АСТРОНОМИЧЕСКИЙ ЖУРНАЛ, 2022, том 99, № 11, с. 1040–1057

О НЕКОТОРЫХ ВОПРОСАХ КРОСС-ИДЕНТИФИКАЦИИ АСТРОНОМИЧЕСКИХ КАТАЛОГОВ

© 2022 г. Д. А. Ладейщиков[1,*], А. М. Соболев[1,]**

[1] *Коуровская астрономическая обсерватория, Уральский федеральный университет, Екатеринбург, Россия*

**E-mail: dmitry.ladeyschikov@urfu.ru*

***E-mail: andrej.sobolev@urfu.ru*

Поступила в редакцию 12.09.2022 г.
После доработки 30.09.2022 г.
Принята к публикации 30.09.2022 г.

В работе рассматриваются вопросы кросс-идентификации источников из различных астрономических каталогов. Одна из главных рассматриваемых проблем – как кросс-идентифицировать большое количество каталогов в выбранной области на небе, когда нет опорного каталога источников? Для кросс-идентификации больших объемов данных предлагается использовать алгоритм поиска групп DBSCAN. Разработан специальный программный код cross-match.online, работающий в режимах онлайн и оффлайн, который позволяет автоматизировать процесс кросс-идентификации источников по множеству каталогов. В работе рассматриваются вопросы сравнения каталогов с различной плотностью источников, разрешения неопределенностей при кросс-идентификации, а также учета неопределенности положений источников. Предложена методика, позволяющая кросс-идентифицировать каталоги с разной плотностью источников и различным значением неопределенности положений. Открытый доступ к системе предоставлен по адресу https://cross-match.online.

Ключевые слова: кросс-идентификация больших объемов данных, сравнение каталогов с различной плотностью источников, статистика данных

DOI: 10.31857/S0004629922110111

1. ВВЕДЕНИЕ

Кросс-идентификация источников из различных астрономических каталогов является в настоящее время актуальной задачей, так как наиболее полное исследование астрофизических объектов зачастую возможно только при использовании большого количества архивных данных, полученных на разных инструментах. Кросс-идентификация также является важным инструментом для решения задач виртуальной обсерватории [1]. В последнее время в открытом доступе появились результаты крупномасштабных обзоров неба в широком диапазоне длин волн: обзоры Gaia [2], Pan-STARRS [3], UKIRT Hemisphere Survey [4], SDSS [5], LSST [6], TESS [7], Hi-GAL 360 [8] и другие. Объемы данных по этим обзорам превышают сотни терабайт, поэтому хранение локальных копий этих каталогов затруднено. Частично обзоры размещены в едином Страсбургском центре данных (CDS) VizieR [9], но иногда там размещаются не самые последние версии каталогов, а некоторые каталоги и вовсе отсутствуют. Поэтому для получения данных по новейшим обзорам требуется посещать официальные сайты обзоров и скачивать данные через веб-формы. Зачастую это требует усилий и знания технических деталей. К счастью, существует возможность автоматизации данного процесса с помощью протокола TAP (Table Access Protocol) – многие крупные обзоры предоставляют возможность получения данных через данный протокол.

Одним из успешных примеров реализации работы с астрономическими каталогами через протокол TAP является приложение TOPCAT [10], которое является графическим интерфейсом для пакета STIL [11]. Программа обладает широчайшим функционалом для анализа астрономических каталогов. С ее помощью также возможно получение данных из различных источников, в том числе из Страсбургского центра данных. Тем не менее процесс получения данных не автоматизирован и возможны затруднения с пониманием необходимых SQL-запросов. Другой успешный пример программы для работы с протоколом TAP – Aladin [12]. В данной программе реализован удобный способ получения и визуализации данных из множества различных источников, но возможности по анализу астрономических каталогов ограничены по сравнению с TOPCAT. К примеру, в Aladin нет возможности

выполнить кросс-идентификацию более чем двух каталогов.

Одной из целей настоящей работы является создание исходного кода и веб-интерфейса для решения следующей задачи: простое получение и кросс-идентификация наиболее актуальных архивных фотометрических данных об источниках или областях на небе, которые интересны пользователю. Программный код по замыслу автора должен сочетать все лучшие стороны описанных ранее программ для более быстрого решения повседневных задач, связанных с астрономическими каталогами, чем при использовании существующих решений.

Проблема кросс-идентификации каталогов возникает сразу после загрузки данных. Задача кросс-идентификации может быть решена как с помощью локальных пакетов, которые выполняют всю работу на компьютере пользователя, так и с помощью онлайн-приложений, где расчеты производятся удаленно. К локальным решениям относятся TOPCAT/STIL [10, 11], C^3 [13], Xmatch [14], метод Малкова и Карпова [15] и другие. Локальные утилиты нацелены в первую очередь на увеличение скорости кросс-идентификации с помощью применения различных инструментов и технологий. К примеру, в методе Xmatch [14] применяются возможности систем с множеством графических процессоров (GPU), а в работе [16] разработан эффективный алгоритм распараллеливания кросс-идентификации для работы в кластере. Но все-таки эффективное использование алгоритмов с применением параллельных вычислений или GPU нельзя назвать доступным для каждого.

Применяются также технологии по сокращению вычислений с помощью разбиения всего неба на отдельные элементы и последующего выполнения кросс-идентификации только для отдельных элементов. В качестве схем разбиения зачастую используются HEALPix [17] и HTM (Hierarchical Triangular Mesh [18]). Такие технологии использованы в TOPCAT [10], CDS Xmatch [19], методе Малкова и Карпова [15] и других.

К решениям в виде веб-приложений в первую очередь относится сервис CDS Xmatch [19], позволяющий эффективно кросс-идентифицировать два каталога из Страсбургского центра данных или от пользователя. Недостатком сервиса является отсутствие возможности кросс-идентификации большего количества каталогов. Другой менее известный сервис ARCHES [20] позволяет кросс-идентифицировать несколько каталогов, но работа с ним доступна не каждому из-за необходимости писать достаточно сложные скрипты.

Таким образом, в настоящее время существует достаточное количество программ и сервисов для получения данных и их кросс-идентификации, но сложно найти единый интерфейс, который бы объединял удобство получения данных по множеству различных каталогов и их эффективную кросс-идентификацию. Целью настоящей работы является создание такого инструмента, который позволяет без необходимости написания скриптов и SQL-запросов быстро получить и кросс-идентифицировать данные по определенной области неба или по списку источников для множества астрономических каталогов. Для кросс-идентификации источников предлагается использовать алгоритм поиска групп DBSCAN [21]. Преимуществом этого алгоритма кросс-идентификации является работа сразу по нескольким каталогам, симметричность результата при любом количестве каталогов, а также скорость работы, что немаловажно при обработке больших объемов данных.

Результаты кросс-идентификации источников зачастую используются для построения спектрального распределения энергии и его моделирования. В настоящее время наиболее полным и развитым инструментом для данной цели является сервис VOSA [22]. Он содержит богатую коллекцию спектральных фильтров и моделей для различных типов объектов. Тем не менее процесс создания каталога измерений для множества объектов в полной мере не автоматизирован, а система получения фотометрических данных из виртуальной обсерватории, встроенная в VOSA, имеет ограничения – составление списка источников является нетривиальной задачей. Зачастую при планировании наблюдений необходимо знать наблюдательные характеристики целого ряда объектов в различных спектральных диапазонах. Поэтому возникает задача автоматизации создания каталога измерений при максимальном охвате объектов по различным фотометрическим данным. Спектральное распределение энергии позволяет построить модель излучения источника, что в свою очередь позволяет оценивать фотометрические характеристики источников для будущих наблюдений.

В настоящее время существует большое количество астрономических каталогов в различных спектральных диапазонах. С большой долей вероятности источники, которые планируется наблюдать в будущем, уже наблюдались ранее в этих каталогах. Информация из архивных наблюдений может быть очень полезна для построения модели спектрального распределения энергии. Модель позволяет оценить потоки в определенных фильтрах для будущих наблюдений. Построение модели спектрального распределения энергии для множества источников может быть полезно в том числе для планирования будущих наблюдений. Данная задача может быть решена с помощью описанной в работе системы совместно с сервисом VOSA.

2. МЕТОДИКИ КРОСС-ИДЕНТИФИКАЦИИ

На практике существуют две основные задачи кросс-идентификации. Первая – кросс-идентификация источников из списка пользователя (U) с некоторыми астрономическими каталогами (A_i). Вторая – кросс-идентификация астрономических каталогов A_i для определенной области R на небе (обычно окружность, прямоугольник или диапазон координат). В последующих разделах две эти задачи будут рассмотрены более подробно.

2.1. Кросс-идентификация списка источников пользователя с различными каталогами

В данном случае применение классического метода кросс-идентификации с заданным радиусом r оправдано, так источники U из списка пользователя могут считаться опорным каталогом. В том случае, когда пространственная плотность источников в опорном каталоге выше, чем в каталоге сравнения, имеет смысл поменять направление кросс-идентификации и сделать опорным каталогом не каталог пользователя, а каталог сравнения. Именно так было сделано при кросс-идентификации каталога Gaia с астрономическими каталогами, имеющими низкую пространственную плотность источников [23].

Классическая методика кросс-идентификации такова. Для каждого источника из опорного каталога U ищутся все источники-соседи из каталога сравнения A. В том случае, когда угловое расстояние между источниками U и A будет меньше радиуса кросс-идентификации r, то такие источники считаются кросс-идентифицированными. В более сложном случае вместо окружности может выступать эллипс, параметры которого соответствуют неопределенности положения источника, что, к примеру, реализовано в коде C^3 [13].

Проблема неоднозначной кросс-идентификации возникает, когда с источником из каталога пользователя (U) ассоциируются несколько источников из каталога A. Обычно применяется два стандартных решения: match all и match best. Вывод всех связанных источников – match all. Но зачастую необходимо вывести единственный связанный источник. Обычно выводится тот источник, который имеет наименьшее угловое расстояние до источника пользователя. Такое решение называется match best. Но, если источники в каталоге U имеют неопределенность положения более, чем неопределенность положения источников из каталога A, при использовании метода match best возможны ложные кросс-идентификации. Для уменьшения их количества нужна априорная информация о природе источников в списке U, чтобы иметь возможность наложить дополнительные критерии.

Пусть источники из списка пользователя U имеют неопределенность положения со среднеквадратичным отклонением σ. В таком случае радиус для кросс-идентификации может быть задан как $r = 3\sigma$. В таком случае с вероятностью 99.72% мы можем утверждать, что необходимый источник находится в данном радиусе. Если каталоги сравнения имеют высокую плотность источников (к примеру, UKIDSS, Gaia или Pan-STARRS), то в окружность с радиусом $r = 3\sigma$ может попасть множество источников из A.

Для получения правильной и однозначной кросс-идентификации в первую очередь стоит уточнить положения источников из каталога пользователя U, а затем уже выполнять кросс-идентификацию по уточненным положениям. В зависимости от природы источников в списке пользователя, следует выбрать каталог, который лучше всего отражает данный тип источников с достаточной точностью по положению. Если, к примеру, источники U являются яркими звездами в ближнем инфракрасном (ИК) диапазоне, следует кросс-идентифицировать список U с каталогом 2MASS и выбрать самые яркие из них. Если список источников содержит список молодых звездных объектов, то его можно кросс-идентифицировать с каталогом ATLASGAL или Hi-GAL для уточнения их положений. Далее для уточненных положений возможна кросс-идентификация источников с другими каталогами с уменьшенным значением радиуса кросс-идентификации r, что дает меньше ошибок при кросс-идентификации.

2.2. Кросс-идентификация нескольких каталогов по выбранной области на небе

Задача кросс-идентификации нескольких каталогов по областям может возникать, к примеру, при построении карт поглощения или при анализе звездных скоплений. Входными параметрами являются только параметры области на небе, по которой необходимо кросс-идентифицировать несколько каталогов. В таком случае неясно, какой каталог необходимо считать опорным, так как не один из каталогов не является исчерпывающим по полноте источников во всех диапазонах длин волн, а выбор только одного из этих каталогов может привести к потере источников, которые видны в других каталогах.

Для кросс-идентификации источников в этом случае предлагается использовать метод DBSCAN [21]. Данный алгоритм позволяет искать группы источников в однородном наборе пространственных координат. Наиболее важным параметром для формирования групп является порог ϵ – минимальное угловое расстояние меж-

ду двумя источниками для того, чтобы они были отнесены к одной группе.

Преимуществом метода является возможность группировки сразу всех источников из различных каталогов, которые находятся в непосредственной пространственной близости друг к другу (расстояние не более ϵ). Таким образом, алгоритм является асимметричным – нет необходимости выбирать опорный каталог. В том случае, когда для некоторого источника не найдено других связанных источников на расстоянии ϵ, он считается изолированным. Если в некоторой группе содержится несколько источников, то все источники получают одинаковый идентификатор группы.

Скорость работы алгоритма не зависит от количества каталогов, а зависит только от суммы источников из всех каталогов, что является важным преимуществом по сравнению с другими алгоритмами. Для алгоритма DBSCAN неважно, будут ли проанализированы 2 каталога по 500 тыс. источников или 10 каталогов по 100 тыс. источников. Еще одной особенностью метода DBSCAN является отсутствие необходимости ввода априорной информации о количестве групп источников – оно является выходным параметром. С другой стороны, в одну группу могут попасть несколько источников даже из одного каталога в том случае, если источники расположены на близких угловых расстояниях (менее ϵ). В этом случае важен правильный выбор параметра ϵ, о чем подробнее написано в разделе 2.3.1.

В качестве входных данных для DBSCAN вводятся координаты всех источников из различных каталогов для выбранной области на небе. Параметр ϵ обычно выбирается в соответствии с минимальным расстоянием между ближайшими источниками в исследуемых каталогах. Далее выполняется поиск пространственных групп. Каждый источник получает идентификатор группы. Он равен нулю для тех источников, которые не имеют пространственных ассоциаций с другими каталогами. Напротив, для источников из одной группы идентификатор больше нуля и является общим для всей группы.

2.3. Случай неоднозначной кросс-идентификации

После объединения источников в группы с помощью DBSCAN каждому источнику присваивается идентификатор, уникальный для каждой группы. По данному идентификатору возможен вывод данных из различных каталогов. Случай, когда в каждую группу попадает по одному источнику из каждого каталога, является наиболее простым для анализа. Вывод параметров по этой группе является вполне однозначным. Но в том случае, когда несколько источников из одного каталога являются частью одной группы, вывод параметров группы неоднозначен. Данная проблема является общей для всех методов кросс-идентификации, и в классическом случае она решается с помощью выбора всех источников (match all) или самого близкого источника (match best) по отношению к источнику из опорного каталога. В случае кросс-идентификации с помощью DBSCAN есть свои особенности: в данном алгоритме априори нет опорного каталога. Тем не менее можно выстроить каталоги в порядке увеличения минимального расстояния между источниками в каталоге.

Рассмотрим конкретный пример. Пусть в некоторую группу, найденную с помощью DBSCAN, попало 3 источника Pan-STARRS, 2 источника Gaia и 1 источник WISE. Учитывая, что источники WISE часто являются неразрешенными, использовать их в качестве опорных в данном случае затруднительно. Напротив, каталог с минимальным расстоянием между источниками (в данном случае Pan-STARRS) можно считать опорным, так как более высокая плотность источников может быть достигнута максимальной разрешающей способностью. Наиболее яркий источник из такого каталога можно считать основным для данной группы. Далее для данной группы выводятся параметры таких источников, которые расположены наиболее близко к опорному источнику.

Описанный подход работает наиболее эффективно при сравнении каталогов на близких диапазонах длин волн. Но он не учитывает физические особенности излучения источников в различных диапазонах длин волн. К примеру, молодые звездные объекты зачастую не видны в оптическом диапазоне, но могут быть яркими в инфракрасном диапазоне. Более того, количество точечных источников в дальнем ИК диапазоне значительно меньше, чем в ближнем ИК и оптическом диапазоне, поэтому для отождествления молодых звездных объектов не всегда следует использовать каталог с максимальной плотностью источников. Другой пример – проэволюционировавшие звезды. Они имеют инфракрасный избыток света и хорошо видны в инфракрасном диапазоне длин волн, но отождествить их в оптическом диапазоне бывает трудно. В этих и подобных случаях не имеет смысла использовать каталог с максимальной разрешающей способностью в качестве опорного. Необходимо вручную задать опорный каталог в зависимости от решаемой задачи. Эта априорная информация поможет частично разрешить неоднозначность в выборе опорного источника в каждой группе. При этом не в каждую группу может попасть источник из опорного каталога. В зависимости от решаемой задачи такие группы можно либо отбросить, либо записать параметры источников в таких группах в

отдельный список источников. Может случиться другая ситуация, когда в группу попадает сразу несколько источников из опорного каталога. В таком случае необходимо выбрать наиболее яркий из них по определенному пользователем параметру.

Таким образом, используя каталог с максимальной разрешающей способностью, или каталог, выбранный пользователем для решения конкретной задачи, можно решить вопрос о неоднозначности кросс-идентификации при группировке источников с помощью DBSCAN.

2.3.1. Выбор размера порога для кросс-идентификации многих каталогов. Параметр ϵ соответствует минимальному угловому расстоянию между двумя ближайшими источниками для того, чтобы они считались связанными в единую группу. Если в ближайшей окрестности находится больше двух источников, то для каждого последующего источника $N+1$ вхождение в группу G происходит в том случае, если источник $N+1$ находится на угловом расстоянии менее ϵ к любому источнику из группы G.

Выбор оптимального значения ϵ является решающим для корректного отождествления источников из различных каталогов. При выборе избыточных значений ϵ могут быть сгруппированы слишком большое количество источников, а при недостаточных значениях ϵ каждый источник будет изолированным от других источников, и кросс-идентификация фактически выполнена не будет.

Для проверки корректности выбора значения ϵ можно выполнить внутреннюю проверку каждого исследуемого каталога. Для этого метод DBSCAN запускается независимо для каждого каталога при некотором значении ϵ. Если в каталоге нет множественных измерений одного источника, то количество источников, которые будут объединены в группы при данном значении ϵ, могут служить мерой ошибки выбранного порога при кросс-идентификации. Решающим значением является характерное расстояние между ближайшими источниками в данном каталоге, которое в свою очередь зависит от плотности источников в каталоге. Если при некотором значении ϵ значительная часть источников в каталоге будет объединяться в группы со своими ближайшими соседями, то такая группировка будет бесполезна при кросс-идентификации с другими каталогами, т.к. вместо объединения с источниками из других каталогов будут созданы группы из соседних источников по одному каталогу, поэтому возникнет много случаев неоднозначной кросс-идентификации. Таким образом, верхняя граница значения ϵ ограничена необходимостью уменьшения числа бесполезных групп, образованных из-за близости источников в каталоге с наибольшей плотностью источников. С другой стороны, необходимо выбрать максимально большое значение ϵ, чтобы не исключить создание полезных групп, когда источники из различных каталогов являются одним объектом, но имеют небольшие сдвиги по положению в пределах эллипса неопределенности. Проблема заключается в том, что для различных каталогов имеет место различное минимальное расстояние между соседними источниками. Поэтому выбор фиксированного значения ϵ может привести к тому, что для источников с большой неопределенностью положения (более ϵ) кросс-идентификация выполнена не будет и источники с большой неопределенностью положения останутся “изолированными”, или могут быть кросс-идентифицированы ложно. Решение этой проблемы заключается в дополнительной проверке источников с большой неопределенностью положения. Необходимо для каждого такого источника проверить его ближайших соседей с увеличенным радиусом поиска и отметить все найденные таким образом кросс-идентификации. Подробнее данная проблема рассматривается в разделе 2.3.2.

Пусть метод DBSCAN запускается независимо для каждого из рассматриваемых каталогов. Отношение числа изолированных источников, которые не имеют соседей при данном значении ϵ, к общему числу источников дает представление о количестве “бесполезных” групп, при условии, что на каждый источник приходится только одно измерение. Мы будем считать данное отношение точностью работы алгоритма DBSCAN для данного каталога при данном значении ϵ. Для получения оптимальных результатов кросс-идентификации значение ϵ должно быть выбрано таким образом, что количество изолированных источников в рамках одного каталога должно быть не менее 99.72% (по аналогии с правилом трех сигм). В табл. 1 представлен пример верхних порогов ϵ для различных астрономических каталогов, которые обеспечивают точность кросс-идентификации не хуже 99.72%. Выбрано направление $(l, b) = (31.5^\circ, 0.0^\circ)$, для которого доступны данные в рассматриваемых каталогах. Величины порогов могут варьироваться для различных направлений, так как плотность источников в каталоге является функцией их положения, т.к. значения в таблице не являются абсолютными и показаны только для ознакомления. В веб-приложении cross-match.online доступна возможность оценки минимального порога для кросс-идентификации в режиме онлайн для любых наборов данных.

Следует отметить, что для многих каталогов не выполняется условие, при котором на один источник приходится только одно измерение. Зачастую множественные измерения возникают из-за повторения наблюдений источников в некоторых областях. Для таких каталогов для оценки верхне-

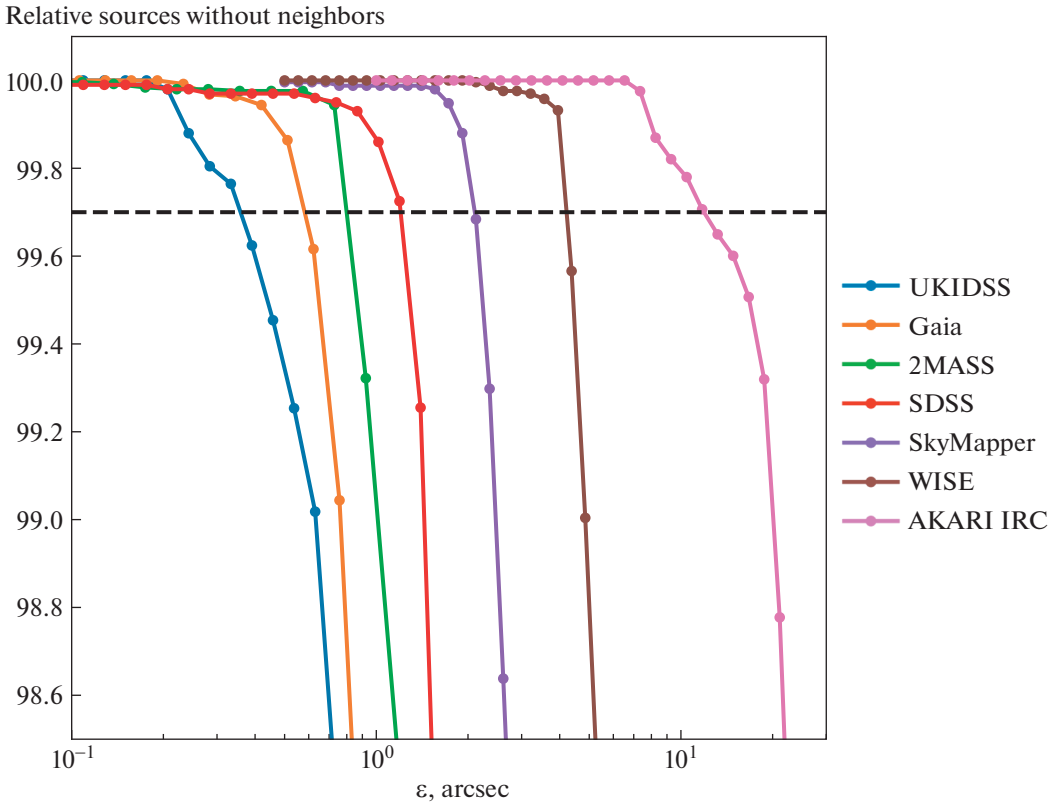

Рис. 1. Зависимость относительного числа изолированных источников после группировки DBSCAN для серии каталогов при различных значениях порога ϵ. Горизонтальной линией обозначен уровень изолированности источников 99.72%.

го порога ϵ необходимо сперва выполнить внутреннюю группировку с порогом, соответствующим неопределенности положения источников. Подробнее об этом – в разделе 3.3.1. Другой выход – наложить фильтр на исходный каталог таким образом, чтобы выводить данные только за одну эпоху наблюдений.

На рис. 1 представлен пример зависимости относительного числа изолированных источников при разных значениях порога ϵ для различных астрономических каталогов. Из рис. 1 следует, что для большинства каталогов относительное число изолированных источников падает медленно до некоторого порога, а после него идет быстрое падение, связанное с достижением характерного расстояния между источниками в каталоге. Значение ϵ, при котором количество изолированных источников для данного каталога составляет 99.72%, будем в дальнейшем называть верхним пределом (ϵ_{max}). Для практического получения значения ϵ_{max} достаточно рассчитать несколько опорных точек, а более точное значение ϵ определяется с помощью интерполяции между двумя расчетными значениями.

Так как в методе DBSCAN все каталоги оцениваются единым значением порога ϵ, то результат с наименьшими неоднозначностями получится в случае выбора порога, который не превышает значения ϵ_{max} для всех рассматриваемых каталогов. При выборе бóльших значений порога может возникнуть ситуация, когда отдельные источники из каталога с наибольшей плотностью источников будут объединены в крупные группы и возникнет проблема неоднозначности параметров групп. К примеру, при сравнении каталогов Gaia и WISE выбор порога сводится к поиску такого значения ϵ, при котором источники Gaia не будут объединяться в группы сами с собой. При этом источники WISE будут кросс-идентифицированы с источниками Gaia только в том случае, если они будут к ним на достаточно близком расстоянии, соответствующему параметру ϵ. В случае, если источник WISE будет на значительном удалении от источника Gaia, необходимо учитывать неопределенность положения для решения вопроса о том, являются ли данные источники кросс-идентифицированными (вопрос об этом будет рассмотрен в разделе 2.3.2).

Таким образом, если требуется минимизация неоднозначности при кросс-идентификации, выбор значения ϵ может быть выполнен в следующей последовательности: (1) для каждого каталога рассчитываем верхний предел ϵ_{max}, при котором реализуется количество изолированных источников не менее 99.72%; в случае множественных измерений надо предварительно выполнить внутреннюю группировку источников; (2) порог для кросс-идентификации ϵ надо выбрать таким образом, чтобы он не превышал значение ϵ_{max} для любого из каталогов.

Если для исследования наличие ложных групп не является критичным, а более важным является поиск всех возможных корреляций источников в различных каталогах, то в таком случае можно выбрать уровень ϵ_{max}, при котором реализуется изоляция источников на 95% (2σ). Тогда значение ϵ, соответственно, может быть увеличено.

2.3.2. Учет неопределенности положения источников. Пусть значение ϵ для DBSCN было выбрано таким образом, что оно не превышает значения ϵ_{max} для любого из рассматриваемых каталогов. Пример значений ϵ_{max} для некоторых оптических и инфракрасных каталогов представлен в табл. 1.

В результате выполнения алгоритма DBSCAN для всех исследуемых каталогов при некотором выбранном значении ϵ все наблюдения будут поделены на два типа: источники, которые получили идентификатор группы, и источники, которые его не получили ("изолированные"). Первый тип источников представляет собой найденные с помощью DBSCAN кросс-идентификации источников из различных каталогов. Второй тип источников соответствует "изолированным" источникам, которые при данном значении ϵ не имеют ближайших соседей. Но так как неопределенность положения источника может быть больше, чем значение ϵ, то его "изолированность" может быть связана со смещением из-за неопределенности положения. При сравнении каталогов с существенно разной неопределенностью положения источников (к примеру Gaia и WISE), количество "изолированных" источников вследствие неопределенности положения может быть достаточно велико.

Таблица 1. Верхние границы порога DBSCAN ϵ_{max} для некоторых оптических и инфракрасных каталогов, которые обеспечивают изоляцию источников как минимум на 99.72%

Каталог	Радиус, °	Источники, тыс.	ϵ_{max}, ″
Оптические каталоги			
Gaia [2]	1	625.5	0.58
SDSS [5]	1	384.7	1.2
SkyMapper [24]	1	94.5	2.1
Инфракрасные каталоги			
UKIDSS [25]	0.3	470.9	0.35
MASS [26]	1	370.3	0.80
WISE [27]	1	72.2	4.1
Akari IRC [28]	5	24.5	11

Примечание. Представлены оценки для направления $(l, b) = (31.5°, 0°)$. Радиус показывает размер области, которая выбрана для анализа. Столбец "Источники" показывает количество источников в выбранной области.

Для учета неопределенности положения источников необходимо выполнить дополнительные шаги после кросс-идентификации методом DBSCAN. Для каждого изолированного источника (который не был отнесен ни к какой группе по результатам DBSCAN) необходимо найти N ближайших соседей. В наиболее простом случае для каждого каталога задается фиксированная величина (радиус) неопределенности положения. Пусть для изолированного источника A имеется некоторый сосед B. Источники A и B считаются кросс-идентифицированными в том случае, если $r_A + r_B < d$, где r_A и r_B – радиус неопределенности ($r = 3\sigma$) положения источников A и B, d – угловое расстояние между источниками A и B. Таким образом, если "изолированный" источник ассоциируется с одним или несколькими другими источниками, тогда его можно считать уже не "изолированным", а кросс-идентифицированным с этими источниками (см. левую часть рис. 2). Данную процедуру необходимо провести для всех источников, которые были помечены как "изолированные" в методе DBSCAN при использовании порога ϵ. Процедура также может быть проведена для проверки кросс-идентификации источников в группах, но в таком случае один источник с большой неопределенностью положения может быть кросс-идентифицирован сразу с несколькими ранее найденными группами.

В более сложном случае нужно учитывать неопределенность положения каждого источника в отдельности. Для этого необходимо знать три величины для каждого источника: θ_{maj}, θ_{min} – размер большой и малой оси эллипса ошибки, θ_{PA} – позиционный угол эллипса ошибки. Источники являются кросс-идентифицированными в том случае, если их эллипсы неопределенности имеют пересечение (см. правую часть рис. 2).

Указанная процедура позволяет исправить главный недостаток метода DBSCAN, который заключается в фиксированном значении ϵ для всех каталогов. Сначала выбирается значение ϵ, которое обеспечивает минимальный уровень ошибок при группировке источников: $\epsilon < \epsilon_{max}$;

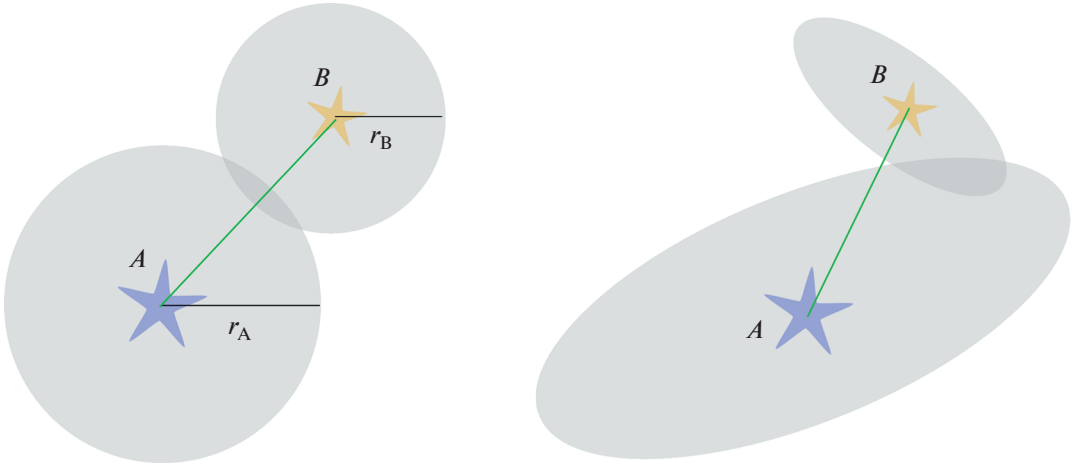

Рис. 2. Схематичное изображение источника A и его соседа B в двух случаях учета неопределенности положения источников. Слева – случай фиксированного размера неопределенности для каталогов, к которым относятся источники A и B. Справа – случай учета эллипса ошибки для каждого источника.

затем производится дополнительная проверка кросс-идентификаций с учетом радиуса или эллипса неопределенности. Таким образом, можно достичь лучших результатов кросс-идентификации, чем просто с помощью группировки всех источников с фиксированным значением ϵ. Практическая проверка корректности кросс-идентификаций при данном способе поиска будет представлена в разделе 4.1.

3. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ МЕТОДИКИ

Методика кросс-идентификации источников из различных каталогов, которая описана в настоящей работе, реализована в графическом веб-интерфейсе[1], а также с помощью скрипта для языка Python[2], который может быть запущен на локальном компьютере. Скрипт предназначен только для выполнения процедуры кросс-идентификации, а веб-интерфейс включает в себя полный цикл необходимых утилит для выбора, загрузки и кросс-идентификации различных астрономических каталогов. Веб-интерфейс включает в себя несколько этапов работы, которые будут рассмотрены в последующих разделах.

3.1. Выбор каталогов и режима работы

Общий вид главной страницы сервиса cross-match.online представлен на рис. 3. Реализовано два режима работы сервиса: (1) кросс-идентификация списка источников пользователя с астрономическими каталогами и (2) кросс-идентификация области исследования. Режим (1) описан подробно в разделе 2.1, а режиму (2) посвящена основная часть работы и описана в разделе 2.2 и других разделах.

В самую первую очередь необходимо выбрать каталоги, которые необходимо кросс-идентифицировать. В систему cross-match.online уже встроены следующие астрономические каталоги (в будущем этот список может быть расширен).

1. Ультрафиолетовые: GALEX GUVcat_AIS [29], каталог 10 лет работы XMM и SWIFT [30], MSX UV [31];

2. Оптические: APASS DR9 [32], Gaia DR3 [2], SDSS DR16 [5], SkyMapper DR1.1 [24], Pan-STARRS DR1 [3], IPHAS DR2 [33];

3. Инфракрасные: 2MASS и 2MASX [26], UKIDSS [25], WISE AllSky [27], DENIS DR3 [34], IRAS v2.0 [35], AKARI IRC [28], AKARI FIS [36], GLIMPSE I+II+3D [37], ATLASGAL [38], Hi-GAL 360 [8].

Кроме перечисленных выше, есть возможность включить в работу любое количество произвольных каталогов, включенных в Страсбургский центр данных. Для этого необходимо указать внутренний идентификатор таблицы (к примеру II/321/iphas2) в системе CDS VizieR [9]. Есть также возможность присвоить каталогу произвольное название через символ @ после идентификатора таблицы (к примеру, II/363/unwise@unWISE). В таком случае упрощается идентификация каталога. Для встроенных каталогов возможен автоматизированный вывод фотометрических данных для сервиса VOSA, что необходимо для построения и моделирования спектрального распределения энергии. В настоящей работе данная функция подробно не рассматривается, и она будет рассмотрена подробнее в будущих работах.

[1] https://cross-match.online
[2] https://cross-match.online/source_code.zip

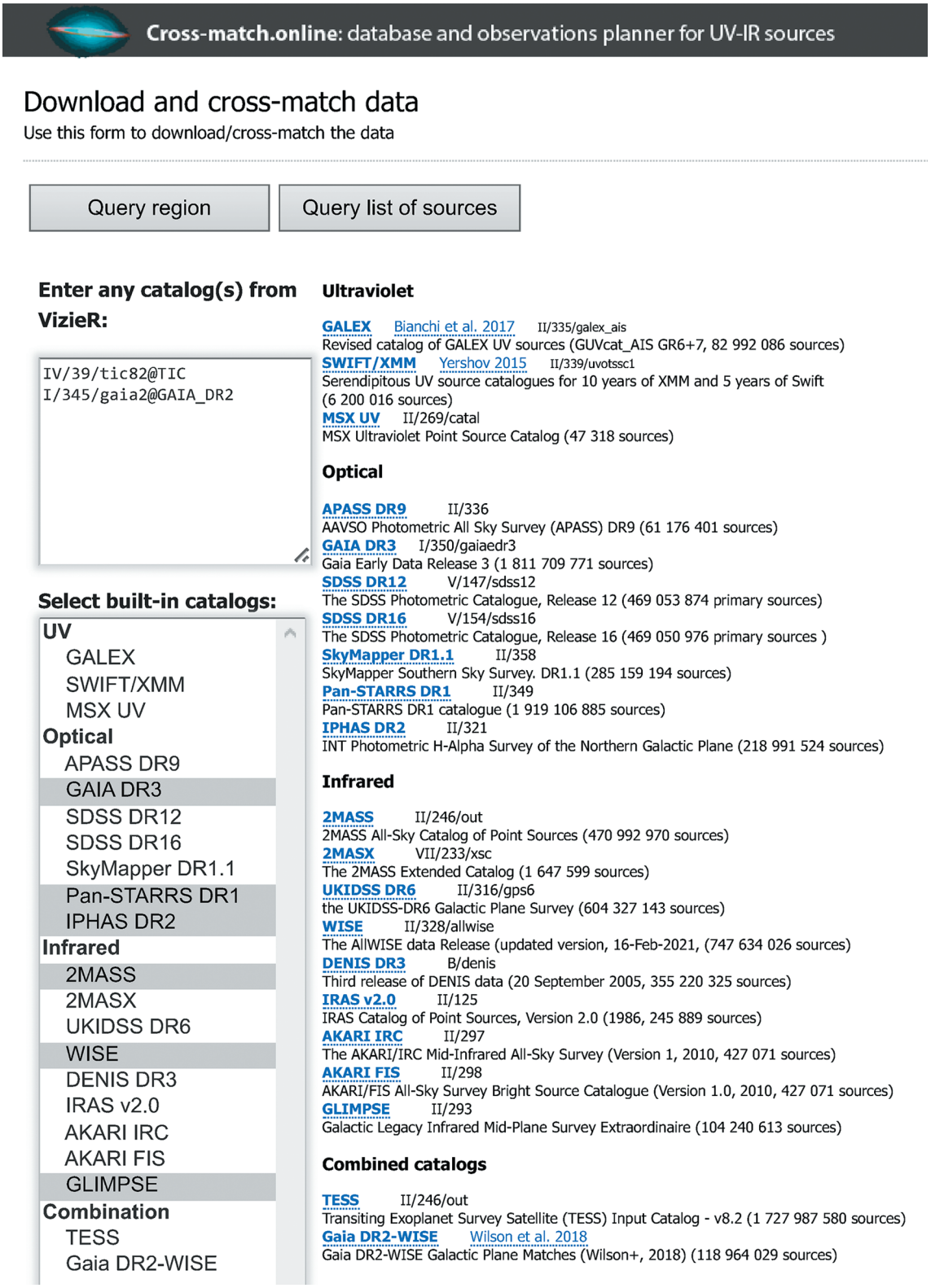

Рис. 3. Первый шаг: выбор каталогов и режима работы.

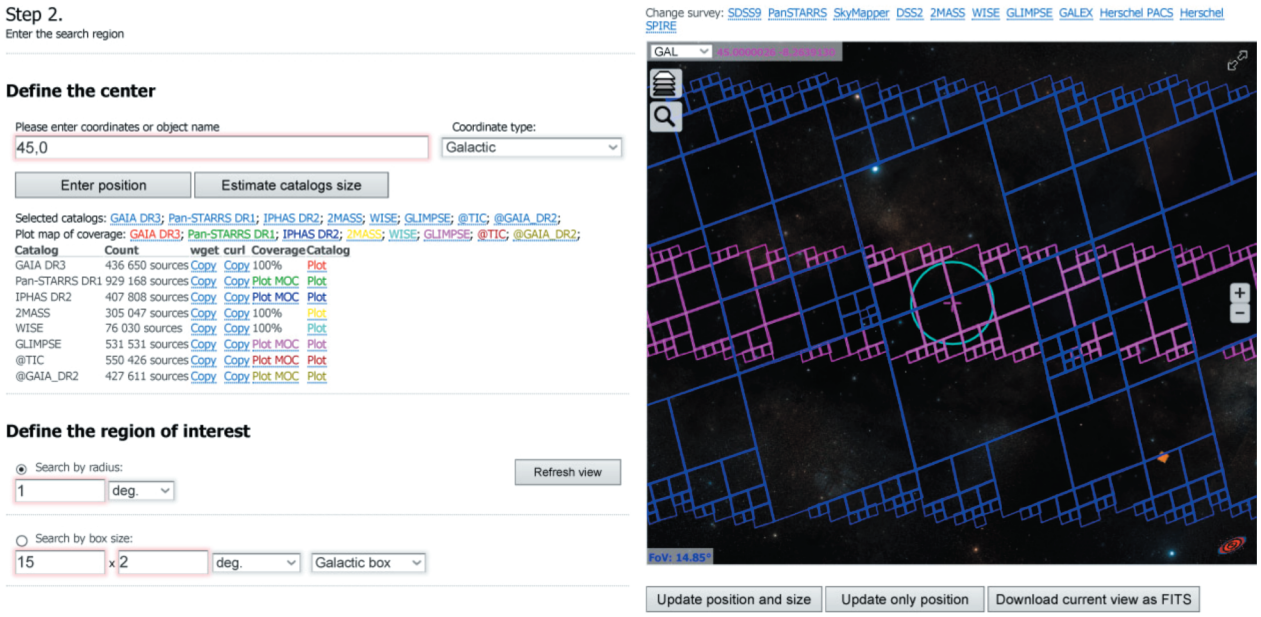

Рис. 4. Второй шаг: выбор области для исследования.

Выбор каталогов ничем не ограничен, но следует учесть, что не все каталоги имеют полное покрытие неба. В системе предусмотрена специальная функция, которая загружает карту покрытия любого каталога с помощью сервиса MOCServer [39], которой можно воспользоваться на следующем этапе.

3.2. Выбор области для исследования

На данном этапе появляется возможность выбрать параметры исследования.

В режиме 1 (область) задается центр области с помощью астрономических координат, либо названия источника (с помощью сервиса Simbad [40]). Дополнительно задаются размеры области – радиус окружности или размер прямоугольника. Выбранная область отображается визуально с помощью программы Aladin Lite API [42]. Реализована возможность визуального выбора области для исследования: при перемещении положения центра и масштаба окна Aladin Lite, положение или размер области можно обновить в соответствии с текущим положением на небе и выбранным масштабом. Общий вид формы в данном режиме показан на рис. 4.

В режиме 2 (список источников) пользователь задает список источников с помощью астрономических координат. Для этого необходимо загрузить файл в текстовом формате (с указанием разделителя) и выбрать, какие именно столбцы нужно использовать для названий источников и их координат. Возможны следующие системы координат: FK4 B1950, FK5 J2000, Галактическая. Ввод данных возможен как в градусах, так и в формате (часы, минуты, секунды).

На этом же шаге появляется возможность оперативно оценить параметры каталогов для выбранной области или списка объектов. Для проверки есть возможность построить в режиме онлайн все источники из выбранных каталогов, а также построить карты покрытия каталогов, что может быть полезно для поиска областей пересечений различных каталогов.

3.3. Загрузка данных, выбор порога и выполнение кросс-идентификации

На данном этапе происходит загрузка каталогов и их дальнейшая кросс-идентификация. Общий вид формы показан на рис. 5. Область для исследования на данном этапе фиксирована и не может быть изменена (для ее изменения нужно вернуться на предыдущий этап). Перед загрузкой возможно применение фильтров. Фильтры возможны двух типов: (1) ограничение максимального числа источников для загрузки; (2) наложение условий по параметрам каталогов. Условия являются математическими выражениями, в которых можно использовать все доступные параметры каталога. Подробное описание параметров каталогов доступно во всплывающем окне по ссылке “Description”. Нажатие на кнопку “Estimate source size” позволяет оценить число источников в выбранной области с применением фильтров (при необходимости).

При нажатии на кнопку “Download all” происходит параллельная загрузка всех каталогов в хра-

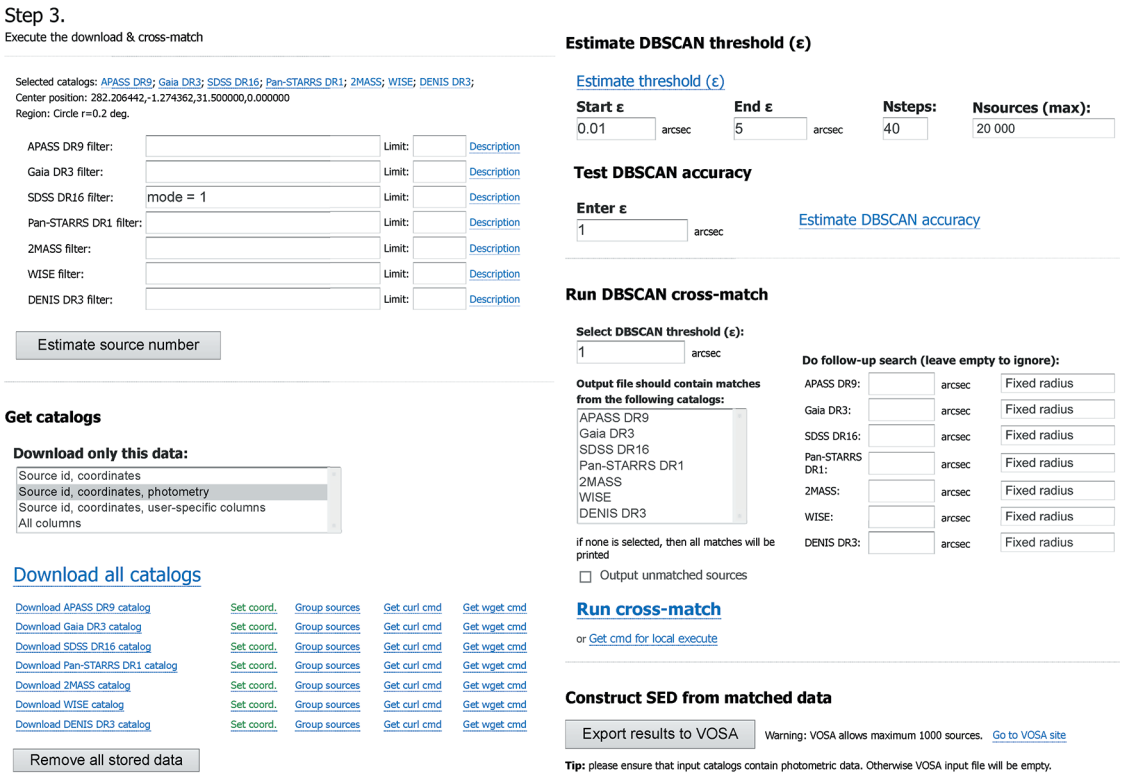

Рис. 5. Третий шаг: загрузка данных, выбор порога и выполнение кросс-идентификации. Форма на данном рисунке может отличаться от последней версии, представленной по адресу https://cross-match.online.

нилище данных на сервере, где расположен сервис cross-match.online, с учетом выбранной области и заранее введенных фильтров. Перед загрузкой можно выбрать объем скачиваемой информации. Доступно четыре режима: (1) только названия и координаты источников; (2) названия, координаты и фотометрические данные – режим доступен только для встроенных каталогов; (3) параметры, выбранные пользователем, и (4) все данные. Обычно для работы нет необходимости загружать все доступные данные, поэтому более эффективным будет ограничить набор загружаемых данных в зависимости от решаемой задачи.

Координаты для всех каталогов, которые загружаются для кросс-идентификации, должны быть представлены в одной системе координат. К примеру, разность между системами координат FK5 и ICRS может достигать 0.08″, а разность между системами FK4 и FK5 – десятки угловых минут. Поэтому в системе cross-match.online есть возможность вручную указать параметры с координатами в нужной системе координат, если указанные параметры по умолчанию не соответствуют требованиям. Для этого нужно нажать на ссылку “Set coord” и выбрать необходимые столбцы во всплывающем окне. Обычно в Страсбургском центре данных для большинства каталогов представлен пересчет координат в системе FK5 (J2000), поэтому достаточно просто выбрать нужные столбцы. Но при отсутствии такой возможности пересчет координат можно выполнить прямо в системе cross-match.online. Для этого в окне с выбором координат для каталога нужно выбрать столбцы с координатами и нажать на галочку “Convert sky coordinates” и выбрать режим работы: “ICRS → FK5 (J2000)”, “FK5 (J2000) → ICRS”, “FK4 (B1950) → FK5 (J2000)”, “FK4 (B1950) → → ICRS”, “Galactic → FK5 (J2000)”, “Galactic → → ICRS”, после чего нужно загрузить каталог. В этом случае координаты будут пересчитаны в соответствии с выбранным режимом и будут в дальнейшем использованы при кросс-идентификации.

После загрузки каталогов, при работе в режиме области, необходимо выполнить оценку оптимального значения порога для кросс-идентификации. При нажатии на кнопку “Estimate threshold (ϵ)” происходит оценка верхнего предела для значения ϵ для каждого каталога. Оценка производится с помощью многократного запуска метода DBSCAN при различных значениях ϵ. Затем с помощью интерполяции находится такое значение ϵ, при котором относительное число изолированных источников составляет не менее 99.72% (3σ). Перед запуском оценки нужно выбрать ин-

тервал проверки значений порога (Start ϵ, End ϵ), а также число шагов сетки (N_{steps}). В сетке используется логарифмическая шкала с основанием 10. Так как для поиска оптимального значения ϵ метод DBSCAN запускается многократно, то использовать полный набор данных нежелательно, т.к. это может привести к существенному увеличению времени расчетов. Для ускорения вычислений при большом количестве источников программа ограничивает ввод данных по каждому каталогу до 5–50 тыс. источников (данный параметр N_{sources} может быть настроен). Такое ограничение существенно не изменяет параметры группировки, так как даже часть источников из каталога обладают свойствами всей выборки, но при этом существенно сокращается время расчетов. Но следует иметь в виду, что ограничение по числу источников приводит к тому, что данные загружаются из ограниченной области, которая может быть смещена относительно центра выбранной области. Если плотность источников в центре области существенно выше, чем на периферии, то это может привести к неправильной оценке порога. Для решения следует уменьшить размер области таким образом, чтобы число источников для каждого каталога не превышало 50 тыс. После оценки порога размер области можно обратно увеличить. После завершения оценки выводится график, подобный рис. 1, а также численные значения верхних пределов ϵ для каждого каталога. Если начальное или конечное значение ϵ выбрано с недостаточным запасом, то программа выдаст предупреждение о необходимости увеличения верхнего порога или уменьшения нижнего порога.

Далее нужно принять решение о том, с каким порогом выполнять первичную кросс-идентификацию. Для этого могут быть полезны две функции. Первая: можно вывести загруженные источники в окне программы Aladin Lite и визуально оценить размер порога, нажав на кнопку “Display threshold size”, вводя необходимое значение ϵ. В окне Aladin Lite появится окружность соответствующего размера, которая позволяет оценить количество источников, которые будут объединены в группы с данным выбранным порогом. Можно также проверить число создаваемых групп в каждом каталоге без ограничений по максимальному числу источников. Для этого в разделе “Test DBSCAN accuracy” необходимо ввести значение порога в текстовое поле и нажать на “Estimate DBSCAN accuracy”. В результате появится таблица, в которой будет показано количество групп и изолированных источников для каждого каталога. Слишком большое число групп может привести к большим неоднозначностям при кросс-идентификации. При этом не следует принимать решение о пороге для группировки, ориентируясь на каталоги с множественными измерениями. Для таких каталогов нужно либо предварительно выполнить внутреннюю группировку, либо использовать для оценки другие каталоги, где на каждый источник приходится только одно измерение (см. подробнее раздел 3.3.1).

Непосредственный запуск процесса кросс-идентификации начинается при нажатии на “Run Cross-match”, причем предварительно нужно ввести порог для группировки в соответствующее текстовое поле. Результат кросс-идентификации доступен для загрузки в формате CSV, а параметры расчетов сохраняются в файле в формате LOG.

Для улучшения количества кросс-идентификаций предусмотрена возможность дополнительного поиска в направлении на источники, которые были отмечены как “изолированные” в DBSCAN. При этом для таких источников можно выбрать увеличенный радиус поиска, который будет соответствовать неопределенности положения источников в данном каталоге. Дополнительный поиск будет произведен только для источников из тех каталогов, для которых указан радиус дополнительного поиска. Вместо указания фиксированного радиуса возможно указать параметры эллипса неопределенности для положения источников. Два режима поиска соответствуют двум режимам учета неопределенности положения источников (см. рис. 2).

3.3.1. Каталоги с множественными измерениями одинаковых источников. Для некоторых каталогов имеются множественные измерения одинаковых источников в некоторых областях. К таким каталогам, к примеру, относится каталоги SDSS [5], UKIDSS [25], Pan-STARRS [3], IPHAS [33] и другие. Множественные измерения не мешают их кросс-идентификации с другими каталогами, но для правильной оценки порога для кросс-идентификации необходимо выполнить внутреннюю группировку по положению. Она производится с помощью того же метода DBSCAN [21] независимо от других каталогов.

При наличии значений σ_{RA} и σ_{Dec} в самом каталоге можно оценить порог для группировки следующим образом. Необходимо вычислить 95-й процентиль от значения $3\sigma_{\text{RA}}$ и $3\sigma_{\text{Dec}}$ для всех источников в выборке. Далее выбрать максимальное значение и использовать его в качестве порога для группировки. Вычисление 95-го процентиля для всех числовых параметров доступно в системе cross-match.online после загрузки каталога при нажатии на ссылку “Get statistics”.

Для выполнения предварительной группировки каталога необходимо нажать на ссылку “Group sources” для уже загруженного каталога и ввести порог для группировки. При необходимости можно визуально отобразить источники в Aladin

Lite, которые были объединены в группы, нажав на ссылку “Plot grouped sources”. После группировки число источников в каталоге будет уменьшено.

Следует отметить, что для кросс-идентификации выполнять предварительную группировку источников не обязательно, так как алгоритм cross-match.online не выводит в качестве кросс-идентификаций те группы, которые содержат источники только из одного каталога. Такие группы попадают в разряд изолированных источников. Если в некоторую группу, кроме множества источников из одного каталога, попал хотя бы один источник из другого каталога, то такая группа уже войдет в результаты кросс-идентификации. Из множества источников в группе будет выбран только один источник для каждого каталога. Аналогичный результат может быть получен при предварительной группировке каталога со множественными измерениями и последующей кросс-идентификации с другими каталогами.

3.4. Экспорт данных для построения спектрального распределения энергии источников

С помощью разработанной системы есть возможность экспорта данных для системы VOSA, что упрощает построение спектрального распределения энергии для выбранных источников или области исследования. Для этого необходимо после выполнения кросс-идентификации нажать на кнопку “Export results to VOSA” и сохранить файл на диск. Этот файл далее можно использовать в качестве входного для системы VOSA. Данная функция доступна только для встроенных в систему cross-match.online каталогов. Необходимо также обеспечить наличие фотометрических данных в загруженных каталогах, в противном случае выходной файл не будет содержать данных. В системе VOSA есть ограничение на максимальное количество источников во входном файле, равное 1000.

3.5. Работа в режиме оффлайн

Так как решение определенных задач (к примеру, построение карт поглощения) предполагает кросс-идентификацию больших объемов данных (более 10 млн. источников), то такой процесс целесообразнее выполнять на локальной машине пользователя, чтобы иметь возможность использовать имеющиеся вычислительные ресурсы и лучше контролировать процесс работы. Для этого в рамках настоящей работы доступен исходный код для кросс-идентификации различных каталогов, который может быть запущен на локальном компьютере с установленным языком программирования Python 2.7+. Код доступен по адресу cross-match.online[3].

Размер каталогов в этом случае ограничен только производительностью компьютера и объемом оперативной памяти. Следует учесть, что при кросс-идентификации каталогов с количеством источников более 10–15 млн. желателен объем оперативной памяти более 8 Гб. При этом систему cross-match.online можно использовать для получения скриптов для загрузки каталогов на локальную машину пользователя. Возможен удобный выбор области для загрузки необходимых данных. Пользователю предоставляются готовые текстовые команды для утилит “wget” или “curl”, позволяющие загрузить каталоги в необходимом объеме на локальной машине.

Для получения соответствующих команд необходимо нажать на ссылку “Get curl cmd” или “Get wget cmd”. Будут выведены команды с учетом выбранных фильтров. Название выходных файлов следует сохранить, если требуется дальнейшая кросс-идентификация каталогов. Для локального запуска кросс-идентификации для каждого каталога необходим также специальный файл с расширением *.cols – он содержит информацию о типах данных в каталоге, а также указывает на столбцы с названием источника и его координатами. Для получения этого файла необходимо нажать на ссылку “Get column information file (.cols)” после вывода скрипта для загрузки каталога.

После загрузки каталогов в формате CSV на ПК пользователя необходимо запустить скрипт на языке Python, который последовательно читает все загруженные каталоги и выполняет их кросс-идентификацию с выбранным значением порога. Значение порога лучше всего предварительно оценивать в онлайн-версии системы cross-match.online, ограничив число источников до ~50 тыс. в каждом каталоге для ускорения расчетов. Также исходный код выполняет уточняющий поиск кросс-идентификаций с использованием неопределенности положения источников (задается отдельно для каждого каталога). Для получения правильной команды для запуска скрипта можно выбрать необходимые параметры в онлайн-версии, а затем нажать на “Get cmd for local execute”. Будет выведена строка, которую нужно ввести в терминале на локальном компьютере для корректного запуска кросс-идентификации.

Если требуется выполнить группировку одного каталога по определенному значению ϵ на локальном компьютере, то данный каталог указывается как единственный входной каталог для скрипта. В таком случае код будет работать в режиме одного каталога. Пример запуска скрипта

[3] https://cross-match.online/source_code.zip

для группировки одного каталога (для всех файлов подразумевается расширение CSV) с порогом 0.5″:

```
python cross-match-dbscan.py 0.5
output_ukidss_dr6 input_ukidss_dr6
```

Пример запуска кросс-идентификации одновременно трех каталогов (2MASS, UKIDSS, WISE) с порогом 1.1″:

```
python cross-match-dbscan.py 1.1 output_file
4PHh4_2mass,4PHh4_ukidss_dr6,4PHh4_wise
,,3 1
```

Для каталога WISE в данном примере выполняется дополнительный поиск ассоциаций в изолированных источниках с фиксированным радиусом 3″; выводятся только те ассоциации, которые содержат источник 2MASS (порядковый номер каталога 1). “4PHh4” – уникальный идентификатор загрузки (отличается для разных пользователей), присваивается автоматически при загрузке каталога с помощью cross-match.online.

3.6. Технические особенности исходного кода

Веб-сервис cross-match.online написан с использованием языка Perl/CGI. Сама процедура кросс-идентификации разработана с использованием языка Python. После тестирования различных модулей были использованы наиболее эффективные решения для ускорения работы алгоритма кросс-идентификации как в режиме онлайн, так и оффлайн. Для увеличения скорости работы при кросс-идентификации не используются реляционные базы данных (MySQL, PostgreSQL), все астрономические данные хранятся в формате CSV. Как показала практика, скорость работы базы данных ниже по сравнению со скоростью работы при прямом чтении, обработке и записи файлов в формате CSV.

Для чтения каталогов примеряется процедура быстрого чтения файлов CSV “read_csv” из библиотеки DASK для Python. Поиск групп источников выполняется с помощью процедуры DBSCAN из библиотеки “sklearn” для Python. Поиск соседей выполняется с помощью процедуры “match_coordinates_sky” из пакета “astropy” для Python. Из этого же пакета используются процедуры для преобразования координат. Для загрузки каталогов используется сервис TAP Vizier [41], а визуализация выбранной области на небе выполняется с помощью Aladin Lite API [42]. Построения карты покрытия любого каталога выполняется с помощью сервиса MOCServer [39]. Загрузка данных происходит в фоновом асинхронном режиме с помощью AJAX-библиотеки SACK[4] (Simple Ajax Code Kit).

[4] http://tuandc.0fees.net/Sack/

4. АНАЛИЗ РЕЗУЛЬТАТОВ КРОСС-ИДЕНТИФИКАЦИИ

4.1. Сравнение качества кросс-идентификации

Существуют определенные работы, в которых авторы выполнили кросс-идентификацию различных крупных каталогов с учетом неопределенности положения источников в них. К таким работам относится, к примеру Gaia DR2×WISE [43], Gaia×IPHAS/KIS [44], TESS Input catalog [45]. Во многих каталогах уже встроены кросс-идентификации с другими каталогами. К примеру, в каталоге Gaia [23] представлены кросс-идентификации с бо́льши́м числом других каталогов.

Возникает вопрос: насколько кросс-идентификация, выполненная с помощью DBSCAN, отличается от кросс-идентификации, выполненной в данных работах? В качестве основы мы использовали результаты кросс-идентификации каталогов WISE AllSky и Gaia DR2 в работе Вильсона [43] (далее W18), где для кросс-идентификации учитывалась функция распределения сигнала для точечного источника (Point Spread Function, PSF).

Для сравнения выбраны направление (l, b) = = (150.0°, 0.0°) и область с радиусом 1.5°. В этой области в каталоге WISE содержится 160 598 источников, в каталоге Gaia DR2 – 332 232 источника. Проведена кросс-идентификация каталогов с помощью метода DBSCAN, описанного в настоящей работе. Для DBSCAN использован порог $\epsilon = 0.5''$, при котором 99.96% источников Gaia являются изолированными относительно других источников Gaia. Для расширенного поиска кросс-идентификаций использовался радиус поиска 3″ для каталога WISE, который соответствует максимальному угловому расстоянию между источниками Gaia и WISE в работе W18 по выбранной области. В результате было найдено 144 107 ассоциаций, в том числе 84 222 с помощью DBSCAN и 59 886 с помощью расширенного поиска. Далее были загружены результаты кросс-идентификации каталогов WISE и Gaia DR2 из работы W18 в этой области – 127 580 ассоциаций, что на 11% меньше – это связано с применением фильтров в их выборке данных (см. [43, табл. 2]). После этого был произведен поиск таких ассоциаций, которые отсутствуют в выборке DBSCAN, но присутствуют в работе W18. Таких источников было найдено всего 7. Эти источники были рассмотрены более подробно. Оказалось, что причина их отсутствия не связана с работой DBSCAN. 4 источника WISE, указанные в работе W18 (J040553.96+505611.2, J035507.62+525453.0, J035517.09+515412.8, J035446.14+524153.9), отсутствовали в каталоге WISE All-Sky в архиве Страсбургского центра данных, а для 3 источников (J040107.96+512310.9, J040300.46+523759.4, J035836.69+520122.8) из-за пересвета по данным

WISE в работе W18 ассоциация Gaia была указана неверно – вместо ближайшего источника Gaia был ассоциирован следующий по счету ближайший источник. Таким образом, с помощью метода DBSCAN все возможные ассоциации были найдены.

Установлено, что в 98.1% случаев источники WISE ассоциируются с идентичными источниками Gaia в двух рассматриваемых методах. Визуальный осмотр оставшихся 1.9% случаев показал, что разность в выборе источника Gaia возникает в том случае, когда по данным Gaia в некоторой области имеется два или более источников, а по данным WISE имеется лишь один неразрешенный источник. В результате выбор источника Gaia для такого неразрешенного источника WISE становится неоднозначным, поэтому возникает разница в результатах кросс-идентификации.

4.2. Тестирование производительности

Вопрос о производительности и эффективности использования ресурсов неизбежно возникает при выполнении кросс-идентификации различных каталогов. Подобно работе [13], для тестирования метода DBSCAN выбрано направление $(l, b) = (45°, 0°)$, размер области сравнения $5° \times 2°$. Кросс-идентификация выполнена для каталогов UKIDSS GPS (DR6) и GLIMPSE. Для тестирования количество источников GLIMPSE было ограничено 1 млн., а количество источников UKIDSS варьировалось от 1 тыс. до 10 млн. Исходные данные для тестирования аналогичны [13, панель (d) на рис. 7]. Кросс-идентификация DBSCAN проводилась с величиной $\epsilon = 0.5''$.

Работа программы для кросс-идентификации складывается из следующих этапов: чтение каталогов, выполнение процедуры DBSCAN, поиск соседей для всех источников (NN, опционально), пост-обработка и запись результатов в файл. Наиболее затратные по времени этапы – это DBSCAN и поиск соседей (NN). Все остальные этапы составляют 20–30% от общего времени выполнения всех этапов работы программы. При тестировании учитывалось только время на выполнение алгоритмов DBSCAN и поиска соседей, другие этапы не учитывались аналогично работе [13].

В первом подходе тестировался только алгоритм DBSCAN. Во втором подходе (DBSCAN+NN) дополнительно к DBSCAN тестировался алгоритм поиска соседей для всех источников. Результаты представлены на рис. 6. Для сравнения на рис. 6 представлены результаты скорости работы алгоритмов C^3 и STILTS/TopCat, полученные из работы [13] по тем же параметрам [13, панель (d) на рис. 7].

Анализ времени выполнения показал, что алгоритм DBSCAN выполняется в 5–9 раз быстрее, чем STILTS/TopCat, и в 3–5 раз быстрее, чем C^3 для количества источников $\mathrm{Nrows}_{\mathrm{Catalog1}} > 10^5$. Наиболее существенная разница проявляется при большем количестве источников: при $\mathrm{Nrows}_{\mathrm{UKIDSS}} = 10^7$ скорость выполнения алгоритма C^3 составляет ~600 с, а DBSCAN – 110 с. Если к выполнению алгоритма DBSCAN добавить выполнение алгоритма поиска соседей (NN), то разница несколько сократится, но тем не менее в целом ситуация не изменится: DBSCAN+NN быстрее C^3 в 1.5–3 раза при $\mathrm{Nrows}_{\mathrm{Catalog1}} > 10^{5.5}$, и быстрее STILTS/TopCat в 3–5 раз. При этом у алгоритма C^3 есть преимущество при $\mathrm{Nrows}_{\mathrm{Catalog1}} < 10^5$.

Таким образом, DBSCAN показал хорошие результаты по производительности, особенно при анализе большого числа источников (от 100 тыс. до 10 млн. и более). При этом DBSCAN не требует наличия GPU и не требует распараллеливания вычислений. Пиковое использование оперативной памяти составило 5.7 Гб при кросс-идентификации ~10 млн. источников.

5. ЗАКЛЮЧЕНИЕ

В настоящей работе рассмотрены различные методики кросс-идентификации астрономических каталогов и особенности их применения для решения широкого круга задач. В настоящее время существуют эффективные инструменты для кросс-идентификации двух каталогов, но работа с кросс-идентификацией одновременно трех и более каталогов с большим количеством источником детально не рассматривалась. В настоящей работе рассматриваются возможности метода DBSCAN для кросс-идентификации множества каталогов по выбранной области на небе. В результате анализа установлено следующее.

1. Преимуществами DBSCAN являются асимметричность алгоритма и возможность сравнения большого числа каталогов и источников без ущерба для скорости работы алгоритма: решающим значением для скорости является общее количество источников во всех каталогах вне зависимости от количества сравниваемых каталогов.

2. Недостатком метода DBSCAN является фиксированное значение порога ϵ для кросс-идентификации. В случае сравнения каталогов с существенно разной пространственной плотностью источников одного метода DBSCAN недостаточно, чтобы найти все возможные кросс-идентификации.

Для обхода последнего недостатка в настоящей работе была разработана методика, которая позволяет эффективно комбинировать результаты работы DBSCAN с классической методикой

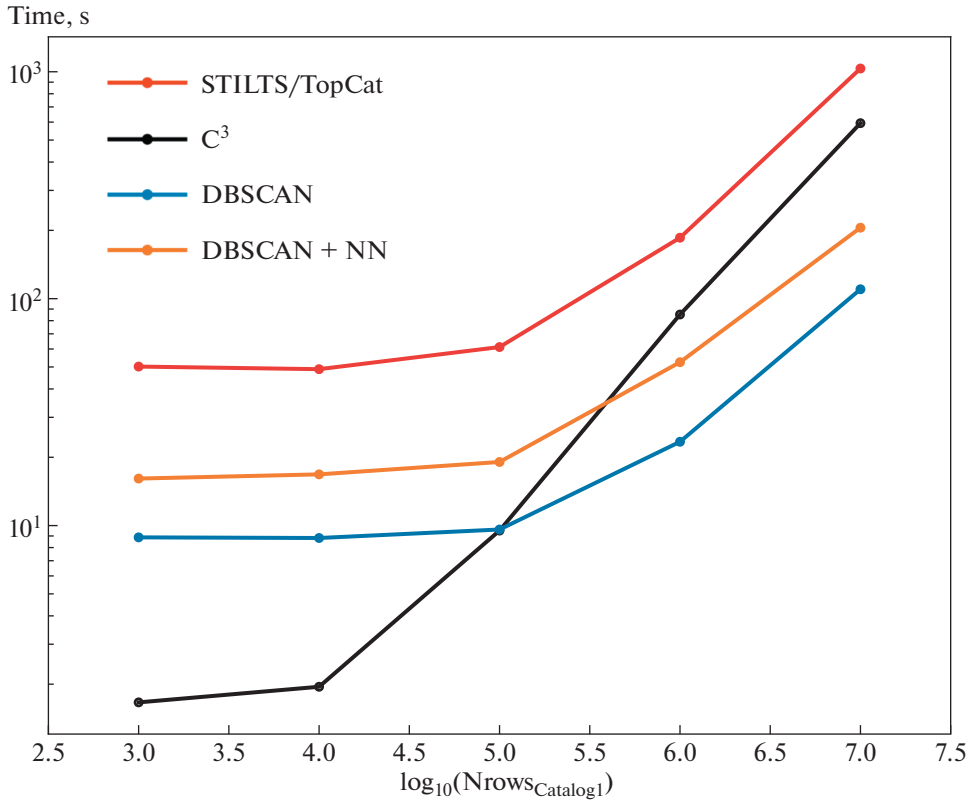

Рис. 6. Зависимость времени выполнения кросс-идентификации от количества источников в каталоге сравнения (UKIDSS). Для исходного каталога (GLIMPSE) число источников фиксировано и составляет 1 млн. Значения времени выполнения для методов C^3 и STILTS/TopCat получены из работы [13]. Значения для методов DBSCAN и DBSCAN+NN рассчитаны в настоящей работе.

кросс-идентификации. Она состоит из нескольких шагов.

1. Для каждого каталога нужно оценить верхний предел группировки $\epsilon_{\max,i}$, при котором значительное число источников не будет объединяться в группы со своими ближайшими соседями из одного каталога.

2. Выполнить алгоритм DBSCAN со значением порога $\epsilon < \epsilon_{\max,i}$.

3. Для источников, которые по выполнении DBSCAN были отмечены как “изолированные” (т.е. не имеющие соседей при выбранном значении ϵ), выполнить дополнительную проверку ближайших соседей из других каталогов на возможность их кросс-идентификаций с “изолированными” источниками. Проверка возможна как по фиксированным значениям неопределенности положения источников, так и по эллипсу неопределенности положения каждого отдельного источника.

В целом данная методика позволяет выполнять кросс-идентификацию сразу большого числа каталогов с минимальными затратами по времени. Результаты могут быть использованы в том числе для построения и моделирования спектрального распределения энергии для множества объектов.

Описанный в настоящей работе инструмент доступен в режиме онлайн и может быть использован для оценки выполнения наблюдательных программ на наземных и космических инструментах, в том числе для будущих внеатмосферных обсерваторий Спектр-УФ [46] и Миллиметрон [47]. Открытый доступ к системе предоставлен по адресу cross-match.online[5].

ФИНАНСИРОВАНИЕ

Работа выполнена при поддержке проекта Министерства науки и высшего образования “Теоретические и экспериментальные исследования формирования и эволюции внесолнечных планетных систем и характеристик экзопланет” (№ 075-15-2020-780, контракт 780-10).

[5] https://cross-match.online

СПИСОК ЛИТЕРАТУРЫ

1. *O. Malkov, O. Dluzhnevskaya, S. Karpov, E. Kilpio, A. Kniazev, A. Mironov, and S. Sichevskij*, Open Astronomy **21**, 319 (2012).
2. *A. Vallenari, A. G. A. Brown, T. Prusti, J. H. J. de Bruijne, et al.*, arXiv:2208.00211 [astro-ph.GA](2022).
3. *H. A. Flewelling, E. A. Magnier, K. C. Chambers, J. N. Heasley, et al.*, Astrophys. J. Suppl. Ser. **251**, id. 7 (2020), arXiv:1612.05243 [astro-ph.IM].
4. *S. Dye, A. Lawrence, M. A. Read, X. Fan, et al.*, Monthly Not. Roy. Astron. Soc. **473**, 5113 (2018), arXiv:1707.09975 [astro-ph.IM].
5. *R. Ahumada, C. A. Prieto, A. Almeida, F. Anders, et al.*, Astrophys. J. Suppl. Ser. **249**, id. 3 (2020), arXiv:1912.02905 [astro-ph.GA].
6. *Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, et al.*, Astrophys. J. **873**, id. 111 (2019), arXiv:0805.2366 [astro-ph].
7. *K. G. Stassun, R. J. Oelkers, M. Paegert, G. Torres, et al.*, Astron. J. **158**, id. 138 (2019), arXiv:1905.10694 [astro-ph.SR].
8. *D. Elia, M. Merello, S. Molinari, E. Schisano, et al.*, Monthly Not. Roy. Astron. Soc. **504**, 2742 (2021), arXiv:2104.04807 [astro-ph.GA].
9. *G. Landais and F. Ochsenbein*, in *Astronomical Data Analysis Software and Systems XXI*, edited by P. Ballester, D. Egret, and N. P. F. Lorente, Astron. Soc. Pacific Conf. Ser. **461**, 383 (2012).
10. *M. B. Taylor*, in *Astronomical Data Analysis Software and Systems XIV*, edited by P. Shopbell, M. Britton, and R. Ebert, Astron. Soc. Pacific Conf. Ser. **347**, 29 (2005).
11. *M. B. Taylor*, in *Astronomical Data Analysis Software and Systems XV*, edited by C. Gabriel, C. Arviset, D. Ponz, and S. Enrique, Astron. Soc. Pacific Conf. Ser. **351**, 666 (2006).
12. *F. Bonnarel, P. Fernique, O. Bienaymé, D. Egret, et al.*, Astron. and Astrophys. Suppl. Ser. **143**, 33 (2000).
13. *G. Riccio, M. Brescia, S. Cavuoti, A. Mercurio, A. M. di Giorgio, and S. Molinari*, Publ. Astron. Soc. Pacific **129**(972), 024005 (2017), arXiv:1611.04431 [astro-ph.IM].
14. *T. Budavari and M. A. Lee*, *Xmatch: GPU Enhanced Astronomic Catalog Cross-Matching*, Astrophys. Source Code Library, record ascl:1303.021 (2013).
15. *O. Malkov and S. Karpov*, in *Astronomical Data Analysis Software and Systems XX*, edited by I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots, Astron. Soc. Pacific Conf. Ser. **442**, 583 (2011).
16. *X. Jia, Q. Luo, and D. Fan*, in *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, p. 617 (2015).
17. *K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann*, Astrophys. J. **622**, 759 (2005), arXiv:astro-ph/0409513.
18. *P. Z. Kunszt, A. S. Szalay, and A. R. Thakar*, in *Mining the Sky*, edited by *A. J. Banday, S. Zaroubi, and M. Bartelmann*, p. 631 (2001).
19. *F. X. Pineau, T. Boch, and S. Derriere*, in *Astronomical Data Analysis Software and Systems XX*, edited by I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots, Astron. Soc. Pacific Conf. Ser. **442**, 85 (2011).
20. *C. Motch* and Arches Consortium, in *Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV)*, edited by A. R. Taylor and E. Rosolowsky, Astron. Soc. Pacific Conf. Ser. **495**, 437 (2015).
21. *M. Ester, H.-P. Kriegel, J. Sander, and X. Xu*, in *Proc. of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI Press, 1996), p. 226.
22. *A. Bayo, C. Rodrigo*, D. Barrado Y Navascués, E. Solano, R. Gutiérrez, M. Morales-Calderón, and F. Allard, Astron. and Astrophys. **492**, 277 (2008), arXiv:0808.0270 [astro-ph].
23. *P. M. Marrese, S. Marinoni, M. Fabrizio, and G. Altavilla*, Astron. and Astrophys. **621**, id. A144 (2019), arXiv:1808.09151 [astro-ph.SR].
24. *C. Wolf, C. A. Onken, L. C. Luvaul, B. P. Schmidt, et al.*, Publ. Astron. Soc. Australia **35**, id. e010 (2018), arXiv:1801.07834 [astro-ph.IM].
25. *A. Lawrence, S. J. Warren, O. Almaini, A. C. Edge, et al.*, Monthly Not. Roy. Astron. Soc. **379**, 1599 (2007), arXiv:astro-ph/0604426.
26. *M. F. Skrutskie, R. M. Cutri, R. Stiening, M. D. Weinberg, et al.*, Astron. J. **131**, 1163 (2006).
27. *E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer, M. E. Ressler, et al.*, Astron. J. **140**, 1868 (2010), arXiv:1008.0031 [astro-ph.IM].
28. *D. Ishihara, T. Onaka, H. Kataza, A. Salama, et al.*, Astron. and Astrophys. **514**, id. A1 (2010), arXiv:1003.0270 [astro-ph.IM].
29. *L. Bianchi, B. Shiao, and D. Thilker*, Astrophys. J. Suppl. Ser. **230**, id. 24 (2017), arXiv:1704.05903 [astro-ph.GA].
30. *V. N. Yershov*, Astrophys. Space Sci. **354**, 97 (2014).
31. *J. F. Carbary, E. H. Darlington, T. J. Harris, P. J. McEvaddy, M. J. Mayr, K. Peacock, and C. I. Meng*, Applied Optics **33**, 4201 (1994).
32. *A. A. Henden, S. Levine, D. Terrell, and D. L.Welch*, Amer. Astronomical Society, AAS Meeting, Abstracts **225**, 336.16 (2015).
33. *G. Barentsen, H. J. Farnhill, J. E. Drew, E. A. González-Solares, et al.*, Monthly Not. Roy. Astron. Soc. **444**, 3230 (2014), arXiv:1406.4862 [astro-ph.SR].
34. *N. Epchtein, B. de Batz, L. Capoani, L. Chevallier, et al.*, Messenger **87**, 27 (1997).
35. *C. A. Beichman, G. Helou, and D. W. Walker*, *Infrared Astronomical Satellite (IRAS) Catalogs and Atlases, Vol. 7: The Small Scale Structure Catalog* (Scientific and Technical Information Division, National Aeronautics and Space Administration, 1988).
36. *M. Kawada, H. Baba, P. D. Barthel, D. Clements, et al.*, Publ. Astron. Soc. Japan **59**, S389 (2007), arXiv:0708.3004 [astro-ph].
37. *E. Churchwell, B. L. Babler, M. R. Meade, B. A. Whitney, et al.*, Publ. Astron. Soc. Pacific **121**, 213 (2009).
38. *T. Csengeri, J. S. Urquhart, F. Schuller, F. Motte, et al.*, Astron. and Astrophys. **565**, id. A75 (2014), arXiv:1312.0937 [astro-ph.GA].
39. *P. Fernique, T. Boch, A. Oberto, and F. X. Pineau*, in *Astronomical Data Analysis Software and Systems XXV*, edited by N. P. F. Lorente, K. Shortridge, and R. Wayth, Astron. Soc. Pacific Conf. Ser. **512**, 133 (2017), arXiv:1611.01374 [astro-ph.IM].

40. *M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, et al.*, Astron. and Astrophys. Suppl. Ser. **143**, 9 (2000), arXiv:astro-ph/0002110.

41. *G. Landais, F. Ochsenbein, and A. Simon*, in *Astronomical Data Analysis Software and Systems XXII*, edited by D. N. Friedel, Astron. Soc. Pacific Conf. Ser. **475**, 227 (2013).

42. *T. Boch and P. Fernique*, in *Astronomical Data Analysis Software and Systems XXIII*, edited by N. Manset and P. Forshay, Astron. Soc. Pacific Conf. Ser. **485**, 277 (2014).

43. *T. J. Wilson and T. Naylor*, Monthly Not. Roy. Astron. Soc. **481**, 2148 (2018), arXiv:1809.00018 [astro-ph.SR].

44. *S. Scaringi, C. Knigge, J. E. Drew, M. Monguió, et al.*, Monthly Not. Roy. Astron. Soc. **481**, 3357 (2018), arXiv:1809.04086 [astro-ph.SR].

45. *M. Paegert, K. G. Stassun, K. A. Collins, J. Pepper, G. Torres, J. Jenkins, J. D. Twicken, and D. W. Latham*, arXiv:2108.04778 (2021).

46. *B. M. Shustov, M. E. Sachkov, S. G. Sichevsky, R. N. Arkhangelsky, et al.*, Solar System Res. **55**, 677 (2021).

47. *N. S. Kardashev, I. D. Novikov, V. N. Lukash, S. V. Pilipenko, et al.*, Physics Uspekhi **57**, 1199 (2014), arXiv:1502.06071 [astro-ph.IM].