ОТ ПРОСОПОГРАФИИ УНИВЕРСИТЕТСКОЙ ПРОФЕССУРЫ ДО ЦИФРОВОГО СЛЕДА ФИЛОСОФСКОГО ПАРОХОДА: «СРЕДНИЕ ДАННЫЕ» И ФОРМАЛЬНЫЕ ПОДХОДЫ В ИСТОРИИ НАУКИ[1]

Алексей Куприянов[2]

Abstract

A concept of 'medium-sized' data is introduced to complement 'Big' data used in many projects in quantitative history. Like Big data, medium-sized data are disaggregated, machine-readable, represent 'natural' populations rather than samples, and are 'shallow' (the number of variables is usually small). Unlike 'Big' data they are not accumulated routinely in a machine-readable format and require a lot of manual work, which puts certain limits to the size of datasets. General principles of dataset formation for the analysis of populations of persons and organizations are discussed. Two datasets (one, for the 19th c. Russian University professors and instructors, and another, for Russian philosophical periodicals of the first half of the 20th c.) are used to demonstrate techniques of stepwise data aggregation (which helps to partly overcome the original shallowness of the medium-sized data) and visualization of historical processes. The role of novel descriptive and representative techniques in comparative studies is discussed.

Keywords: digital humanities, big data, R scripting, quantitative history, descriptive theory, historical demography, history of universities, history of philosophy, history of Russia.

Эта статья написана не для специалистов в области *Digital Humanities*, с легкостью входящих в обсуждение тонких вопросов архитектуры баз данных или методов компьютерного анализа текстов. Я обращаюсь прежде всего к коллегам-историкам, которые располагают значительными массивами данных и не решаются перейти к их компьютерной обработке.

Наиболее очевидный выход в такой ситуации – обратиться к программистам или к специалистам по «анализу

[1] Исследования по кадровой динамике университетов поддержаны грантом РФФИ (грант No. 15-06-04531) и программой НИУ ВШЭ по софинансированию проектов РФФИ и РГНФ (No. 15-09-0289).

[2] Алексей Куприянов – кандидат биол.наук, доцент Департамента социологии Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург, Россия).

данных». Здесь перед нами неизбежно возникает проблема коммуникации. Традиционному историку и «аналитику данных» придется выработать общий язык, позволяющий переходить от характерных для историков интуитивных озарений и смутных желаний к формальным моделям, и переводить технический жаргон программистов и аналитиков в обычную человеческую речь. К несчастью для историков, учиться придется в основном им, поскольку со стороны аналитиков возможности для ведения переговоров ограничены особенностями применения статистических методов и структурными свойствами реляционных баз данных. Историкам предстоит вникнуть в логику этих ограничений и научиться выживать в ней. В худшем случае они смогут формулировать задачи на понятном аналитикам языке и интерпретировать результаты их работы, а в лучшем – обходиться без помощи аналитиков в решении относительно несложных задач.

В данной статье я расскажу о налаживании мостов между историей науки и высшего образования с одной стороны и *Data science* с другой, осуществляемом нашей эфемерной исследовательской группой[3]. Содержательно наши проекты связаны с изучением

[3] Организационно эта неформальная группа была связана с кафедрой гуманитарных наук НИУ ВШЭ (СПб) и включала М. Демина, меня и нескольких студентов бакалавриата, работавших под нашим руководством. Эфемерность обусловлена, с одной стороны, неустойчивым положением преподавателей в современном корпоративном университете, а с другой – тем, что большую ее часть всегда составляли студенты. К сожалению, на настоящий момент исследовательский семинар, бывший организационным центром группы, ликвидирован решением рководителя программы по социологии, и перспектив для его восстановления в ближайшее время нет. Предварительные результаты работы группы – см. А. Куприянов: Реструктуризация и общая депрессия: предварительные замечания о природе библиометрических кризисов в истории Советской науки // *Социология науки и технологий.* 4/4 (2013), 80–98; Е. Иванова: Что может дать единая база данных по профессуре дореволюционной России исследователям академической мобильности? // *Социология в действии — 2014. Избранные материалы VI социологической межвузовской конференции студентов и аспирантов* / отв. ред. М. Демин. СПб.: НИУ ВШЭ, 2014, 131–141; Е. Иванова: Попытка построения каузальной модели кадровой динамики профессорско-преподавательского состава университетов дореволюционной России // *Социология в действии – 2015. Избранные материалы VII социологической межвузовской конференции студентов и аспирантов* / отв. ред. М. Демин. СПб.: НИУ ВШЭ, 2015, 118–136; М. Фотиади: Философия революционной эпохи: наукометрический анализ русской философской периодики первой половины XX века // *Социология в действии – 2014. Избранные материалы VI социологической межвузовской конференции студентов и аспирантов* / отв. ред. М.Р. Демин. СПб.: НИУ ВШЭ, 2014, 120–131; A. Kouprianov: Beyond the Humanities: A Comparison of two Bibliometric Crises in the Domain of Soviet Biological Periodicals (1917–1950), in *Russian Journal of Communication.* 6/1 (2014), 52–66; M. Demin, A. Kouprianov: Studying Kanonbildung: an exercise in a distant reading of contemporary self-descriptions of the 19th century German philosophy, in *Social Epistemology* [in press]; М. Демин,

истории университетов дореволюционной России и российских научных и философских журналов, методически – с приложением строгих формальных методов к анализу исторической динамики.

В первом разделе я остановлюсь на специфике данных, используемых в наших проектах, во втором – на примере проекта по анализу кадровой динамики российских университетов опишу формальные методы, которые мы используем, в заключительных разделах попытаюсь ответить на самые острые и неприятные вопросы: зачем это нужно, и что это все нам дает. Тот, кто боится утонуть в частностях, может обратиться к последним двум разделам. Для тех, кому интересны детали, во втором разделе не только описывается последовательность формальных процедур, но и приводятся примеры реализации их в виде фрагментов скриптов, написанных в среде статистического программирования и анализа данных R[4]. Надеюсь, их простота вдохновит начинающих аналитиков.

Данные

Средние и бедные

Digital humanities – гетерогенное образование, объединяющее под своими знаменами исследователей, выполняющих столь разнообразные проекты, что придать им общую методологическую оформленность пока трудно (в особенности при наличии конкурирующих идентичностей, вроде *Computational social science*, исторической информатики, количественной истории или математической лингвистики)[5]. Несмотря на эти затруднения, я попытаюсь

А. Куприянов: Digital Humanities на службе истории философии: методическое послесловие // *Логос* [in press]). Первые результаты работы группы оказались настолько впечатляющими, что побудили коллег, историков науки и историков образования, к сотрудничеству в области использования баз данных для анализа исторической динамики (см. Л. Жмудь, А. Куприянов: Социологический анализ античной науки: проблемы и перспективы // *Социология науки и технологий.* 7/ 1 (2016), 23–45, T. Kostina, A. Kouprianov: Growth or stagnation? Historical dynamics of the growth patterns of Dorpat University (1803–1884), in *Vestnik of Saint Petersburg University. History.* 3 (2016), 31–45.

4 См. *R: A Language and Environment for Statistical Computing*, in The R Development Core Team. Vienna, 2015. Режим доступа: https://www.r-project.org/

5 Уже введенный в обращение русский перевод «цифровые гуманитарные науки» на первый взгляд уродлив, а на второй – спорен, поэтому я намеренно его избегаю. О рождении маркера *Digital humanities* в переговорах с издателями и кулуарных обсуждениях после конференций см. подробнее в M. Kirschenbaum: What is Digital Humanities and what's it doing in English departments?, in: *Debates in the Digital Humanities* / ed. M. K. Gold. Minneapolis; L.: Univ. of Minnesota Press, 2012, 3–11. Что касается концептуального содержания, то определения DH колеблются между двух полюсов, на одном из которых подчеркивается роль новых медиа, преимущественно крупных онлайновых коллабораций с открытыми данными, включающими массивы текстов, изобра-

определиться с пониманием специфики *Digital humanities*, чтобы на этом фоне описать ансамбль практик, придающих своеобразие проектам нашей неформальной группы.

Я буду исходить из несколько механического представления об исследовательском процессе как комплексе процедур, связанных с постановкой задач, сбором и преобразованием данных и их обработкой. Если выбирать параметры для сравнения изо всей этой производственной цепочки (или производственной сети, поскольку процедуры не обязательно выполняются в строгой последовательности), то специфика *Digital humanities* будет отчетливее всего видна в особой природе данных и приемов работы с ними.

В чем состоят эти особенности? Беглый обзор проектов, результаты которых публикуют в изданиях, связанных с продвижением новой дисциплинарной идентичности, показывает, что почти везде мы имеем дело с вторжением технологий, применяемых в работе с «большими данными», в предметную область гуманитарных исследований. Когда мы говорим о «больших данных», речь идет не только о размерах массивов (*datasets*) самих по себе, хотя они, конечно, имеют значение, и не только о переводе данных в машиночитаемый вид. Более важные свойства «больших данных» – их принципиальная дезагрегированность и, где это возможно, их автоматический «захват» (см. табл. 1)[6].

жений, а порой аудио- и видеозаписей, и сложной техносоциальной инфраструктурой, а на другом – центральное значение инновативных аналитических подходов. Мне ближе второй, аналитический полюс, хотя мейнстрим *Digital humanities* явно тяготеет к первому. От необходимости подробно рассматривать историю вопроса меня избавляют недавно опубликованные обзоры, см., например, М. Таллер: Дискуссии вокруг Digital Humanities // *Историческая информатика*. 1 (2012), 5–13; А. Володин: Digital humanities (цифровые гуманитарные науки): в поисках самоопределения // *Вестник Пермского университета, серия История*. 3/26 (2014), 5–12; И. Гарскова: Информационное обеспечение гуманитарных исследований в цифровую эпоху: модели формирования и развития // *Вестник Пермского университета, серия История*. 3/26 (2014), 76–86.

6 В методологической литературе уже проскакивает новое словцо *capta*, см. J. Drucker: Humanities Approaches to Graphical Display, in: *Digital Humanities Quarterly*. 5/1 (2011). Режим доступа: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html. В отличие от слова *данные* (*data*), в котором просвечивает *данность* их пассивному исследователю, *capta – взятые? схваченные?* – подчеркивает активную роль исследователя, избирательно захватывающего нечто, исходя из своей специфической исследовательской оптики. Для меня важна не столько активность исследователя, сколько ее опосредованный «машиной» характер. Все большее значение приобретают гигантские массивы данных, генерируемые в результате взаимодействия мобильных устройств с коммуникационными *hubs* сетей, в которые они включены. *Capta* накапливаются в результате работы добровольно устанавливаемого программного обеспечения, отслеживающего в автоматическом или полуавтоматическом режиме виды активности пользователей компьютера или иного устройства. Массивы *capta* для третьей категории проектов создают программы-роботы, взаимо-

Табл. 1. Типы данных.

	Большие	Средние	«Традиционные»
Агрегированные	–	–	+
Машино-читаемые	+	+	±
Машино-генерируемые	+	–	–

При всей тривиальности, различение агрегированных и дезагрегированных данных нуждается в пояснении. Под агрегированными данными здесь подразумеваются обобщенные показатели популяций или выборок[7], будь то общая численность, средний возраст или какие-то более ухищренные статистики, например, стандартное отклонение или доверительный интервал для среднего значения. Под дезагрегированными данными – совокупность описаний индивидов популяции или выборки по интересующим нас параметрам.

Например, в нашем проекте по истории университетов к *агрегированным данным* можно отнести то, что можно извлечь из сводных таблиц, публиковавшихся в отчетах университетов или всеподданнейших отчетах Министра народного просвещения. Данные о составе профессоров и преподавателей университета за определенный год представлены в таком случае *одной цифрой* (общая численность) или *группой цифр* (численность по различным категориям профессорско-преподавательского состава, например, отдельно численность ординарных и экстраординарных профессоров или численность по факультетам). *Дезагрегированными данными* был бы *список* профессоров и преподавателей, числившихся при университете в данном году[8].

действующие с интернет-ресурсами через web-API (API – *application programing interface* – интерфейсы, позволяющие роботу захватывать информацию непосредственно из доступной онлайновой базы данных или с интернет-страниц ресурса).

7 Под популяцией здесь понимается «естественно» ограниченная совокупность индивидов, под выборкой – вычлененное по определенным правилам подмножество индивидов популяции. К выборочному методу прибегают в тех случаях, когда изучение популяции во всем ее объеме невозможно или нецелесообразно. В проектах, построенных на анализе «больших» (и, забегая вперед, «средних») данных, обычно работают с популяциями (например, «все пьесы У. Шекспира» или, в нашем случае, «все профессора и преподаватели университетов дореволюционной России»).

8 Эта информация может быть получена из различных источников. В отношении большинства категорий преподавателей (кроме приват-доцентов) – из адрес-календарей, списков лиц, служащих по Министерству народного просвещения, или памятных книжек по губерниям. Более широкий охват с включением приват-доцентов дает публиковавшиеся ежегодно объявления о лекциях, погодные академические списки или биографические словари профессоров и преподавателей университетов. Отсутствие технологий работы с данными такого рода долго ограничивало сферу их использования нарративным анализом.

Преимущество дезагрегированных данных состоит в том, что с использованием современных технологий они могут быть легко агрегированы, причем, сообразно количеству атрибутов, в различных отношениях, а вот «разобрать обратно» агрегированные данные уже невозможно. Например, сравнивая численность преподавателей университета за два соседних года по министерским отчетам невозможно без дополнительных изысканий понять, что стоит за этими двумя цифрами. Речь идет, разумеется, даже не об отдаленных причинах изменений численности (например, политика Министерства, давление со стороны попечителя, рутинные действия Совета, смертность и т.д.), а о самых приземленных вопросах. Следует ли объяснить прирост по сравнению с предыдущим годом только приходом новых преподавателей? Падение численности – только выбытием «старых»? Комбинацией этих двух процессов? Можно ли ожидать, что при сохранении численности остался неизменным состав преподавательского корпуса? Сравнение списков позволяет дать ответ на эти вопросы и сформулировать массу новых, которые невозможно поставить в отношении данных агрегированных.

Данные, с которыми идет работа в наших проектах, трудно назвать «большими». Счет обычно идет на сотни или тысячи записей (максимальный объем имеет *dataset* по *преподаванию наук* в Московском университете, насчитывающий около 6 000 записей) – довольно скромные масштабы по меркам *Data science*. Отчасти это связано с трудностями автоматизации захвата данных из имеющихся в нашем распоряжении источников. Тексты разнообразны (от биографических словарей до таблиц) и не имеют четкой структуры. В большинстве случаев затраты времени на сканирование, распознавание и выверку текста, а также на написание скрипта, который извлекал бы конкретно из этого текста необходимую информацию, сопоставимы с затратами на ввод тех же данных вручную. Ручной сбор добавляет человеческих ошибок и избавляет от некоторых машинных, однако главная проблема в том, что он требует значительных трудозатрат.

Желание сохранить ключевые принципы работы с «большими данными» в отсутствие машинного «захвата» приводит к компромиссам. Например, очевидно, что серьезная количественная история кадров высшей школы дореволюционной России подразумевала бы работу со всеми высшими учебными заведениями, однако эта задача для малой группы неподъемна, поэтому приходится ограничивать объемы популяции (анализируются только «классические» университеты)[9] и количество фиксируемых параметров.

[9] Эта условность заметна на университетах, возникших в результате преобразования ранее существовавших учреждений, например Санкт-Петербургского университета, возникшего на основе Главного педагогического института, Новороссийского университета в Одессе – на основе Ришельевского лицея, или даже Казанского – на основе местной гимназии.

В результате в распоряжении исследователей появляется то, что я предложил бы назвать «средними данными». Они наследуют у «больших» ряд их сильных сторон (дезагрегированность и ориентацию на работу с популяциями, а не выборками), но ограничены в объеме и «мощности». Этот же недостаток характерен и для «больших данных». При всех своих впечатляющих объемах в десятки и сотни тысяч записей, они «бедны» содержательно. Незначительное количество переменных ограничивает возможности создания каузальных моделей[10]. «Бедность» данных, как будет показано ниже, может быть частично компенсирована в процессе анализа за счет нетривиальных способов агрегирования. Однако избавиться от нее невозможно.

Структура массива данных

Каким образом выглядят данные, подготовленные для анализа? В проекте по истории научных журналов используется традиционный формат библиографического описания, к которому добавляются дополнительные переменные. Библиографические описания составляются на основе журналов либо заимствуются из сводных библиографий. Я не буду останавливаться на этом типе данных подробно. Любой историк в состоянии реконструировать формат исходных данных из приводимых ниже «журнальных» примеров даже при минимальном усилии воображения.

В проекте по кадровой динамике университетов Российской империи все несколько сложнее. Исходные данные в нем более гетерогенны, объекты содержательно богаче. В связи с этим нам пришлось выработать практический стандарт представления био-

[10] Наиболее знакомый мне в этом отношении пример лежит в иной области, но удачно иллюстрирует суть проблемы. У нас имеются довольно детальные данные по голосованию на выборах (иногда с точностью до микрорайона, которым примерно соответствуют избирательные участки). Массив по участковым избирательным комиссиям РФ насчитывает более 95 000 записей. При этом каждый участок весьма точно геопозиционирован. Вместе с тем до недавнего времени исследователям было нечего добавить к результатам голосования, поскольку ни одна социально-экономическая или демографическая характеристика населения не была картирована с той же степенью детализации. Лишь в самое последнее время в открытом доступе появились данные кадастра недвижимости (см. проект Александра Кукушкина. Режим доступа: https://github.com/alexanderkuk/analyze-reformagkh), но они, по сути, еще не введены в научный оборот. Это ограничивало возможности аналитиков поиском статистических аномалий в распределениях цифр без попытки связать хотя бы какие-то параметры голосования с особенностями местного электората (D. Kobak, S. Shpilkin, M. Pshenichnikov: Integer percentages as electoral falsification fingerprints, in: *The Annals of Applied Statistics*. 10/1 (2016), 54–73.). В результате при поражающих воображение масштабах фальсификаций, на которые указывают статистические аномалии, у лояльных властям социологов остается лазейка для того, чтобы списать многое на естественную гетерогенность населения.

графической информации. Я остановлюсь на нем подробнее, поскольку на его примере можно сформулировать некоторые общие принципы организации массивов данных такого рода.

Следует оговориться, что база данных, в которой хранится информация о преподавателях, устроена гораздо сложнее *dataset*'а, передаваемого для дальнейших преобразований и подсчетов в среду анализа данных[11].

Обмен между базой и средой анализа данных производится через промежуточный текстовый файл в формате CSV — *comma separated values* (массив данных, или *dataset*). Это текстовый файл, строки которого структурированы таким образом, что каждая из них становится строкой таблицы, разбитой знаками-разделителями (в нашем случае — запятыми) на ячейки. Человеческий глаз не всегда опознает такой текст как таблицу, но компьютерные алгоритмы «обращают внимание» не на визуальную презентацию текста, а на разделители строк и ячеек. В этом разделе речь пойдет именно о структуре *dataset*'а, поскольку он непосредственно используется в качестве объекта анализа и обмена данными между различными средами и участниками проекта.

Основа нашего обменного формата – набор из восьми обязательных ячеек или «полей»: *фамилия – имена – дата вступления в должность – дата увольнения от должности – должность – кафедра – факультет – университет*. Например:[12]

LASTNAME, NAMES, STARTYEAR, ENDYEAR, POSITION, DEPARTMENT, FACULTY, UNIVERSITY

"Аристов","Евмений Филипович",1837,1848,"Экстраординарный профессор",...

"Аристов","Евмений Филипович",1848,1864,"Ординарный профессор",...

На агрегировании по этим полям построена большая часть аналитических процедур.

В плане организации массива данных необходимо различать два типа характеристик индивида: *атрибуты* и *состояния*. Под *атрибутом* я подразумеваю характеристику, которая, будучи приписанной индивиду, может быть признана на уровне модели постоянной, под *состоянием* – характеристику индивида, которая в рамках модели подразумевает изменение. Строго говоря, идеальным атрибутом может быть только искусственно сгенерированный уникальный идентификатор индивида, так называемый

[11] В качестве платформы для создания базы данных используется *PostgreSQL*, а в качестве среды для анализа и статистической обработки – R.

[12] Фрагмент для примера заимствован из *dataset*'а по Казанскому университету. LASTNAME, NAMES, STARTYEAR, ENDYEAR, POSITION, DEPARTMENT, FACULTY, UNIVERSITY – имена полей, отточия символизируют продолжения строк, опущенные в примере и не вмещающиеся в ширину печатной страницы.

primary key базы данных, однако отклонение многих реальных характеристик от идеала может быть признано в рамках модели несущественным. Очевидно, что индивид может поменять в течение жизни многие «атрибуты» – подданство, вероисповедание, имя и даже паспортный пол (и все это может быть важно для понимания особенностей его или ее карьерного роста). Однако эти изменения, особенно на больших популяциях, пренебрежимо редки по сравнению с рутинной сменой «состояний», таких как должность в университете или чин в Табели о рангах[13].

С практической точки зрения это различение важно для структуры *dataset*'а. Две записи в приведенном выше примере описывают не индивида, а именно его преходящие состояния. Атрибуты (имя и фамилия) повторяются в разных строках, описывающих состояния одного и того же индивида. Состояния же, описываемые крайними датами и статусом, в котором индивид пребывал между этими крайними датами (например, должностью в университете), уникальны.

Помимо этих полей в каждую строку могут быть добавлены дополнительные поля с *атрибутами*. Например, датой и местом рождения или смерти, местом получения высшего или среднего образования, родом занятий родителей, вероисповеданием и т.д. Атрибуты, вносимые в *dataset* непосредственно из базы данных, я буду называть *первичными атрибутами*. На основе первичных атрибутов и состояний в ходе анализа могут быть сгенерированы *вторичные атрибуты*. Например, на основе первичного атрибута «год рождения» может быть автоматически рассчитан вторичный атрибут «возраст на определенный момент». Как это сделать, будет рассказано позже.

Анализ исторической динамики: процедура ступенчатого агрегирования

Итак, диагноз нашим данным поставлен – они «средние» и «бедные». Охарактеризована в общих чертах структура массива. Пришла пора показать, что позволяют сделать с этим среднего размера бедным массивом разработанные нами методы формального анализа. Этот раздел будет целиком посвящен процедурным моментам ступенчатого агрегирования – от создания подмножеств по категориям до извлечения показателей временной динамики.

Суть ступенчатого агрегирования состоит в том, что мы для целей анализа группируем наши «средние данные» последовательно по различным атрибутам и по временным срезам. Например, на первом этапе мы выделяем из всего массива только приват-доцентов, на втором – только приват-доцентов, окончивших какой-то определенный университет, или только тех приват-доцентов, которые позже занимали в этом же университете позиции экстра-

[13] В нашем примере все характеристики, кроме должностей, имеют статус атрибутов.

ординарных профессоров и т.д. По завершении этих операций мы рассчитываем численность профессорско-преподавательского состава погодно как в целом, так по вычлененным категориям, получая при этом агрегированные данные, непосредственно используемые при построении диаграмм и в расчетах.

Выделение категорий

С точки зрения программирования на R, большинство этих задач довольно тривиально. Вот как выделяется подмножество по атрибутам или состояниям при помощи команды subset()[14]:

```
kaz.pd <- subset(kaz, kaz$POSITION == "Приват-доцент")
```

Здесь kaz – имя объекта, содержащего массив данных по Казанскому университету, POSITION – имя переменной внутри этого объекта, значения которой проверяются на совпадение с паттерном «Приват-доцент», а kaz.pd – подмножество, включающее записи, соответствующие этому условию.

Выделение среди приват-доцентов тех, кто далее занимал в университете должность экстраординарного профессора, – более сложная задача. Сначала выделяется два подмножества – приват-доцентов и экстраординарных профессоров. Затем извлекается список общих для двух подмножеств уникальных имен. И, наконец, из подмножества приват-доцентов – только те строки, в которых встречаются имена, совпадающие с этим списком:

```
# Создание уникальных имен:
kaz$UN <- as.factor(paste(kaz$LASTNAME, kaz$NAMES, sep=", "))

# Выделение подмножеств:
kaz.pd <- subset(kaz, kaz$POSITION == "Приват-доцент")
kaz.eop <- subset(kaz, kaz$POSITION == "Экстраординарный профессор")

# Создание списка уникальных имен для пересечения подмножеств:
kaz.pd_eop.UN <- intersect(kaz.pd$UN, kaz.eop$UN)

# Отбор приват-доцентов, попавших в область пересечения:
kaz.pd_eop <- NULL
i <- 1
while (i <= length(kaz.pd_eop.UN)){
kaz.pd_eop <- rbind(kaz.pd_eop,
                    subset(kaz.pd, kaz.pd$UN == kaz.pd_eop.UN[i]))
i <- i+1
}
```

[14] Здесь и ниже примеры заимствованы из реального скрипта, задействованного в проекте по истории университетов.

В результате появляется ряд подмножеств исходного массива, полностью сохраняющих его структуру полей, но с меньшим количеством записей.

Создание временной последовательности

Сердце анализа временной динамики – фрагмент скрипта, генерирующего то, что мы называем *временной последовательностью*. В качестве материала для работы он берет исходный *dataset* или выделенные из него подмножества. Этот фрагмент скрипта создает серию подмножеств, каждое из которых содержит таблицу состояний и атрибутов для заданного временного интервала. Временной интервал может быть любым (например, при анализе *dataset*'а по математикам древности[15] он был равен 25 или 50 годам, в зависимости от вида анализа), но при анализе кадровой динамики университетов используется интервал длительностью в один год. В каждой итерации цикла из исходного *dataset*'а сначала отбрасывают тех, кто оставил службу до наступления этого шага, а затем тех, кто заступил на службу после его окончания. В результате в таблице для данного временного интервала остаются те, кто в его течение состоял на службе. Фрагмент скрипта для Петербургского университета выглядит так:

```
spb.ts <- NULL
spb.ts <- as.list(spb.ts)

i <- min(spb$STARTYEAR, na.rm=T)
while (i <= max(spb$ENDYEAR, na.rm=T)){
spb.ts[[i]] <- subset(spb, spb$ENDYEAR >= i & spb$STARTYEAR <= i)
i <- i+1
}
```

Он начинается с создания пустого объекта spb.ts, которому присваивается тип «list». Затем располагается цикл while(){}, в котором для каждого года с начального (в нашем случае – 1819) по конечный (в нашем случае – 1916, совпадающий с верхней границей данных, внесенных в базу) включительно поочередно выделяются подмножества (*subsets*) из исходного массива spb. Эти подмножества хранятся в объекте spb.ts под номерами, начиная с spb.ts[[1819]] и заканчивая spb.ts[[1916]], соответственно STARTYEAR и ENDYEAR – имена полей, в которых содержатся год вступления в должность и год увольнения от должности.

Такого рода временные последовательности могут быть созданы на основе любого подмножества исходного массива, сохраняющего структуру его полей. После того как создана временная последовательность, можно приступать к формированию временных рядов для различных показателей.

[15] Жмудь, Куприянов, op. cit.

Абсолютная и относительная численность по временным срезам

Переход к агрегированным данным происходит на этом этапе. Для формирования временных рядов используется цикл, в каждой итерации которого на основании уникальных имен рассчитывается численность определенной категории лиц для данного временного среза. Например, для расчета временного ряда общей численности профессоров и преподавателей Петербургского университета цикл будет иметь вид:

```
spb.dyn.TOTAL <- NULL

i <- min(spb$STARTYEAR, na.rm=T)
while (i <= max(spb$ENDYEAR, na.rm=T)){
spb.dyn.TOTAL <- c(spb.dyn.TOTAL, length(unique(spb.ts[[i]]$UN)))
i <- i+1
}
```

Таким образом могут быть сформированы *абсолютные* параметры динамики: численность по определенным категориям, выделенным на основании первичных или вторичных атрибутов. На их основе могут быть рассчитаны *относительные* параметры динамики: доли одних категорий в общей численности по другим (например, доля приват-доцентов, ставших впоследствии экстраординарными профессорами, среди всех приват-доцентов, числившихся на данный год).

Показатели движения и преемственности корпуса

Помимо моментальной численности, сформированная *временная последовательность* дает возможность рассчитать показатели движения и преемственности корпуса: количество вновь появившихся или исчезнувших представителей той или иной категории на данный год и преемственность от года к году. Все эти показатели рассчитываются путем сопоставления списков уникальных имен за соседние годы при помощи команды setdiff(), определяющей разность множеств, и intersect(), выделяющей область пересечения. Расчет показателей движения тривиален. Важнее показать расчет показателя преемственности[16]. Принцип здесь тот же, что и при работе с показателями движения, но расчетная часть сложнее. Формула для исчисления сходства по Жаккару в приложении к нашим

[16] Изначально эта мера сходства была предложена для сравнения списков видов растений различных регионов, см. P. Jaccard: The Distribution of the Flora in the Alpine Zone, in: *New Phytologist*. 11/2 (1912), 37–50. Для оценки преемственности корпуса она использовалась в работе по библиометрическим кризисам, см. Куприянов, op. cit.; Kouprianov, op. cit.

данным о составе преподавателей за соседние годы (например, 1819 и 1820) будет выглядеть следующим образом:

$$J_{(N_{1819},N_{1820})}=\frac{|N_{1819}\cap N_{1820}|}{|N_{1819}\cup N_{1820}|},$$

где N_{1819} – список имен по категории на 1819, а N_{1820} – на 1820 г.

Один из возможных вариантов реализации в коде R представлен ниже.

```
spb.dyn.JACCARD <- NULL

i <- min(spb$STARTYEAR, na.rm=T)+1
while (i <= max(spb$ENDYEAR, na.rm=T)){
spb.dyn.JACCARD <- c(spb.dyn.JACCARD,
        length(intersect(spb.ts[[i]]$UN, spb.ts[[i-1]]$UN))/
        length(unique(c(spb.ts[[i]]$UN, spb.ts[[i-1]]$UN)))
        )
i <- i+1
}
```

При дальнейшей работе с этим показателем важно помнить, что сходство с предшествующим годом не может быть рассчитано для начального года, поэтому получившийся числовой ряд будет на один член короче.

Повторение: общая схема анализа исторической динамики корпуса преподавателей

Общая схема ступенчатого агрегирования в упрощенном виде представлена на рис. 1. Упрощения эти касаются далеко не полного набора возможных первичных и вторичных атрибутов и схемы расчета показателей движения и преемственности корпуса (показаны только для всех профессоров и преподавателей в целом).

Взяв за основу исходный массив, мы начинаем с выделения подмножеств по первичным и вторичным атрибутам. Затем для всего массива и для каждого из подмножеств строятся временные последовательности, из которых извлекаются числовые ряды, отражающие абсолютные показатели численности для временных срезов и показатели движения и преемственности корпуса от одного среза к другому. В результате формируется числовой массив, описывающий в виде ряда переменных кадровую динамику университета за охваченный период. Ряды для относительной численности не генерируются при агрегировании, поскольку их легко получить уже в процессе работы с итоговым числовым массивом, разделив члены одного числового ряда на члены другого.

В следующем разделе я постараюсь объяснить, какое применение можно найти этому «мешку с цифрами», сформированному

на основе исходного массива, фиксировавшего сегменты карьерных траекторий.

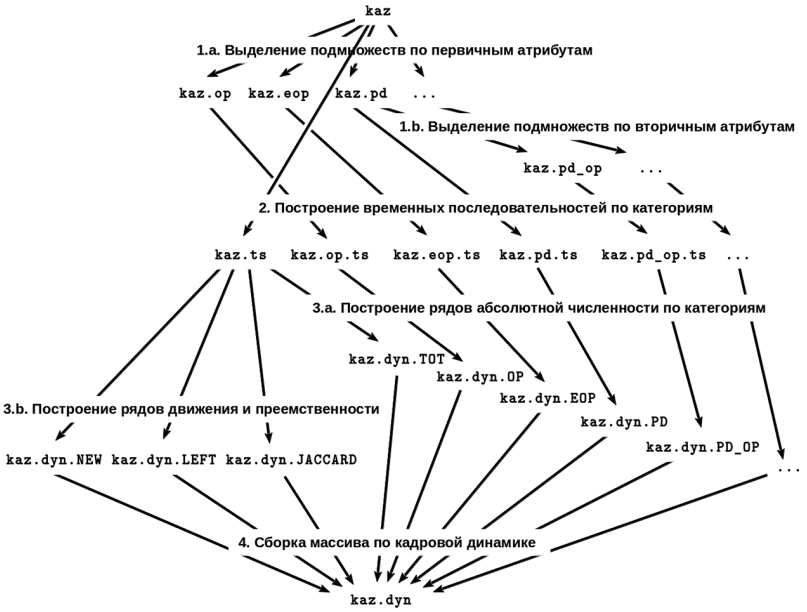

Рис. 1. Общая схема преобразования данных при анализе исторической динамики корпуса профессоров и преподавателей на примере массива по Казанскому университету. Многоточия обозначают потенциально возможные объекты, ради экономии места не показанные на схеме. Имена объектов: kaz – исходный массив дезагрегированных данных; подмножества, выделенные по первичным атрибутам: kaz.op – ординарные профессора, kaz.eop – экстраординарные профессора, kaz.pd – приват-доценты; подмножества, выделенные по вторичным атрибутам: kaz.pd_op – приват-доценты, ставшие впоследствии ординарными профессорами в этом же университете; временные последовательности: kaz.ts – для всех профессоров и преподавателей в целом, kaz.op.ts, kaz.eop.ts, kaz.pd.ts, kaz.pd_eop.ts – для перечисленных ранее категорий профессоров и преподавателей по отдельности; ряды абсолютной численности по временным интервалам: kaz.dyn.TOT – для всех профессоров и преподавателей в целом, kaz.dyn.OP, kaz.dyn.EOP, kaz.dyn.PD, kaz.dyn.PD_EOP – для перечисленных ранее категорий профессоров и преподавателей по отдельности; ряды показателей движения и преемственности: kaz.dyn.NEW – вступившие в должность, kaz.dyn.LEFT – оставившие службу, kaz.dyn.JACCARD – сходство с предшествующим временным интервалом по Жаккару; kaz.dyn – итоговый массив агрегированных данных.

Преимущества формального анализа

Прежде чем перейти к обсуждению вопроса о том, что дают историкам все эти упражнения в программировании, хотелось бы обратить внимание читателей на одно обстоятельство. В цепи преобразований исходных данных, описанных в предыдущих разделах,

нет ни одного звена, в котором мы не могли бы обойтись без компьютера. Вся процедура ступенчатого агрегирования может быть проделана вручную. Да, на это уйдет больше времени. Возможно, добавятся человеческие ошибки. Я говорю об этом не для того, чтобы объяснить, что компьютер, по большому счету, *не нужен*. Я говорю об этом, чтобы было ясно, что он *не страшен*. Он не делает ничего противоестественного. Команды R – не таинственные заклинания, это просто доведенные до известного предела краткости описания формальных процедур. Нет никакой сущностной разницы между командой «отобрать имена, общие для двух списков», данной на естественном языке, и командой intersect(x, y).

Самое очевидное преимущество компьютерной обработки – скорость. Легкость, с которой производится агрегирование исходных данных на различных основаниях, позволяет ставить все менее и менее тривиальные задачи и решать их в обозримые сроки. Написание и отладка скрипта для новой ветки анализа занимает от нескольких минут до нескольких часов, но это намного быстрее, чем ручной подсчет. Кроме того, существенные затраты времени потребуются только на первый раз. Запись последовательности команд в отдельном текстовом файле позволяет вернуться к анализу, воспроизвести его, выявить ошибки, двинуться дальше, перенести алгоритм анализа, разработанный для одного массива данных, на другой, аналогичный.

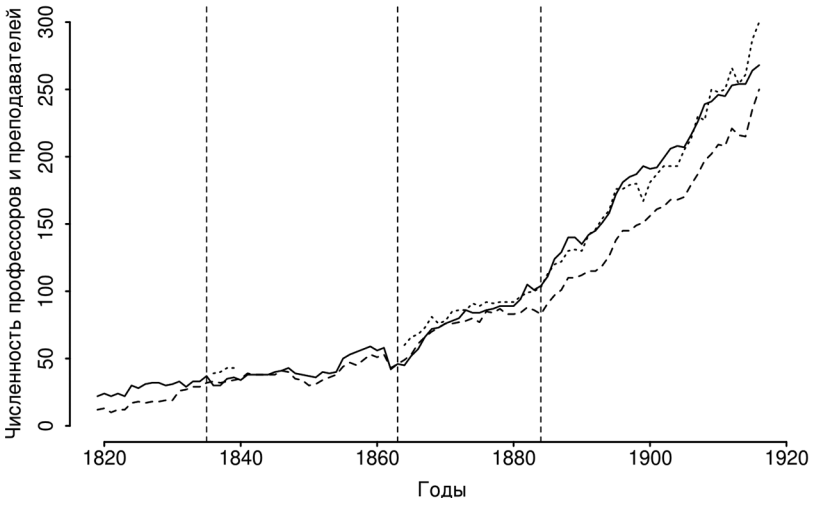

Рис. 2. Пример использования системы ступенчатого агрегирования для оценки эффекта альтернативных операционализаций: сопоставление данных о численности профессоров и преподавателей Санкт-Петербургского университета по различным источникам. Сплошная линия – опубликованные списки, прерывистая – онлайновая база данных по профессорам и преподавателям ИСПбУ, пунктирная – опубликованные отчеты (данные имеются за 1836–1839 и 1864–1916 гг.); прерывистые вертикальные линии – университетские Уставы 1835, 1863 и 1884 гг.

Рис. 3. Пример использования системы ступенчатого агрегирования для оценки полноты данных: доступность данных о месте получения образования профессоров и преподавателей Санкт-Петербургского университета. По вертикальной оси – доля ППС на данный год, для которых имеются данные о месте получения образования; сплошная линия – данные о высшем, пунктирная – о среднем образовании; прерывистые вертикальные линии – здесь и далее – университетские Уставы 1835, 1863 и 1884 гг.

Легкость агрегирования не только позволяет значительно обогатить анализ. Она помогает решить две методические проблемы подготовительного этапа исследований. Речь идет об альтернативных операционализациях и оценке меры незнания или неполноты данных.

С проблемой альтернативных операционализаций мы сталкиваемся постоянно, в том числе в исторических исследованиях. Как, например, оценить долю профессоров-иностранцев в российских университетах? Кого считать иностранцем? Что определяет статус? Город или страна рождения? Место получения образования? Подданство? Подданство родителей? Язык, на котором профессор вел преподавание? Обсуждение критериев включения в эти и подобные категории – рутинная процедура при определении границ популяции. С новыми возможностями эта проблема из экзистенциальной превращается в чисто техническую. Мы можем не только оперативно провести альтернативную операционализацию, но и оценить ее вклад в конечный результат (рис. 2). То же касается «слепых зон». Можем ли мы использовать при оценке географической мобильности преподавателей данные об их образовании? Один график (рис. 3) дает нам ответ на этот вопрос. Например, для Санкт-Петербургского университета данные за первую половину XIX в. как по высшему, так и по среднему образованию использовать почти бессмысленно. Только во второй половине XIX в., когда сведения о высшем образовании становятся доступны, их можно

анализировать, и то с осторожностью, поскольку мы не в состоянии оценить систематическую ошибку, которая могла возникнуть из-за неизвестного фактора, повлиявшего на доступность данных. Вместе с тем ситуация с опубликованными данными о высшем образовании много лучше, чем с данными о среднем, которые на большинстве временных срезов известны менее, чем для 60% преподавателей.

К скорости преобразования данных и расчетов я бы добавил возможности визуализации. Речь идет не только о завершающем презентационном этапе, хотя R трудно найти замену при изготовлении иллюстраций, пригодных для качественной печати в научном журнале. Возможность быстро генерировать стандартные наборы графиков – важный инструмент эксплораторного этапа анализа[17]. Например, на ранних этапах работы над количественной характеристикой временной динамики ряда параметров корпуса немецкоязычной литературы по истории философии скрипт генерировал около полутора – двух сотен изображений, анализ которых позволил быстро выявить некоторые тенденции и провести предварительное тестирование ряда рабочих гипотез[18]. В процессе работы над статьей по количественной истории древнегреческой науки[19] автоматически генерировались десятки графиков исторической динамики и сотни географических карт. Визуализация превращается из экзотического, требующего большого мастерства ремесла в обыденный аналитический инструмент.

Помимо облегчения работа с компьютером полезна тем, что дисциплинирует и приучает к формализации. Так, алгоритмизация анализа кадровой динамики российских университетов позволила выработать особый язык, открывающий широкие перспективы для формального описания и строгого сравнения того, что раньше даже не приходило в голову ни описывать, ни сравнивать. Принципиальное значение имеют возможность охарактеризовать количественно любое подмножество преподавателей, выделенное на основании первичных или вторичных атрибутов, и возможность представить временную динамику в виде числовых рядов с любой степенью детализации.

Мне не раз приходилось слышать от историков, которых я, демонстрируя графики, знакомил с предварительными выводами, что они «приблизительно этого и ожидали», поскольку, например, было интуитивно понятно, что университеты будут расти со временем и что в какой-то момент этот рост будет происходить преимущественно за счет приват-доцентов. Я согласен с тем, что многие из выявляемых тенденций ожидаемы (будь это не так, следовало бы тревожиться либо за историков образования, совершенно не-

[17] Концепция эксплораторного анализа данных восходит к работам американского математика Дж. Тьюки. См. J.W. Tukey: *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977.

[18] Demin, Kouprianov, op. cit.

[19] Жмудь, Куприянов, op. cit.

способных заметить очевидные вещи, либо за компьютерные алгоритмы, выявляющие нечто, ни с чем не сообразное). Вместе с тем, во-первых, изменения некоторых показателей трудно, а каких-то – практически невозможно заметить на глаз, ничего не считая и не рисуя. Во-вторых, интуитивные ощущения (даже верные), в отличие от числового ряда, трудно подставить в регрессионную модель. Перевод сравнения на количественную основу позволяет не только строже подойти к выявлению наличия или отсутствия какого-либо эффекта, но и оценить его размер. Мы делаем первые попытки применения формальных методов к тестированию гипотез о причинах кадровой динамики, но даже первые, скромные результаты вселяют надежду[20]. В-третьих, созданный нами инструмент построения временных рядов позволяет по-новому подойти к изучению истории в процессуальном измерении. Одна из фундаментальных проблем исторического описания – трудность перехода от описания стабильных состояний, сменяющих друг друга, к описанию процессов трансформации. Наши диаграммы позволяют визуализировать некоторые процессы практически в кинематографическом режиме. Да и как выделить эти стабильные состояния? В той же истории университетов – на что положиться при периодизации? Можно ли считать университетские уставы или отдельные постановления, их дополнявшие, за вехи, отмечающие границы периодов относительной стабильности? Или пульс, взлеты и падения в развитии университетов задают свою, «внутреннюю» периодизацию, подобно тому как не вполне совпадают изменения формальных статусов и внутреннее ощущение «взросления» человека? Переход к периодизации с опорой на «внутреннее» время развивающейся системы, измеряемое сменой ее состояний, – вот одна из наиболее амбициозных задач нашего проекта.

Увидеть невидимое, посчитать едва ощутимое

Позволю себе привести пару примеров, в которых формальный анализ дал не вполне тривиальные результаты. Первый из них, заимствованный из «университетского» проекта, важен сразу в нескольких отношениях. Во-первых, он показывает, как, располагая лишь бедными данными, можно создать объемную картину исторической динамики. Во-вторых, в нем задействовано сразу несколько невидимых невооруженным глазом производных показателей. Второй взят из «журнального» проекта и демонстрирует неожиданные побочные преимущества работы с популяциями.

[20] См. Иванова, op.cit.; Kostina, Kouprianov, op. cit.

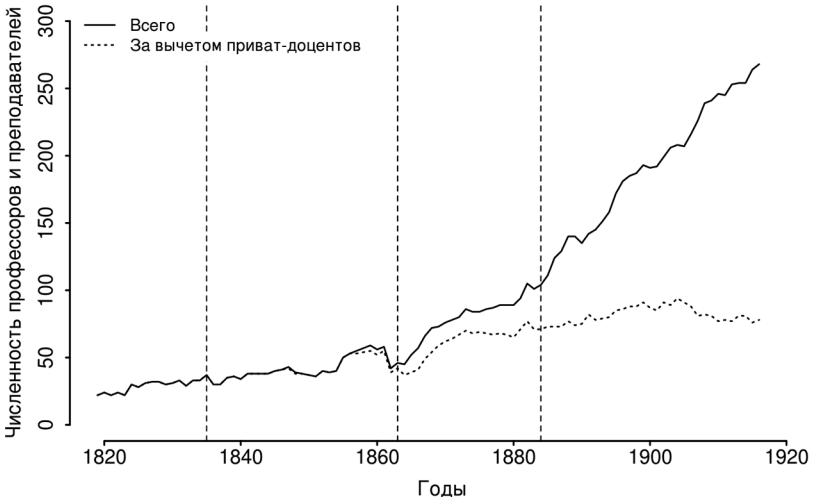

Рис. 4. Динамика численности корпуса профессоров и преподавателей Санкт-Петербургского университета. Обратите внимание на роль приват-доцентов во взрывном росте университета после принятия устава 1863 г.

Я не буду разворачивать полную картину временной динамики университетской корпорации в широкой сравнительной перспективе, ограничусь лишь необходимым минимумом иллюстраций. Начнем с самого простого – динамики общей численности (см. рис. 4). Мы видим, что рост численности преподавателей университета происходил неравномерно. Медленное и постепенное увеличение численности в течение первых нескольких десятилетий сменяется «взрывным» ростом после введения в действие Устава 1863 г. Устав 1884 г. приводит к еще большему ускорению роста. Анализ по категориям преподавателей показал, что эти процессы связаны с возникновением института приват-доцентуры и стремительным ростом численности приват-доцентов. Если до начала 1860-х гг. самой многочисленной категорией преподавателей были ординарные профессора, то уже в середине 1880-х приват-доценты по численности сравнялись с ними и превзошли их. В середине 1890-х приват-доценты составляли уже половину всего профессорско-преподавательского состава, и в дальнейшем их доля только увеличивалась. Сравнительный анализ показал, что различная скорость роста университетских корпораций была обусловлена, в том числе, различиями в статусах приват-доцентов в разных университетах[21].

Как мы помним, при обращении к дезагрегированным данным мы получили возможность посмотреть не только на колебания общей численности и численности преподавателей по категориям, но и на показатели преемственности и движения корпуса. Из них я продемонстрирую только динамику преемственности, оценива-

[21] Подробнее см. Kostina, Kouprianov, op. cit.

емую при помощи сходства по Жаккару между списочными составами университета за соседние годы. Добавление этого параметра позволяет выявить кризисные моменты в истории университетов, которые становятся легко заметны благодаря резким кратковременным падениям значений коэффициента Жаккара (рис. 5). Для Казанского и Санкт-Петербургского университетов это были, прежде всего, массовые увольнения 1819–1821 гг., произведенные в Казани М. Л. Магницким, а в Петербурге – Д. П. Руничем (и последовавший затем найм преподавателей), и перемены, связанные с введением в действие устава 1835 г. Заметим, что введение устава 1863 г. сказалось уже гораздо менее, а устав 1884 г. и вовсе не выделяется на фоне окружающих его случайных флуктуаций.

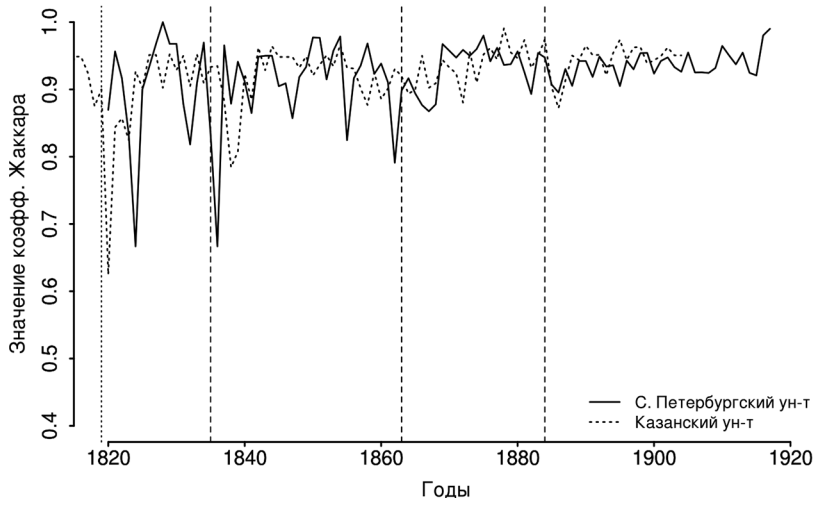

Рис. 5. Динамика преемственности корпуса профессоров и преподавателей Санкт-Петербургского и Казанского университетов (сходство по Жаккару для данного года по сравнению с предыдущим). Обратите внимание на разрывы преемственности в годы, следующие за «ревизиями» Магницкого и Рунича и принятием университетского устава 1835 г.

Еще одно измерение динамики открывается, когда мы начинаем анализировать вторичные атрибуты, связанные с карьерным продвижением в рамках какого-то одного университета. Речь идет об оценке количества переходов между позициями (например, о том, сколько приват-доцентов смогло стать экстраординарными и впоследствии ординарными профессорами). Я покажу для примера динамику только одной группы параметров из целого семейства. Речь пойдет об источниках пополнения корпуса ординарных профессоров Санкт-Петербургского университета (рис. 6). На протяжении исследуемого периода заметна устойчивая тенденция к снижению доли людей, пришедших на позицию ординарного профессора со стороны. Первый резкий скачок доли «инсайдеров» (до 40%) при

введении в действие Устава 1835 г. связан с производством в ординарные профессора большого количества экстраординарных.

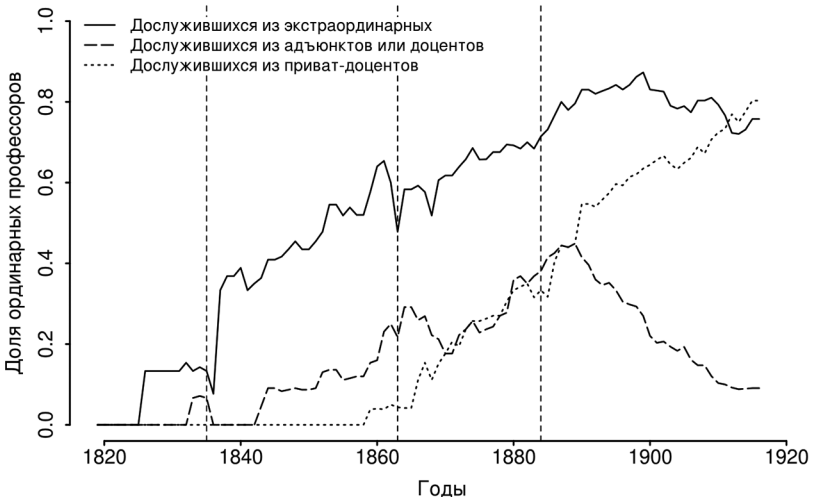

Рис. 6. Формирование замкнутой системы карьерного продвижения в Императорском Санкт-Петербургском университете. Динамика долей ординарных профессоров, ранее бывших в этом же университете экстраординарными профессорами, адъюнктами или доцентами и приват-доцентами. Обратите внимание на рост значения приват-доцентов как источника кадров для профессуры.

К середине 1860-х гг. доля ординарных профессоров, бывших ранее здесь же экстраординарными, стабильно превышает 50% и к концу исследуемого периода держится в диапазоне 70–80%. Нарастание замкнутости системы карьерного продвижения в отдельно взятом университете становится еще более заметным, если мы обратимся к более низким уровням иерархии преподавателей: адъюнкт-профессорам, штатным доцентам и приват-доцентам. До 1860-х гг. лица, дослужившиеся «на месте» с позиций адъюнкта, составляли ничтожно малую (порядка 10–15%) часть ординарных профессоров. К 1890-м уже половину ординарных профессоров составляли лица, начавшие карьеру в данном университете с позиций доцентов или приват-доцентов, а к концу исследуемого периода они уже находились в подавляющем большинстве (порядка 80%). Следует отметить, что в других университетах наблюдаются похожие тенденции, но степень замкнутости их профессорских корпораций различна. Например, в Казани она выше, а в Дерпте – ниже, чем в Петербурге.

Как мы видим, уже базовый набор временных рядов дает объемную картину количественной истории университетов. Добавление в *dataset* хотя бы одного дополнительного первичного атрибута позволяет значительно оживить ее. Попробуем добавить к ба-

зовому набору высшие учебные заведения, оконченные нашими героями и годы их рождения.

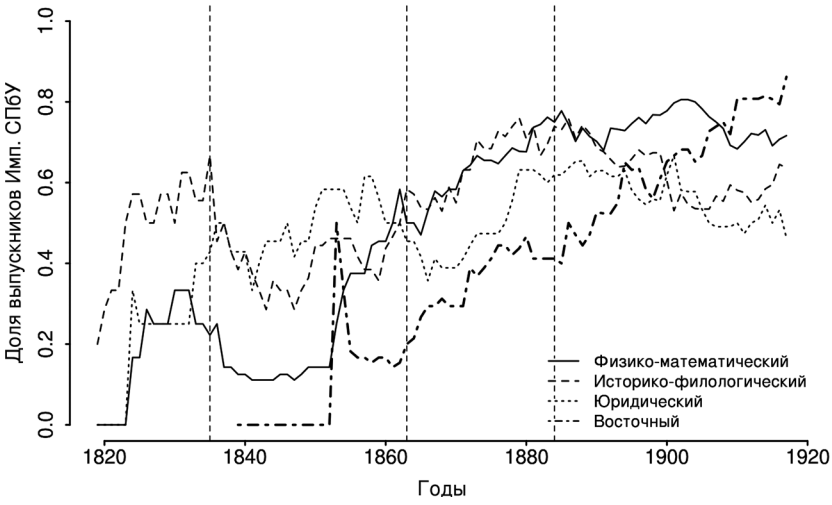

Рис. 7. Динамика академического инбридинга в сравнительной перспективе: четыре факультета Императорского Санкт-Петербургского университета. Обратите внимание на относительно согласованную динамику физико-математического и историко-филологического факультетов и на своеобразие юридического и восточного.

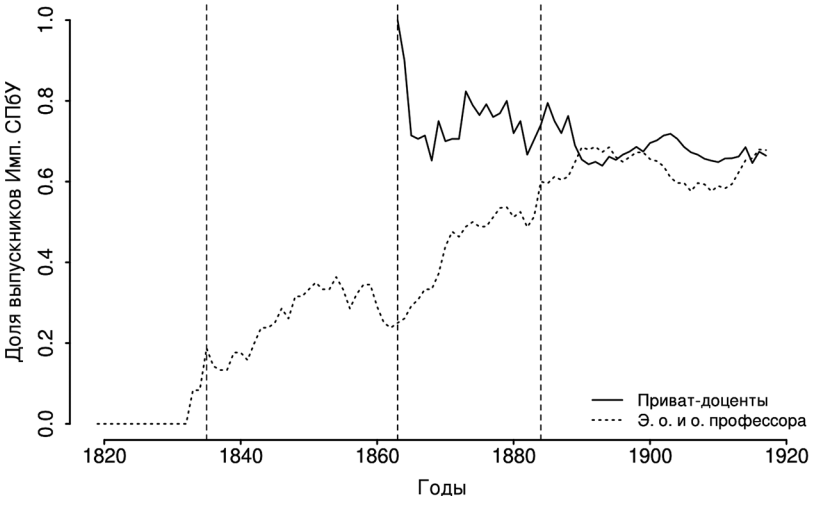

Рис. 8. Динамика академического инбридинга в сравнительной перспективе: приват-доценты и профессора Императорского Санкт-Петербургского университета. Обратите внимание на систематически более высокую долю выпускников ИСПбУ среди приват-доцентов.

Первый дополнительный атрибут позволяет углубить наши представления о системе карьерной мобильности преподавателей университетов. Один из важных показателей при ее изучении –

уровень академического инбридинга (доля выпускников данного университета в числе его преподавателей). В современных исследованиях по образовательной политике это один из важных показателей оценки институциональной среды вуза[22]. Нам удалось установить, что уровень инбридинга значительно изменялся на протяжении истории университетов дореволюционной России. Более того, были выявлены как общие особенности его динамики, так и систематические отличия университетов (например, доля собственных выпускников среди преподавателей Петербургского и Казанского университета была выше, чем в Дерпте), факультетов (например, восточный факультет Петербургского университета выделяется среди прочих стабильно восходящим трендом инбридинга, к 1917 г. он почти полностью комплектовался своими собственными выпускниками, см. рис. 7) и категорий преподавателей (среди приват-доцентов доля выпускников местного университета, как правило, была выше, чем среди профессоров, см. рис. 8).

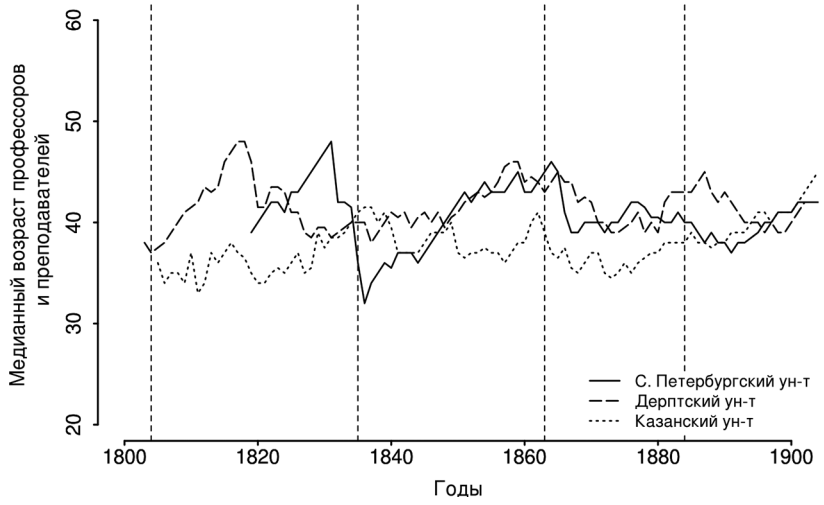

Рис. 9. Динамика возрастного состава: медианный возраст преподавателей трех университетов.

Второй дополнительный атрибут позволяет оценить динамику возрастного состава. Зная даты рождения, мы можем получить распределения по возрастам на каждый год и, соответственно, визуализировать динамику основных показателей этих распределений. Более того, мы можем получить эти данные не только по всем преподавателям, но и по отдельным их категориям. Поначалу график, отражавший динамику возрастного состава, казался нам не более чем очередным курьезом, демонстрирующим возможность на-

[22] См.: *Академический инбридинг и мобильность в высшем образовании: Глобальные перспективы* / Под ред. М.М. Юдкевич, Ф.Дж. Альтбаха и Л. Рамбли, пер. Г.С. Петренко. М.: Изд. дом Высшей школы экономики, 2016.

глядно отображать какие-то труднопредставимые показатели, еще одним штрихом, придающим объем картине исторической динамики. Однако отношение к нему изменилось после того, как стало ясно, что для разных университетов характерны различные паттерны динамики возрастного состава. В то время как в Казани и Петербурге экстраординарные профессора составляли особую страту, по своим возрастным характеристикам промежуточную между младшими штатными преподавателями и приват-доцентами с одной стороны и ординарными профессорами с другой, в Дерпте экстраординарные профессора ничем не отличались по возрасту от адъюнктов, доцентов и приват-доцентов. Точно так же неожиданно обнаружилось, что профессорско-преподавательский корпус Казанского университета был, за исключением двух относительно непродолжительных периодов, систематически младше, чем в Петербургском и Дерптском университетах (см. рис. 9). Установление причин и содержательная интерпретация выявленных систематических отличий – задача дальнейших исследований. Но, если бы не новые методы анализа и визуализации, мы вряд ли вообще увидели бы эту разницу и обратили бы на нее внимание.

Отмечу, что почти все показатели, кроме общей численности преподавателей, при всей их простоте трудно, если вообще возможно, оценивать на глаз. Тем более трудно уловить колебания, связанные как с активными интервенциями в области кадровой политики, так и с рутинными процедурами найма.

Наконец, работа с популяциями, а не выборками позволяет расширить спектр возможных исследовательских задач еще в одном направлении. Есть целый раздел в области анализа данных, опирающийся на алгоритмы анализа социальных сетей, который не столько по формальным, сколько по содержательным причинам не может работать с выборками и требует данных о популяциях. Исходно анализ социальных сетей был создан для изучения структуры контактов между индивидами. Когда мы говорим о применении методов анализа социальных сетей, речь идет не только о визуализации структуры связей в сообществе в виде графа, в узлах которого находятся люди, а ребра обозначают наличие связи того или иного рода, но и о целой системе метрик, позволяющей количественно оценить параметры получившейся сети. Значительная часть этих метрик содержательно осмысленна только в тех случаях, когда мы используем аппарат анализа социальных сетей для изучения инфраструктуры распространения информации и принятия решений и аналогичных процессов. Вместе с тем некоторые метрики более общего характера могут быть содержательно осмыслены и в тех случаях, когда узлами сети выступают не люди, а связи, не интерпретируемые в терминах потоков информации. Хотя некоторые элементы анализа социальных сетей задействованы и

при изучении университетов, бóльших результатов мы достигли в рамках одного из «журнальных» проектов[23].

Сетевой анализ позволил не только визуализировать структуру поля русскоязычной философской периодики начала XX в., но и более-менее строго охарактеризовать его трансформацию в результате событий, последовавших за революциями 1917 г.[24] В частности, удалось показать, что, несмотря на сам факт наличия связи между дореволюционной русской философской периодикой и послереволюционными изданиями как Советской России, так и «русского зарубежья», характер связей между этими тремя субдоменами различался. Уже визуальная инспекция графа (см. рис. 10) показывала, что русское зарубежье гораздо сильнее связано с дореволюционной традицией (сюрпризом было наличие преемственности между дореволюционной и послереволюционной советской периодикой). Формальный количественный анализ подтвердил это наблюдение. При этом ситуация казалась несколько парадоксальной, поскольку авторов, общих для дореволюционной и послереволюционной советской периодики, было немногим меньше, чем авторов, общих для дореволюционной периодики и периодики русского зарубежья. Дальнейший анализ позволил понять, что различия были обусловлены как «качеством» авторов (Советская Россия и СССР унаследовала от дореволюционных философских журналов маргинальных авторов с низкой публикационной активностью, русское зарубежье – много центральных фигур, активно публиковавшихся как до, так и после революции), так и их публикационными стратегиями (в Советской России и СССР наиболее продуктивными авторами были те, кто публиковался только в одном журнале, в русском зарубежье – те, кто публиковался в широком спектре изданий).

Интересны не столько полученные результаты, сколько то, что эти результаты добыты на чрезвычайно «бедных» данных – базовом *dataset'e* по университетской профессуре с добавлением пары переменных, оглавлениях журналов, переведенных в табличный формат. Очевидно, для оживления этих «скелетных» картин прошлого нам потребуется дополнительная работа интерпретации. Однако скелет уже богат деталями.

Благодарности

Материалы, положенные в основу статьи, были представлены в докладах на коллоквиумах «Усложнение гуманитария:

[23] Результаты частично опубликованы, см. Фотиади, op. cit., их резюме и приводится ниже.

[24] В обсуждаемом примере речь идет о построении сети, в которой вершины обозначают журналы, а связи между ними — наличие общих для связываемой пары журналов авторов. Сила связи определялась количеством общих для двух журналов авторов: чем больше их было, тем сильнее связь.

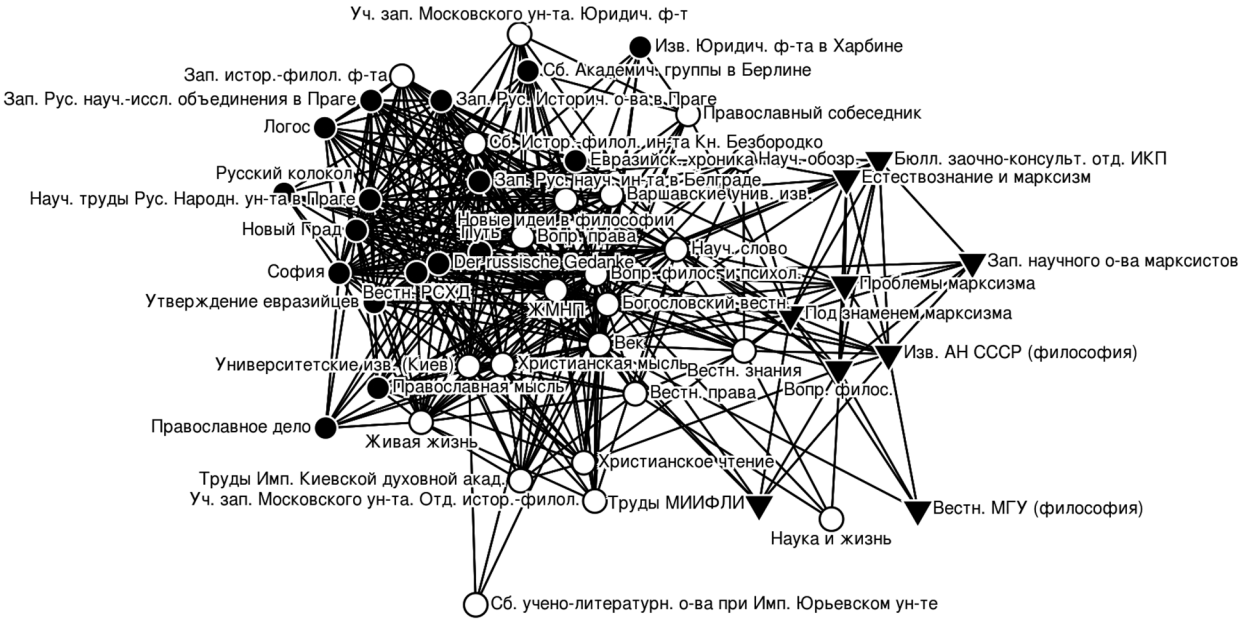

Рис. 10. Сеть русских журналов, в которых публиковались работы по философии. Белые круги – издания дореволюционной России (1901–1919, $N = 24$); черные круги – издания в эмиграции (1919–1939, $N = 17$); черные треугольники – Советская Россия и СССР (1921–1950, $N = 10$). Изоляты не показаны. Связь между журналами (узлами) означает наличие общих для этих журналов авторов. Расшифровка названий не приводится, поскольку нас интересует общая форма сети и характер взаимного расположения узлов трех вышеперечисленных групп.

дигитальная учёность в эмпирическом исследовании» (Москва, РАНХиГС, 12–13 сентября 2014) и «Топология цифровых расширений» (Москва, РАНХиГС, 11–12 декабря 2015) и на семинаре «Big Data Approaches to Intellectual and Linguistic History» (Хельсинки, Helsinki Collegium for Advanced Studies, 1–2 декабря 2014). Хотел бы поблагодарить всех, кто принимал участие в обсуждениях. Автор считает своим приятным долгом выразить особую благодарность А.Л. Зорину (РАНХиГС, Москва) – за стимулирующие критические замечания; Г.А. Орловой (РАНХиГС, Москва; ЕГУ, Вильнюс) – за приглашение принять участие в работе «цифровых» коллоквиумов и непоколебимую веру в то, что я смогу дописать этот текст до конца; моим коллегам и соавторам М.Р. Демину (НИУ ВШЭ, С. Петербург), Л.Я. Жмудю (ИИЕТ РАН, С. Петербург) и Т.В. Костиной (СПФ АРАН, С. Петербург), сотрудничество с которыми показало, что в сообществе историков философии, науки и образования есть живой интерес к возможностям, открывающимся при использовании формальных методов анализа; наконец, бывшим и нынешним студентам Петербургской школы социальных и гуманитарных наук НИУ ВШЭ и, в особенности, Е.В. Ивановой, В.М. Комовой и М.Ф. Фотиади, с которыми мы вместе работали над количественной историей журналов и университетов.