1. ИЗМЕРЕНИЕ И АНАЛИЗ ДАННЫХ

Об институционализации истории методов социологического исследования

Толстова Юлиана Николаевна, *НИУ ВШЭ*

1. Как мы понимаем самостоятельность рассматриваемой области знания и институционализацию этой самостоятельности.

1.1. О чем доклад

В данном докладе мы не будем строго определять все те термины, которые присутствуют в заглавии. Поясним, как мы намереваемся обойти необходимость такого определения.

Без доказательства утверждаем, что в нашем случае речь может идти только об *относительной* самостоятельности интересующей нас области научного знания, поскольку полагаем очевидным (в данной заметке; в принципе все надо доказывать), что она не может не быть сильно связанной с историей социологии[1] (а когда речь идет об использовании математического языка, что очень важно и о чем мы будем говорить ниже, то и с историей математики)[2].

Не будем говорить и о критериях *научности* тех фактов, которые составляют основу рассматриваемой отрасли знания; без доказательства примем утверждение о том, что рассматриваемая область знания есть часть науки (хотя и это, вообще говоря, требует доказательства). Приведенные ниже, в п. 1.2 свойства этой области можно рассматривать как фрагмент такого доказательства, но этот фрагмент касается как бы внешней, логической стороны науки, но не сути «наполняющих» её фактов.

Цель доклада — обоснование *самостоятельности* (хотя и относительной) рассматриваемой области научного знания, дающей этой области право называться *научной ветвью*, и о необходимости *институционализации* этой самостоятельности. Наше понимание самостоятельности и институционализации мы опишем ниже, но не будем подробно обосновывать целесообразность принятия именно предлагаемых определений.

1.2. Наше понимание самостоятельности рассматриваемой области научного знания

Итак, полагаем, что упомянутая *самостоятельность* (относительная) рассматриваемого фрагмента научного знания, дающая ему возможность называться автономной научной ветвью, проявляется в следующем.

1. Эта область имеет свой *объект* исследования: способы получения нового научного социологического (предсоциологического) знания, использовавшиеся на всем протяжении развития науки об обществе. *Предметом* рассматриваемой научной ветви служат закономерности, по которым развивались и развиваются способы получения нового социологического знания. Полагаем очевидным, что эти предмет и объект, во-первых, отличаются от таковых для истории социологии, хотя и связаны с ними; и, во-вторых, представляют собой интерес для изучения.

2. Эта область знания имеет свои собственные *методы* изучения предмета исследования. О них более подробно пойдет речь ниже в п. 3.

3. Рассматриваемая область знания требует *особой квалификации специалистов*, занимающихся ею. Эти специалисты должны быть профессионалами и в области социологии и её истории, и в области психологии, что позволит им адекватно общаться с респондентами для получения нового знания в настоящем и оценивать соответствующие правила, использовавшиеся в

[1] Под «социологией» мы понимаем и то, что обычно называют предсоциологией; здесь мы имеем в виду знания, накопленные европейской наукой в течение последних четырех веков.

[2] Рассматриваемая нами область знания не является подветвью истории социологии, поскольку в нее входят история математики и история психологии.

прошлом[1]; и в области моделирования социальных явлений, что нужно для эффективного выделения в реальности того, что интересует исследователя, что подлежит отражению в более или менее формальных конструкциях[2]; и в области математического аппарата, что должно обеспечивать возможность строить новые и эффективно использовать известные математические конструкции для получения нового знания. И всеми этими качествами специалисты должны обладать не только для решения социологических задач в настоящем, а и для оценки сути, значимости для настоящего того, что имело место в прошлом. Наверное, всё сказанное в настоящем пункте тоже можно считать более-менее очевидным

4. Работа в рассматриваемой области требует *особой психической организации специалистов*, что обусловлено сказанным в предыдущем пункте.

5. Конечно, необходимо также, чтобы в рассматриваемой области знания было накоплено достаточно фактов для того, чтобы имело смысл обсуждать все сказанное выше. Сюда вероятно, следует добавить и факты, касающиеся наличия в историческом развитии методов определенных закономерностей. К вопросу о соответствующих «накоплениях» в области истории развития социологических методов мы вернемся в п. 3.

1.3. Наше понимание институционализации области научного знания

Интересующая нас *институционализация* означает, что научным сообществом должна быть признана необходимость изучения истории методов, разработки более детальных представлений о предмете и объекте рассматриваемой ветви науки, методах их изучения; должна быть осознана потребность в целенаправленной подготовке кадров. Ясно, что это — неконструктивная формулировка, но вряд ли сейчас надо добиваться полной конструктивности. Своим выступлением я хочу призвать слушателей заниматься историей развития социологических методов, увеличивать количество соответствующих публикаций, давать соответствующие темы для научных работ студентам, создавать условия, при которых формировалось бы внимательное отношение к рассматриваемой проблематике у лиц, принимающих решение при выделении грантов, принятии статей в журналах, публикации книг и, конечно, при формировании учебных планов вузов, готовящих социологов.

Для принятия моего призыва, естественно, у специалистов-социологов должно сформироваться убеждение в том, что усиление работы в области изучения истории методов социологического исследования, действительно нужно для продвижения вперед социологии. Перейдем к обоснованию этого тезиса.

2. Зачем нужно институционализировать научную область «история методов социологического исследования»

Сразу дадим ответ, а потом попытаемся его в какой-то мере обосновать: институционализазация указанной научной области как относительно самостоятельной ветви науки требуется для того, чтобы повысить уровень методологически-методического обеспечения проводимых в стране социологических исследований, который сейчас оставляется желать много лучшего.

2.1. Неблагополучие с методным обеспечением социологических исследований

Положение с методами в проводящихся в стране социологических исследованиях нельзя назвать благополучным. Об этом было сказано, в частности, в меморандуме IV конференции памяти А.О. Крыштановского. Меморандум был помещен на сайте РОС, и отклики на него показали, что с констатацией этого факта согласны очень многие российские социологи. Особенно плохо обстоит дело с методами, использующими математический язык[3].

---

[1] Термином «методы» мы обозначаем все методы, так или иначе использующиеся в социологии.

[2] Имеется в виду отнюдь не только использование математических методов. Моделирование, формулировка априорных «аксиом» требуется в любом научном исследовании.

[3] Мы намеренно говорим об использовании в социологии не математики, а математического языка (будем также говорить в том же смысле о математических методах). В математике мы выделяем две составляющие: ту, которая связана с её пониманием как науки, изучающей формальные объекты по формальным правилам; и ту, которая касается гносеологической сути математики, отражающей понимание того, какое абстрагирование от реальных си-

В чем состоит неблагополучие. Из существующего в мировой науке арсенала методов используется крайне незначительная их часть. Это касается и методов сбора данных (имеются в виду процедуры всех уровней формализации), и методов анализа собранной информации. Если что-то используется, то без особых раздумий, без тщательного обоснования выбора того или иного подхода. При анализе данных это бывает сопряжено с т.н. «кнопочной» психологией: компьютер сам «довезет», знай, нажимай кнопки. Существующие методические разработки редко используются, а для многих известных подходов их просто нет (никто о них не думал со времени создания метода).

Причины неблагополучия ситуации: слабое обеспечение исследователей методической литературой; отсутствие достаточного количества методических разработок, указаний на то, в каких ситуациях имеет смысл использовать тот или иной подход; отсутствие традиции связывать методы сбора и анализа данных со смыслом решающейся задачи; слабая подготовка исследователей в области моделирования социальных явлений (мы имеем в виду в первую очередь не конкретные методы моделирования, а принципы выделения в реальности тех ее аспектов, которые подлежат изучению). Естественно, сказанное тесно связано с относительно низким уровнем преподавания методов, с отсутствием достаточного количества соответствующих специалистов как среди исследователей, так и среди преподавателей.

Последствия неблагополучия ситуации. Стоящие перед социологией, как и перед любой другой наукой основные задачи — описание, объяснение, предсказание — часто решаются весьма некорректно. Приведем лишь отдельные (и относительно простые) примеры, делая упор на использование т.н. математических методов, насчет использования которых вопрос по понятным причинам стоит наиболее остро. Для описания ситуации часто используются распределения наблюдаемых признаков. При этом, как правило, не учитывается целый веер методических моментов. Так, отдельный наблюдаемый признак может интересовать исследователя лишь как показатель чего-то латентного и, вследствие этого, наблюдаемая шкала должна быть переделана. Скажем, возраст может интересовать социолога как показатель социальной зрелости респондента, и в таком случае «расстояние», скажем между 10 и 20 годами должно стать больше, чем «расстояние» между 30 и 40 годами, а «расстояние» между 70 и 80 годами — обнулиться, о чем обычно вопрос даже не ставится. Для объяснения, или, как чаще говорят, для поиска причин социолог в лучшем случае считает какой-нибудь парный коэффициент связи (а часто и этого не делает, опираясь только на смутные интуитивные соображения о том, что один признак можно считать причиной, другой — следствием), в то время как наука предлагает серьезные методы изучения каузальных структур (скажем, методы моделирования связей с помощью структурных уравнений). Для решения задач прогноза крайне редко используются современные методы анализа временных рядов. И объясняется это не только тем, что социолог нечасто имеет возможность получить числовые ряды, на анализ которых рассчитано большинство методов, но и слабым знанием социологов соответствующих подходов. И, конечно, неадекватным делается использование любого метода в силу отсутствия четкой отработанной методики по установлению связи формализма с содержанием, о чем мы говорили выше.

2.2. Извлечение уроков из обращения к истории социологических методов

Прежде чем говорить об обращении к истории, ответим на естественно возникающий вопрос: почему мы говорим об институционализации научной ветви, называемой не «Методы социологического исследования», а «История методов социологического исследования»? Не имея возможности подробно на него ответить, укажем коротко на следующие моменты.

Во-первых, научная ветвь «Методы социологических исследований» в значительной мере институционализирована. Соотваетствующая учебная дисциплина преподается в каждом вузе, готовящем социологов, в ряде вузов имеется соответствующая специализация, издаются книги, в журналах есть рубрики. Этого нельзя сказать о научной ветви «История методов…».

Во-вторых, мы полагаем, что, говоря об исправлении ситуации с методным обеспечением проводимых в стране социологических исследований, следует делать больший упор именно на

туаций приводит к рождению той или иной математической конструкции. Социолога интересует, прежде всего, вторая составляющая.

историю методов, а не просто на методы. Причина по существу совпадает с причиной того, почему теория и история социологии обычно объединяются в одну дисциплину. Единой современной теории социологии не существует. Теорий много, и каждая существует как фрагмент история социологической мысли. Чтобы понять теорию, надо понять, как она рождалась, в каких условиях и почему возникла.

То же и с методами. Единого подхода к решению социологических задач какого-либо типа не существует. Для сложных задач вообще не существует универсальных (для разных задач) методов их удовлетворительного (тут, конечно, требуется разработка критериев удовлетворительности, да и критериев сложности тоже, не будем здесь об этом говорить) решения. Такие задачи уникальны, и для решения каждой из них требуется разработка уникального подхода, своеобразных методов моделирования ситуации. Но то же верно и для относительно простых задач. Скажем, если мы хотим проверить, можем ли мы считать один признак причиной другого, то для этого можем использовать какой-либо из многочисленных парных коэффициентов связи, регрессионный анализ, дисперсионный анализ. И не исключено, что получим разные ответы на наш вопрос. Каков же выход из положения? Во всяком случае, один из выходов — изучение истории появления каждого метода; корней, из которых он вырос; путей его дальнейшего развития и осуществление и выбора метода, и способов интерпретации результатов его применения, исходя из «советов» истории.

Как известно, изучение истории вопроса способствует решению очень многих проблем. Это имеет место и для интересующей нас ситуации.

О необходимости изучения истории науки много полезного говорится, например, в работах В.И. Вернадского (1863–1945), крупнейшего русского и советского учёного, одного из основателей института истории естествознания и техники. Приведем цитату: «Прошлое научной мысли рисуется нам каждый раз в совершенно иной и новой перспективе. Каждое научное поколение открывает в прошлом новые черты… Случайное и неважное в глазах ученых одного десятилетия получает в глазах другого нередко крупное и глубокое значение… История научной мысли… никогда не может дать законченную неизменную картину, реально передающую действительный ход событий… историк сам создает материалы своего исследования, оставаясь, однако, все время в рамках точного научного наблюдения. Поэтому в истории науки постоянно приходится возвращаться к старым сюжетам, пересматривать историю вопроса, вновь ее строить и переделывать»[1]. Мы еще вернемся к соображениям Вернадского о целях изучения истории науки, говоря о том, что такое социология знания о прошлом в п. 3.

Любая профильная дисциплина, преподаваемая в вузах, сопровождается преподаванием ее истории. Все социологи изучают историю социологии. Если вспомнить еще раз о «больной» проблеме объединения социологии с математикой, то можно напомнить, что и каждый будущий математик, конечно, изучает историю математики. А вот вопроса об изучении истории методов социологического исследования пока даже не стоит, хотя, как мы уже упоминали, дисциплина «Методы социологических исследований» включена в учебные планы всех студентов-социологов России. Не стоит, несмотря на то, что соответствующие аспекты не включаются полностью ни в одну из преподаваемых «историй».

2.3. Что способствует и что мешает институционализации

В п. 1.2, мы описали наше представление о том, выполнение каких условий даёт нам основания говорить о возможности институционализации области знания, касающейся истории социологических методов. И обещали вернуться к двум из этих условий (2 и 5), поскольку выполнение их представляется нам неочевидным.

Начнем с условия 5. Хорошо ли нам известна история развития методов социологии? Наберется ли достаточное количество известных «методных» фактов, чтобы в принципе можно было говорить, скажем, о выявлении каких-то закономерностей в развитии методов? Нам представляется, что в литературе отражено достаточное количество фактов, но, к сожалению, лишь небольшое их число отражено в работах, ориентированных непосредственно на социологов.

[1] Вернадский В.И. Избранные труды по истории науки. М.: Наука, 1981.

Много занимались вопросами социологической методологии и методики, социологическими методами, к примеру, классики социологии М. Вебер[1] и Э. Дюркгейм[2] (и многие другие известные ученые). На наш взгляд, их работы мало изучены именно с точки зрения развития методов социологического исследования. А дать такое изучение может много. Представляется, что наш современник, желающий использовать, к примеру, идеи причинного анализа, очень много мог бы почерпнуть из работы Дюркгейма по изучению проблемы самоубийства[3]. Анализ приемов классика может помочь грамотно применять современную компьютерную технологию, избежать механического использования известных пакетов программ.

В отечественной литературе по истории социологии часто встречаются фрагменты, касающиеся так называемой эмпирической социологии. К сожалению, о методах в этих фрагментах говорится редко. Одним из исключений является работа[4]. Много материала можно найти в работах русских статистиков конца XIX – первой половины XX века[5]. Кладезем информации являются работы русских земских статистиков. Они очень много работали над методами исследования (как «качественными», так и «количественными»), но, к сожалению, многое мы забыли[6]. Некоторые интересующие нас факты можно найти даже в современной литературе по истории теории вероятностей и математической статистике[7].

Итак, в целом представляется, что существующие в наше время публикации говорят о том, что история развития методов социологии достаточно богата фактами для того, чтобы можно было говорить о ее институционализации.

Какими могут быть методы изучения истории развития социологических методов? Какие закономерности уже были выявлены? Соответствующих разработок, насколько нам известно, в литературе весьма мало. Предложим свои соображения и затем приведем пример некоторой закономерности в развития социологических методов, выявленной нами с помощью предлагаемых методов изучения истории. Итак, каковы могут быть способы познания закономерностей в процессе развития методов социологического исследования?

3. Принципы построения «История методов социологических исследований»

3.1. Предложение выделения двух логических линий в развитии истории науки, отражающих социологические аспекты развития науки: «социология прошлого знания» (СПЗ) и «социология знания о прошлом» (СЗоП)

Искать закономерности, связывающие разные факты из истории науки, можно по-разному, выделяя разные аспекты развития исследовательской мысли. О двух выделяемых нами возможных способах поиска пойдет речь ниже. Свяжем их, соответственно, с двумя предлагаемыми на рассмотрение читателя понятиями: социологией знания о прошлом (СЗоП) и социологией прошлого знания (СПЗ)[8]. Характер искомых закономерностей и принципы их поиска — свои в рамках каждой из названных «социологий». Понятия СЗоП и СПЗ широки, касаются истории отнюдь не только методов социологического исследования, не только социологии и даже не только науки. Однако мы будем их связывать с историей той узкой части науки, которая каса-

---

[1] Вебер М. Избранные произведения. М.: Прогресс, 1990. С. 603–625.

[2] Дюркгейм Э. Социология. Её метод, предмет, предназначение. М.: Канон, 1995.

[3] Дюркгейм Э. Самоубийство. СПб.: Союз, 1998.

[4] Ковалёва М.С. Предыстория эмпирической социологии // История теоретической социологии. М.: Наука, 1995. С. 173–189.

[5] Птуха М. Очерки по истории статистики XVII–XVIII веков. М.: Госполитиздат, 1945; Чупров А.А. Вопросы статистики. М.: Госстатиздат ЦСУ СССР, 1960.

[6] О некоторых методических результатах земских статистиков см.: Ермолаев А.В., Забаев И.В. К вопросу о методике земских статистических исследований // Социс. 2001. №11.

[7] См., например, о том, как Лаплас изучал проблему парижских подкидышей: Лаплас. Опыт философии теории вероятностей. Популярное изложение основ теории вероятностей и ее приложений. М., 1908.

[8] Оба термина встречаются в литературе, но мы придадим им свой смысл. Так, в работе И.М. Савельевой и А.В. Полетаева (Савельева И.М., Полетаев А.В. Социология знания о прошлом. М.: ГУ–ВШЭ, 2005) трактовка термина «СЗоП» имеет мало общего с нашей. Заметим, что введение нами этих терминов говорит о том, что при изучении истории методов мы предлагаем учитывать принципы социологии знания — науки, занимающейся проблематикой социальной природы знания.

ется социологических методов.

СЗоП применительно к нашей проблематике — это наука о том, как бытующие в современном научном мире взгляды на те или иные свойства арсенала социологических методов (специфика отдельных методов, набор методов, тенденции развития арсенала и т.д.) определяют видение исследователями истории развития методов, как это видение может помочь современной работе социолога (эта линия науки достигает тех целей, о которых шла речь в приведенной выше цитате из Вернадского).

Базовым принципом, определяющим методы СЗоП, мы считаем следующий: выделение актуальной, требующей решения проблемы в настоящем, анализ её генезиса и решение соответствующих задач в прошлом; желательно априорное формирование определенной гипотезы о развитии методов и проверка этой гипотезы при изучении исторического материала.

СПЗ применительно к нашей проблематике — это наука о том, как социальные условия прошлого (и, в частности, совокупность интересующих нас научных представлений) и личностные качества ученых определяли то, какого рода знания в области методов социологических исследований были эти учеными получены и каким образом соответствующие сведения могут быть обобщены на настоящее, могут помочь современному социологу.

3.2. Пример реализации основного принципа СЗоП: развитие представлений о роли признака в социологическом исследовании

Выделение проблемы в настоящем. Мы выделили в настоящем состоянии анализа данных проблему, состоящую в выяснении того, почему в последние десятилетия в совокупности методов анализа данных все шире внедряются методы, связанные с поиском т.н. взаимодействий, т.е. сочетаний значений признаков, детерминирующих какое-нибудь явление (например, известный алгоритм CHAID из SPSS). Насколько важна эта тенденция? Говорит ли она о каких-то принципиально новых поворотах в постановке социологических задач и способах их решения? Не является ли она проявлением непривычного отношения социолога к самому понятию признака? Мы сформировали гипотезу: использование понятия признака (заимствованного социологами из естествознания) подходит далеко не для всех социологических задач.

Рождение понятия признака в естественных науках. Само понятие признака было предложено Декартом. По существу он же предложил и понятие признакового пространства (без использования такой терминологии). Эта идея оказалась очень плодотворной для математики и естественных наук. Говоря о координатном пространстве, Декарт прежде всего говорил о размещении в нем физических тел.

Положительные аспекты использования этого понятия в социологии. Вряд ли можно сомневаться в колоссальном значении для социологов роли признака и признакового пространства. Подавляющая часть методов анализа данных основана на представлении объектов в таком пространстве. В анализе данных активно используются идеи математической статистики, объектов которой служит случайная величина — результат «скрещивания» понятий «признак» и «вероятность». Но в этом социология следовала примеру естественных наук и демографии. Многие методы (в том числе и использующие другие способы получения первичных данных) направлены на поиск признакового пространства в результате производного (вторичного) измерения (многомерное шкалирование, совместный анализ, анализ соответствий). Но в этом социология следовала примеру психологии. А какие методы родились именно в социологии? Остановимся на тех, которые связаны с уходом от естественнонаучной (и психологической) трактовки понятия признака.

«Сопротивление» социологов внедрению понятия признака в социологию. История показывает, что социологи с большим «скрипом» воспринимали идеи математической статистики[1]. Любимый подход социологов к получению нового знания — это анализ частотных таблиц (классический пример — «Самоубийство» Э. Дюркгейма). Построение таблицы — это виртуозная работа, в процессе которой решается много задач выбора: признаков, способов разбиения диапазона изменения каждого признака на интервалы, сочетаний признаков, сочетаний отдельных

[1] Чупров А.А. Вопросы статистики. М.: Госстатиздат, 1960.

интервалов разных признаков. В рамках математической статистики и анализа данных тем временем происходили следующие интересующие нас сдвиги: в дисперсионном анализе родились методы множественного сравнения, позволяющие искать не статистические закономерности «в среднем», а конкретные сочетания значений независимых признаков, в определенном смысле детерминирующие отдельные значения зависимого; стали появляться алгоритмы, в название которых входило сочетание AID (Automatic interaction detector)[1], затем в пакет SPSS был включен алгоритм CHAID и разработана основанная на нем методология поиска деревьев решений и т.д. Другими словами, в течение более чем 100 лет явно пробивала себе дорогу тенденция не рассматривать признаки целиком, а искать сочетания значений разных признаков (взаимодействия), детерминирующие интересующие социолога процессы. Признаки же стали играть чисто номинальную роль: с их помощью удобно собирать информацию, а анализ сводится к «рассыпанию» признаков на отдельные альтернативы и поиск их комбинаций. Значение этой тенденции помог осуществить исторический поиск.

Об оценке качества процедур анализа данных

Орлов Александр Иванович,
Институт высоких статистических технологий и эконометрики МГТУ им. Н.Э. Баумана

Конференция посвящена актуальным тенденциям в развитии методов сбора и анализа полевых данных, методологическим подходам к исследованию современной российской реальности, а также вопросам оценки качества процедур исследования. Обсудим, с какими научными областями связаны научно-исследовательские работы по тематике конференции.

Полевые данные — это, прежде всего, данные выборочных исследований. Выборочные методы широко применяются в научных медицинских исследованиях, при изучении качества продукции, в производственном менеджменте, в биологических, химических и других научных и прикладных исследованиях. Другими словами, выборочные методы — часть прикладной статистики, т. е. науки о том, как обрабатывать данные[2]. Социологическая специфика в выборочных исследованиях, на наш взгляд, в большинстве случаев не выражена.

Отметим, что ряд прикладных областей — маркетинг (изучение предпочтений потребителей, рекламное дело), управление персоналом, отношения с общественностью и другими заинтересованными сторонами и проч. — рассматриваются и социологией, и экономикой в качестве своих составных частей. По крайней мере, подготовка студентов и защита диссертаций, относящихся к этим областям, проводятся как в экономических, так и в социологических учебных структурах и диссертационных советах. Это не случайно. Согласно современным воззрениям управленческие решения следует принимать на основе всей совокупности социальных, технологических, экономических, экологических, политических факторов. Все эти факторы следует рассматривать совместно.

Может показаться, что речь идет об очевидном: о единстве окружающего нас мира и вытекающего из этого тривиального факта единстве науки. Однако отечественная наука искусственно разбита на отдельные части: социологию, экономику, математику, медицину и т.д. Сначала деление было создано для удобства управления, затем перегородки укрепились и стали почти непроницаемыми. Давно придуманное деление мешает развитию науки. В частности, тем, что оно игнорирует самостоятельное существование такой области, как «статистические методы», если угодно, «прикладная статистика».

Между тем хорошо известно, что в США число специалистов по статистическим методам существенно больше, чем математиков. А вот в нашей стране математика, как самостоятельная

[1] Messenger R.S., Mandell J.M. A model search technique for predictive nominal scale multivariate analysis // J. Amer. Stat. Ass. 1972. Vol. 67. P. 768–773; Morgan J.N., Messenger R.S. THAID — a sequential analysis program for nominal dependent variables. Ann Arbor: Institute for social research, 1973.

[2] Орлов А.И. Прикладная статистика: Учебник. М.: Экзамен, 2006.

наука, есть (со своими факультетами, институтами, журналами, диссертационными советами), а прикладная статистика официально отсутствует.

Есть математическая специальность «теория вероятностей и математическая статистика», её представители заняты доказыванием теорем, не имеющих отношения к реальному миру и ничего не дающих тем, кто обрабатывает конкретные данные. Внутри экономической науки есть специальность «статистика», ее представители заняты удовлетворением ведомственных интересов Росстата и действуют на уровне позапрошлого (XIX) века. А вот статистических методов в технике, управлении, экономике, социологии, медицине, истории и т.д., с точки зрения официальных органов, в России нет.

Статистические методы развиваются на нелегальном положении. Конечно, нелегалы находят те или иные легальные прикрытия. В СССР статистические методы в медицине выступали под знаменем медицинской кибернетики (не знаю, как обстоит дело сейчас). Статистические методы в технических науках маскировались под «математические методы в научных исследованиях» (конечно, смешно быть доктором технических наук, ничего не понимающим в технике, как автору этих строк). Статистические методы в экономике, т.е. эконометрика, проходят в рамках специальности «Математические и инструментальные методы экономики» (именно по этой специальности автор защитил свою вторую докторскую диссертацию). Связь с экономической практикой может полностью отсутствовать.

Вопрос «Кто виноват?» не является актуальным. Проанализируем последствия уродливого развития отечественной науки и обсудим возможные пути исправления ситуации. Последствия достаточно понятны. В каждой камере, выделенной делениями официальной науки, оказалось некоторое число лиц, применяющих (а иногда и развивающих) статистические методы. В области социологии это многие из тех, кто собрался на настоящую конференцию. Веря в обоснованность официального деления, они замкнулись внутри своей группы. Всё, что делается в других камерах, их не интересует. Они ничего не знают, и главное, не хотят знать о внешнем мире.

К чему это приводит? К замкнутости внутри своего круга, т.е. к групповщине. Читают только своих, цитируют только своих. А поскольку группа небольшая, и, так уж случилось, нет внутри нее специалистов по математическим методам статистики и анализа данных, происходит понижение общего научного уровня. Ошибки накапливаются, попадая в учебники, энциклопедические издания. Готовя доклад, я долго думал, давать ли здесь перечень ошибок известных участников настоящей конференции, подобный содержанию интернет-ресурса «Профессора-невежды готовят себе на смену новых невежд»[1]. Хотелось порассуждать о том, что кусочно-постоянная функция отличается от непрерывной. Или вот рассуждение, попавшее в энциклопедическое издание. Бесспорно совершенно, что многие переменные признаки, изучаемые в ходе социологических исследований, интегрируют в себе большую совокупность не связанных друг с другом экономических, психологических и других факторов, каждый из которых оказывает на итоговую переменную сравнительно слабое влияние. Автор рассматриваемого утверждения почему-то считает, что согласно Центральной предельной теореме теории вероятностей такие итоговые переменные должны быть распределены нормально. Это верно, если справедлива аддитивная модель, — когда влияющие на нее факторы складываются. А если верна мультипликативная модель, — когда влияющие факторы перемножаются, — то из той же Центральной предельной теоремы следует, что итоговые переменные должны быть распределены логарифмически нормально (т.е. не они сами, а их логарифмы распределены нормально).

Несколько иной эффект встретился мне в теории классификации. Она — междисциплинарна, и ясно это было давно. В 1980-е годы активно действовала Комиссии по классификации Всесоюзного совета научно-технических обществ во главе с член-корр. АН СССР Г.Б. Бокием, включавшая около тысячи специалистов. И вот социологи пишут про типологию и классификацию, игнорируя всё, что «за забором». Чтобы чуть-чуть исправить положение, укажу на обзор[2] со 126 литературными ссылками, неизвестными социологам.

---

[1] Профессора-невежды готовят себе на смену новых невежд // http://forum.orlovs.pp.ru/viewtopic.php?t=548

[2] Орлов А.И. О развитии математических методов теории классификации // Заводская лаборатория. 2009. Т. 75. № 7. С. 51–63.

Впрочем, судя по учебникам (!) для студентов-социологов, некоторые их авторы еще не освоили аксиоматику теории вероятностей по А.Н. Колмогорову (1933 год!) и не понимают, что для описания, изучения, сравнения методов анализа данных надо сначала ввести вероятностно-статистическую модель порождения данных в терминах современной теории вероятностей. Смешно читать про какой-то «комплекс условий», который должен повторяться… Достаточно вспомнить про проверку статистических гипотез, которая обычно проводится на основе всего лишь одного случайного значения. Если это значение попадает в определенную область, то принимается первая гипотеза, если не попадает — вторая. Никакого повторения «комплекса условий»…

Игнорируется не только сделанное «снаружи», но и сделанное здесь же, но давно. Мне об этом уже приходилось писать[1] и выступать здесь же на предыдущей конференции[2], полемическое выступление включено в «Дискуссию о социологии» на официальном сайте Российского общества социологов[3]. Этот сюжет можно было бы развить, называя конкретные фамилии, проводя ссылки на источники. Однако я не стал этого делать, поскольку опасаюсь, что моя критика была бы адресована наиболее достойным ученым, с которыми меня связывают десятилетия движения по параллельным путям в науке, а деятельность менее квалифицированных лиц осталась бы без адекватной оценки.

Подводя итоги первой части настоящей работы, констатируем, что методы сбора и анализа полевых данных — это предмет не только социологии, но и теории выборочных исследований как части прикладной статистики.

Вторая часть тематики конференции — методологические подходы к исследованию современной российской реальности — также междисциплинарна. Изучение поведения потребителей (маркетинг) и работников (управление персоналом) тяготеет к экономике, исследование динамики семейных отношений — к демографии, политических процессов — к политологии, самоубийств — к криминальной статистике и т.д. и т.п.

Знаковой фигурой XXI в. является американский исследователь Дэниэл Канеман, получивший Нобелевскую премию по экономике за 2002 г. В ходе ряда экспериментов ему удалось доказать, что в своей повседневной жизни большинство людей не руководствуются здравым смыслом, и тем самым впервые удалось ввести в экономику понятие человеческого фактора, объединив психологию и экономику. Интересно, что лауреат Нобелевской премии по экономике никогда экономике не учился, а всю жизнь занимался психологией принятия повседневных экономических решений. Экономисты, начиная с А. Смита, делали одну и ту же ошибку; они предполагали, что человек руководствуется элементарной логикой и собственной выгодой: покупает там, где дешевле, работает там, где больше платят, из двух товаров одинакового качества выберет тот, который дешевле. Исследования Д. Канемана показали, что все не так просто. Люди, оказывается, не хотят думать. Они руководствуются не логикой, а эмоциями, случайными импульсами; тем, что вчера слышали по телевизору, от соседа, устоявшимися предрассудками, рекламой и т.д. Работы Д. Канемана показывают вектор дальнейшего развития общественных наук, необходимость разрушения искусственных перегородок между науками.

Третья часть тематики конференции — вопросы оценки качества процедур исследования. Здесь мы обратим внимание на брошюру[4] по прикладной статистике, посвященную основным характеристикам и требованиям к методам обработки данных, и на методику сравнительного анализа родственных эконометрических моделей, помещенную в качестве приложения 3 к

---

[1] Орлов А.И. Статистические методы в российской социологии (тридцать лет спустя) // Социология 4М. 2005. № 20. С. 32–53.

[2] Орлов А.И. Отечественные достижения: теория устойчивости и нечисловая статистика // Материалы IV конференции «Современные проблемы формирования методного арсенала социолога» (Москва, 16 февраля 2010 г.). М.: Институт социологии РАН, 2010. CD диск ISBN 978-5-89697-181-8 http://www.ssa-rss.ru/index.php?page_id=259

[3] Орлов А.И. Черная дыра отечественной социологии: Выступление 09-01-2011 в «Дискуссии о социологии» на сайте Российского общества социологов http://www.ssa-rss.ru/index.php?page_id=19&id=456

[4] Орлов А.И., Миронова Н.Г., Фомин В.Н., Черчинцев А.Н. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. М.: ВНИИСтандартизации, 1987.

учебнику «Эконометрика»[1]. В рамках настоящего доклада сосредоточимся на двух примерах: на оценке качества процедур усреднения и алгоритмов диагностики.

Среди актуальных направлений, в которых развиваются математические методы исследования, обычно выделяют статистику объектов нечисловой природы, а в ней, как одну из важнейших составных частей, — теорию измерений. За последние десятилетия теория измерений прошла путь от малоизвестного раздела математической психологии до общенаучной концепции, знакомство с которой признается обязательным для исследователей и студентов самых разных специальностей.

Теория измерений исходит из того, что арифметические действия с используемыми в практической работе числами не всегда имеют смысл. Например, зачем складывать или умножать номера телефонов? Далее, не всегда выполнены привычные арифметические соотношения. Например, сумма знаний двух двоечников не равна знаниям «хорошиста», т.е. для оценок знаний 2+2 не равно 4. Приведенные примеры показывают, что практика использования чисел для описания результатов наблюдений (измерений, испытаний, анализов, опытов) заслуживает методологического анализа.

Основные шкалы измерений — наименований (номинальная), порядковая, интервалов, отношений, разностей, абсолютная — подробно описаны в литературе[2]. Используются и иные типы шкал[3]. В настоящее время считается необходимым перед применением тех или иных алгоритмов анализа данных установить, в шкалах каких типов измерены рассматриваемые величины.

Выяснение типов используемых шкал необходимо для адекватного выбора методов анализа данных. Основополагающим требованием является независимость выводов от того, какой именно шкалой измерения воспользовался исследователь (среди всех шкал, переходящих друг в друга при допустимых преобразованиях). Например, если речь о длинах, то выводы не должны зависеть от того, измерены ли длины в метрах, аршинах, саженях, футах или дюймах. Другими словами, выводы должны быть инвариантны относительно группы допустимых преобразований шкалы измерения. Только тогда их можно назвать адекватными, т.е. избавленными от субъективизма исследователя, выбирающего определенную шкалу из множества шкал заданного типа, связанных допустимыми преобразованиями. Требование инвариантности выводов накладывает ограничения на множество возможных алгоритмов анализа данных. В качестве примера рассмотрим порядковую шкалу. Одни алгоритмы анализа данных позволяют получать адекватные выводы, другие нет. Например, в задаче проверки однородности двух независимых выборок алгоритмы ранговой статистики (т.е. использующие только ранги результатов измерений) дают адекватные выводы, а статистики Крамера-Уэлча и Стьюдента — нет. Значит, для обработки данных, измеренных в порядковой шкале, критерии Смирнова и Вилкоксона можно использовать, а критерии Крамера-Уэлча и Стьюдента нет.

Оказывается, требование инвариантности является достаточно сильным. Из многих алгоритмов анализа статистических данных ему удовлетворяют лишь некоторые. Покажем это на примере сравнения средних величин.

Пусть X_1 , X_2 ,..., X_n — выборка объема n. Наиболее общее понятие средней величины введено французским математиком первой половины XIX в., академиком О. Коши. Средней величиной (по Коши) является любая функция $f(X_1, X_2,...,X_n)$ такая, что при всех возможных значениях аргументов значение этой функции не меньше, чем минимальное из чисел $X_1, X_2,...,X_n$, и не больше, чем максимальное из этих чисел. Средними по Коши являются среднее арифметическое, медиана, мода, среднее геометрическое, среднее гармоническое, среднее квадратическое.

Средние величины используются обычно для того, чтобы заменить совокупность чисел (выборку) одним числом, а затем сравнивать совокупности с помощью средних. Пусть, например, $Y_1, Y_2,...,Y_n$ — совокупность оценок экспертов, «выставленных» одному объекту экспертизы, Z_1, $Z_2,...,Z_n$ — второму. Как сравнивать эти совокупности? Очевидно, самый простой способ — по средним значениям.

---

[1] Орлов А.И. Эконометрика: Учебник для вузов. Изд. 3-е, перераб. и дополн. М.: Экзамен, 2004. Приложение 3.

[2] Орлов А.И. Организационно-экономическое моделирование: теория принятия решений. М.: КноРус, 2011.

[3] Толстова Ю.Н. Измерения в социологии. М.: Инфра-М, 1998.

При допустимом преобразовании шкалы значение средней величины, очевидно, меняется. Но выводы о том, для какой совокупности среднее больше, а для какой меньше, не должны меняться (в соответствии с требованием инвариантности выводов, принятом как основное требование в теории измерений). Сформулируем соответствующую математическую задачу поиска вида средних величин, результат сравнения которых устойчив относительно допустимых преобразований шкалы.

Пусть $f(X_1, X_2,...,X_n)$ — среднее по Коши. Пусть среднее по первой совокупности меньше среднего по второй совокупности:

$$f(Y_1, Y_2,...,Y_n) < f(Z_1, Z_2,...,Z_n).$$

Тогда согласно теории измерений для устойчивости результата сравнения средних необходимо, чтобы для любого допустимого преобразования g (из группы допустимых преобразований в соответствующей шкале) было справедливо также неравенство

$$f(g(Y_1), g(Y_2),...,g(Y_n)) < f(g(Z_1), g(Z_2),...,g(Z_n)),$$

т.е. среднее преобразованных значений из первой совокупности также было меньше среднего преобразованных значений для второй совокупности. Причем сформулированное условие должно быть выполнено для любых двух совокупностей $Y_1, Y_2,...,Y_n$ и $Z_1, Z_2,...,Z_n$. И, напомним, для любого допустимого преобразования. Средние величины, удовлетворяющие сформулированному условию, назовем *допустимыми* (в соответствующей шкале). Согласно теории измерений только допустимыми средними величинами можно пользоваться при анализе мнений экспертов и иных данных, измеренных в рассматриваемой шкале.

С помощью разработанной математической теории[1] удается описать вид допустимых средних величин в основных шкалах. Рассмотрим обработку, для определенности, мнений респондентов или экспертов, измеренных в порядковой шкале. Справедливо следующее утверждение:

Теорема 1. Из всех средних по Коши допустимыми средними в порядковой шкале являются только члены вариационного ряда (порядковые статистики).

Теорема 1 справедлива при условии, что среднее $f(X_1, X_2,...,X_n)$ является непрерывной (по совокупности переменных) и симметрической функцией. Последнее означает, что при перестановке аргументов значение функции $f(X_1, X_2,...,X_n)$ не меняется. Это условие является вполне естественным, ибо среднюю величину находим для *совокупности* (*множества*) чисел, а не для *последовательности*. Множество не меняется в зависимости от того, в какой последовательности мы перечисляем его элементы.

Согласно теореме 1, в качестве среднего для данных, измеренных в порядковой шкале, можно использовать, в частности, медиану (при нечетном объеме выборки). При четном же объеме следует применять один из двух центральных членов вариационного ряда, как их иногда называют, левую медиану или правую медиану. Моду тоже можно использовать, она всегда является членом вариационного ряда. Можно применять выборочные квартили, минимум и максимум, но никогда нельзя рассчитывать среднее арифметическое, среднее геометрическое и т.д.

Естественная система аксиом (требований к средним величинам) приводит к так называемым ассоциативным средним. Их общий вид нашел в 1930 г. А.Н. Колмогоров[2], и теперь их называют «средними по Колмогорову». Для чисел $X_1, X_2,...,X_n$ средним по Колмогорову является

$$G\{(F(X_1) + F(X_2) +...+ F(X_n))/n\},$$

где F — строго монотонная функция (т.е. строго возрастающая или строго убывающая), G — функция, обратная к F. Среди средних по Колмогорову много хорошо известных персонажей. Так, если $F(x) = x$, то среднее по Колмогорову — это среднее арифметическое, если $F(x) = \ln x$, то среднее геометрическое, если $F(x) = 1/x$, то среднее гармоническое, если $F(x) = x^2$, то среднее квадратическое и т.д. (в последних трех случаях усредняются положительные величины).

Среднее по Колмогорову — частный случай среднего по Коши. Однако такие популярные средние, как медиана и мода, нельзя представить в виде средних по Колмогорову.

Справедливы следующие утверждения:

---

[1] Орлов А.И. Устойчивость в социально-экономических моделях. М.: Наука, 1979.

[2] Колмогоров А.Н. Избранные труды. Математика и механика. М.: Наука, 1985. С. 136–138.

Теорема 2. В шкале интервалов из всех средних по Колмогорову допустимым является только среднее арифметическое. Таким образом, среднее геометрическое или среднее квадратическое температур (в шкале Цельсия), потенциальных энергий или координат точек не имеют смысла. В качестве среднего надо применять среднее арифметическое. А также можно использовать медиану или моду.

Теорема 3. В шкале отношений из всех средних по Колмогорову допустимыми являются только степенные средние с $F(x) = x^c,\ c \neq 0$ и среднее геометрическое.

Есть ли средние по Колмогорову, которыми нельзя пользоваться в шкале отношений? Конечно, есть. Например, с $F(x) = e^x$. Среднее геометрическое является пределом степенных средних при $c \to 0$. Теоремы 2 и 3 справедливы при выполнении некоторых внутриматематических условий регулярности.

На наш взгляд, теоремы 1–3 должны быть известны всем студентам-социологам. Как и все, я не могу претендовать на полное знание литературы. Однако буду благодарен за указание учебников для социологов, в которых приведены теоремы 1–3.

Аналогично средним величинам могут быть изучены и другие статистические характеристики: показатели разброса, связи, расстояния и др. Нетрудно показать, например, что коэффициент корреляции не меняется при любом допустимом преобразовании в шкале интервалов, как и отношение дисперсий. Дисперсия не меняется в шкале разностей, коэффициент вариации — в шкале отношений и т.д.

Перейдем к оценке качества алгоритмов диагностики. Результаты обработки реальных данных с помощью некоторого алгоритма диагностики в случае двух классов описываются долями правильной диагностики a в первом классе и b во втором, с учетом долей классов в объединенной совокупности $c(i)$, $i = 1, 2$, $c(1) + c(2) = 1$.

Нередко как показатель качества алгоритма диагностики (прогностической «силы») используют долю правильной диагностики $m = c(1)a + c(2)b$[1]. Однако показатель m определяется, в частности, через характеристики $c(1)$, $c(2)$, частично заданные исследователем (например, на них влияет тактика отбора респондентов или образцов для изучения).

В аналогичной медицинской задаче величина m оказалась больше для тривиального прогноза, согласно которому у всех больных течение заболевания будет благоприятно. Тривиальный прогноз сравнивался с алгоритмом выделения больных с прогнозируемым тяжелым течением заболевания. Он был разработан группы под руководством академика АН СССР И.М. Гельфанда. Применение этого алгоритма с медицинской точки зрения вполне оправдано[2]. Но по доле правильной классификации m алгоритм группы И.М. Гельфанда оказался хуже тривиального — объявить всех больных легкими, не требующими специального наблюдения. Этот вывод очевидно нелеп. И причина появления нелепости вполне понятна. Хотя доля тяжелых больных невелика, но смертельные исходы сосредоточены именно в этой группе больных. Поэтому целесообразна гипердиагностика: рациональнее часть легких больных объявить тяжелыми, чем сделать ошибку в противоположную сторону.

Разберем ситуацию подробнее. Пусть имеется некоторый алгоритм диагностики на два класса с долями правильной диагностики a в первом классе и b во втором. Сравним его с двумя тривиальными алгоритмами диагностики. Первый тривиальный алгоритм относит все классифицируемые объекты к первому классу, для него $a = 1$ и $b = 0$, следовательно, $m = c(1)$. Второй тривиальный алгоритм относит все классифицируемые объекты ко второму классу, для него $a = 0$ и $b = 1$, следовательно, $m = c(2)$.

В качестве показателя качества алгоритма диагностики будем использовать долю правильной диагностики. Когда первый тривиальный алгоритм лучше исходного? Когда $c(1) > c(1)a + c(2)b$, т.е. $b / (1 - a + b) < c(1)$ (с учетом того, что $c(1) + c(2) = 1$). Когда второй тривиальный алгоритм лучше исходного? Когда $c(2) > c(1)a + c(2)b$, т.е. $c(1) < (1 - b) /(1 + a - b)$. Таким образом, для любого заданного алгоритма диагностики существуют границы $d(1)$ и $d(2)$ для доли

[1] Горелик А.Л., Скрипкин В.А. Методы распознавания. М.: Высшая школа, 1984.

[2] Гельфанд И.М., Алексеевская М.А., Губерман Ш.А. и др. Прогнозирование исхода инфаркта миокарда с помощью программы «Кора-3» // Кардиология. 1977. Т. 17. № 6. С. 19–23.

первого класса $c(1)$ в объединенной контрольной выборке такие, что при $c(1) < d(1)$ рассматриваемый алгоритм хуже второго тривиального алгоритма, а при $c(1) > d(2)$ он хуже первого тривиального алгоритма.

Поэтому мы полагаем, что использовать в качестве показателя качества алгоритма диагностики долю правильной диагностики нецелесообразно.

Предлагаем применять *метод пересчета на модель линейного дискриминантного анализа*[1], согласно которому показателем качества алгоритма диагностики является т.н. «прогностическая сила», а статистической оценкой «прогностической силы» h является «эмпирическая прогностическая сила» $h^* = \Phi(d^*/2)$, $d^* = G(a) + G(b)$, где $\Phi(x)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1, а $G(y)$ — обратная ей функция.

Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для классификации применяется классический линейный дискриминантный анализ Р. Фишера, то величина d^* представляет собой состоятельную статистическую оценку расстояния Махаланобиса между двумя рассматриваемыми совокупностями, причем независимо от порогового значения, определяющего конкретное решающее правило. В общем случае показатель h^* вводится как эвристический. Распределение статистики h^* является асимптотически нормальным, что позволяет строить доверительные интервалы для прогностической силы h[2].

Как проверить обоснованность пересчета на модель линейного дискриминантного анализа? Допустим, что классификация состоит в вычислении некоторого прогностического индекса y и сравнении его с заданным порогом c. Объект относят к первому классу, если $y \le c$, ко второму, если $y > c$. Прогностический индекс — это обычно линейная функция от характеристик рассматриваемых объектов. Возьмем два значения порога c_1 и c_2. Если пересчет на модель линейного дискриминантного анализа обоснован, то, как можно показать, «прогностические силы» для обоих правил совпадают: $h(c_1) = h(c_2)$. Выполнение этого равенства можно проверить как статистическую гипотезу. Расчетные алгоритмы предложены нами в цитированной работе 1987 г. и включены в наши учебники, цитированные на протяжении настоящей статьи.

Организационные меры, которые целесообразно использовать для исправления явно ненормальной ситуации в той научной области, которой посвящена настоящая конференция, достаточно очевидны. Надо проанализировать сделанное в самой этой области и в прилегающей к ней, в которых развиваются и применяются статистические методы. Не только необходима научная и учебная специальность «Математические методы в социологии», но и обеспечивающая ее инфраструктура — сеть научных учреждений и подразделений, журналов, конференций, диссертационных советов и т.д. Иначе можно умереть от жажды, сидя на каменном островке посередине реки и громко крича: «Ау, мы отстали! Где вода?».

Выбор типичных объектов в классификационных задачах

Буховец Алексей Георгиевич,
Бирючинская Татьяна Яковлевна,
Воронежский государственный аграрный университет им. К.Д. Глинки
Лаврова Екатерина Олеговна, *НИУ ВШЭ*

1. Разделение понятий типологии и классификации позволило акцентировать внимание исследователей на связи содержательной постановки задачи с общеметодологическими понятиями и методами получения типологии как таковой.

[1] Орлов А.И. О сравнении алгоритмов классификации по результатам обработки реальных данных // Доклады Московского Общества испытателей природы 1985 г. Общая биология: Новые данные исследований структуры и функций биологических систем. М.: Наука, 1987. С. 79–82.

[2] См.: Орлов А.И. Организационно-экономическое моделирование: учебник в 3 ч. Часть 1: Нечисловая статистика. М.: Изд-во МГТУ им. Н.Э. Баумана. 2009 и др. учебники.

Классификацию мы будем рассматривать как знание об объективно существующей внешней системе, функционирующей в рамках целостности и состоящей из подобных друг другу представителей целостностной общности. Проблема системности в данном случае — это проблема соотнесения наблюдаемого подобия с общей сущностью. В рамках такого подхода выделяемые классы соответствуют некоторым типам. Известны два принципиально различных подхода к пониманию и описанию типа: (1) тип как *среднее* (предельно обобщенное) и (2) тип как *крайнее* (предельно своеобразное)[1]. В первом случае типичным является объект со свойствами, близкими по своей выраженности к среднему значению класса выборки, во втором — с максимально выраженными свойствами. Эти различия, на наш взгляд, в большей степени связаны с представлением исследователя о рассматриваемой совокупности. Если класс рассматривается отдельно и изолировано от других выделенных классов, то в качестве типичного объекта обычно берется объект, значения показателей которого наиболее близки к среднему значению, т.е. тип интерпретируется как *среднее* (предельно обобщенное). А если классы и анализируются в рамках системного подхода, т.е. как некоторые паттерны целостной совокупности, то в качестве типичного объекта предпочтительнее взять объект, значения признаков которого *максимально* выражают свойства выделенного класса. Таким образом, выбор типичного объекта в значительной степени определяется точкой зрения наблюдателя.

Рассматривая классификационную задачу как задачу исследования структуры многомерных данных, направленную на выделение пространственно неоднородных стационарных распределений объектов[2], отметим, что сам результат, без глубокого понимания причин, его породивших, отражает не более чем образ мышления наблюдателя, его стереотипы или начальные идеологические установки, а не объективное положение дел. Таким образом, выявление и понимание отраженных в понятии *тип* противоречий позволяет развивать это понятие.

Мы разделяем полностью представленные выше взгляды на понятие *типа*. Однако использование таких математических процедур анализа многомерных данных, как кластерный анализ (или его различные модификации), позволяет получать типичные объекты только в виде средних значений, т.е. использовать понятие типа как некой усреднённой величины. В арсенале математических средств, используемых в анализе социологических данных, практически не были представлены методы, позволяющие производить конструирование (выделение) типа, соответствующего второму из отмеченных выше понятий. Такие типы выделялись на этапе теоретического анализа (т.е. при использовании абстрактно-теоретических методов). Однако этот подход всегда оставляет открытым вопрос о существовании таких (даже логически допустимых типов) в конкретном эмпирическом материале. Для определения типичных объектов нами предлагается подход, в основе которого лежит представление о фрактальной структуре многомерных данных и метод (алгоритм), позволяющий проводить выделение типичных объектов, отвечающих второму из приведенных выше определений.

2. Понятие фрактальных множеств (фрактальных структур) было введено сравнительно недавно, в 1975 г., американским математиком польского происхождения Б. Мандельбротом. Не вдаваясь в утомительные подробности строгого математического определения, и не пытаясь охватить все возможные случаи, представляющие фрактальные образования, мы лишь отметим, что для нашей работы вполне достаточно рассмотрения фрактальных объектов, представленных совокупностью точек некоторого признакового пространства. Такие пространства обычно встречаются в работах, где используются методы многомерного статистического анализа: факторного, кластерного или регрессионного. В этом случае фрактальное множество обычно представимо в виде совокупности точек, обладающей некоторыми свойствами, основными из которых являются дробная (фрактальная) размерность множества, меньшая, чем размерность признакового пространства, и наличие свойства самоподобия.

Большая часть работ Б. Мандельброта так или иначе связана с изучением фрактальных структур, их свойств, областей их распространения. В частности, приведем одну цитату из его

[1] Ганзен В.А., Фомин А.А. О понятии типа в психологии // Вестник СПбГУ. Сер. 6. 1993. Вып. 1 (№6).

[2] Типология и классификация в социологических исследованиях / Отв. ред. В.Г. Андреенков, Ю.Н. Толстова. М.: Наука, 1982. С. 156.

широко известной книги: «Я считаю совершенно необходимым — параллельно с продолжением попыток *объяснить* кластеризацию — найти способ *описать* ее и смоделировать реальность чисто геометрическими средствами (*курсив***. — Авторы**)[1]». Фактически здесь идет обсуждение одной важной проблемы современной науки — необходимости не только описывать результаты классификации (кластеризацию), но и моделировать эти структуры. Современный подход к исследованию каких-либо структур — физических, биологических или социальных — нацелен на описание процессов, порождающих эти структуры. Об этом направлении в современном системном мышлении хорошо сказано в известной книге Ф. Капра «Дао физики».

Для моделирования структур многомерных данных нами была предложена процедура, основанная на рандомизированной системе итерирующих функций. Оставляя в стороне подробное описание этой процедуры, мы отсылаем за подробностями читателя к работам, в которых приводится описание самой процедуры и исследование свойств получаемых в результате ее выполнения множеств[2]. Эту процедуру относительно легко можно реализовать в системе Mathcad, что и было сделано авторами.

Для выполнения этой процедуры необходимо указать количество генерируемых точек, точки, которые участвуют в порождении фрактального множества, *точки протофрактала*, вместе с заданным на них распределением вероятностей, а также значение входного параметра, характеризующего степень близости точек одного класса к соответствующей точке протофрактала. Было показано, что построение фрактального множества (предфрактал — это еще то предельное множество, которое называется фракталом, и поэтому предпочтительнее говорить о предфрактале, а не о фрактале) можно выполнить и другим способом, при котором учитывается указанное выше значение входного параметра, число классов в классификации и численности классов[3]. Возможность выполнения процедуры предфрактального множества построения двумя различными способами — с и без использования множества протофрактала — позволяет изменить (обратить) саму схему выполнения алгоритма. Для этого достаточно предположить, что имеющиеся в распоряжении исследователя данные представляют собой некоторое фрактальное образование — предфрактал. Проведенное исследование структуры многомерных данных с помощью алгоритмов кластерного анализа позволяет получить классификационное разбиение, т.е. выделить классы, указать их состав, численность и оценить степень их рассеяния в признаковом пространстве. Полученные результаты выполнения классификационных алгоритмов в дальнейшем будут использоваться для оценки входных параметров указанной выше процедуры построения протофрактала. Результат выполнения процедуры — в данном случае в обратном порядке, т.е. не для построения фрактального множества, а для определения протофрактала — позволяет получить значения признаков объектов, которые в процедуре при выполнении прямого хода играют роль точек протофрактала. Другими словами, если бы выделенные объекты рассматривались как объекты, порождающие все множество посредством процедуры построения фрактала, то эти объекты следует признать как типичные во втором указанном выше определении типичного объекта. В ходе исследования свойств предложенной процедуры было выявлено, что эти объекты могут быть рассмотрены как типичные объекты выделенных классов. При этом понятие типичного объекта, выделенного этой процедурой, будет соответствовать именно второму из рассмотренных нами типов[4].

3. В качестве примера, иллюстрирующего изложенные выше теоретические положения, рассмотрим задачу о выделении типичных объектов при классификации регионов России. Мы будем использовать именно этот термин, т.к. содержательный анализ, выполненный нами в ходе построения типологии, в этой работе не приводится ввиду ограниченности объема.

---

[1] Мандельброт Б. Фрактальная геометрия природы. М: Институт компьютерных исследований, 2002. С. 127.

[2] Буховец А.Г. Системная интерпретация результатов классификации // Социология: методология, методы, математические модели. 2006. № 22. С. 114–144; Буховец А.Г. Математические модели и методы типологического анализа // Социологические методы в современной исследовательской практике: Материалы Всероссийской научной конференции. М., 2007. С. 16–31.

[3] Буховец А.Г. Математические модели и методы…

[4] Буховец А.Г. Системная интерпретация…; Буховец А.Г. Математические модели….

Как известно, задача выделения типичных объектов играет большую роль в построении классификационной (районированной) выборки. Кроме этого, эта задача может иметь самостоятельное значение, поскольку в классификационных задачах типичный объект не всегда совпадает с некоторым усреднением по классу, а скорее определяет главную тенденцию эволюционного развития выделенного класса.

Исходя из анализа основных теоретических работ, посвященных типологии российских регионов были выбраны следующие блоки показателей:

— уровень жизни в регионе;

— инвестиционная активность;

— экономический потенциал.

Были получены статистические данные по 83 регионам, каждый регион характеризовался следующими 8 показателями:

- доля населения с доходами ниже прожиточного минимума (индикатор бедности);
- отношение среднедушевых доходов к прожиточному минимуму;
- отношение среднедушевых расходов к прожиточному минимуму;
- доля инвестиций в валовом региональном продукте;
- темпы роста инвестиций в 2009 г. по отношению к среднероссийскому уровню на соответствующий период 2008 г.;
- отношение иностранных инвестиций к валовому региональному продукту;
- уровень безработицы в регионе (доля безработных от экономически активного населения региона);
- темпы роста валового регионального продукта по отношению к ВВП.

Все показатели, кроме прожиточного минимума (2010 г.) и доли инвестиций в основной капитал (2007 г.), были рассчитаны за 2008 г. (ввиду того, что большинство из них имеют привязку к ВРП, а последнее значение ВРП есть только за 2008 г.).

Все статистические данные были предварительно стандартизированы.

При построения классификации регионов использовались методы кластерного анализа: алгоритмы иерархической классификации и метод k-средних. Как показали исследования, структура данных была такова, что ее хорошо можно было представить по результатам метода k-средних.

Исследование структуры данных дополнялось изучением проекций данных в пространстве первых двух главных компонент, которые представляли 67,4 % обобщенной дисперсии.

Итоговое классификационное разбиение на 8 кластеров приведено в Приложении 1. Средние значения выделенных кластеров приведены в таблице 1.

Приведем очень краткие характеристики выделенных кластеров.

1-й кластер условно можно назвать **«самым благополучным»**, если исходить из высокого уровня жизни и экономического потенциала Ближе всего к кластерному центру находится **Свердловская область**.

2-й кластер составили регионы, которые можно определить как «**потенциально неблагополучные**». Эти регионы быстрыми темпами приближаются к «неблагополучным». Самый яркий представитель — **Амурская** область.

3-й кластер образуют два региона, которые условно можно назвать «самыми **неблагополучными»**. Здесь самая высокая доля населения с доходами ниже прожиточного минимума, отношение среднедушевых доходов и расходов также самое низкое, причем отношение расходов к прожиточному минимуму меньше 1.

4-й кластер состоит из одного региона, который выделен в отдельный кластер из-за низких темпов роста инвестиций и высокой доли иностранных инвестиций.

5-й кластер — «**неблагополучные**» — один из самых многочисленных. Регионы этого кластера имеют довольно высокую долю населения с доходами ниже прожиточного минимума, низкие среднедушевые доходы и расходы. Темпы роста инвестиций в основной бюджет по сравнению со среднероссийским уровнем невысоки. Регионы данного кластера похожи на регионы второго кластера, однако во втором кластере наблюдается другая структура распределения данных, и эта структура, видимо, обусловлена главным различием этих кластеров — высо-

кими темпами роста во втором кластере. Самые яркие представители «**неблагополучных**» регионов — **Забайкальский** край, **Чувашская** республика и **Ульяновская** область.

6-й кластер составили регионы **«потенциально благополучные»**. Самый типичный пример — **Калужская** область.

7-й кластер — самый многочисленный — составляют регионы, занимающие положение между «неблагополучными» и «потенциально благополучными». Можно сказать, что это **«средние»** по уровню развития регионы без ярко выраженных тенденций перехода к неблагополучным или благополучным регионам. Среднедушевые доходы и расходы выше среднего уровня по России, но ниже, же у «благополучных» и «потенциально благополучных» регионов. Типичными представителями кластера служат **Смоленская о**бласть и **Пермский** край.

В отдельный 8-й кластер выделен г**. Москва**, где наблюдается самый высокий уровень жизни и экономического потенциала.

Для определения типичных объектов в рамках фрактального подхода была построена матрица А размером 83х8 при значении параметра процедуры, равного 17,8. Значение параметра выбиралось экспертным путем.

С помощью матрицы А были получены точки признакового пространства представляющие протофрактал. Результаты построения приводятся в Приложении 2. Каждую строку этой матрицы Z можно соотносить с ближайшим объектом классификации.

Точки выделенного протофрактала отражают структуру распределения объектов — регионов России — в пространстве выбранных для классификации признаков. Результаты вычислений для сравнения их со средними значениями приводится в таблице 2.

Можно показать, что в рамках фрактального подхода складывается ситуация, когда значения точек протофрактала, соответствующие кластерам, где средние значения высокие, будет еще выше, а в тех кластерах, где они низкие — еще ниже. Так, к примеру, при разбиении населения по уровню дохода типичным представителем состоятельных людей будет самый богатый, а представителем бедных — самый бедный регион, и лишь среди людей со средним доходом самым ярким представителем будет человек со средним доходом[1]. Однако, такая упрощенная ситуация при анализе классификационного разбиения складывается не часто. Обычно кластеров бывает больше трех, и главной задачей аналитика при содержательной интерпретации является сравнение кластеров между собой, т.е. определение отношения кластеров друг к другу, их взаимосвязи. В нашем случае использование фрактального подхода позволяет оценить различия в структуре кластеров, обнаружить скрытые закономерности в общей структуре и оценить, насколько «потенциальные» регионы отличаются от более однородных.

Ближайшие к точкам протофрактала объекты кластеров в некоторых случаях отличаются от объектов, ближайших к кластерным центрам (см. Таблица 2). Так, самым ярким представителем «самых благополучных» регионов с точки зрения фрактального подхода является г. Санкт-Петербург (была Свердловская область), «потенциально благополучных» — Приморский край (была Амурская область), «неблагополучных» — республика Хакасия и Чувашская республика (был Забайкальский край и Чувашская республика). В кластерах «потенциально благополучных» и «средних» типичные представители не изменились (Калужская и Смоленская области соответственно).

Как видно, различия в положении средних выборочных значений кластеров и вычисленных значений признаков точек протофракалов связано с тем положением, которое занимает кластер во всей исследуемой системе объектов.

В заключении еще раз отметим, что рассмотрение (исследование) эмпирических данных с точки зрения фрактального подхода позволяет практически реализовать метод выделения типичных объектов, наиболее ярко и полно выражающих свойства классов.

Результаты классификационных построений

Приведены: (численность / процентный состав кластеров), список входящих в кластер регионов.

[1] Буховец А.Г. Математические модели….

Жирным шрифтом выделены регионы, типичные в смысле близости к средним значениям показателей кластера; подчеркиванием выделены регионы, типичность которых была установлена в рамках фрактального подхода.

Кластер 1 (9 / 11) Московская область, г. Санкт-Петербург, Татарстан, Самарская, **Свердловская**, Тюменская области, Ханты-Мансийский и Ямало-Ненецкий автономные округа, Челябинская область.

Кластер 2 (3 / 4) **Амурская область**, Приморский край, Якутия.

Кластер 3 (2 / 2) **Ингушетия** и **Чеченская Республика**.

Кластер 4 (1 / 1) ***Ненецкий автономный округ***.

Кластер 5 (27 / 33) Владимирская, Воронежская, Ивановская, Костромская, Рязанская области, Карелия, Калмыкия, Кабардино-Балкария, Карачаево-Черкессия, Марий Эл, Мордовия, **Чувашия**, Кировская, Саратовская, **Ульяновская** области, республика Алтай, Бурятия, Тыва, Хакасия, Алтайский край, **Забайкальский край**, Иркутская и Томская области, Камчатский край, Хабаровский край, Магаданская область, Еврейская автономная область.

Кластер 6 (7 / 8) Чукотский автономный округ, Сахалинская, Вологодская, Архангельская, Липецкая, **Калужская**, Белгородская области.

Кластер 7 (33 / 40) Брянская, Курская, Орловская, **Смоленская**, Тамбовская, Тверская, Тульская и Ярославская области, Республика Коми, Калининградская, Ленинградская, Мурманская, Новгородская и Псковская области, Адыгея, Краснодарский край, Астраханская, Волгоградская и Ростовская области, Дагестан, Северная Осетия-Алания, Ставропольский край, Башкортостан, Удмуртия, **Пермский край**, Нижегородская, Оренбургская, Пензенская и Курганская области, Красноярский край, Кемеровская, Новосибирская и Омская области.

Кластер 8 (1 / 1) **г. Москва**.

Таблица 1

Средние значения показателей кластеров

№ кластера	Численность кластеров	Доля населения с доходами ниже прожиточного минимума	Отношение среднедушевого дохода к прожиточному минимуму	Отношение среднедушевых расходов к прожиточному минимуму.	Темпы роста инвестиций в % к 2008 г. по сравнению со среднерос. уровнем 2009 г.
1	9	10,59	3,25	2,55	96,27
2	3	20,67	1,71	1,31	196,86
3	2	36,10	1,10	0,52	169,99
4	1	10,20	1,74	1,23	43,44
5	27	21,19	1,69	1,36	100,56
6	7	12,49	2,29	1,83	96,80
7	33	15,08	2,23	1,88	109,75
8	1	10,00	4,17	3,84	92,84

Продолжение таблицы 1

№ кластера	Численность кластеров	Доля инвестиций в основной капитал в ВРП_2007	Отношение иностранных инвестиций к ВРП	Отношение темпов роста ВРП и ВВП	Уровень безработицы
1	9	27,80	0,07	0,03	6,46
2	3	35,47	0,05	0,01	9,07
3	2	70,05	0,00	0,00	43,95
4	1	92,70	0,37	0,00	9,70
5	27	27,07	0,02	0,00	10,81
6	7	34,69	0,27	0,01	6,59
7	33	26,88	0,03	0,01	9,17
8	1	11,50	0,12	0,20	2,70

$$Z = \begin{pmatrix} 10.276 & 3.313 & 2.59 & 94.907 & 27.899 & 0.065 & 0.035 & 6.253 \\ 20.638 & 1.688 & 1.289 & 200.858 & 34.995 & 0.046 & 0.006 & 8.455 \\ 37.262 & 0.442 & 0.164 & 173.095 & 71.949 & -0.003 & -0.001 & 45.915 \\ 9.955 & 1.709 & 1.191 & 39.852 & 96.355 & 0.389 & -0 & 9.738 \\ 21.437 & 1.667 & 1.335 & 100.208 & 26.923 & 0.023 & 0.002 & 10.876 \\ 12.33 & 2.296 & 1.827 & 96.794 & 35.277 & 0.279 & 0.007 & 6.367 \\ 15.02 & 2.23 & 1.883 & 109.989 & 26.779 & 0.029 & 0.007 & 9.177 \\ 9.695 & 4.28 & 3.953 & 91.895 & 10.636 & 0.125 & 0.211 & 2.321 \end{pmatrix}$$

Рис. Матрица протофрактала, полученная после выполнения процедуры построения фрактального множества.

Таблица 2

Сравнение значений признаков средних значений (левая колонка) и типичных объектов, полученных в рамках фрактальной теории (правая колонка)

№ кластера	Доля населения с доходами ниже прожиточного минимума		Отношение среднедушевого дохода к прожиточному минимуму		Отношение среднедушевых расходов к прожиточному минимуму		Темпы роста инвестиций в % к 2008 г. по сравнению со среднероссийск. уровнем 2009 г.		Доля инвестиций в основной капитал в ВРП_2007	
1	10,59	10,29	3,25	3,30	2,55	2,59	96,27	95,03	27,80	27,81
2	20,67	20,87	1,71	1,69	1,31	1,29	196,86	201,2	35,47	35,98
3	36,10	36,64	1,10	0,55	0,52	0,21	169,99	174,1	70,05	70,29
4	10,20	9,62	1,74	1,75	1,23	1,23	43,44	35,51	92,70	95,5
5	21,19	21,35	1,69	1,67	1,36	1,34	100,56	100,3	27,07	27,04
6	12,49	12,09	2,29	2,31	1,83	1,84	96,80	96,16	34,69	34,78
7	15,08	15,02	2,23	2,23	1,88	1,88	109,75	109,6	26,88	26,8
8	10,00	9,75	4,17	4,27	3,84	3,94	92,84	91,99	11,50	10,71

Продолжение таблицы 2

№ кластера	Отношение иностранных инвестиций к ВРП		Отношение темпов роста ВРП и ВВП		Уровень безработицы	
1	0,065	0,067	0,033	0,034	6,46	6,17
2	0,047	0,049	0,006	0,006	9,07	8,97
3	0,00	0,001	0,001	0,001	43,95	44,91
4	0,371	0,386	0,002	0,003	9,70	9,65
5	0,025	0,023	0,003	0,002	10,81	10,87
6	0,266	0,279	0,006	0,007	6,59	6,19
7	0,029	0,029	0,007	0,007	9,17	9,17
8	0,121	0,125	0,204	0,21	2,70	2,37

Получение представительных данных по группам на основе представительного обследования входящих в них индивидов

Вейхер Андрей Алексеевич,
СПб. филиал НИУ ВШЭ

Проведение представительного обследования объектов, в которые входит несколько людей (домохозяйств-семей, малых предприятий и т.п.) в большинстве случаев оказывается значительно более трудным, чем представительное обследование общей совокупности людей, входящих во все эти группы.

Как известно, создание выборочных совокупностей адресов мест проживания, по которым проводятся опросы населения, считающиеся наиболее представительными, не являются опросами домохозяйств. При таком опросе населения многоступенчатая выборка с различными спо-

собами рандомизации отбора объектов обследования есть только средство осуществления максимально возможной математической случайности такого отбора. На последней ступени объектом оказывается одиночный респондент, практически, в последнее время, отбираемый по квотам пола, возраста, образования. Параметр «семейность» замеряется при этом как признак индивида. Результатами таких опросов оказывается информация о распределениях индивидов и связях признаков-переменных, описывающих индивидов. Мы получаем, например, информацию о том, сколько людей входят в состав семьей, стоящих на очереди по улучшению жилищных условий, но какая часть семей города стоит на очереди, сказать по собираемым данным нельзя. Соответственно, нельзя узнать, как эти семьи распределены по семейному доходу, т.е. каковы их шансы на приобретение жилья при тех или иных условиях ипотечного кредитования. В то же время, расчеты по обеспечению жильем делаются в единицах квартир, чему адекватен расчет численности домохозяйств.

Обследования домохозяйств (например, для проведения обследований бюджетов семей) предполагают существенно более сложную и дорогостоящую выборочную процедуру, которая по силам очень немногим организациям. Необходимой для построения таких (квотных!) выборок информацией располагают только органы государственной статистики в первые годы после проведения переписи населения, т.е. проведения сплошного опроса. (В данном сообщении не рассматривается проблема качества переписи и определяемого её данными качества соответствующих квотных выборок.) Большие затраты на формирование таких выборок побуждают создавать их в панельном формате. При этом за экономию (от многократности использования одной выборки и содержательное преимущество отслеживания непосредственной динамики состояния домохозяйств) приходится платить нарастанием погрешностей из-за социально неравномерной «смертности» панелей и расхождением обследуемого состава панели (даже при восстановлении панели до исходных пропорций) с теми изменениями состава населения, которые происходят в период использования одной панели.

Появление хорошо актуализируемых регистров населения, объединяющих сведения из разных источников, конечно, существенно облегчило построение выборок домохозяйств однократного использования, однако такие выборки остаются «тяжелой артиллерией» в арсенале выборочных обследований.

Для решения компактных задач по выяснению состава домохозяйств по нескольким признакам разработан метод[1] получения представительных данных в размерности домохозяйств по информации, собираемой в представительных опросных обследованиях индивидов (населения).

Исследования, проводимые с его использованием, ведутся, как правило, в технологии придомового опроса: респонденты отбираются внутри кварталов, в удалении от транспортных узлов и торговых зон, с фильтрацией на проживание в домах в пределах видимости или прямо на расстоянии 30–60 метров от жилого дома, откуда вышел или куда направляется потенциальный респондент. Опрос людей вблизи конкретных домов применяется там, где налицо высокий уровень автомобилизации и есть опасность потерять автомобилизированную часть респондентов. (Потери потенциальных респондентов, проживающих в особо охраняемых и трущобных домах, неизбежны и в этом случае, как и при других технологиях опросов.)

В «паспортичку» опросника включается вопрос о фактическом составе домохозяйства (с указанием возраста и внутрисемейного статуса всех его членов). Это позволяет при обработке данных определить вероятность попадания в опрос респондентов из домохозяйств с разной численностью людей в возрастах, подлежащих опросу. (Чаще всего, это лица старше 18 лет, но при обследованиях транспортного и покупательского поведения к этой категории относились лица старше 14 лет.) Очевидно, что вероятность попадания в выборку представителей семьи, где четыре человека в возрасте, подлежащем интервьюированию, в два раза больше, чем для семьи, где таких двое, и в четыре раза больше, чем для одиноко проживающего.

Дальнейшим перевзвешиванием по признакам, которые характеризуют домохозяйство (душевой доход, численность детей разных возрастов, включенность в очередь на улучшение жи-

[1] Метод был разработан и развивается под руководством автора с 2001 года в представительных опросах населения Санкт-Петербурга, проводимых исследовательской фирмой «Крона Корсинто».

лья, проживание в отдельной, коммунальной квартире или общежитии, наличие автомобиля, стационарного телефона, участка за городом, наличие в семье человека, который не выходит из дому, «льготника» и т.п.), доля респондентов с таким значением признака легко пересчитывается в долю домохозяйств с этим же свойством.

Расчет проводится по формуле:

$$X_j = \frac{\sum_{j=1,2,\ldots}^{n} \left(n_{ij} : I \right)}{\sum_{j=1,2,\ldots}^{n} \sum_{i=1,2,\ldots}^{n} \left(n_{ij} : I \right)} \times 100 \quad , в \ \%$$

где:

X_j — процентная доля домохозяйств с признаком *J;*

I — количество членов домохозяйства у респондента, включая его самого, которые попадают в круг лиц подлежащих опросу, I = 1,2,3…;

nij — количество респондентов в опросе, которые имеют признак J и входят в домохозяйство с I числом лиц, могущих быть опрошенными.

Расчет удобнее всего проводить в матрицах JxI, в которых маргинальный столбец J в относительных величинах является искомым процентным распределением.

Представительность получаемых распределений определяется представительностью исходных данных. В то же время, рассчитанные распределения, как в относительных величинах, так и в абсолютных, могут служить эффективным индикатором для оценок внешней валидности по показателям государственной статистики и опросов с минимальной фальсифицируемостью. К таким, например, относится численность детей разных возрастов, базирующаяся на статистике рождений и детской смертности, а также ведомственной медицинской статистике детских поликлиник. Правда, ответы на вопрос о числе детей в семье и их возрасте очень редко вызывают затруднения у респондентов. Поэтому отклонение рассчитанной по результатам опроса (2007 год) численности детей в Петербурге 752 тыс. чел. от данных статистики — 742 тыс. чел., погрешность 1,3% — позволяет считать, что пропорции распределения домохозяйств по наличию детей, полученные в опросе, адекватны реальности.

С помощью этого метода пересчетов решались четыре типа задач:

1) расчет распределений «групп из лиц, включенных в генеральную совокупность опроса индивидов» при том, что выборка из генеральной совокупности таких групп (домохозяйств, небольших (до 20 человек) трудовых коллективов, соседских общин и т.п.) для исследователя недоступна, либо информации о ней нет вообще (например, выяснилось, что 14–15% опрошенных, ответивших, что они живут одиноко, составляют треть всех домохозяйств в Петербурге);

2) определение абсолютной численности указанных «групп», имеющих интересующее исследователя значение признака при том, что известна только общая численность индивидов, входящих во все эти группы вместе взятые (так выяснилось, что домохозяйств, где есть один ребенок, немногим более 0,45 млн, несколько меньше, 200 тысяч, с двумя детьми, и лишь 20 тысяч с тремя и более детьми в возрасте до 18 лет);

3) определение неизвестного объема генеральной совокупности «групп» на основе знания общей численности людей, входящих в эти «группы». Так, для определения реальной численности функционирующих в городе на момент опроса малых предприятий, сперва, по имеющимся статистическим данным определяется общее число людей, работающих на малых предприятиях (вычитанием из численности всех работающих работников крупных и средних предприятий и работников госучреждений). Далее из ответов на вопрос «Сколько людей работает в Вашей организации?» получаем оценку вероятности попадания в опросную совокупность людей, работающих в трудовых коллективах с разной численностью (прямо пропорциональную этой численности), и, по приведенной формуле, определяем численность предприятий с численностью работников, например, до 20 человек. Таким путем было показано, что из зарегистрированных в Санкт-Петербурге более чем 200 тыс. малых предприятий, реально работают немногим больше 100 тысяч, что существенно отличается от официальных цифр, на основе которых Петербург называют одним из лидеров развития малого бизнеса;

4) определение средних значений, доли и абсолютной величины объема общего признака, например, для выяснения распределения доходов между типами домохозяйств. Для этого к описанию каждого респондента добавляем новые переменные, в соответствии с задачами исследования: тип домохозяйства, в которое он входит (номинальная шкала), и общий доход семьи (умножение среднедушевого дохода, названного респондентом, на число людей, входящих в домохозяйство — количественная шкала). Потом по общей формуле рассчитывается вероятность попадания каждого респондента, как представителя типа домохозяйства, в выборочную совокупность и, с учетом полученного респондентом веса, определяется взвешенная средняя величина общего дохода домохозяйства для каждого типа. Если задачей было только выявление различия средних величин общего дохода, то на этом расчет заканчивается. Если же необходимо определить доли домохозяйств в общей массе доходов населения, то проводится по той же общей формуле расчет численностей каждого типа домохозяйств, и полученные величины умножаются на ранее рассчитанные средние величины общего дохода домохозяйств, с последующим обычным расчетом долей.

Рассмотренный метод относится к методам взвешивания (weighting) в таблицах сопряженности. Однако он отличается простотой интерпретации каждого шага и требует для своего применения минимальных расчетов.

В развитие изложенного заметим, что возможность перехода от представительного описания индивидов к представительному описанию групповых образований индивидов значительно расширяет сферу применения внешней валидизации представительности самого опроса индивидов. Это существенно в актуальной ситуации роста числа отказов от участия в опросе вообще и отказов отвечать на отдельные вопросы, что заставляет всё чаще признавать невозможность полноценной реализации случайной плановой выборочной совокупности. Гипотеза о том, что социальная структура «отказников» (unit-nonresponse и item- nonresponse) не отличается от социальной структуры участников опроса, оправдывается только в отношении части вопросов обследований, а по другим вопросам «отказники» могут отличаться по параметрам (например, уровень доходов), которые существенны для формирования распределений ответов на эти вопросы. В тоже время, говорить о прямой валидизации распределений ответов по внешним источникам для многих вопросов бессмысленно, так как, если бы информация такого рода имелась, то опрос не надо было бы и проводить. Методы же косвенной оценки позволяют хоть както оценить возможные погрешности за счет различий генеральной и выборочной совокупностей по показателям, которые являются факторами ответов на целевые вопросы обследования. Например, если надо выяснить готовность населения к страхованию квартир, то полезным было бы знать, насколько полученная по опросу доля застрахованных квартир (приблизительно равна числу домохозяйств), совпадает с реальной статистикой страхования по обследованному городу или региону. Расчеты величин, которые можно использовать для внешней валидизации, проводимые по данным опроса могут потребовать использования данных ответов по нескольким вопросам опросника. Это, конечно, делают такие расчеты более сложными, но в то же время расширяет круг одновременно валидизируемых результатов.

Социальная напряженность: концептуальная схема

Воронина Наталья Дмитриевна, *НИУ ВШЭ*

Социальная напряженность в российской и в западной литературе

Всплеск интереса к изучению социальной напряженности (СН) в российской науке возник в начале 1990-х годов. Это закономерно, так как именно в период с 1987 года в Советском Союзе нарастала волна протестов, и, что самое главное, информация об этих протестах начала транслироваться СМИ. До этого времени СН советскими авторами практически не изучалась, так как считалось, что в социалистическом обществе ее не может быть[1]. В 2000-е годы интерес к СН в

[1] Подробнее см.: Рукавишников В.О. Социальная напряженность // Диалог. М., 1990. №3. С. 6–11.

России идет на спад, вероятно, в связи с пропагандируемой государственной установкой на стабильность.

В отечественных работах под СН обычно понимается весьма широкий спектр негативных социальных явлений, сопутствующих тем переменам, которые стали происходить в общественно-политической сфере российского общества на рубеже 1980–1990-х гг, начиная от ажиотажного спроса на товары и заканчивая забастовками и акциями протеста. Эти явления можно назвать признаками, или симптомами СН.

В западной литературе редко рассматривается именно термин «СН». Как правило, соответствующие аспекты социальной жизни изучаются при анализе других, хотя и близких по смыслу явлений. Прежде всего, таким явлениям выступает аномия; именно с такой точки зрения СН объясняется, например, у Р. Мертона[1]. Также то, что у нас подразумевают под СН (недовольство, неудовлетворенность, протестные действия и др.), на Западе в основном изучается в рамках теорий общественных движений[2], коллективного действия[3] и социальных конфликтов[4]. Наконец, в западной литературе используется широко трактуемое понятие «социальный стресс», и в рамках этого направления акцент делается на психологические аспекты СН[5].

Иными словами, явления, которые в российских источниках квалифицируются как признаки СН и объединяются принадлежностью к СН, могут рассматриваться в рамках различных областей социологического знания.

Среди нескольких десятков определений СН одни авторы, которые концентрируют внимание на каком-либо одном аспекте СН (в российской литературе таким аспектом чаще всего выступает неудовлетворенность), другие подчеркивают ее многоаспектность и тем самым репрезентируют неконструктивную позицию с точки зрения измерения[6].

Поскольку конечной целью разработки концептуальной схемы СН является измерение, в качестве рабочего определения мы останавливаемся на следующем: *социальная напряженность есть реакция совокупности* (социальной группы, трудового коллектива и т.п.) *на некоторое негативное событие внешней среды*. Негативное событие внешней среды трактуется нами предельно широко: фактически, мы подразумеваем под ним все те события или явления, которые перечисляется в литературе в качестве причин социальной напряженности, «широкий круг факторов различной природы (экономических, политических, природных, национальных и т.п.)»[7]. Не менее широко мы понимаем и реакцию людей: мы относим к ней все симптомы СН, которые так или иначе описываются авторами работ по социальной напряженности (они будут перечислены ниже).

Проблемы измерения СН

Анализ литературы показывает, что основным (и практически единственным) способом является измерение выраженности явлений, которые относятся к симптомам социальной напряженности. По выраженности одного или несколько таких симптомов делается вывод об уровне социальной напряженности[8]. Такой подход имеет существенные недостатки, о которых во многих случаях упоминают и сами авторы.

Во-первых, показатели, основанные на измерении симптомов, описывают состояние ситуации на текущий момент (и только с точки зрения выраженности тех симптомов, которые изме-

---

[1] Мертон Р. Социальная структура и аномия // Социология преступности: Современные буржуазные теории. М.: Прогресс, 1966.

[2] Здравомыслова Е.А. Парадигмы западной социологии общественных движений. СПб.: Наука, 1993.

[3] Парсонс Т. Теория коллективного поведения. М.: Наука, 1972; Смелзер Н. Социология / Пер. с англ. под ред. В.А. Ядова. М.: Феникс, 1994.

[4] Зайцев А.К. Парадигмы насилия и ненасилия в конфликтологии // Социальный конфликт. 1999. № 4. С. 3–22.

[5] Давыдов А.А., Давыдова Е.В. Измерение социальной напряженности / РАН, Институт социологии. М., 1992.

[6] См. напр.: Рукавишников В.О. Указ. соч.; Пирогов И.В. Социальная напряженность: теория, методология и методы измерения: Дисс. … канд. социол. наук: 22.00.01. Иваново, 2002.

[7] Рукавишников В.О. Указ. соч.

[8] Пирогов И.В. Социальная напряженность: теория, методология и методы измерения: Дисс. … канд. социол. наук: 22.00.01. Иваново, 2002; Воронина Н.Д. Критический анализ известных способов измерения социальной напряженности // Актуальные проблемы современной социологии: Сборник статей аспирантов. М.: МАКС Пресс, 2001. Ч. 1. С. 100–108.

ряются). Измерив уровень социальной напряженности по выраженности некоторых симптомов нельзя ответить на вопрос о том, как будет развиваться ситуация дальше: то ли выраженность симптомов пойдет на спад, то ли наоборот, усилится, то ли сами симптомы изменятся: например, на смену неудовлетворенности придут протестные действия. Помимо этого, сами измерительные процедуры, предлагаемые авторами, зачастую несовершенны[1].

Во-вторых, остается без ответа вопрос о том, почему в случаях, когда происходит одно и то же негативное событие, реакция совокупности может быть разной. Например, почему в ответ на повышение коммунальных платежей в одних случаях недовольство выражается «на кухне», а в других — в протестных действиях?

Отсюда видится необходимость построить концептуальную схему явления СН, на основании которой можно было бы проводить измерение. Схема должна отражать факторы, определяющие форму и силу выраженности симптомов социальной напряженности, которые, в свою очередь, являются реакцией совокупности на неблагоприятные события внешней среды (причины СН), см рис. 1.

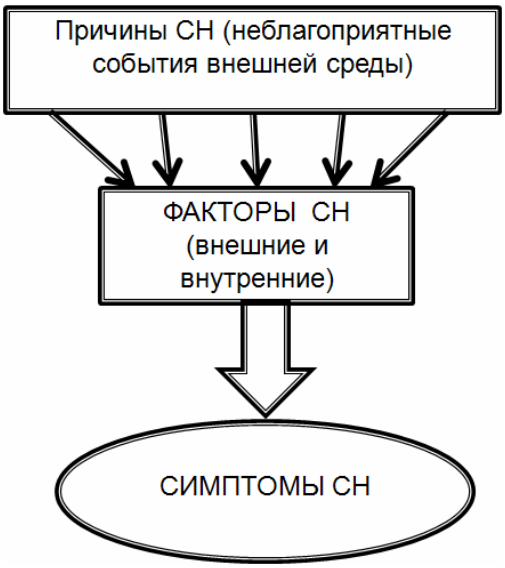


Рис. 1. Соотношение причин, симптомов и факторов СН.

Представляется, что эти факторы делятся на 2 группы: внутренние, присущие индивиду, и внешние — факторы микро- и макросреды. В результатах измерения, основанного на такой схеме, должна быть заложена возможность прогноза симптомов СН, которые проявятся при наступлении неблагоприятного события в микро- и макросреде (появления причины СН), и, как максимум, — прогноза вероятности структурных изменений в совокупности в результате реакции на породившие СН события. С этой точки зрения, наиболее удобной формой результата измерения будет вероятностная форма: итоговый показатель СН будет вероятностью (набором вероятностей) проявления СН в виде тех или иных симптомов.

Концептуальная модель СН

Выше, давая рабочее определение СН, мы уже упоминали, что различные неблагоприятные события внешней среды, называемые нами причинами СН, вызывают некоторую реакцию совокупности в форме симптомов социальной напряженности. Задача концептуальной схемы — выделить факторы, которые определяют форму и выраженность симптомов СН в совокупности.

Для выделения внутренних факторов нами был произведен анализ симптомов социальной напряженности, указанных различными авторами. В приведенной ниже *таблице* сведены все возможные симптомы социальной напряженности, которые встречались нам в литературе, в тех формулировках, в которых они были употреблены авторами. Среди них встречаются как те, которые тесно связаны с социальной напряженностью (например, протестные действия), так и те,

[1] Воронина Н.Д. Критический анализ известных способов измерения социальной напряженности // Актуальные проблемы современной социологии: Сборник статей аспирантов. М.: МАКС Пресс, 2001. Ч. 1. С. 100–108.

которые связаны менее тесно и имеют множество причин своего возникновения помимо социальной напряженности (например, уровень самоубийств и т.п.).

Внутренние факторы, определяющие систему симптомов СН

При анализе перечисленных симптомов обращает на себя внимание тот факт, что практически все из них так или иначе связаны с фрустрацией, являются, можно так сказать, реакцией индивидов на фрустрацию, вызванную негативным внешним событием. Некоторые авторы[1] рассматривают фрустрацию как важнейшую составляющую социальной напряженности. На основании этого первым фактором, определяющим выраженность СН (под выраженностью условимся понимать конкретные симптомы и их уровень в сообществе), может являться фрустрация. Такие признаки СН, как протестные действия и массовые действия деструктивного характера, насилие в каком-то смысле стоят особняком, как наиболее опасные и разрушительные по своим последствиям. Следовательно, необходимо обратить внимание на факторы, которые связаны с вероятностью проявления именно этих симптомов.

Возникновение вышеуказанных симптомов может в частности, определяться тем, сформирован ли в сознании индивида образ того, кто, условно говоря, «виноват» в сложившейся неблагоприятной ситуации, или кто не совершает необходимых действий по преодолению этой ситуации, или препятствует ее разрешению, или кого-то, кто воспринимается как несущий угрозу, опасный. При наличии кого-то, против кого можно предпринимать какие-то действия, социальная напряженность с большей вероятностью будет выражаться в протестных действиях, при отсутствии такового более вероятны будут иные признаки социальной напряженности. Добавим, что С. Тэрроу[2] в компонентах протестного действия отмечает его обращенность на другие группы и элитные группировки.

Следовательно, вторым фактором, определяющим форму, в которой выразиться социальная напряженность, будет являться наличие четко сформированного образа «виноватого». Мы считаем, что образ «виноватого» должен быть физическим лицом (группой лиц), предельно конкретным и доступным для каких-то действий. Так, например, умерший человек может быть виноватым («Во всем виноват Ельцин»), но протестные действия по его адресу будут бессмысленными. Также не будет способствовать росту вероятности проявления социальной напряженности в форме протестных действий неодушевленные объекты («менталитет», «традиции») и обращенность вины на самих себя («мы сами виноваты, что таких выбрали»).

Третий фактор — внутренняя готовность индивидов и групп участвовать в протестных, насильственных действиях; на это обращает внимание, в частности, Ю.М. Плюснин[3]. Отсутствие готовности действовать повышает вероятность «психологических» симптомов социальной напряженности — депрессий, подавленного состояния, апатии и т.п. Готовность к действиям может выражаться в стремлении избежать, уйти от неблагоприятной ситуации; в этом случае, мы предполагаем, будет повышаться вероятность миграции. Еще одна форма — готовность к деликвентному, криминальному поведению. Наконец, последней формой готовности к действиям может служить гражданская активность, под которой мы подразумеваем готовность принимать участие в общественных движениях, в протестных действиях, защищать свои права с помощью социальных институтов и т.п.

Помимо трех вышеперечисленных факторов представляется необходимым рассмотреть еще один, связанный с положением концепции мобилизации ресурсов, разработанной в рамках теорий коллективных действий, которое говорит о том, что общественные движения являются не просто индикатором социальных патологий, но представляют собой осознанную деятельность участников, направленную на достижение конкретных целей[4].

---

[1] См., например: Бородкин Ф.М., Володина Н.П. Социальная напряженность и агрессия // Мир России. 1997. Т. 6. №4. С. 107–150.

[2] Tarrow S. Democracy and Disorder: Protest and Politics in Italy 1965–1975. Oxford: Clarendon Press, 1989.

[3] Плюснин Ю.М. Социальная напряженность в Новосибирске. 1999 год. Новосибирск: ЦСА, 1999.

[4] Цепилова О.Д. От социальной проблемы к коллективному действию: экологические движения в районах экологической опасности // Общественные движения в современной России: от социальной проблемы к коллективному действию. М.: ИС РАН, 1999. С. 101–118.

Таблица

Симптомы социальной напряженности

Связанные с демографической и миграционной сферами	**Связанные с политической сферой**	**Связанные с агрессией**	**Связанные с психологической сферой**	**Иные**
Активизация миграционных процессов	Протестные действия (во всем спектре форм)	Агрессия	Стрессы	Аномия
Снижение рождаемости	Декларируемая готовность к протестным действиям	Насилие	Психическая усталость	Недовольство
Увеличение числа разводов	Массовые действия деструктивного характера (социальный взрыв, гражданская война, революция)	Девиантное поведение (в т.ч. алкоголизм, наркомания)	Раздражительность	Изменения социального самочувствия
Повышение уровня смертности	Активизация социально-политических движений	Изменения уровня преступности	Тревожность	Падение дисциплины
Изменения доли самоубийств	Изменение уровня политической активности населения	Уровень конфликтности	Апатия	Негативные лингвистические символы в текстах масс-медиа
	Утрата доверия к властям		Отсутствие ощущения безопасности	Ажиотажный спрос
			Пессимистические оценки будущего	Снижение сплоченности, обособленность членов совокупности
			Панические настроения	Разрыв хозяйственных связей
			Ощущение дискомфорта	Ухудшение показателей экономической деятельности
				Социальная дезинтеграция

Некоторые авторы обращают внимание на своеобразную «энергетическую» составляющую социальной напряженности[1]. Представляется, что фактором, сочетающим эти стороны, может служить понятие пассионарности, введенное Л.Н. Гумилевым в рамках его этнологических

[1] Губина Н.В. Социальная напряженность в трудовом коллективе // Социс. 1998. № 11. С. 17–25.

концепций. Сразу отметим, что термин пассионарность мы в данном случае используем вне его этнологического контекста, в котором, прежде всего, его использовал Л.Н. Гумилев и который является дискуссионным. Нас интересует возможная «социологическая» сторона этого явления. Пассионарность, по Л.Н. Гумилеву[1], — это особое свойство характера людей, необоримое внутреннее (осознанное или неосознанное) стремление к деятельности, направленной на осуществление какой-либо цели, зачастую иллюзорной.

Высокая пассионарность, связанная с активной деятельностью по достижению цели, может быть связана с такими проявлениями СН как протестные и деструктивные действия, участие в общественных движениях, а низкая пассионарность — с апатией, социальной дезинтеграцией, ростом разобщенности и т.п.

Таким образом, можно выделить 4 внутренних фактора, связанные с формой и силой выраженности симптомов социальной напряженности: фрустрация, сформированный образ «виноватого», готовность действовать, пассионарность. Мы предполагаем, что вышеуказанные факторы действуют совместно, следовательно, для определения симптомов социальной напряженности необходимо рассматривать сочетания значений этих факторов.

Различные сочетания значений внутренних факторов формируют специфические «типы» личности, для каждого из которых наиболее вероятен тот или иной вариант реакции на негативное событие внешней среды. Распределение долей этих типов в изучаемой совокупности будет служить оценкой вероятности проявления того или иного симптома социальной напряженности. Итоговым результатом измерения по описанной концептуальной схеме будет являться набор вероятностей проявления каждого из симптомов социальной напряженности.

Подходы к измерению внутренних факторов СН

Отдельно рассмотрим вопрос об измерении перечисленных факторов. Фрустрацию индивида можно замерять любым из многочисленных разработанных для этой цели тестов, однако при условии применения в массовых опросах, этот тест должен быть по возможности кратким. Определение «образа виноватого» будет основано на социально-психологическом подходе к измерению угрозы групп, т.е. определения индивидов или групп, которые несут опасность для респондента, являются неприятными, враждебными и т.п. (Методика в настоящий момент разрабатывается автором). Измерение готовности к действиям будет основано на приминении методики проективных ситуаций. Самым сложным моментом во всей схеме является измерение пассионарности. Автором была разработана специальная методика измерения пассионарности, основанная на приминении латентно-структурного анализа Лазарсфельда[2].

Внешние факторы, определяющие систему симптомов СН

Вкратце рассмотрим вторую группу факторов, которые мы назвали внешними. В рамках данной статьи мы ограничимся их перечислением. Отметим, что это перечисление является результатом сведения воедино всех факторов микро- и макросреды, так или иначе указывавшихся в различных работах по социальной напряженности в качестве влияющих на уровень социальной напряженности и на тип и интенсивность проявления ее симптомов. К таковым могут быть отнесены следующие характеристики:

— наличие в совокупности некоторой организованной структуры, которая может аккумулировать и направлять негативную реакцию людей на неблагоприятные события (политическая партия, общественное движение, правозащитные организации и т.п.);

— наличие в совокупности харизматического лидера, способного организовать проявления негативной реакции;

— СМИ, и как источник, из которого становится известно о негативных событиях внешней среды, и как транслятор идеологических установок в совокупности. Специфическим СМИ является интернет, пока неподконтрольный государственным структурам, так как помимо указанных функций играет роль площадки для общения;

[1] Гумилев Л.Н. Этногенез и биосфера земли. М.: Танаис ДИ-ДИК, 1994.

[2] Воронина Н.Д. Использование ЛСА для измерения пассионарности // Современные проблемы формирования методного арсенала социолога: Материалы Всероссийской научной конференции памяти А.О. Крыштановского. М.: ГУ–ВШЭ, 2009.

— господствующие в совокупности традиции;

— господствующая в совокупности культура;

— принятая в совокупности идеология или ее отсутствие;

— особенности политической и управленческой систем, выражающиеся в степени терпимости по отношению к проявлениям СН (напрмер, к протесту) и в способности контролировать симптомы социальной напряженности (например, уровень преступности, миграционные процессы, и конечно, протестные действия);

— ближайшее социальное окружение индивида: семья, референтные группы и т.п.;

— ценностные ориентации индивида, хотя и являются внутренними качествами индивида, однако рассматриваются отдельно, потому что оказывают влияние на все стороны его жизни, а не только на то, что связано с реакцией на неблагоприятные события внешней среды.

Влияние внешних факторов является опосредованным: они могут изменять значения внутренних факторов, что, в свою очередь, изменит тот тип, к которому на данный момент принадлежит индивид, и как следствие, изменится долевое соотношение типов в изучаемой совокупности и итоговое распределение вероятностей проявления каждого из симптомов социальной напряженности.

На рис. 2 приведена концептуальная схема социальной напряженности, в которой отражена модель связи уже перечисленных внутренних и внешних факторов и симптомов социальной напряженности.

Рис. 2 Модель связи внутренних и внешних факторов и симптомов СН.

Подводя итог, можно сказать, что СН является многоаспектным, многофакторным явлением. Измерение такого рода явлений не должно ограничиваться замером текущих симптомов, а требует разработки концептуальных моделей, связывающих воедино факторы, лежащие в их основе, и позволяющие разрабатывать адекватный измерительный инструментарий.

Изучение ресурсной обеспеченности российских школ с помощью методов, основанных на решетках понятий

Игнатов Дмитрий Игоревич, *НИУ ВШЭ*
Хавенсон Татьяна Евгеньевна, *НИУ ВШЭ*

В области изучения образования (educational studies) существует несколько крупных международных проектов по изучению качества школьного образования. Один из этих проектов — международное исследование по оценке качества математического и естественнонаучного образования TIMSS (Trends in Mathematics and Science Study), осуществляемое Международной Ассоциацией по оценке учебных достижений IEA (International Association for the Evaluation of Educational Achievements). Исследование проводиться с 1995 года, раз в 4 года, в исследовании принимают участие около 60 стран мира (волна 2007 г.). Изучается подготовка выпускников начальной школы (4 класс) и учащихся 8 классов.

По результатам этих исследований оценивается качество общего образования в стране в целом, ее положение относительно других стран, изучается ход образовательных реформ и т.д. Помимо оценки знаний школьников в исследовании собирается большое количество контекстной информации как по ученикам (социально-экономическое положение семьи, отношение к учебе и т.п., проводится также опрос родителей), так и по школам (оснащенность школы различными ресурсами важными для учебного процесса, роль различных органов власти, администрации, учителей и родителей в жизнедеятельности школы, правила и принципы существования школы, школьный климат и т.д.). На вопросы анкет второй группы отвечает директор или другой администратор школы.

В нашей работе мы исследуем, насколько школам хватает или не хватает различных ресурсов для ведения образовательной деятельности. Ключевой вопрос: влияет ли на учебный процесс несоответствие современным требованиям или недостаточное количество… (варианты ответа: Совершенно не влияет / Мало влияет / Влияет в некоторой степени / Влияет значительно). Список ресурсов: канцелярские товары; учебные материалы; компьютеры и компьютерное программное обеспечение; школьные здания и территории; системы отопления и освещения; аудиовизуальные пособия; книги в библиотеке и др.

Обычно на основе этих вопросов строят индекс ресурсообеспеченности, который принимает три значения: высокий, средний и низкий уровень и смотрят связь с достижениями. Достаточное количество исследований показывает, что в большинстве стран нет сильной связи между ресурсами школы и достижениями учеников.

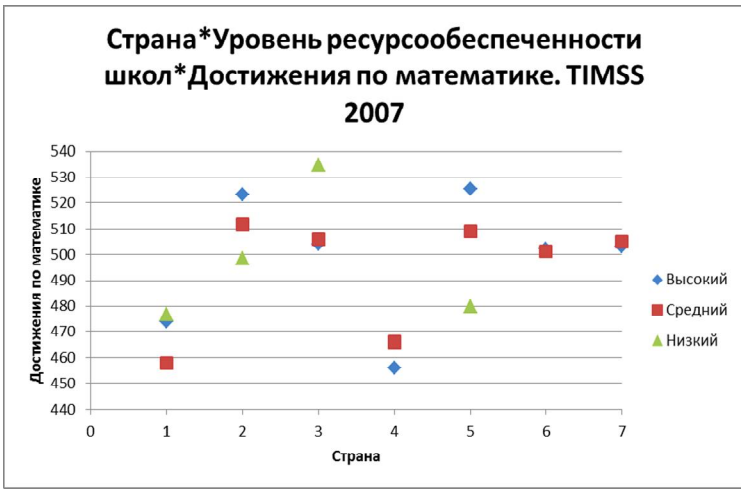
Например, график на рисунке 1 показывает распределение достижений в школах с разным уровнем ресурсообеспеченности в некоторых странах Восточной Европы и постсоветских республиках[1]. Видно, что особой связи нет. В противном случае, ученики хорошо обеспеченных школ показывали бы результаты выше, чем ученики среднеобеспеченных, а те в свою очередь выше, чем ученики плохо оснащенных школ. Мы наблюдаем такую картину только для Венгрии и России, тогда как сравнение средних значений показывает, что статистически эти различия незначимы. Однако в описании TIMSS 2007 сказано, что в целом учащиеся школ с более высоким индексом оснащенности имеют лучше результаты[2].

Оба приведенных выше примера рассматривают связь на агрегированном уровне, во-первых, по странам, во-вторых, ответы респондентов об обеспечении ресурсами сведены в один индекс. Нам представляется интересным выяснить, существуют ли некоторые ресурсы, являющиеся более значимыми для реализации учебного процесса. Нам кажется, что для решения этой задачи необходимо исследовать исходные данные, то есть ответы респондентов на вопросы о каждом ресурсе в отдельности, и выделить среди них важные ресурсы и/или их сочетания. Наше иссле-

[1] Тюменева Ю.А., Хавенсон Т.Е. Тренды некоторых показателей образовательных систем в странах Восточной Европы и достижения школьников в TIMSS и PISA: Доклад на семинаре «Актуальные исследования и разработки в области образования», ИРО ГУ–ВШЭ, ноябрь 2010.

[2] Mullis I.V.S, Martin M.O., Ruddock G.J., O'Sullivan C.Y., Arora A., Erberber E. TIMSS 2007 Assessment Frameworks. Boston: TIMSS & PIRLS International Study Center et al., 2005. P. 86.

дование проходит в два этапа. Сначала мы исследуем данные об обеспечении ресурсами агрегировано по странам, а затем только данные по российским школам, где каждый показатель описывает школу. Таким образом, мы сможем выявить группы стран или школ, которые испытывают или не испытывают проблем с одними и теми же ресурсами, а также найти сочетания ресурсов, нехватка которых сопряжена друг с другом.



1	Болгария
2	Венгрия
3	Латвия
4	Румыния
5	Россия
6	Словения
7	Чехия

Рис. 1. Связь между уровнем оснащенности школы и достижениями по математике в TIMSS 2007.

Метод исследования

Предлагаемый нами подход основан на анализе формальных понятий (АФП), алгебраической дисциплине представляющей собой прикладную ветвь теории решеток и нашедшей широкое применение в анализе объектно-признаковых данных[1].

АФП и ранее применялся в социологических исследованиях, в частности, в анализе социальных сетей с помощью решеток понятий[2], при выявлении эпистемических сообществ и построение их таксономии[3], для изучения групп посетителей Интернет-ресурсов[4], в анализе опросных данных[5].

Суть применения АФП состоит в том, что в полученных результатах опросов респонденты могут быть рассмотрены как объекты, а ответы на вопросы анкет как признаки, которыми они обладают. По этим данным выявляется множество групп объектов, обладающих общими признаками. Далее на множестве таких пар вида (объекты, признаки), называемых формальными понятиями, задается частичный порядок по отношению вложения первой компоненты. Это отношение является частичным порядком и определяет т.н. решетку понятий. Граф покрытия этого отношения удобно использовать для визуализации выявленных групп, что дает возможность исследователю сделать выводы об их размерах, пересечениях, общих признаках и наличии некоторых других закономерностей. Помимо выделения групп респондентов и их визуализации,

[1] Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. Berlin: Springer, 1999.

[2] White D.R., Duquenne V. Social network & discrete structure analysis: Introduction to a special issue // Social Networks. 1996. Vol. 18. P. 169–172; Freeman L. Cliques, Galois lattices, and the structure of human social groups // Social Networks. 1996. Vol. 18. P. 173–187.

[3] Roth C., Obiedkov S., Kourie D.G. Towards Concise Representation for Taxonomies of Epistemic Communities // LNCS. 2008. Vol. 4923. p. 240–255.

[4] Кедров С.А., Кузнецов С.О. Исследование групп пользователей Интернет-ресурсами методами анализа формальных понятий и разработки данных (Data Mining) // Бизнес-информатика. 2007. №.1. С. 45–51; Kuznetsov S.O., Ignatov D.I. Concept Stability for Constructing Taxonomies of Web-site Users // Proc. Satellite Workshop «Social Network Analysis and Conceptual Structures: Exploring Opportunities» at the 5th International Conference Formal Concept Analysis (ICFCA'07). Clermont-Ferrand, 2007. P. 19–24.

[5] Игнатов Д.И., Кононыхина О.Н. Решетки формальных понятий для анализа данных социологических опросов // Интегрированные модели и мягкие вычисления в искусственном интеллекте: Сборник научных трудов V Международной научно-технической конференции (Коломна, 20–30 мая 2009 г.). В 2-х томах. Т 1. М.: Физматлит, 2009.

АФП предоставляет возможность поиска признаковых зависимостей в виде импликаций, что позволяет делать выводы о взаимосвязи исследуемых социальных явлений.

Анализ формальных понятий используется для обнаружения и исследования изначально присутствующей в данных информации. В терминах теории формальных понятий сформулированы разнообразные методы анализа данных, такие как поиск ассоциативных правил и машинное обучение на основе положительных и отрицательных примеров и др.

В АФП предпринята попытка математически формализовать наиболее простую единицу человеческого мышления — понятие. «Формальное понятие» (formal concept) характеризуется объемом и содержанием. Объем понятия — это множество объектов, которые подпадают под понятие, а содержание — общее описание всех таких объектов.

Исходные объектно-признаковые данные в АФП описываются с помощью формальных контекстов. Запись K= (G, M, I) обозначает формальный контекст (formal context), где G — изучаемые объекты (в нашем случае школы или страны), M — их свойства (в нашем исследовании используемые ресурсы), и I — бинарное отношение между ними (установление факта, что определенный объект обладает определенным свойством, например, есть или нет в школе/стране нехватка данного ресурса)[1]. **Формальный контекст** может быть представлен в виде объектно-признаковой таблицы.

Таблица 1

Формальный контекст «Обеспеченность ресурсами школ по странам за 2007 год»

id	Страна	Учебные мат.	Канц. товары	Здания, террит.	Отопление	Помещения для занятий	Аудио-виз. ресурсы	Компьютеры	ПО	Книги в библ.
1	Литва	1	1	1	1	1	1	1	1	1
2	Россия	0	0	0	1	0	1	1	1	0
3	Болгария	0	1	1	0	0	1	1	1	1
4	Венгрия	0	1	0	0	0	1	1	1	0
5	Румыния	1	1	1	0	0	1	1	1	1
6	Словения	0	0	0	0	0	1	1	1	0
7	Чехия	0	0	0	0	0	1	1	1	1

Формальное понятие формального контекста (*G, M, I*) — это пара (*A, B*), где *A* — это подмножество объектов из *G* (т.е. некоторая часть исходной выборки), а *B* — это подмножество признаков из *M*, при этом *A* — это в точности все объекты, обладающие всеми признаками из *B*, а *B*, в свою очередь, — в точности множество всех признаков, которыми обладают все объекты из *A*. *A* называется объемом формального понятия, *B* — содержанием.

В нашем случае формальное понятие группирует (в своем объеме) страны или школы, которые испытывают проблемы с нехваткой одних и тех же ресурсов, и (в своем содержании) те ресурсы, по которым нехватка у разных стран/школ одинакова.

Решетка понятий. Понятия, упорядоченные по отношению «быть более общим понятием чем»[2], образуют алгебраическую структуру, называемую *решеткой*. Решетку понятий обычно изображают при помощи линейной диаграммы (диаграммы Хассе): (1) более общие понятия помещаются над менее общими; (2) два понятия соединяются линией, если одно из них является более общим, чем другое и нет понятия, которое было бы одновременно менее общим, чем первое понятие, и более общим, чем второе понятие. Узлы диаграммы часто снабжаются метками, с помощью которых можно оценить размеры объема и содержания понятий, понять какие объекты и признаки образуют понятие и т.п.

[1] Wille R. Communicative Rationality, Logic, and Mathematics // Medina R., Obiedkov S. (eds.). Formal Concept Analysis – 6th International Conference, ICFCA 2008, Montreal, Canada, February 25-28, 2008, Proceedings. Vol. 4933. Berlin; Heidelberg: Springer, 2008. P. 1–13; Ganter B., Wille R. Applied Lattice Theory: Formal Concept Analysis. Dresden: TU Dresden, 1997. P. 1.

[2] Формальное понятие (*A*, *B*) является *более общим*, чем понятие (*C*, *D*), если *C* является собственным подмножеством *A*, то есть под понятие (*A*, *B*) попадают все объекты, которые подпадают под (*C*, *D*) и некоторые другие.

На рис. 2 каждый узел диаграммы — это формальное понятие, названия стран под узлом показывает, что данная страна (объект) принадлежит объему этого понятия и всем более общим понятиям (когда объектов много тут отображается число объектов в понятии, но не перечисляются их имена). Содержание понятия включает в себя признаки, указанные при соответствующем узле на диаграмме, а также при всех узлах, расположенных ниже на диаграмме решетки (т.е. при более частных понятиях). Самый верхний узел диаграммы — формальное понятие, включающее в себя все множество объектов G (как правило, с пустым содержанием, так как нет такого признака, которым бы обладали все объекты).

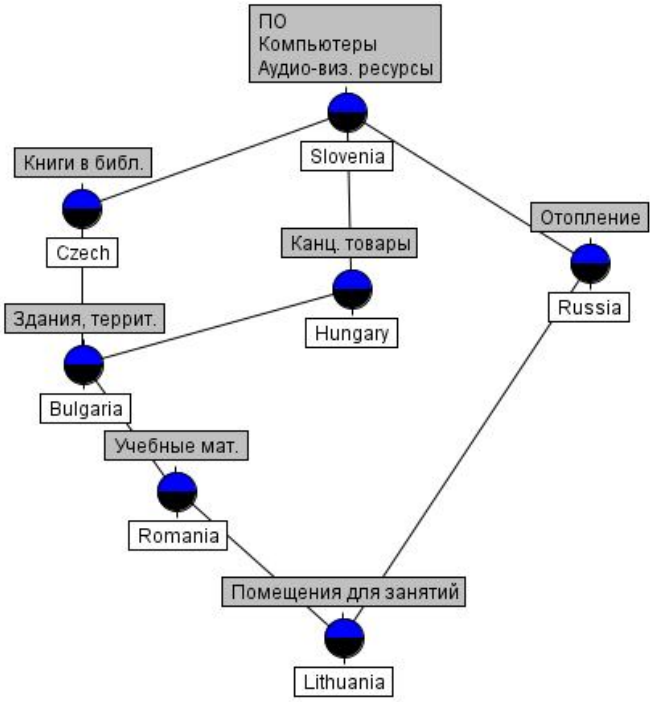


Рис. 2. Пример диаграммы решетки понятий с сокращенными пометками. Данные за 2003 год

Довольно очевидно, что даже для небольшого формального контекста количество порожденных формальных понятий может быть значительным (например, $2^{|G|}$ или $2^{|M|}$ в худшем случае). Существуют различные способы отбора наиболее устойчивых и релевантных понятий, которые упрощают работу с данными, в частности облегчают визуализацию. При этом, оставляя только устойчивые понятия, мы отбираем только наиболее релевантные понятия, устойчивые к шуму. Наиболее устойчивые понятия отбираются на основе порога на индекс устойчивости понятия (число от 0 до 1, задаваемое аналитиком). Его можно выразить через вероятность сохранения содержания понятия при удалении из его объема одного или нескольких объектов. То есть, понятие устойчиво, если его объем не сильно зависит от подмножеств своих объектов.

Также важным аспектом в применении метода является процесс преобразования исходных данных в дихотомические; такой процесс в терминах АФП называется шкалированием. Вопрос о том, каким образом шкалировать данные, является скорее содержательным, чем техническим. Также необходимо учитывать тип исходных данных (номинальные, порядковые, и др.) и другие свойства, которые важны исследователю. В данном случае, мы преобразовали исходную 4-балльную шкалу в дихотомическую. Если респондент отвечал, что нехватка определенного ресурса «Совершенно не влияет» на качество учебного процесса в школе, то ему присваивалось значение 0, если отвечал «Мало влияет», «Влияет в некоторой степени» или «Влияет значительно», — значение 1. Таким образом, мы не учитывали изначально порядковую природу данных, но зато получили простые и легко интерпретируемые результаты. Мы полагаем, что в случае применения данной группы разведывательных методов анализа данная такая стратегия от простого к сложному наиболее приемлемая.

В качестве программного средства использовалась программа Concept Explorer, дополненная внешними утилитами для преобразования данных, шкалирования и вычисления устойчивости.

Предварительные результаты

Опишем первые результаты на примере данных за 2007 год[1] (рис. 3 и таблица 1). В связи с небольшим количеством объектов и признаков, итоговая решетка содержит только 7 формальных понятий, поэтому они все отображены на диаграмме.

Самое верхнее понятие диаграммы – ({Словения}, {ПО, Компьютеры, аудио-визуальные ресурсы}), то есть Словения из всех изучаемых стран наиболее благополучна, словенские школы в целом испытывают затруднения только с этими тремя типами ресурсов. То, что эти ресурсы расположены сверху, в самом общем понятии, означает, что эти признаки свойственны[2] всем рассматриваемым странам. Спускаясь вниз по диаграмме, можно заметить, что далее страны делятся на три типа: первый — Россия, в которой к перечисленным выше трем проблемам добавляется проблема с отоплением, второй — Чехия, с нехваткой книг в библиотеках, и третий — Венгрия, с проблемой обеспечения школ канцелярскими товарами. Далее Венгрия и Чехия объединяются в понятие, которое формально можно записать так: ({Словения, Венгрия, Чехия, Болгария}, {ПО, компьютеры, аудио-визуальные ресурсы, книги в библиотеках, канцелярские товары}).

Все признаки «скапливаются» в содержании самого нижнего понятия, в объеме которого только одна страна — Литва. Это говорит о том, что в Литве наблюдается самая неблагоприятная картина с точки зрения ресурсной обеспеченности школ. Для сравнения, в 2003 году таких стран было значительно больше: Словакия, Румыния, Болгария, Россия и Литва.

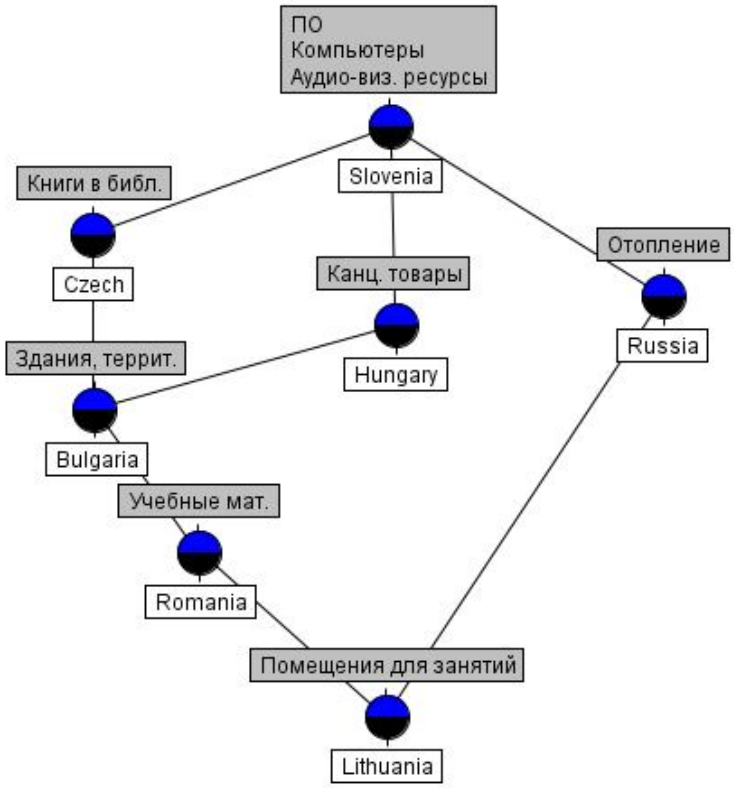


Рис. 3. Решетка понятий для контекста о ресурсной обеспеченности школ по странам в 2007 г.

Опишем дальнейшие возможности исследования с применением методологии АФП. После выделения таких классов объектов, можно продолжать интерпретацию, описывая их с помощью других признаков (переменных). Причем, при большем количестве объектов, классы по-

[1] Диаграмма для 2003 года изображена на рис. 2.

[2] Напомним, что фраза «объект обладает таким-то признаком» или «объекту свойственен такой-то признак» в нашем случае означает, что школа испытывает затруднения с данным ресурсом.

лучаются более наполненными, а их интерпретация становится более сложной. Можно сказать, что возникают две взаимосвязанные классификации объектов (стран) и признаков (ресурсов).

Следующим этапом нашего исследования будет рассмотрение российских школ в качестве объектов и несколько расширенного списка ресурсов. Мы предполагаем использовать техники отбора релевантных понятий по размеру объема понятия, индексу устойчивости и некоторые другие. Для поиска «сильных» признаковых зависимостей предполагается использовать т.н. импликации — признаковые зависимости вида A→B (наличие признаков A у объектов влечет наличие признаков B).

Таким образом, нами продемонстрирована возможность проведения социологического исследования с помощью средств Анализа Формальных Понятий. Полученные диаграммы и выводы свидетельствует о том, что данный метод и поддерживающие его программные средства потенциально полезны при анализе социологических данных опросного характера.

Социологическое измерение инновационной активности с использованием психофизиологических шкал

Кутенков Рудольф Петрович,
Институт аграрных проблем РАН

Предлагается единый методологический подход к математическому моделированию и социологическому оцениванию процессов распространения инноваций. Известно, что социологическое измерение можно определить как многоуровневый процесс, который включает представление изучаемого объекта в виде системы одномерных характеристик (индикаторов), доступных количественной оценке, последующее их измерение и объединение полученных значений с целью получения оценки объекта в целом. Представляет интерес, что, несмотря на наличие общетеоретических разработок по теории измерений (квалиметрии[1]), измерения в конкретных областях знаний имеют различный уровень теоретической обоснованности и практической значимости.

В частности, методология и методика социологических измерений, хотя и не представляет к настоящему времени завершенной теории (этот вопрос достаточно подробно обсуждался в трудах докторов наук Ю.Н. Толстовой, Г.Г. Татаровой[2]), но дает ряд хорошо зарекомендовавших себя на практике рецептов построения оценок с помощью шкал, графов, проективных технологий и других специфических приемов[3].

Близкие по содержанию вопросы рассматриваются в теории психофизиологических измерений[4]. Используемые в этой теории подходы, основанные на использовании лингвистических переменных (переменная, которая может принимать значения понятий (фраз) и использоваться при описании объектов и явлений с помощью нечетких множеств[5]) и шкал, устанавливающих соответствие между значениями лингвистической переменной и некоторых психофизиологических параметров (чувств) коммуникантов (респондентов) по сути аналогичны описанным выше подходам социологических измерений. Лингвистические шкалы и их оцифровки используются в последнее время достаточно широко, в том числе, в областях, близких к социологии. В числе сфер применения, как показывает анализ, можно отметить методы оценки уровня социально-экономического развития региона, оптимизации маркетингового бизнеса, оценки инвестиционных проектов, построения интегрального показателя конкурентного статуса предприятия, формирования универсальной шкалы оценки уровня его экономической безопасности и оценки добротности социального партнерства. Несмотря на столь широкие пересечения, методы пси-

[1] Калейчик М.М. Квалиметрия: Учебное пособие. М.: МГИУ, 2007.

[2] См., например, Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М. Научный мир, 2000; Татарова Г.Г. Методология анализа данных в социологии. М.: NOTA BENE, 1999.

[3] См., например, Ядов В.А. Стратегия социологического исследования: Описание, объяснение, понимание социальной реальности. М.: Добросвет, 2003.

[4] Безруких М.М., Фарбер Д.А. Психофизиология: Энциклопедический словарь. М.: ПЭР СЭ, 2006.

[5] Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990.

хофизиологических исследований к настоящему времени не нашли широкого применения в социологии. Как представляется, это связано с определенной замкнутостью социологов в рамках социологических методов анализа данных и недостаточной осведомленностью об уровне развития инструментария смежных дисциплин.

Чтобы продвинуться к устранению отмеченного методологического недостатка, мы в данной работе обосновываем возможность применить к задачам социологического оценивания инновационной активности одну из основных психофизиологических шкал — шкалу Харрингтона.

Рассмотрим типичную шкалу для оценки выраженности какого-либо свойства респондента (или исследуемого объекта) со стандартным набором возможных ответов: «очень слабо – слабо – средне – сильно – очень сильно». Возможны и другие наборы ответов, связанные, например, с противопоставлением «важно – неважно», «плохо – хорошо» и др. В зависимости от задачи в построенную шкалу можно ввести вариант ответа «затрудняюсь ответить», или отсеять респондентов, не имеющих мнения по этому вопросу, с помощью фильтра.

В социологии подобная шкала называется шкалой с полным описанием альтернатив, применительно к теории психофизиологических измерений может интерпретироваться как лингвистическая, по отношению к измерительным возможностям представляется не выше ранговой. Оцифровка подобной шкалы (приписывание числовых меток каждому вербальному значению признака с целью обеспечения дальнейшего использования методов статистической обработки, рассчитанных на количественные данные), как известно, теоретически обоснована лишь применительно к определенному классу оптимизационных задач[1].

Во всех других случаях переход от языковых обозначений градаций к числовым допустим лишь на основе «правдоподобных рассуждений», конкретизирующих тип шкалы с учетом специфики решаемой задачи, объекта исследования, механизма формирования оценки респондентом. При построении шкал на основе теории нечетких множеств возможны следующие допущения. Искомая шкала представляется как объединение непересекающихся интервалов, целиком покрывающих отрезок [0,1] числовой оси. Каждой вербальной градации ставится в соответствие один и только один интервал. Длины интервалов различны, что связано с особенностями психологического восприятия шкалы применительно к исследуемому свойству. Например, область оценки «средне» воспринимается более протяженной, чем области оценок «слабо» и «сильно», так как области последних ограничены соседними градациями «очень слабо» и «очень сильно». Учет отмеченной психологической особенности повышает обоснованность применяемого метода измерений.

Далее мы покажем, что предлагаемая модель измерения инновационной активности идентична модели распространения (диффузии) инноваций или, более общо, модели процессов, отображающих динамику роста кумулятивного значения показателя. Развитие подобных процессов характеризуется медленным ростом на старте и в конце и более быстрым ростом на средних этапах, что хорошо описывается функцией Гомперца, которая может иметь вид[2]:

$$y(t) = Le^{-ae^{-t}}, \quad L,a>0,\ t>0,$$

где $y(t)$ — например, кумулятивная функция распространения инновации во времени (суммарное число внедрений, или доля от гипотетически возможного общего числа внедрений конкретного инновационного технологического решения к моменту t, общий объем (или доля) продукции, произведенной к моменту t на основе инновации, доля рынка, охваченная исследуемой инновацией и т.п.). L — максимальное значение функции y (можно интерпретировать как, например, емкость рынка для конкретной инновации, или, соответственно, 100 % в случае измерения долей), a — параметры, который конкретизирует вид функции $y(t)$ и оцениваются по специальным методикам.

Функция Гомперца со значениями параметров L, $a=1$ была использована Харрингтоном для оцифровки рассматриваемых шкал оценки выраженности свойства. Введенная им в 1965 г. функция желательности устанавливала соответствие между вербальными и шкальными оцен-

[1] Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. М.: Статистика, 1985. Т. 3.

[2] Плотницкий Ю.М. Модели социальных процессов: учебное пособие. М.: Логос, 2001.

ками интенсивности (желательности) в соответствии с определенной методикой[1]. В наиболее распространенных случаях вербальным оценкам устанавливаются следующие числовые интервалы:

«очень слабо» («очень плохо»)	[0,00–0,20)
«слабо» («плохо»)	[0,20–0,37)
«средне» («удовлетворительно»)	[0,37–0,63)
«сильно» («хорошо»)	[0,63–0,80)
«очень сильно» («очень хорошо»)	[0,80–1,00]

Из приведенного соответствия видно, что оценке «средне» соответствует интервал наибольшей длины — от значения функции Гомперца в точке перегиба (0,37, что приблизительно равняется *1/e*) до симметричной точки, равноудаленной от максимального значения функции желательности. Наименьшие длины у интервалов, соответствующих оценкам «слабо» и «сильно», что согласуется с приведенными выше допущениям. Определенную сложность представляет идентификация вербальных оценок числовыми (точечными) значениями из интервалов шкалы желательности. Наиболее распространенный прием состоит в использовании в качестве оценок граничных значений интервалов (например, 0 – 0,20 – 0,37 – 0,64 – 0,80 – 1) или их середин. Если известна ранжировка подмножества исследуемых объектов, то числовые значения вербальных оценок могут быть определены математически, например, по критерию максимизации коэффициента корреляции между результатами ранжирования и кодировки.

Сопоставление функций Гомперца и Харрингтона свидетельствует об идентичности моделей распространения инноваций и измерения инновационной активности. Это дает основание считать, что процессы оценивания в рассмотренных задачах характеризуются изоморфностью психофизиологических механизмов принятия решений и является косвенным подтверждением корректности предложенной оцифровки вербальной шкалы интенсивности при обработке результатов социологических измерений.

Для снижения систематической ошибки, связанной с неоднозначностью индивидуальных трактовок респондентами введенных вербальных оценок желательности (в шкале «плохо–хорошо»), последние следует расписать более подробно, гарантируя тем самым однозначность понимания. Например, можно перечислить, какие характерные проявления инновационной активности соответствуют каждой вербальной оценке. Систематическая ошибка будет сведена к минимуму, если вербальные оценки для каждого респондента будут строиться исследователем по результатам опроса, на основе включенных в анкету специально разработанных индикаторов, характеризующих изучаемую предметную область. Так, при оценке распространения инновационных практик в сельском социуме можно использовать индикаторы, характеризующие уровень образования респондента, количество и значимость освоенных инноваций в ведении домашнего хозяйства, наличие современных бытовых приборов и т.д.

После нахождения по описанной методике частных оценок инновационной активности встает задача их агрегирования. С учетом многомерности задача сводится к анализу векторных переменных. Подобная задача достаточно сложна, ее результаты трудно интерпретируемы, и обычно в социологии многомерные оценки сводятся к скалярным. В качестве примера можно упомянуть известный метод суммирования частных оценок нескольких составляющих (шкала Лайкерта).

В рассматриваемом нами случае агрегирование производится с учетом следующих особенностей шкал Харрингтона, обусловливающих их успешное применение в задачах оценивания инновационной активности. Все частные шкальные оценки построены по единой методике, являются безразмерными, однонаправленными и могут (с определенной натяжкой!) интерпретироваться как количественные. Это дает возможность использовать при их агрегировании достаточно широкий спектр математических преобразований. Диапазон возможного изменения значений каждого индикатора одинаков ([0–1]), что существенно упрощает технику агрегирования

[1] Адлер Ю.П., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий М.: Наука, 1976.

и интерпретацию результатов. В частности, позволяет построить шкалу для агрегированной переменной, удовлетворяющую следующим естественным требованиям. Максимальное значение агрегированной переменной должно равняться единице и достигаться тогда и только тогда, когда все частные показатели также равняются единице. Иными словами, наивысшая инновационная активность ассоциируется с наивысшей активностью всех частных составляющих одновременно. Минимальное значение агрегированной переменной должно равняться нулю и достигаться при равенстве нулю хотя бы одного из объединяемых частных показателей. Для реализации сформулированных требований в качестве обобщенной (агрегированной) оценки может использоваться показатель Y, определяемый как среднее геометрическое частных показателей желательности y_i, $i=1,\ldots,n$, оценивающих каждое из исследуемых проявлений инновационной активности:

$$Y = (y_1 * \ldots * y_n)^{1/n}.$$

Легко установить непосредственным вычислением, что введенная с помощью последней формулы функция удовлетворяет сформулированным выше условиям ее равенства 0 и 1. Кроме того, если все частные показатели y_i принимают одно и то же значение, то оно будет соответствовать и обобщенному показателю Y. Это дает основание для использования в качестве базовых отметок обобщенной функции желательности тех же значений, что и для частных желательностей y_i.

Естественно, что в зависимости от постановки задачи возможны и другие способы обработки и обобщений шкал желательности, в том числе широко применяемое социологами суммирование оценок по разным шкалам, построение одно- и многомерных распределений для определенных наборов индикаторов и др.

Предложенный нами подход существенно расширяет методный арсенал социолога, так как рассмотренный способ оцифровки допускает новые интерпретации результатов, а также представляет возможности применения количественных методов анализа, например, таких, как метод главных компонент, и построение латентных переменных для объяснения наиболее существенных проявлений инновационной активности.

О логико-комбинаторных методах в причинном анализе

Кученкова Анна Владимировна, *РГГУ*

Под причинным анализом понимают «методологическую процедуру анализа данных, имеющую специфическую языковую структуру и предназначенную для поиска знаний о причинно-следственных отношениях между социальными феноменами. На эмпирическом уровне причинный анализ — анализ структуры связей между признаками-причинами и признаками-следствиями»[1]. Для реализации причинного анализа как одного из основных видов анализа в эмпирической социологии используются различные методы многомерного анализа: регрессионный, факторный, дисперсионный анализ и т.д.

В изучении причинно-следственных отношений выделяют[2] два направления: построение структурных уравнений и индуктивное направление. К последнему относят методы, формализующие идеи Дж.С.Милля по поиску причинных отношений. Таковыми являются ДСМ-метод и метод сравнительного качественного анализа (Qualitative comparative analysis).

Второй из них (условно обозначим как СКА) был предложен американским социологом Ч. Рейджином[3] в конце 1980-х гг. для сравнительного анализа небольшого количества объектов путём установления комбинаций значений независимых переменных, детерминирующих определенное значение зависимой переменной. С точки зрения математической формализации СКА опирается на булеву алгебру (двузначную логику). Процедурно метод реализуется следующим

[1] Татарова Г.Г. Основы типологического анализа в социологических исследованиях. М.: Высшее образование и наука, 2007. С. 216.

[2] Толстова Ю.Н. Математико-статистические модели в социологии. М.: ГУ–ВШЭ, 2008. С. 154.

[3] Ragin C.C. The comparative method: Moving beyond qualitative and quantitative strategies. Berkley, etc.: University of California Press. 1987.

образом: исследователь выделяет различные объекты анализа (например, национальные меньшинства, организации[1]) и целевой признак (признак-следствие). Выдвигается гипотеза о влияние на этот признак различных факторов (независимых переменных дихотомического типа). Далее для каждого объекта определяется, ярко ли выражен каждый признак (присущ он объекту или нет), и в соответствии с этим, объектам приписывались коды 0 или 1. Если несколько объектов с одинаковым значением целевого признака обладают одинаковым набором характеристик (значений зависимых переменных), то это сочетание интерпретируется как совокупность условий, детерминирующих соответствующее значение целевого признака. Метод позволяет изучать «множественную причинность»[2], т.е. «причинами» (условиями возникновения) одного и того же значения зависимой переменной выступают различные сочетания значений независимых переменных.

ДСМ-метод[3] (названный в честь Дж. С. Милля) — метод обнаружения причинно-следственных связей между «структурой объекта исследования и его свойствами»[4] — применяется для выявления связей между социальными, биографическими (и другими) характеристиками респондентов и их мнением, поведенческими установками[5]. Этот метод был предложен В.К. Финном в конце 1970-х гг.

В основу ДСМ-метода положен особый способ «обнаружения причинно-следственных зависимостей на основе правдоподобного рассуждения»[6]. Метод успешно апробирован в социологических исследованиях[7], включая изучение электоральных предпочтений[8].

Применение ДСМ-метода основывается на определённых допущениях. Одним из них является «постулат поведения»[9], который заключается в утверждении, что мнение и установки респондента определяются в большей или меньшей степени его биографическими данными, социальными характеристиками, индивидуальными психологическими чертами. Если каждая из трёх составляющих представлена множеством характеристик, то их подмножество будет являться причиной наличия или отсутствия у респондента определённого мнения. Этот постулат является основой для постановки задач по поиску детерминант мнений респондента (типичных черт характера, фактов биографии, психологических характеристик) относительно конкретной проблеме. Безусловно, не отрицается возможное влияние ситуации. Если, например, обнаружена «причина» (детерминанта) определённого мнения (в виде сочетания значений независимых переменных, описывающих респондента), а респондент выражает другую точку зрения, значит, есть ситуация, которая повлияла на его позицию, ситуация, которую необходимо формализовать. В этом заключается принцип ситуативности.[10]

ДСМ-метод применяется для анализа жёстко структурированных данных. Специфика построения инструментария для проведения опроса заключается в том, что вопросник должен состоять из двух блоков. Первый блок содержит вопросы, раскрывающие характеристики респондентов (признаки-причины, потенциальные «причины» того или иного мнения); второй блок

---

[1] Отметим, что на практике метод не использовался для анализа мнений, поведений людей. В качестве объектов анализа выступали крупные организации, объединения и т.п.

[2] Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques / Ed. by B. Rihoux, Ch. Ragin. London: Sage, 2009. P. 8.

[3] Автоматическое порождение гипотез в интеллектуальных системах / Сост. Е.С. Панкратова, В.К. Финн; под общ.ред. В.К. Финна. М.: Книжный дом «ЛИБРОКОМ», 2009.

[4] Шашкин О.О. Приближенные средства установления сходства для ДСМ-метода автоматического порождения гипотез: автореферат диссертации … канд.техн.наук. М.: [б.и.], 2010. С. 3.

[5] С этой целью ДСМ-метод используется в социологии. Он успешно применяется в медицине, фармакологии, криминалистике и т.д.

[6] Данилова Е.Н., Михеенкова М.А., Климова С.Г. Возможности применения логико-комбинаторных методов для анализа социальной информации // Социология 4М. 1999. №1.

[7] См.: Данилова Е.Н., Михеенкова М.А., Московский С.С., Финн В.К. Об одной модели детерминации социального поведения // Интеллектуальные системы и общество. М.: РГГУ, 2000. С. 263–272.

[8] Бурковская Ж.И., Михеенкова М.А., Финн В.К. О логических принципах анализа электорального поведения // НТИ. Сер. 2. 2004. №8. С. 18–22.

[9] Михеенкова М.А., Финн В.К. О логических средствах концептуализации анализа мнений // НТИ. Серия 2. 2002. №6. С. 5.

[10] Там же. С. 7.

вопросов раскрывает мнение респондентов по теме исследования, и может быть представлен в виде набора суждений, с которыми респонденту предстоит согласиться или нет. Тем самым респонденты и их мнения описываются (представляются) через набор переменных. На основе этих описаний выявляются группы респондентов со сходным мнением (одинаковым сочетанием значений переменных, описывающих мнение). У этих респондентов обнаруживаются сходства в их характеристиках (биографических данных, социальном характере, психологических чертах – сочетания значений переменных, описывающих респондентов). Исходя из этого, формулируется гипотеза о том, что обнаруженные сходства в чертах респондентов являются детерминантами сходств их мнений. Порождение такого рода гипотез является промежуточным этапом применения ДСМ-метода. Подробно принцип работы метода раскрыт в работах В.К. Финна, М.А. Михеенковой.

Методы СКА и ДСМ имеют много общего, хотя и разрабатывались независимо друг от друга. (1) Оба метода основаны на использовании языка математической логики: в СКА используется булева алгебра (двузначная логика), в ДСМ — логика аргументации.

(2) Эпистемологическим основанием обоих методов являются идеи Дж. С. Милля: логические правила причинного вывода, опирающиеся на понятия «сходства», «различия» и т.д. Ппостулируется, что сходства, обнаруженные у объектов, обладающих определённым одинаковым свойством (чертой), являются причинами проявления у них этого свойства. Однако на практике эти методы использовались для анализа разных объектов: СКА — для сравнения стран, городов, стран, организаций, наций (в сравнительном анализе на макроуровне), для которых ищутся причины наличия или отсутствия у них каких-либо свойств. ДСМ-метод использовался для выявления детерминант мнений респондентов. Безусловно, это не означает, что их нельзя применять для изучения других объектов, но пока таких прецедентов не было[1].

(3) Оба метода реализуют индуктивную, восходящую стратегию анализа данных. ДСМ-метод является средством автоматического порождения гипотез, тем самым подчёркивается, что логика анализа не гипотетико-дедуктивная (заключающаяся в проверке выдвинутых гипотез), а индуктивная. Отметим, что разработчики рассматриваемых методов относят их к «формализованному качественному анализу данных». Это словосочетание не является устоявшимся и требует пояснения. Под ним понимается формализованная реализация «восходящей» стратегии анализа. Логико-комбинаторные методы направлены на отыскание всех существующих в данных зависимостей, в том числе неочевидных для исследователя. Такая стратегия характерна именно для «качественного» подхода, поэтому рассматриваемые эти методы называют методами «формализованного качественного анализа данных».

(4) СКА и ДСМ-метод схожим образом трактуют понятие причины, а именно, как сочетание условий, вызывающих определенное явление. Здесь подразумевается «множественная причинность»: что одно и то же явление может детерминироваться несколькими различными сочетаниями независимых переменных. При этом в ходе анализа переменные рассматриваются не по отдельности, напротив, изучаются сочетания значений переменных.

Рассматриваемые логико-комбинаторные методы отличаются от статистических по ряду параметров. В них не предполагаются вероятностная природа данных и высокий уровень измерения переменных. Другое отличие заключается в способе реализации «причинного анализа». Статистические методы направлены на отыскание некого универсального для всех рассматриваемых объектов объяснения (на эмпирическом уровне выделения небольшого количества переменных, «объясняющих» изменения значений зависимой переменной), они позволяют изучать связь между переменными. Логико-комбинаторные методы обнаруживают комбинации (сочетания) значений независимых переменных, детерминирующие определённое значение зависимой переменной, тем самым они направлены на установление связи между отдельными значениями переменных (а не между переменными). Тем самым они выступают альтернативой по отношению к традиционным статистическим методам, предоставляя средства формализации рассуждений о причинно-следственных отношениях.

[1] Имеется в виду сфера социологии. ДСМ-метод успешно использовался в фармакологии, медицине и других областях, которые в данной работе не рассматриваются.

В завершении следует отметить, что несмотря на наличие различных методов и подходов к анализу причинно-следственных отношений, «никакой формальный логико-математический анализ не может нам доказать, что какой-то признак (признаки) являются причиной такого-то явления. Тем не менее, использование логико-математического формализма — это «единственный подход, позволяющий… изучать причинно-следственные отношения»[1].

Обобщение модели Изинга для анализа поляризации мнений в сообществе с тремя превалирующими предпочтениями

Рыжова Анастасия Валентиновна, *НИУ ВШЭ*

В качестве математического аппарата для моделирования социальных процессов традиционно применяются преимущественно дифференциальные уравнения. Также очень широкое распространение имеет метод аналогий, когда в качестве основы берется модель физического процесса и с естественными уточнениями используется при моделировании социальных процессов. В настоящее время в рамках этой схемы быстро набирает силу статистическое направление, в котором для моделирования социальных и экономических явлений применяются методы статистической физики, т.е. за основу берется модель физического процесса, в качестве математического аппарата используются статистические методы. Направление достаточно новое, основной импульс оно получило после выхода работ Мантенья и Стэнли[2].

В частности, статистические методы оказываются очень продуктивными в задачах, где необходимо ответить на вопрос, каким образом возникает упорядоченное состояние из первоначально неупорядоченного. Под упорядоченным состоянием мы понимаем состояние общества, при котором существуют четко выделенные культурные и политические предпочтения, взгляды, интересы, которые разделяются большинством представителей данной группы. Общие взгляды формируются за счет обмена мнениями между составляющими группу индивидами (взаимодействия между ними), которое при некоторых условиях имеет тенденцию делать этих индивидов более похожими в определенных аспектах друг на друга. Очевидно, что при отсутствии взаимодействия каждый индивид сделает свой собственный выбор, не обусловленный влиянием его окружения. Это состояние мы называем неупорядоченным.

Одной из наиболее простых моделей формирования единого мнения является модель Изинга[3]. Пусть группа состоит из N объектов; имеется дихотомическая переменная S (два возможных состояния), обозначим ее значения через 1 и –1, тогда уравнение модели можно записать следующим образом:

$$H = -J\sum_{i<j} S_i S_j - h\sum_i S_i \qquad (1)$$

В этой формуле J — параметр, характеризующий интенсивность контактов между объектами внутри группы (взаимодействие между ними), h — внешнее воздействие, H — энергия системы. Рассмотрим немного подробнее интерпретацию некоторых параметров. Мы указали, что переменная S характеризует два возможных состояния, в которых могут находиться объекты системы. Это может применяться как для простых ситуаций, где под состояниями подразумеваются значения наблюдаемой переменной (скажем, посещает ли студент спецкурсы), так и для более интересных задач, в которых переменная рассматривается как латентная и представляет собой некоторую общую направленность, установку. Параметр h характеризует внешнее воздействие. При этом подразумевается, что оно является внешним не по отношению к системе в целом, а по отношению к объектам системы. Т.е. источник воздействия может находиться как вне группы,

---

[1] Толстова Ю.Н. Математико-статистические модели в социологии. М.: ГУ–ВШЭ, 2008. С. 154–155.

[2] Mantegna R.N., Stanley H.E. An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge: Cambridge University Press, 1999.

[3] Давыдов А.А.. Инновационный климат в стране и инновационная энергия предпринимателей // http://www.ssa-rss.ru/index.php?page_id=22&id=53#5 (2010); Castellano C., Fortunato S., Loreto V. Statistical physics of social dynamics // Reviews of Modern Physics. 2007. Vol. 81. Iss. 2. P. 591–646 (http://arxiv.org/pdf/0710.3256); Chakrabarti B., Chakraborti A., Chatterjee A. (eds.) Econophysics and Sociophysics: Trends and Perspectives. Berlin: Wiley-VCH, 2006.

так и внутри, во втором случае можно говорить об определенных *условиях среды*. Если в качестве примера вновь рассмотреть студенческое сообщество, то под внешним воздействием можно подразумевать как влияние преподавателей (источник находится вне студенческих групп), так и общий *климат* внутри системы, скажем, располагает ли он к тому, чтобы хорошо учиться, или хорошая учеба не ценится.

Отметим, что в состоянии равновесия энергия системы минимальна. Рассмотрим, при каких сочетаниях значений параметров уравнения (1) H будет стремиться к минимуму. Во-первых, при совпадении состояний (мнений) составляющих систему объектов. В этом случае величина $-J\sum_{i<j} S_i S_j$ принимает свое наименьшее значение; чем больше пар взаимодействующих объектов, находящихся в различных состояниях, чем выше значение данной величины. Второе слагаемое, $-h\sum_i S_i$ принимает свое минимальное значение, когда все объекты системы находятся в состоянии 1; при повышении количества объектов в состоянии –1 будет увеличиваться значение этой величины. Это можно проинтерпретировать таким образом, что объекты, находящиеся в состоянии 1, сонаправлены с внешним воздействием, а объекты, находящиеся в состоянии –1, направлены противоположным образом. При h<0, наоборот, состояние –1 сонаправлено с внешним воздействием. Таким образом, мы получаем картину, которая полностью согласуется с представлениями, основанными на здравом смысле: система находится в стабильном состоянии или близка к нему, если большинство составляющих ее индивидов придерживаются одного и того же мнения, и направление внешнего воздействия с ним совпадает.

Вероятность реализации состояния $(S_1, S_2, ..., S_N)$ подчиняется закону распределения Гиббса:

$$P(S_1, S_2, ..., S_N) = A\exp(-H(S_1, S_2, ..., S_N)/T) \qquad (2)$$

В формуле (2) A — нормировочная постоянная, T — температура. Применительно к нашей задаче температура интерпретируется как способность объектов хаотично и самопроизвольно менять свое состояние. Таким образом, высокие значения параметра T будут соответствовать ситуации «тревожности», когда большинство составляющих группу индивидов не имеют постоянного мнения.

Модель Изинга позволяет рассматривать только те задачи, где для объектов есть только два потенциальных состояния. Это серьезное ограничение, поэтому, на наш взгляд, полезно и интересно рассмотреть обобщение модели на случай, когда число возможных состояний больше чем два; также есть много ситуаций, когда эти состояния не «равноправны», на них накладывается определенная структура, исходя из содержательного смысла задачи. Нас интересует обобщение модели Изинга на случай, когда в системе существует выбор из трех возможных вариантов, при этом два из этих вариантов равноправны, а третий выделен. Когда мы говорим о том, что варианты равноправны, мы подразумеваем, что они эквивалентны с точки зрения одной или нескольких характеристик, на которые опирается индивид, когда делает выбор. В качестве примера можно рассмотреть ситуацию, когда при выборе ВУЗа абитуриент рассматривает два института естественнонаучного направления и один институт гуманитарного направления.

При построении обобщенной модели вместо изинговских переменных S_i мы возьмем новые переменные вида $Q_i = Z_i^2 - q$, где Z_i может принимать значения 0, 1, или –1; параметр q характеризует, насколько сильно с содержательной точки зрения два эквивалентных состояния отличаются от выделенного. В этом случае характеристическая функция имеет вид:

$$H = -J\sum_{i<j} Q_i Q_j - h_1\sum_i Z_i - h_2\sum_i Q_i \qquad (3)$$

Параметр h_1 характеризует внешнее воздействие, сонаправленное с одним из состояний $Z = \pm 1$ (с состоянием $Z = 1$ при h_1>0 и состоянием $Z = -1$ при h_1<0), параметр h_2 характеризует воздействие, направленное на оба эквивалентных состояния одновременно (в соответствующее слагаемое входит Q_i а не Z_i).

С содержательной точки зрения наиболее интересен вопрос о том, какое состояние в системе преобладает. Сначала проанализируем зависимость количества объектов в каждом из состояний

от параметра взаимодействия J при различных значениях параметра q, затем рассмотрим, как меняется ситуация при учете внешнего воздействия (h_1 и h_2).

Рассмотрим зависимость относительного количества объектов в каждом из состояний от J при q=0.5 и q=1 (рис. 1 и 2). При малых значениях J система находится в неупорядоченном состоянии ($N_{+1}/N=N_{-1}/N=N_0/N$). При возрастании J в зависимости от величины q начинает превалировать один из возможных вариантов: в первом случае возрастают N_{+1} и N_{-1} и убывает N_0, во втором случае ситуация обратная. Было найдено критическое значение q=2/3 (рис. 3), при котором есть довольно большой диапазон J, в рамках которого изменение значений этого параметра не оказывает влияния на поведение системы и в ней сохраняется неупорядоченное состояние.

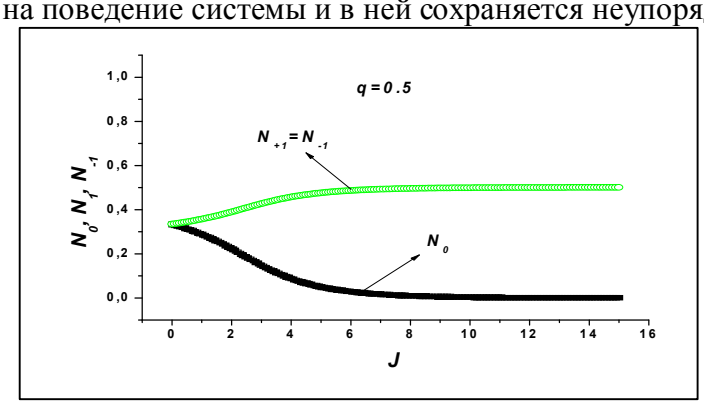


Рис. 1. Зависимость от J относительного количества объектов в каждом из состояний при q=0.5, $h_1=h_2=0$.

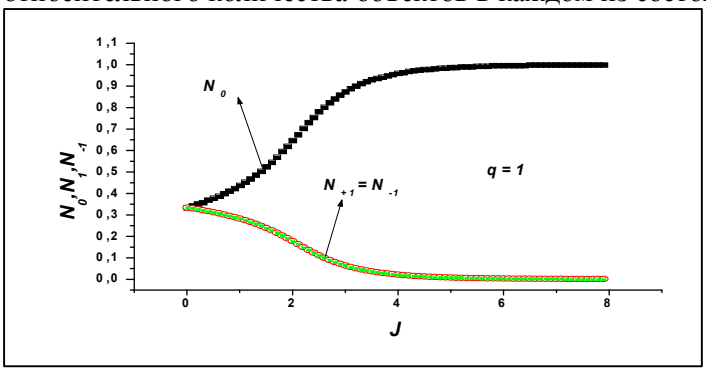


Рис. 2. Зависимость от J относительного количества объектов в каждом из состояний при q=1, $h_1=h_2=0$.

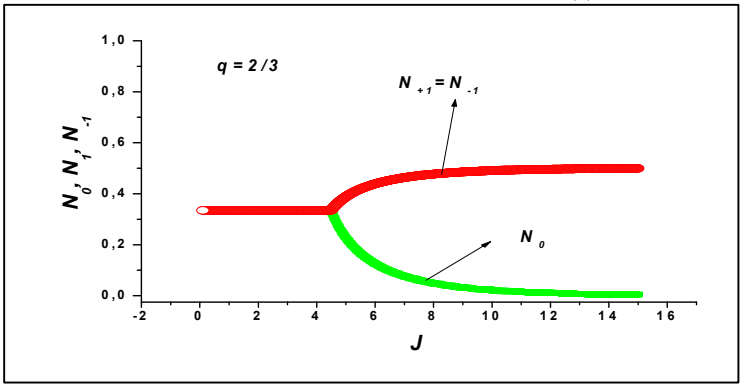


Рис. 3. Зависимость от J относительного количества объектов в каждом из состояний при q=2/3, $h_1=h_2=0$.

Далее рассмотрим воздействие на систему параметра h_1 (рис. 4). Здесь ситуация отличается от предыдущих, так как уже при малых значениях J система находится в упорядоченном состоянии, при этом преобладает N_{+1}. Это согласуется со смыслом параметра h_1, который характеризует внешнее воздействие, сонаправленное с одним из состояний +1, –1. В нашем случае h_1=1>0, т.е. это воздействие «поддерживает» состояние Z=1, что мы и наблюдаем. При возрастании значений параметра J преобладание состояния Z=1 ослабевает, и ситуация сводится к представленной на рис. 2. Таким образом, эффект от взаимодействия между объектами оказывается более сильным, чем эффект внешнего воздействия.

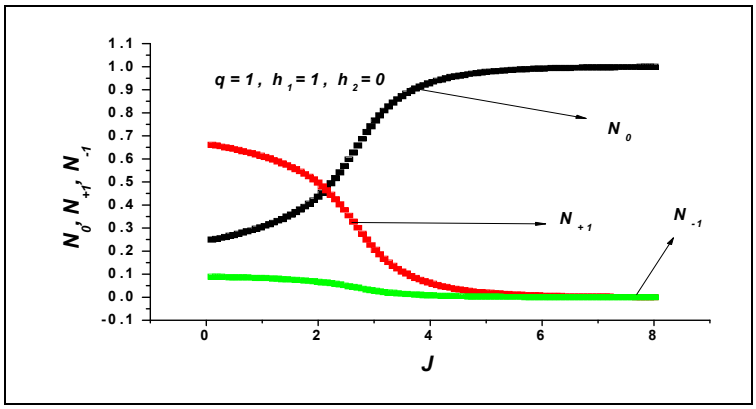


Рис. 4. Зависимость от J относительного количества объектов в каждом из состояний при q=1, h_1=1,h_2=0.

Теперь перейдем к рассмотрению воздействия на систему параметра h_2 (рис. 5). Здесь ситуация очень похожа на рассмотренную на рис. 2, за исключением того что при малых значениях J преобладают состояния $Z = \pm 1$, в соответствии со смыслом параметра h_2 (характеризуемое этим параметром воздействие при h_2>0 «поддерживает» указанные состояния). Как и при рассмотрении эффекта параметра h_1, здесь при увеличении интенсивности взаимодействия внутри группы эффект от внешнего воздействия ослабевает.

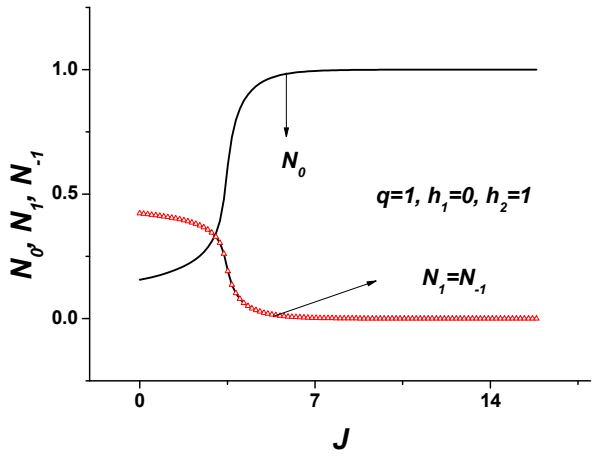


Рис. 5. Зависимость от J относительного количества объектов в каждом из состояний при q=1, h_1=0,h_2=1.

Таким образом, нами было получено точное решение для случая, когда в системе все индивиды взаимодействуют друг с другом с одинаковой интенсивностью. В рамках полученного решения проанализировано поведение системы в зависимости от интенсивности взаимодействия между объектами, степени внешнего воздействия и разницы между потенциальными состояниями.

Классификация простых и сложных семей

Савельев Лев Яковлевич,
Новосибирский государственный университет
Гончарова Галина Савитовна
Институт философии и права СО РАН

Семья является основной ячейкой общества. Все демографические и социальные процессы так или иначе связаны с семейной структурой общества. Для того чтобы их эффективно исследовать, необходимо знать семейную структуру населения. Для ее описания нужна четкая структурная классификация семей. Описываемая матричная классификации учитывает пол, возраст и некоторые другие основные характеристики членов семьи. Эта классификация формализована и ее можно использовать при компьютерной обработке данных. Она была успешно использована при исследовании структуры населения некоторых регионов РФ[1]. Здесь описываются некоторые модификации модели и ее приложения.

Предлагается структурное определение семьи, основанное на точно описываемом понятии *простая* семья. Остальные — *сложные* семьи — являются объединением простых. Такой подход позволяет дать подробные классификации семей, которые удобно использовать для социального, экономического, демографического и этнического описания различных групп населения. Индивид, как правило, тесно связан с семьей, которая существенно определяет его потребности и поведение. В задачах социально-экономического планирования и управления реалистичный прогноз таких важных показателей, как распределение семей по величине совокупного и душевого дохода, объем и структура спроса на товары потребления, объем и структура жилищного фонда, развитие сферы услуг, должен опираться на демографические данные о численности населения, их возрастно-половом и родственном составе.

1. Деление сложных семей на простые

В основе предлагаемых классификаций лежит простая семья. Сложные семьи классифицируются по числу и характеру составляющих их простых семей. Это позволяет выявить важные особенности сложных семей, не проявляющиеся при классификации их по размеру. Простой семьей, кроме брачной пары с детьми, мужчины или женщины с детьми считается и одинокий взрослый (18 лет и старше).

При исследовании процессов, в которых родственные связи не играют существенной роли, удобно делить семьи на простые части и рассматривать совокупность простых семей и простых частей сложных семей. Тогда можно считать, что исследуемая семейная структура состоит только из простых семей. Это значительно упрощает ее изучение и позволяет выявить закономерности, которые без такого деления незаметны.

2. Матричная классификация простых семей

Принцип, положенный в основу рассматриваемых классификаций, поясняет следующее описание одной из самых простых таких классификаций. Это описание удобно сделать с использованием матриц и их формальных линейных комбинаций. Матрицы делают его наглядным. Тип каждой простой семьи описывается матрицами:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \qquad B = \begin{pmatrix} \operatorname{sgn} a_{11} & \operatorname{sgn} a_{12} \\ \operatorname{sgn} a_{21} & \operatorname{sgn} a_{22} \\ \operatorname{sgn} a_{31} & \operatorname{sgn} a_{32} \end{pmatrix}$$

Элемент a_{ij} в i-той строке и j-том столбце (i = 1,2,3 и j =1,2) равен числу элементов семьи i-го поколения (1-старшее, 2-среднее, 3-младшее) и j-го пола (1-мужской, 2-женский).

[1] Гончарова Г.С., Савельев Л.Я. Семейно-брачные отношения у народов Сибири. Новосибирск: Нонпарель, 2004. С. 286. (http://www.tuva.asia/news/ruregions/2865-goncharova-savelev.html)

Вместо матрицы A можно взять индикаторную матрицу B. Элемент $b_{ij} = \operatorname{sgn} a_{ij}$ в i-той строке и j-том столбце матрицы равен знаку элемента a_{ij}. Если $a_{ij} = 0$, то $b_{ij} = 0$, а если $a_{ij} > 0$, то $b_{ij} = 1$. Матрица $B = \operatorname{sgn} A$ имеет двоичный номер (от 0 до 63):

$$N(B) = b_{32} \times 2^0 + b_{22} \times 2^1 + b_{12} \times 2^2 + b_{31} \times 2^3 + b_{21} \times 2^4 + b_{11} \times 2^5 .$$

Номера выбраны так, чтобы они возрастали вместе с возрастом членов семьи и присутствием женщин. Так как в простой семье по определению не может быть трех поколений, то не все матрицы A и B описывают семьи. Число семейных индикаторных матриц B равно 6+2(6+4+1)=28. Каждая из них описывает укрупненный тип простой семьи, в котором учитывается только пол и возраст (поколение) членов семьи, существование детей, но не учитывается их число.

Столбец $C = (c_i = a_{i1} + a_{i2})$, i=1,2,3, полученный сложением элементов матрицы A в каждой строке, описывает численности поколений в простой семье. А строка $D = (d_1 = a_{11} + a_{21} + a_{31}$, $d_2 = a_{12} + a_{22} + a_{32})$, полученная сложением элементов матрицы A в каждом столбце, описывает числа членов каждого пола в простой семье. Соответствующие индикаторные матрицы $\operatorname{sgn} C$ и $\operatorname{sgn} D$ описывают заполненность поколений и полов. Этим матрицам тоже можно присвоить двоичные номера:

$$N(\operatorname{sgn} C) = \operatorname{sgn} c_3 + \operatorname{sgn} c_2 \times 2 + \operatorname{sgn} c_1 \times 4 , \quad N(\operatorname{sgn} R) = \operatorname{sgn} r_2 + \operatorname{sgn} r_1 \times 2 .$$

Номер $N(\operatorname{sgn} C) = 1$ имеют простые семьи, в которых есть только младшее поколение, а номер $N(\operatorname{sgn} C) = 6$ – простые семьи, в которых есть среднее и старшее поколения. Простых семей с номерами $N(\operatorname{sgn} C) = 5$ и 7 не существует. Номер $N(\operatorname{sgn} R) = 1$ имеет чисто женская семья, номер $N(\operatorname{sgn} R) = 2$ – чисто мужская семья, номер $N(\operatorname{sgn} R) = 3$ имеют смешанные по полу семьи.

Возрастные поколения членов семьи определяются по выбранной шкале их возрастом и не зависят от родственных отношений. Во многих случаях представляется целесообразным рассматривать семейные поколения, при определении которых, прежде всего, учитываются родственные отношения членов семьи. Семейные поколения могут отличаться от возрастных. Во многих социальных процессах важны именно семейные поколения.

3. Матричная классификация сложных семей

По определению сложная семья состоит из простых и поэтому тип сложной семьи естественно определяется типами составляющих ее простых семей. Формально тип сложной семьи удобно записывать линейной комбинацией матриц этих составляющих:

$$A = \sum \alpha(k) A(k), \qquad B = \sum \beta(l) B(l)$$

Здесь $A(k)$ и $B(l) = \operatorname{sgn} A(k)$ — семейные матрицы k и l типов, $\alpha(k)$ и $\beta(l)$ — числа простых семей типов $A(k)$ и $B(l)$, составляющих сложную семью.

Вместо индикаторных матриц $B(l)$ можно использовать их двоичные номера и заменять рассматриваемые формальные линейные комбинации матриц мультииндексными произведениями:

$$M(B) = \prod m^{j(m)} .$$

Здесь множитель $m^{\gamma(m)}$ означает, что в сложной семье есть $\gamma(m)$ простых частей с демографическим номером m. Сложные семьи с большим числом простых частей разного типа встречаются редко. Поэтому описывающих типы реальных сложных семей комбинаций матриц и мультииндексных произведений немного.

4. Демографические номера

Присвоим каждому типу простой семьи четырехзначный демографический номер $m = ijkl$, где i обозначает номер поколения мужчины, j — номер поколения женщины, k — число мальчиков, l — число девочек. Если мужчины или женщины нет, то соответственно $i = 0$ или $j = 0$. Точно так же $k = 0$ или $l = 0$, если в семье нет мальчиков или девочек. Размер семьи типа

$m = ijkl$ (число членов в ней) определяется по формуле $n = \operatorname{sgn} i + \operatorname{sgn} j + k + l$.

Пара (i,k) характеризует мужскую часть семьи, а пара (j,l) – женскую. Если $i = k = 0$, то семья чисто женская. Если $j = l = 0$, то семья чисто мужская. Пара (i, j) описывает взрослую часть семьи, а пара (k,l) — детскую. Если $k = l = 0$, то семья бездетная. Если $i \neq 0$ или $j \neq 0$, то семья неполная.

Подходящий выбор количества и качества поколений позволяет паре (i, j) давать достаточно точную демографическую характеристику взрослой части семьи. Детская часть семьи характеризуется парой (k,l) хуже. При необходимости можно ввести кроме взрослых еще детские поколения и заменить числа k,l векторами $k = k(1)k(2)$, $l = l(1)l(2)$ или $k = k(1)k(2)k(3)$, $l = l(1)l(2)l(3)$ соответственно для 2-х или 3-х поколений детей. Получаются шести или восьмизначные демографические номера. В них $k(s)$ и $l(s)$ будут обозначать числа мальчиков и числа девочек 3-го детского поколения. Например, для задач, связанных с жилищными проблемами, целесообразно делить детей на младших (дошкольников) и старших (школьников).

По демографическому номеру семьи можно определить число поколений в ней, номера самого младшего и самого старшего из имеющихся семейных поколений можно вычислить абсолютную и относительную асимметрию семьи по полу (ее феминизацию): $f = (\operatorname{sgn} j - \operatorname{sgn} i) + (1 - k)$ и $f / n = (\operatorname{sgn} j - \operatorname{sgn} i + 1 - k)/(\operatorname{sgn} i + \operatorname{sgn} j + k + 1)$.

Демографические номера сложных семей являются линейными комбинациями номеров составляющих их простых семей.

5. Формализация списка членов семьи

По определению *простая семья* может состоять из (1) *одиночки*, (2) *одного из родителей с невзрослыми детьми*, (3) *брачной пары без детей*, (4) *брачной пары с невзрослыми детьми*. Независимо от возраста дети считаются невзрослыми, если они не состоят в браке и не имеют своих детей в семье. С точки зрения структуры это логично: имеется в виду *семейная невзрослость*. Включение одиночек позволяет составлять из простых различные типы сложных семей и говорить о семейной структуре всего населения данной группы. При необходимости одиночки легко исключаются.

Данное определение простой семьи основывается на понятиях *родители, дети, брачная пара*. Каждое из этих понятий нуждается в определении. Естественно разделяются фактические и юридические отношения: кровное родство и усыновление, гражданский и юридический браки. Для структуры семьи эти различия несущественны. Можно считать, что входящие в определение простой семьи понятия имеют точный смысл. При анкетировании обычно используется самоидентификация или указания респондента.

Сложная семья состоит из нескольких простых, члены которых связаны между собой родственными отношениями. Простые семьи, составляющие сложную семью, называются её *простыми частями*.

Структура сложной семьи определяется структурой её простых частей. Поэтому каждая сложная семья имеет общие и составные структурные характеристики: общее число членов сложной семьи и вектор чисел членов её простых частей, общее число брачных пар в сложной семье и вектор чисел брачных пар в её простых частях. Рассматриваются также общие и составные гендерные и генерационные характеристики сложных семей. Составные характеристики позволяют проводить более подробный качественный и количественный анализ сложных семей и семейной структуры населения.

В рассматривавшихся массивах большинство сложных семей состояло из двух простых частей. Благодаря своим размерам сложные семьи охватывали значительную часть населения. Структура сложной семьи отражает кроме демографических и социально-экономические процессы. Поэтому сложные семьи заслуживают отдельного исследования.

Будем термином *семья* обозначать любую группу родственников, представленную данным списком. Требуется, чтобы каждые два члена семьи были связаны отношением родства, выражаемым некоторой композицией основных отношений *ребенок, родитель, супруг*. Определения этих и других, связанных с ними понятий, зависят от условий задачи и здесь не обсуждаются.

Каждые два члена семьи связаны отношением родства, выражаемым некоторой композицией основных операторов: *ребенок* — c, *родитель* — p, *супруг* — q.

Предполагается, что в списке семьи указаны: (1) *порядковый номер члена семьи*; (2) *родство с первым в списке*; (3) *номера детей*; (4) *номера родителей*; (5) *номер супруга*; (6) *пол*; (7) *возраст*. Родство с первым в начальном списке может описываться в обиходных терминах. Но для предлагаемого исследования структуры семьи оно должно быть представлено в операторной форме с помощью кодовой таблицы. Отсутствие детей, родителей, супруга отмечается нулем в соответствующем пункте. Мужской пол кодируется +1 или просто 1, женский — –1. Возраст определяется числом полных лет. Все эти сведения о семье удобно представлять в матричном виде.

Будем представлять список семьи X из $n[x]$ человек в виде обозначаемой также буквой X матрицы с $n[x]$ строками и 7-ю столбцами. Столбцы $X[*, j]$ $(1 \le j \le 7)$ матрицы X содержат информацию о соответствующем признаке для всех членов семьи: 1-й столбец $X[*,1]$ — порядковую нумерацию членов семьи, 2-й $X[*,2]$ — родство с первым в списке, 3-й $X[*,3]$ — номера детей, 4-й $X[*,4]$ — номера родителей, 5-й $X[*,5]$ — номера супругов, 6-й $X[*,6]$ — пол, 7-й $X[*,7]$ — возраст. Строку с номером i матрицы X обозначим $X[i,*]$ $(1 \le i \le n[x])$. Она содержит всю информацию об i-ом по списку члене семьи X. Элемент, принадлежащий i-ой строке и j-му столбцу матрицы X, обозначается $X[i, j]$. Равенство $X[i, j] = 0$ означает, что у i-го члена семьи в списке нет родственников j-ой категории (детей, родителей, супруга).

Для социологического анализа в более подробном списке целесообразно указать дополнительно: (8) *образование*; (9) *доход*; (10) *национальность*; (11) *религию*. Образование удобно записывать числом лет обучения, доход — в выбранных единицах, национальность и религию — по специальным кодам. Если рассматривается совокупность семей, то к этому списку добавляется еще номер данной семьи. Можно добавить и любые другие характеристики. Программа позволяет обрабатывать и длинные списки.

Пример. При анкетировании был составлен следующий список членов семьи:

1. женщина,–1,53; 2. муж,+1,60; 3. дочь,–1,33; 4. муж дочери,+1,35; 5. внучка (дочь 3),–1,12; 6. внучка (дочь 3),–1,6.

После кодирования и добавления номеров детей, родителей и супругов эти списки преобразуются в следующую матрицу:

$$X = \begin{pmatrix} 1 & a & 3 & 0 & 2 & -1 & 53 \\ 2 & q & 3 & 0 & 1 & +1 & 60 \\ 3 & c & \{5,6\} & \{1,2\} & 4 & -1 & 33 \\ 4 & qc & \{5,6\} & 0 & 3 & +1 & 35 \\ 5 & cc & 0 & \{3,4\} & 0 & -1 & 12 \\ 6 & cc & 0 & \{3,4\} & 0 & -1 & 6 \end{pmatrix}$$

Используются операторные обозначения: a — первый в списке, q — ее супруг, c — их ребенок, qc — супруг c, cc — ребенок c.

Столбцы матриц содержат информацию о семье: 1-й столбец — порядковую нумерацию членов семьи, 2-й — родство с первым в списке, 3-й — номера детей, 4-й — номера родителей, 5-й — номера супругов, 6-й — пол, 7-й — возраст. При разделении семьи на простые части, прежде всего, выделяются одиночки и брачные пары без невзрослых детей. Все остальные простые семьи получаются из одиночек или брачных пар добавлением их невзрослых детей с указанием пола каждого из членов семьи.

6. Алгоритм деления сложной семьи на простые

Делить семью на простые части можно различными способами. Представляется целесообразным делать это по следующему алгоритму.

(1) Находится самый младший по возрасту член семьи (a). Если членов семьи минимального возраста несколько, то берется имеющий наименьший номер в списке семьи.

(2) Если у (a) в семье есть супруг (b), у которого нет в семье своих невзрослых детей, то брачная пара (a,b) составляет первую простую часть семьи: $s=\{a,b\}$. Если у (b) в семье есть невзрослые дети (C), то эти дети добавляются к брачной паре (a,b), и все вместе они составляют первую простую часть семьи: $s\{a,b,C\}$.

(3) Если у (a) в семье нет супруга, но есть родители (P), то (a), его родители и все их невзрослые дети (D), имеющиеся в семье, составляют первую простую часть семьи: $s\{a,P,D\}$.

(4) Если супруга и родителей у (a) в семье нет, то первая простая часть семьи состоит из самого (a): $S=\{a\}$.

(5) Из списка семьи исключается первая простая часть и получается первый остаток (R). Применение к нему правил (1) – (4) дает вторую простую часть и второй остаток. И так далее, пока список семьи не будет исчерпан.

Ясно, что получаемые по правилам (1) – (5) части являются простыми семьями, не имеют общих членов и исчерпывают всю семью, то есть составляют *разбиение на простые части* рассматриваемой семьи. Для регулярных семей это разбиение определяется правилами (1) – (5) однозначно.

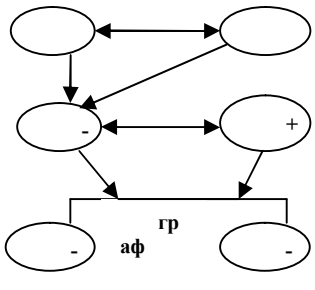
Замечание. Если начинать не с самого младшего по возрасту члена семьи, а с какого-нибудь другого соответствующим образом изменив правила (например, с самого старшего или с первого в списке), то разбиение одной и той же семьи на простые части может быть другим.

7. Граф семьи[1]

Часто бывает удобно изображать некоторую ситуацию в виде рисунка (схемы, графа), состоящего из точек (вершин), представляющих основные элементы ситуации, и линий (ребер), соединяющих определенные пары этих вершин и представляющих связи между ними. Такие рисунки известны под общим названием *графов.* Графы встречаются во многих областях под разными названиями: «структуры» в гражданском строительстве, «сети» в электротехнике, «социограммы» в социологии и экономике, «молекулярные структуры» в химии. Есть также «дорожные карты», электрические и газовые «распределительные сети».

Модели коллективов и групп, используемые в социологии, основываются на представлении людей или их групп в виде вершин, а отношений между ними (например, отношений прямого или бокового родства) — в виде ребер или дуг. Подобными описаниями успешно решаются многие задачи исследования структуры социальных групп, их сравнения, определения агрегированных показателей, отражающих степень напряженности, согласованности, взаимодействия.

Графическое представление дает наглядную картину семейной структуры исследуемой группы населения, позволяет проводить ее качественный и количественный анализ. На рисунке изображен граф семьи, описанной в примере.



По алгоритму эта семья делится на простые части следующим образом. Выбирается самый младший член семьи (a), имеющий номер 6. Его родители имеют номера 3 и 4: $P=\{3,4\}$. Их невзрослые дети имеют номера 5 и 6: D = $\{5,6\}$. Все вместе они составляют первую простую часть семьи: $S_1=\{a,P,D\}=\{3,4,5,6\}$. Остаются члены семьи с номерами 1 и 2: $R=\{1,2\}$. Они составляют вторую простую часть семьи $S_2=\{1,2\}$.

На графе хорошо видны эти простые части и старшее, среднее, младшее поколения: $\{1,2\}$, $\{3,4\}$, $\{5,6\}$. Знаки перед номерами описывают пол: плюс — мужской, минус — женский.

[1] См.: Харари Ф. Теория графов. Москва: Мир, 1973.

8. Применение классификации семей при определении потребности в жилье[1].

Понятие жилища обычно включает не только жилую ячейку, место пространственного размещения семьи, но и сферу повседневного обслуживания населения и прилегающие территории. В данном исследовании под жилищем понимается жилая ячейка, и изучается только соответствие жилища структуре семьи. Оно представляется сейчас наиболее важным.

Соответствие жилища структуре семьи, ее потребностям и возможностям можно определять по-разному. Самым простым является принцип: каждой семье — отдельное жилище. Он естественно дополняется другим принципом: каждому члену семьи отдельная комната. Объединение этих двух принципов может характеризовать определенное соответствие между структурой семьи и ее жилищем. Такая характеристика много не учитывает и ее нужно уточнять. Кроме того, если считать такое соответствие нормой, нужно определить возможные отклонения от этой нормы. Важны размер семьи и занимаемая площадь. Можно, например, требовать отдельные комнаты только для взрослых членов семьи. Можно добавить еще общую комнату для семьи, как это делается в некоторых странах. Можно исключить из этого требования супругов. При объединении детей естественно учитывать их пол и возраст. Вообще пол и возраст членов семьи играют важную роль при определении для семьи соответствующего жилища.

В данном исследовании при определении соответствия семьи ее жилищу структура семьи выбирается в качестве главного критерия. Как и во всех других случаях сначала рассматриваются простые семьи. Решение вопроса о сложных семьях сводится к их простым частям. Применяется, в частности, принцип: каждой простой семье — отдельную комнату. К простым частям сложных семей присоединялись взрослые члены семьи, не состоящие в браке.

Размер семьи является главным признаком, определяющим потребность в жилище. Ясно, что каждому члену семьи нужна какая-то площадь. Потребности членов семьи в жилой площади в зависимости от пола и возраста обычно учитываются во вторую очередь.

Считается, что каждая простая семья и каждая простая часть сложной семьи нуждается как минимум в отдельной комнате, а как максимум — в отдельном жилище.

Для определения минимальной потребности семьи в жилище были предложены следующие правила:

1. Брачной паре полагается 1 большая комната.

2. Каждой из пар: *взрослый + ребенок, подросток + ребенок, ребенок + ребенок, мужчина + юноша, женщина + девушка, юноша + юноша, девушка + девушка* полагается одна большая комната. (Мужчина и женщина имеются ввиду холостые.)

3. Каждому из оставшихся членов семьи, которых не удается объединить в указанные в пунктах 1 и 2 пары, полагается 1 стандартная комната.

В соответствии с принятыми тогда нормами и допустимыми отклонениями здесь большой (1Б) называется комната на двух человек, стандартной (1С) — комната на одного человека.

Замечание. Сформулированные правила позволяют только вычислить минимальную потребность семьи в жилище.

В соответствии с правилами 1–3 предлагается следующий алгоритм для определения минимальной потребности простой семьи в жилище.

1-й шаг. Выделяются *брачные пары* (если они есть). Юноши, девушки и дети делятся на пары *юноша + юноша, девушка + девушка, ребенок + ребенок*.

После выделения указанных пар могут остаться: холостые мужчины, женщина и непарные юноша, девушка, ребенок (не больше чем по одному). Из оставшихся членов семьи при соответствующем их составе можно выделить еще пары для больших комнат. Этот процесс описывают следующие два шага.

2-й шаг. Если есть непарный ребенок и один из непарных подростков или холостой взрослый, то соответственно образуется пара *ребенок + подросток* или пара *ребенок + взрослый.*

Если есть непарные ребенок, один из подростков, холостые мужчина или женщина либо ребенок, оба подростка и холостые мужчина или женщина, то соответственно образуются пары

[1] Более подробно изложено в заключительном отчете по проекту «Семья в Новосибирской области» № Н1-15-97, грант областной администрации Новосибирской области.

ребенок + взрослый и *подросток + взрослый* либо *ребенок + подросток* и *подросток + взрослый.* (Объединяются подросток и взрослый одного пола.) Если есть непарные ребенок, оба подростка, холостые мужчина и женщина, еще один холостой взрослый, то образуются пары *ребенок + взрослый, юноша + мужчина, девушка + женщина.*

3-шаг. Если непарного ребенка нет, то оставшийся непарный юноша объединяется с холостым мужчиной (если он есть), а оставшаяся непарная девушка объединяется с холостой женщиной (если она есть).

Четвертый шаг позволяет подсчитать нужное семье число стандартных комнат.

4-й шаг. Выделяются все оставшиеся не объединенные в пары члены семьи.

Предлагаемый алгоритм позволяет выразить удобными для расчетов формулами полагающиеся по правилам 1–3 количества больших и стандартных комнат для простой семьи. *Обозначения:* s — размер семьи; p — число брачных пар; m — число холостых мужчин, n — число холостых женщин; i — число юношей, j — число девушек; k — число детей; $[a]$, $\{a\}$ — целая и дробная части числа a; $signa = 0$ при $a = 0$ и $signa = 1$ при $a > 0$; x — число больших комнат, y — число стандартных комнат.

Каждая семья $F = F(p,m,n,i,j,k)$ описывается набором чисел p,m,n,i,j,k. Полагающееся ей количество $x = x(F) = x(p,m,n,i,j,k)$ больших комнат и количество $y = y(F) = y(p,m,n,i,j,k)$ стандартных комнат в соответствии с описанным алгоритмом определяются следующими равенствами:

$x = x(1) + x(2) + x(3)$, где $x(1) = p + [i/2] + [j/2] + [k/2]$,

$x(2) = sign\{k/2\}[sign(\{i/2\} + \{j/2\} + m + n) + sign(\{i/2\}\{j/2\}(m+n) + (\{i/2\} + \{j/2\})mn) +$

$+ sign(\{i/2\}\{j/2\}mn(m+n-2))]$,

$x(3) = (1 - sign\{k/2\})[sign(\{i/2\}m) + sign(\{j/2\}n)]$; $y = s - 2x$.

Пара чисел $K = (x,y)$ описывает полагающееся по сформулированным правилам жилище для семьи F. Вместо $K = (x,y)$ можно писать $K = xБ + y$, или более подробно — $K(F) = x(F)Б + y(F)$.

Кроме алгоритма вычисления минимальной потребности семей в жилье были разработаны алгоритмы для определения среднего и максимального варианта потребностей семей в жилье. В них были изменены правила формирования потребностей семьи в жилище в сторону расширения возможности расселения по комнатам.

По минимальному варианту холостые взрослые объединяются с детьми и подростками из любых частей. В среднем варианте холостые взрослые объединяются в пары с детьми и подростками только из своей части, а дети и подростки могут при объединении браться и из разных простых частей сложной семьи. По максимальному варианту объединяются в пары дети из любых частей, а оставшийся непарный ребенок объединяется с любым подростком или холостым взрослым. Если рассматривать каждую простую часть как отдельную семью и для нее рассчитывать потребность в жилище по максимальному варианту, то суммарная оценка потребности для сложной семьи увеличится.

На основе первичных данных микропереписи 1994 г. были вычислены для Новосибирской области распределение потребностей в жилище семей различных типов по минимальному, среднему и максимальному вариантам. Производилось разбиение сложных семей на простые части, поэтому различных типов семей получилось много. В таблицах были указаны полагающиеся по каждому из вариантов типы квартир. Из таблиц, охватывающих почти 95 % выборки, можно было извлечь и много другой полезной информации о структуре семей Новосибирской области. Почти 35 % семей составляли брачные пары без детей и одинокие взрослые. По минимальному и среднему вариантам им полагается соответственно квартиры типов 1Б и 1С. Этим объяснялась высокая потребность в однокомнатных квартирах. По максимальному варианту брачной паре без детей полагается квартира 2С (две стандартные комнаты). В этом варианте потребность в однокомнатных квартирах значительно уменьшалась, а в 2-комнатных — увеличивалась. Потребность в 2-комнатных квартирах по минимальному и среднему вариантам со-

ставляла почти 30 % из наиболее часто встречающихся типов семей (брачная пара с 1–2 детьми или 1–2 подростками и 2 холостяка.)

Анализ таблиц позволял оценивать потребности по предлагаемым трем вариантам: минимальному, среднему и максимальному. Можно предложить и более жесткие методы оценки, по которым потребность семьи в жилище будет еще меньше. Или, наоборот, менее жесткие, по которым она будет еще больше. Например, можно применить формулу s + t, по которой каждому члену семьи размера s полагается отдельная комната и добавляется t комнат на семью: спальни, гостиная, столовая. Или использовать европейские стандарты.

Квартиры делятся по числу больших и стандартных комнат. Это позволяет точнее сравнить потребности семей различных типов в жилище по минимальному, среднему и максимальному вариантам. В частности, по среднему варианту нужно меньше квартир типа 2Б и больше типов Б+2С, 2Б+2С, Б+3С (число больших + число стандартных комнат).

Анализ полученных результатов показал, что больше всего требовались 1-комнатные и 2-комнатные квартиры. Потребность в них составляла почти 80% общей: 36,5 % для 1-комнатных и 42% для 2-комнатных. При этом среди 1-комнатных выделяется тип 1Б, а среди 2-комнатных — Б+С. Деление комнат на большие и стандартные позволило выявить значительное различие потребностей в 2-комнатных квартирах типов 2Б, Б+С, 2С. Таким образом, подробное формализованное описание структуры семей позволило провести детальный качественный и количественный анализ потребностей в жилье. Он мог быть использован при планировании жилищного строительства.

Использование лонгитюдных данных в причинном моделировании

Терещенко Ольга Викентовна,
Белорусский государственный университет, Минск
Королева Илзе, *Университет Латвии, Рига*

Применение в социологии регрессионных причинных моделей долгие годы являлось предметом ожесточенных дискуссий. Философы — методологи науки — в принципе отрицали возможность статистического изучения причинных связей как не соответствующего логическому определению причинности, согласно которому следствие наступает в том и только в том случае, когда в наличии имеется причина. Позитивисты, допускавшие возможность статистического рассмотрения причинности как тенденции, в качестве единственного инструмента анализа причинных связей рассматривали эксперимент, в соответствии с идущей от О. Конта традицией. Наконец, аналитики, в принципе не отрицающие возможности использования регрессионных моделей, к которым принадлежит и А. О. Крыштановский[1], неоднократно указывали на существенные ограничения их применения, приводя соответствующие примеры. Аргументы в пользу регрессионного моделирования причинных связей (включая логит-, пробит- и логистические модели для номинальных и порядковых переменных) наиболее систематично изложены Г. Блейлоком, одним из разработчиков путевого анализа, в статье 1991 г.[2], после выхода которой дискуссия в целом пошла на спад.

Здесь мы рассмотрим только одно из возражений против применения регрессионных моделей к данным, полученным в ходе социологических опросов. Оно заключается в том, что в рамках одного опроса при выборе зависимых и независимых переменных часто невозможно установить, что является причиной, а что следствием, поскольку все переменные измеряются одновременно. Например, уровень доходов в качестве зависимой переменной можно объяснять имеющимися у респондентов ценностями и установками, но можно, наоборот, ценности и установки рассматривать как следствие уровня доходов и материального благополучия. Для решения подобных проблем М. Борном в 1949 г. были предложены постулаты статистической при-

[1] Крыштановский А. О. Ограничения метода регрессионного анализа // Социология: 4М. 2002. № 12.

[2] Blalock H. M. Are there really any *constructive* alternatives to causal modeling? // Sociological Methodology. 1991. Vol. 21.

чинности[1], проверка которых считается обязательной при использовании регрессионных моделей. Один из них гласит, что причина должна предшествовать следствию во времени или, применительно к социологическим данным, она должна быть *измерена* раньше, чем следствие. Исключения обычно делаются для аскриптивных характеристик респондентов[2], которые по умолчанию считаются сложившимися *до* проведения опроса, а также для данных, собранных ретроспективно, например, в исследованиях внутрипоколенной и межпоколенной социальной мобильности[3].

Лонгитюдные исследования, крайне редкие в постсоветских странах, предоставляют замечательную возможность решить данную проблему. Здесь равно обоснованно можно изучать как влияние ценностей, имевшихся при окончании школы, на доходы в 30-летнем возрасте, так и влияние доходов в родительской семье в год окончания школы на ценности, сложившиеся к 30 годам.

В данной работе мы стремимся показать, что использование лонгитюдных данных в регрессионном моделировании позволяет не только корректно соотносить причины и следствия при выборе зависимых и независимых переменных, но также исследовать *процессы* жизненного пути респондентов, используя в качестве переменных произошедшие между этапами лонгитюда изменения. Это касается как объективных показателей (изменение образовательного, профессионального, семейного статусов, уровня доходов и т. п.), так и субъективных оценок (удовлетворенность различными сторонами жизни, мотивация и т. п.)

Для построения моделей использованы данные лонгитюдного исследования «Пути поколения»[4] (руководитель проф. М. Титма). Этот международный проект с 1983 г. осуществлялся в республиках бывшего СССР, а с 1991 г. уже в независимых государствах. Генеральная совокупность — образовательная когорта, закончившая средние учебные заведения в 1983–1985 гг. (1965–1967 гг. рождения). Модели построены на данных, полученных в Латвии в ходе третьего (1993 г.) и четвертого (1998 г.) этапов исследования, когда респондентам, в среднем, было 27 и 32 года соответственно. Обычно это возраст стабилизации профессионального и семейного статуса, решения жилищных проблем и т. п. Однако для данного поколения он пришелся на трансформационный период, который в Латвии протекал крайне остро, и процессы социальной мобильности отличались большим разнообразием. Соответственно изменялись и оценки субъективного благополучия (subjective well-being), показателем которой может выступать, например, общая удовлетворенность жизнью.

Целью предпринятого нами анализа данных было изучение возможностей моделирования изменения удовлетворенности жизнью в зависимости от изменения объективных и субъективных факторов. Поскольку удовлетворенность может изменяться как в сторону улучшения, так и в сторону ухудшения, выборка была разделена на две части – тех, кто в 1993 г. в целом был удовлетворен своей жизнью (группа 1), и тех, кто не был (группа 2). Для каждой группы дихотомическая зависимая переменная показывала изменение оценки удовлетворенности жизнью в 1998 г. по сравнению с 1993 г.: 1 – удовлетворенность изменилась, 0 – осталась прежней. Другими словами, для первой группы изучалась возможность изменить удовлетворенность жизнью в целом на неудовлетворенность, для второй — наоборот.

Для построения независимых переменных выбирались только объективные и субъективные показатели качества жизни, которые могли за 5 лет реально измениться и в результате повлиять на стабильность / изменение удовлетворенности жизнью. Учитывались также интеркорреляции между потенциальными независимыми переменными и количество пропущенных значений. Окончательный список рассматриваемых объективных факторов включает брачный статус, условия жизни, недвижимую собственность и занятость; субъективных факторов — оценку здоровья, семейной жизни и сексуальной жизни. Каждая независимая переменная представлена

[1] Цит по: Sowa, J. S. Processes and Causality // http://www.jfsowa.com/ontology/causal.htm, 2001.

[2] См. напр.: Терещенко О. В., Титма М. Х. Дифференциация доходов в когорте тридцатилетних // Социологический журнал, 1996. № 3/4.

[3] Напр.: Елисеева И. И., Рукавишников В. О. Логика прикладного статистического анализа. М.: Финансы и статистика, 1982. С. 132–133.

[4] Социальное расслоение возрастной когорты / Отв. ред. М. Х. Титма. М.: ИС РАН, 1997.

тремя категориями: за 5 лет ситуация улучшилась / осталась без изменения / ухудшилась. Таким образом, в конструкции зависимых и независимых переменных реализован *процессуальный потенциал* лонгитюдных данных.

Для прогнозирования использовалась модель бинарной логистической регрессии[1]:

$$\ln\left(p/(1-p)\right) = b_0 + \sum b_i x_i \ ,$$

где p — вероятность того, что изменение произойдет, $1-p$ — вероятность того, что изменение не произойдет, $\ln\left(p/(1-p)\right)$ — логарифм отношения вероятностей (логит), x_i — независимые переменные, b_0 и b_i — свободный член и коэффициенты регрессии для независимых переменных.

Уравнение может быть также представлено в виде:

$$p = e^{b_0+\sum b_i x_i} \Big/ \left(1 + e^{b_0+b_i x_i}\right) \text{ или } p/(1-p) = e^{b_0} \prod e^{b_i x_i} \ .$$

Для включения независимых переменных в регрессионные модели использовалась техника фиктивных или dummy-переменных[2]. В качестве референтного везде использовалось значение «без изменений». Результаты моделирования представлены в табл. 1.

Таблица 1

Уравнения бинарной логистической регрессии

	Группа 1. Зависимая переменная: изменение удовлетворенности жизнью на неудовлетворенность			Группа 2. Зависимая переменная: изменение неудовлетворенности жизнью на удовлетворенность		
x_i	b_i	Sig.	Exp(b_i)	b_i	Sig.	Exp(b_i)
Брачный статус		0,665			0,219	
хуже	0,212	0,504	1,237	-1,196	0,082	0,302
лучше	0,403	0,519	1,497	0,018	0,982	1,018
Условия жизни		0,078			0,433	
хуже	0,841	0,088	2,319	-0,078	0,946	0,925
лучше	-0,658	0,156	0,518	1,667	0,197	5,299
Недвижимость		0,098			0,318	
хуже	0,738	0,033	2,093	-0,376	0,775	0,687
лучше	0,065	0,755	1,068	0,764	0,145	2,147
Занятость		0,086			0,450	
хуже	0,505	0,152	1,656	-1,115	0,226	0,328
лучше	-0,293	0,175	0,746	0,013	0,978	1,013
Оценка здоровья		0,008			0,828	
хуже	0,619	0,009	1,858	-0,095	0,872	0,910
лучше	-0,237	0,367	0,789	-0,339	0,540	0,713
Оценка семейной жизни		0,000			0,173	
хуже	2,263	0,000	9,611	-1,133	0,163	0,264
лучше	0,031	0,943	1,013	0,514	0,318	1,672
Оценка сексуальной жизни		0,776			0,753	
хуже	0,159	0,594	1,173	-0,426	0,481	0,653
лучше	0,187	0,609	1,205	0,049	0,930	1,050
Constant	-1,902	0,000	0,149	0,070	0,888	1,072

[1] Напр.: Бююль А., Цёфель П. SPSS: Искусство обработки информации: Анализ статистических данных и восстановление скрытых закономерностей. СПб: ДиаСофтЮП, 2002; Наследов А. SPSS: Компьютерный анализ данных в психологии и социальных науках, 2-е изд. СПб.: Питер, 2007.

[2] Напр.: Терещенко О. В. Dummy-кодирование // Социология: Энциклопедия / Сост. А. А. Грицанов и др. Минск: Книжный Дом, 2003. С. 304; Бююль А., Цёфель П. Указ. соч. С. 280, 293; Малхотра Н. К. Маркетинговые исследования: Практическое руководство, 4-е изд. М.: И. Д. Вильямс, 2007. С. 806.

В группе 1, удовлетворенной жизнью в 1993 г., представлено 764 респондента, из которых в 1998 г. 166 удовлетворены жизнью не были. Группа 2, не удовлетворенная жизнью в 1993 г., включает 110 респондентов, из которых в 1998 г. удовлетворены жизнью 53. Небольшим объемом этой группы, в частности, объясняется отсутствие статистически значимых коэффициентов в соответствующей модели. Из первоначальной выборки объемом 1202 человека данные 328 респондентов не могли быть использованы при построении моделей из-за пропущенных значений в одной или нескольких переменных.

Не вдаваясь в подробности интерпретации коэффициентов логистической регрессии и использования их для прогнозирования, что не входит в задачи данной работы[1], отметим следующее. Коэффициенты при фиктивных переменных показывают, как изменится вероятность события, представленного зависимой дихотомической переменной, при соответствующем значении независимой переменной *по сравнению с референтным значением.* А именно, положительное значение коэффициента показывает, что соответствующая вероятность увеличится, отрицательная — что уменьшится. Например, для группы 1, в целом удовлетворенной своей жизнью в 1993 г., ухудшение условий жизни повышает вероятность неудовлетворенности жизнью в 1998 г. (положительный коэффициент 0,841), а улучшение условий жизни такую вероятность снижает (отрицательный коэффициент –0,658), по сравнению с ситуацией, когда условия жизни остаются стабильными.

Эффективность уравнения логистической регрессии можно оценить, например, процентом правильной классификации. Для группы 1 модель правильно классифицирует 81,5 % респондентов. Однако более подробное рассмотрение итогов классификации показывает, что правильно классифицированы, главным образом, респонденты, сохранившие удовлетворенность жизнью (табл. 2); правильная классификация потерявших удовлетворенность составляет только 34,9 %.

Таблица 2

Процент правильной классификации по логистической модели для группы 1 (в 1993 г. удовлетворены)

По результатам опроса	Предсказание удовлетворенности по результатам моделирования		
	удовлетворены	не удовлетворены	% правильной классификации
удовлетворены в 1998 г.	565	33	94,5
не удовлетворены в 1998 г.	108	58	34,9
Общий %			81,5

Для группы 2 модель правильно классифицирует только 69,1 %, однако процент правильной классификации для тех, у кого удовлетворенность жизнью изменилась (повысилась), здесь составляет 75,5 %, что является гораздо лучшим результатом по сравнению с группой 1 (табл. 3).

Таблица 3

Процент правильной классификации по логистической модели для группы 2 (в 1993 г. не удовлетворены)

По результатам опроса	Предсказание удовлетворенности по результатам моделирования		
	не удовлетворены	удовлетворены	% правильной классификации
не удовлетворены в 1998 г.	36	21	63,2
удовлетворены в 1998 г.	13	40	75,5
Общий %			69,1

Таким образом, мы показали, *как* лонгитюдные данные позволяют изучать в динамике не только объективные, но и субъективные социальные процессы. Субъективное благополучие, представленное в данном примере удовлетворенностью жизнью, считается одним из наиболее сложных для моделирования и предсказания социологических показателей. Невысокое качество моделей объясняется в значительной мере отбором независимых переменных, носившем во многом дидактический характер. Однако наши результаты дают возможность предположить, что изменения субъективных и объективных факторов позволяют предсказать скорее повышение удовлетворенностью жизнью, нежели его снижение. Для проверки данного предположения

[1] См.: Бююль А., Цёфель П. Указ. соч.

в модели были дополнительно включены некоторые аскриптивные показатели: пол, знание титульного языка, образование. Интересно, что качество модели для первой группы фактически осталось тем же, в то время как для второй группы оно значительно улучшилось (табл. 4).

Таблица 4

Процент правильной классификации по логистической модели, включающей пол, знание титульного языка и образования, для группы 2 (в 1993 г. не удовлетворены)

По результатам опроса	Предсказание удовлетворенности по результатам моделирования		
	не удовлетворены	удовлетворены	% правильной классификации
не удовлетворены в 1998 г.	44	12	78,6
удовлетворены в 1998 г.	12	41	77,4
Общий %			78,0

Соответственно, можно предположить, что ухудшение субъективного благополучия намного более индивидуальный и сложный процесс, не поддающийся удовлетворительному моделированию с использованием тривиального набора независимых переменных. В то же время доля респондентов, у которых удовлетворенность жизнью снизилась, в полтора раза превышает долю респондентов, у которых она повысилась. Соответственно, определение факторов, влияющих на снижение субъективного благополучия, остается важной социологической задачей.

Модель квазипериодических колебаний в анализе эволюции интенсивности искусства

Харуто Александр Витальевич
Московская государственной консерватории им. П. И. Чайковского

Одной из важных задач анализа социокультурной сферы является выявление циклов, которые регулярно повторяются через определенный промежуток времени, т. е. периодических составляющих эволюции этой сферы[1]. Циклические изменения стиля изобразительного искусства с периодом порядка 500 лет были обнаружены П.А. Сорокиным[2]; С.Ю. Маслов показал наличие 50-летних волн в социально-политическом климате России и синхронных с ними стилевых волн в русской архитектуре[3]. Исследования циклических изменений стиля в разных видах искусства Европы и США провел К. Мартиндейл[4]. В работах Г. Лемаршана и др.[5] было показано, что количество европейских композиторов, одновременно участвовавших в творческом процессе в Европе в 950–2000 гг., имеет колебания с периодом около 160 лет при общем экспоненциальном росте кривой.

В последнее время при участии автора проводились более широкие исследования эволюции *интенсивности творчества* в разных областях художественной жизни России и Западной Европы[6]. В качестве исходных данных в таких исследованиях используются *количественные показатели* — например, число деятелей культуры, родившихся в *некоторый небольшой период времени* и описанных в специальных энциклопедиях, т. е. деятелей, впоследствии ставших (по

[1] Петров В. М. Волнообразные социальные процессы: к методике прогнозирования // Социология: 4М. 2004. № 18. С. 130–153.

[2] Сорокин П. А. Социальная и культурная динамика / Пер. с англ., вст. ст. В. В. Сапова. М.: Астрель, 2006.

[3] Маслов С. Ю. Асимметрия познавательных механизмов и ее следствия // Семиотика и информатика. М., 1983. Вып. 20. С. 3–34.

[4] Martindale C. The Clockwork Muse: The Predictability of Artistic Change. NY: Basic Books, 1990.

[5] Mallmann C.A., Lemarchand G.A. Generational explanation of long-term 'billow-like' dynamics of societal processes // Technological Forecasting and Social Change. 1998. Vol. 58. P. 1–30; Lemarchand G. A cyclic model of long-term recurrences in societal processes: Application to the millenary behavior of classical music (950–2000) / Искусствознание и теория информации: Сб. науч. ст. / Под ред. В. М. Петрова, А. В. Харуто. М.: Красанд, 2009. С. 234–244.

[6] См.: Харуто А. В., Коваленко Т. В. и др. Интенсивность российской художественной жизни: анализ периодических компонент // Социология 4М. 2007. № 25. С. 142–166; Харуто А. В., Коваленко Т. В. Исследование интенсивности русской и западноевропейской театральной жизни: анализ периодических компонент // Математическое моделирование социальных процессов. М.: МГУ, 2007. Вып. 9. С. 126–144.

мнению экспертов) «лицом» данной ветви художественной жизни (такой показатель использовался в работах Г. Лемаршана). Другим важным показателем *интенсивности творчества* на данном отрезке времени является *объем текста*, посвященного в сумме всем этим деятелям в данном справочном издании. При этом предполагается, что чем весомее вклад деятеля культуры, чем больше *интенсивность* его влияния на соответствующую ветвь художественной жизни, тем больший объем текста будет выделен ему в энциклопедии.

Ранее в работах автора было проведено сравнение различных методов выявления периодичности и показано, что наиболее эффективным является анализ на основе спектра Фурье, позволяющий выявить основные гармонические компоненты[1]. При этом сначала определяется «оптимальная» степень многочлена, аппроксимирующего немонотонный тренд, содержащийся в данных, и линия тренда вычитается из исходного ряда. Поскольку для анализа доступны обычно только несколько циклов колебаний, оценка спектра дает приближенный результат, который может быть далее уточнен путем численной оптимизации параметров аппроксимирующего тригонометрического ряда по среднеквадратичному критерию.

Пример исходных социокультурных данных, содержащих медленный немонотонный тренд и «быстрые» колебания, показан на рис. 1 (здесь представлена интенсивность композиторского творчества Италии с 1480 по 1920 гг. по данным словаря Гроува[2], использованного П. А. Куличкиным[3]; данные суммировались по композиторам, родившимся в каждое 10-летие). Линия тренда построена путем вычисления аппроксимирующего многочлена 9-й степени.

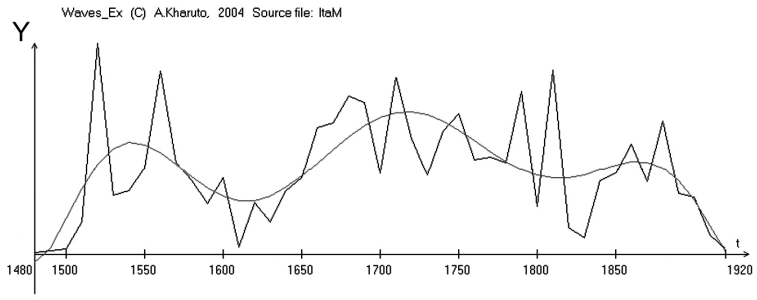


Рис 1. Эволюционная кривая интенсивности композиторского творчества Италии и линия тренда (полином 9-й степени)

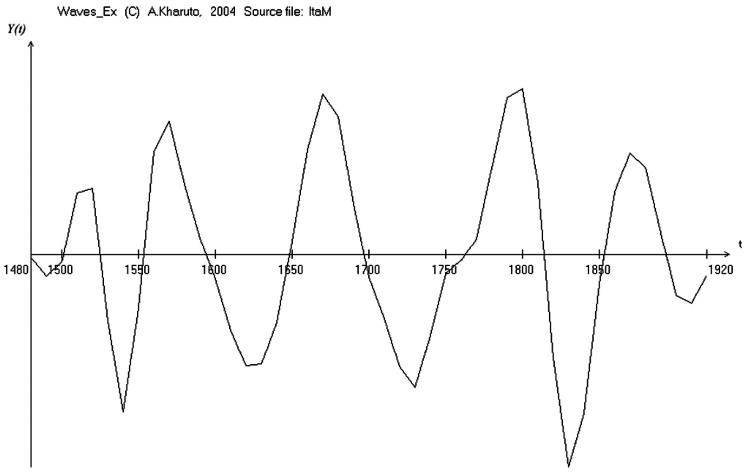


Рис. 2. Интенсивность композиторского творчества Италии: центрированная и сглаженная эволюционная кривая

[1] Харуто А. В. Методы анализа периодических компонент в социокультурных данных // Математическое моделирование социальных процессов / Под ред. А. П. Михайлова. М.: КДУ, 2009. Вып. 10. С. 439–455; Kharuto A. V. Periodical waves in the evolution of art: methods of study // Key Engineering Materials. 2010. Vol. 437. P. 530–534.

[2] Grove G. Grove's Dictionary of Music and Musicians: 5th edition / Ed. by Eric Blom. L.: Macmillan, 1954.

[3] Куличкин П. А. Эволюция художественной жизни и стиля мышления (опыт количественного исследования): Автореф… дисс.... канд. культурологи / Гос. ин-т искусствознания. М., 2004.

Для дальнейшего анализа путем вычитания тренда и сглаживания (треугольной взвешивающей функцией) выделяется «колебательная часть» эволюционной кривой, показанная на рис. 2. Сглаживание позволяет подавить «слишком быстрые» колебания в рядах данных — те, которые не представляют интереса в данной работе.

Расчет спектра Фурье для этой «колебательной части» эволюционной зависимости дает многогорбую кривую, показанную на рис. 3.

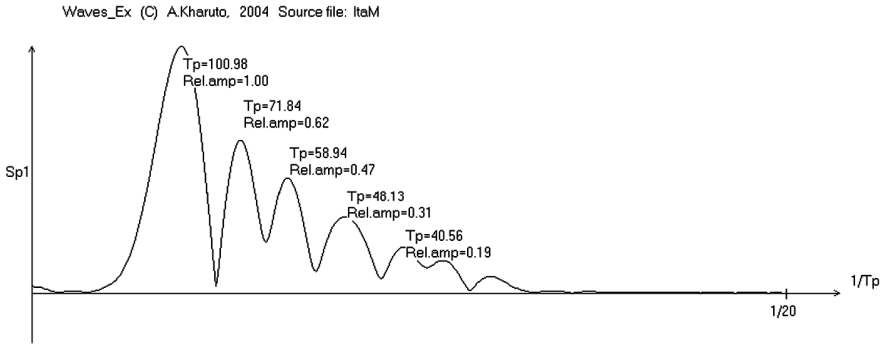


Рис. 3. Интенсивность композиторского творчества Италии:
спектр колебательной части центрированной эволюционной кривой

I. Модель с постоянными параметрами

Первоначальный расчет спектра, как видно из рис. 3, обнаруживает наличие пяти-семи основных «пиков», соответствующих гармоническим составляющим эволюционной кривой. Однако синтез этой кривой по таким начальным оценкам параметров компонент дает среднеквадратичную погрешность около 48%, что объясняется упоминавшейся уже значительной погрешностью оценки из-за очень слишком короткого (по сравнению с оцениваемыми периодами) ряда исходных данных[1]. При дальнейшей оптимизации независимо варьировались частоты, начальные фазы и амплитуды каждой гармонической составляющей. Поиск минимума среднеквадратичной погрешности осуществлялся путем чередования методов случайного и градиентного спуска. В результате было получено приближение, относительная погрешность которого составила около 29 %.

Далее, сопоставление основного по мощности «пика» (соответствующего периоду около 100 лет) с теоретической шириной главного лепестка спектра $\Delta F = 2/T_A$, определяемой интервалом анализа T_A, показало, что реальная ширина вычисленного спектрального лепестка заметно превышает ΔF. Это может объясняться наличием двух близких по частоте спектральных составляющих, которые при имеющемся разрешении (ограниченном величиной T_A) невозможно различить. Исходя из этих соображений, была добавлена гармоническая компонента с периодом около 130 лет (в процессе оптимизации эта величина изменилась). В результате было получено приближение с погрешностью аппроксимации около 9 % (рис. 4).

Спектральные составляющие показаны здесь треугольниками, повернутыми вниз от оси частот; высоты треугольников пропорциональны амплитудам, а основания — ширине главных лепестков спектра, т. е. равны $\Delta F = 2/T_A$. Следует отметить, что периоды обнаруженных гармонических составляющих не находятся здесь в кратных отношениях, т. е. исследуемое колебание не проявляет свойств периодичности на исследуемом временном участке. Число использованных для аппроксимации составляющих равно семи, что не позволяет предложить какую-либо очевидную трактовку их происхождения и непосредственно использовать результат в построении социокультурной модели.

[1] См.: Харуто А. В. Методы анализа периодических…

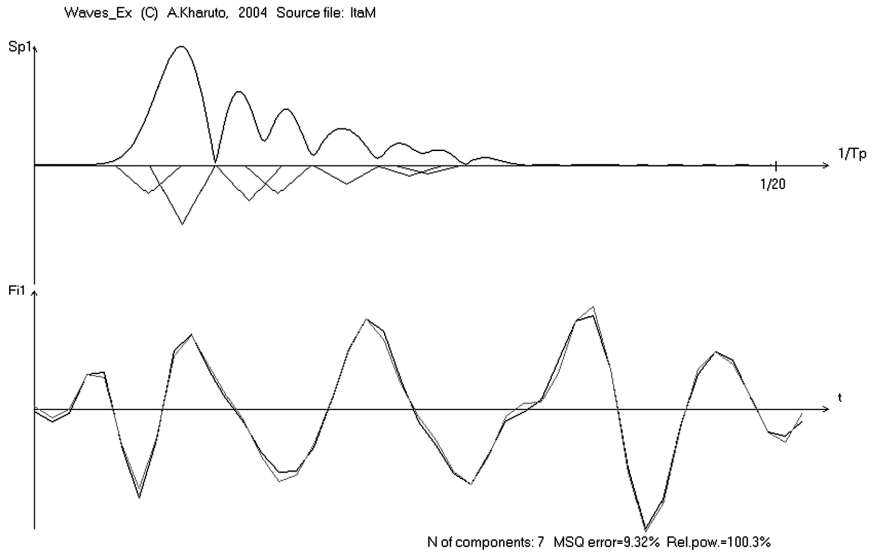


Рис. 4. Результат аппроксимации эволюционной кривой (7 гармонических компонент).

II. Модель с переменными параметрами колебания

Большое число параметров составляющих, получаемых в модели с постоянными параметрами, а также отсутствие строгой периодичности в исследуемых данных (переменная длительность цикла колебаний) дает основание для построения *квазипериодических* моделей таких колебаний, где величина «периода» зависит от времени. Поскольку в данной задаче основной интерес представляет именно временная структура колебаний, вариациями амплитуды можно пренебречь. Тогда модель колебаний может быть описана в виде

$$Y(t) = A_0 \sin[2\pi \times \theta(t) \times t + \varphi_0], \qquad (1)$$

где A_0 — постоянная амплитуда, φ_0 — начальная фаза колебаний, $\theta(t) = 1/T_Q(t)$ — мгновенная частота, $T_Q(t)$ — квазипериод колебаний.

Мгновенная частота как функция времени может быть, в свою очередь, представлена суммой «постоянной составляющей» f_0 и переменной части $f_m(t)$:

$$\theta(t) = f_0 + f_m(t). \qquad (2)$$

Здесь предполагается, что колебание моделируются всего одной гармоникой (1), параметры которой изменяются во времени, но в принципе число гармоник может быть увеличено. Характер изменений мгновенной частоты колебаний заранее неизвестен и должен быть выявлен при обработке эмпирических данных. Для аппроксимации функции $\theta(t)$ можно использовать, например, степенной многочлен либо тригонометрический ряд. В последнем случае мгновенная частота представляется как

$$\theta(t) = f_0 + \sum_{k=1}^{k=N} B_k \sin(2\pi k f_1 t + \varphi_k), \qquad (3)$$

где B_k и φ_k — искомые параметры k-й гармоники ряда, представляющего изменение мгновенной частоты колебаний эволюционной кривой во времени, f_1 — частота первой гармоники в (3).

Оценку $\hat{\theta}(t)$ — зависимости длительности цикла колебаний от времени — можно в первом приближении построить, используя точки пересечения оси абсцисс с колебательной частью эволюционной кривой $Y(t)$. Расстояния между каждыми двумя соседними точками дают оценку половины «текущего» квазипериода на соответствующем временном интервале. Следует отметить, что такая оценка $\hat{\theta}(t)$ оказывается весьма приближенной хотя бы потому, что заменяет истинную функцию последовательностью «ступенек»; подстановка первоначальной аппроксимации $\theta(t)$ в формулу (1) приводит к наличию «скачков» (разрывов 1-го рода) в модели коле-

бания, которых в эволюционной кривой нет. В результате погрешность первого приближения обычно превышает 100 %, т. е. исходная и моделирующая кривая эволюционных колебаний в значительной своей части не совпадают.

С целью уточнения параметров модели при использовании Фурье-аппроксимации (3) производится минимизация функции среднеквадратичной погрешности $\varepsilon(\vec{x})$ приближения колебательной части эволюционной кривой функцией (1); поиск производится по параметрам B_k, φ_k всех гармоник (3), частоте f_1 и параметрам «базового» синусоидального колебания — A_0, f_0, φ_0. Амплитудные вариации предварительно нивелируются путем кусочно-постоянного масштабирования исходной кривой между точками пересечения ею оси абсцисс. Кроме того, с целью выделения только «длинных» волн (60–100 лет) сглаживание треугольной функцией проводилось дважды, что позволяет увеличить относительное подавление «слишком быстрых» колебаний с периодами 20–40 лет, которые ярко выражены в исходной эволюционной кривой, но не исследуются на данном этапе анализа. Все необходимые операции выполняются в разработанной автором программе Wave_Ex.

В рассматриваемом примере (композиторское творчество Италии) оптимизация производилась при интервале представления данных $T_A = 440$ лет; оценка среднего периода колебаний интенсивности дала величину $T_0 = 67$ лет. Основной период изменения частоты колебаний интенсивности получился равным $T_1 = 1014$ лет; для представления изменений мгновенной частоты использован один член ряда (3). При этом среднеквадратичная погрешность аппроксимации исходной кривой составляет 20,3%. Графически результат аппроксимации для такой модели с переменной частотой колебаний представлен на рис. 5.

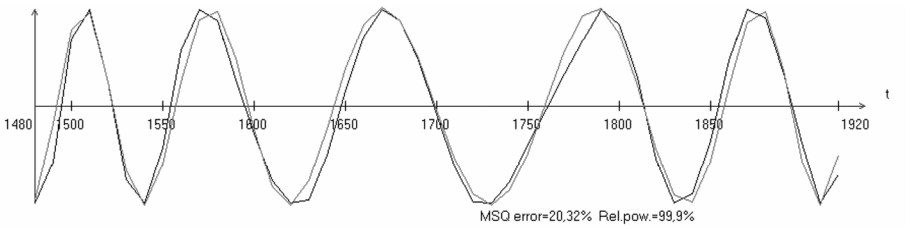


Рис. 5. Результат оптимизация модели с аппроксимацией мгновенной частоты колебаний рядом Фурье (интенсивность композиторского творчества Италии).

Аналогичный анализ был проведен также для данных по интенсивности композиторского творчества Англии и Австрии с Германией (традиционно составляющих одну композиторскую школу), а также для Европы в целом. В первых двух случаях аппроксимация колебательной части эволюционной кривой была получена с погрешностью соответственно 16 % и 14 % при использовании одного члена ряда (3). Для Европы в целом понадобились как минимум 3 гармоники такого ряда; при этом погрешность аппроксимации составила 30 %.

График изменения квазипериода исследуемых эволюционных колебаний в историческом времени представлен на рис. 6. Итальянская школа обнаруживает в целом нарастающий период эволюционных колебаний, в то время как английская и австро-германская — волнообразный (с небольшим размахом). Зависимость общего квазипериода эволюционных колебаний для Европы в целом имеет колебательный характер (пределы вариаций от 90 лет до 110 лет); на интервале в 500 лет квазипериод в среднем возрастает. Следует отметить, что эта кривая не является ни суммой, ни взвешенной суммой трех других: для получения эволюционной кривой интенсивности композиторского творчества по Европе суммировались оценки интенсивности (т. е. объемы текстов, посвященных композиторам в словаре Гроува), и исследованию на квазипериодические составляющие подвергалась эта кривая суммарной интенсивности.

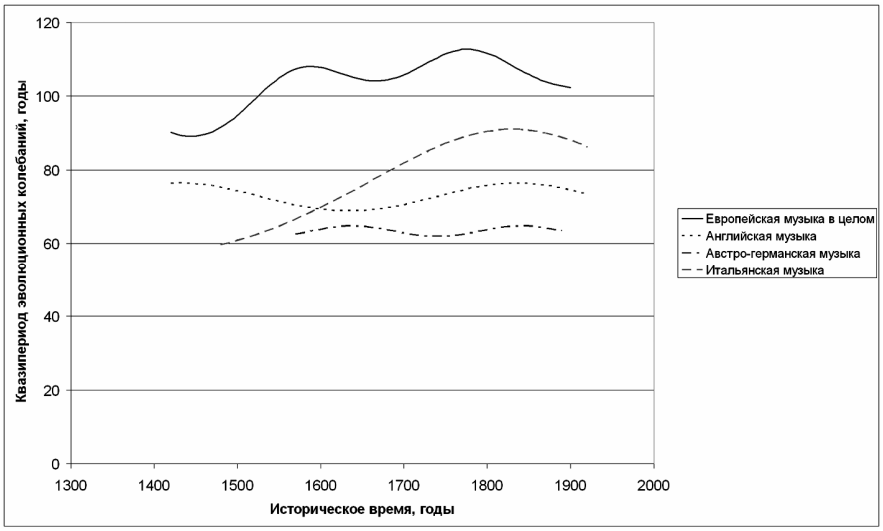


Рис. 6. Изменение квазипериода колебаний интенсивности композиторского творчества в Италии, Англии, Австрии и Германии и Европе в целом.

По результатам расчетов можно также заметить, что в функции (3), описывающей изменения квазипериода, период основной гармоники во всех случаях в несколько раз превышает величину самого квазипериода эволюционных колебаний (63–73 года), т. е. ряд Фурье используется как средство аппроксимации исследуемой функции на сравнительно небольшом отрезке собственного периода этого ряда. Зависимости, представленные на рис. 6, можно непосредственно использовать для построения моделей исследуемых социокультурных колебаний в хронологических рамках 1420–1920 гг.

Рассмотренные выше два типа моделей эволюционных кривых — с постоянными параметрами и с переменными — предполагают разную структуру соответствующих социокультурных подсистем. В случае модели с постоянными параметрами в социуме должны присутствовать несколько *параллельно действующих* «социокультурных осцилляторов» одного иерархического уровня, т. е. несколько социальных подсистем, проявляющих колебательный характер гармонического типа, причем число таких осцилляторов оказывается довольно велико (5–8), а параметры их колебаний довольно близки (периоды составляют 50–100 лет), но все же различны, что затрудняет трактовку «физической сущности» компонентов модели. Эволюционной же модели с «переменными параметрами» соответствуют всего *два* «осциллятора» *разного иерархического уровня*, т. е. один «осциллятор», непосредственно описывающий колебания интенсивности творческой деятельности (со средним периодом 60–70 лет в случае рассмотренных национальных композиторских школ и 90–110 лет для Европы в целом), и второй, с гораздо большим периодом колебаний, который *управляет* параметрами (частотой колебаний) первого, т. е. находится выше в иерархии. Колебания «управляющего» осциллятора могут содержать гармоники (как в случае Европы в целом), т. е. иметь выраженную несинусоидальную форму. При этом число параметров в последнем классе моделей оказывается заметно меньшим, чем в моделях с постоянными параметрами, а структура модели допускает более ясную содержательную интерпретацию. В дальнейшем планируется исследовать аналогичным образом и другие имеющиеся ряды экспериментальных данных — для живописи, театральной и драматургической деятельности, литературы.

Алгоритм деревьев решений C4.5 в оценке качества образования

Чудова Олеся Владимировна,
Алтайский государственный университет

В настоящее время повышение качества образования — одна из актуальных проблем для всего мирового сообщества. Современное определение качества не только определяет его оценку, но и связывается с потребителем. Работодатели как потребители результатов образователь-

ных систем оценивают качество образования и подготовленность специалистов по уровню их компетентности.

В связи с внедрением государственных образовательных стандартов высшего профессионального образования третьего поколения образовательная политика и практика работы всех высших учебных заведений будет перестроена в соответствии с компетентностным подходом. Одним из важнейших вопросов является построение методики оценки уровня сформированности как отдельных компетенций, так и групп компетенций. В федеральных государственных образовательных стандартах высшего профессионального образования (ФГОС ВПО) по каждому направлению подготовки определен список профессиональных компетенций (уровень частных компетенций), которыми должен обладать выпускник. Все профессиональные компетенции объединены в группы по виду деятельности (уровень промежуточных компетенций): проектная деятельность, организационно-управленческая и производственно-технологическая деятельность, аналитическая деятельность, научно-исследовательская деятельность. Уровень подготовки к осуществлению перечисленных видов деятельности и определяет уровень сформированности профессиональной компетентности выпускника ВУЗа. Состояние всех элементов системы будет определять комплексную оценку профессиональной компетентности.

Частные и промежуточные профессиональные компетенции можно формировать постепенно в процессе обучения и уровень их сформированности может быть измерен по мере реализации дисциплин учебного плана направления подготовки, с использованием методов математической статистики[1].

Оценка общего уровня сформированности профессиональной компетентности достаточно сложна и плохо формализуема, она зависит от специфики специальности, потребностей рынка труда и не может быть осуществлена без учета мнения работодателей, специалистов в данной области, выпускников. Для ее нахождения могут быть использованы продукционные системы[2]. Причинами выбора данного метода явилась необходимость объяснения причин полученной оценки, небольшая размерность входных показателей.

Продукционные модели близки к логическим моделям, что позволяет организовывать на них эффективные процедуры вывода, а с другой стороны, более наглядно отражают знания, чем классические логические модели. В них отсутствуют жесткие ограничения, характерные для логических исчислений, что дает возможность изменять интерпретацию элементов продукции. Продукционная модель, или модель, основанная на правилах, позволяет представить знания в виде предложений типа: Если (условие), то (следствие): $(Условие\ 1)\ \&\ (Условие\ 2)...(Условие\ T) \rightarrow (Следствие)$.

Правила продукции разрабатываются с помощью экспертов в данной предметной области. Экспертная деятельность в области образования — система действий, выполняемых с привлечением экспертов, для анализа и оценки качества образовательного процесса с целью повышения обоснованности принимаемых решений в условиях частичной неопределенности, противоречий или конфликтов. В качестве экспертов могут выступать студенты, выпускники вуза, потенциальные работодатели и преподаватели. Одним из существенных условий повышения надежности экспертных оценок социальных объектов является научно обоснованный отбор и формирование экспертной группы. Точность групповой оценки экспертов зависит от численности экспертной группы: уменьшение числа экспертов ведет к снижению точности оценок, т. к. оценка каждого эксперта приобретает больший вес, увеличивается роль субъективного фактора. Однако при большом количестве участников экспертизы усложняется обработка их суждений, становится сложным выявить согласованность их мнений[3].

---

[1] Чудова О.В. Применение методов многомерной классификации в оценке компетентности выпускников // Известия АлтГУ. Барнаул, 2009. С. 93–95; Чудова О.В. Применение методов свертки для оценки уровня сформированности компетенций выпускника ВУЗА // Образование и наука в третьем тысячелетии. Барнаул, 2009. Т. 5. С. 87–91.

[2] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2001.

[3] Берестнева О.Г., Марухина О.В. Компьютерная система принятия решений по результатам экспертного оценивания в задачах оценки качества образования // Казань, 2002. Вып. 3. С. 216–230.

Целью проведения экспертного опроса является изучение мнения специалистов о предложенных ФГОС ВПО компетенциях и необходимом уровне сформированности этих компетенций для успешной профессиональной деятельности. Эксперту предлагается определить минимальный уровень сформированности частной профессиональной компетенции для выполнения профессиональных задач на заданном уровне. Введем следующие обозначения: y — показатель оценки общего уровня сформированности профессиональной компетентности, y_1 — показатель оценки проектной деятельности; y_2 — показатель оценки организационно-управленческой и производственно-технологической деятельности; y_3 — показатель оценки аналитической деятельности; y_4 — показатель оценки научно-исследовательской деятельности; y_{ij} — показатель оценки частной профессиональной компетенции, формирующей промежуточную компетенцию y_i ($i=\overline{1,4}$, $j=\overline{1,n_i}$, где n_i — количество частных профессиональных компетенций, формирующей промежуточную компетенцию y_i); z_{ijl}^k — оценка k-м экспертом минимального уровня сформированности частной профессиональной компетенции y_{ij} для выполнения профессиональной деятельности на уровне l. Тогда пороговые значения промежуточных компетенций вычислим по формуле:

$$z_{il}^k = \frac{\sum_{j=1}^{n_i} z_{ijl}^k}{n_i}.$$

В результате получим прямоугольную таблицу, на основании которой будет построена система правил продукции.

Таблица 1

Исходные данные для построения правил продукции

№ эксперта	Проектная деятельность	Организационно-управленческая и производственно-технологическая деятельность	Аналитическая деятельность	Научно-исследовательская деятельность	Профессиональная компетентность
1	z_{11}^1	z_{21}^1	z_{31}^1	z_{41}^1	z_1
1	z_{12}^1	z_{22}^1	z_{32}^1	z_{42}^1	z_2
……					
n	z_{11}^n	z_{21}^n	z_{31}^n	z_{41}^n	z_1
n	z_{12}^n	z_{22}^n	z_{32}^n	z_{42}^n	z_2
……					

В качестве инструментария для построения продукционных моделей может быть использован метод построение деревьев решений C4.5. Деревья решений являются наиболее распространенным в настоящее время подходом к выявлению и визуализации логических закономерностей в данных. Каждому узлу сопоставлен некоторый признак, а ветвям – либо конкретные значения для качественных признаков, либо области значений для количественных признаков.

По результатам экспертного опроса имеем множество примеров $\dot{O}$, где каждый элемент этого множества описывается $m=5$ атрибутами. Количество примеров в множестве T будем называть мощностью этого множества и будем обозначать $|T|$. Пусть метка класса принимает следующие значения $C_1, C_2 \ldots C_1$.

Задача заключается в построении иерархической классификационной модели в виде дерева из множества примеров T. Процесс построения дерева происходит сверху вниз. Сначала создается корень дерева, затем потомки корня и т.д.

На первом шаге мы имеем пустое дерево (имеется только корень) и исходное множество T (ассоциированное с корнем). Требуется разбить исходное множество на подмножества. Это можно сделать, выбрав один из атрибутов в качестве проверки. Тогда в результате разбиения

получаются n (по числу значений атрибута) подмножеств и, соответственно, создаются n потомков корня, каждому из которых поставлено в соответствие свое подмножество, полученное при разбиении множества T. Затем эта процедура рекурсивно применяется ко всем подмножествам (потомкам корня) и т.д. В результате получаем правила вида:

$$Если\ y_1 > 85\ \ и\ y_2 > 80\ и\ y_3 > 87\ и\ y_4 > 81 \quad то\ y = высокий.$$

Построенное дерево решений используется для распознавания нового объекта. Обход дерева решений начинается с корня дерева. На каждом внутреннем узле проверяется значение объекта по атрибуту, который соответствует проверке в данном узле, и, в зависимости от полученного ответа, находится соответствующее ветвление, и по этой дуге двигаемся к узлу, находящему на уровень ниже и т.д. Обход дерева заканчивается, как только встретится узел решения, который и дает название класса объекта.

Такая же методика применяется, когда дерево используется для классификации новых примеров. Если на каком-то узле дерева при выполнении проверки выясняется, что значение соответствующего атрибута классифицируемого примера пропущено, то алгоритм исследует все возможные пути вниз по дереву и определяет, с какой вероятностью пример относится к различным классам. В этом случае, «классификация» — это скорее распределение классов. Как только распределение классов установлено, то класс, имеющий наибольшую вероятность появления, выбирается в качестве ответа дерева решений. Процедура построения деревьев решений была проведена с использованием аналитической платформы Deductor (см.: http://basegroup.ru/).

В результате система правил продукции позволила исследовать возможности оценки и анализ компонентов профессиональной компетентности студента. Преимуществом построенной модели является то, что после каждой процедуры контроля оценки уровня сформированности профессиональных компетенций преподаватель и студент может получать индивидуальную диаграмму уровней сформированности компетенций.

Градуировка коэффициента Джини
(Памяти В.И. Арнольда (1937–2010))[1]

Шмерлинг Дмитрий Семенович, *НИУ ВШЭ*

Проблема неравенства в доходах хорошо известна, по крайней мере, с работы Макса Лоренца об измерении концентрации богатства[2]. Один из наиболее распространенных методов измерения неравенства стал расчет коэффициента (индекса) Джини:

$$\Delta_1 = \frac{1}{N(N-1)} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \left| x_j - x_k \right| f(x_j) f(x_k) \qquad (1)$$

(для дискретного случая, без учета совпадений), где x_1, x_2, … — величина доходов, $f(x_1)$, $f(x_2)$ — вероятность (или частота по выборке) людей с доходами x_1, x_2, … соответственно.

Величину Δ_1 обычно нормируют так, чтобы $\Delta_1^* = \Delta_1 / \Delta_1^{max} \in [0,1]$, при этом, чем она больше (ближе к 1), тем значительней неравенство населения.

Как известно, удвоенная площадь между диагональю и кривой рассеяния равно коэффициенту Джини. Заметим, что площадь над кривой рассеяния (Лоренца) и под диагональю[3] равна $\Delta_1/4$

[1] Автор благодарит В.И. Арнольда, А.Я. Кируту, Я.Ю. Никитина, А.И. Орлова, Ю.Н. Толстову, Ю.Н. Тюрина, В.В. Ульянова за содействие и обсуждение.

[2] Lorenz M.D. Methods of Measuring the Concentration of Wealth // Publ. Amer. Statist. Ass. 1905. Vol. 9. No. 70. P. 209–219; Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. § 2.25. Впрочем, о неравенстве писал и В.И. Ленин в работе «Развитие капитализма в России» (1899).

[3] Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. С. 75. Рис. 2.2.

μ_1, где $\mu_1 = \int_{-\infty}^{\infty}(x-a)f(x)dx$, f(x) — плотность распределения, обычно организуют a=0[1]. Собственно кривая рассеяния есть «неполный первый момент распределения»[2]:

$$\phi(x) = \frac{1}{\mu_1} \int_{-\infty}^{x} xf(x)dx \quad (2)$$

Коэффициент Джини вычисляется и публикуется для большинства стран уже десятки лет, в т.ч. со группированными данными[3]. К примеру в Норвегии он составлял 0,25 (2008 год), во Франции 0,32 (2008 год), в России 0,42 (2008 год), в США 0,45 (2007 год), в Мексике 0,48 (2008 год), в Южной Африке 0,65 (2005 год), в Намибии 0,71 (2003 год)[4].

Коэффициент Джини измеряет величину дифференциации доходов населения, «богатств», расходов, ВВП регионов или стран и тому подобных показателей, которые по-английски все вместе называются «Size» (эквивалентный русский термин отсутствует). Значения, большие, чем 0,3–0,4, по мнению большинства специалистов, свидетельствуют о высоком неравенстве и приводят к замедлению темпов развития стран, например, вследствие «ловушки бедности» [5]. Читатель-экономист, может быть, уже привык к таким данным, но насколько обществом понят смысл значений коэффициента Джини?

Существует обширная литература о вреде высокого (> 0,3) коэффициента Джини[6]. Однако граница 0,3 выбрана достаточно произвольно и представляет что-то около среднего общеевропейского коэффициента Джини. И было бы полезно поискать за значениями нормированного коэффициента G', $0 \le G' \le 1$, какой-нибудь «предметный» (например, экономический) смысл.

Рассмотрим следующую модель. Пусть x_i — доход лиц, относящихся к i-му уровню иерархии применительно к компании, населению территории и т.п. Модель P такова, что доход на i-м уровне (i = 1 — лица с наименьшим, а i = n — с наибольшими доходами) равен $x_i = ki^m$, m = 1, 2, 3, …, k > 0.

Теорема: коэффициент Джини $G_m'(n)$ для модели P равен асимптотически при $n \to \infty$

$$G_m'(n) = \frac{m}{m+2} \quad (1a)$$

Набросок доказательства:

$$G'(n) = \frac{"\sigma"(n)}{\max\limits_x "\sigma"(n)}, \quad (2a)$$

где maximum берется по всем возможным $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, таким, что

$$\sum_{1 \le i \le n} x_{(i)} = C(n), \quad (3)$$

$$"\sigma"'(n) = \frac{2\sqrt{\pi}}{n(n-1)} \sum_{1 \le i \le n} (i - \frac{n+1}{n}) x_{(i)}, \quad (4)$$

---

[1] Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. § 2.3.

[2] Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. С. 75–77. § 2.31.

[3] Gastwirth Y.L. The Estimation of the Lorenz Curve and Gini Index // Rev. of Econ. Statistics. 1972. Vol. 52. No. 3. P. 306–316; Moderres R., Gastwirth J.L. A Cautionary Note on Estimating the Standard Error of the Gini Index of Inequality // Oxford Bull. of Econ. Statist. 2006. Vol. 68. P. 385-390; Morgan J. The Anatomy of Income Distribution // Rev. of Econ.Statist. 1962. Vol. 44. No 3. P. 270–283.

[4] Wikipedia. List of Countries by income equality // http://en.wikipedia.org/wiki/List_of_countries_by_income_equality

[5] См., например: Atkinson A.B., Bourguignon F. (ed.) Handbook of Income Distribution. Amsterdam et al.: Elsevier, 2000. Vol. 1.

[6] См., например: Atkinson A.B., Bourguignon F. (ed.) Handbook of Income Distribution. Amsterdam et al.: Elsevier, 2000. Vol. 1.

$X_{(i)}$ – i-ая порядковая статистика[1]

$$"\sigma"\ (n) = \frac{1}{2}\sqrt{\pi}G(n), \quad (5)$$

$$G(n) = \frac{1}{n(n-1)} \sum_{1 \le i.j \le n} \left| x_i - x_j \right|. \quad (6)$$

Используя другую форму "σ"(n)= "σ"

$$"\sigma" = \frac{\widetilde{\kappa} 2 * \sqrt{\pi}}{n(n-1)} \left\{ \sum_{1}^{n} {}_i x_{(i)} - \frac{n+1}{n} \sum_{1}^{n} x_{(i)} \right\} \quad (7)$$

и вычисляя

$$\max_x "\sigma" = \frac{\sqrt{\pi} * \widetilde{\kappa}}{n} S_m(n) \quad (8)$$

где $S_m(n) = \sum_{k=0}^{n-1} k^m$ [2]

сумма целых чисел в степени m=1, 2, 3,…, k = 0, 1, 2,…., n–1, можно получить выражение для (3). Именно, при

$$\widetilde{S}_m(n) = S_m(n) + n^m,$$

$$"\sigma" = \frac{2\widetilde{\kappa}\sqrt{\pi}}{n(n-1)} \left\{ \widetilde{S}_{m+1}(n) - \frac{n+1}{n} \widetilde{S}_m(n) \right\} / \frac{\widetilde{\kappa}\sqrt{\pi}}{n} \widetilde{S}_m(n) \quad (9)$$

Откуда (здесь m – напоминает о степени многочлена в формуле $x_i = ki^m$, m = 1, 2, 3, …, k > 0)

$$G'_m(n) = \frac{2}{n-1} \left\{ \frac{\widetilde{S}_{m+1}(n)}{\widetilde{S}_m(n)} - \frac{n+1}{2} \right\} \quad (10)$$

Теперь нам понадобятся выражения $S_i(n)$, приведенные в удобной форме в уже цитировавшейся великолепной книге Р. Грехэма с соавторами[3]:

$$\widetilde{S}_m(n) = \frac{1}{m+1} \sum_{0 \le k \le n} \binom{m+1}{k} B_k n^{m+1-k} \quad (11),$$

где B_k, k = 0, 1, 2,…., числа Якова Бернулли, а именно:

k	1	2	3	4	5	6	7	8	9	10	11	12	…
B_k	1	$-\frac{1}{2}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	0	$\frac{5}{66}$	0	$-\frac{691}{2730}$	…

[1] См.: Дейвид Г. Порядковые статистики: Пер. с англ. М.: Наука, 1978. С. 187–189 (§7.4), 214 (§9.6), где обсуждается асимптотическая нормальность "σ". При этом $E"\sigma" = 2\sqrt{\pi \int_{-\infty}^{+\infty} x\left[P_{(x)} - \frac{1}{2}\right] dP_{(x)}}$, "σ" — несмещенная оценка для σ в случае нормальных выборок, $P_{(x)}$ — функция распределения.

[2] См.: Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики / Пер. с англ.; 3-е изд. М.: БИНОМ; Лаборатория знаний; СМР, 2009.

[3] Там же. § 61.78.

Формулы Sm(n), m = 0, 1, 2,…, 10 см. в упомянутой книге с. 314. Из (10) легко получается при n→∞

$$G'_m(n) \cong \frac{2}{n-1}\left\{\frac{m+1}{m+2}n - \frac{n+1}{2}\right\} \approx \frac{m}{m+2}, QED.$$

Приведем таблицу 2 для $G_m'(n)$[1]

m	1	2	3	4	5	6	7	8	9	10	…
$G_m'(n)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{5}$	$\frac{2}{3}$	$\frac{5}{7}$	$\frac{3}{4}$	$\frac{7}{9}$	$\frac{4}{5}$	$\frac{9}{11}$	$\frac{5}{6}$	…

Теперь проинтерпретируем наши результаты. Леопольд Кронекер (1823–1891) не зря говорил, что целые числа придумал Бог, а остальное — люди. В нашем случае для модели P с помощью целых значений *m* любые совокупности сообществ, компаний, стран, регионов можно аналитически (условно) разделить на линейные (m≈1), квадратичные (m≈2), кубические (m≈3), «тетричные» (m≈4), «пентальные» (m≈5) и т.д.

Таким образом, Норвегию, у которой G'=0,25 и m<1, по распределению доходов можно отнести к сублинейным странам, Францию при G'=0,327 и *m*≈1 — к линейным, Мексику при G'=0,482 и m≈2 — к квадратичным; Россия с G'=0,423 попадает между Францией и Мексикой (1<*m*<2), существенно отставая, например, от Гаити (G'=0.538, 2<*m*<3), Сьерра-Леоне (G'=0,629, 3<*m*<4) и Намибии (5<*m*<6, по разным данным 0,707< G'<0,750).

В этой градуировке Москва может претендовать на кубический тип распределения доходов, поскольку, по данным официальной статистики, G' доходил до 0,62, а по мнению многих специалистов, реальные значения G' в Москве находятся в интервале 0,60–0,70.

Аналогичные расчеты по ведущим российским компаниям, проведенные по данным годовых отчетов, публикуемым газетой «Ведомости»[2], указывают на величину 2<m<3 в 2009 г. При величине ежемесячной зарплаты топ-менеджера в $1,5–3,0 тыс. внутрикорпоративная *m* может указывать на линейность в распределении доходов, однако с учетом бонусов порядка в $1–3 млн. в год *m* может достигать и 4.

Здесь требуется обсуждение. Можно увязать обсуждаемую модель с традиционными статистическими распределениями. Для распределения Парето с таким же G^, как в нашей модели P, лишь степень m<1 обеспечивает конечную дисперсию, а для лог-логистического распределения для того же требуется m<2.

Что касается лог-нормального распределения, то дисперсия не стремится к бесконечности, а лишь медленно растет при росте m. Заметим, что распределения Парето и лог-логистическое хорошо описывают правый (верхний) хвост распределения доходов, а лог-нормальное хорошо описывает не слишком большие доходы, но плохо описывает правый хвост.

Интерпретация может быть следующей: при высокой степени неравенства в модели P малая (богатая) часть общества стремится увеличить свои доходы, так что верхние хвосты распределения утяжеляются и дисперсия стремится к бесконечности. В тоже время средняя по доходам часть общества медленно реагирует на рост степени модели m.

[1] При m=1 выражении для индекса Джини точное, при всех n = 1, 2, 3….

[2] См.: Милек О. Изучение распределения дохода с помощью распределения с тяжелыми хвостами: магистерская диссертация / ГУ–ВШЭ. М., 2010.