РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

На правах рукописи

Ермолаева Анна Михайловна

Механизмы кумулятивного преимущества в наукометрии

Специальность 1.2.2. Математическое моделирование, численные методы и комплексы программ

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель
д.ф.-м.н.
Д. С. Кулябов

Москва — 2026

Оглавление

Введение

Актуальность темы исследования

В области компьютерных наук и ряда смежных дисциплин конференции являются основным каналом представления и публикации научных результатов — более 60% работ выходят в трудах конференций. В то же время не существует единой общепринятой метрики для оценки качества конференций. Существующие рейтинги (CORE, CCF, Qualis) носят региональный или экспертный характер и не всегда отражают объективную наукометрическую картину. Особую проблему представляет прогнозирование рейтинга новых конференций, для которых еще не накоплены данные по цитированию. Кроме того, в наукометрии конференций проявляется эффект кумулятивного преимущества (закон Матфея), когда уже известные площадки получают непропорционально большее внимание и цитирование. Диссертационная работа посвящена развитию мультимодельного подхода к моделированию наукометрических показателей, в частности динамики научных конференций, что является актуальной научной задачей.

Исследование влияния конференций играет важную роль в оценке научного вклада ученых в компьютерных науках. Это обусловлено тем, что большая часть результатов исследований представляется на конференциях и, затем, публикуется в сборниках трудов конференций [45]. Из-за этого возникает интерес к разработке новых методов оценки качества конференций, как метрик, для оценки качества конференций, так и рейтингов конференций.

Также интерес к исследованию ранжирования конференций обусловлен и тем, что не существует единой метрики для их оценивания, как, например, журнальный импакт-фактор. Существует несколько рейтингов конференций, но большинство из них региональные, например австралийский рейтинг CORE, рейтинг конференций китайской компьютерной федерации CCF. Так как эти рейтинги региональные, то они ориентированы на рекомендации для исследователей конкретного региона, CORE для ученых из Австралии, а CCF для ученых из Китая. Они составлены экспертами на основе наукометрических показателей. Также была предпринята попытка Microsoft Academic составить универсальный рейтинг конференций по областям (Microsoft Academic’s field ratings for conferences), но на данный момент этот рейтинг удален с сайта.

Также проблема оценивания конференций обусловлена тем, что участие в конференциях связано с некоторыми трудностями, например внесение платы за участие, траты на дорогу и проживание. Кроме того, конференции не так высоко ценятся, как журнальные статьи при оценке трудов ученых как государством, так и ведущей организацией, а в такой области как компьютерные науки, ошибочно недооценивать влияние конференций на научный вклад.

Еще одной проблемой в исследовании ранжирования конференций является присвоения рейтинга новой конференции. Новые конференции появляются довольно часто и их не сразу включают в рейтинги, поскольку некоторые наукометрические показатели невозможно рассчитать до появления результатов публикации трудов конференций, основанные на цитировании. Срок появления первых результатов может достигать довольно большого временного периода, в некоторых случаях до 60 лет [30]. Для анализа возраста цитирования был введен показатель «cited half life», который показывает с какого «возраста» публикации его начинают цитировать. Для каждой области исследований этот показатель разный, например [76] рассчитали его для журналов из разных областей, и для естественнонаучной области в среднем составил 8,5 лет, а для медицинской – 6 лет. Следовательно, чтобы рассчитать рейтинг новой конференции, необходимо достаточно много времени. Отсюда следует необходимость «предсказания» или прогнозирования рейтинга конференции на основе данных, которые можно получить сразу после проведения первой конференции, такие как количество публикаций, количество членов программного комитета, их средняя цитируемость на момент проведения конференции и так далее.

Степень разработанности темы

Проблема оценки качества и ранжирования научных конференций привлекает внимание исследователей в области наукометрии, особенно в контексте компьютерных наук, где материалы конференций являются основным каналом научной коммуникации. Существенный вклад в анализ роли конференций внесли работы М. Франчесчета (M. Franceschet), Л. Мехо (L. I. Meho), Г. Вреттаса и М. Сандерсона (G. Vrettas, M. Sanderson), в которых показано, что более 60% результатов в информатике публикуются в трудах конференций,

а ведущие конференции по цитируемости сопоставимы с журналами первого квартиля.

Практические подходы к ранжированию конференций реализованы в виде региональных и отраслевых рейтингов: австралийский CORE (создан при участии Л. Батлер и др.), китайский CCF, бразильский Qualis, а также ныне не поддерживаемый Microsoft Academic. В академической литературе активно обсуждаются методы построения таких рейтингов: экспертные, библиометрические (на основе h-индекса, импакт-фактора Google Scholar, CiteScore Scopus) и гибридные. Работы С. Эффенди и Р. Япа (S. Effendy, R. H. C. Yap), И. Джахья и соавторов (I. Jahja et al.) посвящены применению алгоритмов случайного блуждания и графовых методов для выявления взаимосвязей между конференциями. Возможности машинного обучения для прогнозирования рейтингов рассматривались в исследованиях В. Альмендры (V. S. Almendra), Г. Чоудхури (G. R. Chowdhury) и П. Удупи (P. K. Udupi).

Отдельное направление составляют работы, посвящённые проявлению кумулятивного преимущества (закона Матфея) в наукометрии. Теоретические основы этого феномена заложены в трудах Р. Мертона (R. K. Merton), а его количественное моделирование в контексте цитирования и научной конкуренции развивалось в исследованиях А.-Л. Барабаши (A.-L. Barabási), Д. Прайса (D. J. de Solla Price) и их последователей.

Вместе с тем, в существующем научном заделе остаётся ряд нерешённых вопросов. Существующие математические модели динамики рейтинга конференций обычно не учитывают механизмы кумулятивного преимущества. Недостаточно разработаны методы прогнозирования рейтинга новых конференций по данным, не зависящим от времени накопления цитирований. Существующие метрики, как правило, не позволяют оценить, насколько конференция способствует продвижению результатов исследователей из конкретной страны.

Настоящая диссертационная работа направлена на восполнение указанных пробелов путём модификации модели конкуренции конференций путём явного введения кумулятивного преимущества, введения нового странового показателя MNCS и построения прогностических статистических моделей на основе дискриминантного анализа.

Цели исследования

Цель диссертационной работы является разработка гибридного мультимодельного подхода к моделированию оценки качества научных конференций и выявление эффективного метода прогнозирования рейтинга конференций на основе данных, не привязанных к временной шкале, с учётом эффекта кумулятивного преимущества.

Задачи исследования

— Модификация математической модели динамики рейтинга конференций с учётом кумулятивного преимущества, конкуренции и внешних воздействий.
— Введение наукометрического показателя, позволяющего оценить, насколько публикация на конкретной конференции для исследователя из определённой страны превзойдёт ожидаемый уровень цитирования; апробация метрики на примере России, Китая и США.
— Построение статистических моделей (линейная регрессия, дискриминантный анализ, нейронная сеть) для прогнозирования квартиля конференции по данным, доступным сразу после проведения мероприятия (количество участников, средняя цитируемость оргкомитета и др.).
— Сравнительный анализ методов прогнозирования и выбор наилучшей модели.

Научная новизна

— Разработан метод описания динамики развития научных конференций на основе модифицированной модели Ферхюльста с учётом кумулятивного преимущества и асимметричной конкуренции.
— Введён операциональный показатель кумулятивного преимущества в наукометрии, позволяющий моделировать эффект «победитель получает всё».
— Предложен и апробирован новый показатель MNCS (средняя нормализованная цитируемость) для ожидаемого уровня цитирования в зависимости от страны аффилиации автора.

— Впервые для прогнозирования рейтинга конференций применён дискриминантный анализ по данным, не зависящим от времени (без показателей цитирования).
— Построены регрессионные функции и нейросетевая модель, проведено их сравнение с дискриминантным подходом.

Теоретическая и практическая значимость работы

Предложенные модели и метрики могут использоваться исследователями и научными организациями для обоснованного выбора конференций с целью повышения видимости результатов. Организаторы конференций получают инструмент ранней оценки потенциального рейтинга. Разработанные программные комплексы зарегистрированы в качестве объектов интеллектуальной собственности.

Методы исследования

В работе используются методы математического моделирования, качественной теории дифференциальных уравнений, статистического анализа (корреляционный, регрессионный, дискриминантный анализ, проверка статистических гипотез), элементы машинного обучения (нейронная сеть типа многослойный персептрон). Данные извлечены из международных наукометрических баз Scopus, CORE, Microsoft Academic.

Положения, выносимые на защиту

— Модифицирована модель динамики рейтинга научных конференций путём включения механизмов кумулятивного преимущества и асимметричной конкуренции.
— Впервые применён метод прогнозирования квартиля новой конференции на основе дискриминантного анализа с использованием независимых от времени показателей.
— Получены результаты сравнительного анализа трёх статистических методов (регрессия, дискриминантный анализ, нейронная сеть) для задачи

ранжирования конференций, показывающие преимущество дискриминантного подхода по точности классификации (84,6% совпадений с экспертными рейтингами).

— Получены эмпирически обоснованные рекомендательные списки конференций по искусственному интеллекту для исследователей из России, Китая и США на основе введённого показателя MNCS.

Степень достоверности результатов проведённых исследований

Степень достоверности результатов обеспечивается корректным применением математического аппарата, методов теории кинетических уравнений, математической статистики, методов дифференциальных уравнений при выводе аналитических соотношений и доказательстве утверждений; валидацией моделей на реальных данных Scopus, использованием статистических критериев значимости, сравнением с независимыми экспертными рейтингами, регистрацией программ для ЭВМ. обоснованностью принятых допущений, перекрестной верификацией методов, численными экспериментами с применением численного анализа, валидацией математических моделей.

Апробация работы

Основные результаты докладывались на международных конференциях: AIST (Москва, 2020), DCCN (Москва, 2021), «Математическое и программное обеспечение информационных, технических и экономических систем» (Томск, 2020).

Основные результаты опубликованы в ведущих научных сборниках и журналах – Lecture Notes in Computer Science, Communications in Computer and Information Science, Discrete and Continuous Models and Applied Computational Science, Heliyon, а также в трудах международных конференций, индексируемых WoS (Web of Science), Scopus и РИНЦ.

Зарегистрировано два объекта интеллектуальной собственности на программы ЭВМ.

Также основные результаты докладывались на научном семинаре «Математическое моделирование» РУДН.

Личный вклад соискателя

Все результаты, представленные в диссертационной работе, получены автором самостоятельно. В публикациях, выполненных в соавторстве, личный вклад соискателя выражается в исследовании математических моделей и методов их анализа, доказательстве положений, разработке алгоритмов и создании программных инструментов для проведения численных экспериментов. Программное обеспечение, применяемое при численном и графическом анализе, создано непосредственно автором.

Соответствие паспорту специальности

Диссертационное исследование соответствует следующим разделам паспорта специальности 1.2.2. Математическое моделирование, численные методы и комплексы программ, а именно:

— п. 1. «Разработка новых математических методов моделирования объектов и явлений» в части адаптации кинетических моделей к разработке описания динамики развития научных конференций.
— п. 5. «Разработка новых математических методов и алгоритмов валидации математических моделей объектов на основе данных натурного эксперимента или на основе анализа математических моделей» в части развития операциональных методов анализа исследуемых моделей.
— п. 8. «Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента» в части применения мультимодельного подхода.

Публикации

Основные результаты, выводы и рекомендации диссертационного исследования отражены в 3 работах [23; 38; 63], в том числе в изданиях, входящих в базу данных Scopus, Web of Science, список ВАК категорий К-1, К-2 и в 2 свидетельствах о государственной регистрации программ для ЭВМ [83; 86]. А также в других рецензируемых изданиях [22]. Авторский вклад 87%.

Глава 1. Анализ теории и практик измерения качества конференций

1.1. Анализ существующих рейтингов и введение новых метрик

Проблема оценки конференций сейчас очень актуальна для ученых в области наукометрии, поскольку не существует универсальной методологии оценки всех конференций во всех областях. Существует несколько рейтингов конференций, таких как Австралийский CORE, китайский рейтинг конференций CCF, бразильский QUALIS, отраслевые рейтинги конференций Microsoft Academic. Все эти рейтинги составлены для конференций в области компьютерных наук, это связано с тем, что конференции чрезвычайно важны для этой области, поскольку более 60% результатов исследований публикуются в материалах конференций [45]. Далее подробно рассмотрим, как ранжируются конференции в каждом рейтинге.

Рейтинг CORE составляется экспертами на основе следующих показателей: цитируемость документа, с учетом места проведения конференции, на основе данных с различных ресурсов, таких как Google Scholar и Elsevier; участие в конференции ведущих ученых, на основе данных h-index и Elsevier; вовлеченность и уровень членов программного комитета, также на основе индекса Elsevier h-index; показатели принятия [13].

Рейтинг конференции составленный Китайской Компьютерной Федерацией (CCF) составлен экспертами и включает наиболее значимые конференции по компьютерным наукам по разделам для ученых из Китая. Рейтинг разделен на уровни A, B и C и пронумерованы, так что конференция может иметь рейтинг, например, B(8).

Qualis (2012) — это бразильский рейтинг конференций, который использует h-индекс в качестве критерия ранжирования.

Рейтинг MSAR составляется с использованием алгоритма, который позволяет узнать информацию об объектах графа. В открытом доступе не удалось найти информацию о том, какие данные и какой алгоритм используется для составления этого рейтинга, но, изучив веб-сайт Microsoft [50], можно предположить, что рейтинг основан на следующих данных: количество выпущенных документов (статей, тезисов и т.д.), объем памяти, ссылки (возраст статей,

цитируемых в материалы конференции, опубликованные в этом году), исходящие ссылки, входящие цитаты, наиболее цитируемые авторы, ведущие университеты и ведущие авторы.

В открытом доступе нет информации о том, как в настоящее время составляется рейтинг конференций Google Scholar. Ученые в статье [6] выполнили первые шаги по реинжинирингу алгоритма ранжирования Google Scholar. Результаты их исследования показали, что рейтинг в наибольшей степени основан на цитировании, также учитывается возраст статей, новым статьям придается больший вес, чтобы уравнять их со старыми статьями.

Согласно описанию представленных рейтингов и методов их составления, можно сделать вывод, что они составляются тремя способами, либо экспертной группой, либо на основе наукометрических показателей (h-индекс, цитирования), либо экспертными группами на основе наукометрических показателей. Сложность заключается в том, что рейтинги составляются для определенных научных сообществ и затрагивают географический аспект. Таким образом, CORE в основном включает конференции, которые проходят в Австралии и США, CCF включает крупные конференции в Азии, конференции Qualis в Латинской Америке. Только MSAR пытался составить рейтинг, не опираясь на географию конференций, но с 2022 года рейтинг был удален с официального академического сайта Microsoft.

Такие рейтинги, как CORE, CCF и Quails, могут влиять на поведение ученых при выборе конференций, и, например, ученые из Китая будут представлять свои работы только на азиатских конференциях, из Бразилии - на латиноамериканских конференциях, что в свою очередь может привести к замедлению научной коммуникации и торможению развития. По этой причине ученые, занимающиеся наукометрией, все еще пытаются разработать инструмент для оценки конференций.

Исследование рейтингов конференций в основном имеет два направления. Первое — это анализ существующих рейтингов. Второе направление — это создание собственных рейтингов или метрик или применение журнальных метрик к конференциям.

Авторы в статье [32] используют алгоритм случайного блуждания, основанный на базе данных DBLP, полученной в 2013 году. С помощью этого алгоритма была выбрана эталонная конференция, другие конференции, участвующие в исследовании, были сопоставлены с этой конференцией, таким образом,

фактически был построен рейтинг конференций с использованием дерева графов. Авторы рассмотрели набор наиболее рейтинговых сводных конференций, чтобы обеспечить хорошее освещение широкого круга конференций. Рейтинг взаимосвязанности конференции определяется как ее минимальный ранг по отношению к контрольным точкам. Для составления списка конференций авторы использовали рейтинги CORE и CCF, а также рейтинг Microsoft Academic search (LIBRA).

Результаты исследования показали, что используемый алгоритм дает хорошую корреляцию с рейтингами, включенными в исследование, как с CORE, так и с CCF.

Поскольку единой системы оценки конференций не существует, а существующие рейтинги являются как официальными, так и неофициальными, поэтому в [19] авторы изучают возможность расчета объективных показателей для оценки конференций, которые хорошо коррелировали бы с существующими рейтингами. Эта работа продолжает исследование [32]. Работа направлена на выявление взаимосвязанности между конференциями, для этого авторы используют адаптированный подход случайного блуждания с использованием Microsoft Academic Graph. Метод, представленный в этой статье, отличается тем, что он использует оценку с независимым пороговым значением и классифицирует конференции на основе набора комбинированных оценок с использованием двух комбинированных пороговых значений. Таким образом, сводная таблица конференций составляется для каждого поддомена computer science, и последующие конференции сравниваются в этой сводной таблице. Авторы также рассчитали коэффициент корреляции тау-Кендалла между рейтингами Libra и CCF, который показал довольно хорошую взаимосвязь между этими рейтингами, но по поводу некоторых конференций существуют разногласия.

Результаты этого исследования показывают, что составленный рейтинг имеет сильную корреляцию с рейтингом CCF, кроме того, нет разногласий в классификации конференций, т.е. рейтинг не классифицирует конференции CCF с рейтингом C, как A, как это было с рейтингами CCF и Libra. Это говорит о том, что представленный алгоритм может быть лучше, чем алгоритм, используемый при составлении рейтинга Libra. Исследование показывает, что как оценка взаимосвязанности, так и цитируемость хорошо коррелируют с рейтингом конференции, но оценка взаимосвязанности, по-видимому, лучше

справляется с ошибками, которые могут возникнуть при ранжировании по показателям, основанным на цитируемости.

В ходе исследования [73] авторы хотели выяснить, оказывают ли статьи из материалов конференций такое же влияние, как журнальные статьи, отличается ли практика публикации в информатике от других областей, влияет ли объем статьи на ее цитирование. Исследование было основано на базах данных ERA, MSAR и DBLP за 2007-2011 годы и 195 000 материалах конференций 108 000 журнальных статей. Для проведения исследования авторы составили алгоритм, который сравнивал названия журналов и конференций из рейтинга ERA и рейтинга MSAR. Также, используя этот алгоритм, статьи были разделены на три типа в зависимости от их длины. Далее, для оценки публикаций авторы использовали два показателя: среднее количество цитирований на статью и h-индекс за 5 лет, последний взят из Google Scholar. Для анализа был использован корреляционный анализ, в ходе которого было выявлено, что, если место публикации оценивается с использованием h5, существует риск того, что размер места публикации окажет заметное влияние на оценку; поэтому было решено измерять публикации, используя среднюю цитируемость на статью. Для дальнейшего анализа был применен статистический анализ.

Результаты исследования [73] показали, что конференции A* получили значительно более высокий средний показатель цитируемости, чем журналы A*. Конференции и журналы с рейтингом A не показали статистической разницы. Журналы с рейтингом B и C оказали значительно более высокое влияние на цитируемость по сравнению с конференциями с таким же рейтингом. Как для A*, так и для A, скорее всего, существует большее количество журнальных статей с низким процентом цитирования и относительно небольшим числом конференции с особенно высоким процентом цитирования. Следующий вывод, который был получен в ходе этого исследования, заключается в том, что объем статей не повлиял на показатели цитируемости.

В конце статьи [73] авторы сформулировали ответы на поставленные вопросы. Хотя существует разница между цитированием статей в журналах и на конференциях, она несущественна. Для небольшого числа элитных конференций средние показатели цитирования были заметно выше, чем для элитных журналов, но для остальных конференций эта тенденция была обратной. Также авторы не отрицают важности участия в любых конференциях, даже на

низком уровне, поскольку это ценный опыт для ученого, кроме того, конференции — это площадка для обсуждения идей и взаимодействия.

В работе [46] автор выяснял, достаточно ли хорош инструмент CiteScore для оценки конференций, эффективен ли он для оценки конференций по информатике, совпадают ли результаты метода CiteScore с экспертными рейтингами конференций. Поскольку Scopus по разным причинам не учитывает CiteScore для всех конференций, автор решил вручную рассчитать этот показатель для всех конференций, которые включены в базу данных Scopus. Для проведения этого исследования были взяты данные за 2013-2016 годы и за 395 конференций по информатике. Список конференций был составлен с использованием CCF, CORE, Google Scholar Metrics, рейтинга Microsoft Academic, а также конференций, отобранных двумя экспертными группами (CSRANKINGS5 и Рочестерским технологическим институтом).

Результаты этого исследования [46] показывают, что показатели цитируемости этих 154 конференций можно сравнить с журналами из верхнего квартиля, и 67 из них можно сравнить с 10% лучших журналов в каждой из рассматриваемых областей информатики. Более того, эти 154 конференции имеют несколько более высокие показатели цитируемости, чем ведущие журналы. Эти результаты подтверждают утверждение о том, что участие и последующая публикация в ведущих конференциях не менее важны и влиятельны, чем публикации в ведущих журналах.

Сравнение инструментария CiteScore с существующими рейтингами и экспертными оценками, которые были составлены для этого исследования, показало, что этот инструмент очень хорошо согласуется с рейтингами, особенно для топовых конференций. CiteScore также позволяет оценивать новые конференции и предоставляет равные возможности для всех конференций, независимо от их размера. Также важным преимуществом является относительная простота и прозрачность этого метода.

Авторы в [67] исследовали проблему влияния рейтингов конференций (CORE и CCF) на публикации авторов из разных стран. Данные были взяты из баз данных DBLP и MSAR для 700 000 статей с 521 конференции и 800 000 авторов из Китая, Австралии, США, Германии и Индии за период с 2005 по 2016 год. В ходе исследования была предпринята попытка выявить влияние таких факторов, как среднегодовое количество статей, опубликованных на конференции; количество авторов на статью; максимальный и средний h-индекс

всех авторов; Нормализованное количество статей (NPC) на картинке публикаций. Была предпринята попытка выявить преимущества и недостатки систем ранжирования. Авторы выявили следующие тенденции: за исследуемый период времени больше всего работ было произведено в Соединенных Штатах, а меньше всего - в Германии, в то время как в Индии наблюдался постоянный положительный рост. Также был рассчитан коэффициент Джини для конференций всех пяти стран, и рассмотрение этого коэффициента с течением времени показало тенденцию к сближению, то есть более низкие коэффициенты Джини увеличиваются, в то время как более высокие коэффициенты Джини уменьшаются. Авторы приходят к выводу, что такое изменение коэффициентов указывает на то, что исследователи по всему миру более единодушны в отношении репутации конференций, чем раньше. Авторы проанализировали конференции, используя такие методы, как модель отрицательной биномиальной регрессии со случайным эффектом (количество опубликованных статей, NPC и т.д.); регрессионный анализ. Кроме того, авторы ввели две новые метрики для этого анализа: Индекс трансформационной активности (TAI) (этот показатель был введен ранее, но для подобластей, и здесь авторы применили его к наборам конференций) и нормализованное количество статей (NPC).

Результаты этого анализа [67] показали, что рейтинг CCF оказывает высокое влияние на поведение китайских ученых и ученых из Индии, но не такое сильное. CORE также оказывает влияние на исследователей из Австралии, но не столь значительное, как в случае CCF. Таким образом, можно сказать, что исследуемые рейтинги оказывают влияние на публикации ученых, но его сила варьируется от страны к стране. Двумя общими факторами для всех стран являются место проведения конференции и среднее количество докладов, опубликованных на конференции. Все страны, как правило, публиковали больше докладов на конференциях, которые проводились внутри страны или имели более высокое среднее количество докладов. Это является следствием относительно низкой стоимости участия в местной конференции и относительно лучшей репутации крупных конференций. Авторы статьи отмечают, что азиатские страны расширяют свои международные научные связи и пытаются привлечь к участию ученых с высокой репутацией. Авторы также отмечают тот факт, что авторы из азиатских стран предпочитают публиковаться в составе научных групп, в то время как западные публикуются либо

самостоятельно, либо с одним соавтором. В ходе исследования авторы пришли к выводу, что CCF принес значительный прогресс Китаю, и влияние публикаций на ведущих конференциях улучшается. Список, опубликованный CCF, на протяжении многих лет служил ориентиром исследователям в Китае, и это ускорило процесс публикации большего количества статей на ведущих конференциях, а также улучшило качество этих публикаций.

В статье [58] авторы, основываясь на цитировании Google Scholar, вводят новую метрику, рассчитываемую как импакт-фактор ISI Web of Knowledge для оценки конференции по информатике. Также введенный показатель был сопоставлен с импакт-фактором ISI Web of Knowledge. В ходе исследования по тестированию новой метрики авторы протестировали большой объем публикаций на конференциях и в журналах (8 764), которые были взяты за период с 2005 по 2007 год. Результаты исследования показали следующее: конференции, включенные в исследование, показали хорошие результаты по сравнению с журналами; корреляция между импакт-фактором ISI и импакт-фактором Google Scholar высока; корреляция Пирсона между импакт-фактором Google Scholar и частотой отказов невысока. Авторы также утверждают, что импакт-фактор Google Scholar лучше, чем импакт-фактор ISI, поскольку он оценивает не только журналы, но и конференции.

В статье [17] авторы проанализировали три ведущих журнала и три лучшие конференции по компьютерному зрению и выяснили, были ли предыдущие статьи, опубликованные в ведущих журналах, основаны на статьях, опубликованных на ведущих конференциях. Исследование основано на данных о цитировании (2005 и 2007 годы) из Google Scholar и Microsoft Academic Search и времени между отправкой и принятием статьи, а также был проведен опрос. Данные о цитировании анализировались с использованием статистических методов: непараметрических тестов (критерий знакового ранга Уилкоксона, U-критерий Манна-Уитни), вероятностного ранжирования.

Результаты исследования [17] показали, что в среднем 30% статей из ведущих журналов основаны на материалах конференций. Результаты опроса показали, что ученые считают необязательным участие в конференциях перед публикацией в топовом журнале, но это условие увеличивает шансы на принятие статьи в журнал. Кроме того, около половины респондентов считают, что доклад на топовой конференции пользуется большим авторитетом и получает больше цитирований. Анализ цитирования показал, что журнальные

статьи получают больше цитирований, чем аналогичные или предыдущие доклады на конференции, и вероятность того, что статья в журнале получит столько же цитирований, сколько предыдущая, так же высока, как и наоборот. Подсчет цитирований без учета окон показывает, что в краткосрочной перспективе (3–5 лет) статьи конференций получают больше цитирований, чем последующая журнальная статья. Анализ времени между отправкой статьи и ее принятием показал, что публикация в топ-3 первой конференции не сокращает время процесса рецензирования журнальной статьи, основанной на этих предыдущих статьях.

Также для анализа рейтингов исследователи используют не только стандартные математические методы, такие как корреляционно-регрессионный анализ, статистические методы, теория графов, но и элементы машинного обучения, теории принятия решений, нейронных сетей, рекомендательные системы. В исследовании [54] на основе факторов (место проведения конференции, регистрационный взнос, язык конференции, тематика конференции и крайний срок подачи), с помощью процесса аналитической сети были ранжированы десять значимых научных конференций Бангладеша. После этого исследователи с помощью машинного обучения спрогнозировали рейтинг предстоящих конференций. Для определения лучшей конференции из всех участвующих в исследовании использовался алгоритм машинного обучения вместе с множественной линейной регрессией.

В статье [20] был предложен новый метод прогнозирования научных статей, который основан на одном из методов автоматической классификации, который применяется в научных исследованиях в качестве задачи обучения под наблюдением. Алгоритм основывался на наборе данных, включающий идентификатор статьи, оценку автора, количество опубликованных статей, среднюю скорость загрузки, среднее количество цитирований. Результаты этого исследования показывают, что использования алгоритма машинного обучения в дисциплине ранжирования является эффективным.

В исследовании [53] исследуется подход Quacquarelli Symonds для оценки глобальных рейтингов университетов и разрабатываются модели машинного обучения для прогнозирования глобальных рейтингов. Также в этой работе используется поисковый анализ данных для анализа набора данных, а затем оцениваются алгоритмы машинного обучения с использованием методов регрессии для прогнозирования глобального рейтинга.

1.2. Важность влияния конференций на наукометрическую оценку результатов исследования

Во многих странах (например, в Китае [51], Индии [43], России, Турции [71], Великобритания [39]) оценка исследований основана на показателях источников, т.е. журналов, материалов конференций, серий книг, в которой публикуются результаты. Это часто приводит к присвоению источникам публикаций нескольких предопределенных классов и оценке важности публикации на основе класса источника. Поскольку во многих областях исследований оригинальные результаты публикуются в журналах [73], политика оценки исследований часто предвзята по отношению к журналам.

В качестве примера такой политики мы можем упомянуть продолжающееся обсуждение методики расчета качественного показателя государственного задания “Комплексный балл публикационной результативности” (КБПР) для научных организаций в России. Методология основана на оценке публикаций в зависимости от квартиля импакт-фактора (IF) журналов в Web of Science: журнальная статья, опубликованная в журнале первого квартиля, получает 20 баллов, во втором квартиле - 10 баллов, в третьем квартиле - 5 баллов, а в третьем квартиле - 5 баллов. Четвертый квартиль - 2,5 балла. Все остальные публикации, включая материалы конференций, главы книг, журнальные статьи, индексируемые только в Scopus или Российском индексе научного цитирования, получают 1 балл. Эта шкала применима к естественным наукам, инженерии и наукам о жизни. Для социальных и гуманитарных наук существует единая шкала: все публикации в Scopus или Web of Science получают 3 балла независимо от квартиля. К преимуществам этого метода относятся его простота и возможность оценить рукопись при повторном поиске непосредственно в момент публикации, поскольку процесс сбора цитат требует времени. Недостатки метода обсуждаются в Декларации об оценке исследований (DORA) [10] и включают возможность манипулирования количественными показателями журнала [18], хотя это справедливо для всех чисто количественных методов оценки. Однако основными недостатками методологии являются возможное несоответствие между количеством цитирований конкретной статьи и значением журнала, в котором она опубликована, и высокая разница в цитировании статей в одном и том же журнале [61]. Более поздние исследования [75] показывают, что, возможно, импакт-фактор является более точным

показателем значимости статьи, чем количество полученных ею цитирований. Однако дизайн исследования не позволяет авторам делать выводы о том, действительно ли на практике импакт-фактор является более точным, чем количество цитирований.

Импакт-фактор и другие показатели журнального уровня широко используются при оценке исследований, чтобы судить о влиянии вклада в исследование или самих исследователей. Если статья была опубликована в журнале первого квартиля, то, скорее всего, вероятность того, что она будет превосходной, выше, чем статья в журнале четвертого квартиля по той же дисциплине. Однако можно было бы найти контрпримеры, например, слабую статью в журнале первого квартиля или превосходную статью в журнале четвертого квартиля. Поэтому наряду с количественными показателями на уровне журнала следует использовать показатели уровня статьи или, что еще лучше, качественные методы исследования. Утверждение о том, что для оценки качества необходимо использовать не только количественные, но и качественные методы, подтверждается Руководством REF по материалам [31], в котором рекомендуется больше полагаться на качественные оценки исследований, чем на цитаты. Можно заметить, что подход КБПР, как и большинство подходов журнального уровня, пренебрегает конференциями: даже самые престижные конференции получают более низкий рейтинг, чем статьи в журналах четвертого квартиля.

Перечислим основные преимущества конференций. Во-первых, процесс публикации обычно происходит быстрее, что может быть особенно важно в быстро меняющихся областях, таких как информатика. Во-вторых, отчеты о конференциях могут быть хорошим способом получить раннюю обратную связь о текущей работе или продвинуть ее результаты. В-третьих, посещение конференций - отличный способ наладить связи в своей области, узнать о последних работах других групп и познакомиться с людьми, которые могут быть заинтересованы в ваших результатах или работать с вами в будущем. Проблемы, которые могут возникнуть при участии в конференциях, включают оплату участия, подготовку презентации и выступления с речью, что занимает довольно много времени, а также тот факт, что публикации в материалах конференций часто имеют меньший вес для исследователя в его резюме, чем журнальные статьи. Проблема недооценки конференций в резюме исследователя должна быть решена, поскольку в областях, где большинство научных результатов публикуется в материалах конференции, они имеют большой вес

для научных сообществ. Вышеуказанные преимущества и недостатки перечислены в [24]. Кроме того, одним из недостатков может быть быстрый процесс рецензирования, который может привести к менее тщательному отбору работ [7].

Нежелание участвовать в конференциях, которое спровоцировано вышеперечисленными факторами, может привести к более серьезным последствиям. Так [40] называет отсутствие навыков научной коммуникации одним из факторов, негативно влияющих на эффективность издательской деятельности исследователей. Отсутствие научной коммуникации в некоторых странах приводит к созданию платформ для обмена научной информацией, таких как [5; 33], которые не всегда работают эффективно. Система исследовательского сотрудничества и модель SECI среди ученых Таиланда были изучены в [33]. Это был первый шаг к улучшению обмена знаниями. В свою очередь, в исследовании [5] авторы изучили, используют ли турецкие ученые академический сервис социальных сетей, такой как ResearchGate, Academia.edu, Профиль Google Scholar, LinkedIn, ResearcherID и Mendeley. Исследование показало, что эти сервисы широко используются, но большинство ученых не используют их для обмена знаниями и сотрудничества.

Авторы [52] подчеркивают важность участия в конференциях для исследователей в области компьютерных наук и инженерии, а скорость передачи информации перевешивает многие недостатки конференций. А тот факт, что ученых или преподавателей оценивают традиционным способом (через публикации в журналах), мешает их карьере [52], и несмотря на то, что эта работа была опубликована более 20 лет назад, тенденция сохраняется и по сей день.

Конференции играют важную роль в некоторых областях науки - например, в информатике более 60% результатов исследований публикуются в материалах конференций [45]. Также в этом исследовании были выявлены несколько проблем, связанных с оценкой влияния материалов конференций: во-первых, такие рейтинги, как Google Scholar и Microsoft Academic, пропускают или недооценивают некоторые конференции из-за молодого возраста конференции или непостоянного размера, что вынуждает исследователей искать новые методы ранжирования конференций. Для конференций, которым не менее четырех лет (более конкретно, “конференции, материалы которых индексировались в Scopus не менее 4 лет”, но поскольку нас в основном интересуют

конференции, на которых есть материалы, мы будем использовать “конференции” и “материалы конференций” взаимозаменяемо в документе), CiteScore предоставляет равные возможности для проведения всех конференций, симпозиумов и семинаров, независимо от их размера или возраста. Во-вторых, авторы показывают, что CiteScore Scopus является эффективным методом оценки конференций по информатике. В-третьих, метод CiteScore показывает, что материалы конференций также оказывают большое влияние, как и статьи в ведущих журналах по информатике. Исследование [73] показывают, что небольшое количество элитных конференций имеют более высокий средний показатель цитируемости, чем элитные журналы. Авторы [27] проанализировали библиометрию конференций и пришли к выводу: конференции более популярны, чем журналы; около 78% публикаций, отражающих наиболее актуальные исследования, публикуются в материалах конференций.; публикации в журналах цитируются чаще, чем публикации на конференциях (57% против 43% соответственно); распределение публикаций на конференциях по сравнению с журналами зависит от страны. Несмотря на то, что существует ряд показателей для оценки журналов, например, импакт-фактор, рейтинг журнала SCImago (SJR), не существует общей метрики для оценки конференций [3].

Первые попытки создания рейтингов конференций появились в области компьютерных наук. Наиболее часто используемыми рейтингами, основанными на опыте авторов, являются Ассоциация компьютерных исследований и образования Австралии, CORE и ERA Ranking [8], бразильский рейтинг Qualis (2012), рейтинг Китайской компьютерной федерации (CCF) (2012), MSAR (2014) и GII-GRIN-SCIE (2014). Рейтинг конференций CORE (`http://www.core.edu.au/conference-portal`) содержит оценку крупных конференций по информатике. Решения принимаются академическими комитетами на основе данных, запрошенных в рамках процесса подачи заявок. Рейтинг ERA (2010) был создан в рамках программы Excellence in Research in Australia (ERA) (`https://www.arc.gov.au/excellence-research-australia`). Он включает данные из предыдущей попытки ранжирования, проведенной CORE. Qualis (https://qualis.ic.ufmt.br/) был опубликован министерством образования Бразилии и использует h-индекс (строго говоря, процентили h-индекса) в качестве показателя эффективности конференций. Последнее издание было выпущено в 2016 году. Пятое издание списка журналов и конференций, рекомендованных CCF (`https://www.ccf.org.cn/c/2019-04-25/663625.shtml`)

был выпущен в апреле 2019 года и присваивает конференциям ранги A, B, C в каждой подобласти CS. Интересно, что в пресс-релизе упоминается, что “влияние журналов и конференций напрямую не связано с влиянием отдельной статьи, опубликованной там. Поэтому не рекомендуется использовать этот список в качестве основы для академической оценки” – см. наше предыдущее обсуждение показателей на уровне журнала и статьи. MSAR (https://academic.microsoft.com/home) — это отраслевой рейтинг конференций Microsoft Academic. Он аналогичен h-индексу и вычисляет количество публикаций автора и распределение цитирований по публикациям. Отраслевой рейтинг рассчитывает публикации и цитируемость только в определенной области и показывает влияние ученого или журнала в этой конкретной области. Рейтинг конференции GII-GRIN-SCIE (`http://gii-grin-scie-rating.scie.es/`) представляет собой попытку разработать единый рейтинг конференций по информатике, возглавляемых GII (Группой итальянских профессоров компьютерной инженерии), GRIN (Группой итальянских профессоров компьютерных наук) и SCIE (Испанское общество компьютерных наук) [9]. Последняя версия ранжирует все конференции по четырем уровням (высший уровень, очень высокое качество, хорошие и незавершенные) на основе рейтингов CORE, Qualis и MSAR и была выпущена в 2018 году. И последнее, но не менее важное: существует также инструмент Google Scholar Top publications [28], который выводит как конференции, так и журналы в одном рейтинге.

Теперь, когда было рассмотрено, почему конференции важны в информатике и как это решается с помощью различных рейтингов, давайте проверим, важны ли конференции только в информатике. Как и в предыдущем исследовании, посвященном конференциям [45; 70], был использован Scopus для расчета доли материалов конференций в общем количестве публикаций по конкретным тематическим категориям в 2015-2019 годах. На рисунке 1.1 показаны тематические категории, в которых доля материалов конференций составляет более 10% от всех источников публикаций. Самый высокий процент конференций наблюдается в области компьютерных наук, а также математики и наук о принятии решений. Обратите внимание, что, учитывая, что публикации в Scopus могут принадлежать к нескольким категориям, возможно дублирование. Значительная доля публикаций содержится в материалах конференций по машиностроению и энергетике.

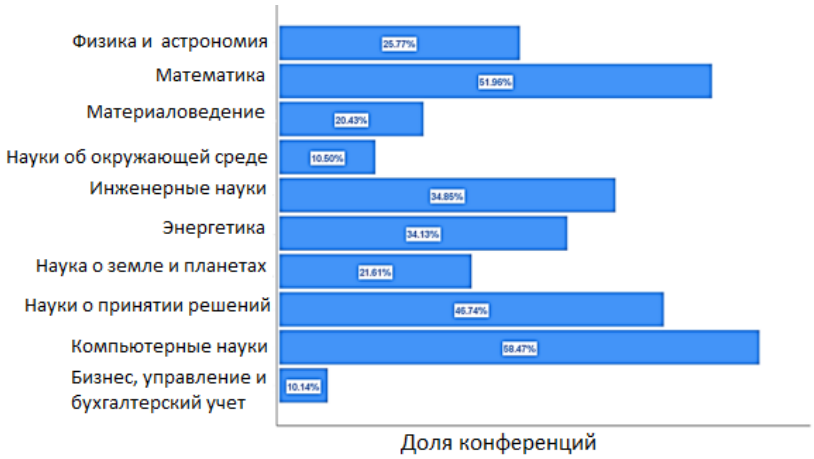

Рис. 1.1. Доля материалов конференций в общем количестве публикаций по отдельным тематическим категориям в 2015–2019 годах. Основано на базе данных Scopus

Упомянутые выше рейтинги конференций обеспечивают первоначальную основу для оценки исследований, опубликованных в материалах конференции. В то время как большинство текущих рейтингов основаны на цитировании, для определения влиятельных статей и авторов могут использоваться альтернативные показатели, например алгоритмы, подобные PageRank [16]. Многообещающим направлением исследований является вопрос о том, можно ли использовать такие рейтинги для борьбы с феноменом хищнических или фальшивых конференций [64]. Однако большинство существующих рейтингов были созданы для конкретных местных сообществ и никогда не предназначались для глобального использования, как в случае с журнальным импакт-фактором. Далее будет предложен показатель для конференций, индексируемых в Scopus, и сравнен его с наиболее часто используемым рейтингом конференций компьютерных наук, CORE. Таким образом, общая цель данного исследования - использовать методы библиометрии для оценки роли материалов конференций в различных дисциплинах.

На первом этапе в списке источников Scopus (https://www.scopus.com/) были выбраны те, которые представлены в качестве материалов конференции (Материалы конференции после 1995 года). С точки зрения времени сбора данных и полноты, Scopus является оптимальным инструментом для проведения исследований [47]. Авторы исследования [12] провели аналогичное исследование, но гораздо позже первого, и пришли к тем же выводам. Затем мы сосредоточились на тех, которые в настоящее время проиндексированы

(т.е. имеют текущий статус) и для которых SJR (https://www.scimagojr.com/) оценка доступна. В результате этого отбора был получен 171 источник с материалами конференции. Обратите внимание, что способ индексации конференций Scopus зависит от конференции и источника публикации (журнал, книжная серия, материалы конференции). Таким образом, 171 выбранный нами источник содержал гораздо большее количество конференций, поскольку такие источники, как ACM International Conference Proceedings Series или CEUR Workshop Proceedings, публикуют несколько сотен материалов конференций в год.

Рейтинг журнала SCImago (SJR) — это не просто показатель цитируемости, такой как импакт-фактор или CiteScore; он основан на алгоритме, подобном PageRank, который представляет собой итеративный процесс передачи престижа среди источников публикации. Расчет представляет собой итеративный процесс, в котором престиж каждого источника зависит от престижа источников, которые на него ссылаются. Конечное значение SJR нормализуется по количеству документов, опубликованных во временном окне цитирования [14].

Учитывая, что SJR вычисляется на основе данных Scopus, мы также использовали эту базу данных в нашем анализе. Из 171 источника материалов конференции 153 были отнесены к одной или нескольким тематическим категориям (третий уровень ASJC (Все классификационные коды научных журналов [21])) в Scopus. Для источников материалов 18 конференций, которым не была присвоена какая-либо тематическая категория, мы вывели категории на основе публикаций в Scopus.

Затем для каждой из тематических категорий мы вычислили пороговые значения SJR для квартилей таким же образом, как SCImago вычисляет их для журналов. Это было необходимо, поскольку SCImago не присваивает квартили источникам материалов конференций, а только журналам и книжным сериям. Это позволило нам присвоить каждому источнику конференции соответствующий квартиль (Q1, Q2, Q3, Q4) в каждой тематической категории. Например, минимальный SJR для журналов и книжных серий первого квартиля составляет 0,261, второго - 0,139, третьего - 0,104 и четвертого - 0,1. Серия конференций IOP: Материаловедение и инженерия имеет SJR 0,195, поэтому мы можем классифицировать источник как Q2. Мы подчеркиваем, что это не

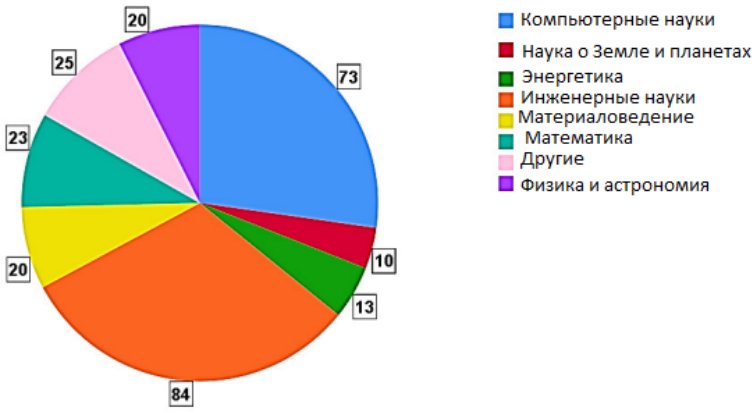

Рис. 1.2. Распределение источников материалов конференций по тематическим категориям

квартиль сам по себе; это условное присвоение источника материалов конференции квартилю на основе значения SJR. Для журналов, охватывающих несколько тем, в рекомендациях КБПР по оценке исследований рекомендуется использовать максимальное количество квартилей по этим темам. Однако важность одной и той же конференции в разных сообществах неодинакова, о чем также упоминается в примечаниях к выпуску CCF. Поэтому мы хотели бы подчеркнуть важность использования тематических квартилей для конференций, т.е. конференция может принадлежать к нескольким тематическим категориям и иметь там разные квартили.

Распределение источников материалов конференций по тематическим категориям показано на рис. 1.2. Обратите внимание, что одна конференция может принадлежать к нескольким тематическим категориям. Из 171 источника один был отнесен к пяти тематическим категориям, один - к четырем, 13 - к трем, 66 - к двум, а 90 конференций имели только одну тематическую категорию.

На рисунке 1.3 показано распределение материалов конференции по квартилям в контексте тематических категорий. Если рассматривать все источники (журналы, серии книг, материалы конференций), то доля каждого квартиля, очевидно, составляет 25%. Однако, поскольку наш выбор ограничен материалами конференции, распределение между категориями Q1–Q4 сильно отличается для каждой тематической категории.

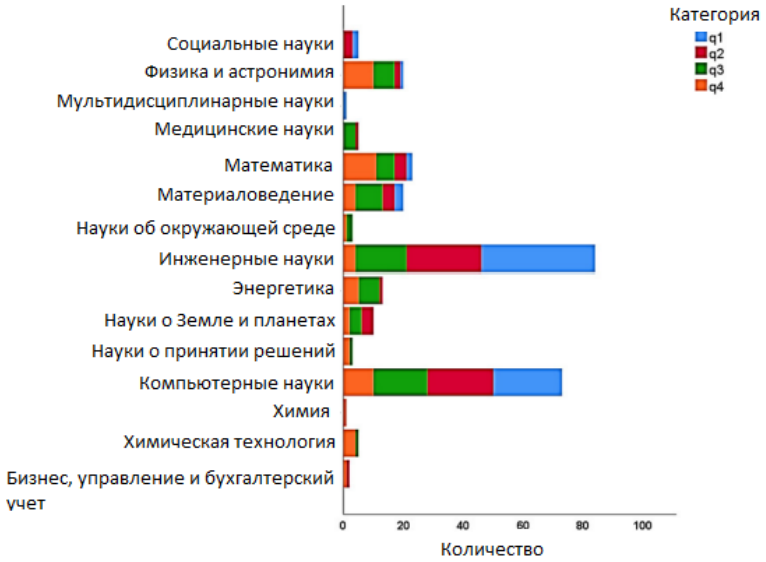

Рис. 1.3. Распределение источников материалов конференции по категориям

Из графика видно, что на инженерные и компьютерные науки приходится не только самая высокая доля материалов конференций, но и наибольшее количество материалов конференций с высокой отдачей. Это еще раз подтверждает тезис о том, что материалы конференций необходимо учитывать при оценке исследований в этих областях. Наши результаты согласуются с результатами более ранних исследований [45]. Разница в количестве и качестве материалов конференций между этими тематическими категориями и остальными существенна.

Из 73 материалов конференций по информатике 45 конференций (62%) находятся в рейтинге CORE; 10 источников являются агрегаторами, которые публикуют материалы многих конференций (например, Procedia Computer Science, ACM International Conference Proceedings Series и т.д.); и 18 материалов конференций не являются основными конференциями компьютерных наук, но относятся к смежным областям (например, сборник IEEE MTT-S International Microwave Symposium Digest). Последние фигурируют в нашем наборе данных, поскольку в соответствии с классификацией ASJC конференции могут одновременно подразделяться на несколько предметных областей/категорий. Однако такие конференции выходят за рамки CORE, которая фокусируется исключительно на конференциях по информатике.

SJR 2018/CORE 2018	A*	A	B	C	NA
Q1	11	4	4	1	3
Q2	5	7	2	1	7
Q3	-	2	6	1	9
Q4	-	-	-	1	9
NA	51	407	402	793	

Рис. 1.4. Сравнительный анализ распределения конференций по категориям

Для 45 конференций из нашего списка, которые также присутствуют в CORE, мы сравнили распределение по категориям (рис. 1.4, Q1 для SJR соответствует A* для CORE, Q2 - A, Q3 - B, Q4 - C). Коэффициент ранговой корреляции Спирмена составил 0,452, что предполагает среднюю корреляционную зависимость. Это интересный факт, учитывая принципиально разные подходы к формированию списков, библиометрические и экспертные.

Существует несколько ограничений. Во-первых, были оценены источники материалов конференций, а не сами конференции. Если оценивать конференции, необходимо учитывать не только библиометрические данные, но и различные другие параметры: тематический охват, программный комитет, авторов, процесс рецензирования, культуру публикации материалов и т.д. Однако представленная здесь количественная оценка может быть удобным вспомогательным инструментом, даже если она не устраняет необходимость в экспертной оценке. Во-вторых, в списке отражены только источники, не относящиеся к журналам и книгам. Материалы конференций, опубликованные в журналах и книжных сериях (например, серия конференций Journal of Physics, Lecture Notes in Computer Science), могут использовать квартиль SJR соответствующего журнала или книжной серии. В-третьих, список отражает только материалы конференций с серийным номером ISSN; некоторые конференции не получают его по недосмотру оргкомитета. Такие конференции не включены в список Scopus и не могли быть включены в анализ. В-четвертых, список включает в себя не только материалы отдельных конференций, но и агрегаторы, такие как материалы семинаров CEUR, Международные труды Лейбница по информатике (LIPIcs). Уровень конференций в рамках таких изданий может существенно различаться. К сожалению, степень детализации данных в Scopus не позволяет проводить анализ на уровне конференций в рамках этих источников.

Несмотря на то, что CCF рекомендует не использовать рейтинги конференций для академической оценки, [68] показывает, как такие рейтинги влияют на публикационное поведение ученых. Поэтому важно обеспечить большую прозрачность в том, как создаются рейтинги, что включается, какие показатели используются и т.д. Предложенная методология представляет собой шаг в этом направлении, поскольку она сочетает в себе прозрачные библиометрические показатели и коррелирует с экспертными заключениями.

В этом исследовании была предпринята попытка рассмотреть роль конференций в оценке исследований и определить научные дисциплины, в которых материалы конференций являются важным средством публикации оригинальных результатов исследований. Наряду с “обычными подозреваемыми”, то есть информатикой, материалы конференций часто используются для публикации результатов в области инженерии, математики, энергетики, наук о принятии решений. Также была представлена новая методология применения данных Scopus и Scimago Journal Rank (SJR) для оценки материалов конференций и показали, что она дает аналогичные результаты с экспертным ранжированием, таким как CORE. Методология показывает, что материалы некоторых конференций по информатике, инженерии, материаловедению, физике и астрономии, а также математике сопоставимы с журналами за 1-2 квартал.

1.3. Моделирование в наукометрии

1.3.1. Кинетические модели в наукометрии

Модели, изначально созданные для описания динамики биологических популяций, нашли свое применение и в наукометрии. Модель Ферхюльста и модель «хищник-жертва» (Лотки–Вольтерры) используются для описания и прогнозирования различных аспектов этого развития. Модели, пришедшие из биологии, играют важную роль в наукометрии. Модель Ферхюльста стала мощным количественным инструментом для описания и прогнозирования роста научного знания. Модель «хищник-жертва» чаще выступает в роли концептуальной метафоры, позволяя описывать динамику конкуренции в науке через аналогии.

Модель Ферхюльста

В наукометрии чаще всего применяется модель Ферхюльста (логистическая модель) [72]. Она описывает рост, который начинается стремительно, но со временем замедляется и выходит на плато насыщения.

Дерек Джон де Солла Прайс (один из основоположников наукометрии) предложил модель роста науки [55]. Изучая рост количества научных журналов и публикаций, он заметил, что экспоненциальный рост не может продолжаться вечно, и кривая роста со временем принимает S-образную (сигмоидную) форму, стремясь к некоторому пределу. Этот предел, или «емкость среды», в наукометрии может означать ограниченность ресурсов: финансирования, числа ученых или исчерпание легкодоступных открытий в рамках научной парадигмы [85].

В статье [66] валидируется теория Прайса на основе эмпирических исследований, опубликованных с 1913 по 2018 год. Большая часть рассмотренных работ подтверждают логистический закон в наукометрии.

В исследовании [65] для описания жизненного цикла научных тем используются логистическая модель и модель Гомпертца.

Модель «хищник-жертва»

Классическая модель «хищник-жертва» (Лотки-Вольтерры) [41; 78] описывает циклическое взаимодействие двух видов, численность которых колеблется в противофазе: рост популяции жертв приводит к росту числа хищников, что ведет к сокращению жертв и, как следствие, последующему сокращению хищников. Эта модель также нашла свое отражение в наукометрии, но ее применение часто бывает более концептуальным.

В наукометрии эта модель служит скорее концептуальной метафорой, помогая описывать процессы конкуренции, чем точным инструментом прогнозирования. Модель может описывать, как новая технология или научная парадигма («хищник») со временем вытесняет старую («жертву»). Это похоже на колебания популярности или доли рынка технологий. Также в условиях ограниченного финансирования и внимания научного сообщества, различные научные направления могут конкурировать между собой подобно видам в экосистеме.

В исследовании [77] построена симуляция Лотки–Вольтерры для моделирования симбиотических отношений между учеными и платформами социальных сетей в рамках экосистемы научной коммуникации.

Закон Матфея (или эффект Матфея, эффект кумулятивного преимущества) — один из фундаментальных механизмов, описывающих социальное неравенство в науке. Его суть, отраженная в библейской метафоре, заключается в том, что известные ученые получают непропорционально большее признание за свои работы, чем их менее именитые коллеги. Это ставит под вопрос валидность библиометрических показателей как объективных измерителей научного качества.

1.3.2. Закон Матфея

Термин и лежащую в его основе теорию кумулятивного преимущества в науке сформулировал и развил в своих классических работах американский социолог Роберт К. Мертон (Robert K. Merton). Ключевой работой считается статья 1968 года «The Matthew Effect in Science» [49], идеи которой были позже развиты в статье 1988 года [48].

Практически одновременно Дерек де Солла Прайс (Derek J. de Solla Price), наблюдая за сетями цитирования, описал то же явление, но использовал термин «кумулятивное преимущество» (Cumulative advantage) [55].

Мертон показал, что закон Матфея — это не просто частный случай, а системное свойство научной коммуникации, прочно связанное с неравномерным распределением научной продуктивности, также известным как закон Лотки [42].

Эффект Матфея можно обнаружить в нескольких ключевых областях библиометрического анализа.

Цитируемость отдельных ученых Это самый прямой и очевидный способ проявления эффекта. Если два исследователя публикуют работы сопоставимого качества, статья более известного автора с высокой вероятностью соберет значительно больше цитирований.

Исследования показывают, что распределение научных карьер с точки зрения признания может быть описано с помощью эффекта Матфея.

Эффект Матфея в цитировании можно декомпозировать:

Сетевой эффект (Networking) Автор получает дополнительные цитирования просто потому, что его новая работа привлекает внимание к его прошлым публикациям.

Эффект престижа (Prestige) Цитирования приходят благодаря высокой репутации автора, учреждения или журнала, а не только благодаря качеству самой работы.

Эффект уместности (Appropriateness) Цитирование происходит потому, что работа действительно является релевантной и подходящей ссылкой.

Стоит заметить, что первые два эффекта ставят под сомнение объективность метрик, основанных только на цитированиях.

Роль научных журналов Журнал, в котором опубликована статья, сам по себе является мощным фактором, инициирующим эффект Матфея. Когда одни и те же статьи публиковались в журналах с разным импакт-фактором, версии в более престижных изданиях получали в среднем в два раза больше цитирований. То есть у статьи есть дополнительная ценность, связанная исключительно с журналом, в котором она вышла.

Уровень стран и институтов Эффект Матфея действует и в глобальном масштабе, создавая неравенство между странами. Работы, аффилированные с ведущими западными странами, получают больше цитирований, чем статьи из стран с менее развитой научной инфраструктурой, даже при публикации в одних и тех же журналах.

Глава 2. Статистическая модель и метод анализа рейтингов конференций по искусственному интеллекту

2.1. Анализ статистических методов в оценке конференций по искусственному интеллекту

Искусственный интеллект (ИИ) - динамично развивающаяся область исследований, которая является междисциплинарной, но имеет прочные корни в компьютерных науках, где более 60% результатов исследований публикуются в материалах конференций. Scopus обеспечивает хорошее освещение материалов конференции [56]. Двумя ведущими странами, публикующими материалы конференций в области искусственного интеллекта, по данным Scopus, являются Китай (23368 материалов конференций или 77% их оригинальных исследований за последние 10 лет) и США (21194, 79%). Доля материалов конференций обеих стран, а также общемировая доля материалов конференций по ИИ, составляющая 75%, показывает, что публикации в материалах конференций имеют больший вес и в сообществе искусственного интеллекта.

Лидирующие позиции Китая по количеству научных исследований в области искусственного интеллекта обусловлены тем, что Китай пытался обогнать Соединенные Штаты в технологической гонке, поэтому он прилагает огромные усилия в этой области [11]. Таким образом, Китай принял стратегическую государственную программу развития сектора искусственного интеллекта до 2030 года. Его реализация поддерживается крупномасштабным государственным финансированием, а также средствами частных технологических компаний, действующих в Китае. Главным преимуществом Китая является огромный объем генерируемых данных [11]. США также вкладывают много бюджетных средств в развитие этой области компьютерных наук [25].

В работе [35] мы предложили методологию оценки количества и качества докладов конференции из конкретной страны. В нем анализируется количество публикаций и цитирований на высокопоставленных конференциях и сравнивается с мировыми тенденциями. Мы протестировали эту методологию на материалах конференций по искусственному интеллекту из России.

В этом исследовании мы проводим сравнительный анализ, используя ту же методологию, для публикаций конференций в Соединенных Штатах и Китае в 2011-2020 годах. Цель анализа - определить конференции, на которых научные статьи из конкретной страны, скорее всего, получат более высокий, чем средний, уровень цитирования. Это помогает сделать стратегию публикации более эффективной с точки зрения наукометрии, а также предоставляет ценную информацию исследователям искусственного интеллекта из других стран. Мы составили список конференций на основе данных Ассоциации компьютерных исследований и образования Австралии (CORE), отраслевых рейтингов конференций Microsoft Academic (MSAR) и списка международных научных конференций и периодических изданий, рекомендованных China Computer Federation (CCF), и используем информацию о цитировании из Scopus. Далее мы разделили рейтинг по квартилям, как это обычно делается в журнальных рейтингах, и проанализировали публикации исследователей из США и Китая. Результаты показывают, что, хотя в Китае больше публикаций по искусственному интеллекту, публикации в США цитируются чаще и чаще превышают ожидаемый уровень цитируемости для конкретных конференций.

Существует несколько объяснений того, почему конференции играют такую важную роль в информатике и часто считаются более важными, чем журналы. Одним из наиболее распространенных является то, что исследования в этой области имеют краткосрочную применимость [3]. Поэтому разрабатываются новые методы оценки конференций с использованием различных методов, например, в [3] был предложен метод ранжирования конференций на основе машинного обучения. Они также показывают, что авторы, вошедшие в первую десятку рейтинга цитируемости, опубликовали около 60% своих исследований в сборниках материалов конференций. Другой пример метода ранжирования конференций представлен в [62], где модифицированный алгоритм PageRank используется для ранжирования статей на основе сети цитирования. Предлагаемый алгоритм ранжирования также включает новую метрику, которая учитывает фактор времени, чтобы не наказывать статьи, опубликованные недавно. Так, [57] предложил алгоритм ранжирования, с помощью которого авторы составили рейтинг финансовых конференций и пришли к выводу, что конференции являются важной составляющей фундамента научной коммуникации и карьеры ученого. В статье [60] также обсуждается роль цитирования

в оценке воздействия статей и их восприятия, а также используются методы машинного обучения для анализа цитирования.

Хотя проблема ранжирования конференций очень важна, многие из существующих рейтингов имеют различные плюсы и минусы, обсуждаемые в академическом сообществе, и универсального рейтинга, признанного всеми, не существует. Существует также ряд применений методов оценки исследований журналов или авторов на конференциях. Например, в [4] авторы использовали индекс DS для ранжирования конференций, который ранее использовался для ранжирования авторов. Этот индекс присваивает каждой конференции уникальный рейтинг, что является его главным преимуществом перед методами, которые присваивают один и тот же рейтинг нескольким конференциям. Авторы приходят к выводу, что индекс DS обеспечивает лучшую дифференциацию конференций по сравнению с другими показателями, такими как h-индекс, g-индекс и R-индекс. Другим примером является использование индекса h5 для оценки цитируемости материалов конференции в [73]. Здесь авторы сравнили цитируемость журнальных статей и материалов конференций по информатике, используя h5 и среднюю цитируемость. Авторы [73] сделали следующие выводы: а) для информатики как дисциплины конференции играют более значительную роль, чем для других дисциплин; б) хотя, в целом, показатели цитируемости конференций не выше, чем у журналов, существует ряд элитных конференций, которые имеют самые высокие средние показатели цитируемости.

В исследовании [44] сравнивается публикационная активность Китая в области биоинформатики с другими ведущими странами в этой области - Соединенными Штатами, Великобританией, Германией, Японией и Индией. Результаты этого исследования показали, что Китай имеет самую низкую международную репутацию в этой области из шести стран, изученных в данной работе, и предложили возможные решения этой проблемы. В другом исследовании [29] авторы сравнивают публикационную активность североафриканских исследователей в области биотехнологий, энергетики, астрономии и палеонтологии и сравнивают ее с активностью ученых из стран БРИК (Бразилия, Россия, Индия и Китай) и Египта в тех же областях. В ходе исследования были выявлены области, в которых исследователи показывают относительно высокие результаты по сравнению с другими странами, участвующими в исследовании, а также университетами и организациями, занимающими

лидирующие позиции в каждой из областей исследований. В исследовании [29] анализируется взаимосвязь между уровнем высшего образования и публикационной активностью стран Организации исламского сотрудничества (ОИС) и их позицией по сравнению со странами-лидерами по количеству публикаций. В [56] автор рассмотрел глобальные и региональные тенденции, которые отражают репрезентацию материалов конференций в международной научной литературе. В исследовании приняли участие 10 стран Юго-Восточной Азии. Результат исследования показал, что из всех стран, участвовавших в исследовании, Индонезия показала хороший результат в пользу увеличения количества публикаций в материалах конференций, что может быть связано с увеличением количества местных конференций. Также, в результате этого исследования автор пришел к выводу, что материалы конференций все чаще индексируются основными базами данных для составления тезисов и индексирования.

В исследовании [45] автор исследует, подходит ли показатель CiteScore от Scopus для выбора конференций по информатике. Сравнивая CiteScore с такими рейтингами, как Google Scholar и Microsoft Academic, по 395 значимым конференциям, автор [45] определил 154 конференции, которые соответствуют диапазонам CiteScore ведущих квартильных журналов. Также важным выводом является то, что 154 конференции составляют 30% от всех 515 лучших мест публикации в области компьютерных наук, что подтвердило тезис о важности и влиянии публикации топовых конференций как публикации в топовых журналах. Метод CiteScore, реализованный здесь, показывает, что он очень эффективен в качестве ориентира для оценки и сравнения мест публикации в области компьютерных наук. Однако Scopus необходимо усовершенствовать некоторые из своих методов индексации, прежде чем база данных и метод CiteScore смогут стать стандартными инструментами для оценки качества конференций.

В [34] авторы разработали новый алгоритм ранжирования 15 финансовых конференций, основанный на сочетании трех факторов, которые измеряют качество конференций. Чтобы оценить качество полученного рейтинга, они провели различные оценки надежности, которые показали, что рейтинг был достаточно стабильным. В исследовании также использовался метод, основанный на восприятии качества опрошенными участниками конференции.

Авторы [2] предложили метод ранжирования новых мест публикации (конференций, журналов) на основе социальных показателей (научных ссылок с сайтов академических социальных сетей), которые также могут выступать в качестве раннего индикатора влияния. Также был проведен сравнительный анализ между новым методом ранжирования и методами, использующими традиционные показатели цитируемости. Результаты показали, что новая система, которая была разработана авторами на основе социальных ссылок, имеет значительную корреляцию с традиционными методами, но в то же время обладает потенциалом обеспечить ранний интеллектуальный индикатор влияния научных сайтов, снижая при этом ограничения метрик, основанных на цитировании.

И последнее, но не менее важное: прозрачное и объективное ранжирование конференций может помочь предотвратить участие авторов в хищнических конференциях. Эта проблема очень важна, например, в России с 2011 года значительно увеличилось количество статей, опубликованных в хищнических журналах [1]. Эта проблема распространилась на конференции с хищническими журналами и процветает, поскольку подобные мероприятия привлекают нечестных организаторов с целью получения финансовой выгоды. Статьи с таких конференций часто не публикуются и не индексируются, а если и публикуются и индексируются, то впоследствии могут быть удалены из индексируемых баз данных. Эта проблема также распространена в разных странах, таких как Дания и Южная Корея [26].

2.2. Методы статистического для анализа качества конференций и введение новой метрики

Чтобы рассчитать процент конференций по странам, мы использовали базу данных реферативания и индексации Scopus и выполнили поиск по предметному полю “Искусственный интеллект” (1702), периоду времени 2011-2020 гг. и стране (Китай против США). Мы считали, что публикация была из США или Китая, если хотя бы один из авторов имел отношение к США или Китаю. В таблице 2.1 показано количество публикаций по основным типам документов.

На первом этапе мы определили список конференций по искусственному интеллекту, на которых исследователи из Соединенных Штатов и Китая

Таблица 2.1.

Количество документов по типам

Страна	Документы конференции	Журнальные статьи	Обзорные статьи	Общее количество
Китай	89 791	50 787	570	143 275
США	54 430	22 875	684	82 187

опубликовали статьи. Мы рассмотрели 100 лучших конференций по искусственному интеллекту из рейтинга Microsoft Academic conference ranking, все 176 конференций, занесенных в Australian CORE 2021 как AI (код 0801), и все 40 конференций рейтинга Китайской компьютерной федерации в области искусственного интеллекта. Поскольку аббревиатуры конференций могут отличаться в разных рейтингах, вручную было установлено соответствие по полному названию конференции.

На втором этапе мы подсчитали количество цитирований, полученных из статей, опубликованных в трудах этих конференций. Мы использовали количество цитирований, поскольку исследование [59] показывает, что библиометрические показатели дают надежные результаты при выявлении конференций высшего уровня. Были использованы данные Scopus за период 2011-2020 гг. Документы и цитаты были извлечены, используя следующую строку поиска CONF (“Полное название конференции” OR аббревиатура) AND PUBYEAR AFT 2010 AND PUBYEAR BEF 2021 AND DOCTYPE (cp) AND SUBJTERMS (1702). В случае, если аббревиатуры конференций совпадали, мы проверяли полное название конференции вручную и выполняли поиск по полному названию, а там, где это было необходимо, мы вручную проверяли источники. Это позволило выполнить поиск по всем статьям, опубликованным в трудах указанной конференции. После выбора конференции мы устанавливаем фильтр по требуемому периоду и стране.

На третьем этапе мы ввели показатели для анализа цитируемости. Мы рассчитали ожидаемый уровень цитирования (e_i) и фактическое количество цитирований на документ (c_i) за каждый год i для обеих стран. e_i определяется в [36] как ожидаемый коэффициент цитирования, который рассчитывается делением количества цитирований на количество документов за каждый год.

Таблица 2.2.

Введённые метрики

Метрика	Определение
Total output	Общее количество публикаций
Total citation score (TCS)	Общее количество цитирований
Citations per paper (CPP)	Общее количество цитирования, деленный на общее количество публикаций
Mean normalized citation score (MNCS)	Среднее количество цитирований на публикацию, нормированное по году публикации, названию и стране принадлежности

Этот показатель был определен в [74] следующим образом: ожидаемое количество цитирований публикации определяется как среднее количество цитирований всех публикаций в одной и той же области (за один и тот же год и одного и того же типа документа). Как упоминалось выше, публикация включалась в расчет, если хотя бы один автор был связан с Китаем или Соединенными Штатами.

MNCS — это независимый от размера индикатор цитирования, ориентированный на статьи, который был определен в [69]:

$$\text{MNCS} = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i}{e_i}, \tag{2.1}$$

где n — количество лет, c_i — фактический уровень цитирования, а e_i — ожидаемый уровень цитирования. Таким образом, это помогает выявлять публикации, которые превзошли ожидания.

Формула (2.1) была применена к нашему набору данных и рассчитан ожидаемый уровень цитирования как среднее значение цитирований по годам для всех документов каждой конференции, что помогло определить ожидаемый уровень цитирования для каждой конференции, а не для всех конференций в нашей выборке.

Для анализа в таблице 2.2 были введены следующие показатели.

Перед проведением анализа мы протестировали выборки MNCS Китая и MNCS США на нормальность, используя критерий Колмогорова-Смирнова

в IBM SPSS Statistics 21. Проведенная проверка показала, что все три выборки не соответствуют нормальному распределению.

Критерий Колмогорова-Смирнова предназначен для проверки гипотезы о том, что две независимые выборки принадлежат одному и тому же закону распределения, то есть что два эмпирических распределения не соответствуют одному и тому же закону. Даны две независимые выборки $\vec{X} = (X_1, X_2, \dots, X_n)$ и $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$ из неизвестных распределений $\mathfrak{F}$ и $\mathfrak{G}$ соответственно. Проверяется сложная гипотеза $H_1 = \{F = G\}$ при альтернативе $H_2 = \{H_1 \text{неверна}\}$.

Критерий Колмогорова–Смирнова используют, если $\mathfrak{F}$ и $\mathfrak{G}$ имеют непрерывные функции распределения.

Пусть $F_n^*(y)$ и $G_m^*(y)$ — эмпирические функции распределения, построенные по выборкам $\vec{X}$ и $\vec{Y}$:

$$\rho(\vec{X}, \vec{Y}) = \sqrt{\frac{mn}{m+n}} \sup |F_n^*(y) - G_m^*(y)|.$$

Теорема. Если гипотеза H_1 верна, то $\rho(\vec{X}, \vec{Y}) \to \eta$ при $n, m \to \infty$, где η имеет распределение Колмогорова.

В таблице распределения Колмогорова по заданному ε найдем C такое, что $\varepsilon = P(\eta \geqslant C)$, и построим критерий Колмогорова–Смирнова:

$$\delta(\vec{X}) = \{H_1\}, \quad \rho(\vec{X}) < CH_2, \quad \rho(\vec{X}) > C.$$

Подробнее об этом критерии можно прочитать в [87].

Кроме того, для проведения исследования с использованием методов, описанных ниже, необходимо условие линейности. Из графиков мы можем заключить, что данные не являются линейными. Если существует линейная зависимость и выборки соответствуют нормальному закону распределения, мы можем применить коэффициент корреляции Пирсона. Если эти два условия не выполняются, то мы применим коэффициент корреляции Спирмена. Поскольку предположение о существовании линейной зависимости и принадлежности к одному и тому же закону распределения между рассматриваемыми выборками не подтвердилось, мы далее рассмотрим применение коэффициента корреляции Спирмена.

Коэффициент корреляции Спирмена был рассчитан между следующими выборками — MNCS Китая и MNCS США, MNCS Китая/MNCS США и рейтингом MSAR, MNCS Китая/MNCS США и рейтингом CORE. Мы сделали это для того, чтобы определить взаимосвязь между рассчитанными значениями MNCS и рейтингами MSAR и CORE. Чтобы рассчитать коэффициент корреляции между MNCS и CORE, мы сопоставили каждый показатель CORE следующим образом: $A^* - 1$, $A - 2$, $B - 3$, $C - 4$ и национальные или не ранжированные, но включенные в рейтинг — 5. Также для каждой выборки MNCS мы рассчитали квартили 25%, 50% и 75% и в соответствии с ними разделили конференции на 4 части и присвоили им номера. Кроме того, для каждого коэффициента корреляции мы рассчитали значимость коэффициента корреляции.

Затем мы оценили значимость коэффициентов корреляции. Мы ввели две гипотезы по данным [79]:

$$\begin{cases} H_0 : & r = 0, \\ H_1 : & r \neq 0, \end{cases}$$

где r — коэффициент корреляции.

Также значимость коэффициентов корреляции r была проверена. Если нулевая гипотеза принята, это означает, что данные не коррелированы, в противном случае они коррелированы.

Далее было рассчитано наблюдаемое значение критерия по формуле:

$$t = t(\alpha, k) = \frac{\sqrt{1 - \rho^2}}{n - 2},$$

где n — размер выборки; ρ — выборочный коэффициента ранговой корреляции Спирмена; $t(\alpha, k)$ — критическая точка двусторонней критической области, которая находится по таблице критических точек распределения Стьюдента, по данным на уровне значимости α и числа степеней свободы $k = n - 2$.

Весь процесс анализа был автоматизирован от сбора данных до расчета MNCS и статистического анализа, включая проверку гипотез, анализа коэффициента корреляции и построения графиков, проверки выборок на нормальность и так далее. Листинг кода представлен в приложении В.

2.3. Применение статистического анализа и введенной метрики для анализа конференций по искусственному интеллекту

2.3.1. Применения метрики MNCS для анализа конференции по искусственному интеллекту: Россия

На первом этапе мы определили список ведущих конференций в области искусственного интеллекта. Мы использовали два источника данных: рейтинг конференций Australian CORE 2018 и рейтинг конференций Microsoft Academic field. CORE предоставляет оценки основных конференций по компьютерным дисциплинам и включает в себя 1626 конференций. Мы отфильтровали конференции по предметной области ’Искусственный интеллект’ (для кода 0801, 364 конференции) и рангу A* (19 конференций), последний обозначает лучшие конференции. Мы также взяли 20 лучших конференций по искусственному интеллекту из академической базы данных Microsoft, в которой перечислены 4472 конференции, из которых 1601 посвящена искусственному интеллекту. Поскольку оба списка перекрываются, в результате мы получили список из 31 конференции (таблица 2.1).

На втором этапе мы подсчитали количество цитирований, полученных статьями, опубликованными в материалах этих конференций. Это было основано на данных Scopus за период с 2010 по 2019 год. Данные были взяты по годам (2010, 2011,...,2019) в виде цитируемых статей с той же конференции за этот период времени (2010-2019). Мы использовали Scopus, потому что он охватывал все конференции в нашем анализе, за исключением IEEE InfoVis.

Далее, для выбранных конференции на основе количества публикаций и цитирований по ним был рассчитан показатель MNCS по формуле (2.1).

Материалы конференции IEEE Information Visualization Conference не были проиндексированы в Scopus, поэтому нам пришлось их исключить. В таблице 2.3 представлены результаты расчетов по остальным 30 конференциям.

Прежде всего, действительно бросается в глаза крайне малое количество публикаций российских ученых в материалах анализируемой конференции. Почти во всех случаях это составляет менее 1% от общего количества публикаций. Объяснений может быть несколько: а) рост публикационной активности в России начался с запуска проекта 5-100 в 2013 году. Фактически, результаты

Таблица 2.3.
Показатели цитирования для конференций по искусственному интеллекту (в алфавитном порядке). Жирным шрифтом выделены 8 конференций, входящие в оба рейтинга, а * обозначает конференции с MNCS (RU)> 1

Conference	Total output	Output (RU)	Average output per year	MNCS (RU)	TCS	TCS (RU)	CPP	CPP (RU)
AAAI	3958	8	439.78	0.73	34172	39	8.63	4.88
AAMAS	3400	8	340.00	0.59	20170	22	5.93	2.75
ACL*	3842	21	384.20	1.98	74351	473	19.35	22.52
COLT	402	2	57.43	0.83	4429	17	11.02	8.50
CVPR	6250	21	694.44	0.56	415899	981	66.54	46.71
EC	394	2	65.67	0.00	2231	0	5.66	0.00
ECCV*	3107	17	443.86	1.86	68784	786	22.14	46.24
EMBC	12953	32	1619.13	0.95	48334	121	3.73	3.78
EMNLP	1813	4	201.44	0.31	56109	29	30.95	7.25
FOGA	90	1	18.00	0.00	635	0	7.06	0.00
ICAPS*	626	1	62.60	2.39	5150	8	8.23	8.00
ICASSP	14625	20	1462.50	0.83	116017	94	7.93	4.70
ICCV*	3329	12	665.80	1.57	144935	757	43.54	63.08
ICIP*	8678	6	867.80	1.52	45414	53	5.23	8.83
ICLR	1696	15	282.67	0.20	41690	72	24.58	4.80
ICML*	3653	26	365.30	1.59	82020	807	22.45	31.04
ICPR	5580	21	558.00	0.50	30606	46	5.48	2.19
ICRA*	8974	12	897.40	2.07	126521	131	14.10	10.92
IJCAI	5281	25	586.78	0.49	45140	111	8.55	4.44
IJCAR	252	3	42.00	0.55	2057	4	8.16	1.33
INTER-SPEECH*	7861	41	786.10	2.07	54804	379	6.97	9.54
IROS	9023	20	902.30	0.35	87510	53	9.70	2.65
ISMAR*	1105	2	110.50	2.42	8891	9	8.05	4.50
KDD	2416	3	241.60	0.13	71748	13	29.70	4.33
KR*	417	1	83.40	1.97	3290	23	7.89	23.00
NIPS/ NeurIPS	3466	19	433.25	0.58	199271	473	57.49	24.89
RSS*	363	1	51.86	3.42	4177	23	11.51	23.00
SIG-GRAPH	10789	22	1078.90	0.74	83862	167	7.77	7.59
SMC*	6842	22	760.22	1.14	21185	86	3.10	3.91
UAI*	1459	5	145.90	1.53	8456	6	5.80	1.20

этого проекта можно увидеть только с 2014 года, то есть только во второй половине анализируемого нами периода. б) культура публикаций в большинстве дисциплин (за исключением информатики) основана на журналах. Например, система оценки научных исследований в России, введенная в 2019 году, то есть “Комплексный балл публикационной результативности” (КБПР), основана на квартиле журнала, определяемом его импакт-фактором (IF) в Web of Science: журнальная статья, опубликованная в журнале первого квартиля, получает 20 баллов, второй квартиль - 10 баллов, третий квартиль - 5 баллов и четвертый квартиль — 2,5 балла. Все остальные публикации, включая материалы конференций, главы из книг, журнальные статьи, индексируемые только в Scopus или Российском индексе научного цитирования, получают 1 балл. Эта шкала применима к естественным наукам, инженерии и наукам о жизни. Такая система значительно снижает стимулы для публикаций в сборниках конференций для российских ученых и противоречит международной практике публикации в области компьютерных наук. Предыдущее исследование [37] показывает, что если мы рассмотрим рейтинг журналов Scimago (SJR), то для ведущих конференций он будет выше, чем у многих журналов первого квартиля (например, Proceedings of the IEEE International Conference on Computer Vision).

Что касается цитирования, то в 13 случаях из 30 (включая 3 профильные конференции) среднее количество цитирований на статью для российских документов больше или равно среднему показателю для этой конференции, т.е. фактический уровень цитирования российских авторов больше или равен ожидаемому уровню цитирования. Однако оценить уровень цитируемости на основе такого небольшого числа публикаций довольно сложно. Корреляционный анализ показал относительно высокий уровень корреляции Пирсона между средним процентом цитирования материалов конференций в целом и процентом цитирования публикаций российских авторов (рис. 2.1).

2.3.2. Статистический анализ и применение метрики MNCS для анализа конференций по искусственному интеллекту: Китай и США

Основываясь на полученных данных, было обнаружено, что исследователи из Китая не опубликовали свои статьи на 17 конференциях, в то время как исследователи из Соединенных Штатов никогда не выступали только на 7

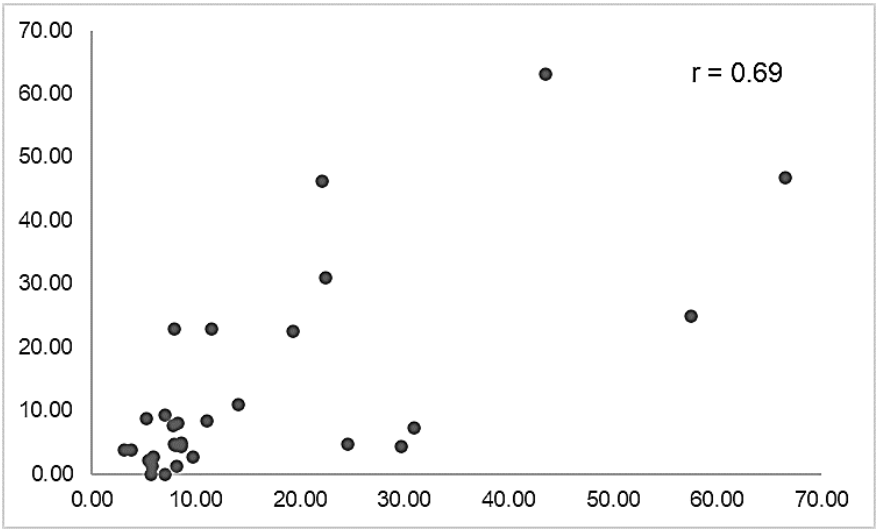

Рис. 2.1. Корреляционный анализ CPP (всего) и CPP статей российских авторов

конференциях из 83 в нашей выборке. В представленных таблицах 2.4 и 2.5 мы отсортировали конференции в порядке убывания показателя MNCS.

В таблице 2.4 показаны те конференции, которые получили значение MNCS, большее или равное 1 для исследователей из Китая. В таблице 2.5 показаны конференции, получившие значение MNCS, большее или равное 1 для исследователей из Соединённых Штатов. Столбец MSAR показывает рейтинг конференций в рейтинге Microsoft Academic conference в области искусственного интеллекта (1-100). В столбце CORE показан рейтинг конференции, который был присвоен ей австралийским CORE 2021 в области искусственного интеллекта (A*,A, B, C,n/r - не ранжирован, это означает, что конференция находится в рейтинге, но ей не присвоили никакого ранга, потому что она национальная/региональный или не накопил достаточного количества данных). В столбце CCF представлен рейтинг конференции в рейтинге конференций Китайской компьютерной федерации, который разделен на 3 группы (A, B, C), а число в круглых скобках указывает место конференции в каждой части рейтинга.

Из этих результатов можно сделать следующие выводы. Во-первых, исследователи из Соединённых Штатов участвовали почти во всех конференциях в списке (не участвовали в 7 конференциях из 83). Ученые из Китая не участвовали в 17 конференциях из 83. Во-вторых, было больше конференций, на которых статьи американских исследователей получили оценки MNCS выше среднего, чем статьи исследователей из Китая. Это связано с тем, что

исследователи из Китая приняли участие в 18 конференциях, где цитируемость их статей превысила ожидания, в то время как в Соединенных Штатах цитируемость на 37 конференциях превысила ожидания. Исследователи из Соединенных Штатов и Китая не получили цитат на 4 конференциях.

Таблица 2.4.

Показатели цитирования для Китая

Конференция	Total output	TCS	CPP	Output China	TCP China	CPP China	MNCS China	MSAR	CORE	CCF
IE	891	5790	6.498	37	633	17.108	5.681	-	B	-
FIAIRS	1009	4649	4.608	7	86	12.286	4.104	58	-	-
ACL	910	54821	60.243	134	14863	110.918	3.069	-	A*	A (3)
COPLAS	36	109	3.028	1	1	1	3	-	B	-
PACLIC	106	727	6.858	21	421	20.048	2.923	-	B	-
FedCSIS	232	898	3.871	2	24	12	2.887	-	multi	-
ASRU	204	5744	28.157	13	655	50.385	2.118	-	C	-
IEEE SIS	122	1206	9.885	18	309	17.167	1.606	-	C	-
IJCAI	5670	132398	23.351	1853	63467	34.251	1.547	2	A*	A(7)
AAAI	8491	269243	31.709	2459	109413	44.495	1.447	1	A*	A(1)
SAMI	584	2916	4.993	2	15	7.5	1.404	-	nat.	-
ICPR	11	98	8.909	1	11	11	1.235	-	-	C(17)
SNPD	844	4031	4.776	209	1269	6.072	1.233	-	C	-
ICARV	1467	6845	4.666	582	1985	3.411	1.193	-	C	-
GECCO	750	3668	4.891	31	169	5.452	1.105	13	A	C(7)
IRI	279	2294	8.222	11	129	11.727	1.079	-	nat.	-
ECAI	1027	3041	2.961	10	426	42.6	1.018	12	A	B(3)
NAACL	279	5658	20.279	6	84	14	1.016	-	A	C(21)

Таблица 2.5.

Показатели цитирования для США

Конференция	Total output	TCS	CPP	Output USA	TCS USA	CPP USA	MNCS USA	MSAR	CORE	CCF
ICARCV	1467	6845	4.666	59	947	16.051	2.629	-	C	-
CSIT	117	932	7.966	3	8	2.667	2.509	-	nat.	-
SYSY	620	2591	4.179	10	55	5.5	2.419	-	nat.	-
CIS	1699	6332	3.727	23	279	12.130	2.382	-	C	-
ICAPS	549	6792	12.372	174	4298	24..701	2.105	23	A*	B(6)
ICPR	11	98	8.909	2	36	18	2.020	-	-	C(17)
AAAI	7815	102634	13.133	3364	104046	30.929	2.991	1	A*	A(1)
IEEE HPCS	149	618	4.148	23	158	6.869	1.656	-	B	-
ASRU	204	5744	28.157	92	4223	45.908	1.649	-	C	-
SST	165	701	4.248	2	10	5	1.576	-	nat.	-
IE	891	5790	6.498	50	769	15.38	1.549	-	B	-
CoNLL	407	13984	34.359	133	7334	55.142	1.646	-	-	C(6)
ACRA	436	2545	5.837	13	99	7.615	1.424	-	nat.	-
ISARC	1599	7751	4.847	262	1721	6.569	1.389	-	C	-
RANLP	602	4948	8.219	67	865	12.91	1.317	-	nat.	-
CLEI	350	1016	3.903	1	8	8	1.125	-	C	-
ICINCO	518	1230	2.375	26	66	2.538	1.221	-	C	-
ICTAI	1734	11220	6.471	299	2487	8.317	1.208	88	B	C(8)
IEEE IS	659	2413	3.662	16	66	4.125	1.205	-	C	-
AAMAS	3294	32626	9.905	1153	13832	11.997	1.203	6	A*	B(11)
GECCO	750	3668	4.891	104	621	5.971	1.198	13	A	C(7)
IAAI	3401	95407	28.053	1504	42941	28.551	1.197	-	B	-
ALIFE	321	1362	4.243	111	550	4.955	1.166	-	C	-
CDC	1213	9994	8.239	574	5500	9.582	1.163	40	-	-
CIKM	19	429	22.583	12	307	25.583	1.133	4	A	-
IRI	279	2294	8.222	174	1581	9.086	1.122	-	nat.	-

Продолжение на следующей странице

Таблица 2.5.

Показатели цитирования для США (продолжение)

CogSci	5305	17703	3.337	3166	11870	3.749	1.118	-	A	-
UAI	1172	12161	10.376	659	8002	12.143	1.111	7	A	B(10)
PACLIC	106	727	6.859	5	38	7.6	1.108	-	B	-
MMAR	1413	6374	4.511	19	94	4.947	1.095	-	nat.	-
BigData	1151	8641	7.507	659	5455	8.278	1.088	-	B	-
SMC	2698	17251	6.394	344	2479	7.206	1.075	19	-	-
ICAIL	190	1933	10.174	59	592	10.034	1.069	35	C	-
FIAIRS	1009	4649	4.608	601	2875	4.784	1.041	58	nat.	-
FG	437	11144	25.501	175	5868	33.531	1.041	-	-	C(12)
TIME	822	3349	4.074	102	653	6.402	1.037	91	B	-
ICAART	1384	4477	3.235	108	361	3.343	1.028	-	B	-
IJCAI	5670	132398	23.351	1669	39572	23.71	1.016	2	A*	A(7)

Есть 9 конференций, которые включены в обе таблицы: ICARCV, ICPR, ASRU, IE, GECCO, IRI, PACLIC, FlAIRS и IJCAI. Интересно, что конференции, получившие значение MNCS больше 1 для Китая, были в основном из рейтинга CORE, а для Соединенных Штатов конференции с MNCS> 1 были общими во всех трех рейтингах (MSAR, CORE и CCF). Это может свидетельствовать о том, что ученые из Китая при выборе конференций были больше ориентированы на рейтинг CORE, в то время как на ученых из Соединенных Штатов рейтинги конференций не влияют. Интересным фактом является то, что в таблице 2.4 (для Китая) было всего семь конференций из рейтинга CCF, а в таблице 2.5 (для Соединенных Штатов) было 10 конференций из этого рейтинга. Рейтинг CCF включает важные конференции для китайского научного сообщества, и все же исследователи из Соединенных Штатов получают на этих конференциях больше цитирований, чем ожидалось, больше, чем ученые из Китая. Это также подтверждает тот факт, что, хотя исследователи из Китая публикуют больше статей об искусственном интеллекте, публикации американских исследователей имеют большее количество цитирований и известность.

Чтобы наглядно представить динамику публикаций и цитирования исследователей из Китая и Соединенных Штатов по сравнению со средними значениями, мы создали столбчатые диаграммы. На рис. 2.2 показано годовое количество публикаций по всем конференциям в наборе данных и отдельно

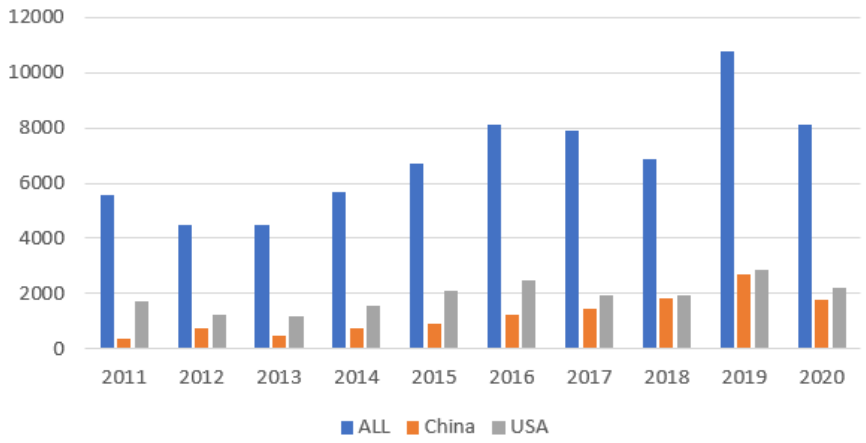

Рис. 2.2. Количество публикаций по годам, 2011–2020. Источник: собственные расчеты авторов, основанные на данных Scopus

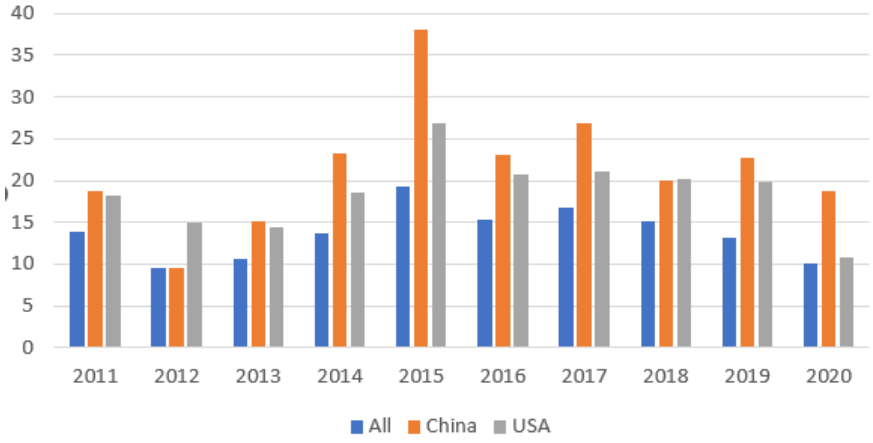

Рис. 2.3. Цитируемость на статью по годам, 2011-2020. Источник: собственные расчеты авторов, основанные на данных Scopus

для исследователей как из Китая, так и из Соединенных Штатов. На рис. 2.3 показана динамика количества цитирований в расчете на одну статью для одних и тех же групп.

Основываясь на графиках, мы можем сделать вывод, что, несмотря на тот факт, что американские исследователи публикуют больше статей на конференциях с высоким рейтингом, представленных в выборке, показатель цитируемости на одну статью у китайских исследователей выше почти во все периоды времени (исключая 2012 и 2018 годы). Значения цитируемости как для Китая, так и для Соединенных Штатов значительно превышают средний показатель цитируемости по выборке.

Таблица 2.6.

Корреляционные метрики

Показатели	Коэф. корреляции Спирмена	Значимость коэффициента	Доверительный интервал
MNCS Китая /MNCS США	-0.005	Не значим	(-0.259; 0.264)
MNCS Китая/MSAR	0.251	Не значим	(-0.182; 0.597)
MNCS США/MSAR	0.066	Не значим	(-0.374; 0.485)
MNCS Китая/CORE	0.147	Не значим	(-0.282; 0.557)
MNCS США/CORE	0.259	Не значим	(-0.140; 0.632)

Используя коэффициент корреляции Спирмена, мы определили близость (силу) и направление корреляционной связи между парами выборок: MNCS Китая и MNCS США, MNCS Китая/MNCS США и рейтинг MSAR, MNCS Китая/MNCS США и рейтинг CORE.

Коэффициент корреляции для пары MNCS Китая и MNCS США составил -0,005. При проверке значимости коэффициента он оказался незначимым, что указывает на отсутствие связи между этими двумя выборками. Мы провели один и тот же анализ для каждой пары сравниваемых данных, и результаты представлены в таблице 2.6. Мы провели наши расчеты с уровнем значимости 95%.

Основываясь на полученных значениях коэффициентов корреляции, можно сделать следующие выводы:

— MNCS Китая и MNCS США имеют слабую обратную зависимость, что указывает на слабую зависимость между ними;
— MNCS Китая и рейтинг MSAR имеют слабую связь;
— MNCS США и рейтинг MSAR не имеют связи и независимы друг от друга;
— MNCS Китая и рейтинг CORE не имеют связи;
— MNCS США и рейтинг CORE также не имеют корреляционной связи.

Из вышеизложенного можно сделать вывод, что как MNCS Китая, так и MNCS Соединенных Штатов не коррелируют ни друг с другом, ни с рейтингами CORE и MSAR.

Основываясь на проведенном анализе, можно сделать вывод, что между данными существует существенная взаимосвязь, и стратегия выбора конференций для публикации результатов, основанная на методах и выводах этого исследования, может быть эффективной и применимой для ученых из разных стран.

Таким образом, по результатам исследования мы пришли к выводу, что, несмотря на то что количество документов в материалах конференций в Китае выше (89 791) по сравнению с Соединенными Штатами (54 430), Соединенные Штаты по-прежнему лидируют по количеству цитирований и количеству конференций, на которых американские исследователи получили более высокие оценки (цитирований больше, чем ожидалось). Также можно сделать вывод, что ученые из Соединенных Штатов в большей степени ориентированы на участие в конференциях с высоким рейтингом, поскольку количество публикаций на конференциях из нашей выборки приходится на Соединенные Штаты (19 120) и Китай (12 179).

2.4. Методология прогнозирования показателя цитируемости методами дискриминантного анализа

Присвоения рейтинга или его прогнозирование для новых конференций является актуальной проблемой для исследователей в этой области. Для ее решения применяются различные статистические методы, о которых было упомянуто выше. При этом, в ходе изучения предыдущих исследований, не удалось найти таких, которые бы применяли дискриминантный анализ для этой цели, хотя этот метод статистического анализа одним из наиболее подходящих для этой цели, также являясь предпосылкой к машинному обучению.

Дискриминантный анализ заключается в разделении признаков на однородные группы (классификация) или отнесении новых объектов к уже известным группам. В зависимости от известных данных классификацию можно разделить на две задачи: классификация с обучением (есть обучающая выборка) и классификация без обучения (нет обучающей выборки). В ходе проведения дискриминации (интерпретации межгрупповых различий) необходимо найти линейную или нелинейную комбинацию переменных, которая оптимально разделяет интересующие нас группы, которая называется дискриминантной

функцией. В дискриминантном анализе, помимо задачи классификации, также ставится задача прогнозирования.

Алгоритм построения дискриминантной функции, отнесение не сгруппированных значений к определённой группе и оценка значимости функции будет разобран далее.

Сначала рассчитываются оценки векторов средних значений для каждой выборки:

$$\hat{a}_1^{(l)} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}^{(l)}, \hat{a}_1^{(l)} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{1i}^{(l)}, \dots, l = 1, \dots, p,$$

где i — номер обучающей выборки, j — номер измерения в каждой выборке, n_i — количество измерений в каждой обучающей выборке.

Векторов средних значений должно быть столько сколько представлено обучающих выборок.

Далее рассчитываются ковариационные матрицы для каждой обучающей выборки по формуле

$$\hat{\sigma}_{lq} = \frac{1}{n_1+n_2-2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}^{(l)}-\hat{a_j}^{(l)})\left(x_{ji}^{(q)}-\hat{a_j}^{(q)}\right), l, q = 1, \dots, p, \hat{\Sigma} = \{\sigma_{lq}\}$$

и находим обратную матрицу к матрице $\hat{\Sigma}$, которую обозначим $\hat{\Sigma}^{-1}$.

Также рассчитываем векторы $(\hat{a_1}-\hat{a_2})$ и $\frac{1}{2}(\hat{a_1}+\hat{a_2})$.

Рассчитываем коэффициенты дискриминантной функции

$$w = \hat{\Sigma}^{-1}(\hat{a_1}-\hat{a_2})$$

Вычислив

$$\left[X_0-\frac{1}{2}(\hat{a_1}+\hat{a_2})\right]^T \hat{\Sigma}^{-1}(\hat{a_1}-\hat{a_2}),$$

классифицируем наблюдение так, если значение функции ≥ 0, то наблюдение X_0 относится к первой обучающей выборке, иначе ко второй.

Дискриминантный анализ проводится при предположении, что объекты одному из двух классов. Также существуют ограничения по свойствам переменных, подлежащих дискриминантному анализу. Перечислим эти свойства:

1. $0<p<(n-2)$, т.е. количество дискриминантных переменных не должно превышать число объектов (n-2);

2. переменные должны быть измерены в метрической шкале;
3. дискриминантные переменные должны быть линейно независимы.

Чтобы проверить качество дискриминантных функций, которые были построены при соблюдении всех упомянутых условий, применяется статистическая проверка следующей гипотезы о равенстве математических ожиданий внутри классов или о незначимости дискриминации:

$$H_0 \,:\, a_1 = a_2 = \ldots = a_k$$

С целью проверки этой гипотезы обычно используют обобщенное расстояние Махаланобиса

$$D^2 = \sum_{j=1}^{k} n_j \left(a_j\text{-}a\right)^T \Sigma^{\text{-}1} \left(a_j\text{-}a\right),$$

D^2 – взвешенная сумма расстояний от вектора средних каждого класса (a_j) до общего вектора средних a. В том случае, если нулевая гипотеза принимается, то D^2 аппроксимируется F распределением Фишера.

Вторым методом проверки гипотезы H_0 – проверка с использованием статистики лямбда Уилкса:

$$\lambda = \frac{|\Sigma|}{|T|},$$

где $\hat{\Sigma} = \{\hat{\sigma}_{lq}\}$– оценки матрицы внутригрупповой ковариации, $\hat{T} = \left\{\hat{t}_{lq}\right\}$ – полная ковариационная матрица и $\hat{t}_{lq} = \frac{1}{n\text{-}k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji}^{(l)}\text{-}\hat{a}_j^{(l)}) \left(x_{ji}^{(q)}\text{-}\hat{a}_j^{(q)}\right), l, q = 1, \ldots, p$.

Эта статистика также аппроксимируется F распределением Фишера:

$$\frac{n\text{-}k\text{-}1}{p} \cdot \frac{1\text{-}\sqrt{\lambda}}{\sqrt{\lambda}} \backsim F_\alpha \left(2p; 2n\text{-}2k\text{-}2\right).$$

Если $\lambda = 0$, то дискриминация считается наилучшей, если же $\lambda = 1$, то она считается наихудшей.

Глава 3. Прогнозирование качества конференций с помощью дискриминантного анализа

3.1. Модель с учётом кумулятивного преимущества

3.1.1. Базовая модель

Пусть в некоторой научной области имеется n конференций [22]. Обозначим через $R^i(t) \geqslant 0$ числовую меру рейтинга i-й конференции в момент времени t. Рейтинг рассматривается как единая агрегированная величина.

Пусть динамика рейтинга каждой конференции определяется следующими механизмами:

— внутренним ростом;
— конкуренцией;
— естественным затуханием;
— внешними воздействиями.

Внутренний рост связан со стремлением увеличивать рейтинг за счёт внутренних усилий (привлечение известных докладчиков, повышение качества рецензирования, улучшение организации). Конкуренция вызвана взаимным торможением конференций, поскольку ресурсы (люди, деньги) ограничены. Естественное затухание (диссипация) связана с устареванием, потеря актуальности и т.д. Внешние воздействия вызваны непредсказуемыми факторами (чёрными лебедями) (например, смена программного комитета, публикация прорывных результатов, скандалы).

В качестве основы будем использовать модель Ферхюльста для изолированной конференции:

$$\frac{dR}{dt} = rR\left(1 - \frac{R}{K}\right) - \delta R,$$

где r — максимальная скорость роста, K — ёмкость (максимально достижимый рейтинг при отсутствии конкурентов), δ — коэффициент затухания.

Тогда равновесный рейтинг:

$$R^* = K(1 - \delta/r), \qquad r > \delta.$$

Для нескольких конференций введём конкурентное торможение. Рост каждой конференции замедляется не только за счёт её собственного рейтинга, но

и за счёт рейтингов других конференций.

$$\frac{dR^i}{dt} = r_i R^i \left(1 - \frac{\sum_{j=1}^{n} \alpha_{ij} R^j}{K_i}\right),$$

где α_{ij} — коэффициент влияния конференции j на конференцию i. Естественно предположить $\alpha_{ii} = 1$. α_{ij} для $i \neq j$ показывает, насколько сильно конкуренты подавляют рост i-й конференции.

Добавим затухание и внешнее воздействие:

$$\frac{dR^i}{dt} = r_i R^i \left(1 - \frac{\sum_{j=1}^{n} \alpha_{ij} R^j}{K_i}\right) - \delta_i R^i + \gamma_i F_i(t),$$

где:

- $r_i > 0$ — потенциальная скорость роста,
- $K_i > 0$ — максимально возможный рейтинг при отсутствии конкурентов,
- $\alpha_{ij} \geqslant 0$ — коэффициенты конкуренции,
- $\delta_i \geqslant 0$ — скорость естественного затухания,
- $\gamma_i \geqslant 0$ — чувствительность к внешним воздействиям,
- $F_i(t) \geqslant 0$ — функция внешнего импульса.

В более компактной форме можно переписать:

$$\frac{dR^i}{dt} = R^i \left(r_i - \frac{r_i}{K_i} \sum_{j=1}^{n} \alpha_{ij} R^j - \delta_i\right) + \gamma_i F_i(t).$$

Слагаемое $-\frac{r_i}{K_i} \alpha_{ij} R^i R^j$ описывает взаимное торможение.

3.1.2. Учёт закона Матфея

Закон Матфея [48; 49] есть проявление кумулятивного преимущества. Чем выше рейтинг конференции, тем легче ей привлекать лучших авторов, получать больше цитирований и, соответственно, ещё больше увеличивать свой рейтинг.

Зависимость от текущего рейтинга

Добавим член $\beta_i R^i$, который увеличивает скорость роста пропорционально текущему рейтингу:

$$\frac{dR^i}{dt} = r_i R^i \left(1 - \frac{\sum \alpha_{ij} R^j}{K_i}\right) + \beta_i R_\theta^i - \delta_i R^i + \gamma_i F_i(t),$$

где $\beta_i \geqslant 0$ — интенсивность кумулятивного преимущества.

Зависимость ёмкости от рейтинга

Ёмкость K_i можно сделать зависящей от рейтинга: чем выше рейтинг, тем больше места под солнцем:

$$K_i = K_i^{(0)} + \varkappa_i R^i,$$

где $\varkappa_i \geqslant 0$.

Тогда логистическое ограничение становится менее жёстким для лидеров, что способствует их дальнейшему росту.

Асимметричная конкуренция

Для учёта закона Матфея можно сделать α_{ij} зависящими от разницы рейтингов:

$$\alpha_{ij} = \alpha_{ij\,\text{base}} \cdot \exp\bigl(-\lambda(R^i - R^j)\bigr), \qquad R^i > R^j,$$

где $\lambda > 0$.

Конференция с большим рейтингом R^i испытывает меньшее торможение от конференции с меньшим R^j (и наоборот, слабая конференция сильнее подавляется лидером). Это порождает положительную обратную связь: лидеры становятся менее уязвимыми для конкурентов.

Пороговый эффект

Если кумулятивное преимущество очень сильное, система может проявлять бистабильность. Конференция становится лидером или остаётся на низком

уровне. Для учёта этого можно добавить член с порогом:

$$\frac{dR^i}{dt} = R^i \left(\rho_i (R^i - R_{\text{th}})\right) - \delta_i R^i + \dots$$

Кумулятивное преимущество

Кумулятивное преимущество означает, что прирост рейтинга пропорционален текущему рейтингу (или его степени) с положительным коэффициентом. В этом случае получаем:

$$\frac{dR^i}{dt} \approx (r_i + \beta_i) R^i.$$

Это приводит к экспоненциальному росту, который ограничивается только ёмкостью K_i и конкуренцией. Если у двух конференций изначально близкие параметры, но одна получила небольшое преимущество, оно будет усиливаться со временем, и система может сойтись к равновесию с сильным доминированием лидера. Равновесие становится неустойчивыми, возникает режим «победитель получает всё».

3.1.3. Семантика параметров

Для применения модели необходимо конкретизировать параметры $r_i, K_i, \alpha_{ij}, \delta_i, \gamma_i$ и функцию $F_i(t)$. Они должны выражаться через наблюдаемые характеристики конференций.

Скорость роста r_i

r_i отражает способность конференции быстро наращивать рейтинг. Может завить от разных показателей:

— качества программного комитета (например, средний h-индекс);
— финансовой поддержки;
— эффективности продвижения среди авторов.

Можно считать r_i постоянным.

Ёмкость K_i

K_i — максимальный рейтинг, которого конференция могла бы достичь в отсутствие конкурентов. Можно связать K_i с размером научного сообщества в данной области и долей, которую конференция может охватить. Также можно рассматривать K_i как предел, определяемый историческими максимумами рейтинга.

Коэффициент конкуренции α_{ij}

Коэффициент α_{ij} показывает, насколько присутствие конференции j снижает потенциал роста конференции i.

Способы задания:

— Тематическое перекрытие. Если две конференции близки по тематике, они конкурируют сильнее. Можно определить

$$\alpha_{ij} = \exp(-\lambda \cdot d_{ij}),$$

где d_{ij} есть семантическое расстояние между тематиками конференций, $\lambda > 0$ — масштабный коэффициент.

— Пересечение аудиторий.

$$\alpha_{ij} = \frac{|A_i \cap A_j|}{|A_i|}$$

или

$$\alpha_{ij} = \frac{|A_i \cap A_j|}{\sqrt{|A_i||A_j|}},$$

где A_i — множество авторов, публиковавшихся на конференции i за последние годы. Большое пересечение означает прямую конкуренцию за одних и тех же учёных.

— Симметричность. Примем $\alpha_{ij} = \alpha_{ji}$. То есть будем считать, что взаимное влияние одинаково.

— Нормировка. Будем считать, что $\alpha_{ii} = 1$.

Затухание δ_i

Затухание δ_i есть скорость снижения рейтинга при отсутствии новых достижений. Можно задать как величину, обратную характерному периоду полураспада рейтинга. Можно считать δ_i постоянным или связанным с возрастом конференции. Старые конференции быстрее теряют актуальность, если не обновляются.

Внешние воздействия $F_i(t)$ и чувствительность γ_i

Параметр $F_i(t)$ описывает воздействия, которые могут мгновенно изменить рейтинг. Приведём примеры.

— Дельта-импульсы в годы, когда произошло важное событие (например, смена главного редактора, получение престижной награды участниками и т.д.):

$$F_i(t) = \sum_k \delta(t - t_k).$$

— Временной всплеск:

$$F_i(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t - t_0)^2}{2\sigma^2}\right).$$

Параметр γ_i отражает то, как эффективно конференция способна воспринимать такие события.

3.1.4. Способы определения параметров по экспериментальным данным

Параметры можно получить r_i, K_i, α_{ij}, δ_i, γ_i на основе реальных временных рядов. Источником данных могут служить наукометрические базы.

Построение временного ряда рейтинга $R^i(t)$

В моделях рейтинг — абстрактная величина. Необходимо задать наблюдаемый показатель. Возможные варианты:

— Суммарное цитирование всех докладов конференции за год (или кумулятивно).
— Импакт-фактор конференции.

— Собственный композитный индекс — например, взвешенная сумма: доля принятых заявок, среднее число цитирований, индекс Хирша программного комитета.
— Нормализованный показатель (например, доля от максимального значения среди всех конференций).

$R^i(t)$ должен быть определён для каждого года существования конференции с достаточной частотой (обычно ежегодно).

Оценка коэффициентов конкуренции α_{ij}

Коэффициенты α_{ij} можно вычислить независимо от временной динамики, используя статические данные:

— По пересечению множеств авторов за несколько лет:

$$\alpha_{ij} = \frac{|A_i \cap A_j|}{\sqrt{|A_i||A_j|}}.$$

— По пересечению ключевых слов (тематическое сходство).
— По сетям цитирования. Если конференции часто ссылаются друг на друга, это может указывать не только на конкуренцию, но и на взаимное влияние. Сильное цитирование может означать и тематическую близость, и научный диалог.

Полученные α_{ij} затем используются как фиксированные величины при идентификации остальных параметров.

Определение параметров $r_i, K_i, \delta_i, \gamma_i$

Параметры определяются путём аппроксимации временных рядов решениями системы ОДУ. Задача сводится к минимизации функции потерь:

$$J(\theta) = \sum_{i=1}^{n} \sum_{t \in T} \left(R^i_{\text{мод}}(t, \theta) - R^i_{\text{набл}}(t) \right)^2,$$

где θ — вектор всех неизвестных параметров, $R^i_{\text{мод}}$ — численное решение системы ОДУ.

Методы оптимизации:

— Градиентный спуск с вычислением чувствительности для систем ОДУ.

— Эволюционные алгоритмы.
— Байесовская идентификация (позволяет получить не только оценки, но и интервалы неопределённости).

Можно сначала упростить систему: считать α_{ij} известными, а для каждой конференции подбирать параметры отдельно, игнорируя связи, а затем уточнять совместно.

3.1.5. Модель для двух конференций

Рассмотрим реализацию модели для двух конференций.

Симметричная конкуренция

Примем следующие упрощения:
— Две конференции: $R^1(t)$ и $R^2(t)$.
— Одинаковые внутренние параметры: $r_1 = r_2 = r$, $K_1 = K_2 = K$, $\delta_1 = \delta_2 = \delta$.
— Симметричная конкуренция (полное перекрытие аудиторий): $\alpha_{12} = \alpha_{21} = 1$.
— Внешние воздействия отсутствуют: $\gamma_i F_i(t) = 0$.
— Кумулятивное преимущество в простейшей линейной форме: $\beta_i R_\theta^i$ с $\theta = 1$ и $\beta_1 = \beta_2 = \beta \geqslant 0$.

Тогда система принимает вид:

$$\begin{cases} \frac{dR^1}{dt} = rR^1\left(1 - \frac{R^1+R^2}{K}\right) + \beta R^1 - \delta R^1, \\ \frac{dR^2}{dt} = rR^2\left(1 - \frac{R^1+R^2}{K}\right) + \beta R^2 - \delta R^2. \end{cases}$$

Вынесем общие множители:

$$\frac{dR^i}{dt} = R^i\left[r\left(1 - \frac{R^1+R^2}{K}\right) + \beta - \delta\right], \qquad i = 1, 2.$$

Обозначим через $\rho = r + \beta - \delta$ эффективную скорость роста при малых R^i. Условие $\rho > 0$ необходимо для нетривиальной динамики.

Система примет вид:

$$\frac{dR^i}{dt} = R^i\left(\rho - \frac{r}{K}(R^1 + R^2)\right), \quad i = 1, 2.$$

Кумулятивное преимущество β увеличивает эффективную скорость роста. Член βR^i не вносит асимметрии. Обе конференции получают одинаковое усиление. Здесь закон Матфея проявляется лишь как ускорение роста, но не как механизм усиления различий.

Для создания асимметричного кумулятивного преимущества необходимо ввести зависимость β_i от текущего рейтинга или сделать α_{ij} асимметричными.

Асимметричная модель с эффектом «победитель получает всё»

Введём асимметричную конкуренцию, где влияние конференции j на i зависит от разницы рейтингов:

$$\alpha_{ij} = \exp\bigl(-\lambda(R^i - R^j)\bigr), \qquad \lambda > 0.$$

Тогда для двух конференций:

$$\alpha_{11} = 1, \quad \alpha_{22} = 1, \quad \alpha_{12} = e^{-\lambda(R^1 - R^2)}, \quad \alpha_{21} = e^{-\lambda(R^2 - R^1)} = e^{\lambda(R^1 - R^2)}.$$

Система (без внешних воздействий, с учётом кумулятивного преимущества βR^i):

$$\begin{cases} \frac{dR^1}{dt} &= rR^1\left(1 - \frac{R^1 + \alpha_{12}R^2}{K}\right) + \beta R^1 - \delta R^1, \\ \frac{dR^2}{dt} &= rR^2\left(1 - \frac{\alpha_{21}R^1 + R^2}{K}\right) + \beta R^2 - \delta R^2. \end{cases}$$

Подставим выражения для α_{12}, α_{21}:

$$\begin{cases} \frac{dR^1}{dt} &= rR^1\left(1 - \frac{R^1 + e^{-\lambda(R^1 - R^2)}R^2}{K}\right) + (\beta - \delta)R^1, \\ \frac{dR^2}{dt} &= rR^2\left(1 - \frac{e^{\lambda(R^1 - R^2)}R^1 + R^2}{K}\right) + (\beta - \delta)R^2. \end{cases}$$

Для удобства введём $\mu = \beta - \delta$ (чистая скорость роста за счёт кумулятивного преимущества и затухания).

Система принимает вид:

$$\begin{cases} \dot{R}_1 &= rR^1\left(1 - \frac{R^1 + e^{-\lambda(R^1 - R^2)}R^2}{K}\right) + \mu R^1, \\ \dot{R}_2 &= rR^2\left(1 - \frac{e^{\lambda(R^1 - R^2)}R^1 + R^2}{K}\right) + \mu R^2. \end{cases}$$

Параметры:

- $r > 0$,
- $K > 0$,
- $\mu \in \mathbb{R}$,
- $\lambda \geqslant 0$.

Стационарные состояния

Ищем точки (R^{1*}, R^{2*}), где $\dot{R}^1 = \dot{R}^2 = 0$.

Возможны три типа решений:

- Тривиальное $(0, 0)$.
- Граничные $(R^{1*}, 0)$ и $(0, R^{2*})$. Одна конференция вытесняет другую.
- Внутреннее (R^*, R^*). Симметричное равновесие (при $\lambda = 0$).

Граничное состояние $(R^{1*}, 0)$ Положим $R^2 = 0$. Уравнение для R^1:

$$\dot{R}_1 = rR^1\left(1 - \frac{R^1}{K}\right) + \mu R^1 = R^1\left(r + \mu - \frac{r}{K}R^1\right) = 0.$$

Ненулевое решение: $R^{1*} = \frac{K}{r}(r + \mu) = K\left(1 + \frac{\mu}{r}\right)$. Для существования положительного равновесия требуется $r + \mu > 0$ (т.е. $\mu > -r$).

Аналогично, $(0, R^{2*})$ с $R^{2*} = K\left(1 + \frac{\mu}{r}\right)$.

Внутреннее симметричное равновесие (R^*, R^*) При $R^1 = R^2 = R$ имеем $\alpha_{12} = e^{-\lambda(0)} = 1$, $\alpha_{21} = 1$.

Уравнения становятся одинаковыми:

$$\dot{R} = rR\left(1 - \frac{2R}{K}\right) + \mu R = R\left(r + \mu - \frac{2r}{K}R\right) = 0.$$

Ненулевое решение:

$$R^* = \frac{K}{2r}(r + \mu) = \frac{K}{2}\left(1 + \frac{\mu}{r}\right).$$

Решение существует при $r + \mu > 0$. R^* равно половине граничного равновесия.

Асимметричные внутренние равновесия (при $\lambda > 0$) В общем случае система может иметь два асимметричных равновесия (седловые узлы), возникающие при бифуркации из симметричного. При $\lambda > 0$ симметричное равновесие теряет устойчивость, и появляются два устойчивых граничных равновесия (эффект «победитель получает всё»).

Линеаризованная устойчивость

Проведём линеаризацию в окрестности симметричного равновесия (R^*, R^*). Вычислим элементы матрицы Якоби.

Обозначим $f_1(R^1, R^2)$ и $f_2(R^1, R^2)$ правые части системы. В силу симметрии достаточно найти производные в точке (R^*, R^*).

Сначала выпишем производные от экспоненциальных членов:

$$\begin{cases} \frac{\partial}{\partial R^1} e^{-\lambda(R^1-R^2)} & = -\lambda e^{-\lambda(R^1-R^2)}, \\ \frac{\partial}{\partial R^2} e^{-\lambda(R^1-R^2)} & = \lambda e^{-\lambda(R^1-R^2)}. \end{cases}$$

В точке (R^*, R^*) $e^{-\lambda(R^*-R^*)} = 1$.

Для функции f_1:

$$f_1 = rR^1 - \frac{r}{K}R^{12} - \frac{r}{K}R^1R^2e^{-\lambda(R^1-R^2)} + \mu R^1.$$

Вычислим частные производные в (R^*, R^*):

$$\frac{\partial f_1}{\partial R^1}\Big|_* = r - \frac{2r}{K}R^* - \frac{r}{K}R^* e^{-\lambda(R^1-R^2)}\Big|_* - \frac{r}{K}R^1R^2 \cdot (-\lambda)e^{-\lambda(R^1-R^2)}\Big|_* + \mu.$$

Упрощаем, учитывая $e^{-\lambda(R^1-R^2)}\big|_* = 1$:

$$\frac{\partial f_1}{\partial R^1}\Big|_* = r - \frac{2r}{K}R^* - \frac{r}{K}R^* + \frac{r}{K}R^{2*}\lambda + \mu.$$

Подставляем $R^* = \frac{K}{2r}(r+\mu)$:

$$\frac{r}{K}R^* = \frac{r}{K}\cdot\frac{K}{2r}(r+\mu) = \frac{r+\mu}{2};$$
$$\frac{2r}{K}R^* = r+\mu;$$
$$\frac{r}{K}R^{2*} = \frac{r}{K}\cdot\frac{K^2}{4r^2}(r+\mu)^2 = \frac{K}{4r}(r+\mu)^2.$$

Тогда:

$$\left.\frac{\partial f_1}{\partial R^1}\right|_* = r-(r+\mu)-\frac{r+\mu}{2}+\frac{K}{4r}(r+\mu)^2\lambda+\mu.$$

Слагаемые r и $-r$ сокращаются, $-\mu+\mu=0$.

Остаётся:

$$\left.\frac{\partial f_1}{\partial R^1}\right|_* = -\frac{r+\mu}{2}+\frac{K\lambda}{4r}(r+\mu)^2 = (r+\mu)\left(-\frac{1}{2}+\frac{K\lambda}{4r}(r+\mu)\right).$$

Теперь $\frac{\partial f_1}{\partial R^2}$ в той же точке:

$$\frac{\partial f_1}{\partial R^2} = -\frac{r}{K}R^1 e^{-\lambda(R^1-R^2)} - \frac{r}{K}R^1R^2\cdot\lambda e^{-\lambda(R^1-R^2)}.$$

При (R^*, R^*):

$$\left.\frac{\partial f_1}{\partial R^2}\right|_* = -\frac{r}{K}R^* - \frac{r}{K}R^{2*}\lambda = -\frac{r+\mu}{2}-\frac{K\lambda}{4r}(r+\mu)^2.$$

В силу симметрии:

$$\left.\frac{\partial f_2}{\partial R^2}\right|_* = \left.\frac{\partial f_1}{\partial R^1}\right|_*, \quad \left.\frac{\partial f_2}{\partial R^1}\right|_* = \left.\frac{\partial f_1}{\partial R^2}\right|_*.$$

Таким образом, матрица Якоби в симметричном равновесии:

$$J = \begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

где

$$a = (r+\mu)\left(-\frac{1}{2}+\frac{K\lambda}{4r}(r+\mu)\right),$$
$$b = -\frac{r+\mu}{2}-\frac{K\lambda}{4r}(r+\mu)^2.$$

Собственные значения:

- $\lambda_1 = a + b$ (собственный вектор $(1, 1)$);
- $\lambda_2 = a - b$ (собственный вектор $(1, -1)$).

Вычислим:

$$a + b = (r + \mu)\left(-\frac{1}{2} + \frac{K\lambda}{4r}(r + \mu)\right) + \left(-\frac{r + \mu}{2} - \frac{K\lambda}{4r}(r + \mu)^2\right) = -(r + \mu).$$

Таким образом, $\lambda_1 = -(r + \mu)$. Поскольку для существования равновесия требуется $r+\mu > 0$, то $\lambda_1 < 0$, получаем устойчивость вдоль диагонали (синфазные колебания).

Второе собственное значение:

$$a - b = (r + \mu)\left(-\frac{1}{2} + \frac{K\lambda}{4r}(r + \mu)\right) - \left(-\frac{r + \mu}{2} - \frac{K\lambda}{4r}(r + \mu)^2\right) =$$
$$= (r + \mu)\left(-\frac{1}{2} + \frac{K\lambda}{4r}(r + \mu) + \frac{1}{2} + \frac{K\lambda}{4r}(r + \mu)\right) = (r + \mu) \cdot \frac{K\lambda}{2r}(r + \mu) = \frac{K\lambda}{2r}(r + \mu)^2.$$

Следовательно, $\lambda_2 = \frac{K\lambda}{2r}(r + \mu)^2$.

Анализ устойчивости: - $\lambda_2 > 0$ при $\lambda > 0$ и $r + \mu > 0$. - $\lambda_2 = 0$ при $\lambda = 0$. Симметричное равновесие может быть центром. Также из-за члена μR^i оно может быть узлом или седлом. При $\lambda = 0$ получаем $\lambda_2 = 0$, что говорит о нейтральной устойчивости.

При $\lambda > 0$ симметричное равновесие является седловой точкой (одно отрицательное собственное значение, одно положительное). Малые отклонения от симметрии будут расти, и система будет стремиться к одному из граничных устойчивых равновесий. Закон Матфея реализуется через асимметричную конкуренцию ($\lambda > 0$). Если одна конференция получает небольшое преимущество, она начинает меньше тормозиться конкурентами и в итоге полностью вытесняет соперника.

Фазовые портреты

Случай $\lambda = 0$ (модель без кумулятивного преимущества) При $\lambda = 0$ система сводится к:

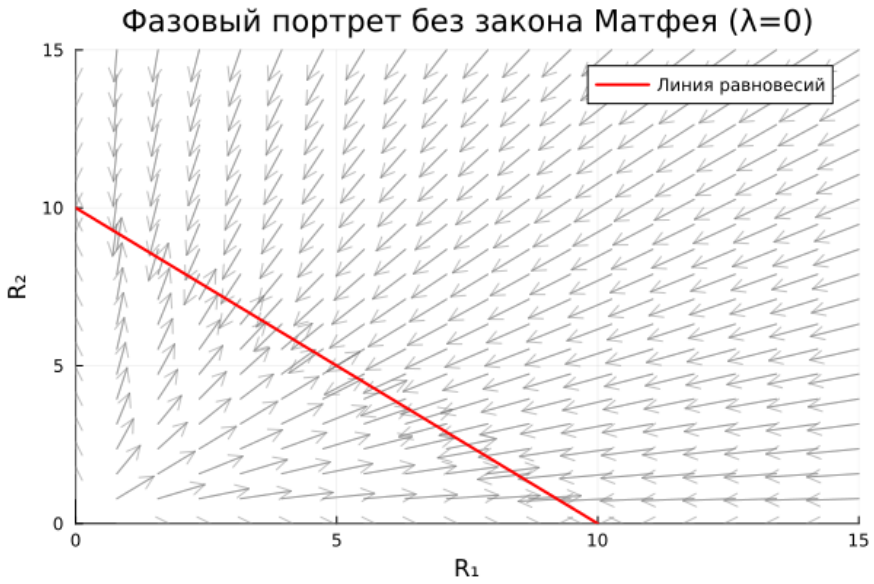

Рис. 3.1. Фазовый портрет для случая без кумулятивного преимущества

$$\begin{cases} \dot{R}_1 &= rR^1\left(1 - \frac{R^1+R^2}{K}\right) + \mu R^1, \\ \dot{R}_2 &= rR^2\left(1 - \frac{R^1+R^2}{K}\right) + \mu R^2. \end{cases}$$

Фазовый портрет (рис. 3.1):

- Изоклины: $\dot{R}_1 = 0$ и $\dot{R}_2 = 0$ совпадают и задаются прямой $R^1 + R^2 = \frac{K}{r}(r + \mu)$.
- Вся прямая является линией равновесия (неизолированные точки).
- Система вырождена: отношение R^1/R^2 сохраняется во времени (движение по лучам).
- Все траектории сходятся к точке на этой прямой, определяемой начальным отношением.
- Нет эффекта усиления различий. Любое начальное неравенство сохраняется.

Случай $\lambda > 0$ (асимметричная конкуренция, закон Матфея) Фазовый портрет качественно меняется (рис. 3.2):

- Симметричное равновесие (R^*, R^*) есть седло. Его устойчивое многообразие — прямая $R^1 = R^2$ (диагональ). Неустойчивое многообразие — направление $(1, -1)$ (отклонения от симметрии).

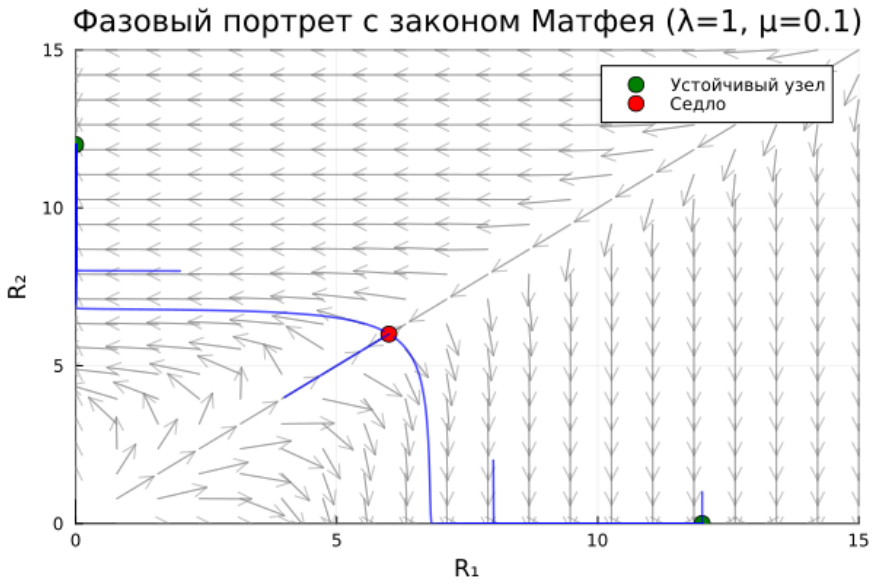

Рис. 3.2. Фазовый портрет для случая с кумулятивным преимуществом

— Два граничных равновесия $(R^{1*}, 0)$ и $(0, R^{2*})$ суть устойчивые узлы (при $\mu > -r$). К ним сходятся траектории из почти всей области фазовой плоскости, кроме сепаратрисы седла.
— Сепаратриса седла (неустойчивое многообразие) разделяет бассейны притяжения двух граничных равновесий. Она проходит через седловую точку и разграничивает начальные условия, приводящие к доминированию первой или второй конференции.

Если начальные рейтинги конференций не равны точно, то даже сколь угодно малая асимметрия приводит к вытеснению одного из конкурентов.

Докажем теорему.

Теорема 3.1 (О локальной устойчивости стационарных состояний). *Рассмотрим систему для двух конференций с идентичными параметрами:*

$$\begin{cases} \dot{R}_1 = rR_1\left(1 - \dfrac{R_1 + e^{-\lambda(R_1 - R_2)}R_2}{K}\right) + \mu R_1, \\ \dot{R}_2 = rR_2\left(1 - \dfrac{e^{\lambda(R_1 - R_2)}R_1 + R_2}{K}\right) + \mu R_2, \end{cases}$$

где $r > 0, K > 0, \mu > -r, \lambda \geq 0$.

Тогда имеем два случая.

Случай 1. При $\lambda = 0$ (симметричная конкуренция) множество точек прямой $R_1 + R_2 = K\left(1 + \frac{\mu}{r}\right)$ является непрерывным семейством устойчивых равновесий (нейтрально устойчивых).

Случай 2. При $\lambda > 0$ (закон Матвея) система имеет три изолированных равновесия: - симметричное (R_*, R_*) с $R_* = \frac{K}{2}\left(1 + \frac{\mu}{r}\right)$; - граничные $(R_{bound}, 0)$ и $(0, R_{bound})$ с $R_{bound} = K\left(1 + \frac{\mu}{r}\right)$.

При этом: граничные равновесия являются локально асимптотически устойчивыми (устойчивые узлы); симметричное равновесие является неустойчивым (седло).

Доказательство. Случай 1

Пусть $\lambda = 0$. Тогда $\alpha_{12} = \alpha_{21} = 1$, и система становится вырожденной:

$$\dot{R}_i = R_i\left(r + \mu - \frac{r}{K}(R_1 + R_2)\right), \quad i = 1, 2.$$

Сумма $S = R_1 + R_2$ удовлетворяет $\dot{S} = S\left(r + \mu - \frac{r}{K}S\right)$, что является логистическим уравнением. Любая точка с $S = K(1 + \mu/r)$ есть равновесие. Линеаризация в точке (R_1^0, R_2^0) на этой прямой даёт матрицу Якоби с собственными значениями $\lambda_1 = 0$ (вдоль прямой) и $\lambda_2 = -(r + \mu) < 0$ (поперёк). Следовательно, равновесия нейтрально устойчивы. Утверждение 1 доказано.

Случай 2.

Пусть $\lambda > 0$. Существование равновесий проверяется непосредственной подстановкой. Для исследования устойчивости проведём линеаризацию в симметричной точке (R_*, R_*). Матрица Якоби имеет вид:

$$J = \begin{pmatrix} a & b \\ b & a \end{pmatrix}, \qquad a = (r + \mu)\left(-\frac{1}{2} + \frac{K\lambda}{4r}(r + \mu)\right), \quad b = -\frac{r + \mu}{2} - \frac{K\lambda}{4r}(r + \mu)^2.$$

Собственные значения:

$$\lambda_1 = a + b = -(r + \mu) < 0, \qquad \lambda_2 = a - b = \frac{K\lambda}{2r}(r + \mu)^2 > 0.$$

Наличие положительного собственного значения $\lambda_2 > 0$ означает, что равновесие (R_*, R_*) является седлом (неустойчивым).

Для граничного равновесия $(R_{\text{bound}}, 0)$ запишем систему в переменных (R_1, R_2). Вблизи этой точки линеаризация даёт диагональную матрицу:

$$J_{\text{bound}} = \begin{pmatrix} -(r+\mu) & * \\ 0 & -(r+\mu) - \frac{r}{K} R_{\text{bound}} e^{\lambda R_{\text{bound}}} \end{pmatrix},$$

где оба диагональных элемента строго отрицательны ($r + \mu > 0$).

Следовательно, $(R_{\text{bound}}, 0)$ – устойчивый узел. Аналогично для $(0, R_{\text{bound}})$.

Таким образом, при $\lambda > 0$ граничные равновесия устойчивы, а симметричное – неустойчиво. Теорема доказана. □

Численный пример

Выберем условные параметры:

- $r = 0.5$ (скорость роста);
- $K = 10$ (ёмкость);
- $\delta = 0.2$ (затухание);
- $\beta = 0.3$ (кумулятивное преимущество);
- $\lambda = 1$ (сильная асимметрия конкуренции).

Тогда $\mu = \beta - \delta = 0.1 > -r$. Равновесия существуют.

Вычислим равновесия:

$$R^* = \frac{K}{2r}(r + \mu) = \frac{10}{2 \cdot 0.5}(0.5 + 0.1) = 10 \cdot 0.6 = 6.$$

Граничное равновесие: $R^{1\,*} = K(1 + \mu/r) = 10 \cdot (1 + 0.1/0.5) = 10 \cdot 1.2 = 12$.

Поведение:

- Симметричное равновесие $(6, 6)$ есть седло.
- Устойчивые узлы: $(12, 0)$ и $(0, 12)$.

Рассмотрим начальные точки.

- Почти симметричная начальная точка: $R^1(0) = 6.1,\ R^2(0) = 5.9$. Из-за положительного λ конкуренция асимметрична. Конференция 1 (с чуть большим рейтингом) испытывает меньшее торможение от конференции 2. Траектория быстро уходит к $(12, 0)$. R^1 растёт, R^2 сначала немного растёт, но затем падает до нуля.
- Почти симметричная, но с другим знаком: $R^1(0) = 5.9,\ R^2(0) = 6.1$. Здесь побеждает вторая конференция.
- Точное симметричное начальное условие $(6, 6)$. Система остаётся в равновесии, но это неустойчивое состояние. Любое возмущение приведёт

к монополизации. Время достижения монополии зависит от λ и начального дисбаланса. При $\lambda = 1$ даже дисбаланс 0.1 приводит к вытеснению за 5 условных единиц времени.

3.2. Программная реализация модели

3.2.1. Общие требования к программной реализации

Для проведения численных экспериментов, калибровки параметров модели по реальным данным, а также визуализации результатов была разработана программная система, реализующая математическую модель динамики рейтингов научных конференций [82].

Система создавалась в соответствии со следующими принципами:

— Целостность. Все компоненты (модель, методы анализа, сценарии обработки данных, визуализация) объединены в единую структуру, используют общие определения параметров и функций.
— Модульность. Каждый функциональный блок выделен в отдельный файл или модуль, что упрощает поддержку, тестирование и расширение.
— Воспроизводимость. Точная фиксация версий всех используемых библиотек, сохранение параметров каждого эксперимента с временной меткой, исключение «магических» путей к файлам.
— Масштабируемость. Возможность работы с произвольным числом конференций, автоматическая адаптация кода при изменении размерности.
— Интеграция с данными. Встроенные средства импорта из распространённых форматов (Excel, CSV), интерполяции пропусков и нормализации рядов.
— Документированность. Каждый модуль содержит комментарии, поясняющие математический смысл и параметры.

Разработанная система распространяется как открытое программное обеспечение.

3.2.2. Инструментальные средства

Язык программирования Julia

Для реализации комплекса выбран язык Julia. Основные причины выбора:

— Высокая производительность, близкая к компилируемым языкам (C, Fortran), при сохранении динамической типизации и удобстве научных вычислений.
— Развитая экосистема для численного решения ОДУ (`DifferentialEquations.jl`), оптимизации (`Optimization.jl`), автоматического дифференцирования (`ForwardDiff.jl`).
— Встроенный пакетный менеджер, обеспечивающий воспроизводимость окружения.
— Удобный синтаксис для работы с многомерными массивами и векторизации.

Фреймворк DrWatson

Для организации проекта и обеспечения воспроизводимости использован специализированный фреймворк DrWatson [15]. Он предоставляет:

— стандартизированную структуру каталогов (`src/`, `scripts/`, `data/`, `plots/`);
— функции для безопасного доступа к данным (`datadir()`, `plotsdir()`, `simulationsdir()`);
— автоматическое сохранение параметров симуляций в формате JLD2;
— средства для быстрого переключения между проектами и управления зависимостями.

Основные библиотеки

Все зависимости зафиксированы в файлах `Project.toml` и `Manifest.toml`, что позволяет развернуть окружение на любом компьютере командой `import Pkg; Pkg.instantiate()`.

Используются следующие основные библиотеки:

— `DifferentialEquations.jl`: численное интегрирование систем ОДУ (методы `Tsit5`, `Rodas5`) (методы Рунге–Кутты);
— `ModelingToolkit.jl`: символьное построение систем (опционально);
— `Optimization.jl` + `OptimizationOptimJL.jl`: решение задач оптимизации (LBFGS, NelderMead);
— `ForwardDiff.jl`: автоматическое дифференцирование для вычисления матрицы Якоби;

— `Plots.jl`: визуализация (временные ряды, фазовые портреты);
— `CSV.jl`, `DataFrames.jl`: работа с табличными данными;
— `XLSX.jl`: импорт из Excel-файлов;
— `Interpolations.jl`: интерполяция пропущенных значений;
— `DrWatson.jl`: управление проектом и воспроизводимость.

3.2.3. Архитектура программного комплекса

Общая структура проекта представлена на рис. и соответствует структуре фреймворка DrWatson.

```
ConferenceRanking/
│
├── Project.toml                # зависимости проекта
├── Manifest.toml               # точные версии всех пакетов
│
├── src/                        # исходный код (ядро)
│
├── scripts/                    # исполняемые сценарии
│
├── data/                       # данные
│
├── plots/                      # выходные графики (PNG, PDF)
│
└── docs/                       # документация
```

Ядро (`src/`) не зависит от сценариев и может использоваться любым скриптом. Сценарии (`scripts/`) импортируют ядро и реализуют конкретные вычислительные эксперименты. Данные и графики хранятся вне исходного кода, что облегчает их версионирование и обмен.

3.2.4. Реализация ключевых компонентов

Модуль модели

Файл: `src/model.jl`

Модель реализована в виде функции, которая вычисляет производные для заданного состояния `u` в момент времени `t` с параметрами `p`. Параметры передаются плоским вектором, что стандартно для `DifferentialEquations.jl`.

Функция поддерживает произвольное число конференций `n`. Вычислительный цикл использует простую вложенную сумму. Для больших `n` (более 50) возможна оптимизация с использованием предвычисленных матриц.

Анализ устойчивости

Файл: `src/stability.jl`.

Для анализа локальной устойчивости стационарных точек реализована функция `compute_jacobian`, вычисляющая матрицу Якоби с помощью автоматического дифференцирования (`ForwardDiff.jl`). На основе собственных значений матрицы производится классификация особой точки.

Предобработка данных

Скрипт `dataset.jl` реализует алгоритм, представленный на рис. 3.3.

Калибровка параметров

Файл: `scripts/calibrate_model.jl`.

Калибровка сводится к минимизации суммы квадратов отклонений между модельным решением и временным рядом рейтингов, полученным из наукометрической базы. Оптимизационная задача решается методом LBFGS (ограниченный вариант BFGS) с использованием пакета `Optimization.jl`.

Алгоритм калибровки:

- Загрузка CSV-файла с годами и значениями рейтинга для выбранных конференций.
- Нормализация рядов делением на глобальный максимум (интервал $[0, 1]$).
- Построение функции потерь, решающей ОДУ с текущими параметрами и сравнивающей результат с данными.
- Запуск оптимизатора с начальным приближением (примерные значения: $r_i = 0.5$, $K_i = 1.0$, $\mu_i = 0.1$, $\lambda = 1.0$).
- Сохранение оптимальных параметров в JLD2-файл вместе с метаданными (временная метка, число итераций, финальная ошибка).

Процесс калибровки изображён на рис. 3.4.

Пример результирующих графиков приведён на рис. 3.5.

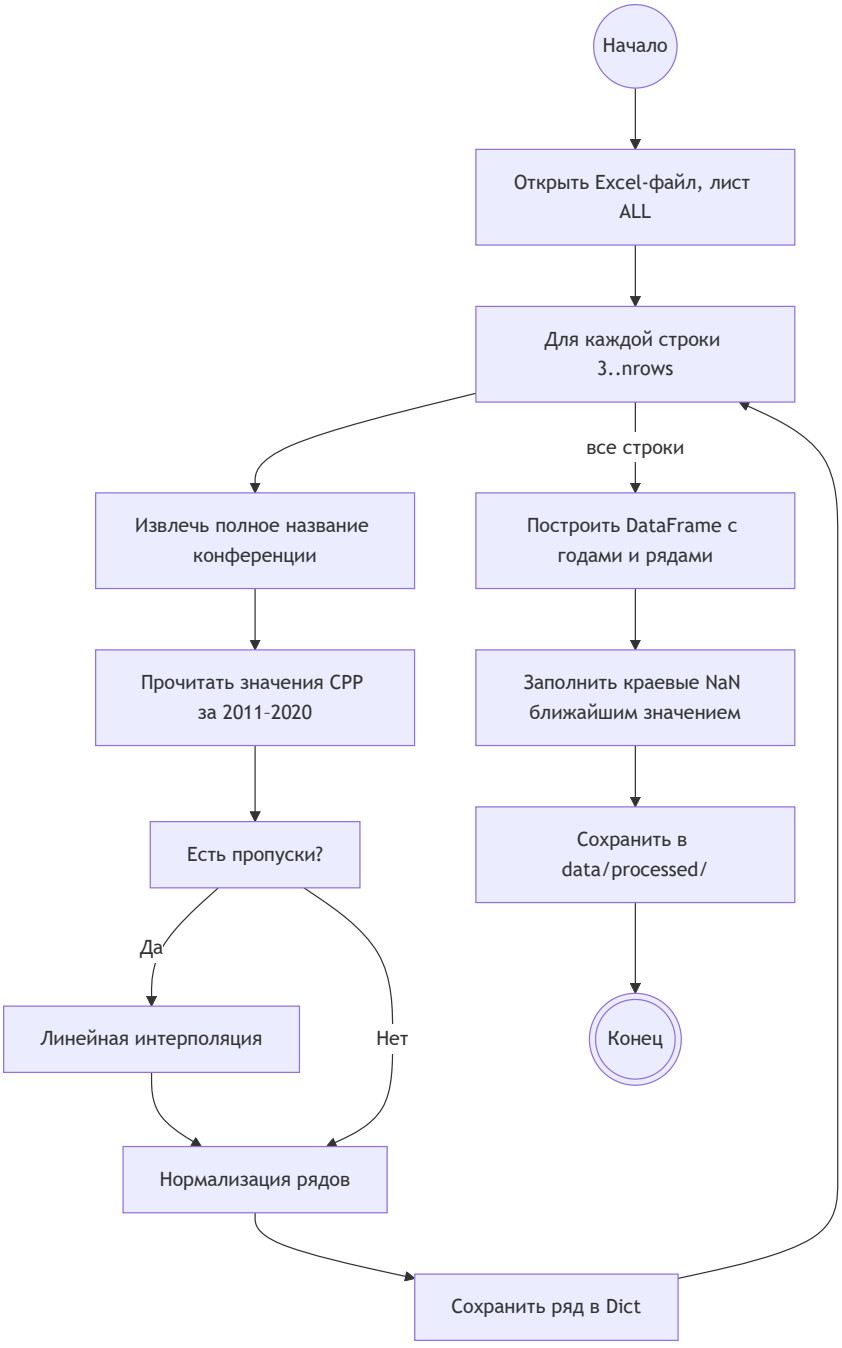

Рис. 3.3. Алгоритм предобработки данных из Excel в CSV с интерполяцией

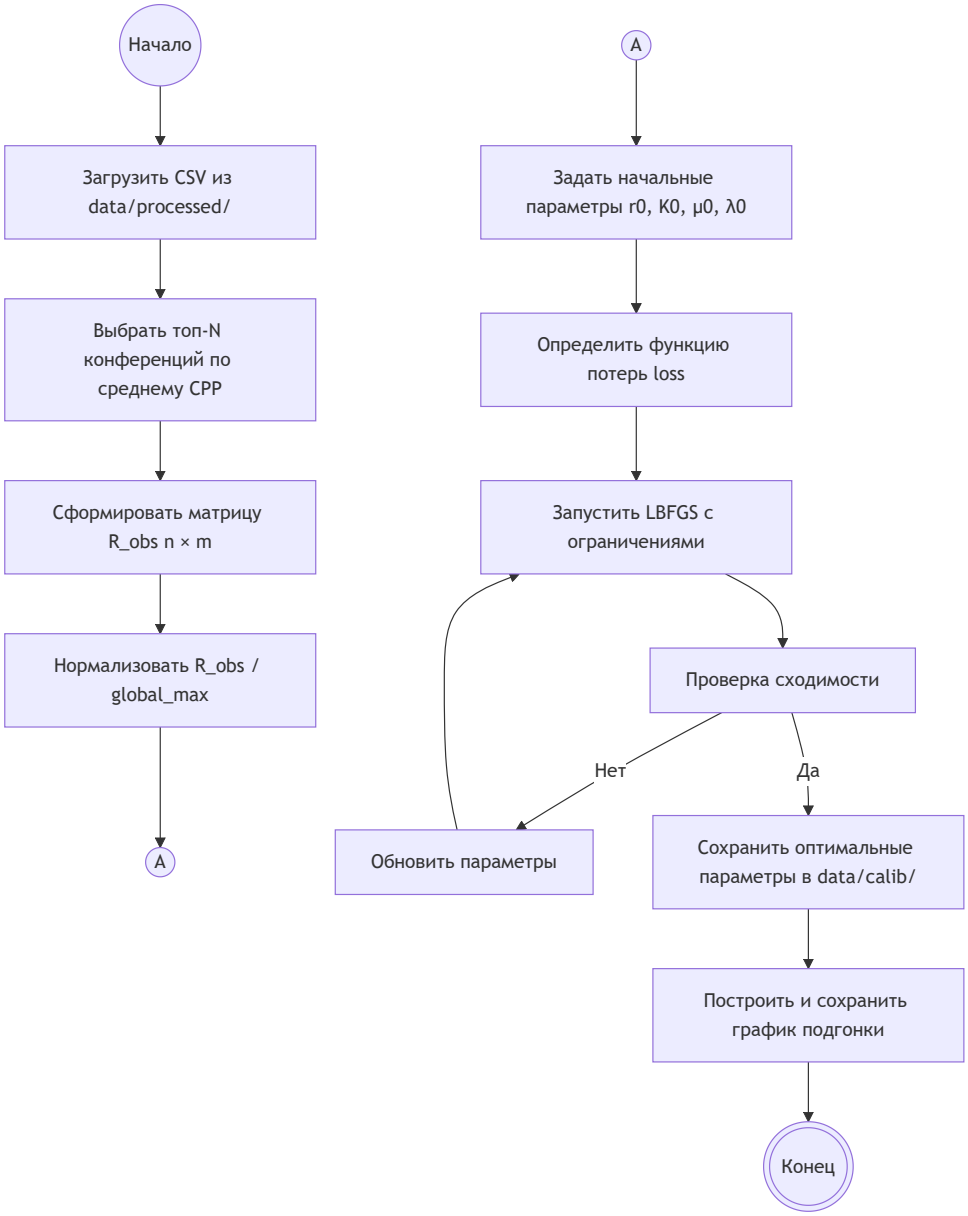

Рис. 3.4. Алгоритм калибровки модели по временным рядам рейтингов

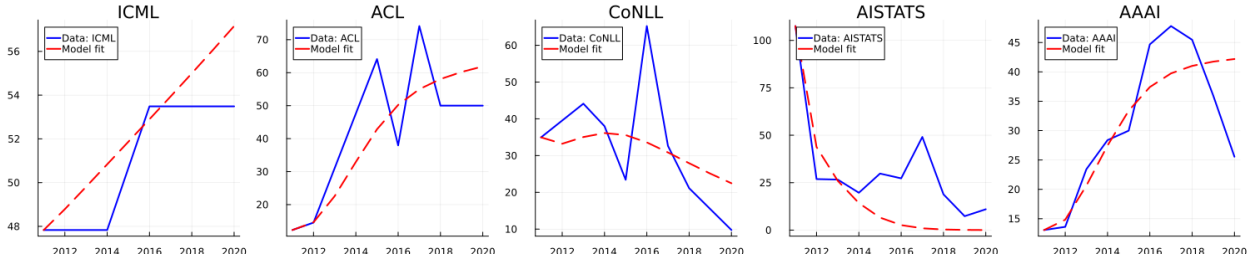

Рис. 3.5. Пример калибровки модели по временным рядам рейтингов

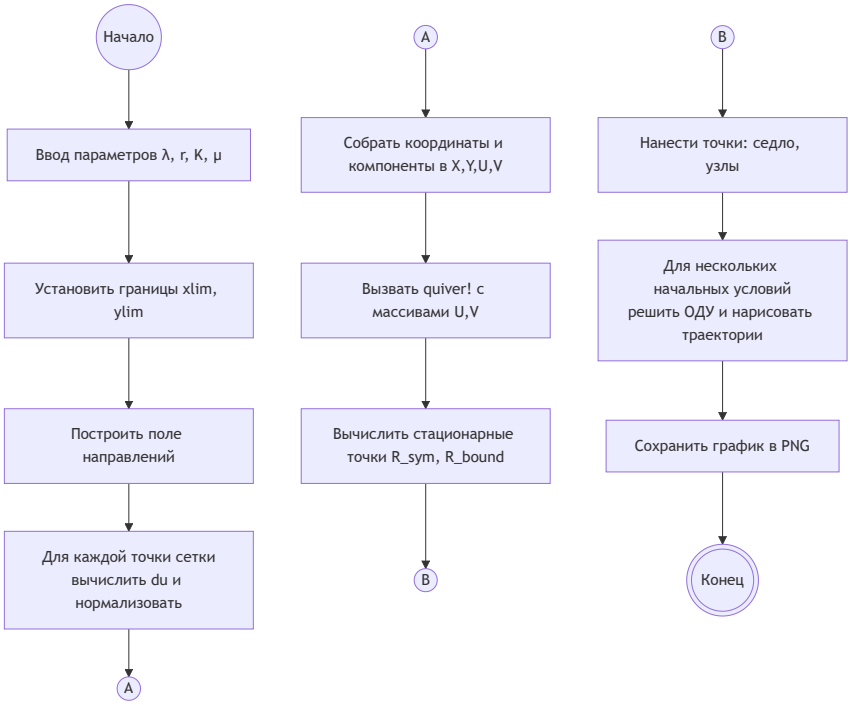

Рис. 3.6. Алгоритм генерации фазового портрета для двух конференций

Визуализация

Для визуализации используется библиотека `Plots.jl`. Реализованы следующие типы графиков:

— Временные ряды. Зависимость $R_i(t)$ для всех конференций на одном рисунке.

— Фазовые портреты. Только для двух конференций. Строятся поле направлений, траектории из нескольких начальных условий и стационарные точки.

— Графики подгонки. Сравнение реальных данных и модельной кривой после калибровки.

Например, построения фазового портрета производится следующим образом (рис. 3.6)

3.2.5. Обеспечение воспроизводимости

Воспроизводимость является ключевым требованием к вычислительным исследованиям. В разработанном комплексе она достигается следующим образом:

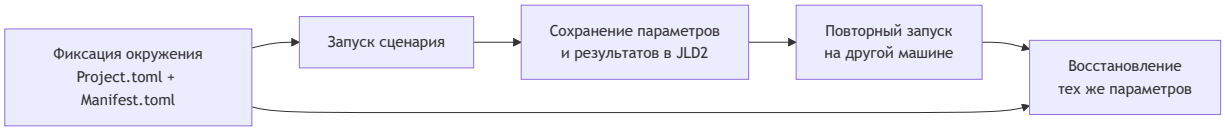

Рис. 3.7. Схема воспроизводимости

— Фиксация окружения. Файлы `Project.toml` и `Manifest.toml` содержат точные версии всех использованных пакетов. Любой пользователь может воссоздать идентичное окружение.
— Сохранение параметров экспериментов. Все результаты калибровки и симуляций сохраняются в формате JLD2 вместе с метаданными. Это позволяет в будущем точно восстановить, какие входные данные и настройки привели к данному результату.
— Использование фреймворка DrWatson. Функции `datadir()`, `plotsdir()`, `simulationsdir()` гарантируют, что пути к файлам не зашиты жёстко, а формируются относительно корня проекта. Это исключает ошибки при переносе на другую машину.

На рис. 3.7 показано, как достигается воспроизводимость экспериментов.

3.2.6. Программный комплекс

Очень часто при выполнении исследований по математическому моделированию разрабатываются отдельные скрипты, которые слабо связаны между собой, используют жёстко закодированные пути и не обеспечивают воспроизводимости. При создании программ мы ориентировались именно на создание программного комплекса [80; 81]. Созданный программный комплекс обладает следующими важными характеристиками (таб. 3.1).

Таким образом, комплекс представляет собой научно-исследовательскую программную систему, пригодную для дальнейшего использования как самим автором, так и другими исследователями, занимающимися наукометрическим моделированием.

3.2.7. Потенциальные направления расширения

Благодаря модульной архитектуре, программный комплекс легко расширять:

Таблица 3.1.

Программный комплекс и набор скриптов

Характеристика	Набор скриптов	Программный комплекс
Модульность	Код дублируется, изменения разрозненны	Единое ядро (`src/`), сценарии только вызывают функции
Воспроизводимость	Зависимости не фиксированы	`Project.toml` + `Manifest.toml`, фиксация версий
Работа с данными	Пути жёсткие, ручное копирование	Настраиваемое размещение
Сохранение экспериментов	Вручную сохраняются графики, параметры теряются	Все параметры и результаты – в JLD2 с метаданными
Документация	Отсутствует	Комментарии в коде, описание модулей

— Добавление стохастического члена (уравнения случайных колебаний).
— Введение запаздывания для описания влияние рейтингов прошлых лет на текущее состояние.
— Иерархические модели для описания конкуренция на нескольких уровнях (страны, научные области).
— Интерактивный графический интерфейс.

Все эти модификации не потребуют переписывания уже работающих сценариев, достаточно расширить ядро.

3.3. Применения методов дискриминантного анализа для прогнозирования качества конференций

Для проведения дискриминантного анализа была собрана обучающая выборка, которая содержит следующие переменные:

Y – случайная величина (с.в.), принимающая значения 1, 2, 3 либо 4 – квартиль научной конференции;

X1 – неотрицательная с.в., принимающая значения из множества действительных чисел, - средняя цитируемость материалов конференции (количество цитирований, приходящихся на один доклад за последние 10 лет);

X2 – целочисленная положительная с.в. – количество участников конференции;

X3 – целочисленная положительная с.в. – количество докладов на конференции;

X4 – целочисленная положительная с.в. – количество участников, представивших более одного доклада;

X5 – целочисленная с.в., принимающая два значения: 0 либо 1 - показатель публикационной активности организаторов конференции (1 – если организаторы представили доклад на конференцию и 0 – в противном случае);

X6 – неотрицательная с.в., принимающая значения из множества действительных чисел – показатель публикационной активности организаторов конференции, равный средней цитируемости научных публикаций, приходящейся на одного организатора конференции.

В таблице 3.2 представлена обучающая выборка значений с.в. Y, X1 – X6, составленная по материалам сайтов [17; 58; 67].

Проведя предварительный анализ выборки, который будет описан далее, дискриминантная функция для независимой переменной Y будет строиться с использованием факторов X1 и X2.

Из таблицы а будем рассматривать столбец с переменными Y, X1, X2. Так как информации о виде распределении нет, то расчеты проводились в соответствии с алгоритмом непараметрического дискриминантного анализа. Первым этапом является расчет оценок векторов средних значений для каждой из четырех обучающих выборок, первая выборка соответствует значению Y=1, вторая – Y=2, третья Y=3 и четвертая Y=4.

$$\hat{a_1} = (39,56; 64)\,, \hat{a_2} = (8,72; 46,18)\,, \hat{a_3} = (4,05; 18,67)\,, \hat{a_4} = (2,3; 127,5)$$

Вторым этапом является расчёт ковариационных матриц для каждой обучающей выборки:

$$\hat{\Sigma}(1) = \begin{pmatrix} 55,2 & 37,39 \\ 55 & -64,21 \end{pmatrix}, \hat{\Sigma}(2) = \begin{pmatrix} 18,93 & 16,65 \\ 74 & 47,99 \end{pmatrix},$$

Таблица 3.2.

Обучающая выборка

№	Y	X1	X2	X3	X4	X5	X6
1	1	55,20	55	146	2	1	36,17
2	1	37,85	58	126	8	0	32,50
3	1	25,62	79	153	3	1	9,33
4	2	18,93	74	139	9	0	16,21
5	2	16,38	48	132	0	1	7,17
6	2	14,39	95	143	6	1	12,21
7	2	7,06	31	48	5	0	8,83
8	2	7,03	30	87	0	1	15,83
9	2	6,87	59	105	8	0	17,32
10	2	6,46	33	94	9	1	16,58
11	2	6,04	30	78	0	0	4,83
12	2	5,95	31	66	1	0	23,21
13	3	5,34	26	33	0	1	8,50
14	2	3,69	25	48	0	1	10,51
15	3	3,42	17	34	0	1	10,67
16	3	3,39	13	26	2	0	18,67
17	4	3,20	95	255	9	1	22,37
18	2	3,07	52	100	5	1	25,87
19	4	2,48	110	301	5	0	14,83
20	4	2,42	157	345	0	1	20,17
21	4	2,04	99	282	8	0	16,71
22	4	1,89	135	380	7	1	25,60
23	4	1,76	169	382	2	1	10,50
24		55,2	55	146	12	1	36,17
25		6,46	33	94	8	1	23,67
26		0,87	27	72	13	1	21,5

$$\hat{\Sigma}(3) = \begin{pmatrix} 5,34 & 0,0003 \\ 26 & 0,03 \end{pmatrix},\ \hat{\Sigma}(4) = \begin{pmatrix} 3,2 & 0,082 \\ 95 & -2,35 \end{pmatrix}$$

Далее рассчитываем попарные суммарные ковариационные матрицы

$$\hat{\Sigma}_{12} = \frac{1}{n_1 + n_2 - 2}\left(n_1\hat{\Sigma}(1) + n_2\hat{\Sigma}(2)\right) =$$

$$= \frac{1}{3+11-4}\left(3\begin{pmatrix} 55,2 & 37,39 \\ 55 & -64,21 \end{pmatrix} + 11\begin{pmatrix} 18,93 & 16,65 \\ 74 & 47,99 \end{pmatrix}\right) =$$

$$= \frac{1}{10}\begin{pmatrix} 373,83 & 295,35 \\ 979 & 335,27 \end{pmatrix} = \begin{pmatrix} 37,38 & 29,54 \\ 9,79 & 33,53 \end{pmatrix}$$

Найдем обратную матрицу к объединённой матрице ковариаций

$$\hat{\Sigma}_{12}^{-1} = \begin{pmatrix} 0,035 & -0,031 \\ -0,01 & 0,039 \end{pmatrix}.$$

Коэффициенты дискриминантной функции

$$w_1 = \hat{\Sigma}_{12}^{-1}(\hat{a_1} - \hat{a_2}) = \begin{pmatrix} 0,035 & -0,031 \\ -0,01 & 0,039 \end{pmatrix}(30,8417,82) = \begin{pmatrix} 0,53 \\ 0,39 \end{pmatrix}$$

Далее рассчитываем:

$$\frac{1}{2}(\hat{a_1} + \hat{a_2}) = \frac{1}{2}\begin{pmatrix} 48,28 \\ 110,18 \end{pmatrix} = \begin{pmatrix} 24,14 \\ 55,09 \end{pmatrix}.$$

Классифицируем первое наблюдение

$$\begin{pmatrix} 55,2 \\ 55 \end{pmatrix} - \begin{pmatrix} 24,14 \\ 55,09 \end{pmatrix} = \begin{pmatrix} 31,06 \\ -0,09 \end{pmatrix}$$

и

$$\begin{pmatrix} 31,06 \\ -0,09 \end{pmatrix}^T \cdot \begin{pmatrix} 0,53 \\ 0,39 \end{pmatrix} = 16,43 > 0,$$

следовательно, наблюдение классифицируем в первую группу. Аналогично проверим второе наблюдение на принадлежность к первой группе.

$$\begin{pmatrix}6,46\\33\end{pmatrix}-\begin{pmatrix}24,14\\55,09\end{pmatrix}=\begin{pmatrix}-17,68\\-22,09\end{pmatrix}$$

и

$$\begin{pmatrix}-17,68\\-22,09\end{pmatrix}^{T}\begin{pmatrix}0,53\\0,39\end{pmatrix}=-17.99<0,$$

второе наблюдение принадлежит ко второй группе.

$$\begin{pmatrix}0,87\\27\end{pmatrix}-\begin{pmatrix}24,14\\55,09\end{pmatrix}=\begin{pmatrix}-23,27\\-28,09\end{pmatrix}$$

и

$$\begin{pmatrix}-23,27\\-28,09\end{pmatrix}^{T}\begin{pmatrix}0,53\\0,39\end{pmatrix}=-23.288<0,$$

третье наблюдение также не принадлежит первой группе.

Теперь рассчитаем попарной ковариационной матрицы для третьей и четвёртой групп.

$$\hat{\Sigma}_{34}=\frac{1}{n_3+n_4-2}\left(n_3\hat{\Sigma}(3)+n_4\hat{\Sigma}(4)\right)=$$

$$=\frac{1}{3+6-4}\left(3\begin{pmatrix}5,34 & 0,0003\\26 & 0,03\end{pmatrix}+6\begin{pmatrix}3,2 & 0,082\\95 & -2,35\end{pmatrix}\right)=$$

$$=\frac{1}{5}\begin{pmatrix}35,22 & 0,49\\648 & -14,124\end{pmatrix}=\begin{pmatrix}7,044 & 0,098\\129,6 & -2,81\end{pmatrix}$$

Найдем обратную матрицу к объединённой матрице ковариаций

$$\hat{\Sigma}_{34}^{-1}=\begin{pmatrix}0,087 & 0,003\\3,988 & -0,217\end{pmatrix}$$

Коэффициенты дискриминантной функции

$$w_3=\hat{\Sigma}_{34}^{-1}(\hat{a_3}-\hat{a_4})=\begin{pmatrix}0,087 & 0,003\\3,988 & -0,217\end{pmatrix}(1,75\ -106,83)=\begin{pmatrix}-0,168\\30,161\end{pmatrix}.$$

Далее рассчитываем

$$\frac{1}{2}(\hat{a_3} + \hat{a_4}) = \frac{1}{2}\begin{pmatrix} 6,35 \\ 146,17 \end{pmatrix} = \begin{pmatrix} 3,18 \\ 73,09 \end{pmatrix}.$$

Классифицируем второе наблюдение

$$\begin{pmatrix} 0,87 \\ 27 \end{pmatrix} - \begin{pmatrix} 3,18 \\ 73,09 \end{pmatrix} = \begin{pmatrix} -2,31 \\ -46,09 \end{pmatrix}$$

и

$$\begin{pmatrix} -2,31 \\ -46,09 \end{pmatrix}^T \begin{pmatrix} -0,168 \\ 30,161 \end{pmatrix} = -1389,732 < 0,$$

следовательно, наблюдение принадлежит к четвёртой группе.

3.4. Сравнение методов статистического анализа

На основе данных, представленных в таблице 3.3, построим модель линейной регрессии, отражающую зависимость Y от перечисленных выше факторов. Построение будем проводить с помощью статистического пакета SPSS.

В начале мы оценим степень линейной зависимости Y от X1 – X6, построив матрицу парных корреляций Пирсона. Исследование показало, что значимая зависимость наблюдается между Y и факторами X1, X2, X3 (таблица 3.4). Факторы X4 - X6 слабо влияют на значения Y, поэтому в дальнейшем мы их учитывать не будем. При этом, сильная связь наблюдается между факторами X2 и X3.

Чтобы избежать негативного влияния мультиколлинеарности, мы исключим из рассмотрения фактор X3 и построим двухфакторную регрессионную модель Y(X1,X2).

В результате получено уравнение (см. табл. 3.5).

Заметим, что в таблице 3.5 даны не только абсолютные значения коэффициентов модели, но и приведены результаты проверки их значимости с помощью T-критерия. Согласно данным из последнего столбца таблицы, все коэффициенты значимы с уровнем значимости, не превышающим 10^{-3}. Во втором столбце таблицы подсчитаны оценки среднеквадратического отклонения σ_j коэффициентов модели, а также их значения после стандартизации. Согласно

Таблица 3.3.

Обучающая выборка

№	Y	X1	X2	X3	X4	X5	X6
1	1	55,20	55	146	2	1	36,17
2	1	37,85	58	126	8	0	32,50
3	1	25,62	79	153	3	1	9,33
4	2	18,93	74	139	9	0	16,21
5	2	16,38	48	132	0	1	7,17
6	2	14,39	95	143	6	1	12,21
7	2	7,06	31	48	5	0	8,83
8	2	7,03	30	87	0	1	15,83
9	2	6,87	59	105	8	0	17,32
10	2	6,46	33	94	9	1	16,58
11	2	6,04	30	78	0	0	4,83
12	2	5,95	31	66	1	0	23,21
13	3	5,34	26	33	0	1	8,50
14	2	3,69	25	48	0	1	10,51
15	3	3,42	17	34	0	1	10,67
16	3	3,39	13	26	2	0	18,67
17	4	3,20	95	255	9	1	22,37
18	2	3,07	52	100	5	1	25,87
19	4	2,48	110	301	5	0	14,83
20	4	2,42	157	345	0	1	20,17
21	4	2,04	99	282	8	0	16,71
22	4	1,89	135	380	7	1	25,60
23	4	1,76	169	382	2	1	10,50

Таблица 3.4.

Матрица парных корреляций Пирсона

	Y	X1	X2	X3
Y	1	-0.678	0.588	0.666
X1	-0.678	1	-0.114	-0.141
X2	0.588	-0.114	1	0.959
X3	0.666	-0.141	0.959	1

Таблица 3.5.

Коэффициенты

Модель	Нестан-дарт.	Ст.ошибка	Стандартизир.	t	Знач.
Константа	2.237	0.246		9.082	0.000
X1	-0.049	0.009	-0.62	-5.244	0.000
X2	0.012	0.003	0.517	4.377	0.000

Таблица 3.6.

Сводка для модели

Модель	R	R - квадрат	Скор. R-квадрат	Ст. ошибка оц.	Дарбин-Уотсон
1	0.851	0.724	0.697	0.572	1.722

этим данным, изменение j-го коэффициента модели на одно σ_j влечет изменение Y примерно на 0,62 σ_j в сторону уменьшения для коэффициента при X1 и на 0,517 σ_j в сторону увеличения для коэффициента при X2.

Согласно данным, представленным в таблице 3.6, построенная модель на 85,1% отражает реальную зависимость квартиля от цитируемости материалов и количества участников конференции. При этом 72,4% вариации Y в нашей модели связано с изменчивостью факторов X1 и X2. Сама модель значима на уровне значимости, не превышающем 10^{-3} (таб. 3.7).

Автокорреляция остатков в построенной модели отсутствует, т.к. статистика Дарбина-Уотсона, равная 1,722 (таб. 3.5), попадает в интервал $(d_u; 4 - d_u)$, где $d_u = 1.33$ (согласно таблице критических значений для уровня значимости $\alpha = 0.05$).

Отсутствие автокорреляции остатков в совокупности с условием независимости результатов наблюдений фактически означает выполнение условий

Таблица 3.7.

Дисперсионный анализ

Модель	Сумма квад.	Ст. своб.	Средн. квадр.	F	Знач.
Регрессия	17.196	2	8.598	26.283	0.000
Остаток	6.543	20	0.327		
Всего	23.739	22			

теоремы Гаусса-Маркова, на основании которой справедливо следующее утверждение: модель (1) является моделью с минимальной дисперсией среди всех линейных моделей фиксированного уровня значимости α.

Определим оценку дисперсии погрешностей модели (1). Для этого сначала решим вопрос о нормальности остатков. Проверку нормальности проведем с помощью критерия Фроцини [84, с. 235]. Для этого необходимо вычислить статистику:

$$B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left| \Phi(z_i) - \frac{i - 0.5}{n} \right|,$$

где $z_i = \frac{x_i - \bar{x}}{s}; \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i; s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2; \Phi(z_i)$ — функция распределения $N(0, 1)$. Результаты вычислений представлены в таблице 3.8.

Таблица 3.8.

Расчет статистики B_n

№	Y	X1	X2	Y*	Y-Y*	F(Zi)	B(i)	(Y-Y*)**2
1	1	25,620	79,000	1,930	-0,930	0,040	0,018	0,864
2	2	3,070	52,000	2,711	-0,711	0,090	0,025	0,505
3	2	14,390	95,000	2,672	-0,672	0,102	0,007	0,451
4	2	6,870	59,000	2,608	-0,608	0,125	0,027	0,370
5	2	3,69	25	2,356	-0,356	0,250	0,054	0,127
6	2	5,95	31	2,317	-0,317	0,274	0,035	0,101
7	2	6,46	33	2,316	-0,316	0,274	0,009	0,100
8	2	6,04	30	2,301	-0,301	0,284	0,042	0,091
9	2	7,06	31	2,263	-0,263	0,309	0,061	0,069
10	2	7,03	30	2,253	-0,253	0,316	0,097	0,064
11	2	18,93	74	2,197	-0,197	0,359	0,097	0,039
12	4	1,76	169	4,179	-0,179	0,370	0,130	0,032
13	1	37,85	58	1,078	-0,078	0,440	0,103	0,006
14	2	16,38	48	2,010	-0,010	0,494	0,093	0,000

Продолжение на следующей странице

Таблица 3.8.

Расчет статистики B_n (продолжение)

15	4	2,42	157	4,002	-0,002	0,500	0,130	0,000
16	4	1,89	135	3,764	0,236	0,670	0,004	0,056
17	4	2,48	110	3,435	0,565	0,857	0,140	0,319
18	4	2,04	99	3,325	0,675	0,898	0,137	0,456
19	3	5,34	26	2,287	0,713	0,910	0,106	0,508
20	3	3,42	17	2,273	0,727	0,915	0,067	0,528
21	3	3,39	13	2,227	0,773	0,928	0,037	0,598
22	4	3,20	95	3,220	0,780	0,929	0,006	0,608
23	1	55,20	55	0,192	0,808	0,936	0,043	0,653
						Сумма=	**1,467**	**6,543**
						B(n)	**0,306**	**0,327**
						Скр(0,01)=	**0,341**	

Фиксируя уровень значимости $\alpha = 0.01$ и учитывая, что $C(0.01) = 0.341$ [84, с. 235], получаем:

$$B_n = 0.306 < (0.01) = 0.341.$$

Следовательно, остатки распределены нормально.

И, наконец, учитывая, что оценка дисперсии σ^2 определяется по формуле:

$$(\sigma^2)^* = \frac{1}{(n-(p+1))}(Y - Y^*)^T(Y - Y^*)$$

и равна в нашем случае 0.327 (см. последний столбец табл. 3.8), получаем доказательство следующей теоремы.

Теорема 3.2. *Модель линейной регрессии квартилей научных конференций,построенная по данным, представленным в таблице 3.3, имеет вид:*

$$Y = -0.0491 + 0.0122 + 2.237 + \varepsilon,$$

Таблица 3.9.
Результаты расчета прогнозных значений квартилей

№	Квартиль	Цитируемость	Кол-во
	(прогноз. знач.)		участников
	Y	X1	X2
24	1.0323	38.30	56
25	2.23101	5.51	22
26	3.55425	2.75	121

где ε — с.в., имеющая нормальное распределение с параметрами $m = 0$ и $\sigma = 0.57$.

Далее, на основе данных для трех «новых» конференций, мы с помощью модели (1) получили прогнозные значения их квартилей (таблица 3.9).

Как мы видим из результатов, представленных в таблице 3.9, конференциям под номерами 24 и 25 следует присвоить 1-ый и 2-ой квартиль соответственно. С конференцией под номером 26 картина не столь однозначна, т.к. прогнозное значение Y лежит примерно по середине между цифрами 3 и 4, что говорит о том, что данной конференции следует присвоить 4-ый квартиль с вероятностью 0,55 либо 3-ий квартиль с вероятностью 0,45.

3.4.1. Дискриминантный анализ

Дискриминантный анализ является методом классификации, цель которого заключается в разделении объектов наблюдения на классы в соответствии со значениями результативного признака, зависящего от ряда контролируемых факторов [51]. В нашем случае результативным признаком является квартиль, а контролируемыми факторами – цитируемость материалов конференции и количество ее участников. Наша дальнейшая цель – с помощью дискриминантного анализа классифицировать новые конференции, данные по которым представлены в таблице 3.9, на основе обучающей выборки, представленной таблицей 3.3. Для решения данной задачи по-прежнему используем статистический пакет SPSS.

В первую очередь обращаем внимание на данные, приведенные в таблице 3.10. В данной таблице представлены результаты проверки значимости различий средних значений дискриминантных функций в группах данных,

Таблица 3.10.

Критерий равенства групповых средних

	Лямбда Уилкса	F	ст.св1	ст.св2	Знч.
X1	.190	26.972	3	19	.000
X2	.232	20.913	3	19	.000

Таблица 3.11.

Собственные значения. В анализе использовались первые 2 канонические дискриминантные функции

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция
1	5.378a	67.5	67.5	.918
2	2.586a	32.5	100.0	.849

соответствующих факторам X1 и X2, с помощью критерия Лямбда Уилкса. В нашем случае уровни значимости по каждому фактору не превосходят 0.05, что доказывает наличие дискриминирующих особенностей этих факторов и подтверждает возможность их использования для проведения дискриминантного анализа.

Согласно данным, представленным в таблице 3.11, первая дискриминантная функция учитывает 67,5 % дисперсии результативного признака, а корреляция между данными обучающей выборки и данными, полученными по модели, равна 0.918, что является довольно высоким показателем. Для второй дискриминантной функции эти показатели равны 32,5% и 0.849 соответственно.

Таблица 3.12.

Ненормированные коэффициенты канонических дискриминантных функций

	Функция	
	1	2
X1	.141	.083
X2	-.028	.034
(Константа)	.352	-3.100

Таблица 3.13.

Статистики Лямбда–Уилкса

Функция	Λ_k	χ_k^2	m_k
1	,044	59,470	6
2	,279	24,265	2

По данным таблицы 3.12 получаем следующие выражения для дискриминантных функций:

$$D1(X1, X2) = 0.141X1 - 0.028X2 + 0.352,$$

$$D2(X1, X2) = 0.083X1 + 0.034X2 - 3.100.$$

Теорема 3.3. *Дискриминантные функции (2) и (3) значимы на уровне значимости* $\alpha = 0.01$.

Доказательство. Оценку значимости дискриминантных функций проведем с помощью критерия Лямбда Уилкса [50], согласно которому необходимо вычислить статистики:

$$\chi_k^2\{m_k\} = -\left[n - \frac{p+g}{2} - 1\right] \ln \Lambda_k, \qquad k = 1, 2,$$

где $\Lambda_1 = \frac{1}{1+\lambda_1}\frac{1}{1+\lambda_2}$, $\Lambda_2 = \frac{1}{1+\lambda_2}$, $p = 2$ — количество дискриминантных признаков, $g = 4$ — количество групп, $m_1 = p + g$; $m_2 = p$ –– количество степеней свободы.

Результаты вычислений представлены в таблице 3.13.

Известно [50], что статистики $\chi_k^2\{m_k\}$ имеют χ^2-распределение с m_k степенями свободы. Фиксируя $\alpha = 0.01$ и учитывая, что $(1-\alpha)$ — квантили χ^2-распределения со степенями свободы $m_1 = 6$ и $m_2 = 2$ равны соответственно 16.8 и 9.21, приходим к следующему результату:

- поскольку $59.470 > 16.8$, принимается гипотеза о значимости дискриминантной функции 3.4.1;
- поскольку $24{,}265 > 9{,}21$, принимается гипотеза о значимсоти дискриминантной функции 3.4.1.

Таким образом, теорема доказано. □

Результаты проведенного анализа представлены в таблице 3.14. В итоге «новым» конференциям были выставлены квартили 1, 2 и 4 соответственно. При этом конференциям под номерами 24 и 26 квартили были предсказаны с вероятностями 1 и 0.996. Для конференции под номером 25 картина получилась не столь однозначной. Ей был предсказан 2-ой квартиль с вероятностью 0.673, либо 3-тий квартиль с вероятностью 0.327.

Кроме этого, были пересчитаны квартили конференций из обучающей выборки. В результате конференции с номерами 3, 13, 15 и 16 получили новые значения квартилей. Квартили остальных конференций, составляющих 82.6%, были признаны корректными.

3.4.2. Нейронная сеть

Для решения задачи классификации лучше всего подходит нейронная сеть, называемая многослойным персептроном [43; 71]. Как правило, сеть состоит из одного входного слоя, одного или нескольких скрытых слоев и одного выходного слоя. Каждый слои состоит из нескольких нейронов. Нейрон обрабатывает свои входы и генерирует одно выходное значение, которое передается нейронам в последующем слое. Каждый нейрон во входном слое представляет значения одного предсказателя из вектора $x = (x_1, x_2)$. В нашем случае x_1 и x_2 — цитируемость и количество участников научной конференции.

Для построения сети используем раздел «нейронные сети» пакета SPSS, в котором указываем в качестве зависимой переменной квартиль конференции, а в качестве ковариант – цитируемость и количество участников, и задаем разделение данных на три подмножества: обучающее, контрольное и проверочное в соотношении 20:3:3. Архитектуру сети задаем в ручном режиме, фиксируя наличие одного скрытого слоя с четырьмя нейронами. В качестве функции активации для скрытого и выходного слоев выбираем сигмоид. Затем выбираем интерактивный тип обучения по методу градиентного спуска и задаем время и правило остановки процесса обучения. Параметры сети отражены в таблице 13, а ее конфигурация представлена на рисунке 3.8.

В отчёте, представленном в таблице 3.16, обращаем внимание на строчки «ошибка суммы квадратов» и «относительная ошибка» в разделе «проверочная выборка». Значения ошибок оказались равными 0,016 и 0,045. Эти значения

Таблица 3.14.

Результаты классификации

		1-ая наиве-	группа	2-ая наиве-	группа
		роят.		роят.	
№	Факт. груп-	Предск.	Вероят-	Предск.	Вероят-
	па	группа	ность	группа	ность
1	1	1	1,000	2	0,000
2	1	1	1,000	2	0,000
3	1	2**	0,524	1	0,473
4	2	2	0,983	3	0,011
5	2	2	0,951	3	0,048
6	2	2	0,828	4	0,166
7	2	2	0,785	3	0,215
8	2	2	0,777	3	0,223
9	2	2	0,927	3	0,069
10	2	2	0,791	3	0,209
11	2	2	0,760	3	0,240
12	2	2	0,767	3	0,233
13	3	2**	0,710	3	0,290
14	2	2	0,667	3	0,333
15	3	2**	0,572	3	0,428
16	3	2**	0,524	3	0,476
17	4	4	0,786	2	0,210
18	2	2	0,868	3	0,128
19	4	4	0,981	2	0,019
20	4	4	1,000	2	0,000
21	4	4	0,905	2	0,094
22	4	4	1,000	2	0,000
23	4	4	1,000	2	0,000
24	не сгруп-	1	1,000	2	0,000
	пир.				
25	не сгруп-	2	0,673	3	0,327
	пир.				
26	не сгруп-	4	0,996	2	0,004
	пир.				

Таблица 3.15.

Параметры сети

Входной слой	Ковариаты	1	X1
		2	X2
		Число нейронов	2
		Метод изменения масштаба для ковариат	Стандартизировано
Скрытые слои		Количество скрытых слоев	1
		Количество нейронов в скрытом слое 1[a]	
		Функция активации	Сигмоид
Выходной слой	Зависимые переменные	1	Y
		Количество нейронов	1
		Метод изменения масштаба для количественных зависимых переменных	Нормализовано
		Функция активации	Сигмоид
		Функция ошибки	Сумма квадратов

достаточно малы, что свидетельствует о том, что нейронная сеть хорошо обучена.

Предсказанные значения квартилей как для «новых» конференций, так и для конференций из обучающей выборки содержаться в четвертом столбце таблицы 3.17. Заметим, что для «новых» конференций квартили, полученные с помощью нейронной сети, совпадают с квартилями, полученными с помощью дискриминантного анализа.

В результате проведенного исследования с помощью трех различных методов нами были рассчитаны квартили научных конференций. Результаты

Таблица 3.16.

Сводка для модели

Обучающая выборка	Ошибка суммы квадратов	0.115
	Относительная ошибка	.100
	Использованное правило остановки	Количество последовательных шагов без уменьшения ошибки: 1
	Время обучения	0:00:00.002
Проверочная выборка	Ошибка суммы квадратов	0.016
	Относительная ошибка	.045

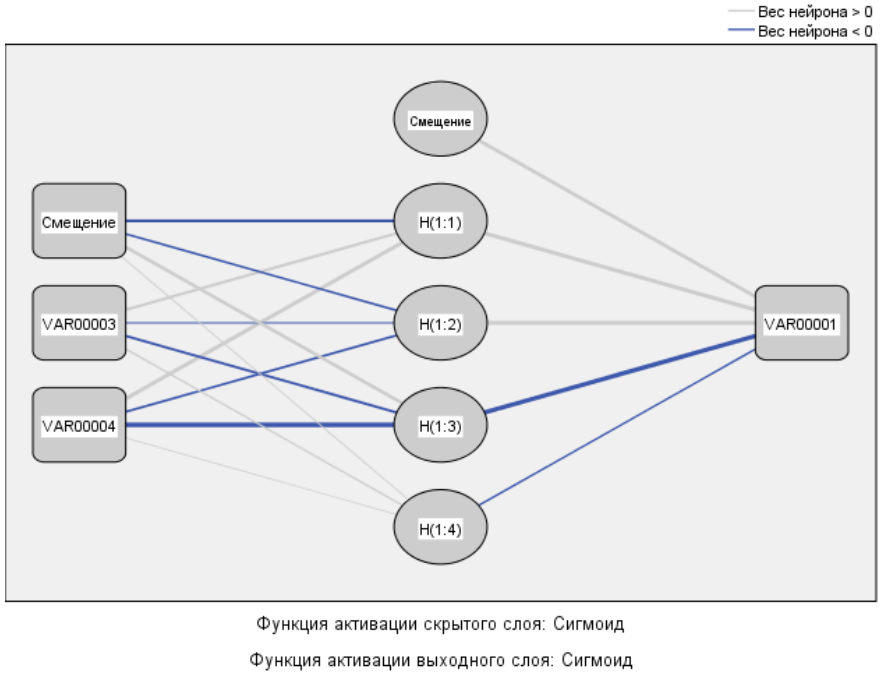

Рис. 3.8. Конфигурация нейронной сети

вычислений отражены в таблице 3.17. Звездочками помечены квартили, не совпадающие с теми, что были выставлены рейтинговыми агентствами и которые мы назвали фактическими.

В последней строке таблицы 3.17 указан процент совпадений фактических квартилей и квартилей, рассчитанных по соответствующему методу. Как мы видим, наилучший показатель у дисперсионного анализа (4 несовпадения). На втором месте с разницей в одну конференцию – нейронная сеть. На третьем месте – метод линейной регрессии, который выявил 6 несовпадений.

Таблица 3.17.

Значения квартилей

№	Фактическое значение квартиля	Значение квартиля по методу регрессии с помощью ДА	Значение квартиля, полученное нейронной сетью	Значение квартиля, предска-занное нейронной сетью
1	1	1	1	1
2	1	1	1	2*
3	1	2*	2**	2*
4	2	2	2	2
5	2	2	2	2
6	2	3*	2	2
7	2	2	2	2
8	2	2	2	2
9	2	2	2	2
10	2	2	2	2*
11	2	2	2	2*
12	2	2	2	2*
13	3	2*	2**	3
14	2	2	2	2
15	3	3	2**	3
16	3	3	2**	3
17	4	4	4	4
18	2	3*	2	2
19	4	3*	4	4
20	4	4	4	4
21	4	3*	4	4
22	4	4	4	4
23	4	4	4	4
24	1	1	1	1
25	2	2	2	2
26	4	4	4	4
Кол-во несовпадений		6	4	5
% совпадений		**76,92**	**84,61**	**80,77**

Заключение

Основные результаты и выводы диссертационной работы, следующие:

В первой главе диссертационной работы представлен обзор литературы по теме исследования, в котором описываются основные рейтинги конференций, методы составления этих рейтингов, описаны попытки ученых в создании новых рейтингов и показателей для оценки качества конференций, кроме того, в первой главе проводится первичный анализ распределения количества конференций по предметным областям и составление рейтинга конференций, разделенного на квартили, аналогично с журнальным рейтингом SJR.

Во второй главе предложен показатель, который ранее не применялся для оценки качества конференций и описаны статистические методы исследования, которые были использованы для анализа качества конференций. Этот показатель был применен к конференциям и составлен рекомендательный список конференций по искусственному интеллекту, для ученых из России, Китая и США. Для каждой страны, выбранной для анализа, этот список свой, так как предложенный показатель позволяет выявить конференции, на которых исследователь из определенной страны сможет получить наибольшее количество цитирований, что делает работу более заметной в научном сообществе. Также представлен статистический анализ показателей цитирования конференций.

В третьей главе на основе не временных данных построены регрессионная и дискриминантная модели для прогнозирования рейтинга конференций, качество этих моделей оценено с помощью статистических методов, также построена простейшая нейронная сеть, которая также прогнозировала рейтинг конференций на основе той же выборки. Все три метода были сравнены друг с другом и выбрана лучшая модель для поставленной цели исследования.

Список литературы

1. *Abalkina A.* Unethical Practices in Research and Publishing: Evidence from Russia. — 2021. — URL: https://scholarlykitchen.sspnet.org/2021/02/04/guest-post-unethical-practices-inresearch-and-publishing-evidence-from-russia/ ; Accessed: 15.12.2021.
2. *Alhoori H., Furuta R.* Can social reference management systems predict a ranking of scholarly venues? // Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Proceedings. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2013. — P. 138–143.
3. *Almendra V. S., Enăchescu D., Enăchescu C.* Ranking computer science conferences using self-organizing maps with dynamic node splitting // Scientometrics. — 2015. — Vol. 102. — P. 267–283.
4. An index-based ranking of conferences in a distinctive manner / M. Farooq [et al.] // The Electronic Library. — 2019. — Vol. 37, no. 1. — P. 67–80.
5. *Bardakcı S., Arslan Ö., Ünver T. K.* How scholars use academic social networking services // Information Development. — 2018. — Vol. 34, no. 4. — P. 334–345.
6. *Beel J., Gipp B.* Google scholar's ranking algorithm: an introductory overview // Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09). Vol. 1. — Rio de Janeiro, Brazil, 2009. — P. 230–241.
7. *Bowyer K. W.* Mentoring advice on "conferences versus journals" for CSE faculty. — 2012. — University of Notre Dame.
8. *Butler L.* ICT assessment: Moving beyond journal outputs // Scientometrics. — 2008. — Vol. 74, no. 1. — P. 39–55.
9. *Cabitza F., Locoro A.* Exploiting the Collective Knowledge of Communities of Experts // KMIS. — 2015. — P. 159–167.
10. *Cagan R.* San Francisco declaration on research assessment // Disease Models & Mechanisms. — 2013. — P. dmm.012955.
11. *Castro D., McLaughlin M., Chivot E.* Who is winning the AI race: China, the EU or the United States : tech. rep. / Center for Data Innovation. — 2019. — No. 19.

12. Comparative analysis of the bibliographic data sources Dimensions and Scopus: An approach at the country and institutional levels / V. P. Guerrero-Bote [et al.] // Frontiers in Research Metrics and Analytics. — 2021. — Vol. 5. — P. 593494.

13. CORE Rankings Portal. — 2022. — URL: https://www.core.edu.au/conference-portal ; Accessed: May 2022.

14. Description of Scimago Journal Rank Indicator. — 2020. — URL: https://www.scimagojr.com/SCImagoJournalRank.pdf.

15. DrWatson: the perfect sidekick for your scientific inquiries / G. Datseris [et al.] // Journal of Open Source Software. — 2020. — Oct. — Vol. 5, no. 54. — P. 2673. — DOI: 10.21105/joss.02673.

16. *Dunaiski M.*, *Visser W.*, *Geldenhuys J.* Evaluating paper and author ranking algorithms using impact and contribution awards // Journal of Informetrics. — 2016. — Vol. 10, no. 2. — P. 392–407.

17. *Eckmann M.*, *Rocha A.*, *Wainer J.* Relationship between high-quality journals and conferences in computer vision // Scientometrics. — 2012. — Vol. 90, no. 2. — P. 617–630.

18. *Editors P. M.* The impact factor game: It is time to find a better way to assess the scientific literature // PLoS Medicine. — 2006. — Vol. 3, no. 6. — e291.

19. *Effendy S.*, *Yap R. H. C.* Investigations on rating computer sciences conferences: An experiment with the Microsoft Academic Graph dataset // Proceedings of the 25th International Conference Companion on World Wide Web. — 2016. — P. 425–430.

20. *El Mohadab M.*, *Bouikhalene B.*, *Safi S.* Predicting rank for scientific research papers using supervised learning // Applied Computing and Informatics. — 2019. — Vol. 15, no. 2. — P. 182–190. — DOI: 10.1016/j.aci.2018.02.004.

21. *Elsevier*. What is the Complete List of Scopus Subject Areas and All Science Journal Classification Codes (ASJC)? — URL: https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/.

22. *Ermolayeva A. M.* A model of cumulative advantage for conference dynamics // Discrete and Continuous Models and Applied Computational Science. — 2026. — Vol. 34, no. 1. — P. 145–149. — DOI: 10.22363/2658-4670-2026-34-1-145-149.

23. *Ermolayeva A. M.* Statistical methods for estimating quartiles of scientific conferences // Discrete and Continuous Models and Applied Computational Science. — 2024. — June. — Vol. 32, no. 1. — P. 5–17. — DOI: 10.22363/2658-4670-2024-32-1-5-17.

24. *Ernst M.* Choosing a venue: Conference or journal. — 2006. — URL: https://homes.cs.washington.edu/~mernst/advice/conferences-vs-journals.html.

25. Federal AI spending to top $6 billion. — 2021. — URL: https://www.nationaldefensemagazine.org/articles/2021/2/10/federal-ai-spending-to-top-6-billion.

26. First predatory journals, now conferences: The need to establish lists of fake conferences / C. Sonne [et al.] // Science of the Total Environment. — 2020. — Vol. 715. — P. 136990.

27. *Franceschet M.* The role of conference publications in CS // Communications of the ACM. — 2010. — Vol. 53, no. 12. — P. 129–132.

28. Google Scholar Conference Ranking. — URL: https://scholar.google.com/citations?view_op=top_venues%5C&hl=en.

29. *Haq I. U.*, *Tanveer M.* Status of research productivity and higher education in the members of Organization of Islamic Cooperation (OIC) // Library Philosophy and Practice e-journal. — 2020. — Vol. 3845.

30. *Hughes K.* The half-life of citations. — 07/2020. — URL: https://brushingupscience.com/2020/07/01/the-half-life-of-citations/.

31. Index of revisions to the 'guidance on submissions'. — 2019. — URL: https://www.ref.ac.uk/media/1447/ref-201901-guidance-on-submissions.pdf; May 2019.

32. *Jahja I.*, *Effendy S.*, *Yap R. H. C.* Experiments on rating conferences with CORE and DBLP // D-Lib Magazine. — 2014. — Vol. 20, no. 11/12.

33. *Julpisit A.*, *Esichaikul V.* A collaborative system to improve knowledge sharing in scientific research projects // Information Development. — 2019. — Vol. 35, no. 4. — P. 624–638.

34. *Kerl A.*, *Miersch E.*, *Walter A.* Evaluation of academic finance conferences // Journal of Banking & Finance. — 2018. — Vol. 89. — P. 26–38.

35. *Kochetkov D.*, *Birukou A.*, *Ermolayeva A.* Russia on the Global Artificial Intelligence Scene // Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Selected Papers. — Cham : Springer International Publishing, 2021. — P. 369–378.

36. *Kochetkov D.*, *Birukou A.*, *Ermolayeva A.* Russia on the Global Artificial Intelligence Scene // Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Selected Papers. — Cham : Springer International Publishing, 2021. — P. 369–378.

37. *Kochetkov D.*, *Birukou A.*, *Ermolayeva A.* The importance of conference proceedings in research evaluation: a methodology for assessing conference impact // Distributed Computer and Communication Networks: 24th International Conference, DCCN 2021, Revised Selected Papers. — Cham : Springer International Publishing, 2022. — P. 359–370.

38. *Kochetkov D.*, *Birukou A.*, *Ermolayeva A. M.* The Importance of Conference Proceedings in Research Evaluation: A Methodology for Assessing Conference Impact // Distributed Computer and Communication Networks. — Springer International Publishing, 2022. — DOI: 10.1007/978-3-030-97110-6_28.

39. *Koya K.*, *Chowdhury G.* Metric-based vs peer-reviewed evaluation of a research output: Lesson learnt from UK's national research assessment exercise // PLoS ONE. — 2017. — Vol. 12, no. 7. — e0179722.

40. *Kwanya T.* Publishing and perishing? Publishing patterns of information science academics in Kenya // Information Development. — 2020. — Vol. 36, no. 1. — P. 5–15.

41. *Lotka A. J.* Contribution to the Theory of Periodic Reaction // The Journal of Physical Chemistry A. — 1910. — Vol. 14, no. 3. — P. 271–274. — DOI: 10.1021/j150111a004.

42. *Lotka A. J.* The frequency distribution of scientific productivity // Journal of the Washington Academy of Sciences. — 1926. — Vol. 16, no. 12. — P. 317–324.

43. *Madhan M.*, *Gunasekaran S.*, *Arunachalam S.* Evaluation of research in India–Are we doing it right // Indian Journal of Medical Ethics. — 2018. — Vol. 3, no. 3. — P. 221–229.

44. *Makhoba X.*, *Pouris A.* Scientometric assessment of selected R&D priority areas in South Africa: A comparison with other BRICS countries // African Journal of Science Technology Innovation and Development. — 2016. — Vol. 8, no. 2. — P. 187–196.

45. *Meho L. I.* Using Scopus's CiteScore for assessing the quality of computer science conferences // Journal of Informetrics. — 2019. — Vol. 13, no. 1. — P. 419–433.

46. *Meho L. I.* Using Scopus's CiteScore for assessing the quality of computer science conferences // Journal of Informetrics. — 2019. — Vol. 13, no. 1. — P. 419–433.

47. *Meho L. I.*, *Yang K.* Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar // Journal of the American Society for Information Science and Technology. — 2007. — Vol. 58, no. 13. — P. 2105–2125.

48. *Merton R. K.* The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property // Isis. — 1988. — Dec. — Vol. 79, no. 4. — P. 606–623. — DOI: 10.1086/354848.

49. *Merton R. K.* The Matthew Effect in Science: The reward and communication systems of science are considered // Science. — 1968. — Jan. — Vol. 159, no. 3810. — P. 56–63. — DOI: 10.1126/science.159.3810.56.

50. Microsoft Academic. — 2022. — URL: https://www.microsoft.com/en-us/research/project/academic/articles/www-conference-analytics/ ; Accessed: May 2022.

51. *Nature E. in*. China's research-evaluation revamp should not mean fewer international collaborations // Nature. — 2020. — Vol. 579. — P. 8.

52. *Patterson D.*, *Snyder L.*, *Ullman J.* Evaluating computer scientists and engineers for promotion and tenure // Computing Research News. — 1999.

53. Predicting global ranking of universities across the world using machine learning regression technique / P. K. Udupi [et al.] // SHS Web of Conferences. — 2023. — Vol. 156. — P. 04001. — DOI: 10.1051/shsconf/202315604001.

54. Prediction of upcoming conferences ranking in Bangladesh based on analytic network process and machine learning / G. R. Chowdhury [et al.] // 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET). — Chittagong, Bangladesh : IEEE, 2018. — P. 463–467. — DOI: 10.1109/ICISET.2018.8745611.

55. *Price D. J. d. S.* Little Science, Big Science. — New York : Columbia University Press, 1963.

56. *Purnell P. J.* Conference proceedings publications in bibliographic databases: a case study of countries in Southeast Asia // Scientometrics. — 2021. — Vol. 126, no. 1. — P. 355–387.

57. *Reinartz S. J., Urban D.* Finance conference quality and publication success: A conference ranking // Journal of Empirical Finance. — 2017. — Vol. 42. — P. 155–174.

58. Relative status of journal and conference publications in computer science / J. Freyne [et al.] // Communications of the ACM. — 2010. — Vol. 53, no. 11. — P. 124–132.

59. Reverse-engineering conference rankings: what does it take to make a reputable conference? / P. Küngas [et al.] // Scientometrics. — 2013. — Vol. 96. — P. 651–665.

60. *Saier T., Färber M.* A Large-Scale Analysis of Cross-lingual Citations in English Papers // Digital Libraries at Times of Massive Societal Transition: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, Proceedings. — Cham : Springer International Publishing, 2020. — P. 122–138.

61. *Seglen P. O.* Why the impact factor of journals should not be used for evaluating research // BMJ. — 1997. — Vol. 314, no. 7079. — P. 497.

62. *Singh A. P., Shubhankar K., Pudi V.* An efficient algorithm for ranking research papers based on citation network // 2011 3rd Conference on Data Mining and Optimization (DMO). — IEEE, 2011. — P. 88–95.

63. Statistical model and method for analyzing AI conference rankings: China vs USA / A. M. Ermolayeva [et al.] // Heliyon. — 2023. — Nov. — Vol. 9, no. 11. — e21592. — DOI: 10.1016/j.heliyon.2023.e21592.

64. Taking the con out of conferences. — 2017. — URL: https://www.crossref.org/blog/taking-the-con-out-of-conferences/ ; February 2017.

65. *Tattershall E., Nenadic G., Stevens R. D.* Modelling trend life cycles in scientific research using the Logistic and Gompertz equations // Scientometrics. — 2021. — Oct. — Vol. 126, no. 11. — P. 9113–9132. — DOI: 10.1007/s11192-021-04137-0.

66. *Teli S., Dutta B.* Revisiting De Solla Price: growth dynamics studies of various subjects over last one hundred years // Annals of Library and Information Studies. — 2020. — Mar. — Vol. 67, no. 1. — P. 17–35.

67. The impact of conference ranking systems in computer science: A comparative regression analysis / X. Li [et al.] // Scientometrics. — 2018. — Vol. 116. — P. 879–907.

68. The impact of conference ranking systems in computer science: A comparative regression analysis / X. Li [et al.] // Scientometrics. — 2018. — Vol. 116. — P. 879–907.

69. Towards a new crown indicator: An empirical analysis / L. Waltman [et al.] // Scientometrics. — 2011. — Vol. 87, no. 3. — P. 467–481.

70. Trends and characteristics of global medical informatics conferences from 2007 to 2017: a bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences / Y. Jia [et al.] // Computer Methods and Programs in Biomedicine. — 2018. — Vol. 166. — P. 19–32.

71. URAP Methodology. — URL: https://www.urapcenter.org/Methodology; Accessed: 03.07.2020.

72. *Verhulst P. F.* Notice sur la loi que la population suit dans son accroissement. T. 10. — 1838. — 113–117.

73. *Vrettas G., Sanderson M.* Conferences vs. journals in computer science // Journal of the Association for Information Science and Technology. — 2014. — Vol. 23349. — https://doi.org/10.1002/asi.

74. *Waltman L.* A review of the literature on citation impact indicators // Journal of Informetrics. — 2016. — Vol. 10, no. 2. — P. 365–391.

75. *Waltman L., Traag V. A.* Use of the journal impact factor for assessing individual articles: Statistically flawed or not? — 2017. — arXiv preprint arXiv:1703.02334.

76. What makes top 20 JIF journals ”top”?: Exploring characteristics of journals indexed in the Journal Citation Reports / G. Doğan [et al.] // 26th International Conference on Science, Technology and Innovation Indicators (STI 2022). — Granada, Spain, 09/2022. — DOI: 10.5281/zenodo.6906891.

77. *Xia M., He X., Zhou Y.* Symbiosis Evolution of Science Communication Ecosystem Based on Social Media: A Lotka–Volterra Model-Based Simulation // Complexity / ed. by J. Jiang. — 2021. — Jan. — Vol. 2021, no. 1. — DOI: 10.1155/2021/6655469.

78. *Вольтерра В.* Математическая теория борьбы за существование : пер. с фр. — М. : Наука, 1976. — 288 с.

79. *Гмурман В.* Теория вероятностей и математическая статистика. — 12-е изд. — Litres, 2022.

80. *Горбунов-Посадов М. М.* Расширяемые программы. — М. : ООО «Полиптих», 1999. — 336 с.

81. *Горбунов-Посадов М. М., Корягин Д. А., Мартынюк В. В.* Системное обеспечение пакетов прикладных программ / под ред. А. А. Самарский. — М. : Наука, 1990. — 205 с. — (Библиотечка программиста).

82. Прогр. обесп. А. М. Ермолаева, Механизмы кумулятивного преимущества в наукометрии 2026. — DOI: 10.5281/zenodo.20025331.

83. Извлечение и статистический анализ наукометрических показателей конференций в области распределенных вычислений на основе международной реферативной научной базы данных Scopus : Свидетельство о государственной регистрации программы для ЭВМ 2022665930 / А. М. Ермолаева, А. А. Давтян ; Ф. государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов имени Патриса Лумумбы» (РУДН). — Заявл. 23.08.2022 ; опубл. 23.08.2022.

84. *Кобзарь А. И.* Прикладная математическая статистика: для инженеров и научных работников. — 2-е изд. — М. : Физматлит, 2012. — 813 с. — (Современные методы в математике). — Библиогр.: с. 737-759.

85. *Кун Т.* Структура научных революций / под ред. В. Ю. Кузнецов ; пер. И. З. Налетов. — М. : АСТ, 2003. — 605 с. — (Philosophy).

86. Сбор, обработка и конвертация данных для анализа наукометрических показателей конференций на основе международной реферативной научной базы данных Scopus : Свидетельство о государственной регистрации программы для ЭВМ 2022665773 / А. М. Ермолаева ; Ф.

государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов имени Патриса Лумумбы» (РУДН). — Заявл. 22.08.2022 ; опубл. 22.08.2022.

87. *Чернова Н. И.* Математическая статистика. — 2007.

Список иллюстраций

Список таблиц

Приложение А. Сбор, обработка и конвертация данных для анализа наукометрических показателей конференций на основе международной реферативной научной базы данных Scopus

Данная программа зарегистрирована как программа ЭВМ [86].

```
import pandas as pd
import requests

API = ""
# название конференции
CONFNAME = ""
# год от (включительно)
PUBYEAR_FROM = 2011
# год до (включительно)
PUBYEAR_TO = 2020
# страна
AFFILCOUNTRY = ""


def _search_scopus(key, query, view, index):
    """
    Функция делающая запрос к API
    """


par = {
    "apikey": key,
    "query": query,
    "start": index,
    "httpAccept": "application/json",
    "view": view,
}
r = requests.get("http://api.elsevier.com/content/search/scopus", params=par)

js = r.json()
total_count = int(js["search-results"]["opensearch:totalResults"])
entries = js["search-results"]["entry"]
result_df = pd.DataFrame([_parse_article(entry) for entry in entries])
```

```
    return (result_df, total_count)


def _parse_article(entry):
    """Функция обрабатывающая возвращаемый json"""


try:
    coverdate = entry["prism:coverDate"]
except:
    coverdate = None
    try:
        citationcount = int(entry["citedby-count"])
    except:
        citationcount = None
    return pd.Series({"cover_date": coverdate, "citation_count": citationcount})


def search(query, view="STANDARD"):
    """Функция которая выполняет все действия по скачианию и обработке"""
    i = 1
    result_df, total_count = _search_scopus(API, query, view=view, index=i)

    while True:
        index = 25 * i
        print(index, "из", total_count)
        result_df = result_df.append(
            _search_scopus(API, query, view=view, index=index)[0], ignore_index=True
        )
    if result_df.shape[0] >= total_count:
        return result_df[:total_count]
    i += 1
    return result_df


def get_file(CONFNAME, PUBYEAR_FROM, PUBYEAR_TO, AFFILCOUNTRY, file_name="exit.csv"):
    """Скачивает данные и сохраняет в файл"""
    print("Начало загрузки данных")
    search_df = search(
        f"CONFNAME({CONFNAME}) and ((PUBYEAR < {PUBYEAR_TO} or PUBYEAR =
        ↪  {PUBYEAR_TO}) and (PUBYEAR > {PUBYEAR_FROM} or PUBYEAR = {PUBYEAR_FROM}))
        ↪  and AFFILCOUNTRY ({AFFILCOUNTRY})"
    )
    print("Начало обработки данных")
```

```
61      search_df.cover_date = list(map(lambda x: x[:4], search_df.cover_date))
62      count = pd.DataFrame(
63          search_df.value_counts(subset=["cover_date"]), columns=["count"]
64      )
65      citation_count = search_df.groupby("cover_date").sum("citation_count")
66      final = citation_count.merge(count, on="cover_date")
67      final.to_csv(file_name)
68      print(f"Данные сохранены в файл {file_name}")


69  get_file(CONFNAME, PUBYEAR_FROM, PUBYEAR_TO, AFFILCOUNTRY, file_name="exit.csv")
```

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

RU 2022665773

(12) **ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ**

Номер регистрации (свидетельства): **2022665773**	Автор: **Ермолаева Анна Михайловна (RU)**
Дата регистрации: **22.08.2022**	Правообладатель: **Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов» (РУДН) (RU)**
Номер и дата поступления заявки: **2022665266 11.08.2022**	
Дата публикации и номер бюллетеня: **22.08.2022 Бюл. № 9**	
Контактные реквизиты: **нет**	

Название программы для ЭВМ:
Сбор, обработка и конвертация данных для анализа наукометрических показателей конференций на основе международной реферативной научной базы данных Scopus

Реферат:
Программа предназначена для сбора, обработки и конвертации данных для анализа наукометрических показателей конференций с помощью API базы данных Scopus. Программа включает функции, делающие запрос к API и обрабатывающие возвращаемые данные в формате json, выполняющие действия по скачиванию, обработке и сохранению данных в файл формата csv для последующего анализа.

Язык программирования: Python

Объем программы для ЭВМ: 4 КБ

Рис. А.1. Регистрационное свидетельство № 2022665773

Приложение В. Извлечение и статистический анализ наукометрических показателей конференций в области распределенных вычислений на основе международной реферативной научной базы данных Scopus

Данная программа зарегистрирована как программа ЭВМ [83].

```
import pandas as pd
from pybliometrics.scopus import ScopusSearch, CitationOverview
import matplotlib.pyplot as plt

plt.style.use("hello.mplstyle")
import numpy as np

df1 = pd.read_csv("CORE.csv")
df2 = pd.read_excel("CCF.xlsx")
m1 = pd.merge(df1, df2, on="conf")  # общее в CORE и CCF
# CCF и Microsoft не имеют общих конференций, поэтому не берем
res = pd.concat([df1, df2, m1]).drop_duplicates(subset=["conf"], keep="last")
# объединяем CORE, CCF, Microsoft и общее; убираем дубли и оставляем last, потому что там общее.
res = (
    res.drop(
        ["2189", "ANT", "CORE2020", "Yes", "4606", "Unnamed: 7", "Unnamed: 8"], axis=1
    )
    .replace(np.nan, "-")
    .reset_index(drop="True")
)
# убираем ненужные столбцы /////////// меняем NaN на "-" // восстанавливаем правильный порядок res
# списки для MultiIndex
years = [i for i in range(2011, 2021)]
count = [("Works") for i in range(2011, 2021)]
cit_count = [("Citations") for i in range(2011, 2021)]
# пустой датафрейм, чтобы было, к чему сшивать
begin = pd.DataFrame([], columns=years)
resultw = pd.DataFrame([])
for eid in res["conf"].tolist():
    # убираем скобки, из-за которых ломается поиск в Скопусе
```

```
    disallowed_characters = "()"
    for character in disallowed_characters:
        eid = eid.replace(character, "")
        a = "CONF ( {} ) AND PUBYEAR > 2010 AND PUBYEAR < 2021 AND AFFILCOUNTRY ( russian AND federation )".format(
            eid
        )
        s = ScopusSearch(a)
        df = pd.DataFrame(pd.DataFrame(s.results))

# пропуск, если результатов нет
if df.empty != True:
    df["coverDate"] = pd.DatetimeIndex(df["coverDate"]).year
    # меняем данные на формат .year
    temp1 = pd.DataFrame(
        df.value_counts(subset=["coverDate"]), columns=["count"]
    )  # считаем публикации
    temp1 = temp1.rename_axis(
        None
    ).transpose()  # переворачиваем, потому что костыли надо вертеть
    temp1.columns = temp1.columns.get_level_values(
        0
    )  # лечим то, что программа считает, что это мультииндекс
    temp2 = begin.merge(temp1, how="outer")  # объединяем результаты
    temp2.index = ["{}".format(eid)]  # подписываем конференцию
    resultw = pd.concat([resultw, temp2]).replace(
        np.nan, "-"
    )  # меняем NaN на - для ГОСТа
    resultw2 = pd.DataFrame([])
    for eid in res["conf"].tolist():
        disallowed_characters = "()"
    for character in disallowed_characters:
        eid = eid.replace(character, "")
        a2 = "CONF ( {} ) AND PUBYEAR > 2010 AND PUBYEAR < 2021".format(eid)
        s2 = ScopusSearch(a2)
        df = pd.DataFrame(pd.DataFrame(s2.results))
    if df.empty != True:
        df["coverDate"] = pd.DatetimeIndex(df["coverDate"]).year
        temp1 = pd.DataFrame(
            df.value_counts(subset=["coverDate"]), columns=["count"]
        )  # считаем публикации
        temp1 = temp1.rename_axis(None).transpose()
        temp1.columns = temp1.columns.get_level_values(0)
        temp2 = begin.merge(temp1, how="outer")
```

```
        temp2.index = ["{}".format(eid)]
    resultw2 = pd.concat([resultw2, temp2]).replace(np.nan, "-")

resultw.columns = pd.MultiIndex.from_arrays(
    [count, resultw.columns]
)  # две строчки-названия
esultw2.columns = pd.MultiIndex.from_arrays(
    [count, resultw2.columns]
)  # ==== цитирования!
esultc = pd.DataFrame([])
for eid in res["conf"].tolist():
    disallowed_characters = "()"
for character in disallowed_characters:
    eid = eid.replace(character, "")
    a = "CONF ( {} ) AND PUBYEAR > 2010 AND PUBYEAR < 2021 AND AFFILCOUNTRY ( russian
     ↪ AND federation )".format(
        eid
    )
    s = ScopusSearch(a, subscriber=False)
    df = pd.DataFrame(pd.DataFrame(s.results))
if df.empty != True:
    df["coverDate"] = pd.DatetimeIndex(df["coverDate"]).year
    temp1 = (
        df.groupby("coverDate").sum("citedby_count").drop("openaccess", axis=1)
    )  # убираем что-то, что непонятно откуда взялось и мешает
    temp1 = temp1.rename_axis(None).transpose()  # переворачиваем
    temp1.columns = temp1.columns.get_level_values(0)  # лечим MultiIndex
    temp2 = begin.merge(temp1, how="outer")
    temp2.index = ["{}".format(eid)]
    resultc = pd.concat([resultc, temp2]).replace(np.nan, "-")
    resultc2 = pd.DataFrame([])
    for eid in res["conf"].tolist():
        disallowed_characters = "()"
    for character in disallowed_characters:
        eid = eid.replace(character, "")
        a2 = "CONF ( {} ) AND PUBYEAR > 2010 AND PUBYEAR < 2021".format(eid)
        s2 = ScopusSearch(a2)

df = pd.DataFrame(pd.DataFrame(s2.results))
if df.empty != True:
    df["coverDate"] = pd.DatetimeIndex(df["coverDate"]).year
temp1 = df.groupby("coverDate").sum("citedby_count").drop("openaccess", axis=1)
temp1 = temp1.rename_axis(None).transpose()
temp1.columns = temp1.columns.get_level_values(0)
```

```
112 temp2 = begin.merge(temp1, how="outer")
113 temp2.index = ["{}".format(eid)]
114 resultc2 = pd.concat([resultc2, temp2]).replace(np.nan, "-")
115 resultc.columns = pd.MultiIndex.from_arrays([cit_count, resultc.columns])
116 resultc2.columns = pd.MultiIndex.from_arrays(
117     [cit_count, resultc2.columns]
118 )  # ====== объединяем!!
119 result1 = resultw.join(resultc).replace(np.nan, "-")
120 result1.index.name = None
121 result2 = resultw2.join(resultc2).replace(np.nan, "-")
122 result2.index.name = None
123 ranks = [("Rankings") for i in range(3)]  # MultiIndex
124 addr = res.set_index("conf").rename_axis(None)
125 addr.columns = pd.MultiIndex.from_arrays([ranks, addr.columns])
126 russia = pd.merge(addr, result1, left_index=True, right_index=True)
127 # Russia
128 world = pd.merge(addr, result2, left_index=True, right_index=True)
129 # World
130 # из ГОСТ в удобный для подсчётов
131 world = world.replace("-", np.nan)
132 russia = russia.replace("-", np.nan)
133 world_in_ru = world[world.index.isin(russia.index)]
134 # WORLD
135 TO_w = pd.DataFrame(
136     [world_in_ru["Works"].sum(axis=1)], index=["Total output"]
137 ).transpose()  # Total output 51

138 TCS_w = pd.DataFrame([world_in_ru["Citations"].sum(axis=1)],
    ↪  index=["TCS"]).transpose()
139 CPP_w = pd.DataFrame([TCS_w["TCS"] / TO_w["Total output"]],
    ↪  index=["CPP"]).transpose()
140 W_res = pd.merge(
141     pd.merge(TO_w, TCS_w, left_index=True, right_index=True),
142     CPP_w,
143     left_index=True,
144     right_index=True,
145 )
146 # Считаем MNCS для России
147 MNCS_ci = russia["Citations"] / russia["Works"]
148 MNCS_ei = world_in_ru["Citations"] / world_in_ru["Works"]
149 MNCS_sum = (MNCS_ci / MNCS_ei).sum(axis=1)
150 MNCS_list = MNCS_sum * 0.1
151 # RUSSIA
152 TO_r = pd.DataFrame(
```

```
    [russia["Works"].replace("-", np.nan).sum(axis=1)], index=["Total output (Russia)"]
).transpose()
TCS_r = pd.DataFrame(
    [russia["Citations"].replace("-", np.nan).sum(axis=1)], index=["TCS (Russia)"]
).transpose()
CPP_r = pd.DataFrame(
    [TCS_r["TCS (Russia)"] / TO_r["Total output (Russia)"]], index=["CPP (Russia)"]
).transpose()
MNCS_r = pd.DataFrame([MNCS_list], index=["MNCS (Russia)"]).transpose()
R_res = pd.merge(
    pd.merge(TO_r, TCS_r, left_index=True, right_index=True),
    CPP_r,
    left_index=True,
    right_index=True,
)
R_res = pd.merge(R_res, MNCS_r, left_index=True, right_index=True)
Result = pd.merge(
    pd.merge(W_res, R_res, left_index=True, right_index=True),
    world_in_ru["Rankings"],
    left_index=True,
    right_index=True,
).rename(columns={"Unranked": "CORE"})
from scipy import stats


# проверяем, подчиняется ли CPP нормальному распределению
def kstest(x):
    if stats.kstest(x, "norm")[1] < 0.05:
        print(
            "Значение p = "
            + str(stats.kstest(x, "norm")[1])
            + " <  0.5\nКритерий Колмогорова-Смирнова не выполняется, распределение не подчиняется нормальному закону"
        )
    else:
        print(
            "Значение p = "
            + str(stats.kstest(x, "norm")[1])
            + " > 0.5\nКритерий Колмогорова-Смирнова не выполняется, распределение не подчиняется нормальному закону"
        )
        print(
            "\033[1m"
```

```
            + "Проверка критерия Колмогорова-Смирнова для мира (конференции, где есть
             ↪  Россия):"
            + "\033[0m"
        )
    kstest(Result["CPP"])


print(
    "\033[1m"
    + "Проверка критерия Колмогорова-Смирнова для мира (все конференции):"
    + "\033[0m"
)
kstest(CPP_General)
print("\033[1m" + "Проверка критерия Колмогорова-Смирнова для России" + "\033[0m")
kstest(Result["CPP (Russia)"])


def spearman(x, y):
    if stats.spearmanr(x, y)[1] <= 0.01:
        print(
            "Значение p = "
            + str(stats.spearmanr(x, y)[1])
            + " < 0.01\nГипотеза H0 не выполняется, корреляция есть\nУровень
             ↪  корреляции = "
            + str(stats.spearmanr(x, y)[0])
        )
    elif stats.spearmanr(x, y)[1] >= 0.1:
        print(
            "Значение p = "
            + str(stats.spearmanr(x, y)[1])
            + " > 0.1\nГипотеза H0 выполняется, корреляции нет"
        )
    else:
        print(
            "Значение p = "
            + str(stats.spearmanr(x, y)[1])
            + "\nНужна дополнительная проверка"
        )
    spearman(Result["CPP"], Result["CPP (Russia)"])
    stats.spearmanr(Result["CPP"], Result["CPP (Russia)"])
    good_confs = Result.loc[Result["MNCS (Russia)"] >= 1].replace(np.nan, "-")
    # Корреляция CPP мира к России
    fig01 = plt.figure(num=0)
    ax01 = fig01.add_subplot()
```

```
230     ax01.set_title("Анализ корреляции")
231     ax01.set_xlabel(r"Мировой СРР")
232     ax01.set_ylabel(r"Российский СРР")
233     ax01.plot(Result["CPP"], Result["CPP (Russia)"], "o", color="maroon")
234     fig02 = plt.figure(num=0)
235     ax02 = fig02.add_subplot()
236     ax02.set_title("Распределение СРР России")

237     ax02.hist(Result["CPP (Russia)"], bins=20, width=2.5, color="maroon")
238     fig03 = plt.figure(num=0)
239     ax03 = fig03.add_subplot()
240     ax03.set_title("Распределение СРР Мира")
241     ax03.hist(Result["CPP"], bins=20, width=2.15, color="maroon")
```

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

RU 2022665930

(12) **ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ**

Номер регистрации (свидетельства):
2022665930

Дата регистрации: **23.08.2022**

Номер и дата поступления заявки:
2022664928 11.08.2022

Дата публикации и номер бюллетеня:
23.08.2022 Бюл. № 9

Контактные реквизиты:
нет

Авторы:
Ермолаева Анна Михайловна (RU),
Давтян Артур Арменович (RU)

Правообладатель:
Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов» (РУДН) (RU)

Название программы для ЭВМ:
Извлечение и статистический анализ наукометрических показателей конференций в области распределенных вычислений на основе международной реферативной научной базы данных Scopus

Реферат:
Программа предназначена для извлечения наукометрических показателей конференций из базы данных Scopus и проведения статистического анализа полученных данных. Программа включает универсальный алгоритм получения наукометрических данных, вычисление наукометрических показателей, анализ корреляции по критерию Спирмена и проверку на подчинение нормальному закону по критерию Колмогорова-Смирнова, а также вывод графиков для наглядной демонстрации работы алгоритма.

Язык программирования: Python

Объем программы для ЭВМ: 246 КБ

Рис. В.1. Регистрационное свидетельство № 2022665930