладывались на российских и международных конференциях:

1. Авдеева А.С., Ковалева Д.А., Малков О.Ю, Некрасов А.Д. “Определение параметров межзвездного поглощения в высоких галактических широтах”, Всероссийская астрономическая конференция, 23 августа – 28 августа 2021, устный доклад, онлайн
2. Авдеева А.С., Ковалева Д.А., Малков О.Ю, Некрасов А.Д. “Процедура аппроксимации для оценки межзвездного поглощения на высоких галактических широтах”, Конференция молодых ученых и специалистов Института астрономии РАН, 16 ноября 2021, устный доклад, онлайн
3. Avdeeva A.S., Kovaleva D.A., Malkov O.Yu. “Combining Gaia, LAMOST and RAVE data for the determination of interstellar extinction”, Stellar Spectroscopy and Astrophysical Parameterization from Gaia to Large Spectroscopic Surveys, 21 сентября – 23 сентября 2021, доклад, онлайн
4. Avdeeva A.S., Karpov S.V., Malkov O.Yu. “Cross-matching of high proper motion L, T & Y brown dwarfs with large photometric surveys” Data Analytics and Management In Data Intensive Domains Conference

(DAMDID), 4 октября – 7 октября 2022, доклад, Санкт-Петербург, ITMO

5. Avdeeva A.S. “Machine learning methods for the search for L&T brown dwarfs in the data of modern sky surveys”, 31 октября – 4 ноября 2022, Astronomical Data Analysis Software and Systems Conference, доклад, онлайн
6. Авдеева А.С., Карпов С.В., Малков О.Ю. “Поиск коричневых карликов в больших фотометрических обзорах”, Современная звездная астрономия, 8 ноября – 11 ноября 2022, устный доклад, КГО, ГАИШ, Карачаево-Черкесия
7. Авдеева А.С., Ковалева Д.А., Малков О.Ю. “Независимая проверка оценок эффективной температуры в каталоге Gaia DR3”, Физика звёзд: теория и наблюдения, 26 июня – 30 июня 2023, устный доклад, ГАИШ МГУ, Москва
8. A. Avdeeva, D. Kovaleva, O. Malkov “Assessing the Reliability of Gaia DR3 Effective Temperatures”, XXV International Conference on Data Analytics and Management in Data Intensive Domains, 24 октября – 27 октября 2023, Москва, HSE University

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 5 из которых опубликованы в рецензируемых научных изданиях, индексируемых в базах данных Web of Science и Scopus.

Публикации в Web of Science:

1. *Avdeeva A. S.*, *Kovaleva D. A.*, *Malkov O. Y.*, *Zhao G.* Quality flags for GSP-Phot Gaia DR3 astrophysical parameters with machine learning: effective temperatures case study // Monthly Notices of the Royal Astronomical Society. — 2024. — Янв. — Т. 527, № 3. — С. 7382—7393.
2. *Avdeeva A.* Machine learning methods for the search for L&T brown dwarfs in the data of modern sky surveys // Astronomy and Computing. — 2023. — Окт. — Т. 45. — С. 100744.
3. *Avdeeva A. S.*, *Karpov S. V.*, *Malkov O. Y.* Searching for Brown Dwarfs in Large Photometric Surveys: WISE, 2MASS, and DES // Astrophysical Bulletin. — 2023. — Июнь. — Т. 78, № 2. — С. 209—216.

4. *Malkov O. Y.*, *Avdeeva A. S.*, *Kovaleva D. A.*, *Nekrasov A. D.* Interstellar Extinction at High Galactic Latitudes: An Analytical Approximation // Astronomy Reports. — 2022. — Июль. — Т. 66, № 7. — С. 526—534.
5. *Avdeeva A.*, *Kovaleva D.*, *Malkov O.*, *Nekrasov A.* Fitting procedure for estimating interstellar extinction at high galactic latitudes // Open Astronomy. — 2021. — Дек. — Т. 30, № 1. — С. 168—175.

Личный вклад.

Работа 2 была опубликована соискателем самостоятельно без соавторов. В работах 1, 3, 4 и 5 автор участвовал в постановке задачи, анализе данных, интерпретации и обсуждении полученных результатов, подготовке публикации. При этом в работе вклад соискателя определяющий и составляет не менее 80%.

Объем и структура работы. Работа состоит из введения, четырех глав и заключения. Полный объём работы составляет 122 страницы, включая 33 рисунка и 16 таблиц. Список литературы содержит 99 наименований.

Во **Введении** обсуждается актуальность работы и личный вклад автора, описывается достоверность результатов, их апробация, практическая значимость и методы, использованные для их достижения, сделан обзор литературы по теме диссертации, приведен список публикаций автора, а также выписаны решаемые задачи и выносимые на защиту положения.

В **Главе 1** представлено описание разработки фотометрических правил для поиска коричневых карликов в данных больших фотометрических обзоров методом отбора по цвету. В качестве списка известных коричневых карликов используется список, приведенный в работе [10]. Данный список представляет собой перепись коричневых карликов в ближайших 20 парсеках от Солнца, с известными данными фотометрии в полосах 2MASS и WISE (CatWISE). Производится тщательное кросс-сопоставление коричневых карликов из списка с объектами обзора Pan-STARRS. По найденным сопоставлениям описываются области положения коричневых карликов в пространстве параметров - цветов. По описанию областей затем производится поиск коричневых карликов в пересечении обзоров DES, WISE, 2MASS. Результаты этого поиска анализируются на предмет достоверности. 11 объектов из найденных не обнаруживаются в базе данных SIMBAD, что позволяет сделать вывод о том, что эти коричневые карлики обнаружены впервые.

В **Главе 2** описываются и исследуются модели машинного обучения для решения задачи о поиске коричневых карликов в данных больших фотометрических обзоров. Описано составление набора данных для дальнейшего обучения. Данные об известных коричневых карликах взяты из работы [66], данные об объектах других спектральных классов и классов светимости собраны из открытой базы данных SIMBAD. Четыре модели машинного обучения – случайный лес, XGBoost, метод опорных векторов и нейронная сеть TabNet – были использованы для двоичной классификации между объектами положительного класса (коричневых карликов) и объектами других спектральных классов и классов светимости. Эффективность моделей сравнивалась с эффективностью фотометрических правил из литературы, использующихся для отбора коричневых карликов по цвету. На данном наборе данных все модели машинного обучения превзошли методы классификации из литературы. Также были проанализированы распределения важности признаков, используемых моделями. Показатели цвета, которые модели считают важными для классификации, в целом сходятся с отмеченными в литературе признаками. Однако в процессе изучения важности признаков моделей было обнаружено, что показатель цвета $(i - y)_{PS1}$ является наиболее перспективным из исследуемых, ранее важность этого признака в литературе не отмечалась.

Глава 3 посвящена созданию флагов качества для эффективных температур Gaia GSP-Phot. В этой главе рассматривается методика и процедура получения флагов качества для эффективных температур с использованием моделей классического машинного обучения. Эффективные температуры Gaia GSP-Phot сравниваются с эффективными температурами обзоров APOGEE/GALAH. Затем модели классического машинного обучения тренируются отличать эффективные температуры Gaia GSP-Phot, которые сходятся с температурами, усредненными между APOGEE и GALAH, от тех, которые отличаются сильно. Рассматриваются два порога качества температур: температуры сходятся в пределах 125 K и 250 K. Модели показали лучшую результативность при установлении порога в 250 K, что может указывать на то, что случайные ошибки в данных эффективных температур Gaia GSP-Phot превосходят порог в 125 K. Результаты были оценены на дополнительных наборах данных, которые модель не видела при обучении. В каждом случае эффективные температуры объектов, выбранных моделями, сходятся в среднем лучше с эталонными температурами, чем эффективные температуры полного набора

данных, таким образом, использование флагов качества действительно помогает выбрать более качественные значения температур. Флаги качества также вычислены для всего набора данных Gaia GSP-Phot, по данным моделей качественными являются 66% эффективных температур Gaia GSP-Phot. В главе обсуждаются возможные ограничения данного метода и задачи, требующие дальнейшего исследования.

В **Главе 4** приводится эмпирическая модель поглощения излучения, основанная на данных спектроскопического обзора RAVE и фотометрических и астрометрических данных обзора Gaia. Поглощение света в полосе V вычисляется для отдельных объектов в 40 площадках южного неба, в области покрытия обзора RAVE. Внутри каждой площадки полученные поглощения аппроксимируются в зависимости от расстояния законом косеканса (1), аппроксимация была успешно проведена для 36 из 40 исследуемых площадок. Для 4 оставшихся площадок аппроксимацию не удалось произвести, вероятно, из-за низкого качества данных. Полученные результаты сравниваются с двумерной картой поглощения [67]. Параметры закона косеканса, полученные в 36 площадках, аппроксимируются затем полиномом из сферических функций по всей небесной сфере.

В **Заключении** излагаются итоги выполненного исследования, выводы, рекомендации, перспективы дальнейшей разработки темы, а также приводятся **Благодарности**.

Глава 1. Фотометрические правила поиска коричневых карликов в каталогах

Эта глава посвящена разработке фотометрических правил для поиска коричневых карликов в обзорах 2MASS, WISE и DES и поиску коричневых карликов в этих обзорах на основе разработанных правил. Основные результаты исследований опубликованы в работе Avdeeva A. S., Karpov S. V., Malkov O. Y. Searching for Brown Dwarfs in Large Photometric Surveys: WISE, 2MASS, and DES // Astrophysical Bulletin. — 2023. — Июнь. — Т. 78, № 2. — С. 209—216. Фотометрические полосы обзоров представлены на Рис. 1.1.

2MASS — это инфракрасный обзор всего неба, наблюдения которого проводились с 1997 по 2001 год. Обзор 2MASS сделан с помощью двух высокоавтоматизированных 1.3-метровых телескопов: один на горе Хопкинса, Аризона и один в Межамериканской обсерватории Серро-Тололо, Чили. Каждый телескоп оснащен трехканальной камерой, каждый канал состоит из матрицы детекторов HgCdTe (ртутно-кадмиево-теллуриевых) размером 256 × 256, способных одновременно наблюдать небо на J (1.25 микрона), H (1.65 микрона) и Ks (2.17 микрона). Северный телескоп 2MASS начал штатную работу в июне 1997 года, а южный - в марте 1998 года. Обзор был завершен для обоих полушарий 15 февраля 2001 г.

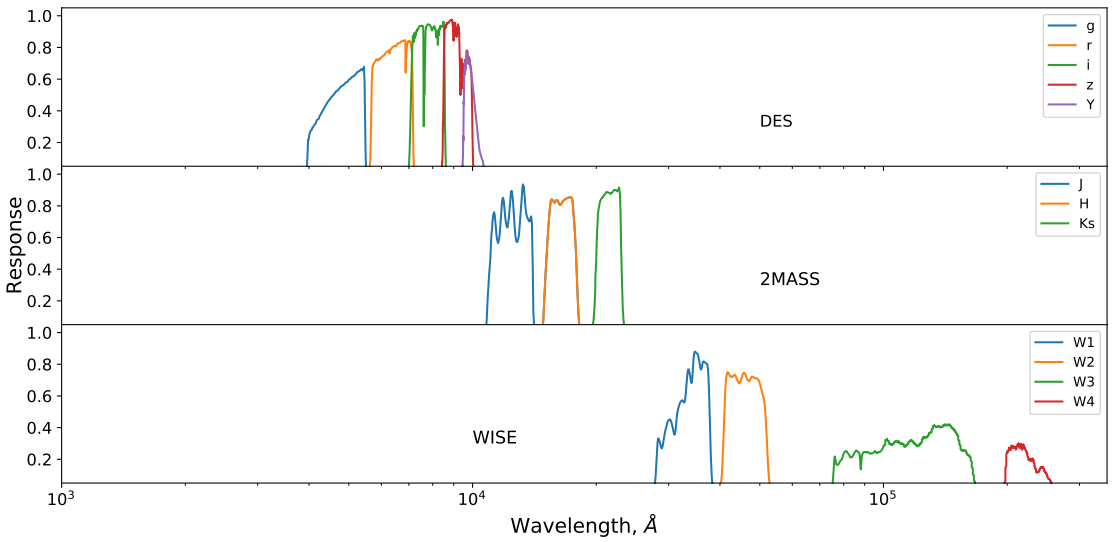

Рисунок 1.1 — Кривые пропускания фильтров обзоров 2MASS, WISE и DES.

Одним из достижений миссии 2MASS стало создание статистической основы для поиска редких, но астрофизически важных объектов, которые либо холодны и, следовательно, являются чрезвычайно красными (например, коричневые карлики), либо скрыты в оптических длинах волн (например, затененные пылью активные ядра галактик и шаровые скопления, находящиеся в плоскости Галактики).

Миссия WISE была разработана для наблюдения всего неба в инфракрасном диапазоне. Основной инструмент обсерватории WISE — это криогенный пятизеркальный афокальный телескоп диаметром 40 сантиметров с фокусным расстоянием 1.35 метра и полем зрения 47 угловых минут. Телескоп оснащён четырьмя камерами, каждая из которых работает в своём диапазоне: W1 (3.3 мкм), W2 (4.7 мкм), W3 (12 мкм) и W4 (23 мкм). Полоса W2 разработана для обнаружения теплового излучения субзвёздных объектов, таких как коричневые карлики. Телескоп начал свою работу в декабре 2009 года. Данные All-sky WISE стали доступны в марте 2012 года. Финальный выпуск данных включает в себя все данные, полученные с декабря по август 2010 года.

Обзор Dark Energy Survey (DES) был разработан для изучения динамики расширения Вселенной. С этой целью обзор наблюдал большое количество сверхновых, гравитационных линз и скоплений галактик, а также получал данные по распределению галактик по небу. DES исследует небо в оптическом и ближнем инфракрасном диапазоне. Для этого используется пять фотометрических фильтров: g,r,i,z,Y. Наблюдения проводятся с помощью 4-метрового телескопа, расположенного в Межамериканской обсерватории Серро-Тололо, Чили. Выпуск данных 1 (DR1) [68] включает в себя данные, полученные с августа 2013 г. по февраль 2016 г. Объекты, наблюденные DES располагаются на почти 5000 квадратных градусах южного неба.

Как было отмечено во Введении, зачастую для поиска новых коричневых карликов, из-за их низкой светимости, используется метод отбора коричневых карликов по цвету. Этот метод основан на том, что различные объекты имеют различные спектральные характеристики, которые могут проявляться через их цветовые индексы или отношения между их фотометрическими измерениями на различных длинах волн. В большинстве работ используется критерий по цвету $(i-z) > 1.2$ так как этот показатель цвета наиболее вариативен при переходе от M карликам к L карликам. Также показатели цвета $(z - J)$ и $(Y - J)$ показывают хороший результат в задаче разделения M и L карликов. Показатель

цвета $W1 - W2$ показывает сильную вариацию между классами L, T и Y, как и было задумано при дизайне фильтра W2, и может использоваться для дифференциации классов коричневых карликов друг от друга. Однако, каждое из правил по отдельности может быть неэффективным. Тем же самым правилам могут удовлетворять и другие типы объектов, такие как галактики, молодые звездные объекты, мириды и красные карлики.

Для создания специфических критериев выделения коричневых карликов необходима фотометрическая информация в разных фильтрах, покрывающих разные диапазоны спектра. Наиболее информативными с этой точки зрения являются фильтры покрывающие ближний и дальний инфракрасный диапазон.

1.1 Кросс-идентификация объектов с каталогом DES и определение границ в пространстве параметров

В качестве надежного источника коричневых карликов мы используем список, составленный в работе [10] (K2021). Ключевые результаты этой работы включают тригонометрические параллаксы Spitzer для 361 карликов L, T и Y. Эти данные были объединены с данными ранее проведенных исследований для составления списка из 525 известных карликов L, T и Y в пределах 20 парсек от Солнца, в том числе 38 ранее неопубликованных объектов. Авторы работы утверждают, что для объектов спектрального класса раньше T8 и с эффективной температурой выше 600 K, список является статистически полным. Этот список включает в себя данные о фотометрии (звездные величины из 2MASS и WISE), астрометрии (из CatWISE) и спектральной классификации для 496 коричневых карликов.

Как отмечается в работе [23], коричневые карлики могут занимать в пространстве параметров (блесков и цветов) три разных области. Все дело в явлении, называемом L/T - переходом: из-за сложной структуры атмосфер холодных карликов их фотометрические свойства меняются с изменением температуры (или спектрального класса) нелинейно.

На Рис. 1.2 представлены диаграммы: температура - спектральный класс и цвет-спектральный класс для 496 коричневых карликов из K2021. Спектральный класс здесь закодирован числом: SpAd=0...9 для L0...L9, SpAd=10...19 для

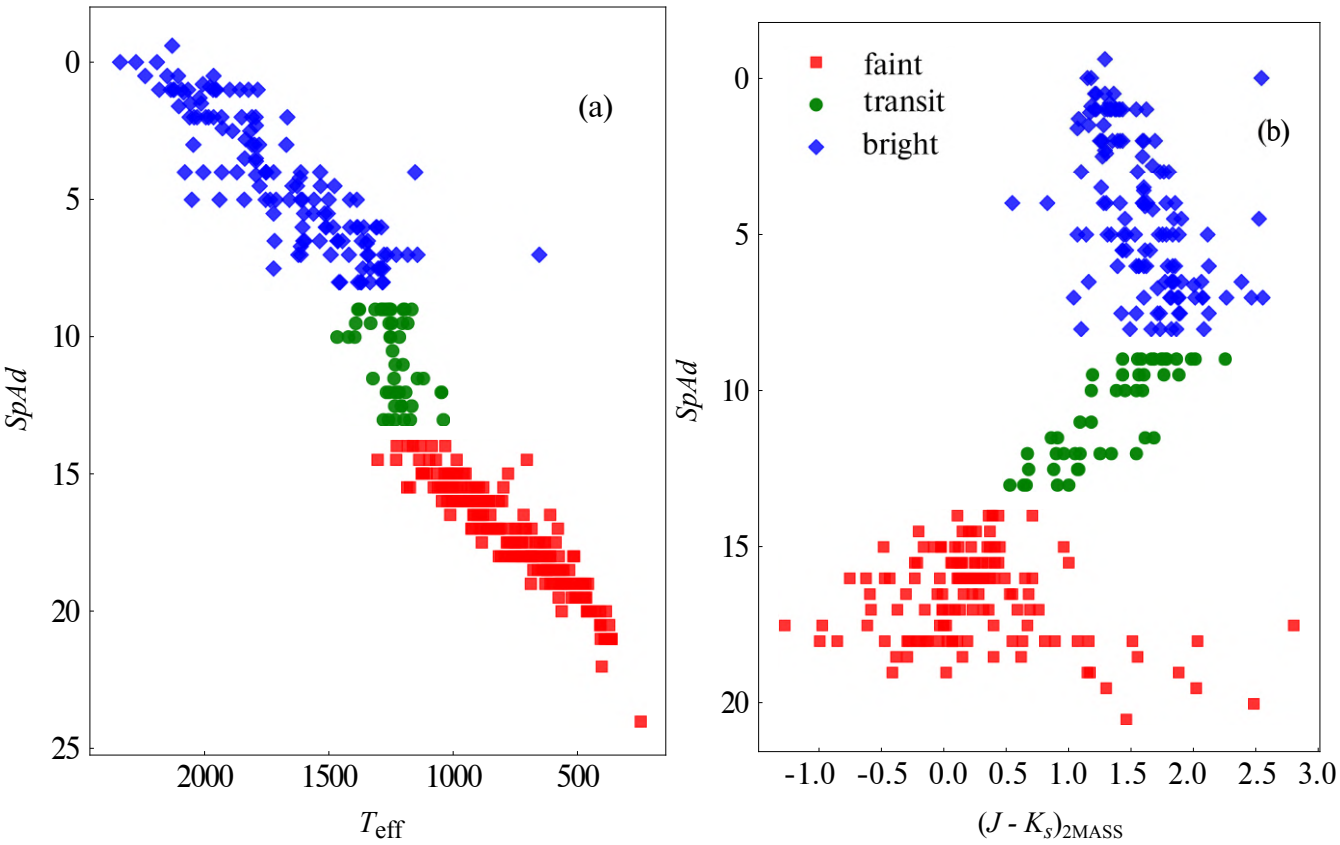

Рисунок 1.2 — Фотометрические, цветовые и спектральные характеристики объектов трех групп. Яркие объекты обозначены синим цветом, транзитные — зеленым, слабые — красным. Описание см. в тексте.

T0...T9, SpAd=20...24 для Y0...Y4. На диаграмме цвет - спектральный класс наблюдается два излома: между L8 и L9, и следующий между T3 и T4. По мере уменьшения эффективной температуры коричневые карлики сначала становятся более красными (что соотносится со смещением максимума функции Планка при уменьшении температуры), а затем, после первого излома, их показатель цвета $(J - K_s)$ начинает смещаться в сторону более голубого цвета.

Определение границы следующего излома - сложная задача из-за значительной дисперсии показателей цвета в этом диапазоне. При этом в литературе объекты спектрального класса позднее T3 не выделяются в отдельную категорию. Однако, сравнивая диаграммы “спектральный класс - температура” и “цвет - спектральный класс”, можно заметить, что коричневые карлики поздних спектральных классов, начиная с $T3$, проявляют поведение, отличное от карликов в процессе L/T перехода. Соответственно, они занимают другие области в пространстве цветовых параметров. Таким образом, мы классифицируем объекты на три группы на основе значения $SpAd$: объекты с $SpAd < 9$ мы называем яркими, с $9 \leqslant SpAd < 14$ - транзитными, и с $SpAd \geqslant 14$ - слабыми.

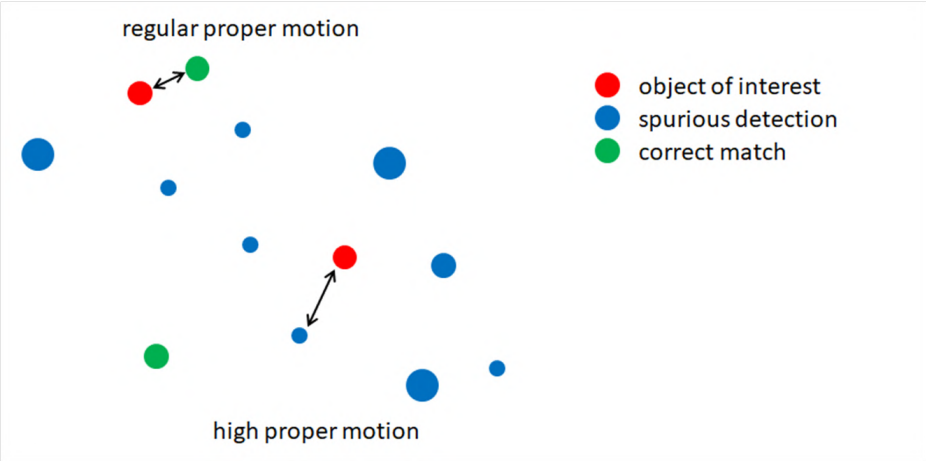

Рисунок 1.3 — Иллюстрация к процессу кросс-сопоставления объектов с большим собственным движением.

Для каждой из этих трех групп устанавливаются фотометрические правила для кросс-идентификации и поиска в обзорах независимо друг от друга. Это крайне желательно, поскольку объекты разных групп имеют различные типичные показатели цвета и зависимости этих показателей друг от друга, хотя фотометрические правила и могут частично пересекаться. Поиск объектов по группам позволяет провести первичную оценку их спектрального класса, так как каждой группе соответствует определенный диапазон спектральных классов.

Кросс-идентификация объектов в различных обзорах осуществляется путем установления однозначного соответствия между наблюдениями тех же самых объектов в других обзорах. Зачастую ключевым аспектом этого процесса является определение радиуса кросс-идентификации - углового расстояния, которое типично для соответствующих объектов в паре обзоров. Однако исследуемые в данной работе объекты - коричневые карлики - являются близкими к Солнцу относительно большинства объектов в больших обзорах неба (список K2021 включает объекты, находящиеся на расстоянии от Солнца не более чем на 20 парсек), и обладают значительными собственными движениями. В результате за время, прошедшее между эпохами наблюдений различных обзоров, использованных в данной работе, объекты могут менять своё видимое положение на несколько десятков угловых секунд.

Для кросс-идентификации объектов K2021 с обзором DES, в этом обзоре был выполнен поиск в радиусе 10 $''$ по координатам, указанным в [10], на эпоху MJD 57170.5. При этом рассматривались все объекты, попавшие в об-

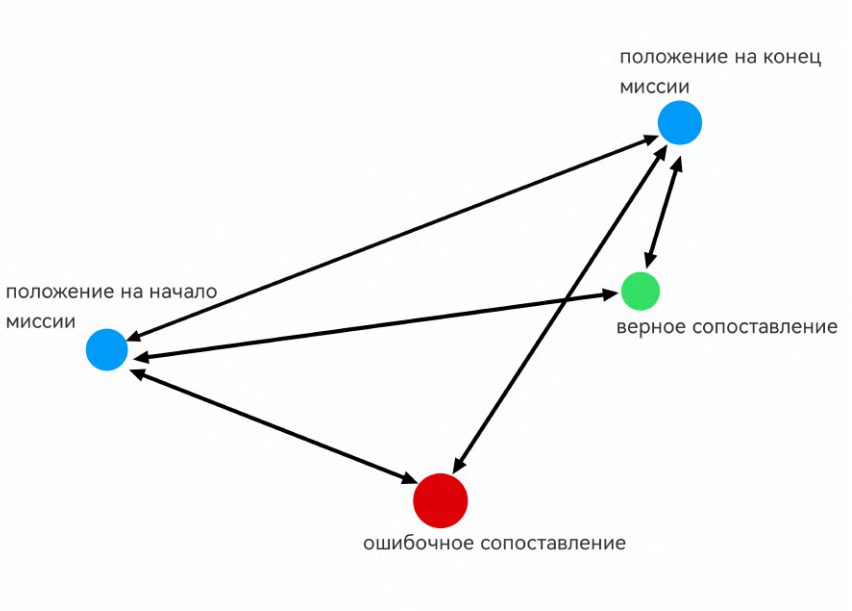

Рисунок 1.4 — иллюстрация к выбору правильного сопоставления методом треугольника.

ласть указанного радиуса, поскольку предполагалось, что в случае больших собственных движений ближайший объект из интересующего нас обзора не обязательно будет правильным сопоставлением. На Рис. 1.1 представлена иллюстрация кросс-сопоставления объектов в разных обзорах. Следовательно, каждому объекту из списка K2021, для которого в радиусе 10 ″ было найдено сопоставление в каталоге DES, может соответствовать несколько записей в этом каталоге, из которых предстоит определить правильное сопоставление.

Неприятной особенностью каталога DES является отсутствие информации о времени проведения наблюдения конкретного объекта или записи. Несмотря на то, что в списке K2021 присутствует информация о координатах и собственных движениях, точное сопоставление координат, эпох наблюдения и величины собственного движения записей из двух каталогов: списка K2021 и DES – невозможно. Для определения наиболее вероятного сопоставления мы вычисляем предполагаемое положение объекта на небесной сфере в начале и конце периода наблюдений (эпоха начала, MJD 56519, и конца, MJD 58492, наблюдений миссии DES), см. Рис. 1.1.

Для каждого кандидата на правильное отождествление вычисляется следующее значение:

$$\Delta = d_{ci} + d_{cf} - d_{if} \tag{1.1}$$

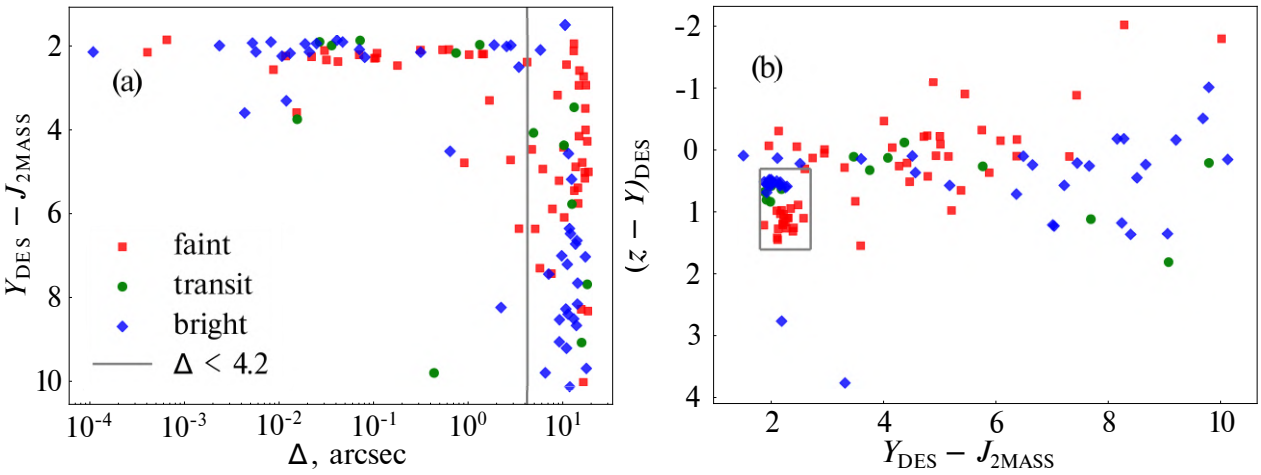

Рисунок 1.5 — Первичная фильтрация всех объектов, попавших в радиус поиска в обзоре DES. Отсечка $\Delta < 4.2$ на графике слева проводится таким образом, чтобы все объекты из отмеченной области на диаграмме $(z - Y, Y - J)$ справа удовлетворяли этому критерию фильтрации.

где d_{ci} - угловое расстояние между положением кандидата на правильное отождествление и предполагаемое положение объекта на небесной сфере в начале наблюдений, d_{cf} - угловое расстояние между положением кандидата на кросс-сопоставление и предполагаемое положение объекта на небесной сфере в конце наблюдений и d_{if} - угловое расстояние между предполагаемыми положениями объекта на небесной сфере в начале и конце периода наблюдений. Чем меньше вычисленное таким образом значение Δ, тем более вероятно, что отождествление проведено верно.

На Рис. 1.5 представлен график зависимости показателя цвета $Y - J$ от Δ (панель a) и диаграмма "цвет - цвет"$(z - Y, Y - J)$ (панель b). Выбор критерия $\Delta < 4.2$ для всех объектов, среди которых осуществляется поиск правильных сопоставлений, производится таким образом, чтобы область повышенной плотности точек на диаграмме $(z - Y, Y - J)$ однозначно проходила этот фильтр. Предполагается, что повышенная плотность объектов в области, попадающей в выделенную область на диаграмме $(z - Y, Y - J)$, свидетельствует о более вероятном правильном их идентифицировании.

После этого проводится анализ положений объектов, которые остались после первой фильтрации, на пяти диаграммах “цвет - цвет”: $(z - Y, Y - J)$, $(r - i, Y - J)$, $(Y - J, J - H)$, $(r - i, i - z)$, $(i - z, z - Y)$ с целью выявления объектов, сильно отличающихся по цветовым характеристикам, для каждой

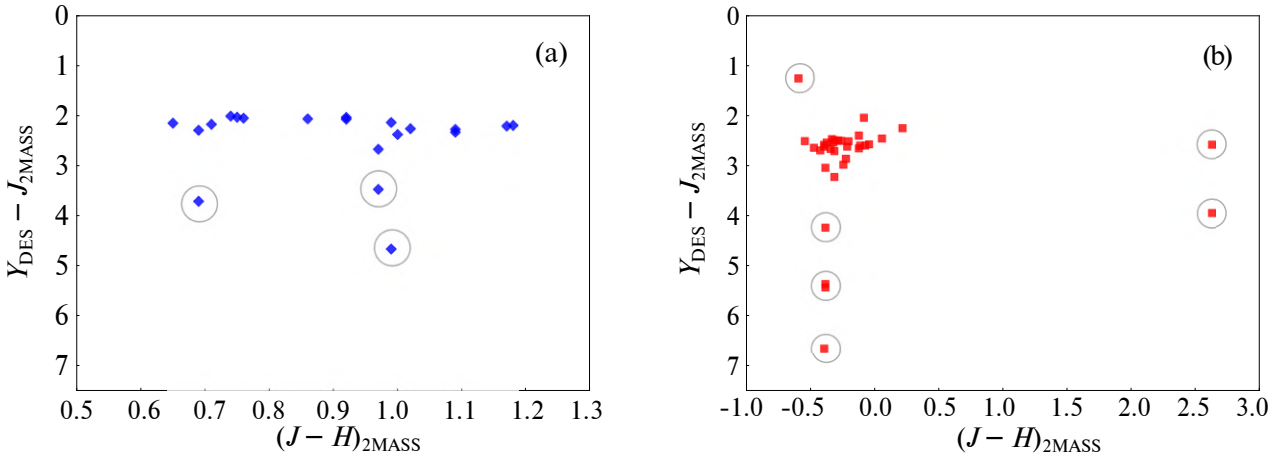

Рисунок 1.6 — Пример выбросов на диаграммах для случая ярких (a) и слабых (b) семейств. Объекты, которые обведены кружками, мы считаем “подозрительными” (вероятнее всего, неправдоподобными) сопоставлениями и отмечаем их специальными флагами.

группы отдельно. На данном этапе исследуются только диаграммы, составленные с использованием показателей цвета с блесками DES.

На Рис. 1.6 приведен пример таких объектов на диаграмме $(Y-J, J-H)$ для яркого и слабого семейств. Кругами выделены объекты, которые считаются выбросами и, следовательно, неправильными сопоставлениями. Каждый объект, отнесенный к выбросам на какой-либо диаграмме, получил соответствующий флаг (для каждой диаграммы свой). Таким образом, сопоставление, отмеченное хотя бы одним флагом, считалось неудачным и не учитывалось в дальнейшей работе. После анализа пяти цветовых диаграмм осталось 56 объектов из обзора DES, соответствующих объектам из K2021, из которых 33 слабых, 18 ярких и 5 транзитных.

Исключив выбивающиеся точки на всех диаграммах, мы описали все области на 9 диаграммах “цвет - цвет”: $(J-H, H-K)$, $(H-K, K-W1)$, $(K-W1, W1-W2)$, $(r-i, i-z)$, $(i-z, z-Y)$, $(z-Y, Y-J)$, $(r-i, Y-J)$, $(Y-J, J-H)$, чтобы получить фотометрические правила для поиска коричневых карликов в трех обзорах. Каждое правило представляет собой набор прямых (вертикальных, горизонтальных и наклонных), ограничивающих область, в которой на диаграмме находятся объекты.

На Рис. 1.7 показаны примеры таких фотометрических правил. Поскольку на предыдущем этапе были удалены все объекты, считаемые ненадежными, прямые проводятся таким образом, чтобы все объекты, для которых можно

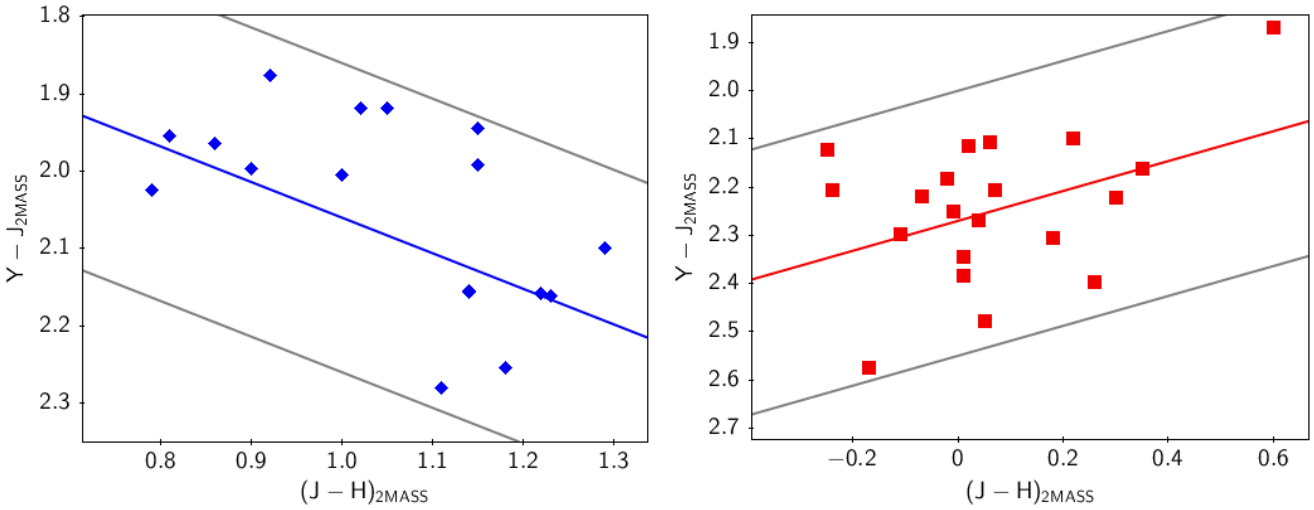

Рисунок 1.7 — Примеры фотометрических правил для поиска коричневых карликов яркого и слабого семейства в данных DES и 2MASS. Прямыми определяется область с “надежными” кандидатами (подробнее см. в тексте).

вычислить соответствующие показатели цвета, попадали внутрь области. Угол наклона границ области определяется линейной аппроксимацией показателей цвета. Если угол наклона $k < 0.1$, то считается, что линейную аппроксимацию проводить нецелесообразно, и область ограничивается только горизонтальными и вертикальными линиями. В Таб. 1 приведен весь список полученных правил.

Нужно отметить, что одни и те же показатели цвета могут повторяться в разных диаграммах, например цвет $Y - J$ повторяется на диаграммах $(z - Y, Y - J)$, $(r - i, Y - J)$ и $(Y - J, J - H)$. Соответственно, на них несколько раз накладываются ограничения на минимальное и максимальное значение. Из-за того, что мы исследуем набор данных, в котором значения (блески в фильтрах и, соответственно, показатели цвета) могут отсутствовать, набор точек, по которому формируется правило, не обязан быть идентичным для разных диаграмм. Поэтому ограничения на минимальное и максимальное значение могут на разных диаграммах отличаться. В таблице мы во всех таких случаях указываем самый широкий из возможных диапазонов везде, где повторяющийся показатель цвета встречается несколько раз.

Таблица 1 — Сводная таблица фотометрических правил для поиска коричневых карликов

Яркие	**Транзитные**	**Слабые**
JHK		
$-0.01 < (J-H)-(H-K) < 0.75$ $0.56 < J-H < 1.62$ $0.2 < H-K < 1.05$	$0.28 < (J-H)-0.9(H-K) < 1.03$ $0.4 < J-H < 1.6$ $0 < H-K < 0.84$	 $-0.9 < J-H < 1$ $-1.4 < H-K < 2.7$
$HKW1$		
$0.05 < (H-K)-0.42(K-W1) < 0.6$ $0.26 < K-W1 < 1.24$ $0.2 < H-K < 1.05$	$-0.42 < (H-K)-0.78(K-W1) < 0.16$ $0.35 < K-W1 < 1.15$ $0 < H-K < 0.84$	$0.55 < (H-K)+0.83(K-W1) < 2.15$ $-1.7 < K-W1 < 2$ $-1.4 < H-K < 2.7$
$KW1W2$		
$-0.3 < (K-W1)-1.62(W1-W2) < 0.45$ $0.26 < K-W1 < 1.24$ $0.17 < W1-W2 < 0.67$	$0.7 < (K-W1)+0.44(W1-W2) < 1.4$ $0.35 < K-W1 < 1.15$ $0.3 < W1-W2 < 1.32$	$-0.75 < (K-W1)+0.29(W1-W2) < 3$ $-1.7 < K-W1 < 2$ $0.7 < W1-W2 < 4.7$
$W1W2W3$		
 $0.17 < W1-W2 < 0.67$ $-0.44 < W2-W3 < 1.29$	 $0.3 < W1-W2 < 1.32$ $0.54 < W2-W3 < 1.68$	$0.3 < (W1-W2)-0.46(W2-W3) < 3.9$ $0.7 < W1-W2 < 4.7$ $0.7 < W2-W3 < 3.5$
riz		
$0.01 < (r-i)-0.69(i-z) < 1.4$ $1 < r-i < 2.65$ $1.2 < i-z < 2.25$	 $2.05 < r-i < 4.45$ $2.15 < i-z < 3.05$	$-2.8 < (r-i)-0.6(i-z) < 5.2$ $-0.2 < r-i < 7.4$ $0.3 < i-z < 4.45$
izY		
$1.2 < i-z < 2.25$ $0.45 < z-Y < 0.7$	$2.15 < i-z < 3.05$ $0.57 < z-Y < 0.84$	$0.3 < i-z < 4.45$ $0.7 < z-Y < 1.55$
zYJ		
 $0.45 < z-Y < 0.7$ $1.86 < Y-J < 2.3$	 $0.57 < z-Y < 0.84$ $1.85 < Y-J < 2.2$	$1.57 < (z-Y)+0.29(Y-J) < 2.07$ $0.7 < z-Y < 1.55$ $1.8 < Y-J < 2.6$
$riYJ$		
$1 < r-i < 2.65$ $1.86 < Y-J < 2.3$	$2.05 < r-i < 4.45$ $1.85 < Y-J < 2.2$	$-0.2 < r-i < 7.4$ $1.8 < Y-J < 2.6$
YJH		
$1.4 < (Y-J)-0.46(J-H) < 1.8$ $1.86 < Y-J < 2.3$ $0.56 < J-H < 1.62$	 $1.85 < Y-J < 2.2$ $0.4 < J-H < 1.6$	$2 < (Y-J)+0.31(J-H) < 2.55$ $1.8 < Y-J < 2.6$ $-0.9 < J-H < 1$

1.2 Поиск коричневых карликов в каталогах 2MASS, WISE и DES

По разработанным правилам был проведен поиск коричневых карликов в обзорах AllWISE, 2MASS и DES DR1. Из обзоров были отобраны объекты, находящиеся друг от друга не далее, чем 50 $''$, и соответствующие вышеприведенным правилам. При этом все правила считались обязательными, а значит если у объекта отсутствует блеск в какой-либо из полос от r до $W3$, он автоматически не проходит наш фильтр из критериев.

Поиск привел к обнаружению 174 записей, удовлетворяющих нашим условиям. Затем мы проверили соответствие координат предположительно одного и

того же объекта в разных обзорах его собственному движению. Для этого было определено расстояние между положениями по координатам в AllWISE и большим кругом, проведенным через положения в DES и 2MASS. Были отброшены все объекты, где это расстояние превышало 1 ″. В результате осталось 137 из 174 объектов. Затем мы сравнили расстояния от точки в DES до AllWISE и 2MASS. Расстояние до положения в AllWISE должно находиться в интервале от 0.2 до 0.35 расстояния от 2MASS (учитывая позиционную точность 2MASS и AllWISE в 1 ″, а также условную абсолютную точность DES). Это привело к отбрасыванию еще двух объектов. В результате у нас осталось 135 кандидатов, позиционно согласующихся с гипотезой о собственном движении.

Для анализа полученных объектов было использовано кросс-сопоставление по координатам с радиусом 1.5 ″, доступное в NoirDataLab, которое присутствует для 96 из 135 объектов. Учитывая, что простое кросс-сопоставление по координатам может быть недостаточно эффективным, было проведено сравнение блесков объектов в DES и их соответствий в Gaia (см. Рис. 1.8a). Большинство объектов показывает хорошее согласие в блесках i_{DES} и G_{Gaia}. Заметно, что расхождение увеличивается с уменьшением блеска объекта в обоих обзорах. На Рис. 1.8b представлено сопоставление расстояний между положениями объектов в DES и 2MASS и их собственных движений из Gaia. Сравниваемые величины в целом согласуются качественно, а учитывая среднюю разницу в эпохах обзоров, и количественно.

Также было проведено сравнение объектов, для которых нашлись и не нашлись сопоставления в Gaia, по величине собственного движения и блескам (см. Рис. 1.9). Собственные движения вычислялись следующим образом:

$$\mu_{tot} = d_{dt}/15 \tag{1.2}$$

где d_{dt} - расстояние между положениями объектов в обзорах 2MASS и DES в миллисекундах дуги, а μ_{tot}, соответственно, величина собственного движения в миллисекундах дуги в год. Средняя разница в эпохах между 2MASS и DES взята 15 лет.

Из распределения собственных движений видно, что большие собственные движения не являются причиной, по которой для 39 объектов не было найдено данных в архиве Gaia. В отличие от собственных движений, блески объектов, для которых было найдено сопоставление в Gaia и для которых сопоставления не было найдено, значительно отличаются. Несмотря на некоторые

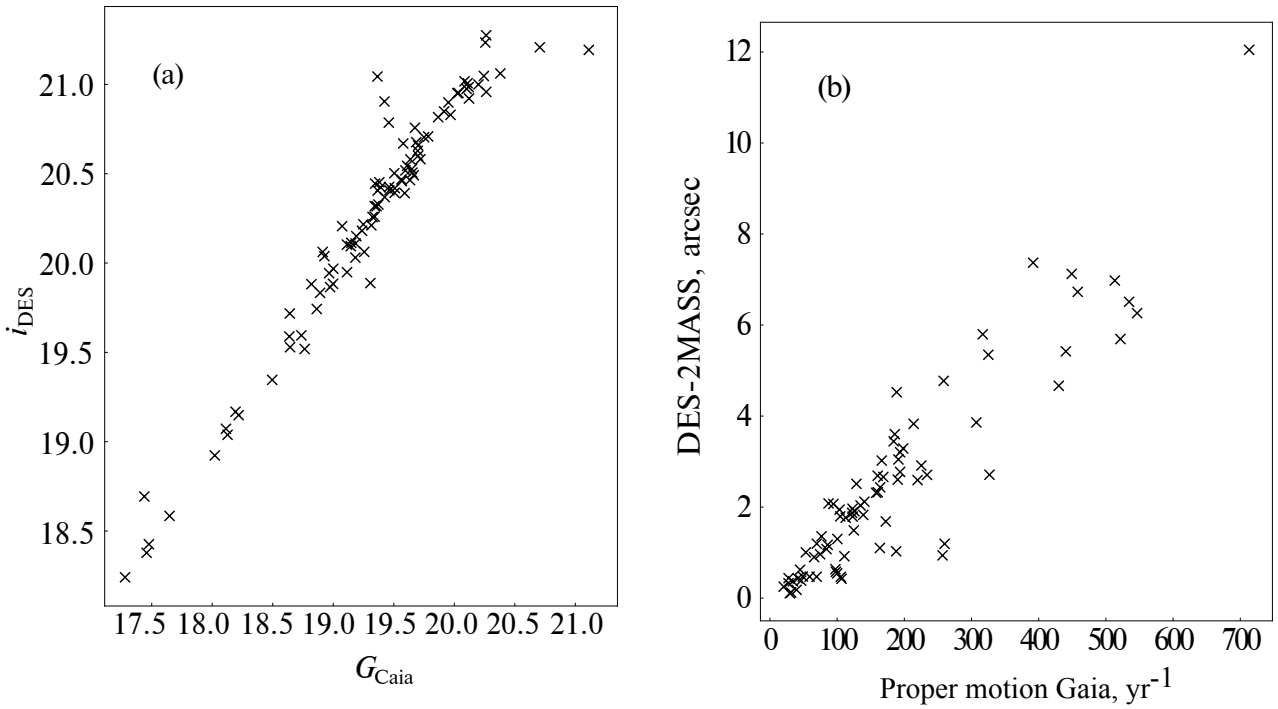

Рисунок 1.8 — Сравнение данных потенциальных коричневых карликов и отождествления с Gaia: фотометрия (a) и собственное движение (b)

исключения, блески "найденных" объектов в среднем ярче, чем блески остальных объектов. На основании этого делается вывод о том, что глубина обзора Gaia может быть недостаточной для идентификации, по крайней мере, трети коричневых карликов.

Каждый из 39 объектов, которые не удалось найти в обзоре Gaia, мы попробовали найти в базе данных SIMBAD. В базе данных удалось найти 25 объектов, большинство из которых имеют подтвержденную спектральную классификацию, относящуюся к коричневым карликам, что подтверждает правильность выбранных правил. При этом 11 объектов не обнаруживаются в базе данных SIMBAD, параметры этих объектов представлены в Таб. 2.

Также 13 объектов, для которых существует сопоставление из обзора Gaia, не обнаруживаются в Simbad. Для 19 обнаруженных объектов спектральный тип в Simbad не определен, что также может свидетельствовать о том, что это новые кандидаты в коричневые карлики. Кроме того, среди найденых нами объектов, 10 имеют в Simbad статус кандидатов в коричневые карлики, и наша работа, таким образом, подтверждает этот статус.

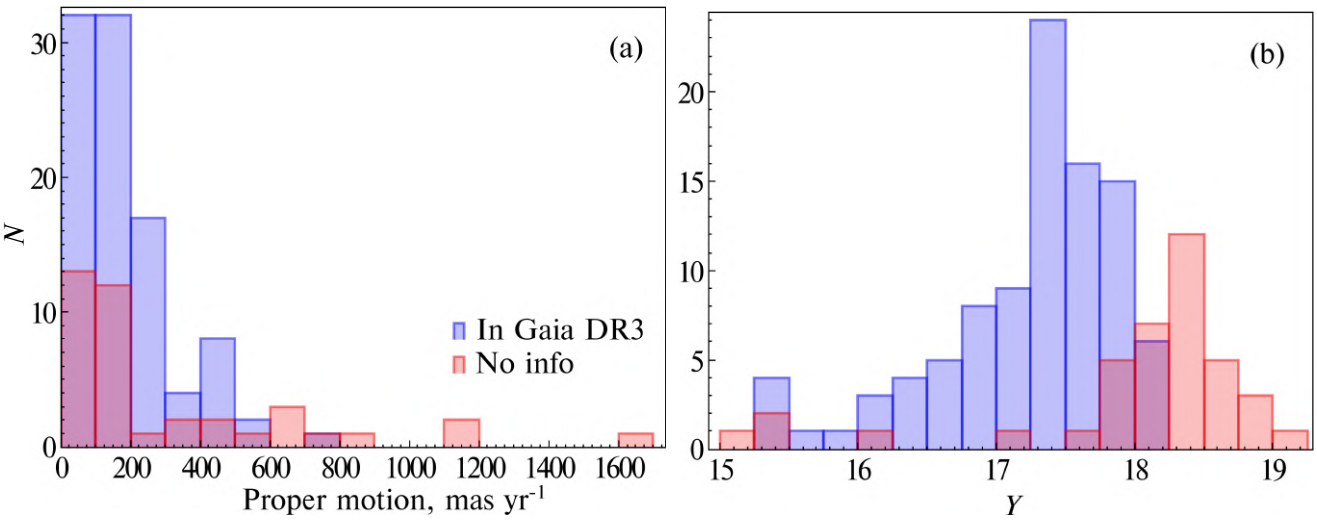

Рисунок 1.9 — Распределения для объектов, имеющихся и отсутствующих в архиве Gaia: по собственным движениям, вычисленным по расстоянию между положениями в обзорах DES и 2MASS, и блескам согласно данным DES, панели (a) и (b) соответственно.

Таблица 2 — Данные о найденных коричневых карликах, не обнаруженных в базе данных SIMBAD

id_2mass	id_wise	id_des	ra_des	dec_des	z_des	J_2mass	H_2mass
01540148-0637137	0288m061_ac51-007963	261505169	28.50612	-6.62037	18.65281	16.099	15.135
05211425-3733317	0795m379_ac51-049013	425288238	80.30951	-37.55902	19.03073	16.554	15.525
05480851-3752031	0871m379_ac51-030641	446411451	87.03542	-37.86712	19.31803	16.54	15.42
04543338-3458339	0730m349_ac51-029700	493456714	73.63929	-34.97611	18.90746	16.237	15.513
04554443-2501073	0733m258_ac51-053558	396620130	73.93524	-25.01873	19.57393	16.944	15.818
05090167-2017083	0768m197_ac51-002792	413129201	77.2571	-20.28571	18.99485	16.243	15.144
03085441-6214377	0473m621_ac51-031675	327445094	47.22682	-62.24364	18.83113	16.093	15.235
05174170-5158282	0786m515_ac51-002683	419155737	79.42339	-51.97399	18.46643	15.91	14.905
05210460-4955309	0812m500_ac51-036722	427542833	80.26883	-49.92517	18.86474	16.423	15.34
04260037-5558408	0666m561_ac51-031412	506500754	66.50152	-55.97804	19.15256	16.596	15.845
02211279-4053437	0356m409_ac51-034238	111688863	35.30309	-40.89558	19.16188	16.489	15.328

1.3 Обсуждение результатов главы

В данной главе представлены разработка фотометрических правил для поиска коричневых карликов в обзорах WISE, 2MASS и DES и поиск объектов, основанный на этих правилах. Для того чтобы разработать фотометрические правила, предварительно было проведено кросс-отождествление переписи коричневых карликов в ближайших 20 пк (K2021) с обзором DES. Фотометрические правила были разработаны для трех семейств коричневых карликов:

слабых, транзитных и ярких, в соответствии с их фотометрическими проявлениями.

По разработанным правилам был проведен пробный поиск коричневых карликов в трех обзорах с учетом выполнения всех правил. Для 96 из 135 объектов, удовлетворяющих нашим критериям, в радиусе $1''.5$ нашлось сопоставление из каталога Gaia DR3. Собственные движения наших объектов, вычисленные исходя из разности положений в обзорах 2MASS и DES, качественно и количественно совпадают с измерениями Gaia. Еще 39 объектов, для которых не нашлось сопоставления в Gaia, по-видимому, являются слишком слабыми для этого обзора. По разработанным правилам обнаружено в общей сложности 43 объекта, которые можно считать новыми кандидатами в коричневые карлики.

Глава 2. Машинное обучение для идентификации коричневых карликов в каталогах

При использовании методов выбора по цвету для поиска коричневых карликов применение машинного обучения может принести значительные преимущества. Техники машинного обучения могут улучшить точность и эффективность процесса выбора по цвету, используя большие наборы данных и сложные алгоритмы для выявления закономерностей. Методы машинного обучения могут помочь в раскрытии тонких взаимосвязей и корреляции в многомерном цветовом пространстве, что позволяет выявлять характерные цветовые признаки, связанные с коричневыми карликами. Это может быть особенно ценно при работе со сложными и перекрывающимися распределениями цветов между различными объектами.

Цель работы, которая является основой данной главы, заключается в разработке инструмента для поиска коричневых карликов в больших фотометрических обзорах с использованием методов машинного обучения. То есть на основе набора величин и цветов объекта модель должна определить, является ли данный объект коричневым карликом или нет. Мы также сравниваем наши результаты с некоторыми правилами цветового отбора из литературы, пользующимися наибольшей популярностью: [6] и [7]. Сводка этих правил приведена в Табл. 4. Основные результаты данного исследования опубликованы в работе Avdeeva A. Machine learning methods for the search for L&T brown dwarfs in the data of modern sky surveys // Astronomy and Computing. — 2023. — Окт. — Т. 45. — С. 100744.

2.1 Построение набора данных и предварительная подготовка

Набор данных для применения машинного обучения основан на каталоге коричневых карликов типов L и T из работы [66]. В каталоге содержится информация о 1601 коричневом карлике типов L и T, а также 8287 красных карликах типа M, спектральный класс которых наиболее близок по физическим характеристикам к коричневым карликам. Предоставляются величины в 12 фо-

тометрических полосах и их погрешности: g, r, i, z, y из обзора Pan-STARRS 1 [69], J, H, Ks от обзора 2MASS [70] и $W1, W2, W3, W4$ от космической миссии WISE [71]. В каталоге также содержится астрометрическая информация: координаты, параллакс и собственное движение. Кроме того, приведены ссылки на литературу, из которой были взяты данные о собственном движении и параллаксе.

Для целей машинного обучения мы считаем коричневые карлики объектами целевого положительного класса. Чтобы создать репрезентативное распределение объектов отрицательного класса, мы изучили распределение 100 тысяч звезд из Gaia DR3 [35; 36] по абсолютной звёздной величине M_G (рис. 2.1a). Мы выбрали 1791 объект от спектрального класса A0 до K9 в пропорциях, наблюдаемых на Рис. 2.1a, из базы данных астрономических объектов SIMBAD[1]. Объекты, представленные в SIMBAD, обычно хорошо изучены и имеют надёжные спектральные классификации. Данные Gaia кажутся недостаточными для M-карликов, особенно позднее M3, поэтому мы берем их распределение из работы [66]. Распределение полученного набора данных показано на Рис. 2.1b.

Объекты, выбранные из SIMBAD, которые не являются объектами целевого класса, были сопоставлены с данными из каталогов Pan-STARRS DR1, 2MASS и AllWISE. Мы выбрали радиус сопоставления $1''$, что является разумным значением для большинства обзоров и объектов с низкими собственными движениями, включая те, которые использовались в данной работе.

Полученный набор данных содержит 5669 объектов, из которых 1601 относятся к положительному классу целевых объектов. Полученный набор данных доступен онлайн[2]. Поскольку пик интенсивности коричневых карликов приходится на инфракрасную часть спектра, их величины в оптических фотометрических полосах (g, r, i), вероятно, находятся за пределами чувствительности телескопа и, следовательно, отсутствуют в данных. Величины g и r из данных Pan-STARRS отсутствуют практически у всех объектов, поэтому мы не используем эти величины как признаки. Значения величины i отсутствуют примерно у трети объектов положительного класса и небольшого числа объектов отрицательного класса. Значения величины в этой полосе важны для нас, а также для сравнения с классическими правилами, поэтому мы их оставля-

[1] http://simbad.cds.unistra.fr/simbad/

[2] https://github.com/iamaleksandra/ML-Brown-Dwarfs/

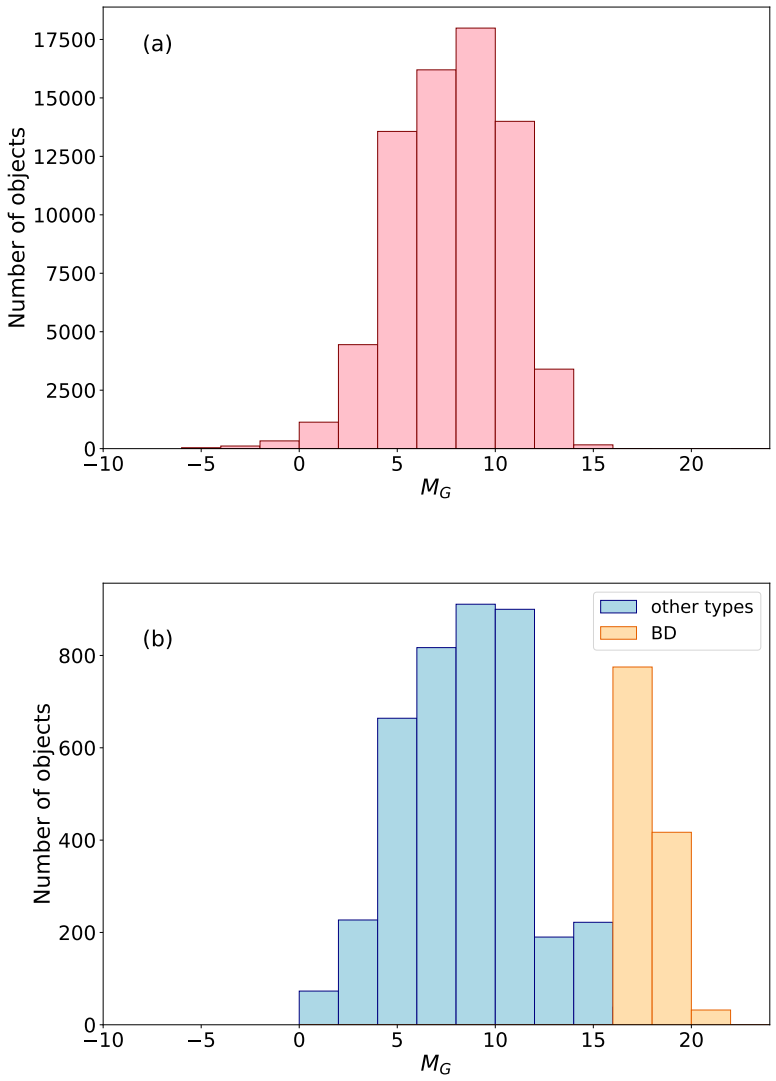

Рисунок 2.1 — Распределение по абсолютным звездным величинам для выборки Gaia (a) и набора данных, использованного в данной работе (b).

ем. Мы также удаляем $W3$ и $W4$, так как они имеют низкое качество: 90-й процентиль ошибки измерения величины для обеих величин составляет около половины звездной величины, что очень много для задачи классификации, которую мы хотим решить. В результате у нас есть 7 величин для каждого объекта: i_{PS1}, z_{PS1}, y_{PS1}, J, H, Ks, $W1$, $W2$.

Данные были разделены на тренировочный (60%), валидационный (20%) и тестовый (20%) наборы в стратифицированном порядке с использованием метода *train_test_split* из библиотеки sci-kit learn. Признаки были масштабированы с помощью *StandardScaler*, а гиперпараметры модели были выбраны на вали-

дационном наборе с использованием *optuna* [72]. Итоговая производительность модели оценивается на тестовом наборе.

Поскольку количество объектов отрицательного класса почти в 2.5 раза превышает количество объектов положительного класса, мы используем аугментацию с добавлением гауссовского шума как метод увеличения выборки для того, чтобы сделать набор данных более сбалансированными. Аугментацией называется дополнение исходного набора данных другими данными, полученными из исходных с помощью различных методов. Мы аугментируем данные положительного и отрицательного классов отдельно, используя все объекты положительного класса и только объекты отрицательного класса с значениями i_{PS1} в диапазоне от 12^m до 15^m. Для каждого признака вычисляем среднее и стандартное отклонение ошибки. Затем генерируем значения шума, распределенные нормально с теми же параметрами. Шум добавляется ко всем значениям соответствующих признаков, не имеющих пропущенных значений. Данные были разделены на тренировочные, валидационные и тестовые подвыборки до аугментации, чтобы модели не видели аугментированные данные в тестовой выборке, а обучались только на оригинальном прототипе этих аугментированных объектов. Таким образом, у нас есть 8364 объекта, из которых 4155 положительных и 4209 отрицательных.

В классификации астрономических объектов, как и в различных типах астрономических задач, цвета объектов даже более важны, чем величины. Цвета характеризуют распределение энергии в спектре и практически независимы от расстояния. Чтобы учесть это при классификации, мы добавили несколько дополнительных признаков - цветовых индексов: $(i-z)_{PS1}$, $(i-y)_{PS1}$, $(z-y)_{PS1}$, $z_{PS1}-J$, $y_{PS1}-J$, $J-H$, $H-Ks$, $Ks-W1$, $W1-W2$. Они также часто используются в качестве цветовых критериев для различения коричневых карликов от других объектов. Мы используем цвета только для наиболее спектроскопически близких фильтров. При этом у нас есть два исключения: $z_{PS1}-J$, так как он часто используется в качестве цветового критерия в литературе, и $(i-y)_{PS1}$, так как он оказался чрезвычайно полезным для классификации.

После этой процедуры в таблице содержится 17 признаков для каждого из 8364 объектов, перечисленных в Табл. 3. Рис. 2.2 показывает, как выглядят объекты целевого класса по сравнению с объектами всех остальных классов в двумерном срезе пространства признаков. Мы используем все величины и цвета одновременно (хотя последние являются линейной комбинацией первых),

поскольку обрабатываем пропущенные значения независимо друг от друга, что иногда нарушает эти соотношения между цветами и блесками.

Как видно из Рис. 2.2a, верхний предел величины i_{PS1} существенно отличается для наших объектов положительного и отрицательного классов. Это связано с процедурой построения набора данных: в SIMBAD в четыре раза больше объектов с $i_{SDSS} > 20$ и спектральным типом позднее $L0$, чем объектов со спектральным типом ранее $L0$. Однако каталоги содержат большое количество тусклых объектов, которые не являются коричневыми карликами. Таким образом, мы по возможности должны избегать моделей, полагающихся в основном на блески Pan-STARRS. Стоит также отметить, что хотя наш набор данных можно назвать сбалансированным по блескам в фотометрических полосах обзоров 2MASS и WISE, видимая звездная величина объекта зависит не только от яркости объекта, но и от расстояния до него. Поэтому предпочтительным будет получить модели, опирающиеся в принятии решений на показатели цвета.

Таким образом, для каждой модели рассматриваются три подхода: все блески и показатели цвета используются в качестве признаков (мы называем это “все признаки”), не используются блески Pan-STARRS (“без величин PS”) и вообще не используются никакие блески (“только цвета”).

Как уже упоминалось ранее, в наборе данных присутствует значительное количество пропущенных значений. В более длинноволновой части спектра это, вероятно, связано с пределом чувствительности телескопа: коричневые карлики являются достаточно тусклыми объектами, и их максимум излучения приходится на инфракрасную часть спектра. Пропущенные значения в части спектра с более короткой длиной волны, по-видимому, возникают из-за низкого качества измерений или артефактов.

Для работы с пропущенными значениями [24] использовали метод заполнения средними значениями и Iterative Imputer из библиотеки Scikit-learn. В их работе было показано, что Iterative Imputer обеспечивает более устойчивые и эффективные результаты с точки зрения классификации.

Мы проверяем выбранный метод, временно исключая значения величин для 5 процентов объектов. Затем искуственно созданные пропущенные значения заполняются с использованием Iterative Imputer и производится сравнение результатов с исходными значениями признака для объекта. В Таб. 3 показаны результаты заполнения с использованием следующих параметров Iterative Imputer:

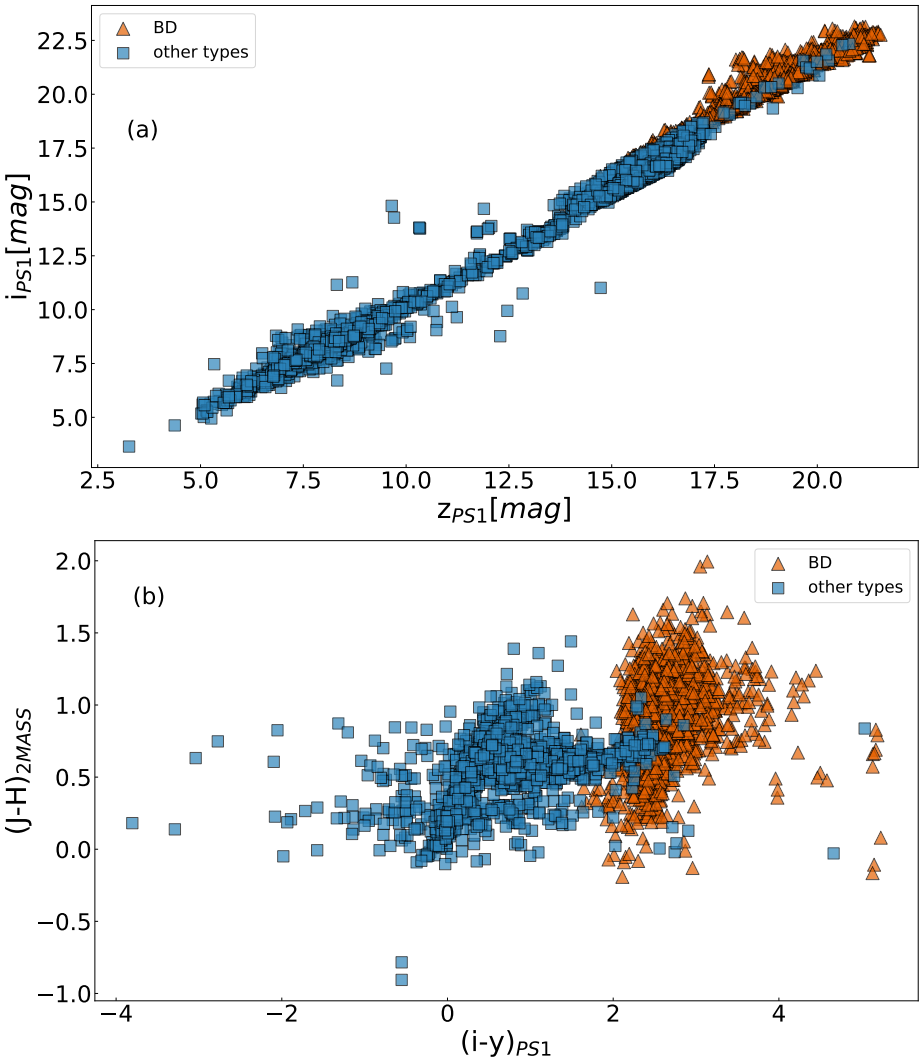

Рисунок 2.2 — Данные с заполненными пропущенными значениями: объекты различных классов на диаграммах блеск-блеск (a) и цвет-цвет (b). Коричневые треугольники представляют объекты положительного класса (коричневые карлики L&T), синие квадраты - объекты отрицательного класса. Здесь синие квадраты отображены поверх коричневых треугольников.

```
estimator=ExtraTreesRegressor
(n_estimators=150
max_features=14
max_depth=15
min_samples_split=12
initial_strategy='median'
max_iter=20)
```

Таб. 3 содержит информацию о доле пропущенных значений определенного признака в наборе данных и количестве объектов, которые были скрыты в целях тестирования метода. Мы также сравниваем 90-й процентиль ошибки измерения признака (ошибка величины обычно указывается в каталоге, а ошибка показателя цвета вычисляется как квадратный корень из суммы квадратов ошибок используемых блесков) с 90-м процентилем расхождения фактического значения признака и значения, предсказанного моделью Iterative Imputer.

В большинстве случаев 90-й процентиль расхождения между заполненными значениями и исходными сопоставим с 90-м процентилем ошибки признака. Несмотря на то, что значения показателей цвета, вычисленные ранее, напрямую связаны с блесками, было решено применять к ним Iterative Imputer независимо. Это позволяет достичь лучших результатов заполнения пропущенных значений и избежать больших ошибок при вычислении показателей цвета, как видно из Таб. 3.

На Рис.2.4 можно видеть пример заполнения пропущенных значений. На верхнем рисунке показана диаграмма звездных величин, содержащая как исходные, так и вставленные моделью данные. Следует отметить, что значительная часть пропущенных значений данных находится в области более слабых звездных величин в полосе i. В нижней панели сравниваются исходные данные с заполненными значениями для тех же объектов. Хотя наблюдаются некоторые расхождения вполть до 0.5^{m} в цвете $y_{PS1} - J$, большинство значений предсказаны достаточно точно. В частности, для 90% исследуемых звезд значение $y_{PS1} - J$ имеет ошибку менее 0.063^{m}.

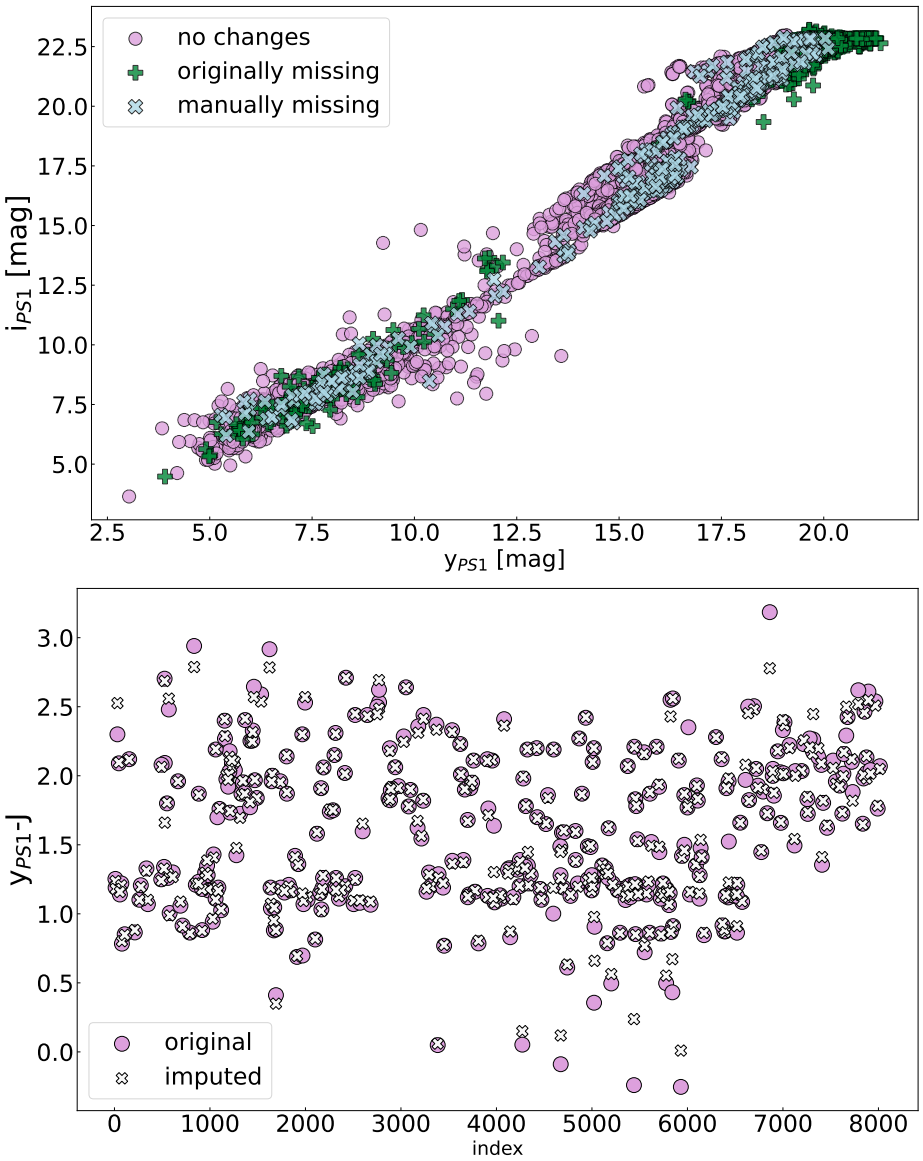

Рисунок 2.3 — Пример заполненных значений для показателя цвета y_{PS1}−J и y_{PS1}. (Верхняя панель) Исходные данные (розовые круги), изначально отсутствующие данные (зеленый знак плюс) и вручную скрытые от модели значения (синие кресты) на диаграмме цвет-величина. (Нижняя панель) Сравнение исходных данных (розовые круги) и значений, заполненных с использованием метода Iterative Imputer (белые крестики).

2.2 Применение машинного обучения

В ходе работы было протестировано четыре подхода: Случайный лес (RF), Метод опорных векторов (SVM), XGBoost и TabNet. Мы исследуем три случая для каждого подхода: “все признаки”, “без блесков Pan-STARRS” и “только цвета”. При этом вычисляется оценка результативности моделей и важность признаков для каждой из моделей на каждом наборе признаков. Для этого используется метод *SHAP* [73]. Несмотря на то, что у TabNet есть встроенные возможности для вычисления важности признаков на основе механизма внимания к признакам, мы также используем *SHAP* для этой модели, чтобы можно было корректно сравнить результаты.

Метод *SHAP* работает путем оценки прогноза модели для каждого случая с перестановкой значений определенного признака. Эта процедура включает в себя перетасовку значений выбранного признака при сохранении остальных признаков неизменными. Разница между прогнозом модели с оригинальными значениями признаков и прогнозом с переставленными значениями признаков используется для вычисления значения чисел Шэпли.

Коэффициент корреляции Мэтьюса (MCC) был выбран в качестве основной метрики, так как он учитывает как ложноположительные, так и ложноотрицательные предсказания. Он может быть вычислен следующим образом:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})}}$$

где:

– TP - количество истинно положительных предсказаний,

– TN - количество истинно отрицательных предсказаний,

– FP - количество ложноположительных предсказаний,

– FN - количество ложноотрицательных предсказаний.

Коэффициент корреляции Мэтьюса принимает значения от -1 до +1, где:

– $+1$ соответствует идеальному прогнозу,

– 0 соответствует случайному прогнозу,

– -1 соответствует полностью неправильному прогнозу.

Эта метрика особенно полезна для несбалансированных классов и позволяет оценить качество двоичной классификации, учитывая все аспекты матрицы ошибок.

Мы также предоставляем оценки точности (precision) и полноты (recall) работы модели, поскольку они более интуитивны для интерпретации. Точность - это доля верно классифицированных объектов среди всех предсказанных как положительные, а полнота - это доля верно классифицированных объектов среди всех истинно положительных объектов. Эти два показателя определяются следующим образом:

$$\textsc{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\textsc{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Классические правила отбора коричневых карликов по цвету из литературы и результат их применения к тестовому набору данных после аугментации и заполнения пропущенных значений собраны в Таб. 4. Оценка MCC при этом рассчитывалась только на тестовой части набора данных. Следует отметить, что хотя фильтры в различных обзорах имеют схожие названия (например, y_{PS1} и Y_{DES}), они не идентичны друг другу. Поэтому здесь важно отметить, что не совсем корректно применять правила, созданные для одного обзора, к звездным величинам и блескам другого обзора. Однако мы оценили, что для нашего набора данных блески в фильтрах с одинаковыми названиями различаются менее чем на 0.2 звездной величины, что ни в каком из случаев не изменяет значительно оценки для правил отбора из литературы. Также стоит отметить, что правило отбора по цвету, предложенное в работе [6], изначально было посвящено исключительно коричневым карликам типа T, однако оно показывает также хорошие результаты на коричневых карликах типа L, поэтому мы используем его в качестве правила принятия отбора как для коричневых карликов типа L, так и для типа T.

Хотя производительность правил отбора по цвету достаточно высока, количество ложноположительных и ложноотрицательных классификаций растет с количеством объектов. Это становится важным, когда у нас есть миллионы объектов, как в большинстве современных обзоров неба. Например, в PanSTARRS содержится 1,9 миллиарда объектов, в 2MASS — 470 миллионов

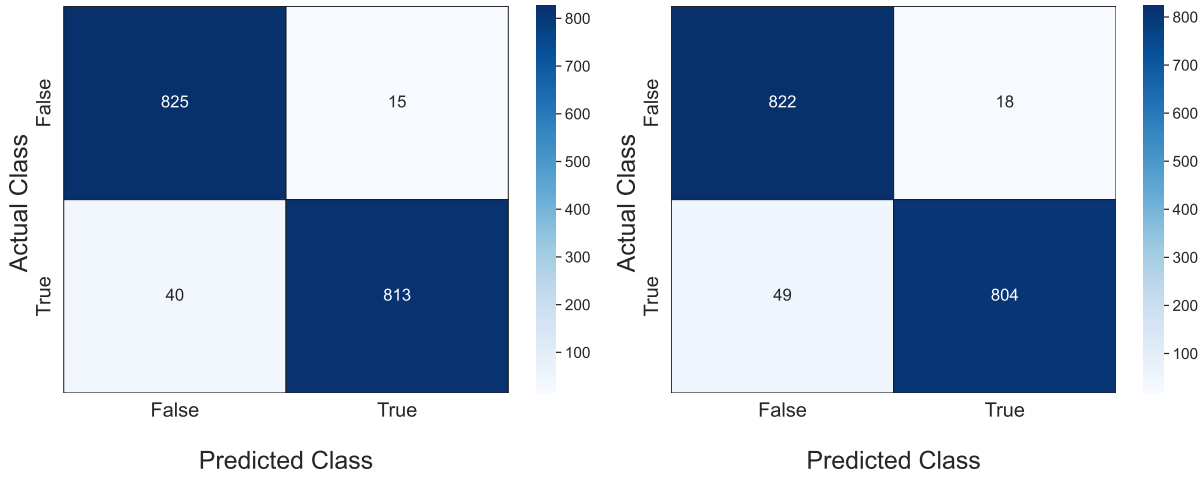

Рисунок 2.4 — Матрицы ошибок для правил отбора по цвету [7] слева и [6] справа, примененные к тестовой части набора данных.

объектов, а в AllWISE — 560 миллионов объектов. Поэтому имеет смысл приложить усилия для повышения производительности. Таким образом, эти значения MCC (Таб. 4) являются базовыми показателями, которые мы хотим превзойти.

2.2.1 Модели

Метод решающего дерева похож по своей концепции с классическими правилами отбора по цвету, которые традиционно используются в астрономии для классификации объектов. И хотя автоматизированные деревья решений могут быть намного более эффективными, чем классические правила, они склонны к переобучению, т. е. слишком хорошо изучают данные, на которых обучаются, и могут дать сбой при применении к данным, которые они раньше не видели. Решением такой проблемы может служить случайный лес (RF) - ансамбль из решающих деревьев. В таком случае решение о том, к какому классу принадлежит объект принимается на основании того, за какой класс проголосовало большее число деревьев.

Бустинг - очень популярный алгоритм машинного обучения. Он представляет собой тип ансамблевого обучения, который использует результат предыдущей модели в качестве входных данных для следующей модели. Вместо того, чтобы обучать модели по отдельности, бустинг обучает модели последо-

вательно, причем каждая новая модель обучается для исправления ошибок предыдущих. На каждой итерации результатам, предсказанным правильно, присваивается меньший вес, а результатам, предсказанным неправильно - больший вес. Затем метод использует средневзвешенное значение для получения окончательного результата.

Также бустинг нам интересен с той точки зрения, что зачастую в алгоритм встроен базовый принцип обработки пропущенных значений, что в нашей работе очень важно. Две популярных модели бустинга - это CatBoost и XGBoost. В CatBoost встроено только заполнение пропущенных значений некоторым конкретным числом, а вот XGBoost использует более хитрую стратегию: каждому узлу приписывается решение по умолчанию и это во многих случаях хорошо работает. Поэтому для работы мы выбрали XGBoost и также сравнили результативность модели на заполненных методом по умолчанию пропущенных значениях и на заполненных IterativeImputer.

На тестовом наборе данных с использованием алгоритма обработки пропущенных значений по умолчанию, XGBoost дает значение MCC = 0.96. Когда обучение и тестирование проводятся на данных, в которых пропущенные значения заполняются с использованием метода Iterative Imputer, производительность достигает MCC = 0.986. Таким образом, мы приходим к выводу, что в этом случае Iterative Imputer не только является более надежным методом, но также оказывает положительный эффект на производительность модели.

Метод опорных векторов (Support Vector Machine, SVM) [74] является еще одним широко используемым и хорошо разработанным методом. Принцип SVM заключается в поиске линии, поверхности или гиперповерхности, которая разделит классы в пространстве признаков. Процесс обучения модели максимизирует расстояние от каждой точки до границы решения (опорного вектора).

TabNet [75] — это нейронная сеть глубокого обучения, которая использует механизм внимания для выбора важных признаков на каждом этапе процесса принятия решений, так что используются только наиболее важные признаки. В этом случае выбор признаков зависит от объекта и, например, может быть разным для каждой строки набора данных обучения. В конечном итоге можно увидеть, на какие признаки модель сосредоточила внимание больше всего.

TabNet состоит из нескольких этапов, каждый из которых представляет собой блок компонентов, при этом количество этапов является гиперпараметром. Каждый этап дает свой голос в окончательной классификации, что

имитирует ансамблевую классификацию. Другие гиперпараметры включают ширину слоя прогнозирования решений (N_d), ширину внедрения внимания для каждой маски (N_a), количество общих блоков с линейными элементами управления на каждом этапе (N_shared) и коэффициент повторного использования признаков в масках (Gamma). Мы подогнали гиперпараметры TabNet с использованием *optuna* и метрики MCC. Подобранные параметры приведены в Табл. 8.

Модель обучается с использованием алгоритма градиентного спуска, при этом используется оптимизатор Adam (Adaptive Moment Estimation), который автоматически регулирует скорость обучения для каждого параметра. Подробнее про метод можно прочитать в работе [76]. Этот метод помогает ускорить процесс обучения и улучшить сходимость модели. В процессе обучения используется часть валидационных данных из набора данных, чтобы предотвратить переобучение, поэтому результатом обучения модели является конфигурация, которая дает лучшие показатели как на обучающих, так и на валидационных данных.

С использованием метода *optuna* мы подобрали параметры для каждой из трех оставшихся моделей. Для модели случайного леса мы выбрали максимальную глубину дерева, минимальное количество образцов, необходимое для разделения внутреннего узла, критерий и максимальное количество признаков в узле. Настроенные параметры и соответствующие оценки представлены в Табл. 5. Для модели XGBoost количество оценивающих деревьев было зафиксировано на уровне 500. С помощью optuna были подобраны максимальная глубина дерева, темп обучения (шаг уменьшения, используемый во время обновления для предотвращения переобучения), коэффициент подвыборки экземпляров обучения, reg_alpha (регуляризационный терм) и gamma - минимальное уменьшение потерь, необходимое для создания дополнительного разделения на листе узла дерева. Оптимизированные значения представлены в Табл. 6. Для модели SVM мы настраивали параметр регуляризации C, тип ядра и коэффициент ядра (‘gamma’ для ядра ‘rbf’). Функция принятия решений была установлена как one-versus-one (‘ovo’), поскольку это бинарная классификация, и веса классов автоматически настраивались обратно пропорционально частотам классов во входных данных. Для настроенных параметров в каждом случае см. Таб. 7. Из таблицы видно, что модели очень похожи для всех случаев, как и наиболее важные признаки (см. Рис. 2.7). Поскольку SVM в основном

опирается на индексы цвета, распределение важности практически не меняется, когда какие-либо или все величины исключаются.

2.2.2 Результаты

На Рис. 2.5 приведен пример двумерного среза, показывающего разделяющую границу между классами, определенную моделями случайного леса, опорных векторов и нейронной сети TabNet. Следует отметить, что это срез в многомерном пространстве параметров, поэтому он не отражает в полной мере производительность модели. Производительность моделей количественно очень хорошая, что означает, что базовые правила принятия решений этой модели могут быть довольно сложными и не могут быть представлены на двумерной диаграмме. Граница разделения в случае с моделью TabNet, как видно, более сложна по сравнению с границами других моделей.

Интервалы достоверности для оценок MCC полученных моделей представлены на Рис. 2.6. Доверительные интервалы рассчитываются методом бутстрэпа со 100 выборками, длина которых составляет половину тестового набора данных. Цветная рамка представляет интервал от 25-го перцентиля до 75-го перцентиля, а медианное значение обозначено черной линией. Полосы ошибок показывают минимальное и максимальное значения, а выбросы обозначены ромбами. Базовые значения, полученные с использованием правил отбора по цвету из литературы, представлены на Рис. 2.6 пунктирными линиями.

Все модели на полном наборе признаков и признаках без блесков PS1 показывают сопоставимые результаты, однако некоторые работают немного лучше. Если использовать только признаки, основанные на показателях цвета, производительность снижается. Тем не менее, такие модели все равно превосходят референсные показатели производительности правил отбора, представленных в литературе.

Хотя результативность моделей практически одинакова, можно сказать, что они отличаются по надежности. Для того, чтобы оценить, насколько модели действительно работают верно, мы анализируем важность отдельных признаков для каждой модели в случаях трех наборов признаков. Важность признаков в случае каждой модели представлена на Рис. 2.7. Случайный лес (Random

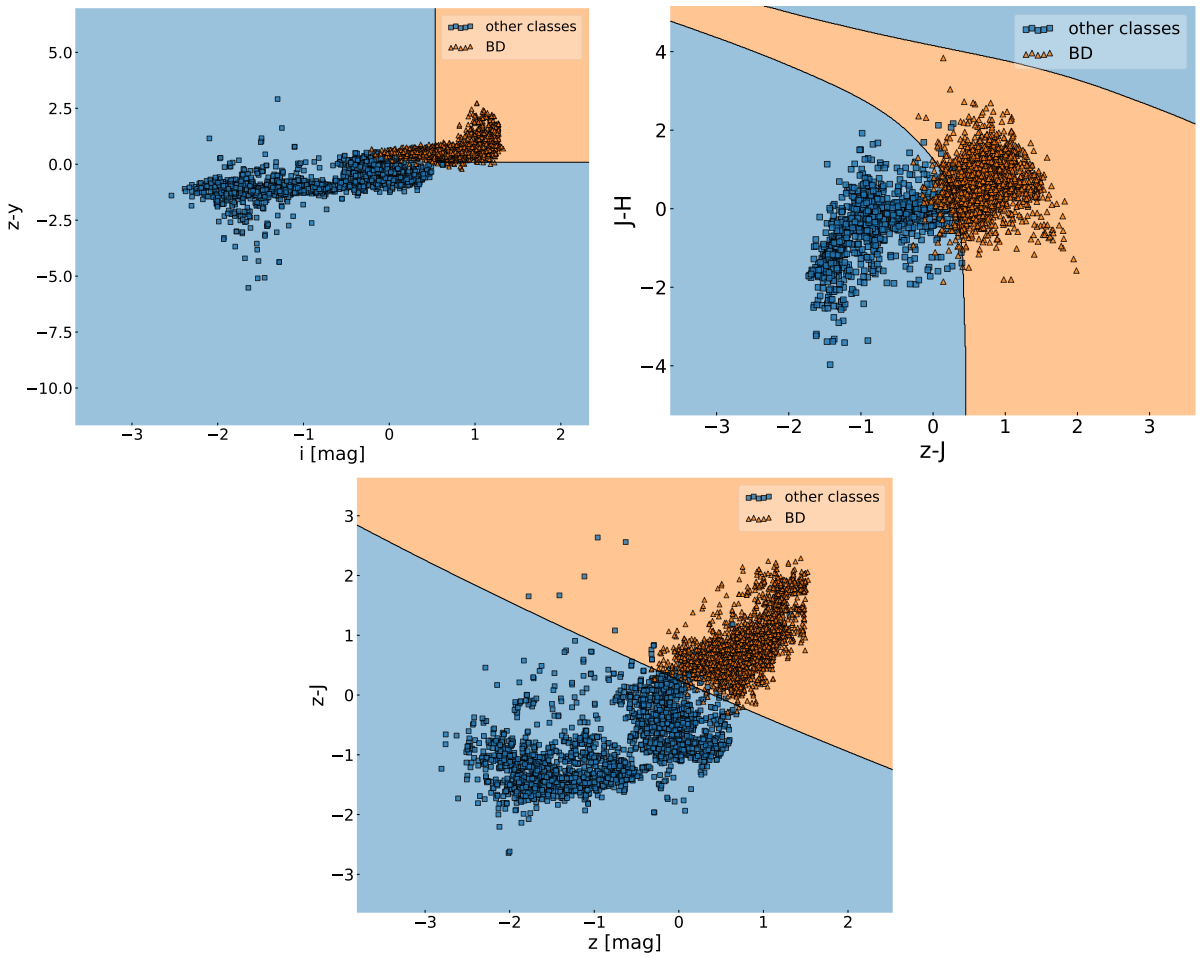

Рисунок 2.5 — Срез разделяющей границы в пространстве признаков в соответствии с моделями случайного леса, модели глубокого обучения TabNet и модели опорных векторов.

Forest) и XGBoost в основном полагаются на блеск в полосе i при принятии решений, если такой присутствует в наборе признаков. В то же время SVM и TabNet на полном наборе признаков представляются более надежными, так как они в основном опираются на цветовые индексы. TabNet также обращает много внимания на блески, хотя в случае использования только цветов в качестве признаков, модель показывает наилучшие результаты среди других моделей. Таб. 9 показывает истинно положительные, истинно отрицательные, ложно положительные и ложно отрицательные значения для всех полученных моделей на тестовой части набора данных.

Исходя из данных работ [7] и [77], ожидалось, что цветовые индексы $(i - z)_{PS1}$ и $y_{PS1} - J$ будут наиболее важными признаками, поскольку они имеют наибольшее изменение при переходе от M к L карликам. В то время как цветовой индекс $(i - z)_{PS1}$ важен для классификатора SVM и в некоторых

случаях для XGBoost и TabNet, большинство моделей не считают его наиболее значимым в решении задачи классификации.

Ожидалось также, что $z_{PS1} - J$ будет самым важным признаком, так как [6], который мы использовали для сравнения, почти целиком основан на этом цвете. Однако для большинства моделей и наборов признаков он играет второстепенную роль. В данной работе мы выявили важность показателя цвета $(i - y)_{PS1}$ для классификации коричневых карликов от объектов других классов, чего не было отмечено в работах ранее. Этот показатель цвета является наиболее важным признаком в большинстве случаев. Более того, дополнительное исследование показало, что условие выбора по цвету $(i - y)_{PS1} > 1.88$ в одиночку дает коэффициент корреляции Мэттьюса (MCC) 0.968 на тестовых данных. Другие показатели цвета, однако, могут быть важны в случае не двоичной классификации, например, карлики типа L и T существенно отличаются по значению показателя цвета $W1 - W2$.

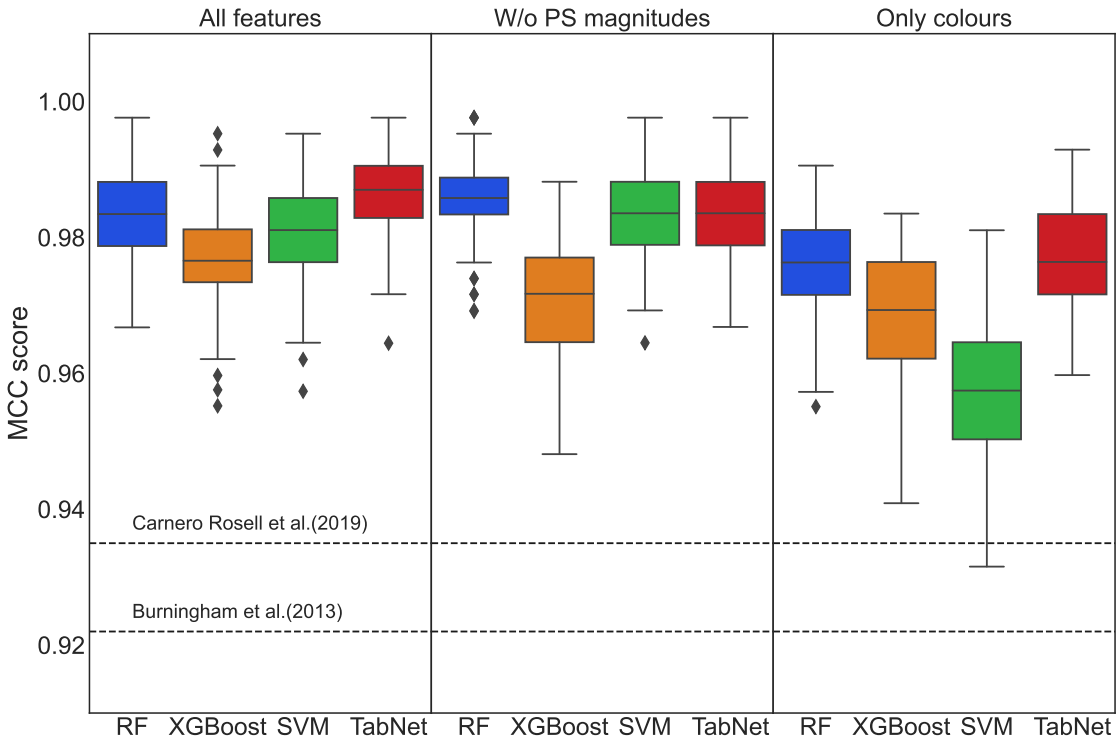

Рисунок 2.6 — Доверительные интервалы для оценок моделей.

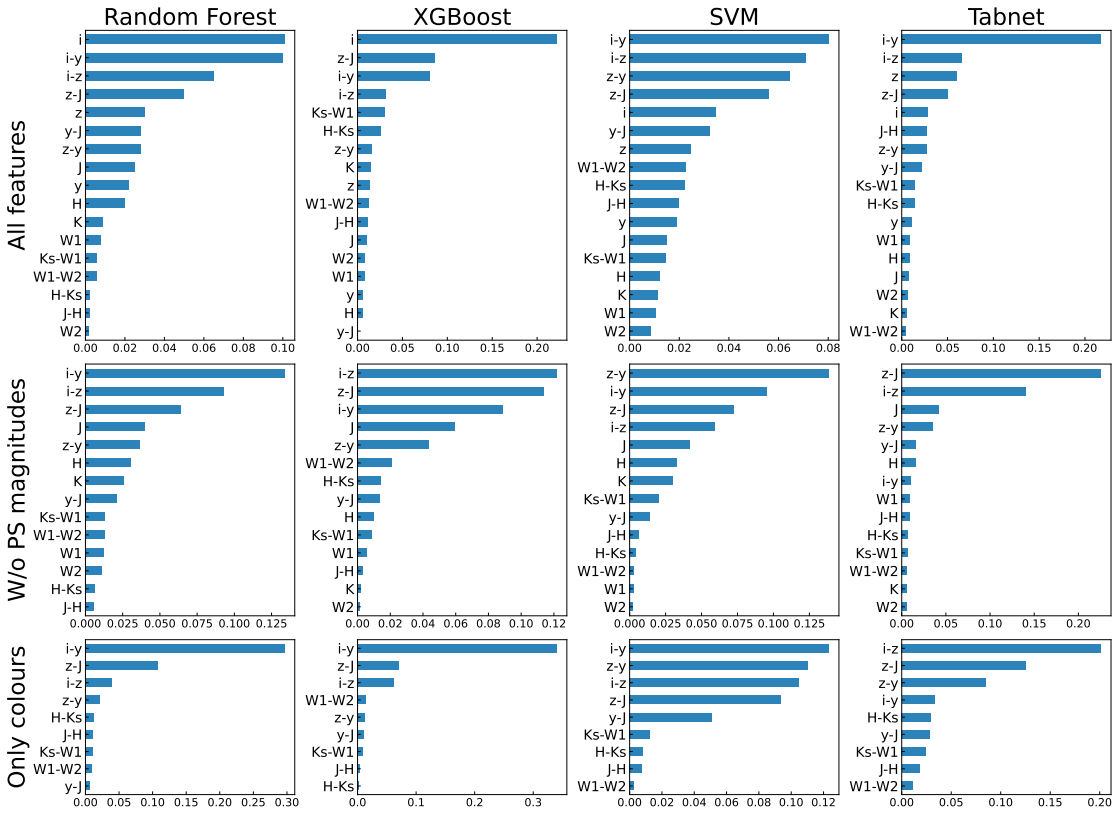

Рисунок 2.7 — Важность признаков для всех моделей. Для моделей RF, XGBoost и SVM мы рассчитываем важность каждого признака с использованием *SHAP*.

2.3 Обсуждение результатов главы

В этой главе мы составили из каталогов и открытой базы данных Simbad набор данных, содержащий фотометрическую информацию о коричневых карликах типов L и T (отмеченных как положительный класс) и объектов других спектральных классов (отмеченных как отрицательный класс). Хотя мы собрали набор данных таким образом, чтобы воспроизвести наблюдаемое распределение по абсолютным звездным величинам, распределение по видимым величинам может быть не репрезентативным и сдвинутым в сторону более ярких объектов. Скорее всего, это не повлияет на производительность модели в случае режима “только показатели цвета”, но может повлиять на модели, обученные на звездных величинах. Неполнота выборки карликов типа M, возни-

кающая из-за ограничений каталога [66], приводит к смещению набора данных в сторону объектов, попавших в выборку, что потенциально влияет на обобщаемость результатов.

Коричневые карлики – слабые астрономические источники, пик интенсивности которых попадает в инфракрасную часть спектра, поэтому проводить наблюдения таких объектов в оптических фильтрах, таких как i и z, сложнее. Вследствие этого, для многих объектов значения блесков, а также соответствующих цветов, отсутствуют. Мы заполнили пропущенные значения блесков с помощью метода Iterative Imputer и исследовали результат. Мы также заполнили значения показателей цвета независимо от соответствующих звездных величин, чтобы уменьшить ошибку заполнения пропущенных значений. Для большинства блесков ошибка заполнения в звездных величинах сопоставима с ошибкой измерения величины, представленной в каталоге. Для показателей цвета ошибка заполнения обычно намного ниже, чем соответствующая ошибка измерения. Таким образом, эта часть предварительной обработки, работа с пропущенными значениями, считается проведенной успешно.

Четыре модели, а именно: Случайный лес, Метод опорных векторов, XGBoost и TabNet Classifier, были обучены отличать коричневые карлики среди всех объектов по их фотометрическим данным. Результаты классификации у всех моделей стабильно высокие, все модели превосходят результативность классификации методами, известными из литературы. Однако модели, основанные на деревьях решений (RF и XGBoost), как правило, используют в значительной степени блески объектов, что менее предпочтительно с точки зрения надежности модели. Напротив, SVM и TabNet в первую очередь опираются на показатели цвета, которые уменьшают возможность неправильной классификации слабых источников других типов.

Изучая особенности моделей, мы обнаружили, что показатель цвета $(i - y)_{PS1}$ является наиболее важным признаком в большинстве случаев. Этот показатель цвета можно использовать впоследствии и независимо в задачах отбора коричневых карликов по цвету. Этот результат не был ранее отмечен в других работах и является совершенно новым. Результаты работы также подтверждают важность показателей цвета $z_{PS1} - J$, $(i - z)_{PS1}$ и $(z - y)_{PS1}$, часто использующихся в задаче классификации коричневых карликов.

Таблица 3 — Характеристики набора данных и результаты тестирования заполнения пропущенных значений для каждого признака в обучающей части набора данных. В таблице представлены доля пропущенных значений в наборе данных и количество объектов, которые были временно скрыты для тестирования (5% объектов, для которых значение признака было представлено). Мы сравниваем 90-й процентиль расхождения между заполненными значениями и исходными значениями с 90-м процентилем ошибки измерения значения соответствующего признака. Ошибка показателя цвета рассчитывается как квадратный корень из суммы квадратов ошибок блесков.

Признак	Доля пропущенных значений	Количество отложенных объектов	90-й процентиль ошибки	90-й процентиль расхождения
i_{PS1}	17%	208	0.050	0.070
z_{PS1}	5.5%	237	0.050	0.091
y_{PS1}	2.2%	245	0.060	0.109
J	8.8%	228	0.100	0.101
H	8.8%	228	0.120	0.107
Ks	8.9%	228	0.110	0.075
W1	2.1%	245	0.141	0.085
W2	2.0%	242	0.080	0.096
$(i\text{-}z)_{PS1}$	18.4%	204	0.067	0.050
$(i\text{-}y)_{PS1}$	18.0%	205	0.072	0.038
$(z\text{-}y)_{PS1}$	6.7%	234	0.073	0.077
z_{PS1}-J	12.3%	219	0.122	0.048
y_{PS1}-J	10.9%	223	0.130	0.063
J-H	8.8%	228	0.164	0.154
H-Ks	9.1%	225	0.170	0.145
Ks-W1	9.9%	226	0.183	0.143
W1-W2	2.2%	245	0.168	0.176

Таблица 4 — Правила отбора по цвету из литературы.

Автор	Правило	MCC
Carnero Rosell et al. (2019)	$(i-z)>1.2$, $(z-Y)>0.15$, $(Y_{AB}-J_{Vega})>1.6$, z<22	0.935
Burningham et al. (2013)	$(z-J)_{Vega}>2.5$, $J<17.5$	0.921

Таблица 5 — Гиперпараметры случайного леса для трех наборов признаков. Количество деревьев составляет 500 для всех моделей. Число в максимальных признаках - это доля всех доступных признаков.

Гиперпараметр	Все признаки	Без признаков PS	Только цвета
Максимальная глубина	11	13	12
Минимальное число объектов для разделения	20	9	8
Максимальные признаки	sqrt	sqrt	0.7
Критерий	энтропия	энтропия	джини
Оценка MCC на тестовом наборе	0.983	0.986	0.975
Оценка MCC на обучающем наборе	0.987	0.990	0.990
Точность	0.992	0.992	0.988
Полнота	0.992	0.994	0.987

Таблица 6 — Гиперпараметры XGBoost для трех наборов признаков.

Гиперпараметр	Все признаки	Без звездных величин PS	Только цвета
Макс. глубина	15	15	10
Темп обучения	0.340	0.126	0.033
Подвыборка	0.05	0.04	0.11
Gamma	0.323	0.548	0.996
Reg alpha	0.82	0.48	0.03
MCC на тесте	0.978	0.972	0.969
MCC на тренировке	0.980	0.978	0.974
Precision	0.992	0.989	0.985
Recall	0.985	0.982	0.985

Таблица 7 — Гиперпараметры классификатора SVM для трех наборов признаков.

Параметр	Все признаки	Без PS-величин	Только цвета
Ядро	rbf	линейное	rbf
C	1.150	0.729	0.792
Gamma	0.298	0.066	0.453
MCC на тесте	0.981	0.984	0.958
MCC на обучении	0.982	0.980	0.968
Точность	0.989	0.986	0.972
Полнота	0.992	0.998	0.986

Таблица 8 — Гиперпараметры TabNet для разных наборов признаков.

Гиперпараметр	Все признаки	Без блесков Pan-STARRS	Только цвета
N_d	16	12	12
N_a	16	28	12
Количество этапов	3	3	2
Gamma	1.2	1.6	1.6
N_shared	2	2	1
MCC на тестовой выборке	0.983	0.986	0.975
MCC на обучающей выборке	0.987	0.990	0.990
Precision	0.992	0.992	0.988
Recall	0.992	0.994	0.987

Таблица 9 — Значения TP (истинно положительные), TN (истинно отрицательные), FP (ложно положительные) и FN (ложно отрицательные), а также показатели Precision и Recall для четырех моделей: случайный лес (RF), XGBoost, метод опорных векторов (SVM) и TabNet. Каждая модель была обучена на трех наборах признаков, обозначенных как “All features” (все признаки), “w/o PS magnitudes” (без блесков Pan-STARRS) и “only colours” (только показатели цвета).

Модель	TP	TN	FP	FN	Precision	Recall
Random Forest						
Все признаки	846	7	846	7	0.992	0.992
Без блесков PS	848	7	833	5	0.992	0.994
Только цвета	842	10	830	11	0.988	0.987
XGBoost						
Все признаки	840	6	834	13	0.992	0.985
Без блесков PS	838	9	831	15	0.989	0.982
Только цвета	840	13	827	13	0.985	0.985
SVM						
Все признаки	846	9	831	7	0.989	0.992
Без блесков PS	851	12	828	2	0.986	0.998
Только цвета	841	24	816	12	0.972	0.986
TabNet						
Все признаки	850	9	831	3	0.992	0.992
Без блесков PS	851	12	828	2	0.992	0.994
Только цвета	846	13	827	7	0.988	0.987

Глава 3. Оценка надежности и флаги качества для температур Gaia DR3 GSP-Phot

В этой главе проводится оценка качества эффективных температур каталога Gaia DR3 модуля GSP-Phot. Мы провели сравнение эффективных температур Gaia DR3 с эффективными температурами из каталогов APOGEE и GALAH для звезд, которые присутствуют как в Gaia DR3, так и в этих каталогах. Пользуясь значениями эффективных температур из каталогов APOGEE и GALAH, мы обучили модели определять, надежным или нет является каждый конкретный результат температуры из представленных в каталоге Gaia DR3 модулем GSP-Phot. Основные результаты, которым посвящена данная глава, опубликованы в работе Avdeeva A. S., Kovaleva D. A., Malkov O. Y., Zhao G. Quality flags for GSP-Phot Gaia DR3 astrophysical parameters with machine learning: effective temperatures case study // Monthly Notices of the Royal Astronomical Society. — 2024. — Янв. — Т. 527, № 3. — С. 7382—7393.

3.1 Сравнение температур Gaia GSP-Phot с температурами APOGEE и GALAH

Каталоги APOGEE DR17 и GALAH DR3, последние версии на момент проведения исследования, предоставляют предварительно выполненную кросс-идентификацию с каталогом Gaia DR3. Для того, чтобы сравнение с данными спектроскопических обзоров было более объективным, мы используем флаги качества, рекомендованные для этих каталогов. Таким образом, мы сравниваем эффективные температуры Gaia GSP-Phot только с надежными значениями из референсных каталогов. Для APOGEE критерии отбора включают ASPCAPFLAG $\neq$ STAR_BAD, что обеспечивает использование только измерений высокого качества с предпочтительным значением отношения сигнал к шуму и отсутствием проблем, связанных с обработкой данных. Для каталога GALAH критерии включают FLAG_SP $= 0$, FLAG_FE_H $= 0$ и SNR_C3_IRAF > 30. Кроме того, мы исключаем записи, не содержащие эффективные температуры

GSP-Phot из Gaia DR3. После этой процедуры у нас осталось 433097 записей, общих между APOGEE и Gaia, и 291065 записей, общих между GALAH и Gaia.

На Рис. 3.1 продемонстрировано сравнение эффективных температур Gaia DR3 с эффективными температурами из обзоров APOGEE(a) и GALAH(b). На графике показано различие между эффективными температурами из обзоров высокого разрешения и эффективными температурами Gaia DR3 GSP-Phot. Здесь и далее под эффективными температурами Gaia подразумеваются только эффективные температуры Gaia DR3 модуля GSP-Phot. Наблюдается значительное расхождение между эффективными температурами Gaia и APOGEE. Эффективные температуры Gaia систематически выше, чем эффективные температуры APOGEE в диапазоне температур APOGEE от 3000 K до 7500 K. Для более горячих звезд выявляется тренд на систематическое занижение эффективных температур Gaia.

Однако, эффективные температуры Gaia показывают в среднем хорошее соответствие с большинством температур GALAH, несмотря на то, что у небольшого числа звезд все-таки наблюдаются сильные отклонения. Это лучшее соответствие, возможно, объясняется критериями выбора целей, наблюдаемых в GALAH, а также применяемыми в обзоре GALAH флагами качества. То есть для тех звезд, для которых эффективные температуры в обзоре GALAH определены и считаются определенными надежно, температуры в каталоге Gaia также показывают хорошие оценки.

Для объяснения значительной разницы в эффективных температурах между Gaia и APOGEE мы провели детальное исследование отличия температур в зависимости от различных параметров. Разница эффективных температур между обзорами представлена, как функция от блеска в фильтре G, показателя цвета $BP - RP$, логарифма ускорения свободного падения на поверхности звезды $\log g$, предоставленного в Gaia и APOGEE, галактической широты b и вычисленного в каталоге Gaia DR3 поглощения на длине волны 541.4 нм, A_0. Визуальное график этих зависимостей представлен на Рис. 3.2.

Некоторые звезды в диапазоне блеска от 12 до 18 G звездной величины проявляют значительные отличия в эффективных температурах. Различия в этой области могут доходить до 35000 K между эффективными температурами Gaia и APOGEE. Возможные систематические ошибки в области слабых звездных величин упоминались также в Главе 11 Документации Gaia DR3 [78]. Авторы указывают, что расхождения могут возникать для $\log g$ при $G > 16$

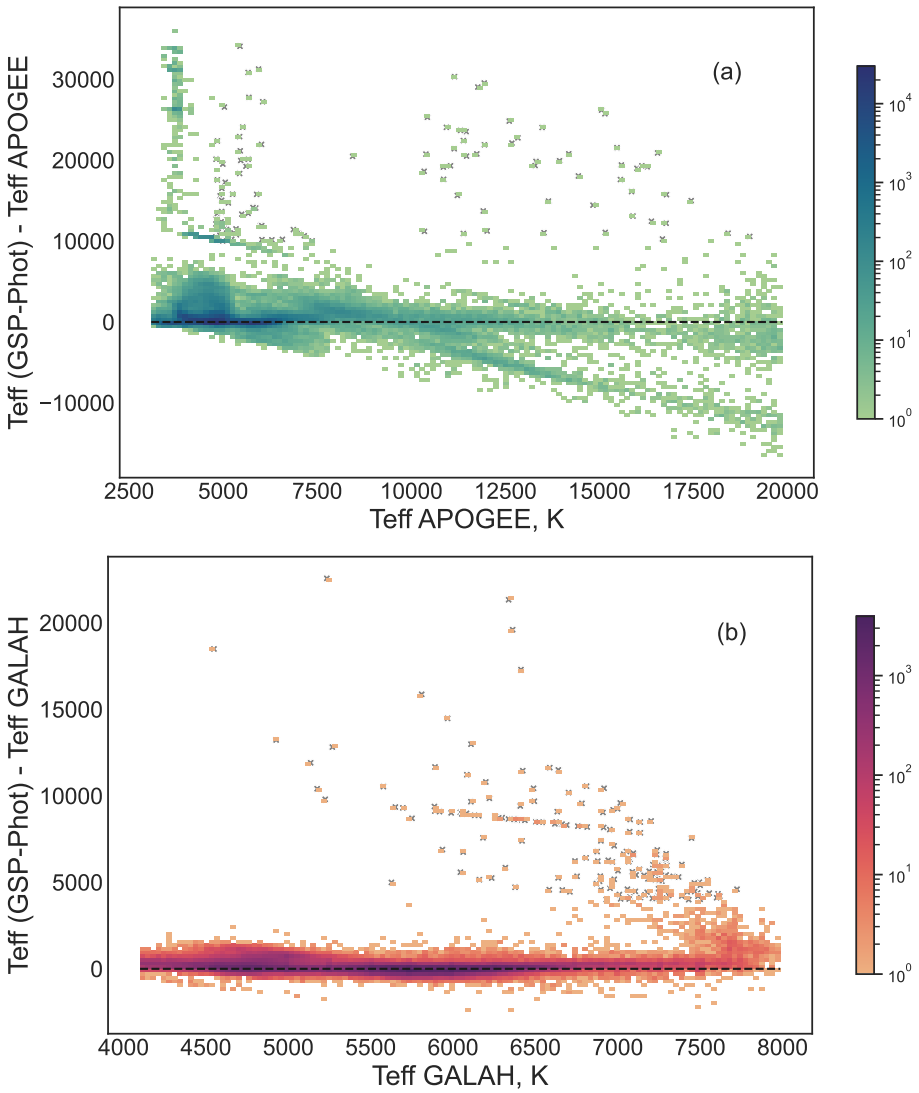

Рисунок 3.1 — Сравнение эффективных температур Gaia с эффективными температурами APOGEE (панель а) и GALAH (панель б). Цвет точек обозначает плотность звезд, а пунктирная линия показывает нулевое отклонение в температурах.

звездной величины. Напротив, звезды с $G < 7$ звездной величины демонстрируют заметное согласие в эффективных температурах между двумя обзорами.

При рассмотрении зависимости от логарифма ускорения свободного падения на поверхности, определенного в APOGEE ($\log g$) и определенного в GSP-Phot ($\log g$(GSP-Phot)) видно, что большое расхождение в эффективных температурах сопровождается также сдвигом в значениях $\log g$. А именно, звезды, для которых эффективная температура GSP-Phot превышает соответствующую температуру на 10000 K и более, $\log g$(GSP-Phot) определяется в пределах от 3 до 4.5 dex, в то время как значения $\log g$ занимают весь диапазон от 0 до 5 dex. Необходимо отметить, что у звезд с наибольшими отклонениями

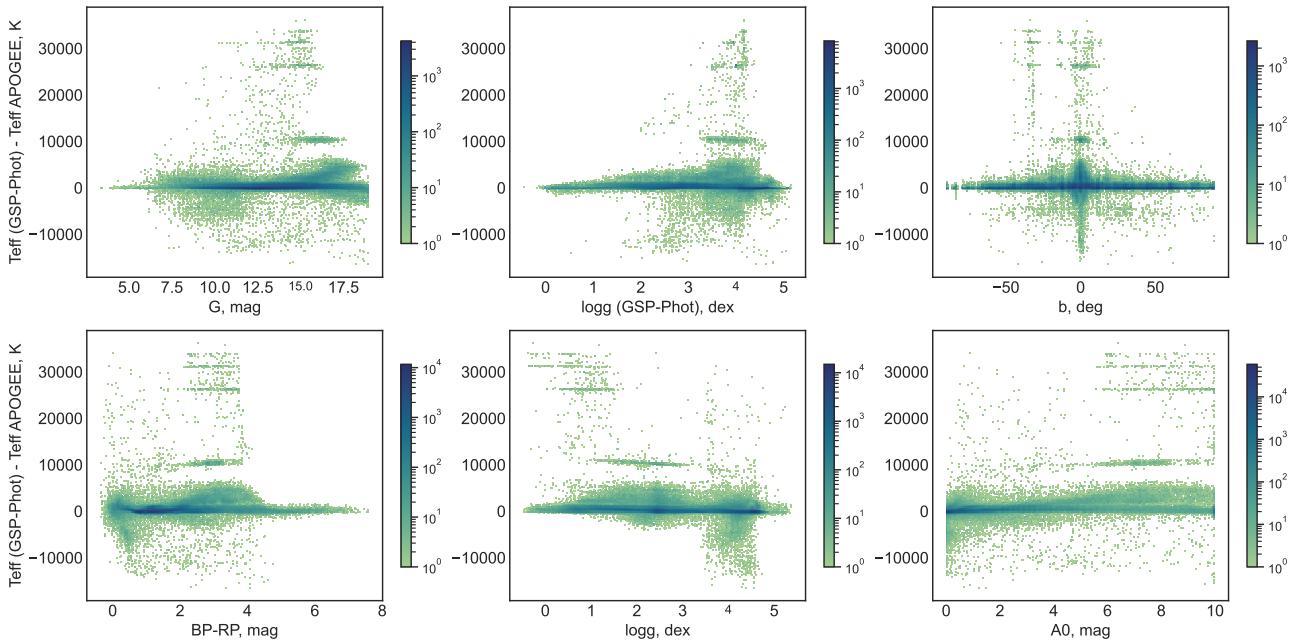

Рисунок 3.2 — Разница в эффективных температурах между Gaia DR3 и APOGEE в зависимости от различных параметров: звездной величины G, показателя цвета $BP - RP$, log$g$ от Gaia DR3 и APOGEE, галактической широты b и A_0. Точки окрашены в соответствии с плотностью звезд на диаграмме. Подробности в тексте.

в T_{eff} значения log$g$ в Gaia DR3 колеблются от 3.3 до 4.4 dex, в то время как в APOGEE для тех же звезд этот диапазон составляет от 0 до 1.5 dex.

Как отмечается в документации Gaia DR3 [78], наличие высоких различий в астрофизических параметрах в области галактической плоскости связано с тем, что влияние эффективной температуры и поглощения на наблюдаемые величины взаимозаменяемо в наблюдаемом диапазоне длин волн. Это особенно заметно при значениях $|\text{b}| < 15^\circ$ и $A_0 > 5$ звездных величин.

Также нужно отметить, что наилучшее согласие между эффективными температурами Gaia DR3 и APOGEE наблюдается для звезд с показателем цвета $BP - RP$ более 4.5. Напротив, наименее удачное согласие наблюдается для звезд с $BP - RP$ в диапазоне от 2 до 4.5. Это может быть связано с ошибкой алгоритма, применяемого в GSP-Phot при определении спектрального класса звезды, а также с худшим определением эффективной температуры для соответствующих спектральных классов.

Полученные из сравнения выводы предоставляют основу для понимания причин лучшего согласования эффективных температур Gaia DR3 с данными от GALAH. Это согласие объясняется наблюдательной стратегией обзора

GALAH, которая настроена на относительно близкие звезды вне галактической плоскости с блеском $G < 15$ звездных величин. Такой подход естественным образом исключает звезды с высоким поглощением, что также связано с выраженными сдвигами по температуре. Выбор целей обзора GALAH исключает также значительное количество плохо рассчитанных эффективных температур в Gaia DR3 и способствует видимому улучшению согласования между двумя обзорами.

Хотя ограничения, полученные в ходе этого анализа, могут помочь выбрать объекты, для которых астрофизические параметры каталога Gaia DR3 определены наилучшим образом, важно принимать во внимание, что применение этих ограничений может привести к исключению значительного числа звезд с температурами высокого качества, особенно в галактической плоскости. Строгие ограничения, применяемые для решения таких проблем, как неправильное определение эффективной температуры из-за поглощения и других систематических ошибок, могут непреднамеренно отсеивать также звезды с надежными оценками температур, пригодные для дальнейшего анализа.

3.2 Применение машинного обучения для создания флагов качества эффективных температур GSP-Phot

Чтобы решить эту проблему потенциального отсеивания звезд с температурами высокого качества из-за строгих ограничений, в данной работе предлагается использовать методы машинного обучения. Обученные на данных спектроскопических обзоров высокого разрешения, модели могут научиться идентифицировать и учитывать систематические ошибки и неопределенности, присутствующие в эффективных температурах Gaia DR3. Этот подход позволяет избежать грубого применения ограничений, которые могли бы исключить ценные для дальнейшего анализа данные. Подход, основанный на данных, способный улавливать взаимосвязи между разными параметрами в многомерном пространстве, поможет сохранить наиболее ценные данные, полученные модулем GSP-Phot.

Выбор правильного набора тренировочных данных для обучения модели имеет решающее значение для получения наилучшего результата. Для обес-

печения наиболее высокого качества референсных данных, мы используем набор данных для обучения, выбранный из пересечения обзоров APOGEE и GALAH. Решение использовать пересечение этих обзоров обусловлено желанием использовать сильные стороны обоих обзоров. С помощью использования их пересечения мы стремимся улучшить устойчивость нашего обучающего набора данных и минимизировать возможные искажения и систематические погрешности, специфичные для каждого из обзоров в отдельности. После проведения отбора по качеству с помощью фильтров, упомянутых в предыдущем разделе, в пересечении APOGEE DR17 и GALAH DR3 у нас осталось 17501 звезд.

Нужно отметить, что эффективные температуры в APOGEE DR17 калибруются с использованием фотометрических данных. Калибровка эффективных температур производится путем сравнения их с фотометрическими эффективными температурами, следуя методологии, изложенной в работе [79]. В отличие от APOGEE, в GALAH DR3 температуры по умолчанию получаются из наилучшего соответствия спектров. Это различие подчеркивает разнообразие методов, используемых в обзорах, каждый из которых имеет свои сильные стороны и ограничения.

Для установления надежных референсных значений для нашей модели машинного обучения, мы вычисляем взвешенное среднее эффективных температур, полученных из APOGEE DR17 и GALAH DR3. Полученное значение эффективной температуры сочетает надежность и точность обоих обзоров, обеспечивая более сбалансированное представление об истинных эффективных температурах. Такой подход также помогает ослабить влияние возможных искажений, которые могли бы быть из-за особенностей каждого обзора в отдельности.

Одним из ограничений этой методологии является верхний предел яркости звезд, общих для обзоров, который составляет около $G = 15.9$ звездной величины. Кроме того, большинство звезд в полученном наборе данных имеют эффективную температуру ниже 7000 K. Как утверждается в работе [32], калибровка эффективных температур в APOGEE выполнялась на основе звезд, которые также в основном холоднее 7000 K, поэтому калибровка для более горячих звезд может давать менее надежные результаты. Тем не менее, мы сохраняем все звезды в APOGEE, для которых установлен соответствующий критерий качества. Все эти ограничения влияют на то, насколько мы можем экстраполировать результаты за пределами этих диапазонов.

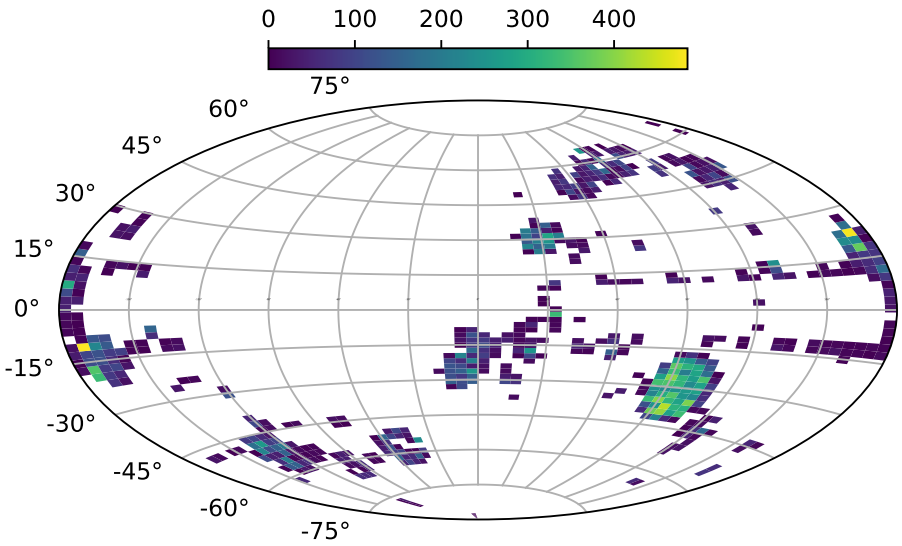

Рисунок 3.3 — Распределение набора обучающих данных по небу в проекции Аитова - модифицированной азимутальной проекции с центром Галактики в начале координат. Точки на карте кодируются цветом в зависимости от плотности звезд, отражая концентрацию звезд в различных регионах. Одним из наиболее плотно населенных регионов является Большое Магелланово Облако, наблюдаемое как в обзоре GALAH, так и в южной части обзора APOGEE, имеющей название APOGEE-2S.

Пространственное распределение объектов в наборе данных показано на Рис. 3.3. Большинство точек сосредоточены вдоль экваториальной плоскости, а также выделяется область, совпадающая с положением Большого Магелланова Облака. Отметим, что в наборе данных содержится некоторое количество точек, располагающееся в плоскости Галактики. Это особенно значимо, поскольку обеспечивает модели материалом, необходимым для эффективного отличия точных эффективных температуры от ненадежных в этой области пространства параметров.

Хотя Gaia предоставляет обширный набор различных параметров для каждого из объектов, практические соображения требуют выборочного использования этих параметров. Поэтому мы должны выбрать характеристики, которые считаются значимыми или потенциально ценными для оценки качества эффективных температур. В данной работе значимыми были выбраны следующие параметры объектов из основного каталога Gaia DR3 и дополнительных таблиц астрофизических параметров:

- *ra* и *dec*: прямое восхождение и склонение соответственно.

- l и b: галактическая долгота и широта.

- *parallax* и *parallax_error*: изменение видимого смещения положения звезды из-за движения Земли вокруг Солнца, параллакс, с соответствующей неопределенностью.

- *pm*, *pmra* и *pmdec*: параметры собственного движения, где *pm* - полное собственное движение, а *pmra* и *pmdec* - собственные движения по прямому восхождению и склонению соответственно.

- *ruwe*: Renormalized Unit Weight Error - оценка надежности астрометрических решений Gaia.

- *ipd_frac_multi_peak* и *ipd_frac_odd_win*: параметры, связанные с интегрированной функцией плотности вероятности (IPD). Эти значения содержат информацию о вероятности того, что источник является двойной звездой или данные фотометрии искажены вследствие наличия другого источника.

- *phot_g_mean_mag*, *bp_rp*, *bp_g* и *g_rp*: параметры фотометрии, включая звездную величину в полосе G и показатели цвета (*bp_rp*, *bp_g* и *g_rp*), предоставляющие информацию о цвете объекта. Мы не используем величины в полосах BP и RP, поскольку они являются линейной комбинацией величины в полосе G и соответствующего показателя цвета.

- *teff_gspphot*, *teff_gspphot_lower* и *teff_gspphot_upper*: эффективная температура, полученная с помощью модуля GSP-Phot Gaia с нижним и верхним пределами.

- *logg_gspphot*, *logg_gspphot_lower* и *logg_gspphot_upper*: ускорение свободного падения на поверхности звезды, полученное с помощью модуля GSP-Phot Gaia с нижним и верхним пределами.

- *mh_gspphot*, *mh_gspphot_lower* и *mh_gspphot_upper*: металличность, полученная с помощью модуля GSP-Phot Gaia с нижним и верхним пределами.

- *azero_gspphot*: монохроматическое поглощение (A_0) на длине волны 541.4 нм, в предположении, что звезда является одиночным источником, определенное с помощью модуля GSP-Phot Aeneas с использованием спектров BP/RP, видимой величины G и параллакса.

- C*: модифицированная версия фактора *bp_rp_excess*, введенная в работе [80]. Когда C* положительно, это означает, что совокупный поток из полос BP и RP превышает поток из полосы G, что может указывать на засвечивание от близких источников. Напротив, когда C* отрицательно, это указывает

на противоположную ситуацию, возможно, вызванную чрезмерной вычиткой фона в полосах BP или RP.

Выбранные характеристики в совокупности составляют профиль каждой звезды. Хотя наш предварительный анализ указал на зависимость различия температур только от галактической широты (отметим, что она не совсем симметрична), мы решили также оставить долготу в качестве признака в нашей модели. Это решение обусловлено известными систематическими эффектами, присутствующими в Gaia DR3, см., например, раздел 3.3 в работе [81].

Для определения качественных температур в данных Gaia GSP-Phot, мы решаем задачу двоичной классификации. Для разделения данных на положительные и отрицательные классы мы применяем следующий критерий:

$$\mid T_{\rm eff}^{Gaia} - \overline{T}_{\rm eff} \mid < \delta T_{\rm eff}^{crit} \tag{3.1}$$

$$\overline{T}_{\rm eff} = \frac{\sum (T_{\rm eff}^{i} \cdot \frac{1}{\sigma^2(T_{\rm eff}^{i})})}{\sum \frac{1}{\sigma^2(T_{\rm eff}^{i})}} \tag{3.2}$$

Здесь $T_{\rm eff}^{Gaia}$, $\overline{T}_{\rm eff}$ представляют собой эффективную температуру из Gaia DR3, определенную модулем GSP-Phot и средневзвешенное значение эффективных температур из обзоров APOGEE и GALAH для одного и того же объекта, соответственно. Следует отметить, что ошибки эффективных температур, предоставленные в двух каталогах, имеют не одинаковый масштаб. На Рис. 3.4 показано распределение ошибок $T_{\rm eff}$ для каждого обзора в области их пересечения. Можно видеть, что ошибки в APOGEE систематически ниже для всех объектов, что может привести к тому, что модели будут предпочитать температуры, сходные с температурами APOGEE.

Порог $\delta T_{\rm eff}^{crit}$ представляет собой желаемый уровень точности результата, который мы хотим достичь, и является предметом обсуждения. В работе [82] утверждается, что разница между эффективными температурами APOGEE и GALAH может быть охарактеризована с помощью стандартного отклонения в 126.6 K. В нашей работе для эффективных температур ниже 7000 K стандартное отклонение между эффективными температурами APOGEE и GALAH было оценено на уровне 130 K. Точное значение в работе [82] и в нашей оценке отличается, вероятно, вследствие различных критериев качества, налагаемых на данные. В то же время стандартное отклонение для всей обучающей выборки (включая также звезды горяее 7000 K) составляет около 270 K. Хотя у нас нет

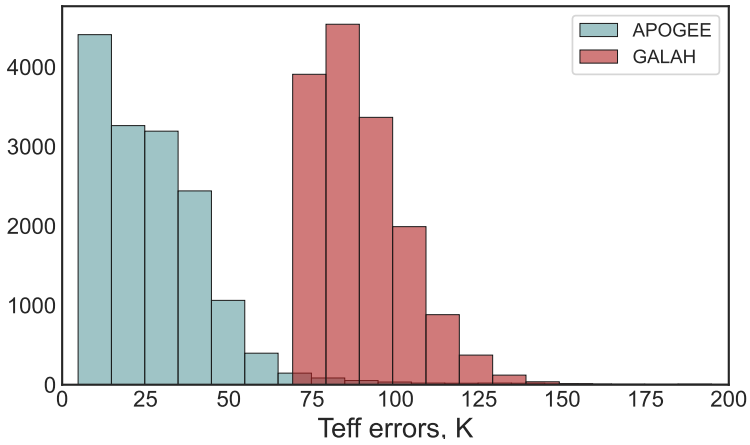

Рисунок 3.4 — Распределение ошибок температур, предоставленных обзорами APOGEE и GALAH.

Таблица 10 — Количество объектов положительного (надежные температуры) и отрицательного классов для каждого порога.

Порог	N_{pos}	N_{neg}	Процент прохождения критерия
125 K	9497	8004	54.3%
250 K	13707	3794	78.3%

надежных эталонных температур для звезд с температурой выше $\sim$ 7000 K, мы сохраняем подход, основанный на усреднении значений между двумя обзорами даже в случае значительных различий в температурах между двумя обзорами.

Мы используем точность в 125 K в качестве желаемого идеального порога точности, а также исследуем порог в 250 K (в дальнейшем обозначаемые как Порог-125 и Порог-250) в качестве более реалистичного сценария. В каждом из этих случаев критерии позволяют выделить объекты, принадлежащие к положительному классу. Порог в 125 K, вероятно, является теоретическим пределом точности для этой обучающей выборки и сопоставим со случайными ошибками при определении эффективной температуры.

Таб. 10 показывает количество объектов положительного и отрицательного классов для каждого из двух случаев. В то время как разница между количеством объектов в разных классах не существенна в случае с порогом 125 K, наблюдается значительный дисбаланс в данных с порогом 250 K. Этот дисбаланс в основном обусловлен стратегией выбора целей, используемой в GALAH, как обсуждалось ранее, поскольку объекты в нашем наборе данных также подчиняются этому процессу отбора.

Мы используем метод SMOTE [83], чтобы сбалансировать классы в каждом из случаев. В данной работе балансировка классов производится с помощью техники увеличения количества образцов в менее представленном классе. Метод SMOTE генерирует синтетические образцы для менее представленного класса путем интерполяции между существующими образцами. Введение этих синтетических образцов направлено на создание более сбалансированного распределения классов и смягчения негативных эффектов дисбаланса классов на обучение модели.

Набор данных делится на три части с учетом стратификации с соотношением классов в оригинальной выборке: обучающая выборка (60 процентов), валидационная выборка (20 процентов) и тестовая выборка (20 процентов). Разделение данных производится с использованием метода TRAIN_TEST_SPLIT из библиотеки SCIKIT-LEARN. Метод SMOTE применяется к обучающей части набора данных, при этом валидационная и тестовая части остаются неизменными, чтобы оценка модели производилась на неизмененных данных. Гиперпараметры модели выбираются на основе валидационной выборки с использованием модуля OPTUNA [84]. Окончательная оценка качества модели производится на тестовой выборке.

В данном исследовании изучаются три классических модели машинного обучения: XGBoost [85], CatBoost [86] и LightGBM [87]. Обзор эффективности различных моделей, проведенный [88], подчеркивает наилучшую производительность этих трех алгоритмов бустинга на табличных данных.

Алгоритмы бустинга уже упоминались в Главе 2 настоящей диссертации и представляют собой мощное семейство методов машинного обучения, разработанных для улучшения предсказательной точности моделей путем комбинирования предсказаний нескольких более слабых моделей, часто называемых базовыми обучающими моделями или слабыми обучающими моделями. Основная идея бустинга заключается в последовательном обучении этих базовых моделей, каждая из которых фокусируется на ошибках, допущенных предыдущеми в цепочке моделями. Приписывая больший вес объектам, классифицированным неправильно, алгоритмы бустинга итеративно улучшают производительность модели, в конечном итоге создавая сильную и точную ансамблевую модель.

Мы используем стандартные метрики классификации, а именно точность (precision) и полноту (recall). Точность характеризует долю релевантных объ-

ектов среди извлеченных, в то время как полнота измеряет эффективность захвата релевантных объектов среди всех существующих. Формулы для расчета precision и recall приведены в Главе 2. Также вычисляется f1-мера, которая является гармоническим средним между величинами точности и полноты.

Следует отметить, что из-за стратегии выбора целевого класса, применяемой как GALAH, так и APOGEE, доля положительного класса в тестовой части набора данных может значительно отличаться от доли этого класса в полном наборе данных Gaia DR3. Этот эффект менее выражен в случае Порога в 125 K, но в случае Порога в 250 K количество объектов положительного класса значительно превышает количество объектов отрицательного класса. В то время как значение полноты (recall) должно оставаться неизменным независимо от различий в дисбалансе классов, поскольку она учитывает только объекты положительного класса, точность (precision) более чувствительна к таким изменениям. Если доля положительных объектов действительно больше, чем в тестовой выборке, то оцениваемая в данном исследовании точность может быть занижена. Напротив, в случае недооцененного отрицательного класса в тестовой выборке точность будет завышена.

Более того, когда требуется вычислить условную вероятность того, что температура действительно является надежной при условии, что модель классифицировала ее как таковую, априорные вероятности становятся ключевым фактором. Представим ситуацию, в которой все температуры, предоставленные Gaia GSP-Phot, являются надежными. В этом случае, насколько бы не были низки точность и полнота моделей, любой выбор, который они сделают, даст нам только точные температуры. Напротив, когда хорошие температуры редки, даже модели с высокими значениями точности и полноты не могут гарантировать высокую вероятность того, что выбранная температура действительно точна. Это обычно происходит при классификации крайне редких классов объектов, таких, например, как квазары [89]. На данный момент у нас нет надежных априорных вероятностей для температур Gaia GSP-Phot. Доли хороших оценок температур в Таб. 10 подвержены влиянию функции отбора целей наблюдения APOGEE и GALAH и могут не быть репрезентативными по отношению к более широкому набору данных Gaia GSP-Phot. Тем не менее, эти оценки указывают на то, что ни один из двух классов, положительный или отрицательный, не является чрезвычайно редким.

В каждой из моделей, как это уже было отмечено в Главе 2, гиперпараметры играют ключевую роль в получении эффективного решения. В Таб. ?? показаны гиперпараметры, которые были подобраны для каждой модели при двух случаях порога, обозначенных как Порог-125 и Порог-250. Мы оставили фиксированное количество базовых слабых моделей (estimators), а именно 500, для моделей XGBoost и LightGBM, в то время как другие параметры (не указанные в таблице) использовались при значениях по умолчанию. Для подбора гиперпараметров также как и ранее был использован фреймворк OPTUNA. Для каждой модели в каждом случае порога мы создали исследование OPTUNA с целью максимизации значения f1-меры на валидационной части набора данных.

В Таб. 12 представлены оценки, полученные на тестовом наборе данных для каждой модели в двух случаях: Порог-125 и Порог-250. Очевидно, что все метрики производительности, включая точность, полноту и f1-меру, демонстрируют значительно более высокие значения в случае Порога в 250 K. Это расхождение может быть обусловлено ограничениями метода определения эффективных температур в Gaia DR3, даже в его наилучших реализациях. То есть случайные ошибки метода могут лежать в пределах от 125 K до 250 K, даже без учета систематических ошибок. Кроме того, это может быть вызвано изначальным дисбалансом в данных в случае Порога-250, который сохраняется в тестовом подмножестве и может способствовать упрощению классификации. Этот дисбаланс в наборе данных может потенциально привести к более явному различию между положительными и отрицательными классами, что формально улучшает метрики производительности модели без улучшения реальной эффективности.

В случае Порога-125 значение точности ниже значения полноты для одной и той же модели, что указывает на то, что модели затрудняются в различении эффективных температур разного качества, определенных соответствующим критерием в 125 K. Особенно отмечается недостаток полноты решения для модели CatBoost в этом случае, в то время как различие в точности меньше по сравнению с моделями XGBoost и LightGBM.

В случае Порога-250 точность и полнота достаточно сбалансированы, что является предпочтительным в большинстве сценариев. Кроме того, почти одинаковые оценки для всех моделей указывают на то, что все модели легко справляются с классификацией при пороге 250 K, что делает ее более достижимой для реальных данных во всем наборе Gaia DR3.

Таблица 11 — Настроенные гиперпараметры для каждой модели с различными порогами. Количество базовых слабых моделей зафиксировано на уровне 500 для моделей XGBoost и LightGBM, а количество итераций зафиксировано на уровне 500 для модели CatBoost. Другие параметры оставлены по умолчанию.

Модель	Гиперпараметры (Настроенные значения)
XGBoost	Порог-125: max_depth (13), learning_rate (0.014), subsample (0.753), gamma (0.593), reg_alpha (0.457) Порог-250: max_depth (10), learning_rate (0.061), subsample (0.997), gamma (0.750), reg_alpha (0.048)
CatBoost	Порог-125: learning_rate (0.652), depth (5), bootstrap_type ("Bernoulli"), objective ("Logloss"), subsample (0.960), boosting_type("Ordered"), colsample_bylevel (0.830) Порог-250: learning_rate (0.225), depth (11), bootstrap_type ("Bernoulli"), objective ("CrossEntropy"), subsample (0.282), boosting_type("Ordered"), colsample_bylevel (0.293)
LightGBM	Порог-125: learning_rate (0.002), max_depth (15), num_leaves (891), feature_fraction (0.637), bagging_fraction (0.808), bagging_frec (7) Порог-250: learning_rate (0.237), max_depth (12), num_leaves (316), feature_fraction (0.931), bagging_fraction (0.959), bagging_frec (3)

Таблица 12 — Производительность моделей на тестовой части наборов данных с различными порогами. Лучшие оценки выделены жирным шрифтом.

	Порог-125			Порог-250		
	Точность	Полнота	F1-мера	Точность	Полнота	F1-мера
XGBoost	**0.796**	0.844	**0.819**	**0.939**	0.922	0.930
CatBoost	0.781	0.797	0.789	0.934	0.930	**0.932**
LightGBM	0.785	**0.845**	0.814	0.926	**0.939**	**0.932**

Для получения более глубокого понимания того, как модель будет работать за пределами пространства параметров обучающего набора данных, мы оцениваем ее производительность на трех дополнительных наборах данных. Это полные наборы данных APOGEE и GALAH (за исключением звезд, использованных при обучении и оценке модели), а также база данных PASTEL [90] – библиографическая компиляция атмосферных параметров звезд, основанная на спектроскопии высокого разрешения, исследованной и полученной разными авторами.

Как было отмечено выше, из наборов данных APOGEE и GALAH мы исключаем объекты, которые уже были использованы при обучении, верификации и проверке моделей. Мы применяем те же самые критерии качества, что и для обучающего набора, и исключаем звезды с отсутствующими эффективными температурами из Gaia DR3. После этой процедуры остается 451772 объекта из пересечения APOGEE и Gaia и 273564 объекта из соответствующего пересечения с GALAH.

База данных PASTEL во многих случаях содержит несколько записей для одного и того же объекта. Чтобы избежать неоднозначности, мы сначала вычисляем взвешенное среднее эффективных температур, если для объекта имеется более одной записи. Затем мы переходим к перекрестному сопоставлению полученных записей с каталогом Gaia DR3. Мы использовали сервис CDS X-Match с максимальным радиусом перекрестного сопоставления равным $2''$, чтобы найти соответствующие объекты из PASTEL в Gaia DR3. На Рис. 3.5 показано угловое расстояние между записями из PASTEL и соответствующими объектами, найденными в Gaia DR3. Сопоставления для объектов с эффективными температурами из Gaia DR3 найдены в пределах $2''$ для 28588 объектов из PASTEL.

Объекты в этих трех наборах данных маркируются аналогично обучающему набору данных. Мы используем два порога, 125 K и 250 K, чтобы классифицировать объекты как положительные или отрицательные. Затем мы применяем соответствующие модели и извлекаем объекты с высококачественными эффективными температурами в Gaia DR3 в соответствии с каждой моделью. Мы оцениваем результаты с помощью тех же метрик, что были использованы ранее, но также изучаем разницу между эффективными температурами Gaia DR3 и теми, которые получены из эталонного набора данных для объектов, извлеченных моделью. Следует отметить, что, хотя все эти эталонные наборы данных являются источниками атмосферных параметров высокого

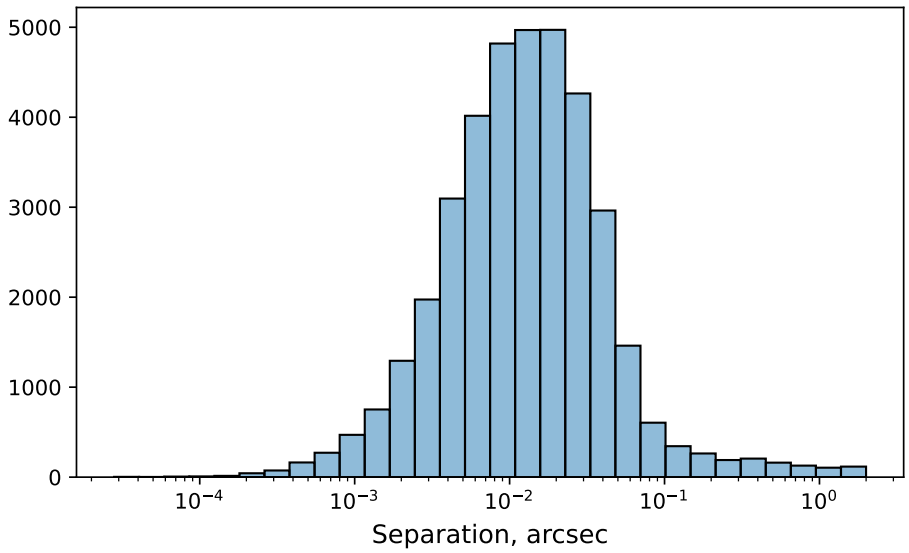

Рисунок 3.5 — Распределение значений углового расстояния между объектами из базы данных PASTEL и их ближайшими соседями из Gaia DR3. Большинство соответствующих объектов находятся на расстоянии менее $0.1''$.

качества, значения в этих наборах данных также могут систематически различаться друг от друга.

Результаты применения моделей машинного обучения к наборам данных представлены в Таб. 13 для случая с порогом 125 K и в Таб. 14 для случая с порогом 250 K. Эти таблицы представляют обзор эффективности моделей в различных сценариях, включая параметры точности, полноты и f1-меры. Кроме того, мы оцениваем модели, вычисляя медиану разницы и 90-й процентиль абсолютной разницы между эффективными температурами Gaia DR3 и эффективными температурами из соответствующего референсного набора данных для объектов, выбранных каждой моделью, как хорошие. Эти статистические характеристики предоставляют ценные сведения о способности моделей точно идентифицировать высококачественные эффективные температуры в Gaia DR3. Для сравнения в таблицу также добавлены медиана и 90-й процентиль для полного набора данных.

В случае Порога-125 K наши модели демонстрируют относительно низкие как точность, так и полноту и значение f1-меры. При этом точность получается выше в случае, когда модели применяются к данным APOGEE. Медиана разницы температур в этих данных относительно близка к нулю, а 90-й процентиль абсолютной разницы заметно ниже относительно полного набора данных, хотя желаемый порог в 125 K был достигнут только для примерно 77 процентов выбранных объектов. Эта тенденция, скорее всего, обусловлена различиями в

Таблица 13 — Производительность моделей машинного обучения в случае порога 125 K. Для каждой модели приведены метрики precision, recall и f1-меру. Кроме того, представлены медиана разницы и 90-й процентиль абсолютных различий между эффективными температурами Gaia DR3 и теми из эталонного набора данных для объектов, классифицированных как высококачественные температуры каждой моделью. Для сравнения приведены медиана и 90-й процентиль значений для всего набора данных под заголовком "Без модели".

Модель	Точность	Полнота	F1-мера	Медиана	90-й перцентиль
APOGEE					
Без модели	-	-	-	110.4	849.1
XGBoost	0.778	0.768	0.773	-2.3	139.8
CatBoost	0.766	0.643	0.699	1.6	151.6
LightGBM	0.760	0.776	0.768	1.0	156.0
GALAH					
Без модели	-	-	-	22.9	391.0
XGBoost	0.710	0.674	0.692	-27.0	204.3
CatBoost	0.691	0.674	0.683	-16.5	212.7
LightGBM	0.710	0.684	0.697	-27.2	204.7
PASTEL					
Без модели	-	-	-	-64.3	524.0
XGBoost	0.609	0.584	0.596	-91.8	245.8
CatBoost	0.600	0.497	0.544	-92.8	248.9
LightGBM	0.608	0.580	0.593	-89.3	242.7

Таблица 14 — Та же таблица, что и Таб. 13, но для случая порога 250 K.

Модель	Точность	Полнота	F1-мера	Медиана	90-й перцентиль
APOGEE					
Без модели	-	-	-	110.4	849.1
XGBoost	0.891	0.850	0.870	15.9	260.6
CatBoost	0.872	0.887	0.879	22.4	285.2
LightGBM	0.867	0.875	0.871	22.7	294.9
GALAH					
Без модели	-	-	-	22.9	391.0
XGBoost	0.907	0.821	0.862	-10.5	244.4
CatBoost	0.900	0.850	0.874	-6.3	250.1
LightGBM	0.900	0.827	0.862	-8.5	249.8
PASTEL					
Без модели	-	-	-	-64.3	524.0
XGBoost	0.875	0.865	0.870	-79.7	274.0
CatBoost	0.872	0.872	0.872	-77.6	278.2
LightGBM	0.873	0.829	0.851	-79.8	275.9

ошибках эффективных температур между GALAH и APOGEE, как показано на Рис. 3.3. Учитывая, что ошибки, предоставленные APOGEE, всегда меньше, средние значения, используемые во время обучения, склонны больше соответствовать эффективным температурам в APOGEE.

Когда мы применяем модели к набору данных PASTEL, медианная абсолютная разница сдвигается в сторону больших отклонений. Это можно объяснить неоднородным характером взвешенных эффективных температур PASTEL (см. обсуждение расхождений между значениями из различных источников в [91]). Тем не менее, стоит отметить, что 90-й перцентиль абсолютной разницы в эффективных температурах уменьшается более чем в два раза по сравнению с полным набором данных без какой-либо фильтрации. Это указывает на то, что, несмотря на смещение к более высоким средним значениям отклонений в температурах, модели по-прежнему способствуют улучшению общего качества выборки температур.

Как уже обсуждалось ранее, отбор целей наблюдений GALAH DR3 способствует лучшей сходимости температур между обзорами gaia и GALAH. Это

подтверждается параметрами полного набора данных. Тем не менее, даже в данном сценарии все модели способствуют улучшению статистического качества температур Gaia DR3. С другой стороны, средние значения полноты решения по всем моделям указывают на то, что модели могут отвергать большое количество хороших эффективных температур в погоне за достижением желаемой точности. Хотя в некоторых исследованиях это не играет решающей роли, полнота извлечения хороших эффективных температур является важным параметром.

При использовании Порога-250 K модели показывают заметное улучшение на всех наборах данных. Достижение желаемого порога в 250 K в этом случае оказывается значительно более реальным, хотя от 9 до 14 процентов звезд, классифицированных моделями как положительные, превышают этот порог по абсолютной разнице в эффективных температурах. Оценка полноты, которая отражает способность моделей извлекать качественные данные, значительно улучшается по сравнению с использованием порога в 125 K. Однако модели способны восстанавливать только до 89 процентов хорошо оцененных звезд в соответствующих наборах данных. Это можно частично объяснить эффективными температурами, на которых были обучены модели. Как уже отмечалось, большинство звезд в обучающем наборе данных имеют эффективную температуру менее 7000 K, что делает модели в основном нацеленными на изучение особенностей в этом диапазоне. Следовательно, они могут испытывать трудности при работе с более горячими звездами.

Следует отметить, что все модели при пороге в 250 K не привязаны исключительно к эффективным температурам APOGEE, как это получается при пороге в 125 K. Вместо этого они показывают сбалансированный выбор значений между температурами GALAH и APOGEE. Это более сбалансированное решение снижает выраженность смещений в медианных значениях при применении к набору данных PASTEL по сравнению с использованием порога в 125 K.

Хотя 90-й процентиль абсолютной разницы в эффективных температурах выше, чем при пороге в 125 K, модели с порогом в 250 K более универсальны и лучше подходят для обобщения результатов. Это указывает на более широкие возможности их использования и способность обеспечивать более надежные результаты при работе с широким спектром параметров звезд.

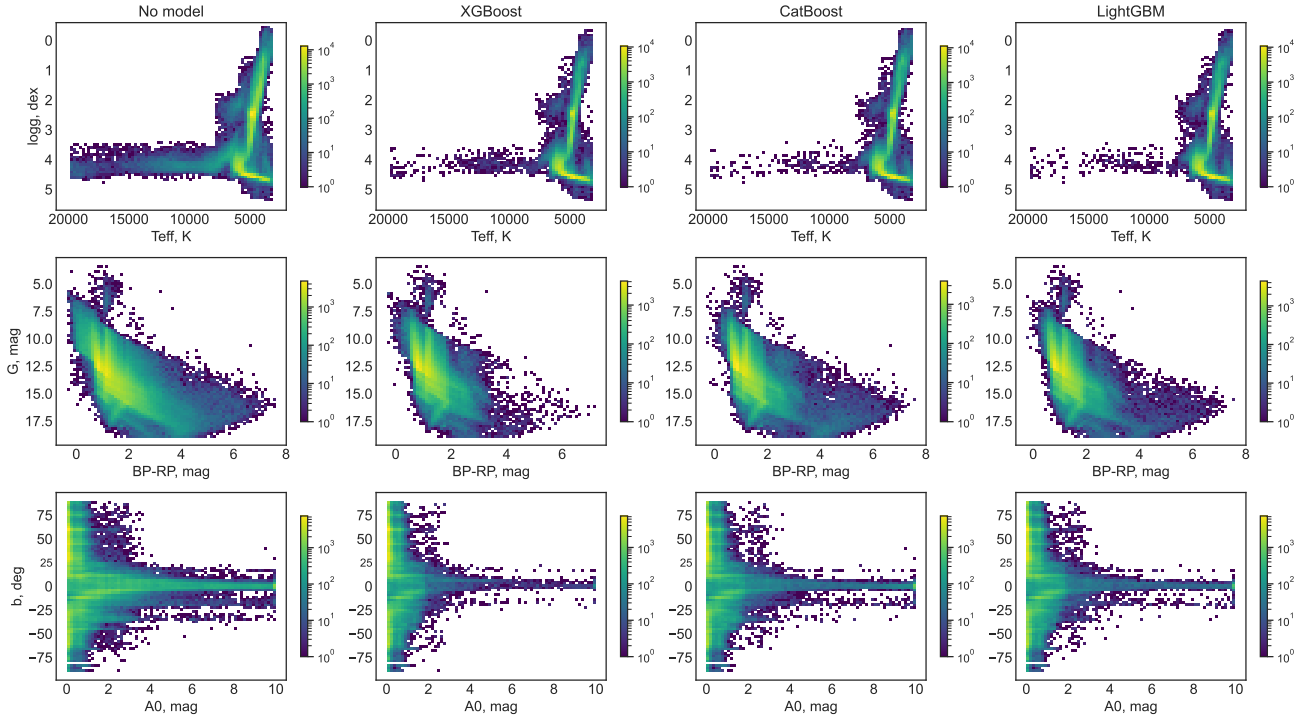

Рисунок 3.6 — Распределение звезд из набора данных APOGEE DR17 на трех различных диаграммах: log$g$-T_{eff}, G-$(BP-RP)$ и b-A_0. Самая левая колонка показывает полный набор данных APOGEE, в то время как другие колонки отображают только данные, признанные моделями как качественные. Точки окрашены в зависимости от плотности звезд.

3.3 Изучение звезд, отобранных моделью из APOGEE в пространстве параметров

В данном разделе изучается распределение выбранных моделями звезд по различным параметрам, таким как галактическая широта, блеск, показатели цвета и другие. Мы представляем (см. Рис. 3.6) распределение звезд из APOGEE на трех диаграммах: log$g$-T_{eff}, G-$(BP-RP)$ и b-A_0. Для данного анализа были выбраны данные APOGEE, так как они охватывают более широкий диапазон параметров и включают большее количество звезд, что обеспечивает наиболее полное представление о распределении. Первая колонка показывает полный набор данных APOGEE, в то время как последующие колонки представляют только данные, которые соответствующая модель классифицирует как качественные. Все примененные модели представлены в версии Порог-250.

Данные представленные на Рис. 3.6 показывают, что применяемые методы хоть и снижают количество звезд, но не являются селективными по

каким-либо параметрам вдоль осей, за исключением области горячих звезд. Модели способны выделять надежные значения температур даже в сложных областях параметров, таких как низкая видимая звездная величина и область галактической плоскости. Это указывает на потенциал различения надежных эффективных температур в сложных областях пространства параметров.

3.4 Применение модели ко всем эффективным температурам GSP-Phot

Мы создали флаги качества для всех звезд с эффективными температурами Gaia, используя модель XGBoost с порогом 250 K. XGBoost был выбран в качестве модели с наименьшим 90-м процентилем абсолютной разницы между эталонными эффективными температурами и теми, которые выбираются моделью. Модель определяет 313 миллионов звезд модуля GSP-Phot как качественные. В данном подразделе мы проводим статистический анализ и исследуем разницу между эффективными температурами с флагами 0 (плохие температуры) и 1 (хорошие температуры) в пространстве параметров.

Рис. 3.7 показывает относительную разницу в плотности звезд между хорошими и плохими эффективными температурами, предсказанными моделью. График нормирован на распределение всех объектов Gaia DR3 с известными температурами GSP-Phot. Светло-желтые части графика указывают на области, в которых преобладают плохие эффективные температуры. Видно, что эти области в основном соответствуют областям высокого поглощения в Галактике. На график также нанесены контуры обучающего набора данных белым цветом. Важно отметить, что не наблюдается заметной корреляции между положением объектов в обучающем наборе данных и распределением хороших или плохих эффективных температур в результатах. Это наблюдение указывает на то, что модель хорошо обобщает пространственное распределение по всему небу. Однако некоторые артефакты тоже заметны. Заметно резкое изменение доли хороших температур между молекулярными облаками Тельца и Цефея слева на графике. Подобная граница также заметна к северу от молекулярного облака Волка, примерно между 240 и 330 градусами галактической долготы.

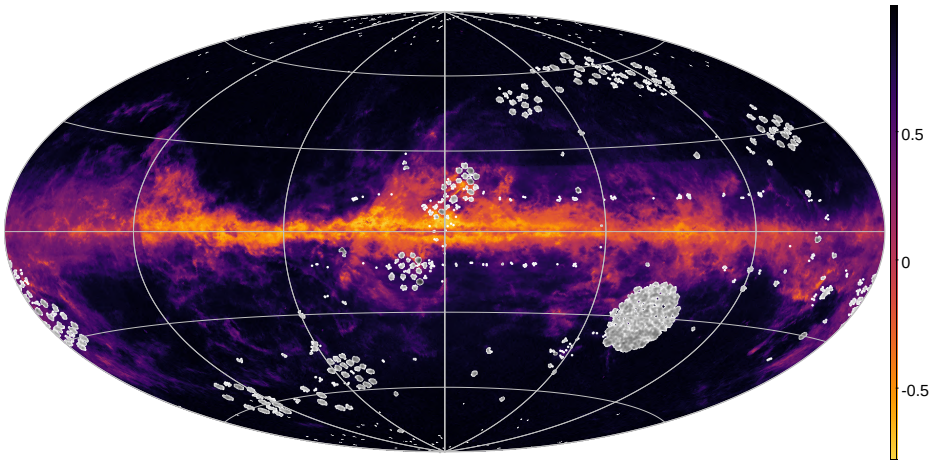

Рисунок 3.7 — Относительная разница между плотностью звезд с хорошими и плохими эффективными температурами, поделенная на общую плотность объектов. Белые области представляют собой области обучающего набора данных, в светло-желтых областях доминируют плохие температуры, а темно-фиолетовые области указывают на наличие большого количества хороших температур. Обсуждение представлено в тексте.

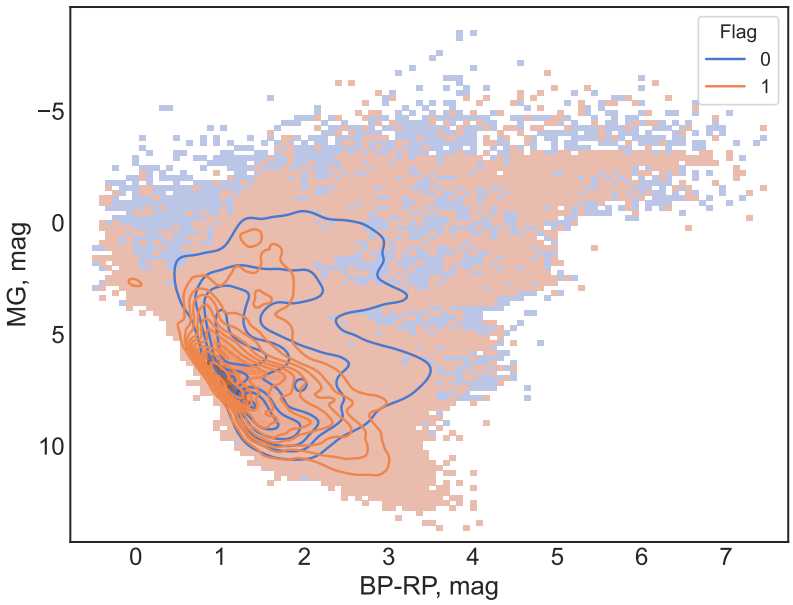

Рисунок 3.8 — Распределение двух различных подвыборок на диаграмме Герцшпрунга-Рассела: одна с плохими эффективными температурами (флаг 0), другая с хорошими эффективными температурами (флаг 1).

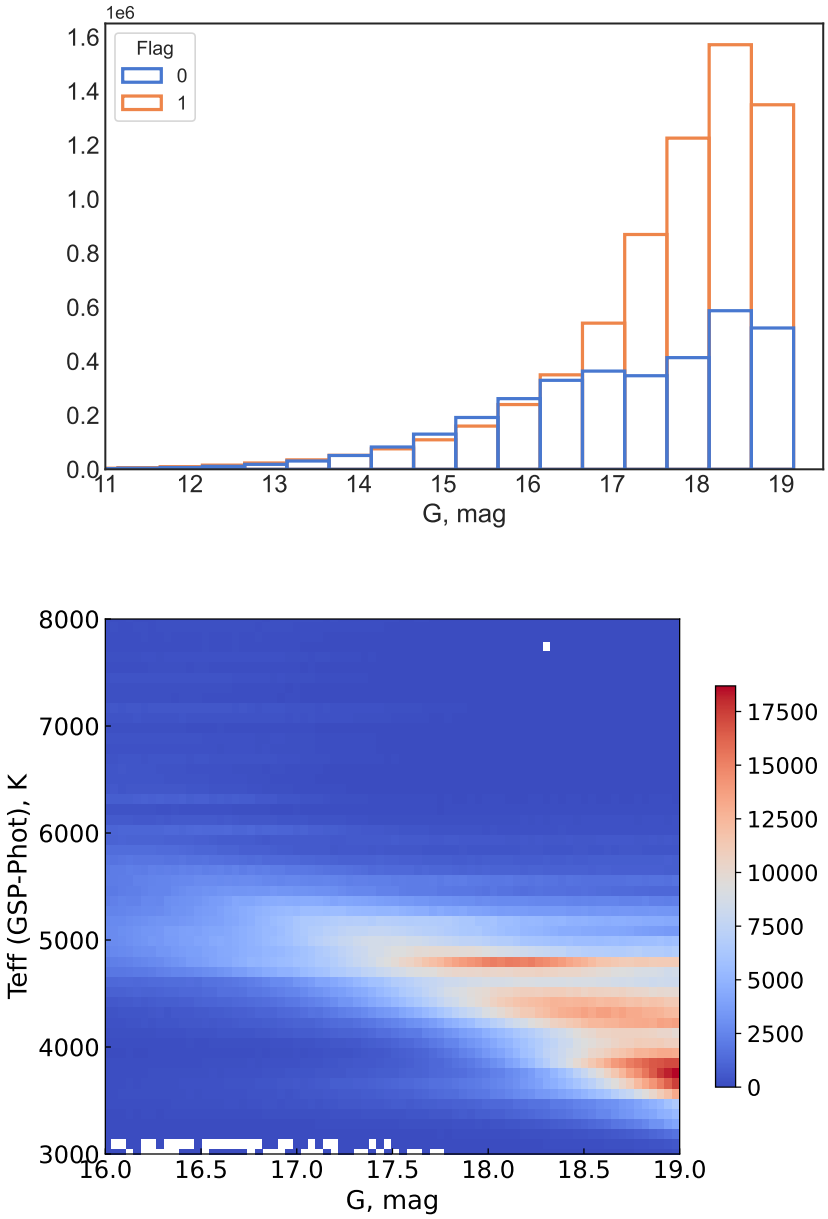

Рисунок 3.9 — Верхний график иллюстрирует распределение двух различных подвыборок, обозначенных флагом 0 для плохих эффективных температур и флагом 1 для хороших эффективных температур, на основе видимой звездной величины. Нижний график отображает распределение объектов Gaia DR3 в зависимости как от видимой звездной величины, так и от эффективных температур. Оба графика созданы на основе выборки из 10 миллионов объектов Gaia, и плотность звезд закодирована цветом.

Мы сравниваем распределение объектов с хорошими и плохими эффективными температурами на диаграмме Герцшпрунга-Рассела на Рис. 3.8. Эта диаграмма была построена на основе выборки из одного миллиона объектов, случайным образом выбранных из полного набора данных GSP-Phot. Заметно, что распределение объектов разных классов качества на диаграмме Герцшпрунга-Рассела показывает значительные различия. В частности, объекты с хорошими температурами формируют отдельную область хорошо определенную вдоль главной последовательности. Напротив, объекты с плохими температурами занимают все пространство на диаграмме, что может указывать на то, что полученные абсолютные звездные величины для этой группы могут быть неточными. Кроме того, стоит отметить, что выборка звезд с хорошим качеством не содержит горячих звезд, а гиганты недостаточно представлены, возможно, из-за ограниченного числа гигантов в обучающем наборе данных.

На Рис. 3.9 изображено распределение звезд по видимой звездной величины в полосе G. На верхнем графике распределение разделено на классификацию по хорошим и плохим температурам, в то время как нижний график отображает общее распределение по величине, без учета классификации по группам. Эти распределения основаны на выборке из 10 миллионов объектов из Gaia DR3. Заметно увеличение доли объектов с хорошими температурами при увеличении видимой звездной величины (то есть для более слабых объектов). Этот рост числа звезд, сопровождающийся более высокой долей хороших звезд, можно объяснить появлением значительной популяции звезд с эффективными температурами GSP-Phot в диапазоне от 3500 K до 5000 K для источников с блеском $G \geqslant 17.5$ звездных величин, как видно из нижнего графика.

Повышенная доля хороших звезд среди более слабых объектов является дискуссионным результатом. С одной стороны, важно отметить, что обучающая выборка, использованная для модели, исключает объекты с блеском слабее $G = 15.7$ звездной величины. Следовательно, модель имеет ограниченный опыт работы со звездами в этом диапазоне звездных величин. Кроме того, качество других параметров, таких как астрометрические решения и фотометрия, обычно ухудшается с увеличением звездной величины. Таким образом, это снижение качества данных может привести к менее точным астрофизическим решениям, предоставляемым GSP-Phot для более слабых звезд. Кроме того, может быть актуальна проблема вырождения температуры и поглощения, когда поглоще-

ние для более слабых объектов может быть недооценено, что делает звезды визуально более "холодными"в оценках GSP-Phot.

С другой стороны, сравнение с данными APOGEE не показывает специфических различий в температурах для объектов со звездной величиной $G = 17.5$ и слабее (см. Рис. 3.2a). Для пересечения данных APOGEE и Gaia мы дополнительно сравнили распределение абсолютной разницы температур для звезд с блеском ярче $G = 17.5$ и для менее ярких звезд. Для каждого диапазона звездных величин мы сравнили два диапазона температур GSP-Phot, а именно звезды с температурой ниже 5000 K и выше этой температуры (см. Рис. 3.10). В то время как распределения для $G < 17.5$ практически одинаковы для более холодных и более горячих звезд, распределения для $G \geqslant 17.5$ значительно различаются между двумя диапазонами температур. Можно предположить, что для звезд с $G \geqslant 17.5$ более распространены холодные звезды, и они также обладают более высокой долей температур, близких по оценкам к APOGEE. Таким образом, мы не можем ни подтвердить, ни опровергнуть точность результатов модели для более слабых звезд на имеющихся на данный момент данных.

3.5 Обсуждение результатов главы

В данной главе мы сравнили оценки эффективной температуры, полученные из Gaia Data Release 3 (DR3), с данными, полученными из спектроскопических обзоров высокого разрешения, в частности, APOGEE и GALAH. Кроме того, мы исследовали применение методов машинного обучения для идентификации эффективных температур хорошего качества в данных Gaia DR3. Все обученные модели доступны для загрузки онлайн[1].

Сравнение показало, что, хотя Gaia DR3 предоставляет точные оценки эффективной температуры для значительного числа звезд, имеется систематическое расхождение между оценками Gaia DR3 и данными от GALAH/APOGEE. В частности, звезды, близкие к галактической плоскости с высокими значениями A_0 в Gaia DR3, могут показывать смещения в эффективной температуре до 30000 K. Крайне большие смещения также связаны с определенным диапазоном показателей цвета $BP - RP$, а именно от 2 до 4 звездных величин. Тем не

[1]https://github.com/iamaleksandra/Gaia-Teff/

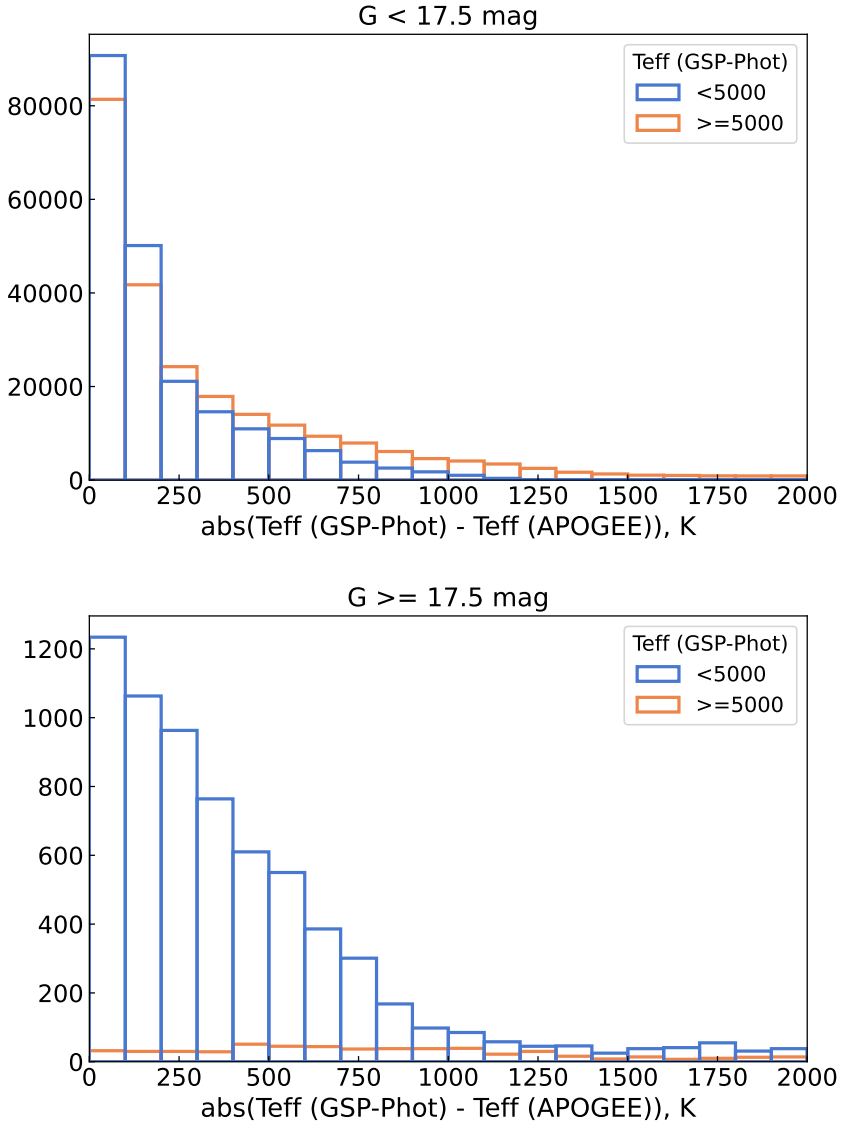

Рисунок 3.10 — Распределение разницы между эффективными температурами APOGEE и GSP-Phot для объектов с блеском ярче 17 звездной величины и слабее этого значения. На каждом графике сравнивается распределение для объектов с эффективными температурами, оцененными GSP-Phot, холоднее 5000 K и горячее этой температуры. Ось x обрезана на значениях различия температур до 2000 K.

менее, звезды с хорошими и плохими оценками температуры в Gaia DR3 иногда занимают одни и те же области пространства параметров.

Для установления надежного эталона для эффективных температур мы использовали пересечение наборов данных из каталогов APOGEE и GALAH, взяв средневзвешенное значение их эффективных температур в качестве эталонного значения. Мы рассмотрели два критерия для температур хорошего качества: $\delta T_{\rm eff}^{crit}$ равное 125 K или 250 K, как определено в уравнении 3.1, и применили три алгоритма машинного обучения: XGBoost, CatBoost и LightGBM для классификации хороших и плохих температур.

Результаты показали, что сценарий с порогом в 250 K дает более высокие оценки производительности по всем моделям, возможно, из-за ограничений в определении температуры в Gaia DR3 связанных со случайными ошибками. В случае с порогом в 125 K точность оказалась ниже полноты, что указывает на сложности в различении качества температур, тогда как в случае с порогом в 250 K достигнут желаемый баланс между точностью и полнотой.

Для оценки производительности моделей в более широком диапазоне данных мы применили модели к трем различным наборам данных: APOGEE, GALAH и PASTEL. В случае с APOGEE и GALAH мы рассматривали исключительно звезды, не входящие в обучающие наборы данных моделей. Мы обнаружили, что точность оценок температуры в случае с порогом в 125 K является относительно низкой при применении моделей машинного обучения к данным Gaia. Это указывает на сложности в точной классификации температур с порогом в 125 K. Оценка полноты также низкая для всех моделей в этом случае, что означает, что множество температур хорошего качества в процессе отбрасывается. В случае с порогом в 250 K мы наблюдаем значительное улучшение оценок моделей по всем наборам данных. Особенно стоит отметить, что модели в случае с порогом в 250 K проявляют более сбалансированное предпочтение между значениями температур GALAH и APOGEE.

Распределение звезд в пространстве параметров показало, что модели машинного обучения, которые мы использовали, не ограничены жесткими рамками по параметрам. Вместо этого они используют гибкий подход, позволяющий сохранять значительную долю объектов с температурами хорошего качества. Эта адаптивная способность подчеркивает потенциал использования этого подхода для ответа на вопрос о том, достаточно ли надежна оценка эффективной температуры конкретного объекта, даже в сложных регионах.

Ограничение данного подхода заключается в том, что модели в основном извлекают объекты с эффективными температурами менее 7000 K, что связано с отсутствием существенного количества более горячих звезд в обучающих данных. В дальнейшей работе необходимо разработать подход, позволяющий работать в более широком диапазоне температур с более высокой точностью.

Также мы создали флаги качества для всех звезд Gaia DR3 с эффективными температурами GSP-Phot, используя модель XGBoost Порог-250. Набор данных с флагами доступен по ссылке https://doi.org/10.5281/zenodo.8325377. Согласно проведенному исследованию, 313 миллионов звезд (66% всех звезд Gaia DR3 с атмосферными параметрами из модуля GSP-Phot) имеют температуры хорошего качества.

При ближайшем рассмотрении результатов полного набора данных GSP-Phot стало очевидно, что модель выделила значительно большее количество звезд, как имеющих хорошее качество, среди тех, у которых $G \geqslant 17.5$. Это наблюдение высвечивает несколько потенциальных проблем, таких как то, что модель в основном обучалась на более ярком наборе данных, и наличие вырождения между температурой и поглощением. Однако сравнение с данными APOGEE показывает, что для более слабых звезд сходимость температур получается лучше в случае, если звезды является холодными. Это наблюдение подтверждает идею о том, что результаты модели для звезд с $G \geqslant 17.5$ могут быть дастаточно точными. Тем не менее, важно отметить существующие ограничения такого вывода. В настоящее время звезды с таким слабым блеском лишены надежных эталонов температуры, что затрудняет оценку точности модели для этой конкретной группе звезд. Следовательно, требуются дальнейшие исследования в этой области.

Глава 4. Эмпирическая модель межзвездного поглощения на основе данных спектроскопического обзора LAMOST

В данной главе описывается построение эмпирической модели поглощения, основанной на вычислении избытков цвета и поглощения из данных спектроскопических обзоров. Основные результаты опубликованы в статьях Avdeeva A., Kovaleva D., Malkov O., Nekrasov A. Fitting procedure for estimating interstellar extinction at high galactic latitudes // Open Astronomy. — 2021. — Дек. — Т. 30, № 1. — С. 168—175 и Malkov O. Y., Avdeeva A. S., Kovaleva D. A., Nekrasov A. D. Interstellar Extinction at High Galactic Latitudes: An Analytical Approximation // Astronomy Reports. — 2022. — Июль. — Т. 66, № 7. — С. 526—534. В данной главе в качестве источника эффективных температур используются данные спектроскопического обзора RAVE, а также обобщаются результаты из работы [29].

4.1 Наблюдательные данные и оценка поглощения

Мы выбрали 40 различных площадок в зоне покрытия спектроскопического обзора RAVEб расположенных в высоких ($b > 20°$) галактических широтах (см. Рис. 4.1). Предел $|b|$ был выбран исходя из тех соображений, что закон косеканса удовлетворительно описывает зависимость $A_V(d)$ только в высоких широтах, вдали от галактической плоскости. Каждая область представляет собой конус с радиусом 80 секунд дуги, определяемый центральным направлением.

Обзор RAVE (Radial Velocity Experiment) представляет собой спектроскопическое исследование звёзд Галактики, отобранных случайным образом в южном полушарии и ограниченных по звездной величине ($9 < \mathrm{I} < 12$). Данные RAVE основаны на спектрах среднего разрешения ($R \sim 7500$), покрывающих область триплета кальция (8410 - 8795 A). В шестом выпуске данных RAVE, который используется в данной работе, представлены калиброванные по длине волны и нормализованные по потоку спектры для 518387 наблюдений 451783 уникальных звёзд. Финальный на момент проведения исследования выпуск

данных включает в себя спектры, спектральную классификацию, полученные радиальные скорости и перекрестные сопоставления с другими соответствующими каталогами.

Дополнительно каталог предоставляет информацию об атмосферных параметрах звёзд, в частности, спектроскопически полученных атмосферных параметрах звёзд, параметрах звёзд, полученных с использованием байесовского метода исходя из астрометрических данных Gaia DR2, и астеросейсмически калиброванных атмосферных параметрах звёзд гигантов на основе астеросейсмических наблюдений для 699 звёзд K2. Мы использовали каталог BDASP RAVE DR6 (https://www.rave-survey.org/metadata/ravedr6/dr6_bdasp/) в качестве источника спектроскопических параметров звезд.

Было проведено кросс-сопоставление объектов RAVE DR6 и Gaia EDR3 в каждой выбранной области. Для каждого объекта были взяты расстояния из каталога [92]. Для отдельных звезд в каждой области мы вычислили визуальное поглощение, используя следующую формулу:

$$A_V = \frac{c1}{c2} \cdot \Big((BP - RP) - (BP - RP)_0 \Big), \tag{4.1}$$

где $c1$ - это отношение A_G к $E(BP - RP)$, а $c2$ - отношение A_G к A_V. A_G - это межзвездное поглощение в полосе G у Gaia. Согласно [93], $c2 = 0.84$. Коэффициент $c1 = 2.02$ был рассчитан согласно [94], предполагая, что значение R_V равно 3.1.

Стоит отметить, что $c1$ и $c2$ на самом деле не являются постоянными из-за широты полосы G Gaia [95] (см. также https://www.cosmos.esa.int/web/gaia/edr3-extinction-law). Если за A_0 обозначить межзвездное поглощение при $\lambda = 550$ нм, то $A_G/A_0, A_{BP}/A_0, A_{RP}/A_0$ могут меняться в зависимости от эффективной температуры и ускорения свободного падения на поверхности звезды $\log g$, а также от самого значения A_0. Зачастую, $A_k/A_0 < 1$. Однако, оба коэффициента $c1, c2$ изменяются согласованно, поэтому изменения их отношения в Ур. 4.1 можно считать незначительными. В данной работе мы пренебрегаем непостоянством этих параметров.

Собственные показатели цвета $(BP-RP)_0$ были рассчитаны согласно [96] (таблица также опубликована на http://www.pas.rochester.edu/~emamajek/EEM_dwarf_UBVIJHK_colors_Teff.txt). В таблице [96] представлено однозначное соответствие эффективных температур и собственных показателей цвета

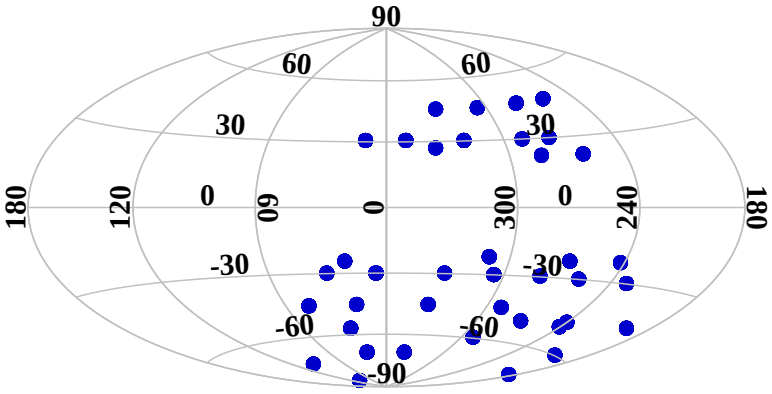

Рисунок 4.1 — Распределение выбранных областей по небесной сфере в галактических координатах в проекции Аитова - модифицированной азимутальной проекции.

для звезд главной последовательности. В интервалах между значениями, приведенными в таблице, была проведена линейная интерполяция.

В некоторых выбранных областях общий тренд поглощения с увеличением расстояния оказался отрицательным из-за низких значений поглощения у далеких объектов. Мы связываем это с ненадежным определением эффективной температуры для некоторых объектов в данных RAVE. Чтобы исключить звезды с чрезмерно низкими поглощениями на больших расстояниях, мы установили следующие феноменологические ограничения на данные, используемые в работе. Объекты, по которым было рассчитано поглощение должны удовлетворять следующим критериям: $1.6 > BP - RP > 0.8$ и $\log g > 3.5$. Другие проблемы с данными и возможные причины обсуждаются в разделе 4.2.

Полученные по оставшимся после фильтрации данные поглощения сравнивались с тремя различными моделями или трехмерными картами межзвездного поглощения в четырех различных областях небесной сферы. Это сравнение представлено на Рис. 4.2.

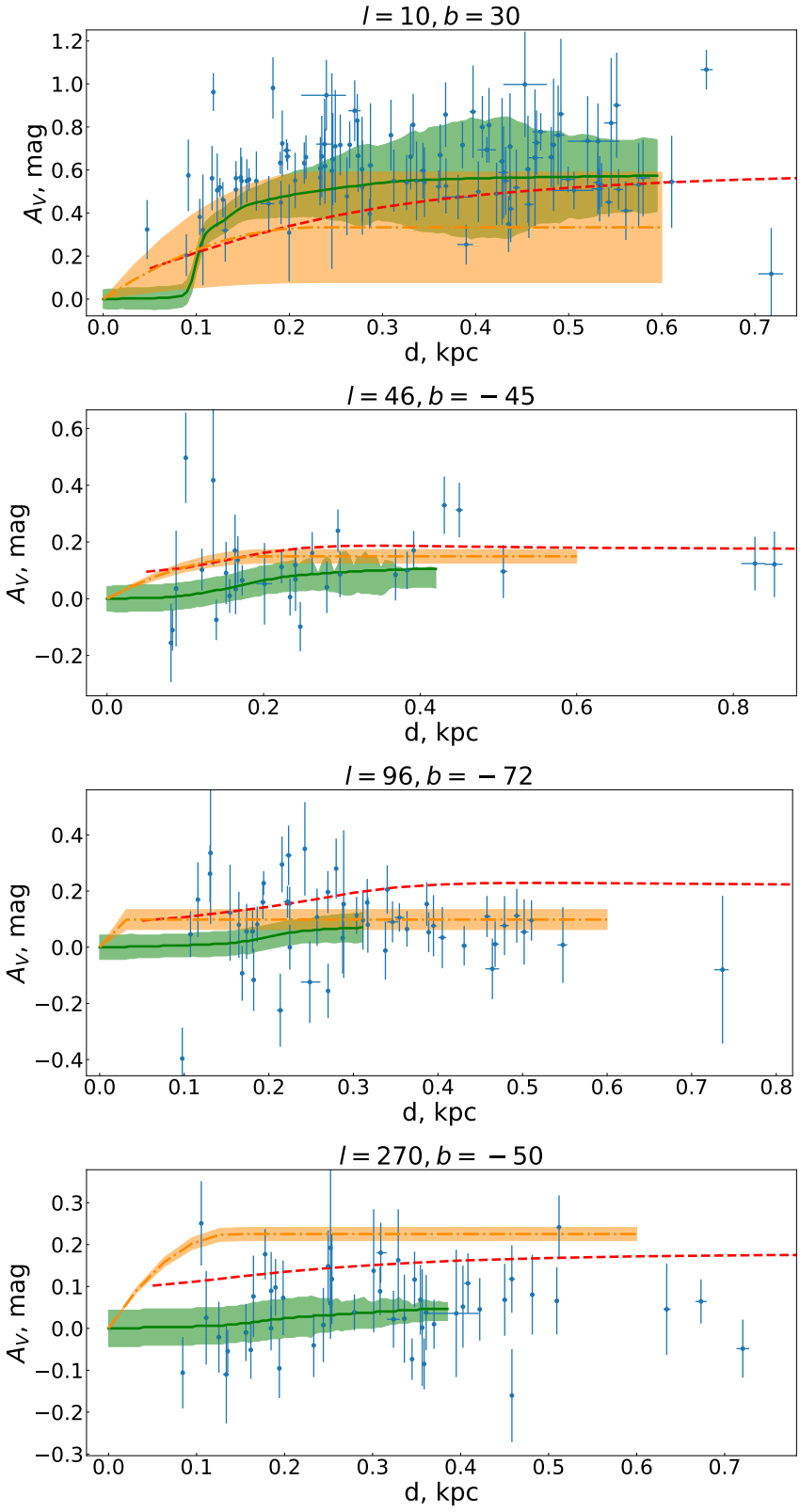

Рисунок 4.2 — Сравнение поглощений, полученных в данной работе, с тремя различными моделями или трехмерными картами межзвездного поглощения. Зеленая сплошная линия построена на основе значений, предоставленных картой Stilism ([54]). Оранжевая штрихпунктирная линия построена по значениям из работы [63]. Красная пунктирная линия отражает данные модели [65], в которой погрешности A_V составляют менее 0.03 звездной величины.

4.2 Вычисление параметров закона косеканса в отдельных областях

В каждой отдельно взятой площадке мы аппроксимируем полученную зависимость поглощения от расстояния барометрическим законом, описанном во введении (см. Ур. 1). Для определения параметров a_0 и β закона косеканса, минимизировалась следующая функция:

$$\chi^2 = \sum_{n=1}^{N} \left(\frac{A_V(d_n) - A_{V,n}}{\varepsilon_n(d_n, A_{V,n})} \right)^2 \tag{4.2}$$

Здесь $A_{V,n}$ - значения поглощения, которые вычислены по данным RAVE и Gaia, $A_V(d_n)$ - значения, предсказанные законом косеканса (1) для расстояния d_n; ε_n - погрешность поглощения и расстояний, N - общее число объектов в области.

Были использованы два метода для нахождения минимума функции: минимизация методом наилучшего соответствия и сканирование области параметров или метод грубого перебора. Описание методов приводится ниже.

4.2.1 Минимизация функции χ^2 методом наилучшего соответствия

Подбор параметров методом наилучшего соответствия для выражения 4.2 был проведен с использованием пакета lmfit для языка программирования Python [97]. Мы использовали алгоритм Левенберга-Марквардта из предложенных стандартных методов. Алгоритм Левенберга-Марквардта используется для решения задачи минимизации при аппроксимации функций. Он сочетает в себе методы Гаусса-Ньютона и градиентного спуска, обеспечивая устойчивость и эффективность в случае сложных нелинейных проблем. Алгоритм позволяет находить локальные минимумы функции, минимизируя сумму квадратов остатков между наблюдаемыми и модельными значениями. Стандартные ошибки представляют собой оценку неопределенности 1σ, рассчитанную как отклонение при котором значение χ^2 уменьшается на 1. Примеры успешной оценки параметров a_0 и β представлены на Рис. 4.3.

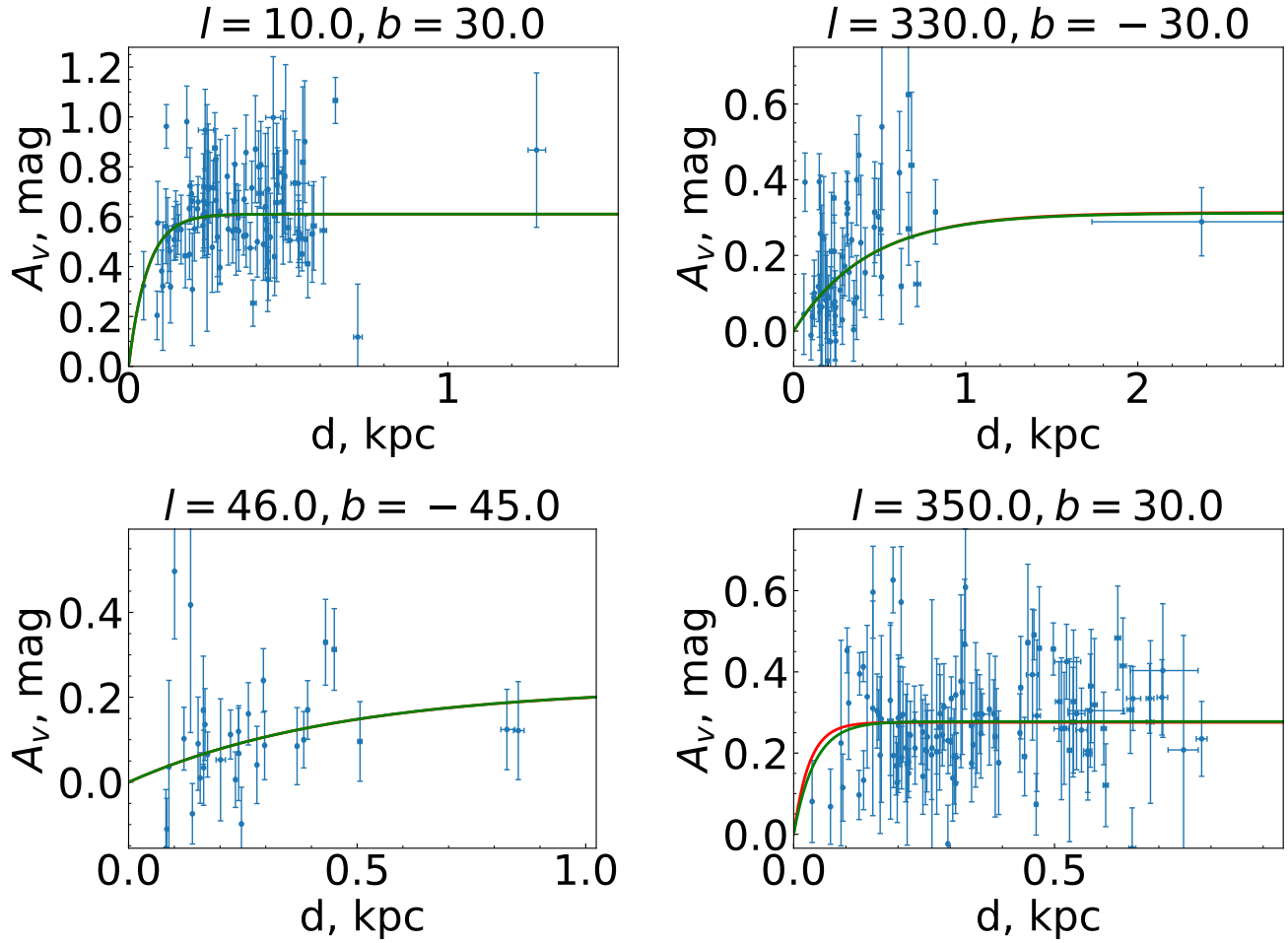

Рисунок 4.3 — Примеры минимизации χ^2. Галактические координаты центра области указаны вверху каждого графика. Зеленая линия - результат аппроксимации с использованием метода наилучшего соответствия χ^2, а красная - решение, полученное с помощью сканирования χ^2.

Аппроксимация была успешно выполнена для 36 из 40 областей. В трех других областях решение не было найдено. Это могло быть вызвано как значительным количеством отрицательных значений поглощения на луче зрения, так и несоответствующим трендом поглощения, а именно немонотонным или убывающим. Так как барометрическая функция предполагает неубывающую последовательность неотрицательных поглощений, аппроксимация таких данных барометрической функцией может не обладать решением. Еще одну область мы исключили из анализа из-за сомнительно низкого значения a_0. Мы сохранили решения с высокими значениями ошибок, такие как области под номерами 3, 20 и 37 (эти номера можно найти в Таб. 15), поскольку они имеют разумные значения a_0 и β и могут внести положительный вклад в аппроксимацию всего неба.

Можно предположить, что основной проблемой при аппроксимации в "плохих" областях являются систематические эффекты, влияющие на значения $T_{\rm eff}$ в каталоге RAVE. Поскольку температуры, взятые из данных LAMOST и данных RAVE для одних и тех же объектов, пересекающихся в обзорах, систематически отличаются, мы предполагаем, что проблема кроется именно в температурах RAVE. Исследование [**2021EPJST.tmp..185N**], проведенное на основе данных LAMOST, не показало такого сильного расхождения с законом косеканса, в то время как в исследовании, основанном на данных RAVE DR6, встречаются распределения поглощения с расстоянием, несовместимые с выражением (1). Ранее уже были отмечены отличия в оценках радиальных скоростей для пересечения обзоров LAMOST и RAVE [98].

4.2.2 Сканирование области параметров a_0 и β

Помимо алгоритма Левенберга-Марквардта для минимизации, мы вычислили значения χ^2 на сетке (a_0, β) и нашли минимум χ^2 методом грубого перебора. Стандартные ошибки были вычислены как контур 1σ. На Рис. 4.4 показаны карты χ^2 для тех же областей на небе, что представлены на Рис. 4.3. Результаты аппроксимации, полученные методом грубого перебора, также отображены на Рис. 4.3 красными линиями. Если на графике видна только одна зеленая линия, то это значит, что оба решения совпадают.

Размер и форма синей области на карте χ^2 отражают вырожденность параметров, то есть диапазон параметров a_0 и β, соответствующий значениям χ^2, близким к минимуму, может быть достаточно широким. Это может приводить к большим ошибкам и в результатах отражено высоким уровнем 1σ. Следует отметить, что в этом методе ошибки естественно ограничены размером сетки и не всегда отражают фактическую неопределенность. Размер синих областей на картах χ^2 также может быть связан с низкой надежностью данных, которая обсуждалась выше.

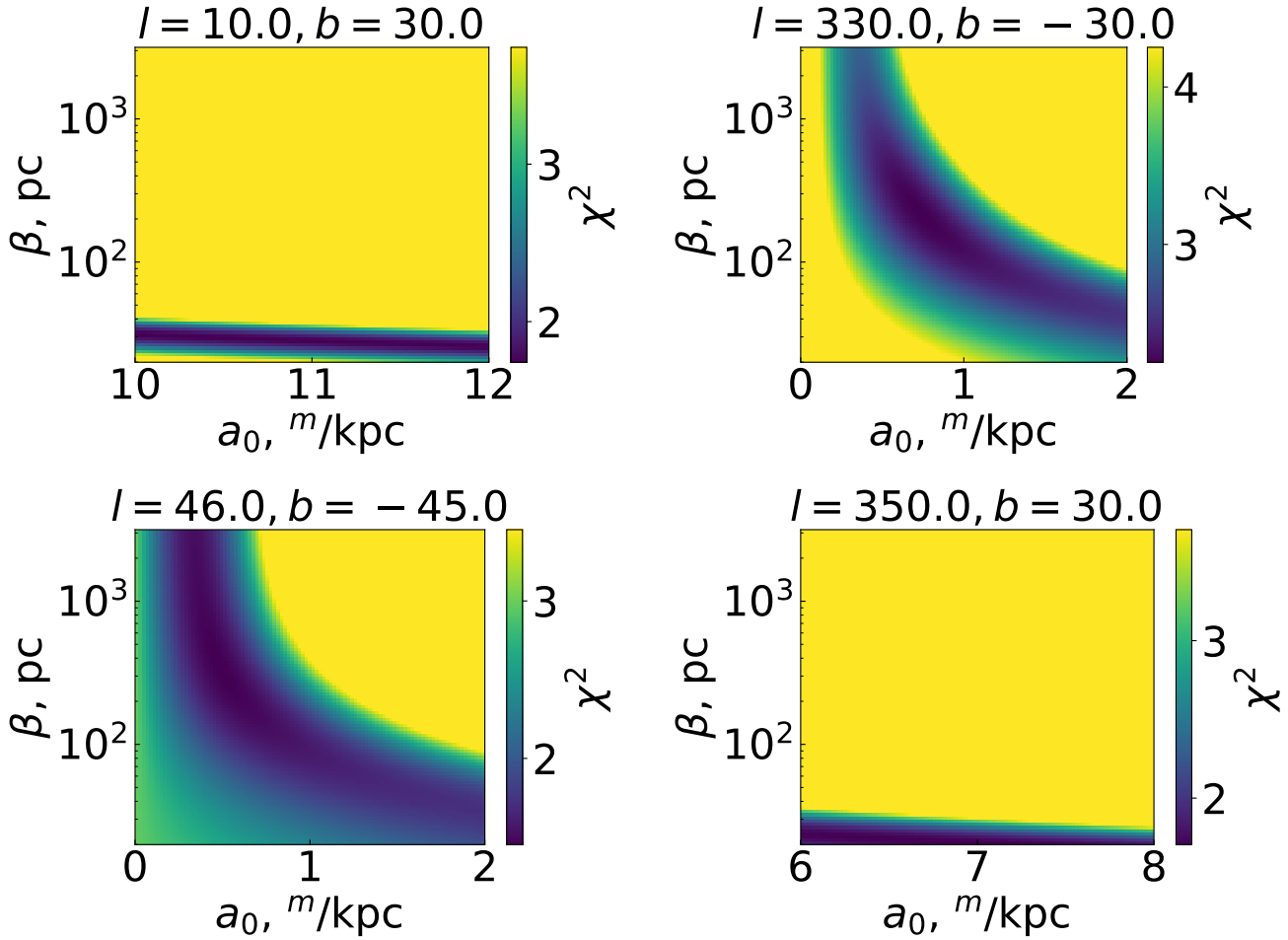

Рисунок 4.4 — Примеры решений метода грубого перебора для тех же областей, что и на Рис. 4.3.

4.3 Результаты аппроксимации в пределах областей

В обоих методах было вычислено значение χ^2. Мы выбирали для каждой области тот метод, решение которого обеспечивает минимальное значение χ^2. Окончательные результаты представлены в Таб. 15.

Также была произведена оценка общего галактического поглощения. Общее галактическое поглощение в полосе V, A_{Gal}, является особенно важным в контексте внегалактических исследований. Более того, оно играет ключевую роль в масштабировании расстояний, поскольку масштаб зависит от расстояния до эталонных объектов, которое в свою очередь зависит от их абсолютных величин, для определения которых необходимо знание общего галактического поглощения.

Общее галактическое поглощение можно оценить по формуле Ур. 1, если положить расстояние стремящимся к бесконечности:

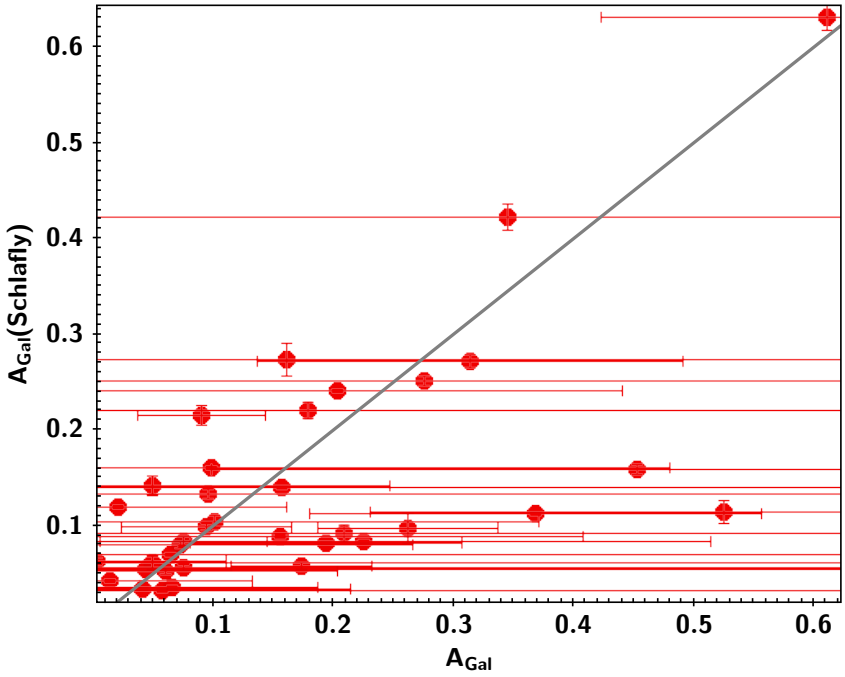

Рисунок 4.5 — Галактическое поглощение $A_{\rm Gal}$ в сравнении с данными из работы [67]. Идеальное соответствие показано сплошной линией.

$$A_{\rm Gal}(l,b) = \frac{a_0 \cdot \beta}{\sin|b|} \tag{4.3}$$

Мы вычислили значения $A_{\rm Gal}$ для каждой из наших областей. Результаты также представлены в Таб. 15. Мы сравнили значения $A_{\rm Gal}$ со значениями, предсказанными в работе [67] (см. Рис. 4.5).

4.4 Аппроксимация параметров закона косеканса по всему небу и окончательная формула

Наконец, параметры a_0, β и полное галактическое поглощение $A_{\rm Gal}$ были аппроксимированы по всему небу полиномом, состоящим из сферических гармоник порядка и степени 2:

$$\begin{aligned} f(l,b) = A_{00}Y_0^0\left(l,\frac{\pi}{2}-b\right) + A_{10}Y_1^0\left(l,\frac{\pi}{2}-b\right) + A_{11}Y_1^1\left(l,\frac{\pi}{2}-b\right) + \\ + A_{20}Y_2^0\left(l,\frac{\pi}{2}-b\right) + A_{21}Y_2^1\left(l,\frac{\pi}{2}-b\right) + A_{22}Y_2^2\left(l,\frac{\pi}{2}-b\right) \end{aligned} \tag{4.4}$$

Таблица 15 — Параметры межзвездного поглощения в выбранных областях

Area	l	b	a_0, mag/kpc	β, pc	A_{Gal}
1	5.0	-30.0	6 ± 16	14 ± 32	0.18 ± 0.66
2	30.0	-30.0	0.8 ± 0.4	55 ± 32	0.09 ± 0.03
3	230.0	-30.0	7 ± 9e5	3 ± 4e5	0.05 ± 900
5	280.0	-30.0	5 ± 13	15 ± 37	0.16 ± 0.56
6	305.0	-30.0	15 ± 54	11 ± 41	0.35 ± 1.75
7	330.0	-30.0	0.7 ± 0.2	216 ± 110	0.31 ± 0.18
8	17.0	-45.0	3.4 ± 31.4	19 ± 172	0.09 ± 1.17
9	46.0	-45.0	0.5 ± 0.2	330 ± 397	0.23 ± 0.29
10	195.0	-45.0	3.5 ± 8.5	32 ± 79	0.16 ± 0.55
12	270.0	-50.0	0.22 ± 0.24	210 ± 440	0.06 ± 0.14
13	290.0	-45.0	0.9 ± 2.4	40 ± 115	0.05 ± 0.19
15	335.0	-45.0	0.2 ± 0.9	70 ± 360	0.02 ± 0.14
16	26.0	-56.9	0.05 ± 2.2	50 ± 2000	0.003 ± 0.18
17	96.7	-72.9	1.6 ± 2.9	60 ± 120	0.1 ± 0.3
18	200.0	-60.0	2 ± 15	30 ± 220	0.08 ± 0.7
19	290.0	-60.0	5 ± 520	12 ± 1300	0.1 ± 10
20	20.0	-70.0	6 ± 7e4	7 ± 8e4	0.05 ± 800
22	340.0	-70.0	0.21 ± 0.04	190 ± 700	0.04 ± 0.17
23	275.0	30.0	3.2 ± 8.5	15 ± 43	0.1 ± 0.4
24	290.0	30.0	0.9 ± 0.5	115 ± 114	0.2 ± 0.2
25	320.0	30.0	6 ± 390	9 ± 590	0.1 ± 9.5
26	350.0	30.0	9.2 ± 8.1	15 ± 13	0.28 ± 0.35
27	10.0	30.0	10.5 ± 2.2	29.3 ± 6.6	0.61 ± 0.19
28	113.0	-85.0	0.1 ± 0.5	140 ± 1000	0.014 ± 0.119
29	260.3	46.3	0.5 ± 0.9	100 ± 200	0.07 ± 0.19
30	279.5	45.8	0.42 ± 0.45	110 ± 170	0.07 ± 0.12
31	305.0	45.0	6 ± 30	20 ± 100	0.17 ± 1.25
32	330.0	45.0	2.50 ± 1.15	55 ± 28	0.19 ± 0.13
33	242.0	-22.0	2.3 ± 25.0	10 ± 110	0.06 ± 0.92
34	269.0	-23.0	6.4 ± 21.0	13 ± 45	0.2 ± 1.0
35	310.0	-22.0	4.8 ± 1.6	41 ± 19	0.5 ± 0.3
36	19.9	-24.8	6.9 ± 3.5	28 ± 16	0.5 ± 0.4
37	262.3	22.3	19 ± 7e6	1.5 ± 5e5	0.07 ± 4e5
38	335.0	-26.5	7.5 ± 2.6	22 ± 8	0.37 ± 0.19
39	283.9	-22.4	1.91 ± 1.98	31 ± 39	0.16 ± 0.25
40	245.4	-20.3	4.3 ± 7.6	20 ± 38	0.26 ± 0.68

Таблица 16 — Коэффициенты полиномиальной аппроксимации для a_0, β и $A_{\rm Gal}$. Результаты аппроксимации действительны только для областей с $|b| > 20°$. Кроме того, они не могут использоваться в областях, обозначенных пустыми местами на Рис. 4.6.

$f(l,b)$	a_0	β	$A_{\rm Gal}$
A_{00}	10.0 ± 1.5	811.4 ± 210.2	0.967 ± 0.113
A_{10}	-0.9 ± 1.5	495.3 ± 212.4	-0.148 ± 0.114
A_{11}	-6.64 ± 2.18	715.7 ± 298.9	0.02 ± 0.16
A_{20}	-4.0 ± 1.8	336.3 ± 250.6	-0.51 ± 0.13
A_{21}	-2.6 ± 1.9	520.2 ± 266.8	-0.43 ± 0.14
A_{22}	0.77 ± 2.19	32.4 ± 300.5	0.56 ± 0.16

Для аппроксимации мы использовали параметры для 36 из 40 областей, перечисленных в Таб. 15. Также были использованы данные, полученные в работе [29]. Коэффициенты полинома для аппроксимации a_0, β и $A_{\rm Gal}$ представлены в Таб. 16.

Подставляя значения коэффициентов A_{ij} в Ур. 4.4, а затем в закон косеканса (Ур. 1), можно получить аналитическое выражение для оценки межзвездного поглощения в определенном направлении на определенном расстоянии.

Следует отметить, что из-за гармонической формы функции аппроксимации в решении возникают области, где значения параметров становятся отрицательными. Они показаны на Рис. 4.6 пустыми белыми участками. В этих областях мы либо не можем ничего сказать о поглощении, либо оно приблизительно равно нулю (как в областях, близких к полюсу). Использование результатов аппроксимации нежелательно в этих областях. Кроме того, нужно подчеркнуть, что данная аппроксимация дает оценки только для высоких галактических широт $|b| > 20°$.

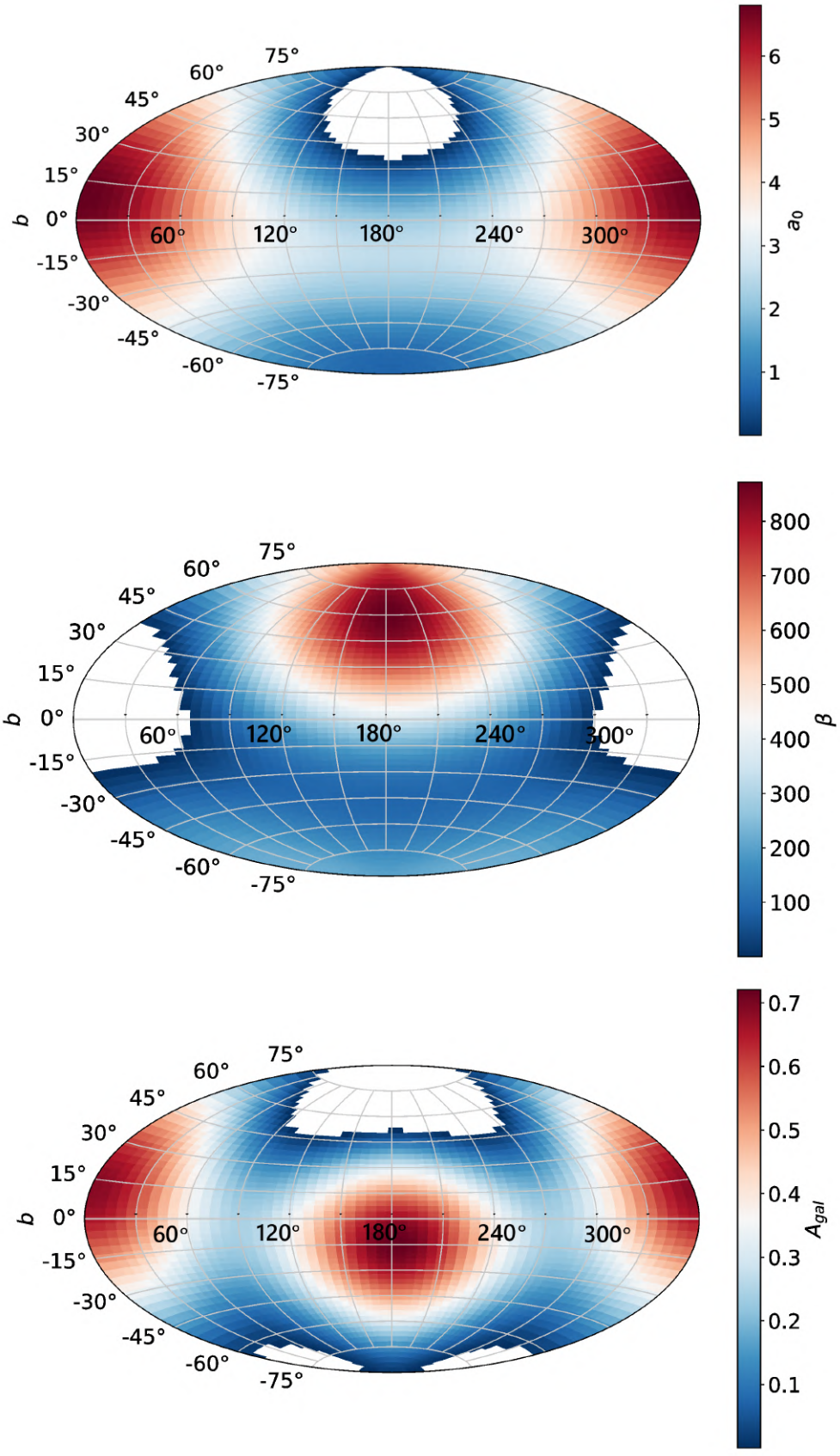

Рисунок 4.6 — Аппроксимация параметров по всему небу. Области с отрицательными значениями обозначены пустыми пространствами.

4.4.1 Ошибки полученных формул

Неопределенности a_0, β и A_{Gal} зависят от широты и долготы луча зрения и могут быть рассчитаны следующим образом:

$$\delta f(l,b) = \Big[\delta A_{00}^2 \cdot k_{00}^2 + \delta A_{10}^2 \cdot k_{10}^2 + \delta A_{11}^2 \cdot k_{11}^2 + \\ + \delta A_{20}^2 \cdot k_{20}^2 + \delta A_{21}^2 \cdot k_{21}^2 + \delta A_{22}^2 \cdot k_{22}^2\Big]^{1/2} \quad (4.5)$$

где $k_{i,j} = Y_i^j(l, \frac{\pi}{2} - b)$.

По этой формуле также можно оценить относительную неопределенность поглощения A_V, которая зависит от широты, долготы и расстояния:

$$\frac{\delta A_V}{A_V}(l,b,d) = \left[\frac{\delta a_0^2}{a_0^2} + \frac{\delta\beta^2}{\beta^2} \cdot \left(1 + \frac{d \cdot \sin(|b|)}{\beta} \cdot \frac{\exp(\frac{-d \cdot \sin(|b|)}{\beta})}{(1 - \exp(\frac{-d \cdot \sin(|b|)}{\beta})}\right)\right]^{1/2} \quad (4.6)$$

Примеры межзвездного поглощения A_v, рассчитанные с учетом неопределенности, показаны на Рис. 4.7. Как правило, относительная неопределенность имеет тенденцию быть больше в более высоких широтах, хотя она также зависит и от долготы.

4.5 Обсуждение результатов главы

С помощью данных Gaia EDR3 и RAVE DR6 мы получили межзвездное поглощение в полосе V для объектов в 40 областях южного неба. Сравнение рассчитанных поглощений с различными моделями и картами межзвездного поглощения показывает качественное сходство, что указывает на адекватность выбранного метода определения поглощения.

Аппроксимация в площадках показала, что качество данных об эффективных температурах из обзора RAVE неудовлетворительное для решения данной задачи. В некоторых областях успешная аппроксимация не была выполнена

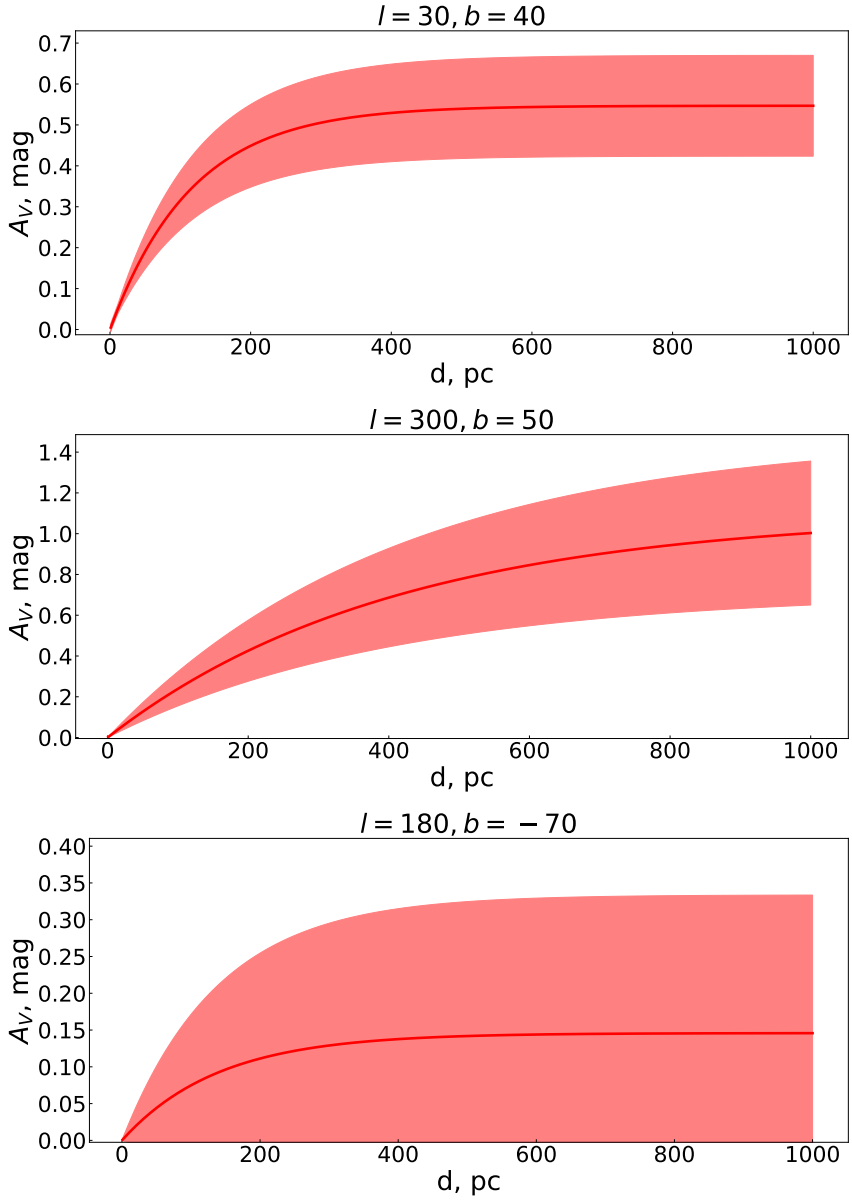

Рисунок 4.7 — Примеры зависимости поглощения на луче зрения от расстояния для трех лучей зрения. Закрашенные области демонстрируют рассчитанные неопределенности в решении.

из-за тенденции поглощения быть ниже на более далеких расстояниях. Мы предполагаем, что эта проблема вызвана систематическими эффектами, которым подвержены эффективные температуры, представленные в каталоге обзора RAVE.

Для областей, в которых аппроксимация была проведена успешно, выполнено приближение параметров закона косеканса полиномом из сферических гармоник и получено приближенное аналитическое описание распределения поглощения по небу. Тем не менее, существуют области, где решение стремится к нулю в пределах погрешности (области с отрицательными значениями параметров), поэтому разумного решения для таких областей найти не представляется возможным. Для получения разумных решений для областей с низким полным поглощением необходимо уменьшить ошибку аппроксимации на порядок. Кроме того, настоящее исследование основано на данных для высоких галактических широт, и его результаты применимы только к областям с $|b| > 20$.

Проблема с несоответствием эффективных температур, принятых по данным исследований LAMOST и RAVE, остается. Использование других обзоров (таких как APOGEE или SEGUE), вероятно, поможет нам получить дополнительную информацию и сделать выводы о достоверности используемых данных путем сравнения температур.

Заключение

Основные результаты работы заключаются в следующем.

1. Были разработаны фотометрические правила для поиска коричневых карликов в обзорах WISE, 2MASS и DES с учетом разделения на три подгруппы, яркие, транзитные и слабые коричневые карлики. Применяя эти правила, мы провели поиск коричневых карликов в области пересечения трех обзоров и обнаружили 135 объектов, удовлетворяющих условиям. Из них 96 объектов были дополнительно исследованы с использованием данных Gaia. Собственные движения этих объектов, рассчитанные по разнице положений в обзорах 2MASS и DES, оказались качественно и количественно согласованными с измерениями Gaia. Было продемонстрировано, что по крайней мере треть коричневых карликов не может быть обнаружена с помощью обзора Gaia. Кроме того, 11 из найденных объектов не обнаруживаются в базе данных SIMBAD, что позволяет сделать вывод о том, что эти коричневые карлики обнаружены впервые.
2. Была исследована возможность и перспективы использования моделей машинного обучения для выделения коричневых карликов в фотометрических обзорах WISE, 2MASS и Pan-STARRS. Для этого был создан обучающий набор данных, включающий коричневые карлики, красные карлики и звезды других спектральных классов. На этом наборе были обучены три модели классического машинного обучения и одна нейронная сеть. Обученные модели на тестовых данных превзошли существующие в литературе методы поиска коричневых карликов, основанные на показателях цвета (точность 0.97-0.98 против 0.935). Это демонстрирует большой потенциал использования методов машинного обучения для решения данной задачи. Исследование также выявило перспективность использования показателя цвета $(i-y)_{PS1}$ для фильтрации коричневых карликов среди других объектов. Было показано, что применение одного правила $(i-y)_{PS1} > 1.88$ на тестовом наборе данных обеспечивает результативность классификации на уровне 0.968. Этот показатель цвета ранее не рассматривался в литературе для

задачи выделения коричневых карликов, что подчеркивает новизну и значимость проведенного исследования.

3. Произведено сравнение оценок эффективных температур из обзора Gaia DR3 модуля GSP-Phot с эффективными температурами из спектроскопических обзоров высокого разрешения, APOGEE и GALAH. Сравнение показало, что, несмотря на то, что для большинства объектов из пересечения Gaia DR3 и APOGEE/GALAH эффективные температуры показывают хорошее согласие, для значительного числа звезд наблюдаются большие отклонения в оценках эффективных температур между оценками Gaia GSP-Phot и спектроскопическими обзорами высокого разрешения. В частности, звезды, близкие к плоскости Галактики, с высокими значениями A0 в Gaia DR3 могут иметь отличие эффективных температур от эталонных до 30 000 K.
4. Обучены несколько моделей классического машинного обучения для выделения из полной выборки эффективных температур Gaia GSP-Phot в 471 миллион объектов только объектов с температурами, совпадающих с эталонными температурами в пределах 250K. По данным моделей в Gaia GSP-Phot порядка 66% (313 миллионов) объектов обладают надежными оценками эффективной температуры. Верификация результатов была проведена с помощью каталогов APOGEE/GALAH, а также на каталоге PASTEL, сборнике атмосферных параметров звезд, полученных из различных спектроскопических исследований высокого разрешения.
5. На южном небе, по данным спектроскопического обзора RAVE, фотометрическим и астрометрическим данным Gaia EDR3 и соотношению эффективная температура - собственный показатель цвета, была исследована зависимость поглощения от расстояния для 40 участков. В 36 из этих участков, с использованием закона косеканса, удалось получить оценки полного галактического поглощения. Сравнение полученных данных с ранее известными оценками позволило сделать вывод о недостаточной точности эффективных температур RAVE для адекватной оценки межзвездного поглощения.

В дальнейшем планируется расширить метод поиска коричневых карликов на оптический обзор Pan-STARRS, поскольку он покрывает примерно в пять раз большую площадь неба, хотя и является менее глубоким. Предполага-

ется также использовать обзоры APOGEE и GALAH, а также планирующиеся обзоры, такие как WEAVE и 4MOST. В настоящее время также проводится работа по гомогенизации параметров атмосферы, полученных по данным разных обзоров [99], результаты которой также могут оказаться полезными для данной задачи.

Благодарности

Автор хотела бы выразить благодарность своему научному руководителю, Малкову Олегу Юрьевичу, за доверие, всестороннюю поддержку и ценные советы, а также Ковалевой Дане Александровне за помощь на всех этапах выполнения диссертационной работы.

Также хотелось бы выразить благодарность коллективу Института астрономии РАН за теплый прием, несмотря на то, что я проходила подготовку в другой аспирантуре.

Особая благодарность выражается моей семье и в особенности моему мужу, Авдееву Никите Алексеевичу, за безоговорочную веру в то, что у меня все получится.

Список литературы

1. *Kumar S. S.* The Structure of Stars of Very Low Mass. // The Astrophysical Journal. — 1963. — Т. 137. — С. 1121—1125.
2. *Hayashi C.*, *Nakano T.* Evolution of Stars of Small Masses in the Pre-Main-Sequence Stages // Progress of Theoretical Physics. — 1963. — Т. 30, № 4. — С. 460—474.
3. *Rebolo R.*, *Zapatero Osorio M. R.*, *Martin E. L.* Discovery of a brown dwarf in the Pleiades star cluster // Nature. — 1995. — Т. 377, № 6545. — С. 129—131.
4. *Nakajima T.* [и др.]. Discovery of a cool brown dwarf // Nature. — 1995. — Т. 378, № 6556. — С. 463—465.
5. *Luhman K. L.* Discovery of a Binary Brown Dwarf at 2 pc from the Sun // The Astrophysical Journal Letter. — 2013. — Т. 767, № 1. — id.L1, 6 pp. — arXiv: 1303.2401.
6. *Burningham B.* [и др.]. 76 T dwarfs from the UKIDSS LAS: benchmarks, kinematics and an updated space density // Monthly Notices of the Royal Astronomical Society. — 2013. — Т. 433, № 1. — С. 457—497. — arXiv: 1304.7246.
7. *Carnero Rosell A.* [и др.]. Brown dwarf census with the Dark Energy Survey year 3 data and the thin disc scale height of early L types // Monthly Notices of the Royal Astronomical Society. — 2019. — Т. 489, № 4. — С. 5301—5325. — arXiv: 1903.10806 [astro-ph.SR].
8. *Kirkpatrick J. D.* [и др.]. Dwarfs Cooler than "M": The Definition of Spectral Type "L" Using Discoveries from the 2 Micron All-Sky Survey (2MASS) // The Astrophysical Journal. — 1999. — Т. 519, № 2. — С. 802—833.
9. *Skrzypek N.*, *Warren S. J.*, *Faherty J. K.* VizieR Online Data Catalog: Photometric brown-dwarf classification (Skrzypek+, 2016) // VizieR Online Data Catalog. — 2016. — J/A+A/589/A49.
10. *Kirkpatrick J. D.* [и др.]. The Field Substellar Mass Function Based on the Full-sky 20 pc Census of 525 L, T, and Y Dwarfs // Astrophysical Journal Supplement Series. — 2021. — Т. 253, № 1. — id.7, 85 pp.

11. *Mužić K.* [и др.]. The low-mass content of the massive young star cluster RCW 38 // Monthly Notices of the Royal Astronomical Society. — 2017. — Т. 471, № 3. — С. 3699—3712. — arXiv: 1707.00277 [astro-ph.SR].

12. *Smith L.* [и др.]. High proper motion objects from the UKIDSS Galactic plane survey // Monthly Notices of the Royal Astronomical Society. — 2014. — Т. 443, № 3. — С. 2327—2341. — arXiv: 1406.6698 [astro-ph.SR].

13. *Lodieu N.* [и др.]. Binary frequency of planet-host stars at wide separations. A new brown dwarf companion to a planet-host star // Astronomy and Astrophysics. — 2014. — Т. 569. — id.A120, 14 pp. — arXiv: 1408.1208 [astro-ph.EP].

14. *Artigau É., Bouchard S., Doyon R., Lafreniére D.* Photometric Variability of the T2.5 Brown Dwarf SIMP J013656.5+093347: Evidence for Evolving Weather Patterns // The Astrophysical Journal. — 2009. — Т. 701. — С. 1534—1539. — URL: https://api.semanticscholar.org/CorpusID:3123957.

15. *Gillon M.* [и др.]. Fast-evolving weather for the coolest of our two new substellar neighbours // Astronomy & Astrophysics. — 2013. — Т. 555. — id.L5, 4 pp. — arXiv: 1304.0481 [astro-ph.SR].

16. *Khandrika H.* [и др.]. A Search for Photometric Variability in L- and T-type Brown Dwarf Atmospheres // The Astronomical Journal. — 2013. — Т. 145, № 3. — id.71, 16 pp. — arXiv: 1301.0545 [astro-ph.SR].

17. *Marley M. S., Robinson T. D.* On the Cool Side: Modeling the Atmospheres of Brown Dwarfs and Giant Planets // Annual Review of Astronomy and Astrophysics. — 2015. — Т. 53. — С. 279—323. — arXiv: 1410.6512 [astro-ph.EP]. — URL: https://ui.adsabs.harvard.edu/abs/2015ARA&A..53..279M.

18. *Charnay B.* [и др.]. A Self-consistent Cloud Model for Brown Dwarfs and Young Giant Exoplanets: Comparison with Photometric and Spectroscopic Observations // The Astrophysical Journal. — 2018. — Т. 854, № 2. — id.172, 20 pp. — arXiv: 1711.11483.

19. *Tremblin P.* [и др.]. Thermo-compositional Diabatic Convection in the Atmospheres of Brown Dwarfs and in Earth's Atmosphere and Oceans // The Astrophysical Journal. — 2019. — Т. 876, № 2. — id. 144, 13 pp. — arXiv: 1902.03553.

20. *Tan X.*, *Showman A. P.* Atmospheric Variability Driven by Radiative Cloud Feedback in Brown Dwarfs and Directly Imaged Extrasolar Giant Planets // The Astrophysical Journal. — 2019. — Т. 874, № 2. — id.111, 18 pp. — arXiv: 1809.06467.

21. *Burningham B.* [и др.]. Retrieval of atmospheric properties of cloudy L dwarfs // Monthly Notices of the Royal Astronomical Society. — 2017. — Т. 470, № 1. — С. 1177—1197. — arXiv: 1701.01257.

22. *Saumon D.*, *Marley M. S.* The Evolution of L and T Dwarfs in Color-Magnitude Diagrams // The Astrophysical Journal. — 2008. — Т. 689, № 2. — С. 1327—1344. — arXiv: 0808.2611.

23. *Vos J. M.* [и др.]. Astro2020 White Paper: The L/T Transition. — 2019. — eprint: 1903.06691v1 (astro-ph.SR).

24. *Maravelias G.* [и др.]. A machine-learning photometric classifier for massive stars in nearby galaxies. I. The method // Astronomy & Astrophysics. — 2022. — Т. 666. — id.A122, 26 pp. — arXiv: 2203.08125 [astro-ph.SR].

25. *Pedregosa F.* [и др.]. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Т. 12. — С. 2825—2830.

26. *Mohammadi M.*, *Mutatiina J.*, *Saifollahi T.*, *Bunte K.* Detection of extragalactic Ultra-compact dwarfs and Globular Clusters using Explainable AI techniques // Astronomy and Computing. — 2022. — Т. 39. — С. id.100555.

27. *Dréau G.*, *Lebreton Y.*, *Mosser B.*, *Bossini D.*, *Yu J.* Characterising the AGB bump and its potential to constrain mixing processes in stellar interiors // Astronomy & Astrophysics. — 2022. — Т. 668. — id.A115, 20 pp. — arXiv: 2207.00571 [astro-ph.SR].

28. *Grunblatt S. K.* [и др.]. Age-dating Red Giant Stars Associated with Galactic Disk and Halo Substructures // The Astrophysical Journal. — 2021. — Т. 916, № 2. — id.88, 19 pp. — arXiv: 2105.10505 [astro-ph.SR].

29. *Nekrasov A.*, *Grishin K.*, *Kovaleva D.*, *Malkov O.* Approximate analytical description of the high latitude extinction // European Physical Journal Special Topics. — 2021. — Т. 230, № 10. — С. 2193—2205. — arXiv: 2106.03081 [astro-ph.GA].

30. *Sun M.*, *Jiang B.*, *Yuan H.*, *Li J.* The Ultraviolet Extinction Map and Dust Properties at High Galactic Latitude // The Astrophysical Journal Supplement Series. — 2021. — Т. 254, № 2. — id.38, 12 pp. — arXiv: 2104.08505 [astro-ph.GA].

31. *Buder S.* [и др.]. The GALAH+ survey: Third data release // Monthly Notices of the Royal Astronomical Society. — 2021. — Т. 506, № 1. — С. 150—201. — arXiv: 2011.02505.

32. *Jönsson H.*, *Holtzman J. A.*, *al et.* APOGEE Data and Spectral Analysis from SDSS Data Release 16: Seven Years of Observations Including First Results from APOGEE-South // Astronomical Journal. — 2020. — Т. 160, № 3. — id.120, 32pp. — arXiv: 2007.05537 [astro-ph.GA].

33. *Abdurro'uf* [и др.]. The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data // Astrophysical Journal Supplement Series. — 2022. — Т. 259, № 2. — id.35, 39 pp. — arXiv: 2112.02026 [astro-ph.GA].

34. *Gilmore G.*, *Randich S.*, *al. et.* The Gaia-ESO Public Spectroscopic Survey // The Messenger. — 2012. — Т. 147. — С. 25—31.

35. *Collaboration G.* The Gaia mission // Astronomy and Astrophysics. — 2016. — Т. 595. — id.A1, 36 pp. — arXiv: 1609.04153 [astro-ph.IM].

36. *Gaia Collaboration* [и др.]. Gaia Data Release 3. Summary of the content and survey properties // Astronomy & Astrophysics. — 2023. — Т. 674. — id.A1, 22 pp. — arXiv: 2208.00211 [astro-ph.GA].

37. *Creevey O. L.* [и др.]. Gaia Data Release 3. Astrophysical parameters inference system (Apsis). I. Methods and content overview // Astronomy & Astrophysics. — 2023. — Т. 674. — id.A26, 35 pp.

38. *Carrasco J. M.* [и др.]. Internal calibration of Gaia BP/RP low-resolution spectra // Astronomy & Astrophysics. — 2021. — Т. 652. — id.A86, 20 pp. — arXiv: 2106.01752 [astro-ph.IM].

39. *De Angeli F.* [и др.]. Gaia Data Release 3. Processing and validation of BP/RP low-resolution spectral data // Astronomy & Astrophysics. — 2023. — Т. 674. — id.A2, 28 pp. — arXiv: 2206.06143 [astro-ph.IM].

40. *Montegriffo P.* [и др.]. Gaia Data Release 3. External calibration of BP/RP low-resolution spectroscopic data // Astronomy & Astrophysics. — 2023. — Т. 674. — id.A3, 33 pp. — arXiv: 2206.06205 [astro-ph.IM].

41. *Andrae R.* [и др.]. Gaia Data Release 3. Analysis of the Gaia BP/RP spectra using the General Stellar Parameterizer from Photometry // Astronomy & Astrophysics. — 2023. — Июнь. — Т. 674. — id.A27, 22 pp. — arXiv: 2206.06138 [astro-ph.SR].

42. *Gustafsson B.* [и др.]. A grid of MARCS model atmospheres for late-type stars. I. Methods and general properties // Astronomy & Astrophysics. — 2008. — Т. 486, № 3. — С. 951—970. — arXiv: 0805.0554 [astro-ph].

43. *Brott I., Hauschildt P. H.* A PHOENIX Model Atmosphere Grid for Gaia // The Three-Dimensional Universe with Gaia. Т. 576 / под ред. С. Turon, K. S. O'Flaherty, M. A. C. Perryman. — 2005. — С. 565—568. — (ESA Special Publication). — arXiv: astro-ph/0503395 [astro-ph].

44. *Lanz T., Hubeny I.* A Grid of Non-LTE Line-blanketed Model Atmospheres of O-Type Stars // The Astrophysical Journal Supplement Series. — 2003. — Т. 146, № 2. — С. 417—441. — arXiv: astro-ph/0210157 [astro-ph].

45. *Lanz T., Hubeny I.* A Grid of NLTE Line-blanketed Model Atmospheres of Early B-Type Stars // The Astrophysical Journal Supplement Series. — 2007. — Т. 169, № 1. — С. 83—104. — arXiv: astro-ph/0611891 [astro-ph].

46. *Borisov S. B.* [и др.]. New Generation Stellar Spectral Libraries in the Optical and Near-infrared. I. The Recalibrated UVES-POP Library for Stellar Population Synthesis // The Astrophysical Journal Supplement Series. — 2023. — Т. 266, № 1. — id.11, 20 pp. — arXiv: 2211.09130 [astro-ph.IM].

47. *Brandner W., Calissendorff P., Kopytova T.* Benchmarking Gaia DR3 Apsis with the Hyades and Pleiades open clusters // Astronomy & Astrophysics. — 2023. — Т. 677. — id.A162, 8 pp. — arXiv: 2306.03132 [astro-ph.SR].

48. *Andrae R., Rix H.-W., Chandra V.* Robust Data-driven Metallicities for 175 Million Stars from Gaia XP Spectra // The Astrophysical Journal Supplement Series. — 2023. — Т. 267, № 1. — id.8, 15 pp. — arXiv: 2302.02611 [astro-ph.SR].

49. *Zhang X.*, *Green G. M.*, *Rix H.-W.* Parameters of 220 million stars from Gaia BP/RP spectra // Monthly Notices of the Royal Astronomical Society. — 2023. — Т. 524, № 2. — С. 1855—1884. — arXiv: 2303.03420 [astro-ph.SR].

50. *Schlegel D. J.*, *Finkbeiner D. P.*, *Davis M.* Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds // The Astrophysical Journal. — 1998. — Т. 500, № 2. — С. 525—553. — arXiv: astro-ph/9710327 [astro-ph].

51. *Marshall D. J.*, *Robin A. C.*, *Reylé C.*, *Schultheis M.*, *Picaud S.* Modelling the Galactic interstellar extinction distribution in three dimensions // Astronomy & Astrophysics. — 2006. — Т. 453, № 2. — С. 635—651.

52. *Sale S. E.* [и др.]. A 3D extinction map of the northern Galactic plane based on IPHAS photometry // Monthly Notices of the Royal Astronomical Society. — 2014. — Т. 443, № 4. — С. 2907—2922.

53. *Green G. M.*, *Schlafly E.*, *Zucker C.*, *Speagle J. S.*, *Finkbeiner D.* A 3D Dust Map Based on Gaia, Pan-STARRS 1, and 2MASS // The Astrophysical Journal. — 2019. — Т. 887, № 1. — id.93, 27pp. — arXiv: 1905.02734 [astro-ph.GA].

54. *Lallement R.* [и др.]. 3D maps of the local ISM from inversion of individual color excess measurements // Astronomy & Astrophysics. — 2014. — Т. 561. — id.A91, 17 pp.

55. *Vergely J.*, *Valette B.*, *Lallement R.*, *Raimond S.* Spatial distribution of interstellar dust in the Sun's vicinity. Comparison with neutral sodium-bearing gas // Astronomy & Astrophysics. — 2010. — Т. 518. — id.A31, 12 pp. — arXiv: 1002.4578 [astro-ph.GA].

56. *Cramer N.* Calibrations for B-type stars in the Geneva photometric system // New Astronomy Reviews. — 1999. — Т. 43, № 5. — С. 343—387.

57. *Nordström B.* [и др.]. The Geneva-Copenhagen survey of the Solar neighbourhood. Ages, metallicities, and kinematic properties of ∼14 000 F and G dwarfs // Astronomy & Astrophysics. — 2004. — Т. 418. — С. 989—1019. — arXiv: astro-ph/0405198 [astro-ph].

58. *Casagrande L.* [и др.]. New constraints on the chemical evolution of the solar neighbourhood and Galactic disc(s). Improved astrophysical parameters for the Geneva-Copenhagen Survey // Astronomy & Astrophysics. — 2011. — Т. 530. — id.A138, 21 pp. — arXiv: 1103.4651 [astro-ph.GA].

59. *Lallement R.* [и др.]. Gaia-2MASS 3D maps of Galactic interstellar dust within 3 kpc // Astronomy & Astrophysics. — 2019. — Т. 625. — id.A135, 16 pp.

60. *Chen B.-Q.* [и др.]. A three-dimensional extinction map of the Galactic anticentre from multiband photometry // Monthly Notices of the Royal Astronomical Society. — 2014. — Т. 443, № 2. — С. 1192—1210.

61. *Gontcharov G. A.* Influence of the Gould belt on interstellar extinction // Astronomy Letters. — 2009. — Т. 35, № 11. — С. 780—790. — arXiv: 1606.09624 [astro-ph.SR].

62. *Parenago P. P.* On interstellar extinction of light // Astron. Zh. — 1940. — Т. 13. — С. 3.

63. *Arenou F., Grenon M., Gomez A.* 16. A tridimensional model of the galactic interstellar extinction. // Astronomy & Astrophysics. — 1992. — Т. 258. — С. 104—111.

64. *Gontcharov G. A., Mosenkov A. V.* Gaia DR2 giants in the Galactic dust - I. Reddening across the whole dust layer and some properties of the giant clump // Monthly Notices of the Royal Astronomical Society. — 2021. — Т. 500, № 2. — С. 2590—2606. — arXiv: 2011.11113 [astro-ph.GA].

65. *Gontcharov G. A., Mosenkov A. V.* Gaia DR2 giants in the Galactic dust - II. Application of the reddening maps and models // Monthly Notices of the Royal Astronomical Society. — 2021. — Т. 500, № 2. — С. 2607—2619. — arXiv: 2011.13811 [astro-ph.GA].

66. *Best W. M. J.* [и др.]. Photometry and Proper Motions of M, L, and T Dwarfs from the Pan-STARRS1 3π Survey // Astrophysical Journal Supplement Series. — 2018. — Т. 234, № 1. — id.1, 37 pp. — arXiv: 1701.00490.

67. *Schlafly E. F., Finkbeiner D. P.* Measuring reddening with Sloan Digital Sky Survey stellar spectra and recalibrating SFD. // The Astrophysical Journal. — 2011. — Т. 737, № 2. — id.103, 13 pp.

68. *Abbott T. M. C.* [и др.]. The Dark Energy Survey: Data Release 1 // The Astrophysical Journal Supplement Series. — 2018. — Т. 239, № 2. — id.18, 25 pp. — arXiv: 1801.03181 [astro-ph.IM].

69. *Chambers K. C.* [и др.]. The Pan-STARRS1 Surveys // arXiv e-prints. — 2016. — arXiv:1612.05560. — arXiv: 1612.05560.

70. *Cutri R. M.* [и др.]. VizieR Online Data Catalog: 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003) // VizieR Online Data Catalog. — 2003. — С. II/246.

71. *Cutri R. M.* [и др.]. VizieR Online Data Catalog: AllWISE Data Release (Cutri+ 2013) // VizieR Online Data Catalog. — 2021. — С. II/328.

72. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A Next-generation Hyperparameter Optimization Framework. — 2019. — arXiv: 1907.10902v1 [cs.LG].

73. *Lundberg S. M., Lee S.-I.* A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems 30 / под ред. I. Guyon [и др.]. — Curran Associates, Inc., 2017. — С. 4765—4774.

74. *Cortes C., Vapnik V.* Support-vector networks // Machine learning. — 1995. — Т. 20, № 3. — С. 273—297.

75. *Arik S. O., Pfister T.* TabNet: Attentive Interpretable Tabular Learning // arXiv e-prints. — 2019. — arXiv:1908.07442.

76. *Kingma D. P., Ba J.* Adam: A Method for Stochastic Optimization // arXiv e-prints. — 2014. — arXiv:1412.6980. — arXiv: 1412.6980 [cs.LG].

77. *Skrzypek N.* [и др.]. Photometric brown-dwarf classification // Astronomy & Astrophysics. — 2015. — Т. 574. — id.A78, 14pp.

78. *Ulla A.* [и др.]. Gaia DR3 documentation Chapter 11: Astrophysical parameters. — 2022. — URL: https://ui.adsabs.harvard.edu/abs/2022gdr3.reptE..11U. Gaia DR3 documentation, European Space Agency; Gaia Data Processing and Analysis Consortium.

79. *González Hernández J. I., Bonifacio P.* A new implementation of the infrared flux method using the 2MASS catalogue // Astronomy & Astrophysics. — 2009. — Т. 497, № 2. — С. 497—509. — arXiv: 0901.3034 [astro-ph.SR].

80. *Riello M.* [и др.]. Gaia Early Data Release 3. Photometric content and validation // Astronomy & Astrophysics. — 2021. — Т. 649. — id.A3, 33 pp. — arXiv: 2012.01916 [astro-ph.IM].

81. *Fabricius C.* [и др.]. Gaia Early Data Release 3. Catalogue validation // Astronomy & Astrophysics. — 2021. — Т. 649. — id.A5, 25 pp. — arXiv: 2012.06242.

82. *Hegedűs V.* [и др.]. Comparative analysis of atmospheric parameters from high-resolution spectroscopic sky surveys: APOGEE, GALAH, Gaia-ESO // Astronomy & Astrophysics. — 2023. — Т. 670. — id.A107, 22 pp. — arXiv: 2211.03416 [astro-ph.SR].

83. *Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P.* SMOTE: Synthetic Minority Over-sampling Technique // Journal Of Artificial Intelligence Research. — 2002. — Т. 16. — С. 321—357. — arXiv: 1106.1813 [cs.AI].

84. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A Next-Generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. — Anchorage, AK, USA : Association for Computing Machinery, 2019. — С. 2623—2631. — (KDD '19).

85. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System. — 2016. — arXiv: 1603.02754v3 [cs.LG].

86. *Prokhorenkova L., Gusev G., Vorobev A., Veronika Dorogush A., Gulin A.* CatBoost: unbiased boosting with categorical features // arXiv e-prints. — 2017. — arXiv:1706.09516.

87. *Ke G.* [и др.]. Lightgbm: A highly efficient gradient boosting decision tree // Advances in neural information processing systems. — 2017. — Т. 30. — С. 3146—3154.

88. *Borisov V.* [и др.]. Deep Neural Networks and Tabular Data: A Survey // arXiv e-prints. — 2021. — arXiv:2110.01889. — arXiv: 2110.01889 [cs.LG].

89. *Bailer-Jones C. A. L., Fouesneau M., Andrae R.* Quasar and galaxy classification in Gaia Data Release 2 // Monthly Notices of the Royal Astronomical Society. — 2019. — Т. 490, № 4. — С. 5615—5633. — arXiv: 1910.05255.

90. *Soubiran C.*, *Le Campion J.-F.*, *Brouillet N.*, *Chemin L.* VizieR Online Data Catalog: The PASTEL catalogue (Soubiran+, 2016-) // VizieR Online Data Catalog. — 2020. — B/pastel.

91. *Soubiran C.*, *Le Campion J.-F.*, *Brouillet N.*, *Chemin L.* The PASTEL catalogue: 2016 version // Astronomy & Astrophysics. — 2016. — T. 591. — id.A118, 7 pp. — arXiv: 1605.07384 [astro-ph.SR].

92. *Bailer-Jones C. A. L.*, *Rybizki J.*, *Fouesneau M.*, *Demleitner M.*, *Andrae R.* Estimating Distances from Parallaxes. V. Geometric and Photogeometric Distances to 1.47 Billion Stars in Gaia Early Data Release 3 // The Astrophysical Journal. — 2021. — T. 161, № 3. — id.147, 24pp. — arXiv: 2012.05220 [astro-ph.SR].

93. *Bono G.* [и др.]. On a New Method to Estimate the Distance, Reddening, and Metallicity of RR Lyrae Stars Using Optical/Near-infrared (B, V, I, J, H, K) Mean Magnitudes: ω Centauri as a First Test Case // The Astrophysical Journal. — 2019. — T. 870, № 2. — id. 115, 21 pp. — arXiv: 1811.07069 [astro-ph.SR].

94. *Cardelli J. A.*, *Clayton G. C.*, *Mathis J. S.* The Relationship between Infrared, Optical, and Ultraviolet Extinction // The Astrophysical Journal. — 1989. — T. 345. — C. 245—256.

95. *Danielski, C.*, *Babusiaux, C.*, *Ruiz-Dern, L.*, *Sartoretti, P.*, *Arenou, F.* The empirical Gaia G-band extinction coefficient // Astronomy & Astrophysics. — 2018. — T. 614. — id.A19, 8 pp.

96. *Pecaut M. J.*, *Mamajek E. E.* Intrinsic Colors, Temperatures, and Bolometric Corrections of Pre-main-sequence Stars // The Astrophysical Journal Supplement Series. — 2013. — T. 208, № 1. — id.9, 22pp. — eprint: 1307.2657.

97. *Newville M.* [и др.]. Lmfit: Non-Linear Least-Square Minimization and Curve-Fitting for Python. — 06.2016. — Astrophysics Source Code Library, record ascl:1606.014.

98. *Vickers J. J.*, *Smith M. C.* The Lives of Stars: Insights from the TGAS-RAVE-LAMOST Data Set // The Astrophysical Journal. — 2018. — T. 860, № 2. — id.91, 16pp. — arXiv: 1805.02332 [astro-ph.GA].

99. *Tsantaki M.* [и др.]. Survey of Surveys. I. The largest compilation of radial velocities for the Galaxy // Astronomy & Astrophysics. — 2022. — Т. 659. — id.A95, 24 pp. — eprint: 2110.09316 (astro-ph.GA).

Список рисунков

Список таблиц