МЕТОДЫ И МЕТОДОЛОГИЯ

DOI: 10.14515/monitoring.2019.2.02

Правильная ссылка на статью:
Жучкова С. В., Ротмистров А. Н. Поиск многомерной связи категориальных признаков: сравнение CHAID, логлинейного анализа и множественного анализа соответствий // Мониторинг общественного мнения: Экономические и социальные перемены. 2019. № 2. С. 32—53. https://doi.org/10.14515/monitoring.2019.2.02.
For citation:
Zhuchkova S. V., Rotmistrov A. N. (2019) In search of multivariate associations: comparison of CHAID, log-linear analysis, and multiple correspondence analysis. *Monitoring of Public Opinion: Economic and Social Changes.* No. 2. P. 32—53. https://doi.org/10.14515/monitoring.2019.2.02.

С. В. Жучкова, А. Н. Ротмистров

ПОИСК МНОГОМЕРНОЙ СВЯЗИ КАТЕГОРИАЛЬНЫХ ПРИЗНАКОВ: СРАВНЕНИЕ CHAID, ЛОГЛИНЕЙНОГО АНАЛИЗА И МНОЖЕСТВЕННОГО АНАЛИЗА СООТВЕТСТВИЙ

ПОИСК МНОГОМЕРНОЙ СВЯЗИ КАТЕГОРИАЛЬНЫХ ПРИЗНАКОВ: СРАВНЕНИЕ CHAID, ЛОГЛИНЕЙНОГО АНАЛИЗА И МНОЖЕСТВЕННОГО АНАЛИЗА СООТВЕТСТВИЙ

ЖУЧКОВА Светлана Васильевна — студентка 1 курса магистратуры факультета компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: lana_lob@mail.ru
https://orcid.org/0000-0002-4425-725X

IN SEARCH OF MULTIVARIATE ASSOCIATIONS: COMPARISON OF CHAID, LOG-LINEAR ANALYSIS, AND MULTIPLE CORRESPONDENCE ANALYSIS

Svetlana V. ZHUCHKOVA[1] — Master Student at the Faculty of Computer Science
E-MAIL: lana_lob@mail.ru
https://orcid.org/0000-0002-4425-725X

[1] National Research University Higher School of Economics, Moscow, Russia

РОТМИСТРОВ Алексей Николаевич — кандидат социологических наук, доцент, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: alexey.n.rotmistrov@gmail.com
http://orcid.org/0000-0003-2386-8710

Аннотация. В работе затрагивается проблема отсутствия разработанных концепций анализа многомерных связей между категориальными признаками при том, что такие признаки и многомерные связи между ними довольно распространены в социологических исследованиях. Об этом свидетельствует ряд методологических работ, в которых делается вывод о необходимости анализа многомерных связей, а не только парных, поскольку многомерные связи не сводятся к набору парных связей. Тем не менее опыт изучения многомерных связей между категориальными признаками в социологии остается довольно ограничен и практически отсутствует его теоретическое обобщение. Настоящим исследованием авторы попытались восполнить этот пробел через сравнение трех методов, подходящих для поиска многомерной связи между категориальными признаками: дерева решений CHAID, логлинейного анализа и множественного анализа соответствий. Сравнение методов происходило на теоретическом и эмпирическом уровнях. Содержательной задачей эмпирического этапа выступило составление портрета типичного представителя электората различных российских политических партий на основе базы восьмой волны Европейского социального исследования, проведенного в 2016 г., и социологического теоретико-методологического подхода к

Alexey N. ROTMISTROV[1] — Cand. Sci. (Soc.), Associate Professor
E-MAIL: alexey.n.rotmistrov@gmail.com
http://orcid.org/0000-0003-2386-8710

[1] National Research University Higher School of Economics, Moscow, Russia

Abstract. The paper addressed to the problem of the elaborated concepts shortage which deals with the analysis of multivariate associations among categorical variables. Meanwhile, such associations are rather common in sociological research what is argued by a corpus of methodological works. In them, it is grounded the necessity of the analysis of multivariate associations among categorical variables. Nevertheless, sociological experience in such an analysis is pretty poor as well as its theoretical generalization. In this study, we have tried to fill this gap by comparing the three methods: CHAID, log-linear analysis, and multiple correspondence analysis. The methods were compared at both theoretical and empirical levels. The empirical objective was to create a portrait of various Russian political parties' electorate using the data of the European social research conducted in 2016. By bringing the results of the application of methods to the form of categories combinations and by formulating numerical criteria for the comparison, the study allowed to identify the most effective method in two types of analytical tasks: description and forecasting. According to the results of the study, multiple correspondence analysis was the most effective in descriptive tasks, and log-linear analysis was the most effective in forecasting. The latter conclusion contradicts the currently predominating opinion regarding the CHAID's efficiency in cases when a

изучению электорального поведения. Результаты применения этих методов приведены к форме комбинаций категорий; введены числовые критерии сравнения, благодаря чему выделен наиболее эффективный метод в двух типах аналитических задач: описании и прогнозировании. Согласно результатам исследования, в описательных задачах наиболее эффективен множественный анализ соответствий, а в задачах прогноза — логлинейный анализ. Последний вывод противоречит сложившемуся мнению о преимуществе CHAID в случаях наличия в данных какого-либо целевого признака и в связи с этим обладает высокой практической значимостью для дальнейшего развития идеи построения высокоточных прогностических моделей в социологических исследованиях.

Ключевые слова: категориальные переменные, многомерная связь, прогностические модели, электоральное поведение, эффекты взаимодействия

Благодарность. Публикация подготовлена в ходе проведения исследования «Обоснование преимуществ поиска эффектов взаимодействия и их учета в социологических регрессионных моделях» (№18-05-0031) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018 г. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

target variable is presented in data and, therefore, has high practical significance for further development of the idea of building high-precision predictive models in sociological research.

Keywords: categorical variables, electoral behavior, interaction effects, multivariate associations, predictive models

Acknowledgments. The publication has been prepared during the study "Substantiation of the benefits of searching for interaction effects and their consideration in sociological regression models" (№18-05-0031) under "HSE Academic Fund Programme" in 2018 and under the state support programme of leading universities of the Russian Federation "5-100".

Введение

Использование статистическихметодов характерно для многих областей науки, однако именно в социологических исследованиях, оперирующих данными массовых опросов, преобладают категориальные переменные, которые накладывают ограничения на применение привычных статистических методов. Категориальные переменные — это признаки, измеренные на уровне номинальных или порядковых шкал. Упомянутое ограничение имеет разные проявления: во-первых, к категориальным переменным исходно не применимы многие методы статистического анализа (например, классическая линейная регрессия), во-вторых, в большинстве случаев даже при использовании адаптированных вариантов классических методов (например, регрессии с фиктивными переменными) анализ категориальных признаков заканчивается рассмотрением двумерных связей между ними. Тем не менее методы, позволяющие обнаружить многомерную связь категориальных признаков, существуют, и именно в социологических исследованиях актуальность их использования особенно велика. Эта актуальность обосновывается как теоретическими, так и эмпирическими предпосылками.

Так, впервые эту актуальность обосновали авторы первого алгоритма автоматического поиска взаимодействий, заявив о распространенности в данных массовых опросов интеркорреляций и взаимодействий — двух разновидностей многомерной связи признаков [Morgan, Sonquist, 1963]. При интеркорреляции, или совместной изменчивости (встречаемости) признаков, становится невозможно «разделить» их и определить вклад каждого при моделировании какого-либо явления. Взаимодействие — иной случай проявления многомерной связи — возникает, когда вклад каких-либо предикторов (независимых переменных) в отклик (зависимую, целевую переменную) опосредуется другой переменной, изменяется при ее введении. В обоих случаях речь идет об учете трех и более переменных сразу. Обозначив множество примеров интеркорреляций и взаимодействий в эмпирических исследованиях, авторы связали распространенность многомерных связей с понятием «латентного признака»: зачастую необходимые в социальных науках признаки невозможно измерить напрямую, а наблюдаемые индикаторы представляют собой лишь индикаторы латентных переменных — тогда именно совместное рассмотрение индикаторов и позволяет изучить необходимый теоретический конструкт.

Кроме того, необходимость рассмотрения именно многомерных связей имеет свое обоснование и в самой статистике: при анализе агрегированных, двумерных данных исследователь рискует столкнуться с парадоксом Симпсона — явлением, при котором связь между признаками «исчезает» или меняет свой характер на противоположный, если в анализе не учтены важные переменные [Simpson, 1951]. Примеры проявления парадокса можно встретить во множестве исследований из области демографии, образования, психологии, социологии, экономики [Kievit et al., 2013; Goltz, Smith, 2010; Fabris, Freitas, 2006]. Существование парадокса Симпсона свидетельствует о том, что отказ от рассмотрения многомерной связи может обернуться неверными или неполными выводами, а следовательно, и смещенными представлениями об изучаемой социальной реальности. Однако несмотря на обозначенные аргументы в пользу анализа многомерной связи кате-

гориальных признаков, использование соответствующих методов в эмпирических исследованиях — скорее исключение, чем правило. Отчасти такое противоречие может быть объяснено отсутствием универсальной концепции многомерного анализа категориальных данных (известные методы существенно различаются своими алгоритмами и реализацией, поэтому только аналитик, владеющий обширным арсеналом методов поиска многомерных связей, может ставить перед собой задачи их поиска), а также тем, что возможности и ограничения существующих методов в подобных задачах остаются не изучены и не обобщены.

Целью настоящего исследования стало комплексное сравнение трех методов поиска многомерной связи категориальных признаков: дерева решений CHAID, логлинейного анализа и множественного анализа соответствий. Выбранные методы хотя и основаны на расчете одинакового статистического критерия (хи-квадрата), однако значительно различаются алгоритмом поиска связи внутри таблицы сопряженности. Приведение результатов применения методов к форме комбинаций категорий и введение числовых критериев сравнения позволило выделить наиболее эффективный метод в двух типах аналитических задач: описании и прогнозировании. Содержательной задачей в исследовании выступило составление портрета типичного представителя электората различных политических партий по российским данным восьмой волны Европейского социального исследования 2016 г. и его интерпретация в категориях социологического анализа электорального поведения.

Описание и теоретическое сравнение методов

Для исследования были отобраны три наиболее доступных метода поиска многомерной связи категориальных признаков: дерево решений CHAID, логлинейный анализ и множественный анализ соответствий. Основанием для выбора методов стал заложенный в них статистический критерий — хи-квадрат, универсальный инструмент для статистического сравнения нескольких распределений и анализа таблиц сопряженности. Каждый из методов включает в себя расчет указанного критерия, однако задачи, выполняемые ими, вид получаемых результатов и области применения методов существенно различаются. Так, множественный анализ соответствий используется для визуализации «похожести» категорий изучаемых признаков и часто встречается в маркетинговых исследованиях; логлинейный анализ используется для моделирования частот в ячейках таблицы сопряженности, позволяет получить результат в виде соответствующего уравнения с откликом-частотой и распространен главным образом в эконометрике; CHAID, наиболее универсальный из трех методов, используется для множества задач (кластеризации, классификации, регрессии, поиска взаимодействий) и применяется в различных областях социальных и компьютерных наук. В основе всех перечисленных различий находятся разные алгоритмы поиска связи, реализуемые внутри каждого метода. Рассмотрим подробнее, что они из себя представляют.

Дерево решений CHAID

Принцип любого алгоритма построения деревьев решений — разбить имеющуюся выборку на группы по какому-либо основанию. В CHAID таким основани-

ем становится связь предикторов и отклика, и алгоритм нацелен на максимизацию этой связи через отбор предикторов с наибольшим значением статистики хи-квадрат.

На первом этапе работы алгоритма рассматриваются двумерные таблицы сопряженности между откликом и каждым предиктором. В каждой таблице вычисляется значение хи-квадрата, и предиктор, для которого в паре с откликом это значение оказывается наибольшим[1], становится первым из разделяющих, образует первый ярус дерева. Следующей задачей для машины становится объединение в узел таких категорий этого предиктора, связь которых с откликом одинакова (статистически значимо не различается). Математически это происходит через расчет того же хи-квадрата в таблицах, где представлены все категории отклика и отдельные пары категорий предиктора (если в такой таблице статистика не значима — рассматриваемая пара категорий объединяется). Причем в случае, когда предиктор измерен на уровне номинальной шкалы, объединяться категории могут в любом порядке, если же речь идет о переменных порядкового типа и выше — объединяться могут только соседние категории. Наконец, после объединения необходимых категорий рассматриваемого предиктора алгоритм «спускается» на ярус ниже — переходит к повторению всех описанных этапов уже в каждом из полученных узлов: снова рассматривает все возможные предикторы и их связь с откликом, выбирает из них наиболее сильно связанный, при необходимости объединяет категории этого предиктора в узлы. На этом этапе можно говорить о рассмотрении именно многомерной связи, поскольку теперь учитывается не только текущий предиктор, но и предиктор, который был выделен ранее — иными словами, теперь рассматривается условное распределение отклика. Описанные шаги повторяются в новых узлах, и построение дерева заканчивается либо естественным образом — в случае, когда значимо связанных с откликом предикторов не остается в наборе данных, либо по воле исследователя — когда достигнуты заданные глубина дерева (число ярусов) или минимальный объем наблюдений в узле. Итогом анализа становится схема в виде дерева.

Узлы, из которых не исходят другие узлы, называются терминальными, и именно они представляют наибольший интерес для исследователя: эти узлы содержат информацию о полученных группах наблюдений, включая признаки, по которым эти группы можно описать, и распределение отклика внутри этих групп, описывающее характер связи между ним и выделенными признаками. Подробное описание алгоритма CHAID можно найти в [Ritschard, 2013].

Логлинейный анализ

Задача алгоритма в логлинейном анализе — смоделировать частоты в многомерной таблице сопряженности через частоты отдельных категорий признаков и комбинаций этих категорий. Так, «на вход» метода подается таблица профилей — всех возможных комбинаций категорий переменных, в которой также ука-

[1] Строго говоря, учитывается не абсолютное значение хи-квадрата, а значение p-value — величины, нормирующей хи-квадрат на количество категорий и позволяющей таким образом сравнивать переменные с разным их числом. Разделяющей становится переменная с наименьшим p-value (что часто, но не всегда соответствует наибольшему хи-квадрату).

зываются частоты, с которыми эти комбинации встречаются. Вопрос, который стоит за процедурой проведения логлинейного анализа, заключается в следующем: почему конкретная комбинация признаков встречается именно столько раз? Обусловлено ли это только особенностями формирования выборки или же конкретным категориям действительно свойственно «встречаться вместе» (или, напротив, не свойственно)? Логлинейный анализ способен разграничивать эти ситуации — определять, какой вклад в формирование частоты некой комбинации вносят частоты составляющих эту комбинацию категорий по отдельности, попарно, по трое и т. д. Каждый такой вклад называется эффектом. Кроме того, выясняется направление эффектов: каким категориям свойственно встречаться вместе, а какие как бы отталкиваются друг от друга.

Каждый эффект представляет собой числовое значение (коэффициент уравнения), соответствующее параметру распределения Пуассона и вычисляемое по специальным формулам после логарифмирования моделируемых частот (подробнее см. [Vermunt, 2005]). При этом расчет параметров для построения модели происходит в строгой последовательности: вначале рассчитываются одномерные эффекты, затем на их основании рассчитываются двумерные, далее с учетом одномерных и двумерных эффектов рассчитываются трехмерные и т. д. — таким образом, каждый эффект уже «учитывает» вклады, которые приходятся на эффекты более низких уровней, и «объясняет» только то отклонение, которое не было «объяснено» предыдущими эффектами. С помощью расчета стандартной ошибки полученные значения эффектов можно ранжировать по силе, разграничив при этом статистически значимые и незначимые эффекты. Наконец, значимые эффекты интерпретируются содержательно: главные, или одномерные, эффекты идентифицируют смещения в распределении исходных переменных, т. е. указывают на особенности выборки, а не на связь признаков, а эффекты второго и выше уровней содержат информацию о двумерной или многомерной связи конкретных категорий переменных соответственно. При этом каждый эффект виртуально сравнивает частоту с базовой частотой, которая находилась бы в ячейке таблицы в случае отсутствия связи между категориями признаков (или в случае отсутствия значимых смещений в распределении — если речь идет об одномерных эффектах).

Статистика хи-квадрат в логлинейном анализе используется для предварительной процедуры отбора нужной спецификации модели. Модель, в которую включены все возможные эффекты (насыщенная модель), на практике оказывается мало полезна, поскольку не учитывает, что некоторые колебания частот могут быть случайными, несущественными. Для восполнения этого недостатка перед расчетом параметров отбирается модель, которая содержит наименьшее число эффектов и при этом все еще не значимо отклоняется от имеющихся данных — и оценка значимости этого отклонения происходит через оценку значимости хи-квадрата. Так, после выбора нужной спецификации, машина моделирует уже не исходные эмпирические частоты, а так называемые ожидаемые — частоты, «очищенные» от лишних колебаний, которые не объясняются никакими эффектами. Полная пошаговая схема применения логлинейного анализа представлена, например, в [Tabachnick, Fidell, 2012].

Множественный анализ соответствий

Суть анализа соответствий состоит в расположении категорий на карте соответствий: похожих близко, а непохожих — далеко. «Похожесть» определяется как сходство условных распределений признаков. При этом в методе предполагается процесс снижения размерности — переход от исходного числа анализируемых категорий переменных к меньшему числу агрегированных измерений-осей, в связи с чем метод часто определяют как аналог метода главных компонент, применяемый к категориальным переменным [Franco, 2016].

Множественный анализ соответствий ориентирован на особый тип таблицы сопряженности — матрицу Берта [2] [Greenacre, Blasius, 2006]. В ней по строкам и столбцам располагаются категории всех анализируемых переменных — так, что матрица оказывается симметричной, в итоге включает в себя несколько одинаковых таблиц сопряженности, а по диагонали у нее располагаются безусловные частоты категорий. К этой таблице далее применяется «простой» анализ соответствий, и на первом шаге рассчитывается инерция — нормированная на сумму частот в матрице величина хи-квадрата.

В дальнейшем алгоритм использует не абсолютные частоты из матрицы, а относительные частоты, которые формируют профили строк и столбцов. Профиль — это вектор, или набор значений относительных частот каждой строки и каждого столбца матрицы. Средний профиль по строке и столбцу отражает «центр тяжести» — начало координат будущего пространства. Кроме самих профилей, в построении модели задействован и показатель массы строк и столбцов матрицы, который вычисляется как доля соответствующей категории относительно общей частоты.

После получения профилей строк и столбцов матрицы, а также их масс, алгоритм переходит к расчету расстояния хи-квадрат между профилями. Эта процедура похожа на привычный расчет хи-квадрата в таблице сопряженности с тем отличием, что вместо абсолютных частот используются относительные (образующие профили), наблюдаемыми частотами считаются профили одной строки (столбца), ожидаемыми — профили другой строки (столбца), а поправка идет на средний профиль строки или столбца. С помощью полученных расстояний можно альтернативным способом представить инерцию, и тогда она обретает физический смысл — показывает, насколько в целом точки далеки от имеющегося центра тяжести, т. е. насколько профили строчных и столбцовых категорий отличаются от среднего профиля по строкам и столбцам соответственно. Затем с помощью процедуры сингулярного разложения инерции находится некоторое пространство, оптимальным образом описывающее точки, используемые в матрице. В ходе этого процесса точкам задаются координаты в найденных осях, согласованные с ранее полученным расстоянием, а также рассчитываются показатели качества модели — доля объясненной инерции и разнообразные вклады осей и точек в инерцию, которые затем используются для интерпретации найденных связей. Наконец, с помощью рассчитанных координат строится искомая карта. Формально-математическая составляющая всего описанного процесса представлена в [Greenacre, 2007].

[2] Существует также реализация множественного анализа соответствий с использованием другого вида таблицы — индикаторной матрицы. Однако этот вариант метода рассчитан на анализ связи объектов, а не категорий, поэтому в этой статье не рассматривается.

Теоретическое сравнение методов

В каждом из методов используется существенно отличный от остальных алгоритм поиска связи, однако можно обнаружить и основания для их сравнения. Как уже было отмечено, главное общее основание заключается в анализе таблиц сопряженности, основанном на расчете статистики хи-квадрат. Но принципы анализа этих таблиц посредством статистики хи-квадрат в каждом методе свои. Так, в CHAID анализируются привычные двумерные таблицы сопряженности, но соответствующие конкретным условным распределениям, множественный анализ соответствий на входе имеет матрицу Берта, объединяющую несколько двумерных таблиц сопряженности, и лишь логлинейный анализ работает с полной многомерной таблицей. Чтобы детально рассмотреть различие этих принципов, мы предлагаем четыре критерия теоретического сравнения методов. Различие именно по этим критериям, на наш взгляд, ведет к получению неодинаковых результатов при применении рассматриваемых методов к одним и тем же данным.

1. **Понимание самой связи — формальное и содержательное**. Формально в CHAID многомерная связь обнаруживается в терминальных узлах дерева, т. е. в тех условных распределениях, в которых статистика хи-квадрат между откликом и предиктором оказывается значимой. Содержательно эта многомерная связь интерпретируется как «склонность» категорий встречаться совместно чаще или реже, чем в случае отсутствия связи между ними. Аналогичным образом можно интерпретировать и многомерную связь, найденную с помощью логлинейного анализа, но формально она выявляется по-другому — через расчет параметров распределения Пуассона. Во множественном анализе соответствий связанными оказываются те категории, расстояние хи-квадрат между профилями которых минимально. Поскольку профиль в контексте множественного анализа соответствий — это относительные частоты какой-либо категории «в разрезе» категорий других признаков, то содержательно близость координат двух категорий означает сходство их условных распределений.
2. **Способность разделять переменные по уровню измерения и роли в анализе**. С этой точки зрения «выделяется» CHAID: он требует обязательного наличия в данных какого-либо отклика, а также адаптирует используемые статистики под разные типы шкал отклика и изменяет алгоритм объединения категорий предиктора в узлы в зависимости от типов шкал предикторов. В отличие от CHAID, оставшиеся методы рассматривают все переменные «на одном уровне», не разделяя на отклик и предикторы, и анализируют их исключительно как номинальные. Однако это нельзя назвать недостатком этих методов: и в логлинейном анализе, и во множественном анализе соответствий исследователь может ограничиться интерпретацией только тех многомерных комбинаций, которые включают в себя интересующий отклик. Тем не менее практики использования всех трех методов сложились таким образом, что CHAID (как и прочие модели деревьев решений) традиционно противопоставляется другим методам поиска многомерной связи как метод, который наилучшим образом подходит в ситуации наличия в данных некоторого отклика [Agresti, 2002; Han, Kamber, Pei, 2012; Tabachnick, Fidell,

2012; Ratner, 2017]. Более того, практика поиска дополнительных многомерных предикторов для предсказательных моделей распространяется только на различные алгоритмы деревьев решений (такой подход глубоко распространен в компьютерных науках в рамках так называемой feature engineering — работы по отбору признаков [Ratner, 2017], однако почти не применяется в социологии). По указанным причинам рассматриваемые методы в целом крайне редко сравниваются друг с другом.

3. **Возможность производить статистический вывод о связи признаков и их категорий (показатели частного качества модели)**. Статистические выводы из сравниваемых методов различаются по степени «глубины»: CHAID и логлинейный анализ позволяют оценивать значимость выявленных связей наиболее «глубоко» — через оценку значимости связи не только отдельных переменных из всего набора, но и конкретных комбинаций их категорий. В CHAID это реализуется с помощью оценки значимости стандартизованных остатков, а в логлинейном анализе — значимости параметров модели (эффектов) или стандартизованных остатков. Благодаря этому исследователь получает возможность не только выявлять значимые комбинации, но и ранжировать их по силе связи. В отличие от этих двух методов, множественный анализ соответствий позволяет статистически оценить наличие связи во всем наборе переменных в целом — и только ее — через оценку значимости хи-квадрата в матрице Берта. С одной стороны, это открывает более широкие возможности для поиска многомерных комбинаций, с другой — затрудняет анализ и интерпретацию в связи с тем, что у исследователя не имеется какого-либо индикатора для выбора наиболее существенных из них.
4. **Показатели общего качества модели**. В CHAID как методе классификации показателем общего качества модели оказывается процент правильных предсказаний категорий отклика (и прочие аналогичные метрики). Однако значения подобных показателей не всегда непосредственно соотносятся с результатом применения CHAID как метода поиска взаимодействий. В логлинейном анализе о приемлемом качестве модели свидетельствует незначимое отклонение смоделированных частот от эмпирических. Наконец, наиболее легко интерпретируемым показателем общего качества модели обладает множественный анализ соответствий — в нем в этой роли выступает доля объясненной инерции. Так как инерция в этом методе — это нормированная величина хи-квадрата в матрице Берта, то указанная метрика в самом явном виде показывает, насколько «успешно» удалось эту связь описать.

Методология исследования

Как было отмечено, отсутствие единой концепции анализа многомерной связи категориальных признаков, предположительно, служит одной из причин слабой распространенности подобного анализа в социологических исследованиях — хотя именно для этой предметной области такие связи наиболее характерны. Целью исследования стало сравнение возможностей выбранных методов в решении двух

широких типов аналитических задач: описании и прогнозировании. Представим поэтапно процедуру проведения этого сравнения.

1. Выбор и подготовка переменных для анализа. Поскольку генерирование многомерных связей должно опираться на конкретный алгоритм поиска этой связи (а как мы показали выше, алгоритмы в трех сравниваемых методах существенно различаются), в своем исследовании мы используем реальные, а не сгенерированные данные. Выбор же самого предмета для иллюстрации сравнения методов ограничивается двумя обстоятельствами: во-первых, согласно требованию CHAID, в данных должен присутствовать какой-либо отклик; во-вторых, обращаться следует к такому феномену, для которого существуют теоретические основания к поиску многомерной связи. Таким феноменом в настоящей работе стало электоральное поведение, в теоретических и эмпирических работах по изучению которого содержится множество указаний и примеров наличия такой связи. Откликом стало голосование респондента за определенную партию на выборах в Государственную думу 2016 г., а набор предикторов составили переменные, соответствующие так называемому социологическому теоретико-методологическому подходу к изучению электорального поведения [Мелешкина, 2001]. Согласно этому подходу, выбор избирателя объясняется его принадлежностью к некоторой социальной группе и выражением солидарности с ней (цит. по: [Голосов, 1997]), а в эмпирических исследованиях предикторами выступают различные социально-демографические, профессиональные и экономические характеристики респондентов [Страхов, 2000]. В силу естественных ограничений на возможное число используемых переменных, в настоящем исследовании предикторами послужили шесть признаков: пол, возраст, уровень образования респондента, тип населенного пункта, в котором он проживает, характер его трудовых отношений и самооценка дохода его семьи. Для того чтобы избежать искажений, связанных с наличием нулевых частот в ячейках многомерной таблицы сопряженности, было введено следующее требование к подготовке выбранных переменных: при наличии больших смещений в распределении (если доли каких-либо категорий переменных составляли меньше 5 % от выборки), схожие по смыслу категории этих переменных должны были быть сгруппированы таким образом, чтобы все итоговые категории были представлены не менее чем 5 % наблюдений и при этом итоговая многомерная таблица содержала не более 6000 ячеек (такое число связано с техническими ограничениями, определяющими возможность расчета параметров модели в логлинейном анализе и время этого расчета). Переменные в итоговом виде представлены в табл. 1. Источником данных стала база Европейского социального исследования 2016 г., содержащая все необходимые переменные[3].

[3] ESS Round 8: European Social Survey Round 8 Data. (2016) Data file edition 2.0. NSD — Norwegian Centre for Research Data, Norway — Data Archive and distributor of ESS data for ESS ERIC.

Таблица 1. ***Информация об используемых переменных***

Переменная	**Итоговые категории**
За какую партию Вы проголосовали на выборах в Госдуму в сентябре 2016 г.?	«Единая Россия» КПРФ «Справедливая Россия» ЛДПР Другие*
Пол	Мужской Женский
Возраст	=< 35 лет 36—54 лет >= 55 лет
Тип населенного пункта	Большой город или его окраина Небольшой город или поселок городского типа Сельская местность, ферма, хутор
Уровень образования	Основное общее и ниже Среднее общее Начальное профессиональное Среднее профессиональное Высшее профессиональное или ученая степень
Характер трудовых отношений	Наемный работник Работает не по найму Не работает
Самооценка уровня дохода семьи	Живем на этот доход, не испытывая материальных затруднений Этого дохода нам в принципе хватает Жить на такой доход довольно трудно Жить на такой доход очень трудно

* К этой категории были отнесены партии, не получившие в 2016 г. мест в Госдуме: «Родина», «Гражданская сила», «Коммунисты России», «Яблоко», Российская партия пенсионеров за справедливость, «Патриоты России», «Зеленые», «Гражданская платформа», Партия народной свободы (Парнас), Партия роста.

2. Применение методов и получение первичных результатов. Стратегии применения выбранных методов могут сильно варьироваться, поскольку в каждом из них предусмотрена процедура отбора конкретной спецификации модели — и выбор ее итогового варианта так или иначе зависит от самого исследователя. В настоящем исследовании использовались такие стратегии, которые потенциально позволяют обнаружить максимально возможное число многомерных связей. Так, при применении логлинейного анализа были учтены результаты построения как насыщенной, так и редуцированных моделей [Tabachnick, Fidell, 2012]. При использовании CHAID применялись методики, направленные на получение максимально разветвленного, но при этом устойчивого дерева: это исключение пропущенных значений [Жучкова, Ротмистров, 2018], работа с поправкой Бонферонни [Ritschard, 2013], проведение проверки на непереобученность с помощью процедуры кросс-валидации. При реализации множественного анализа соответствий была выбрана модель, объясняющая не менее половины исходной инерции, а для интерпретации самих осей и положения точек в них было снижено пороговое значение показателя корреляции точки с осью — им стала средняя корреляция для всех категорий переменной о политических партиях по всем осям.

3. Приведение полученных результатов к единой форме. В силу описанных ранее отличий методов для их последующего сравнения мы предлагаем процедуру «стандартизации» их результатов за счет приведения этих результатов к форме комбинаций категорий признаков. Таким образом, ***многомерная связь категориальных признаков*** определяется нами как *комбинация категорий трех и более переменных, которым свойственно или не свойственно «встречаться вместе» по результатам применения конкретного метода*. Для CHAID и логлинейного анализа это означает, что категориям свойственно совместно наблюдаться чаще или реже, чем в случае отсутствия связи, а для множественного анализа соответствий — что категориям свойственно или не свойственно иметь высокую корреляцию с полюсами оси. Поскольку в явном виде такие комбинации можно обнаружить только в логлинейном анализе, в двух других методах мы прибегаем к дополнительным процедурам для их выявления: в CHAID производится расчет стандартизованных остатков хи-квадрата в терминальных узлах дерева и отбор только значимых из них, а во множественном анализе соответствий — расчет уже упомянутых аналогов факторных нагрузок (показателей корреляции точки с осью) по алгоритму, предложенному в [Шафир, 2011], отбор наиболее тесных связей и объединение их в комбинации категорий.

Подобное приведение результатов применения методов к форме комбинаций категорий позволяет операционализировать и понятия «описания» и «прогнозирования» в контексте использования выбранных методов. Так, под описанием понимается способность метода выделять максимально «насыщенные», многомерные комбинации, т. е. такие, с помощью которых можно наиболее полно описать какую-либо группу наблюдений и ее содержательно проинтерпретировать. Под прогнозированием понимается способность метода выделять такие комбинации признаков, с помощью которых можно с наибольшей точностью предсказывать категории одной из анализируемых переменных. На практике для последней задачи в абсолютном большинстве случаев используется CHAID, а два других метода даже не рассматриваются — в силу того, что эти методы не способны разделять переменные на отклик и предикторы. Однако, как уже было отмечено, это обстоятельство не является ограничением: в каждом из методов можно рассматривать только те комбинации, которые содержат категории гипотетического отклика.

Исходя из проведенного теоретического сравнения методов, мы выдвинули две основные гипотезы об эффективности тех или иных методов в решении двух упомянутых задач:

*(1) С точки зрения **описания** наиболее «насыщенные» результаты будет давать логлинейный анализ, поскольку этот метод исходно анализирует все возможные комбинации категорий переменных на всех уровнях. В отличие от него CHAID всегда ограничен первой разделяющей дерево переменной (следовательно, анализирует не все возможные комбинации категорий), а множественный анализ соответствий в целом не содержит эксплицитных указаний на значимость эффектов тех или иных комбинаций.*

*(2) С точки зрения **прогнозирования** лучшие результаты покажет CHAID, поскольку этот метод исходно нацелен на поиск взаимодействий (а не интеркорреляций, как два других), разграничивая роли отклика и предикторов. Иными словами,*

поскольку CHAID действует в контексте конкретного отклика, этот метод исходно ориентирован именно на его предсказание.

4. Расчет числовых критериев сравнения. Для проверки гипотез используются следующие числовые критерии сравнения: при описании — число и размерность получаемых комбинаций признаков, где размерность — это среднее число переменных, участвующих в найденных комбинациях, при прогнозировании — значение псевдокоэффициента детерминации (псевдо-R^2) во вспомогательных моделях мультиномиальной логистической регрессии, в которой откликом выступает переменная о выборе определенной политической партии, а предикторами — найденные с помощью каждого метода комбинации[4]. Кроме того, поскольку методы не имеют единого показателя качества модели, в своем исследовании мы вводим дополнительный числовой критерий — это доля наблюдений из выборки, описанных найденными комбинациями. В отличие от указанных ранее критериев, этот показатель позволяет в целом оценить «охват» найденной многомерной связи, а не ее отдельные свойства. Эталоном по этому критерию выступает CHAID — метод, для которого такая доля исходно равна 100%. Использование всех указанных критериев, таким образом, позволяет в полной мере охарактеризовать найденную многомерную связь и в удобном виде сравнить используемые методы.

Результаты исследования

В целом попытку рассмотреть портрет российских политических партий в «многомерном» разрезе можно признать успешной: в данных, описывающих российские выборы 2016 г., действительно удалось обнаружить множество комбинаций признаков (65 комбинаций, описывающих группы, которым более или менее свойственно голосовать за определенные партии), однако описательные возможности методов существенно различаются.

Таблица 2. ***Число и размерность найденных комбинаций признаков***

Метод	Логлинейный анализ	CHAID	Множественный анализ соответствий
Число комбинаций	23	14	28
Размерность комбинаций	2,1	2,9	3,7
95% доверительный интервал для размерности	1,96—2,24	2,54—3,26	3,31—4,09

Вернемся к первой гипотезе: «*С точки зрения* ***описания*** *наиболее «насыщенные» результаты будет давать логлинейный анализ…*». Результаты применения методов дают основания отвергнуть эту гипотезу (см. табл. 2). Так, объективно среди всех методов наиболее «насыщенные» результаты были получены с помощью ***множественного анализа соответствий***: во-первых, с его помощью было выделено наибольшее число комбинаций-эффектов; во-вторых, размерность найденных

[4] При этом во избежание смещения оценок регрессионных коэффициентов из моделей были удалены комбинации, охватывающие менее 30 наблюдений, а также тесно скоррелированные сочетания (с коэффициентом корреляции Пирсона выше 0,7 по модулю).

комбинаций для множественного анализа соответствий статистически значимо превышает таковую для двух других методов. Таким образом, множественный анализ соответствий позволяет получить и более глубокую интерпретацию существующих связей. Однако следует отметить, что с этими результатами связаны как минимум три ограничения.

Во-первых, размерность выявляемых в рамках множественного анализа соответствий комбинаций полностью зависит от тех пороговых значений, которые выставляет исследователь при интерпретации аналогов факторных нагрузок: в нашем случае порог был достаточно низким (0,242 — средняя нагрузка для категорий переменных о партии), что и позволило выделить пятимерные (а учитывая также сами категории отклика — шестимерные) связи.

Во-вторых, чем более многомерными оказываются выявленные комбинации, тем труднее становится с их помощью описать весь имеющийся объем выборки, т. е. обнаружить респондентов, которые сочетали бы в себе все выявленные признаки. Это иллюстрируют и полученные результаты: комбинации, выявленные с помощью множественного анализа соответствий, описывают наименьшую долю наблюдений в выборке — 51% по сравнению с 92% наблюдений, описанных эффектами из логлинейного анализа, и 100% наблюдений, описанных терминальными узлами CHAID и принятых исходно за эталон (см. табл. 3).

Таблица 3. ***Доля наблюдений, описанных найденными комбинациями***

Метод	Логлинейный анализ	CHAID	Множественный анализ соответствий
Доля наблюдений	92%	100%	51%
95% доверительный интервал для доли	91—93%	100%	49—53%

Наконец, в-третьих, как отмечено и в самой гипотезе, несмотря на содержательную насыщенность результатов множественного анализа соответствий, степень «уверенности» в них остается неопределенной, поскольку в методе не предусмотрено построение статистического вывода о значимости найденных эффектов. Однако это ограничение компенсируется за счет второго этапа сравнения — оценки прогностической способности найденных эффектов с помощью мультиномиальной логистической регрессии.

Для проверки второй гипотезы, посвященной этому этапу: «*с точки зрения* ***прогнозирования*** *лучшие результаты покажет CHAID…*» — были построены четыре модели мультиномиальной логистической регрессии с откликом про политические партии (опорной категорией была выбрана «Единая Россия»): одна модель (далее — исходная), которая в качестве предикторов содержит исходные переменные в виде наборов фиктивных переменных, и три модели, в каждой из которых предикторами выступают найденные с помощью выбранных методов комбинации.

Первая модель, содержащая только главные эффекты, использовалась, чтобы в целом оценить разницу между прогностической способностью одномерных

и многомерных эффектов — иными словами, чтобы понять, существует ли вообще потенциал использования именно многомерных связей для предсказания рассматриваемого отклика. Как было упомянуто, в качестве меры для сравнения результатов были использованы показатели псевдо-R^2: они, в отличие от процента правильных предсказаний, учитывают неокругленные значения исходных и предсказанных моделью вероятностей (а следовательно, точнее иллюстрируют прогностическую способность предикторов), а также оказываются менее чувствительны к смещениям в распределении отклика [Field, 2009]. Рассчитанные показатели приведены в табл. 4.

Таблица 4. ***Показатели псевдо-R 2***

Псевдо-R^2	**Исходная модель**	**Модель с эффектами**		
		MCA	**CHAID**	**LLA**
Кокса и Снелла	0,124	0,053	0,103	0,181
Нагелькерке	0,139	0,06	0,115	0,203
95 % доверительный интервал для псевдо-R^2 Нагелькерке	0,11—0,153	0,023—0,071	0,092—0,144	0,175—0,217
Макфаддена	0,059	0,024	0,049	0,089

Эти показатели были дополнительно оценены статистически с помощью расчета доверительных интервалов через процедуру bootstrap (для псевдо-R^2 Нагелькерке интервальная оценка представлена в табл. 4), и из результатов следует, что лишь одна из моделей — с эффектами, обнаруженными с помощью ***логлинейного анализа,*** — по своей прогностической способности превосходит модель с одномерными эффектами, и этой же модели в целом соответствуют наивысшие показатели псевдо-R^2.

Этот результат позволяет сделать несколько важных методологических выводов. Во-первых, вторая гипотеза, в рамках которой лучшие результаты приписывались методу CHAID, также ***не подтвердилась*** — и это ***противоречит*** сложившейся практике использования CHAID как метода, наиболее подходящего к случаю наличия в наборе переменных какого-либо отклика. Согласно полученным результатам, среди рассматриваемых методов логлинейный анализ, а не CHAID, смог выявить комбинации, наилучшим образом предсказывающие выбранный отклик. Предположительное объяснение такому «противоречию» может быть следующим: практика перехода после методов поиска многомерной связи к использованию выявленных комбинаций в регрессионном уравнении очень слабо распространена, в связи с чем реальный потенциал использования этого метода в задачах прогноза до сих пор мог быть не выявлен, а потому — и недооценен. Объяснение же самому факту «превосходства» логлинейного анализа отчасти уже было дано в первой гипотезе: CHAID всегда рассматривает ограниченное число комбинаций

из всех возможных — по той причине, что анализирует последовательно только выделенные условные распределения. В отличие от него логлинейный анализ рассматривает все возможные комбинации категорий переменных на всех уровнях, не ограничиваясь каким-то заранее выделенным «путем». Следовательно, логлинейный анализ потенциально может выявить больше значимых многомерных комбинаций.

Во-вторых, неуверенность в результатах применения множественного анализа соответствий в связи с отсутствием в методе процедуры статистического вывода оправдалась: хотя выявленные комбинации и получились наиболее «насыщенными» из всех, их предсказательная способность оказалась невысокой — по этому критерию метод показал худшие результаты. Связь категорий в выделенных с помощью этого метода комбинациях слабо распространяется на генеральную совокупность.

В-третьих, тот факт, что модель с эффектами из логлинейного анализа по своему качеству значимо превзошла, в том числе, исходную модель с одномерными эффектами, подтверждает потенциал использования многомерных связей в изучении электорального поведения.

Наконец, следует и содержательно проинтерпретировать полученные результаты. Так как в общей сложности было выявлено более 60 многомерных комбинаций, ограничимся интерпретацией лишь тех из них, которые в рамках каждого из трех методов оказались наиболее наполненными, значимыми в регрессии и при этом соответствовали бы как положительным, так и отрицательным значениям регрессионных коэффициентов. В контексте построенных регрессионных моделей, в которых опорной категорией выступила партия «Единая Россия», отрицательные коэффициенты будут повышать вероятность проголосовать за эту партию, а положительные — за некоторую другую.

Так, **повышает** вероятность проголосовать на выборах за «Единую Россию» принадлежность россиян к группам, которые образуются следующими комбинациями признаков:

— «наемный работник», «дохода в принципе хватает» (по результатам логлинейного анализа, 34 % выборки). Причем для таких респондентов особо нехарактерно голосовать за «Справедливую Россию» и ЛДПР;
— «женщина», «менее 55 лет», «работает», «не имеет высшего образования» (по результатам CHAID, 16 % выборки). Для таких респондентов особо не характерно голосовать за КПРФ;
— «женщина», «проживает в большом городе», «наемный работник» (по результатам множественного анализа соответствий, 20 % выборки). Причем для таких респондентов нехарактерно голосование за КПРФ и ЛДПР.

Напротив, **понижает** вероятность проголосовать на выборах за «Единую Россию» принадлежность россиян к группам, которые образуются следующими комбинациями признаков:

— «проживает в большом городе», «имеет высшее образование или ученую степень» (по результатам логлинейного анализа, 17 % выборки). Таким респондентам в большей степени свойственно голосовать за «Справедливую Россию» или за партии, которые в 2016 г. не получили мест в Госдуме;

— «мужчина», «старше 55 лет» (по результатам CHAID, 10 % выборки). Таким респондентам в большей степени свойственно голосовать за КПРФ или за партии, которые в 2016 г. не получили мест в Госдуме;
— «женщина», «старше 55 лет», «проживает в большом городе», «имеет высшее образование или ученую степень» (по результатам множественного анализа соответствий, 3 % выборки). Как и в предыдущей модели, таким респондентам в большей степени свойственно голосовать за КПРФ или за партии, которые в 2016 г. не получили мест в Госдуме.

Примечательно, что интерпретация даже этих шести эффектов позволяет сделать некоторые содержательные выводы. Так, в целом полученные комбинации соотносятся с двумя наиболее распространенными стереотипами о составе электората партий «Единая Россия» и КПРФ: первые три из перечисленных эффектов соответствует портрету работника бюджетных организаций, который может голосовать за правящую партию в силу действий административного ресурса, а последние два эффекта описывают людей старшего возраста, — каким предстает типичный избиратель КПРФ в глазах общественности. Тем не менее представленные в тексте эффекты — лишь десятая часть от всех полученных результатов, и для составления более детального портрета электората партий следует проинтерпретировать их все — такая задача, однако, находится за рамками настоящего исследования.

В целом эффекты, выделенные разными методами, не противоречат друг другу, а скорее дополняют, уточняют один другой. Конкретные вероятности проголосовать за ту или иную партию при принадлежности к указанным группам приведены в табл. 5.

Таблица 5. ***Выделенные комбинации признаков***

Метод	Логлинейный анализ	CHAID	Множественный анализ соответствий
Комбинация 1: наиболее наполненная значимая комбинация с отрицательным коэффициентом (*повышает вероятность проголосовать за «Единую Россию»*)	Наемный работник, дохода в принципе хватает	Женщина, менее 55 лет, работает, не имеет высшего образования	Женщина, проживает в большом городе, наемный работник
Доля наблюдений, описанных комбинацией 1	0,34	0,16	0,2
Вероятностьпроголосовать за «Единую Россию» при комбинации 1	71 %	77 %	72 %
Вероятность проголосовать за КПРФ при комбинации 1	18 %	4 %	7 %
Вероятность проголосовать за «Справедливую Россию» при комбинации 1	1 %	4 %	6 %

Метод	Логлинейный анализ	CHAID	Множественный анализ соответствий
Вероятность проголосовать за ЛДПР при комбинации 1	5%	9%	6%
Вероятность проголосовать за другие партии при комбинации 1	5%	5%	9%
Комбинация 2: наиболее наполненная значимая комбинация с положительным коэффициентом (*понижает вероятность проголосовать за «Единую Россию»*)	Проживает в большом городе, имеет высшее образование или ученую степень	Мужчина, старше 55 лет	Женщина, старше 55 лет, проживает в большом городе, имеет высшее образование или ученую степень
Доля наблюдений, описанных комбинацией 2	0,17	0,1	0,03
Вероятность проголосовать за «Единую Россию» при комбинации 2	50%	52%	40%
Вероятность проголосовать за КПРФ при комбинации 2	13%	28%	32%
Вероятность проголосовать за «Справедливую Россию» при комбинации 2	17%	3%	3%
Вероятность проголосовать за ЛДПР при комбинации 2	7%	6%	8%
Вероятность проголосовать за другие партии при комбинации 2	13%	11%	17%

Заключение

В настоящей работе затронута тема многомерных связей между категориальными признаками, часто встречающимися в социологических исследованиях. Несмотря на наличие различных аргументов в пользу анализа именно многомерных связей таких признаков, практика использования соответствующих методов слабо распространена отчасти потому, что единой концепции для подобного анализа не существует, как и не существует актуальных работ, обобщающих подходы к нему. Настоящим исследованием мы попытались восполнить этот пробел через сравнение трех методов, подходящих для решения задач поиска такой связи.

Гипотезы, поставленные в исследовании и предполагающие эффективность логлинейного анализа в описательных задачах и эффективность CHAID в задачах прогноза, не подтвердились. Так, наиболее эффективным с точки зрения описания оказался множественный анализ соответствий, а наиболее полезные для прогнозирования комбинации удалось выявить с помощью логлинейного анализа. И если с результатом проверки первой гипотезы связаны ограничения, обозначенные в тексте, то результаты проверки второй гипотезы оказались не только контринтуитивными, но и противоречащими сложившемуся мнению об эффективности

CHAID в случаях наличия в данных какого-либо отклика и практике использовать комбинации, полученные именно с помощью деревьев решений, как предикторы в будущей регрессионной модели. Было показано, что потенциал логлинейного анализа в этой задаче недооценен и метод способен выдавать результаты более точные, чем CHAID.

У проведенного исследования есть ограничения. Одно из ограничений было подробно описано при проверке первой гипотезы: результаты применения множественного анализа соответствий во многом зависят от конкретной исследовательской ситуации, в частности, от тех пороговых значений аналогов факторных нагрузок, на основе которых категории отбираются в комбинации (и впоследствии происходит интерпретация связей). Второе, более общее ограничение связано с тем, что в рамках исследования методы сравнивались (пусть и с проведением статистической оценки различий по всем показателям) только на одном эмпирическом примере, и нет оснований утверждать, что методы «поведут себя» так же на других данных. Тем не менее даже такое сравнение позволило добиться значительных результатов.

Для дальнейшего развития проведенного исследования видятся следующие направления:

— разработать критерии определения такого оптимального порога аналога факторной нагрузки во множественном анализе соответствий, который позволил бы привести в баланс размерность получаемых комбинаций с их точностью и полнотой охвата — для компенсации первого обозначенного выше ограничения,
— провести аналогичное исследование, но с применением методологии статистического эксперимента — для преодоления второго обозначенного выше ограничения,
— адаптировать логлинейный анализ для его эффективного использования в задачах по прогнозированию — для дальнейшего развития полученного в ходе исследования нетривиального вывода об эффективности указанного метода,
— применить рассматриваемые методы поиска многомерной связи к переменным, соответствующим иным теоретическим подходам к изучению электорального поведения — для дополнительной иллюстрации возможностей многомерного анализа категориальных данных.

Список литературы (References)

Голосов Г. В. Поведение избирателей в России: теоретические перспективы и результаты региональных выборов // Полис. 1997. № 4. С. 44—56.
Golosov G. V. (1997) Electoral Behaviour in Russia: Theoretical Prospects and the Results of the Regional Elections. *Polis. Political Studies*. No. 4. P. 44—56. (In Russ.)

Жучкова С. В., Ротмистров А. Н. Возможность работы с пропущенными данными при использовании CHAID: результаты статистического эксперимента // Социология: методология, методы, математическое моделирование. 2018. № 46. С. 85—122.

Zhuchkova S. V., Rotmistrov A. N. (2018) Handling Missing Data with CHAID: Results of a Statistical Experiment. *Sociology: Methodology, Methods, Mathematical Modeling*. No. 46. P. 85—122. (In Russ.)

Мелешкина Е. Ю. Исследования электорального поведения: Теоретические модели и проблемы их применения // Политическая наука. 2001. № 2. С. 187—212.
Meleshkina E. Yu. (2001) Studies of Electoral Behavior: Theoretical Models and Problems of Their Application. *Political Science*. No. 2. P. 187—212. (In Russ.)

Страхов А. П. Изучение электорального поведения россиян: социокультурный подход // Полис. 2000. № 3. С. 90—96.
Strakhov A. P. The Study of Russians' Electoral Behaviour: Socio-Cultural Approach. *Polis. Political Studies*. 2000. No. 3. P. 90—96. (In Russ.)

Шафир М. А. Новый способ интерпретации результатов анализа соответствий // Социологические методы в современной исследовательской практике: Сборник статей, посвященный памяти первого декана факультета социологии НИУ ВШЭ А. О. Крыштановского. М. : Издательский дом НИУ ВШЭ, 2011. С. 223—231.
Shafir M. A. (2011) A New Way to Interpret the Results of Correspondence Analysis. In: Sociological methods in contemporary research practice: Collection of articles in memoriam of the first Dean of the Faculty of Sociology at NRU HSE, A. O. Kryshtanovsky. Moscow: NRU HSE. P. 223—231.

Agresti A. (2002) Categorical Data Analysis (Second edition). New York: John Wiley & Sons.

Fabris C., Freitas A. (2006) Discovering Surprising Instances of Simpson's paradox in Hierarchical Multidimensional Data. *International Journal of Data Warehousing and Mining (IJDWM).* Vol. 2. No. 1. P. 27—49. https://doi.org/10.4018/jdwm.2006010102.

Field A. (2009) Discovering statistics using SPSS. London: Sage.

Franco G. (2016) Multiple Correspondence Analysis: One Only or Several Techniques? *Quality & Quantity*. Vol. 50. No. 3. P. 1299—1315. https://doi.org/10.1007/s11135-015-0206-0.

Goltz H., Smith M. (2010) Yule–Simpson's Paradox in Research. *Practical Assessment, Research & Evaluation.* Vol. 15. No. 15. P. 1—9.

Greenacre M., Blasius J. (2006) Multiple Correspondence Analysis and Related Methods. London: Chapman and Hall/CRC.

Greenacre M. (2007) Correspondence Analysis in Practice (Second Edition). London: Chapman & Hall/CRC.

Han J., Kamber M., Pei J. (2012) Data Mining: Concepts and Techniques (Third edition). Elsevier/Morgan Kaufmann Series in Data Management Systems.

Kievit R., Frankenhuis W., Waldorp L., Borsboom D. (2013) Simpson's Paradox in Psychological Science: a Practical Guide. Frontiers in Psychology. Vol. 4. No. 513. P. 1—14. https://doi.org/10.3389/fpsyg.2013.00513.

Morgan J., Sonquist J. (1963) Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*. Vol. 58. No. 302. P. 415—434. http://dx.doi.org/10.1080/01621459.1963.10500855.

Ratner B. (2017) Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. London: Chapman and Hall/CRC.

Ritschard G. (2013) CHAID and Earlier Supervised Tree Methods. In: *Contemporary issues in exploratory data mining in the behavioral sciences*. New York: Routledge. P. 70—96.

Simpson E. (1951) The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*. Vol. 13. No. 2. P. 238—241. http://dx.doi.org/10.1111/j.2517-6161.1951.tb00088.x.

Tabachnick B., Fidell L. (2012) Using Multivariate Statistics. Boston, Mass.: Pearson.

Vermunt J. (2005) Log-Linear Models. In: *Encyclopedia of Statistics in Behavioral Science.* New York: John Wiley & Sons. P. 1082—1093.