

SRGZ: МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ И СВОЙСТВА КАТАЛОГА ОПТИЧЕСКИХ КОМПАЬОНОВ ТОЧЕЧНЫХ РЕНТГЕНОВСКИХ ИСТОЧНИКОВ СРГ/ЕРОЗИТА В ОБЛАСТИ ПОКРЫТИЯ DESI LEGACY IMAGING SURVEYS

© 2023 г. А. В. Мещеряков^{1,2*}, В. Д. Борисов^{2,1}, Г. А. Хорунжев¹,
П. А. Медведев¹, М. Р. Гильфанов^{1,3}, М. И. Бельведерский¹,
С. Ю. Сазонов¹, Р. А. Буренин¹, Р. А. Кривонос¹, И. Ф. Бикмаев⁴,
И. М. Хамитов⁴, С. В. Герасимов², И. В. Машечкин², Р. А. Сюняев^{1,3}

¹Институт космических исследований РАН, Москва, Россия

²Московский государственный университет им. М.В. Ломоносова, Москва, Россия

³Институт им. Макса Планка, Гаршинг, Германия

⁴Казанский федеральный университет, Казань, Россия

Поступила в редакцию 08.12.2022 г.

После доработки 14.07.2023 г.; принята к публикации 11.08.2023 г.

Описаны методы системы SRGz для физического отождествления точечных рентгеновских источников обзора еРОЗИТА по фотометрическим данным в области покрытия обзора DESI Legacy Imaging Surveys. Рассмотрены модели, входящие в состав системы SRGz (версия 2.1), позволившие для 87% площади всей восточной внегалактической области ($0 < l < 180^\circ$, $|b| > 20^\circ$) получить точные измерения космологического красного смещения и класса рентгеновского объекта (квазар/галактика/звезда) по данным многоволновых фотометрических обзоров неба (DESI LIS, SDSS, Pan-STARRS, WISE, еРОЗИТА). Важной особенностью системы SRGz является то, что ее модель работы с данными (алгоритмы отождествления, классификации, photo-z) целиком опирается на эвристические подходы машинного обучения. Для стандартного выбора параметров SRGz полнота отождествления оптических компаньонов в области DESI LIS составляет 95% (с точностью отбора оптических компаньонов 94%); полнота классификации рентгеновских источников, не имеющих оптических партнеров в обзоре DESI LIS, составляет 82% (точность — 85%). В SRGz достигается высокое качество фотометрической классификации оптических компаньонов рентгеновских источников: >99% полнота фотометрической классификации внегалактических объектов (квазар или галактика) и звезд на тестовой выборке источников со спектрами SDSS и астрометрическими звездами GAIA. Представлен анализ важности различных фотометрических признаков для оптического отождествления и классификации рентгеновских источников еРОЗИТА. Мы показали, что ИК-величина W_2 , отношения рентген/оптика (ИК), цвета в оптическом (например, $(g-r)$) и ИК-цвет ($W_1 - W_2$), а также, введенные нами цветовые расстояния играют существенную роль при разделении классов рентгеновских объектов. Мы используем наиболее важные фотометрические признаки для интерпретации прогнозов SRGz в настоящей работе. Точность фотометрических красных смещений SRGz (по фотометрическим данным DESI LIS, SDSS, Pan-STARRS, WISE) тестировалась в поле Stripe82X на выборке 3/4 оптических компаньонов точечных рентгеновских источников еРОЗИТА (для которых в Stripe82X доступны спектроскопические измерения) и составила: $\sigma_{NMAD} = 3.1\%$ (нормализованное медианное абсолютное отклонение прогноза) и $n_{>0.15} = 7.8\%$ (доля катастрофических выбросов). Представленные результаты photo-z для источников еРОЗИТА в поле Stripe82X более чем в 2 раза по обоим метрикам точности (σ_{NMAD} и $n_{>0.15}$) превосходят результаты photo-z других групп, опубликованные в каталоге Stripe82X.

Ключевые слова: рентгеновские источники, СРГ/еРОЗИТА.

DOI: 10.31857/S0320010823070021, **EDN:** YDWFAX

*Электронный адрес: mesch@cosmos.ru

1. ВВЕДЕНИЕ

Космическая обсерватория СРГ (Сюняев и др., 2021) с рентгеновскими телескопами еРОЗИТА (Предел и др., 2021) и АРТ-ХС им. Павлинского (Павлинский и др., 2021) на борту была успешно запущена 13 июля 2019 г. с космодрома Байконур, а уже 12 декабря 2019 г. (после фазы проверки работоспособности и полетной калибровки, англ. CalPV) спутник начал рентгеновский обзор всего неба с гало-орбиты вокруг точки либрации L2 Солнце–Земля. Полный обзор неба СРГ/еРОЗИТА должен был состоять из восьми последовательных полных сканирований неба, каждое продолжительностью 6 мес. К концу 2021 г. еРОЗИТА завершила двухлетние обзорные наблюдения, в которых зарегистрировала более 2 млн точечных рентгеновских источников (с лог-правдоподобием детектирования $\mathcal{L} > 6$) в восточной галактической полусфере $0 < l < 180^\circ$ ¹ (см., Медведев и др., 2022).

Для подавляющего большинства источников еРОЗИТА по данным из рентгеновского диапазона (в котором новые астрономические объекты были открыты) могут быть получены только координаты объекта и оценка его рентгеновского потока. Так, надежное измерение наклона рентгеновского спектра может быть получено лишь для небольшого процента всех рентгеновских источников еРОЗИТА. Таким образом, информации доступной непосредственно из рентгеновских наблюдений оказывается критически мало для определения физического класса объектов, их светимости и расстояния (космологического красного смещения для внегалактических источников). К счастью, на современном этапе развития наблюдательной астрофизики, данные СРГ/еРОЗИТА на всем небе могут быть дополнены широким набором спектральных, астрометрических и фотометрических обзоров неба. В то же время, бурное развитие технологий машинного обучения в последние годы дает исследователям эффективные инструменты для интеллектуального анализа больших массивов фотометрических данных из открытых астрономических архивов.

Современные широкоформатные спектроскопические и астрометрические обзоры неба важны тем, что позволяют получить физический класс оптического компаньона рентгеновского источника практически напрямую (по измерениям линий излучения/поглощения в случае спектрального обзора; по измерениям собственного движения и/или параллакса в случае астрометрического обзора)

для значительной части достаточно ярких оптических источников, у которых могут быть выполнены соответствующие измерения. В настоящей работе мы будем использовать данные спектрального обзора SDSS DR17 (Абдуллоуф и др., 2022) и каталога оптических квазаров DR16Q (Люк и др., 2020), а также данные астрометрического обзора GAIA DR2 (коллаборация GAIA и др., 2018). Спектральный обзор SDSS DR17 содержит информацию о более чем 5 млн оптических галактик, квазаров и звезд, отобранных выборочно по фотометрическим признакам вплоть до ~ 22 видимой звездной величины в рамках различных наблюдательных программ SDSS (APOGEE, BOSS/eBOSS, SEQUELS, eBOSS, MARVELS, MaNGA, SEGUE и LEGACY). SDSS DR16Q является наиболее полным на данный момент каталогом спектроскопических квазаров SDSS и содержит более 750 тыс. объектов. Астрометрический обзор GAIA содержит информацию о параллаксах и собственных движениях более миллиарда звезд на всем небе.

Наиболее крупными по площади покрытия неба и наиболее глубокими по чувствительности в оптическом и ИК-диапазонах фотометрическими обзорами неба для внегалактических источников в восточной галактической полусфере ($0 < l < 180^\circ$) являются DESI Legacy Imaging Surveys (Дей и др., 2019), Pan-STARRS1 (Чамберс и др., 2016), SDSS (Йорк и др., 2000) и WISE (Райт и др., 2010). Данные фотометрических обзоров неба значительно расширяют набор доступной информации по зарегистрированным точечным рентгеновским источникам еРОЗИТА. В отличие от спектральных наблюдений SDSS и астрометрических данных GAIA, фотометрические данные из обзоров неба доступны для всех оптических объектов в окрестности рентгеновских источников (оптических объектов выше порога чувствительности выбранного фотометрического обзора и в области его покрытия). Фотометрические данные, как правило, не позволяют непосредственно определить класс и физические параметры астрономических объектов. Применение методов машинного обучения открывает новые возможности для эффективного использования доступных широкополосных фотометрических наблюдений из обзоров неба. Целью таких моделей является (как можно более точное) автоматическое восстановление зависимости $X_{\text{ph}} \rightarrow (\text{Class}, Z, \dots)$ — функции отображения из пространства фотометрических признаков астрономических объектов (X_{ph}) в пространство их физических классов, расстояний и других характеристик, необходимых для понимания физической природы астрономических источников и проверки физических теорий. Для решения этой задачи

¹За анализ данных еРОЗИТА в восточной половине неба (в галактических координатах) отвечает российский консорциум еРОЗИТА.

применительно к выборкам точечных рентгеновских источников ePOZITA, в научной группе по рентгеновскому каталогу Российского Консорциума СРГ/ePOZITA, была создана система SRGz, алгоритмы которой, в области покрытия DESI LIS, описаны в настоящей работе.

2. ДАННЫЕ

Модели SRGz, представленные в данной работе, базируются на данных 11 обзоров неба в разных спектральных диапазонах. Во-первых, это рентгеновский каталог точечных источников по данным двух лет обзора неба ePOZITA в восточной галактической полусфере. Во-вторых, это вспомогательные рентгеновские каталоги CX2, 4XMM, 2RXS, полученные по данным обсерваторий Chandra, XMM-Newton, Swift, которые будут использоваться нами (см. п. 5.1.3) для обучения и оценки точности моделей оптического отождествления рентгеновских источников ePOZITA. Далее, это фотометрический каталог DESI Legacy Imaging Surveys, в котором собраны данные ряда оптических обзоров (MzLS, DECaLS, BASS), инфракрасные данные (принудительная фотометрия WISE) и астрометрические данные обзора GAIA. Наконец, для моделей photo-z мы будем использовать также данные двух крупнейших фотометрических обзоров неба: Pan-STARRS1 и SDSS.

Далее мы рассмотрим используемые нами данные более подробно.

2.1. Рентгеновский обзор СРГ/ePOZITA

Каталог точечных рентгеновских источников СРГ/ePOZITA по итогам двух лет сканирования неба в восточной галактической полусфере. В SRGz используется информация о координатах рентгеновского источника на небе, ошибке локализации источника и рентгеновском потоке в диапазоне 0.5–2 кэВ.

2.2. Вспомогательные рентгеновские данные обсерваторий XMM-Newton, Chandra, Swift

Для создания тренировочной и тестовой выборки модели отождествления рентгеновских источников ePOZITA нами использовались объединенные данные рентгеновских каталогов крупнейших обзоров неба, выполненных тремя рентгеновскими обсерваториями: XMM-Newton, Chandra, Swift. Во-первых, мы рассматривали рентгеновские источники из каталога 4XMM-DR12² (Вебб и др.,

2020). Из рентгеновского каталога уникальных источников 4XMM в области покрытия обзора DESI Legacy Imaging Survey были исключены источники с флагами $SC_SUM_FLAG > 0$ (имеют проблемы с детектированием или измерением параметров) и протяженные рентгеновские объекты с $SC_EXTENT > 0$. Во-вторых, мы рассматривали рентгеновские объекты из Chandra Source Catalog 2.0³ (CSC2, Эванс и др., 2010) за исключением источников с флагами: *conf_flag*, *dither_warning_flag*, *extent_flag*, *pileup_flag*, *sat_src_flag*, *streak_src_flag*. В-третьих, мы рассматривали рентгеновские объекты из каталога 2SXPS (Эванс и др., 2020) обсерватории Swift им. Нила Герелса. Из рентгеновского каталога точечных рентгеновских источников были исключены объекты со следующими флагами: *DetFlag*, *FieldFlag*, *OpticalLoadingWarning*, *StrayLightWarning*, *NearBrightSourceWarning*, *IsPotentialAlias*.

2.3. Оптические/ИК обзоры неба

DESI Legacy Imaging Surveys DR9. Каталог фотометрического обзора неба в оптическом диапазоне в трех фильтрах (g , r , z), полученный по данным трех обзоров неба, сделанных различными телескопами (Beijing-Arizona Sky Survey (BASS), Dark Energy Camera Legacy Survey (DECaLS), Mayall z-band Legacy Survey (MzLS)) и приведенных к единой фотометрической системе (Дей и др., 2019). Отметим, что в настоящей работе мы не отмечаем систематических сдвигов в результатах моделей SRGz, вызванных возможным остаточным различием фотометрических систем данных внутри обзора DESI LIS, поэтому далее мы не будем различать фотометрические измерения объектов DESI LIS, сделанных на разных телескопах.

WISE. Принудительная фотометрия по данным изображений всего неба спутника WISE в фильтрах W_1 и W_2 (после шести лет обзора NEOWISE-Reactivation) приведена в фотометрическом каталоге DESI LIS DR9.

GAIA. Данные по источникам GAIA DR2, сопоставленным с оптическими объектами, приведена в каталоге DESI LIS DR9.

SDSS. Данные фотометрического обзора неба в оптическом диапазоне в пяти фильтрах (u , g , r , i , z) для источников, ассоциированных (в 1") с оптическими объектами DESI LIS.

Pan-STARRS1. Данные фотометрического обзора неба Pan-STARRS1 DR2 (Чамберс и др., 2016) в оптическом диапазоне в пяти фильтрах (g , r , i , z , y) для источников, ассоциированных (в 1") с оптическими объектами DESI LIS.

²http://xmmssc.irap.omp.eu/Catalogue/4XMM-DR12/4XMM_DR12.html

³<https://cdsarc.cds.unistra.fr/viz-bin/cat/IX/57>

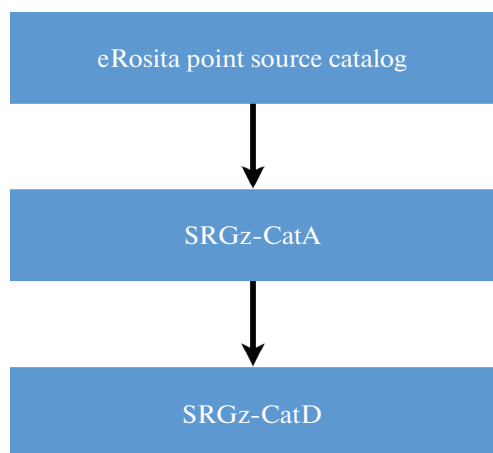


Рис. 1. Схема конвейера SRGz.

3. СИСТЕМА SRGz

Система SRGz была задумана нами как реализация возможности для всех рентгеновских источников в каталоге ePOZITA получить измерения физического класса и затем поместить их на трехмерную карту Вселенной. SRGz целиком базируется на анализе методами машинного обучения многоволновых обзоров неба в окрестности рентгеновских источников. Данная задача сама по себе является чрезвычайно обширной, а астрономические данные, на которые она опирается, в полной мере обладают характеристиками больших данных (англ. Big Data). Это так называемые 5V (от англ. Volume, Variety, Velocity, Veracity, Value, см. Ломотей, Дитерс, 2014), т.е. речь идет о больших объемах и разнообразии типов доступных астрономических данных, большой скорости возникновения новых массивов данных в астрономии, о наличии в сырых данных и каталогах большого числа артефактов, пропущенных значений и шума различной природы (что многократно усложняет их автоматический анализ), и, наконец, астрономические массивы данных внутри себя содержат ценную информацию для определения физических свойств астрономических объектов, которую позволяют извлекать техники машинного обучения. В рамках системы SRGz мы разрабатываем алгоритмы для извлечения информации из больших астрономических данных в интересах науки рентгеновского обзора ePOZITA.

Алгоритмы системы SRGz (версия 2.1), описанные в данной работе, позволяют отождествлять оптические компаньоны точечных рентгеновских источников ePOZITA в области покрытия фотометрического обзора DESI Legacy Imaging Surveys, классифицировать их на три основных класса (звезды, квазары, галактики), а также измерять фотометрические красные смещения до внегалактических объектов. Модели машинного обу-

чения для отождествления и классификации рентгеновских источников опираются на данные трех обзоров: рентгеновского (ePOZITA), оптического (DESI LIS) и инфракрасного (WISE, в рамках принудительной фотометрии в DESI LIS). Модели машинного обучения для измерения фотометрических красных смещений внегалактических объектов опираются на данные оптических и ИК-обзоров неба: DESI LIS DR9, WISE (в составе DESI LIS), Pan-STARRS1 DR2, SDSS DR17.

3.1. Схема работы SRGz

Работа системы SRGz представляет собой (см. рис. 1) конвейер (англ. pipeline) с предобработкой фотометрических данных и применением к ним моделей машинного обучения для оптического отождествления, классификации и измерения красных смещений рентгеновских источников.

На вход SRGz поступают рентгеновские данные в виде каталога точечных рентгеновских источников ePOZITA. Для удобства применения SRGz на всем небе весь каталог рентгеновских источников был разбит на фрагменты площадью 53.7 кв. градусов по схеме разбиения неба HEALPix (NESTED) с индексом разрешения (англ. resolution index) $Res=3$: в общей сложности 768 площадок на всей сфере, 415 содержат рентгеновские источники в восточной внегалактической области, и из них 266 (64%) покрыты оптическими данными обзора DESI LIS. В каталоге двух лет обзора ePOZITA в восточной галактической полусфере ≈ 2 млн точечных рентгеновских источников, из них ≈ 1.27 млн приходится на область покрытия DESI LIS DR9.

На этапе предобработки данных мы создаем каталог SRGz-CatA всех оптических объектов (DESI LIS, Pan-STARRS, SDSS), находящихся в окрестности точечных рентгеновских источников ePOZITA. Сначала для каждого объекта из рентгеновского каталога (на этапе первичного оптического сопоставления) отбираются все известные источники в оптическом диапазоне в радиусе $R_{\text{match}}^X = 30''$ из вспомогательных фотометрических, спектральных и астрометрических обзоров неба. В данной работе в качестве вспомогательных оптических/ИК данных используются: DESI Legacy Surveys DR9 (DESI LIS), Pan-STARRS1 DR2, WISE (в составе DESI LIS), SDSS DR17 и GAIA DR2 (в составе DESI LIS). Оптические источники из фотометрических каталогов DESI LIS, Pan-STARRS и SDSS соединяются между собой по расстоянию $R_{\text{match}}^{\text{Opt}} < 1''$ (при этом DESI LIS используется в качестве базовой оптической выборки). Данные WISE нами используются в двух фильтрах (W_1 , W_2) из каталога DESI LIS (полученные в результате принудительной фотометрии

на изображениях обзора WISE). Данные GAIA DR2 берутся из каталога DESI LIS. На этапе предобработки фотометрических данных мы получаем полные (в рамках выбранного обзора DESI LIS) списки оптических источников в окрестности R_{match}^X от рентгеновских объектов с доступными фотометрическими, спектральными, астрометрическими признаками из вспомогательных оптических/ИК данных.

На следующем этапе с помощью моделей машинного обучения для всех рентгеновских источников отождествляются наиболее вероятные оптические компаньоны или утверждается отсутствие оптического компаньона у рентгеновского объекта (“бездомный” (англ. *hostless*) рентгеновский источник). У рентгеновских источников с оптически-ми партнерами прогнозируется класс и измеряется фотометрическое красное смещение для внегалактических объектов. Полученный таким образом каталог SRGz-CatD содержит в себе рентгеновские объекты, их наиболее вероятные оптические компаньоны, физические характеристики объектов. Каталог SRGz-CatD далее может использоваться для решения различных научных задач, например, для анализа широкополосного спектра рентгеновских источников, исследования эволюции функции светимости квазаров, поиска уникальных объектов с большим красным смещением и/или светимостью (Хорунжев и др., 2021, 2022) и прочее.

Далее мы подробно опишем фотометрические признаки объектов и алгоритмы машинного обучения, на которых построены модели физической классификации и измерения красных смещений по фотометрическим данным об оптических источниках.

3.2. Признаки объектов

Модели SRGz для оптического отождествления, классификации и измерения *photo-z* точечных рентгеновских источников целиком базируются на методах машинного обучения и используют фотометрические признаки объектов из пяти широкоформатных обзоров неба в разных спектральных диапазонах (от рентгеновского до инфракрасного): ePOZITA, DESI Legacy Imaging Surveys, SDSS, Pan-STARRS1, WISE. В табл. 1 приведены наборы фотометрических признаков, которые используют различные модели системы SRGz (версия 2.1). Рассмотрим выбор признаков для моделей машинного обучения более подробно.

Первым признаком, который мы используем в моделях оптического отождествления рентгеновских источников, является расстояние до рассматриваемого оптического объекта (d), выраженное в

единицах ошибки локализации (σ_X) рентгеновского источника:

$$\text{offset} = \frac{d}{\sigma_X}. \quad (1)$$

Для всех рентгеновских источников каталога ePOZITA нам доступна откалиброванная инструментальная ошибка локализации рентгеновских объектов σ_X . Калибровка $\sigma_X(\sigma_{\text{det}})$ относительно инструментальной ошибки σ_{det} (полученной алгоритмом детектирования eSASS) была выполнена путем сопоставления рентгеновского каталога ePOZITA на восточной галактической полусфере с данными астрометрического каталога GAIA, как описано в работе Гильфанова и др. (2023).

Следующим фактором, влияющим на сопоставление объектов из рентгеновского и оптического каталогов, является поверхностная плотность оптических объектов поля в направлении на рентгеновский источник. Мы добавили в модель отождествления SRGz два фотометрических признака: n_{20} и n_{30} — число оптических объектов DESI LIS в кружках радиуса $20''$ и $30''$ соответственно с центром на рентгеновском объекте.

В моделях SRGz мы будем использовать рентгеновские, оптические, ИК-потоки объектов, выраженные в логарифмической шкале звездных величин. Рентгеновская “величина” рассчитывается нами из рентгеновского потока объекта в диапазоне 0.5–2 кэВ по следующей формуле:

$$\text{mag}_X = -2.5 \lg(F_{X,0.5-2}) + C, \quad (2)$$

где аддитивная константа $C = -18$ выбрана так, чтобы вычисленные с помощью mag_X отношения рентген/оптика (см. далее) для $X/O = 0$ хорошо разделяли классы рентгеновских квазаров и рентгеновских звезд. Отметим, что выбор константы C никак не влияет на работу моделей SRGz (при условии, что C фиксируется до обучения моделей и далее остается неизменной).

Оптические и ИК-измерения, приведенные в фотометрических каталогах, содержат как измерение потока объекта, так и ошибку на поток. Для части фотометрических объектов в каталоге величина потока может принимать отрицательное значения (это означает, что источник не детектируется в выбранном фильтре, и измеряется только верхний предел на поток). Чтобы учитывать данные особенности оптических измерений при построении моделей SRGz, мы рассчитали так называемую гиперболическую звездную величину по каждому фотометрическому измерению потока объекта и его ошибке (в каталогах обзоров DESI LIS, Pan-STARRS1, SDSS). Потоки объекта в оптических/ИК фильтрах были приведены к единым единицам измерений — наномагги (англ.

Таблица 1. Фотометрические признаки для моделей SRGz

#	Фотометрические признаки
Признаки лучшей модели отождествления оптического компаньона — $Model_{match}$	
3_X	$offset, n_{20}, n_{30}, g_{lim}, r_{lim}, z_{lim}, W_{1,lim}, W_{2,lim},$ $type, sersic, mag_X, X/g, X/r, X/z, X/W_1, X/W_2, (X/W_1)_{salvato},$ $g, r, z, W_1, W_2, (g - r), (g - z), (g - W_1), (g - W_2),$ $(r - z), (r - W_1), (r - W_2), (z - W_1), (z - W_2), (W_1 - W_2),$ $(g - W_1) \text{ distance}, (z - W_1) \text{ distance},$ $(z - W_2) \text{ distance}$
Признаки моделей классификации (звезда/квазар/галактика) — $Model1_{class}, Model2_{class}$	
$3_{X,col}$	$type, sersic, X/g, X/r, X/z, X/W_1, X/W_2, (X/W_1)_{salvato},$ $(g - r), (g - z), (g - W_1), (g - W_2), (r - z),$ $(r - W_1), (r - W_2), (z - W_1), (z - W_2), (W_1 - W_2),$ $(g - W_1) \text{ distance}, (z - W_1) \text{ distance},$ $(z - W_2) \text{ distance}$
Признаки моделей измерения фотометрических красных смещений — $Model_{photoz}$	
2	$g, r, z, W_1, W_2, (g - r), (g - z), (g - W_1), (g - W_2), (r - z),$ $(r - W_1), (r - W_2), (z - W_1), (z - W_2), (W_1 - W_2)$
3_{PS}	$features : 2 + g_{PS,psf}, r_{PS,psf}, i_{PS,psf}, z_{PS,psf}, y_{PS,psf}, i_{PS,kron}, y_{PS,kron}, g_{PS,kron},$ $r_{PS,kron}, z_{PS,kron}, (g_{PS,psf} - i_{PS,psf}), (g_{PS,psf} - y_{PS,psf}), (r_{PS,psf} - i_{PS,psf}),$ $(r_{PS,psf} - y_{PS,psf}), (i_{PS,psf} - z_{PS,psf}), (i_{PS,psf} - y_{PS,psf}), (z_{PS,psf} - y_{PS,psf}),$ $(i_{PS,psf} - i_{PS,kron}), (y_{PS,psf} - y_{PS,kron}), (g_{PS,psf} - r_{PS,psf}), (g_{PS,psf} - z_{PS,psf}),$ $(r_{PS,psf} - z_{PS,psf}), (g_{PS,psf} - g_{PS,kron}), (r_{PS,psf} - r_{PS,kron}), (z_{PS,psf} - z_{PS,kron}),$ $(g_{PS,kron} - g), (r_{PS,kron} - r), (z_{PS,kron} - z), (g_{PS,kron} - W_1), (g_{PS,kron} - W_2),$ $(r_{PS,kron} - W_1), (r_{PS,kron} - W_2), (i_{PS,kron} - W_1), (i_{PS,kron} - W_2),$ $(z_{PS,kron} - W_1), (z_{PS,kron} - W_2), (y_{PS,kron} - W_1), (y_{PS,kron} - W_2)$
3_{SDSS}	$features : 2 + u_{psf}, g_{psf}, r_{psf}, i_{psf}, z_{psf}, u_{cmodel}, i_{cmodel}, g_{cmodel}, r_{cmodel}, z_{cmodel},$ $(u_{psf} - g_{psf}), (u_{psf} - r_{psf}), (u_{psf} - i_{psf}), (u_{psf} - z_{psf}), (u_{psf} - u_{cmodel}), (g_{psf} - i_{psf}),$ $(g_{psf} - g_{cmodel}), (r_{psf} - i_{psf}), (i_{psf} - z_{psf}), (i_{psf} - i_{cmodel}), (g_{psf} - r_{psf}),$ $(g_{psf} - z_{psf}), (r_{psf} - z_{psf}), (r_{psf} - r_{cmodel}), (z_{psf} - z_{cmodel}), (g_{cmodel} - g),$ $(r_{cmodel} - r), (z_{cmodel} - z), (u_{cmodel} - W_1), (u_{cmodel} - W_2), (g_{cmodel} - W_1),$ $(g_{cmodel} - W_2), (r_{cmodel} - W_1), (r_{cmodel} - W_2), (i_{cmodel} - W_1), (i_{cmodel} - W_2),$ $(z_{cmodel} - W_1), (z_{cmodel} - W_2)$
4	$SET(features : 2 + features : 3_{SDSS} + features : 3_{PS})$

nanomaggie, $1 \text{ nanomaggie} = 10^{-(48.6+22.5)/2.5} = 3.63 \times 10^{-29} \text{ [эрг см}^{-2} \text{ с}^{-1}\text{]}$). Полученный поток f и его ошибка σ_f преобразуются в гиперболическую величину mag_O по следующей формуле:

$$\text{mag}_O = \left[\text{asinh}\left(\frac{f}{2\sigma_f}\right) + \log(\sigma_f) \right] \left(\frac{-2.5}{\log 10} \right). \quad (3)$$

Отметим, что введенная здесь формула для гиперболических звездных величин отличается от формулы для гиперболических величин в фотометрическом каталоге обзора SDSS (так называемые луптитуды, англ. *luptitudes*). В SDSS параметр “мягкости” (англ. *softening parameter*, $b = \sigma_f$) рассчитывается сразу для всего обзора (см. табл. 21, Стоутон и др., 2002), мы же используем при расчете гиперболических величин индивидуальные ошибки на поток σ_f , приведенные для каждого объекта в фотометрическом каталоге. Отметим, что в пределе $|f/\sigma_f| \rightarrow 0$ гиперболическая величина стремится к погсоновской (Pogson) видимой звездной величине со значением потока $f = \sigma_f$. В пределе больших потоков $f \gg \sigma_f$ введенная нами гиперболическая величина стремится к стандартной погсоновской видимой звездной величине. Далее в тексте статьи, если специально не оговорено обратное, под звездной величиной оптического объекта мы везде будем понимать гиперболические величины, полученные на основе формулы (3).

Важной характеристикой для каждого рентгеновского источника, которая влияет на возможность наличия компаньона в выбранном оптическом обзоре, является предельная величина оптического обзора в заданном фильтре в направлении на рентгеновский объект. Поскольку DESI LIS представляет собой фотометрический каталог, составленный по данным нескольких обзоров с неоднородной картой экспозиции по небу, его предельная звездная величина в разных фильтрах и в разных направлениях на небе может заметно отличаться. Чтобы учесть эти факторы в модели SRGz, мы оценили максимальные величины объектов в обзоре DESI LIS в разных фильтрах в направлениях на рентгеновские источники следующим образом. В выбранном фильтре мы рассчитывали максимальную гиперболическую величину среди оптических объектов в $30''$ -окрестности рентгеновского источника:

$$\text{mag}_{O_{\text{lim}}} = \max_{d_i < 30''} \text{mag}_{O_i}, \quad (4)$$

где mag_{O_i} — гиперболическая звездная величина i -го оптического объекта в окрестности рентгеновского объекта, d_i — угловое расстояние между рентгеновским объектом и i -м оптическим источником. Мы будем использовать определенные выше максимальные звездные величины в трех

оптических ($g_{\text{lim}}, r_{\text{lim}}, z_{\text{lim}}$) фильтрах DESI LIS и двух ИК-фильтрах ($W_{1,\text{lim}}, W_{2,\text{lim}}$) WISE в качестве признаков модели оптического отождествления SRGz, что позволит при оптическом отождествлении рентгеновского источника учитывать глубину оптического обзора в заданном направлении.

Дополнительно для каждого источника еРОЗИТА в восточной галактической полусфере в области покрытия DESI LIS мы рассчитали вероятность, с которой рентгеновский источник с заданным потоком $F_{X,0.5-2}$ и для заданных значений максимальных звездных величин оптического обзора DESI LIS в голубом (g_{lim}) и красном ($z < z_{\text{lim}}$) фильтрах будет иметь оптический компаньон с $g < g_{\text{lim}}$ и $z < z_{\text{lim}}$:

$$P_c = P_c(F_{X,0.5-2}, g_{\text{lim}}, z_{\text{lim}}). \quad (5)$$

Для этого нами была составлена сетка значений по рентгеновскому потоку $F_{X,0.5-2}$ и максимальным величинам g_{lim} и z_{lim} . По этой сетке, опираясь на совокупные данные о точечных рентгеновских источниках в каталогах CXC2 и 4XMM, а также на данные об оптических объектах DESI LIS в $30''$ окрестности рентгеновских источников, мы оценивали вероятность оптического компаньона рентгеновского источника (5). Для объектов в интервале по рентгеновскому потоку $\lg(F_{X,0.5-2}) \pm 0.1$ и заданными максимальными величинами оптического обзора $g_{\text{lim}}, z_{\text{lim}}$, оценка P_c делалась по формуле

$$P_c = \frac{N_{R98} - N_{20-30} C_S}{0.98 N_X}, \quad (6)$$

где N_X — полное число точечных рентгеновских источников CXC2 и 4XMM в полях DESI LIS в выбранном интервале по рентгеновскому потоку; N_{R98} — число оптических объектов DESI LIS с величинами $g < g_{\text{lim}}$ и $z < z_{\text{lim}}$ на расстоянии $R < R_{98}$ от рентгеновских источников (R_{98} — радиус кружка, в котором с 98% вероятностью находится истинное положение рентгеновского объекта на небе); N_{20-30} — число оптических объектов DESI LIS с величинами $g < g_{\text{lim}}$ и $z < z_{\text{lim}}$ в кольце $R = 20-30$ угл. сек от рентгеновских источников; нормирующая константа C_S в формуле (6) отражает отношение площадей областей $R < R_{98}$ и внешних колец $20'' < R < 30''$. Таким образом, нами была построена табличная функция $P_c(F_{X,0.5-2}, g_{\text{lim}}, z_{\text{lim}})$, с помощью которой мы рассчитали вероятность компаньона для всех объектов из каталога точечных рентгеновских источников по данным двух лет обзора еРОЗИТА в восточной галактической полусфере в области покрытия DESI LIS. Мы будем использовать величину P_c как один из признаков модели оптического отождествления SRGz (см. табл. 1). Кроме того, полученные здесь оценки P_c для объектов тестовой

выборки позволяют моделировать в ней долю рентгеновских объектов без оптических компаньонов (“бездомных”), что позволяет более точно оценивать результаты модели оптического отождествления SRGz (см. п. 6.2.2).

Поверхностную плотность оптических источников в окрестности рентгеновских объектов мы будем рассчитывать по следующей формуле:

$$\Sigma_O = \frac{n_{30} - P_c \left[1 - e^{-30^2/(2\sigma_X^2)} \right]}{\pi 30^2}, \quad (7)$$

где величины n_{30} , P_c , σ_X были введены нами ранее и измеряются отдельно для каждого рентгеновского источника. Отметим, что второй член в числителе формулы (7) учитывает вклад от оптических компаньонов в круге радиуса $30''$ с центром на рентгеновском источнике.

Рассчитав рентгеновские, оптические и ИК-величины объектов, как описано выше, далее мы составили из них все возможные комбинации цветов. При этом цвета рассчитывались как внутри одного фотометрического обзора, так и между величинами в разных обзорах. Дополнительно для обзоров с доступными несколькими видами звездных величин, которые отражают как полную яркость объекта, так и яркость в предположении звездобразного профиля (например, CMODEL- и PSF-величины в SDSS, Kron- и PSF-величины в Pap-STARRS1), мы рассчитывали цвета между разным типом величин внутри одного обзора в том же фильтре. Данные цвета являются хорошими индикаторами протяженности оптического объекта и широко применялись для определения фотометрического типа объекта в обзоре SDSS (Стоутон и др., 2002).

Для того чтобы охарактеризовать отношение потоков объектов в рентгеновском диапазоне 0.5–2 кэВ ePOZITA и в оптических/ИК фильтрах DESI LIS и WISE, мы вводим ряд отношений “рентген/оптика” и “рентген/ИК” (по аналогии с цветами между оптическими фильтрами), следующим образом:

$$(X/O) = \text{mag}_O - \text{mag}_X, \quad (8)$$

где mag_O может измеряться в фильтрах g, r, z, W_1, W_2 (оптические и основные ИК-фильтры, доступные в DESI LIS). Отметим, что введенная нами формула (8) для отношения рентген/оптика отличается от принятой в (Маккаро и др., 1988) главным образом основанием логарифма. Введенные здесь формулы X/O дают нам возможность считать цвета между рентгеновским и оптическим/ИК диапазонами полностью аналогичным образом, как это принято делать в оптической астрономии (по разнице звездных величин).

Заметим, что в работе (Сальвато и др., 2018) была предложена эмпирическая формула для отношения X/W_1 , которая позволяет лучше разделять (чем классическое отношение X/W_1 принятое нами или в Маккаро и др., 1988) рентгеновские звезды и галактики, с одной стороны, и рентгеновские квазары, с другой, в диапазоне шести звездных величин W_1 . Мы добавили данный признак в число рассматриваемых нами отношений рентген/оптика:

$$(X/W_1)_{\text{salvato}} = W_1 - 1.625 \times \lg(F_{X,0.5-2}) + 6.101, \quad (9)$$

где переход величины W_1 из системы Веги к AB-системе (принятой в данной работе) выполнен стандартным образом⁴.

В каталоге DESI LIS (Дей и др., 2019) имеются специальные признаки, которые характеризуют модель радиального профиля яркости оптического объекта. В первую очередь категориальный признак фотометрического морфологического типа (*type*) и соответствующий ему признак параметра наклона радиального профиля яркости оптического объекта (*serisic*). Параметр *type* может принимать одно из шести значений:

- **PSF.** Звезды и квазары, компактные галактики.
- **EXP.** Спиральные галактики с экспоненциальным радиальным профилем яркости диска.
- **REX.** Круглые галактики с экспоненциальным профилем яркости.
- **DEV.** Эллиптические галактики с девокулеровским $I(r) \propto r^{1/4}$ радиальным профилем яркости.
- **SER.** Галактики с радиальным профилем яркости вида $I(r) \propto r^{1/n}$, где параметр модели профиля Серсика приводится в признаке *serisic* из фотометрического каталога.
- **DUP.** Яркие звезды GAIA, профили яркости и потоки которых оценивались в DESI LIS из данных GAIA.

Последней группой фотометрических признаков, которые мы введем ниже для лучшего разделения классов звезд Млечного Пути и внегалактических объектов, являются признаки, названные нами “цветовыми расстояниями”.

По мере проведения ИК-обзоров всего неба, а также значительного увеличения чувствительности

⁴<https://www.legacysurvey.org/dr9/description/>

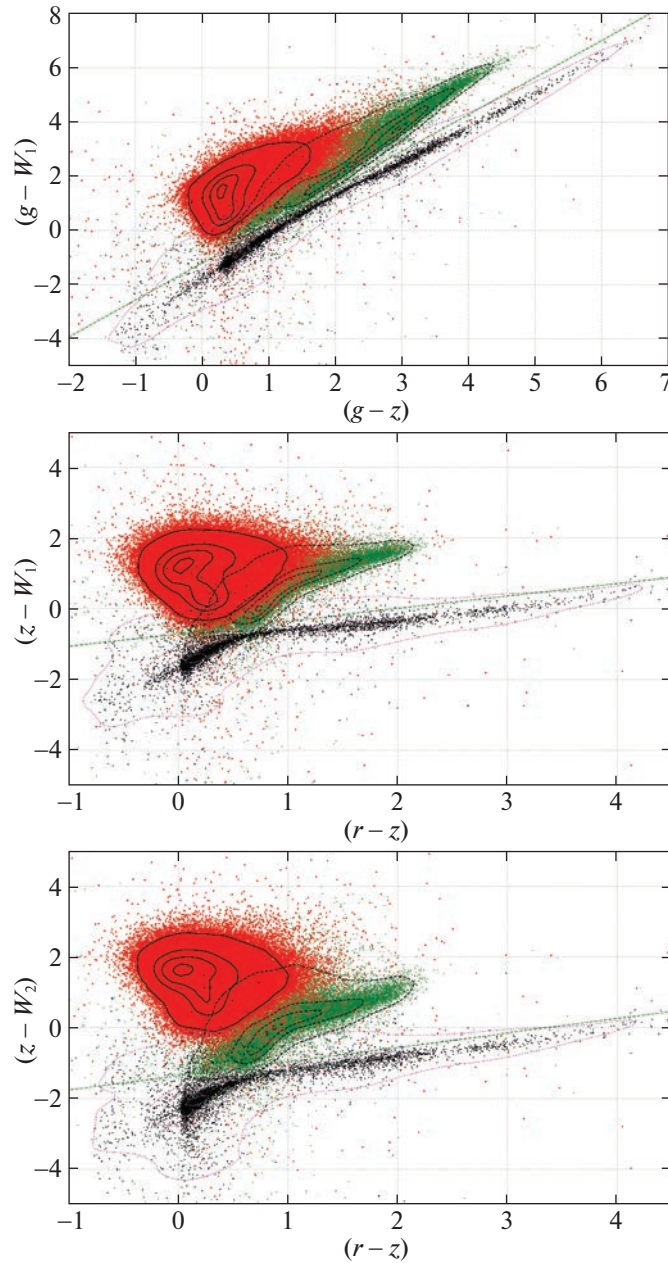


Рис. 2. Фотометрические диаграммы цвет-цвет для спектральных звезд (SDSS DR17 с астрометрическим подтверждением GAIA DR2), квазаров (SDSS DR17) и галактик (SDSS DR17) в $30''$ от точечных рентгеновских источников ePOZITA в тестовой области. Верхняя панель — цвета $(g - r)$, $(g - W_1)$; средняя панель — цвета $(r - z)$, $(z - W_1)$; нижняя панель — цвета $(r - z)$, $(z - W_2)$. Спектральные квазары, галактики и звезды показаны красными, зелеными и черными точками соответственно. Сплошные, штриховые, пунктирные контуры соответствуют уровням квантилей 0.05, 0.35, 0.65, 0.95 для распределений квазаров, галактик и звезд. Прямые линии разделяют области звезд и внегалактических объектов. В данной статье мы ввели цветовые расстояния, которые представляют собой разность цвета по оси ординат от выбранной точки до соответствующей прямой разделения классов.

фотометрических данных WISE за счет многократных сканирований небесной сферы по программе NEOWISE-R (Мейнзер и др., 2011), и доступности техники принудительной (англ. forced) фотометрии для большого числа оптических объектов на небе, стало возможным достаточно точно и массово измерять цвета астрономических объектов между

оптическим диапазоном (0.3–1 мкм) и средним ИК (3–5.5 мкм). При этом было отмечено (см., например, Буренин, 2022), что области звезд и внегалактических объектов хорошо разделяются на диаграмме (оптический цвет) — (оптический/ИК цвет) и, таким образом, разделяющая их линия может

использоваться для классификации звезд в данных больших фотометрических обзоров.

На рис. 2 мы построили три варианта диаграммы (оптический цвет) — (оптический/ИК цвет). На графиках показаны распределения спектроскопических объектов SDSS DR17 в тестовой области (при этом спектроскопические звезды были дополнительно сопоставлены со звездами GAIA, чтобы исключить из рассмотрения объекты с ошибочной спектральной классификацией), найденных в окрестности $30''$ от точечных рентгеновских источников двухлетнего обзора ePOZITA. На верхней панели — зависимость цвета $(g - W_1)$ от цвета $(g - r)$; средняя панель — зависимость $(r - z) - (z - W_1)$; нижняя панель — зависимость $(r - z) - (z - W_2)$. Спектральные квазары, галактики и звезды показаны красными, зелеными и черными точками соответственно. Как можно видеть, на всех трех панелях locus звезд хорошо отделяется от локуса квазаров (в первую очередь по оптическому/ИК цвету). Это неудивительно, так как в ИК-области для звезд ожидается рэлей-джинсовский наклон спектра черного тела фотосферы звезды, в то время как для квазаров и галактик излучение пыли может давать существенный вклад в ИК-поток. Как можно видеть из графиков, визуальный зазор между границами классов квазаров и звезд оказывается наибольшим при использовании цвета $(z - W_2)$. На всех трех графиках, мы построили прямые линии, приблизительно параллельные областям звезд и разделяющие классы звезд и внегалактических объектов (показаны сплошными черными линиями на графиках).

Введем цветовое расстояние объекта как разницу в оптическом/ИК цвете между положением объекта на соответствующей диаграмме и линией разделения классов галактических (звезд) и внегалактических (квазаров, галактик) оптических источников. Ниже мы приводим выражения, по которым будем вычислять признаки цветовых расстояний объектов:

$$(g - W_1)_{distance} = 1.37 \times (g - z) - 1.2, \quad (10)$$

$$(z - W_1)_{distance} = 0.35 \times (r - z) - 0.7, \quad (11)$$

$$(z - W_2)_{distance} = 0.40 \times (r - z) - 1.35. \quad (12)$$

3.3. Тестовые данные

Для анализа точности моделей SRGz для отождествления, классификации и измерения фотометрических красных смещений (photo-z) оптических компаньонов, мы использовали две тестовые области, одна из которых — большая область спектроскопического обзора SDSS площадью ≈ 4000 кв. градусов с наибольшей плотностью наблюдений

спектроскопических квазаров в каталоге SDSS DR16Q (см. подробнее ниже п. 3.3.1).

В выбранную тестовую область SDSS вошли два глубоких рентгеновских обзора средней площади: XMM-XXL-N (18 кв. градусов, Мензель и др., 2016) и Stripe82X (31.3 кв. градусов, ЛаМаса и др., 2016). Для этих областей ранее были опубликованы каталоги оптических компаньонов рентгеновских источников с данными о спектроскопических измерениях для значительной части объектов. В качестве второй тестовой области, для тестирования моделей SRGz, мы будем использовать область Stripe82X как обзор большей площади с большей глубиной оптической фотометрии (на основе данных SDSS Stripe82, предельная чувствительность которых в оптическом диапазоне превосходит DESI LIS). Также для данного поля в литературе доступны прогнозы photo-z, полученные разными группами (Ананна и др., 2017; Брешиа и др., 2019). Все это делает Stripe82X удобной тестовой выборкой для оценки моделей отождествления, классификации и photo-z рентгеновских объектов.

Далее мы рассмотрим выбранные тестовые области более подробно.

3.3.1. Тестовая область SDSS. Чтобы иметь возможность максимально эффективно использовать спектроскопическую информацию обзора SDSS для анализа природы рентгеновских источников ePOZITA, мы выбрали большую область на небе площадью несколько тысяч кв. градусов с самой высокой плотностью объектов из каталога спектроскопических квазаров SDSS DR16Q (Люк и др., 2020). Мы разбили небо на площадки равной площади ≈ 0.84 кв. градуса (HEALPIX, resolution = 6) и оценили в них плотность спектроскопических квазаров. Затем мы последовательно отбирали в тестовую выборку площадки с максимальной плотностью спектроскопических квазаров, пока не достигли площади тестовой области в ≈ 4179 кв. градусов (≈ 550 тыс. спектроскопических квазаров DR16Q). В рассматриваемой нами тестовой области были зарегистрированы ≈ 464 тыс. точечных рентгеновских источников (с лог-правдоподобием детектирования $\mathcal{L} > 6$) по данным двух лет обзора СРГ/ePOZITA. Отметим, что двухлетний рентгеновский обзор ePOZITA (eRASS:4) в этой области имеет предельную чувствительность для точечных источников $F_{X,0.5-2}^{\text{lim}} \sim 1.5 \times 10^{-14}$ эрг/с/см². Пространственная маска выбранной тестовой области на небе приводится нами на странице результатов SRGz⁵.

⁵<https://www.srg.cosmos.ru/srgz2023>

3.4. Тестовое поле *Stripe82X*

Выборка *Stripe82X* (Ананна и др., 2017) состоит из рентгеновских объектов *Chandra*, *XMM-Newton* в области *Stripe82*, их оптических компаньонов, результатов оптической спектроскопии. Эта выборка сочетает в себе значительную площадь покрытия рентгеновскими наблюдениями, глубину оптических данных *Stripe82* и достаточно высокую полноту спектроскопии оптических компаньонов рентгеновских объектов. Все это делает ее удобной тестовой выборкой для сравнения моделей отождествления, классификации и *photo-z* рентгеновских объектов. В п. 6.4 мы сравним точность моделей SRGz с результатами, опубликованными в литературе (Ананна и др., 2017; Брешиа и др., 2019) для данного поля.

4. МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

4.1. Выбор алгоритмов машинного обучения

При создании SRGz перед нами встал вопрос выбора базовых моделей машинного обучения для решения задач фотометрической классификации и вероятностной регрессии *photo-z* астрономических объектов. Данные фотометрических каталогов представляют собой характерный пример табличных гетерогенных данных с большим числом скоррелированных числовых признаков объектов (звездные величины и разнообразные цвета) и категориальных признаков (например, фотометрический тип объекта в каталоге DESI LIS). Для данных такого рода в индустрии нейросетевые модели классификации (например, многослойные перцептроны или специально разработанные архитектуры для табличных данных *TabNet*, см. Арик, Пфистер, 2019), как правило, уступают по точности алгоритмам машинного обучения из семейства древовидных ансамблей (например, градиентный бустинг). Пока ведется активный поиск новых нейросетевых архитектур для табличных данных (см. Борисов и др., 2021), ансамбли деревьев решений являются наиболее подходящим выбором алгоритма для построения моделей классификации по данным фотометрических каталогов астрономических объектов. В SRGz для отождествления оптических компаньонов рентгеновских источников и классификации звезда/квазар/галактика мы будем использовать одну из наиболее эффективных реализаций алгоритма градиентного бустинга — библиотеку *LightGBM* (Ке и др., 2017).

В задаче измерения *photo-z* внегалактических объектов алгоритмы машинного обучения при наличии достаточно большой и репрезентативной тренировочной выборки превосходят по точности методы на основе шаблонов (см., например, свежий

обзор Ньюман, Груен, 2022). При прогнозировании *photo-z* на табличных данных (фотометрические каталоги), древовидные ансамбли превосходят нейросетевые модели по точности оценок красных смещений как для галактик (см., например, Ли и др., 2023; Шмидт и др., 2020), так и для рентгеновских квазаров (Борисов и др., 2022; Мещеряков и др., 2018). Необходимо отметить, однако, что нейросетевые подходы демонстрируют более точную калибровку вероятностных прогнозов *photo-z* галактик (см., например, сравнение калибровок на рис. 2 в работе Шмидт и др., 2020). Видно, что на выборке галактик для обзора LSST модель случайного леса TPZ (Карраско Кинд, Бруннер, 2013) заметно уступает по калибровке моделям нейросетей ANNz2 (Садах и др., 2016) и METAPhoR (Кавуоти и др., 2017). С другой стороны, древовидные ансамбли градиентного бустинга (модель *FlexZBoost*, см. Избицки, Ли, 2017; рис. 12 из работы Шмидт и др., 2020) демонстрируют качество калибровки *photo-z*, аналогичное нейросетям. Авторы отмечают, что *FlexZBoost* на основе градиентного бустинга при обучении небольших тренировочных выборках уступает по точности прогнозов алгоритму случайного леса.

Описанные выше общие закономерности поведения различных моделей машинного обучения справедливы и для *photo-z* квазаров (являющихся доминирующим классом источников в рентгеновском обзоре ePOZITA). Исходя из сравнения методов *photo-z* для рентгеновских источников (Борисов и др., 2022; Мещеряков и др., 2018), мы выбрали модель квантильного случайного леса (англ. *quantile random forest*) как наиболее точную для основной массы объектов, так и применительно к далеким объектам на красных смещениях $z > 3-5$ (для которых доступны только небольшие тренировочные выборки). Можно упомянуть, что модели *photo-z* на основе квантильного случайного леса, используемые в SRGz, позволили открыть в данных обзора ePOZITA наиболее мощный рентгеновский квазар во Вселенной на красном смещении $z \approx 5.5$ (Хорунжев и др., 2021) и множество далеких рентгеновских квазаров на $z > 3-4$ (Хорунжев и др., 2020, 2022; Додин и др., 2020, 2021; Бикмаев и др., 2020, 2021).

Далее мы подробно рассмотрим применяемые нами древовидные ансамблевые алгоритмы машинного обучения.

4.1.1. Фотометрическая классификация методом градиентного бустинга. Алгоритм градиентного бустинга над деревьями решений (GBDT, англ. *Gradient Boosted Decision Trees*; Фридман, 2001) является одним из наиболее точных методов машинного обучения. Данный алгоритм поэтапно

строит аддитивную модель и является частным случаем метода градиентного бустинга (использующего произвольные базовые алгоритмы и позволяющего оптимизировать целевую функцию, использующую произвольные функции потерь). Каждая следующая модель в ансамбле GBDT строится для антиградиента выбранной функции потерь.

В настоящей работе мы используем программную библиотеку LightGBM (англ. Light Gradient Boosting Machine, см. Ке и др., 2017), являющуюся одной из самых популярных, точных и эффективных по вычислительным ресурсам реализаций метода градиентного бустинга. LightGBM обладает высокой точностью благодаря использованию деревьев, ограниченных по числу листьев, а не по глубине, как во многих других реализациях GBDT. Устойчивость алгоритма к переобучению достигается за счет использования нескольких основных параметров для регуляризации модели, а высокая производительность достигается за счет использования аппроксимации второго порядка (метод Ньютона) вместо градиента и рядом других алгоритмических оптимизаций (подробнее см. Ке и др., 2017).

В качестве функций потерь на этапе тренировки модели градиентного бустинга нами используются метрики бинарной и многоклассовой кросс-энтропии (см. Бишоп, 2007). Оптимальные гиперпараметры градиентного бустинга мы находим на этапе кросс-валидации с использованием метрики ROC AUC (для бинарной классификации) и многоклассовой кросс-энтропии (в случае многоклассовой классификации). Отметим, что в случае задачи многоклассовой классификации звезда/квазар/галактика оказывается полезным строить несколько отдельных моделей бинарной классификации (звезда/квазар+галактика, квазар/звезда+галактика, галактика/звезда+квазар, звезда/квазар, галактика/звезда, квазар/галактика) и затем комбинировать их результаты. Зачастую точность таких комбинированных моделей классификации превосходит точность одной модели трехклассовой классификации.

Необходимо отметить, для эффективного использования метода градиентного бустинга, ключевую роль играет контроль переобучения алгоритма. В первую очередь, это достигается выбором оптимального числа решающих деревьев в ансамбле (по значению метрики качества модели на этапе кросс-валидации). Мы подбирали число деревьев с помощью процедуры раннего останова (англ. early stopping) — рост ансамбля прерывался в случае прекращения улучшения метрики точности на кросс-валидационной выборке за последние несколько итераций. Далее, фундаментальным

параметром LightGBM, определяющим сложность модели и контролирующим переобучение, является максимальное число листьев в дереве (гиперпараметр num_leaves). Современные реализации GBRT, в том числе LightGBM, содержат множество дополнительных способов контроля переобучения, например, минимальное число примеров в листе, минимальный прирост целевой функции при разбиении узла, коэффициенты L1- и L2- регуляризации, параметр сжатия (англ. shrinkage) — большинство из них являются взаимозаменяемыми (так, на значительной части эталонных наборов данных современные реализации GBDT демонстрируют хорошую точность прогнозов со значениями гиперпараметров по умолчанию⁶). В настоящей работе для тонкой настройки алгоритма градиентного бустинга мы дополнительно оптимизировали минимальное число примеров в листе (гиперпараметр min_data), а также гиперпараметры рандомизации деревьев (feature_fraction и bagging_fraction) и использовали перевзвешивание примеров для учета дисбаланса классов (гиперпараметр scale_pos_weight).

4.1.2. Вероятностная регрессия photo-z методом квантильного случайного леса. Алгоритм случайного леса для решения задач регрессии и классификации был предложен в работе (Брейман, 2001). Впервые для задачи вероятностной регрессии был использован в работе Майнсхаузен (2006) для точечной оценки квантилей распределения.

Модель случайного леса, применяемая в данной работе, представляет собой ансамбль k деревьев решений $\{A^j\}$, которые строятся нами без ограничения глубины на случайной выборке N обучающих примеров с M фотометрическими признаками (случайный отбор объектов и признаков управляется гиперпараметрами рандомизации алгоритма). Построение модели производится следующим образом. Пусть дана обучающая выборка $\mathcal{D} = (\mathcal{D}_X, \mathcal{D}_z)$, где $\mathcal{D}_X = (x_{i,j})_{j=1}^M$ и $\mathcal{D}_z = (z_i)$ — фотометрические признаки и спектроскопическое красное смещение объектов обучающей выборки соответственно; $i = \overline{1, N}$. Каждое дерево A^j строит свое разбиение пространства признаков X на непересекающиеся подпространства A_i^j , которые в сумме дают все пространство, т.е.

$$\sum_i A_i^j = X, \quad A_n^j \cap A_m^j = \emptyset, \quad m \neq n. \quad (13)$$

Для построения каждого дерева:

- методом бутстреп создается случайная подвыборка $\mathcal{D}^j$ обучающей выборки;

⁶https://github.com/catboost/benchmarks/tree/master/quality_benchmarks

- при каждом разбиении в j -м дереве всех признаков случайно выбирается подвыборка m признаков ($m \leq M$);
- на каждом шаге ветвления дерева по критерию MSE определяются оптимальный признак и пороговое значение, по которому делается разбиение пространства;
- мы строим разбиение пространства, пока не исчерпается вся выборка $\mathcal{D}^j$ (деревья без ограничения глубины), так что в листьях оказывается по одному объекту обучающей выборки со спектральным красным смещением.

В работе Танг и др. (2018) проводится анализ поведения случайного леса при различных гиперпараметрах рандомизации. Глубокие деревья решений дают практически несмещенные прогнозы ($Bias \sim 0$), но оценки отдельных деревьев имеют высокую дисперсию и плохо обобщаются на новые выборки данных. Дисперсия прогноза одного дерева существенно понижается в случайном лесу путем усреднения по ансамблю независимых деревьев.

Мы используем реализацию алгоритма случайного леса в программной библиотеке `scikit-learn` (версия 1.1.3; Педрегоса и др., 2011). Деревья регрессии строятся без ограничения глубины с гиперпараметрами по умолчанию (см. `sklearn.ensemble.RandomForestRegressor`). Пространство признаков каждого дерева разбивается до того момента, пока во всех листьях не останется по одному объекту тренировочной выборки (на базе этих спектральных объектов в листьях дерева и будут формироваться прогнозы красного смещения). Как было отмечено в работе (Ли, Джеон, 2006), случайный лес с k деревьями без ограничения глубины представляет собой “адаптированный” алгоритм ближайших соседей (kNN). При таком рассмотрении, “близость” объектов в многомерном пространстве признаков определяется не как в kNN (через метрику расстояния, которая подвержена проклятию размерности и плохо работает в случае сильно скоррелированных фотометрических признаков). Вместо метрики расстояния степень близости объектов в случае случайного леса определяется разбиением пространства признаков алгоритмом дерева решений (на области с минимальной дисперсией тренировочных объектов по спектральному красному смещению). Так как на каждом шаге дерево решений (жадно) выбирает какой-то один признак для разбиения пространства, то степень близости объектов менее подвержена корреляции между признаками в тренировочной выборке. В итоге ансамбль прогнозов $(z_j)_{j=1}^k$ случайного

леса (построенного без ограничения глубины) представляет собой бутстреп-выборку ближайших спектральных соседей (со сложноопределенной функцией близости) для каждого целевого объекта в фотометрическом каталоге.

Как показано на рис. 3, параметрическую оценку $\hat{p}(z|X)$ условного распределения прогноза мы восстанавливаем из прогнозов случайного леса с помощью гауссовой ядерной оценки плотности (с подбором параметра полосы пропускания, Баштанник, Гайндман, 2001). Независимость и случайность построения деревьев позволяют рассматривать ансамбль предсказаний деревьев $(z_j)_{j=1}^k$ как случайную выборку, порожденную истинным условным распределением $p(z|X)$, а $\hat{p}(z|X)$ будет представлять собой параметрическую оценку данного распределения.

Отметим, что данная выше интерпретация прогноза $\hat{p}(z|X)$ опирается на предположение о независимости деревьев из случайного леса, что никогда полностью не достигается (так как деревья ансамбля используют перекрывающиеся выборки тренировочных объектов и признаков), поэтому измерение $\hat{p}(z|X)$, получаемое данным методом, не является в общем случае несмещенной оценкой условного распределения прогноза $p(z|X)$ (как было отмечено в обзоре Ньюман, Груен, 2020). Это также означает, что калибровка прогноза $\hat{p}(z|X)$, полученного случайным лесом, может быть хуже, чем у других подходов машинного обучения, в частности, искусственных нейронных сетей, что и наблюдается на практике (см., например, Шмидт и др., 2020). Следует отметить, что сильной стороной измерений $photo-z$ внегалактических источников при помощи алгоритма случайного леса остается высокая точность его точечных прогнозов $photo-z$ (что было многократно показано на разных выборках, см., например, Шмидт и др., 2020; Мещеряков и др., 2018).

5. МОДЕЛИ SRGz

Базовая версия системы SRGz (версия 2.1), описанная в данной работе, использует для отождествления (модель 3_X) и классификации (модель $3_{X,col}$) оптических компаньонов рентгеновских источников только данные обзоров ePOZITA, DESI LIS и WISE (принудительная фотометрия приведена в каталоге DESI LIS). Выбор только двух каталогов (ePOZITA, DESI LIS) обусловлен тем, что DESI LIS является наиболее глубоким из рассматриваемых нами оптических обзоров, в то время как данные SDSS и/или Pan-STARRS1 доступны не для всех оптических объектов DESI LIS. Мы хотим, чтобы отождествление оптических

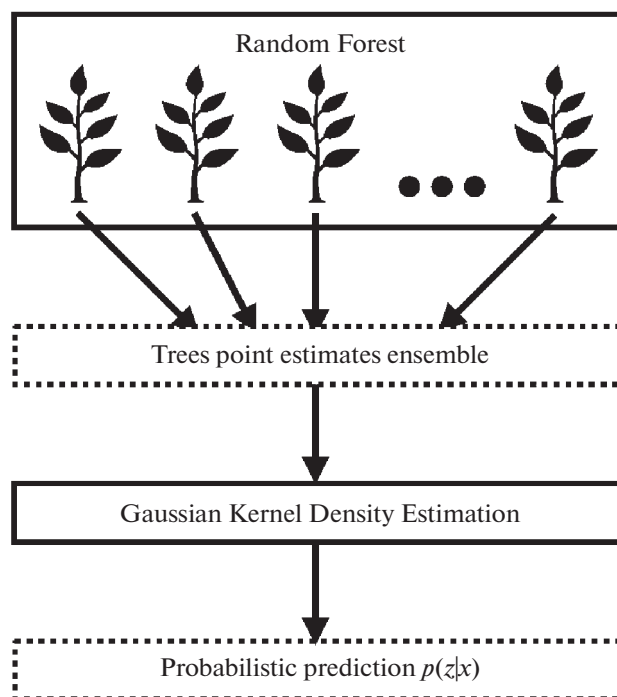


Рис. 3. Схема получения прогнозов photo-z методом случайного леса.

компаньонов рентгеновских объектов и их классификация были сделаны единообразно во всей рассматриваемой области.

В отношении моделей photo-z оптических компаньонов в SRGz мы придерживались следующего подхода. Во-первых, здесь мы не использовали рентгеновские признаки ePOZITA и связанные с ними цвета (также не использовались *offset*, *type*, *serisic* и введенные нами ранее цветовые расстояния). Рентгеновские данные исключались из рассмотрения для того, чтобы увеличить тренировочную выборку за счет оптических квазаров SDSS ($\approx 2/3$ которых не имеют рентгеновских измерений ePOZITA). Во-вторых, в моделях photo-z нами использовался широкий набор звездных величин и цветов оптических объектов (см. табл. 1 и описание признаков в п. 3.2).

На выходе конвейера SRGz мы получили каталог SRGz-CatD вероятных оптических компаньонов точечных рентгеновских источников ePOZITA с фотометрической классификацией и измерениями photo-z. В табл. 2 представлено описание базовых колонок каталога SRGz-CatD (см. также пояснения к содержанию каталога в п. 5.1.5, п. 5.2.5 и п. 5.3.3).

5.1. SRGz: модель отождествления оптических компаньонов рентгеновских источников

Задача отождествления оптических компаньонов рентгеновских источников ePOZITA реша-

ется нами в два этапа. Сначала мы обучим модель классификации пар рентгеновский источник ePOZITA — оптический источник DESI LIS. Она позволит для любого i -го оптического объекта в каталоге SRGz-CatA в окрестности заданного рентгеновского источника спрогнозировать вероятности ассоциации p_i . Далее мы преобразуем p_i в вероятности оптического отождествления следующим образом.

Пусть в окрестности заданного рентгеновского источника в каталоге SRGz-CatA содержатся n оптических объектов. Предполагая, что у рентгеновского источника имеется не более одного оптического компаньона и оптические объекты независимы, мы можем составить полную систему $(n + 1)$ гипотез оптического отождествления для заданного рентгеновского объекта с вероятностями: P_0, P_{i1}^n . Вероятность P_i (гипотеза о том, что i -й оптический объект является компаньоном, а все остальные оптические источники — объектами поля) будет равна

$$P_i = \frac{p_i \prod_{j \neq i} (1 - p_j)}{\sum_{k=1}^n \left[p_k \prod_{j \neq k} (1 - p_j) \right] + \prod_{j=1}^n (1 - p_j)}, \quad (14)$$

где второй член в знаменателе формулы определяет вероятность того, что ни один из оптических объектов не является компаньоном. После упрощения

Таблица 2. Описание базовых полей каталога SRGz-CatD

Поле в каталоге	в статье	Описание поля
ero_srcname		Название рентгеновского источника
ero_RA		Прямое восхождение рентгеновского объекта
ero_DEC		Склонение рентгеновского объекта
ls_ra		Прямое восхождение оптического объекта
ls_dec		Склонение оптического объекта
srg_match_flag		Флаг отождествления оптического компаньона
srg_match_p	p_i	Вероятность отождествления пары рентгеновский—оптический объект
srg_match_pi	P_i	Вероятность отождествления с учетом всех оптических объектов в $30''$
srg_match_p0	P_0	Вероятность “бездомного” рентгеновского объекта
srg_match_pstar		Вероятность звезды
srg_match_pqso		Вероятность квазара
srg_match_pgal		Вероятность галактики
srg_match_SQG.		Метка фотометрического класса
srg_match_warning		Предупреждения о проблемах в модели оптического отождествления
srgz_z_max	z_{ph}	Фотометрическое красное смещение
srgz_z_maxConf	$zConf$	Надежность фотометрической оценки красного смещения
srgz_z_merr68		68% ошибка снизу на z_{ph}
srgz_z_perr68		68% ошибка сверху на z_{ph}
srgz_z_model		Название модели photo-z
srgz_z_warning		Предупреждения о проблемах в модели photo-z

формула выше примет вид

$$P_i = \frac{p_i}{(1 - p_i)} P_0, \quad (15)$$

где вероятность P_0 (гипотеза, что рентгеновский источник не имеет оптического компаньона) выражается формулой:

$$P_0 = \frac{1}{\sum_{k=1}^n [p_k / (1 - p_k)] + 1}. \quad (16)$$

Далее мы рассмотрим две модели классификации пар рентгеновский—оптический объект. Первая модель — простая модель ближайшего оптического соседа, которую мы будем использовать в качестве базовой для сравнения. Вторая модель — классификация методом градиентного бустинга с использованием доступных астрометриче-

ских и фотометрических признаков рентгеновского и оптического объектов.

5.1.1. Базовая модель отождествления рентгеновского источника с ближайшим оптическим соседом. Наиболее простой моделью оптического отождествления является ассоциация рентгеновского источника с ближайшим оптическим объектом. Ниже мы приведем формулы вероятности ассоциации p_i для данной модели, которая в дальнейшем будет использоваться как базовая для сравнения с SRGz.

Мы будем считать, что в окрестности рентгеновских источников расположены оптические объекты поля согласно заданной поверхностной плотности Σ_O (которая определяется по формуле (7) индивидуально для каждого рентгеновского объекта). Рентгеновский источник характеризуется гауссовой ошибкой локализации σ_X и вероятностью P_c

наличия оптического компаньона в DESI LIS (см. формулы (5)–(6)). Тогда вероятность ассоциации i -й пары рентгеновский–оптический объект можно записать в следующем виде (см. также формулу (11) в работе Бельведерский и др., 2022):

$$p_i = \frac{P_c e^{-d_i^2/(2\sigma_X^2)}}{P_c e^{-d_i^2/(2\sigma_X^2)} + 2\pi \Sigma_O \sigma_X^2}, \quad (17)$$

где d_i — угловое расстояние от рентгеновского источника до i -го оптического объекта. Как можно заметить из формулы (17), в данной модели ближайший оптический объект всегда имеет наибольшую вероятность ассоциации с рентгеновским источником.

Определив p_i в модели ближайшего оптического соседа, как это описано выше, вероятности оптического отождествления P_i и $P_\emptyset$ могут быть получены с помощью формул (15)–(16).

5.1.2. Классификация пар рентгеновский–оптический объект с помощью машинного обучения (*Model_{match}*). Описанная в предыдущей части базовая модель классификации пар рентгеновский–оптический источник использует информацию об ошибке локализации рентгеновского источника, вероятности наличия у рентгеновского объекта оптического компаньона в DESI LIS и плотности оптических объектов в окрестности рентгеновского источника. Между тем набор фотометрических признаков оптических объектов в полях рентгеновских источников значительно шире (см. п. 3.2), и многие из этих признаков оказываются важными при поиске оптических партнеров рентгеновских источников (см. далее п. 6.1).

Преобразовать доступную информацию в более точные оценки вероятностей оптического отождествления (p_i , P_i , $P_\emptyset$) можно разными способами. В рентгеновской астрофизике получил широкое распространение метод поиска оптических компаньонов, основанный на объединении по теореме Байеса астрометрической и фотометрической информации об объектах (NWAY, Сальвато и др., 2018). Основными недостатками данного метода являются ограниченное использование в нем фотометрической информации и необходимость расчета фотометрических априорных вероятностей для каждой конкретной задачи. Так, в работе Сальвато и др. (2018) использовались фотометрические априорные вероятности, рассчитанные по двум фотометрическим признакам. В работах (Быков и др., 2023) и (Тэаро, Мещеряков, 2023) сделаны успешные попытки использования алгоритмов машинного обучения для расчета фотометрических априорных вероятностей в методе NWAY с применением к задачам оптического отождествления рентгеновских источников SRG/ePOZITA и XMM-

Newton соответственно. В работе (Тэаро, Мещеряков, 2023) также показано, что применение машинного обучения напрямую на всем пространстве астрометрических и фотометрических признаков дает лучшую точность оптического отождествления, чем объединение астрометрической и фотометрической информации по теореме Байеса (в NWAY).

Эвристические алгоритмы машинного обучения позволяют наиболее эффективно извлекать информацию об ассоциации объектов в выбранной паре рентгеновский–оптический источник из многомерного пространства фотометрических признаков. В рамках SRGz мы используем алгоритм градиентного бустинга (см. п. 4.1.1) для расчета вероятностей p_i по всем рассматриваемым признакам, как астрометрическим, так и фотометрическим (см. п. 3.2 и табл. 1): *offset*, n_{20} , n_{30} , P_c ; предельные величины g_{lim} , r_{lim} , z_{lim} , $W_{1,\text{lim}}$, $W_{2,\text{lim}}$; рентгеновские, оптические и ИК-величины, их производные цвета, а также цветовые расстояния.

В следующем разделе мы рассмотрим тренировочную выборку для обучения алгоритма классификации пар рентгеновский источник ePOZITA — оптический объект DESI LIS.

5.1.3. Выборка оптических компаньонов рентгеновских источников Chandra, XMM-Newton, Swift в полях обзора ePOZITA и обучение модели. Для обучения и оценки качества моделей отождествления мы создали тренировочную выборку с помощью данных наиболее крупных публичных рентгеновских обзоров, выполненных рентгеновскими космическими обсерваториями Chandra (Chandra Source Catalog 2.0, далее CXC2), XMM-Newton (XMM-Newton Serendipitous Source Catalogue Data Release 12, далее 4XMM) и Swift им. Нила Гепелза (Second Swift-XRT Point Source Catalog, далее 2SXPS). Идея использования вспомогательных рентгеновских каталогов для определения оптических компаньонов объектов ePOZITA состоит в следующем. Для объектов из каталогов CXC2, 4XMM, 2SXPS (удовлетворяющих набору стандартных критериев отбора точечных рентгеновских источников, изложенных в разделе 2.2) в области покрытия обзора DESI LIS в восточной галактической полусфере для каждого рентгеновского источника мы определили радиус оптического отождествления r_{false} по следующей формуле:

$$r_{\text{false}} = \left(-\frac{\log(1 - \alpha) \times 30^2}{n_{30}} \right)^{1/2} \times \left[1 - \frac{\beta}{2\sqrt{n_{30}}} \right], \quad (18)$$

где n_{30} — число оптических объектов в $30''$ -окрестности рентгеновского источника (из обзоров

Таблица 3. Выборка оптических компаньонов точечных рентгеновских источников ePOZITA, найденных с помощью данных рентгеновских каталогов Chandra, XMM-Newton и Swift

Оптические компаньоны	CXC2	4XMM	2SXPS	Всего	Пересечения	% ошибок
eRASS4	5833	11322	2325	17880	1550	0.84
CXC2	5833	1188	124		1262	0.55
4XMM	1188	11322	338		1476	0.88
2SXPS	124	338	2325		412	1.46
Тренировочная выборка	2820	5898	1303	9200	793	1.01
Тестовая выборка SDSS	3013	5424	1022	8680	757	0.66

CXC2/4XMM/2SXPS). Параметры α и β определяют чистоту получившейся выборки оптических компаньонов рентгеновских источников. Идея формулы (18) состоит в том, чтобы для каждого рентгеновского объекта из вспомогательных рентгеновских обзоров консервативно определить радиус, в котором вероятность встретить оптический источник поля $P < \alpha$. В предположении постоянной плотности оптических объектов поля в окрестности рентгеновского источника выражение для пуассоновской вероятности найти один или более оптических объектов поля в кружке радиуса r_{false} имеет вид $\alpha = 1 - \exp(-n_{30}(r_{\text{false}}/30)^2)$. Множитель в квадратных скобках в формуле (18) в первом приближении, учитывает ошибку измерения радиуса кружка (в условиях конечного числа оптических объектов в $30''$ -окрестности рентгеновского источника):

$$\frac{\Delta r_{\text{false}}}{r_{\text{false}}} \approx \frac{\Delta n_{30}}{2n_{30}}.$$

Таким образом, параметр β управляет уменьшением радиуса в масштабе неопределенности: $r_{\text{false}} \rightarrow r_{\text{false}} - \beta \Delta r_{\text{false}}$. В данной работе мы используем $\alpha = 0.03$ и $\beta = 2$ для построения чистой выборки оптических компаньонов рентгеновских источников. Таким образом, были отобраны 48 269, 55 356 и 7088 объектов DISI LIS, связанных с рентгеновскими объектами CXC2, 4XMM и 2SXPS соответственно.

Далее полученные выше списки оптических компаньонов рентгеновских источников были соотнесены с каталогом точечных рентгеновских источников по данным двух лет обзора SRG/ePOZITA в восточной галактической полусфере, за исключением областей Stripe 82X (Ананна и др., 2017), XMM-XXL-N (Мензель и др., 2016) и Lockman Hole (Бельведерский и др., 2022). Оставляя в рассмотрении только оптические компаньоны с единственным объектом ePOZITA

в $30''$ -окрестности и на расстоянии $d < 1.7R_{98}$ (где R_{98} — радиус кружка, в котором с вероятностью 98% находится истинное положение рентгеновского источника), мы получили финальную выборку оптических компаньонов источников ePOZITA для обучения и тестирования моделей SRGz.

В табл. 3 представлена подробная информация о выборке оптических компаньонов точечных рентгеновских источников ePOZITA. В общей сложности выборка состоит из 17 880 оптических объектов DESI LIS в полях рентгеновских источников ePOZITA, из которых 5833, 11 322 и 2325 оптических объектов связаны с рентгеновскими источниками CXC2, 4XMM и 2SXPS соответственно. Между объектами обзоров CXC2, 4XMM и 2SXPS имеется в общей сложности 1550 пересечений по рентгеновским источникам ePOZITA, что позволяет оценить чистоту полученной выборки оптических компаньонов (эти данные также приведены в табл. 3). Так, для 5833 оптических компаньонов рентгеновских источников CXC2 в $30''$ -окрестности соответствующих объектов ePOZITA находятся 1188 и 124 оптических компаньонов рентгеновских объектов 4XMM и 2SXPS соответственно; всего 1262 пересечения ($1262 < 1188 + 124$ с учетом тройных пересечений). В идеально чистом каталоге компаньонов все вспомогательные рентгеновские каталоги в полях пересечений должны указывать на одни и те же оптические объекты (% ошибок = 0). В построенном нами каталоге доля ошибок составляет 0.6, 0.9, 1.5% для оптических компаньонов из CXC2, 4XMM и 2SXPS соответственно. На оптических компаньонах из всех трех вспомогательных рентгеновских каталогов доля ошибок составляет 0.8% (1.0 и 0.7% на тренировочной и тестовой выборке соответственно). Таким образом, мы можем сделать вывод, что полученная нами выборка обладает очень высокой (>99%) точностью отбора

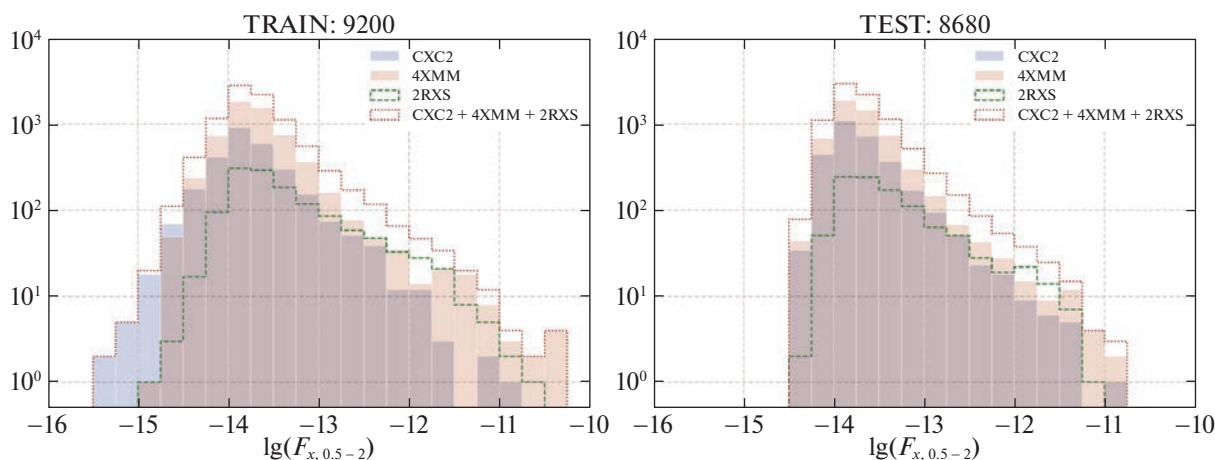


Рис. 4. Распределение по рентгеновскому потоку источников еРОЗИТА из тренировочной (левая панель) и тестовой (правая панель) выборок с оптическими компаньонами найденных из пересечения с рентгеновскими обзорами CXC2, 4XMM, 2RXS. Показаны распределения для всех источников вместе и для выборок объектов из каждого вспомогательного рентгеновского обзора по отдельности.

оптических компаньонов точечных рентгеновских источников еРОЗИТА.

На рис. 4 показано распределение по рентгеновскому потоку источников еРОЗИТА из тренировочной (левая панель) и тестовой (правая панель) выборок с оптическими компаньонами найденных из пересечения с рентгеновскими объектами обзоров CXC2, 4XMM, 2RXS. Показаны распределения для всех источников вместе и для выборок объектов из каждого вспомогательного рентгеновского обзора, по отдельности.

Заметим, что выборки объектов поля (тренировочная и тестовая) формируются из оптических объектов в $30''$ -окрестности рентгеновских источников еРОЗИТА после исключения из нее найденных оптических компаньонов. Каждый источник еРОЗИТА из описанных выше тренировочной и тестовой выборок имеет измерение вероятности оптического компаньона P_c (по формулам (5)–(6)). Мы проверили у оптических объектов наличие всех фотометрических признаков, необходимых для моделей (см. табл. 1), отфильтровали объекты с пропущенными значениями оптических признаков (менее 1% оптических компаньонов рентгеновских источников) и получили финальный список объектов для тренировочной и тестовой выборки (см. табл. 4).

Таким образом, наконец, нами были получены чистые выборки оптических компаньонов рентгеновских источников и оптических объектов поля в окрестности рентгеновских источников еРОЗИТА, которые мы использовали для обучения базовой модели ($2x$) отождествления SRGz.

Модель классификации по ближайшему оптическому соседу рентгеновского источника (см.

п. 5.1.1) не требует обучения, мы оцениваем ее точность на тестовой выборке. Модель градиентного бустинга (см. п. 4.1.1, п. 5.1.2) обучалась в задаче бинарной классификации (компаньоны рентгеновских источников еРОЗИТА / объекты поля) на тренировочной выборке, описанной выше. Необходимо отметить большой дисбаланс классов в тренировочной выборке, который учитывался в алгоритме LightGBM путем взвешивания объектов при помощи гиперпараметра `is_unbalance=True`. Мы также делали различные эксперименты с уменьшением размера тренировочной выборки объектов поля (и дисбаланса классов соответственно), но лучшие результаты в итоге показала модель с полным числом объектов.

Далее мы рассмотрим метрики оценки качества моделей отождествления оптических компаньонов рентгеновских источников.

5.1.4. Метрики оценки качества оптического отождествления рентгеновских источников.

Мы будем использовать несколько величин для оценки качества отождествления оптических компаньонов рентгеновских источников (в зависимости от порога по P_0):

- **Полнота и точность отождествления оптических компаньонов рентгеновских источников еРОЗИТА ($Recall_c$, $Precision_c$).** Полноту и точность отождествления оптических компаньонов рентгеновских источников на n -м значении порога t_n по P_0 можно вычислить по формулам:

$$Recall_c^n = \frac{\hat{N}_c^*(P_0 \leq t_n)}{N_c}, \quad (19)$$

Таблица 4. Тренировочные и тестовые выборки для базовых моделей SRGz

#	Тренировочная область	Тестовая область SDSS
$Model_{match}: 3_X$	Модель отождествления оптического компаньона	
оптические компаньоны	9135	8635
объекты поля	160 959	132 057
ближайшие объекты поля	9135	8635
$Model_{1,class}, Model_{2,class}: 3_{X,col}$	Модели классификации (звезда/квазар/галактика)	
QSO	39 132	163 218
GALAXY	10 957	17 371
STAR	78 958	33 660
$Model_{photoz}: 2, 3_{PS}, 3_{SDSS}, 4$	Модели измерения красных смещений (photo-z)	
	586 035	

$$Precision_c^n = \frac{\hat{N}_c^*(P_\emptyset \leq t_n)}{\hat{N}_c(P_\emptyset \leq t_n)}, \quad (20)$$

где N_c — полное число оптических компаньонов в тестовой выборке, а величины $\hat{N}_c^*$ и $\hat{N}_c$ представляют собой соответственно число правильно классифицированных и число всех оптических компаньонов для выборки объектов с $P_\emptyset \leq t_n$.

- **Полнота и точность отождествления рентгеновских источников ePOZITA, не имеющих оптического компаньона в обзоре DESI LIS** ($Recall_h$, $Precision_h$). Полноту и точность классификации рентгеновских источников как “бездомных” на n -м значении порога t_n по $P_\emptyset$ можно вычислить по формулам:

$$Recall_h^n = \frac{\hat{N}_h^*(P_\emptyset \geq t_n)}{N_h}, \quad (21)$$

$$Precision_h^n = \frac{\hat{N}_h^*(P_\emptyset \geq t_n)}{\hat{N}_h(P_\emptyset \geq t_n)}, \quad (22)$$

где величины $\hat{N}_h^*$ и $\hat{N}_h$ представляют собой соответственно число правильно классифицированных “бездомных” и число всех отмеченных моделью как “бездомные” рентгеновских источников для выборки объектов с порогом отождествления t_n ; N_h — полное число “бездомных” рентгеновских источников в тестовой выборке.

- **Точность оптического отождествления рентгеновских источников ePOZITA** ($Precision_X$).

$$Precision_X^n(t) = \frac{\hat{N}_c^*(P_\emptyset \leq t_n) + \hat{N}_h^*(P_\emptyset > t_n)}{N_c + N_h}. \quad (23)$$

Представленные выше метрики полноты и точности зависят от выбранного порога по величине $P_\emptyset$ и, таким образом, позволяют рассмотреть качество модели оптического отождествления рентгеновских источников в различных режимах и для различных задач.

Часто представляется интересным иметь интегральные показатели качества модели отождествления, выраженные одним или несколькими числами. В качестве таких показателей мы будем использовать:

- **Площадь под рабочей характеристикой приемника (ROC AUC).** Построение и анализ ROC-кривой (англ. Receiver Operating Characteristic) является стандартным способом оценки качества модели бинарной классификации. Площадь под ROC-кривой (ROC AUC, Ханлей, Мак-Нейл, 1982) является количественной мерой качества ранжирования и позволяет сравнивать между собой модели классификации в целом, вне зависимости от конкретного порога по вероятности. Оценка ROC AUC нечувствительна к дисбалансу классов в выборке. В настоящей работе мы будем использовать метрику ROC AUC для оценки

качества ранжирования пар рентгеновский—оптический объект. Положительный класс составят оптические компаньоны рентгеновских источников, отрицательный класс — ближайшие к рентгеновским источникам еРОЗИТА оптические объекты поля.

- **Максимальная полнота отождествления оптических компаньонов рентгеновских источников еРОЗИТА** ($Recall_{c,max}$). При условии $P_0 \leq 1$ каждому рентгеновскому источнику в тестовой выборке модель оптического отождествления ставит в соответствие определенный оптический объект. Этому случаю соответствует максимально возможная в модели полнота отождествления оптических компаньонов:

$$Recall_{c,max} = Recall_c(P_0 \leq 1). \quad (24)$$

Величина $Recall_{c,max}$ может быть меньше 1 в случаях, когда некоторые оптические компаньоны рентгеновских источников не идентифицируются моделью как наиболее вероятные среди оптических объектов поля (см. обсуждение работы модели отождествления в п. 5.1.3).

- **Максимальная точность оптического отождествления рентгеновских источников еРОЗИТА** ($Precision_{X,max}$).

$$Precision_{X,max} = \max_n Precision_X^n \quad (25)$$

есть величина характеризует решение задачи отождествления рентгеновских источников в целом (как имеющих, так и не имеющих оптические компаньоны). Точность $Precision_{X,max}$ достигается на некотором значении порога $P_{0,best}$.

- **Средняя точность отождествления рентгеновских источников еРОЗИТА с компаньонами** (AP_c) **и без оптических компаньонов** (AP_h). Средняя точность отождествления оптических компаньонов рентгеновских источников вычисляется по формуле

$$AP_c = \frac{\sum_n (Recall_c^n - Recall_c^{n-1}) Precision_c^n}{\sum_n (Recall_c^n - Recall_c^{n-1})}, \quad (26)$$

где $Recall_c^n$ и $Precision_c^n$ представляют собой полноту и точность отождествления оптических компаньонов рентгеновских источников на n -м значении порога P_0 .

Средняя точность отождествления рентгеновских источников без оптического компаньона в DESI LIS вычисляется по формуле:

$$AP_h = \frac{\sum_n (Recall_h^n - Recall_h^{n-1}) Precision_h^n}{\sum_n (Recall_h^n - Recall_h^{n-1})}, \quad (27)$$

где $Recall_h^n$ и $Precision_h^n$ представляют собой полноту и точность отождествления рентгеновских источников, не имеющих оптического компаньона (для n -го значения порога по P_0).

5.1.5. Данные об оптическом отождествлении рентгеновских источников в каталоге SRGz. В конвейере SRGz отождествление оптических компаньонов рентгеновских источников еРОЗИТА делается в два этапа. Сначала с помощью обученной модели классификации вычисляются вероятности ассоциации p_i для всех возможных пар рентгеновский источник еРОЗИТА — оптический объект DESI LIS в окрестности 30'' (собранных в каталоге SRGz-CatA). Затем по формулам (15)–(16) для каждого рентгеновского объекта мы рассчитываем вероятность отсутствия оптического компаньона P_0 и вероятности возможных оптических компаньонов P_i из SRGz-CatA.

В каталоге SRGz-CatD для каждого точноного рентгеновского источника еРОЗИТА мы приводим величину P_0 и данные по вероятным оптическим компаньонам рентгеновских источников еРОЗИТА. Приводятся данные по наиболее вероятному оптическому компаньону с максимальным p_i (такой объект в каталоге имеет значение 1 в столбце `src_match_flag`, см. табл. 2). Кроме того, в каталог SRGz-CatD также попадают другие вероятные оптические компаньоны — оптические объекты, у которых величина p_i менее чем в 2 раза отличается от вероятности наиболее вероятного компаньона: $p_i / \max_j p_j > 0.5$ (такие объекты в каталоге оптических отождествлений имеют значение 2 в столбце `src_match_flag`).

5.2. SRGz: модели классификации оптических компаньонов

В отношении моделей SRGz для отождествления и классификации оптических компаньонов рентгеновских источников следует обратить внимание на различие в наборе использовавшихся фотометрических признаков для этих моделей (см. табл. 1 и определение фотометрических признаков в п. 3.2).

В модели фотометрической классификации $Z_{X,col}$ не используются признаки *offset*, n_{20} , n_{30} , P_c , а также максимальные величины в полях

рентгеновских источников ($g_{\text{lim}}, r_{\text{lim}}, z_{\text{lim}}, W_{1,\text{lim}}, W_{2,\text{lim}}$), так как они не являются информативными при определении класса объекта.

Далее, в модели фотометрической классификации не используются непосредственно звездные величины объектов, а только их производные цвета, цветовые расстояния и признаки фотометрического типа (*type, sersic*). Это сделано для того, чтобы модели фотометрической классификации SRGz были применимы также для слабых оптических объектов. Тренировочные выборки для модели классификации состоят из объектов спектрального каталога SDSS и звезд GAIA, которые существенно ограничены по звездной величине (спектроскопия SDSS и астрометрия GAIA доступны только для достаточно ярких объектов). Кроме того, функция отбора по величине для разных классов объектов SDSS (например, звезд, квазаров) заметно отличается. Таким образом, мы не рассматривали в модели классификации признаки, связанные с абсолютными потоками источников.

Мы построили две модели фотометрической классификации рентгеновских источников: модель трехклассовой классификации (звезда/галактика/квазар) и составная модель на основе набора бинарных классификаторов. Ниже мы рассмотрим модели фотометрической классификации SRGz более подробно.

5.2.1. Трехклассовая модель (*Model_{1,class}*).

Мы обучили модель фотометрической классификации рентгеновских источников на три класса, определенных в спектральном каталоге SDSS: звезда/квазар/галактика. Гиперпараметры модели оптимизировались с помощью метрики отрицательного лог-правдоподобия в трехклассовом случае (использовался гиперпараметр `objective=multiclass` в алгоритме LightGBM). В результате обучения алгоритма градиентного бустинга мы получили модель, которая предсказывает тройку вероятностей на основе набора фотометрических признаков объекта:

$$P_{\text{STAR}}, P_{\text{QSO}}, P_{\text{GAL}} = F^{\text{SvQvG}}(X_{\text{ph}}), \quad (28)$$

где $P_{\text{STAR}} + P_{\text{QSO}} + P_{\text{GAL}} = 1$.

Трехклассовая модель фотометрической классификации, несмотря на свою простоту и логичность, может обладать рядом недостатков. Отметим две возможных проблемы. Во-первых, использование слоя `softmax` на выходе модели градиентного бустинга в паре с функцией ошибок вида многоклассовой кросс-энтропии в обучении моделей многоклассовой классификации не гарантирует правильной калибровки вероятностей (как обсуждалось в литературе, например, для случая нейросетевых классификаторов, см. Гуо и

др., 2017) и требует, например, дополнительной настройки параметра температуры в слое `softmax`. Второй возможной проблемой является сильная несбалансированность тренировочной и тестовой выборки по числу объектов разных классов. В случае бинарной классификации LightGBM имеет гиперпараметры для работы с несбалансированными выборками (например, `is_inbalance`, `scale_pos_weight`, `pos_bagging_fraction`), они не могут быть использованы для многоклассовых моделей кросс-энтропийной функции ошибок. Модели бинарной классификации на базе градиентного бустинга представляются более гибкими при решении указанных выше проблем. Далее мы рассмотрим составную модель фотометрической классификации рентгеновских источников, составленную из набора бинарных классификаторов.

5.2.2. Составная модель на основе бинарных классификаторов (*Model_{2,class}*).

Здесь мы независимо решали 6 задач бинарной классификации (звезда+галактика/квазар, звезда+квазар/галактика, звезда/галактика+квазар):

$$P_{\text{QSO}}^{\text{QvSG}}, P_{\text{STAR+GAL}}^{\text{QvSG}} = F^{\text{QvSG}}(X_{\text{ph}}), \quad (29)$$

$$P_{\text{GAL}}^{\text{GvSQ}}, P_{\text{STAR+QSO}}^{\text{GvSQ}} = F^{\text{GvSQ}}(X_{\text{ph}}), \quad (30)$$

$$P_{\text{STAR}}^{\text{SvQG}}, P_{\text{QSO+GAL}}^{\text{SvQG}} = F^{\text{SvQG}}(X_{\text{ph}}). \quad (31)$$

Затем мы получили вероятности для фотометрической классификации на три класса путем объединения прогнозов моделей бинарной классификации, следующим образом:

$$P_{\text{STAR}} = \frac{P_{\text{STAR}}^{\text{SvQG}}}{P_{\text{STAR}}^{\text{SvQG}} + P_{\text{QSO}}^{\text{QvSG}} + P_{\text{GAL}}^{\text{GvSQ}}}, \quad (32)$$

$$P_{\text{QSO}} = \frac{P_{\text{QSO}}^{\text{QvSG}}}{P_{\text{STAR}}^{\text{SvQG}} + P_{\text{QSO}}^{\text{QvSG}} + P_{\text{GAL}}^{\text{GvSQ}}}, \quad (33)$$

$$P_{\text{GAL}} = \frac{P_{\text{GAL}}^{\text{GvSQ}}}{P_{\text{STAR}}^{\text{SvQG}} + P_{\text{QSO}}^{\text{QvSG}} + P_{\text{GAL}}^{\text{GvSQ}}}. \quad (34)$$

Модель составной классификации по схеме один-против-всех, представленная выше, имеет ряд недостатков. Основной трудностью является возможное различие калибровок вероятностей, полученных разными моделями (что может быть особенно заметным для несбалансированных выборок, подробнее см. Бишоп, 2007).

5.2.3. Тренировочная выборка и обучение моделей классификации. Для оптических объектов в 30''-окрестности точечных рентгеновских источников ePOZITA мы применили модель 3x

отождествления оптических компаньонов и отобрали наиболее вероятные оптические компаньоны ($srg_match_flag = 1, 2$). Полученный список оптических компаньонов мы сопоставили в $1''$ со спектроскопической выборкой звезд, квазаров, галактик из обзора SDSS DR17, а также с выборкой звезд GAIA DR2 ($S/N > 5$ по параллаксу или собственному движению) в области DESI LIS. Таким образом, была получена выборка для обучения и тестирования моделей фотометрической классификации рентгеновских источников. В табл. 4 мы приводим число объектов в тренировочной и тестовой части выборки.

Гиперпараметры модели градиентного бустинга для трехклассовой классификации (28) оптимизировались при помощи метрики многоклассовой кросс-энтропии. Гиперпараметры моделей градиентного бустинга для бинарной классификации (29)–(31) оптимизировались при помощи метрики ROCAUC, несбалансированность тренировочных выборок здесь учитывалась при помощи настройки соответствующих гиперпараметров алгоритма LightGBM.

5.2.4. Метрики оценки качества фотометрической классификации. Для оценки качества фотометрической классификации рентгеновских источников на выборке с известными (из астрометрических или спектроскопических данных) классами мы будем использовать метрику полноты классификации. Для выбранного класса C_i и предсказанного класса C_j :

$$Recall_{ij} = \frac{\hat{N}_{C_j^{i*}}}{N_{C_i}}, \quad (35)$$

где N_{C_i} — число объектов истинного класса C_i в выборке, $\hat{N}_{C_j^{i*}}$ — число объектов истинного класса C_i , классифицированного как объекты фотометрического класса C_j . Набор значений $Recall_{ij}$ представляет собой так называемую матрицу путаницы (англ. confusion matrix).

В разделе 6 мы приведем результаты сравнения по этим метрикам двух описанных выше подходов фотометрической классификации рентгеновских источников. Результаты классификации лучшими моделями приводятся в каталоге SRGz (см. далее п. 5.2.5).

5.2.5. Данные о фотометрической классификации рентгеновских источников в каталоге SRGz. Для каждого объекта в каталоге SRGz-CatD мы приводим вероятности различных классов и метку наиболее вероятного класса (в столбцах srg_match_pstar , srg_match_pqso , srg_match_pgal , srg_match_SQG — см. табл. 2).

5.3. SRGz: модель измерения фотометрических красных смещений

Нами обучались четыре модели photo-z на данных из различных фотометрических обзоров.

- **Модель 2.** По данным двух обзоров — DESI LIS и WISE (принудительная фотометрия W1 и W2 в каталоге DESI LIS).
- **Модель 3_{SDSS}.** По данным трех обзоров — DESI LIS, WISE и SDSS.
- **Модель 3_{PS}.** По данным трех обзоров — DESI LIS, WISE и Pan-STARRS1.
- **Модель 4.** По данным четырех обзоров — DESI LIS, WISE, SDSS и Pan-STARRS1.

На основе указанных выше моделей photo-z далее была получена комбинированная модель SRGz, для всех найденных оптических компаньонов рентгеновских источников. Для этого все описанные выше модели были отранжированы по достигаемой ими точности следующим образом:

$$4 \rightarrow 3_{SDSS} \rightarrow 3_{PS} \rightarrow 2, \quad (36)$$

где модель 4 (photo-z по данным четырех обзоров) — наиболее точная. Далее для каждого найденного оптического компаньона рентгеновского источника мы выбирали одну из четырех моделей в соответствии с приоритетом и доступностью фотометрических данных рассматриваемых обзоров для каждого конкретного объекта.

В SRGz мы будем вычислять точечный прогноз photo-z по моде распределения $\hat{p}(z|X)$

$$z_{ph} = \underset{z}{\operatorname{argmax}} \hat{p}(z|X) \quad (37)$$

и его левые и правые 68% доверительные интервалы. Наконец, мы будем измерять стандартную величину уверенности прогноза photo-z следующим образом:

$$zConf = \int_{\delta z_{\text{norm}} < 0.06} \hat{p}(z|X) dz, \quad (38)$$

$$\text{где } \delta z_{\text{norm}} = \frac{z - z_{ph}}{1 + z_{ph}}.$$

5.3.1. Тренировочная выборка и обучение моделей photo-z. Тренировочная выборка для обучения моделей photo-z формировалась из следующих составных частей.

- 449 751 объект каталога оптических квазаров SDSS DR14Q (Парис и др., 2018). Из каталога SDSS DR14Q были случайным образом удалены часть квазаров в районе пиков распределения, чтобы сделать более ровным распределение объектов по красному смещению).

- Выборка VHzQ далеких оптических квазаров $z > 5$ из работы (Росс, Кросс, 2020) была ограничена по области покрытия обзора DESI LIS.
- 136 428 спектральных галактик из обзора SDSS DR14 (Аболфати и др., 2018). Из спектральной выборки галактик SDSS DR14 были выбраны все объекты спектрального подкласса AGN, а для объектов прочих спектральных подклассов делались случайные подвыборки таким образом, чтобы соотношение числа объектов соответствующего спектрального подкласса к числу объектов подкласса AGN соответствовало показателям, измеряемым у спектроскопических оптических компаньонов рентгеновских источников в поле XMM-XXL-N (Мензель и др., 2016). Таким способом мы приблизили распределение по подклассам спектроскопической выборки галактик SDSS (среди которых в основном большие красные галактики, снятые по программам обзора BOSS) к ожидаемому распределению спектральных подклассов рентгеновских галактик (многие из которых являются галактиками с активными ядрами).

Тренировочная выборка для моделей photo- z покрывает всю область обзора DESI LIS (в табл. 4 приведено суммарное число объектов). Координаты и красные смещения объектов выборки приведены на странице результатов SRGz⁷.

В качестве тестовой выборки для оценки качества измерений photo- z рентгеновских источников ePOZITA использовалась выборка рентгеновских источников с измеренными спектроскопическими красными смещениями в поле Stripe82X (см. п. 3.4).

5.3.2. Метрики оценки качества моделей photo- z . Для оценки качества точечных прогнозов photo- z мы используем две метрики:

- нормализованное среднее абсолютное отклонение

$$\sigma_{NMAD} = 1.48 \times \text{median}(|\delta z_{\text{norm},i}|), \quad (39)$$

- долю катастрофических выбросов

$$n_{>0.15} = \frac{\#\{i = \overline{1, N} | \delta z_{\text{norm},i} > 0.15\}}{N}, \quad (40)$$

где N — размер тестовой выборки, а нормированная невязка прогноза красного смещения (у i -го объекта) равна

$$\delta z_{\text{norm},i} = \frac{\hat{z}_{\text{ph},i} - z_{\text{spec},i}}{1 + z_{\text{spec},i}}.$$

Доля катастрофических выбросов рассчитывается для порога $|\delta z_{\text{norm},i}| > 0.15$, что является традиционной метрикой оценки доли аномальных измерений в прогнозах фотометрических красных смещений рентгеновских квазаров (см., например, Сальвато и др., 2022; Мещеряков и др., 2018).

5.3.3. Данные о фотометрических красных смещениях внегалактических рентгеновских источников в каталоге SRGz. Для каждого объекта в каталоге SRGz-CatD мы вычисляем точечную оценку красного смещения объекта srgez_z_max (полученную как мода вероятностного распределения $\hat{p}(z|X)$, см. формулу (37), приводим для нее левые и правые 68% доверительные интервалы, а также стандартную величину уверенности прогноза z_{Conf} (см. формулу (38)).

6. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Для получения каталога SRGz-CatD наиболее вероятных оптических компаньонов точечных рентгеновских источников ePOZITA в области обзора DESI LIS мы обучили лучшие модели SRGz на данных соответствующих тренировочных выборок (см. табл. 4 и описание обучения моделей в п. 5.1.3, п. 5.2.3, п. 5.3.1). Во всех трех задачах, которые последовательно решает SRGz (оптическое отождествление, фотометрическая классификация и измерения photo- z), мы сравниваем лучшие и базовые модели. Для оптического отождествления базовой моделью является отождествление по ближайшему оптическому соседу (см. п. 5.1.1). Для фотометрической классификации базовой является трехклассовая модель (п. 5.2.1). Для измерения photo- z мы будем сравнивать результаты лучшей модели SRGz с результатами моделей из литературы для тестовой области Stripe82X (Ананна и др., 2017; Брешиа и др., 2019).

Описание результатов работы организовано следующим образом. В п. 6.1 выполнен анализ фотометрических признаков (из рассматриваемых в п. 3.2), которые модели SRGz используют для отождествления и классификации оптических компаньонов точечных рентгеновских источников ePOZITA, выделены наиболее важные фотометрические признаки. Затем в п. 6.2 мы будем обсуждать результаты моделей оптического отождествления SRGz в отношении рентгеновских объектов с оптическими компаньонами и без оптических компаньонов в DESI LIS (“бездомные”

⁷<https://www.srg.cosmos.ru/srgz2023>

рентгеновские источники). Далее в п. 6.3 обсуждаются результаты моделей фотометрической классификации SRGz. В разделе 6.4 обсуждаются применение SRGz для оптического отождествления точечных рентгеновских источников ePOZITA, а также их классификация и измерение фотометрических красных смещений, в поле Stripe82X и делается сравнение с результатами, доступными в литературе по этому полю.

6.1. Анализ важности фотометрических признаков для отождествления и классификации оптических компаньонов точечных рентгеновских источников ePOZITA

Перед тем как представить результаты автоматической классификации SRGz, мы задались вопросом, какие фотометрические признаки объектов являются наиболее важными для разделения оптического компаньона и объекта поля, для разделения между различными физическими классами объектов. Отметим, что обученные модели машинного обучения позволяют оценить важность признаков автоматически, однако эти оценки могут плохо работать в условиях большого количества скоррелированных фотометрических признаков. Оценка важности признаков модели, встроенная в метод градиентного бустинга, также страдает от этого недостатка, и, показав один признак более важным, важность других признаков может быть искусственно понижена.

Наиболее точные и подробные оценки важности признаков дает использование метода векторов Шэпли (англ. Shapley additive explanations, SHAP — Лундберг, Ли, 2017). Подробный анализ важности фотометрических признаков рентгеновских объектов выходит за рамки данной работы, поэтому мы воспользуемся здесь более простой оценкой важности, в которой все признаки рассматриваются независимо. Мы будем вычислять для каждого фотометрического признака, входящего в модель, величину площади под ROC-кривой (англ. Receiver Operating Characteristic, рабочая характеристика приемника), являющейся стандартной мерой качества ранжирования в задачах классификации (Фосетт, 2006).

В табл. 5 приведено значение площади под ROC-кривой, которое отражает важность фотометрического признака при ранжировании объектов разных классов: (i) оптических компаньонов и объектов поля в 30''-окрестности, (ii) рентгеновских квазаров и звезд, (iii) рентгеновских квазаров и галактик, (iv) рентгеновских звезд и галактик. В таблице признаки отсортированы в порядке важности, приведены 10 наиболее важных признаков для каждой задачи.

На основе важности признаков табл. 5 можно сделать несколько интересных выводов:

- Для всех задач, кроме задачи разделения квазар/галактика, отношения рентген/оптика являются наиболее важными признаками, разделяющими классы. Так, для указанных задач различные варианты отношения рентген/оптика обязательно входят в TOP3 признаков.
- Для разделения квазар/галактика большое значение имеют введенные нами выше (формулы 10–12) цветовые расстояния, а также цвета объектов в оптическом и ИК-диапазоне.
- Цвет ($W_1 - W_2$) наиболее важен для задач разделения классов рентгеновский квазар/звезда и рентгеновский квазар/галактика (для этих задач ($W_1 - W_2$) находится на первом/втором месте по важности).
- Для разделения галактика/звезда и галактика/квазар важную роль имеют признаки *seraic* и *type* — они определяются морфологией и формой профиля яркости объекта на оптических изображениях DESI LIS. Важность данных признаков при разделении протяженных/точечных объектов в оптическом диапазоне вполне понятна.
- В задаче разделения оптического компаньона и объекта поля наиболее важную роль играют ИК-величина W_2 и отношение рентген/ИК X/W_2 . По важности эти признаки входят в TOP3 вместе с признаком расстояния между рентгеновским и оптическим источником в единицах ошибки локализации (*offset*).

Опираясь на описанные выше закономерности, мы отобрали группу признаков, которые представляют наибольший интерес с точки зрения визуализации и интерпретации фотометрических данных рентгеновских источников. Сначала мы разбили признак *type* на две категории: (i) *type* = *PSF*, *DUP* (характерные для звезд и квазаров) и (ii) *type* = *REX*, *SER*, *DEV*, *EXP* (характерные для радиальных профилей яркости галактик). Далее мы выбрали две звездных величины (оптическую g и ИК W_2), два цветовых расстояния ($(z - W_2) distance$, $(g - W_1) distance$), два цвета (оптический $(g - r)$, ИК $(W_1 - W_2)$) и два отношения рентген/оптика(ИК) — (X/z) и $(X/W_1)_{salvato}$.

Мы разбили указанные выше признаки на пары и будем анализировать распределения объектов разных физических типов на плоскостях (отдельно для каждой из двух категорий оптической морфологии источника).

Таблица 5. Важность фотометрических признаков при разделении рентгеновских объектов различных классов

#	Фотометрический признак	ROC AUC	#	Фотометрический признак	ROC AUC
Оптический компаньон/объект поля			Галактика/звезда		
3_X	<i>offset</i>	0.97	$3_{X,col}$	<i>seraic</i>	0.98
	$\mathbf{W_2}$	0.94		<i>type</i>	0.98
	X/W_2	0.94		$\mathbf{X/z}$	0.90
	$(\mathbf{X/W_1})_{salvato}$	0.91		X/W_1	0.88
	W_1	0.90		$(\mathbf{X/W_1})_{salvato}$	0.88
	X/W_1	0.90		X/r	0.87
	X/g	0.90		X/W_2	0.87
	$\mathbf{g}$	0.89		X/g	0.86
	X/r	0.88		$(\mathbf{W_1 - W_2})$	0.84
	r	0.87		$(z - W_2)$	0.80
Квazar/звезда			Квazar/галактика		
$3_{X,col}$	$\mathbf{X/z}$	0.99	$3_{X,col}$	$(\mathbf{W_1 - W_2})$	0.97
	$(\mathbf{W_1 - W_2})$	0.99		$(g - z)$	0.97
	X/W_1	0.99		$(\mathbf{g - r})$	0.97
	$(\mathbf{X/W_1})_{salvato}$	0.99		$(\mathbf{z - W_2})$ distance	0.96
	X/r	0.98		<i>seraic</i>	0.95
	X/W_2	0.98		$(\mathbf{g - W_1})$ distance	0.94
	X/g	0.97		$(r - z)$	0.94
	$(z - W_2)$	0.94		$(r - W_2)$	0.94
	$(\mathbf{z - W_2})$ distance	0.92		$(\mathbf{X/W_1})_{salvato}$	0.93
	$(r - W_2)$	0.89		<i>type</i>	0.92

Примечание. Значение площади под ROC-кривой (ROC AUC) отражает важность признака при ранжировании: (i) оптических компаньонов и объектов поля в $30''$ -окрестности, (ii) квазаров и звезд, (iii) квазаров и галактик, (iv) галактик и звезд. В таблице признаки отсортированы по ROC AUC и приведены 10 наиболее важных признаков для каждой задачи. Жирным выделены восемь признаков, которые мы будем использовать на графиках интерпретации моделей.

1. Диаграммы цвет—величина:

- оптический цвет $(g - r)$ — ИК-величина W_2 ,
- цветовое расстояние $(g - W_1)$ distance — оптическая величина g .

2. Диаграммы цвет—цвет:

- инфракрасный цвет $(W_1 - W_2)$ — отношение рентген—оптика (X/z) ,

- цветовое расстояние $(z - W_2)$ — отношение рентген—ИК $((X/W_1)_{salvato})$.

Следует отметить, что на таком большом объеме данных о рентгеновских источниках в обзоре неба подобный детальный анализ фотометрических свойств рентгеновских источников делается впервые и может быть чрезвычайно полезен для дальнейших исследований в области рентгеновской астрофизики, опирающихся на анализ многоволновой фотометрии.

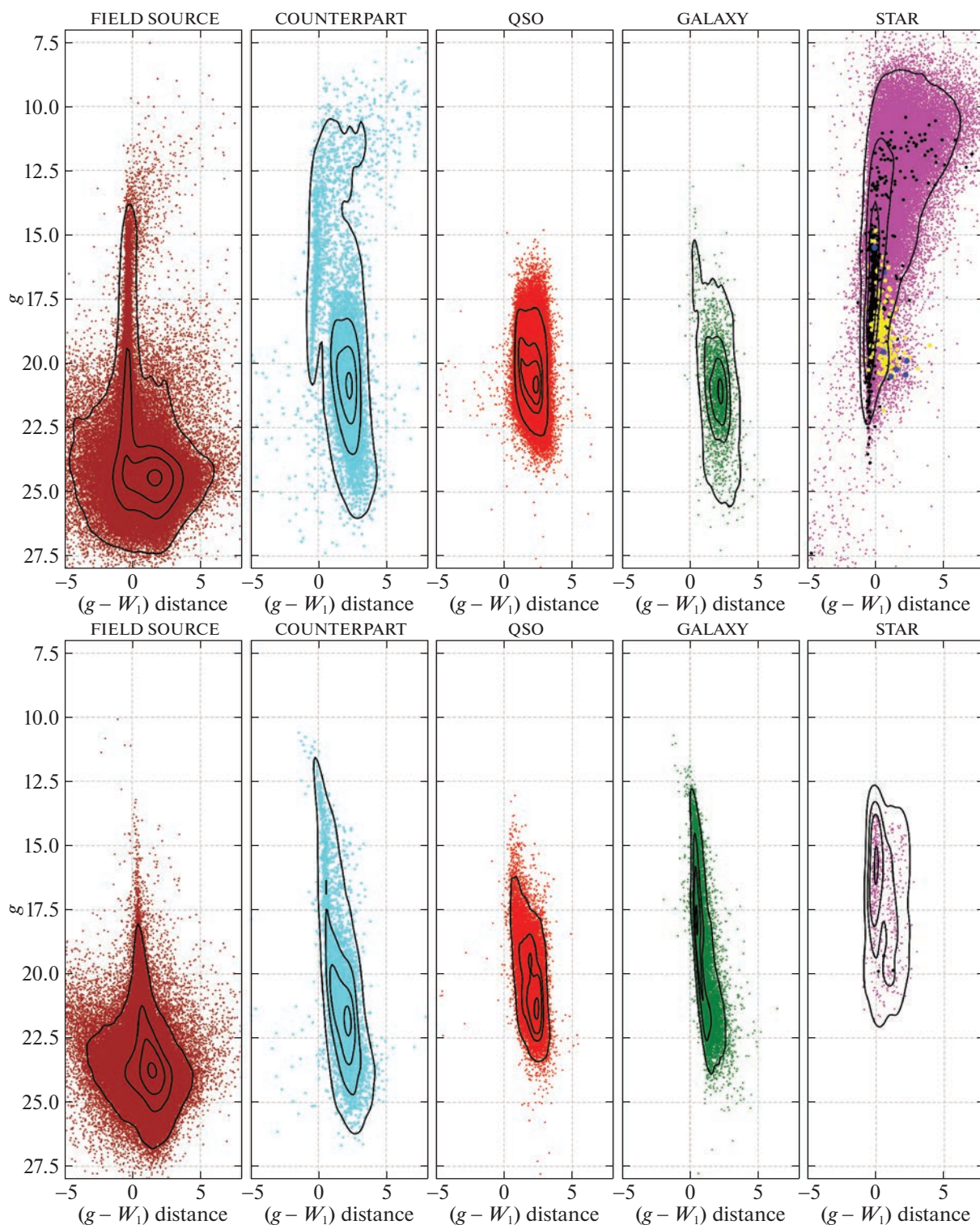


Рис. 5. Разграничение оптических компаньонов точечных рентгеновских источников двухлетнего обзора ePOZITA для спектральных/астрометрических классов на плоскости цветное расстояние $(g - W_1)$ — величина g . Данные приведены для звездообразных оптических объектов (верхняя панель) и протяженных источников (нижняя панель) в обзоре DESI LIS. В левом столбце показано разделение между оптическими компаньонами (голубые точки) и объектами поля (коричневые точки), следующие столбцы графиков — квазары, галактики, звезды. На графиках справа показаны звезды GAIA (пурпурные точки), катаклизмические переменные (CV — желтые точки), белые карлики (WD — синие точки), прочие спектроскопические звезды SDSS (черные точки).

Далее мы рассмотрим подвыборку объектов каталога SRGz-CatD в области DESI LIS, у которых имеется спектральная (SDSS) или астрометрическая (GAIA) классификация. Также мы рассмотрим выборку объектов из каталога SRGz-CatA, для которых имеются указания (по данным Chandra, Chandra, Swift) о том, какой именно оптический объект является компаньоном источника ePOZITA, а какие источники не связаны с соответствующим рентгеновским объектом (см. подробнее в п. 5.1.3).

Здесь мы будем рассматривать следующие физические классы оптических объектов в окрестности рентгеновских источников:

1. **Оптические компаньоны.** Класс оптических компаньонов точечных рентгеновских источников ePOZITA, отождествленных при помощи данных Chandra. На графиках объекты показаны голубыми (англ. cyan) точками.
2. **Объекты поля.** Класс оптических источников, не связанных с рентгеновскими источниками ePOZITA и отождествленных при помощи данных Chandra. На графиках объекты показаны коричневыми (англ. brown) точками.
3. **Оптические компаньоны — звезды GAIA.** Астрометрический класс звезд по GAIA DR2 (у оптического компаньона обнаружен параллакс или собственное движение с $S/N > 5$). На графиках объекты показаны пурпурными (англ. magenta) точками.
4. **Оптические компаньоны — звезды SDSS.** Спектроскопический класс звезд по SDSS DR17 (для всех объектов мы дополнительно потребовали, чтобы они также являлись звездами по GAIA). Кроме того, мы дополнительно разделили спектроскопические звезды на три категории на базе приведенного подкласса в спектроскопическом каталоге SDSS — катаклизмические переменные (CV — желтые точки), белые карлики (WD — синие точки) и все остальные подклассы звезд (черные точки на графиках).
5. **Оптические компаньоны — квазары SDSS.** Спектроскопический класс квазаров по SDSS DR17. Квазары показаны красными точками на графиках.
6. **Оптические компаньоны — галактики SDSS.** Спектроскопический класс галактик по SDSS DR17. Галактики показаны зелеными точками на графиках.

На рис. 5–8 для всех выбранных плоскостей цвет–величина и цвет–цвет показаны по две панели графиков. На верхней панели — оптические источники с морфологическим типом $type = PSF, DUP$. На нижней панели графиков показаны объекты всех остальных морфологических типов ($type = REX, SER, DEV, EXP$) — протяженные оптические источники. На каждом столбце графиков рис. 5: оптические объекты поля, компаньоны рентгеновских источников, рентгеновские квазары, рентгеновские галактики, рентгеновские звезды.

Из представленных рисунков может быть сделан ряд важных выводов:

- Как можно видеть на левом верхнем графике рис. 5, основная масса рентгеновских звезд (не CV) и квазаров хорошо разделяется по оптической звездной величине в силу кардинально различающегося у этих классов отношения рентген/оптика (см. также левые верхние графики на рис. 7–8). С другой стороны, оптическая величина плохо разделяет компаньоны–квазары и объекты поля.
- Распределение компаньонов–квазаров и объектов поля сильно перекрывается по оптическим величинам (см. левый верхний график рис. 5 и по отношению рентген/оптика (см. рис. 7). Как следует из рис. 6 и 8 (левые верхние графики), разделение компаньонов и объектов поля по ИК-величине или отношению рентген/ИК становится хорошо заметным и работает как для точечных, так и для протяженных оптических объектов.
- Классы рентгеновских квазаров и галактик заметно перекрываются в пространстве признаков и могут быть надежно разделены только по оптическому морфологическому типу.
- Классы катаклизмических переменных (желтые точки на графиках) и квазаров (красные точки) значительно перекрываются в пространстве признаков. Звезды класса CV, как и квазары, имеют высокие значения отношения рентген/оптика–ИК, что серьезно затрудняет их фотометрическое разделение от квазаров.

В следующей части мы представим и проанализируем результаты автоматической фотометрической классификации SRGz в тестовой области SDSS.

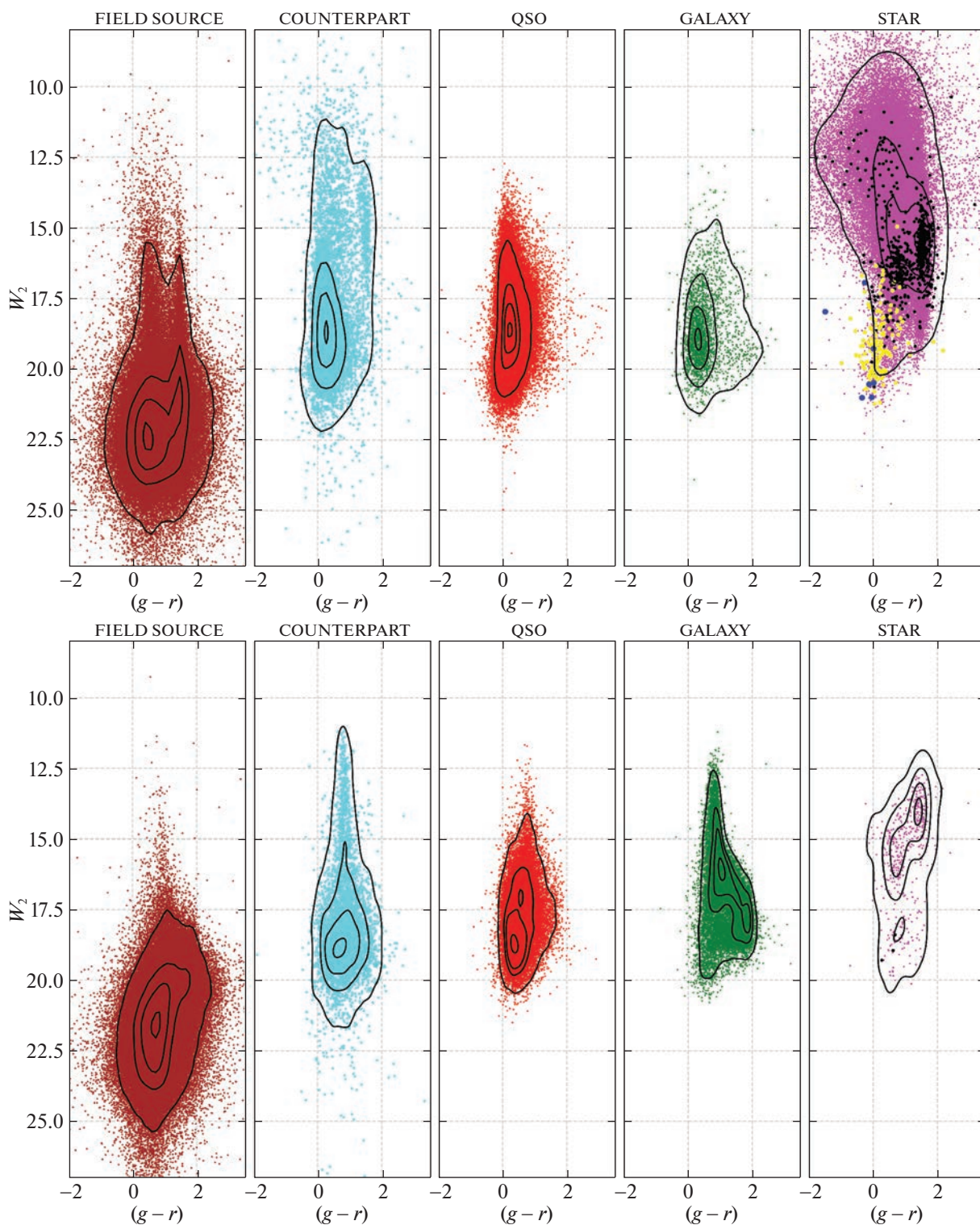


Рис. 6. Разграничение оптических компаньонов точечных рентгеновских источников двухлетнего обзора еРОЗИТА для спектральных/астрометрических классов на плоскости цвет $(g-r)$ — величина W_2 (подробности по обозначениям и графикам см. текст и подпись к рис. 5).

6.2. Отождествление оптических компаньонов и “бездомных” рентгеновских объектов

В каталоге SRGz-CatD сформирована выборка вероятных оптических компаньонов точечных

рентгеновских источников еРОЗИТА. Для каждого рентгеновского объекта в каталоге содержится, во-первых, наиболее вероятный оптический ком-

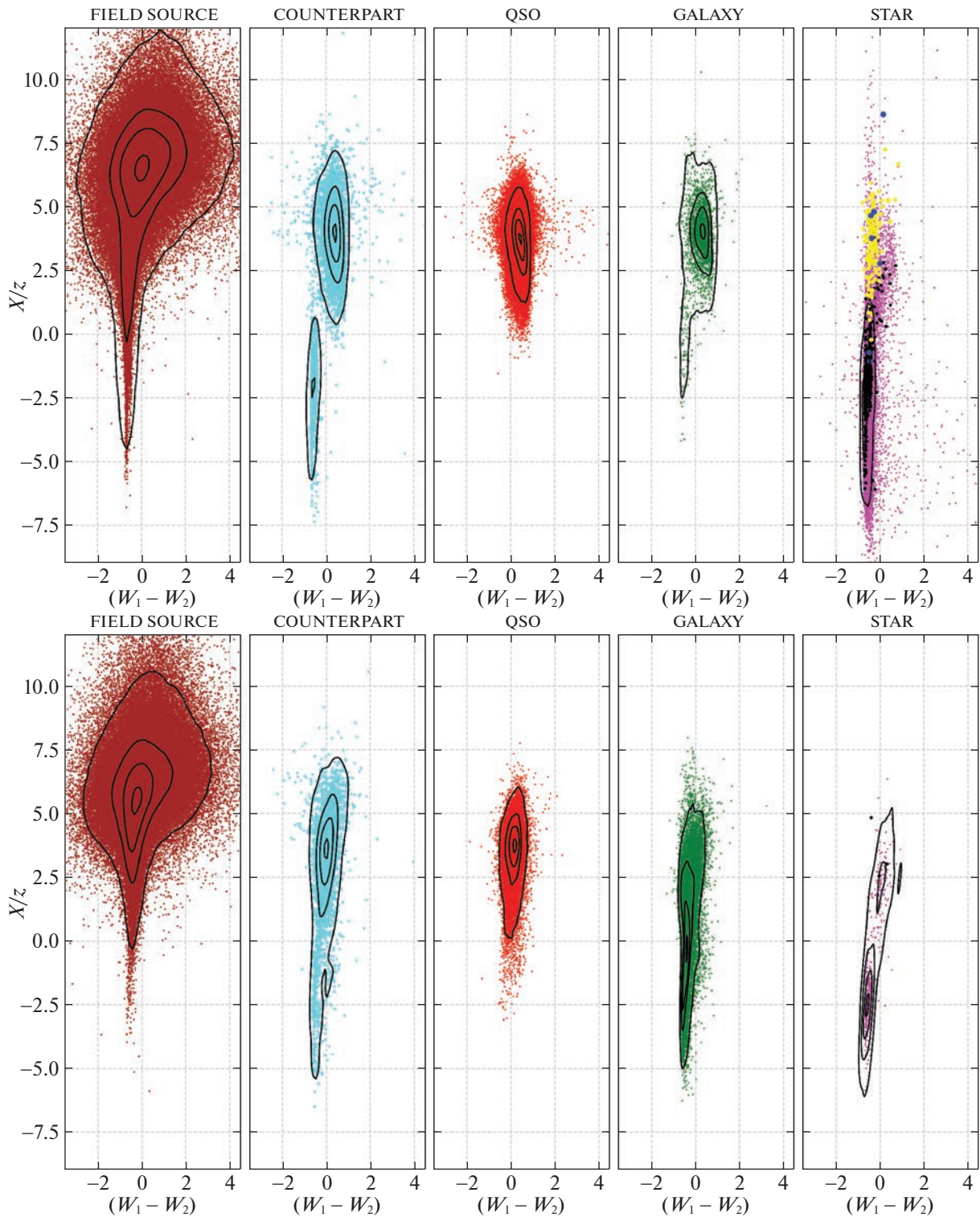


Рис. 7. Разграничение оптических компаньонов точечных рентгеновских источников двухлетнего обзора еРОЗИТА для спектральных/астрометрических классов на плоскости цвет $(W_1 - W_2)$ — отношение рентген/оптика X/z (подробности по обозначениям и графикам см. текст и подпись к рис. 5).

паньон и, во-вторых, другие возможные компаньоны с высоким p_i (значения “1” и “2” в столбце `_match_flag` соответственно). Для всех объектов в

каталоге SRGz-CatD приводятся вероятности P_i , P_i , P_0 , введенные нами ранее (см. п. 5.1.5 и табл. 2).

Какая-то часть рентгеновских источников

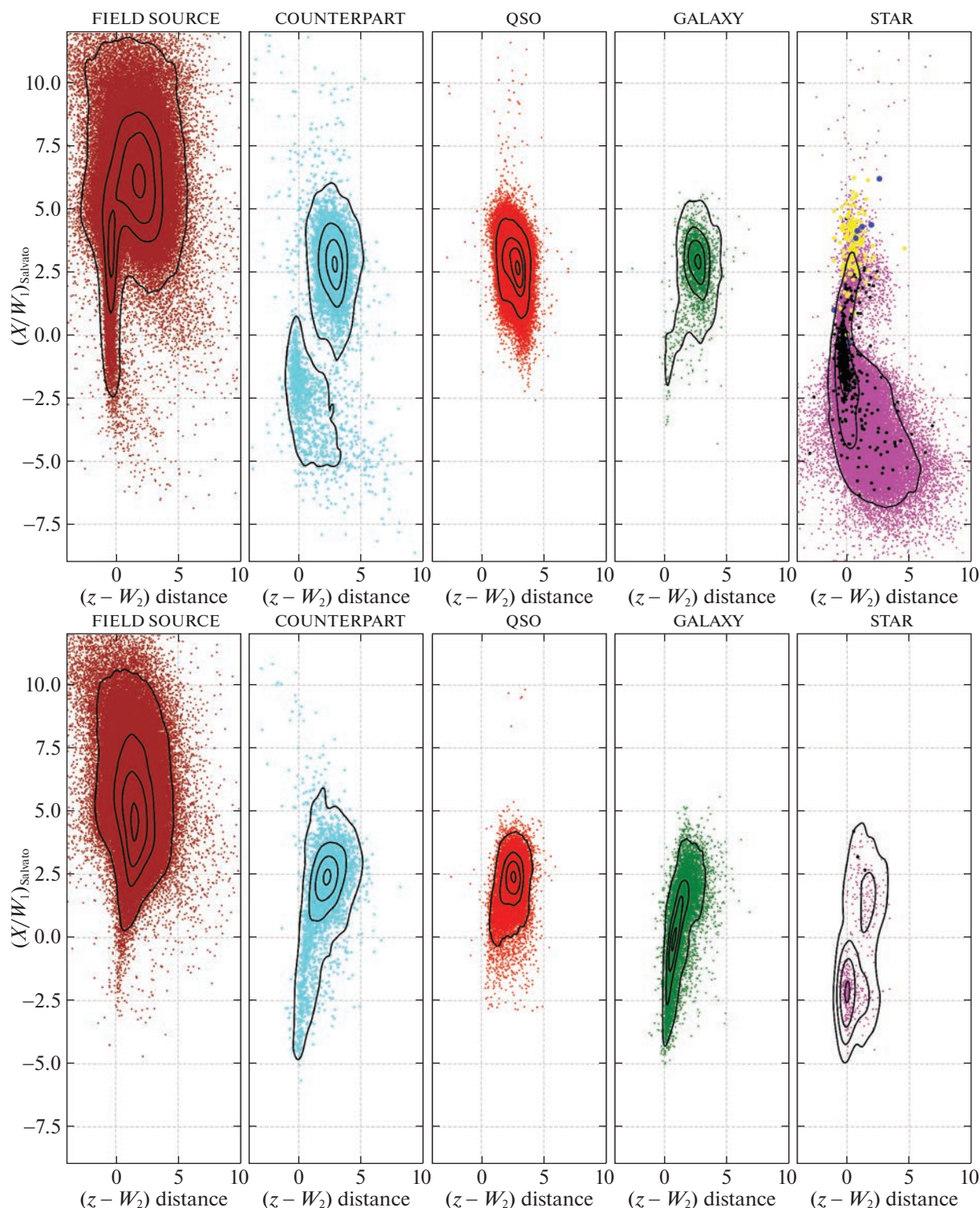


Рис. 8. Разграничение оптических компаньонов точечных рентгеновских источников двухлетнего обзора ePOZITA для спектральных/астрометрических классов на плоскости цветное расстояние $(z - W_2)$ — отношение рентген/ИК $(X/W_1)_{\text{Salvato}}$ (подробности по обозначениям и графикам см. текст и подпись к рис. 5).

ePOZITA может не иметь оптического компаньона в каталоге DESI LIS. Ранее мы сделали оценку доли “бездомных” рентгеновских объектов в зависимости от рентгеновского потока по наблюдениям ePOZITA в поле Дыры Локмана

(доля “бездомных” объектов составляет величину $(1 - P_c)$, см. рис. 15 в работе Бельведерский и др., 2022). Как можно видеть, в поле Дыры Локмана оптическая глубина DESI LIS достаточна

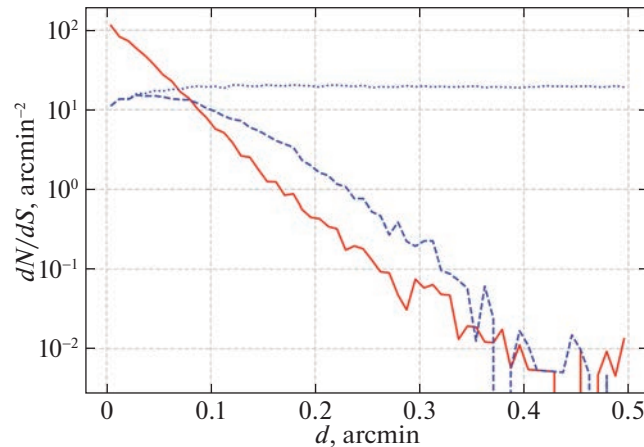


Рис. 9. Поверхностная плотность оптических объектов в зависимости от расстояния до точечного рентгеновского источника в тестовой выборке оптического отождествления. Сплошная, пунктирная, штриховые линии — поверхностная плотность оптических компаньонов, всех оптических объектов поля и ближайших объектов поля соответственно.

для регистрации оптических компаньонов у более чем 80% источников ePOZITA с рентгеновским потоком $F_{X,0.5-2} \gtrsim 7 \times 10^{-15}$ эрг $\text{с}^{-1} \text{см}^{-2}$.

Идентифицировать “бездомные” рентгеновские объекты в SRGz-CatD позволяет величина $P_\emptyset$ (столбец `srg_match_p0`), приведенная для каждого рентгеновского источника в каталоге. Оптическое отождествление рентгеновских источников для заданного порога $P_{\emptyset, \text{thres}}$ выполняется в 2 этапа (см. также рис. 3 в работе Бельведерский и др., 2022) следующим образом:

1. Из выборки рентгеновских источников отбираются объекты с $P_\emptyset \geq P_{\emptyset, \text{thres}}$, которые считаются рентгеновскими объектами, не имеющими оптического компаньона (“бездомные”) в обзоре DESI LIS.
2. Для оставшихся рентгеновских объектов с $P_\emptyset < P_{\emptyset, \text{thres}}$ в качестве наиболее вероятного оптического компаньона выбирается оптический объект в окрестности рентгеновского источника с максимальным значением p_i .

При получении выборок оптических компаньонов рентгеновских источников и “бездомных” рентгеновских объектов из каталога SRGz-CatD встает вопрос выбора наилучшего порога $P_{\emptyset, \text{thres}}$. При исследовании рентгеновских объектов с компаньонами и без них значение оптимального порога $P_{\emptyset, \text{thres}}$ может выбираться разным. Так, уменьшая значение порога, мы делаем выборку оптических компаньонов более чистой (при этом также уменьшая ее полноту). С другой стороны, увеличение $P_{\emptyset, \text{thres}}$ делает более чистой выборку “бездомных” объектов (одновременно увеличивая полноту выборки оптических компаньонов).

6.2.1. Сравнение качества моделей оптического отождествления рентгеновских источников. На рис. 9 показана зависимость поверхностной плотности оптических объектов от расстояния до точечного рентгеновского источника в тестовой выборке оптического отождествления (сплошная, пунктирная, штриховые линии — поверхностная плотность оптических компаньонов, всех оптических объектов поля и ближайших к рентгеновским источникам объектов поля соответственно). Как следует из рис. 9, с ростом расстояния до рентгеновского объекта, на $d > 5''$ плотность ближайших оптических объектов поля начинает превосходить плотность оптических компаньонов рентгеновских источников, что может привести к уменьшению точности модели оптического отождествления по ближайшему оптическому соседу.

Далее мы сравним показатели качества модели отождествления по ближайшему оптическому соседу и лучшей модели SRGz (описанные в п. 5.1.1 и п. 5.1.2 соответственно) по метрике *ROCAUC*. Мы определим, насколько хорошо выбранные модели позволяют ранжировать два класса оптических объектов в тестовой выборке: (i) оптические компаньоны точечных рентгеновских источников (сплошная линия на рис. 9) и (ii) ближайшие к рентгеновским источникам оптические объекты поля (штриховая линия на рис. 9).

Как можно видеть из табл. 6, модель отождествления SRGz значительно лучше ранжирует оптические компаньоны и ближайшие оптические объекты поля, чем оптическое отождествление по ближайшему оптическому соседу. Метрика *ROCAUC* рассчитывалась по всем оптическим компаньонам и ближайшим оптическим объектам поля в тестовой выборке. Для расчета $ROCAUC_{R>5''}$ были использованы оптические компаньоны, расположенные далее $5''$ от рентгеновского источника, и все

Таблица 6. Результаты сравнения модели отождествления по ближайшему оптическому соседу и лучшей модели SRGz в задаче ранжирования ближайших оптических источников

Модель	$ROCAUC$	$ROCAUC_{R>5''}$
Отождествление по ближайшему оптическому соседу	0.797	0.532
SRGz $Model_{match}(3X)$	0.981	0.957

Примечание. Метрика ROCAUC рассчитывалась по всем оптическим компаньонам и ближайшим оптическим объектам поля в тестовой выборке. Для расчета $ROCAUC_{R>5''}$ были использованы оптические компаньоны, расположенные далее 5'' от рентгеновского источника, и все ближайшие оптические объекты поля в тестовой выборке.

ближайшие оптические объекты поля в тестовой выборке. Здесь различие в качестве моделей становится особенно заметно — модель ближайшего оптического соседа оказывается близка к случайному выбору оптического источника ($ROCAUC_{R>5''} \sim \sim 50\%$), в то время как модель SRGz все еще демонстрирует достаточно высокое качество ранжирования ($ROCAUC_{R>5''} \approx 97\%$) оптических объектов.

6.2.2. Точность и полнота отождествления оптических компаньонов и классификации рентгеновских объектов как “бездомных”. В этом разделе мы рассмотрим зависимость метрик точности и полноты оптического отождествления рентгеновских источников от выбранного порога по параметру отождествления P_0 лучшей модели SRGz (для подготовленной тестовой выборки в области наивысшей плотности объектов спектроскопического обзора SDSS, см. п. 3.3.1 и п. 5.1.4).

Тестовая выборка рентгеновских источников с точно определенными оптическими компаньонами была подготовлена нами ранее (см. п. 5.1.3). Из нее мы получили выборку рентгеновских источников, имеющих также “бездомные” объекты в DESI LIS. Для этого для каждого рентгеновского источника мы случайным образом разыгрывали событие отсутствия у рентгеновского источника оптического компаньона в соответствии с распределением Бернулли с параметром вероятности P_c (величины P_c были ранее измерены нами для каждого рентгеновского объекта, см. формулу (6)). Если оптический компаньон “выпадал” отсутствующим, то он удалялся из выборки. Полученная таким образом тестовая выборка содержала рентгеновские источники с оптическими компаньонами, “бездомные” рентгеновские объекты и все оптические объекты поля в 30''-окрестности рентгеновских источников. Далее мы применили лучшую модель оптического отождествления SRGz к данной тестовой выборке и для каждой пары рентгеновский—оптический источник определили величины p_i , P_i , P_0 и srg_match_flag (см. табл. 2).

Регулируя порог отождествления $P_{0,thres} = t_n$ и отбирая пары рентгеновский—оптический объект с

$P_0 \leq t_n$, мы имеем возможность получать выборки (обладающие различной полнотой и точностью) рентгеновских источников ePOЗИТА с оптическими компаньонами. Отбирая объекты с $P_0 \geq t_n$, мы получаем выборки “бездомных” рентгеновских источников.

На рис. 10 показаны зависимости точность — полнота (слева) и $P_{0,thres}$ — полнота (справа) отождествления оптических компаньонов (верхняя панель) и “бездомных” в DESI LIS (нижняя панель) рентгеновских объектов на тестовой выборке. Сплошными линиями показана лучшая модель отождествления SRGz, штриховыми — модель отождествления по ближайшему оптическому соседу. В подписях к графикам указаны показатели средней точности моделей (AP_c и AP_h). Как можно видеть, модель SRGz показывает высокую точность оптического отождествления рентгеновских источников ePOЗИТА.

На рис. 11 показана зависимость $P_{0,thres}$ — полная точность оптического отождествления рентгеновских источников ePOЗИТА. Сплошной линией показана лучшая модель отождествления SRGz, штриховой — модель отождествления по ближайшему оптическому соседу. Максимальная точность отождествления в модели SRGz составляет $Precision_{X,max} = 0.93$ (соответствует $P_{0,thres} = 0.24$).

Мы делаем вывод, что близкий к оптимальному отбор оптических компаньонов рентгеновских источников может быть сделан по следующим критериям SRGz: $srg_match_flag=1$ и $P_{0,thres} = 0.24$. В отобранную с $P_0 \leq P_{0,thres}$ выборку попадают $\approx 95\%$ оптических компаньонов рентгеновских источников (из тестовой выборки) с точностью отбора компаньонов $\approx 94\%$. Отметим, что максимальная полнота отождествления оптических компаньонов рентгеновских источников ePOЗИТА для модели SRGz составляет $Recall_{c,max} \approx 96\%$. В отобранную с $P_0 \geq P_{0,thres}$ выборку попадут $\approx 82\%$ всех рентгеновских источников без оптического

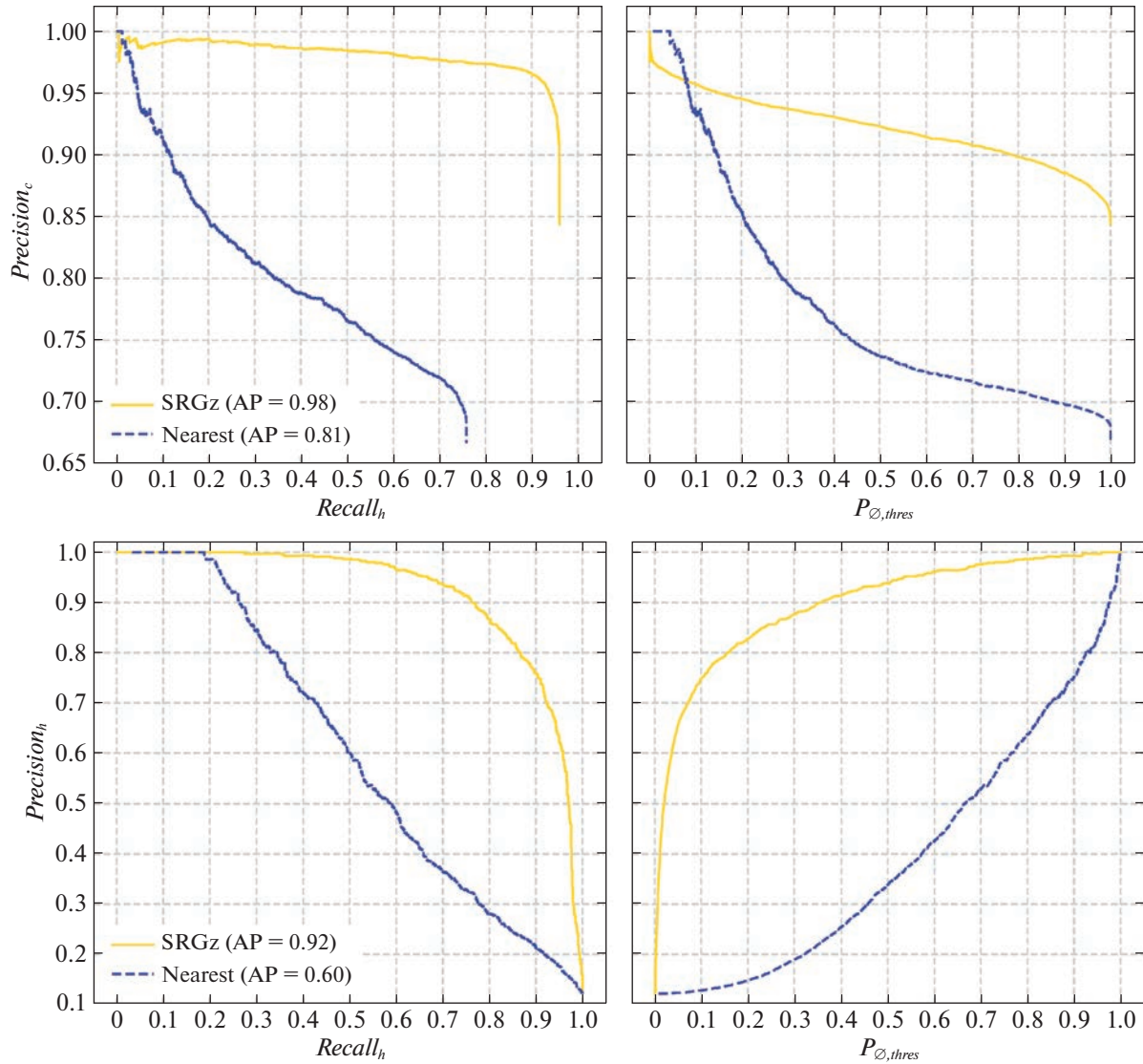


Рис. 10. Зависимости точность — полнота (слева) и $P_{0,thres}$ — полнота (справа) отождествления оптических компаньонов (верхняя панель) и “бездомных” в DESI LIS (нижняя панель) рентгеновских объектов на тестовой выборке. Сплошными линиями показана лучшая модель отождествления SRGz, штриховыми — модель отождествления по ближайшему оптическому соседу. Максимальная полнота отождествления оптических компаньонов рентгеновских источников ePOZITA для представленной модели SRGz составляет $Recall_{c,max} \approx 96\%$.

компаньона в DESI LIS, при этом с точность отбора “бездомных” рентгеновских объектов составит $\approx 85\%$.

Полученные нами результаты по оптическому отождествлению источников в обзоре ePOZITA можно попробовать сравнить с результатами отождествления точечных рентгеновских объектов в открытых данных ePOZITA (ранний выпуск данных телескопа ePOZITA и ART-XC им. М.Н. Павлинского) в поле eFEDS (см. Сальвато и др., 2022). Отметим, что рентгеновский обзор eFEDS имеет предельный поток $\approx 6.5 \times 10^{-15}$ эрг/с/см², что ~ 2 раза глубже рассматриваемых здесь данных двухлетнего обзора всего

неба ePOZITA. Отождествление оптических компаньонов в поле eFEDS проводилось в обзоре DESI LIS (DR8). В работе Сальвато и др. (2022) использовались независимо два различных метода оптического отождествления точечных рентгеновских источников — NWAY (Сальвато и др., 2018) и ASTROMATCH (Пуиз и др., 2018). На тестовой выборке, моделирующей рентгеновскую (eFEDS) и оптическую (DESI LIS) выборки, авторы оценили максимальную полноту отождествления $Recall_c = 0.95$ и $Recall_c = 0.89$ (для методов NWAY и ASTROMATCH соответственно). Таким образом, мы можем заключить, что показатели SRGz ($Recall_{c,max} = 0.96$) для двухлетнего обзора ePO-

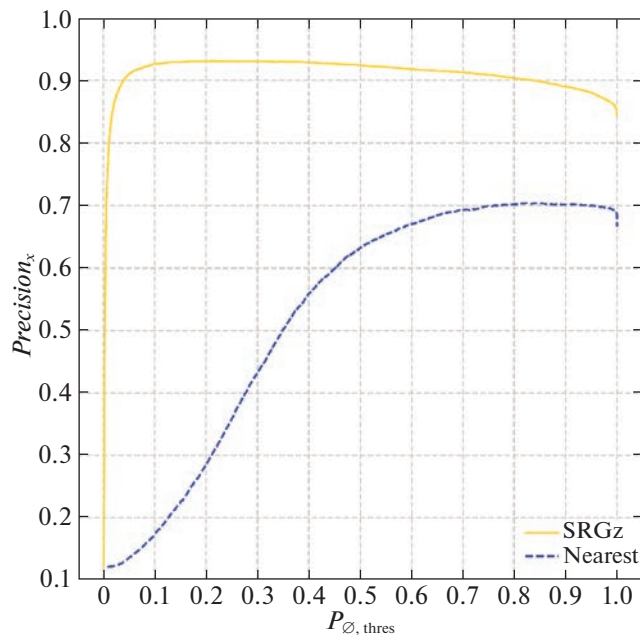


Рис. 11. Зависимость $P_{0,thres}$ — полная точность оптического отождествления рентгеновских источников еРОЗИТА. Сплошной линией показана лучшая модель отождествления SRGz, штриховой — модель отождествления по ближайшему оптическому соседу. Максимальная точность отождествления в модели SRGz составляет $Precision_{x,max} = 0.93$ (соответствует $P_{0,thres} = 0.24$).

ЗИТА соответствуют показателям отождествления лучших методов, представленных в литературе для поля eFEDS. Мы отмечаем, что прямое сравнение различных методов отождествления на одних и тех же рентгеновских и оптических данных представляет определенный интерес, но выходит за рамки данной работы.

Наконец, стоит отметить, что, хотя обзор DESI LIS состоит из данных трех фотометрических обзоров (DECaLS для $\delta \lesssim 32^\circ$ и MzLS, BASS для $\delta \gtrsim 32^\circ$), его данные имеют достаточно хорошую фотометрическую кросс-калибровку (применительно к задачам, решаемым в данной работе). Показатели оптического отождествления рентгеновских источников, фотометрической классификации и photo-z не показывают систематических отличий в зависимости от используемого обзорного телескопа в составе DESI LIS. В то же время на метрики оптического отождествления оказывают влияние вычисляемые нами максимальные оптические величины (g_{lim} , z_{lim} , см. формулу (5)) фотометрического обзора в окрестности рентгеновского источника.

6.3. Фотометрическая классификация SRGz

В этом разделе мы оценим качество фотометрической классификации SRGz рентгеновских источников в обзоре еРОЗИТА, опираясь на спектроскопические (SDSS DR17) и астрометрические (GAIA DR2) данные. Мы будем использовать выборку рентгеновских источников в тестовой области SDSS, которая не участвовала

в обучении моделей классификации. Из каталога SRGz-CatD нами были отобраны точечные рентгеновские источники с оптическими компаньонами ($P_0 \leq 0.24$ для лучшей модели отождествления SRGz, см. п. 6.2). Мы оставляли в выборке только наиболее вероятные оптические компаньоны ($srg_match_flag = 1$ в SRGz-CatD).

Результаты фотометрической классификации рентгеновских объектов с оптическими компаньонами, имеющих спектроскопическую/астрометрическую классификацию, представлены в табл. 7. Из ≈ 464 тыс. точечных рентгеновских источников еРОЗИТА (с лог-вероятностью детектирования $\mathcal{L} > 6$) в тестовой области SDSS около 426 тыс. объектов имеют оптический компаньон согласно критерию $P_0 \leq 0.24$. Для наиболее вероятных ($srg_match_flag = 1$) оптических компаньонов в данной выборке мы имеем 179 582 внегалактических объектов (спектральный класс GALAXY или QSO в каталоге SDSS DR17), 30 999 астрометрических звезд (объекты с $S/N \geq 5$ по параллаксу или собственному движению в каталоге GAIA DR2). Среди звезд GAIA имеется 406 спектральных звезд из каталога SDSS, 85 из которых являются катаклизмическими переменными (по спектральному подклассу SDSS). Таким образом, $\approx$ половина оптических компаньонов точечных рентгеновских источников обзора еРОЗИТА в тестовой обла-

Таблица 7. Результаты фотометрической, спектроскопической и астрометрической классификации источников ePOZITA в тестовой области SDSS

#	N_X	Photo-QSO	Photo-GALAXY	Photo-STAR
Все объекты $\mathcal{L} > 6$	464 404			
$P_0 \leq 0.24, srg_match_flag = 1$	426 370			
$Model_{1,class}$	426 370	343 656	48 225	34 489
QSO/GALAXY	179 582	0.9976		0.0024
QSO	165 531	0.9908	0.0073	0.0019
GALAXY	14 051	0.2681	0.7232	0.0087
STAR GAIA	30 999	0.0061	0.0010	0.9929
STAR SDSS (non-CV)	406	0.0296	0.0049	0.9655
STAR SDSS (CV)	85	0.2118	0.0000	0.7882
$Model_{2,class}$	426 370	349 962	42 322	34 086
QSO/GALAXY	179 582	0.9982		0.0018
QSO	165 531	0.9927	0.0060	0.0013
GALAXY	14 051	0.2731	0.7188	0.0080
STAR GAIA	30 999	0.0049	0.0007	0.9944
STAR SDSS (non-CV)	406	0.0172	0.0025	0.9803
STAR SDSS (CV)	85	0.2000	0.0000	0.8000
Точечные оптические компаньоны рентгеновских источников				
$Model_{2,class}$	273 578	239 747	466	33 365
QSO/GALAXY	149 453	0.9980		0.0020
QSO	147 519	0.9985	0.0001	0.0013
GALAXY	1934	0.9364	0.0140	0.0496
STAR GAIA	30 822	0.0040	0.0000	0.9960
STAR SDSS (non-CV)	403	0.0149	0.0000	0.9851
STAR SDSS (CV)	85	0.2000	0.0000	0.8000
Протяженные оптические компаньоны рентгеновских источников				
$Model_{2,class}$	152 792	110 215	41 856	721
QSO/GALAXY	30 129	0.9989		0.0011
QSO	18 012	0.9446	0.0544	0.0009
GALAXY	12 117	0.1673	0.8313	0.0014
STAR GAIA	177	0.1638	0.1186	0.7175
STAR SDSS (non-CV)	3	0.3333	0.3333	0.3333

сти SDSS имеет измеренный спектроскопический/астрометрический класс.

В верхних разделах табл. 7 содержится сравнение двух ранее представленных моделей фотометрической классификации $Model_{1,class}$ и $Model_{2,class}$ (см. п. `refsec:class3model`, п. 5.2.2) на основе показателей матрицы путаницы $Recall_{ij}$. В качестве истинных классов рассматриваются:

- 1) внегалактические объекты по спектроскопии — QSO/GALAXY SDSS,
- 2) спектроскопические квазары SDSS — QSO,
- 3) спектроскопические галактики SDSS — GALAXY,
- 4) астрометрические звезды — STAR GAIA,
- 5) астрометрические звезды, имеющие спектральную классификацию (не катаклизмические переменные) — STAR SDSS (non-CV),
- 6) астрометрические звезды, имеющие спектральную классификацию катаклизмических переменных — STAR SDSS (CV).

Фотометрическая классификация выполнялась моделями SRGz по трем классам (см. соответствующие столбцы в табл. 7):

- 1) фотометрические квазары — Photo-QSO,
- 2) фотометрические галактики — Photo-GALAXY,
- 3) фотометрические звезды — Photo-STAR.

Таблица 7 показывает, что составная модель фотометрической классификации $Model_{2,class}$ превосходит модель $Model_{1,class}$ по большинству показателей. В частности, $Model_{2,class}$ правильно классифицирует 99.1% внегалактических объектов, 98% звезд, 80% катаклизмических переменных (как класс Photo-STAR) в тестовой выборке. Для $Model_{1,class}$ эти показатели составляют 99.8, 99.3 и 78.8% соответственно. Далее для фотометрической классификации в каталоге SRGz мы будем рассматривать только лучшую модель $Model_{2,class}$.

Наши результаты можно сравнить с результатами фотометрической классификации искусственной нейросетью оптических объектов в каталоге SDSS из работы (Брейша и др., 2015). Для лучшей двухклассовой модели на основе фотометрических данных SDSS (см. табл. 12 в Брейша и др., 2015), полнота классификации оптических звезд и квазаров составила 88 и 95% соответственно. В нашей работе (как можно видеть из табл. 7) модели SRGz демонстрируют очень высокую (>99%) полноту фотометрической классификации для звезд, внегалактических объектов (без разделения на классы QSO/GALAXY) и квазаров (спектральный класс QSO). Для сравнения лучшая трехклассовая модель оптических объектов поля, по данным SDSS

(см. табл. 13 в Брейша и др., 2015), демонстрирует следующие значения метрики полноты: 90% (спектральный класс QSO), 97% (GALAXY), 86% (STAR).

Следует отметить, что, хотя основная масса рентгеновских звезд очень хорошо распознается фотометрическим классификатором SRGz, для некоторых важных типов звезд полнота классификации заметно снижается. Речь идет о катаклизмических переменных (подкласс CV в спектральном каталоге SDSS). Как можно видеть из табл. 7, полнота классификации катаклизмических переменных составляет ≈80%, при этом 20% катаклизмических переменных в тестовой выборке классифицируются как фотометрические квазары (Photo-QSO). Неоднозначность фотометрической классификации катаклизмических переменных связана с тем, что данный тип рентгеновских звезд часто демонстрирует высокое отношение рентген/оптика, близкое по величине к рентгеновским квазарам (см. далее).

Другой проблемой представленной модели SRGz является неоднозначность фотометрической классификации квазаров и галактик. В тестовой области 426 370 точечных рентгеновских источников ePOZITA были классифицированы лучшей моделью SRGz на следующие фотометрические классы: 349 962 квазара (Photo-QSO), 42 322 галактики (Photo-GALAXY) и 34 086 звезд (Photo-STAR). Модель SRGz показывает следующие значения метрик полноты классификации по спектроскопическим/астрометрическим классам: 99% (спектральный класс QSO), 72% (GALAXY), 99% (STAR). Как можно видеть из табл. 7, снижение полноты классификации галактик связано с “перетеканием” объектов спектроскопического класса GALAXY в фотометрический класс Photo-QSO, 27% спектральных галактик идентифицируются моделью как фотометрические квазары. В нижних разделах табл. 7 мы приводим результаты фотометрической классификации отдельно для точечных и протяженных оптических источников. Как можно видеть, 94% точечных оптических объектов спектрального класса GALAXY классифицируются как фотометрические квазары (Photo-QSO). С другой стороны, 83% протяженных оптических компаньонов спектрального класса GALAXY классифицируются как фотометрические галактики (Photo-GALAXY), и только 17% — как фотометрические квазары (Photo-QSO). Разделение между спектральными классами QSO и GALAXY в каталоге SDSS базируется на выборе оптимального шаблона спектра центральных областей оптических объектов (Болтон и др., 2012). Фотометрическая путаница между спектральными квазарами и галактиками носит объективный

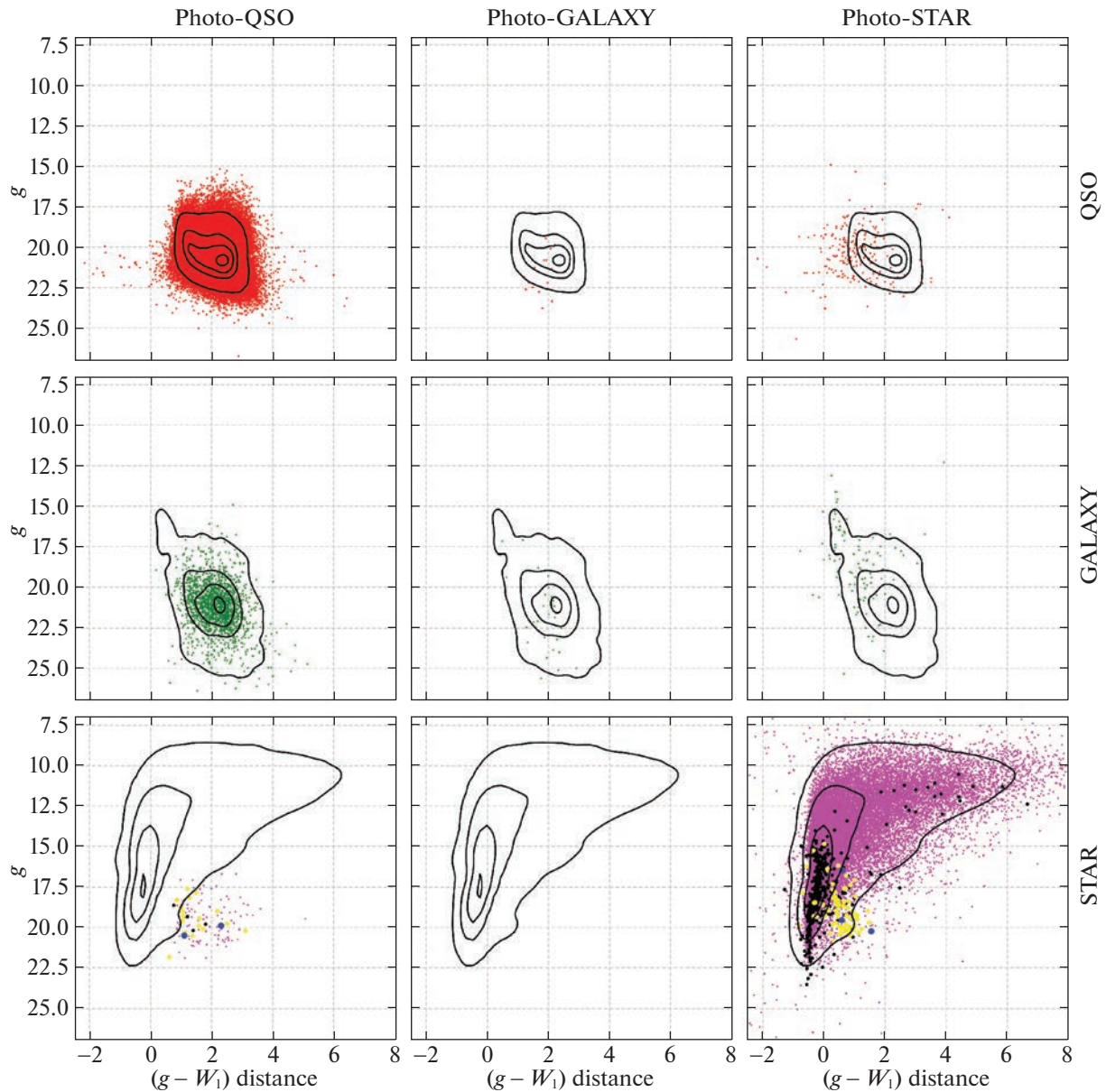


Рис. 12. Фотометрическая классификация SRGz точечных оптических объектов ($type = PSF, DUP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цветовое расстояние $(g - W_1)$ — величина g . Левый столбец графиков — фотометрические звезды, средний — фотометрические галактики, правый столбец — фотометрические квазары. Верхняя, средняя и нижняя строки графиков демонстрируют рентгеновские источники, имеющие компаньоном спектроскопические квазары, галактики и спектроскопические/астрометрические звезды (соответственно). Красные и зеленые точки — спектральные квазары и галактики SDSS, пурпурные точки — звезды GAIA, желтые, синие, черные точки — CV, WD и все прочие классы спектроскопических звезд SDSS. Контуры на графиках соответствуют 0.05, 0.35, 0.65, 0.95 квантилям распределений спектроскопических квазаров и галактик, астрометрических звезд (взяты с рис. 5–8 соответственно).

характер и связана с тем, что мы пытаемся предсказать оптическую спектральную классификацию ядра галактики на основе отношения полных потоков объекта в нескольких широких фильтрах в ИК, оптическом и рентгеновском диапазонах.

Ранее в п. 6.1 мы отобрали три пары фотометрических признаков, которые наилучшим образом разделяют различные классы: оптические объекты

поля, оптические компаньоны, рентгеновские квазары, галактики, звезды. На рис. 5–8 мы показали распределения спектральных квазаров, галактик и астрометрических звезд — компаньонов точечных рентгеновских источников ePOZITA. Матрицы путаницы фотометрической классификации, приведенные в двух нижних панелях табл. 7, могут быть показаны наглядно на графиках наиболее

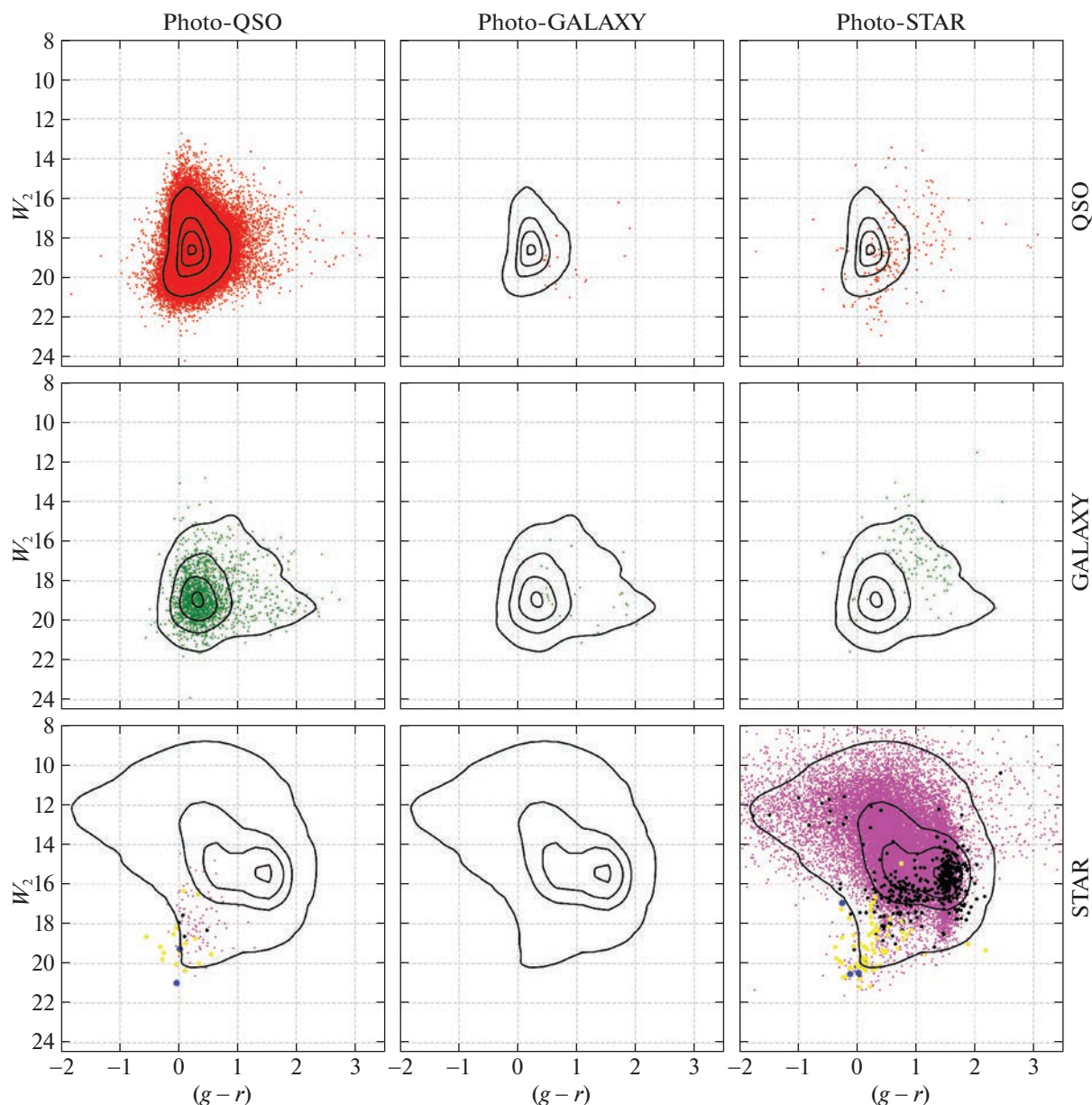


Рис. 13. Фотометрическая классификация SRGz точечных оптических объектов ($type = PSF, DUP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цвет $(g-r)$ — величина W_2 (подробное описание см. рис. 12).

важных фотометрических признаков. Графическая демонстрация матриц путаницы между спектральной/астрометрической и фотометрической классификациями рентгеновских источников (в рассматриваемой тестовой области SDSS) приведена на рис. 12–15 (точечные оптические компаньоны) и рис. 16–19 (протяженные оптические компаньоны). На каждом рисунке левый столбец графиков — фотометрические квазары, средний — фотометрические галактики, правый столбец — фотометрические звезды. Верхняя, средняя и нижняя строки графиков демонстрируют рентгенов-

ские источники, имеющие компаньоном спектроскопические квазары, галактики и спектроскопические/астрометрические звезды (соответственно). Красные и зеленые точки — спектральные квазары и галактики SDSS; пурпурные точки — звезды GAIA; желтые, синие, черные точки — CV (катаклизмические переменные), WD (белые карлики) и прочие классы спектроскопических звезд SDSS соответственно. Контуры на графиках соответствуют 0.05, 0.35, 0.65, 0.95 квантилям распределений спектроскопических квазаров и галактик, астрометрических звезд (взяты с рис. 5–8). На рисун-

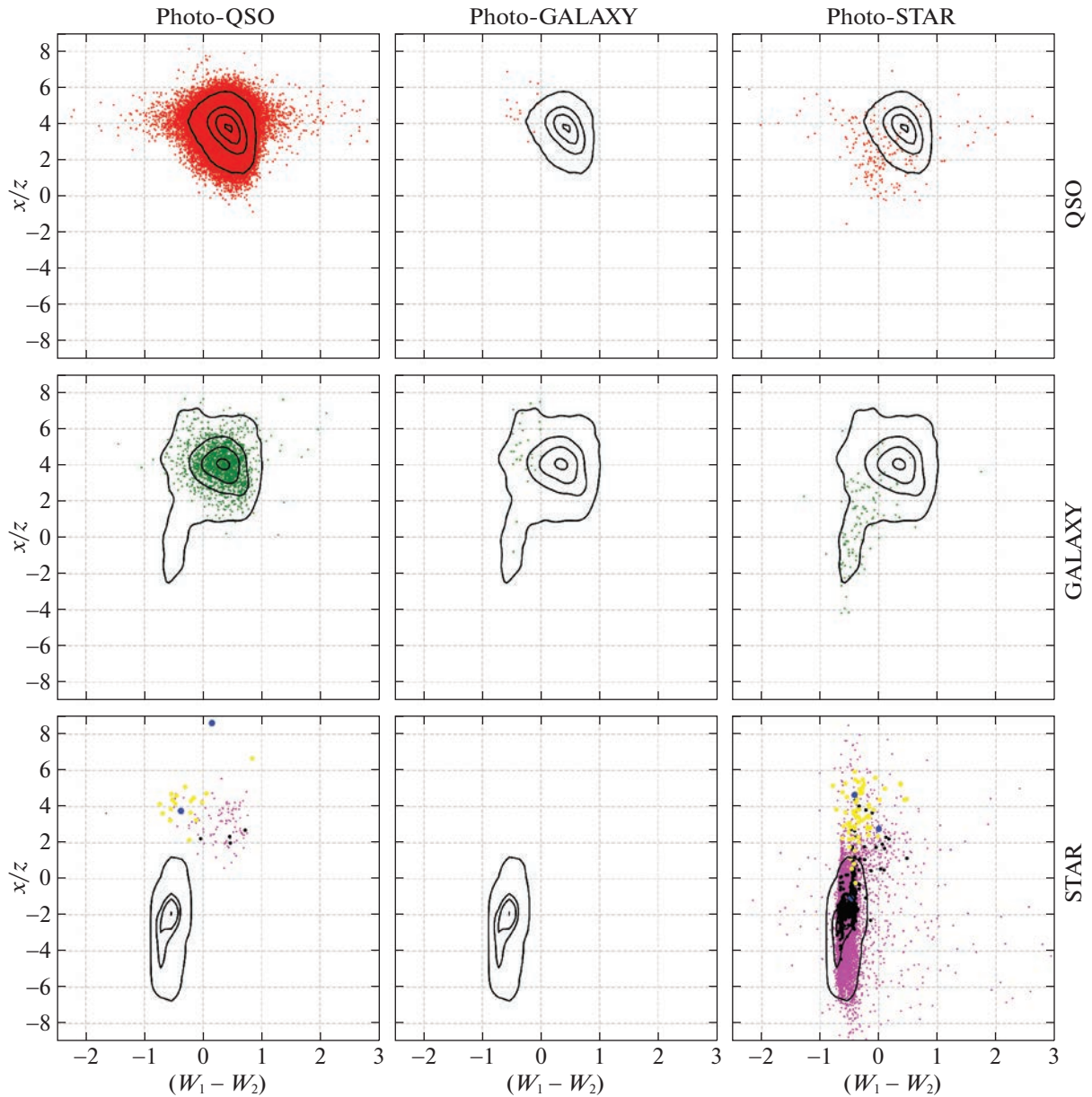


Рис. 14. Фотометрическая классификация SRGz точечных оптических объектов (*type* = *PSF*, *DUP*) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цвет $(W_1 - W_2)$ — отношение рентген/оптика X/z (подробное описание см. рис. 12).

ках можно увидеть наиболее часто встречающиеся ошибки фотометрической классификации, и в каких областях пространства параметров они происходят.

Приведенная графическая интерпретация матриц путаницы фотометрической классификации (см. рис. 12–19) подтверждает все сказанное нами ранее в отношении сложности разделения спектроскопических классов оптических квазаров и галактик применительно к рентгеновским источникам. Так, на графиках хорошо видно, что спектральные галактики — точечные оптические источники

лежат в области спектральных квазаров, и, таким образом, данные типы оптических компаньонов не могут быть разделены в выбранном пространстве фотометрических признаков. Проведенные здесь исследования демонстрируют необходимость выработки новой схемы для автоматической фотометрической классификации внегалактических рентгеновских источников, которая лучше отражает физическую природу рассматриваемых рентгеновских объектов. Данная задача выходит за рамки представленной работы. Необходимо отметить, что обнаруженный нами переток объектов между

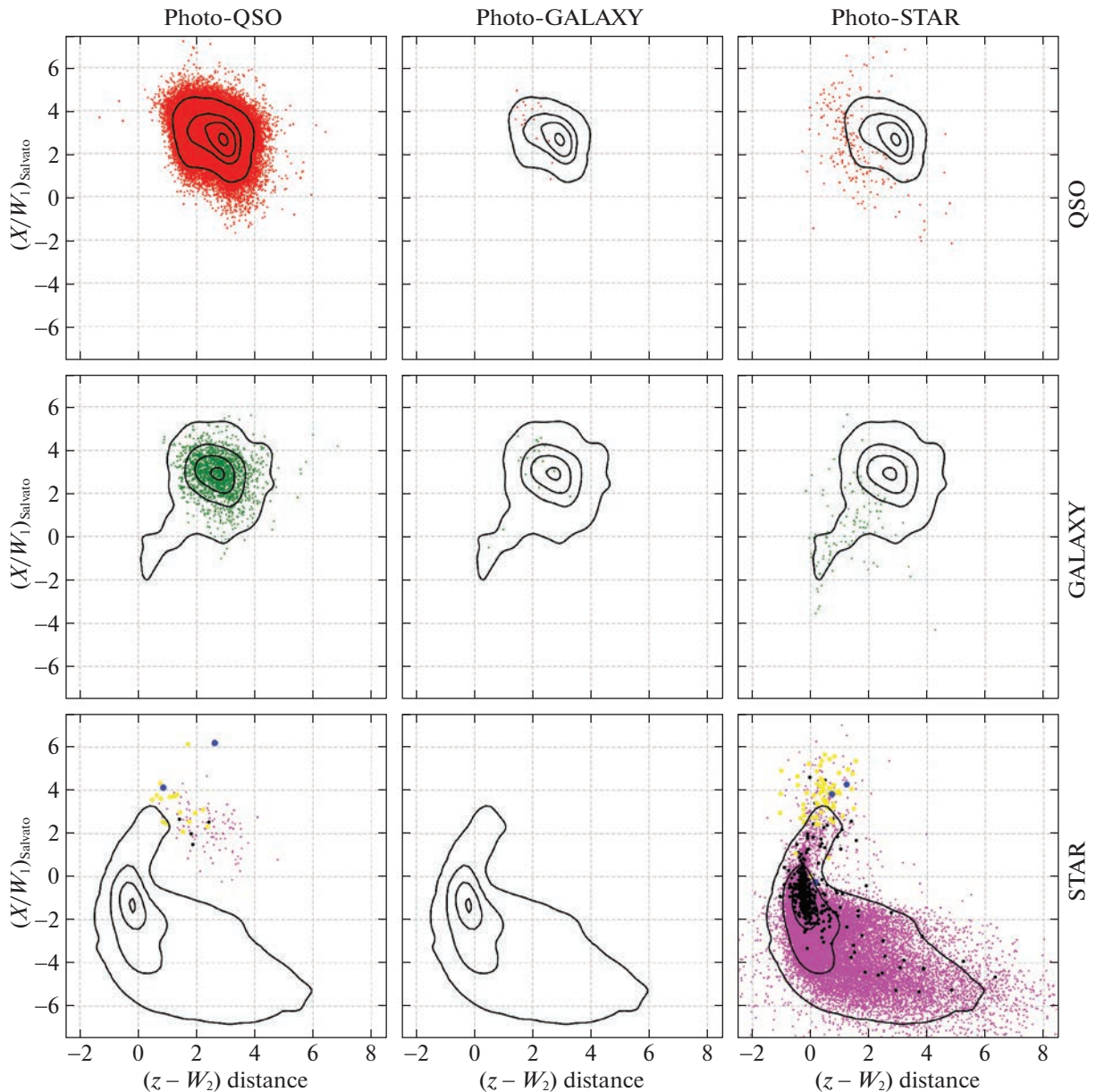


Рис. 15. Фотометрическая классификация SRGz точечных оптических объектов ($type = PSF, DUP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цветное расстояние $(z - W_2)$ — отношение рентген/ИК $(X/W_1)_{Salvato}$ (подробное описание см. рис. 12).

фотометрическими классами квазаров и галактик оказывается несущественен для многих задач, где важно определить, является ли рентгеновский источник звездой Галактики или внегалактическим объектом.

Сложность классификации катаклизмических переменных, близких к рентгеновским квазарам в выбранном фотометрическом признаковом пространстве, хорошо видна из приведенной выше графической интерпретации матриц путаницы классификации. Точное отождествление CV-звезд (а также других классов рентгеновских двойных) требу-

ет привлечения дополнительных фотометрических данных, информации о переменности оптических объектов, использования обзоров неба в узких спектральных линиях (например, VPHAS+ (Дрю и др., 2014) и подобные проекты). Мы планируем подробнее остановиться на задаче фотометрической классификации катаклизмических переменных в последующих статьях, а также представить специализированные модели классификации для отождествления рентгеновских двойных звезд среди точечных источников обзора ePOZITA.

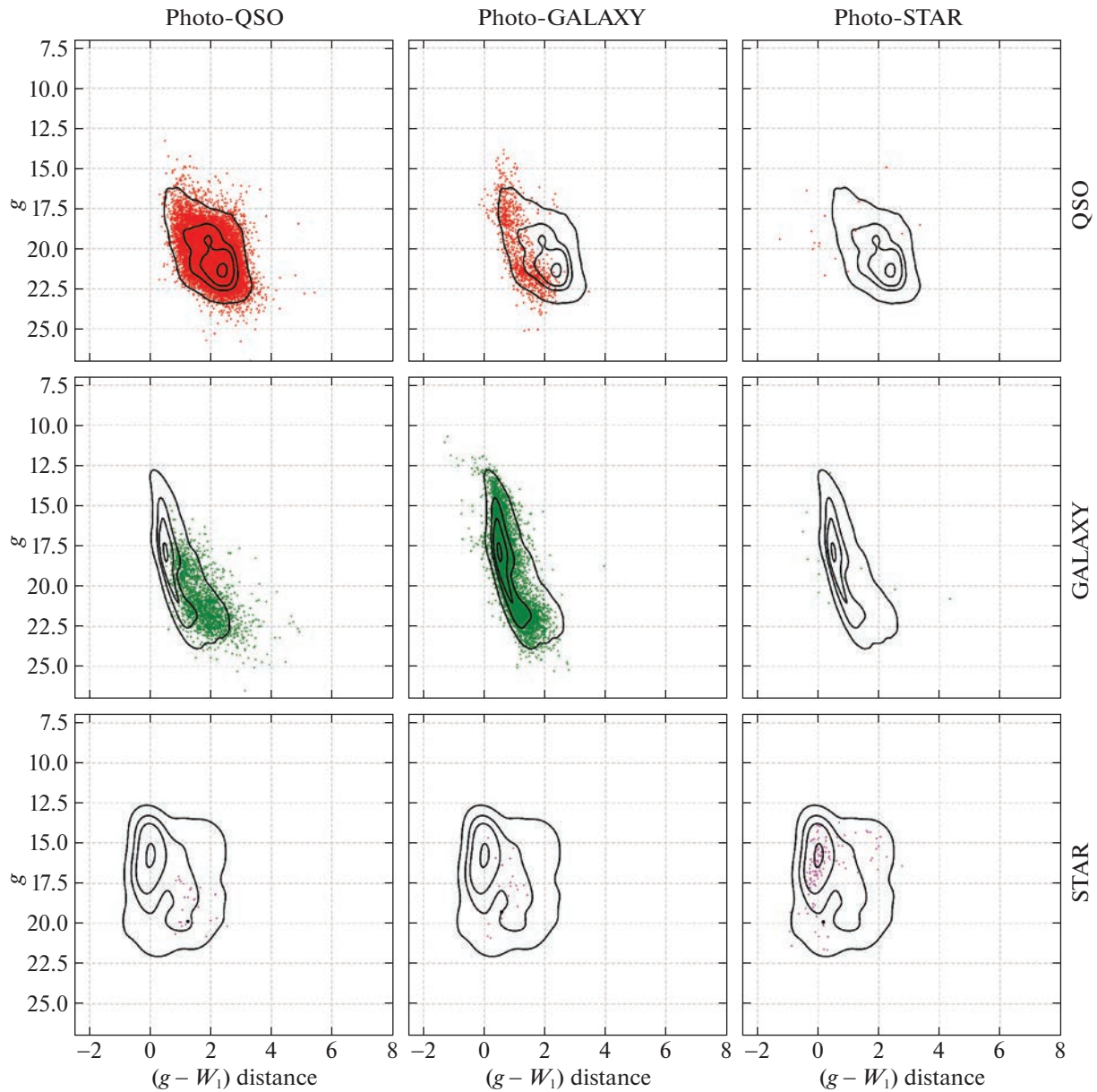


Рис. 16. Фотометрическая классификация SRGz протяженных оптических объектов ($type = REX, SER, DEV, EXP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цветное расстояние $(g - W_1)$ — величина g (подробное описание см. рис. 12).

На рис. 20–23 показана фотометрическая классификация SRGz для всех оптических компаньонов точечных рентгеновских источников ePOZITA из рассматриваемой тестовой области. Набор рисунков демонстрирует результаты классификации в проекции на четыре фотометрических плоскости (четыре пары фотометрических признаков, наилучшим образом разделяющих классы рентгеновских объектов, см. выше). На каждом рисунке, левый столбец графиков — фотометрические квазары, средний — фотометрические галактики, правый столбец — фотометрические звезды.

Верхняя строчка графиков — точечные оптические компаньоны рентгеновских источников, нижняя строчка — протяженные оптические компаньоны. Сплошные контуры на графиках соответствуют 0.05, 0.35, 0.65, 0.95 квантилям распределений спектроскопических квазаров, галактик, астрометрических звезд (взяты с рис. 5–8). Штриховые контуры соответствуют 0.05, 0.35, 0.65, 0.95 квантилям распределений фотометрических объектов на графиках.

Различие сплошных и штриховых контуров на представленных графиках показывает рассогла-

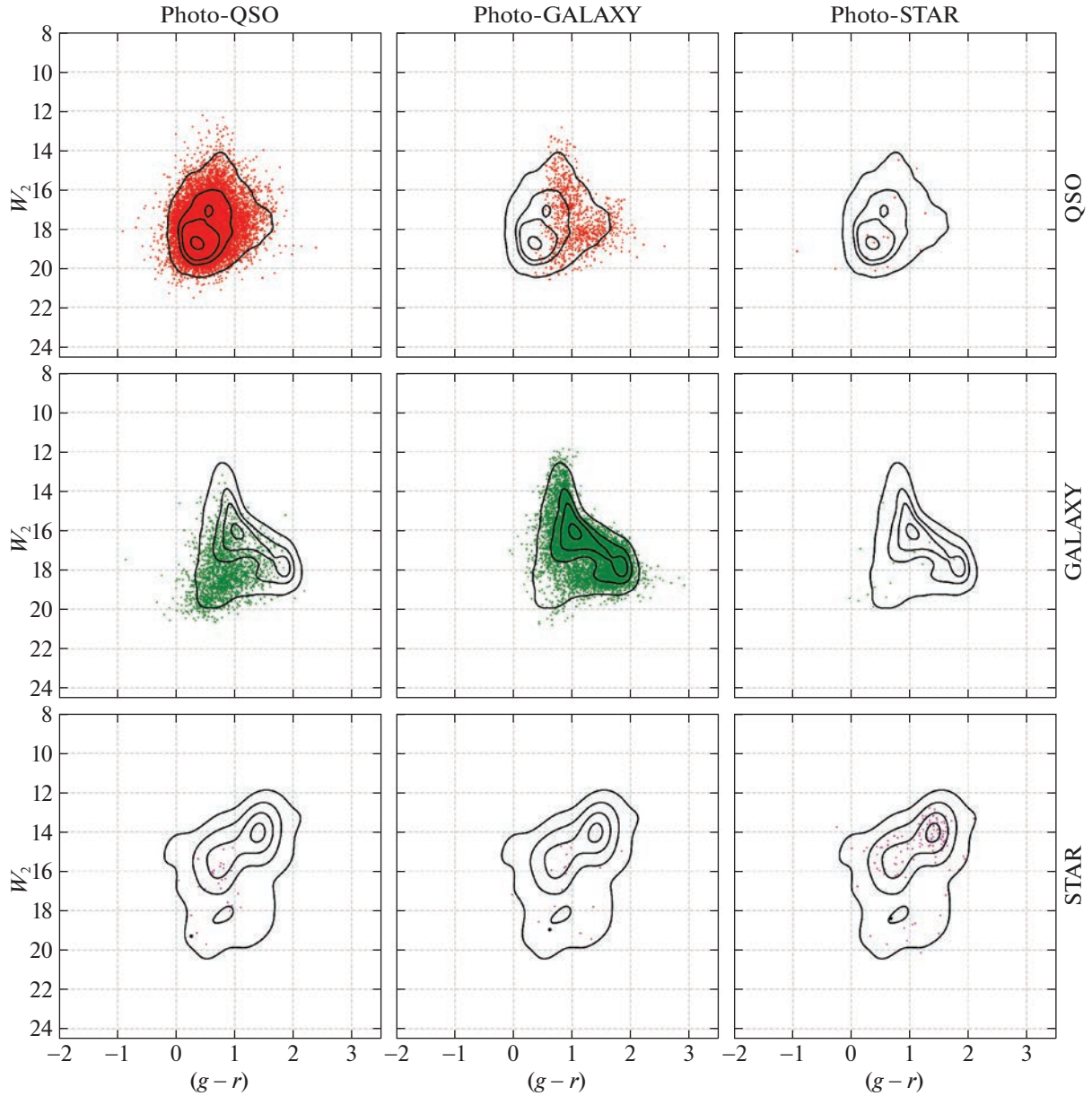


Рис. 17. Фотометрическая классификация SRGz протяженных оптических объектов ($type = REX, SER, DEV, EXP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цвет $(g-r)$ — величина W_2 (подробное описание см. рис. 12).

сование между спектральной/астрометрической (тестовой) и полной фотометрической выборками рентгеновских объектов в рассматриваемой области. В спектральных выборках отсутствует значительное число слабых $g > 22.5$ и $(g-r) > 1$ оптических объектов (квазаров и галактик), которые видны в полной фотометрической выборке. Мы также наблюдаем заметное различие сплошных и штриховых контуров на значениях показателя рентген/оптика $X/z > 2$ (на графиках фотометрических звезд — точечных оптических объектов). Также на верхней правой панели рис. 23 обращает

на себя внимание интересная деталь в области $(X/W_1)_{\text{salvato}} > 0$, $(z - W_2)_{\text{distance}} \sim 1$, которая, по-видимому, обусловлена рентгеновскими квазарами, ложно классифицированными SRGz как звезды. Аналогичным образом можно объяснить особенность на верхней правой панели рис. 20 в области цветовых расстояний $(g - W_1)_{\text{distance}} \sim 1-2$, и величин $g > 20$. В качестве практической рекомендации мы рекомендуем с осторожностью относиться к результатам классификации фотометрических объектов, существенно выходящих за пределы внешнего сплошного контура по величине

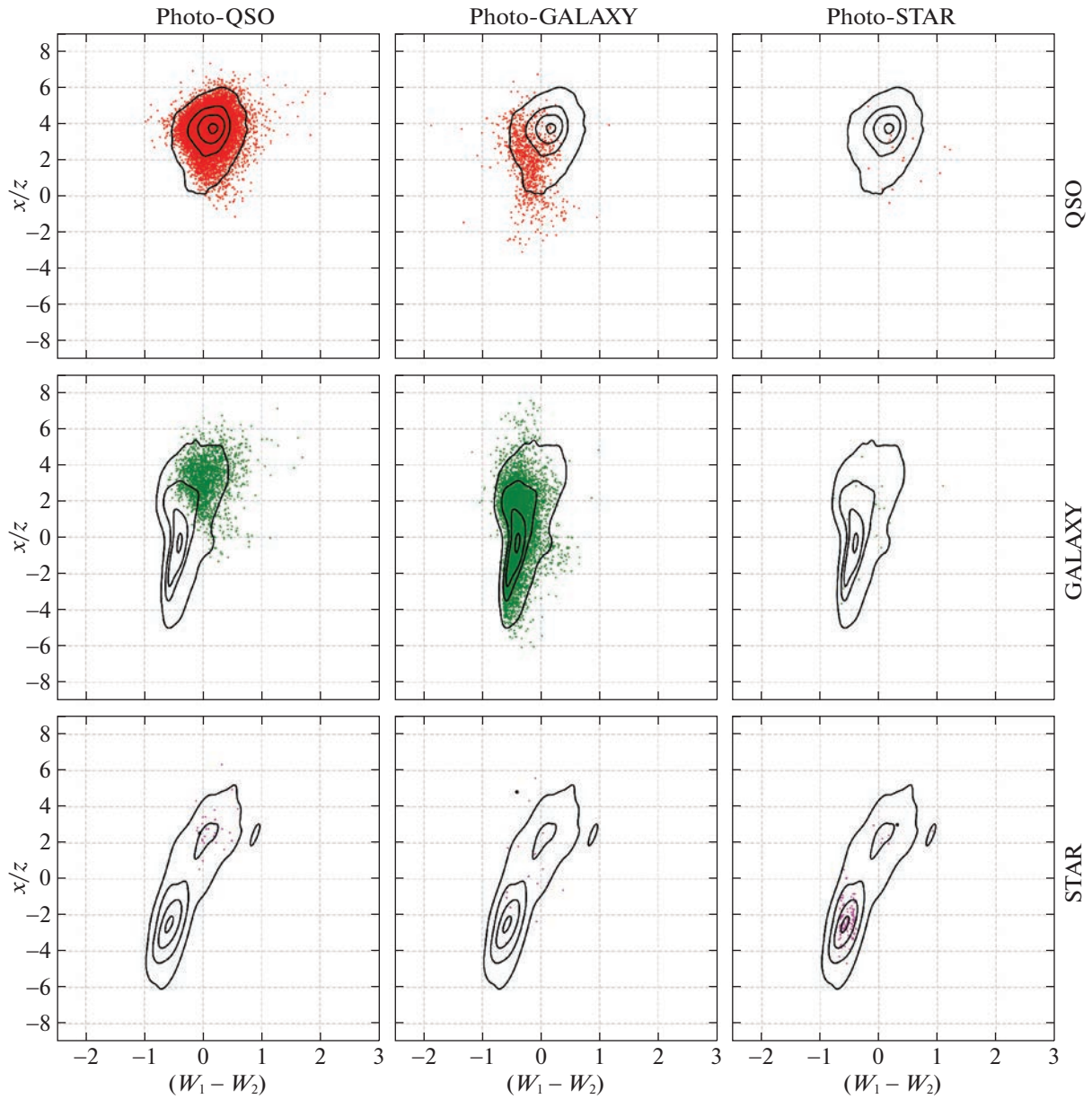


Рис. 18. Фотометрическая классификация SRGz протяженных оптических объектов ($type = REX, SER, DEV, EXP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора ePOZITA (в тестовой области SDSS), на плоскости цвет $(W_1 - W_2)$ — отношение рентген/оптика X/z (подробное описание см. рис. 12).

нам цветов, цветовых расстояний или отношений рентген/оптика(ИК).

Заметим, что указанные выше особенности классификации рентгеновских источников являются общими при работе с фотометрическими данными обзоров неба. Ценность фотометрической классификации — в ее массовости: многоволновые фотометрические данные доступны для всех рентгеновских источников ePOZITA, в то время как спектроскопическая/астрометрическая классификация в настоящее время ограничена выборкой

достаточно ярких оптических объектов. Как уже было сказано, в выбранной тестовой области SDSS $\approx 50\%$ оптических компаньонов точечных рентгеновских источников не имеют спектроскопической/астрометрической классификации. Для других областей обзора ePOZITA спектроскопическое покрытие заметно ниже (за исключением отдельных полей). Поэтому наши высокоточные модели фотометрической классификации SRGz, представляют большой интерес.

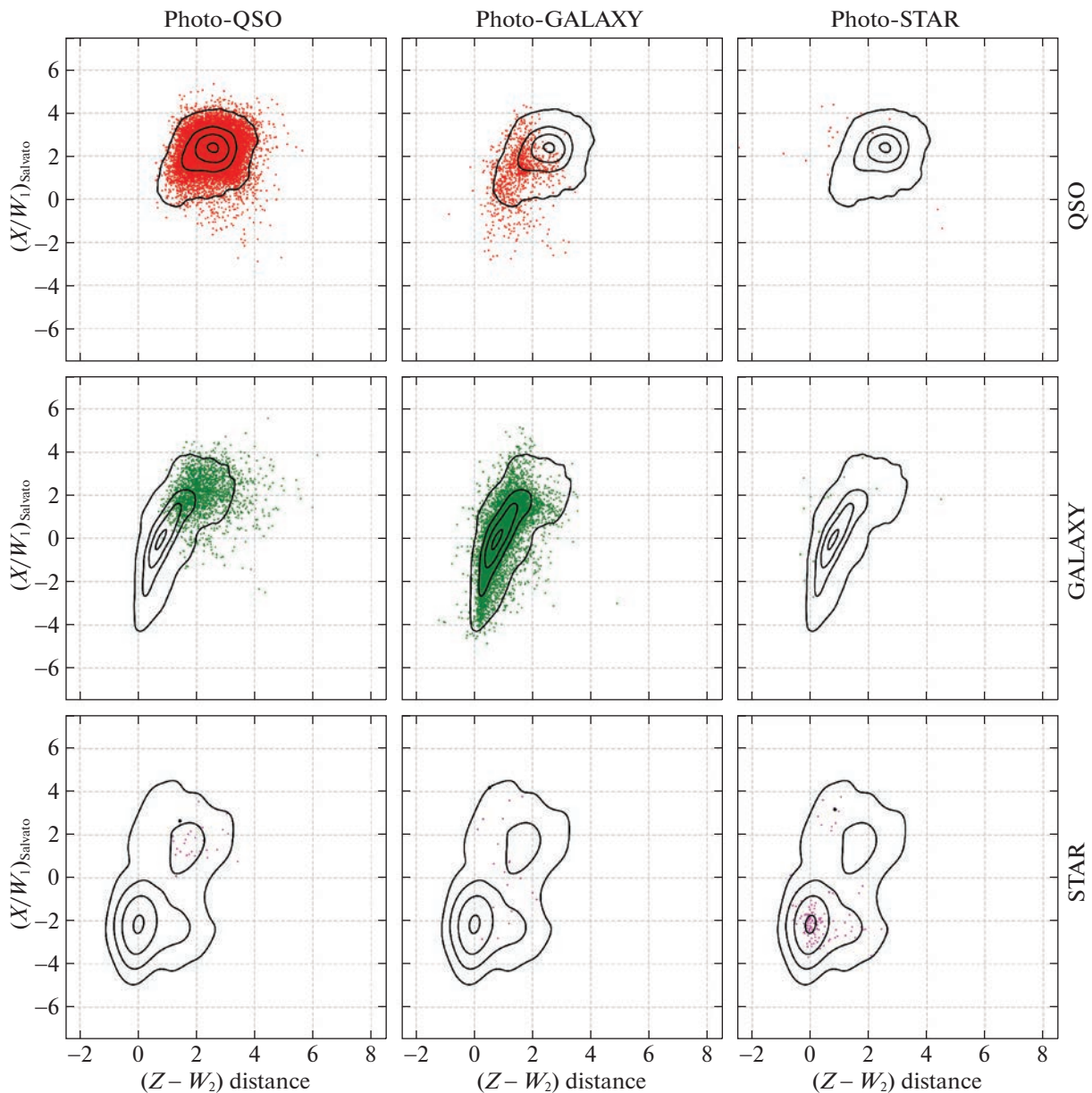


Рис. 19. Фотометрическая классификация SRGz протяженных оптических объектов ($type = REX, SER, DEV, EXP$) обзора DESI LIS, являющихся компаньонами точечных рентгеновских источников двухлетнего обзора еРОЗИТА (в тестовой области SDSS), на плоскости цветное расстояние ($z - W_2$) — отношение рентген/ИК $(X/W_1)_{\text{Salvato}}$. Левый столбец графиков — фотометрические звезды, средний — фотометрические галактики, правый столбец — фотометрические квазары (подробное описание см. рис. 12).

6.4. Отождествление, классификация и *photo-z* точечных рентгеновских источников еРОЗИТА в области Stripe82X

Рентгеновский обзор в области Stripe82X и составленный на его основе каталог оптических компаньонов рентгеновских источников, их фотометрических красных смещений, имеющихся для части объектов спектроскопических красных смещений (Ананна и др., 2017; ЛаМасса и др., 2016; Брейша и др., 2019) дают нам возможность комплексно

сравнить результаты SRGz для обзора еРОЗИТА и результаты отождествления, классификации и *photo-z*, полученные другими командами на одной и той же площадке неба, достаточно большой площади (31.3 кв. градусов).

Сначала мы сопоставили весь каталог Stripe82X (6181 рентгеновский источник Chandra и XMM-Newton) с каталогом точечных рентгеновских объектов еРОЗИТА (по данным двух лет обзора неба) в радиусе 30'' и получили 1936 объектов Stripe82X в окрестности источников еРОЗИТА.

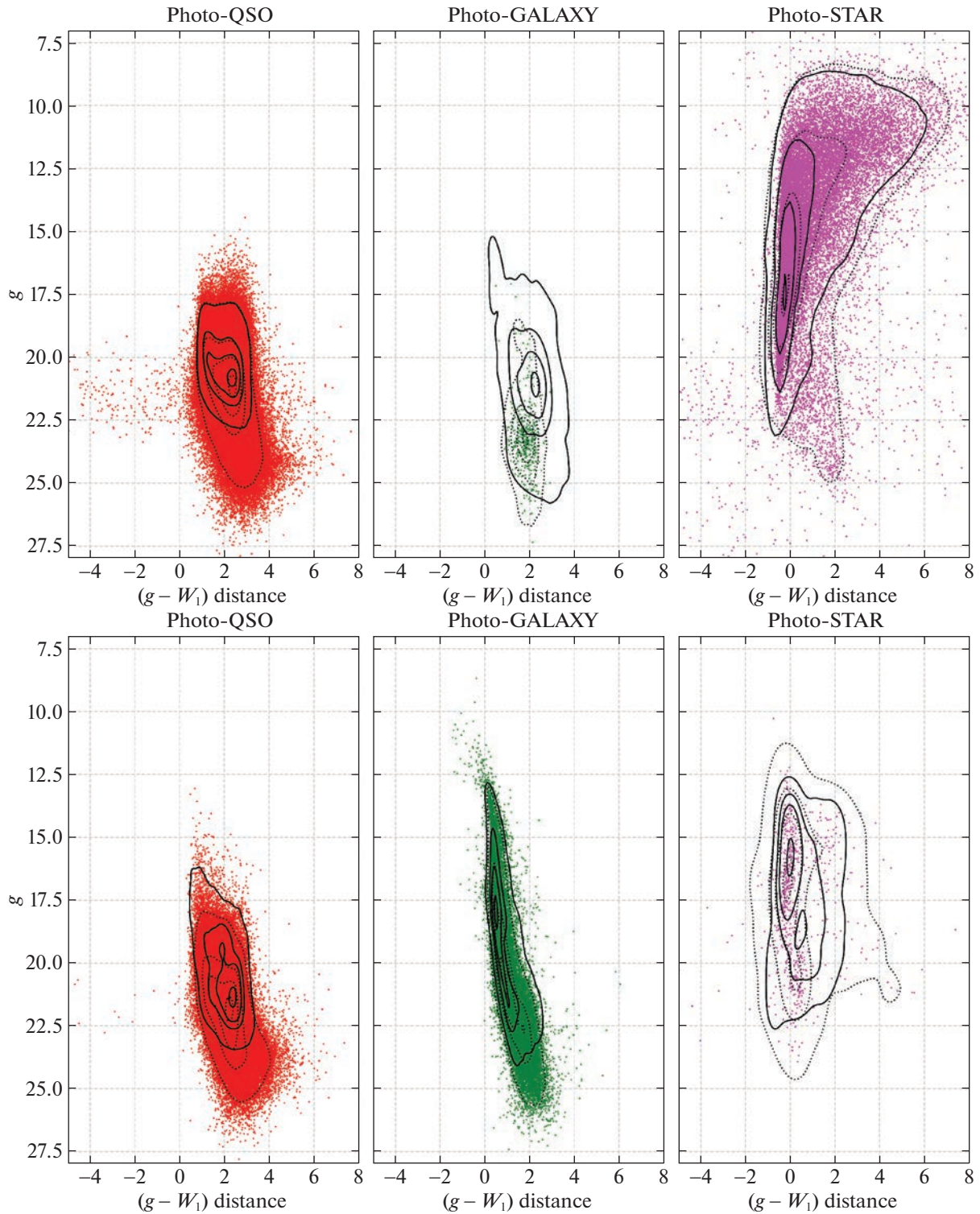


Рис. 20. Фотометрическая классификация SRGz оптических компаньонов точечных рентгеновских источников двухлетнего обзора ePOZITA ($srg_match_flag = 1$, $P_{empty} \leq 0.24$) на плоскости цветовое расстояние $(g - W_1)distance$ — величина g . Левый столбец графиков — фотометрические квазары, средний — фотометрические галактики, правый столбец — фотометрические звезды. Верхняя строка графиков — точечные ($type = PSF, DUP$) оптические компаньоны рентгеновских источников из обзора DESI LIS, нижняя строка — протяженные ($type = REX, SER, DEV, EXP$) оптические компаньоны. Сплошные контуры на графиках соответствуют 0.05, 0.35, 0.65, 0.95 квантилям распределений спектроскопических квазаров, галактик, астрометрических звезд (взяты с рис. 5–8). Штриховые контуры соответствуют 0.05, 0.35, 0.65, 0.95 квантилям распределений фотометрических объектов на графиках.

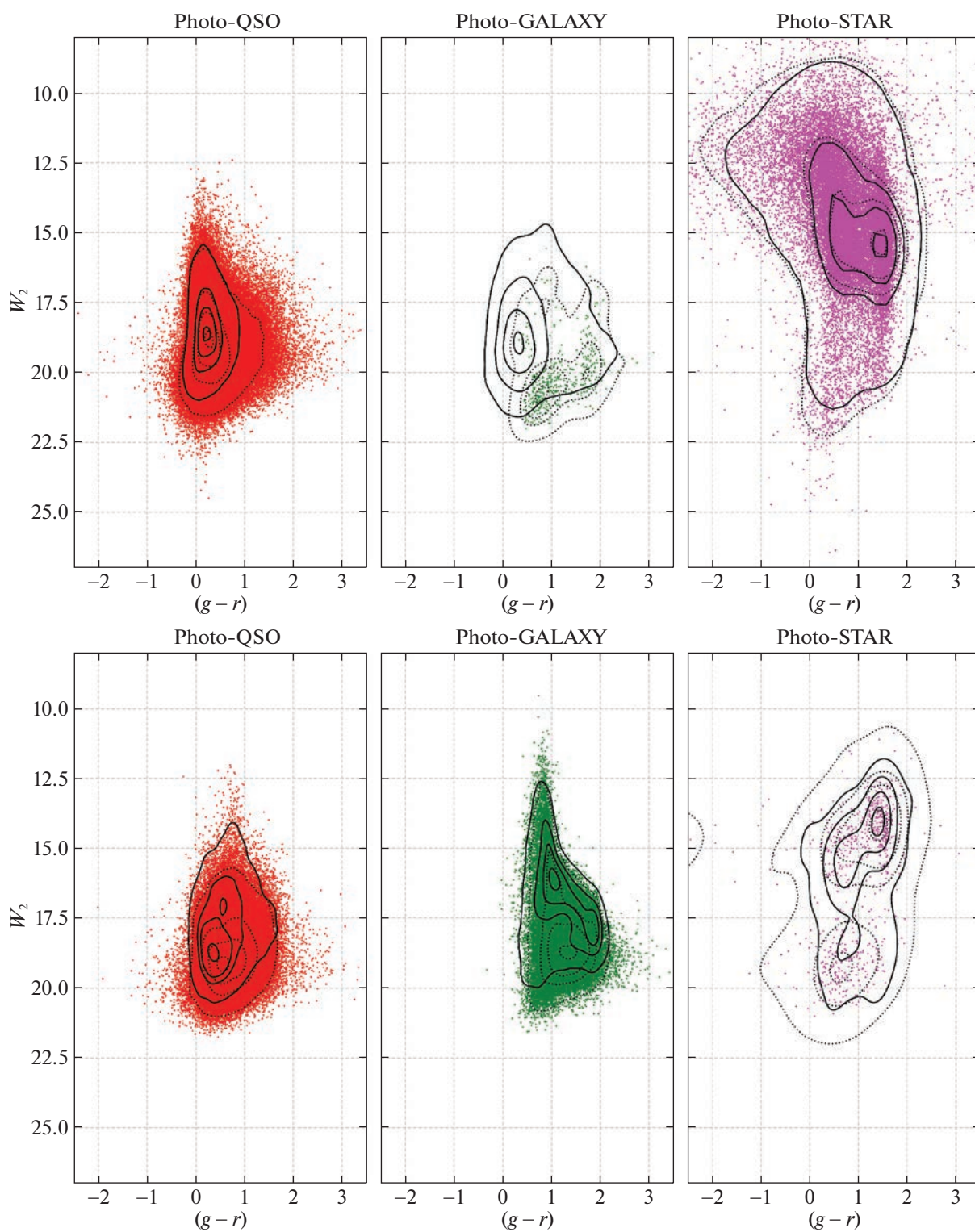


Рис. 21. Фотометрическая классификация SRGz оптических компаньонов точечных рентгеновских источников двухлетнего обзора eROSITA на плоскости $(g-r)$ — величина W_2 (подробнее см. подпись к рис. 20).

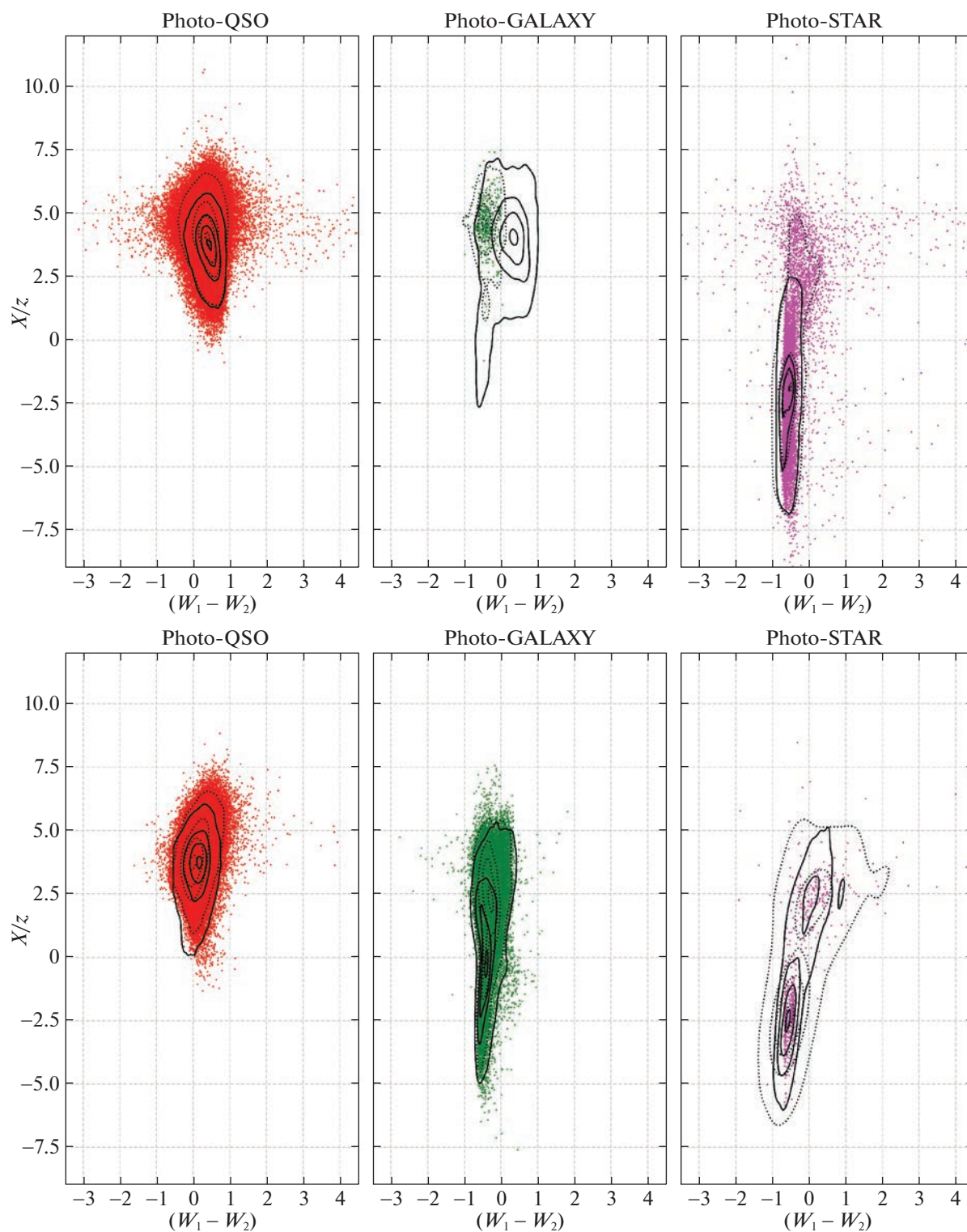


Рис. 22. Фотометрическая классификация SRGz оптических компаньонов точечных рентгеновских источников двух-летнего обзора eROSITA на плоскости цвет $(W_1 - W_2)$ — отношение рентген/оптика X/z (подробнее см. подпись к рис. 20).

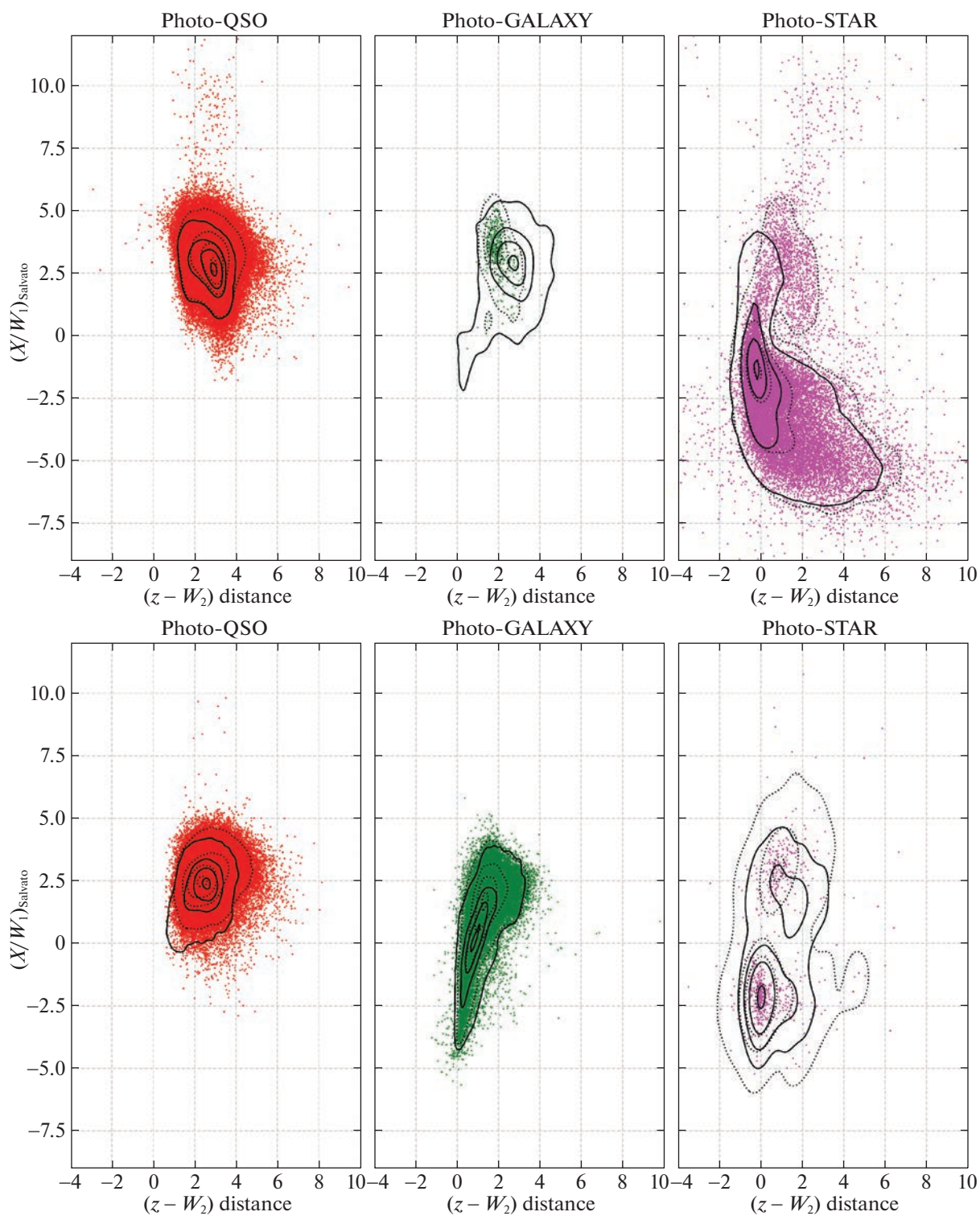


Рис. 23. Фотометрическая классификация SRGz оптических компаньонов точечных рентгеновских источников двухлетнего обзора eROZITA на плоскости цветное расстояние $(z - W_2)$ — отношение рентген/ИК $(X/W_1)_{\text{Salvato}}$ (подробнее см. подпись к рис. 20).

Таблица 8. Результаты оптического отождествления точечных источников ePOZITA в области Stripe82X

#	N_X	<i>Recall</i>
Все объекты S82X	6181	
Выборка S82X×eRASS:4 (30'')	1936	
- нет надежного оптического отождествления в S82X	24	
- оптическое отождествление в S82X ($0 < QF \leq 4$)	1912	
- оптическое отождествление в S82X ($0 < QF \leq 4$) × DESI LIS (1'')	1850	
- оптическое отождествление в S82X ($QF = 1$) × DESI LIS (1'')	1650	
Наиболее надежные ($QF = 1$) компаньоны S82X×eRASS:4 в DESI LIS (1650)		
Выборка $P_0 \leq 0.24$	1551	0.968
Выборка $P_0 \leq 0.24$ и <i>match_flag</i> =1,2	1508	0.941
Выборка $P_0 \leq 0.24$ и <i>match_flag</i> =1	1472	0.919
Оптические компаньоны ($QF > 0$) S82X×eRASS:4 в DESI LIS (1850)		
Выборка $P_0 \leq 0.24$	1785	0.965
Выборка $P_0 \leq 0.24$ и <i>match_flag</i> =1,2	1707	0.923
Выборка $P_0 \leq 0.24$ и <i>match_flag</i> =1	1649	0.891
- SRGz QSO	1426	
- SRGz GALAXY	118	
- SRGz STAR	105	
Оптические компаньоны S82X×eRASS:4 в DESI LIS с <i>zspec</i> > 0 (1384)		
Выборка $P_0 \leq 0.24$	1359	0.982
Выборка $P_0 \leq 0.24$ и <i>match_flag</i> =1,2	1338	0.967
Выборка $P_0 \leq 0.24$ и <i>match_flag</i> =1	1315	0.950

Из них 1912 источников имеют надежное оптическое отождествление в Stripe82X (с параметром качества отождествления в пределах $0 < QF \leq 4$, см. Ананна и др., 2017) и 24 не имеют оптического отождествления. Далее мы соотнесли 1912 оптических компаньонов Stripe82X с каталогом оптического обзора DESI LIS в кружке радиуса 1'' и получили выборку оптических компаньонов (1850 объектов).

Из 1850 источников выборки оптических компаньонов 1649 объектов (89%) отбираются SRGz в соответствии с оптимальными критериями отбора ($P_0 \leq 0.24$ и *srg_match_flag*=1, см. п. 6.2). Отметим, что доля компаньонов, ложно отброшенная нами как “бездомные” при использовании фильтра $P_0 \leq 0.24$, составила $\approx 4\%$ (см. табл. 8).

Доля отброшенных компаньонов хорошо согласуются со сделанными ранее оценками по тестовой области SDSS (5% потерь при отборе $P_0 \leq 0.24$ и *srg_match_flag*=1, см. п. 6.2). Суммарные потери оптических компаньонов на Stripe82X при отборе ($P_0 \leq 0.24$ и *srg_match_flag*=1) составили $\approx 11\%$, что в 2 раза больше, чем ожидаемые по тестовой области SDSS $\approx 5\%$). Различие в суммарных потерях оптических компаньонов может быть обусловлено возможными неточностями при отборе компаньонов в Stripe82X по критерию $0 < QF \leq 4$ (Ананна и др., 2017). При отборе оптических компаньонов по более строгому критерию $QF = 1$, 1650 объектов имеют единственный оптический источник DESI LIS в кружке радиуса 1''; потери оптических компаньонов при отборе ($P_0 \leq 0.24$ и

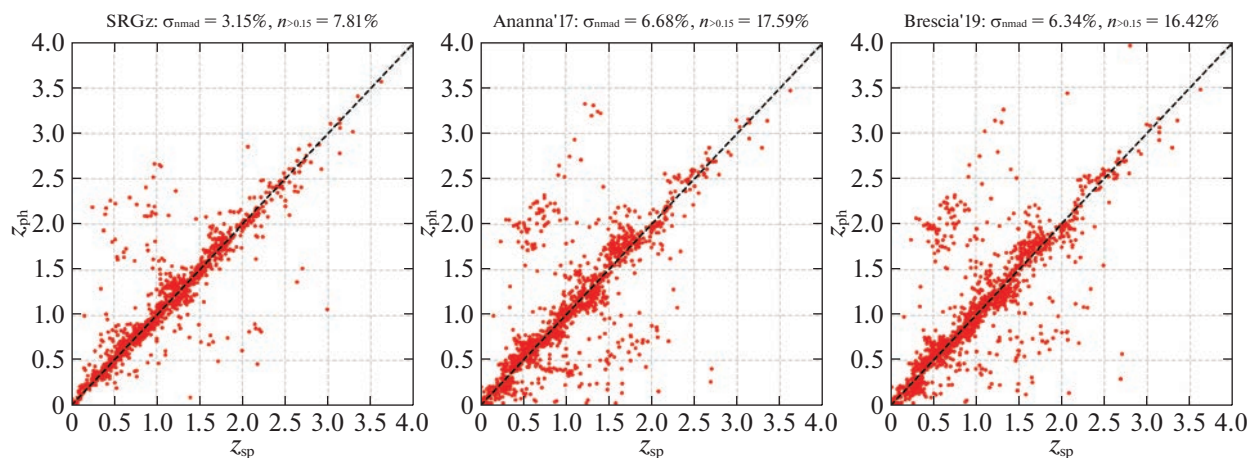


Рис. 24. Точность фотометрических красных смещений оптических компаньонов точечных рентгеновских источников еРОЗИТА в области Stripe82X разными методами: слева — SRGz, в центре — Ананна и др. (2017), справа — Брешиа и др. (2019).

$\text{srg_match_flag}=1$) сокращаются до $\approx 8\%$, что уже довольно близко к оценке 5%, полученной нами на тестовой выборке SDSS (см. п. 6.2).

На выборке 1850 оптических компаньонов точечных рентгеновских источников еРОЗИТА в поле Stripe82X, 58 объектов идентифицированы как звезды ($\text{Morph}=1$) на основе аппроксимации фотометрическими шаблонами Ананна и др. (2017), а 21 является спектроскопической звездой ($\text{SpClass}=\text{STAR}$). Мы сравнили результаты классификации звезд из статьи (Ананна и др., 2017) с фотометрической классификацией SRGz для указанных оптических источников и получили хорошее согласие: модель SRGz распознает как фотометрические звезды 54 из 58 объектов с $\text{Morph}=1$ и 20 из 21 спектроскопической звезды.

Из 1850 источников выборки оптических компаньонов, 1384 объекта (75%) имеют спектроскопические измерения, при этом не являются спектроскопическими звездами и имеют измерение спектрального красного смещения в каталоге Stripe82X (см. табл. 8). 1315 (95%) из 1384 источников отбираются SRGz при использовании оптимальных параметров оптического отождествления ($P_0 \leq 0.24$, $\text{srg_match_flag}=1$). Далее мы сравним точность фотометрических красных смещений SRGz для внегалактических рентгеновских источников еРОЗИТА в поле Stripe82X, имеющих спектроскопические измерения красного смещения.

В каталоге Stripe82X собраны измерения фотометрических красных смещений, выполненные двумя подходами: (i) методом шаблонов (Ананна и др., 2017), (ii) с применением искусственных нейронных сетей (Брешиа и др., 2019). Для 1370 (из 1384) оптических компаньонов рентгеновских

источников в каталоге Stripe82X имеются спектральные измерения красного смещения, измерения photo-z обоими методами, указанными выше, а также измерение photo-z моделью SRGz. Мы будем использовать эту выборку рентгеновских источников в области Stripe82X для анализа точности фотометрических красных смещений SRGz в обзоре неба еРОЗИТА в сравнении с результатами других групп.

На рис. 24 показаны диаграммы рассеяния (по оси абсцисс — спектральное красное смещение, по оси ординат — фотометрическое красное смещение) для 1370 оптических компаньонов точечных рентгеновских источников еРОЗИТА в поле Stripe82X. На панели слева показаны фотометрические красные смещения SRGz, посередине — модель шаблонов из работы (Ананна и др., 2017), правая панель — машинное обучение на базе нейросетей (Брешиа и др., 2019). В заголовке графиков приведены стандартные метрики оценки качества photo-z квазаров: дисперсия прогнозов по величине колокола распределения σ_{NMAD} и доля выбросов $n_{>0.15}$ (см. п. 5.3.2).

Как можно видеть, SRGz существенно превосходит другие модели photo-z рентгеновских источников, представленные в литературе для области Stripe82X. На выборке оптических компаньонов точечных рентгеновских источников еРОЗИТА измерения фотометрических красных смещений системой SRGz более чем в 2 раза лучше измерений photo-z, полученных другими группами (Ананна и др., 2017; Брешиа и др., 2019). Можно отметить высокую точность измерений фотометрических красных смещений SRGz как с точки зрения

дисперсии прогноза красного смещения ($\sigma_{NMAD} \approx 3.1\%$), так и по величине выбросов ($n_{>0.15} \approx 7.8\%$). Качество моделей SRGz обусловлено, в значительной степени, большой тренировочной выборкой квазаров и большим разнообразием используемых в SRGz фотометрических данных (см. также сравнение точности photo-z рентгеновских квазаров для разного набора данных фотометрических обзоров неба в работах (Мещеряков и др., 2018; Борисов и др., 2022)).

Следует отметить, что представленный здесь результат по точности photo-z в поле Stripe82X можно считать характерным для всех рентгеновских источников в области пересечения обзоров DESI LIS, SDSS, Pan-STARRS1. Точность photo-z в поле Stripe82X оценивалась для 3/4 оптических компаньонов рентгеновских источников ePOZITA. Для оставшейся 1/4 объектов, как можно ожидать, точность измерений photo-z будет снижаться при рассмотрении более слабых оптических объектов.

7. ВЫВОДЫ

В работе описаны методы системы SRGz для физического отождествления точечных рентгеновских источников обзора ePOZITA по фотометрическим данным в области покрытия обзора DESI Legacy Imaging Surveys DR9. Рассмотрены модели машинного обучения, входящие в состав системы SRGz (версия 2.1), позволившие для 87% площади всей внегалактической области ($0 < l < 180^\circ, |b| > 20^\circ$) получить надежные измерения космологического красного смещения и физического класса рентгеновских объектов (звезда/квазар/галактика), опираясь на многоволновые фотометрические данные обзоров неба DESI LIS, SDSS, Pan-STARRS, WISE, ePOZITA.

Уникальной особенностью системы SRGz является то, что ее модель работы с данными (алгоритмы отождествления, классификации, photo-z) целиком опирается на эвристические подходы машинного обучения, что позволяет значительно увеличить точность измерения физических характеристик рентгеновских объектов по фотометрическим данным.

Для стандартного выбора параметров SRGz, полнота отождествления оптических компаньонов в области DESI LIS составляет 95% (точность отбора оптических компаньонов 94%), полнота 82% (точность 85%) классификации рентгеновских источников, не имеющих оптического партнера (“бездомных”, англ. *hostless*). В SRGz достигается высокая полнота фотометрической классификации оптических компаньонов рентгеновских источников (>99% для внегалактических объектов: квазар

или галактика) на тестовой выборке объектов со спектрами SDSS и звезд GAIA.

В статье выполнен анализ важности фотометрических признаков для отождествления оптических компаньонов рентгеновских источников ePOZITA, а также разделения на классы рентгеновских квазаров галактик и звезд. Показано, что отношения рентген/оптика-ИК, введенные нами признаки цветовых расстояний, а также оптические цвета (например, $(g - r)$), цвет ($W_1 - W_2$) и ИК-величина W_2 играют большую роль при разделении на классы рентгеновских объектов.

Точность фотометрических красных смещений SRGz тестировалась на выборке 3/4 всех оптических компаньонов (для которых доступны спектроскопические измерения) рентгеновских источников ePOZITA в поле Stripe82X. Точность photo-z рентгеновских источников составила: $\sigma_{NMAD} = 3.1\%$ (нормализованное медианное абсолютное отклонение прогноза) и $n_{>0.15} = 7.8\%$ (доля катастрофических выбросов). Представленные результаты photo-z для источников ePOZITA в поле Stripe82X более чем в 2 раза (по обоим метрикам — σ_{NMAD} и $n_{>0.15}$) превосходят результаты других групп (Ананна и др., 2017; Брешиа и др., 2019), опубликованные в каталоге Stripe82X.

В заключение отметим, что в соответствии с политикой Российского консорциума телескопа СРГ/ePOZITA публикация данных каталога SRGz-CatD в открытый доступ будет происходить в процессе публикации каталогов рентгеновских источников телескопа ePOZITA для различных участков неба, классов объектов и т.д.

В наших планах — дальнейшее развитие системы SRGz в интересах науки рентгеновского обзора неба СРГ/ePOZITA, которое будет идти в нескольких направлениях.

Мы исследуем ряд нейросетевых алгоритмов для применения их в моделях классификации и вероятностной регрессии на табличных данных фотометрических каталогов. Как следует из обзора Борисова и др. (2021), нейросетевые модели на гетерогенных табличных данных светлую приблизились по точности классификации к древовидным ансамблям градиентного бустинга. В первую очередь это связано с совершенствованием моделей глубокого обучения на основе алгоритмов нейросетей-трансформеров (англ. *transformer neural architecture*) и механизма внимания (см. Сампалли и др., 2021). Включение подобных моделей в систему SRGz представляет большой интерес и потенциально позволит дополнительно улучшить измерения физических характеристик рентгеновских объектов.

Также в рамках системы SRGz в группе по рентгеновскому каталогу нами развивается инфраструктура для хранения и доступа к изображениям открытых фотометрических обзоров с последующим их анализом с помощью нейросетей. Следующим шагом развития SRGz в этом направлении являются нейросетевые модели отождествления, классификации и photo-z, работающие непосредственно с сырыми данными обзоров неба в рентгеновском, оптическом, ИК и других диапазонах спектра. Это позволит более полно использовать потенциал нейросетевых подходов, заключающийся в возможности автоматического извлечения сжатых векторных представлений из сырых данных обзоров неба (в первую очередь, в виде массивов изображений, таблиц).

Это исследование основано на наблюдениях телескопа ePOZITA на борту обсерватории СРГ. Обсерватория СРГ изготовлена Роскосмосом в интересах Российской академии наук в лице Института космических исследований (ИКИ) в рамках Российской федеральной научной программы с участием Германского центра авиации и космонавтики (DLR). Рентгеновский телескоп СРГ изготовлен консорциумом германских институтов во главе с Институтом внеземной астрофизики Общества им. Макса Планка (МРЕ) при поддержке DLR. Космический аппарат СРГ спроектирован, изготовлен, запущен и управляется НПО им. Лавочкина и его субподрядчиками. Прием научных данных осуществляется комплексом антенн дальней космической связи в Медвежьих озерах, Уссурийске и Байконуре и финансируется Роскосмосом. Использованные в настоящей работе данные телескопа eРозита обработаны с помощью программного обеспечения eSASS, разработанного германским консорциумом ePOZITA, и программного обеспечения, разработанного российским консорциумом телескопа СРГ/ePOZITA. Система SRGz создана в научной рабочей группе по поиску рентгеновских источников, их отождествлению и составлению каталога по данным телескопа СРГ/ePOZITA в отделе астрофизики высоких энергий ИКИ РАН.

Работа выполнена при поддержке гранта РНФ 21-12-00343. Работа И.Ф. Бикмаева и И.М. Хамитова частично поддержана субсидией Минобрнауки РФ FZSM-2023-0015, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности.

СПИСОК ЛИТЕРАТУРЫ

1. Абдуллоуф и др. (Abdurro'uf, K. Accetta, C. Aerts, V. Silva Aguirre, R. Ahumada, N. Ajgaonkar, et al.), *Astrophys. J. Suppl. Ser.* **259**, 35 (2022).
2. Аболфати и др. (B. Abolfathi, D.S. Aguado, G. Aguilar, C. Allende Prieto, A. Almeida, T.T. Ananna, et al.), *Astrophys. J. Suppl. Ser.* **235**, 42 (2018).
3. Ананна и др. (T.T. Ananna, M. Salvato, S. LaMassa, C.M. Urry, N. Cappelluti, C. Cardamone, et al.), *Astrophys. J.* **850**, 66 (2017).
4. Арик, Пфистер (S.O. Arik and T. Pfister), arXiv e-prints, p. arXiv:1908.07442 (2019).
5. Баштанник, Гайндман (D. Bashtannyk and R. Hyndman), *Comput. Statist. and Data Analys.* **36**, 279 (2001).
6. Бельведерский и др. (M.I. Belvedersky, A.V. Meshcheryakov, P.S. Medvedev, and M.R. Gilfanov), *Astron. Lett.* **48**, 109 (2022).
7. Бикмаев и др. (I.F. Bikmaev, E.N. Irtuganov, E.A. Nikolaeva, N.A. Sakhibullin, R.I. Gumerov, A.S. Sklyanov, et al.), *Astron. Lett.* **46**, 645 (2020).
8. Бикмаев и др. (I.F. Bikmaev, E.N. Irtuganov, E.A. Nikolaeva, N.A. Sakhibullin, R.I. Gumerov, A.S. Sklyanov, et al.), *Astron. Lett.* **47**, 277 (2021).
9. Бишоп (C.M. Bishop), *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 1st Ed. (2007).
10. Болтон и др. (A.S. Bolton, D.J. Schlegel, É. Aubourg, S. Bailey, V. Bhardwaj, J.R. Brownstein, et al.), *Astron. J.* **144**, 144 (2012).
11. Борисов и др. (V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci), arXiv e-prints, p. arXiv:2110.01889 (2021).
12. Борисов и др. (V. Borisov, A. Meshcheryakov, and S. Gerasimov), *Astron. Soc. Pacific Conf. Ser. V. 532 of Astron. Soc. Pacific Conf. Ser.*, p. 231 (2022).
13. Брейман (L. Breiman), *Machine Learn.* **45**, 5 (2001).
14. Брешиа и др. (M. Brescia, S. Caviuoti, and G. Longo), *MNRAS* **450**, 3893 (2015).
15. Брешиа и др. (M. Brescia, M. Salvato, S. Caviuoti, T.T. Ananna, G. Riccio, S.M. LaMassa, et al.), *MNRAS* **489**, 663 (2019).
16. Буренин (R.A. Burenin), *Astron. Lett.* **48**, 153 (2022).
17. Быков и др. (S.D. Bykov, M.I. Belvedersky, and M.R. Gilfanov), arXiv e-prints, p. arXiv:2302.13689 (2023).
18. Вебб и др. (N.A. Webb, M. Coriat, I. Traulsen, J. Ballet, C. Motch, F.J. Carrera, et al.), *Astron. Astrophys.* **641**, A136 (2020).
19. Гильфанов М.Р. и др., Письма в Астрон. журн., в печати (2023).
20. Гуо и др. (C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger), arXiv e-prints, p. arXiv:1706.04599 (2017).
21. Дей и др. (A. Dey, D.J. Schlegel, D. Lang, R. Blum, K. Burleigh, X. Fan, et al.), *Astron. J.* **157**, 168 (2019).
22. Додин и др. (A.V. Dodin, S.A. Potanin, N.I. Shatsky, A.A. Belinski, K.E. Atapin, M.A. Burlak, et al.), *Astron. Lett.* **46**, 429 (2020).
23. Додин и др. (A.V. Dodin, N.I. Shatsky, A.A. Belinski, K.E. Atapin, M.A. Burlak, S.G. Zheltoukhov, et al.), *Astron. Lett.* **47**, 661 (2021).

24. Дрю и др. (J.E. Drew, E. Gonzalez-Solares, R. Greimel, M.J. Irwin, A. Küpcü Yoldas, J. Lewis, et al.), *MNRAS* **440**, 2036 (2014).
25. Избицки, Ли (R. Izbicki and A.B. Lee), arXiv e-prints, p. arXiv:1704.08095 (2017).
26. Йорк и др. (D.G. York, J. Adelman, J. Anderson, S.F. Anderson, J. Annis, N.A. Bahcall, et al.), *Astron. J.* **120**, 1579 (2000).
27. Кавуоти и др. (S. Caviuoti, V. Amaro, M. Brescia, C. Vellucci, C. Tortora, and G. Longo), *MNRAS* **465**, 1959 (2017).
28. Карраско Кинд, Бруннер (M. Carrasco Rind and R.J. Brunner), *MNRAS* **432**, 1483 (2013).
29. Ке и др. (G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al.), in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 3149–3157, Curran Associates Inc., Red Hook, NY, USA (2017).
30. Коллаборация GAIA и др. (Gaia Collaboration, G.A. Brown, A. Vallenari, T. Prusti, J.H.J. de Bruijne, C. Babusiaux, et al.), *Astron. Astrophys.* **616**, A1 (2018).
31. Ла Масса и др. (S.M. LaMassa, C.M. Urry, N. Cappelluti, H. Böhringer, A. Comastri, E. Glikman, et al.), *Astrophys. J.* **817**, 172 (2016).
32. Ли и др. (C. Li, Y. Zhang, C. Cui, D. Fan, Y. Zhao, X.-B. Wu, et al.), *MNRAS* **518**, 513 (2023).
33. Лин, Джеон (Y. Lin and Y. Jeon), *J. Am. Statistic. Associat.* **101**, 578 (2006).
34. Ломотей, Дитерс (R. Lomotey and R. Deters), Proc. – IEEE 8th Inter. Symposium on Service Oriented System Engineering, SOSE 2014, pp. 181–191, IEEE Computer Society, United States, 8th IEEE Inter. Symp. on Service Oriented System Engineering, SOSE 2014 ; Conference date: 07-04-2014 Through 11-04-2014 (2014).
35. Лундберг, Ли (S.M. Lundberg and S.-I. Lee), in I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems* 30, pp. 4765–4774, Curran Associates, Inc. (2017).
36. Люк и др. (B.W. Lyke, A.N. Higley, J.N. McLane, D.P. Schurhammer, A.D. Myers, A. J. Ross, et al.), *Astrophys. J. Suppl. Ser.* **250**, 8 (2020).
37. Майнсхаузен (N. Meinshausen), *J. Mach. Learn. Res.* **7**, 983 (2006).
38. Маккаро и др. (T. Maccacaro, I.M. Gioia, A. Wolter, G. Zamorani, and J.T. Stocke), *Astrophys. J.* **326**, 680 (1988).
39. Медведев П.С. и др., Письма в Астрон. журн., в печати (2022).
40. Мейнзер и др. (A. Mainzer, J. Bauer, T. Grav, J. Masiero, R. M. Cutri, J. Dailey, et al.), *Astrophys. J.* **731**, 53 (2011).
41. Мензель и др. (M.L. Menzel, A. Merloni, A. Georgakakis, M. Salvato, E. Aubourg, W.N. Brandt, et al.), *MNRAS* **457**, 110 (2016).
42. Мещеряков и др. (A.V. Meshcheryakov, V.V. Glazkova, S.V. Gerasimov, and I.V. Mashechkin), *Astron. Lett.* **44**, 735 (2018).
43. Ньюман, Грюен (J. Newman), *Ann. Rev. Astron. and Astrophys.* **60**, 363 (2022).
44. Павлинский и др. (M. Pavlinsky, A. Tkachenko, V. Levin, N. Alexandrovich, V. Arefiev, V. Babyshkin, et al.), *Astron. Astrophys.* **650**, A42 (2021).
45. Парис и др. (I. Pâris, P. Petitjean, É. Aubourg, A.D. Myers, A. Streblyanska, B.W. Lyke, et al.), *Astron. Astrophys.* **613**, A51 (2018).
46. Педрегоса и др. (F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al.), *J. Mach. Learn. Res.* **12**, 2825 (2011).
47. Предель и др. (P. Predehl, R. Andritschke, V. Arefiev, V. Babyshkin, O. Batanov, W. Becker, et al.), *Astron. Astrophys.* **647**, A1 (2021).
48. Райт и др. (E.L. Wright, P.R.M. Eisenhardt, A.K. Mainzer, M.E. Ressler, R.M. Cutri, T. Jarrett, et al.), *Astron. J.* **140**, 1868 (2010).
49. Росс, Кросс (N.P. Ross and N.J.G. Cross), *MNRAS* **494**, 789 (2020).
50. Руиз и др. (A. Ruiz, A. Corral, G. Mountrichas, and I. Georgantopoulos), *Astron. Astrophys.* **618**, A52 (2018).
51. Садех и др. (I. Sadeh, F.B. Abdalla, and O. Lahav), *PASP* **128**, 104502 (2016).
52. Сальвато и др. (M. Salvato, J. Buchner, T. Budavári, T. Dwelly, A. Merloni, M. Brusa, et al.), *MNRAS* **473**, 4937 (2018).
53. Сальвато и др. (M. Salvato, J. Wolf, T. Dwelly, A. Georgakakis, M. Brusa, A. Merloni, et al.), *Astron. Astrophys.*, **661**, A3 (2022).
54. Сампалли и др. (G. Somepalli, M. Goldblum, A. Schwarzschild, C. Bayan Bruss, and T. Goldstein), arXiv e-prints, p. arXiv:2106.01342 (2021).
55. Стоутон и др. (C. Stoughton, R.H. Lupton, M. Bernardi, M.R. Blanton, S. Burles, F.J. Castander, et al.), *Astron. J.* **123**, 485 (2002).
56. Сюняев и др. (R. Sunyaev, V. Arefiev, V. Babyshkin, A. Bogomolov, K. Borisov, M. Buntov, et al.), *Astron. Astrophys.* **656**, A132 (2021).
57. Танг и др. (C. Tang, D. Garreau, and U. von Luxburg), in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc. (2018).
58. Тэаро, Мещеряков, in preprint (2023).
59. Фосетт (T. Fawcett), *Pattern Recognition Letters*, **27**, 861, *ROC Analysis in Pattern Recognition* (2006).
60. Фридман (J.H. Friedman), *Annals of statistics*, p. 1189–1232 (2001).
61. Ханлей, МакНейл (J.A. Hanley and B.J. McNeil), *Radiology* **143**, 29 (1982).
62. Хорунжев и др. (G.A. Khorunzhev, A.V. Meshcheryakov, R.A. Burenin, A.R. Lyapin, P.S. Medvedev, S.Y. Sazonov, et al.), *Astron. Lett.* **46**, 149 (2020).

63. Хорунжев и др. (G.A. Khorunzhev, A.V. Meshcheryakov, P.S. Medvedev, V.D. Borisov, R.A. Burenin, R.A. Krivonos, et al.), *Astron. Lett.* **47**, 123 (2021).
64. Хорунжев и др. (G.A. Khorunzhev, S.N. Dodonov, A.V. Meshcheryakov, A.V. Moiseev, A. Grokhovskaya, S.S. Kotov, et al.), *Astron. Lett.* **48**, 69 (2022).
65. Чамберс и др. (K.C. Chambers, E.A. Magnier, N. Metcalfe, H.A. Flewelling, M.E. Huber, C.Z. Waters, et al.), *arXiv e-prints*, p. arXiv:1612.05560 (2016).
66. Шмидт и др. (S.J. Schmidt, A.I. Malz, J.Y.H. Soo, I.A. Almosallam, M. Brescia, S. Caviuoti, et al.), *MNRAS* **499**, 1587 (2020).
67. Эванс и др. (I.N. Evans, F.A. Primini, K.J. Glotfelty, C.S. Anderson, N.R. Bonaventura, J.C. Chen, et al.), *Astrophys. J. Suppl. Ser.* **189**, 37 (2010).
68. Эванс и др. (P.A. Evans, K.L. Page, J.P. Osborne, A.P. Beardmore, R. Willingale, D.N. Burrows, et al.), *Astrophys. J. Suppl. Ser.* **247**, 54 (2020).