

АКАДЕМИЯ НАУК СССР  
ИНСТИТУТ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

# **Анализ нечисловой информации в социологических исследованиях**

Ответственные редакторы  
кандидат философских наук  
В. Г. АНДРЕЕНКОВ,  
кандидат физико-математических наук  
А. И. ОРЛОВ,  
кандидат экономических наук  
Ю. Н. ТОЛСТОВА

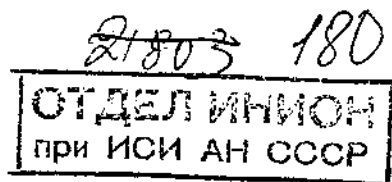


Москва  
«НАУКА»  
1985

Монография охватывает широкий круг математических методов, разработанных в различных научных центрах страны и предназначенных для анализа нечисловых данных. Статьи сборника носят аналитический характер и представляют собой описание различных классов методов, являющихся перспективными с точки зрения их применения для анализа социологических данных.

Рецензенты:

Г. Б. БОКИЙ, Ю. Н. ГАВРИЛЕЦ,  
О. В. СТАРОВЕРОВ, Л. Л. РЫБАКОВСКИЙ



## ПРЕДИСЛОВИЕ

В решениях XXVI съезда КПСС, последующих Пленумов ЦК КПСС поставлены задачи по дальнейшему повышению эффективности и качества социологических исследований и внедрению их результатов в практику социально-экономического управления и прогнозирования. А это требует овладения всем арсеналом средств, предоставляемых современной математической наукой для обработки и анализа социологической информации.

Что же нужно для достижения поставленной цели? Не будем касаться организационных проблем — статуса социологических подразделений, обеспеченности техникой, обучения специалистов и вспомогательного персонала. Обратимся к процессу использования математического метода в социологических исследованиях. Этот процесс, с нашей точки зрения, в общем виде состоит из следующих этапов:

1. Постановка конкретной содержательной задачи, ее обобщение, формализация.
2. Теоретическая разработка метода решения задачи.
3. Разработка конкретных алгоритмов, реализующих метод (с учетом специфики больших массивов, пропущенных данных и т. д.).
4. Разработка программ для ЭВМ.
5. Внедрение метода в практику.
6. Разработка рекомендаций относительно того, при решении каких конкретных задач имеет смысл применять метод.

Если придерживаться принятой в настоящее время классификации научной тематики, то можно сказать, что сборник посвящен современным методам математической статистики, ориентированным на применение в социологических исследованиях.

В название сборника вынесены слова «Нечисловая информация». Что это такое? Результаты социологических исследований зачастую связаны с обработкой данных на ЭВМ, а ЭВМ, как известно, имеет дело с числами. Нет ли здесь противоречия? Дело в том, что социологические данные зачастую существенно отличаются от результатов физических измерений, таких, как измерения длины, массы и т. д. Привычные нам числа вошли в человеческий обиход на заре истории для того, чтобы выражать результаты физических измерений. Массы можно складывать, длины — перемножать. Соответствующие операции были определены и для чисел.

После того как подобные, вытекающие из свойств реального мира, свойства чисел были формализованы, посвященная числам ветвь математики получила самостоятельное развитие. Были сформулированы известные математические закономерности, уже не являющиеся непосредственным слепком с какой-то конкретной реальности. И человечество в течение многих веков с успехом пользовалось (и в науке, и в обыденной жизни) этими закономерностями. С их помощью открывались новые свойства окружающей человека действительности.

Так было до тех пор, пока числами не стали пользоваться в общественных науках. Будем говорить о социологии. Эта наука имеет дело, скажем, с оценками привлекательности профессий. Если респондент оценил одну профессию баллом 5, а другую — 10, то не имеет смысла говорить, что одна из них нравится ему ровно в два раза больше, чем другая. Более того, оперирование с этими баллами как с действительными числами может привести к неправильным выводам.

Встает вопрос: а можем ли мы вообще пользоваться какими бы то ни было известными математическими свойствами таких «чисел»? Подобные вопросы дали толчок к развитию новой науки — теории измерений. Говоря языком этой теории, данные в естественных науках обычно бывают получены по интервальным шкалам, либо шкалам отношений, а социологические данные — по номинальным и порядковым шкалам. Сформулированный вопрос соответствует известной из теории измерений проблеме адекватности.

Нечисловые данные в социологических исследованиях могут иметь и другой вид — вместо привычных для естествоиспытателя чисел исходной информацией социолог может использовать бинарные отношения, вершины графа и т. д.

Для обработки нечисловых данных не годятся классические статистические методы. Нужны новые модели и алгоритмы. И они стали разрабатываться. В СССР за последние 10 лет сформировалась новая область математической статистики — статистика объектов нечисловой природы. Почти синонимом является также широко используемый термин «анализ нечисловой информации».

Как почти все новое, рождалась эта область в противоречиях. Например, довольно много относящихся к ней работ публиковалось в сборниках, в названиях которых имелись слова «экспертные оценки», хотя содержащиеся в этих работах результаты могли применяться в самых разных областях, в том числе (но не исключительно!) и в столь специфичной области, как экспертные оценки. Причина этого противоречия проста: неформальный коллектив авторов был организован для разработки математической теории экспертных оценок, но пошел по другому пути.

Соответствующие результаты можно найти также в книгах, посвященных прикладной статистике, анализу данных, анализу качественных признаков, квалиметрии (анализу качества продукции),

математическим методам в социологии, экономике психологии и т. д.<sup>1</sup>

Термин, используемый нами как название направления, нельзя признать окончательным. Но он используется в названиях книг и статей, конференций и конкурсов лекций.

Статистика объектов нечисловой природы молода. Почти все имеющиеся публикации содержат лишь оригинальные результаты. Многие из этих результатов опубликованы в работах, мало доступных широкому кругу заинтересованных читателей.

Настоящая книга — первая попытка дать широкому читателю представление о теории анализа нечисловой информации и о перспективах ее применения в социологических исследованиях. Она содержит результаты разработок, ведущихся в ряде научных центров страны. Большинство статей представляет собой обзоры отдельных направлений нечисловой статистики, написанные ведущими специалистами в этих направлениях. В книге отражена деятельность нескольких имеющихся в нашей стране «школ», соответствующих различным подходам к анализу данных. Это прежде всего новосибирская школа, которая, в свою очередь, представлена двумя направлениями. Первое направление основано на особом матричном представлении исходных данных (статья Б. Г. Миркина), второе — на поиске логических закономерностей (статья Г. С. Лбова и Т. М. Пестуновой).

В третьем разделе сборника содержатся две статьи, посвященные многомерному шкалированию. Это сравнительно новый метод анализа данных, получивший распространение с 70-х годов. Здесь можно выделить две школы — московскую и ленинградскую. Представители первой (В. С. Каменский, М. С. Косолапов, Г. А. Сатаров, А. Ю. Терехина, Е. Л. Шрайбер и др.) развили ряд методов, предложенных западными специалистами, обратив основное внимание на вопросы, связанные с методикой практического использования методов многомерного шкалирования. Прикладная направленность московской школы нашла свое отражение в статье Г. А. Сатарова.

<sup>1</sup> Примерами появившихся в последние годы сборников, содержащих соответствующие результаты, служат: Алгоритмизация анализа данных социально-экономических исследований. Новосибирск, 1982; Анализ нечисловых данных в системных исследованиях. М., 1982; Математические методы в социологическом исследовании. М., 1981; Методы анализа многомерной экономической информации. Новосибирск, 1981; Прикладная статистика. М., 1982; Экспертные оценки в задачах управления. М., 1982. См. также книги: Айвазян С. А., Мешалкин Л. Д., Енюков И. С. Прикладная статистика. М., 1983; Гаврилец Ю. Н. Целевые функции социально-экономического планирования. М., 1983; Лбов Г. С. Методы обработки разнотипных экспериментальных данных. Новосибирск, 1981; Миркин Б. Г. Анализ качественных признаков и структур. М., 1980; Типология и классификация в социологических исследованиях. М., 1981; Тюрин Ю. Н., Литвак В. Г., Орлов А. И., Сатаров Г. А., Шмерлинг Д. С. Анализ нечисловой информации. М., 1981; Хованов Н. В. Математические основы теории шкал измерения качества. Л., 1982; Чесноков С. В. Детерминационный анализ социально-экономических данных. М., 1982.

Ленинградская школа развивается на основе оригинального подхода, предложенного несколько лет назад В. Т. Перекрестом. С ней также можно познакомиться в этой книге.

Во втором разделе отражена существующая в практике тенденция в области внедрения математических методов анализа социологических данных, состоящая в разработке нескольких методов, предназначенных для решения одной и той же содержательной задачи. Разработка таких «дублей» обусловлена сложностью интересующих социолога закономерностей, трудностями с формализацией соответствующих положений. Раздел посвящен наиболее актуальным для социологии методам — методам группировок, которых имеется огромное количество (строго говоря, мощности континуума)<sup>2</sup>. В настоящее время нет обоснованного ответа на вопрос о том, в каких конкретных случаях имеет смысл пользоваться тем или иным методом. Соответствующие методические вопросы только начинают разрабатываться.

В книге не нашли отражения достижения советской науки в области разработки социометрических методов, теории нечетких множеств, динамических моделей, непараметрической статистики и некоторых других представляющихся перспективными для социологии методов анализа нечисловой информации. Публикация их — дело будущего.

Книга подготовлена отделом методического и информационного обеспечения Института социологических исследований АН СССР и подкомиссией «Статистика объектов нечисловой природы», Научного Совета АН СССР по комплексной проблеме «Кибернетика». Сборник рассчитан на лиц, занимающихся математическим анализом социологических данных. Большинство статей сборника предполагает наличие у читателя профессиональной математической подготовки. Однако каждая статья снабжена рассчитанной на неспециалиста в области математики преамбулой, в которой раскрываются смысл предлагаемых в статье методов и их возможная роль в повышении эффективности социологических исследований.

Авторский коллектив надеется, что сборник будет полезен всем, кто заинтересован в развитии и применении новых методов анализа социологической информации.

В. Г. Андреевков,  
А. И. Орлов,  
Ю. Н. Толстова

<sup>2</sup> Интересная классификация методов классификации недавно опубликована в работе: Бухштабер В. М., Маслов В. К., Зеленюк Е. А. Методы анализа и построения алгоритмов автоматической классификации на основе математических моделей. — В кн.: Прикладная статистика. М., 1983, с. 126 — 144.

# Общие вопросы анализа нечисловой информации

В. Г. АНДРЕЕНКОВ, Ю. Н. ТОЛСТОВА

## ОСОБЕННОСТИ ПРИМЕНЕНИЯ МАТЕМАТИЧЕСКИХ МЕТОДОВ В СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

В статье предпринимается попытка проанализировать процесс применения математических методов в социологических исследованиях с целью выявления таких особенностей этого процесса, которые отличают его от аналогичных процессов в естественных и некоторых других гуманитарных науках. Анализ осуществляется таким образом, чтобы соответствующие положения могли внести некоторый элемент конструктивности в решение проблемы выбора математического метода применительно к конкретным социологическим задачам. Авторы выделяют этапы названного процесса и указывают принципы соотнесения возможных путей реализации каждого выделенного этапа с содержательным характером решаемых задач, характером исходных данных, априорным знанием исследователя об изучаемом объекте. Практическое использование этих принципов в процессе решения конкретных социологических задач позволяет выбирать математические методы для наиболее эффективного решения. Авторы не претендуют ни на исчерпывающее решение вопроса об особенностях математического анализа социологических данных (эти особенности требуют дальнейшего изучения), ни на полноту охвата математических методов, представляющихся полезными для социологии. В основном имеются в виду методы, затрагиваемые в настоящей книге.

Прежде чем говорить о специфике применения математики в социологии, необходимо сформулировать ряд положений, характеризующих основные этапы процесса применения математического метода для решения любой социологической задачи (эти положения, вообще говоря, не несут ничего нового с точки зрения рассмотрения общих принципов применения математики в любой сфере знаний), однако применительно к социологии такие принципы не были до сих пор сформулированы достаточно четко, чтобы ими можно было практически пользоваться при выборе математического метода).

Вот эти этапы [4].

1. Выделение некоторого фрагмента изучаемой реальности, так называемой эмпирической системы с отношениями (ЭСО) [23]. Процесс выделения ЭСО состоит в определении изучаемой совокупности объектов, абстрагировании от их многочисленных свойств и указании интересующих нас отношений между ними.

2. Осуществление отображения элементов ЭСО вместе с упомянутыми отношениями в совокупность некоторых математических конструкторов (например, чисел) с соответствующими отноше-

ниями между ними. Такую систему можно назвать математической системой с отношениями (МСО).

Второй этап, в процессе реализации которого с каждым эмпирическим объектом сопоставляется некоторый математический конструкт, будем называть процессом измерения.

Традиционно понятие измерения связывается с числовыми системами не только в естественных, но и в гуманитарных науках. Например, измерение в социальном исследовании понимается как «способ упорядочения социальной информации, при котором системы чисел и отношений между ними ставятся в соответствие ряду измеряемых социальных фактов» [33, с. 202]. Однако известен ряд работ, в которых высказывается мысль о целесообразности связывания понятия измерения не только с числами. Мы подчеркиваем это обстоятельство, поскольку оно представляется нам принципиальным с точки зрения рассмотрения специфики применения математики в социологии. Именно в социологии очень часто адекватными исходным данным являются нечисловые МСО. Если используемая МСО — числовая, то будем употреблять термин «шкалирование». Осуществляя измерение, мы ставим своей целью сопоставить с каждым объектом некий математический конструкт, а с каждым эмпирическим отношением — некоторое математическое отношение таким образом, чтобы эмпирические отношения между какими-либо объектами «переходили» в соответствующие им математические. Ясно, что такое отображение есть моделирование структуры исходных данных, если под такой структурой понимать совокупность эмпирических объектов вместе с интересующими исследователя отношениями между ними. В дальнейшем мы будем называть структурой эмпирических данных (СЭД) ту часть ЭСО, которая непосредственно отображается в МСО (т. е. выборку вместе с рассматриваемыми отношениями между ее элементами). Математической же моделью структуры эмпирических данных (ММСЭД) — ту часть используемой МСО, которая является отображением СЭД.

3. Осуществление непосредственного математического анализа ММСЭД. Этот этап соответствует как бы второму (после измерения) уровню моделирования — моделированию «развития» изучаемой ЭСО. Термин «развитие» понимается здесь в значительной мере условно. Под моделированием «развития» мы понимаем установление закономерностей между элементами построенной ММСЭД. Говорить о «развитии» представляется возможным потому, что целью нахождения каждой такой закономерности, как правило, является решение определенной содержательной задачи, связанной с некоторым прогнозом.

Третий этап тоже в определенном смысле можно рассматривать как моделирование некоторых отношений между элементами этой системы, и в силу этого можно говорить об определенном сходстве между двумя рассматриваемыми уровнями моделирования.

4. Интерпретация результатов применения математического метода.

В литературе существует ряд работ, отражающих анализ специфики применения математики для решения социологических задач, причин, обуславливающих соответствующие трудности. В этих работах предмет рассматривается с разных точек зрения. Однако при этом, как правило, не выделяется единого «основания» для классификации анализируемых особенностей применения математических методов [2; 17; 8; 9, с. 9; 16]. На наш взгляд, это мешает конструктивному использованию анализа таких особенностей для выбора математического метода, адекватного решаемой социологической задаче. Более того, представляется, что соответствующие рекомендации могут быть разработаны только в том случае, если в качестве упомянутого «основания» будут использованы выделенные выше этапы применения математического метода. Обосновать такое положение можно следующим образом: особенности математических конструктов, являющихся результатами измерения (II этап), как и особенности строящихся математических моделей «развития» рассматриваемой социальной системы (III этап), своим источником имеют специфику изучаемой реальности, отражают ее. При этом на III этап такая специфика может влиять не только непосредственно, но и «через» II этап. И только четкое выделение первопричин, обуславливающих специфику измерения и моделирования «развития» системы, может помочь решить вопрос о выборе и адаптации математического метода.

Перейдем к рассмотрению некоторых особенностей социальных явлений, изучаемых с помощью математических методов — вернее, особенностей соответствующих ЭСО — с точки зрения того, какую роль эти особенности играют в процессе выбора ММСЭД и построения моделей ее «развития». Говоря точнее, рассмотрим некоторые особенности процесса применения математических методов для решения социологических задач, указывая при этом, какими свойствами изучаемых ЭСО эти особенности обуславливаются. Здесь мы ограничиваемся рассмотрением только тех особенностей применения математических методов в социологических исследованиях, которые так или иначе затрагиваются в книге.

Прежде чем говорить об указанных особенностях, заметим, что анализ причин частой некорректности применения традиционных математических методов в социологии приводит к необходимости рассмотрения понятия методик такого применения. Это понятие включает в себя рекомендации, в соответствии с которыми на основе анализа содержательного характера решаемой задачи следует осуществить выбор исходных данных. В соответствии с выделенными первыми двумя этапами применения математического метода термин «социологические данные» может пониматься двояко: либо как часть реальности (ЭСО), либо как совокупность результатов измерения (ММСЭД) и различных формальных компонент используемого алгоритма, сочетание метода с другими способами математического анализа исходных данных, интерпретация результатов применения метода и т. д. Принципы разработки

таких методик по существу являются отражением особенностей применения математики в социологии [29].

Основной причиной, вызывающей наличие всех интересующих нас особенностей изучаемых социологами ЭСО, служит то, что источником социологической информации, чаще всего является человек. Как пишет А. Г. Аганбегян [2, с. 5], все трудности математического моделирования социальных явлений обусловлены тем, что «главной особенностью социальных процессов является то, что познаваемый объект является здесь субъектом — носителем».

Нам представляется, что вызванными этой «главной особенностью» свойствами социальных явлений, обуславливающими трудности применения математики в социологии, являются, во-первых, совокупность свойств, вызывающих нечисловой характер используемых ММСЭД; во-вторых, вероятностные свойства упомянутых явлений; в-третьих, сложность интересующих социолога закономерностей, слабая изученность социальных явлений. Все перечисленное может вызвать и специфичный вид ММСЭД, и особенности процесса их анализа. Именно указанные свойства и рассматриваются. Параллельно обсудим также некоторые общие принципы построения и анализа ММСЭД в социологии.

Прежде всего представляется целесообразным остановиться на нашем понимании термина «нечисловой» в сравнении с известным термином «качественный», который используется для характеристики исходных социологических данных многими авторами. При этом признаки, значения которых используются в качестве исходной (для применения метода) информации, подразделяются на «качественные» и «количественные» в соответствии с тем, соответствует ли им номинальная либо порядковая шкала («качественные» признаки) или же интервальная шкала либо шкала более высокого типа («количественные» признаки). Подобный подход представляется нам неудачным с точки зрения используемых терминов и не отражающим в достаточной мере специфику социологических данных с точки зрения рассмотрения существа вопроса. Поясним это более подробно.

Использование указанных терминов представляется нам неудачным потому, что за различиями в значении «количественного» признака стоят определенные качественные различия объектов и, напротив, разница в значении «качественного» признака обычно отражает какое-то количественное отличие объектов друг от друга.

Рассматривая сущность вопроса, прежде всего отметим, что в силу высказанных соображений о «взаимодействии» причин, обуславливающих трудности применения математики в социологии, представляется нецелесообразным, говоря об этих причинах, в первую очередь выделять соответствующие свойства используемых в качестве ММСЭД совокупностей значений задействованных в исследовании признаков. Характер таких совокупностей является вторичным по отношению к тем связям между изучаемыми объектами, которые интересуют исследователя. Характер последних (точнее, наши гипотезы о нем) обуславливают и выбор используе-

мого признака, и способ получения его значений, и то, по какой шкале мы считаем эти значения полученными. Специфичность исходной для применения математического метода информации является следствием специфичности первичных эмпирических данных, которую и нужно изучать в первую очередь и о которой прежде всего необходимо говорить, анализируя причины пригодности или непригодности того или иного метода. Именно специфичность встречающихся в социологии ЭСО и является первопричиной сложности применения математики в социологических исследованиях.

Второе, что хотелось бы отметить, состоит в следующем. Конечно, «качественные» признаки в социологии широко распространены. Объясняется это тем, что очень часто единственными отношениями в изучаемой ЭСО, которые исследователь с достаточной степенью надежности может «измерять» (т. е. устанавливать, имеет ли место рассматриваемое отношение для той или иной конкретной совокупности объектов), являются отношения равенства (неравенства) и порядка. Но, несмотря на такую распространенность «качественных» признаков, специфичность интересующих социолога ММСЭД (по сравнению с теми моделями, с которыми имеет дело ученый-естествоиспытатель) может проявляться не только в том, что в качестве ее элементов выступают значения какого-либо признака, полученные по порядковой или номинальной шкале, но и в каком-либо виде.

Так, несмотря на трудности с построением интервальной шкалы в социологии (что обуславливается «бедностью» соответствующих ЭСО; интервальная шкала, как известно, требует упорядочения интервалов между объектами), иногда она все же используется социологами. Но нельзя считать, что числа, полученные по интервальной шкале (которую обычно относят к «количественным» шкалам), — это такие же числа, с которыми имеет дело, например, физик, измерив вес или объем каких-то тел. Анализируя такие числа, мы не можем пользоваться, например, коэффициентом вариации, имеющем смысл для веса или объема, значения которых получаются по шкале отношений<sup>1</sup>. Причина в том, что соответствующая ЭСО недостаточно «богата», т. е. содержит недостаточное количество отношений между ее элементами. Для этих элементов, интервалов между ними определено отношение порядка, но, скажем, не разности или отношения между элементами. В ряде случаев мы можем установить отношение порядка не между всеми интересующими нас объектами, а только между некоторыми из них. А значит, не можем получить даже порядковую шкалу. Здесь эмпирической информации будет соответствовать ММСЭД, яв-

<sup>1</sup> Коэффициент вариации не инвариантен относительно всех линейных преобразований, являющихся допустимыми преобразованиями интервальных шкал. То же можно сказать о простейших арифметических выражениях, связывающих два (или больше) коэффициента вариации (их сумму, отношения и т. д.).

ляющаяся частично упорядоченным множеством<sup>2</sup>. Для некоторых социологических задач адекватным характеру эмпирической информации является отображение ее в такие математические конструкции, в которых вообще не имеют смысла никакие отношения, «похожие» на отношение порядка: в графы, матрицы и т. д. [6, 19, 1, 28].

Другие виды ММСЭД интересующего нас характера (бинарные отношения, нечеткие и случайные множества и т. д.) рассматриваются в работе [31]. Авторы останавливаются на указанных видах нечисловой информации потому, что для них, в отличие от многих других видов подобной информации, существуют «содержательные и специфические математические средства» анализа.

О том, что специфичными для социальных исследований являются данные более широкого круга, чем совокупности значений номинальных и порядковых признаков, говорится также в [15, 30]. Их авторы предлагают расширить понятие «качественный» (так, в последней названной работе выделяется 6 классов качественных свойств: номинальные, ранговые, порядковые, кусочно-порядковые, выделенные, структурные). Можно сослаться также на учебник [37], где речь идет еще и об определенном виде диаграммах, выступающих в качестве ММСЭД, и о некоторых имитационных ММСЭД.

Потребность в использовании такого рода ММСЭД возникает потому, что множество действительных чисел не является адекватным отражением интересующих социолога ЭСО. Значит, целесообразно считать основной особенностью социологических данных, ограничивающей применение традиционного математического аппарата, их нечисловой характер. При этом первичным является нечисловой характер ЭСО, означающий, что множество действительных чисел не может служить ее моделью из-за специальных свойств отношений между ее элементами. Следствием указанной специфики ЭСО является использование нечисловых ММСЭД, не являющихся совокупностью действительных чисел.

Вероятностный характер определенной части изучаемых социологией явлений в настоящее время не вызывает сомнений у подавляющей части исследователей. Однако еще недавно в литературе велись дискуссии по поводу того, правомерно ли использование теории вероятностей и математической статистики в социальных науках и если в принципе правомерно, то каким образом оно может быть осуществлено [17].

Вероятностный характер получаемой от респондентов информации может быть объяснен двумя причинами. Во-первых, случайной величиной является указываемое каждым респондентом его мнение относительно интересующего социолога вопроса. От-

<sup>2</sup> Социологические данные, адекватным представлением которых являются частично упорядоченные множества, встречаются довольно часто [25, с. 10; 36]. Систематизацию используемых в социально-психологических исследованиях частично упорядоченных множеств и соответствующие содержательные примеры можно найти в работе [38].

клонение высказываемого им суждения от «истинного» может быть вызвано влиянием настроения респондента, его неумением разобраться в собственном отношении к каким-либо явлениям, непониманием заданного исследователем вопроса, недоверием к личности интервьюера, конкретной ситуацией, в которой находится респондент. Не меньшее количество факторов может повлечь и неправильную оценку ответа исследователем. В таких случаях в качестве «истинного» суждения можно принять математическое ожидание соответствующего распределения, получающегося при многократном опросе респондента<sup>3</sup>. Подобный подход, сопряженный с вероятностными моделями, часто используется при решении тех или иных задач с помощью математических методов и, по существу, отражает часть того общего, что имеется в социологическом и психологическом аспектах социальной реальности.

Как известно, эти аспекты не тождественны друг другу [3]. И это в интересующем нас отношении выражается в наличии второго фактора, вызывающего вероятностный характер социальных процессов. Массовость социальных явлений дает возможность интерпретировать изучаемые качественные различия между объектами как различия значений некоторой случайной величины (некоторого признака). И если в первом случае эта величина соответствовала каждому респонденту, то при рассматриваемом подходе она соответствует всей изучаемой совокупности респондентов и каждая ее реализация — это значение какого-либо признака для отдельного респондента. Такой подход широко используется в социологических исследованиях.

Именно этот подход мы фактически реализуем при применении большинства традиционных математико-статистических методов, начиная с расчета доверительных интервалов математического ожидания значений рассматриваемого признака для изучаемой генеральной совокупности (на основе расчета соответствующего среднего арифметического) и кончая традиционными способами статистического анализа: регрессионного, факторного и т. д. Именно в связи с широкой распространенностью такого подхода на практике представляется целесообразным упомянуть о необходимости проверки условия, обеспечивающего принципиальную возможность его использования, проверки самого существования, осмысленности для совокупности изучаемых объектов рассматриваемой случайной величины. Другими словами, изучаемая совокупность должна быть в определенном смысле однородной [18, 21].

Возможны ситуации, когда оба названных подхода фактически сливаются воедино. Так, при достаточной однородности изучаемой совокупности респондентов можно полагать, что распределе-

<sup>3</sup> Такое предположение, однако, является проблематичным, если причиной отклонения результата наблюдения от истины служит сознательное искажение респондентом своего действительного мнения. Кроме того, отметим, что существуют распределения, для которых не имеет смысла понятие математического ожидания. Соответствующий момент распределения просто не существует [12].

ние, соответствующее многократной оценке какого-либо объекта одним респондентом, совпадает с распределением, соответствующим однократной оценке того же объекта всеми респондентами рассматриваемой совокупности<sup>4</sup>.

Как отмечается в статье А. И. Орлова, привлечение теории вероятностей для анализа любых статистических данных всегда желательно, поскольку позволяет более глубоко обосновать используемый математический метод, изучив его возможности, сравнить их с возможностями других методов.

Обращение к теории вероятностей в социологии иногда возникает и в ситуациях, не связанных с предположением о том, что ответ каждого респондента — это реализация некоторой случайной величины (см., например, статью С. А. Петровского).

Естественно, что учет вероятностного характера социальных явлений, желание привлечь для их анализа теорию вероятностей заставляет прибегать к применению ММСЭД специального вида. Как известно, в основе математических моделей, используемых в теории вероятностей и математической статистике, лежат три понятия: пространство  $\Omega$  элементарных событий, класс подмножеств  $\Omega$  (событий) и определенная на этих подмножествах функция  $P$  — распределение вероятностей<sup>5</sup>. Возможность использования при решении социологических задач теории вероятностей возникает тогда, когда в качестве элементов ЭСО выступают пары  $\langle$  событие; степень возможности его появления  $\rangle$ , а элементами соответствующей МСО являются пары  $\langle a, p \rangle$ , где  $a$  — математический конструкт, соответствующий событию (например, область признакового пространства, соответствующая некоторым интервалам изменения рассматриваемых признаков), а  $p$  — отвечающая ему вероятность. Иногда, однако, МСО, включающие вероятности, могут иметь несколько иной вид, т. е., помимо пар  $\langle a, p \rangle$ , содержать и другие математические объекты или состоять только из вероятностей (так как случайным событиям могут не приписываться математические конструкты, как это делается в статье С. А. Петровского). Подобное понимание ММСЭД в определенной мере можно считать аналогом известного подхода к пониманию измерения, в рамках которого строятся так называемые статистические измерительные системы, т. е. системы, регистрирующие число элементов в различных точках фазового пространства и вычисляющие параметры распределения вероятностей случайных процессов [33, с. 170—198].

Подчеркнем, что специфичным для социологических данных является то, что в рассмотренных выше парах  $\langle a, p \rangle$  в качестве  $a$  выступают объекты нечисловой природы. Это вызывает большие

<sup>4</sup> Такое предположение делает Л. Л. Терстоун при использовании метода парных сравнений.

<sup>5</sup> Как известно, в математике вероятностью называется числовая величина, удовлетворяющая определенной совокупности аксиом, в виду которых формулируются свойства вероятности, необходимые на данном этапе развития науки [22].

трудности в использовании для анализа подобных пар теории вероятностей, поскольку последняя разработана в основном для тех случаев, когда  $a$  — некоторый набор чисел (значений рассматриваемых случайных величин). Подробнее об этом пойдет речь ниже.

Перейдем к рассмотрению использующихся в социологии способов анализа, описанных ММСЭД. Прежде всего отметим, что недостаточная изученность социальных явлений приводит к тому, что одной и той же СЭД могут соответствовать несколько ММСЭД, и использование каждой такой модели может дать свои возможности изучения свойственных реальным явлениям закономерностей. Приведем пример В. И. Панютто, когда представляется возможность моделирования одних и тех же отношений в малой группе с помощью социоматрицы и с помощью графа. На основе социоматрицы строятся различные индексы, характеризующие те или иные особенности изучаемой группы. Однако общим недостатком всех этих индексов является то, что они учитывают число отданных и полученных выборов (одних членов группы другими — той же группы), не учитывая свойства структуры взаимоотношений в группе как целого. Избавиться от этого недостатка позволяет использование для формирования социометрических индексов теории графов.

Можно выделить две группы (отнюдь не равновеликие) математических методов, использующихся в социологии. Большая часть их является, по существу, методами идентификации параметров некоторой модели изучаемого явления. Выбор такой модели и означает выбор метода. Так, применяя линейный регрессионный анализ, мы идентифицируем параметры  $a_1, \dots, a_n$  модели  $y = \sum_{i=1}^n a_i x_i$ .

То же, хотя и не всегда в столь очевидном виде, происходит и при использовании других методов. Например, для кластерного анализа в качестве модели изучаемого явления выступает набор предположений о характере распределения соответствующих изучаемым объектам точек в исходном признаковом пространстве, о виде используемой меры близости и т. д. Идентифицируемыми же параметрами служат данные о принадлежности каждой конкретной точки к тому или иному таксону.

В порядке альтернативы описанному подходу можно указать на прямое использование (доказательство и содержательную интерпретацию) тех теорем, которые выполняются для МСО. Заметим, однако, что методы этой группы очень редко применяются на практике для получения более или менее серьезных выводов, хотя, конечно, в процессе применения любого метода мы часто в неявном виде пользуемся известными математическими теоремами (так, целая серия теорем о действительных числах бывает задействована, когда, скажем, вычисляется среднее арифметическое или дисперсия). Примером реализации рассматриваемого подхода может служить использование теорем теории графов для

получения выводов о взаимоотношениях в малых группах [20] <sup>6</sup>.

Еще одним направлением, в рамках которого могли бы использоваться методы второй группы (но пока, к сожалению, не используются), можно назвать применение правил формальной логики [35].

Ниже мы будем рассматривать лишь первую группу методов. Используя их, социолог сталкивается со множеством трудностей, из которых, на наш взгляд, можно выделить следующие три, обусловленные сложностью, слабой изученностью социальных явлений.

Ниже пойдет речь о выборе соответствующей используемому методу математической модели изучаемого явления. В связи с этим представляется уместным упомянуть об одной существенной, но не затрагиваемой в настоящей статье проблеме: выбор алгоритма, оптимального с точки зрения затрат на его реализацию (в том числе, затрат на обработку данных с помощью ЭВМ).

Как сообщил авторам А. И. Орлов, в его практике был такой случай. К одним и тем же данным были применены два алгоритма: один из алгоритмов, предложенных Б. Г. Миркиным (лежащий в русле подхода, описанного в помещенной в настоящем сборнике статье этого автора), и алгомеративный иерархический алгоритм, в котором расстояние между кластерами измерялось по принципу «ближайшего соседа». Оба алгоритма дали практически одинаковые результаты. Но первый из них пришлось полгода реализовывать на ЭВМ, а второй потребовал лишь час ручного счета. К сожалению, исследователи сначала обратились к первому алгоритму, что привело к непроизводительным затратам людских и материальных ресурсов.

Итак, о трудностях. Во-первых, выбор той модели, параметры которой должны идентифицироваться в процессе применения метода, является сложным процессом, включающим в себя соотнесение различных формальных сторон рассматриваемого метода с содержательным характером решаемой задачи [29, 27].

Во-вторых, исследователь очень редко в достаточной мере может априори судить о характере той модели, параметры которой должны идентифицироваться в процессе применения того или иного метода. Это обуславливает ряд сложностей и в первую очередь — необходимость параллельного использования разных методов для решения одной и той же задачи, разных способов анализа одной и той же ММСЭД, т. е. «опробования» разных моделей одного и того же явления. Только сопоставляя результаты, полученные с помощью

---

<sup>6</sup> Желая продемонстрировать, что рассматриваемый подход может быть полезным (а также то, что нечисловые ММСЭД эффективно могут быть использованы в социологии), отметим, что в названной работе приводится пример использования одной из теорем теории графов, относительно которой автор говорит, что она «замечательна тем, что демонстрирует возможность применения математики в социологии: из элементарных посылок делается далеко не очевидный вывод» [20, с. 89].

разных методов, исследователь может достаточно глубоко проанализировать реальность, составить представление об «истинной» модели изучаемого явления.

Отметим, что уже само наличие нескольких математических методов решения одной и той же социологической задачи отражает сложность процесса внедрения таких методов в социологию. Ту или иную последующую закономерность оказывается возможным понимать по-разному и вследствие этого прибегать к разным способам формализации соответствующих представлений. О том же говорит и принципиальная неоднозначность решений, получаемых с помощью многих используемых в социологии методов (факторный анализ, многомерное шкалирование и т. д.)<sup>7</sup>.

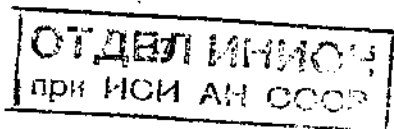
В-третьих, процесс анализа социологических данных настолько сложен, что часто возникает потребность последовательного применения ряда математических методов. В таких случаях ММСЭД, исходная для применения какого-либо метода, сама может быть получена с помощью математики. На практике подобный подход чаще всего применяется, когда мы, выбрав какой-либо математический метод решения интересующей нас социологической задачи, используем другие методы с целью либо получения исходных данных для выбранного метода, либо интерпретации полученных с его помощью результатов<sup>8</sup>. Естественно, что при реализации такого подхода возникает много методических вопросов, требующих решения. В конце статьи мы поясним это на примере рассмотрения того, как с помощью математических методов в социологических исследованиях может осуществляться процесс измерения. Здесь же предварительно заметим следующее.

Как известно, указанный процесс обычно весьма сложен. Те данные, которые исследователь получает непосредственно в результате сбора информации (например, в результате опроса респондентов), очень часто являются лишь отправным пунктом для получения той ММСЭД, которая отражает интересующую исследователя ЭСО. Эту искомую ММСЭД в таких случаях часто бывает возможным получить с помощью математики — применения различных методов поиска латентных переменных, многомерного шкалирования и т. д.<sup>9</sup> И тогда уже при измерении встает вопрос о выборе математической модели, соответствующей используемому способу перехода от непосредственно наблюдаемых данных к искомой ММСЭД. О примерах подобных моделей будет идти речь ниже.

<sup>7</sup> О том же свидетельствует и то, что, как было отмечено выше, одной СЭД может соответствовать несколько ММСЭД.

<sup>8</sup> В качестве примера такого последовательного использования математических методов, которое позволяет решать задачи несколько иного типа, можно привести предложенный в работе [7] способ последовательного применения методов классификации, позволяющий достаточно глубоко проанализировать структуру расположения объектов в исходном признаковом пространстве и построить на основе такого анализа типологию этих объектов.

<sup>9</sup> В таких случаях говорят о шкалировании в процессе анализа социологической информации и отличают его от шкалирования в процессе ее сбора [13].



Предлагаемые в статьях сборника методы анализа описанных выше специфичных для социологии ММСЭД мы будем рассматривать в рамках трех использующихся обычно в социологических исследованиях путей такого анализа, которые представляется целесообразным выделить.

Первый путь состоит в использовании математических свойств самой ММСЭД (точнее, соответствующей МСО) и только их. Этот подход активно используется при анализе ММСЭД, являющихся совокупностями значений признаков, полученных по разным шкалам. Это могут быть либо только порядковые и номинальные шкалы, либо еще и шкалы более высоких типов. В последнем случае часто говорят о разнотипных признаках. Приведем примеры способов анализа подобных ММСЭД.

Методы поиска закономерностей в пространстве признаков, полученных по разным шкалам, разрабатываются рядом авторов из ИМ СО АН СССР и НГУ, работающих под руководством Н. Г. Загоруйко и Г. С. Лбова. Общее понятие закономерности обсуждается в ряде работ Н. Г. Загоруйко и других авторов [10, 11]. Некоторая конкретная реализация соответствующих определений используется в статье Г. С. Лбова и Т. М. Пестуновой настоящей работы.

Авторы названной статьи для одного класса задач — задач группировки объектов — обобщают некоторые результаты, лежащие в русле подхода к анализу разнотипных признаков, предложенного в [14]. Этот подход позволяет решать широкий круг часто встающих перед социологом задач дискриминантного, регрессионного, кластерного анализов, динамического прогнозирования, упорядочения объектов по их перспективности и т. д. Полученные с помощью такого подхода результаты естественным образом легко интерпретируются. Подчеркнем, что логика, заложенная в соответствующих алгоритмах, по существу, отвечает той логике, которую обычно использует социолог при решении соответствующих задач без привлечения какого бы то ни было формализма. И в этом смысле рассматриваемый подход представляется естественным, что можно считать основным его достоинством.

Естественность указанного подхода косвенным образом подтверждается и тем, что подобные подходы (являющиеся, по существу, частным случаем рассмотренного) разрабатываются и другими авторами. Среди них упомянем способы отыскания связей между признаками (точнее, между градациями, отдельными значениями признаков) и группировки объектов, предложенные в [34]. Опыт практического применения этих методов описывается в опубликованной в настоящей книге статье В. А. Малахова и других.

Иной подход к анализу данных, полученных по шкалам низких типов, разработан в ИЭ и ОПП СО АН СССР под руководством Б. Г. Миркина. Методы, предложенные им<sup>10</sup>, требуют другого

---

<sup>10</sup> См. статью Б. Г. Миркина в данном сборнике.

исходного представления социологических данных — матричного. Исходные данные представляются в виде матрицы связи между объектами либо признаками. Задачи конструирования новых признаков и описания одних признаков в терминах других (к которым Б. Г. Миркин сводит все многочисленные возникающие в социологии задачи анализа данных) формулируются как аппроксимационные задачи в пространстве рассматриваемых матриц связи.

На анализ матриц данных произвольной природы (это могут быть матрицы коэффициентов связи между признаками, частотные таблицы и т. д.) рассчитаны методы решения ряда часто встречающихся в социологии задач, предложенные П. С. Ростовцевым. Эти методы дают возможность отыскивать такие блоки исходной матрицы данных, элементы которых имеют «примерно» одинаковые значения, что позволяет, в частности, решать задачи поиска взаимосвязей между отдельными градациями рассматриваемых признаков.

Поскольку природа анализируемых с помощью таких методов матриц может быть различной, то диапазон задач, которые можно решать этими методами, довольно широк. Ясно, что их можно использовать для анализа качественной информации.

Оригинальный подход к анализу широкого диапазона социологических данных предлагается в статье Е. Н. Кузнецова и других. Оригинальными являются и предлагаемая ММСЭД, и способы ее анализа. При этом речь идет о решении задач структуризации социально-экономической информации. Авторы претендуют на создание достаточно универсального математического аппарата (теории монотонных систем), позволяющего снять ряд традиционных ограничений на процесс применения математических методов при решении таких задач.

При использовании этой теории предполагается, что совокупность изучаемых объектов — множество произвольной природы. В качестве ММСЭД выступает система его подмножеств вместе с некоторой функцией  $\lambda$ , заданной на парах вида (подмножество (произвольный его элемент)).

Обзор различных мер близости, пригодных для измерения степени «похожести» интересующих социолога объектов, дан в статье Г. В. Раушенбаха. Вычисление той или иной меры близости (между объектами, либо между признаками) является составной частью большинства применяемых в социологии математических методов, а потому статья носит общий характер с точки зрения круга задач, решаемых с помощью рассматриваемых в статье методов. Традиционные меры близости, задействованные в большинстве используемых в социологии алгоритмах, очень часто бывают не адекватными содержательному характеру социологических задач. В частности, на практике используется весьма мало мер близости, рассчитанных на данные, полученные по шкалам низких типов. Поэтому обзор, в котором предлагаются меры близости адекватные весьма широкому кругу интересующих социолога данных, представляется полезным.

Предполагается, что каждый данный объект описывается каким-то подмножеством некоторого множества (например, множества свойств рассматриваемых объектов). Каждое такое множество можно рассматривать как алгебраическую решетку<sup>11</sup> относительно некоторых операций (объединения и пересечения, максимума и минимума и т. д.). Таким образом, можно сказать, что автор полагает, что в качестве ММСЭД фигурирует некоторая решетка.

В статье не дается конкретных рекомендаций по выбору меры близости при решении тех или иных социологических задач. Однако ряд моментов, отраженных в статье, все же позволяет в какой-то мере ограничить круг пригодных для решения конкретных задач мер близости.

В этой связи, во-первых, упомянем единообразное представление всех мер близости через некоторые «оценки», что дает возможность более глубоко проанализировать сходство и различие разных мер близости. Во-вторых, в тех случаях, когда требуется оценить близость между объектами, заданными разнородными признаками, автор рекомендует измерять ее в виде обобщенного среднего соответствующих частных близостей. В-третьих, в статье описываются две известные попытки ограничения чрезмерного многообразия предложенных в литературе мер близости. Первая — с точки зрения допустимости этих мер в смысле теории измерений<sup>12</sup>, вторая — возможности задания меры близости с помощью системы акспом.

Целый ряд проблем возникает при попытке использования рассматриваемого пути анализа ММСЭД для изучения описанных выше ММСЭД, состоящих из пар вида  $\langle a, p \rangle$ , позволяющих учитывать вероятностный характер интересующих социолога процессов. Как отмечалось выше, традиционные методы теории вероятностей и математической статистики разработаны для тех случаев, когда в качестве  $a$  фигурируют значения (или наборы значений) числовых случайных величин. Встает вопрос о распространении соответствующих понятий на нечисловые величины. Такое распространение и будет означать реализацию рассматриваемого пути анализа ММСЭД.

Работа в этом направлении, по существу, только начата. Соответствующие результаты, отражены в статье А. И. Орлова. В качестве элементов ММСЭД он фактически рассматривает пары  $\langle a, p \rangle$ , где  $a$  — объект пространства достаточно произвольной природы (результат измерения по шкале низкого типа, ранжировка, бинарное отношение и т. д.). Роль формулировки и доказательства подобных соотношений для повышения эффективности применения математических методов в социологии очевидна. Так, относительно доказанного Орловым аналога закона больших чисел

<sup>11</sup> В советской математической литературе бытуют два термина для обозначения этого понятия — «решетка» и «структура».

<sup>12</sup> О такой допустимости пойдет речь ниже при описании второго пути анализа ММСЭД.

для случая, когда в качестве  $a$  выступают произвольные бинарные отношения, можно сказать следующее. Если, к примеру, речь будет идти об усреднении классификаций рассматриваемой совокупности объектов, осуществляемых отдельными экспертами, то закон больших чисел дает нам уверенность в том, что, увеличивая количество экспертов, мы будем, выясняя их «среднее» мнение, стремиться к какому-то пределу.

Закон больших чисел обобщается им на тот случай, когда изучаемые объекты — элементы произвольного множества, а также когда объект, играющий роль «среднего арифметического», ищется среди конструкций иной природы, чем множество исходных объектов. Это дает возможность решения широкого класса встречающихся в социологии задач, включающих анализ двух разнородных совокупностей: например, респондентов и оцениваемых ими объектов (анализируемых, к примеру, с целью поиска объекта, наиболее «близкого» всем рассматриваемым респондентам).

Ряд допущений, сделанных А. И. Орловым в процессе доказательства, не всегда может быть оправдан. Так, упомянутый выше закон больших чисел для бинарных отношений доказывается в предположении, что эти отношения суть реализации одинаково распределенных случайных величин. Однако такое условие не всегда отвечает действительности. Например, если рассматриваемые бинарные отношения — осуществляемые отдельными экспертами классификации, то указанное предположение означает однородность совокупности экспертов в смысле их отношения к оцениваемым объектам.

Насущная потребность учета вероятностного характера социологических явлений проявилась в характере работы, ведущейся в области развития многомерного шкалирования в ИСЭП АН СССР и отраженной в статье В. Т. Перекреста. Прежде чем пояснить, что конкретно имеется в виду, напомним: методы многомерного шкалирования предназначены для нахождения координат изучаемых объектов в некотором отыскиваемом в процессе применения метода многомерном пространстве (предполагается, что его оси соответствуют признакам, значения которых получаются по крайней мере по интервальной шкале). Исходной информацией служит матрица близостей между этими объектами. Если элементы этой матрицы можно считать полученными по интервальной шкале, то методы шкалирования называются метрическими. Если же такая шкала — порядковая, то говорят о неметрическом шкалировании.

В основе алгоритмов многомерного шкалирования лежит требование, состоящее в том, чтобы структура евклидовых расстояний между изучаемыми объектами в искомом пространстве в определенном смысле соответствовала структуре исходной матрицы близостей.

В упомянутой статье рассматриваются так называемые функциональные модели метрического шкалирования, предполагающие, что множество исследуемых объектов является пространством

с двумя структурами: топологической, определяемой некоторой функцией близости, и статистической, определяемой некоторой вероятностной мерой, задающей распределение изучаемых объектов. В качестве исходной матрицы близостей выступает матрица значений функции близости. Таким образом, фактически используемая ММСЭД не совпадает с матрицей близостей, как это происходит в традиционных алгоритмах многомерного шкалирования: во-первых, традиционная матрица близостей заменяется определенной на парах точек рассматриваемого пространства функцией близости; во-вторых, кроме значений этой функции, в ММСЭД входит совокупность плотностей вероятности во всех подмножествах точек рассматриваемого пространства.

Цель решения задачи функционального шкалирования — нахождение вектор-функции, заданной во всех рассматриваемых точках, по разнице значений которой в каких-то двух точках можно судить о близости между этими точками. При построении соответствующего оптимизируемого функционала больший вес придается тем различиям между этими разностями и отвечающими им значениями функции близости, для которых плотности вероятностей в рассматриваемых точках больше. Таким образом, основное отличие рассматриваемого критерия сходства двух названных выше структур от традиционных критериев подобного рода состоит в том, что каждое подмножество элементов исходного пространства рассматривается как бы с учетом его «веса», определяемого заданной мерой распределения вероятностей. Такое привлечение анализа вероятностей к процессу многомерного шкалирования, очевидно, вполне согласуется с высказанными выше положениями о необходимости более широкого использования теории вероятностей для анализа социологической информации.

Как уже отмечалось, потребность в привлечении теории вероятностей возникает не только тогда, когда предметом математического анализа являются значения (многомерных) случайных величин, реализации которых характеризуют отдельные изучаемые объекты (отдельных респондентов). Пример вероятностных ММСЭД другого рода можно найти в статье С. А. Петровского. В ней рассматривается специфический класс задач — так называемый анализ на проблемных сетях. Предполагается, что имеется совокупность взаимосвязанных проблем (объектов, явлений, тенденций и т. д.), каждой из которых соответствует ряд возможных допустимых состояний. Задача состоит в осуществлении прогноза, т. е. сравнительной оценки возможности реализации в будущем различных состояний прогнозируемой проблемы (нахождении соответствующих вероятностей). Предполагается, что оценки ряда вероятностей дают эксперты. Автор убедительно показывает, что без привлечения формальных математических методов, на основе лишь экспертных оценок удовлетворительное решение поставленной задачи невозможно: прогноз оказывается либо слишком неопределенным, либо не учитывающим большое число вариан-

тов возможного развития ситуаций и, значит, ненадежным и т. д.

Автор полагает, что полученные от экспертов оценки вероятностей заведомо достаточно хорошо отражают реальность и прибегает к использованию определенной вероятностной модели сложного процесса «взаимодействия» рассматриваемых проблем. Решение соответствующих уравнений позволяет получить значения недостающих вероятностей. Искомые вероятности фигурируют в уравнениях в качестве неизвестных. Уравнения служат моделями, связывающими все интересующие исследователя вероятности.

Таким образом, можно сказать, что в рассматриваемом случае множеством-носителем изучаемой ЭСО является множество всех возможных состояний всех рассматриваемых проблем. Элементы соответствующей ММСЭД — приписываемые этим состояниям вероятности. Как было сказано, часть таких вероятностей — это непосредственные экспертные оценки. Другая часть получается с помощью решения предлагаемых в статье уравнений.

В той же статье описываются и такие подходы к нахождению требующихся вероятностей (подходы, использующие модели с условными вероятностями), при которых в качестве элементов ЭСО принимаются наборы состояний проблем; причем каждое предшествующее состояние рассматривается как влияющее на последующее состояние проблемы. В качестве элементов ММСЭД им будут отвечать соответствующие условные вероятности. Как и выше, часть таких вероятностей указывается экспертами, часть же находится в результате решения предлагаемых уравнений.

Очевидно, предлагаемые методы можно рассматривать как некоторый подход к шкалированию на этапе анализа социологической информации и как таковой он в значительной мере принадлежит к этапу моделирования «развития» изучаемой ММСЭД. В этой связи небезынтересно отметить, что место упомянутых экспертных оценок в моделях (уравнениях) могут фигурировать вероятности, в свою очередь, полученные в результате каких-то формальных построений (например, если каждой проблеме соответствует некоторая переменная, а оценка указанных вероятностей происходит на основе изучения взаимозависимостей таких переменных). Допускается также, что далее формальные построения могут анализироваться и корректироваться экспертами.

Второй путь анализа построенной ММСЭД состоит в использовании свойств некоторой МСО, более широкой (т. е. содержащей более широкий круг отношений между ее элементами), чем эта ММСЭД. В этом случае возникает проблема, называемая в теории измерений проблемой адекватности [23]: насколько и в каком смысле для получения содержательных выводов можно пользоваться теми свойствами МСО, которые включают в себя соотношения, не определенные в построенной ММСЭД, т. е. не моделировавшиеся при измерении. Именно такой подход применяется, когда мы, например, используем некоторые известные свойства

действительных чисел для анализа данных, полученных по порядковой и даже по интервальной шкале<sup>13</sup>.

Основная проблема, возникающая в таких случаях, состоит в доказательстве инвариантности утверждений, делаемых на основе применения метода относительно допустимых преобразований используемых шкал (это условие отражает наиболее часто используемое и единственное конструктивно проверяемое понимание адекватности метода). В сборнике вопросы, связанные с реализацией рассматриваемого пути анализа ММСЭД, затрагиваются мало. В статье А. И. Орлова приводятся полученные автором результаты, касающиеся инвариантности отношения неравенства между значениями разного вида средних относительно допустимых преобразований наиболее часто используемых в социологии шкал. Об адекватности (допустимости) различных мер близости идет речь в статье Г. В. Раушенбаха.

Третий путь анализа ММСЭД состоит в ее обогащении и разработан для моделей, являющихся результатами шкалирования по шкалам низких типов. При реализации этого пути обосновываются (на основе каких-то априорных, может быть, дополнительно проверяемых допущений) такие не моделировавшиеся в процессе измерения отношения, которые позволяют «превратить» ММСЭД в часть более богатой МСО, свойства которой хорошо изучены и известно, в какой мере ими можно пользоваться. Именно это делается, например, при использовании так называемых методов повышения уровня измерения. В этой связи прежде всего упомянем методы оцифровки и метризации, позволяющие осуществлять переход от порядковой шкалы к интервальной (хотя шкальные значения, полученные по интервальной шкале, не обладают всеми свойствами действительных чисел, но для нее относительно ряда математических методов решена проблема адекватности).

Методы оцифровки и метризации относятся к числу методов, позволяющих осуществить измерение по интервальной шкале, не отображая непосредственно в соответствующее числовое отношение интересующее исследователя отношение порядка расстояний между измеряемыми объектами (напомним, именно такое отображение приводит к интервальной шкале по самому ее определению). Как правило, предполагается, что лишь порядок между изучаемыми объектами непосредственно отображается в числовое отношение порядка (естественно, исследователь при этом пола-

<sup>13</sup> При использовании числовых ММСЭД (шкал) проблема адекватности возникает вследствие того, что не всем операциям, входящим в аксиомы, определяющие совокупность действительных чисел, соответствуют какие-либо эмпирические аналоги. Чаще всего это касается операции сложения. Подчеркнем, что входящие в определение используемой МСО аксиомы, включающие в себя только те соотношения, для которых существуют эмпирические аналоги, обязательно должны выполняться для изучаемой ЭСО. Иначе пользоваться какими бы то ни было свойствами действительных чисел в принципе становится невозможным.

тает, что интересующий его порядок между расстояниями трудно определить сразу, например, на основе опроса респондентов, поскольку последние могут затрудняться дать соответствующий ответ и т. д.).

Методы оцифровки, как правило, предполагают выполнение ряда условий для искомых шкальных значений (эти условия, по существу, являются некоторыми предположениями о характере изучаемой ЭСО). Используя эти условия, исследователь приписывает каждому эмпирическому объекту некоторое число таким образом, что в результате получает уверенность в том, что отношение порядка между получающимися числовыми интервалами отражает описанное выше эмпирическое отношение между изучаемыми объектами (о методах оцифровки см. [5]).

Близкими по своей сути к методам оцифровки являются методы так называемой метризации. Отличие последних состоит в следующем: при их использовании первоначальной целью исследователя является установление отношения порядка для расстояний между объектами или, что практически является тем же самым, между градациями рассматриваемого порядкового признака. И уже на основе такого отношения устанавливаются шкальные значения каждого объекта (о некоторых используемых в социологии методах метризации см. [27, гл. IV]). При использовании же методов оцифровки мы на основе каких-то «внешних» соображений приписывали объектам искомые шкальные значения и уже затем определяли, каково соотношение расстояний между объектами.

Когда применяются методы метризации, может и не ставиться задача нахождения шкальных значений. В этом ярко проявляется сущность интервальной шкалы. Ведь, как известно, при ее использовании исследователь в действительности определяет лишь отношение порядка для расстояний между объектами (и соответственно между искомыми шкальными значениями). Находящиеся обычно сами шкальные значения на самом деле определены лишь с точностью до произвольного положительного линейного преобразования. Использование методов, подобных методам оцифровки, иногда заставляет исследователя забывать об этом факте. Применение же методов метризации без нахождения шкальных значений как бы дает нам интервальную шкалу в «чистом» виде.

Рассматриваемые способы получения шкальных значений — оцифровка и метризация — относятся к числу методов шкалирования, реализуемых не при сборе, а при анализе уже собранной каким-то образом информации и как таковые могут расцениваться не только как способы измерения, но и как способы анализа некоторой первичной ММСЭД. Поэтому при их использовании большую роль приобретает соответствующая математическая модель, отражающая представление исследователя об изучаемом явлении. Для иллюстрации того, что мы имеем в виду под такой моделью, сравним два подхода к шкалированию.

В статье В. Т. Перекреста предлагается способ метризации значений порядкового признака <sup>14</sup>, в соответствии с которым расстояние между двумя градациями этого признака считается тем больше, чем больше респондентов дали ответы, соответствующие этим градациям и градациям, лежащим между ними (относительно рассматриваемого отношения порядка). С другой стороны, иногда предлагается такой способ оцифровки градаций порядкового признака, в соответствии с которым «вес» некоторой градации (и, следовательно, получаемое в результате оцифровки соответствующее этой градации числовое значение) тем больше, чем меньше респондентов дали соответствующие ей ответы. Подобная идея взвешивания опирается на предположение, что чем больше людей указывают некоторую градацию, тем меньше интенсивность «чувств», соответствующих этой градации. Нетрудно представить себе задачи, которым отвечают описанные методы. Легко такие заметить, что применение этих методов к одним и тем же данным приведет к прямо противоположным соотношениям между интересующими нас расстояниями. Каждому методу соответствует своя модель изучаемого явления. Роль выбора такой модели ярко видна.

В качестве еще одного способа реализации третьего пути анализа построенной ММСЭД можно назвать использование методов многомерного шкалирования для получения пространства признаков, значения которых получены по интервальной шкале. В основе каждого метода этой группы лежит определенная модель реальности, включающая в себя априорные предположения об изучаемой ЭСО. Матрица близостей, служащая исходной для методов многомерного шкалирования информации, часто является результатом оценок респондентами сходства или различия интересующих исследователя объектов друг с другом. Упомянутые предположения в таких случаях касаются соответствующих «объективной» и «субъективных» ЭСО. Для обеспечения корректности применения метода в подобных ситуациях требуется глубокий анализ «взаимодействия» этих ЭСО. Ряд предположений о подобных системах и их «взаимодействии» используется, в частности, при применении одного из подходов к многомерному шкалированию — техники развертывания [26].

Описанные предположения являются отражением одной из основных причин, обуславливающих трудности применения математики в социологии — причины, состоящей в том, что социолога могут интересовать одновременно и «объективная», и «субъективная» ЭСО. И тогда исследователь, по-существу, имеет дело с ЭСО, элементами которых служат пары типа (объект, наблюдатель).

Если матрицу близости получить, оценив близость изучаемых

<sup>14</sup> Рассматривается более общий случай, когда расстояния вводятся не для градаций одного признака, а для сочетаний значений нескольких признаков.

объектов (например, респондентов) на основе известных для них значений признаков, полученных по шкалам низких типов, тогда методы многомерного шкалирования можно использовать для повышения уровня измерения. Однако к этому вряд ли стоит стремиться. Подобный подход не отвечает основой, на наш взгляд, цели внедрения многомерного шкалирования в социологические исследования: повышения надежности измерения, когда источником информации является человек. Возможность повышения этой надежности связана с тем, что, как показывают некоторые социально-психологические исследования, человеку свойственно оценивать свое собственное мнение более объективно не путем прямой оценки того или иного явления, а путем сравнительного оценивания различных пар явлений. В результате же такой сравнительной оценки мы можем получить исходную для многомерного шкалирования матрицу близостей.

Социологические данные из-за своей нечисловой природы, редко удовлетворяют требованиям метрического многомерного шкалирования. Иногда, однако, возникает необходимость применения соответствующих методов именно к специфическим для социологии нечисловым данным. Это можно сделать, если к исходным данным предварительно применить какой-либо способ метризации (как это делается, например, в статье В. Т. Перекреста). В таком случае мы будем иметь пример последовательного применения нескольких математических методов с целью получения координат интересующих исследователя объектов в некотором «хорошем» пространстве. Это отражает сложность социальных процессов, при изучении которых даже процесс измерения может быть столь трудоемким.

Ряд других примеров последовательного использования нескольких математических методов, центральным из которых является метод многомерного шкалирования, приведены в статье Г. А. Сатарова. При этом другие методы используются как для подготовки данных для многомерного шкалирования (используется кластерный анализ для сокращения объема изучаемой совокупности объектов), так и для интерпретации результатов последнего (имеются в виду методы, близкие к методам множественной регрессии).

В той же статье дается пример параллельного использования нескольких методов на этапе интерпретации результатов многомерного шкалирования. Потребность в таком подходе возникает вследствие того, что априорное отсутствие модели, параметры которой идентифицируются в процессе применения методов многомерного шкалирования, «заложено» в самих методах: эти методы не позволяют получить однозначный результат.

Поскольку методы многомерного шкалирования являются методами шкалирования не при сборе, а при анализе социологических данных, то при их применении во весь рост встает проблема выбора соответствующей модели. Поскольку в понятие такой модели входят трудно поддающиеся проверке предположения

о поведении респондента в процессе получения от него требуемой информации, то приходится говорить об априорном отсутствии у исследователя четких представлений о модели изучаемого явления. Как уже отмечалось, подобное положение обуславливает необходимость одновременного использования нескольких методов многомерного шкалирования.

### Литература

1. Аванесов В. С. Достоверность различий и применение матричных методов в социологических исследованиях. — Социол. исслед., 1975, № 4, с. 146—158.
2. Аганбегян А. Г. О некоторых особенностях применения математических моделей в социологических исследованиях. М., 1968.
3. Актуальные задачи развития методологии и методики социологических исследований. — Социол. исслед., 1975, № 3, с. 3—9.
4. Андреев В. Г., Толстова Ю. Н. Основные этапы процесса применения математических методов в социологии. — В кн.: Применение математических методов и ЭВМ в социологических исследованиях. М., 1982, с. 5—21.
5. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. М., 1982.
6. Белых О. В., Беляев Э. В. Возможности применения теории графов в социологии. — В кн.: Человек и общество. Л., 1968, вып. 1.
7. Бугосец А. Г. Последовательное применение алгоритмов многомерной классификации при решении задач типологии. — В кн.: Многомерный анализ социологических данных: (Методические рекомендации, алгоритмы и описание программ). М., 1981, с. 24—73.
8. Гаврилец Ю. Н. Введение — В кн.: Модели социально-экономических процессов и социальное планирование. М., 1979, с. 3—5.
9. Гаврилец Ю. Н. Целевые функции социально-экономического планирования. М., 1983.
10. Загоруйко Н. Г. Эмпирическое предсказание. Новосибирск, 1979.
11. Загоруйко Н. Г., Самохвалов К. Ф., Свириденко Д. И. Логика эмпирических исследований. Новосибирск, 1978.
12. Золотарев В. М. Одномерные устойчивые распределения. М., 1983.
13. Клигер С. А., Косолапов М. С., Толстова Ю. Н. Шкалирование при сборе и анализе социологической информации. М., 1978.
14. Лбов Г. С. Методы обработки разнотипных экспериментальных данных. Новосибирск, 1981.
15. Миркин В. Г., Высоцкая Н. В., Куперштох В. Л., Полищук Л. И., Трифимов В. А., Черный Л. Б. Шкалы упорядочения. — В кн.: Моделирование в экономических исследованиях. Новосибирск, 1978.
16. Моисеев Н. И. Математика в социальных науках. — В кн.: Математические методы в социологическом исследовании. М., 1981, с. 10—24.
17. Осипов Г. В. Основные направления применения математических методов в конкретных социальных исследованиях. — Социол. исслед., 1976, № 3, с. 131—140.
18. Павловский В. Введение в математическую статистику. М., 1967.
19. Паниотто В. И. Социометрические методы изучения малых социальных групп. — Социол. исслед., 1976, № 3, с. 141—152.
20. Паниотто В. И. Структура межличностных отношений: Методика и мат. методы исследования. Киев, 1975.
21. Парнюк М. А. Принципы детерминизма в системе материалистической диалектики. Киев, 1972.
22. Прохоров Ю. В., Севастьянов Б. А. Вероятностей теория. — МЭ, М., 1977, т. 1, с. 661.
23. Пфанцгаль Н. Теория измерений. М., 1976.

24. Саганенко Г. И. Надежность результатов социологического исследования. Л., 1983.
25. Саганенко Г. И. Социологическая информация. Л., 1979.
26. Сатаров Ю. А., Каменский В. С. Общий подход к анализу экспертных оценок методами неметрического многомерного шкалирования. — В кн.: Статистические методы анализа экспертных оценок. М., 1977, с. 251 — 266.
27. Типология и классификация в социологических исследованиях. М., 1982.
28. Тишин А. И. Теория категорий и системные исследования в социологии. — В кн.: Математические методы в социологическом исследовании. М., 1981, с. 37—46.
29. Толстова Ю. Н. Методика применения математических методов для анализа социологических данных: постановка проблемы и ее решение в статьях сборника. — В кн.: Многомерный анализ социологических данных: (методические рекомендации, алгоритмы и описание программ). М., 1981, с. 11—23.
30. Трофимов В. А. Экспериментальное обоснование методов качественного факторного анализа. — В кн.: Методы анализа многомерной экономической информации. Новосибирск, 1981, с. 30—48.
31. Тюрин Ю. И., Литвак Б. Г., Орлов А. И., Сатаров Г. А., Шмерлинг Д. С. Анализ нечисловой информации. М. 1981.
32. Измерение. — В кн.: Философский энциклопедический словарь. М., 1983.
33. Цапенко М. П. Измерительные информационные системы. М., 1974.
34. Чесноков С. В. Детерминационный анализ социально-экономических данных. М., 1982.
35. Чумаков Б. И. Моделирование ответственности средствами формальных логик. — В кн.: Математическое моделирование и применение вычислительной техники в социологических исследованиях. М., 1980, с. 128—133.
36. Coombs C. H. Mathematical models in psychological scaling. — J. Amer. Statist. Assoc., 1951, vol. 46, p. 480—489.
37. Coombs C. H., Dawes R. H., Tversky A. Mathematical psychology: An elementary introduction. N. Y., 1970.
38. Coombs C. H., Raiffa H., Thrall R. M. Some views on mathematical models and measurement theory. — Psychol. Rev., 1954, vol. 61, p. 132 — 144.

## Е. Н. КУЗНЕЦОВ, И. Б. МУЧНИК, Л. В. ШВАРЦЕР

### МОНОТОННЫЕ СИСТЕМЫ И ИХ СВОЙСТВА

В статье впервые дается систематическое описание нового метода структуризации социально-экономической информации, основанного на теории так называемых монотонных систем. Она нацелена на построение системного описания анализируемых данных. В этой теории структура данных описывается в виде последовательности математических объектов — квазидадер, что придает методологическому подходу авторов некоторое сходство с применением классического агломеративного иерархического алгоритма типа «дендрограмма» (однако в теории монотонных систем начальный член упомянутой последовательности — ядро — определяется здесь из соображений самой теории). Статья посвящена в основном изложению математической основы предлагаемого метода. Это изложение отвечает первому выделенному в Предисловии этапу внедрения математического метода в практику.

# 1. Задачи структуризации социально-экономической информации и теория монотонных систем

В области разработки математических методов структуризации социально-экономической информации наметились следующие пути. Это, во-первых, принятие в качестве искомой структуры той или иной классификации обрабатываемых данных, хотя, конечно, иногда рассматриваются и некоторые другие формы описания структуры: распределение на оси интегрального индекса, графа макросвязей и т. д. [7]. Во-вторых, выбор в качестве информационной базы для классификации матрицы тех или других парных коэффициентов связей между классифицируемыми элементами [5].

И наконец, следует признать, что задача по структуризации формулируется как та или иная задача комбинаторной экстремизации, решение которой в практически интересных случаях может быть только грубо приближенным, так как соответствующие алгоритмы поиска точного решения принципиально требуют огромного перебора. Правда, часто оказывается так, что или имеется необходимая априорная информация, которая позволяет «правильно» выбрать начальные условия, или обрабатываемые данные оказываются «хорошо организованными». В итоге простой приближенный алгоритм за несколько «проб» по разным начальным условиям обычно позволяет найти приемлемое решение.

Сильное осложнение при практическом решении задач структуризации социально-экономических данных создает еще следующее обстоятельство: формальное описание алгоритма решения такой задачи обычно содержит один свободный «структурный» параметр (или число классов, или порог существенности, который определяет максимально допустимый «радиус» класса в пространстве признаков). По этому структурному параметру алгоритм должен быть настроен.

Указанные особенности современных методов структуризации социально-экономической информации, очевидно, являются весьма жесткими ограничениями для прикладников, которые в своих качественных описаниях пользуются существенно более широкими представлениями.

Отношение большинства разработчиков методов структуризации к отмеченным выше особенностям современных методов структуризации состоит в том, что эти особенности признаются характеристиками проблемы, и потому поиск возможностей их «ревизии» или вовсе не рассматривается, а если и рассматривается, что бывает крайне редко, то по частным поводам [1]. Между тем актуальность такой ревизии хорошо чувствуется в прикладных исследованиях, когда все больше внимания уделяется методам, порождающим серии специально организованных классификаций [3].

Прикладники используют все более сложные комбинированные процессы анализа, чтобы преодолеть ограниченность использования только парных коэффициентов связей [5]. Наконец, активно разрабатываются специальные методы структуризации, ориентированные на частные типы информации, позволяющие находить не приближенные, а точные решения задачи структуризации [2]. Таким образом, возникает проблема поиска такого универсального математического аппарата, который бы позволял снять указанные ограничения на уровне постановки задач в общем виде.

Нам представляется, что одна из возможностей решения этой проблемы лежит в интенсивном привлечении для этих целей теории монотонных систем [6]. Описание основной математической конструкции этой теории позволяет рассматривать обрабатываемые наблюдения не как изолированные объекты или признаки, а как элементы целостной системы, которые взаимодействуют и потому существенны не только сами по себе, но и тем, какое «место» они занимают в кругу других данных. В результате оказывается, как будет видно из дальнейшего, что теория монотонных систем порождает не какую-то одну процедуру структуризации, а в определенном смысле дает исчерпывающее и разностороннее описание структуры обрабатываемого массива.

Во втором параграфе дано точное определение, что такое монотонная система. В рамках этого понятия вводится в рассмотрение конструкция «основной структуры данных» — ядра, которое в рамках представлений о классификации удобно ассоциировать как систему «центров» классов. При этом отыскивается точное решение соответствующей экстремальной задачи. Причем ее основной структурный параметр — число центров — не задается заранее, а определяется автоматически в ходе решения. Получающееся решение обладает рядом важных для его содержательной интерпретации свойств, которые приводятся в этом же параграфе.

В третьем параграфе понятие основной структуры (ядра) расширяется до представления о «полной структуре» обрабатываемых данных. В качестве такой полной структуры выступает так называемая последовательность квазиядер, каждый элемент которой есть определенный (и автоматически определенный!) уровень детализации описания структуры данных, а в целом эта последовательность определяет своего рода «спектр» обрабатываемого массива.

Последний параграф посвящен описанию процедур нахождения этого спектра. Показано, что сложность такой процедуры не превосходит  $N^3$  где  $N$  — число наблюдений в обрабатываемом массиве, причем если исследователя интересует только «основная» структура, а не весь спектр, то сложность процедуры резко уменьшается.

Теория монотонных систем выделяет некоторый новый класс эффективно решаемых комбинаторно-экстремальных задач и в этом смысле имеет более широкое значение, чем просто еще одна

теория методов обработки эмпирической информации. В частности, она может найти применение и в задачах математического моделирования организационных структур, структурных систем распределения, формирования коалиционного взаимодействия и т. п. [4].

## 2. Экстремальные подсистемы монотонных систем

Рассмотрим конечное множество элементов  $W$ ,  $|W| = N$ . Пусть задана скалярная функция  $\pi$ , которая каждой паре  $(i, H)$ , где  $H \subseteq W$  — произвольное подмножество множества  $W$ , а  $i \in H$  — его элемент, ставит число  $\pi(i, H)$ . Содержательно это число может измерять близость элемента  $i$  к подмножеству  $H$ , «значимость», «важность» элемента  $i$  в подмножестве  $H$ , степень влияния элемента  $i$  на подмножество  $H$  и т. д. Так, если на множестве  $W$  задана матрица  $\|\rho_{ij}\|$  расстояний между всеми парами элементов, то простейшей функцией  $\pi(i, H)$  может служить сумма расстояний от элемента  $i$  до всех (остальных) элементов подмножества  $H$ :

$$\pi(i, H) = \sum_{j \in H} \rho_{ij}.$$

Число  $\pi(i, H)$  будем называть весом элемента  $i$  на множестве  $H$ .

Рассмотрим систему  $\langle W, \pi \rangle$ , состоящую из конечного множества элементов  $W$  с заданной на множестве пар  $(i, H)$ ,  $H \subseteq W$ ,  $i \in H$ , функцией  $\pi(i, H)$ .

Систему  $\langle W, \pi \rangle$  называют монотонной, если либо

$$\pi(i, H \setminus j) \leq \pi(i, H), \quad \forall i, j \in H, i \neq j, \forall H \subseteq W, \quad (1)$$

либо

$$\pi(i, H \setminus j) \geq \pi(i, H), \quad \forall i, j \in H, i \neq j, \forall H \subseteq W. \quad (2)$$

Если выполняется (1), то называют ее  $\ominus$ -монотонной и обозначают через  $\langle W, \pi^- \rangle$ , а если (2) — то  $\oplus$ -монотонной и обозначают соответственно через  $\langle W, \pi^+ \rangle$ .

*Определение.*  $\ominus$ -монотонной системой  $\langle W, \pi^- \rangle$  называется конечное множество  $W$  с заданной на нем скалярной числовой функцией  $\pi(i, H)$ ,  $H \subseteq W$ ,  $i \in H$ , обладающей свойством монотонности (1).

Аналогично определяется  $\oplus$ -монотонная система. Неравенство (2) в этом случае также выражает свойство монотонности. Итак, имеются монотонные системы двух типов (или двух «знаков»).

Центральная задача теории монотонных систем — выделение некоторой экстремальной подсистемы монотонной системы — так называемого определяемого множества (или наибольшего ядра). Ее формализация и решение осуществляются так.

На множестве всех подмножеств множества  $W$   $\ominus$ -монотонной системы определяется скалярная функция,  $F^-$ , ставящая в соответствие каждому подмножеству  $H$ ,  $H \subseteq W$ , число  $F^-(H)$  по

следующему правилу:

$$F^-(H) = \min_{i \in H} \pi^-(i, H), \quad \forall H \subseteq W. \quad (3)$$

*Определение.* Ядрами (экстремальными подсистемами)  $\ominus$ -монотонной системы  $\langle W, \pi^- \rangle$  называются такие подмножества множества  $W$ , на которых достигается максимум функции  $F^-(H)$ .

Аналогично определяются ядро  $\ominus$ -монотонной системы, для чего вводится функция  $F^+(H)$ :

$$F^+(H) = \max_{i \in H} \pi^+(i, H), \quad \forall H \subseteq W, \quad (4)$$

которая должна достигать на них минимального значения.

В дальнейшем будем писать просто  $\pi(i, H)$  и  $F(H)$ , если из контекста ясно, о какой именно системе идет речь,  $\ominus$ - или  $\oplus$ -монотонной (либо если имеется в виду любая из них). Аналогично будем поступать и с другими обозначениями.

Упорядочим элементы множества  $W$  произвольным образом. Полученной последовательности элементов  $A = \langle \alpha_1, \dots, \alpha_N \rangle$ , где  $W = \{\alpha_1, \dots, \alpha_N\}$  однозначно соответствует последовательность  $\bar{H}(A)$ , или просто  $\bar{H}$ ,  $\bar{H} = \langle H_1, \dots, H_N \rangle$ , вложенных подмножеств <sup>1</sup> множества  $W$ , где  $H_1 = W$ ,  $H_2 = H_1 \setminus \alpha_1, \dots, H_{k+1} = H_k \setminus \alpha_k, \dots, H_N = \alpha_N$ .

*Определение.* Упорядоченная последовательность  $A$  элементов множества  $W$  называется определяющей последовательностью  $\ominus$ -монотонной системы  $\langle W, \pi^- \rangle$ , если в соответствующей ей последовательности множеств  $\bar{H}$  существует такая подпоследовательность  $\bar{\Gamma} = \langle \Gamma_1, \dots, \Gamma_p \rangle$ , где  $\Gamma_1 = H_1 = W$ , что

$$\pi(\alpha_k, H_k) < F(\Gamma_{j+1}), \quad \forall \alpha_k \in \Gamma_j \setminus \Gamma_{j+1}, \quad j = 1, p-1, \quad (5)$$

$$F(L) \leq F(\Gamma_p), \quad \forall L \subset \Gamma_p. \quad (6)$$

Аналогично находится определяющая последовательность  $\oplus$ -монотонной системы, только неравенства (5) и (6) заменяются соответственно на

$$\pi(\alpha_k, H_k) > F(\Gamma_{j+1}), \quad \forall \alpha_k \in \Gamma_j \setminus \Gamma_{j+1}, \quad (7)$$

$$F(L) \geq F(\Gamma_p), \quad \forall L \subset \Gamma_p. \quad (8)$$

*Определение.* Множество  $G$ ,  $G \subseteq W$  называется определяемым множеством монотонной системы  $\langle W, \pi \rangle$ , если существует такая определяющая последовательность, что  $\Gamma_p = G$ .

Будем использовать для определяемого множества  $\ominus$ -монотонных систем обозначение  $G^-$ , а  $\oplus$ -монотонных —  $G^+$ .

Справедливы следующие две центральные теоремы <sup>2</sup> теории монотонных систем [6].

<sup>1</sup> Для упрощения записи будем писать  $H \cup i$  вместо  $H \cup \{i\}$  и  $H \setminus i$  вместо  $H \setminus \{i\}$ .

<sup>2</sup> Поскольку большинство свойств (фактов) монотонных систем симметричны для  $\ominus$ - и  $\oplus$ -монотонных систем, то теоремы, в которых они формулируются, отмечаются соответствующим знаком «-» или «+». При ссылке на ту или иную теорему знак будет опускаться, если из контекста ясно, о каких системах,  $\ominus$ - или  $\oplus$ -монотонных, идет речь.

*Теорема 1<sup>-</sup>.* На определенном множестве  $G$   $\ominus$ -монотонной системы функция  $F(H)$  достигает глобального максимума. Существует единственное определенное множество. Все множества, на которых достигается глобальный максимум функции  $F$ , т. е. все ядра  $\ominus$ -монотонной системы, лежат внутри определенного множества.

Обозначим через  $X$  множество всех подмножеств множества  $W$ .

*Теорема 2<sup>-</sup>.* Система всех тех множеств из  $X$ , на которых функция  $F$  достигает глобального максимума, замкнута по отношению к бинарной операции объединения множеств.

Теоремы 1<sup>+</sup> и 2<sup>+</sup> для  $\ominus$ -монотонных систем формулируются аналогично, только максимум заменяется на минимум.

Рассмотрим числовую последовательность  $\bar{\varepsilon} = \langle \varepsilon_1, \dots, \varepsilon_p \rangle$ , где  $\varepsilon_j = F(\Gamma_j)$ ,  $j = \overline{1, p}$ .

*Лемма 1.* Для  $\ominus$ -монотонной системы имеет место цепочка неравенств

$$\varepsilon_1 < \varepsilon_2 < \dots < \varepsilon_j < \dots < \varepsilon_p, \quad (9)$$

а для  $\ominus$ -монотонной соответственно

$$\varepsilon_1 > \varepsilon_2 > \dots > \varepsilon_j > \dots > \varepsilon_p. \quad (10)$$

*Доказательство.* Пусть  $\gamma_j$  — первый элемент множества  $\Gamma_j$  ( $j = \overline{1, p-1}$ ) в определяющей последовательности.

Поскольку, по определению, последовательность  $\bar{\Gamma}$  есть подпоследовательность последовательности множеств  $\bar{H}$ , множество  $\Gamma_j$  есть в то же время какое-то множество  $H_k$  из  $\bar{H}$ :  $\Gamma_j = H_k \in \bar{H}$ . При этом, очевидно,  $\gamma_j = \alpha_k$ .

Тогда по свойству (5) определяющей последовательности и по определению (3) функции  $F$  получаем, что

$$\varepsilon_j = F(\Gamma_j) \leq \pi(\gamma_j, \Gamma_j) = \pi(\alpha_k, H_k) < F(\Gamma_{j+1}), \quad j = \overline{1, p-1},$$

откуда и следует соотношение (9). Цепочка (10) доказывается аналогичным образом. Лемма доказана.

*Доказательство теоремы 1<sup>-</sup>.* Предположим, что определенное множество  $G$  существует<sup>3</sup>. Нужно доказать, что

$$F(G) \geq F(H), \quad \forall H \subseteq W.$$

Предположим, что существует такое множество  $L \subseteq W$ , что

$$F(G) \leq F(L). \quad (11)$$

Возможно одно из двух: либо  $L \subseteq G$ , либо  $L \setminus G \neq \emptyset$ .

Рассмотрим первый случай. По определению множества  $G$ , существует такая определяющая последовательность  $A$  элементов

---

Существование определенного множества и определяющей последовательности здесь не доказывается. В разд. 4 определяющая последовательность строится конструктивно и тем самым доказывается существование определенного множества. Поэтому полным доказательством теоремы 1 является данное доказательство вместе с алгоритмами из разд. 3.

множества  $W$  со свойством (6), что строгое неравенство  $F(G) < F(L)$  не выполняется и, следовательно, выполняется лишь равенство в неравенстве (11).

В этом случае первое и третье утверждения теоремы доказаны, а единственность определяемого множества  $G$  доказывается ниже для обоих случаев сразу.

Второй случай ( $L \setminus G \neq \emptyset$ ). Пусть  $H_n$  — наименьшее множество последовательности  $\bar{H}$ , соответствующей определяющей последовательности  $A$ , содержащее множество  $L \setminus G$ , т. е. существует такой элемент  $\alpha_n \in L$ , что  $\alpha_n \in H_n$ , но  $\alpha_n \notin H_{n+1}$ .

Пусть теперь  $\Gamma_S$  есть наименьшее множество последовательности  $\bar{\Gamma}$ , такое, что  $\Gamma_S \supseteq H_n$ , но  $\Gamma_{S-1} \not\supseteq H_n$ . Это означает, что  $\alpha_n \in \Gamma_S$ ,  $L \subseteq \Gamma_S$  (так как  $L \subseteq H_n \subseteq \Gamma_S$ ),  $\alpha_n \notin \Gamma_{S+1}$ .

Тогда, используя свойство (5) определяющей последовательности и утверждение леммы 1, заключаем, что

$$\pi(\alpha_n, H_n) < F(\Gamma_{S+1}) < F(G). \quad (12)$$

Объединяя (11) и (12), получаем, что

$$\pi(\alpha_n, H_n) < F(G) \leq F(L).$$

С другой стороны, по свойству монотонности ( $L \subseteq H_n$ )

$$\pi(\alpha_n, L) \leq \pi(\alpha_n, H_n),$$

откуда следует, что

$$\pi(\alpha_n, L) < F(L) = \min_{i \in L} \pi(i, L).$$

Полученное неравенство означает, что в множестве  $L$  существует элемент  $\alpha_n$  с весом, строго меньшим минимального, что невозможно. Тем самым доказано, что  $L$  может быть лишь подмножеством множества  $G$ , т. е. что все множества, на которых также достигается глобальный максимум функции  $F$ , находятся внутри  $G$ .

Остается доказать, что если определяемое множество существует, то оно единственное. Действительно, в силу доказанного выше можно лишь предполагать, что некоторое отличное от  $G$  определяемое множество  $G'$  включено в  $G$ . С другой стороны, те же рассуждения справедливы и по отношению к множеству  $G'$ , откуда и следует, что  $G \subseteq G'$ . Тем самым теорема доказана.

**Доказательство теоремы 2.** Пусть  $G_1$  и  $G_2$  — два различных ядра  $\ominus$ -монотонной системы. Необходимо доказать, что множество  $G_1 \cup G_2$  также является ядром, т. е. что

$$F(G_1 \cup G_2) = F(G_1) = F(G_2) = \max_{H \subseteq W} F(H). \quad (13)$$

Поскольку на  $G_1$  и  $G_2$  достигается максимум функции  $F$ , то имеют место неравенства

$$F(G_1 \cup G_2) \leq F(G_1), F(G_1 \cup G_2) \leq F(G_2). \quad (14)$$

С другой стороны, пусть  $g, g \in G_1 \cup G_2$ , — такой элемент, на котором достигается значение функции  $F(G_1 \cup G_2)$ :

$$\pi(g, G_1 \cup G_2) = F(G_1 \cup G_2).$$

Пусть, например,  $g \in G_1$ . (Если  $g \in G_2$ , то дальнейшие рассуждения аналогичны). Тогда по свойству монотонности имеем

$$\pi(g, G_1) \leq \pi(g, G_1 \cup G_2)$$

и, следовательно,

$$F(G_1) = \min_{i \in G_1} \pi(i, G_1) \leq \pi(g, G_1) \leq \pi(g, G_1 \cup G_2) = F(G_1 \cup G_2). \quad (15)$$

Объединяя (14) и (15), получаем (13). Теорема доказана.

Теоремы 1 и 2 устанавливают важные особенности структуры монотонной системы: множество всех ее ядер образует замкнутую систему множеств по отношению к операции объединения множеств, причем объединение всех ядер — определимое множество — является самым большим ядром.

Алгоритм построения определяющей последовательности, рассматриваемый в разд. 4, позволяет эффективно (без перебора) выделять (в соответствии с его определением) именно это самое большое ядро, которое далее будем называть просто ядром. Другими словами, ядром монотонной системы служит последнее множество  $G = \Gamma_p$  в последовательности  $\bar{\Gamma}$ , зафиксированное при построении определяющей последовательности.

Таким образом, конструктивно ядро монотонной системы определяется при помощи построения определяющей последовательности ее элементов. В связи с этим рассмотрим множество всех возможных определяющих последовательностей данной системы.

Вообще говоря, может быть много различных определяющих последовательностей одной и той же монотонной системы, удовлетворяющих приведенному выше определению. Однако, как это установлено в теоремах 1 и 2, у монотонной системы есть только одно ядро, т. е. наибольшее множество, на котором функция  $F(H)$  принимает глобально-экстремальное значение. Значит, разные определяющие последовательности одной системы построены так, что последнее множество  $G$  в соответствующих им последовательностях  $\bar{\Gamma}$  одно и то же. Это обстоятельство позволяет ограничить все множество возможных определяющих последовательностей данной монотонной системы.

Действительно, пусть известно, как из некоторой определяющей последовательности получить другую определяющую последовательность. Тогда возможный способ перечисления всех определяющих последовательностей заключается в том, что сначала при помощи какого-либо конструктивного алгоритма строится одна определяющая последовательность элементов, а затем множество всех остальных определяющих последовательностей данной мо-

погонной системы получают из нее путем последовательного преобразования.

Формальное ограничение множества всех определяющих последовательностей, получаемых при помощи такого способа, устанавливает следующая теорема.

*Теорема 3<sup>-</sup>.* Если  $\alpha_S$  и  $\alpha_t$ ,  $S < t$ , такие два элемента определяющей последовательности  $A = \langle \alpha_1, \dots, \alpha_N \rangle$ , что  $\alpha_S, \alpha_t \in G$ , то последовательность  $A'$ , где

$$A' = \langle \alpha_1, \dots, \alpha_{S-1}, \alpha_t, \alpha_{S-1}, \dots, \alpha_{t-1}, \alpha_S, \alpha_{t+1}, \dots, \alpha_N \rangle,$$

также является определяющей. Если же  $\alpha_S$  и  $\alpha_t$ ,  $S < t$ , такие два элемента, что  $\alpha_S \in W \setminus G$ , а  $\alpha_t \in G$ , то последовательность  $A'$  определяющей не является.

*Доказательство.* Докажем первое утверждение теоремы. Пусть  $\alpha_S, \alpha_t \in G$ . Поскольку  $G = \Gamma_p$  — последнее множество в последовательности  $\bar{\Gamma}$  (и  $m$ -е множество в последовательности  $H$ ), последовательности  $A$  и  $A'$ ,  $H$  и  $H'$  вплоть до элемента  $i_{m-1}$  и вплоть до множества  $H_m = G$  соответственно совпадают. Следовательно, последовательности множеств  $\bar{\Gamma}$  и  $\bar{\Gamma}'$  также совпадают, причем соотношение (5), верное, по определению, для последовательности  $\bar{\Gamma}$ , справедливо также и для  $\bar{\Gamma}'$ . Выполнение соотношения (6), очевидно, не зависит от перестановок элементов внутри множества  $G$ .

Второе утверждение (в случае, когда  $\alpha_S \in W \setminus G$ ,  $\alpha_t \in G$ ). Учитывая тот факт, что  $\alpha_S = \alpha_t \in G$ , а  $H_S = H_t \supseteq G$ , и используя определение функции  $F(H)$  и свойство монотонности, получаем соотношение

$$F(G) \leq \pi(\alpha_t, G) \leq \pi(\alpha_t, H_S) = \pi(\alpha_S, H_S),$$

которое противоречит свойству (5) определяющей последовательности по отношению к последовательности  $A'$ . Теорема доказана.

Одновременно доказано, что во втором случае (когда  $\alpha_S \in W \setminus G$ ,  $\alpha_t \in G$ ) множество  $(G \cup \alpha_S) \setminus \alpha_t$  не является ядром.

Легко видеть, что теорема 3<sup>-</sup> не охватывает третий случай, а именно когда  $\alpha_S, \alpha_t$ ,  $S < t$ , — такие два элемента последовательности  $A$ , что  $\alpha_S, \alpha_t \in W \setminus G$ . Можно показать, что перестановка произвольных элементов вне ядра может привести как к сохранению, так и к нарушению свойств определяющей последовательности (а именно соотношения (5)).

Поскольку преобразования определяющих последовательностей друг в друга есть всевозможные перестановки элементов множества  $G$  в сочетании с некоторыми перестановками элементов множества  $W \setminus G$ , число всех возможных определяющих последовательностей монотонной системы, очевидно, ограничено сверху величиной  $|W \setminus G|! \cdot |G|!$ .

### 3. Общие свойства монотонных систем

В данном разделе изучаются общие свойства  $\subseteq$ - и  $\ominus$ -монотонных систем. Поэтому в тех случаях, когда эти свойства для  $\subseteq$ - и  $\ominus$ -монотонных систем симметричны, как правило, дается формулировка свойства только для  $\subseteq$ -монотонных систем. Аналогичное свойство для  $\ominus$ -монотонных систем легко получить, если взаимно заменить знаки неравенств  $>$  и  $<$ ,  $\geq$  и  $\leq$ , операции взятия максимума и минимума и т. п.

Основная цель данного раздела — показать, что любое множество  $\Gamma_j$ ,  $j = \overline{1, p-1}$ , из последовательности  $\bar{\Gamma}$  обладает некоторыми экстремальными свойствами, аналогичными свойствам ядра  $G = \Gamma_p$ . Тем самым будет показано значение элементов последовательности  $\bar{\Gamma}$  для задач структуризации наряду с самим ядром  $G$ . Для этого нам понадобятся разные эквивалентные определения ядра.

Прежде всего в соответствии с замечанием, сделанным в конце предыдущего раздела, дадим более точное определение ядра.

**Определение.** Ядром  $\subseteq$ -монотонной системы  $\langle W, \pi \rangle$  будем называть наибольшее по мощности подмножество  $G$  множества  $W$ , доставляющее максимум функции  $F(H)$ , т. е. множество  $G$ , удовлетворяющее следующим соотношениям:

$$F(G) = \max_{H \subseteq W} F(H), \quad (16)$$

$$|G| = \max_{H \subseteq W, F(H) = F(G)} |H|. \quad (17)$$

Соотношения (16) и (17) из этого определения ядра  $G$  можно переписать в более удобной форме

$$F(H) < F(G), \quad \forall H \subseteq W, |H| > |G|, \quad (18)$$

$$F(H) \leq F(G), \quad \forall H \subseteq W, |H| \leq |G|. \quad (19)$$

С другой стороны, из теорем 1 и 2 непосредственно следует, что для определимого множества  $G$  справедливо

$$F(H) < F(G), \quad \forall H \subseteq W, H \setminus G \neq \emptyset, \quad (20)$$

$$F(H) \leq F(G), \quad \forall H \subseteq W, H \subseteq G. \quad (21)$$

Можно показать, что соотношения (18), (19) следуют из соотношений (20), (21). Действительно, множество подмножеств, включающих элементы, не принадлежащие  $G$ , т. е. множество  $\{H \mid H \setminus G \neq \emptyset\}$ , состоит из двух частей: множество подмножеств  $W$  с числом элементов, большим  $|G|$ , и множество подмножеств  $W$ , для которых одновременно выполняются соотношения  $|H| \leq |G|$  и  $H \setminus G \neq \emptyset$ . Последнее составляет одновременно часть множества  $\{H \mid |H| \leq |G|\}$ , вторую, дополнительную, часть которого составляет множество  $\{H \mid H \subseteq G\}$ . Учитывая эти соотношения, легко получить требуемое утверждение.

Это, в свою очередь, означает, что теорему 1 из предыдущего раздела можно переформулировать следующим образом.

*Теорема 4.* Определемое множество есть единственное ядро.

Иными словами, других наибольших ядер — кроме определенного множества — нет. Значит, понятия определенного множества и ядра, определенные выше независимо друг от друга, обозначают один и тот же объект. Ядро тем самым получает несколько эквивалентных определений:

как определенное множество,

как множество, удовлетворяющее соотношениям (16), (17), или (18), (19), или (20), (21).

Эти разные определения используются в дальнейшем при доказательстве тех или иных свойств монотонных систем. Кроме того, в этих условиях естественно попытаться найти такое определение ядра  $G$ , которое содержало бы наименее жесткие и легко проверяемые теоретически требования к названному множеству. Тогда доказательство факта, что какое-либо заданное множество является ядром монотонной системы, было бы наиболее простым. В частности, чтобы доказать адекватность описываемого в следующем разделе алгоритма выделения ядра монотонной системы, понадобится еще одно определение ядра.

Пусть  $G$  — это множество, удовлетворяющее следующим соотношениям:

$$F(H) < F(G), \forall H \subseteq W, H \supset G, \quad (22)$$

$$F(H) \leq F(G), \forall H \subseteq W, H \subseteq G. \quad (23)$$

*Теорема 5.* Для ядра  $G$   $\in$ -монотонной системы справедливы соотношения (22), (23). Множество  $G$ , удовлетворяющее соотношениям (22), (23), является ядром.

*Доказательство.* Нужно доказать, что из соотношений (16), (17), или (18), (19), или (20), (21) следуют (22), (23) и наоборот, из (22), (23) следует справедливость соотношений (16)–(21).

Легко видеть, что соотношения (21) и (23) совпадают, а соотношение (22) является следствием соотношения (18) или (20). Иначе говоря, для ядра, определенного выше, соотношения (22), (23) справедливы. Покажем теперь, что из (22), (23) следует справедливость (20), (21).

Действительно, пусть  $H$  — такое подмножество множества  $W$ , что  $H \setminus G \neq \emptyset$ . Рассмотрим множество  $H \cup G$ . Поскольку  $H \cup G \supset G$ , то в соответствии с (22) имеем

$$F(H \cup G) < F(G).$$

Пусть  $\alpha$  — такой элемент множества  $H \setminus G$ , на котором достигается значение  $F(H \cup G)$ , т. е.

$$F(H \cup G) = \pi(\alpha, H \cup G).$$

Предположим, что  $\alpha \in G$ . Тогда, используя свойство монотонности и определение функции  $F(H)$ , получаем следующую це-

почку неравенств:

$$F(G) \leq \pi(\alpha, G) \leq \pi(\alpha, H \cup G) = F(H \cup G) < F(G),$$

которая противоречива, ибо величина  $F(G)$  не может быть строго меньше самой себя. Пусть теперь  $\alpha \in H \setminus G$  (Напомним, что, по предположению,  $H \setminus G \neq \emptyset$ ). В этом случае получаем аналогичную цепочку неравенств

$$F(H) \leq \pi(\alpha, H) \leq \pi(\alpha, H \cup G) = F(H \cup G) < F(G),$$

которая и доказывает (20), а следовательно, и (18). Теорема доказана.

Таким образом, доказано, что эквивалентное определение ядра дается при помощи соотношений (22), (23). В то же время, как легко заметить, доказательство выполнения соотношений (22), (23) для некоторого множества  $G$  требует рассмотрения меньшего числа подмножеств  $H$ , чем проверка соотношений (20), (24) или (18), (19).

Кроме того, в дальнейшем исследовании свойств монотонных систем нам понадобится также несколько более узкий класс определяющих последовательностей, который выделяется при добавлении в данное выше (в разд. 2) ее определение еще одного условия. Кроме соотношений (5) и (6).

*Определение.* Упорядоченная последовательность  $A = \langle \alpha_1, \dots, \alpha_N \rangle$  элементов множества  $W = \{\alpha_1, \dots, \alpha_N\}$  называется максимальной определяющей последовательностью, если в соответствующей ей последовательности множеств  $\bar{H}$  существует такая подпоследовательность  $\bar{\Gamma} = \langle \Gamma_1, \dots, \Gamma_p \rangle$ , где  $\Gamma_1 = H_1 = W$ , для которой, помимо свойств (5) и (6), выполняется также соотношение

$$\pi(\alpha_k, H_k) \leq F(\Gamma_j), \quad \forall \alpha_k \in \Gamma_j \setminus \Gamma_{j+1}, \quad j = \overline{1, p}. \quad (24)$$

Смысл этого определения заключается в том, что среди множеств  $H_k$ , таких, что  $\Gamma_j \supset H_k \supset \Gamma_{j+1}$ , нет ни одного, по отношению к которому выполнялось бы свойство (5), а следовательно, никакое множество  $H_k$  из последовательности  $\bar{H}$  не может пополнить последовательность  $\bar{\Gamma}$ , будучи помещенным между  $j - \mu$  и  $(j + 1) - \mu$  ее элементами, т. е. между множествами  $\Gamma_j$  и  $\Gamma_{j+1}$ . В то же время определение определяющей последовательности из разд. 2 допускало такое пополнение соответствующей последовательности  $\bar{\Gamma}$ , удовлетворяющей (5) и (6), при абсолютном сохранении самой последовательности элементов  $A$ .

В дальнейшем в качестве определяющей последовательности, как правило, используется максимальная определяющая последовательность, удовлетворяющая данному определению.

Покажем теперь, что для множеств  $\Gamma_j, j = \overline{1, p-1}$ , имеют место соотношения, аналогичные (18)–(23).

**Теорема 6<sup>а</sup>.** Для множеств  $\Gamma_j$  последовательности  $\bar{\Gamma}$  имеют место следующие соотношения:

$$F(H) < F(\Gamma_j), \quad \forall H \subseteq W, H \setminus \Gamma_j \neq \emptyset, j = \overline{2, p}, \quad (25)$$

$$F(H) \leq F(\Gamma_j), \quad \forall H \subseteq \Gamma_j, H \supset \Gamma_{j+1}, j = \overline{1, p-1}. \quad (26)$$

Легко видеть аналогию между соотношениями (20), (21), с одной стороны, и (25), (26) — с другой. Для того чтобы получить вполне единообразные выражения, примем следующие дополнительные обозначения:  $\Gamma_0 = \Gamma_1 = W$  и  $\Gamma_{p+1} = \emptyset$  ( $\Gamma_p = G$  — последнее непустое множество последовательности  $\bar{\Gamma}$ ). Тогда соотношения (20), (21) и (25), (26) можно объединить следующим образом:

$$F(H) < F(\Gamma_j), \quad \forall H \subseteq W, H \setminus \Gamma_j \neq \emptyset, j = \overline{1, p}, \quad (27)$$

$$F(H) \leq F(\Gamma_j), \quad \forall H \subseteq \Gamma_j, H \supset \Gamma_{j+1}, j = \overline{1, p}. \quad (28)$$

Таким образом, отличие ядра  $G = \Gamma_p$  от других множеств  $\Gamma_j \neq \Gamma_p, j = \overline{1, p-1}$ , последовательности  $\bar{\Gamma}$  заключается в том, что оно (т. е. ядро  $G = \Gamma_p$ ) имеет в качестве непосредственно следующего множества  $\Gamma_{p+1}$  в последовательности  $\bar{\Gamma}$  пустое множество. Кроме того, как это следует из леммы 1,  $F(G) > F(\Gamma_j), j = \overline{1, p-1}$ . В остальном экстремальные свойства ядра  $G = \Gamma_p$  и любого другого множества из последовательности  $\bar{\Gamma}$  совпадают.

**Следствие.** Для любого подмножества  $H$  множества  $W$ , такого, что  $\Gamma_j \supset H \supset \Gamma_{j+1}$ , для некоторого  $j = \overline{1, p-1}$  справедливо

$$\left. \begin{aligned} F(H) &\leq F(\Gamma_j) = \varepsilon_j, \\ F(H) &< F(\Gamma_{j+1}) = \varepsilon_{j+1}, \end{aligned} \right\} \Gamma_j \supset H \supset \Gamma_{j+1}, j = \overline{1, p-1}.$$

**Доказательство теоремы 6<sup>а</sup>.** Докажем выполнение соотношения (25). Предположим противное. Рассмотрим такое множество  $H \subseteq W$ , что  $F(H) \geq F(\Gamma_j)$ , (29)

но  $H \not\subseteq \Gamma_j$ . Последнее, очевидно, означает, что  $H \setminus \Gamma_j \neq \emptyset$ .

Пусть  $\alpha_n$  — первый элемент множества  $H$  в определяющей последовательности, т. е.  $\alpha_n \in H, H_n \supseteq H, H_{n+1} \not\supseteq H, H_n \supset \Gamma_j$ , т. е.  $H_n \supseteq H \cup \Gamma_j$ .

Тогда, используя свойство монотонности и первое свойство (5) последовательности  $\bar{\Gamma}$ , по определению, получаем цепочку неравенств

$$F(H) \leq \mu(\alpha_n, H) \leq \pi(\alpha_n, H_n) < F(\Gamma_j),$$

которая противоречит (29). Первое соотношение доказано <sup>4</sup>.

Выполнение соотношения (26) доказывается аналогичным образом, только вместо (29) необходимо использовать строгое неравен-

<sup>4</sup> Доказательство первого утверждения теоремы 6, следовательно, опирается только на соотношение (5) общего определения определяющей последовательности из разд. 2.

ство  $F(H) > F(\Gamma_j)$ , где  $H \not\subseteq \Gamma_{j+1}$ ,  $H \subseteq \Gamma_j$ ; а вместо (5) — соотношение (24). Теорема доказана.

Утверждение теоремы 6 становятся особенно наглядными, если ввести следующие «геометрические» представления.

*Определение.* Интервалом  $[D, E]$ , где  $D \supset E$  на множестве  $X$  всех подмножеств множества  $W$ ,  $X = \{B, B \subseteq W\}$ , называется семейство всех таких его элементов  $T$ , ( $T \subseteq X$ ), которое удовлетворяет соотношению

$$[D, E] = \{T, D \supseteq T \supseteq E\}. \quad (30)$$

Если в правой или левой части соотношения  $D \supseteq T \supseteq E$  из (30) имеет место строгое включение, то это отражается заменой квадратной скобки на круглую в обозначении интервала:

$$[D, E) = \{T, D \supseteq T \supset E\},$$

$$(D, E] = \{T, D \supset T \supseteq E\},$$

$$(D, E) = \{T, D \supset T \supset E\}.$$

*Определение:* Множество  $L$ ,  $L \subseteq W$ , называется максимумом функции  $F(H)$  на интервале  $[D, E]$ , включающем это множество, т. е.  $L \in [D, E]$ ,  $D \supset E$ , если оно удовлетворяет соотношению

$$F(L) \geq F(T), \quad \forall T \in [D, E]. \quad (31)$$

Аналогично определяется экстремум на интервалах с открытыми концами  $[D, E)$ ,  $(D, E]$ ,  $(D, E)$ .

Максимум называется строгим, если в (31) имеет место строгое неравенство.

Используя эти определения, теорему 6<sup>\*</sup> можно переформулировать следующим образом.

*Теорема 7<sup>\*</sup>.* Каждое множество  $\Gamma_j$ ,  $j = \overline{1, p}$ , последовательно сти  $\bar{\Gamma}$  является максимумом функции  $F(H)$  на интервале  $[\Gamma_1 = W, \Gamma_{j+1})$ .

Все множества  $H$ ,  $H \subseteq W$ , на которых достигается значение  $F(H)$ , равное величине  $F(\Gamma_j)$  для некоторого  $j = \overline{1, p}$ , лежат внутри множества  $\Gamma_j$ . Все множества  $H$ ,  $H \subseteq W$ , на которых достигается значение  $F(H)$ , превышающее величину  $F(\Gamma_j)$  для некоторого  $j = \overline{1, p-1}$ , лежат внутри множества  $\Gamma_{j+1}$ .

*Доказательство* полностью совпадает с доказательством теоремы 6. Действительно, соотношения (25), (26) как раз и означают, что множество  $\Gamma_j$ ,  $j = \overline{1, p}$ , есть максимум функции  $F(H)$  на интервале  $[W, \Gamma_{j+1})$ , причем множество  $\Gamma_j$ ,  $j = \overline{2, p}$ , является строгим левосторонним максимумом на интервале  $[W, \Gamma_j]$  и нестрогим правосторонним — на интервале  $[\Gamma_j, \Gamma_{j+1})$ .

Для полноты описания экстремальных свойств множеств  $\Gamma_j$  введем еще два понятия.

*Определение.* Окрестностью множества  $H$  называется совокупность всех множеств  $L$ , таких, что

$$H \subset L \subseteq W.$$

Иными словами, окрестностью множества  $H$  является интервал  $[W, H)$ .

**Определение.** Множество  $H^0$  называется строгим локальным максимумом функции  $F(H)$ , если оно доставляет строгий максимум этой функции в своей окрестности.

**Следствие теоремы 7.** Множество  $H^0$  тогда и только тогда является строгим локальным максимумом функции  $F(H)$ , когда  $H^0 = \Gamma_j$ , где  $\Gamma_j$  множество из последовательности  $\bar{\Gamma}$ . Итак, в теореме 7 установлено, что функция  $F(H)$  имеет весьма простую структуру: у нее имеется всего  $p \leq N = |W|$  строго возрастающих (см. лемму 1) строгих локальных максимумов. Последний из этих максимумов является в то же время глобальным. Именно эта простота определяет возможность построения быстрого алгоритма, решения задачи выделения ядра: необходимо только построить процедуру последовательного перебора всех локальных максимумов функции  $F(H)$ .

Аналогия между ядром  $G = \Gamma_p$  и другими множествами  $\Gamma_j$  последовательности  $\bar{\Gamma}$  прослеживается также и на теоретико-множественном уровне. Вот как это устанавливается.

**Теорема 8.** Система всех тех подмножеств множества  $W$ , на которых функция  $F$  достигает или превышает значение  $F(\Gamma_j)$  для некоторого  $j = \bar{1}, p$ , замкнута по отношению к бинарной операции объединения множеств.

**Доказательство.** Пусть  $E_1$  и  $E_2$  — два различных множества из  $W$ , такие, что

$$F(E_1) \geq F(\Gamma_j), \quad F(E_2) \geq F(\Gamma_j) \quad (32)$$

для некоторого  $j = \bar{1}, p$ . В соответствии с теоремой 7 заключаем:  $E_1, E_2 \subseteq \Gamma_j$ . Нужно доказать, что для множества  $E_1 \cup E_2$  справедливо неравенство

$$F(E_1 \cup E_2) \geq F(\Gamma_j) \quad (33)$$

(легко видеть, что  $E_1 \cup E_2 \subseteq \Gamma_j$ ).

Пусть  $\alpha, \alpha \in E_1 \cup E_2$  — такой элемент, на котором достигается значение функции  $F(E_1 \cup E_2)$ :

$$\pi(\alpha, E_1 \cup E_2) = F(E_1 \cup E_2).$$

Пусть, например,  $\alpha \in E_1$ . Тогда по свойству монотонности имеем

$$\pi(\alpha, E_1) \leq \pi(\alpha, E_1 \cup E_2), \quad (34)$$

и, следовательно,

$$F(E_1) = \min_{i \in E_1} \pi(i, E_1) \leq \pi(\alpha, E_1) \leq \pi(\alpha, E_1 \cup E_2) = F(E_1 \cup E_2). \quad (34)$$

Объединяя (32) и (34), получаем (33). Теорема доказана <sup>5</sup>.

<sup>5</sup> Легко проверить, что теорема 8 остается в силе, если использовать общее определение определяющей последовательности, приведенное в разд. 1.

Сопоставляя с содержательной точки зрения формулировки теорем 7 и 8 с теоремами 1 и 2, легко заметить их почти полное совпадение. Не хватает лишь утверждения о единственности набора множеств  $\Gamma_j$ , последовательности  $\bar{\Gamma}$ , которое дает следующая теорема.

**Теорема 9'.** Если  $A$  и  $A'$  есть две различные максимальные определяющие последовательности одной и той же  $\ominus$ -монотонной системы  $\langle W, \pi \rangle$ , то соответствующие им последовательности  $\bar{\Gamma}$  и  $\bar{\Gamma}'$  совпадают:

$$p = p', \Gamma_j = \Gamma'_j, \forall j = \overline{1, p}.$$

**Доказательство.** Предположим противное. Пусть  $\Gamma'_S$  — первое множество в последовательности  $\bar{\Gamma}'$ , такое, что  $\Gamma_S \neq \Gamma'_S$ , но  $\Gamma_j = \Gamma'_j, \forall j = \overline{1, S-1}$ . (Напомним, что  $\Gamma_1 = \Gamma'_1 = W$ , по определению.)

Возможны два случая:  $\Gamma_S \subset \Gamma'_S$  или  $\Gamma'_S \setminus \Gamma_S \neq \emptyset$ . Рассмотрим их последовательно.

Первый случай. В соответствии со следствием теоремы 3 для множества  $H = \Gamma_S$ , такого, что  $\Gamma_{S-1} = \Gamma'_{S-1} \supset H \supset \Gamma'_S$ , имеем

$$F(\Gamma_S) \leq F(\Gamma_{S-1}).$$

Это неравенство противоречит соотношению (9) для последовательности  $\bar{\Gamma}$  (лемма 1).

Второй случай,  $(\Gamma'_S \setminus \Gamma_S \neq \emptyset)$ . Пусть  $H_n$  — минимальное множество последовательности  $\bar{H}$ , такое, что  $H_n \supset \Gamma'_S$ . Очевидно,  $H_n \subseteq \Gamma_{S-1}$ ,  $\alpha_n \in \Gamma'_S$ , где  $\alpha_n \in H_n$ ,  $\alpha_n \notin H_{n+1}$ .

Поскольку  $\Gamma_{S-1} = \Gamma'_{S-1}$  по предположению, то

$$F(\Gamma_{S-1}) = F(\Gamma'_{S-1}) < F(\Gamma'_S).$$

С другой стороны, учитывая свойство монотонности для множеств  $\Gamma'_S$  и  $H_n$ ,  $\Gamma'_S \subset H_n$ , определение функции  $F(H)$  и следствие теоремы 6, получаем

$$F(\Gamma'_S) \leq \pi(\alpha_n, \Gamma'_S) \leq \pi(\alpha_n, H_n) \leq F(\Gamma_{S-1}).$$

Снова пришли к противоречию с леммой 1. Теорема доказана.

В соответствии с теоремой 6 последовательность вложенных множеств  $\bar{\Gamma}$  может рассматриваться как характеристика структуры монотонной системы, инвариантная по отношению к преобразованиям ее определяющих последовательностей друг в друга.

Рассмотрим еще раз максимальную определяющую последовательность и связанную с ней последовательность множеств  $\bar{\Gamma} = \langle \Gamma_1, \dots, \Gamma_p \rangle$ . Следует отметить, что множества  $\Gamma_{j-1} \setminus \Gamma_j$ ,  $j = \overline{2, p}$ , также обладают некоторыми экстремальными свойствами. А именно, помимо определенной на  $W$  монотонной системы  $\langle W, \pi \rangle$  введем на  $W \setminus \Gamma_j$  монотонную систему  $\langle W \setminus \Gamma_j, \pi' \rangle$ ,

где

$$\pi'(i, H) = \pi(i, H \cup \Gamma_j), \quad \forall H \subseteq (W \setminus \Gamma_j), \quad i \in H.$$

**Теорема 10<sup>а</sup>.** Ядром системы  $\langle W \setminus \Gamma_j, \pi' \rangle$  является множество  $G' = \Gamma_{j-1} \setminus \Gamma_j$ .

*Доказательство.*

Требуется доказать, что

$$F'(H) < F'(G'), \quad \forall H \subseteq (W \setminus \Gamma_j), \quad H \setminus G' \neq \emptyset.$$

$$F'(H) \leq F'(G'), \quad \forall H \subseteq (W \setminus \Gamma_j), \quad H \subseteq G'.$$

Докажем первое соотношение

$$F'(H) = \min_{i \in H} \pi'(i, H) = \min_{i \in H} \pi(i, H \cup \Gamma_j).$$

В силу того, что

$$\min_{i \in \Gamma_j} \pi(i, \Gamma_j) > \min_{i \in H \cup \Gamma_j} \pi(i, H \cup \Gamma_j), \quad \forall H \subseteq W \setminus \Gamma_j, \quad H \not\subseteq G',$$

и монотонности функции  $\pi(i, H)$  верно следующее соотношение:

$$\min_{i \in H} \pi(i, H \cup \Gamma_j) = \min_{i \in H \cup \Gamma_j} \pi(i, H \cup \Gamma_j), \quad \forall H \subseteq W \setminus \Gamma_j, \quad H \not\subseteq G'.$$

Это означает, что

$$\begin{aligned} F'(H) = F(H \cup \Gamma_j) &< F(\Gamma_{j-1}) = F(G' \cup \Gamma_j) = \\ &= \min_{i \in G' \cup \Gamma_j} \pi(i, G' \cup \Gamma_j) = \min_{i \in G'} \pi(i, G' \cup \Gamma_j) = F'(G'). \end{aligned}$$

Докажем теперь второе соотношение.

С помощью рассуждений, аналогичных тем, которые проводились при доказательстве первого соотношения, получаем

$$F'(H) = F(H \cup \Gamma_j) \leq F(\Gamma_{j-1}) = F'(G').$$

Теорема доказана.

В заключение сравнительного исследования свойств ядра  $G = \Gamma_p$  и других элементов  $\Gamma_j$  последовательности  $\bar{\Gamma}$  сформулируем еще два специальных свойства этих множеств.

**Лемма 2<sup>а</sup>.** Для любого множества  $\Gamma_j$ ,  $j = \overline{2, p}$ , последовательности  $\bar{\Gamma}$  справедливо следующее соотношение:

$$\max_{\alpha_n \in W \setminus \Gamma_j} \pi(\alpha_n, H_n) \leq \max_{\alpha_k \in \Gamma_{j-1} \setminus \Gamma_j} \pi(\alpha_k, H_k). \quad (35)$$

*Доказательство.* Поскольку очевидно, что  $W \setminus \Gamma_j \supset \Gamma_{j-1} \setminus \Gamma_j$ , то, следовательно, нужно доказать, что

$$\max_{\alpha_n \in W \setminus \Gamma_{j-1}} \pi(\alpha_n, H_n) \leq \max_{\alpha_k \in \Gamma_{j-1} \setminus \Gamma_j} \pi(\alpha_k, H_k).$$

Действительно, заметим, что правая часть этого неравенства не меньше величины  $F(\Gamma_{j-1})$ , так как в множестве  $\Gamma_{j-1}$  всегда

существует такой элемент  $\gamma_{j-1}$ ,  $\gamma_{j-1} \in \Gamma_{j-1} \setminus \Gamma_j$ , что

$$\pi(\gamma_{j-1}, \Gamma_{j-1}) = F(\Gamma_{j-1}).$$

Таким образом, имеем

$$\max_{\alpha_k \in \Gamma_{j-1} \setminus \Gamma_j} \pi(\alpha_k, H_k) \geq F(\Gamma_{j-1}).$$

С другой стороны, используя свойство множеств  $\Gamma_j$  последовательности  $\bar{\Gamma}$  по определению и лемму 1, для множества  $\Gamma_{j-1}$  и  $H_n$ ,  $H_n \supset \Gamma_{j-1}$ , получаем

$$\pi(\alpha_n, H_n) < F(\Gamma_{j-1}), \quad \forall \alpha_n \in W \setminus \Gamma_{j-1} \quad (\forall H_n \supset \Gamma_{j-1}).$$

Объединяя полученные два неравенства, приходим к утверждению леммы. Лемма доказана <sup>6</sup>.

В соответствии с леммой 1, в частности, можно записать

$$\max_{\alpha_n \in W \setminus G} \pi(\alpha_n, H_n) \leq \max_{\alpha_n \in \Gamma_{p-1} \setminus G} \pi(\alpha_n, H_n), \quad (36)$$

где  $G = \Gamma_p$  — ядро  $\ominus$ -монотонной системы, а  $\Gamma_{p-1}$  — непосредственно предшествующее ему множество в последовательности  $\bar{\Gamma}$ .

*Лемма 3.* Вес любого элемента  $k$ ,  $k \in W$ , не принадлежащего множеству  $\Gamma_j$ , для некоторого  $j = \overline{2, p}$ ,  $k \in W \setminus \Gamma_j$ , получаемый путем произвольного присоединения его к этому множеству, не превышает значения функции  $F(\Gamma_j)$ , т. е.

$$\pi(k, \Gamma_j \cup k) < F(\Gamma_j), \quad \forall k \in W \setminus \Gamma_j, \quad \forall j = \overline{2, p}. \quad (37)$$

*Доказательство.* Предположим противное. Пусть существует такой элемент  $k \in W \setminus \Gamma_j$ , когда

$$\pi(k, \Gamma_j \cup k) \geq F(\Gamma_j).$$

С другой стороны, используя свойство монотонности для множества  $\Gamma_j \cup k$ , получаем

$$\pi(i, \Gamma_j \cup k) \geq \pi(i, \Gamma_j) \geq F(\Gamma_j), \quad \forall i \in \Gamma_j.$$

Объединяя эти два неравенства, приходим к отношению

$$F(\Gamma_j \cup k) = \min_{i \in \Gamma_j \cup k} \pi(i, \Gamma_j \cup k) \geq F(\Gamma_j),$$

которое противоречит соотношению (25) теоремы 6<sup>-</sup>. Лемма доказана.

В соответствии с леммой 2, в частности, можно записать

$$\pi(k, G \cup k) < F(G), \quad \forall k \in W \setminus G. \quad (38)$$

На этом заканчивается сравнительное исследование свойств ядра  $G = \Gamma_p$  и других элементов  $\Gamma_j$  последовательности  $\bar{\Gamma}$ .

<sup>6</sup> Лемма 2 справедлива для определяющей последовательности, удовлетворяющей общему определению из разд. 2. То же относится и к лемме 3.

Перейдем теперь к явному описанию задач структуризации, в которых наряду с ядром  $G$  целесообразно использовать другие множества последовательности  $\bar{G}$ .

В связи с тем, что теория монотонных систем и, в частности, задача выделения экстремальной подсистемы монотонной системы рассматриваются в этой работе как новый подход к структуризации данных, большой интерес представляет исследование возможности решения исходной задачи с дополнительными априорными ограничениями. Ниже рассматриваются два типа таких ограничений.

Первый тип. В условиях задачи структуризации указан допустимый (требуемый) размер искомого экстремального подмножества или желательный диапазон для его размера.

Второй тип. В условиях задачи структуризации задан перечень особых элементов, которые должны обязательно войти в искомое экстремальное подмножество.

В практических задачах структуризации встречаются и оба этих типа ограничений одновременно. В любом случае они связаны с априорными содержательными соображениями, основанными на дополнительной информации, явным образом не отраженной в исходных данных.

Рассмотрим задачу выделения экстремальной подсистемы монотонной системы в случае первого типа дополнительных ограничений.

Пусть  $\langle W, \pi \rangle$  —  $\ominus$ -монотонная система. Зафиксируем число  $n$ ,  $n < N$  ( $n$  — натуральное). Рассмотрим следующую задачу: найти такое подмножество  $G_n$  множества  $W$ , которое доставляло бы максимум функции  $F(H)$  среди всех подмножеств  $W$  с числом элементов, большим  $n$ :

$$F(G_n) = \max_H F(H), \quad H \subseteq W, \quad |H| > n. \quad (39)$$

Решение этой задачи дает следующая теорема.

*Теорема II'.*

а) Для любого множества  $\Gamma_j$ ,  $j = \overline{1, p}$ , последовательности  $\bar{G}$  и любого множества  $H$ ,  $H \subseteq W$ , такого, что  $|H| > |\Gamma_j|$ , справедливо строгое неравенство

$$F(H) < F(\Gamma_j).$$

б) Для любого множества  $\Gamma_j$ ,  $j = \overline{1, p-1}$ , последовательности  $\bar{G}$  и любого множества  $H$ ,  $H \subseteq W$ , такого, что  $|\Gamma_j| \geq |H| > |\Gamma_{j+1}|$ , справедливо неравенство

$$F(H) \leq F(\Gamma_j).$$

Рассмотрим теперь задачу выделения экстремальной подсистемы монотонной системы в случае второго типа ограничений.

Пусть снова  $\langle W, \pi \rangle$  —  $\ominus$ -монотонная система. Зафиксируем некоторое множество  $T$  элементов множества  $W$ ,  $T \subseteq W$ . Рас-

смотрим следующую задачу: найти такое подмножество  $G_T$  множества  $W$ , которое доставляло бы максимум функции  $F(H)$  среди всех подмножеств  $W$ , содержащих множество  $T$ :

$$F(G_T) = \max_{T \subseteq H \subseteq W} F(H). \quad (40)$$

Элементы множества  $T$  в этом случае естественно назвать ключевыми элементами. Иначе говоря, задача заключается в том, чтобы выделить экстремальную подсистему монотонной системы, содержащую заданные ключевые элементы.

Решение сводится к следующему.

*Теорема 12.* Для любого множества  $T$ ,  $T \subseteq W$ , и такого множества  $\Gamma_S$  последовательности  $\bar{\Gamma}$ , что  $T \subseteq \Gamma_S$ , но  $T \not\subseteq \Gamma_{S+1}$ , справедливы такие соотношения:

$$F(H) < F(\Gamma_S), \quad \forall H \subseteq W, \quad T \subseteq H, \quad H \setminus \Gamma_S \neq \emptyset, \quad (41)$$

$$F(H) \leq F(\Gamma_S), \quad \forall H \subseteq W, \quad T \subseteq H, \quad H \subseteq \Gamma_S. \quad (42)$$

*Доказательство.* Теорем 11 и 12 непосредственно следует из теоремы 6, и поэтому их можно рассматривать как следствия этой теоремы.

Рассмотренные свойства монотонной системы позволяют решать исходную задачу выделения экстремальной подсистемы с дополнительными ограничениями в виде заданного множества ключевых элементов, обязательных для включения в искомое подмножество по каким-либо внешним соображениям, или в виде заданного желательного размера (диапазона) искомого подмножества. В обоих случаях, как это следует из теорем 11 и 12, естественным решением такой задачи с ограничениями будет наименьшее множество  $\Gamma_j$ , удовлетворяющее соответствующему дополнительному условию, например наименьшее множество  $\Gamma_j$ , включающее заданные ключевые элементы. Тогда сравнение  $F(\Gamma_j)$  с  $F(G)$  показывает, насколько искажается идеальное решение введением дополнительных ограничений.

Напомним, что в разд. 2 определяющая последовательность элементов монотонной системы вводилась как некоторая вспомогательная конструкция, предназначенная для определения ядра, т. е. выделения такого наибольшего подмножества  $G$ ,  $G \subseteq W$ , на котором функция  $F(H)$ , определенная в (3), достигает глобального максимума. В настоящем разделе, таким образом, показано, что построение максимальной определяющей последовательности означает не только решение задачи поиска ядра, но и существенно более глубокое вскрытие структуры монотонной системы. А именно, оно позволяет выявить ряд особых подмножеств  $\Gamma_j, j = \overline{1, p}$  исходного множества, специальные свойства которых аналогичны свойствам ядра.

Учитывая вышесказанное, по аналогии с определением последнего множества  $\Gamma_p = G$  последовательности  $\bar{\Gamma}$  как ядра естественно назвать остальные множества этой последовательности квази-ядрами, а саму последовательность  $\bar{\Gamma}$  — последовательностью квазиядер.

#### 4. Алгоритмы выделения ядра монотонной системы

При определении понятий и доказательстве теорем, приведенных выше, всюду предполагается, что определяющая последовательность элементов монотонной системы, а следовательно, и ядро существуют. Для доказательства достаточно предложить конструктивную процедуру построения такой последовательности.

Как уже отмечалось, может быть много различных определяющих последовательностей одной и той же монотонной системы. Все они дают одно и то же ядро  $G$ , а все максимальные определяющие последовательности в соответствии с теоремой 9 имеют одну и ту же последовательность квазиядер  $\bar{G}$ .

Рассмотрим три разных алгоритма выделения ядра монотонной системы. Один из них строит определяющую последовательность общего вида, соответствующую определению из разд. 2, второй — максимальную определяющую последовательность, введенную в разд. 3, а третий — некоторую специальную определяющую последовательность.

Прежде чем приступить к непосредственному описанию алгоритмов, рассмотрим одну итеративную процедуру, используемую в первом и втором алгоритмах [6].

*Процедура СЛОИ ( $u$ ).* Входной информацией для нее является произвольное множество  $H$  с заданной на ней функцией  $\pi(i, H)$  и скалярная величина  $u$ , называемая в дальнейшем порогом. Предварительно вычисляются все значения  $\pi(i, H)$ ,  $i \in H$ , если они ранее не были определены.

Каждая итерация этой процедуры состоит из двух действий.

Первое действие заключается в том, что путем последовательных сравнений значений  $\pi(i, H)$  с величиной порога  $u$  определяется множество элементов множества  $H$ , для которых выполняется неравенство

$$\pi(i, H) \leq u.$$

Эти элементы в произвольном порядке включаются в формируемую последовательность элементов множества  $W$  (при этом они исключаются из множества  $H$ ).

Второе действие заключается в том, что для всех элементов множества  $H'$ ,  $H' \subseteq H$ , оставшихся после первого действия, пересчитываются значения весовой функции, т. е. вычисляются значения  $\pi(i, H')$ ,  $\forall i \in H'$ , вместо  $\pi(i, H)$ ,  $\forall i \in H$ .

После этого приступают к первому действию следующей итерации: отбираются элементы множества  $H'$ , удовлетворяющие неравенству

$$\pi(i, H') \leq u,$$

и в произвольном порядке включаются в формируемую последовательность элементов множества  $W$  и т. д.

Последняя итерация процедуры определяется тем, что среди оставшихся элементов множества  $H$  нет ни одного с весом, меньшим или равным  $u$ . Результатом процедуры представляется некоторое множество  $E$ ,  $E \subseteq H$ , оставшихся элементов с вычисленными значениями  $\pi(i, E)$ .  $\forall i \in E$ , каждое из которых больше заданной величины  $u$ , и участок формируемой последовательности элементов множества  $H \setminus E$  в виде нескольких последовательных групп элементов, произвольно расставленных внутри каждой группы. В частном, но весьма важном случае результатом процедуры СЛОЙ ( $u$ ) является пустое множество  $E = \emptyset$ , а все элементы множества  $H$  образуют участок формируемой последовательности.

*Лемма 4.* Пусть  $u_1 \geq u_2$ ,  $H_2$  — результат процедуры СЛОЙ ( $u_2$ ) на множестве  $W$ ,  $H_1$  — результат процедуры СЛОЙ ( $u_1$ ) на множестве  $W$ .  $H_3$  — результат процедуры СЛОЙ ( $u_1$ ) на множестве  $H_2$ . Тогда  $H_3 = H_1$ .

*Доказательство.* Докажем сначала, что

$$1. H_3 \subseteq H_1.$$

В силу монотонности функции  $\pi(i, H)$  и того, что  $H_3$  — результат процедуры СЛОЙ ( $u_1$ ), выполняется соотношение

$$\pi(i, W) \geq \pi(i, H_3) > u_1, \quad \forall i \in H_3.$$

Это означает, что после выполнения первой итерации процедуры СЛОЙ ( $u_1$ ) образуется множество  $H' \equiv H_3$  и выполняется соотношение

$$\pi(i, H') \geq \pi(i, H_3) > u_1, \quad \forall i \in H_3.$$

Рассуждая далее таким образом, приходим к выводу, что элементы множества  $H_3$  не могут быть выкинуты ни на одном шаге процедуры СЛОЙ ( $u_1$ ) на множестве  $W$ . Значит,  $H_3 \subseteq H_1$ .

Теперь докажем, что

$$2. H_1 \subseteq H_3.$$

Очевидно выполнение следующего соотношения:

$$\pi(i, W) \geq \pi(i, H_1) > u_1 \geq u_2, \quad \forall i \in H_1.$$

То есть на первой итерации процедуры СЛОЙ ( $u_2$ ) на множестве  $W$  не могут быть выкинуты элементы множества  $H_1$ . Проводя далее рассуждения, аналогичные тем, которые проводились при доказательстве первого пункта, получаем, что  $H_1 \subseteq H_3$ .

Итак, выполнение первого и второго пунктов означает

$$H_3 = H_1,$$

что и требовалось доказать.

*Лемма 5.* Пусть  $u_1 > u_2$ .  $H_1$  — результат процедуры СЛОЙ ( $u_1$ ),  $H_2$  — результат процедуры СЛОЙ ( $u_2$ ) на том же множестве  $H$ . Если

$$F(H_2) > u_1, \quad \text{то} \quad H_1 = H_2.$$

Если

$$H_2 = \emptyset, \text{ то } H_1 = \emptyset.$$

*Доказательство.* Легко видеть, что так как  $F(H_2) > u_1$ , то элементы множества  $H_2$  не могли быть выкинуты в результате процедуры СЛОЙ ( $u_1$ ). Значит,  $H_2 \subseteq H_1$ . Но  $F(H_1) > u_2$ , а значит,  $H_1 \subseteq H_2$ . То есть  $H_1 = H_2$ , что и требовалось доказать. Второе утверждение леммы доказывается с помощью аналогичных рассуждений.

*Алгоритм А1* построения общей определяющей последовательности (см. разд. 1). Полагаем  $\Gamma_1 = H = W$ . Вычисляются значения  $\pi(i, W)$  для всех элементов множества  $W$  и определяются две величины

$$u' = \min_{i \in W} \pi(i, W), \quad u'' = \max_{i \in W} \pi(i, W).$$

Фиксируется произвольное значение  $u$  между этими величинами, т. е.  $u : u' < u < u''$  в простейшем случае, следующим образом:

$$u = (u' + u'')/2.$$

Эта же формула используется и для пересчета значения  $u$  после того, как в ходе выполнения алгоритма изменяются величины  $u'$  и  $u''$ .

*Произвольный  $j$ -й шаг.* Запускается процедура СЛОЙ ( $u$ ) на множестве  $H$  (на первом шаге  $H = W$ ). В результате ее выполнения возможны два варианта.

1. Множество  $E$  — пустое. В этом случае, который означает, что порог слишком велик, значение  $u$  присваивается величине  $u''$ , т. е.  $u'' = u$ , снова определяется  $u$  и снова запускается процедура СЛОЙ ( $u$ ). При этом элементы, включенные в формируемую последовательность в процессе предыдущего выполнения этой процедуры, «изымаются» из нее и снова возвращаются в текущее подмножество  $H$ . В результате снова может быть два варианта, которые рассматриваются в соответствии с данным описанием.

2. Множество  $E$  — не пустое. В этом случае включенные в формируемую последовательность элементы фиксируются в ней, а множество  $E$  объявляется очередным множеством последовательности  $\bar{\Gamma}$ , т. е.  $\Gamma_{j+1} = E$ . На этом множестве вычисляется величина  $F(\Gamma_{j+1})$ . Заметим, что по построению  $F(\Gamma_{j+1}) > u$ , и запускается процедура СЛОЙ ( $F(\Gamma_{j+1})$ ) на множестве  $H := \Gamma_{j+1}$ .

В результате в случае 1 множество  $H = \Gamma_{j+1}$  есть ядро  $\Gamma_p$ , т. е.  $p = j + 1$ . В случае 2, когда получаемое новое множество  $E$ ,  $E \subseteq H$ , не пусто ( $E \neq \emptyset$ ), переходят к следующему ( $j + 1$ )-му шагу с множеством  $E$ , т. е. значение  $F(H)$  присваивается величине  $u'$ , снова определяется  $u$  и снова запускается процедура СЛОЙ ( $u$ ) на последнем множестве  $E$ .

*Теорема 13.* Множество  $\Gamma_p$ , получаемое в результате выполнения алгоритма А1, есть определенное множество.

*Доказательство.* Докажем сначала, что множества  $\Gamma_j$ ,  $j = \overline{1, p}$ , получающиеся в результате выполнения алгоритма, образуют соответствующие множества из определяющей последовательности. Для этого достаточно показать, что любое из этих множеств будет локальным максимумом функции  $F(H)$  (см. следствие теоремы 7). Итак:

$$F(H) < F(\Gamma_j), \quad \forall H \supset \Gamma_j, \quad j = \overline{1, p}.$$

Множество  $\Gamma_j$  есть результат применения процедуры СЛОЙ ( $u$ ) к  $F(\Gamma_j) > u$ . Пусть  $F(H) > F(\Gamma_j)$  для некоторого  $H \supset \Gamma_j$ . Тогда  $F(H) > u$ , что в свою очередь означает выполнение соотношения

$$\pi(i, H) > u, \quad \forall i \in H,$$

так как

$$F(H) = \min_{i \in H} \pi(i, H).$$

В силу монотонности функции  $\pi(i, H)$  можно записать

$$\pi(i, W) \geq \pi(i, H) > u, \quad \forall i \in H.$$

Это означает, что в ходе выполнения первой итерации процедуры из СЛОЯ ( $u$ ) не мог быть исключен ни один элемент множества  $H$ . На оставшемся множестве  $H'$ ,  $W \supseteq H' \supseteq H$ , также выполняется соотношение

$$\pi(i, H') \geq \pi(i, H) > u, \quad \forall i \in H.$$

Рассуждая таким же образом, можно убедиться, что результатом процедуры СЛОЕМ ( $u$ ) может являться лишь множество  $H^0 \supseteq H$ . Но, как нам известно, результатом этой процедуры является множество  $\Gamma_j < H$ . Полученное противоречие показывает, что

$$F(H) < u,$$

т. е.

$$F(H) < F(\Gamma_j)$$

что и требовалось доказать.

Доказательство того, что  $\Gamma_R = G \not\subseteq \Gamma_p$ , где  $G$  — определимое множество, очевидным образом следует из определения ядра, а также из описания  $j$ -го шага Алгоритма А1 (второй вариант).

Алгоритм А2 построения максимальной определяющей последовательности (см. разд. 3) состоит из  $p$ ,  $p \leq N = |W|$ , шагов (параметр  $p$  не фиксирован заранее; он определяется автоматически в ходе выполнения алгоритма). Каждый шаг состоит из двух этапов.

Первый этап. Множество оставшихся к  $j$ -му шагу (еще не включенных в формируемую последовательность) элементов объявляется очередным множеством  $\Gamma_{j+1}$ , в частности  $\Gamma_1 = W$ .

Вычисленное на предыдущем шаге значение  $F(\Gamma_j) = \min_{i \in \Gamma_j} \pi(i, \Gamma_j)$

(на первом шаге предварительно вычисляются все значения  $\pi(i, W)$  и величина  $F(W)$  принимается в качестве  $j$ -го значения порога  $\varepsilon_j$ ). Следовательно, первый этап произвольного, кроме первого, шага имеет чисто декларативный характер.

Второй этап заключается в применении процедуры СЛОЙ ( $F(\Gamma_j)$ ) по отношению к множеству  $\Gamma_j$ . В результате ее выполнения возможны два варианта.

1. Множество  $E$  пустое. В этом случае множество  $\Gamma_j$  есть ядро, т. е.  $p = j$ , а алгоритм заканчивает работу.

2. Множество  $E$  не пустое. В этом случае переходят к следующему ( $j + 1$ )-му шагу алгоритма с множеством  $E$  и величиной  $F(E) = \min_{i \in E} \pi(i, E)$ .

*Теорема 14<sup>7</sup>.* 1. Результатом выполнения алгоритма является максимальная определяющая последовательность.

2. Последовательность  $\bar{\Gamma}^1$ , полученная в результате выполнения алгоритма 1 является подпоследовательностью последовательности  $\bar{\Gamma}^2$ , полученной при выполнении алгоритма А2.

*Доказательство.* 1. Тот факт, что построенная последовательность — определяющая, доказывается аналогично тому, как это доказывается для последовательности  $\bar{\Gamma}^1$  (теорема 13). Остается доказать, что в этом случае выполняется соотношение (24). Элемент  $\alpha_k \in \Gamma_j^2 \setminus \Gamma_{j+1}^2$  выкидывается на некоторой итерации процедуры СЛОЙ ( $F(\Gamma_j^2)$ ). В силу монотонности функции  $\pi(i, H)$  и так как множество  $H_k$  есть в данном случае подмножество некоторого  $\theta_l$  результата  $l$ -й итерации процедуры СЛОЙ ( $F(\Gamma_j^2)$ ), на которой элемент  $\alpha_k$  выкидывается, то

$$F(\Gamma_j^2) > \pi(\alpha_k, \theta_l) \geq \pi(\alpha_k, H_k).$$

В силу произвольности выбора множества  $\Gamma_j^2$  и элемента  $\alpha_k$  утверждение доказано.

2. Из леммы 4 результаты процедуры СЛОЙ ( $u$ ) на  $W$  и на любом множестве  $\Gamma_j$  совпадают. Множество  $\Gamma_k^1$ , таким образом, есть результат применения процедуры СЛОЙ ( $u$ ) к множеству  $W$  при некотором  $u$ . Если это  $u = F(\Gamma_j^2)$ , то очевидно, что  $\Gamma_k^1 = \Gamma_{j+1}^2$ . Если же  $u \neq F(\Gamma_j^2)$  ни при каком  $j$ , то существуют множества  $\Gamma_j^2$  и  $\Gamma_{j+1}^2$ , такие, что

$$F(\Gamma_j^2) < u, \quad F(\Gamma_{j+1}^2) > u. \quad (43)$$

Последовательное применение лемм и в этом случае дает

$$\Gamma_k^1 = \Gamma_{j+1}^2.$$

Итак, любое множество из последовательности  $\bar{\Gamma}^1$  есть в то же время и множество из последовательности  $\bar{\Gamma}^2$ , что и требовалось доказать.

Наконец, ниже описывается алгоритм построения так называемой строгой определяющей последовательности, хотя его выполнение требует несколько большего числа операций, чем выполнение алгоритмов А1 или А2, приведенных выше. Его описание, как будет видно, намного проще.

Алгоритм А3 сводится к построению такой последовательности  $I = \langle i_1, \dots, i_N \rangle$  элементов множества  $W$ , что для любого  $k$  ( $1 \leq k \leq N$ ), если  $H_k = \{i_k, i_{k+1}, \dots, i_N\}$ ,

$$\pi(i_k, H_k) = \min_{i \in H_k} \pi(i, H_k). \quad (44)$$

В случае, если минимум в (45) достигается на нескольких элементах множества  $H_k$  одновременно, то выбирается любой из них.

В качестве искомого множества  $G$  выбирается такое наибольшее множество  $H_m$ , что

$$\pi(i_m, H_m) \geq \pi(i_k, H_k), \quad \forall k = \overline{1, N}. \quad (45)$$

Для  $\ominus$ -монотонной системы в (44) вместо минимума необходимо писать максимум, а в (45) вместо знака « $\geq$ » — знак « $\leq$ ».

Иными словами, построение определяющей последовательности осуществляется путем последовательного удаления из оставшихся к  $k$ -му шагу элементов, т. е. из множества  $H_k$  элемента  $i_k$  с минимальным значением функции  $\pi(i_k, H_k)$ . При этом каждый раз пересчитываются веса остающихся элементов. Элементы в порядке их удаления образуют искомую последовательность  $I$  и соответствующую ей последовательность вложенных множеств  $\bar{H} = \langle H_1, \dots, H_N \rangle$ , где

$$H_1 = W, H_2 = H_1 \setminus i_1, \dots, H_{k+1} = H_k \setminus i_k, \dots, H_N = i_N. \quad (46)$$

Алгоритм заканчивает работу, когда все элементы множества  $W$  оказались выстроенными в последовательность  $I$ . При этом фиксируется рекордное значение  $\pi(i_m, H_m)$ . Функция  $\pi(i_k, H_k)$ , соответствующий ему элемент  $g = i_m$  и множество  $G = H_m$ :

$$\pi(i_m, H_m) = F(G) = \max_{k=\overline{1, N}} \pi(i_k, H_k). \quad (47)$$

Тот факт, что в качестве  $G$  выбирается наибольшее множество, удовлетворяющее (47), т. е. множество с максимальным значением функции  $F(H_k)$ , первым встретившееся в последовательности  $\bar{H}$ , удобно записать в виде следующих соотношений<sup>7</sup>:

$$F(H_k) < F(G), \quad \forall H_k \supset G, \quad \forall k = \overline{1, m-1}, \quad (48)$$

$$F(H_k) \leq F(G), \quad \forall H_k \subseteq G, \quad \forall k = \overline{m, N}. \quad (49)$$

<sup>7</sup> Для  $\oplus$ -монотонной системы вместо (48) и (49) необходимо воспользоваться следующими соотношениями:

$$F(H_k) > F(G), \quad \forall H_k \supset G, \quad \forall k = \overline{1, m-1},$$

$$F(H_k) \geq F(G), \quad \forall H_k \subseteq G, \quad \forall k = \overline{m, N}.$$

Рассмотрим несколько более подробно, каким образом при этом выделяется последовательность квазицедер  $\bar{\Gamma} = \langle \Gamma_j, j = \overline{1, p} \rangle$ .

Построение последовательности  $I$  элементов множества  $W$  означает также одновременное получение последовательности вложенных множеств  $\bar{H}$  и числовой последовательности порогов  $\delta = \langle \delta_1, \dots, \delta_N \rangle$ , где

$$\begin{cases} \delta_1 = \pi(i_1, W) = F(H), \\ \delta_2 = \max[\delta_{k-1}, \pi(i_k, H_k) = F(H_k)]. \end{cases} \quad (50)$$

Выражение (50) позволяет особо отметить (т. е. выделить) точки последовательности  $\bar{H}$ , соответствующие изменению (в данном случае увеличению) величины порога. Если обозначить их через  $\Gamma_j, j = \overline{1, p}$  ( $p \leq N = |W|$ ), то получим последовательность вложенных множеств  $\bar{\Gamma}$ , являющуюся подпоследовательностью последовательности множеств  $H$ , для которых (для  $\Gamma_j, j = \overline{1, p}$ ) выполняется соотношение (5) и (24). Теорема 15 показывает, что это и есть та последовательность множеств  $\bar{\Gamma}$ , существование которой предусматривается в определении максимальной определяющей последовательности.

*Теорема 15.* Построенная по описанному выше алгоритму последовательность  $I$  есть максимальная определяющая последовательность  $\ominus$ -монотонной системы  $\langle W, \pi \rangle$ .

*Доказательство.* Нужно доказать, что множества  $\Gamma_j, j = \overline{1, p}$ , полученной при построении последовательности  $I$  последовательности множеств  $\bar{\Gamma}$ , удовлетворяют свойствам (5), (24) и (6) последовательности  $\bar{\Gamma}$ , предусмотренной определением определяющей последовательности.

Выполнение свойств (5) и (24), как уже отмечалось, очевидно, поскольку множества  $\Gamma_j$  в соответствии с описанным алгоритмом фиксируются при каждом увеличении значения порога  $\delta_k$  (см. (50)).

Докажем справедливость (6). В результате выполнения алгоритма построения последовательности  $I$  получаем

$$F(H_k) \leq F(G), \forall H_k \subseteq G (\forall k = \overline{m, N}).$$

Нужно доказать, что

$$F(H) \leq F(G), \forall H \subseteq G,$$

т. е. что это неравенство выполняется для любых подмножеств множества  $G$ , не обязательно входящих в последовательность  $\bar{H}$ .

Предположим противное. Пусть для некоторого  $H, H \subseteq G$ , выполняется строгое неравенство

$$F(H) > F(G). \quad (51)$$

Пусть  $H_n$  — наименьшее множество последовательности  $\bar{H}$ , содержащее множество  $H$ , т. е.  $H_n \supseteq H, H_{n+1} \supseteq H$ . Очевидно также, что  $i_n \in H, i_n \in H_n, i_n \in H_{n+1}$ . Тогда, учитывая опре-

деление функции  $F$ , свойство монотонности и правило (44). Можно записать следующую цепочку неравенств:

$$F(H) \leq \pi(i_n, H) \leq \pi(i_n, H_n) \leq F(G).$$

которая противоречит (51). Теорема доказана.

Описанный алгоритм позволяет построить строгую определяющую последовательность. Ее характеристической особенностью является выполнение равенства (44) и выделение ядра, которое в этом случае определяется выполнением соотношений (48) и (49). Однако применение этого алгоритма требует его уточнения.

*Определение.* Элементы  $i_k$  и  $i_n$ ,  $k > n$ , определяющей последовательности  $I \ominus$ -монотонной системы назовем  $I$ -равными, если

$$\pi(i_k, H_n) = \pi(i_n, H_n) = \min_{i \in H_n} \pi(i, H_n) = F_-(H) \quad (52)$$

(а для  $\oplus$ -монотонной системы соответственно, если в (52) заменить минимум на максимум).

В соответствии с описанным выше алгоритмом построения определяющей последовательности последовательность строится с точностью до  $I$ -равных элементов. При этом заранее неизвестно, насколько будут различаться последовательности, полученные в результате выбора в качестве очередного элемента  $i_n$  того или иного элемента из группы  $I$ -равных. В то же время известно (см. теорему 9), что последовательности квазиядер двух различных определяющих последовательностей одной и той же монотонной системы совпадают, а следовательно, не зависят от выбора  $I$ -равных элементов. Этот результат подтверждается следующим образом.

*Лемма 5.* Если  $i_k$  и  $i_n$ ,  $k > n$  —  $I$ -равные элементы в последовательности  $I \ominus$ -монотонной системы  $\langle W, \pi \rangle$ , то не существует такого множества  $\Gamma_S$  в последовательности  $\bar{\Gamma}$ , что  $i_k \in \Gamma_S$ , но  $i_n \notin \Gamma_S$ . Иначе говоря, в последовательности  $\bar{\Gamma}$  существует такое множество  $\Gamma_j$ , что  $i_k, i_n \in \Gamma_j \setminus \Gamma_{j+1}$ , либо  $i_k, i_n \in \Gamma_p$ .

*Доказательство.* Предположим противное, т. е. пусть в последовательности  $\bar{\Gamma}$  есть такое множество  $\Gamma_S$ , что  $i_k \in \Gamma_S$ , но  $i_n \notin \Gamma_S$ . Тогда, очевидно,  $\Gamma_S \subset H_n$ .

Используя теорему 6 по отношению к множеству  $H_n, H_n \setminus \Gamma_S = \emptyset$ , получаем цепочку неравенств

$$\pi(i_n, H_n) < F(\Gamma_S) \leq \pi(i_k, \Gamma_S) \leq \pi(i_k, H_n).$$

В то же время, по определению  $I$ -равных элементов для  $i_k$  и  $i_n$ , имеем

$$\pi(i_n, H_n) = \pi(i_k, H_n).$$

Полученные соотношения противоречат друг другу. Лемма доказана.

Иначе говоря, выбор того или иного элемента из  $I$ -равных не влияет на построение последовательности квазиядер  $\bar{\Gamma}$  и выделе-

ния ядра  $G$ . В то же время для доказательства теорем о свойствах монотонных систем и для практического применения описанного алгоритма необходимо выбрать произвольный, но фиксированный способ однозначного построения определяющей последовательности. Установим следующее.

*Правило 1.* Если при построении определяющей последовательности  $\ominus$ -монотонной системы по описанному выше алгоритму на  $n$ -м шаге встречаются  $I$ -равные элементы, то для включения в последовательность  $I$  на этом шаге, т. е. в качестве элемента  $i_n$ , выбирается элемент с минимальным исходным номером, а для  $\ominus$ -монотонной системы — с максимальным. Предполагается, что априори все элементы множества  $W$  представлены в виде какого-то списка, т. е. занумерованы.

Последовательность  $I$ , построенную по описанному выше алгоритму при помощи этого правила, будем называть фиксированной.

### Литература

1. Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980.
2. Бородин С. М. Оптимальная группировка взаимосвязанных упорядоченных объектов. — Автоматика и телемеханика, 1980, № 2, с. 165 — 172.
3. Браверман Э. М., Киселева Н. Е., Мучник И. Б., Новиков С. Г. Лингвистический подход к задаче обработки больших массивов информации. — Автоматика и телемеханика, 1974, № 11, с. 73—88.
4. Кузнецов Е. Н., Мучник И. Б. Монотонные системы для анализа организационных структур. — В кн.: Методы анализа многомерной экономической информации. Новосибирск, 1981, с. 71—84.
5. Миркин Б. Г. Анализ качественных признаков и структур. М., 1980.
6. Мулат И. Э. Экстремальные подсистемы монотонных систем. I, II, III. — Автоматика и телемеханика, 1976, № 5, с. 130—139; 1976, № 8, с. 169—178; 1977, № 1, с. 65—83.
7. Мучник И. Б., Ослон А. А. Построение фактора, аппроксимирующего матрицу связей. — Автоматика и телемеханика, 1980, № 4, с. 89—96.
8. Учет факторов неопределенности в отраслевом планировании. Новосибирск, 1979.
9. Mullat J., Vyhandu L. Monotonic systems in scene analysis. — In: Symp. «Mathematical processing of cartographic data». Tallinn, 1979, p. 63—66.

## ОБЩИЙ ВЗГЛЯД НА СТАТИСТИКУ ОБЪЕКТОВ НЕЧИСЛОВОЙ ПРИРОДЫ

Статья — первый подробный обзор по новой области математической статистики, которую автор называет «статистикой объектов нечисловой природы». Ее развитие в основном стимулировано запросами социологии. Как самостоятельная научная область она сформирована в значительной степени работами автора. Поэтому, что статья содержит довольно много ссылок на его работы. Программа развития новой области статистики была им сформулирована в 1979 г. [30]. Здесь же основное внимание уделяется идейно-математической стороне статистики объектов нечисловой природы, а за подробными формулировками, доказательствами и обоснованием возможных приложений читатель отсылается к предыдущим публикациям.

### 1. Статистика объектов нечисловой природы как часть математической статистики

1.1. *Вероятностная статистика и анализ данных.* В этой статье термин «статистика» служит сокращением для термина «математическая статистика». Различные авторы понимают его по-разному. Существует более 200 различных определений [24]. Поэтому целесообразно повторить определение А. Н. Колмогорова и Ю. В. Прохорова, данное в Большой Советской Энциклопедии [16], которого и будем придерживаться.

«Математическая статистика — раздел математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов. При этом статистическими данными называются сведения о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками...»

В соответствии с этим определением к математической статистике относятся не только методы, основанные на вероятностных моделях. В нее входят и такие области, как кластер-анализ, многомерное шкалирование, репрезентативная теория измерений и т. д. Часть математической статистики, не использующую вероятностных представлений, принято называть анализом данных. Поясним соотношение между анализом данных и методами, использующими вероятностные модели. Пусть дана совокупность точек на плоскости. Метод наименьших квадратов позволяет построить прямую, в определенном смысле наилучшим образом приближающую эту совокупность точек. С помощью полученной прямой можно прогнозировать значение ординаты по заданному значению абсциссы. Все это позволяет создать аппарат анализа данных. Однако для оценки точности полученного прогноза необходимо обратиться к одной из вероятностных моделей регрессии.

Для построения прямой, наилучшим образом приближающей совокупность точек, может быть применен и метод наименьших

модулей [23]. С точки зрения анализа данных нет оснований предпочесть один метод другому. Такие основания может представить лишь вероятностная модель. Пусть значения абсцисс точек известны абсолютно точно, а ординат — со случайными ошибками, независимыми в совокупности и одинаково распределенными. Если распределение ошибок нормально, то метод наименьших квадратов — наилучший из возможных. Если же ошибки имеют распределение Лапласа, то наилучшим является метод наименьших модулей.

Разбор примера с регрессией позволяет проиллюстрировать два тезиса о взаимоотношении анализа данных и вероятностной статистики. Во-первых, вероятностная статистика позволяет получать более глубокие выводы, чем анализ данных, однако ее использование предполагает наличие большей информации о структуре последних. Поэтому использование анализа данных целесообразно на начальных этапах изучения реальной ситуации, а вероятностной статистики — на дальнейших, когда есть возможность обосновать предположения, лежащие в основе вероятностных моделей.

Во-вторых, математический аппарат анализа определенного типа реальных ситуаций не может исчерпываться лишь методами анализа данных. Необходима разработка вероятностных моделей, позволяющих обосновать эти методы, изучить их «работу», указать области применения для «конкурирующих» методов.

1.2. *Статистические данные.* В определении математической статистики упоминаются «статистические данные». Что это такое? Исходным материалом для работы алгоритмов математической статистики зачастую является выборка  $x_1, x_2, \dots, x_n$ . Какова природа  $x_i$ ? Обычно это числа. В вероятностной статистике, как правило, под выборкой понимается реализация совокупности  $n$  независимых одинаково распределенных случайных величин.

В многомерном статистическом анализе каждое из  $x_i$  — это конечномерный вектор. В статистике случайных процессов  $x_i$  — реализация случайного процесса. Иногда выборка состоит из единственной реализации, по которой судят о характеристиках случайного процесса (эргодическая теория). Если время дискретно, говорят о временных рядах. Математики рассматривали также случаи, когда  $x_i$  — функция (статистика в гильбертовом пространстве) или элемент некоторой группы (статистика на группах).

Практика привела к необходимости массовой обработки статистических данных иной природы. Это измерения в номинальной, порядковой и иных шкалах, отличных от абсолютной, ранжировки, разбиения, толерантности и иные виды бинарных отношений, множества, нечеткие множества, парные и множественные сравнения. Такой перечень обусловлен лишь историческими причинами, а потому сразу возникает вопрос: «Почему бинарные отношения вошли в перечень, а тернарные нет?» Ответ прост — мы пока не сталкивались с работами в приложениях, для которых были бы нужны тернарные отношения.

Перечисленные виды данных отличаются от классических объектов математической статистики тем, что для них не имеют смысла операции сложения, умножения на число и т. п. Другими словами, статистические данные являются элементами пространств, которые не наделены какими-либо алгебраическими структурами или, зачастую, структурами порядка [6]. Это приводит к необходимости специальных определений для таких исходных понятий классической математической статистики, как среднее и математическое ожидание. Мы ведь не можем сложить несколько разбиений и разделить результат на объем выборки!

Естественным представляется обобщение, в котором статистические данные рассматриваются как элементы пространств общей природы. Именно этому обобщению и посвящена основная часть статьи. К рассматриваемым пространствам предъявляются по возможности минимальные требования, обеспечивающие справедливость нужных нам теорем.

1.3. *Термин «статистика объектов нечисловой природы».* Этот термин мы предложили в 1978 г. Первая часть его — «статистика» — объясняется тем, что рассматриваются классические задачи математической статистики, прежде всего описание материала, оценивание, проверка гипотез. Термин «объекты нечисловой природы» предложен как собирательный для таких объектов, как бинарные отношения, множества, элементы пространств общей природы и т. д. Эти объекты нецелесообразно рассматривать как числовые. Конечно, бинарное отношение можно соотнести с матрицей из 0 и 1, а множество описать характеристической функцией, т. е. в обоих случаях объекты можно выразить с помощью чисел. Однако бессмысленно перемножать матрицы, соответствующие бинарным отношениям, или складывать функции, задающие множества. Другими словами, рассматриваемые нами объекты не обладают алгебраическими структурами матриц, функций и других числовых объектов [6]. Именно поэтому термин «объекты нечисловой природы» представляется уместным, хотя многие из этих объектов и можно выразить с помощью чисел.

Прежде чем переходить к статистике объектов общей природы, рассмотрим кратко ряд конкретных видов объектов нечисловой природы.

## 2. Измерение и шкалирование

2.1. *Измерение привлекательности профессий.* Начнем с теории измерений. Интерес к ней возник у нас при обдумывании результатов конкретных социологических и педагогических исследований. Коллектив новосибирских социологов под руководством В. Н. Шубкина изучал привлекательность различных профессий для выпускников новосибирских школ [49]. Был составлен список из 80 профессий. Опрашиваемых просили оценить каждую из этих профессий одним из баллов 1, 2, ..., 10 по правилу: чем больше нравится, тем выше балл. Для получения содержательных

выводов необходимо было найти единую оценку привлекательности профессии для выпускников школ. В качестве таковой использовали среднее арифметическое выставленных профессии баллов. В частности, физика получила средний балл 7,69, а математика — 7,50. Значит, физика предпочитается математике.

Однако, по данным Г. П. Щукиной [50], ленинградские школьники средних классов больше любят математику, чем физику, — их мнение противоположно мнению новосибирских выпускников.

Чем вызвано это различие? Тем, что ленинградцы и новосибирцы отличались по возрасту? Или пристрастия школьников меняются в зависимости от региона? Тогда различие объективно существует. Однако к исходным данным применялась неадекватная методика обработки. В действительности возможно, что никакого различия не существует.

Разберемся, почему недопустимо использовать среднее арифметическое баллов в качестве единой оценки привлекательности профессии. Для этого взглянем сначала в баллы, которые присваиваются профессиям школьниками. Действительно ли это числа? Одна профессия оценена в 10 баллов, другая — в 2. Значит ли это, что первая ровно в пять раз привлекательнее второй? Можно ли содержательно проинтерпретировать сообщение о том, что одна профессия привлекательнее другой на 2 балла?

Естественно принять за основу при разработке методики обработки данных, что школьник действительно в состоянии сказать, какая из двух профессий ему больше нравится, или же они для него одинаково привлекательны. Но вряд ли он сможет сказать, во сколько раз одна из них лучше другой.

Тогда процедуру оценки школьником привлекательности профессий можно представить себе так. Имеется 10 ящиков с номерами 1, 2, . . . , 10. Чем больше номер ящика, тем более привлекательна профессия; так что ящик № 1 предназначен для наименее нравящихся, а № 10 — для самых привлекательных. Названия профессий написаны на карточках. Каждый школьник опускает карточки в те ящики, которые кажутся ему наиболее подходящими. Ящики и карточки существуют в уме школьника, а социологам он сообщает лишь требуемые от него баллы — номера ящиков.

Возникает вопрос: почему ящикам приписаны номера — баллы от 1 до 10? Ведь имеется ни больше, ни меньше, как десять упорядоченных градаций. Школьнику все равно, какой системой баллов пользоваться — 1, 2, 3, . . . , 10 или 1, 4, 9, . . . , 100 и т. д. Балл для него — лишь код, приписанный ящику. А вот результаты сравнения профессий по привлекательности с помощью вычисления средних арифметических баллов могут зависеть, как мы покажем ниже, от того, какая именно система баллов используется.

То, что можно пользоваться системой баллов, отличной от равномерной шкалы 1, 2, . . . , 10, не выдумка математика. Так давно поступают многие социологи. Например, Ю. Вооглайд еще

в 60-х годах включал в анкету для читателей газеты «Эдази» (г. Тарту) вопрос о читаемости со следующими подсказками: «всегда читаю», «иногда читаю», «иногда — нет», «обычно не читаю», «никогда не читаю». (В нашей терминологии: пять подсказок — пять «ящичков».) После обсуждения с экспертами было решено приписать первой подсказке оценку «8», второй — «5», третьей — «4» и т. д. Как видим, здесь использована неравномерная шкала. Ряд способов приписывания численных значений подсказкам-градациям указан в [19].

Вернемся к опросу о привлекательности профессий. Если школьнику все равно, какой системой баллов пользоваться, то это значит, что распределение профессий по ящикам не изменится, если мы перейдем к другой системе баллов с помощью строго возрастающей функции  $\varphi: R^1 \rightarrow R^1$ , отображающей действительную прямую  $R^1$  на себя. Пусть  $x_1, x_2, \dots, x_n$  — ответы  $n$  школьников, касающиеся математики, а  $y_1, y_2, \dots, y_n$  — физики. Поскольку при преобразовании шкалы с помощью функции  $\varphi$  балл  $z$  заменяется на  $\varphi(z)$  при любом  $z$ , то после перехода к новой системе баллов ответы относительно математики будут иметь вид  $\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)$ , а относительно физики —  $\varphi(y_1), \varphi(y_2), \dots, \varphi(y_n)$ .

Пусть единая оценка привлекательности профессии вычисляется с помощью функции  $f: R^n \rightarrow R^1$ . В разбираемом нами исследовании новосибирских социологов использовалась функция

$$f(x_1, x_2, \dots, x_n) = \frac{x_1 + x_2 + \dots + x_n}{n}. \quad (1)$$

Что необходимо потребовать от функции  $f$ , чтобы полученные с ее помощью выводы не зависели от того, какой именно системой баллов пользовался социолог?

2.2. *Устойчивость результата сравнения средних.* Единая оценка вычислялась для того, чтобы сравнивать профессии по привлекательности. Поэтому потребуем устойчивости результата сравнения, т. е. того, чтобы сравнительная привлекательность профессий не менялась при переходе к иной системе баллов. В математических терминах это требование можно выразить так. Пусть

$$f(x_1, x_2, \dots, x_n) < f(y_1, y_2, \dots, y_n). \quad (2)$$

Тогда необходимо, чтобы

$$f(\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)) < f(\varphi(y_1), \varphi(y_2), \dots, \varphi(y_n)). \quad (3)$$

Наоборот, из справедливости (3) должно вытекать, что (2) также выполнено.

Потребуем, чтобы при любых возможных ответах школьников упорядоченность единых оценок не менялась при переходе к произвольной новой системе баллов. Это значит, что для любых  $x_i, y_i, i = 1, 2, 3, \dots, n$  и любой строго возрастающей функции

ф неравенства (2) и (3) либо одновременно выполнены, либо одновременно не выполнены.

Ясно, что функция (1) не удовлетворяет сформулированному в предыдущем абзаце условию. Действительно, пусть  $n = 2$ ,  $x_1 = 1$ ,  $x_2 = 7$ ,  $x_3 = 4$ ,  $x_4 = 6$ , тогда  $f(x_1, x_2) = 4 < f(y_1, y_2) = 5$ . Пусть функция  $\varphi$  такова, что  $\varphi(1) = 1$ ,  $\varphi(4) = 4$ ,  $\varphi(6) = 6$ ,  $\varphi(7) = 11$ . Тогда  $f(\varphi(x_1), \varphi(x_2)) = 6 > f(\varphi(y_1), \varphi(y_2)) = 5$ , неравенство (2) выполнено, а (3) — нет.

Таким образом, использование среднего арифметического неправомерно. Чем же можно пользоваться? Оказывается, есть ровно  $n$  функций, которые можно использовать в качестве средних. Это порядковые статистики, построенные по выборке  $x_1, x_2, \dots, x_n$ .

Напомним определение порядковых статистик. Числа  $x_1, x_2, \dots, x_n$  расположим в порядке неубывания, получим  $x(1) \leq x(2) \leq \dots \leq x(n)$ . Упорядоченная совокупность чисел  $x(1), x(2), \dots, x(n)$  называется вариационным рядом, построенным по  $x_1, x_2, \dots, x_n$ , а его члены — порядковыми статистиками:  $i$ -я порядковая статистика это  $x(i)$ .

*Теорема 1* [29]. Пусть для функции  $f$  результат сравнения устойчив (в разобранном выше смысле). Пусть функция  $f$  непрерывна по совокупности переменных и симметрична, т. е. не меняется при любой перестановке аргументов. Пусть  $f$  — средняя величина, по Коми [9], т. е.

$$\min_{1 \leq i \leq n} x_i \leq f(x_1, x_2, \dots, x_n) \leq \max_{1 \leq i \leq n} x_i. \quad (4)$$

Тогда найдется натуральное число  $m$ ,  $1 \leq m \leq n$ , такое, что  $f(x_1, x_2, \dots, x_n)$  есть  $m$ -я порядковая статистика, т. е.

$$f(x_1, x_2, \dots, x_n) \equiv x(m). \quad (5)$$

Прокомментируем условия теоремы. О необходимости устойчивости результата сравнения достаточно сказано выше. Естественно считать, что при малом изменении аргументов среднее также мало меняется — отсюда непрерывность. Симметричность означает: мнения всех школьников учитываются в равной мере. Условие (4) — самое слабое, которое можно наложить на  $f$ , чтобы иметь основание называть ее средней величиной.

Какие практические рекомендации вытекают из теоремы 1? В качестве средних целесообразно использовать члены вариационного ряда, прежде всего «левую» медиану  $x\left(\left[\frac{n+1}{2}\right]\right)$  и «правую» медиану  $x\left(\left[\frac{n}{2}\right] + 1\right)$ , т. е. порядковые статистики с номерами  $\left[\frac{n+1}{2}\right]$  и  $\left[\frac{n}{2}\right] + 1$ , где квадратными скобками обозначена операция взятия целой части числа. Если  $n$  — нечетное число, то  $\left[\frac{n+1}{2}\right] = \left[\frac{n}{2}\right] + 1$ , «левая» и «правая» медианы совпадают между собой, а также с медианой, определяемой в обычных курсах по об-

щей теории статистики. Совпадают они и с медианой, получаемой в оптимизационном подходе, о котором речь впереди.

Если же  $n$  четно,  $n = 2k$ , то  $\left\lfloor \frac{n+1}{2} \right\rfloor = k$ , в то время как  $\left\lfloor \frac{n}{2} \right\rfloor + 1 = k - 1$ . В курсах по общей теории статистики медианой называют [15, 45] величину  $\frac{x(k) + x(k+1)}{2}$ . Ясно, что для так определенной медианы результат сравнения не будет устойчив. В оптимизационном подходе к статистике медиана — это отрезок  $\{x(k), x(k+1)\}$ . Таким образом, имеется три определения медианы. Ясно, впрочем, что при большом  $n$  они дают обычно близкие результаты.

Требование устойчивости результата сравнения — довольно сильное: неравенства (2) и (3) должны быть равносильны для любых значений аргументов и любого строго монотонного преобразования шкалы. Между тем в исследовании новосибирских социологов получены вполне определенные значения  $x_i, y_i, i = 1, \dots, n$ . Возможно, что для них упорядочение по медианам совпадает с упорядочением по средним баллам. Но гарантии такого идеального совпадения нет. Поэтому в стандартных пакетах обработки статистических данных наряду с вычислением средних арифметических баллов целесообразно предусмотреть вычисление медиан.

Было бы интересно для различных типов реальных данных установить, насколько часто упорядоченность по средним баллам или рангам отличается от упорядоченности по медианам. Для экспертной оценки картин [5] коэффициент ранговой корреляции между двумя соответствующими ранжированиями оказался равным 0,98.

**2.3. Вероятностная теория средних.** Теорема 1 получена в рамках анализа данных. В вероятностной теории можно указать условия, при которых для среднего арифметического (1) вероятность равносильности неравенств (2) и (3) стремится к 1 при возрастании объема наблюдений.

Пусть  $\xi_1, \xi_2, \dots, \xi_m$  — независимые одинаково распределенные случайные величины с функцией распределения  $F(x)$ , а  $\eta_1, \eta_2, \dots, \eta_n$  — независимые одинаково распределенные случайные величины с функцией распределения  $G(x)$ . Представим, что  $E\xi_1$  и  $E\eta_1$  существуют. Рассмотрим сначала случай  $E\xi_1 < E\eta_1$ . По закону больших чисел [7]

$$\begin{aligned} \frac{\xi_1 + \xi_2 + \dots + \xi_m}{m} &\rightarrow E\xi_1, \\ \frac{\eta_1 + \eta_2 + \dots + \eta_n}{n} &\rightarrow E\eta_1 \end{aligned} \quad (6)$$

при  $m \rightarrow \infty, n \rightarrow \infty$  (сходимость по вероятности). Поэтому вероятность выполнения неравенства

$$\frac{\xi_1 + \xi_2 + \dots + \xi_m}{m} < \frac{\eta_1 + \eta_2 + \dots + \eta_n}{n} \quad (7)$$

стремится к 1 при  $\min(m, n) \rightarrow \infty$ . Значит, для выяснения условий применимости, среднего арифметического необходимо и достаточно установить, когда вероятность справедливости неравенства

$$\frac{\varphi(\xi_1) + \varphi(\xi_2) + \dots + \varphi(\xi_m)}{m} < \frac{\varphi(\eta_1) + \varphi(\eta_2) + \dots + \varphi(\eta_n)}{n} \quad (8)$$

стремится к 1 при  $\min(m, n) \rightarrow \infty$ .

Желательно опять использовать закон больших чисел. Однако может случиться так, что  $E\xi_1$  существует, а  $E\varphi(\xi_1)$  не существует!

Какие условия на непрерывную функцию  $\varphi(x)$  надо наложить, чтобы из существования  $E\xi$  вытекало существование  $E\varphi(\xi)$ , каково бы  $\xi$  ни было? Оказывается [29, с.152], необходимое и достаточное условие выглядит следующим образом:

$$\lim_{|x| \rightarrow \infty} \left| \frac{\varphi(x)}{x} \right| < \infty. \quad (9)$$

Достаточность означает, что для любой случайной величины  $\xi$ , для которой существует математическое ожидание, и для любой непрерывной  $\varphi$ , удовлетворяющей условию (9), существует  $E\varphi(\xi)$ . Необходимость означает, что для любой непрерывной  $\varphi$ , не удовлетворяющей условию (9), найдется случайная величина  $\xi$ , для которой  $E\xi$  существует, а  $E\varphi(\xi)$  не существует.

Пусть выполнено (9). Тогда левая и правая части (8) сходятся к  $E\varphi(\xi_1)$  и  $E\varphi(\eta_1)$  соответственно. Необходимое и достаточное условие справедливости неравенства  $E\varphi(\xi_1) < E\varphi(\eta_1)$  для любой удовлетворяющей (9) непрерывной строго возрастающей функции  $\varphi$  таково [29, с. 154]:  $F(x) \geq G(x)$  при всех  $x$ , причем существует  $x_0$ , для которого  $F(x_0) > G(x_0)$ . Необходимость означает: в случае, когда только что сформулированное условие не выполнено, найдется непрерывная строго возрастающая функция  $\varphi$ , удовлетворяющая (9), для которой  $E\varphi(\xi_1) > E\varphi(\eta_1)$ .

Рассматриваемое условие выполнено, например, для  $G(x) = F(x - \alpha)$ ,  $\alpha > 0$ .

В случае  $E\xi_1 > E\eta_1$  аналогичное условие выглядит так:  $F(x) \leq G(x)$  при всех  $x$ , причем существует  $x_0$ , для которого  $F(x_0) < G(x_0)$ . Ему удовлетворяют, например,  $G(x)$ , полученные из  $F(x)$  сдвигом влево:  $G(x) = F(x + \alpha)$ ,  $\alpha > 0$ .

Приведенные выше результаты получены в предположении, что допустимое преобразование шкалы  $\varphi$  может быть любым строго возрастающим. Это предположение иногда пытаются оспаривать, заявляя, что школьники, возможно, все-таки могут сказать, во сколько раз одна профессия привлекательнее другой. Подобное утверждение можно проверить с помощью эксперимента. Нескольким группам школьников можно предложить ту же задачу оценки привлекательности профессий, но с разными системами баллов (по 10 градаций в каждой), а затем сравнить частоты ответов, пришедшихся на определенные градации. Если частоты отличаются незначимо, то распределение по градациям

не зависит от системы баллов, которые служат лишь кодами для градаций. Если же различие частот значимо, то распределение по градациям, наоборот, зависит от системы баллов и, возможно, кроме порядковых статистик существуют другие виды средних, результат сравнения которых устойчив относительно допустимых преобразований шкалы. Однако необходимо выяснить, какие именно преобразования являются допустимыми.

Таким образом, теория измерений позволяет, во-первых, указать способы обработки данных, пригодные во всех рассматриваемых случаях, во-вторых, поставить ряд новых содержательных задач. С математической точки зрения теорема 1 дает характеристику порядковых статистик.

**2.4. О репрезентативной теории измерений.** Различные результаты теории измерений изложены в [29, 42, 44]. Здесь достаточно отметить, что шкалы определяются группами допустимых преобразований  $\Phi = \{\varphi\}$ . В приведенных выше рассмотренных  $\Phi$  — группа всех строго возрастающих непрерывных преобразований. Это порядковая шкала. Шкалу наименований имеем, если  $\Phi$  — группа всех взаимнооднозначных отображений прямой на себя. Если  $\Phi$  — группа линейных преобразований  $\varphi(x) = ax + b$ ,  $a > 0$ , то измерения проводятся в шкале интервалов. В шкале отношений допустимыми преобразованиями шкалы являются все подобные преобразования  $\varphi(x) = ax$ ,  $a > 0$ . Наконец, в абсолютной шкале допустимо лишь тождественное преобразование  $\varphi(x) = x$ . Используются также шкалы с другими группами допустимых преобразований.

В шкалах интервалов, отношений и некоторых других для средних величин получены характеристические теоремы типа теоремы 1. Рассмотрим обобщенные средние по А. Н. Колмогорову [29, § 3.4]

$$f(x_1, x_2, \dots, x_n) = H^{-1} \left( \sum_{i=1}^n a_i H(x_i) \right), \quad (10)$$

где  $H$  — строго монотонная функция (т. е. строго возрастающая или строго убывающая):  $H^{-1}$  — обратная к ней;  $1 > a_i \geq 0$ ;  $i = 1, 2, \dots, n$ ;  $a_1 + a_2 + \dots + a_n = 1$ . Результат сравнения средних устойчив для любого линейного преобразования, т. е. в шкале интервалов, тогда и только тогда, когда  $f$  в (10) определяется  $H(z) = z$ . Он устойчив в шкале отношений (т. е. для любого подобного преобразования) тогда и только тогда, когда  $H(z) = z^c$ ,  $c \neq 0$ , или  $H(z) = \ln z$  (тогда  $f$  — степенное среднее). Эти результаты получены при некоторых условиях регулярности типа дифференцируемости  $H$  [34]. Было бы желательно попытаться ослабить эти условия, например, до непрерывности  $H$ .

С точки зрения теории измерений какой-либо вывод, полученный по данным, измеренным в шкале с группой допустимых преобразований  $\Phi$ , можно считать обоснованным тогда и только тогда, когда он остается в силе для данных, преобразованных с

помощью любого  $\varphi \in \Phi$ . Другими словами, вывод должен быть инвариантен относительно группы  $\Phi$ . Выше рассматривался вывод о том, что среднее одной совокупности меньше среднего другой, см. (2).

Методы обработки данных должны быть согласованы со шкалами измерения, т. е. выводы, полученные с их помощью, должны быть инвариантны относительно соответствующих групп преобразований. Это положение целесообразно сделать обязательным для разработчиков систем статистического анализа данных, включив его в соответствующие руководящие нормативные документы Госстандарта.

**2.5. Максимальные инварианты в шкале наименований и порядковой шкале.** Пусть  $x_1, x_2, \dots, x_n$  — результаты измерений в шкале наименований (другое название — номинальная шкала). Укажем максимальный инвариант, соответствующий  $x_1, x_2, \dots, x_n$ . (Инвариант  $B = B(x_1, x_2, \dots, x_n)$  называется максимальным, если любой другой инвариант представим в виде функции от  $B$ .) Им является [29, с. 109] матрица  $B$  порядка  $n \times n$  с элементами

$$b_{ij} = \begin{cases} 1, & x_i = x_j, \\ 0, & x_i \neq x_j. \end{cases} \quad (11)$$

Если  $x_1, x_2, \dots, x_n$  — результаты измерений в порядковой шкале, то максимальный инвариант — это [29, с. 111] матрица  $C$  порядка  $n \times n$  с элементами

$$c_{ij} = \begin{cases} 1, & x_i \leq x_j, \\ 0, & x_i > x_j. \end{cases} \quad (12)$$

Максимальные инварианты (11) и (12) можно описать так же, как бинарные отношения. Задать матрицу  $B$  — это значит задать разбиение совокупности  $x_1, x_2, \dots, x_n$  на группы, в каждую из которых входят совпадающие элементы, т. е. задать разбиение на классы эквивалентности. Если у нас имеется несколько наборов данных  $\{x_1, x_2, \dots, x_n\}$ ,  $\{y_1, y_2, \dots, y_n\}$  и т. д., то в силу принципа согласованности методов обработки со шкалами измерений из того, что данные измерены в шкале наименований, следует, что фактически требуется анализировать разбиения, порожденные этими наборами.

С помощью матрицы  $C$  также задается разбиение на классы эквивалентности ( $x_i$  и  $x_j$  входят в один класс тогда и только тогда, когда  $c_{ij} = 1$  и  $c_{ji} = 1$ ). Кроме того, между этими классами вводится отношение линейного строгого порядка. Такое бинарное отношение имеет ряд названий — упорядочение [13, гл. 2], квазисерия [21, с. 31], ранжировка со связями [14]. Если каждый из классов эквивалентности состоит только из одного элемента, то имеем обычную ранжировку.

Таким образом, теория измерений естественным путем приводит нас к другому типу объектов нечисловой природы — би-

нарным отношениям. Напомним, что бинарным отношением на конечном множестве  $X = \{x_1, x_2, \dots, x_k\}$  называется произвольное подмножество его декартова квадрата  $X^2$ . Так же как разбиения и упорядочения, любое бинарное отношение может быть описано с помощью матрицы из 0 и 1. Бинарному отношению  $Q$  соответствует матрица  $\|q_{ij}\|$  порядка  $k \times k$ , где  $q_{ij} = 1$ , если  $(x_i, x_j) \in Q$ , и  $q_{ij} = 0$  — в противном случае.

2.6. *Оценка размерности пространства в многомерном шкалировании.* Прежде чем перейти к статистике бинарных отношений, приведем один новый результат, относящийся к многомерному шкалированию — области, тесно примыкающей к теории измерений. Идея многомерного шкалирования состоит в представлении каждого объекта точкой геометрического пространства (обычно размерности 1, 2 или 3), координатами которой служат скрытые значения факторов, в совокупности достаточно адекватно описывающих объект. При этом отношения между объектами заменяются отношениями между точками. Так, данные о сходстве объектов — расстояниями между точками, данные о превосходстве — взаимным расположением точек [47]. Существует ряд различных моделей многомерного шкалирования. Во всех них встает проблема оценки истинной размерности факторного пространства. Рассмотрим эту проблему на примере обработки данных о сходстве объектов с помощью метрического шкалирования. Пусть имеется  $n$  объектов  $O(1), O(2), \dots, O(n)$ , для каждой пары объектов  $O(i), O(j)$  задана мера их сходства  $s_{ij}$  (считаем, что  $s_{ij} = s_{ji}$ ). Происхождение чисел  $s_{ij}$  не имеет значения для описания работы алгоритма. Они могли быть получены либо непосредственным измерением, либо с использованием экспертов, либо путем вычисления по совокупности описательных характеристик, либо как-то иначе. В евклидовом пространстве рассматриваемые  $n$  объектов должны быть представлены конфигурацией  $n$  точек, причем в качестве меры близости точек-представителей выступает евклидово расстояние  $d_{ij}$  между соответствующими точками. Степень соответствия между совокупностью объектов и совокупностью представляющих их точек определяется путем сопоставления матриц сходства  $\|s_{ij}\|$  и расстояний  $\|d_{ij}\|$ . Метрический функционал сходства имеет вид

$$S = \sum_{i < j} |s_{ij} - d_{ij}|^2. \quad (13)$$

Геометрическую конфигурацию надо выбирать так, чтобы  $S$  достигал своего наименьшего значения [47].

Пусть евклидово пространство имеет размерность  $m$ . Положим,

$$\alpha_m = \frac{2}{n(n-1)} \min S, \quad (14)$$

где минимум берется по всем возможным конфигурациям  $n$  точек в  $m$ -мерном евклидовом пространстве; можно показать, что  $\min$

в (14) достигается. Ясно, что при росте  $m$  величина  $\alpha_m$  монотонно убывает (точнее, не возрастает), а при  $m \geq n - 1$  равна 0. Для возможности содержательной интерпретации желательно действовать в пространстве возможно меньшей размерности. При этом, однако, размерность необходимо выбрать так, чтобы точки представляли объекты без больших искажений. Возникает вопрос: как рационально выбирать  $m$ ?

В рамках анализа данных обоснованного ответа на него, видимо, нет. На наш взгляд, необходимо изучить поведение  $\alpha_m$  в некоторых вероятностных моделях. Если  $s_{ij}$  являются случайными величинами, распределение которых зависит от «истинной размерности»  $m_0$  (и возможно, от каких-либо еще параметров), то можно в классическом математико-статистическом стиле ставить задачу оценки  $m_0$ , искать состоятельные оценки и т. д. Правда, задача оценки «размерности» модели трудна, как можно видеть из рассмотренных ранее примеров регрессии [32, 39] и расщепления смесей в теории классификации [40].

Будем строить вероятностные модели. Примем, что объекты представляют собой точки в евклидовом пространстве размерности  $k$ , где  $k$  достаточно велико. То, что «истинная размерность» равна  $m_0$ , означает, что все эти точки лежат на гиперплоскости размерности  $m_0$ . Примем, что совокупность рассматриваемых точек представляет собой выборку из кругового нормального распределения с дисперсией  $\sigma_0^2$ . Это означает, что точки есть независимые в совокупности случайные вектора, каждый из которых строится как  $\xi(1)e(1) + \dots + \xi(m_0)e(m_0)$ , где  $e(1), \dots, e(m_0)$  — ортонормальный базис в подпространстве размерности  $m_0$ , в котором лежат рассматриваемые точки;  $\xi(1), \dots, \xi(m_0)$  — независимые в совокупности одномерные нормальные случайные величины с математическим ожиданием 0 и дисперсией  $\sigma_0^2$ .

Рассмотрим две модели получения мер близости  $s_{ij}$ . В первой из них  $s_{ij}$  отличаются от евклидова расстояния между соответствующими точками из-за того, что точки известны с искажением. Пусть  $d_1, \dots, d_n$  — рассматриваемые точки. Тогда

$$s_{ij} = \rho(d_i + \varepsilon_i, d_j + \varepsilon_j), \quad i, j = 1, \dots, n, \quad (15)$$

где  $\rho$  — евклидово расстояние между точками в  $k$ -мерном пространстве, вектора  $\varepsilon_1, \dots, \varepsilon_n$  представляют собой выборку из кругового нормального распределения в  $k$ -мерном пространстве с нулевым математическим ожиданием и ковариационной матрицей  $\sigma_1^2 I$ , где  $I$  — единичная матрица. Другими словами,  $\varepsilon_i = \eta_{i1}e(1) + \eta_{i2}e(2) + \dots + \eta_{ik}e(k)$ , где  $e(1), \dots, e(k)$  — базис в  $k$ -мерном пространстве (ортонормальный), а  $\{\eta_{ij}\}$  — совокупность независимых в совокупности одномерных нормальных случайных величин с нулевым математическим ожиданием и дисперсией  $\sigma_1^2$ .

Во второй модели искажения наложены на сами расстояния

$$s_{ij} = \rho(d_i, d_j) + \varepsilon_{ij}, \quad i, j = 1, \dots, n, \quad i \neq j, \quad (16)$$

где  $\{\varepsilon_{ij}, i, j = 1, \dots, n\}$  — независимые в совокупности нормальные случайные величины с математическим ожиданием 0 и дисперсией  $\sigma_1^2$ .

Автором показано, что в моделях (15) и (16) случайная величина  $\alpha_m$  из (13), (14) при  $n \rightarrow \infty$  сходится по вероятности к

$$f(m) = f_1(m) + \sigma_1^2(k - m), \quad m = 1, 2, \dots, k, \quad (17)$$

где

$$f_1(m) = \begin{cases} \sigma_0^2(m_0 - m), & m < m_0, \\ 0, & m \geq m_0. \end{cases} \quad (18)$$

Таким образом, функция  $f(m)$  из (17) линейна на интервалах  $[1, m_0]$  и  $[m_0, k]$ , причем на первом интервале она убывает быстрее, чем на втором. Отсюда следует, что статистика

$$m^* = \underset{m}{\operatorname{Argmax}} (\alpha_{m+1} - 2\alpha_m + \alpha_{m-1}) \quad (19)$$

является состоятельной оценкой истинной размерности  $m_0$ .

Итак, из вероятностной теории вытекает рекомендация — определять размерность факторного пространства по правилу (19). Отметим, что подобная рекомендация была сформулирована как эвристическая одним из основателей многомерного шкалирования Краскалом на основе опыта практического использования многомерного шкалирования и вычислительных экспериментов. Вероятностная теория позволила обосновать эту эвристическую рекомендацию.

### 3. Статистика бинарных отношений

**3.1. Бинарные отношения.** Кроме разбиений и упорядочений, рассмотренных выше, важным видом бинарных отношений являются толерантности, т. е. рефлексивные симметричные отношения. (Бинарное отношение  $Q$  на множестве  $X$  называется рефлексивным, если  $(x, x) \in Q$  при любом  $x \in X$ ; симметричным, если из  $(x, y) \in Q$  вытекает  $(y, x) \in Q$  при любых  $x \in X, y \in X$ .) Толерантность — математическая модель для выражения представления о сходстве (близости). Разбиения, конечно, являются толерантностями, но последние не обязаны удовлетворять условию транзитивности, т. е. для толерантности  $Q$  не обязательно выполняться следующее условие: из  $(x, y) \in Q$  и  $(y, z) \in Q$  следует  $(x, z) \in Q$  для любых  $x, y, z$  из  $X$ .

Основной задачей в статистике бинарных отношений является нахождение среднего отношения для совокупности отношений  $Q_1, Q_2, \dots, Q_n$ . Оно, как и в классической статистике, рассматри-

вается как «показатель центральной тенденции» совокупности  $Q_1, Q_2, \dots, Q_n$ . Если  $Q_i$  отражает точку зрения  $i$ -го респондента, то среднее отношение интерпретируется как групповое мнение. Если  $Q_i$  — мнение  $i$ -го эксперта, то среднее представляет собой обобщенное мнение экспертной комиссии. Если  $Q_i$  есть упорядочение объектов по  $i$ -му показателю качества, то в среднем отношении выражено упорядочение по обобщенному показателю качества.

Как определить среднее отношение? В книге Дж. Кемени и Дж. Снелла [13] предложено находить для упорядочений  $Q_1, Q_2, \dots, Q_n$  «медиану Кемени» [51]

$$Q_{med} = \operatorname{Argmin}_Q \sum_{i=1}^n d(Q_i, Q) \quad (20)$$

и «среднее по Кемени»

$$Q_{cp} = \operatorname{Argmin}_Q \sum_{i=1}^n d^2(Q_i, Q). \quad (21)$$

Здесь  $d$  — «расстояние Кемени» между упорядочениями. Минимум берется в (20) и (21) по всем упорядочениям. (Символом  $\operatorname{Argmin}$  обозначено то отношение, для которого соответствующая сумма достигает минимума.)

Пусть  $A$  и  $B$  — бинарные отношения, определенные на одном и том же конечном множестве, а  $\|a_{ij}\|$  и  $\|b_{ij}\|$  — соответствующие им матрицы из 0 и 1. Положим,

$$d(A, B) = \sum_{i,j} |a_{ij} - b_{ij}|. \quad (22)$$

Легко видеть, что определенная в (22) функция  $d$  является метрикой в пространстве всех бинарных отношений, определенных на рассматриваемом конечном множестве, а также в любых его подпространствах, в частности в подпространствах упорядочений, разбиений, толерантностей. В [13] приведена система аксиом для расстояния между упорядочениями и показано, что ему удовлетворяет расстояние (22) и только оно. В дальнейшем для других типов бинарных отношений были предложены системы аксиом для расстояний между отношениями рассматриваемых типов, которым удовлетворяет расстояние (22) и только оно. Аксиоматически можно ввести и расстояния между отношениями, отличные от (22).

Определенное в (22) расстояние  $d$  называется расстоянием Кемени между бинарными отношениями, в честь основоположника аксиоматического подхода к введению расстояний [47].

Для такого частного случая упорядочений, как ранжировки, расстояние Кемени линейно связано с коэффициентом ранговой корреляции  $\tau$  Кендалла [18]

$$\tau(A, B) = 1 - \frac{2d(A, B)}{k(k-1)}, \quad (23)$$

где  $k$  — число ранжируемых объектов.

Для коэффициента ранговой корреляции  $\rho$  Спирмена подобной связи нет — ни одна линейная функция от  $\rho$  не задает расстояние между ранжировками. Это служит доводом в пользу применения  $\tau$ , а не  $\rho$ , при обработке ранжировок.

Иногда утверждают, будто коэффициент ранговой корреляции  $\rho$  Спирмена не инвариантен в порядковой шкале, в то время как коэффициент  $\tau$  Кендалла инвариантен. Это утверждение неверно, и тот и другой коэффициенты определяются с помощью рангов, а потому они инвариантны в порядковой шкале, и теория измерений не дает оснований для того, чтобы предпочесть один коэффициент другому.

3.2. *Результирующее отношение.* Кроме медианы Кемени и среднего по Кемени, можно предложить много способов получения результирующего отношения. Для этого рассмотрим произвольную функцию  $f(A, B)$ , определенную на совокупности пар бинарных отношений рассматриваемого вида, и положим

$$Q_n(f) = \operatorname{Argmin}_Q \sum_{i=1}^n f(Q_i, Q). \quad (24)$$

Ясно, что

$$Q_{\text{med}} = Q_n(d), \quad Q_{\text{ср}} = Q_n(d^2). \quad (25)$$

Можно рассматривать  $Q_n(f)$  в качестве среднего мнения.

Для лучшего согласования с интуицией целесообразно потребовать, чтобы  $f$  была мерой близости, т. е. для нее были выполнены все аксиомы метрики, кроме, быть может, неравенства треугольника. Меры близости обладают любопытным свойством: если  $f$  — мера близости, а  $h$  — строго возрастающая функция, отображающая луч  $[0, +\infty)$  на себя, то  $h(f)$  — также мера близости. Использование мер близости, а не только расстояний, для нахождения результирующего отношения представляется оправданным, поскольку  $Q_{\text{ср}}$  определено с помощью меры близости  $d^2$ , не являющейся расстоянием.

В частности, рассмотрим

$$f_0(A, B) = 1 - \rho(A, B), \quad (26)$$

где  $\rho(A, B)$  — коэффициент ранговой корреляции Спирмена. Легко видеть, что  $f_0$  — мера близости. Известно [14], что  $Q_n(f_0)$  — упорядочение по средним рангам (это значит, что для каждого объекта нужно сложить ранги, присвоенные ему в ранжировках  $Q_1, \dots, Q_n$ , и в результирующей ранжировке расположить объекты в порядке возрастания полученных сумм).

В силу (20) и (23)  $Q_{\text{med}}$  — это ранжировка, в среднем наиболее коррелированная по Кендаллу с совокупностью  $Q_1, \dots, Q_n$ , т. е.

$$Q_{\text{med}} = \operatorname{Argmax}_Q \sum_{i=1}^n \tau(Q_i, Q), \quad (27)$$

где  $\text{Argmax}$  берется по всем ранжировкам. Аналогично  $Q_n(f_0)$  — это ранжировка, в среднем наиболее коррелированная по Спирмену с той же совокупностью, т. е.

$$Q_n(f_0) = \text{Argmax}_Q \sum_{i=1}^n \rho(Q_i, Q). \quad (28)$$

Чем же пользоваться —  $Q_{med}$  или  $Q_n(f_0)$ ? О предпочтении  $Q_{med}$  говорит то, что оно получено с помощью  $d$  из (22), которое, во-первых, является расстоянием, а во-вторых, выводится из некоторой системы аксиом. Однако нахождение медианы Кемени — сложная вычислительная задача и решать ее можно лишь с помощью ЭВМ [47]. В то же время  $Q_n(f_0)$  определяется элементарно — по средним рангам. Это — сильный довод в пользу применения  $Q_n(f_0)$ . Именно поэтому в [9] для нахождения результирующей ранжировки используется  $Q_n(f_0)$ , а не  $Q_{med}$ .

Мы будем называть медианой Кемени и средним по Кемени величины  $Q_{med}$  и  $Q_{cp}$ , определяемые по (20) и (21), также и в случае, когда  $Q_1, Q_2, \dots, Q_n$  — бинарные отношения произвольной природы. При этом  $d$  определяется по (22). Отношение  $Q_n(f)$  будем называть средним относительно  $f$  совокупности  $Q_1, Q_2, \dots, Q_n$ .

При использовании  $Q_n(f)$  возникают два вопроса. Во-первых, будет ли результирующее мнение устойчивым, т. е. будет ли оно мало меняться при увеличении числа отношений  $n$ ? Другими словами, сходится ли  $Q_n(f)$  к какому-либо пределу при  $n \rightarrow \infty$ ? Требование устойчивости представляется естественным, если предполагается, что все отношения «в равной мере» участвуют в формировании среднего. Во-вторых, каков этот предел?

В рамках анализа данных ответить на эти вопросы невозможно. Ответы даются лишь в вероятностной теории.

3.3. Закон больших чисел для бинарных отношений. Пусть  $Q_1, Q_2, \dots, Q_n$  — независимые одинаково распределенные случайные отношения. Положим,<sup>4</sup>

$$Q_\infty(f) = \text{Argmin}_Q E f(Q_1, Q), \quad (29)$$

где  $\text{Argmin}$  берется по рассматриваемому пространству бинарных отношений. Назовем  $Q_\infty(f)$  математическим ожиданием  $Q_1$  относительно  $f$ . (Обоснование использованию термина «математическое ожидание» будет дано ниже.)

Обратим внимание на то, что  $Q_n(f)$  и  $Q_\infty(f)$  могут определяться не единственным образом. Так, если ранжировка  $Q_1$  равномерно распределена на множестве всех ранжировок, то, как можно показать,  $E f(Q_1, Q)$  принимает одно и то же значение при любом  $Q$ , т. е. любая ранжировка удовлетворяет определению  $Q_\infty(f)$ . Поэтому внесем уточнение:  $Q_n(f)$  и  $Q_\infty(f)$  — это множества всех тех ранжировок, на которых достигают минимума соответствующие выражения в (24) и (29). Непустота этих множеств вытекает из того, что все рассматриваемые сейчас пространства состоят из

конечного числа элементов. Разумеется, каждое из множеств  $Q_n(f)$ ,  $Q_\infty(f)$  может состоять лишь из одного элемента. Более того,  $Q_\infty(f)$ , как правило, состоит из одного элемента. Чтобы уточнить смысл слов «как правило», рассмотрим совокупность вероятностей  $P(Q_1 = Q')$ , где  $Q'$  пробегает все рассматриваемое пространство бинарных отношений. Поскольку

$$Ef(Q_1, Q) = \sum_{Q'} f(Q', Q) P(Q_1 = Q'), \quad (30)$$

то из неоднозначности математического ожидания вытекает, что на указанную выше совокупность вероятностей наложено одно или несколько линейных условий (связей), выделяющих в рассматриваемом симплексе вероятностей многообразие меньшей размерности. Поэтому для почти всех (в смысле меры Лебега в соответствующем симплексе вероятностей) распределений вероятностей математическое ожидание единственно.

Рассмотрим сначала случай, когда  $Q_n(f)$  и  $Q_\infty(f)$  состоят каждое из одного элемента. Допустим вольность, обозначив указанные элементы теми же символами, что и рассматриваемые множества. Тогда [29]

$$Q_n(f) \rightarrow Q_\infty(f) \quad (31)$$

при  $n \rightarrow \infty$  (сходимость с вероятностью 1). Таким образом, выборочное среднее  $Q_n(f)$  сходится к математическому ожиданию  $Q_\infty(f)$ . Это оправдывает использование термина «математическое ожидание» применительно к  $Q_\infty(f)$ . Соотношение (31) естественно назвать законом больших чисел для случайных величин со значениями в пространстве отношений. Он состоит в том, что решение выборочной экстремальной задачи сходится к решению предельной экстремальной задачи. Отметим, что (24) отличается от (29) только тем, что в (29) математическое ожидание берется по некоторой теоретической вероятностной мере, а в (24) — по соответствующей ей эмпирической мере, приписывающей, как известно, каждому наблюдению вес  $1/n$ , где  $n$  — объем выборки (в (24) множитель  $1/n$  опущен, поскольку он не влияет на значение  $\text{Argmin}$ ). Ниже мы увидим, что среднее арифметическое и математическое ожидание одномерных случайных величин, участвующие в обычном законе больших чисел, также можно представить в виде решений экстремальных статистических задач.

Необходимо определить понятие сходимости в (31), поскольку топологическую структуру в пространстве отношений мы не вводили. В силу дискретности значений функции  $f$  и единственности решений задач (24) и (29) справедливы следующие утверждения [29]:

$$\lim_{n \rightarrow \infty} P(Q_n(f) = Q_\infty(f)) = 1; \quad (32)$$

с вероятностью 1 существует (случайное) число  $n_0$ , такое, что  $Q_n(f) = Q_\infty(f)$  при всех  $n > n_0$ .

Если учитывать возможность неединственности решений хотя бы одной из задач (24), (29), то закон больших чисел будет выглядеть следующим образом:

$$\lim_{n \rightarrow \infty} P(Q_n(f) \subseteq Q_\infty(f)) = 1. \quad (33)$$

Из (33), в частности, следует, что в случае единственности решения задачи (29) вероятность единственности выборочного среднего стремится к 1.

Усиленный закон больших чисел в общем случае формулируется так: с вероятностью 1 существует (случайное) число  $n_0$ , такое, что  $Q_n(f) \subseteq Q_\infty(f)$  при всех  $n > n_0$ .

Все сказанное выше относилось к случайным величинам со значениями в пространстве отношений. Мы показали, что выборочное среднее  $Q_n(f)$  устойчиво, и указали предел, к которому оно стремится при возрастании выборки. Он зависит как от распределения случайного отношения  $Q_1$ , так и от меры близости  $f$ .

Как законы больших чисел, так и их доказательства [29] могут без изменений быть установлены для любых случайных величин со значениями в конечном множестве. В (24), (29), (32), (33) и формулировках усиленного закона больших чисел можно считать, что  $Q_1, Q_2, \dots, Q_n$  — выборка из генеральной совокупности со значениями в произвольном конечном множестве.

Укажем одно обобщение законов больших чисел. Пусть  $A = \{a\}$  — конечное множество,  $f$  — функция на  $A^2$ ,  $P$  — вероятностная мера на  $A$ ,  $P_n, n = 1, 2, \dots$ , — последовательность случайных вероятностных мер на  $A$ , такая, что  $P_n(a) \rightarrow P(a)$  при  $n \rightarrow \infty$  для любого  $a \in A$ . Тогда

$$\lim_{n \rightarrow \infty} P\left\{\operatorname{Argmin}_{a'} \sum_a f(a', a) P_n(a') \subseteq \operatorname{Argmin}_{a'} \sum_a f(a', a) P(a')\right\} = 1. \quad (34)$$

Соотношение (34) соответствует сходимости мер по вероятности. Если же они сходятся с вероятностью 1, то указанное в (34) включение справедливо при всех  $n > n_0$ , где  $n_0$  зависит от случая. Частному случаю, когда  $P_n$  — эмпирические меры, соответствуют приведенные выше законы больших чисел.

Ясно, что математическое ожидание  $Q_\infty(f)$  зависит от используемой меры близости  $f$ . Укажем один класс распределений случайного отношения  $Q_1$ , такой, что эта зависимость не столь велика.

3.4. *Монотонные распределения.* Распределение  $P$  случайного отношения  $Q_1$  называется монотонным относительно меры близости  $f$  с центром в  $Q_0$ , если из  $f(Q', Q_0) < f(Q'', Q_0)$  следует, что  $P(Q_1 = Q') > P(Q_1 = Q'')$ . Тогда  $Q_\infty(f) = Q_0$  [29, с. 218]. Таким образом, для монотонных распределений математическое ожидание совпадает с модой. Этот результат справедлив для таких пространств бинарных отношений, которые обладают свойством изотропности [29, с. 218]: пространство  $A$  с мерой близости  $f$  называется изотропным, если для любого  $a \in A$  множество чисел

$\{f(a, b), b \in A\}$  является одним и тем же. Причем, если при некотором  $a \in A$  равенство  $f(x, a) = t$  выполнено ровно для  $k$  элементов  $x \in A$ , то и при любом другом  $b \in A$  равенство  $f(x, b) = t$  выполнено ровно для  $k$  элементов  $x$  из  $A$ , и так для всех неотрицательных  $t$ . Пространства разбиений, упорядочений, ранжировок, толерантностей — изотропные. Математическое ожидание совпадает с модой для случайной величины со значениями в любом изотропном пространстве, не только в пространствах отношений.

Требование изотропности является существенным: если нет изотропности, то математическое ожидание может не совпадать с модой. Соответствующий пример строится так: пусть пространство состоит из большого числа близко расположенных друг к другу точек и точки  $x_0$ , далеко отстоящей от всех остальных. Пусть распределение монотонно с центром в  $x_0$ , причем вероятности весьма мало убывают с увеличением расстояния от  $x_0$ . Тогда математическое ожидание будет лежать не в  $x_0$ , а среди точек, близко друг к другу расположенных.

Понятие монотонности является естественным, если мы рассматриваем наблюдения  $Q_1, Q_2, \dots, Q_n$  как результат измерения истинного значения  $Q_0$ . Измерения проводятся со случайной ошибкой, поэтому  $Q_i$  обычно не совпадают с  $Q_0$ . Однако, как в классической теории ошибок измерений, естественно принять, что чем больше отклонение, тем реже оно встречается. Формализация этого требования — условие монотонности.

Пусть распределение  $P$  монотонно с центром в  $Q_0$  относительно меры близости  $f$ . Пусть  $h$  — строго возрастающая функция, отображающая луч  $[0, +\infty)$  на себя. Тогда, очевидно, распределение  $P$  будет монотонно относительно  $h(f)$  с тем же центром  $Q_0$ . Поскольку математическое ожидание совпадает с центром, то  $Q_\infty(f) = Q_\infty(h(f))$  для любой функции  $h$ , удовлетворяющей указанным выше условиям. В частности, для монотонного распределения медиана Кемени и среднее по Кемени имеют один и тот же предел при  $n \rightarrow \infty$ , а именно центр распределения.

Пусть рассматривается выборка из распределения, монотонно относительно меры близости  $f$  с центром в  $Q_0$ . Тогда для оценки  $Q_0$  по выборке можно использовать  $Q_n(h(f))$  при различных  $h$ . Возникают вопросы, на которые пока нет ответа: с какой скоростью  $Q_n(h(f))$  сходится к  $Q_0$ ? Какое  $h$  наиболее предпочтительно? В частности, чем лучше пользоваться — медианой Кемени или средним по Кемени?

3.5. *Обобщения.* Пусть задано расстояние  $\rho$ . Какие меры близости  $t(\rho)$  разумно использовать? На этот вопрос позволяет ответить теория измерений. Как показано выше, естественно минимизировать по всем возможным элементам  $B$  среднее расстояние от  $B$  до элементов выборки  $Q_1, Q_2, \dots, Q_n$ . Какое среднее для чисел  $d(Q_1, B), d(Q_2, B), \dots, d(Q_n, B)$  вычислять? Естественные аксиомы [8] приводят к так называемым ассоциативным средним. Как показали А. Н. Колмогоров, М. Нагумо и Б. де Финетти

в начале 1930-х годов [8], все ассоциативные средние имеют вид (10) с  $a_i = 1/n$ ,  $i = 1, 2, \dots, n$ . Далее, расстояния естественно считать измеренными в шкале отношений. Поэтому естественно потребовать устойчивости результата сравнения средних расстояний при переходе от  $\rho$  к расстоянию  $a\rho$  при произвольном  $a > 0$ . По соответствующей теореме теории измерений (см. выше) это требование приводит к степенной функции  $t$ . Таким образом, в качестве меры близости  $t(d)$  можно рассматривать лишь положительные степени исходного расстояния  $d$ , т. е.  $t(d) = d^p$ ,  $p > 0$  [29].

Различные наблюдения используются при построении  $Q_n(f)$  симметрично. Однако может оказаться полезным одни наблюдения учитывать с большим весом, чем другие. Это приводит к построению среднего по правилу

$$T_n(f, \mu) = \operatorname{Argmin}_Q \sum_{i=1}^n \mu_i f(Q_i, Q), \quad (35)$$

где  $\mu_i > 0$ ,  $i = 1, 2, \dots, n$ . Константы  $\mu_i$  могут, например, измерять уровень компетенции экспертов. Ясно, что, как правило,  $T_n(f, \mu) \neq Q_n(f)$ . Однако, если наблюдения независимы и одинаково распределены, то справедливы аналогии законов больших чисел:  $T_n(f, \mu)$  сходятся к  $Q_\infty(f)$  при  $n \rightarrow \infty$ .

В рассмотренных результатах типа законов больших чисел функция  $f$  имела смысл меры близости и была определена на декартовом квадрате  $X \times X$ , где  $X$  — пространство, в котором лежат рассматриваемые величины. Но анализ доказательств показывает, что они остаются справедливыми и тогда, когда  $f$  определена на  $X \times Y$ , где  $X$  и  $Y$  — некоторые множества, не обязательно совпадающие. В частности, справедлив следующий аналог закона больших чисел (усиленного).

**Теорема 2** [31]. Пусть  $Z$  и  $W$  конечные множества:  $\xi_1, \xi_2, \dots, \dots$ , — независимые в совокупности случайные величины со значениями в  $Z$ . Пусть  $g: Z \times W \rightarrow R^1$ . Тогда с вероятностью 1 существует (случайное) натуральное число  $n_0$ , такое, что при  $n > n_0$  каждое решение задачи

$$\sum_{1 \leq i \leq n} g(\xi_i, w) \rightarrow \min_w, \quad w \in W, \quad (36)$$

является также решением задачи

$$Eg(\xi_1, w) \rightarrow \min_w, \quad w \in W. \quad (37)$$

## 4. Статистика в пространствах общей природы

4.1. Асимптотика решений экстремальных статистических задач. Можно отказаться от предположения конечности  $Z$  и  $W$ . Пусть  $\xi_1, \xi_2, \dots, \xi_n, \dots$  — случайные величины со значениями в пространстве  $X$ , а действительная функция  $f$  определена на про-

изведении пространств  $X$  и  $Y$ . При этом  $Y$  интерпретируется как пространство параметров. Оценкой минимального контраста называется

$$y_n(f) = \operatorname{Argmin}_{y \in Y} \sum_{i=1}^n f(\xi_i, y). \quad (38)$$

Рассмотрим следующий параметр распределения  $\xi_1$ :

$$y_\infty(f) = \operatorname{Argmin}_{y \in Y} Ef(\xi_1, y). \quad (39)$$

Тогда при  $n \rightarrow \infty$

$$y_n(f) \rightarrow y_\infty(f), \quad (40)$$

т. е.  $y_n(f)$  является состоятельной оценкой параметра  $y_\infty(f)$ . О придании точного смысла понятию сходимости в (40) см. [36].

Если  $X = Y = R^1$ , то при  $f(a, b) = (a - b)^2$  в роли  $y_n(f)$  выступает среднее арифметическое результатов наблюдений, а в роли  $y_\infty(f)$  — их математическое ожидание. Тогда (40) — запись обычного закона больших чисел. Если  $f(a, b) = |a - b|$ , то

$$y_\infty(f) = \{x: F(x) \leq 1/2, F(x+) \geq 1/2\}, F(x) = P(\xi_1 < x). \quad (41)$$

В этом случае  $y_\infty(f)$  естественно назвать медианой распределения  $\xi_1$ . Для некоторых распределений медиана не точка, а отрезок. Решение задачи (48) описывается так: если  $n$  нечетно, то

$$y_n(f) = \left\{ x \left( \left[ \frac{n}{2} \right] + 1 \right) \right\}, \quad (42)$$

если же  $n$  четно, то

$$y_n(f) = \left[ x \left( \left[ \frac{n}{2} \right] \right), x \left( \left[ \frac{n}{2} \right] + 1 \right) \right], \quad (43)$$

где  $x(i)$  есть  $i$ -я порядковая статистика совокупности  $\xi_1, \xi_2, \dots, \xi_n$ . Соотношение (40) выражает сходимость выборочной медианы к теоретической. При этом обе части (40) в общем случае не числа, а множества.

Частными случаями оценок минимального контраста являются оценки максимального правдоподобия, устойчивые оценки Тьюки—Хубера [43], оценки параметров в задаче аппроксимации (параметрической регрессии). Состоятельность оценок минимального контраста означает состоятельность всех перечисленных оценок, а также справедливость законов больших чисел. Поэтому каждая общая теорема о справедливости в том или ином смысле соотношения (40) влечет за собой соответствующие следствия, касающиеся перечисленных конкретных областей. Так, в задаче конструирования факторов [22, гл. 4] (40) описывает поведение отношения, аппроксимирующего систему матриц.

Имеет смысл и дальнейшее обобщение. Как показано в [3], основные задачи прикладной статистики допускают экстремаль-

ную формулировку. Это значит, что они сводятся к нахождению параметра, лежащего в некотором пространстве  $Y$ , по правилу

$$y(f_n) = \underset{y \in Y}{\operatorname{Argmin}} f_n(\xi_1, \xi_2, \dots, \xi_n; y). \quad (44)$$

Часто (44) совпадает с (38), однако минимизируемая функция не всегда допускает аддитивное представление, как в (38). Рассмотрим, например, метод главных компонент.

4.2. *Применение к методу главных компонент.* Исходные данные — набор  $\xi_1, \xi_2, \dots, \xi_n$  случайных векторов, лежащих в евклидовом пространстве  $R^p$ . Цель состоит в снижении размерности, т. е. уменьшении числа рассматриваемых показателей. Для этого берут всевозможные линейные ортогональные нормированные центрированные комбинации исходных показателей, получают  $p$  новых показателей, из них берут первые  $p'$ , где  $p' < p$ . Матрицу преобразования  $C$  выбирают так, чтобы максимизировать информационный функционал [3, с. 38]

$$I_n(C) = \frac{s^2(z(1)) + s^2(z(2)) + \dots + s^2(z(p'))}{s^2(x(1)) + s^2(x(2)) + \dots + s^2(x(p))}, \quad (45)$$

где  $x(i)$ ,  $i = 1, \dots, p$ , — исходные показатели;  $\xi_j = (x_j(1), \dots, x_j(p))$ ,  $j = 1, 2, \dots, n$ , в то время как  $z(i)$ ,  $i = 1, 2, \dots, p'$ , — комбинации исходных показателей, полученные с помощью матрицы  $C$ . Наконец,  $s^2(z(i))$ ,  $i = 1, \dots, p'$ ,  $s^2(x(i))$ ,  $i = 1, \dots, p$ , — выборочные дисперсии переменных, указанных в скобках.

Укажем подробнее, как  $z(i)$  строится по  $x(i)$  с помощью матрицы  $C$ :

$$z_j(i) = \sum_{k=1}^p c_{ik} (x_j(k) - \overline{x(k)}), \quad i = 1, \dots, p', \quad j = 1, \dots, n, \quad (46)$$

где

$$\overline{x(k)} = \frac{1}{n} \sum_{j=1}^n x_j(k). \quad (47)$$

Матрица  $C = \|c_{ik}\|$  порядка  $p' \times p$  такова, что

$$\sum_{k=1}^p c_{ik}^2 = 1, \quad i = 1, 2, \dots, p' \quad (48)$$

(нормированность),

$$\sum_{k=1}^p c_{ik} c_{mk} = 0, \quad i, m = 1, 2, \dots, p', \quad i \neq m \quad (49)$$

(ортогональность). Решением основной задачи метода главных компонент является

$$C_n = \underset{C}{\operatorname{Argmin}} (-I_n(C)), \quad (50)$$

где  $I_n$  определяется по (45), а минимизация проводится по всем матрицам  $C$ , удовлетворяющим условиям (48), (49).

Вычисление матрицы  $C_n$  — задача анализа данных. Однако, как и для рассмотренной выше медианы Кемени, возникает вопрос об асимптотическом поведении  $C_n$ . Является ли решение основной задачи метода главных компонент устойчивым, т. е. существует ли предел  $C_n$  при  $n \rightarrow \infty$ ? Чему равен этот предел?

Ответ, как обычно, может быть дан только в вероятностной теории. Пусть  $\xi_1, \xi_2, \dots$  — независимые одинаково распределенные случайные вектора. Аналогично (46) положим,

$$z_{\infty}(i) = \sum_{k=1}^p c_{ik} (x_1(k) - Ex_1(k)), \quad i = 1, \dots, p', \quad (51)$$

где матрица  $C = \|c_{ik}\|$  удовлетворяет соотношениям (48), (49). Положим,

$$I(C) = \frac{D(z_{\infty}(1)) + D(z_{\infty}(2)) + \dots + D(z_{\infty}(p'))}{D(x(1)) + D(x(2)) + \dots + D(x(p))}. \quad (52)$$

Легко видеть, что при  $n \rightarrow \infty$  и любом  $C$

$$I_n(C) \rightarrow I(C). \quad (53)$$

Рассмотрим решение предельной экстремальной задачи

$$C_{\infty} = \underset{C}{\operatorname{Argmin}} (-I(C)). \quad (54)$$

Естественно ожидать, что

$$\lim_{n \rightarrow \infty} C_n = C_{\infty}. \quad (55)$$

Действительно, (55) вытекает из общих результатов об асимптотическом поведении решений экстремальных статистических задач [36], описываемых ниже. Однако в литературе по методу главных компонент (см., например, обзор [4]) теоремы о справедливости соотношения (55) обнаружить не удалось.

4.3. *Статистика объектов общей природы.* В ряде других задач прикладной статистики решение находится путем минимизации функционала, также не являющегося аддитивным. Иными словами, это решение не является оценкой минимального контраста (38). Таковы различные варианты задач классификации, решаемые путем минимизации функционала качества [3, с. 33—37]. К этому же типу относятся факторный анализ, метод экстремальной группировки признаков, отбор наиболее информативных признаков в модели дискриминантного анализа, отбор наиболее информативных переменных в моделях регрессии, скалярная редукция многокритериальной оптимизационной схемы, т. е. экспертно-статистический метод построения единого интегрального показателя «качества» в случае, когда экспертная информация — ранжировки, разбиения или результаты парных сравнений [3, с. 38—45]. Во всех перечисленных задачах решение ищется по правилу (44).

Представляется целесообразным получить общий результат следующего вида. Пусть  $Y$  — произвольное пространство,  $\Omega = \{\omega\}$  — пространство элементарных событий,  $f_n: Y \times \Omega \rightarrow R^1$ ,  $n = 1, 2, \dots$  — последовательность случайных функций на  $Y$ . Пусть существует функция  $f: Y \rightarrow R^1$ , такая, что для любого  $y \in Y$

$$\lim_{n \rightarrow \infty} f_n(y, \omega) = f(y) \quad (56)$$

(сходимость в (56) можно понимать в двух смыслах: по вероятности или с вероятностью 1). Тогда при  $n \rightarrow \infty$

$$\operatorname{Argmin}_{y \in Y} f_n(y, \omega) \rightarrow \operatorname{Argmin}_{y \in Y} f(y). \quad (57)$$

Сходимость в (57) понимается в том смысле, что и в (56).

Этот результат получен в [36] при некоторых условиях регулярности, наложенных на последовательность  $f_n$ . Удастся получить аналог (57) и в случае, когда  $\operatorname{Argmin} f_n$  и  $\operatorname{Argmin} f$  состоят более чем из одного элемента или пусты. Состоятельность оценок минимального контраста [40] получается как следствие. Результаты [29, 36] о состоятельности оценок минимального контраста и законах больших чисел обобщают результаты предшественников [52, 53].

К статистике в пространствах общей природы мы пришли, исходя из статистики бинарных отношений. Затем оказалось, что и для многих классических постановок результаты типа (57) полезны. Поэтому целесообразно развивать статистику объектов общей природы как самостоятельное направление. Рассмотрим следующие постановки: итеративные устойчивые оценки, непараметрическое оценивание плотности и регрессии, предельная теория статистик интегрального типа.

4.4. *Итеративные устойчивые оценки.* В [46] описан психофизический эксперимент, в котором испытуемые ранжировали объекты. Истинная ранжировка была известна экспериментаторам. При анализе результатов эксперимента в [46] показано, что ответы разбиваются на многочисленное «ядро», расположенное вокруг истинного мнения, и отдельные мнения, разбросанные на периферии. Оказалось, что ранжировка, полученная путем усреднения «периферийных» мнений, противоположна истинной. Это значит, что оценка истинного мнения по «ядру» лучше, чем по всей совокупности, поскольку в последнем случае учитываются грубо ошибочные мнения. Значит, целесообразно развивать методы, позволяющие уменьшать влияние «периферийных» наблюдений на оценку параметра.

В [46] порог, выделяющий ядро, выбран из эвристических соображений. В [37] для обработки результатов парных сравнений и, более общо, дихотомических данных предложены статистически обоснованные методы выбора порога.

В классической статистике есть методы, уменьшающие влияние «периферийных» наблюдений на выводы. В одних из них подобные

наблюдения отбрасываются, в других — учитываются с меньшим весом. Первые методы соответствуют выделению «ядра», примером которых являются устойчивые оценки Тьюки—Хубера [43].

Приведем взвешенный итеративный метод оценки параметра [40] (ВИМОП). Пусть  $\xi_1, \xi_2, \dots, \xi_n$  — наблюдения, лежащие в пространстве  $X$ . Пусть  $Y$  — пространство параметров,  $\rho: X \times Y \rightarrow R^1$  — функция контраста,  $\varphi: R^1 \rightarrow [0, 1]$  — монотонно невозрастающая функция. Рассмотрим итерационный процесс

$$y^{m+1} = \operatorname{Argmin}_{y \in Y} \sum_{i=1}^n \rho(\xi_i, y) \varphi(\rho(\xi_i, y^m)), \quad m=0, 1, 2, \dots, \quad (58)$$

где  $y^0$  — некоторая точка из  $Y$ , например,  $y^0 = y_n(\rho)$  — см. (38). Оценкой по методу ВИМОП называется

$$y_n = \lim_{m \rightarrow \infty} y^m \quad (59)$$

(если, конечно, предел существует).

Укажем два частных случая ВИМОП-оценок. Если  $X$  — конечномерное евклидово пространство,  $\varphi$  — степень нормальной плотности, то ВИМОП-оценки сводятся к  $\lambda$  — моментам [20]. Если  $X$  — конечномерное евклидово пространство,  $\rho$  — квадрат евклидова расстояния,  $X = Y$ , функция  $\varphi$  является пороговой, т. е.  $\varphi(r) = 1$  при  $0 \leq r \leq r_0$  и  $\varphi(r) = 0$  при  $r > r_0$ , то процесс построения ВИМОП-оценки есть процесс нахождения центра тяжести класса с помощью алгоритма классификации «Форель» [2]. Таким образом, столь разные объекты, как устойчивые оценки ( $\lambda$ -моменты) и эталонные алгоритмы классификации, оказываются частными случаями одной и той же общей схемы (58) — (59). Если  $\varphi$  — пороговая, то существование предела в (59) вытекает из результатов [28].

4.5. *Задача аппроксимации зависимости.* Рассмотрим задачу аппроксимации (параметрической регрессии). Пусть  $X$  и  $Y$  — некоторые пространства. Пусть имеется  $n$  пар  $(x_i, y_i)$ , где  $x_i \in X$ ,  $y_i \in Y$ ,  $i = 1, 2, \dots, n$ . Требуется подобрать параметр  $\theta \in \Theta$  так, чтобы  $g(x_i, \theta)$  наилучшим образом приближали  $y_i$ ,  $i = 1, 2, \dots, n$ . Пусть мера близости между  $g(x_i, \theta)$  и  $y_i$  задается с помощью функции  $f_i$ . Здесь  $g(x, \theta)$  — заданное семейство функций, отображающих  $X \times \Theta$  в  $Y$ , а  $f_i$ ,  $i = 1, 2, \dots$ , — заданная последовательность функций. Часто, но не всегда, все  $f_i$  совпадают. В классической постановке [1] функции  $f_i$  различны при неравноточных наблюдениях. При сделанных предположениях параметр естественно подбирать по правилу

$$\theta_n = \operatorname{Argmin}_{\theta \in \Theta} \sum_{i=1}^n f_i(g(x_i, \theta), y_i). \quad (60)$$

В рамках анализа данных остается единственный вопрос — о существовании  $\theta_n$ . Если все участвующие в (60) функции непрерывны, а минимум берется по компакту, то  $\theta_n$  существует. Есть и иные условия существования  $\theta_n$ .

При появлении нового наблюдения  $x$  рекомендуется выбирать оценку для соответствующего  $y$  по правилу

$$y^* = g(x, \theta_n). \quad (61)$$

Обосновать такую рекомендацию в рамках анализа данных невозможно. Это можно сделать только в вероятностной теории, равно как и изучить асимптотическое поведение  $\theta_n$ , доказать состоятельность этой оценки.

Как и в классическом случае, вероятностную теорию целесообразно строить для трех различных постановок.

1. Переменная  $x$  — детерминированная, переменная  $y$  — случайная, ее распределение зависит от  $x$ . Это схема В в терминологии [1].

2. Совокупность  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , — выборка из распределения случайной величины со значениями в  $X \times Y$  (схема  $C_1$  из [1]).

3. Имеется детерминированный набор пар  $(x_i^0, y_i^0)$ ,  $i = 1, 2, \dots, n$ , наблюдение  $(x_i, y_i)$  является случайной величиной, распределение которой зависит от  $(x_i^0, y_i^0)$ ,  $i = 1, 2, \dots, n$  (конфликтный анализ — схема  $C_2$  в [1]).

Во всех трех случаях на основе общих результатов типа (57) можно получить утверждения об асимптотическом поведении  $\theta_n$  из (60) и  $y^*$  из (61). Оценки оказываются состоятельными. Отдельные формулировки для второго случая приведены в [36], для регрессионного анализа нечетких множеств — в [33].

Соотношение (60) дает решение задачи аппроксимации. Поясним, как эта задача соотносится с нахождением регрессии. Согласно [30], для случайной величины  $(\xi, \eta)$  со значениями в  $X \times Y$ , регрессией  $\eta$  на  $\xi$  относительно меры близости  $f$  естественно называть решение задачи

$$Ef(g(\xi), \eta) \rightarrow \min_g, \quad f: Y \times Y \rightarrow R^1, \quad g: X \rightarrow Y, \quad (62)$$

где минимум берется по множеству всех измеримых функций. Можно исходить из другого определения. Для каждого  $x \in X$  рассмотрим случайную величину  $\eta(x)$ , распределение которой является условным распределением  $\eta$  при условии  $\xi = x$ . Положим,

$$E(\eta | \xi = x) = \underset{y \in Y}{\operatorname{Argmin}} Ef(y, \eta(x)). \quad (63)$$

Оказывается, при обычных предположениях измеримости решение задачи (62) совпадает с  $E(\eta | \xi = x)$ . (Оговорки типа «почти всюду» в этих кратких формулировках опущены.) Если заранее известно, что  $E(\eta | \xi = x)$  принадлежит параметрическому семейству  $g(x, \theta)$ , то задача нахождения регрессии сводится к оценке параметра  $\theta$  и решение дается формулами (60), (61) — вторая постановка вероятностной теории. Если же нет оснований считать, что регрессия принадлежит параметрическому семейству,

то можно использовать непараметрические оценки регрессии. Они строятся с помощью непараметрических оценок плотности. Укажем некоторые из них.

4.6. *Непараметрические оценки плотности.* Пусть  $\xi$  — случайная величина со значениями в  $X$ , а  $\mu$  — некоторая мера на той  $\sigma$ -алгебре подмножеств  $X$ , которая участвует в определении  $\xi$  как измеримого отображения вероятностного пространства в  $X$  (случайные величины в пространствах общей природы определяются в соответствии с [41, гл. III]). Функция  $g: X \rightarrow R^1$  называется плотностью  $\xi$  (относительно меры  $\mu$ ), если

$$P(\xi \in A) = \int_A g(x) \mu(dx) \quad (64)$$

для любого  $A$  из рассматриваемой  $\sigma$ -алгебры.

Если есть основания считать, что плотность зависит от некоторого параметра, то оценивание плотности сводится к оцениванию этого параметра, скажем, методом максимального правдоподобия. Если же подобных оснований нет, то необходимо использовать непараметрические методы. Скорость сходимости непараметрической оценки к истинной плотности меньше, чем параметрической оценки.

Пусть  $\xi_1, \xi_2, \dots, \xi_n$  — независимые случайные величины со значениями в  $X$ , имеющие плотность  $g(x)$ . В качестве оценки  $g(x)$  можно использовать оценки

$$g_n(x) = \frac{1}{n} \sum_{i=1}^n h_n(x, \xi_i), \quad (65)$$

называемые линейными, где  $K_n: X^2 \rightarrow R^1$ . В [38] приведены условия состоятельности линейных оценок, а также нелинейных оценок типа Фикс-Ходжеса. Рассмотрим подробнее частный случай линейных оценок — те, которые называем оценками Парзена—Розенблатта, перенося на общий случай название, принятое для конечномерных случайных величин. Они имеют вид

$$g_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{\rho(x, \xi_i)}{h_n}\right), \quad K: R^1 \rightarrow R^1, \quad (66)$$

где  $\{h_n\}$  — некоторая последовательность чисел;  $\rho(x, y)$  — мера близости в  $X$ . Легче описать предельное поведение оценок (66), если непрерывная по  $t$  функция

$$M(x, t) = \mu\{y: y \in X, \rho(x, y) \leq t\}, \quad x \in X, t \geq 0, \quad (67)$$

не зависит от  $x$  (аналог свойства изотропности), поскольку в этом случае можно с помощью монотонного преобразования перейти к мере близости  $\rho_1$ , такой, что

$$M(x, t) = t \quad (68)$$

для любого  $t \in [0, t_0]$ , где  $t_0 \leq +\infty$ . Случай конечного  $t_0$  соответствует тому, что  $\mu(X) = t_0 < \infty$ , в то время как  $t_0 = +\infty$  тогда и только тогда, когда  $\mu(X) = +\infty$ .

*Теорема 3* [38]. Пусть плотность  $g(x)$  ограничена на  $X$  и непрерывна в точке  $x \in X$  (в топологии, порожденной мерой близости  $\rho_1$ ). Пусть функция  $K$  такова, что

$$\int_0^\infty K(u) du = 1, \quad \int_0^\infty |K(u)| du < \infty, \quad \int_0^\infty K^2(u) du < \infty. \quad (69)$$

Пусть

$$h_n \rightarrow 0, \quad nh_n \rightarrow \infty \quad (70)$$

при  $n \rightarrow \infty$ . Тогда для оценки (66) с  $\rho = \rho_1$

$$\lim_{n \rightarrow \infty} E g_n(x) = g(x) \quad (71)$$

и

$$\lim_{n \rightarrow \infty} n h_n D g_n(x) = g(x) \int_0^\infty K^2(u) du. \quad (72)$$

Из (71) и (72) вытекает, что  $g_n(x)$  является состоятельной оценкой  $g(x)$ . Более того,  $g_n(x)$  — асимптотически нормальная оценка. Рассмотрим величину

$$\alpha_n = E(g_n(x) - g(x))^2. \quad (73)$$

Можно показать, что  $\alpha_n$  достигает минимума, равного  $n^{-1/2}$  при  $h_n = n^{-1/2}$ . (В параметрическом случае  $\alpha_n = \text{const} \cdot n^{-1}$ .) Для обычных оценок Парзена—Розенблатта ( $X = R^1$ ) оптимальная скорость сходимости выше: она равна  $n^{-1/3}$  при  $h_n = n^{-1/3}$ . Это связано с тем, что в общем случае величина

$$C = \frac{d^2}{dt^2} P\{\xi \in \{y: y \in X, \rho(x, y) \leq t\}\} |_{t=0} \quad (74)$$

не равна 0, а для обычных оценок она равна 0. Если  $C = 0$ , то оптимальный порядок  $h_n$  и скорость сходимости будут такими же, как в классическом случае, в котором лебегова мера, евклидово расстояние и разложение в ряд Тейлора «удачно» связаны между собой. Равенство 0 величины  $C$  из (74) эквивалентно тому, что при интегрировании по интервалу разложения Тейлора некоторой функции, взятого в центре интервала, члены с нечетными производными сокращаются. Если же для одномерных случайных величин брать плотность не по мере Лебега, а по произвольной, то, вообще говоря,  $C \neq 0$  и оптимальный порядок сходимости достигается при  $h_n = n^{-1/2}$ .

**4.7. Статистики интегрального типа.** Предельная теория статистик интегрального типа была развита автором первоначально для конечномерных случайных величин. Первый конкретный результат касался асимптотики распределения статистики типа

омега-квадрат, используемой для проверки симметрии распределения результатов наблюдений относительно 0 [25]. Общая теория развита в [26, 29, подразд. 2.4]. Статистиками интегрального типа называются статистики, представимые в виде

$$\beta_n = \int_X f_n(x, \omega) \mu_n(dx), \quad n = 1, 2, \dots, \quad (75)$$

где  $x$  — точка пространства  $X$  произвольной природы;  $f_n(x, \omega)$  — случайный процесс;  $\mu_n$  — случайная мера, обычно вероятностная. В соответствии с рассматриваемыми приложениями предполагается, что  $f_n(x, \omega)$  сходится по распределению при  $n \rightarrow \infty$  к случайному процессу  $f(x, \omega)$ , а  $\mu_n$  сходится по вероятности к детерминированной мере  $\mu$ . Изучается сходимость  $\beta_n$  по распределению к

$$\beta = \int_X f(x, \omega) \mu(dx). \quad (76)$$

4.8. *Непараметрические оценки регрессии.* Перейдем к непараметрическим оценкам регрессии. Они основаны на использовании непараметрических оценок плотности. Рассмотрим случайную величину  $(\xi, \eta)$ , где  $\xi \in X$ ,  $\eta \in Y$ ,  $(\xi, \eta) \in X \times Y$ . Пусть  $\nu_1$  — мера в  $X$ ,  $\nu_2$  — мера в  $Y$ , а их прямое произведение  $\nu = \nu_1 \times \nu_2$  — мера в  $X \times Y$ . Пусть  $g(x, y)$  — плотность  $(\xi, \eta)$  по мере  $\nu$ . Тогда условная плотность  $g(y | x)$  распределения  $\eta$  при условии  $\xi = x$  имеет вид

$$g(y | x) = \frac{g(x, y)}{\int_Y g(x, y) \nu_2(dy)} \quad (77)$$

(в предположении, что интеграл в знаменателе отличен от 0). В соответствии с (63) имеем

$$E(\eta | \xi = x) = \underset{y \in Y}{\operatorname{Argmin}} \int_Y f(y, a) g(a | x) \nu_2(da). \quad (78)$$

Заменяя  $g(x, y)$  в (77) непараметрической оценкой плотности  $g_n(x, y)$ , получаем оценку условной плотности

$$g_n(y | x) = \frac{g_n(x, y)}{\int_Y g_n(x, y) \nu_2(dy)}. \quad (79)$$

Если  $g_n$  — состоятельная оценка  $g$ , то числитель (79) сходится к числителю (77). Сходимость знаменателя (79) к знаменателю (77) обосновывается с помощью предельной теории статистик интегрального типа. В итоге получаем состоятельность непараметрической оценки условной плотности.

Непараметрическая оценка регрессии ищется как

$$E_n(\eta | \xi = x) = \underset{y \in Y}{\operatorname{Argmin}} \int_Y f(y, a) g_n(a | x) \nu_2(da). \quad (80)$$

Состоятельность оценки (80) следует из общих результатов об асимптотическом поведении решений экстремальных статистических задач (см. (56), (57)).

Непараметрические оценки плотности используются и в других задачах, в частности, в дискриминантном анализе.

Выше мы показали, что из рассмотрения задач статистики бинарных отношений вытекает целесообразность развития статистики в пространствах общей природы. Актуальной в настоящее время является обратная задача — адаптация общих результатов о задачах аппроксимации, непараметрических оценках плотности (вероятности попасть в точку), регрессии и т. п. применительно к бинарным отношениям.

## 5. Специальные модели

5.1. *Парные сравнения и бинарные отношения.* Выше мы рассматривали бинарные отношения как непосредственно данные. Однако в ряде случаев естественно предполагать существование вероятностной модели, порождающей бинарные отношения. Обзор дан в [47, 48]. Опишем две новые модели.

Пусть  $k$  объектов описываются векторами в конечномерном пространстве. Будем считать, что один объект предпочитается другому, если каждая координата вектора, соответствующего первому объекту, меньше соответствующей координаты вектора, соответствующего второму объекту. Если ни один из двух объектов не предпочитается другому, то считаем, что они эквивалентны. В вероятностной теории считаем, что вектор, соответствующий объекту, является суммой «истинного» вектора и случайной ошибки, имеющей круговое нормальное распределение с нулевым средним и единичной дисперсией каждой координаты. Основная задача в вероятностной теории — оценка «истинных» векторов. Классическая модель Терстоуна [47, 48] — частный случай, в котором объекты описываются числами, т. е. одномерными векторами. Параметры модели оцениваются, как в классическом случае, методом максимального правдоподобия, который дает асимптотически эффективные оценки. Важной является также задача проверки согласованности наблюдаемых данных с моделью. Проверка осуществляется с помощью критерия отношения правдоподобия, который приводит к статистикам типа хи-квадрат.

Вторая модель — модель сходства. Объекты описываются точками в конечномерном пространстве. Они считаются сходными, если расстояние между точками, соответствующими двум объектам, не превосходит фиксированного числа, например 1. Конфигурация точек порождает толерантность на множестве объектов. Беря транзитивное замыкание толерантности, получаем разбиение. В вероятностной модели предполагается, что вектор, соответствующий объекту, есть сумма «истинного» вектора и случайной ошибки. Как и в предыдущей модели, параметры можно

оценивать методом максимального правдоподобия, а согласие проверять с помощью критерия отношения правдоподобия, приводящего к статистикам типа хи-квадрат.

Следующий важный вид объектов нечисловой природы — результаты парных сравнений. В вероятностной теории они рассматриваются как совокупность независимых испытаний Бернулли. В параметрической теории парных сравнений вероятности «успехов» выражаются определенным образом через параметры, число которых мало по сравнению с объемом выборки. Оценка параметров и проверка согласия опытных данных с моделью проводятся обычными статистическими методами [47, 48].

Непараметрическая теория парных сравнений развивается с 1975 г. [27]. Основные результаты изложены в [29, 31, 37]. Они касаются в основном проверки согласованности в асимптотике Колмогорова. А именно, пусть  $\xi_{ij}$ ,  $i = 1, 2, \dots, k$ ,  $j = 1, \dots, n$ , — независимые в совокупности испытания Бернулли с вероятностями успеха  $p_{ij}$ ,  $i = 1, \dots, k$ ,  $j = 1, \dots, n$ , соответственно. Нулевая гипотеза такова:

$$H_0: p_{1j} = p_{2j} = \dots = p_{kj}, j = 1, 2, \dots, n. \quad (81)$$

Требуется проверить гипотезу (81) в асимптотике  $k = \text{const} \geq 4$ ,  $n \rightarrow \infty$ . При этом число неизвестных параметров равно  $n$  и растет при  $n \rightarrow \infty$  пропорционально общему числу наблюдений  $kn$ , что служит причиной неклассических трудностей. Теория случайных толерантностей является частным случаем непараметрической теории парных сравнений.

Как известно [47], в случае ранжировок и разбиений методы проверки согласованности развиты лишь для случая равномерных распределений на рассматриваемом множестве бинарных отношений. Непараметрическая теория парных сравнений и теория толерантностей, как видно из (81), свободны от этого недостатка. Равномерность соответствует приравниванию  $1/2$  всех величин, упомянутых в (81). Это указывает на целесообразность более широкого применения методов проверки согласованности парных сравнений, т. е. гипотезы (81), по сравнению с методами проверки согласованности ранжировок.

Проиллюстрируем последний тезис примером. Пусть требуется найти ранжировку, выражающую групповое мнение экспертов, используя результаты парных сравнений, полученные от экспертов. Необходимо проверить согласованность экспертов. Возможны два способа действий.

Можно взять результаты парных сравнений, данные одним экспертом, и по ним восстановить ранжировку, отражающую его мнение. При этом придется «исправить» результаты некоторых парных сравнений, поскольку последние не обязаны обладать свойством транзитивности. После этого проверить согласованность ранжировок.

Во втором способе сначала проверяется согласованность результатов парных сравнений, т. е. проверяется гипотеза (81),

и затем в случае принятия этой гипотезы находится результирующая ранжировка.

На наш взгляд, второй способ более обоснован, чем первый. Исторически сначала был разработан анализ ранжировок [14], но теперь при проверке согласованности на первый план должны выйти методы проверки гипотезы (81) в непараметрической теории независимых парных сравнений.

Можно развивать теорию зависимых парных сравнений, в которой результат следующих сравнений зависит от результатов предыдущих, что позволяет учесть «тягу» экспертов к транзитивности ответов.

С точки зрения социолога, теория парных сравнений — это теория обработки ответов на закрытые вопросы, имеющие две подсказки. Было бы полезно развить аналогичную теорию для случая вопросов с несколькими подсказками, из которых надо выбрать одну. Однако это пока не сделано. Есть и другие заманчивые постановки. Например, когда из  $k$  подсказок надо выбрать  $m$ . Здесь есть два варианта:  $m$  задано социологом;  $m$  выбирается респондентом.

5.2. *Случайные и нечеткие множества.* Как уже отмечалось, бинарные отношения представляют собой подмножества декартова квадрата множества объектов, между которыми устанавливаются эти отношения, а потому случайные бинарные отношения есть случайные множества. Последние естественным образом возникают во многих прикладных и теоретических задачах. Из теории случайных множеств [29, гл. 4] приведем лишь один результат — аксиоматическое введение расстояния между множествами. Приведем систему аксиом типа использованной Кемени и Снеллом [13].

Пусть  $\mathfrak{M}$  — алгебра множества с единицей  $X$  (другими словами,  $\mathfrak{M}$  — система подмножеств множества  $X$ , такая, что  $X \in \mathfrak{M}$  и из принадлежности множеств  $A$  и  $B$  системе  $\mathfrak{M}$  следует принадлежность к этой системе их объединения  $A \cup B$ , пересечения  $A \cap B$  и разности  $A \setminus B$ ).

Естественно в качестве расстояния между множествами  $A$  и  $B$  использовать лишь функции, удовлетворяющие перечисленным ниже аксиомам.

*Аксиома 1.* Функция  $d: \mathfrak{M}^2 \rightarrow R^1$  является псевдометрикой, т. е. для любых  $A, B$  и  $C$  из  $\mathfrak{M}$  выполнены соотношения:

- а)  $d(A, B) \geq 0$ ;
- б)  $d(A, B) = d(B, A)$ ;
- в)  $d(A, B) + d(B, C) \geq d(A, C)$ ;
- г)  $d(A, A) = 0$ .

*Аксиома 2.* Если  $A \cap C \subseteq B \subseteq A \cup C$ , то  $d(A, B) + d(B, C) = d(A, C)$ .

*Аксиома 3.* Если  $A \cap C = B \cap C = \emptyset$ , то  $d(A, B) = d(A \cup C, B \cup C)$ .

*Теорема 4* [29, § 4.3]. Пусть для функции  $d: \mathfrak{M}^2 \rightarrow R^1$  выполнены аксиомы 1а, 1б, 2, 3. Тогда существует, и притом единствен-

ная, мера  $\mu$  на  $\mathfrak{A}$  такая, что

$$d(A, B) = \mu(A \Delta B) \quad (82)$$

при всех  $A, B$  из  $\mathfrak{A}$ , где  $A \Delta B$  — симметрическая разность множеств  $A$  и  $B$ , т. е.  $A \Delta B = (A \setminus B) \cup (B \setminus A)$ .

Обратно, если по произвольной мере  $\mu$  на  $\mathfrak{A}$  построить функцию  $d: \mathfrak{A}^2 \rightarrow R^1$  с помощью равенства (82), то для так определенной функции  $d$  справедливы аксиомы 1, 2, 3.

Наконец, кратко рассмотрим нечеткие множества. С 1965 г., когда их впервые ввел Л. Заде [12], теория нечеткости проделала большой путь, обогатившись рядом крупных теоретических работ и солидных приложений. Общий обзор, показывающий необходимость использования концепции нечеткости при моделировании реальных явлений, описывающий теоретические постановки и ряд приложений, дан в [35]. Показано, что теория нечеткости в известном смысле сводится к теории случайных множеств [27, 29, 33]. Однако указанная связь не мешает рассматривать теорию нечеткости как самостоятельную область прикладной математики, точно так же, как теория вероятностей не теряет самостоятельности из-за того, что ее можно рассматривать как часть теории меры. Изучен также ряд вопросов статистики нечетких множеств. При этом полезными оказались результаты статистики в пространствах общей природы [33].

Статистика объектов нечисловой природы рассматривается как направление в прикладной статистике. Хотя еще не все основные результаты доведены до практического использования, уже сейчас с ее помощью можно получить ряд полезных рекомендаций для обработки социологических данных.

### Литература

1. Айвазян С. А. Статистическое исследование зависимостей. М., 1968.
2. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. М., 1974.
3. Айвазян С. А. Экстремальная формулировка основных проблем прикладной статистики. — В кн.: Всесоюз. школа «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа»: Тез. докл. Ереван, 1979, с. 24—49.
4. Андрукович П. Ф. Некоторые свойства метода главных компонент. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 189—228. (Учен. зап. по статистике; Т. 26).
5. Андрукович П. Ф., Веселая Г. Н., Козырев В. П., Терезин А. Т. Статистический анализ экспертных оценок. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 168—188. (Учен. зап. по статистике; Т. 26).
6. Бурбаки Н. Архитектура математики. — В кн.: Бурбаки Н. Очерки по истории математики. М., 1963, с. 245—259.
7. Гнеденко Б. В. Курс теории вероятностей. М., 1969.
8. Джини К. Средние величины. М., 1970.
9. ГОСТ 23554. 2—81. Система управления качеством продукции. Экспертные методы оценки качества промышленной продукции. Обработка значений экспертных оценок качества продукции. М., 1982.

10. Дэвид Г. Метод парных сравнений. М., 1978.
11. Загоруйко Н. Г., Орлов А. И. Некоторые нерешенные математические задачи прикладной статистики. — В кн.: Современные проблемы кибернетики: (Прикладная статистика). М., 1981, с. 53—63.
12. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М., 1976.
13. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. М., 1972.
14. Кендал М. Ранговые корреляции. М., 1975.
15. Козлов В. С., Эрлих Я. М., Долгушевский Ф. Г., Полушин П. И. Общая теория статистики. М., 1975.
16. Колмогоров А. Н., Прохоров Ю. В. Математическая статистика. — ВСЭ, 1974, т. 15, с. 1428—1438.
17. Крамер Г. Математические методы статистики. М., 1975.
18. Кузьмин В. Б., Овчинников С. В. Модель для измерений в порядковых шкалах. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 384—388. (Учен. зап. по статистике; Т. 26).
19. Математические методы в современной буржуазной социологии. М., 1966.
20. Мещалкин Л. Д. Параметризация многомерных распределений. — В кн.: Прикладной многомерный статистический анализ. Учен. зап. по статистике. М., 1978, т. 33, с. 11—18.
21. Миркин Б. Г. Проблема группового выбора. М., 1974.
22. Миркин Б. Г. Анализ качественных признаков и структур. М., 1980.
23. Мудров В. И., Кушко В. Л. Метод наименьших модулей. М., 1971.
24. Налимов В. В. Вероятностная модель языка. М., 1974.
25. Орлов А. И. О проверке симметрии распределения. — Теория вероятностей и ее применения, 1972, т. 17, № 2, с. 372—377.
26. Орлов А. И. Асимптотическое поведение статистик интегрального типа. — ДАН СССР, 1974, т. 219, № 4, с. 808—811.
27. Орлов А. И. Основания теории нечетких множеств (обобщение аппарата Заде). Случайные толерантности. — В кн.: Алгоритмы многомерного статистического анализа и их применения. М., 1975, с. 169—175.
28. Орлов А. И. Сходимость эталонных алгоритмов. — В кн.: Прикладной многомерный статистический анализ. М., 1978, с. 361—364. (Учен. зап. по статистике; Т. 33.)
29. Орлов А. И. Устойчивость в социально-экономических моделях. М., 1979.
30. Орлов А. И. Статистика объектов нечисловой природы и экспертные оценки. — В кн.: Экспертные оценки. М., 1979, с. 17—33. (Вопросы кибернетики; Вып. 58).
31. Орлов А. И. Случайные множества с независимыми элементами (мюсианы) и их применения. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 287—308. (Учен. зап. по статистике; Т. 36).
32. Орлов В. И. Оценки размерности модели в регрессии. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 92—99. (Учен. зап. по статистике; Т. 36.)
33. Орлов А. И. Задачи оптимизации и нечеткие переменные. М., 1980.
34. Орлов А. И. Связь между средними величинами и допустимыми преобразованиями шкалы. — Мат. заметки, 1981, т. 30, № 4, с. 561—568.
35. Орлов А. И. Математика нечеткости. — Наука и жизнь, 1982, № 7, с. 60—67.
36. Орлов А. И. Асимптотика решений экстремальных статистических задач. — В кн.: Анализ нечисловых данных в системных исследованиях. М., 1982, с. 4—12.
37. Орлов А. И. Парные сравнения в асимптотике Колмогорова. — В кн.: Экспертные оценки в задачах управления. М., 1982, с. 58—66.
38. Орлов А. И. Непараметрические оценки плотности в топологических пространствах. — В кн.: Прикладная статистика: Учен. зап. по статистике. М., 1983, т. 45, с. 12—40.

39. Орлов А. И. Асимптотика некоторых оценок размерности модели в регрессии. — В кн.: Прикладная статистика: Учен. зап. по статистике. М., 1983, т. 45, с. 260—265.
40. Орлов А. И. Некоторые вероятностные вопросы теории классификации. — В кн.: Прикладная статистика: Учен. зап. по статистике. М., 1983, т. 45, с. 166—179.
41. Прохоров Ю. В., Розанов Ю. А. Теория вероятностей. М., 1973.
42. Пфанцагль И. Теория измерений. М., 1976.
43. Смоляк С. А., Титаренко В. П. Устойчивые методы оценивания. М., 1980.
44. Суннес И., Зинес Дж. Основы теории измерений. — В кн.: Психологические измерения. М., 1967, с. 9—110.
45. Суслев Н. П. Общая теория статистики. М., 1970.
46. Тюрин Ю. Н., Василевич А. П. К проблеме обработки рядов ранжировок. — В кн.: Статистические методы анализа экспертных оценок. М. 1977, с. 96—111 (Учен. зап. по статистике; Т. 29).
47. Тюрин Ю. Н., Литвак Б. Г., Орлов А. И., Сатаров Г. А., Шмерлинг Д. С. Анализ нечисловой информации. М., 1981.
48. Шмерлинг Д. С., Дубровский С. А., Аржанова Т. Д., Френкель А. А. Экспертные оценки. Методы и применение: (Обзор). — В кн.: Статистические методы анализа экспертных оценок: М. 1977, с. 290—382 (Учен. зап. по статистике; Т. 29).
49. Шубкин В. Н. Социологические опыты. М., 1970.
50. Щукина Г. И. Проблема познавательного интереса в педагогике. М., 1971.
51. Kemeny J. Mathematics without numbers. — Daedalus, 1959, vol. 88, p. 571—591.
52. Kemeny J. Generalized random variables. — Pacif. J. Math., 1959, vol. 9, N 4, p. 1179—1189.
53. Reiss R. D. Consistency of minimum contrast estimators in nonstandard case. — Metrika, 1978, vol. 25, N 3, p. 129—142.

**С. А. ПЕТРОВСКИЙ**

## ПРОГНОЗИРОВАНИЕ И АНАЛИЗ НА ПРОБЛЕМНЫХ СЕТЯХ

В социально-экономических исследованиях довольно часто возникает ситуация, когда рассматривается совокупность взаимно связанных проблем, каждой из которых соответствует ряд возможных состояний, и требуется с помощью экспертного опроса указать вероятность реализации в будущем различных состояний рассматриваемых проблем. Поскольку удовлетворительное решение такой задачи на основе прямых экспертных оценок вероятностей, как правило, невозможно, автор предлагает формальный аппарат, помогающий учесть вероятностный характер интересующих социолога процессов.

По мере развития системности и комплексности работ в сфере социального моделирования все более ощущается ограниченность как чисто формальных, так и чисто содержательных (например, экспертных) методов анализа. Вместе с тем, они как бы дополняют друг друга, взаимно компенсируя специфическими достоинствами слабые стороны. Все это говорит о необходимости развивать средства комплексного системного моделирования, которые по возможности более органично объединяли в себе взаимодополняющие формальные и содержательные методы исследования, возла-

гая на каждый из них только такие задачи, которые они решают хорошо, и освобождая от задач, которые они решают хуже.

Настоящая статья содержит краткое изложение одной из методических схем, обычно фигурирующей под названием анализа на проблемных сетях [1, 2, 3]. Он предназначен для проведения многовариантных исследований по совокупности взаимосвязанных проблем (в иной терминологии — объектов, явлений, тенденций и т. п.), которые могут принадлежать различным предметным областям (например, экономической, научно-технической, социальной) и которые, по крайней мере, в некоторой своей части являются трудно формализуемыми.

Последнее, как известно, может иметь место в случае проблем, относящихся к особенно динамичным областям с глубокими и частыми структурными сдвигами или же в случае проблем, по которым отсутствует удовлетворительная статистика.

## 1. Первичная сфера анализа на проблемных сетях: экспертный комплексный прогноз

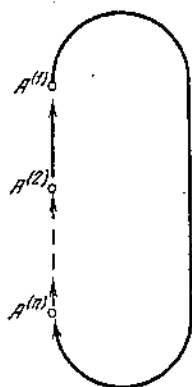
Пояснение того, какие аспекты исследования и в силу каких причин передаются в рамках анализа на проблемных сетях формализму, а какие сохраняются за содержательным мышлением эксперта, целесообразно начать с некоторой заведомо упрощенной, но достаточно характерной ситуации, схематически изображенной на рис. 1.

Кружки на приведенном рисунке обозначают определенные проблемы, т. е. нечто такое, для чего необходимо получить прогноз, а стрелками изображаются отношения влияния-зависимости между проблемами. В данном случае и в последующих аналогичных ситуациях условимся считать, что каждая стрелка исходит из непосредственно влияющей проблемы и втекает в непосредственно зависимую проблему, а подобного рода схематические изображения множества проблем и их взаимосвязей условимся называть проблемными сетями.

Прогноз будем понимать как сравнительную оценку возможности реализации в будущем различных допустимых состояний прогнозируемой проблемы. В случае качественных проблем такие состояния описываются вербально. Если же проблема носит количественный характер, то допустимые состояния могут представлять собой, например, ряд выделенных (с нужной степенью детализации) интервалов, полностью покрывающих отрезок, в котором прогнозируемая величина (показатель) будет с практической достоверностью находиться в течение или в конце прогнозируемого периода.

Вообще говоря, выделение допустимых состояний для каждой проблемы, равно как и фиксирование взаимосвязей между проблемами, составляет важный самостоятельный этап работы. Поскольку он обычно более или менее удовлетворительно выполня-

Рис. 1. Упрощенный иллюстративный вариант проблемной сети



ется экспертными средствами, будем полагать, что указанный этап исследований в нашем случае уже проделан и что, следовательно, помимо проблемной сети, изображенной на рис. 1, для каждой вошедшей в нее проблемы  $A^{(i)}$ ,  $i = 1, 2, \dots, n$ , задано множество  $\{a_{ji}^{(i)}\}$ ,  $j_i = 1, 2, \dots, m_i$ , где  $a_{ji}^{(i)}$  есть  $j_i$ -е состояние  $i$ -й проблемы.

Положим, наконец, что для всех проблем выделены соответствующие эксперты, считая ради простоты, что на каждую проблему приходится один-единственный эксперт. А теперь посмотрим в каком положении окажутся и эти эксперты, и организаторы всего исследования, если они попытаются получить комплексный прогноз по проблемам  $A^{(1)}$ ,  $A^{(2)}$ ,  $\dots$ ,  $A^{(n)}$ , используя лишь обычные средства неформального анализа.

Типичная позиция эксперта, которому надо дать прогноз по проблеме, принадлежащей его сфере компетенции, но объективно зависящей от других проблем, сводится к следующему: «Я смогу оценить возможности реализации состояний моей проблемы, если мне скажут, каково состояние проблем, непосредственно влияющих на мою проблему». Иначе говоря, эксперт готов давать условные прогнозные оценки, где в качестве условий фигурируют состояния непосредственно влияющих проблем.

В случае рассматриваемой нами проблемной сети ситуация является максимально упрощенной в том смысле, что для каждой проблемы в качестве непосредственно влияющей указывается лишь одна проблема. Однако любая такая проблема всегда имеет несколько состояний, и никто, разумеется, не в силах сказать нашему эксперту, какое именно из них будет иметь место в действительности. Эксперт мог бы, конечно, дать прогнозные оценки состояний своей проблемы, предполагая по очереди все ее возможные состояния, т. е. дать полный набор условных прогнозов и этим ограничиться. Нетрудно, однако, представить себе, во что выльется комплексный прогноз по совокупной проблемной сети, в котором все эксперты будут поступать таким образом. Из-за обилия всевозможных «если» он окажется крайне неопределенным. Прогноз, разумеется, должен быть многовариантным или, точнее говоря, множественно условным, но он должен быть таким на промежуточных этапах исследования. В своем же окончательном виде более эффективным и предпочтительным является «безусловный» прогноз, в котором все условные прогнозы соответствующим образом интегрированы и сама итоговая оценка возможности состояний прогнозируемых проблем уже не связана со множеством гипотетических предположений.

Эксперт, однако, не имеет возможности точно выполнить упомянутый синтез и обычно поступает следующим образом. Используя прогноз по непосредственно влияющей проблеме, он выбирает самое вероятное ее состояние, допускает, что именно оно будет иметь место, генерирует для этого предложения прогнозную оценку своей проблемы и выдает полученный результат в качестве окончательного «безусловного» прогноза. Недостатки подобной процедуры очевидны. В этом случае отбрасывается большое число условных прогнозов и вместе с ними большое число вариантов возможного развития ситуации, учет которых способен существенно изменить оценку возможности состояний прогнозируемой проблемы. Ведь выбранный вариант условий всего лишь более вероятен, чем другие, и его собственная вероятность может оказаться небольшой. Во всяком случае, обычно она бывает меньше совокупной вероятности отброшенных вариантов, в рамках которых картина будущего развития прогнозируемой проблемы может оказаться существенно отличной.

Рассмотренный аспект иллюстрирует ограниченные возможности неформального экспертного анализа в выполнении некоторых синтезирующих операций. Другое проявление того же рода ограниченности обнаруживает себя сразу, как только мы примем во внимание, что для выполнения описанной выше огрубленной операции получения «безусловной» прогнозной оценки эксперт вынужден был использовать прогноз по непосредственно влияющей проблеме и что для получения этого последнего совершенно аналогичным образом нужно было бы использовать прогноз по следующей непосредственно влияющей проблеме. Ясно, что действуя таким образом, и далее, мы рано или поздно пришли бы к порочному кругу, поскольку, как это видно из рис. 1, контур связей между проблемами  $A^{(1)}$ ,  $A^{(2)}$ , ...,  $A^{(n)}$  образует замкнутый цикл. В результате обнаружилась бы еще одна характерная слабость неформального мышления — его неспособность чисто развязывать замкнутые циклы зависимостей, которым в действительности соответствует обширный и важный мир обратных влияний, взаимозависимостей и т. п. Метод, к которому вынужден прибегать неформальный анализ для преодоления возникающих при этом трудностей, заключается в разрыве цикла посредством игнорирования какой-либо из связей между проблемами, что, естественно, влечет за собой соответствующее огрубление реальности.

Предположим, однако, что мы воспользовались и этой огрубляющей процедурой, т. е. разорвали замкнутый контур, проигнорировав, например, связь между проблемами  $A^{(n)}$  и  $A^{(1)}$ . В результате мы сможем получить комплексный прогноз по всей совокупности проблем, последовательно выполняя прогнозы, начиная с проблемы  $A^{(n)}$  и каждый раз передавая их на следующий уровень, пока не дойдем до проблемы  $A^{(1)}$ . Ближайшее следствие подобной организации работы заключается в том, что сроки выполнения всего исследования будут представлять собой

по меньшей мере сумму временных затрат, потребовавшихся для выполнения прогнозов по каждой отдельной проблеме сети плюс время, необходимое для передачи прогнозной информации на каждом очередном шаге.

Другое следствие, практически трудно устранимое, состоит в том, что в случае затягивания работ на каком-либо из этапов некоторые из экспертов по остальным проблемам, подстегиваемые сроками собственных плановых заданий, не будут дожидаться получения прогнозов по непосредственно влияющим проблемам, а сами, на свой страх и риск, сделают необходимые оценки по далеким от их деятельности проблемам, выходя тем самым за рамки своей компетенции и привнося в проводимое исследование недоброкачественную информацию.

Итак, рассмотрены некоторые характерные моменты комплексного прогностического исследования, выполняемого неформальными экспертными средствами.

А теперь попытаемся организовать наш гипотетический комплексный прогноз иначе. Будем по-прежнему исходить из проблемной сети, изображенной на рис. 1, и предполагать, что для каждой ее проблемы  $A_{(i)}$  задано множество состояний  $\{a_{j_i}^{(i)}\}$ ,  $j_i = 1, 2, \dots, m_i$ . Наложим на каждое из этих множеств следующие два ограничения:

состояния, принадлежащие множеству  $\{a_{j_i}^{(i)}\}$ , представляют собой попарно несовместимые события; (1)

состояния, принадлежащие множеству  $\{a_{j_i}^{(i)}\}$ , исчерпывают все имеющиеся возможности реализации проблемы  $A_{(i)}$ . (2)

Условия (1) и (2) означают, что в рамках каждой проблемы допустимые состояния образуют полную группу попарно несовместимых событий. Ничего существенно ограничительного в практическом смысле этого слова указанные условия в себе не содержат. Например, обычные формулировки типа «среднегодовые темпы роста производительности труда составят в прогнозируемом периоде от 1 до 2%, либо от 2 до 3%, либо от 3 до 4%» дают описание множества допустимых состояний, практически удовлетворяющее обоим условиям, если есть уверенность, что выделенные интервалы охватывают все реально возможные случаи. Если такой уверенности нет, то число интервалов может быть увеличено или могут быть введены два новых интервала типа «менее 1%» и «более 4%». В случае проблем качественного характера ситуация в принципе аналогичная. Взаимоисключаемость состояний здесь легко обеспечивается смыслом соответствующих формулировок. Если же возникает опасение, что зафиксированный перечень альтернатив не исчерпывает все реальные возможности, то этот перечень либо расширяется, либо добавляется еще одно состояние типа «другие возможные состояния проблемы».

Выполнение условий (1) и (2) позволяет воспользоваться формулой полной вероятности, применение которой к каждому со-

стоянию каждой проблемы сети дает следующую систему соотношений:

$$p(a_1^{(1)}) = \sum_{j_2} p(a_1^{(1)} | a_{j_2}^{(2)}) p(a_{j_2}^{(2)});$$

$$p(a_{m_1}^{(1)}) = \sum_{j_2} p(a_{m_1}^{(1)} | a_{j_2}^{(2)}) p(a_{j_2}^{(2)});$$

$$p(a_1^{(2)}) = \sum_{j_3} p(a_1^{(2)} | a_{j_3}^{(3)}) p(a_{j_3}^{(3)});$$

$$p(a_{m_2}^{(2)}) = \sum_{j_3} p(a_{m_2}^{(2)} | a_{j_3}^{(3)}) p(a_{j_3}^{(3)});$$

$$p(a_1^{(n)}) = \sum_{j_1} p(n_1^n | a_{j_1}^{(1)}) p(a_{j_1}^{(1)});$$

$$p(a_{m_n}^{(n)}) = \sum_{j_1} p(a_{m_n}^{(n)} | a_{j_1}^{(1)}) p(a_{j_1}^{(1)}).$$

Перепишем затем эти соотношения в виде системы уравнений, полагая, что для всех условных вероятностей от экспертов получаются соответствующие оценки, и рассматривая все безусловные вероятности в качестве неизвестных. В матричной форме такая система уравнений выглядит следующим образом:

$$X^{(1)} = C^{(1)} X^{(2)}$$

$$X^{(2)} = C^{(2)} X^{(3)},$$

$$X^{(n)} = C^{(n)} X^{(1)}.$$

(3)

Здесь каждый вектор-столбец  $X^{(i)} = [x_{j_i}^{(i)}]$  замещает набор безусловных вероятностей  $\{p(a_{j_i}^{(i)})\}$ , соответствующий полному комплексу допустимых состояний проблемы  $A^{(i)}$ , а матрицы  $C^{(i)}$  содержат экспертные оценки  $p^*(a_{j_i}^{(i)} | a_{j_{i+1}}^{(i+1)})$  для условных вероятностей  $p(a_{j_i}^{(i)} | a_{j_{i+1}}^{(i+1)})$  и имеют вид

$$C^{(i)} = \begin{bmatrix} p^*(a_1^{(i)} | a_1^{(i+1)}) & p^*(a_1^{(i)} | a_2^{(i+1)}) & \dots & p^*(a_1^{(i)} | a_{m_{i+1}}^{(i+1)}) \\ p^*(a_2^{(i)} | a_1^{(i+1)}) & p^*(a_2^{(i)} | a_2^{(i+1)}) & \dots & p^*(a_2^{(i)} | a_{m_{i+1}}^{(i+1)}) \\ \dots & \dots & \dots & \dots \\ p^*(a_{m_i}^{(i)} | a_1^{(i+1)}) & p^*(a_{m_i}^{(i)} | a_2^{(i+1)}) & \dots & p^*(a_{m_i}^{(i)} | a_{m_{i+1}}^{(i+1)}) \end{bmatrix}.$$

(В случае  $i = n$  во всех оценках  $p^*(a_{j_i}^{(i)} | a_{j_{i-1}}^{(i+1)})$  матрицы  $C^{(i)}$  в качестве состояний проблемы  $A^{(n+1)}$  выступают состояния проблемы  $A^{(1)}$ .)

Предполагается, конечно, что входящие в  $C^{(i)}$  оценки удовлетворяют условиям

$$p^*(a_{j_i}^{(i)} | a_{j_{i+1}}^{(i+1)}) \geq 0, \quad (4)$$

$$\sum_{j_i} p^*(a_{j_i}^{(i)} | a_{j_{i+1}}^{(i+1)}) = 1. \quad (5)$$

С содержательной точки зрения каждый столбец матрицы  $C^{(i)}$  представляет собой условный прогноз состояний проблемы  $A^{(i)}$

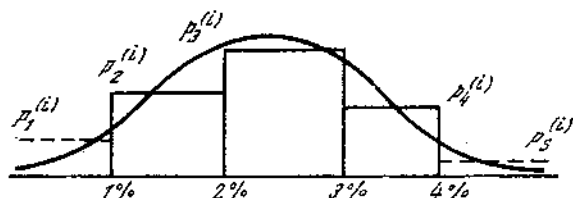


Рис. 2. Графическая иллюстрация решения системы уравнений

при допущении, что имеет место соответствующее состояние непосредственно влияющей проблемы  $A^{(i+1)}$ , а вся матрица  $C^{(i)}$  в целом содержит в себе все возможные условные прогнозы такого рода. Если мы теперь для нашей системы уравнений найдем некоторое конкретное решение  $X^{(1)} = [p_{j_1}^{(1)}]$ ,  $X^{(2)} = [p_{j_2}^{(2)}]$ ,  $\dots$ ,  $X^{(n)} = [p_{j_n}^{(n)}]$ , удовлетворяющее условиям

$$p_{j_i}^{(i)} \geq 0, \quad (6)$$

$$\sum_{j_i} p_{j_i}^{(i)} = 1, \quad (7)$$

то такое решение можно интерпретировать как безусловные вероятностные оценки состояний проблем  $A^{(1)}$ ,  $A^{(2)}$ ,  $\dots$ ,  $A^{(n)}$  или, что то же самое, как «безусловный» прогноз этих проблем, интегрирующий в себе все полученные от экспертов условные прогнозы. В случае качественных проблем это решение дает оценку степени возможности реализации любого из допустимых состояний. Для каждой количественной проблемы оно дает привычную форму измерения будущего уровня прогнозируемой величины в виде соответствующего дискретного распределения. Например, если проблема  $A^{(i)}$  заключается в характеристике средних темпов роста производительности труда и имеет пять допустимых состояний: менее 1%, от 1 до 2%, от 2 до 3%, от 3 до 4%, более 4%, то полученное для нее решение  $X^{(i)} = [p_1^{(i)}, p_2^{(i)}, p_3^{(i)}, p_4^{(i)}, p_5^{(i)}]$  эквивалентно следующему представленному в виде гистограммы распределению (дополнительно сглаженному для лучшей наглядности).

Заметим, наконец, что интересующее нас решение системы (3), точнее говоря решение, удовлетворяющее условиям (6) и (7), всегда существует. Это можно вывести из известной теоремы Брауэра о неподвижной точке, согласно которой такая точка всегда имеется, если непрерывный оператор отображает замкнутое выпуклое ограниченное множество (в нашем случае декартово произведение симплексов  $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ ) в само это множество<sup>1</sup>.

В отличие от рассмотренного ранее неформального экспертного прогнозирования описанная схема получения комплексного прогноза носит промежуточный характер, представляя собой отчасти экспертную, отчасти формализованную процедуру. На экспертов здесь возлагается построение проблемной сети, выделение для каждой проблемы набора допустимых состояний и генерирование условных прогнозных оценок. Соответственно формализму в виде описанной системы уравнений передаются все операции синтеза условных оценок в резюмирующий прогноз.

Моменты, в которых частично формализованная процедура выигрывает по сравнению с простым неформализованным исследованием, очевидны. Такая процедура дает предпочтительный с точки зрения конечного потребителя интегрированный «безусловный» прогноз на основе условных прогнозных оценок, предпочтительных с точки зрения основных исполнителей прогноза — экспертов и реализующих в себе многовариантный анализ проблем на промежуточных этапах исследования. При синтезе условных прогнозов в безусловный ни одна из вариантных оценок не игнорируется и ни одна из связей между циклически зависимыми проблемами не разрывается (системы уравнений, в частности, для того и предназначены, чтобы чисто «развязывать» вычислительные циклические зависимости). Далее, все зависимости между проблемами отображаются теперь не с помощью передачи каких-либо прогнозов, а посредством фиксирования всех возможных состояний непосредственно влияющей проблемы. (В каждой условной оценке  $p^* (a_{ji}^{(i)} | a_{j+1}^{(i+1)})$  это допущение фиксируется в виде состояния  $a_{j+1}^{(i+1)}$ , фигурирующего в записи оценки). Для генерирования своего прогноза ни одному эксперту уже нет необходимости дожидаться получения прогнозов по непосредственно влияющей проблеме. Поэтому после составления проблемной сети и выделения допустимых состояний получение прогнозных оценок можно начинать одновременно и вести независимо по всем вошедшим в сеть проблемам. Все это, с одной стороны, сокращает продолжительность полного исследовательского цикла, а, с другой — снижает организационные нагрузки, устраняя передачу

<sup>1</sup> Более полную информацию о разрешимости системы (3) можно получить, сведя эту систему последовательными подстановками к уравнению типа  $X^{(1)} = CX^{(1)}$ , где  $C = C^{(1)} \times C^{(2)} \times \dots \times C^{(n)}$  — стохастическая матрица. Разрешимость такого уравнения полностью исследована в теории конечных однородных цепей Маркова.

прогнозной информации от одного звена к другому. (При большом числе проблем, а также в случае реального комплексного прогноза, выполняемого с участием различных самостоятельных организаций, оба указанных момента могут оказаться весьма существенными.) Наконец, отсутствие необходимости дожидаться запаздывающего прогноза по непосредственно влияющей проблеме избавляет экспертов от соблазнов генерирования и использования прогнозных оценок, выходящих за рамки их компетенции.

Знакомясь с описанной частично формализованной процедурой нетрудно, конечно, обратить внимание на весьма настораживающее допущение, лежащее, казалось бы, в ее основе. Имеется в виду предположение, что эксперты способны генерировать оценки вероятности допустимых состояний, входящие в матрицы  $S^{(i)}$ . Из психологических исследований известно, насколько сложным является получение интуитивной числовой оценки вероятности одного-единственного события, а ведь количество таких событий в упомянутой процедуре достаточно велико.

В связи с этим заметим, что требование от экспертов соответствующих числовых оценок не является обязательным. Имеется целый ряд возможностей избавиться от необходимости давать прямые числовые оценки, заменив их, например, простыми ранжированиями допустимых состояний по вероятности [2] или операциями с полуколичественными шкалами, где фигурируют оценки типа «вероятность состояния  $a_k^{(i)}$  несколько (значительно, очень существенно) превосходит вероятность состояния  $a_l^{(i)}$ ». Наконец, при достаточно большом числе экспертов по одной и той же проблеме можно было бы обойтись указанием одного-единственного, наиболее вероятного состояния из полной группы событий. Все такие нечисловые оценки поддаются последующему переводу в необходимую числовую форму, предполагаемую моделью.

Интересно, однако, отметить: в реальной практике использования процедур описанного выше типа быстрее прокладывают себе дорогу именно прямые числовые оценки, получение которых представляется на первый взгляд столь нетривиальной задачей. Основная причина указанного обстоятельства заключается, по-видимому, в том, что в подобных процедурах, как правило, не предполагается получение от экспертов числовых оценок вероятности отдельных событий. От них требуют выполнения иной операции, психологически более простой, хотя и приводящей к тем же результатам. Если обычно очень трудно ответить на вопрос, какова вероятность некоторого события, то значительно проще распределить единицу (или сто шансов) между набором имеющихся альтернатив в соответствии с субъективной оценкой их относительной возможности. Именно такая операция и предполагается в рассмотренной выше процедуре. Эксперту всегда дается полный набор состояний проблемы, рассматриваемый при одном и том же условии (каждый такой набор соответствует столбцу в

матрице  $C^{(3)}$ ). И, как показывает опыт применения анализа на проблемных сетях, получение вероятностных оценок в описанной форме не вызывает существенных затруднений, если проблемная сеть составлена хорошо.

Правда, одна готовность экспертов давать соответствующие оценки еще не решает всех проблем; остается еще важный вопрос об адекватности таких оценок. Но, по-видимому, и в этом отношении ситуация более благоприятна, чем может показаться на первый взгляд. Проведенные эксперименты дают основание считать, что по крайней мере в относительном смысле адекватность описанных выше вероятностных оценок является удовлетворительной. Эксперименты показывают, что точность оценок, даваемых в виде распределений, не ниже точности привычных точечных числовых оценок, широко используемых в практике экспертов [2].

В связи с затронутыми вопросами хотелось бы обратить внимание на следующий общеметодологический момент. Если начальная форма оценок, получаемых от экспертов, не является, как мы видели, принципиальной, то этого нельзя сказать о самом факте использования вероятностных распределений в качестве базисного языка модели. Не исключено, что в комплексных социальных исследованиях базисные структуры такого рода являются единственной возможностью говорить на одном языке о проблемах (явлениях, факторах, аспектах и т. п.) как количественных, так и качественных, как хорошо, так и плохо структурируемых, о проблемах, принадлежащих совершенно различным предметным областям. Привлекательная особенность вероятностных распределений состоит именно в том, что они позволяют единообразным способом решать и задачу измерения столь различных сущностей, и задачу их последующего «увязывания» на единой основе, не прибегая при этом к использованию множества специальных шкал, образующих в итоге методическую пестроту, трудно поддающуюся анализу.

Описанная выше схема частично формализованного анализа была изложена применительно к частному случаю прогностического исследования. Разумеется, рассмотрение этой схемы не имело бы особого смысла, если бы лежащие в ее основе принципы не допускали необходимых обобщений, которые мы и попытаемся выполнить. Ясно, что проблемы и характер связей между ними понимались в рассмотренном выше частном случае достаточно широко. От проблем требовалась лишь возможность их представления в виде набора допустимых состояний, удовлетворяющих условиям (1) и (2), а на характер зависимостей между проблемами вообще не накладывалось никаких формальных ограничений. Напротив, структура связей между проблемами предполагалась весьма специфичной. Все проблемы считались влияющими и зависимыми. Каждая проблема зависела только от одной проблемы, и контур связей между ними образовывал единственный замкнутый цикл.

Чтобы получить достаточно общий случай, необходимо освободиться именно от этих структурных ограничений. Понимая проблемы и характер зависимостей между ними как и ранее, будем теперь считать, что для каждой проблемы, вошедшей в сферу комплексного анализа, может быть названа не одна, а произвольное число непосредственно влияющих проблем и что некоторые из них могут при желании не рассматриваться в качестве влияющих или зависимых в рамках проводимого исследования. Возникающие при подобном подходе контуры связей между проблемами могут оказаться, конечно, совершенно произвольными

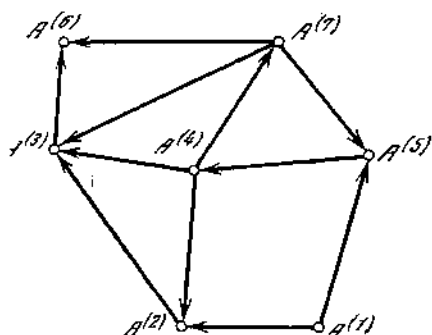


Рис. 3. Условный пример проблемной сети более общего типа

Условный пример проблемной сети такого более общего типа приведен на рис. 3.

Будем называть связкой упорядоченное множество проблем, состоящих из зависимой проблемы и всех проблем, непосредственно на нее влияющих (на рис. 3, например, таких связок шесть:  $(A^{(2)}, A^{(1)}, A^{(4)})$ ,  $(A^{(3)}, A^{(2)}, A^{(4)}, A^{(7)})$ ,  $(A^{(4)}, A^{(5)})$ ,  $(A^{(6)}, A^{(7)}, A^{(1)})$ ,  $(A^{(6)}, A^{(3)}, A^{(7)})$ ,  $(A^{(7)}, A^{(4)})$ ). Рассмотрим некоторую произвольную связку  $A^{(I)}, A^{(II)}, \dots, A^{(N)}$ , где  $A^{(I)}$  — зависимая проблема а  $A^{(II)}, \dots, A^{(N)}$  — проблемы, непосредственно влияющие (для упрощения обозначений нумерация I, II, ..., N произведена здесь в рамках данной связки и не совпадает поэтому с общей нумерацией проблем в рамках всей проблемной сети). Если, как и ранее, эксперт по проблеме  $A^{(I)}$  пожелает давать повариантные условные оценки, то теперь уже в качестве отдельного варианта условий будет выступать конъюнкция  $a_{jII}^{(II)} \wedge a_{jIII}^{(III)} \wedge \dots \wedge a_{jN}^{(N)}$ , включающая по одному состоянию от каждой непосредственно влияющей на  $A^{(I)}$  проблемы. Перебирая все возможные вхождения в конъюнкцию состояний непосредственно влияющих проблем, мы получим все возможные варианты условий для проблемы  $A^{(I)}$ . Поскольку для всех проблем, включенных в сеть, по-прежнему выполняются требования (1) и (2), то каждая пара упомянутых конъюнкций будет представлять собой несовместимые события, а множество всех конъюнкций, соответствующих связке, образует полную группу событий. Поэтому

для связки остается справедливой формула полной вероятности

$$p(a_{j_I}^{(I)}) = \sum_{j_{II} \dots j_N} p(a_{j_I}^{(I)} | a_{j_{II}}^{(II)} a_{j_{III}}^{(III)} \dots a_{j_N}^{(N)}) p(a_{j_{II}}^{(II)} a_{j_{III}}^{(III)} \dots a_{j_N}^{(N)}).$$

Раскрывая <sup>2</sup> элемент  $p(a_{j_{II}}^{(II)} a_{j_{III}}^{(III)} \dots a_{j_N}^{(N)})$ , получаем

$$p(a_{j_I}^{(I)}) = \sum_{j_{II} \dots j_N} p(a_{j_I}^{(I)} | a_{j_{II}}^{(II)} a_{j_{III}}^{(III)} \dots a_{j_N}^{(N)}) \times \\ \times p(a_{j_{II}}^{(II)} | a_{j_{III}}^{(III)} \dots a_{j_N}^{(N)}) \dots p(a_{j_N}^{(N)}). \quad (8)$$

Соотношение (8) синтезирует ковариантные условные оценки  $p(a_{j_I}^{(I)} | a_{j_{II}}^{(II)} a_{j_{III}}^{(III)} \dots a_{j_N}^{(N)})$  с вероятностями компонент, составляющих условие каждого варианта. Заметим, что, поскольку проблемы, входящие в описание вариантов условий (все проблемы или часть их), нередко оказываются практически взаимонезависимыми, соотношение (8) обычно упрощается. В частности, когда все указанные проблемы взаимонезависимы, оно приобретает особенно простой вид:

$$p(a_{j_I}^{(I)}) = \sum_{j_{II} \dots j_N} p(a_{j_I}^{(I)} | a_{j_{II}}^{(II)} \dots a_{j_N}^{(N)}) p(a_{j_{II}}^{(II)}) p(a_{j_{III}}^{(III)}) \dots p(a_{j_N}^{(N)}).$$

Соотношение (8) является обобщением соотношений, лежащих в основе системы (3), и подобно им служит основой построения системы уравнений для любой конкретной проблемной сети более общего вида. Операциональная схема такого построения может быть описана следующим образом:

а) для каждого состояния зависимой проблемы каждой связки сети выписывается соотношение (8);

б) посредством заранее проведенного экспертного анализа в рамках каждой связки выясняется, какие из ее непосредственно влияющих проблем оказываются взаимонезависимыми, и на основе этой информации производится упрощение соотношений (8) заменой соответствующих условных вероятностей на безусловные;

в) для каждого состояния каждой независимой проблемы (т. е., проблемы, для которой не указано ни одной непосредственно влияющей) выписываются тождества  $p(a_{j_i}^{(i)}) = p_i^{(i)}$ ;

г) все безусловные вероятности  $p(a_{j_i}^{(i)})$ , входящие в соотношения (8) после их упрощения, а также в тождества  $p(a_{j_i}^{(i)}) = p_i^{(i)}$  рассматриваются в качестве неизвестных, а для всех остальных условных вероятностей и величин  $p_{j_i}^{(i)}$  получают экспертные оценки (разумеется, неотрицательные и нормированные по единице).

<sup>2</sup> Раскрытие (8) зависит от порядка перечисления влияющих проблем и поэтому не единственно.

После того как система уравнений построена, ищется ее решение, удовлетворяющее требованиям

$$x_{ji}^{(i)} \geq 0; \quad \sum_{ji} x_{ji}^{(i)} = 1.$$

Вследствие уже упоминавшейся теоремы Брауэра о неподвижной точке такое решение всегда существует [1]. Как и ранее, оно интерпретируется в качестве оценки возможности реализации состояний каждой проблемы сети, результирующей в себе весь многовариантный анализ проблем и всю совокупность взаимосвязей между проблемами<sup>3</sup>.

## 2. Включение в сеть частных формальных моделей

Схема анализа на проблемных сетях ассоциировалась до сих пор исключительно с экспертизой. Точнее говоря, предполагалось, что в случае каждой отдельной проблемы сети ее связи с непосредственно влияющими проблемами воспроизводятся экспертами средствами их неформального мышления и что результаты такого воспроизведения в виде соответствующих повариантных оценок вводятся в общую модель анализа на проблемных сетях. Для достаточно широкого комплексного исследования, однако, типична ситуация, когда множество охваченных им проблем, помимо трудноформализуемых и составляющих поэтому объект экспертного анализа, включает также проблемы, допускающие эффективное формальное моделирование. При использовании анализа на проблемных сетях в целях комплексного прогнозирования было бы желательно, таким образом, иметь возможность свободно включать в проблемную сеть как фрагменты, опирающиеся на результаты экспертиз, так и фрагменты, основывающиеся на соответствующих частных моделях. Нетрудно убедиться, что такая возможность предоставляется уже рассмотренной схемой анализа на проблемных сетях без изменения ее формальной стороны.

Рассмотрим произвольную связку сети, состоящую из некоторой зависимой проблемы  $A^{(1)}$  и непосредственно влияющих на нее проблем  $A^{(11)}, \dots, A^{(N)}$  (как и ранее нумерация проблем производится только в рамках данной связки). Пусть все эти проблемы носят количественный характер. В этом случае перечень их допустимых состояний формулируется как полный набор интервалов, в которых может оказаться соответствующая величина. Допустим теперь, что существует некоторая формальная модель, в которой имеется эндогенная переменная  $u_1$ , тождественная проблеме  $A^{(1)}$  и зависящая от экзогенных переменных  $u_2, \dots, u_N$ ,

<sup>3</sup> Описанная выше схема построения модели анализа на проблемных сетях реализует основной, базисный вариант этой модели. Более сложные ее варианты см. в [2].

тождественных проблемам  $A^{(I)}$ , ...,  $A^{(N)}$ . Тогда расчеты, выполненные на основе такой модели, могут быть использованы для получения ковариантных оценок  $p^*(a_{jI}^{(I)} | a_{jII}^{(II)} \dots a_{jN}^{(N)})$ , даваемых согласно прежней интерпретации экспертами. В частности, если модель детерминированная, то  $p^*(a_{jI}^{(I)} | a_{jII}^{(II)} \dots a_{jN}^{(N)}) = 1$  для интервала  $a_{jI}^{(I)}$ , в который попадает значение эндогенной переменной  $y_1$ , вычисленное при значениях экзогенных переменных, соответствующих варианту условий  $(a_{jII}^{(II)}, \dots, a_{jN}^{(N)})$  и  $p^*(a_{jI}^{(I)} | a_{jII}^{(II)} \dots a_{jN}^{(N)}) = 0$  для всех остальных интервалов. Если же модель стохастическая, то оценки  $p^*(a_{jI}^{(I)} | a_{jII}^{(II)} \dots a_{jN}^{(N)})$  определяются из распределения вероятностей для переменной  $y_1$ , полученного при соответствующих значениях экзогенных переменных.

Включение формальной модели в проблемную сеть позволяет получить некоторые полезные эффекты, которые могут отсутствовать при использовании этой модели самой по себе. Например, результат, полученный на основе модели, обычно сводится к набору условных прогнозов, рассчитанных для различных вариантов экзогенных переменных. Включая же модель в проблемную сеть, получаем возможность синтезировать множество условных прогнозов эндогенной переменной в интегрированный безусловный прогноз. Кроме того, включение модели в сеть позволяет повысить уровень системности прогнозов ее экзогенных переменных. Наделение переменных экзогенным статусом обычно означает, что возможности моделирования этих переменных чисто формальными средствами практически исчерпаны и что влияющие на них факторы принадлежат уже плохо структурируемым областям. В результате прогнозы указанных переменных выносятся не только за рамки формальной модели, но и часто за рамки системного исследования вообще. Ситуация меняется, если формальная модель включена в проблемную сеть. В этом случае ее экзогенные переменные становятся эндогенными с точки зрения всей модели анализа на проблемных сетях, охватывающей также и соответствующие трудно формализуемые фрагменты.

Помимо рассмотренного выше способа получения оценок  $p^*(a_{jI}^{(I)} | a_{jII}^{(II)} \dots a_{jN}^{(N)})$ , когда они попросту вычисляются на основе формальной модели, возможен также другой, когда оценки, получаемые в рамках модели, служат лишь начальной информацией, которая затем анализируется экспертами и корректируется ими. В итоге оценки, идущие в модель анализа на проблемной сети, оказываются результатом синтеза формального и экспертного исследования объекта. Формализм анализа на проблемных сетях дает возможность придать экспертной коррекции модельных результатов более систематический характер, чем простая внешняя поправка. Для этого достаточно явно ввести в проблемную сеть те факторы, которые эксперты намерены принимать во внима-

ние в ходе коррекции и которые учитываются формальной моделью. В этом случае каждый вариант условий для соответствующей зависимой проблемы сети, помимо состояний экзогенных переменных модели, будет также содержать состояния вновь введенных факторов (проблем), что позволит экспертам произвести корректировку поочередно с учетом всех возможных состояний этих факторов, явным образом сформулированных. Одновременно совокупная модель анализа на проблемной сети учтет вероятности указанных факторов, выступающих теперь в качестве самостоятельных элементов единого исследования.

### 3. Переход от вероятностей к структурам

Как уже отмечалось, роль исходного стандартного блока при построении модели анализа на проблемных сетях играет соотношение (8), применяемое к каждой связке сети. До сих пор входящие в это соотношение величины мы, естественно, интерпретировали как вероятности событий. Заметим, однако, что когда модель в виде соответствующей системы уравнений сформулирована, нет необходимости постоянно помнить о том, что элементы соотношения (8) являются вероятностями. При решении системы уравнений и ее теоретическом исследовании используются лишь простейшие свойства этих элементов в виде их неотрицательности и «нормированности по единице». С другой стороны, существуют величины иной природы, нежели вероятности (не связанные со случайностью в содержательном ее понимании), которые также обладают такими свойствами и для которых выполняются (имеют четкий содержательный смысл) соотношения, тождественные по своей структуре соотношению (8).

Указанное обстоятельство позволяет расширить интерпретацию и соответственно сферу возможного применения анализа на проблемных сетях.

Введем некоторое обобщение основных понятий рассматриваемого метода. Условимся понимать под проблемой  $A^{(i)} = \{a_j^{(i)}\}$  некоторое целое произвольной природы, для которого  $a_1^{(i)}, a_2^{(i)}, \dots, a_{m_i}^{(i)}$  выступают в качестве его элементов. Будем требовать от них выполнения условий, аналогичных (1) и (2), полагая, что все эти элементы различны и перечень их является полным в том смысле, что все они вместе взятые составляют  $A^{(i)}$ .

Пусть мы имеем перечень проблем  $A^{(1)}, A^{(2)}, \dots, A^{(n)}$ . Будем по-прежнему говорить, что для этих проблем задана проблемная сеть, а если каждой проблеме из перечня поставлено в соответствие некоторое подмножество проблем из того же перечня, которое может быть и пустым. По аналогии с тем, как это делалось ранее применительно к проблемной сети, будем использовать термины «зависимая проблема», «непосредственно влияющая проблема», «независимая проблема», а также «связка проблем».



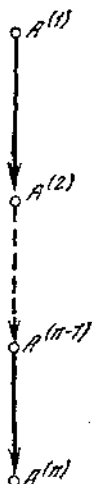


Рис. 4. Условная проблемная сеть

единице» величин. Группируя эти оценки в вектор-столбец  $V^{(1)} = [w^{*(1)}]$  и матрицы  $W^{(i)}$ , составленные по аналогии с матрицами  $C^{(i)}$ , упоминавшимися ранее, получаем в итоге систему уравнений

$$\begin{aligned} X^{(1)} &= V^{(1)}; \\ X^{(2)} &= W^{(2)} X^{(1)}; \end{aligned} \quad (12)$$

$$\vdots$$

$$X^{(n)} = W^{(n)} X^{(n-1)},$$

где  $X^{(i)} = [w(a_{ji}^{(i)})]$  — вектор искомых резюмирующих оценок для элементов проблемы  $A^{(i)}$ . Из (12) имеем, в частности, что для последней  $n$ -й проблемы искомые оценки равны

$$X^{(n)} = W^{(n)} W^{(n-1)} \dots W^{(2)} V^{(1)}. \quad (13)$$

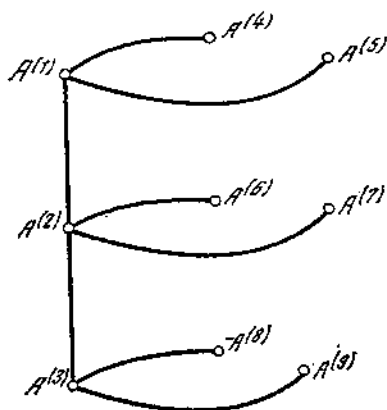
Рассмотрим теперь следующую интерпретацию описанной модели. Будем считать, что каждая вошедшая в сеть проблема  $A^{(i)} = \{a_{ji}^{(i)}\}$  представляет собой набор целей (задач, направлений деятельности, программ, проектов, разработок и т. п.) соответствующего  $i$ -го уровня, а сама проблемная сеть отображает соподчиненность целей различных уровней. Величины  $w(a_1^{(i)})$ ,  $w(a_2^{(i)})$ , ...,  $w(a_{m_i}^{(i)})$  будем рассматривать как интегрированные относительные веса (относительную важность, значимость) целей  $i$ -го уровня, т. е. как веса, интегрирующие в себе все существенные критерии и аспекты анализа. Величины же  $w(a_1^{(i)} | a_{j_{i-1}}^{(i-1)})$ ,  $w(a_2^{(i)} | a_{j_{i-1}}^{(i-1)})$ , ...,  $w(a_{m_i}^{(i)} | a_{j_{i-1}}^{(i-1)})$  будем интерпретировать в качестве условных относительных весов целей  $i$ -го уровня, рассматриваемых с определенной частной точки зрения, а именно с точки зрения их существенности для  $a_{j_{i-1}}^{(i-1)}$ -й цели вышестоящего уровня.

Тогда соотношение (10) оказывается тождественным хорошо известному правилу исследования операций, позволяющему определять интегрированный вес каждой цели  $i$ -го уровня на основе учета ее условных весов, оцениваемых с позиции целей более высокого уровня, а также интегрированных весов для целей этого более высокого уровня<sup>5</sup>. Соответственно смысл всей модели (12) и результата (13) состоит в том, что сложная задача определения относительной важности целей нижнего уровня с позиции всей многоуровневой системы вышестоящих целей сводится к ряду более простых задач, охватывающих каждый раз только связь двух соседних уровней.

<sup>5</sup> В исследовании операций есть и другие правила агрегирования.

Рис. 5. Модификация проблемной сети

При описанной интерпретации модель (12) и результирующее соотношение (13) полностью совпадают с операциями, выполняемыми в рамках известного метода решающих матриц. Заметим, что на языке анализа на проблемных сетях воспроизводится и метод решающих деревьев (деревьев целей). Операции, выполняемые в рамках этого метода, также сводятся к описанной модели (12) с той лишь разницей, что внутренняя структура матриц  $W^{(i)}$  имеет в этом случае специальную форму, отражающую ветвящееся строение дерева.



Разумеется, простая способность воспроизводить на новом языке уже известные процедуры вряд ли представила бы интерес, если бы при этом не появлялись какие-либо дополнительные возможности. В случае анализа на проблемных сетях такие возможности возникают. Применительно к рассмотренной выше интерпретации они заключаются в том, что операциям типа решающих матриц или решающих деревьев при желании можно придать многовариантный характер, а соответствующий этим операциям программно-целевой анализ может быть объединен в контексте единой модели с более широким прогнозным анализом. Проиллюстрируем эти новые моменты на примере проблемной сети, изображенной на рис. 5 и 6, и представляющей собой некоторую модификацию сети, изображенной на рис. 4. Начнем с рис. 5.

Входящая в данную проблемную сеть вертикальная цепочка проблем  $A^{(1)}, A^{(2)}, A^{(3)}$  представляет собой воспроизведение сети на рис. 4, ограниченной ради простоты тремя верхними проблемами. Соответственно и сами эти проблемы, и связи между ними имеют тот же содержательный смысл что и ранее, представляя собой наборы целей трех соподчиненных уровней. Что же касается остальных проблем  $A^{(4)}, \dots, A^{(9)}$ , то они обозначают некоторые объективные факторы, которые существенно влияют на оценки веса целей, входящих в проблемы  $A^{(1)}, A^{(2)}, A^{(3)}$ . Элементами проблем  $A^{(4)}, \dots, A^{(9)}$  являются полные группы попарно несовместимых событий.

Первоначальная система соотношений (9) для данной сети имеет вид

$$w(a_{j_1}^{(1)}) = \sum_{j_4 j_5} w(a_{j_1}^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)}) w(a_{j_4}^{(4)} | a_{j_5}^{(5)}) w(a_{j_5}^{(5)});$$

$$w(a_{j_2}^{(2)}) = \sum_{j_6 j_7} w(a_{j_2}^{(2)} | a_{j_6}^{(6)} a_{j_7}^{(7)}) w(a_{j_6}^{(6)} | a_{j_7}^{(7)}) w(a_{j_7}^{(7)});$$

$$w(a_{j_3}^{(3)}) = \sum_{j_2 j_4} w(a_{j_3}^{(3)} | a_{j_2}^{(2)} a_{j_4}^{(4)}) w(a_{j_2}^{(2)} | a_{j_4}^{(4)} a_{j_3}^{(3)}) w(a_{j_4}^{(4)} | a_{j_3}^{(3)} a_{j_2}^{(2)}) w(a_{j_3}^{(3)} | a_{j_2}^{(2)} a_{j_4}^{(4)});$$

$$w(a_{j_4}^{(4)}) = w_{j_4}^{(4)}; \quad w(a_{j_5}^{(5)}) = w_{j_5}^{(5)}; \quad \dots; \quad w(a_{j_9}^{(9)}) = w_{j_9}^{(9)}.$$

Из структуры сети следует, что проблема  $A^{(1)}$  не зависит от проблем  $A^{(6)}$  и  $A^{(7)}$ , а проблема  $A^{(2)}$  не зависит от проблем  $A^{(3)}$  и  $A^{(8)}$ . Ради простоты будем полагать, что пары проблем  $\{A^{(4)}, A^{(5)}\}$ ,  $\{A^{(6)}, A^{(7)}\}$ ,  $\{A^{(8)}, A^{(9)}\}$  внутри себя также взаимно независимы. Далее, как уже отмечалось, проблемы  $A^{(4)}, \dots, A^{(9)}$  представляют собой наборы событий, и поэтому веса для их элементов можно интерпретировать как вероятности. Учитывая все эти обстоятельства, соотношения можно переписать в виде

$$w(a_{j_1}^{(1)}) = \sum_{j_4 j_5} w^*(a_{j_1}^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)}) p(a_{j_4}^{(4)}) p(a_{j_5}^{(5)});$$

$$w(a_{j_2}^{(2)}) = \sum_{j_3 j_6 j_7} w^*(a_{j_2}^{(2)} | a_{j_3}^{(3)} a_{j_6}^{(6)} a_{j_7}^{(7)}) w(a_{j_3}^{(3)}) p(a_{j_6}^{(6)}) p(a_{j_7}^{(7)}); \quad (14)$$

$$w(a_{j_3}^{(3)}) = \sum_{j_2 j_4 j_5} w^*(a_{j_3}^{(3)} | a_{j_2}^{(2)} a_{j_4}^{(4)} a_{j_5}^{(5)}) w(a_{j_2}^{(2)}) p(a_{j_4}^{(4)}) p(a_{j_5}^{(5)});$$

$$p(a_{j_4}^{(4)}) = p_{j_4}^{*(4)}; \quad p(a_{j_5}^{(5)}) = p_{j_5}^{*(5)}; \quad \dots; \quad p(a_{j_9}^{(9)}) = p_{j_9}^{*(9)}.$$

Переходя к системе уравнений, получаем окончательную форму модели

$$x_{j_1}^{(1)} = \sum_{j_4 j_5} w^*(a_{j_1}^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)}) x_{j_4}^{(4)} x_{j_5}^{(5)};$$

$$x_{j_2}^{(2)} = \sum_{j_3 j_6 j_7} w^*(a_{j_2}^{(2)} | a_{j_3}^{(3)} a_{j_6}^{(6)} a_{j_7}^{(7)}) x_{j_3}^{(3)} x_{j_6}^{(6)} x_{j_7}^{(7)};$$

$$x_{j_3}^{(3)} = \sum_{j_2 j_4 j_5} w^*(a_{j_3}^{(3)} | a_{j_2}^{(2)} a_{j_4}^{(4)} a_{j_5}^{(5)}) x_{j_2}^{(2)} x_{j_4}^{(4)} x_{j_5}^{(5)};$$

$$x_{j_4}^{(4)} = p_{j_4}^{*(4)}; \quad x_{j_5}^{(5)} = p_{j_5}^{*(5)}; \quad \dots; \quad x_{j_9}^{(9)} = p_{j_9}^{*(9)}. \quad (15)$$

Поясним теперь содержательный смысл используемых в (14) и (15) величин, а также смысл всей процедуры в целом. Для этого удобно сопоставить систему (14) с более простой системой (11), воспроизводящей обычный метод решающих матриц или решающих деревьев. Рассмотрим первое выражение из (11)  $w(a_{j_1}^{(1)}) = w_{j_1}^{(1)}$ . Когда эксперту предлагают дать оценки для  $w(a_1^{(1)})$ ,  $w(a_2^{(1)})$ ,  $\dots$ ,  $w(a_{m_1}^{(1)})$ , он оказывается в уже знакомой нам затруднительной ситуации. Он понимает, что такие оценки зависят от состояний ряда объективных факторов, которые будут определять среду или фон для системы целей  $a_1^{(1)}$ ,  $a_2^{(1)}$ ,  $\dots$ ,  $a_{m_1}^{(1)}$  и таким образом существенно влиять на относительную значимость (вес) этих целей. При желании эксперт мог бы дать свои оценки поочередно для всех возможных вариантов условий, сформулированных в терминах состояний упомянутых факторов. Однако он никогда не делает этого в рамках обычного метода решающих матриц или деревьев, поскольку своими собственными средствами он все равно не смог бы синтезировать варианты оценки в единую интегрированную оценку. В результате эксперт вынужден идти

по обычному пути огрублений: он игнорирует все возможные варианты условий, кроме одного, по его мнению, наиболее вероятно, и дает для него свои оценки:  $w_{j_1}^{*(1)}$ .

Соотношение  $w(a_{j_1}^{(1)}) = \sum_{j_4, j_5} w(a_{j_1}^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)}) p(a_{j_4}^{(4)}) p(a_{j_5}^{(5)})$  из (14)

является аналогом выражения  $w(a_{j_1}^{(1)}) = w_{j_1}^{(1)}$  из (11), но в отличие от последнего оно дает возможность выполнить синтез повариантных оценок. Фигурирующие в нем величины  $w(a_1^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)})$ ,  $w(a_2^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)})$ , ...,  $w(a_{m_1}^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)})$  означают относительные

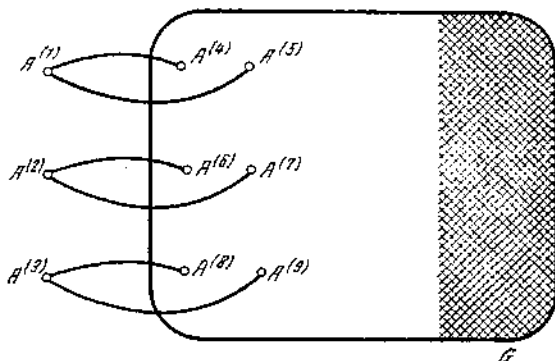


Рис. 6. Расширение проблемной сети

веса целей  $a_1^{(1)}, a_2^{(1)}, \dots, a_{m_1}^{(1)}$  при фиксированном условии  $a_{j_4}^{(4)} a_{j_5}^{(5)}$ , включающем состояния двух влияющих факторов  $A^{(4)}$  и  $A^{(5)}$ , произведение  $p(a_{j_4}^{(4)}) p(a_{j_5}^{(5)})$  дает вероятность этого условия, а суммирование в рамках рассматриваемого соотношения по индексам  $j_4$  и  $j_5$  реализует синтез всех оценок  $w(a_{j_1}^{(1)} | a_{j_4}^{(4)} a_{j_5}^{(5)})$  для всех возможных вариантов условий с учетом вероятности этих вариантов. С остальными выражениями из (11) и (14) ситуация аналогичная.

Остается еще отметить следующую условность проблемной сети, изображенной на рис. 5 и соответствующей ей модели. В этой сети проблемы  $A^{(4)}, \dots, A^{(9)}$  обозначены в качестве независимых, поэтому оценки вероятностей для их состояний задавались в модели экзогенно. В рамках анализа на проблемных сетях естественно, однако, получать такие оценки эндогенно из некоторой более широкой сети. Схематическое изображение подобной ситуации дано на рис. 6, где область  $G$  обозначает некоторую не эксплицированную на рисунке часть проблемной сети, развивающую проблемы  $A^{(4)}, \dots, A^{(9)}$  и обеспечивающую им эндогенный прогноз. Всей такой сети в целом будет соответствовать единая модель в виде определенной системы уравнений, объединяющей элементы программно-целевого исследования с комплексным прогностическим анализом.

#### 4. Заключительные замечания

Поскольку анализ на проблемных сетях объединяет в себе элементы содержательного и формального исследования, естественно организовать его в виде соответствующей человеко-машинной системы, в которой содержательная сторона остается за экспертами, а выполнение всех формальных и многих технических операций (типа сортировки, контроля, размножения, хранения, поиска информации и т. п.) передается ЭВМ. Первая версия такой системы была разработана в ИМЭМО АН СССР [3]. Согласно ей функционирование системы начинается с составления экспертами проблемной сети и фиксирования для каждой проблемы полной группы образующих ее элементов. Полученная информация вводится затем в ЭВМ, которая, опираясь на эту информацию и общие принципы рассматриваемой модели анализа на проблемных сетях, приступает к построению конкретной модели, соответствующей данной проблемной сети. ЭВМ формирует структуру будущей модели, устанавливает, какие оценки следует получить в соответствии с этой структурой, и выдает необходимые запросы, оформленные в виде анкет. Последние заполняются оценками, полученными либо от экспертов, либо с использованием частных формальных моделей, либо комбинированным путем. Все эти оценки вводятся затем в ЭВМ, которая завершает формирование первичного варианта модели, решает соответствующую систему уравнений и выдает первичные результаты. Одновременно ЭВМ исследует модель на чувствительность, выявляя оценки, наиболее существенные для возможного уточнения результатов. Полученная информация анализируется экспертами, существенные оценки уточняются и вновь вводятся в ЭВМ, которая строит на их основе уточненную модель. Такие циклы операций могут быть повторены неоднократно.

#### Литература

1. Анализ на проблемных сетях. Вып. 1. М., 1980.
2. Анализ на проблемных сетях. Вып. 2. М., 1982.
3. Комплексное прогнозирование в экономике и международных отношениях. Вып. 1. Прогнозирование на проблемных сетях. М., 1975.

В. Т. ПЕРЕКРЕСТ

## ФУНКЦИОНАЛЬНЫЙ ПОДХОД В МЕТРИЧЕСКОМ МНОГОМЕРНОМ ШКАЛИРОВАНИИ

В статье излагаются основные идеи нового подхода к многомерному шкалированию. Основным достоинством такого подхода, отличающим его от традиционных алгоритмов многомерного шкалирования, является то, что каждое подмножество элементов исходного пространства рассматривается как бы с учетом его «веса», определяемого заданной мерой распределения вероятностей. Это вполне отвечает вероятностному характеру изучаемых социологических явлений.

Хотя в статье речь идет о метрическом шкалировании (когда предполагается, что элементы исходной матрицы близостей удовлетворяют условиям, делающим их в достаточной мере «похожими» на числа, что редко может быть оправдано в социологии), автор отмечает, что предложенные методы могут использоваться и для неметрических данных, если к последним применить способ порождения метрики в пространстве номинальных и порядковых признаков.

Исходная информация в задачах шкалирования представляется бинарным отношением близости (сходства, различия, превосходства и т. п.) между исследуемыми объектами. В случае, если близость между объектами задается с помощью числовых значений (количественно), многомерное шкалирование называется метрическим. Если же в методах шкалирования учитываются не числовые значения близостей между объектами, а только их порядок, то такие методы называются методами неметрического шкалирования. В последнем случае можно также говорить, что близость между объектами измеряется в порядковой шкале. Методы метрического шкалирования могут использоваться для анализа как количественной, так и качественной (нечисловой) информации.

Для иллюстрации этого утверждения рассмотрим случай нечисловых данных, задаваемых прямоугольной матрицей объект-признак, причем все рассматриваемые признаки — порядковые с конечным числом градаций, которые для определенности закодированы точками соответствующего отрезка натурального ряда чисел. Укажем процедуру вычисления расстояния между точками рассматриваемого признакового пространства и, следовательно, процедуру построения числовой матрицы расстояний между объектами имеющейся совокупности. Обозначим через  $m$  количество рассматриваемых признаков.<sup>1</sup> Тогда признаковое про-

пространство  $X$  конечно и состоит из точек вида  $x = (x_1, \dots, x_m)$ , где  $x_i$  — натуральные числа, значения  $i$ -го признака. Обозначим через  $v(x)$  частоту заполнения точки  $x$  признакового пространства  $X$  объектами имеющейся совокупности (выборки). Очевидно,  $\sum v(x) = 1$ , где суммирование проводится по всем  $x \in X$ . Далее будем говорить, что набор  $\hat{x} = \{x^{(i)}\}_{i=1}^p \subset X$  образует непрерывную кривую в  $X$ , если для любого  $j = 1, \dots, p-1$  точки  $x^{(j)}$  и  $x^{(j+1)}$  являются соседними, т. е.

$$\sum_{l=1}^m |x_l^{(j)} - x_l^{(j+1)}| = 1$$

(заметим, что в сделанных предположениях величина  $|x_l^{(j)} - x_l^{(j+1)}|$  всегда является целым неотрицательным числом)<sup>1</sup>. Точки  $x^{(1)}$  и  $x^{(p)}$  назовем соответственно начальной и конечной точками кривой  $\hat{x}$  ( $\hat{x}$  соединяет точки  $x^{(1)}$  и  $x^{(p)}$ ), а величину

$$\gamma(\hat{x}) = \frac{1}{2} [v(x^{(1)}) + v(x^{(p)})] + \sum_{j=2}^{p-1} v(x^{(j)})$$

назовем длиной кривой  $\hat{x}$ . А теперь для любых двух точек  $y, z \in X$ ,  $y \neq z$ , в качестве расстояния  $r(y, z)$  между ними возьмем длину кратчайшей непрерывной кривой  $\hat{x}$ , соединяющей эти точки, т. е.

$$r(y, z) = \gamma(\hat{x}) = \min \gamma(\omega),$$

где минимум берется по всем непрерывным кривым  $\omega$ , соединяющим точки  $y$  и  $z$ . В случае же  $y = z$  положим  $r(y, z) = 0$ . Легко убедиться, что такое определение расстояния  $r$  корректно. Можно указать аналогичные процедуры определения расстояния и в случае, если рассматриваемые признаки номинальные, и в смешанных ситуациях, когда среди рассматриваемых признаков встречаются как порядковые, так и номинальные, а также количественные признаки. Описанная схема порождения метрики статистическими характеристиками исследуемой совокупности использовалась, в частности, в [7] при построении показателя социально-экономической эффективности труда инженеров.

## 1. Функциональные модели метрического шкалирования

В предложенных функциональных моделях [13, 26] предполагается, что множество исследуемых объектов  $X$  не состоит из конечного числа точек, а является пространством с двумя структурами: топологической, определяемой некоторой числовой функ-

<sup>1</sup> За основание еще одного способа определения соседних точек в  $X$  может быть выбрано более слабое условие, задаваемое равенством

$$\max_{1 \leq l \leq m} |x_l^{(j)} - x_l^{(j+1)}| = 1.$$

цией парной близости  $r(x, y)$ , и статистической — некоторой вероятностной мерой  $\mu$ , задающей распределение объектов в  $X$ .

Под функцией близости понимается такой определенный на  $X^2 = X \times X$  симметричный и принимающий неотрицательные значения функционал  $r$ , который обладает следующими свойствами:

1) для любых  $x, y \in X$ , равенство  $r(x, y) = 0$  эквивалентно  $x = y$ ;

2) для любого  $\varepsilon > 0$  существует такое  $\delta > 0$ , что для любых  $x, y, z \in X$ , из неравенств  $r(x, z) \leq \delta$  и  $r(z, y) \leq \delta$  следует  $r(x, y) \leq \varepsilon$ .

Функция близости  $r$  задает некоторую топологию в  $X$ . Предполагается, что  $X$  как топологическое пространство компактно, под мерой  $\mu$ , заданной на компактном топологическом пространстве  $X$ , понимается мера Родона, т. е. непрерывный линейный функционал в пространстве непрерывных на  $X$  функций с равномерной метрикой. Мера  $\mu$  может рассматриваться как борелевская мера, т. е. как счетно-аддитивная функция на  $\sigma$ -алгебре борелевских множеств. Очевидно, функция близости  $r$ , задающая топологию в  $X$ ,  $\mu^2$  измерима. Предполагается, что функция близости  $r$  вполне непрерывна. Иными словами, при  $r(x, y) \rightarrow 0$  имеет место равномерное по  $x, y \in X$  стремление  $\sigma_r(x, y) \rightarrow 0$ , где

$$\sigma_r(x, y) = \int |r(x, z) - r(y, z)| d\mu(z), \quad (1)$$

а интегрирование проводится по всему множеству  $X$ . Условию полной непрерывности удовлетворяет, например, любая функция близости, являющаяся положительной степенью некоторой метрики.

Упорядоченная тройка  $\mathcal{X} = (X, r, \mu)$  называется генеральной совокупностью исследуемых объектов.

### 1.1. $Z_k$ -шкалирование

Рассмотрим задачу  $k$ -мерного шкалирования для пространства  $\mathcal{X} = (X, r, \mu)$ , заключающуюся в нахождении  $k$ -мерной вектор-функции  $f: X \rightarrow R^k$ , минимизирующей функционал

$$Z_k(f) = \iint \{ |f(x) - f(y)|^2 - r(x, y) \}^2 d\mu(x) d\mu(y),$$

где  $|\cdot|$  — евклидова норма в  $R^k$ . При этом распределение  $\mu$  предполагается неизвестным, но зато имеется выборочная матрица близости  $\{r(x_i, x_j)\}_{i,j=1}^n$  для выборки точек  $x_1, \dots, x_n \in X$ , распределенных в  $X$  в соответствии с мерой  $\mu$  (например, полученной в результате случайных независимых испытаний согласно распределению  $\mu$ ).

В силу ограниченности функции близости  $r$  областью определения функционала  $Z_k$  является пространство  $L = L_k^4(X, \mu)$ , а в силу инвариантности функционала  $Z_k$  относительно преобра-

зований движений в  $R^k$  можно предполагать, что норма  $|\cdot|$  в  $R^k$  диагональна, т. е. для любого  $\xi \in R^k$

$$|\xi|^2 = \sum_{i=1}^k \xi_i^2,$$

и ограничиться рассмотрением задачи минимизации функционала на подмножестве  $L = L_k^4(X, \mu) \subset L$  ортогональных центрированных вектор-функций [16], т. е. вектор-функций  $f = L$ , удовлетворяющих равенствам

$$\langle f_i, 1 \rangle = 0, \langle f_i, f_j \rangle = 0, i, j \in \overline{1, k}, i \neq j,$$

где  $\langle \cdot, \cdot \rangle$  — скалярное произведение в  $L^2(X, \mu)$ , а 1 — функция, тождественно равная единице.

Задача минимизации функционала  $Z_k$  на множестве  $L$  называется задачей  $Z_k$ -шкалирования (функционального шкалирования), а ее решение — отображением  $Z_k$ -шкалирования. Для отображения функционального шкалирования  $f$  получены следующие оценки [12, подразд. 2.3]:

$$\|f\|_4 = \left\{ \int |f(x)|^4 d\mu(x) \right\}^{1/4} < \sqrt{r^*};$$

$$\text{ess sup } |f(x)| \leq D_r;$$

$$x \in (X, \mu)$$

где  $\text{ess sup}$  обозначает существенный  $\sup$ ,  $r^* = \sup r(x, y)$ ,  $D_r = \{[3 + (4k)^{1/2}] r^*\}^{1/2}$ ,  $x, y \in X$ .

## 1.2. Задача $Z_k$ -шкалирования как задача снижения размерности в функциональном пространстве

Применяя известную процедуру преобразования данных о близости в многомерные данные [9, с. 23], согласно которой значения близости  $r(x, y)$  между неким объектом  $x$  и всеми другими объектами рассматриваются как переменная  $z(y) = z_x(y)$ , можно трактовать задачу функционального шкалирования как задачу снижения размерности в функциональном (бесконечномерном) пространстве [12, подразд. 1.1.4]. Компактность пространства  $X$  позволяет аппроксимировать ее некоторой задачей снижения размерности в конечномерном пространстве с аналитически заданной функцией близости (например, евклидовой) [12, подразд. 1.1.5].

## 1.3. Модели группировки параметров

В случае, если  $X = R^m$ , а  $r(\xi, \eta) = \sum_{i=1}^m |\xi_i - \eta_i|^2$ , рассматривается задача  $k$ -мерного шкалирования, заключающаяся в минимизации функционала

$$Z_k^{**}(f, z) = \sum_{l=1}^k \iint \{ |f_l(x) - f_l(y)|^2 - \\ - \sum_{i=1}^m z_i^{(l)} (x_i - y_i)^2 \} d\mu(x) d\mu(y),$$

где  $f \in L_k^4(X, \mu)$ , а  $z = \{z_i^{(l)}\}_{i \in \overline{1, m}; l \in \overline{1, k}}$  — набор вещественных  $z_i^{(l)}$ , удовлетворяющих следующему условию: для любого  $i \in \overline{1, m}$

$$\sum_{l=1}^k z_i^{(l)} = 1, \quad z_i^{(l)} \in [0, 1].$$

Элемент  $z_i^{(l)}$  набора  $z$  естественным образом интерпретируется как степень включенности  $i$ -го признака в  $l$ -ю группу, причем для любого признака его суммарная включенность во все группы равна единице.

Задача минимизации функционала  $Z_k^{**}$  названа задачей размытой группировки параметров [12, подразд. 1.2]. Если же для набора  $z$  вместо включения  $z_i^{(l)} \in [0, 1]$  потребовать  $z_i^{(l)} \in \{0, 1\}$ , то получается задача экстремальной группировки параметров. В рамках корреляционных моделей такие задачи рассмотрены, например, в [2, 10].

#### 1.4. Проекционные модели функционального шкалирования

В [14, 27—29] предложены и исследованы проекционные модели многомерного шкалирования, основанные на максимизации функционала

$$\psi(f) = \frac{(r, \rho_f)}{\|\rho_f\|_2},$$

где  $(\cdot, \cdot)$  — скалярное произведение в  $L_1^2(X, \mu)$  а функция  $\rho_f(x, y)$  определяется равенством

$$\rho_f(x, y) = |f(x) - f(y)|^2.$$

Название модели обусловлено тем, что значение  $\psi(f)$  функционала  $\psi$  можно интерпретировать как величину проекции функции  $r$  на направление  $\rho_f$ . В этом случае величина  $\psi(f)/\|r\|_2$  интерпретируется как коэффициент корреляции между функциями  $r$  и  $\rho_f$ . В [27, 29] показано, что функционал  $\psi$  задает лишь одну из некоторого семейства эквивалентных моделей шкалирования. В это семейство, в частности, входит модель  $Z_k$ -шкалирования. В [30] проекционные идеи используются при разработке алгоритмов построения начального приближения в итеративных схемах метрического шкалирования.

## 2. Параметрические представления в задачах функционального шкалирования

### 2.1. Параметрические представления функции близости

В основе параметрических представлений отображения функционального шкалирования лежат следующие представления функции близости в виде билинейного ряда, используемого в теории линейных интегральных уравнений:

$$r(x, y) = \sum_{i=1}^{\infty} P_i(x) Q_i(y). \quad (2)$$

Как известно, из включения  $r \in L^2(X^2, \mu^2)$  следует существование такой ортонормальной системы  $\{\varphi_i\}_{i=1}^{\infty}$  функций  $\varphi_i \in L^2(X, \mu)$  и такой последовательности  $\{\lambda_i\}_{i=1}^{\infty}$  вещественных  $\lambda_i$ , что имеет место представление (2), где

$$P_i = \varphi_i, \quad Q_i = \lambda_i \varphi_i,$$

и при этом ряд (2) сходится в  $L^2$ ,  $\lambda_i \rightarrow 0$  и

$$\sum_{i=1}^{\infty} \lambda_i^2 = \|r\|_2^2.$$

Функция близости называется вырожденной, если для нее существует билинейный ряд (2), состоящий из конечного числа членов. В частности, функция евклидовой близости  $r_l(x, y) = |x - y|^2$  в  $m$ -мерном евклидовом пространстве  $R^m$  вырождена и для нее при любом положительном  $a$  имеет место следующее конечное билинейное представление:

$$\begin{aligned} r_e(x, y) &\equiv \frac{2}{a} (|x|^2 + a)(|y|^2 + a) - \\ &- \frac{2}{a} (|x|^2 - a)(|y|^2 - a) - 2 \sum_{i=1}^m x_i y_i, \end{aligned}$$

где

$$x = (x_1, \dots, x_m), \quad y = (y_1, \dots, y_m) \in R^m.$$

### 2.2. Параметризация отображения $Z_k$ -шкалирования

Рассмотрим задачу  $Z_k$ -шкалирования для пространства  $\mathcal{X} = (X, r, \mu)$ . Вопросы параметризации отображения  $Z_k$ -шкалирования с помощью билинейного представления (2) функции близости  $r$  рассматривались в [13, 26]. Приведенное ниже решение этой задачи было получено в [12; 15, подразд. 4.2].

Будем предполагать, что для функции близости  $r$  имеет место билинейное представление (2), где  $\{P_i\}_{i \in I_r}$  — система ортонор-

мальных функций,  $P_i = 0$  при  $i \notin I_r$ ,  $Q_i = \lambda_i P_i$ . Обозначим через  $l^1$  пространство суммируемых последовательностей. Пусть  $\omega = \{\omega_i\}_{i=1}^\infty \in l^1$ . Определим наборы  $\omega^{(l)}$ ,  $l \in \overline{0, k}$ , равенствами

$$\omega_i^{(0)} = \omega_i, \quad i \in \overline{1, 2k}, \quad \omega_i^{(l)} = \omega_{k(i+1)} + l, \quad i \in I_r, \quad l \in \overline{1, k},$$

Обозначим через  $\Omega_r$  совокупность таких последовательностей  $\omega = \{\omega_i\}_{i=1}^\infty$ , что

$$|\omega|^2 = \sum_{l=1}^k |\omega^{(l)}|^2 < +\infty,$$

где

$$|\omega^{(l)}|^2 = \sum_{i \in I_r} \lambda_i^{-2} [\omega_i^{(l)}]^2.$$

Имеет место включение  $\Omega_r \subset l^1$ .

Для любой последовательности  $\omega \in \Omega_r$  определим вектор-функции  $z(\omega)$  и  $v(\omega)$  следующим образом: для любых  $l \in \overline{1, k}$  и  $x \in X$

$$[z_l(\omega)](x) = z_l(\omega; x) = \sum_{i=1}^k \omega_i^{(0)} + 2\omega_l^{(0)} - \sum_{i \in I_r} \theta_i P_i(x);$$

$$[v_l(\omega)](x) = v_l(\omega; x) = \omega_{k+l}^{(0)} - \sum_{i \in I_r} \omega_i^{(l)} P_i(x),$$

где  $\theta_i = \langle Q_i, 1 \rangle$ . В силу сделанных относительно функции близости  $r$  предположений для любых  $l \in \overline{1, k}$ ,  $\omega \in \Omega_r$  и  $y \in X$  имеют место неравенства

$$|z_l(\omega; x) - z_l(\omega; y)| \leq \sigma_r(x, y);$$

$$|v_l(\omega; x) - v_l(\omega; y)| \leq |\omega^{(l)}| \sigma_r(x, y),$$

где функция  $\sigma_r$  определяется равенством (1), а также неравенства

$$z_l^*(\omega) = \sup_{x \in X} |z_l(\omega; x)| \leq \sum_{i=1}^k |\omega_i| + 2|\omega_l| + r^*;$$

$$v_l^*(\omega) = \sup_{x \in X} |v_l(\omega; x)| \leq |\omega_{k+l}| + r^* |\omega^{(l)}|.$$

Далее, для любых  $\omega \in \Omega_r$  и  $x \in X$  положим

$$I(\omega; x) = \{i \in \overline{1, k} \mid v_i(\omega; x) \neq 0\}$$

и рассмотрим следующее уравнение относительно вещественного  $u$ :

$$u = \sum_{i \in I(\omega; x)} \frac{v_i^2(\omega; x)}{[u + z_i(\omega; x)]^2}. \quad (3)$$

Обозначим через  $u(\omega; x)$  максимальный вещественный неотрицательный корень этого уравнения<sup>2</sup>. Для функции  $u$  имеет место

<sup>2</sup> Уравнение (3) сводится к алгебраическому уравнению нечетной степени  $2k+1$ , поэтому (3) имеет вещественные корни, число которых не превосходит  $2k+1$ .

следующая оценка: для любого  $\omega \in \Omega_r$ ,

$$0 \leq \sup_{x \in X} u(\omega; x) \leq D_r^2(\omega) < +\infty,$$

где

$$D(\omega) = r^* + \sum_{i=1}^k |\omega_i| + 2 \max_{i \in \overline{1, k}} |\omega_i| + \\ + \{r^* |\omega| + [\sum_{i=1}^k |\omega_{k+1}|^2]^{1/2}\}^{2/3}.$$

Определим теперь отображение  $\hat{h}: \Omega_r \rightarrow L_k^\infty$  следующим образом: для любого  $\omega \in \Omega_r$  положим  $\hat{h}(\omega) = g$ , где  $k$ -мерная вектор-функция  $g = (g_1, \dots, g_k)$  определяется на  $X$  равенством

$$g_i(x) = g_i(\omega; x) = \\ = \begin{cases} 0, & \text{если } v_i(\omega; x) = 0; \\ \frac{v_i(\omega; x)}{u(\omega; x) + \zeta_i(\omega; x)}, & \text{если } v_i(\omega; x) \neq 0. \end{cases}$$

В силу (3) определение вектор функции  $g$  корректно и для любого  $x \in X$  справедливо равенство

$$\|g(\omega; x)\|^2 = u(\omega; x)$$

и, следовательно, неравенство

$$\|\hat{h}(\omega)\|_\infty \leq D_r(\omega),$$

т. е. отображение  $\hat{h}$  переводит любое ограниченное в  $\Omega_r$  множество в множество вектор-функций, ограниченное в  $L_k^\infty$ . Положим

$$D'_r = D_r [D_r + (D_r + r^*) \sqrt{k}];$$

$$\Omega'_k = \{\omega \in \Omega_r \mid \|\omega\|_1 \leq D'_r\}$$

и рассмотрим множество  $\hat{\Phi}(r, \mu) = \hat{h}(\Omega'_r)$ . Можно показать [12, разделы 3, 4], что вектор-функции этого множества кусочно непрерывны и равномерно непрерывны на множествах непрерывности. Заметим, что отображение  $\hat{h}$  параметризует вектор-функции

$$g \in \hat{\Phi}: g = \hat{h}(\omega), \quad \omega \in \Omega_r.$$

Далее, определим для любой вектор-функции  $f \in L_k^4(X, \mu)$  последовательность  $\omega = \overline{\omega}(f)$  равенствами

$$\overline{\omega}_i^{(v)}(f) = \langle f_{i,1}^2, 1 \rangle, \quad \overline{\omega}_{k+i}^{(0)}(f) = \langle f_i, |f|^2 \rangle, \quad i \in \overline{1, k};$$

$$\overline{\omega}_i^{(l)}(f) = \langle f_i, Q_l \rangle, \quad i \in I_r, \quad [l \in \overline{1, k}_*]$$

Тогда [12, подразд. 4.2.3] при естественных предположениях относительно связи функции близости  $r$  и вероятностной меры  $\mu$

имеет место следующее утверждение об  $\omega$ -параметризации отображений  $Z_k$ -шкалирования.

**Теорема 1.** Если центрированная ортогональная вектор-функция  $f$  является отображением  $Z_k$ -шкалирования, то вектор-функция  $g = \hat{f}(\omega)$ , где  $\omega = \bar{\omega}(f) \in \Omega_r$ ,  $\mu$ -эквивалентна  $f$  и, следовательно, также является отображением  $Z_k$ -шкалирования.

Так как для вырожденных функций близости  $r$  пространство  $\Omega_r$  состоит из последовательностей, количество элементов которых конечно и равно величине  $k_q$ , где  $q = |I_r|$  — количество отличных от нуля членов в билинейном ряде (2), то теорема 1 задает конечно-параметрическое представление отображения  $Z_k$ -шкалирования в случае вырожденности рассматриваемой функции близости. В частности, при рассмотрении евклидовой близости размерности  $m$  имеем  $q = m + 2$ . Таким образом, для вырожденных функций близости  $r$  сложность задачи метрического шкалирования (сложность представления ее решения) определяется количеством ненулевых элементов в билинейном ряде функции  $r$ . В общем случае, как будет показано ниже, сложность задачи метрического шкалирования определяется скоростью сходимости к нулю величин  $\lambda_i$  в представлении (2) с попарно ортогональными и нормированными функциями  $P_i$  и с функциями  $Q_i = \lambda_i P_i$ .

Из теоремы 1 также следует, что задача  $Z_k$ -шкалирования эквивалентна задаче минимизации функционала

$$\hat{Z}_k(\omega) = \iint \{ |g(\omega; x) - g(\omega; y)|^2 - r(x, y) \}^2 d\mu(x) d\mu(y)$$

на подмножестве последовательностей  $\Omega_r^0 \subset \Omega_r'$ , определяемом равенством

$$\Omega_r^0 = \{ \omega \in \Omega_r' \mid \langle g_i(\omega), 1 \rangle = \langle g_i(\omega), g_j(\omega) \rangle = 0; i, j \in \overline{1, k}; i \neq j \}.$$

При этом глобальный минимум функционала  $\hat{Z}_k(\omega)$  является решением (в  $\Omega_r^0$ ) следующего уравнения:

$$\hat{\mathfrak{z}}(\omega) = \omega, \quad (4)$$

где оператор  $\hat{\mathfrak{z}}$  определяется равенством  $\hat{\mathfrak{z}}(\omega) = \bar{\omega}(\hat{v}(\omega))$ . Таким образом, любая минимизирующая функционал  $\hat{Z}_k$  последовательность  $(\hat{Z}_k)$  — неподвижная точка оператора  $\hat{\mathfrak{z}}$  в пространстве  $\Omega_r^0$ . В случае вырожденности функции близости  $r$  уравнение (4) является системой нелинейных уравнений, состоящей из  $kq$  уравнений, где  $q$  — число ненулевых членов в билинейном представлении  $r$ .

Задача параметризации отображения шкалирования в функциональных моделях группировки параметров рассматривалась в [3, 4, 12, подразд. 4.2]. Ее решение строится с помощью систем параметров отображений  $Z_1$ -шкалирования.

### 3. Структурная устойчивость<sup>3</sup> моделей функционального шкалирования

Вопросы устойчивости моделей  $Z_k$ -шкалирования рассматривались в [12, разд. 4.3]. Обозначим через  $T_{r, \mu}$  множество точек минимума функционала  $Z_k \equiv Z_k(r, \mu)$  в пространстве  $L$ . Устойчивость (структурная устойчивость) модели  $Z_k(r, \mu)$ -шкалирования понимается как грубость множества  $T_{r, \mu}$  относительно малых возмущений структурных параметров  $r$  и  $\mu$  в рассматриваемой модели  $Z_k(r, \mu)$ -шкалирования. Как и в любой модели анализа эмпирической информации, структурная устойчивость рассматриваемой модели шкалирования является необходимым условием возможности ее применения в прикладных областях. Кроме того, вопросы устойчивости отображения  $Z_k(r, \mu)$ -шкалирования при малых изменениях функции близости  $r$  возникают в связи со следующим обстоятельством. В разд. 2 для вырожденных функций близости, т. е. для функций, допускающих конечное представление

$$r(x, y) = \sum_{i=1}^q P_i(x) Q_i(y), \quad (5)$$

где  $q$  — некоторое натуральное число, получено нелинейное конечно-параметрическое представление для отображения  $Z_k(r, \mu)$ -шкалирования и приведена конечная система (4) нелинейных уравнений для вычисления определяющих эту функцию параметров. В этом случае устойчивость множества  $T_{r, \mu}$  при малых возмущениях функции  $r$  позволяет использовать алгоритмы решения задачи  $Z_k$ -шкалирования для вырожденных функций близости в общем случае, так как любая функция близости  $r$  может быть с любой степенью точности аппроксимирована вырожденной функцией — некоторым отрезком билинейного ряда (2).

При исследовании вопросов структурной устойчивости рассматривается совокупность  $M$  всех вероятностных мер, которая является выпуклым компактом в множестве всех мер, т. е. в пространстве, сопряженном к пространству непрерывных функций, заданных на компакте  $X$ . Множеству  $M$ , в частности, принадлежат меры, сосредоточенные в точках множества  $X$ , а также их взвешенные суммы. Рассмотрим так называемую слабую (широкую) топологию в пространстве мер, которая может быть задана с помощью некоторой метрики следующим образом. Пусть  $\{\varphi_i\}_{i=1}^{\infty}$  — некоторый ортогональный базис пространства  $L^2(X, \mu)$ . Для любой меры  $\mu \in M$  положим

$$\rho_c(\mu) = \max_{l \geq 1} \min \left\{ \frac{1}{l}, \sqrt{\sum_{i=1}^l [\varphi_i, \mu]^2} \right\},$$

где  $[\varphi_i, \mu] = \int \varphi_i(x) d\mu(x)$ . Как легко проверить,  $\rho_c$  — неотри-

<sup>3</sup> Проблемам устойчивости посвящена работа А. И. Орлова «Устойчивость в социально-экономических моделях» (М., 1979).

цательный функционал, удовлетворяющий неравенству треугольника

$$\rho_c(\mu_1 + \mu_2) \leq \rho_c(\mu_1) + \rho_c(\mu_2).$$

Кроме того, сходимость  $\rho_c(\mu_j) \rightarrow 0$  при  $j \rightarrow \infty$  для последовательности  $\mu_j \in M$  означает, что для любого  $i$  имеет место сходимость  $[\varphi_i, \mu_j] \rightarrow 0$  при  $j \rightarrow \infty$ , откуда в силу фундаментальности  $\{\varphi_i\}$  следует, что для любой функции  $\varphi \in L^2$   $[\varphi, \mu_j] \rightarrow 0$ .

Согласно [12, подразд. 4.3.1] модель  $Z_k(r, \mu)$ -шкалирования топологически  $p$ -устойчива в точке  $t \in T_{r, \mu}$ , если для любого  $\varepsilon > 0$  существует такое  $\delta > 0$ , что для любой функции близости  $r'$  из неравенства  $\|r - r'\|_\infty < \delta$  следует существование такой вектор-функции  $f \in T_{r', \mu}$ , что имеет место неравенство

$$\|t - f\|_p < \varepsilon. \quad (6)$$

Модель  $Z_k(r, \mu)$ -шкалирования статистически слабо  $p$ -устойчива в точке  $t \in T_{r, \mu}$ , если для любого  $\varepsilon > 0$  существует такое  $\delta > 0$ , что для любой меры  $\mu' \in M$  из неравенства  $\rho_c(\mu, \mu') < \delta$  следует существование такой вектор-функции  $f \in T_{r, \mu'}$ , когда имеет место неравенство (6). Модель  $Z_k(r, \mu)$ -шкалирования структурно слабо  $p$ -устойчива в точке  $t \in T_{r, \mu}$ , если в этой точке она как топологически, так и статистически слабо  $p$ -устойчива.

Пусть теперь  $X_N = \{x_i\}_{i=1}^N \subset X$  — некоторая конечная выборка объема  $N$  из рассматриваемой генеральной совокупности  $\mathcal{X}$ , а для функции близости  $r$  имеет место билинейное представление (2), где  $P_i = \varphi_i$ ,  $Q_i = \lambda_i \varphi_i$ . Определим для любого натурального  $q$  функцию  $r_q$  равенством

$$r_q(x, y) = \sum_{i=1}^q \lambda_i \varphi_i(x) \varphi_i(y), \quad (7)$$

а меру  $\bar{\mu}$  — равенством

$$\bar{\mu} = \sum_{i=1}^N \beta_i \mu_i, \quad (8)$$

где  $\mu_i$  — вероятностная мера, сосредоточенная в точке  $x_i$  ( $\mu_i\{x_i\} = 1$ ), а неотрицательные  $\beta_i$  удовлетворяют равенству  $\sum_{i=1}^N \beta_i = 1$ . Пусть теперь  $t \in T_{r_q, \bar{\mu}}$ . Тогда в рамках проблемы структурной слабой  $p$ -устойчивости может быть решена следующая задача: для заданного  $\varepsilon > 0$  указать такие  $q$  и  $N$ , чтобы существовала вектор-функция  $f \in T_{r, \mu}$ , удовлетворяющая неравенству (6).

В [12, подразделы 4.3.2, 4.3.3] получены достаточные условия топологической устойчивости, согласно которым для вектор-функции  $t \in T_{r, \mu}$  при выполнении некоторых условий невырожденности при достаточно малой величине  $\varepsilon = \|r - r'\|_\infty$  имеет место неравенство

$$\|f - t - \Gamma_\Delta\|_p \leq c\varepsilon^2,$$

где  $c$  — некоторая положительная константа;  $f$  — некоторая вектор-функция множества  $T_{r', \mu}$ ;  $\Gamma$  — некоторый ограниченный опе-

ратор пространства  $L_k^p$ , а вектор-функция  $\Delta \in L_k^\infty$  определена равенствами

$$\Delta_i(x) = \int [t_i(x) - t_i(y)] [r(x, y) - r'(x, y)] d\mu(y), i \in \overline{1, k}.$$

Заметим, что если в билинейном представлении (2) имеют место равенства  $P_i(x) = \varphi_i(x)$ ,  $Q_i(x) = \lambda_i \varphi_i(x)$ , а значения функций  $\varphi_i$  ограничены в совокупности:

$$\sup_{x \in X} |\varphi_i(x)| \leq \alpha,$$

то для любого натурального  $q$  при  $r' = r_q$ , где функция  $r_q$  определяется равенством (7), имеют место неравенства

$$\|\Delta\|_\infty \leq \alpha^2 C \sum_{i=q+1}^{\infty} |\lambda_i|, \quad \varepsilon \leq \alpha^2 \sum_{i=q+1}^{\infty} |\lambda_i|,$$

где  $C$  — некоторая положительная константа, что доказывает высказанное в подразд. 2.2 утверждение о том, что сложность задачи  $Z_k$ -шкалирования определяется скоростью сходимости к нулю величин  $\lambda_i$ .

При исследовании вопросов структурной слабой устойчивости за счет использования условий топологической устойчивости достаточно ограничиться случаем вырожденности функции близости  $r$ , т. е. случаем существования билинейного представления (5). В этой ситуации для точек множества  $T_{r, \mu}$  имеет место указанное в подразд. 2.2 конечно-параметрическое представление, зависящее от  $kq$  параметров.

С помощью этого представления в [12, подразд. 4.3.4] для любой вектор-функции  $t = g(\bar{\omega}(t)) \in T_{r, \mu}$  указаны в явном виде такие две конечные системы функций  $F(t)$  и  $F_1(t)$ , состоящие соответственно из  $k^2(q+2)^2 + k(q+2) + k^3(k+2)$  и  $k(k^2 + 2k + 3)$  элементов, что при выполнении некоторых условий невырожденности модель  $Z_k$ -шкалирования статистически слабо  $p$  устойчива (при  $p = 2$ ) в точке  $t$ . При этом, если обозначить через  $\bar{\rho}_t$  и  $\bar{\rho}_t$  полуметрики в  $M^2$ ;

$$\bar{\rho}_t(\mu, \nu) = \max_{j \in F(t)} |[f, \mu] - [f, \nu]|;$$

$$\bar{\rho}_t(\mu, \nu) = \max_{f \in F_1(t)} |[f, \mu] - [f, \nu]|,$$

то указываются такие величины  $c_0, \delta_0 > 0$ , что для любого  $\varepsilon > 0$  и для любой вероятностной меры  $\nu \in M$ , удовлетворяющей неравенствам

$$\bar{\rho}_t(\mu, \nu) \leq c_0 \varepsilon, \quad \bar{\rho}_t(\mu, \nu) \leq \delta_0, \quad (9)$$

существует такой набор параметров  $\omega^*$ , что

$$\|\omega^* - \bar{\omega}(t)\| \leq \varepsilon, \quad (10)$$

а для вектор-функции  $f = g(\omega^*)$  имеют место включения

$$f \in L, f \in T_{r, \nu}.$$

Воспользуемся этими результатами в случае, когда имеется некоторая выборка  $X_N = \{x_i\}_{i=1}^N \subset \mathcal{X}$  объема  $N$  из рассматриваемой генеральной совокупности  $\mathcal{X}$ , а мера  $\nu$  имеет вид (8), где для любого  $i \in \overline{1, N}$   $\beta_i = 1/N$ . Тогда получаем, что при решении задачи  $Z_k(r, \mu)$ -шкалирования выборка  $X_N$   $\epsilon$ -представительна, если для меры  $\nu$  имеют место неравенства (9), т. е. достаточно близки средние и выборочные средние значения для функций конечной системы  $F(t)$ . В этом случае, зная, например, вероятностное порождающее правило для элементов выборки  $x_i$ , можно строить статистические оценки для распределения величины (10).

## 4. Алгоритмическое и программное обеспечение моделей функционального шкалирования

### 4.1. Параметрические алгоритмы

В [12, подразд. 4.4] приведены некоторые алгоритмы функционального шкалирования, основанные на построенном параметрическом представлении отображений  $Z_k$ -шкалирования. При этом предполагается, что рассматриваемые модели функционального шкалирования структурно слабо устойчивы. Тогда, не уменьшая общности, можно считать, что функция близости  $r$  вырождена и имеет место билинейное представление (5). Множество  $X$  состоит из  $N$  точек:  $X = X_N = \{x_i\}_{i=1}^N$ , а мера  $\mu$  имеет вид (8). Иначе говоря, задача  $Z_k$ -шкалирования рассматривается как задача шкалирования на конечной выборке с вырожденной функцией близости. Заметим, что построенное по найденному вектору  $\omega$  параметров отображение  $Z_k$ -шкалирования формулой  $f(x) = g(\omega; x)$  определено на всем пространстве  $X$ .

Предложенные параметрические алгоритмы естественным образом разбиваются на две группы. К первой группе относятся алгоритмы, основанные на решении системы уравнений (4) для параметров шкалирования [12, подразд. 4.4.1]. Эта система состоит с учетом условия центрированности и ортогональности из  $k(q + (k - 1)/2)$  нелинейных уравнений для нахождения  $kq$  параметров и решается методом наименьших квадратов с помощью итеративных градиентных схем (наискорейший спуск, метод сопряженных градиентов, оврагов и т. п.). При этом решается оптимизационная задача (минимизации) размерности  $kq$ , а для проведения одной итерации достаточно затратить  $k^2 q^2 N$  элементарных операций. Для сравнения укажем, что в традиционных точечных вычислительных моделях метрического шкалирования необходимо решать оптимизационные задачи размерности  $kN$ , а для проведения одной итерации в градиентных схемах необходимо затратить  $k^2 N^2$  элементарных операций.

Таким образом, преимущество предлагаемых алгоритмов перед традиционными является независимость размерности решаемой оптимизационной задачи от объема выборки (эта размерность

определяется лишь сложностью решаемой задачи шкалирования) и более слабая (линейная вместо квадратичной) зависимость вычислительной трудоемкости реализации одной итерации от объема рассматриваемой выборки.

Особенностью предложенных алгоритмов является то, что оператор  $\mathfrak{z}$  в уравнении (4) не является дифференцируемой функцией вектора и может иметь разрывы 1-го рода на  $N + 1$  явно заданных гиперповерхностях пространства параметров, из которых  $N$ -гиперплоскости. В этой ситуации для применения обычных градиентных схем используется метод Филиппова доопределения градиентного поля в точках разрыва [24].

Ко второй группе параметрических методов относятся алгоритмы прямой минимизации, основанные на минимизации функционала  $\hat{Z}_k(\omega)$  (см. подразд. 2.2), который в сделанных предположениях является функцией конечного числа переменных — координат вектора при условии  $\omega \in \Omega_k^0$  [12, подразд. 4.4.2]. Применяя для решения этой задачи градиентные схемы, получаем итеративные алгоритмы решения оптимизационной задачи размерности  $kq$ , требующие для реализации одной итерации  $k^2 q^2 N$  элементарных операций. Так же, как и выше, градиент функционала  $\hat{Z}_k$  кусочно непрерывен и доопределяется в точках разрыва в соответствии с [24].

## 4.2. Рекурсивные алгоритмы

Рекурсивные алгоритмы функционального шкалирования предложены в [17] и получены за счет применения к задаче минимизации функционала  $Z_k(f)$  итеративных градиентных схем и метода Ньютона (в функциональном пространстве  $L$ ).

В отличие от параметрических в рекурсивных алгоритмах не требуется знание билинейного представления для функции близости, хотя при наличии простого билинейного ряда вычислительная трудоемкость рекурсивных алгоритмов может существенно уменьшаться. Для реализации этих алгоритмов достаточно иметь выборочную матрицу близости (матрицу близостей между объектами имеющейся выборки).

При решении задач функционального шкалирования по имеющейся конечной выборке рекурсивные алгоритмы в вычислительном плане аналогичны традиционным (точечным) алгоритмам метрического шкалирования. Однако в отличие от последних рекурсивные алгоритмы позволяют получать решение в функциональном виде за счет аппроксимации средних значений выборочными средними для функций

$$\bar{r}(x) = \int r(x, y) d\mu(y) \sim \frac{1}{N} \sum_{i=1}^N r(x, x_i);$$

$$\bar{r}_h(x) = \int h(y) r(x, y) d\mu(y) \sim \frac{1}{N} \sum_{i=1}^N h(x_i) r(x, x_i).$$

### 4.3. Пакет программ инвариантного шкалирования

Программное обеспечение рассмотренных выше методов многомерного метрического шкалирования представлено в виде пакета программ инвариантного шкалирования (ПИШ), который является подсистемой общей системы анализа социально-экономических данных ИСЭП АН СССР и реализован на ЭВМ БЭСМ-6 в рамках мониторинжной системы ДУБНА [25]. Особенностью системы управления пакетом ПИШ является возможность описания «плана исследования», реализующего какую-либо методику обработки с помощью методов метрического шкалирования с последующим исполнением задания за один или, возможно, несколько сеансов счета [25]. Эта возможность важна в связи с тем, что в силу вычислительной трудоемкости методов многомерного шкалирования обработка реальных социологических массивов (несколько тысяч анкет и большое количество признаков) по сложным иерархическим методикам типологического анализа может потребовать десятки часов машинного времени высокопроизводительной ЭВМ (класса БЭСМ-6). Поэтому общий счет целесообразно разбивать на кванты приемлемой для используемой вычислительной системы длительности.

Вторая важная особенность системы управления пакетом ПИШ — возможность образования временных и постоянных макродиректив из директив, имеющихся в пакете [25]. Такие макросредства позволяют фиксировать различные сложные и многоэтапные методики обработки в виде одной директивы пакета, что значительно упрощает работу с пакетом широкому кругу пользователей.

И, наконец, специальные системные средства развития пакета позволяют легко расширять возможности пакета за счет введения в него новых проблемных модулей.

## 5. Проблема многоэкстремальности в вычислительных моделях метрического шкалирования

Многоэкстремальность традиционных оптимизационных моделей метрического шкалирования многократно обсуждалась в литературе по многомерному шкалированию (см., например, [1, 21]). Выводы о многоэкстремальности, как правило, делались на основании вычислительных экспериментов, заключавшихся в спуске из различных начальных точек градиентным методом, что не всегда корректно [19, с. 889]. Тем не менее выводы, сделанные в [21, с. 51—52] о многоэкстремальности моделей метрического шкалирования, представляются убедительными, поскольку на основании многочисленных вычислительных экспериментов там сформулирован «физический» механизм порождения локальных

экстремумов. Ниже (подразд. 5.3) приводятся результаты, доказывающие факт многоэкстремальности для одного класса задач одномерного метрического шкалирования. Многоэкстремальность моделей шкалирования приводит к необходимости решения задачи поиска глобального (или хотя бы достаточно глубокого локального) экстремума. Этим вопросам посвящены подразделы 5.1 и 5.2.

### 5.1. Нелокальное необходимое условие функционального шкалирования

Воспользуемся результатами подразд. 2.2 о параметризации отображений  $Z_k$ -шкалирования. Определим оператор  $\mathfrak{h}$  пространства  $L_k^\infty$  равенством

$$\mathfrak{h}(f) = \hat{\mathfrak{h}}(\bar{\omega}(f)).$$

Для этого оператора возможна интерпретация как сужения отображения  $\hat{\mathfrak{h}}$  на последовательности  $\bar{\omega}$  вида  $\bar{\omega}(f)$ . Тогда из теоремы 1.1 следует, что любое отображение  $Z_k$ -шкалирования является решением (в  $L_k^\infty$ ) следующего уравнения:

$$\mathfrak{h}(f) = f, \tag{11}$$

т. е. является неподвижной точкой оператора  $\mathfrak{h}$ .

Равенства (11) и (4) являются необходимыми условиями функционального шкалирования соответственно в функциональной и параметрической формах. При выполнении естественных предположений о связи функции близости  $r$  и вероятностной меры  $\mu$  эти необходимые условия нелокальны. Ограничимся рассмотрением условия (11).

Если пространство  $X$  не имеет точек положительной меры, то в  $L_k^4$  любой минимум функционала  $Z_k$  (удовлетворяющий условию ортогональности и центрированности) является неподвижной точкой оператора  $\mathfrak{h}$ . Если же  $f$  — точка максимума функционала  $Z_k$  в  $L_k^4$  и на некотором множестве положительной меры уравнение (3) имеет более одного вещественного корня, то  $f$  не является неподвижной точкой оператора  $\mathfrak{h}$ . В аналогичной ситуации также не всякий минимум функционала  $Z_k$  является неподвижной точкой оператора  $\mathfrak{h}$ .

Проведенные для случая одномерного шкалирования и конечности пространства  $X$  вычислительные эксперименты показали, что ряд многоэкстремальных в рамках традиционных моделей задач метрического шкалирования являются одноэкстремальными при использовании необходимого условия (4) и, следовательно, (11).

## 5.2. Построение начального приближения в итеративных схемах функционального шкалирования методами квазидинамического программирования

При реализации итеративных методов метрического шкалирования возникает проблема построения начального приближения, важность которой связана также с многоэкстремальностью задач многомерного шкалирования. В [12 разд. 4.5] предложен вычислительно эффективный метод построения начального приближения в функциональном виде, реализующий идеи динамического программирования. Для получения такого начального приближения задача многомерного метрического шкалирования на конечной выборке рассмотрена с позиции динамического программирования как метода последовательного (по объектам) построения образов объектов при отображении шкалирования (координат объектов в пространстве  $R^k$ ). Эта позиция представлена в виде метода квазидинамического программирования, который ранее применялся при решении ряда задач дискретной оптимизации [11, 16]. Применение этого метода позволило построить хорошо интерпретируемые правила приближенного решения задачи многомерного шкалирования, обобщающие известные («локальные») эвристики, и в рамках которых сочетаются «тактические» (локальные) и «стратегические» (глобальные) аспекты. Полученные правила допускают простые вычислительные реализации: порядка  $N^2 - N^3$  элементарных операций для различных реализаций в случае, если данные представлены в виде матрицы близостей  $\{r_{ij}\}_{i,j}^N = 1$ , и  $q^2N - q^2N^2$  элементарных операций, если для функции близости  $r$  известно билинейное представление (5). Применение этих методов для построения начального приближения в итеративных схемах, основанных на нелокальном необходимом условии (4), позволяет, в частности, повысить степень нелокальности этих схем.

## 5.3. Дискретные представления в задачах метрического шкалирования

В [5, 18] рассмотрена задача одномерного метрического шкалирования как задача минимизации функционала

$$S(x) = \sum_{i,j=1}^n (|x_i - x_j| - r_{ij})^2,$$

где  $x = (x_1, \dots, x_n)$  — конфигурация точек на прямой — вектор образов исследуемых объектов в  $R^1$ , а  $r_{ij}$  — величина близости между  $i$ -м и  $j$ -м объектами. Каждой конфигурации  $x \in R^n$  естественным образом сопоставляется перестановка  $l(x) = (l_1, \dots, l_n)$ , где  $l_i = l_i(x)$  — порядковый номер величины  $x_i$  в упорядоченном по неубыванию множестве  $\{x_j\}_j^n = 1$ , т. е.  $x_{l_i} \leq x_{l_j} \leq \dots \leq x_{l_n}$ . В [5] вводится отображение  $Z(l)$  множества

перестановок  $(l_1, \dots, l_n)$  из  $n$  элементов в совокупность всех конфигураций  $R^n$ :

$$z_{li}(l) = \frac{1}{n} \left( \sum_{s=1}^i r_{li l_s} - \sum_{s=i}^n r_{li l_s} \right), \quad i \in \overline{1, n}.$$

В [18] устанавливается взаимнооднозначное соответствие между локальными минимумами функционала  $S$  и подмножеством  $G$  всех перестановок из  $n$  элементов, определяемом равенством

$$G = \{l = (l_1, \dots, l_n) \mid \sum_{s=1}^i (r_{li+1 l_s} - r_{li l_s}) + \\ + \sum_{s=i+1}^n (r_{li l_s} - r_{li+1 l_s}) > 0, \quad i \in \overline{1, n}\},$$

а именно: для того, чтобы конфигурация  $x \in R^n$  была локальным экстремумом функционала  $S$ , необходимо и достаточно, чтобы  $l(x) \in G$ ; последнее условие эквивалентно существованию такой перестановки  $l \in G$ , что имеет место равенство  $x = z(l)$ . Таким образом, необходимое условие экстремальности функционала  $S$  в дискретном виде может быть записано как уравнение

$$g = l(z(g))$$

на множестве перестановок  $G$ .

Кроме того, в [5] показано, что задача минимизации функционала  $S$  эквивалентна задаче минимизации следующего определенного на множестве всех перестановок из  $n$  элементов функционала

$$\varphi(g) = \sum_{i=1}^n V_{gi} \Delta_{gi},$$

где

$$V_{gi} = \sum_{s=1}^n r_{gi g_s}, \quad \Delta_{gi} = \sum_{s=1}^i r_{gi g_s}.$$

Можно показать также, что задача минимизации функционала  $\varphi$  эквивалентна задаче максимизации на множестве всех перестановок функционала  $\varphi'$ :

$$\varphi'(g) = \sum_{i=1}^n [V_{gi}^2 + \Delta_{gi}^2].$$

Полученное дискретное представление непрерывной задачи метрического шкалирования позволяет применять для ее решения методы дискретной оптимизации и организовывать процесс метрического шкалирования как поиск минимума дискретного функционала в пространстве локальных экстремумов функционала шкалирования.

Приведенные в [5] примеры показывают, что множество локальных экстремумов функционала  $S$  может состоять из большого числа элементов (в частности, возможны ситуации, когда  $G$  — множество всех перестановок из  $n$  элементов).

1. Анализ нечисловой информации. М., 1981.
2. Браверман Э. М. Методы экстремальной группировки параметров и задача выделения существенных факторов. — Автоматика и телемеханика, 1970, № 1, с. 123—132.
3. Заболотный А. М. Вычислительная модель размытой группировки параметров. — В кн.: I Всесоюз. совещ. по статистическому и дискретному анализу нечисловой информации, экспертным оценкам и дискретной оптимизации. Москва; Алма-Ата, 1981, с. 422—432.
4. Заболотный А. М. Использование моделей группировки параметров в задачах социально-экономического прогнозирования. — В кн.: Системный анализ показателей социально-экономического развития на разных уровнях народного хозяйства. Минск, 1982, ч. 2, с. 161—162.
5. Заболотный А. М., Плинер В. М. Дискретные представления в задачах метрического шкалирования. М., 1983.
6. Каменский В. С. Методы и модели неметрического многомерного шкалирования. — Автоматика и телемеханика, 1977, № 8, с. 118—156.
7. Канягин Г. В., Мерсон А. Л. Комплексная оценка социально-экономической эффективности труда инженеров. — Социол. исслед., 1983, № 1.
8. Клизер С. А., Косолапов М. С., Толстова Ю. Н. Шкалирование при сборе и анализе социологической информации. М., 1978.
9. Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер-анализом. — В кн.: Классификация и кластер. М., 1980, с. 20—41.
10. Лингвистический подход к задаче обработки больших массивов информации. — Автоматика и телемеханика, 1974, № 11, с. 73—88.
11. Микина Т. Р., Перекрест В. Т. Об одном способе аппроксимации решений задачи коммивояжера. — ДАН СССР, 1975, № 1, с. 31—34.
12. Перекрест В. Т. Нелинейный типологический анализ социально-экономической информации: (Мат. и вычисл. методы). Л., 1983.
13. Перекрест В. Т. Об одной модели одномерного шкалирования. — Автоматика и телемеханика, 1980, № 2, с. 173—181.
14. Перекрест В. Т., Хмельницкая А. Б. Проекционные модели одномерного инвариантного шкалирования. — В кн.: Математические методы в социологическом исследовании. М., 1981, с. 133—146.
15. Перекрест В. Т. Параметрическая вычислительная модель многомерного шкалирования. — В кн.: I Всесоюз. совещ. по статистическому и дискретному анализу нечисловой информации, экспертным оценкам и дискретной оптимизации. Москва, Алма-Ата, 1981, с. 76—78.
16. Перекрест В. Т. Об одном подходе к задачам дискретной оптимизации. — В кн.: Проблемы проектирования и применения дискретных систем в управлении. Минск. 1977, с. 318—320.
17. Плинер В. М. Применение функциональных моделей инвариантного шкалирования в задачах социально-экономического прогнозирования. — В кн.: Системный анализ показателей социально-экономического развития на разных уровнях народного хозяйства. Минск, 1982, ч. 2, с. 152—154.
18. Плинер В. М. Об одном методе решения задачи одномерного метрического шкалирования. — Автоматика и телемеханика, 1983.
19. Поляк Б. Т. Методы минимизации функции многих переменных. — Экономика и мат. методы. 1967, т. 3, № 6, с. 881—902.
20. Терехина А. Ю. Методы многомерного шкалирования и визуализации данных. — Автоматика и телемеханика, 1973, № 7, с. 80—94.
21. Терехина А. Ю. Метрическое многомерное шкалирование. М., 1977.
22. Терехина А. Ю. Неметрическое многомерное шкалирование. М., 1977.
23. Торгерсон В. С. Многомерное шкалирование: Теория и метод. — В кн.: Статистическое измерение качественных характеристик. М., 1972, с. 95—118.
24. Филиппов А. Ф. Дифференциальные уравнения с разрывной правой частью. — Мат. сб., 1960, т. 51, № 1.

25. *Хачатурова Т. В.* Входной язык и управление пакета программ инвариантного шкалирования. — В кн.: Применение многомерного статистического анализа в экономике и оценке качества продукции. Тарту, 1981, с. 360—362.
26. *Хмельницкая А. Б.* Об оптимизационных моделях многомерного инвариантного шкалирования. — В кн.: Математическое моделирование и применение вычислительной техники в социологических исследованиях. М., 1980, с. 55—68.
27. *Хмельницкая А. Б.* О совокупности проекционных моделей инвариантного шкалирования. — В кн.: I Всесоюз. совещ. по статистическому и дискретному анализу нечисловой информации, экспертным оценкам и дискретной оптимизации. Москва — Алма-Ата, 1981, с. 87—89.
28. *Хмельницкая А. Б.* О проекционных моделях многомерного инвариантного шкалирования. — Труды ВЦ Тарт. ун-т, 1980, вып. 44, с. 93—110.
29. *Хмельницкая А. Б.* Об эквивалентности некоторых моделей инвариантного шкалирования. — В кн.: Методы анализа многомерной экономической информации. Новосибирск, 1981, с. 54—61.
30. *Хмельницкая А. Б.* О ливейном методе приближенного решения задач одномерного инвариантного шкалирования. — Автоматика и телемеханика, 1980, № 10, с. 136—145.

**Г. А. САТАРОВ**

## **МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ И ДРУГИЕ МЕТОДЫ ПРИ КОМПЛЕКСНОМ АНАЛИЗЕ ДАННЫХ**

Работа весьма актуальна для социологии. Необходимость комплексного использования различных методов, о которых идет речь, обусловливается сложностью и слабой изученностью социальных явлений. Автор приводит достаточно убедительные аргументы в защиту такой необходимости, иллюстрируя на конкретных примерах все выдвигаемые положения.

Отечественная научная литература содержит достаточно публикаций, познакомившись с которыми исследователи с самой различной математической подготовкой могут получить представление о многомерном шкалировании [2, 3, 6, 7, 10—13]. В этих работах многомерное шкалирование рассматривается как методы «визуализации данных», понижения размерности, преобразования информации, поиска латентных факторов. Ни одна из этих точек зрения в отдельности не может быть верной. Ведь даже в рамках прикладной математики, на математическом уровне строгости, многомерное шкалирование (как и прочие методы) может быть введено и характеризовано не единственным образом. Когда же метод выходит за рамки математики и адаптируется конкретной наукой, то вместе с изменением самого его статуса меняется и статус характеристик. А их многообразие определяется многообразием исследовательских задач, решению которых этот метод может способствовать.

Тем более трудно исчерпывающе описать ситуации, когда многомерное шкалирование соседствует с другими методами. Однако и здесь можно увидеть некоторые закономерности. Представляется очевидным, что все обилие способов комбинирования много-

мерного шкалирования с другими методами сводится к двум основным структурам или их сочетанию: параллельное и последовательное использование методов.

Это утверждение отражает две основополагающие идеи совместного использования различных методов. Первая из них может быть сформулирована, в частности, в следующей форме. «Если некоторая задача может быть решена с использованием двух конкурирующих методов и если нет веских оснований для выбора одного из них, то целесообразно применить оба и считать решением то, что есть общего в результатах, полученных обоими методами». Такое совместное использование методов естественно называть параллельным.

Рассмотрим в качестве примера рекомендации Крускала [5]. Он указывает на то обстоятельство, что иерархический кластерный анализ более надежно передает информацию о небольших различиях между объектами, в то же время неметрическое многомерное шкалирование более устойчиво относительно существенных различий между объектами. Это позволяет, используя оба метода для анализа одной и той же матрицы близости, уточнять результаты, полученные одним методом, с помощью результатов, полученных другим.

Другая идея заключается в цикличности процесса анализа данных о социально-экономических системах. Усложнение задач анализа данных и структур, представляющих эти данные, приводит к тому, что сколько-нибудь серьезные задачи не могут быть решены на основе использования только одного метода. Процесс анализа данных организуется следующим образом: результаты, полученные при использовании одного математического метода, служат исходными данными для использования другого. Причем этот многоступенчатый процесс может повторяться по мере корректировки гипотез и сбора дополнительной информации. В подобных случаях предлагается говорить о последовательном применении методов.

Важно подчеркнуть, что структура совместного использования различных математических методов определяется не набором методов, а главным образом задачами конкретного исследования. Выше указывалось, что многомерное шкалирование и кластерный анализ могут использоваться параллельно при анализе одних и тех же матриц близости. Однако в исследовании расстановки и соотношения сил в сенате США [8, 9] эти два метода применялись последовательно. На первом этапе кластерный анализ использовался для анализа матрицы корреляций (размера  $100 \times 100$ ) между результатами голосований сенаторов. По полученным данным были отобраны сенаторы, адекватно представляющие всю выборку, матрица корреляций результатов голосований которых подвергалась обработке с помощью неметрического многомерного шкалирования. На основе тех же данных сенаторов разбили на группы, и матрица удаленностей этих групп друг от друга так же обрабатывалась методом неметрического многомерного шкалирова-

ния. И в том и в другом случаях были получены одинаково проинтерпретированные двумерные решения, взаимно подтверждающие достоверность друг друга.

Можно выделить два наиболее распространенных способа последовательного применения математических методов, дополняющих метод многомерного шкалирования. В первом случае, как в только что приведенном примере, дополнительный метод применяется для подготовки данных для многомерного шкалирования. Во втором — дополнительный метод используется как средство интерпретации (или обоснования, подтверждения интерпретации) результатов многомерного шкалирования. Этот важный случай требует подробного обсуждения.

Наиболее эффективно метод многомерного шкалирования используется тогда, когда с его помощью решается задача выделения скрытых (латентных) факторов. В этом случае возможно получение интересных нетривиальных результатов, которые, как правило, не могут быть получены другими методами. И именно такое применение в максимальной степени соответствует моделям многомерного шкалирования. Однако при решении этой задачи особенно острым становится вопрос об интерпретации результатов, или, говоря более конкретно, вопрос об идентификации («назывании») выделенных факторов.

Анализируя расположение точек полученной конфигурации в проекции на некоторую координатную ось (если отвлечься от проблемы преобразования решения), исследователь должен понять и объяснить это расположение. Это означает, что, опираясь на свой опыт, интуицию, информированность и т. п., исследователь должен назвать такую характеристику, описывающую шкалируемые объекты, будучи измерены по которой (в предположении такой возможности), эти объекты дали бы такое же (хотя бы в пределах некоторой точности) расположение, что и расположение проекций на анализируемой координатной оси. Значит, идентификация фактора есть неформальный творческий процесс. Это, по существу, выдвижение гипотезы о природе латентного фактора, действующего в данной ситуации, а многомерное шкалирование выступает в качестве инструмента для выдвижения этой гипотезы. Вполне естественно, что следующий шаг должен состоять в проверке выдвинутой гипотезы с привлечением дополнительной информации и других математических (часто — статистических) методов.

Таким образом, необходимость проверки гипотезы о латентном факторе ставит новую задачу — задачу измерения идентифицированного фактора, которая сама может иметь самостоятельное значение. Если она решена, то сравнить ее результаты с результатами многомерного шкалирования можно известными статистическими методами. Однако в многомерном шкалировании часто трудность состоит не только в том, что нужно идентифицировать фактор, но и в том, чтобы найти направление, соответствующее предполагаемому фактору, поскольку в многомерном шкалиро-

вании решение часто находится с точностью до некоторого преобразования (часто — преобразования подобия). Существует статистический метод, который позволяет проверять гипотезу о латентном факторе тогда, когда есть результаты измерения шкалируемых объектов в шкале гипотетического фактора, и необходимо установить, входит ли этот фактор в набор латентных факторов, объясняющих результаты многомерного шкалирования в данном случае. Остановимся на этом методе несколько подробнее.

Предположим, что исследователь располагает гипотезой о латентном факторе, объясняющем многомерное решение задачи шкалирования. Причем эта гипотеза сформирована так, что шкалируемые объекты измерены по этому фактору в некоторой шкале (будем пока полагать, что это шкала уровня не ниже интервальной). В соответствии со сформулированным выше представлением о процессе идентификации фактора латентный фактор можно считать объясняющим решение, если в пространстве координатного представления существует такое направление, что значения проекций точек-объектов на это направление в максимальной степени близки значениям объектов на шкале анализируемого фактора в смысле выбранного критерия близости. Такой подход делает естественной постановку следующей задачи. Пусть в  $R^k$  задана конфигурация  $n$  точек-объектов (решение задачи многомерного шкалирования) матрицей координат  $X = \|x_{ij}\|$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, k$ . Вектор  $Q = (q_1, \dots, q_n)$  содержит значения объектов на шкале исследуемого фактора. Пусть, кроме того,  $P = (p_1, \dots, p_n)$  — вектор проекций точек  $x_i$ ,  $i = 1, \dots, n$ , на направление, задаваемое вектором  $Z = (z_1, \dots, z_k)$ . Вектор  $P$  может рассматриваться как функция вектора  $Z$ . Близость между векторами  $Q$  и  $P$  задается функцией направления  $Z$ :  $T(P(Z), Q)$ . Мы ищем такое направление  $Z$ , которое максимизирует сходство  $T$  между векторами  $Q$  и  $P$ , т. е. решаем экстремальную задачу

$$T(P(Z), Q) \xrightarrow{Z \in R^k} \max. \quad (1)$$

Пусть  $Z_0$  — решение задачи (1), а  $T_0$  — значение функции сродства  $T$  для направления  $Z_0$ . Мы говорим, что направление  $Z_0$  соответствует проверяемому фактору и подтверждает гипотезу о нем, если значение  $T_0$  достаточно велико. Для того чтобы можно было говорить, что «достаточно», а что нет, задача (1) должна быть снабжена средством статистической проверки «достаточности». Рассмотрим с этой целью один частный случай. Пусть в качестве функции  $T$  выступает квадрат коэффициента линейной корреляции между векторами  $Q$  и  $P$ :  $\text{Cor}^2(P(Z), Q)$ . Тогда решение  $Z_0$  задачи (1) определяется по формуле

$$Z_0 = B^{-1/2} W^T, \quad (2)$$

где вектор  $W$  является собственным вектором, соответствующим максимальному собственному значению матрицы  $B^{-1/2} A B^{-1/2}$ ,

а матрицы  $A$  и  $B$  имеют вид

$$\begin{aligned} A &= \| \text{Cov}(Q, x_s) \text{Cov}(Q, x_t) \|_{k \times k}, \\ B &= \| \text{Cov}(x_s, x_t) \|_{k \times k}. \end{aligned} \quad (3)$$

Можно показать, что вектор  $Z_0$  с точностью до множителя совпадает с набором коэффициентов регрессии векторов  $x_j, j = 1, \dots, k$ , на вектор  $Q$ , а квадрат коэффициента корреляции для вектора  $Z_0$  совпадает с квадратом сводного коэффициента корреляции (в задаче множественной регрессии). Если в качестве нулевой гипотезы использовать предположение о том, что компоненты вектора  $Q$  и элементы матрицы  $X$  распределены независимо, одинаково и нормально, то можно воспользоваться тем обстоятельством, что в этих предположениях наш критерий (и сводный коэффициент корреляции, точнее, его квадрат) имеет  $\beta$ -распределение с параметрами, зависящими только от числа точек  $n$  и от размерности  $k$  [4]:

$$p = \frac{k}{2}, \quad q = \frac{n-1-k}{2}. \quad (4)$$

При достаточно больших  $n$  распределение квадрата корреляции, умноженного на  $n$ , сходится к распределению  $\chi^2$  с  $k-1$  степенями свободы [4]. Таким образом, распределение величины  $\text{Cov}^2(P(Z_0), Q)$  в предположениях нулевой гипотезы известно и табулировано. Это позволяет оценивать уровень значимости выводов об исследуемом латентном факторе.

Решение задачи (1) в обсужденном выше частном случае сопряжено с поиском собственных значений. Однако матрицы, собственные значения которых необходимо определять, имеют размер, совпадающий с размерностью  $k$  пространства координатного представления. Поэтому в плоском случае, например, это равносильно решению квадратного уравнения, что позволяет решать задачу (1) простейшими вычислительными средствами. Еще более существенных вычислительных упрощений можно достичь, если в качестве критерия сходства между векторами  $Q$  и  $P$  взять не корреляцию между ними, а ковариацию —  $\text{Cov}(P(Z), Q)$ . В этом случае решение задачи (1) задается простой формулой

$$Z_j^0 = \text{Cov}(Q, x_j), \quad j = 1, \dots, k. \quad (5)$$

Как показали вычислительные эксперименты, в подавляющем большинстве случаев решения, определяемые по формулам (2) и (5), оказываются весьма близки. Пример использования решения задачи (1) для интерпретации результатов многомерного шкалирования можно найти в работах [8, 9]. Как было отмечено, в первом случае задача (1) смыкается с задачей множественной регрессии. Это не случайно. Идея использования аппарата множественной регрессии результатов шкалирования на дополнительно полученные шкалы (или наоборот) естественна и не нова. В обзорах [6, 14] указываются работы, в которых эта идея осу-

ществлена. Некоторые работы будут обсуждаться ниже. Однако, поскольку критерий  $T(P(Z), Q)$  может иметь самый различный вид, эти задачи нельзя считать эквивалентными.

Ниже будет дан небольшой обзор работ, в которых многомерное шкалирование применялось вместе с другими методами. За более подробными сведениями можно обратиться к уже упоминавшимся обзорам [6, 14].

В последнее время появилось немало работ, в которых методы многомерного шкалирования привлекаются к исследованиям окружающей среды и связанных с ней проблем. Так, в работе Уорда [16] методами многомерного шкалирования изучалось восприятие окружающей среды. В ней параллельно использовалось два различных метода многомерного шкалирования, с тем чтобы выявить влияние на конечный результат формы вынесения суждения и выбранной модели шкалирования. С этой целью все исследование было разделено на два эксперимента, в каждом из которых стимулами служили 20 цветных фотографий (среди них: «Вид Сан-Франциско с птичьего полета», «Ферма в Канзасе», «Большой Каньон», «Берег океана с дюнами» и т. п.).

Два разных эксперимента соответствовали двум различным способам определения субъективного сходства между стимулами. В первом эксперименте 15 респондентов классифицировали эти стимулы по заданным инструкциям на разное число классов, с ограничениями или без них. Для каждого испытуемого сходство между стимулами определялось как функция частоты совместного попадания стимулов в один класс. Во втором эксперименте 41 респондент оценивал по семибалльной шкале сходство стимулов (для всех 190 пар).

Данные обоих экспериментов обрабатывались процедурами метрического индивидуального шкалирования и неметрического многомерного шкалирования (по усредненным данным о близости). Индивидуальное шкалирование данных, полученных в обоих экспериментах, дало пятимерные решения с одинаковыми наборами латентных факторов (в порядке их весомости): 1) «искусственное/естественное», 2) «земля (естественное)/вода (естественное)», 3) «открытое (искусственное)/замкнутое (искусственное)», 4) «открытое (естественное)/замкнутое (естественное)», 5) «интересное/скучное». Коэффициенты корреляции между проекциями стимулов на соответственные оси для двух конфигураций от обоих экспериментов равнялись (в том же порядке осей) 0,998, 0,975, 0,962, 0,872, 0,541.

Усредненные по респондентам матрицы сходства анализировались методом неметрического многомерного шкалирования отдельно для каждого эксперимента. Получив решения в размерности от 1 до 5, автор убедился, что в обоих случаях адекватным является трехмерное решение. Однако для удобства сравнения с результатами индивидуального шкалирования анализировались пятимерные решения, преобразованные методом варимакс. Коэффициенты корреляции между соответствующими осями (в порядке

убывания весомости) оказались равными для первого эксперимента 0,986, 0,965, 0,916, 0,624, 0,450, а для второго — 0,999, 0,975, 0,947, 0,818, 0,377. Проведенное исследование показало высокую устойчивость результатов и относительно метода сбора данных, и относительно метода шкалирования по крайней мере для четырех факторов из пяти. Тем самым повышена достоверность полученных результатов.

В цикле статей [15] Оостендорпа и Берлина описано интересное исследование, в котором изучались скрытые факторы, описывающие эстетическое восприятие архитектурных объектов, изображенных на 20 цветных слайдах и охватывающих различные эпохи и стили. В этом исследовании параллельно применялись методы для анализа сходства и для анализа предпочтения стимулов, а последовательно с этим использовалась техника регрессионного анализа и метода главных компонент для интерпретации результатов шкалирования.

В первом эксперименте десяти респондентам предъявлялись 190 пар стимулов на двух экранах, время предъявления — 5 с с такими же интервалами. Респонденты оценивали субъективно воспринимаемую близость между изображениями архитектурных объектов по семибалльной шкале. Полученные матрицы близости всех респондентов обрабатывались методом метрического индивидуального шкалирования, в результате чего было принято четырехмерное решение.

Во втором эксперименте 20 респондентов оценивали те же стимулы по 30 шкалам, описывающим эстетические характеристики архитектурных объектов. Обработка этих результатов позволила получить 13 рейтинговых шкал эстетических характеристик, по которым были измерены все 20 объектов. Эти данные анализировались методом главных компонент. Три главные компоненты (84,6% суммарной дисперсии) авторы назвали: «успокоение/возбуждение», «порядок/хаос», «напряженность». Вслед за этим была проведена множественная регрессия результатов шкалирования на рейтинговые шкалы и факторное решение. Это позволило следующим образом проинтерпретировать часть осей, полученных при шкалировании: фактор 1 — «ясность/неопределенность»; фактор 4 — «необычное/общепринятое»; фактор 2 имел высокую корреляцию с первыми двумя компонентами факторного решения.

Поскольку эстетические характеристики не позволили полностью проинтерпретировать результаты шкалирования, был проведен третий эксперимент, аналогичный второму, но в котором были получены рейтинговые шкалы стилистических и технических характеристик архитектурных объектов. Так же был использован метод главных компонент, и четыре главные компоненты (78,9% суммарной дисперсии) имели следующий смысл: 1) «пышность, вычурность», 2) «монументальность», 3) «традиционность», 4) «декоративность». По совокупности результатов регрессионного анализа во втором и в третьем экспериментах была получена

следующая интерпретация факторов, описывающих восприятие архитектурных объектов: 1) «ясность», 2) «спокойствие/возбуждение», 3) «неопределенность», 4) «привычность».

В четвертом эксперименте в тех же условиях, что и в первом, респонденты должны были выбирать более предпочтительный объект из двух для всех 190 пар. По результатам многомерного неметрического шкалирования предпочтений было выделено два фактора: «спокойствие/возбуждение» и «привычность». Работа Оостендорпа и Берлина поучительна разносторонним подходом к объекту исследования, использованием богатой интерпретационной техники, повышающей достоверность выводов. Заслуживает внимания тот факт, что для описания предпочтений в данной экспериментальной ситуации понадобилось меньше факторов, нежели для описания восприятия.

В работе В. И. Викторова метрическое многомерное шкалирование было включено в исследование параллельно традиционной технике семантического дифференциала. Шкалированию и оценке с помощью СД подвергались 10 эмоциональных состояний.

Оба метода дали трехмерные решения. Коэффициенты ранговой корреляции Спирмена для проекций на соответствующие оси составили 0,92, 0,94 и 0,79. Полученные результаты позволили автору судить об устойчивости канонической факторной структуры семантического дифференциала [1].

Наивно полагать, что успех исследования может быть обеспечен произвольным нагромождением всяческих методов, находящихся в распоряжении исследователя. Выбор методов и структура их совместного использования должны находиться в согласии с природой и структурой анализируемых данных, теоретическими представлениями и гипотезами исследователя. Необходимым условием этого является планирование использования методов математического анализа данных на самых ранних этапах исследования.

### Литература

1. Викторов В. И. Интерпретация канонической факторной структуры семантического дифференциала на основе результатов многомерного шкалирования эмоций. — Рукопись, деп. в ИНИОН, № 10656, 1982.
2. Каменский В. С. Методы и модели неметрического многомерного шкалирования: Обзор. — Автоматика и телемеханика, 1977, № 8.
3. Клигер С. А., Косолапов М. С., Толстова Ю. Н. Шкалирование при сборе и анализе социологической информации. М., 1978.
4. Крамер Г. Математические методы статистики. М., 1975.
5. Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер-анализом. — В кн.: Классификация и кластер. М., 1980.
6. Сатаров Г. А. Многомерное шкалирование: новые идеи и пути использования. — В кн.: Статистические методы в общественных науках. Сб. обзоров ИНИОН. М., 1982.
7. Сатаров Г. А., Каменский В. С. Общий подход к анализу экспертных оценок методами неметрического многомерного шкалирования. — В кн.: Статистические методы анализа экспертных оценок. М., 1977.

8. Сатаров Г. А., Станкевич С. Б. Голосования в конгрессе США: Опыт многомерного анализа.— Социол. исслед., 1983, № 1.
9. Сатаров Г. А., Станкевич С. Б. Применение неметрического многомерного шкалирования при изучении расстановки и соотношения сил в конгрессе США.— В кн.: Анализ нечисловой информации в системных исследованиях. М., 1983.
10. Терехина А. Ю. Метрическое многомерное шкалирование. М., 1977.
11. Терехина А. Ю. Неметрическое многомерное шкалирование. М., 1977.
12. Терехина А. Ю. Многомерный анализ субъективных данных о сходствах или различиях. М., 1978.
13. Тюрин Ю. Н., Литвак В. Г., Орлов А. И., Сатаров Г. А., Шмерлинг Д. С. Анализ нечисловой информации. М., 1984.
14. Шрайбер Е. Л. Примеры сбора данных и интерпретации числовых результатов в процедурах многомерного шкалирования.— В кн.: Статистические методы в общественных науках: Сб. обзоров ИНИОН. М., 1982.
15. Oostendorp A., Berlyne D. E. Dimensions in the perception of architecture.— F. of Psych., 1978, vol. 19, № 1—3
16. Ward Z. M. Multidimensional scaling of the molar physical environment.— Multiv. Beh. Res., 1977, vol. 12, N 1.

Г. С. ЛЬОВ, Т. М. ПЕСТУНОВА

## ГРУППИРОВКА ОБЪЕКТОВ В ПРОСТРАНСТВЕ РАЗНОТИПНЫХ ПРИЗНАКОВ

Группировка объектов осуществляется на основе поиска закономерностей, в качестве которых выступают легко интерпретируемые логические выражения. Метод лежит в русле разработанного Г. С. Льовым [2, 3] подхода, с помощью которого может решаться широкий спектр задач, возникающих при анализе социологической информации, когда исходные данные являются значениями разнотипных признаков. Этот подход, базирующийся на использовании понятия логической закономерности, позволяет учесть также следующие часто встречающиеся в социологии особенности исходных матриц данных, затрудняющие применение для их анализа традиционных математических методов: многомерность, наличие пропусков, отсутствие модели изучаемого явления.

### 1. Содержательная и формальная постановка задачи группировки объектов

В настоящей статье под задачей группировки объектов понимается задача разбиения некоторого множества объектов на несколько непересекающихся групп по принципу «похожести» свойств на основе анализа эмпирической таблицы данных об этих объектах.

Группировка объектов проводится с целью получения более краткого описания больших массивов информации в виде групп данных, обладающих схожими свойствами. Большое количество индивидуальных описаний заменяется значительно меньшим числом описаний групп. Естественно, что при этом неизбежно некоторое огрубление, частичная потеря информации о каждом отдельном объекте, однако это дает возможность получения более наглядной общей картины, позволяет лучше уяснить структуру всего множества данных, определить дальнейшие направления исследований, что особенно важно на ранних этапах обработки информации. Именно тогда в большинстве случаев и возникает задача группировки.

При решении задачи предполагается отсутствие «суперцели», т. е. нет конкретного описания той проблемы или задачи, для которой будет использоваться результат группировки. В этих условиях ставится задача нахождения некоторой «естественной» группировки.

Формальная постановка задачи следующая. Имеется некоторое множество объектов  $A = \{a_1, \dots, a_i, \dots, a_N\}$ , выбранное из генеральной совокупности  $\Gamma: A \subseteq \Gamma$ . Каждому  $a \in \Gamma$  может быть сопоставлен в пространстве  $D = \prod_{j=1}^n D_j$  признаков  $X_1, X_2, \dots, X_n$

вектор  $x = (X_1(a), \dots, X_j(a), \dots, X_n(a))$ . Здесь для любого  $\forall j = 1, \dots, n$  символом  $D_j$  обозначено множество «мыслимых» значений, которые может принимать признак  $X_j$  на объектах из  $\Gamma$ ,  $X_j(a)$  — значение признака  $X_j$  для объекта  $a$ .

Матрица  $v = \{x_{ij}\}$ , где  $x_{ij} = X_j(a_i)$ ,  $i = \overline{1, N}$ ,  $j = \overline{1, n}$ , есть эмпирическая таблица для объектов множества  $A$ . Строки этой таблицы  $x_i = (x_{i1}, \dots, x_{ij}, \dots, x_{in})$ ,  $i = \overline{1, N}$ , назовем реализациями. Таким образом, множеству объектов  $A$  соответствует набор реализаций  $t = \{x_1, \dots, x_i, \dots, x_N\}$  в  $n$ -мерном признаковом пространстве  $D \subseteq \mathcal{D}: t \subseteq D$ ,  $D = \prod_{j=1}^n D_j$ , где  $D_j \subseteq \mathcal{D}_j$  — множество значений признака  $X_j$  на объектах множества  $A$ . Способ формирования множества  $D$  по эмпирической таблице  $v$  описан ниже.

Рассмотрим множество всевозможных разбиений множества  $A$ , обозначим его  $R_A: R_A = \bigcup_{k=1}^N R_A^k$ , где  $R_A^k$  — множество разбиений  $A$  на  $k$  групп.

Пусть  $\alpha^k \in R_A$ ,  $\alpha^k = \{A_1, \dots, A_v, \dots, A_k\}$ ;  $\bigcup_{v=1}^k A_v = A$ ,  $A_v \cap A_\mu = \emptyset$ ,  $\forall v \neq \mu$ ;  $v, \mu = \overline{1, k}$ .

Множество  $A_v$  назовем  $v$ -й группой<sup>2</sup> объектов в разбиении  $\alpha^k$ .

Каждому  $\alpha^k \in R_A$  соответствует вполне определенный набор  $\tau^k = \{\tau_1, \dots, \tau_k\}$ , представляющий собой разбиение на  $k$  подмножеств множества  $t$ :  $\bigcup_{v=1}^k \tau_v = t$ ;  $\tau_v \cap \tau_\mu = \emptyset$ ,  $\forall v \neq \mu$ ;  $v, \mu = \overline{1, k}$ ;  $\tau^k \in R_t^k \subset R_t$ , где  $R_t^k$  — множество всевозможных разбиений на  $k$  подмножеств множества  $t$ ,  $R_t = \bigcup_{k=1}^N R_t^k$ .

Разбиению  $\tau^k$  можно сопоставить (неоднозначно) набор множеств  $T_1, \dots, T_k$  в пространстве  $D$ , удовлетворяющий следующим условиям:

$$\begin{aligned} \tau_v &\subseteq T_v, \\ T_v \cap T_\mu &= \emptyset \quad \forall v \neq \mu; \quad v, \mu = \overline{1, k}, \\ T_1 \cup \dots \cup T_k &\subseteq D. \end{aligned} \quad (1)$$

Для описания множеств  $T_1, \dots, T_k$  используется некоторый класс функций  $\Phi$ . Эти множества, удовлетворяющие (1), назовем

<sup>1</sup> Здесь  $D$  — декартово произведение  $D_1, D_2, \dots, D_n$ .

<sup>2</sup> Поскольку элементы  $\alpha^k$  можно перечислять в произвольном порядке, то номер зависит от этого порядка.

таксонами. По существу,  $\Phi$  — это язык описания таксонов. Так, например, в качестве  $\Phi$  можно использовать класс линейных функций, кусочно-линейных, логических и т. д., в том числе и класс функций произвольной природы.

Обозначим  $\omega^k = \{T_1, \dots, T_k\}$  — конкретный набор из  $k$  таксонов, записанных с помощью функций класса  $\Phi$ . Пусть  $\Omega^k = \{\omega^k\}$  множество таких  $k$ -наборов,  $\Omega = \bigcup_{k=1}^N \Omega^k$ .

Задача группировки состоит в нахождении  $\omega^* \in \Omega$ , максимизирующего некоторый критерий качества  $F$ :  $F(\omega^*) = \max_{\omega \in \Omega} F(\omega)$ .

Набор  $\omega$  однозначно определяет разбивку  $\tau^* \in R_t$  на множестве  $t$ , а  $\tau^*$  соответствует одна и только одна разбивка  $\alpha^*$  множества  $A$ . Таким образом, в классе функций  $\Phi$  с помощью критерия  $F$  определяются наилучшие разбивки множества объектов  $A$  на группы.

Критерий  $F$  в большей или меньшей мере должен представлять собой формальную запись содержательного принципа «похожести» свойств объектов, попавших в один таксон.

## 2. Описание таксонов в классе логических функций

Признак  $X_j$ ,  $j = \overline{1, n}$ , может быть измерен в той или иной шкале. В данной работе рассматриваются следующие типы признаков: номинальные (измеренные в шкале наименований), балльные, порядковые (измеренные в шкале порядка<sup>3</sup>), количественные (измеренные в абсолютной шкале отношений или интервалов).

Рассмотрим множество несовпадающих значений этого признака на эмпирической таблице для множества объектов  $A: x_1, x_2, \dots, x_{N_j}$ , где  $N_j$  — число несовпадающих значений,  $N_j \leq N$ . Кроме того, для количественных и порядковых признаков будем предполагать, что такое множество несовпадающих значений упорядочено, т. е.  $x_1 < x_2 < \dots < x_{N_j}$ .

Множество  $D_j$  формируется следующим образом:  $D_j = \{x_1, x_2, \dots, x_{N_j}\}$ , если  $X_j$  замерен в шкале наименований или порядка;  $D_j = \{x \mid x_1 \leq x \leq x_{N_j}\}$ , если  $X_j$  — количественный признак.

На множестве  $D_j$  определим понятие простейшего множества  $d_j$ : для номинального признака  $X_j$  любое подмножество на  $D_j$  является простейшим, т. е.

$$d_j = \{x_{i_1}, x_{i_2}, \dots, x_{i_q}\},$$

$$1 \leq i_1 < i_2 < \dots < i_q \leq N_j, \quad 1 \leq q \leq N_j.$$

<sup>3</sup> Рассматривается только частный случай шкалы порядка: набор значений такого признака есть упорядоченный набор баллов.

Для балльного порядкового признака  $X_j$  простейшие множества включают в себя одно либо несколько несовпадающих соседних значений из  $D_j$ , т. е.

$$d_j = \{x \mid x \in D_j, x_p \leq x \leq x_q\}, \quad 1 \leq p \leq q \leq N_j.$$

Для количественного признака простейшими множествами являются интервалы, т. е.

$$d_j = \{x \mid x_p \leq x \leq x_q\}, \quad 1 \leq p < q \leq N_j.$$

Пусть  $a \in \Gamma$ ,  $d_j$  — некоторое простейшее множество на  $D_j$ . Определим элементарную высказывательную функцию (э.в.ф.) следующим образом:

$$\begin{aligned} I(a, d_j) &= 1, & \text{если } X_j(a) \in d_j, \\ I(a, d_j) &= 0, & \text{если } X_j(a) \notin d_j \text{ или } X_j(a) \text{ не измерено.} \end{aligned}$$

$I(a, d_j)$  есть двуместный предикат,  $d_j$  — это множество истинности, э.в.ф.  $I(a, d_j)$ .

Обозначим через  $x = \{x_1, \dots, x_j, \dots, x_n\}$  — множество признаков, а через  $x'$  — его подмножество. Рассмотрим декартово произведение

$$D' = \prod_{m=1}^M d_{jm}, \quad M = |X'| \leq n.$$

Конъюнкцией <sup>4</sup>  $S$  назовем выражение вида

$$S(a, D') = \bigwedge_{m=1}^M I(a, d_{jm}),$$

$M$  — длина конъюнкции.

Множество  $D'$  — множество истинности конъюнкции  $S$  — является таксоном для группы объектов  $\bar{A} \subset A$  ( $\forall \bar{a} \in \bar{A} \ S(\bar{a}, D') = 1$ ).

Считается, что значения остальных признаков, не вошедших в  $x'$ , для объектов группы  $\bar{A}$  могут быть любыми ( $d_{jm} = D_{jm}$ ). Таким образом, при необходимости можно считать все конъюнкции имеющими длину  $n$ .

С усложнением «конфигурации» таксонов можно использовать более сложные логические функции — дизъюнкции. Под дизъюнкциями будем понимать функцию следующего вида <sup>5</sup>:

$$L(a, D') = S_1(a, D'_1) \vee S_2(a, D'_2) \vee \dots \vee S_l(a, D'_l),$$

где  $D' = D'_1 \cup D'_2 \cup \dots \cup D'_l$ ;  $l$  назовем длиной дизъюнкции. Для описания таксонов будем использовать дизъюнкции, удовлетворяющие условию:

пусть  $\forall_j = \overline{1, n}$ ,  $d_j^m$  — простейшее множество по признаку  $X_j$ , входящее в  $m$ -ю конъюнкцию  $S_m$  дизъюнкции  $L$  ( $1 \leq m \leq e$ );

<sup>4</sup> Конъюнкция — логическая операция «и»:  $A \& B$  означает « $A$  и  $B$ ».  
Логическая операция «или».

тогда  $d_j = d_j^1 \cup \dots \cup d_j^l$  должно быть также простейшим множеством по этому признаку в соответствующей шкале.

В работах [1—3] показано, что класс логических функций от переменных, измеренных в разных шкалах, обладает рядом ценных свойств. Для задач группировки, возникающих, как правило, на ранних стадиях исследования, одним из наиболее важных свойств является наглядность и хорошая интерпретируемость описания групп, простая форма записи таксонов. Как показывает практика, для специалистов прикладной области при работе с таблицами большой размерности и разнотипными данными предпочтительнее получать хотя и более грубое, но при этом легко воспринимаемое человеком описание. Использование логических функций позволяет строить алгоритмы при наличии пропусков в таблицах, использовать для описания групп не все признаки, выделяя наиболее существенные для группировки.

### 3. Критерий группировки

На содержательном уровне будем придерживаться идей, которые для количественных признаков можно наглядно пояснить так: при группировке объектов обычно стремятся к тому, чтобы были максимальными плотность каждой группы в отдельности и разнесенность групп в признаковом пространстве. Эти идеи в разном виде используются в большинстве подходов и хотя не полностью, но в значительной мере отражают тот смысл, который вкладывается в понятие «естественной» группировки. Следует отметить, что предложенный в данной работе новый критерий группировки объектов в пространстве разнотипных признаков в случае количественных признаков, по существу, аналогичен известному критерию минимальной дисперсии (суммарного объема областей, описывающих группы объектов).

Перейдем к формулировке критерия группировки объектов.

Если в пространстве признаков нет «ступков» реализаций, соответствующих множеству объектов  $\Gamma$ , то эмпирическую таблицу  $u$ , соответствующую любому подмножеству объектов из  $\Gamma$ , можно рассматривать как выборку с равномерным законом распределения<sup>6</sup>.

Каждой дизъюнкции  $L$  сопоставим вероятность ее выполнения  $P_L$ . Она определяется так.

Пусть имеется конъюнкция  $S, I_j(a, d_j)$  — э.в.ф., по  $j$ -му признаку входящая в  $S$ . Если  $X_j$  — номинальный признак, имеющий  $n_0$  значений на множестве объектов  $A$ , то  $d_j$  включает в себя  $n_1$  каких-то из этих значений, а вероятность выполнения  $I_j(a, d_j)$  равна  $p_S^j = n_1/n_0$ . Если  $X_j$  — порядковый признак, принимающий на множестве  $A$ ,  $n_0$  значений, то  $d_j$  включает  $n_1$  соседних значений и  $p_S^j = n_1/n_0$ . Для количественного признака  $X_j, d_j = [d_{\min}^1,$

<sup>6</sup> Это важное содержательное предположение о смысле термина «группировка», с которым можно спорить.

$b_{\max}^1$ ]. и если через  $[b_{\min}^0, b_{\max}^0]$  обозначать минимальный интервал, включающий значения этого признака на объектах множества  $A$ , то искомая вероятность вычисляется по формуле  $p_S^j = (b_{\max}^1 - b_{\min}^1) (b_{\max}^0 - b_{\min}^0)$ . Тогда вероятность выполнения конъюнкции равна  $p_S = p_S^1 \cdot p_S^2 \cdot \dots \cdot p_S^n$ . Вероятность выполнения дизъюнкции  $L = S_{j_1} \vee S_{j_2} \vee \dots \vee S_{j_k}$  определяется по стандартной формуле «включений и исключений».

С другой стороны, имеется вполне конкретное число объектов множества  $A$ , для которых высказывание  $L$  истинно. Обозначим его  $N_L$ . Тогда  $v_L = N_L/N$  есть фактическая частота выполнения дизъюнкции  $L$  на данной эмпирической таблице. Рассмотрим величину  $f(L) = v_L - P_L$ , характеризующую степень превышения частоты выполнения  $L$  над априорной вероятностью. Будем считать, что чем больше  $f(L)$ , тем предпочтительнее рассматривать  $T_L$  (множество истинности  $L$ ) в качестве таксона, а объекты из  $A$ , на которых  $L$  истинно — в качестве группы. Будем использовать  $f(L)$  как характеристику качества таксона вне связи его с остальными таксонами. Максимизируя  $f(L)$ , будем отдавать предпочтение таксонам с «повышенной концентрацией объектов».

Пусть имеется пара таксонов  $T_1$  и  $T_2$ , описываемых дизъюнкциями  $L_1$  и  $L_2$ . Рассмотрим конъюнкцию  $S_{12}$ , удовлетворяющую условиям:

а) множество истинности  $S_{12}$  включает в себя множества истинности для  $L_1$  и  $L_2$ ;

б)  $S_{12}$  имеет наименьшее значение вероятности  $P_{S_{12}}$  среди всех конъюнкций, удовлетворяющих условию а).

При объединении точек, принадлежащих двум таксонам, в один простейший таксон конъюнкция  $S_{12}$  описывает таксон с наилучшей характеристикой качества.

Рассмотрим величину  $\Delta(L_1, L_2) = (P_{L_1} + P_{L_2})/P_{S_{12}}$ , характеризующую степень ухудшения качества описания точек двух таксонов при объединении их в один простейший. Ее будем использовать в качестве характеристики взаимного расположения двух таксонов, отдавая предпочтение паре, имеющей меньшее  $\Delta(L_1, L_2)$  при прочих равных условиях.

В качестве общего критерия качества разбиения на две группы используем

$$F(L_1, L_2) = f(L_1) + f(L_2) - \Delta(L_1, L_2). \quad (2)$$

Если производится разбишка на  $k > 2$  групп, то можно использовать довольно распространенный в литературе метод перебора по всем парам таксонов и в качестве результирующего [5] брать, например, среднее значение:

$$F(L_1, L_2, \dots, L_k) = \frac{1}{N_p} \sum F(L_i, L_j).$$

Здесь  $N_p = k(k-1)/2$  — число пар групп. Задача группировки сводится к максимизации  $F$ .

#### 4. Краткое описание алгоритма и модельного примера

Для нахождения наилучшего в смысле предложенного критерия разбиения разработан алгоритм ALOT с сокращенным перебором, реализующий локально оптимальную процедуру поиска дизъюнкций по принципу «лучший к лучшему». В данный момент программа составлена для случая, когда таксоны описываются конъюнкциями. Описание алгоритма приводится для этого частного случая.

*Выбор 1-го таксона.*

*Шаг 1.* Ищем простейшее множество  $\alpha^1$ , являющееся множеством истинности некоторой э.в.ф.  $S^1$ ;  $f(S^1)$  максимальна. При этом перебор осуществляется по всем признакам и по всем э.в.ф., которые истинны на  $N_1$  объектах исходного множества  $\delta$  ( $1 - \beta$ )  $\leq N_1 \leq \delta(1 + \beta)$ . Величина  $\delta$  зависит от желаемого числа групп  $k_0$ :  $\delta = N/k_0$ . Если это трудно указать, то  $k_0$  полагается равным минимальному числу групп, которое может предположить специалист. Число  $k_0$  является лишь ориентировочным, и в принципе в зависимости от структуры данных может быть выделено больше или меньше, чем  $k_0$  групп. Величина  $\beta$  характеризует предположительную неоднородность таксонов по числу содержащихся в них объектов, например  $\beta = 0,3$ . Если такую  $S^1$  выбрать не удастся (это возможно, когда всем э.в.ф. удовлетворяет либо слишком много, либо слишком мало объектов), то значение  $\beta$  увеличивается. В пределе перебор производится по всем сформированным простейшим множествам (и соответствующим э.в.ф.).

*Шаг n2 ( $\geq 2$ ).* Ищем конъюнкцию длины  $n1$ . После  $(n1 - 1)$ -го шага имеем конъюнкцию  $S^{n1-1}$  длины  $n1 - 1$ , истинную на множестве  $D^{n1-1}$ . Формируем множества  $D^{n1}$ , определяющие конъюнкции длины  $n1$ , путем присоединения к  $D^{n1-1}$  элементарных множеств по признакам, еще не вошедшим в  $S^{n1-1}$ . В качестве  $S^{n1}$  выбираем такую, у которой

$$f(S^{n1}) = \max, f(S^{n1}) \geq f(S^{n1-1}).$$

Если такую выбрать не удастся, то поиск 1-го таксона заканчивается:  $T_1 = D^{n1-1}$ , он описывается конъюнкцией  $S = S^{n1-1}$ .

Если нашли  $S^{n1}$  и  $n1 = n$  (т. е. все признаки задействованы), то  $T_1 = D^{n1}$ , он описывается конъюнкцией  $S = S^{n1}$ .

Определена группа объектов  $A = \{\bar{a} \mid \bar{a} \in A, S(\bar{a}, T_1) = 1\}$ .

*Выбор k1-го таксона,  $k1 \geq 2$ .*

Имеем  $T_1, T_2, \dots, T_{k1-1}$ , описанные конъюнкциями  $S_1, S_2, \dots, S_{k1-1}$ . Выбор  $T_k$  аналогичен выбору 1-го таксона с той разницей, что:

выбирается, исходя из количества объектов, еще не разгруппированных, и количества групп, которое еще можно образовать (максимальное число групп  $K_{\max}$  должно быть задано);

Эмпирическая таблица

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$
$a_1$	4,97	3	43,0	0	3	4	40,0
$a_2$	5,15	3	40,5	1	3	4	41,0
$a_3$	4,80	6	70,0	1	3	4	40,0
$a_4$	4,80	4	73,5	1	3	4	47,0
$a_5$	4,90	6	40,0	0	3	4	45,0
$a_6$	4,85	5	65,0	1	3	4	42,0
$a_7$	5,05	4	55,0	1	3	4	40,0
$a_8$	5,07	4	57,0	1	3	4	38,0
$a_9$	5,00	3	40,0	1	3	4	39,0
$a_{10}$	4,83	5	42,0	1	3	4	43,0
$a_{11}$	5,10	4	50,0	1	3	4	40,0
$a_{12}$	1,10	9	67,0	0	4	1	50,0
$a_{13}$	1,20	8	70,0	0	4	1	55,0
$a_{14}$	1,15	10	65,5	0	4	1	50,0
$a_{15}$	1,17	8	66,0	1	4	2	52,0
$a_{16}$	1,21	8	67,5	1	4	2	57,0
$a_{17}$	1,12	10	65,0	1	4	2	60,0
$a_{18}$	1,20	9	68,0	1	4	2	60,0
$a_{19}$	1,16	9	69,5	1	4	2	52,0
$a_{20}$	10,00	2	50,5	0	5	3	82,0
$a_{21}$	10,10	2	51,0	0	5	3	85,0
$a_{22}$	9,30	1	55,0	0	5	1	80,0
$a_{23}$	9,35	1	56,0	0	5	3	84,0
$a_{24}$	10,05	1	57,5	0	5	3	81,0
$a_{25}$	9,80	1	56,5	0	5	3	80,0
$a_{26}$	9,38	2	55,0	0	5	3	85,0
$a_{27}$	10,03	2	55,5	0	5	1	83,0
$a_{28}$	8,90	2	57,0	0	5	1	86,0
$a_{29}$	8,97	2	56,0	0	5	1	83,0
$a_{30}$	9,01	1	50,0	0	5	1	81,0

порядок предпочтения определяется с помощью  $F$ ;

$$T_{k1} \cap \left( \bigcup_{i=1}^{k1-1} T_i \right) = \emptyset.$$

Для проверки работоспособности программы, реализующей вышеописанный алгоритм, была взята эмпирическая таблица, содержащая информацию по семи признакам о 30 объектах. Признаки 1, 3, 7-й — количественные, 2-й — порядковый, 4, 5, 6-й — номинальные. Требовалось разбить объекты на три — пять групп.

В результате работы программы выделились три таксона.

$$\begin{aligned}
 T_1 = & \{x \mid x \in D_1, 4,8 \leq x \leq 5,15\} \cup \{x \mid x \in D_2, \\
 & x \in \{3; 4; 5; 6\}\} \cup \{x \mid x \in D_5, x = 3\} \cup \\
 & \cup \{x \mid x \in D_6, x = 4\} \cup \{x \mid x \in D_7, 37,99 \leq x \leq 47,00\}, \\
 T_2 = & \{x \mid x \in D_1, 1,10 \leq x \leq 1,21\} \cup \{x \mid x \in D_2, \\
 & x \in \{8, 9; 10\}\} \cup \\
 & \cup \{x \mid x \in D_3, 65,00 \leq x \leq 73,50\} \cup \{x \mid x \in D_5, \\
 & x = 4\} \cup \\
 & \cup \{x \mid x \in D_6, x \in \{1; 2\}\} \cup \{x \mid x \in D_7, 50,00 \leq x \leq \\
 & \leq 60,00\},
 \end{aligned}$$

$$T_3 = \{x \mid x \in D_1, 8,9 \leq x \leq 10,10\} \cup \{x \mid x \in D_2, x \in \{1; 2\}\} \cup \\ \cup \{x \mid x \in D_3, 50,00 \leq x \leq 57,50\} \cup \{x \mid x \in D_4, x = 0\} \cup \{x \mid x \in D_5, x = 5\} \cup \{x \mid x \in D_6, x \in \{1; 3\}\} \cup \{x \mid x \in D_7, 80 \leq x \leq 86\}.$$

Соответственно выделились три группы объектов. Принадлежность объектов к вышеуказанным таксонам определяется истинностью конъюнкций  $S_1, S_2, S_3$ , построенных в соответствии с разд. 2 по множествам истинности  $T_1, T_2, T_3$ :

$$A_1 = \{a \mid a \in A, S_1(a, T_1) = 1\},$$

$$A_2 = \{a \mid a \in A, S_2(a, T_2) = 1\},$$

$$A_3 = \{a \mid a \in A, S_3(a, T_3) = 1\}.$$

Для удобства чтения объекты в прилагаемой таблице перенумерованы в соответствии с их принадлежностью к группам  $A_1, A_2, A_3$ .

Программа ALOT применялась для решения задачи из области медицинской диагностики.

#### Литература

1. Воронин Ю. А. Введение мер сходства и связи для решения геолого-географических задач. — ДАН СССР, т. 199, № 5, 1971, с. 1011—1014.
2. Лбов Г. С. Логические функции в задачах эмпирического предсказания. — В кн.: Эмпирическое предсказание и распознавание образов. Новосибирск, 1978, с. 34—64, (Вычисл. системы: Вып. 76).
3. Лбов Г. С. Методы обработки разнотипных экспериментальных данных. Новосибирск, 1981.

## ОПЫТ ПРИМЕНЕНИЯ ДЕТЕРМИНАЦИОННОГО АНАЛИЗА В ИССЛЕДОВАНИИ ВЛИЯНИЯ СОЦИАЛЬНО-ДЕМОГРАФИЧЕСКИХ ФАКТОРОВ

В статье показывается, что при анализе конкретного социологического материала полезно выделять комбинации значений рассматриваемых признаков, детерминирующих значения некоторых исследуемых характеристик. В качестве средства такого выделения предлагается так называемый детерминационный анализ. Решаемые с его помощью задачи являются, по существу, задачами «локальных» связей, т. е. связей между отдельными градациями признаков, а не совокупностями всех градаций, соответствующих рассматриваемым признакам. Их же можно рассматривать как задачи группировки объектов. Каждая искомая группа состоит из объектов, обладающих найденным сочетанием значений признаков. Отметим, что тот же класс задач решается с помощью методов поиска логических закономерностей, предложенных Г. С. Лбовым (эти методы используются для решения задачи классификации в статье Г. С. Лбова и Т. М. Пестуновской).

Самый распространенный способ анализа информации — поиск зависимостей на основе таблиц сопряженности между выходными признаками, описывающими изучаемое явление (проведение свободного времени, производительность труда, ценностные ориентации), и входными признаками, выбранными для его объяснения (например, социально-демографическими признаками).

Другой способ — применение методов многомерного анализа: факторного, регрессионного, кластерного и других.

У каждого из подходов свои достоинства и недостатки. В литературе они не раз рассматривались. Поэтому не будем на них подробно останавливаться. Отметим еще раз наиболее существенные.

Группировка — наиболее простой и естественный способ анализа информации. На исходные данные не накладывается никаких ограничений. Оценка взаимосвязей между признаками на основе подсчета частот различных комбинаций их значений непосредственно отражает логику эмпирического социологического анализа.

Однако при попытке углубиться в материал, привлечь большее число признаков для «многомерной» оценки явления исследователь сталкивается с необходимостью просмотра огромного числа таблиц. Это часто делает нереальным привлечение более двух-трех признаков для одновременной оценки изучаемого показателя.

Методы многомерного анализа, казалось бы, преодолевают это ограничение, позволяя одновременно использовать существенно большее число признаков, но применение этих методов часто связано с существенными предположениями, ограничивающими возможности их использования, делающими весьма проблематичной интерпретацию результатов.

Весьма перспективны так называемые специальные методы группировки, к числу которых относится и описываемый ниже

метод детерминационного анализа [4, 2]. Вводятся определенные формальные критерии содержательной значимости групп, на основе которых их поиск автоматизируется.

Эти методы позволяют получать комбинации наиболее существенных характеристик объектов, детерминирующих изучаемые показатели. Возможность передать ЭВМ огромную вычислительную работу по поиску содержательно значимых групп, преодолеваемую сейчас обычно исследователем, позволяет глубже проанализировать материал по сравнению с использованием обычных группировок.

Один из наиболее эффективных методов этой группы — детерминационный анализ. Социологи с данным методом практически не знакомы. Представляет интерес накопление и рассмотрение опыта применения данного метода в анализе социологической информации. Процедура анализа рассматривается на примере одного из конкретных социологических исследований. Используются два основных критерия поиска существенных комбинаций характеристик — интенсивность  $I$  и емкость  $C$ . Пусть  $x_\alpha$  — комбинация значений признаков объектов, описывающих искомую группу (например, мужчины, семейные, в возрасте 25—30 лет, живущие в городе);  $y_s$  — исследуемая характеристика, объясняемая через набор остальных признаков (например, время, уделяемое определенному занятию на досуге), тогда

$$I_{x_\alpha y_s} = \frac{N_{x_\alpha y_s}}{N_{x_\alpha}}, \quad C_{x_\alpha y_s} = \frac{N_{x_\alpha y_s}}{N_{y_s}},$$

где  $N_{x_\alpha}$ ,  $N_{y_s}$ ,  $N_{x_\alpha y_s}$  — соответственно численности групп респондентов с характеристиками  $x_\alpha$ ,  $y_s$  и одновременно с характеристиками  $x_\alpha$  и  $y_s$ ;  $I_{x_\alpha y_s}$  характеризует степень, в которой набор значений признаков  $x_\alpha$  детерминирует изучаемый показатель  $y_s$ <sup>1</sup>;  $C_{x_\alpha y_s}$  — измеряет «ценность» набора характеристик  $x_\alpha$ . Детерминация может быть очень высокой, и в то же время набор характеристик  $x_\alpha$  так редко встречается среди респондентов типа  $y_s$ , что становится малоценным для предсказаний.

Поиск детерминирующих характеристик проводится на основе задания:

- 1) набора аргументов — независимых, объясняющих признаков  $x_1, x_2, \dots, x_n$ ;
- 2) набора значений функций — зависимых, объясняемых характеристик  $y_s, s = 1, 2, \dots, m$ ;
- 3) порогов для критериев  $I$  и  $C$  — определяющих отбираемые комбинации значений признаков.

<sup>1</sup> Интенсивность  $I_{x_\alpha y_s}$ , представляющая долю, составляемую респондентами типа  $x_\alpha$  среди респондентов типа  $y_s$ , удобно приводить в процентах и рассматривать в качестве коэффициента детерминации для комбинации характеристик  $x_\alpha$ .

Метод детерминационного анализа применялся нами в исследовании влияния некоторых ключевых социально-демографических показателей на субъективные показатели удовлетворенности респондентов различными аспектами жизненной ситуации<sup>2</sup>.

Эмпирические референты удовлетворенности — ответы респондента на вопросы анкеты «Удовлетворены ли Вы своей работой?»; «Удовлетворены ли Вы своим материальным положением?»; «Удовлетворены ли Вы своей жизнью в целом?».

Социально-демографические показатели: семейное положение, пол, возраст, заработная плата, профессия, служебное положение, образование, жилищные условия.

Дополнительно был привлечен такой показатель, как «Удовлетворенность респондента своим здоровьем». Показатель этот весьма существенно детерминировал удовлетворенность респондента жизненной ситуацией.

Углубленное исследование материалов социологического исследования по относительно небольшому набору ключевых показателей, в частности социально-демографических, имеет как самостоятельный интерес, так и более широкое значение.

При определяющем влиянии этих показателей (таких, например, как пол, возраст, образование, профессия) модели взаимосвязи между признаками могут существенно различаться для разных совокупностей, выделенных по данным признакам. Поэтому целесообразно предварительно разделить совокупность на однородные группы по этим показателям. Эти соображения учитываются сейчас, в частности, в экономических исследованиях при проведении регрессионного анализа [13].

В настоящей работе опробовались некоторые способы выделения детерминирующих комбинаций характеристик. Ставилась задача выявления комбинаций социально-демографических факторов и соответствующих однородных групп, обуславливающих удовлетворенность респондента отмеченными выше аспектами жизненной ситуации.

Выяснялась значимость отдельных факторов и степень уточнения ими коэффициентов детерминации для комбинаций, выявленных по другим факторам. Использовались два основных вида обработки данных (так называемые запросы 1 и 2).

#### *Запрос 1.*

Поиск комбинаций осуществляется в пространстве векторов значений сразу всех независимых признаков. Результат анализа — наборы значений независимых признаков  $x_1, x_2, \dots, x_n$  длины  $n$ , удовлетворяющие заданным порогам по показателям емкости и интенсивности.

#### *Запрос 2.*

Выявляются так называемые базовые детерминирующие комбинации — такие комбинации, которые, во-первых, удовлетворя-

<sup>2</sup> Объект исследования — работники промышленных предприятий крупного индустриального города. Приводимые ниже результаты рассматриваются в методическом плане, иллюстрирующем особенности анализа.

ют заданным порогам по интенсивности  $I$  и емкости  $C$ , и, во-вторых, при отбрасывании любого из значений признаков, входящих в комбинацию, оставшиеся признаки уже не составляют комбинацию, удовлетворяющую заданным порогам по  $I$  и  $C$ . Смысл использования запроса 2 — поиск наиболее простых сочетаний, значений признаков, в достаточной (в соответствии с заданным порогом) степени детерминирующих исследуемое явление. Запрос 2 — более экономный с точки зрения времени работы ЭВМ — может быть использован вначале для поиска «перспективных» наборов признаков, выработки гипотез о степени влияния отдельных признаков и уточнения этого влияния через другие признаки.

Запрос 1 может применяться уже для более ограниченного числа признаков, комбинации значений которых с достаточно высокой интенсивностью детерминировали изучаемые показатели на предыдущем этапе анализа.

Методика использования детерминационного анализа может состоять в виде последовательности следующих шагов.

1. Определение порогов по интенсивности и емкости.
2. Предварительный статистический анализ материала.
3. Выбор наиболее существенных, базовых признаков и решение вопроса о преобразовании признаков (построение комплексных признаков-индексов на основе первичных признаков, объединение соседних градаций).
4. Последовательное уточнение влияния базовых признаков посредством добавления новых признаков.

При задании порога для емкости  $C$  учитывается общий объем совокупности респондентов. Емкость  $C$ , как отмечалось, определяет, какую минимальную долю от объектов данного типа по заданному признаку должны составлять рассматриваемые группы. Ясно, что чем более многочисленной будет группа с высоким коэффициентом детерминации, тем более она для нас интересна. Выявленная зависимость характеризуется большей статистической надежностью, большей ценностью для предсказания интересующего исследователя явления.

В то же время, чем более тонкие «многомерные» зависимости мы рассматриваем, тем меньше будет численность выделяемых групп. Поэтому имеет смысл задавать порог по емкости относительно небольшой, чтобы не упустить интересные комбинации с большим числом признаков. Как и в применении других методов, хороший способ оценки надежности результатов — проведение анализа по одной методике для ряда однородных совокупностей респондентов. Эти совокупности можно получить, например, посредством разбиения случайным образом исходной совокупности на две.

При задании порога для интенсивности, а также при оценке полученных коэффициентов детерминации для разных групп надо учитывать что, во-первых, о детерминирующем влиянии данной комбинации характеристик мы можем говорить только при  $I \geq \geq 50\%$ . Во-вторых, обычно интересен не просто высокий коэффи-

цент детерминации, превышающей 50% (он может быть высок уже для совокупности в целом, как в рассмотренном случае, в частности, из-за известной шаблонности ответов на определенные вопросы анкеты), а его повышение при добавлении тех или иных характеристик.

В том случае, когда цель анализа — получение комбинаций характеристик, позволяющих с высокой надежностью предсказывать явление, имеет смысл задавать высокий порог (например, 0,8; 0,9). Когда имеется в виду более общий анализ, сравнение степени влияния различных характеристик, можно задавать порог по интенсивности 0,5.

В нашем случае были использованы пороги по критериям интенсивности и емкости, равные соответственно 0,5 и 0,01.

На этапе предварительного статистического анализа материала целесообразно получить одномерное распределение по признакам, таблицы сопряженности для пар признаков. Проводится оценка основных параметров статистического распределения. Таблицы сопряженности позволяют дать предварительную оценку степени влияния выбранных факторов на зависимый признак, отобрать наиболее существенные из них.

Часто вопросы анкеты представляют набор излишне дробных градаций. Преследуется цель собрать максимально детальную информацию. В то же время соседние градации могут существенно не различаться по содержанию. В этом случае их целесообразно склеить (например, оценку удовлетворенности респондента работой по 5—7-балльной шкале перевести в оценку по 3-балльной шкале). В то же время при склейке не должна теряться дифференцирующая сила градаций. Например, близкие по заработку группы респондентов могут представлять содержательно различные категории работников.

Рассмотрим некоторые результаты предварительной группировки данных.

### Удовлетворенность работой

В целом среди опрошенных в анализируемом нами массиве данных удовлетворенные работой составляют 78%. Наиболее заметно определяют удовлетворенность работой возраст и заработная плата.

Для зарабатывающих менее 100 руб. удовлетворенность работой существенно ниже средней для совокупности. Для респондентов с заработком от 100 до 250 руб. колеблется рядом со средним значением 77% (100—150 руб.), 82% (150—200 руб.), 81% (200—250 руб.). Для зарабатывающих 250—300 руб. удовлетворенные работой составляют уже 95%.

Заметно повышается удовлетворенность работой с возрастом. Среди респондентов до 20 лет, удовлетворенных работой, — 57%; от 21 до 35 лет — 69—72%; в возрастных границах от 36 до 40, 41—55 и 56—60 лет — соответственно 83, 88 и 85%. Для респондентов старше 60 лет доля удовлетворенных составляет уже 93%.

## Удовлетворенность материальным положением

Общее число удовлетворенных материальным положением среди всей совокупности опрошенных существенно ниже, чем удовлетворенных работой. Так же, как и в случае удовлетворенности работой, наиболее существенное дифференцирующее значение имеют признаки возраст и заработная плата. Хотя коэффициенты детерминации здесь заметно ниже, чем для удовлетворенности работой, они в то же время существенно выше, чем для других рассматриваемых признаков.

Среди зарабатывающих до 200 руб. удовлетворенные составляют менее 50%. В то время как среди зарабатывающих от 200 до 300 руб. и свыше соответствующие цифры 66% и 61%.

Интересно, что, как и в случае удовлетворенности работой, процент удовлетворенных материальным положением несколько падает для зарабатывающих свыше 300 руб. (хотя и остается заметно выше общего среднего значения для совокупности). Это представляет собой количественное подтверждение отсутствия жесткой связи между заработком (несмотря на высокую значимость этого фактора) и удовлетворенностью своим материальным положением и необходимостью привлечения дополнительных факторов.

Для респондентов моложе 40 лет число удовлетворенных материальным положением составляет 50% или ниже. Для респондентов в возрасте от 40 до 55 лет доля удовлетворенных материальным положением поднимается до 63%. Среди респондентов старше 55 лет доля удовлетворенных материальным положением составляет уже 74%. Данный результат может объясняться относительно большим заработком этой группы респондентов, меньшими материальными потребностями людей этого поколения.

## Удовлетворенность жизнью в целом<sup>3</sup>

Общее число удовлетворенных жизнью в целом — 67% (занимает среднее положение между числом удовлетворенных работой — 78% и удовлетворенных материальным положением — 50%).

Как и для рассмотренных ранее аспектов, относительные численности удовлетворенных существенно дифференцируются по градациям заработка и возраста. Наибольшее число удовлетворенных — 80% — в группе с заработком от 200 до 300 руб. Существенно меньшее их число в группах с заработком до 200 руб. и свыше 300 руб. (соответственно 65 и 61%). Удовлетворенность растет с возрастом (от 62—65% для респондентов моложе 40 лет, до 77% для респондентов старше 55 лет).

По сравнению с рассмотренными ранее видами удовлетворенности более заметная дифференциация наблюдается по другим

<sup>3</sup> Здесь и далее имеются в виду респонденты, ответившие утвердительно на соответствующий вопрос анкеты.

независимым признакам. Отметим, что высокие значения коэффициентов детерминации свидетельствуют очень часто не о причинно-следственных связях, а о так называемых связях сопутствия. Так, наибольшее число удовлетворенных жизнью в группах респондентов женатых — 69%. Однако, процент удовлетворенных жизнью среди вдовцов также высок и составляет 64%.

Можно предположить, что последний результат характеризует так называемую связь сопутствия. Наличие достаточно высокой связи между выходным признаком и фактором свидетельствует не о причинном влиянии последнего, а о некоторых других (скрытых) факторах, обуславливающих эту связь. В данном случае большая в среднем удовлетворенность жизнью среди вдовцов может быть связана с их большим возрастом (что подтверждается и результатами последующего анализа) и меньшими притязаниями.

В дальнейшем подобные случаи мы не будем оговаривать. Как и при использовании любого статистического метода анализа выводы о влиянии факторов основываются как на количественных показателях, так и на содержательном рассмотрении взаимосвязи между факторами.

Численность удовлетворенных жизнью в целом среди рабочих — 68%, среди инженерно-технических работников — 56%. Этому до известной степени соответствуют и результаты для различных групп по образованию. Для имеющих среднее образование и ниже относительная численность удовлетворенных составляет 69%, а для имеющих высшее или незаконченное высшее образование — 59%. Для респондентов с более высоким образовательным уровнем это объясняется тем, что удовлетворенность жизнью у них связывается с реализацией более широкого круга потребностей, со служебным продвижением, с более трудно достижимыми целями.

При проведении запроса 2 для таких аспектов удовлетворенности, как «удовлетворенность жизнью в целом» и «удовлетворенность работой» базовые комбинации (для порога интенсивности  $I = 0,50$ ) состояли, как правило, из одного значения признака. Отобранные социально-демографические показатели уже по отдельности достаточно сильно детерминировали данные аспекты удовлетворенности.

Заметно ниже, как отмечалось, были показатели детерминации для «удовлетворенности материальным положением». Базовые комбинации здесь состояли как из одного, так и из ряда значений признаков. А при пороге по интенсивности  $I = 0,60$ , как правило, — из ряда значений признаков. Значит, необходимо было привлекать несколько признаков для отыскания среди них комбинаций значений, для которых доля респондентов, удовлетворенных своим материальным положением, превышала бы 60%.

На основе предварительного анализа среди признаков, наиболее сильно детерминирующих различные аспекты удовлетворенности жизненной ситуацией, были выделены возраст и зара-

ботная плата. Посмотрим на примере характеристики «удовлетворенность жизнью в целом», в какой степени можно повысить детерминацию на основе комбинации этих признаков, а также на основе уточнения их другими признаками. В качестве уточняющих показателей были выбраны: семейное положение, жилищные условия, пол, должность, удовлетворенность здоровьем.

Рассмотрим характерные комбинации, вносящие заметный вклад в первоначальную «базовую детерминацию» удовлетворенности жизнью в целом.

Так как всего удовлетворенных жизнью в совокупности опрошенных 67%, будем рассматривать лишь комбинации характеристик, коэффициент детерминации для которых превышает 67%, т. е. позволяющие уточнить детерминацию удовлетворенности. Особый интерес представляют коэффициенты детерминации, близкие к 100%, позволяющие говорить об определяющем влиянии найденной комбинации факторов.

Так, существенно уточняется коэффициент детерминации для градации по возрасту 41—50 лет. Среди респондентов этого возраста «удовлетворенные жизнью в целом» составляли 71%. Выделение подгруппы с повышенным заработком (200—250 руб.) позволило повысить коэффициент детерминации с 71 до 84%, а выделение подгруппы с высоким заработком (250—300 руб.) довело этот показатель до максимальной величины — 100%. Иначе говоря, все респонденты в возрасте 41—50 лет с высоким заработком были в целом удовлетворены жизнью<sup>4</sup>.

Аналогичные результаты получены по отношению к респондентам в возрасте 51—60 лет, среди которых число «удовлетворенных» составляло 76%. Среди опрошенных этого возраста с повышенным и высоким заработком «удовлетворенные» составляли уже соответственно 80 и 100%.

Очень часто в применении детерминационного анализа добавление определенной характеристики не повышает коэффициент детерминации не потому, что она не значима, а из-за того, что при этом незначительно меняется состав соответствующей группы. Например, комбинации характеристик, выявленные на предыдущем шаге, не уточняются существенно посредством дополнительной характеристики «пол». Это связано с тем, что в основном респондентами в возрасте свыше 40 лет с повышенной и высокой заработной платой были мужчины.

А теперь обратимся к другой комбинации, когда респондентами являются женщины в возрасте от 51 до 60 лет со средним заработком 150—200 руб. Характеристика «средний заработок 150—200 руб.» никак не уточняла нам исходную характеристику «возраст 51—60 лет». В обоих случаях у нас был один и тот же процент удовлетворенных — 76%. Добавление же характеристики «жен-

<sup>4</sup> В этом случае, как и в последующем изложении, мы не констатируем непосредственную причинную взаимосвязь между взятыми факторами и удовлетворенностью, а рассматриваем результаты лишь как определенную эмпирическую информацию по данному вопросу.

щины» позволило повысить коэффициент детерминации сразу на 10 %.

Отметим, что в отличие от мужчин для подавляющей части респондентов-женщин заработок не превышал 200 руб. Можно предположить поэтому, что рассматриваемая группа женщин в большей степени удовлетворена своей заработной платой и работой, чем соответствующая группа мужчин.

В применении детерминационного анализа, да и статистического анализа в целом характерной является ситуация, когда признак из общих содержательных представлений должен был бы заметно влиять на изучаемый показатель в определенном направлении, но такое влияние не имеет места. При более углубленном анализе часто выясняется существенно неоднозначный характер влияния этого фактора для разных групп респондентов. Детерминационный анализ — эффективный метод такого рода дифференцированного анализа.

В рассматриваемом исследовании существенно различалось для разных социально-демографических групп респондентов влияние таких признаков, как должность и удовлетворенность респондента своим здоровьем. Так, для возрастной группы респондентов-мужчин 21—30 лет с повышенным заработком 200—250 руб. работа в руководящей должности (бригадира, руководителя рабочей группы) заметно повышает удовлетворенность жизнью в целом. Например, среди рядовых работников доля удовлетворенных составляет 64 %, а среди руководящих — поднимается до 89 %. Для респондентов-женщин этого же возраста с заработком 100—150 руб. доля удовлетворенных в процентном отношении будет выглядеть соответственно как 53 и 78 %.

В то же время в группах респондентов более старшего возраста мы не наблюдаем влияния руководящей позиции на повышение удовлетворенности. Так, выделение среди респондентов-мужчин в возрасте 41—50 лет со средней заработной платой 150—200 руб. подгруппы респондентов, не имеющих подчиненных, повышает коэффициент детерминации с 77 до 81 %.

Добавление признака «семейное положение» не меняет существенно общей картины влияния признака «должность» в комбинации с другими факторами на удовлетворенность. Отметим лишь группы с особо высоким коэффициентом детерминации — это замужние женщины в возрасте от 51 до 60 лет со средней заработной платой 150—200 руб., не имеющие подчиненных. Введение характеристик «семейное положение» и «должность» позволили здесь повысить коэффициент детерминации с 86 до 92 %. Добавление характеристики «наличие отдельной квартиры» позволяет повысить коэффициент детерминации до 100 %. Все женщины данной группы выразили удовлетворенность жизнью в целом в ответах на соответствующий вопрос анкеты. Это идеальный результат в применении детерминационного анализа, когда посредством последовательного добавления существенных факторов мы добиваемся 100 % детерминации изучаемого явления.

В предварительном анализе было выявлено весьма существенное значение фактора «удовлетворенность здоровьем» на показатели удовлетворенности респондента различными сторонами жизненной ситуации. Как правило, добавление этой характеристики заметно повышало коэффициент детерминации. Интересно, однако, что в ряде случаев добавление данного признака не повышало удовлетворенности жизнью. Характерно это было для групп респондентов младшего возраста. Можно предположить, что такое свойство, как «здоровье», не осознается ими как что-то ценное и имеющее существенное отношение к удовлетворенности жизнью.

Опыт применения детерминационного анализа показывает, насколько осторожно следует делать выводы на основе общих интегральных показателей взаимосвязи между признаками (коэффициенты корреляции, коэффициенты уравнений регрессии), а также таблиц сопряженности для двух признаков (на которых в основном и базируются выводы в эмпирических социологических исследованиях). При общей положительной взаимосвязи между  $x$  и  $y$  в углубленном «локальном» анализе может оказаться так, что с ростом  $x$  значения  $y$  растут лишь до некоторого уровня по  $x$ , а для наиболее высоких значений  $x$  «странным» образом падают. Влияние одного и того же признака, как отмечалось, может быть совершенно различным по степени и даже по направлению для разных социально-демографических групп. Иными словами, имеет место так называемый эффект взаимодействия между признаками. Можно предположить, что такого рода взаимодействия между признаками, когда степень и направление взаимосвязи между признаками  $x$  и  $y$  определяются значениями других признаков, — типичная ситуация для социологической информации.

Значения частот в таблицах сопряженности могут указывать на отсутствие детерминации изучаемого свойства  $y_s$ , выделенного по признаку  $y$ , со стороны характеристики  $x_\alpha$ , выделенной по признаку  $x$ , и, более того, коэффициент взаимосвязи между признаками  $x$  и  $y$  не высок, однако в комбинации с другими характеристиками  $x_\alpha$  может обеспечивать очень высокую детерминацию характеристики  $y_s$ .

Таким образом, детерминационный анализ позволяет провести более детальный анализ, повысить надежность выводов эмпирического социологического исследования. Такой анализ целесообразно применять для углубленного исследования влияния отдельных блоков факторов на изучаемые показатели. Выходными, зависимыми признаками могут быть, например, ценностные ориентации, установки по определенному вопросу, показатели удовлетворенности теми или иными аспектами жизненной ситуации, а входными, независимыми признаками — некоторый их набор, предположительно оказывающий наибольшее влияние на зависимые признаки. Очень часто существенную дифференциацию по выходным признакам обуславливают социально-демографические показатели.

Так же, как во всех областях эмпирического социологического исследования применяются методы группировки, во всех эмпирических исследованиях может быть с большой пользой использован и метод детерминационного анализа. Однако возможности его (как и любого метода анализа данных) ограничены. Исследователь должен задавать конкретные наборы признаков, среди которых осуществляется поиск наборов детерминирующих характеристик. Ясно, что невозможно перебрать все наборы признаков в анкетах, содержащих несколько сот вопросов.

Можно определенным образом формализовать процедуру отбора с использованием адаптивных методик отбора значимых признаков, подобных тем, которые используются в методах распознавания образов. Однако это заведомо ведет к исключению многих значимых комбинаций характеристик.

Целесообразно провести тщательный содержательный анализ материала и сформулировать гипотезы, ограничивающие варианты перебора, выбрать относительно небольшой набор ключевых признаков, в рамках которых и планировать анализ.

Традиционные методы многомерного анализа — дисперсионный, регрессионный, факторный, как и упомянутые методы распознавания образов, позволяют сжимать признаковое пространство при достаточно большом числе исходных признаков. Не отрицая полезности этих методов, надо отметить, что часто это лишь иллюзия отбора (или построения) существенных признаков. Формальные критерии отбора, заложенные в эти методы, часто в весьма общем виде и неполно отражают содержательную значимость признака. Нередко заведомо неполным является перебор признаков, либо модели, на основе которых проводится сжатие признакового пространства, не подходят для имеющихся данных.

К тому же исходная единица информации — признак, представляющий определенный набор характеристик. Исследователя же обычно интересует влияние отдельных характеристик и их комбинаций. Значимость признака в целом — лишь общий косвенный индикатор значимости его градаций (тем более в комбинации с градациями других признаков).

Метод детерминационного анализа работает в определенном диапазоне поиска. Проводится не анализ анкеты в целом, в углубленное исследование в рамках отдельных гипотез. При этом не накладывается каких-либо искусственных моделей на исходные данные. В результате «проблема интерпретации» не встает, а полученная в анализе информация имеет ясное содержательное значение.

### Литература

1. Методы классификации в социологическом исследовании. М., 1982.
2. Чесноков С. В. Детерминационный анализ социально-экономических данных. М., 1982.
3. Розин Б. Б. Распознавание образов в экономических исследованиях. М., 1976.

## О МАТЕМАТИЧЕСКОМ АППАРАТЕ МЕТОДА ГРУППИРОВОК В СОВРЕМЕННЫХ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЯХ

Автор исходит из предположения, что практически все встающие перед социологом задачи могут быть сформулированы как задачи нахождения некоторых группировок, и предлагает подход к их решению. Он основан на матричном представлении исходных данных. Используются две системы матриц: булевские матрицы — для описания данных, полученных по шкалам низких типов, и проекционные — для описания смешанных данных, среди которых могут встретиться и такие, которые получены как по номинальной, так и по интервальной шкале. Все рассматриваемые задачи формулируются как аппроксимационные задачи в пространстве рассматриваемых матриц.

Резюмируя, можно сказать, что статья является описанием подхода, позволяющего решать широкий спектр социологических задач в тех случаях, когда исходные данные получены по произвольным шкалам.

Метод группировок как метод качественного анализа данных является важнейшим инструментом в решении таких основных задач формирования теоретических представлений о конкретных социально-экономических системах, как конструирование новых категорий и установление новых связей между отдельными категориями, поскольку эти представления в силу слабой изученности и сложности систем имеют в значительной степени качественный характер.

Решение указанных задач должно осуществляться в режиме итеративного взаимодействия исследователя и ЭВМ, в котором исследователь играет основную роль. Именно он формирует и проверяет гипотезы, давая ЭВМ те или иные задания по обработке данных, интерпретирует результат и дает на этой основе новые задания для ЭВМ. Математический аппарат при этом играет вспомогательную роль, как средство обеспечения возможности концептуально-исследовательского анализа и в этом качестве должен удовлетворять требованиям, которые при обычном математико-статистическом подходе зачастую остаются в тени.

Во-первых, математический аппарат должен быть универсальным, обеспечивая возможность решения любой задачи рассматриваемого класса, которая может возникнуть в конкретном исследовании. Во-вторых, аппарат призван давать достаточно четкие правила интерпретации получаемых решений. В-третьих, алгоритмы должны допускать автоматическую настройку параметров, если исследователь не может выбрать подходящих значений.

Что касается программного обеспечения, реализующего математический аппарат, оно должно давать средства диалога «исследователь — ЭВМ», максимально приближенные к естественному языку формулирования возникающих проблем конструирования и анализа группировок. Значительная часть программного обеспечения при этом, как показывает опыт, дает исследователю воз-

возможности проверки тех или иных возникающих у него гипотез, что требует развитых средств манипулирования данными.

В данной работе подытоживаются результаты исследований [1, 2] таких слабонзученных систем, как образ жизни населения, мобильность трудовых ресурсов и т. п. Для такого рода исследований характерен примат задач конструирования новых категорий, типологий.

Содержательные задачи, решаемые с помощью группировок, по нашему мнению, могут быть сведены к трем основным типам:

1. Конструирование типологической группировки объектов в терминах тех или иных совокупностей признаков.

2. Исследование связи между различными системами признаков данного множества объектов путем формирования группировок.

3. Исследование структуры связей между объектами путем конструирования группировок.

Способы решения этих задач существенно зависят от имеющихся у исследователя априорных представлений об исследуемом объекте. Мы выделяем две крайние ситуации:

а) исследователь может сформировать необходимые группировки на основе теоретических представлений самостоятельно;

б) исследователь может указать только совокупности признаков (или информацию другого вида), которые должны быть использованы при построении группировок.

В ситуации (а) основные трудности связаны с проблемой перехода от теоретических концепций к эмпирическим группировкам и должны решаться на основе общелогической проработки типичных процедур такого перехода путем формирования проблемно-ориентированного программного обеспечения, позволяющего достаточно простым и наглядным способом формировать и анализировать исследовательские группировки. Некоторые продвижения в этом направлении реализованы М. Л. Суховским в системе преобразования и анализа данных (ПАД) [1]. Основные методические трудности возникают в ситуации (б), когда необходимо разработать универсальный аппарат, моделирующий решение задач 1—3 при минимальной априорной информации. Рассмотрим основные положения.

Прежде всего отметим, что универсальный аппарат решения этих задач должен быть основан на таком математическом представлении группировок, которое бы позволяло фиксировать не только группы объектов, но и связи между группами, и на таком представлении исходных данных, которое бы включало и группировки, задаваемые качественными признаками, и количественные признаки, и информацию, выраженную в терминах связей между объектами. В качестве такого математического представления мы используем понятие  $N \times N$  матрицы связей между объектами ( $N$  — количество объектов) для выборки и двухместных функций, характеризующих связи между парами значений признака —

для генеральной совокупности. Далее рассматривается только тематика, относящаяся к анализу выборок.

Нам не удалось пока выработать единое матричное представление для разных типов данных, так что используются две различные, хотя и связанные, системы матриц:

А) Булевские  $(\alpha, \beta)$  — матрицы.

Этот вид матриц используется для задания качественных данных. Всякую качественную информацию на данном множестве объектов, в том числе и группировку, мы представляем как разбиение  $R = \{R_1, \dots, R_m\}$  множества объектов со структурой  $\kappa \subset \{1, \dots, m\} \times \{1, \dots, m\}$  связей между значениями признака  $s$  (классами  $R_s$ ),  $s = 1, \dots, m$ . Особенно часто рассматриваются структуры  $\kappa = \{(s, s) \mid s = 1, \dots, m\}$  (номинальные признаки),  $\kappa = \{(s, t) \mid s \leq t\}$  (ранговые признаки),  $\kappa = \{(s, t) \mid |s - t| \leq 1\}$  (цепные признаки),  $\kappa = \{(s, s)\}$  (выделенные значения  $s$ ),  $\kappa$  произвольно (структурные признаки) и т. д. Пара  $(R, \kappa)$  эквивалентно задается булевой  $(\alpha, \beta)$  — матрицей  $r = (r_{ij})$  размерности  $N \times N$ , где  $r_{ij} = \alpha$ , если  $(s, t) \in \kappa$  для  $i \in R_s, j \in R_t$  ( $i, j = 1, \dots, N$  — номера объектов). Удобно, в частности, рассматривать обычное  $(1, 0)$  — представление ( $\alpha = 1, \beta = 0$ ), а также такие  $\tilde{\alpha}, \tilde{\beta}$ , при которых матрица  $r$  центрирована и нормирована как вектор  $N \times N$ -мерного пространства. В этом последнем случае, как нетрудно видеть,  $\tilde{\alpha} = -1/\tilde{\beta}$ , причем  $\tilde{\alpha} = (\|\tilde{\rho}\| / \|\rho\|)^{1/2}$ , где  $\rho = \{(i, j) \mid \exists s, t \ i \in R_s \text{ и } j \in R_t \text{ и } (s, t) \in \kappa\}$  — бинарное отношение, соответствующее паре  $(R, \kappa)$ ;  $\bar{\rho}$  — его дополнение;  $|A|$  — количество элементов в множестве  $A$ . Эти значения  $\tilde{\alpha}$  и  $\tilde{\beta}$  можно интерпретировать как характеристики информативности наличия или отсутствия связи между объектами по данному признаку.

Б) Проекционные  $N \times N$  матрицы.

Этот вид матриц используется для одновременного исследования количественных и номинальных признаков. С каждым признаком  $x$  ассоциируется линейное подпространство  $L(x)$ , реализующее множество допустимых преобразований данного признака, и оператор проектирования  $P_x$  на это пространство, представляемый  $N \times N$  матрицей связи между объектами  $p = (p_{ij})$ . Для количественного признака  $x$  элементы матрицы  $p$  равны  $p_{ij} = x_i x_j$ , где  $x_i$  — значение признака на  $i$ -м объекте, а для номинального —  $p_{ij} = 1/N_s$ , если  $x_i = x_j = s$  (где  $N_s$  — количество объектов, имеющих значение  $s$ ,  $s = 1, \dots, m$ ), и  $p_{ij} = 0$ , если  $x_i \neq x_j$ .

К сожалению, пока остается неясным, какое линейное подпространство может быть связано с качественным признаком, имеющим произвольную структуру связей между значениями.

Задачи конструирования (новых признаков) и описания (признаков в терминах других признаков) формулируются как аппроксимационные задачи в пространстве  $N \times N$  матриц связи. Аппроксимация понимается в рамках стандартного метода наименьших квадратов, т. е. в качестве расстояния между матрицами рассма-

тривается квадрат евклидовой метрики в  $N \times N$ -мерном пространстве матриц связи<sup>1</sup>. Это позволяет использовать для решения таких задач как язык матриц связей между объектами, так и язык матриц связей между признаками (скалярных произведений матриц «объект—объект», отвечающих отдельным признакам). Решающее значение при комплексном исследовании слабоизученных систем имеет, на наш взгляд, задача конструирования. Задачи конструирования группировок по данной системе признаков, включая, возможно, матрицы связей между объектами, в той или иной мере сводятся к следующей типовой задаче [1].

В данном классе  $N \times N$  матриц  $E$  найти такую матрицу  $s = (s_{ij})$ , которая минимизирует квадрат евклидовского расстояния до заданной матрицы связей между объектами  $a = (a_{ij})$ :

$$d(a, s) = \sum_{i,j=1}^N (a_{ij} - s_{ij})^2. \quad (1)$$

Рассмотрим частный случай этой задачи, когда  $E$  состоит из булевских  $(\alpha, \beta)$ -матриц, соответствующих тому или иному классу разбиений со структурой, которые мы будем представлять в виде  $(\lambda r_{ij} + \mu)$ , где  $(r_{ij})$  — булевская  $(1, 0)$  — матрица,  $\lambda$  и  $\mu$  — вещественные числа, такие, что  $\lambda + \mu = \alpha$ ,  $\mu = \beta$ . Тогда задача конструирования состоит в том, чтобы минимизировать

$$\Delta(a, r, \lambda, \mu) = \sum_{i,j=1}^N (a_{ij} - \lambda r_{ij} - \mu)^2 \quad (2)$$

по  $r \in E$ ,  $\lambda, \mu$ .

При фиксированных  $\lambda$  и  $\mu$  эта задача сводится к максимизации суммарной величины связей в разбиении со структурой  $(R, \kappa)$ , соответствующем матрице  $r$ :

$$f(r, \pi) = \sum_{i \neq j} (a_{ij} - \pi) r_{ij} = \sum_{i \neq j} a_{ij} r_{ij} - \pi |\rho| = \sum_{(s, t) \in \kappa} \sum_{i \in R_s} \sum_{j \in R_t} (a_{ij} - \pi), \quad (3)$$

где  $\pi = (\alpha + \beta)/2 = \lambda/2 + \mu$  — порог существенности связи (при  $a_{ij} > \pi$  в силу (3) выгодно помещать  $i$  и  $j$  «внутри»  $(R, \kappa)$ , а при  $a_{ij} < \pi$  — невыгодно). Второе представление функции  $f(r, \pi)$  даст еще одну интерпретацию  $\pi$  как характеристики компромисса между максимизацией суммы внутренних связей  $\sum a_{ij} r_{ij}$  и максимизацией количества учтенных связей. Величина  $\pi$  имеет и другие интерпретации в терминах  $\alpha$  и  $\beta$  (вес искомой группировки; доля исходных связей, учтенных в конструируемой группировке), оптимальные значения  $\alpha$  и  $\beta$  характеризуют среднюю внутреннюю связь  $\sum a_{ij} r_{ij} / \sum r_{ij}$  и среднюю внешнюю связь  $\sum a_{ij} (1 - r_{ij}) / \sum (1 - r_{ij})$  соответственно. На этом основана итеративная процедура

<sup>1</sup> Существенным с технической точки зрения является то, что главная диагональ матриц связи (связи объектов с самими собою) не рассматривается, так что правильнее было бы говорить о  $N(N-1)$ -мерном пространстве квадратных матриц с элиминированной главной диагональю.

минимизации (2) с помощью решения задач вида (3), определения для найденной  $(R, \kappa)$  нового значения  $\pi = \alpha + \beta/2$  и т. д.

Задача (3) исследовалась для случаев, когда а)  $(R, \kappa)$  произвольно; б)  $R$  содержит  $m$  классов,  $\kappa$  произвольно; в)  $R$  произвольно,  $\kappa = \{(s, t) \mid s = t\}$  (номинальная группировка).

В ситуации а) решением задачи (3) является так называемый граф  $\pi$ -сходства  $\Gamma_\pi = \{(i, j) \mid a_{ij} > \pi\}$ . В ситуации б) максимизация критерия (3) сводится к максимизации эквивалентного критерия

$$g(R) = \sum_{s=1}^m \left| \sum_{i \in R_s} \sum_{j \in R_t} (a_{ij} - \pi) \right|, \quad (4)$$

не зависящего от  $\kappa$ , что позволяет применять обычные локально-оптимальные алгоритмы построения разбиений (по схемам «объединения» или «размещения»).

В ситуации в) критерий (3) имеет вид

$$f(R, \pi) = \sum_{s=1}^m \sum_{i, j \in R_s} (a_{ij} - \pi) = \sum_{s=1}^m \sum_{i, j \in R_s} a_{ij} - \pi \sum_s N_s^2, \\ N_s = |R_s| \quad (5)$$

суммы связей (с учетом порога) внутри классов  $R_s$ , причем  $\pi$  выступает показателем компромисса у критерия  $\sum N_s^2$ , увеличивающего однородность распределения  $R$ . Это позволяет доказать, что в оптимальном решении задачи (5) все классы компактны в том смысле, что внутренние средние связи в них превышают средние внешние связи. Число классов оптимального разбиения определяется значением  $\pi$ . Но при увеличении  $\pi$  число классов в нем, вообще говоря, может и уменьшиться, а монотонно зависит от характеристики однородности распределения группировки  $|\rho| = \sum_{s=1}^m N_s^2$ . Схемы последовательного объединения классов и перемещения объектов реализуются для критерия (5) очень просто и дают хорошие результаты на реальных данных.

В классе  $E$  проекционных матриц критерий (1) приводит к задаче построения разбиения  $R = \{R_1, \dots, R_m\}$ , максимизирующего известный критерий средней суммарной связи с учетом числа классов

$$f(R) = \sum_{s=1}^m \frac{1}{N_s} \sum_{i, j \in R_s} a_{ij} - \pi m. \quad (6)$$

Любопытно, что минимизация (1) в классе нормированных проекционных матриц количественных признаков сводится к отысканию собственного вектора матрицы  $a + a^T$ , соответствующего максимальному собственному числу (метод главных компонент).

Величина (1) характеризует качество получаемого решения (остаточная дисперсия). Если она не слишком мала, т. е. решение

не достаточно полно, нами предложено применять аналог метода главных компонент для исчерпания матрицы  $a$ . Согласно этому методу после решения задачи (1) рассматривается матрица остаточных связей  $a' = a - s$ , для которой вновь решается задача о построении аппроксимирующей группировки, и так до тех пор, пока дисперсия остаточных связей не станет достаточно мала. Исследование этого метода качественного факторного анализа в классе  $(\alpha, \beta)$ -матриц проведено В. А. Трофимовым, который показал, в частности, что доля дисперсии, снимаемая данным фактором  $\lambda + \mu$ , пропорциональна величине  $\lambda^2 | \rho | | \bar{\rho} |$ . Он же доказал, что в рассматриваемых классах разбиений со структурными<sup>2</sup> этот метод сходится, хотя и за бесконечное число шагов, и предложил конечную модификацию, основанную на оптимальном пересчете весовых коэффициентов  $\lambda$  и  $\mu$  от шага к шагу [1]. Аналогично доказывается сходство метода в классах проекционных матриц. Экспериментальный опыт использования качественного факторного анализа пока невелик.

Интерпретация получаемого решения определяется степенью его близости к тем или иным из исходных признаков. Использование подходящих структур  $\chi$  повышает интерпретируемость. Однако радикальное решение проблемы интерпретируемости видится в переходе на язык признаков и их значений. Возможность использования этого языка определяется тем, что расстояние (1) между матрицами (и их скалярное произведение, ассоциированное с расстоянием) для рассматриваемых классов матриц может быть выражено в терминах распределений соответствующих признаков и, следовательно, может использоваться как мера корреляции между признаками и конструируемыми группировками [2]. В частности, скалярное произведение проекционных матриц равно коэффициенту сопряженности Пирсона  $\chi^2$ , когда оба признака номинальные; коэффициенту детерминации  $\rho^2$ , когда оба признака количественные; и для разнотипных признаков — квадрату корреляционного отношения  $\eta^2$  (т. е. относительному приращению дисперсии количественного признака при переходе от средней его дисперсии в классах номинального признака к дисперсии на всем массиве).

Коэффициент  $\chi^2$ , вопреки распространенному мнению, имеет четкий смысл как оценка качества прогноза одного признака по другому [2]. П. С. Ростовцев с помощью статистического эксперимента показал адекватность этого показателя в задачах группирования [1].

Группировка  $R$  конструируется так, чтобы ее средняя связь с заданной совокупностью «выходных» признаков  $Y$  была максимальной, с помощью алгоритмов последовательных разбиений совокупности объектов признаками совокупности  $X$ , используемой для построения, вообще говоря,  $R (X \neq Y)$ . Алгоритм последовательных разбиений на общем шаге при исходном разбиении

<sup>2</sup> В необходимых случаях матрица  $a$  должна быть симметрична.

$R'$  просматривает возможности разделения отдельных классов  $R'$  по каждому из признаков  $X$ , выбирая в качестве исходной для следующего шага группировки тот вариант  $R''$ , для которого приращение суммарной связи с  $Y$  максимально (Ростовцев, Мучник).

Классы такой группировки характеризуются комбинациями значений признаков  $X$ , которые ориентированы на описание (предсказание значений) признаков  $Y$ , что существенно повышает интерпретируемость результата.

Методы, основанные на мерах связи между признаками, могут применяться при очень больших выборках, поскольку вычисления в них практически не зависят от числа объектов.

Рассмотрим теперь, как изложенные представления используются для решения сформулированных выше основных задач метода группировок в автоматическом режиме.

## 1. Конструирование типологических группировок

Для анализа анкетной информации удобно использовать метод последовательных разбиений, причем исследователь формирует множества «входных» признаков  $X$  и «выходных» признаков  $Y$  (обычно  $X = Y$ ), получая типологическую группировку объектов в терминах входных признаков, построенную так, чтобы как можно лучше прогнозировать  $Y$ . В частности, таким способом строится группировка по данному количественному признаку (интервалирование), а также по данному неальтернативному признаку (в последнем случае в качестве  $X = Y$  используются признаки, соответствующие отдельным неальтернативным значениям).

В том случае, когда такая группировка оказывается слишком грубой, используются связи между объектами на основе критериев вида (3), (5). При этом матрица  $a$  строится по информации о признаках, что допускает такую организацию вычислений, при которой используются только исходные данные объект-признак, а величины связи  $a_{ij}$  вычисляются по мере необходимости. При этом возможен учет показателей веса исходных признаков, в том числе и оптимальных с точки зрения задачи аппроксимации (правда, здесь практически нет экспериментального опыта). При относительно большом количестве объектов (более 500) используются быстрые алгоритмы группировки, аналогичные методу  $k$ -средних.

В некоторых задачах оправдан поиск группировки с фиксированной структурой связей между классами. Например, типы образа жизни предполагаются связанными в цепочку, что характеризует возможные переходы от одного типа образа жизни к другому. Поэтому для конструирования типологии образа жизни используется программа, осуществляющая поиск цепной группировки с помощью соответствующей модификации метода  $k$ -средних.

## 2. Исследование связи между признаками

Исследователь указывает группы признаков  $I$  и  $J$ , между которыми требуется проанализировать связь. Формируется прямоугольная матрица  $|I| \times |J|$  коэффициентов связи  $(w_{kl})$ ,  $k \in I$ ,  $l \in J$ , для которой методами качественного факторного анализа отыскиваются детерминирующие друг друга совокупности признаков  $X \subseteq I$  и  $Y \subseteq J$ . Пара  $(X, Y)$  характеризуется булевой  $(\alpha, \beta)$ -матрицей  $s = (s_{kl})$  размерности  $|I| \times |J|$ , где  $s_{kl} = \alpha$ , если  $(k, l) \in I \times J$ , и  $s_{kl} = \beta$  — в противном случае. Теперь остается использовать критерий вида (2) для отыскания оптимальных  $X$  и  $Y$  по данной матрице  $w$  с дальнейшим переходом к матрице  $w' = w - s$  остаточных связей и повторением процедуры по методу качественного факторного анализа.

П. С. Ростовцев разработал и другие процедуры построения детерминирующих пар  $(X, Y)$ . Для каждой такой пары осуществляется построение типологической группировки методом последовательных разбиений. Комбинация полученной группировки с каждым из признаков  $Y$  (или аналогичным фактором для  $Y$ ) дает картину связей между  $X$  и  $Y$ . Особенно эффективно такое описание в случае, когда  $Y$  состоит из единственного признака  $y$ : каждому  $t$ -му значению  $y$  можно приписать соответствующие ему комбинации значений  $X$  — те, для которых приращение вероятности  $t$  наибольшее [1].

Иногда интерес представляет выявление группы взаимосвязанных признаков в заданной системе  $I$ . Это осуществляется группировкой по  $|I| \times |I|$ -матрице связи признаков на основе критерия вида (5) или (6).

## 3. Исследование связи между объектами

В качестве исходной здесь выступает одна или несколько матриц связи между объектами. Для анализа используется программа, реализующая критерий вида (2)–(5), где  $a$  получена суммированием данных матриц с учетом их весов.

Критерий (5) приводит к группировкам, основанным на обычном принципе компактности: относительно равномерное распределение объектов на относительно далекие друг от друга группы, состоящие из тесно связанных объектов. В том случае, когда нет никаких априорных свидетельств в пользу именно такой структуры связей, ищется разбиение  $R$  с заранее не заданной структурой  $\kappa$  — критерий (3)–(4). К сожалению, в отличие от задачи (5) о поиске номинальной группировки критерий вида (1) дает оптимальное решение тривиальным образом в виде графа  $\lambda$ -сходства, так что здесь необходимо задать еще какой-либо параметр, например число классов, что приводит к критерию (4). Для полной автоматизации процесса выявления структуры основных

связей можно использовать двухфакторную модель: задачу минимизации критерия

$$\sum_{i,j} (a_{ij} - s_{ij} - k_{ij})^2, \quad (7)$$

где  $(s_{ij})$  есть  $(\alpha, \beta)$ -матрица, соответствующая искомому неупорядоченному разбиению  $R$ , а  $(k_{ij})$  есть  $(\alpha', \beta')$ -матрица, соответствующая (искомой) структуре связей между классами  $\kappa$  [2]. Алгоритмы, основанные на критерии (7), используют линейную комбинацию критериев (4) и (5) при различных порогах  $\lambda$ .

Исходя из сказанного, можно утверждать, что проблемно-ориентированное математическое обеспечение метода группировок, должно включать в себя средства автоматического анализа, проверки исследовательских группировок, хорошее обеспечение ввода и вывода данных, а также манипулирования данными, освобождающее исследователя от проведения рутинных технических операций, не включенных в конкретные рассматриваемые проблемы.

### Литература

1. Методы анализа многомерной экономической информации. Новосибирск, 1981.
2. Миркин Б. Г. Анализ качественных признаков и структур. М., 1980.

Г. В. РАУШЕНБАХ

## МЕРЫ БЛИЗОСТИ И СХОДСТВА

Статья представляет собой обзор различных мер близости, пригодных для измерения степени «похожести» объектов, описанных величинами, измеренными в порядковой и номинальной шкалах. Вычисление той или иной меры близости (между объектами либо между признаками) является составной частью многих используемых в социологии математических методов. Поэтому статья носит достаточно общий (с точки зрения круга задач, решаемых с помощью рассматриваемых в статье методов) характер.

В работе не дается исчерпывающих рекомендаций по выбору меры близости при решении тех или иных социологических задач. Однако ряд отраженных в ней моментов все же позволяет ограничить круг пригодных для решения конкретных задач мер близости.

Меры близости — один из важнейших инструментов анализа данных: сопоставление, классификация, выбор средних, поиск закономерностей связаны с оценкой сходства, различия, корреляции, сопряженности изучаемых объектов, т. е. с измерением близости между ними.

В соответствии со сложившейся традицией [22, 110] будем употреблять термин «близость» во всех случаях, когда речь идет о сопоставлении, сравнении данных. «Термин „близость“ ... относится к сходству, к различию, к корреляции, к мере пересечения или

же любой другой переменной, используемой в качестве меры сходства или расстояния между двумя объектами одного вида» [22, с. 21]. Изучение влияния, доминирования и т. п. также может рассматриваться в этом контексте [43, 64, 111], однако основное внимание будет уделено симметричным неотрицательным мерам близости.

Типичными задачами измерения близости при анализе социологической информации являются задачи сопоставления и классификации данных, оценка средних, выбор представителя совокупности наблюдений, аппроксимация данных, проверка качества модели, сравнение различных моделей. В качестве мер близости чаще всего используются некие показатели удаленности (расстояния), коэффициенты корреляции, ассоциации, сопряженности, несколько реже — специально сконструированные индексы сходства, подобия, общности.

В содержательном аспекте задачи измерения близости в социологии весьма разнообразны. В качестве примера укажем задачи анализа социометрических структур [33, 34, 70, 127, 158], построения типологий [31, 55, 56, 59, 60, 67, 120], изучения групповых и индивидуальных предпочтений [1, 32, 43, 61] и социальных сетей [96]. Многие исследователи вводили специально сконструированные меры близости или приспособляли известные ранее для согласования предпочтений [23, 79, 80, 121], сравнения и агрегации разбиений [34—38, 71, 81], измерения социального расстояния [98, 154] и расстояния в пространстве социальных распределений [77, 85], при изучении престижа профессий [43] и целей жизнедеятельности [61], отношений родства [83], влияния и власти [21], измерения социальной зависимости [64] и близости к социальному идеалу [56]. Инструментами исследования матриц близости (сходства, связи) в социологии являются, в частности, такие методы, как многомерное шкалирование [17, 22, 58, 59] и кластер-анализ [2, 12, 36, 45, 59, 60, 70, 78, 93, 145].

## Общие принципы измерения близости

Измерение близости между объектами предполагает наличие определенной информации о них. В наиболее общем случае эта информация может быть представлена в виде вектора количественных характеристик. В более простой ситуации будем считать, что каждому объекту  $X$  из множества объектов  $O = \{X, Y, \dots\}$  можно поставить в соответствие некоторое множество  $\mathcal{X}$ , принадлежащее алгебре множеств  $\mathcal{P}$  с единицей  $E$ :

$$X \xrightarrow{\varphi} \varphi(X) = \mathcal{X} \in \mathcal{P}.$$

Элементы  $e_i$  множества  $E$  могут рассматриваться как характерные свойства (информативные признаки) объектов из множества  $O$ . Отображение  $\varphi$  ставит в соответствие каждому объекту  $X$  набор

его признаков  $X = \varphi(X)$ . Может оказаться, что двум различным объектам  $X_1$  и  $X_2$  соответствует один и тот же образ:  $\varphi(X_1) = \varphi(X_2)$ , и тогда мы будем считать  $X_1$  и  $X_2$  эквивалентными в интересующем нас аспекте измерения близости. В дальнейшем мы, как правило, будем опускать обозначения  $X, Y, \dots$ , отождествляя объекты с их образами  $\mathcal{X}, Y, \dots, \in \mathcal{P}$ . Если известны не только признаки, но и интенсивность (или, возможно, вероятность) их проявления на объекте  $\mathcal{X}$ , то удобной моделью оказывается множество  $\mathcal{X}$  из пространства нечетких множеств  $\mathcal{P}_F$  с носителями, принадлежащими алгебре  $\mathcal{P}$ . Последовательно обобщая нашу модель, мы придем к  $n$ -мерному пространству  $\mathbb{R}^n$  и далее к пространству  $L$  вещественнозначных функций, определенных на множестве  $E$ . Сказанное выше мы можем записать в следующей форме:

$$X \rightarrow (x_1, \dots, x_n),$$

где  $x_i$  — значение индикатора множества  $X \in \mathcal{P} (\chi_X(e_i))$ , функции принадлежности нечеткого множества  $X \in \mathcal{P}_F (\mu_X(e_i))$ , либо просто  $i$ -я компонента вектора  $X \in \mathbb{R}^n$  и т. д.

В задачах анализа нечисловой информации наиболее распространенной моделью является пространство  $\mathcal{P}$ , несколько реже рассматриваются  $\mathcal{P}_F$  и  $\mathbb{R}_+^n$ <sup>1</sup>. В ряде задач, связанных в первую очередь с социологическими приложениями, весьма плодотворным оказывается подход, основанный на изучении частично упорядоченных множеств [75, 76]. Так, измерение близости между разбиениями может рассматриваться в контексте метризации геометрических решеток [11, 81], пространства слабого, частичного, полного и иных порядков полурешетки [73, 74, 131] и т. д. Мы будем иметь в виду, что пространства  $\mathcal{P}$  и  $\mathcal{P}_F$  — дистрибутивные решетки относительно включения, объединения и пересечения, пространства  $\mathbb{R}_+^n, \mathbb{R}^n, L$  — решетки различных типов относительно операций  $\leq, \max, \min$  [6, 11].

Многие меры близости, как будет показано ниже, могут быть описаны с единых позиций, если использовать понятие оценки [6].

**Определение 1.** Пусть  $\mathcal{L}$  — решетка с операциями  $\wedge$  и  $\vee$ . Вещественнозначная функция  $v: \mathcal{L} \rightarrow \mathbb{R}$  называется нижней (верхней) оценкой, если для любых  $X, Y \in \mathcal{L}$  выражение  $v(X \wedge Y) + v(X \vee Y) - v(X) - v(Y)$  неотрицательно (неположительно). Функция  $v$  называется оценкой, если она является верхней и нижней оценкой одновременно. Оценка называется изотонной, если  $X \leq Y \Leftrightarrow v(X) \leq v(Y)$  и положительной, если  $v(X) \geq 0$ , причем  $v(X) = 0$  тогда и только тогда, когда  $\mathcal{X}$  — минимальный элемент  $\mathcal{L}$ .

<sup>1</sup>  $\mathbb{R}_+^n = \{(x_1, \dots, x_n): x_i \geq 0\}$ .

<sup>2</sup> Частично упорядоченное множество  $(\mathcal{L}, \leq)$  называется решеткой, если для любых  $X, Y \in \mathcal{L}$  существует точная нижняя грань  $X \wedge Y \in \mathcal{L}$ , причем  $X \wedge Y \leq X, X \wedge Y \leq Y$  и для любого  $Z \in \mathcal{L}$  ( $Z \leq X$  и  $Z \leq Y$ )  $\Rightarrow Z \leq X \wedge Y$ , а также двойственная ей точная верхняя грань  $X \vee Y$  [6, 11].

Примерами положительных оценок являются мощность (конечного) множества из пространства  $\mathcal{P}$  или  $\mathcal{P}_F$ :

$$|X| = \sum_i \chi_X(e_i), \quad |X|_F = \sum_i \mu_X(e_i)^2,$$

а в более общем случае — мера множества:

$$\nu(X) = \int_E \chi_X d\nu, \quad \nu_F(X) = \int_E \mu_X d\nu,$$

норма матрицы с неотрицательными элементами и т. д. Примеры верхних и нижних оценок, используемых при конструировании мер близости, читатель найдет, например, в [71, 81]. Мы приведем здесь довольно простой результат, иллюстрирующий связь между оценками и верхними (нижними) оценками.

*Теорема 1.* Пусть  $\nu$  — положительная изотонная оценка на решетке  $\mathcal{L}$ ,  $\psi$  — дважды дифференцируемое отображение вещественной прямой в себя. Для того чтобы композиция  $\psi \circ \nu$  была верхней (нижней) оценкой на  $\mathcal{L}$ , необходимо и достаточно, чтобы  $\psi$  была выпуклой вверх (вниз).

## Сходство и несходство

Большинство мер близости, используемых при анализе социологических данных, по существу являются мерами сходства (или, напротив, несходства) либо могут быть сведены к ним. Дефиниции сходства и несходства, предлагаемые различными исследователями, отличаются друг от друга; мы приведем здесь определения, отражающие наиболее общий, традиционный взгляд на меры сходства и несходства.

*Определение 2.* Мерой сходства в пространстве  $Q^4$  называется функция  $S: Q \times Q \rightarrow \mathbb{R}_+$ , обладающая следующими свойствами:

$$S1. \quad 0 \leq S(X, Y) \leq 1.$$

$$S2. \quad S(X, X) = 1.$$

$$S3. \quad S(X, Y) = S(Y, X).$$

Иногда  $S2$  заменяется менее жестким требованием  $S2'$ .  $S(X, X) = \max_Y S(X, Y)$ . Симметричность меры сходства также является не столько обязательным, сколько, пожалуй, более традиционным требованием: так, сходство населенных пунктов  $X$  и  $Y$  несимметрично, если в качестве меры его рассматривать величину потока миграции из  $X$  в  $Y$  [33, 113]. Л. Хьюберт [111]

<sup>3</sup>  $\chi_X, \mu_X$  — соответственно индикатор (характеристическая функция) и функция принадлежности множества  $X$ .

<sup>4</sup> Вместо  $Q$  можно подставить  $\mathcal{P}, \mathcal{P}_F, \mathbb{R}^n$  или обозначение любого другого пространства.

и А. Тверски [155] приводят и другие примеры несимметричного сходства.

По аналогии с мерой сходства определим и меру несходства.

*Определение 3.* Мерой несходства в пространстве  $Q$  называется функция  $D: Q \times Q \rightarrow \mathbb{R}_+$ , обладающая следующими свойствами:

$$D1. 0 \leq D(X, Y) \leq 1.$$

$$D2. D(X, X) = 0 \quad (D2'. D(X, X) = \min D(X, Y)).$$

$$D3. D(X, Y) = D(Y, X).$$

Очевидно, по заданной мере сходства  $S$  всегда можно построить меру несходства  $D = 1 - S$ , и наоборот. Если снять ограничение сверху на величину  $D(X, Y)$  в  $D1$ , то получим систему требований, предъявляемых к так называемым мерам расстояния (удаленности) или квазиметрикам [2, 12, 78]. Дополнительное условие  $D4$ . Для любых  $X, Y, Z \in Q$   $D(X, Z) \leq D(X, Y) + D(Y, Z)$  (неравенство треугольника) означает, что в качестве меры близости рассматривается псевдометрика [15]. Наконец, условие

$$D5. \text{ Если } D(X, Y) = 0, \text{ то } X = Y,$$

в совокупности с  $D1 - D4$  образует известную аксиоматику Фреше [97], которой удовлетворяют истинные расстояния или метрики<sup>5</sup>.

Сходство и несходство, подобие и различие формализуются также и на языке бинарных отношений (в общем случае — нечетких). Обычно отношением сходства называют симметричное рефлексивное отношение  $\sigma$  [102, 107, 114], т. е. толерантность [68]; дополнительные условия типа транзитивности (например, для отношения родства) вводятся реже [5, 21, 127]. Отношения близости [13] и соседства [14, 26] обладают аналогичными свойствами. Отношение несходства  $\delta$  иногда определяется как дополнительное к отношению сходства [21, 42, 127], а иногда — как отношение различимости [147]. Перейти к нечетким отношениям сходства (несходства) [21, 24, 160] тем более удобно, что в этом случае мера сходства  $S$  (мера несходства  $D$ ), удовлетворяющая  $S1 - S3$  ( $D1 - D3$ ), интерпретируется как функция принадлежности отношения  $\sigma_F$  ( $\delta_F$ ).

Когда скоро элементы множества  $E$  рассматриваются как информативные признаки, естественно потребовать, чтобы мера сходства достигала минимального значения тогда, когда объекты общих признаков не имеют:

$$X \cap Y = \emptyset \Rightarrow S(X, Y) = 0 \quad (1)$$

(при этом хотя бы одно из множеств  $X, Y$  предполагается непустым, так как согласно  $S2$   $S(\emptyset, \emptyset) = 1$ ). Если предположить, что

<sup>5</sup> Ограничение сверху в  $D1$  при этом можно оставить в силе, так как любое (псевдо)метрическое пространство гомеоморфно (псевдо)метрическому пространству единичного диаметра [15, с. 165].

приращению меры сходства  $\Delta S = S(X', Y') - S(X, Y)$  соответствует равное по абсолютной величине, но противоположное по знаку приращение меры несходства  $\Delta D = D(X', Y') - D(X, Y)$ , то для любых  $X, Y \in Q$

$$S(X, Y) + D(X, Y) = 1, \quad (2)$$

причем

$$X \cap Y = \emptyset \Rightarrow D(X, Y) = 1 \text{ при } X \cup Y \neq \emptyset. \quad (3)$$

Условие (3) было введено в работе [49] при построении метрики несходства —  $D$ -метрики<sup>6</sup>. Таким образом, требования  $S1$  и  $D1$ , (1)–(3) вполне логично укладываются в ту схему, согласно которой

$$S(X, Y) = \mu_{\sigma_F}(X, Y) = 1 - \mu_{\delta_F}(X, Y).$$

Многие показатели связи, сопряженности, корреляции и т. д., часто используемые в качестве мер близости [2, 8, 16, 36], принимают значения в интервале  $[-1; 1]$  и, следовательно, не удовлетворяют требованию  $S1$ . Более существенное различие между мерами сходства и показателями связи состоит в том, что первые обычно вводятся на множество пар объектов, последние же — на множество пар признаков. Рассуждая формально, любой показатель связи можно рассматривать как меру близости в соответствующем пространстве и наоборот (достаточно транспонировать матрицу «объект—признак»). Величины

$$S(X, Y) = \frac{1}{2} [1 + r(X, Y)] \quad \text{и} \quad D(X, Y) = \frac{1}{2} [1 - r(X, Y)],$$

где  $r(X, Y)$  — некий коэффициент корреляции со значениями в  $[-1, 1]$ , удовлетворяют требованиям  $S1, D1$ . Однако постановка задачи и интерпретация результатов при измерении сходства и связи, как правило, различаются. Объекты  $X, Y, \dots$  обычно являются представителями некоторой однородной в определенном смысле совокупности (например, членами производственного коллектива), в то время как характеризующие их признаки могут быть существенно различной природы, в частности, могут быть измерены в различных шкалах (пол, возраст, служебное положение). При определении связи между признаками, измеренными в шкалах, отличных от номинальной, имеется возможность интерпретации характера связи (положительная — отрицательная), что лишено смысла при измерении близости.

Отмечая различия в постановке задач измерения сходства и связи, мы тем не менее будем иметь в виду, что в качестве математического инструмента показатели связи могут быть использованы как меры близости.

\* От англ. dissimilarity — несходство.

## Меры близости и показатели сходства

В табл. 1 приведены меры близости, наиболее часто употребляемые в задачах анализа нечисловой информации, — показатели сходства, меры несходства (метрики и квазиметрики). Перечень этот легко может быть увеличен [12, 46, 78, 86, 102, 153]. Заметим, что с помощью табл. 1 можно сконструировать «новые» меры близости: так, каждой из первых десяти мер сходства  $S$  соответствует некая мера несходства  $D = 1 - S$  и т. д.

Меры близости объектов, описываемых дихотомическими признаками, принято выражать в терминах четырехклеточной таблицы сопряженности:  $a$  — число признаков, отсутствующих у  $X$  и  $Y$  одновременно,  $d$  — число совпадающих признаков,  $b$  (или  $c$ ) — число признаков, присутствующих у  $X$ , но отсутствующих у  $Y$  (или наоборот),  $a + b + c + d = n$ . Во втором столбце табл. 1 приведены формулы для показателей сходства и несходства, использующие как величины  $a, b, c, d$ , так и «векторную» форму записи  $X = (x_1, \dots, x_n)$ . Более общие выражения, использующие понятие изотонной положительной оценки  $v$  на некоторой решетке  $\mathcal{L}^7$  (тип решетки варьирует в зависимости от характера задачи), можно найти в третьем столбце. При этом употребляем обозначения:

$$d_v(X, Y) = v(X) + v(Y) - 2v(X \wedge Y),$$

если  $v$  — изотонная верхняя оценка;<sup>7</sup>

$$d_v(X, Y) = 2v(X \vee Y) - v(X) - v(Y),$$

если  $v$  — изотонная нижняя оценка;

1 и 0 — соответственно максимальный и минимальный элементы  $\mathcal{L}$  (т. е.  $E$  и  $\emptyset$  в решетках  $\mathcal{P}$  и  $\mathcal{P}_F$ );  $\bar{X}$  — псевдодополнение к  $X$  (теоретико-множественное дополнение к  $X$  в решетках  $\mathcal{P}$  и  $\mathcal{P}_F$ ).

Мы ограничимся пространствами  $\mathcal{P}$ ,  $\mathcal{P}_F$ ,  $\mathbb{R}_+^n$ , наиболее существенными для интересующих нас задач анализа нечисловой информации, добавив в целях обобщения пространство  $L_+$  неотрицательных суммируемых функций на множестве  $E$ . Знак «+» в каком-либо из столбцов 4—7 означает, что в соответствующем пространстве данная мера близости удовлетворяет требованиям  $S1 - S3$  либо  $D1 - D3$ . Прочерк ставится в том случае, когда мера близости либо не определена, либо значения ее могут оказаться неинтерпретируемыми. Если мера близости определена и интерпретируема в некотором пространстве, но при этом принимает значения в интервале  $[-1; 1]$  или  $[0; \infty]$ , проставлен знак «(+»).

Отсутствие ссылок на литературу там, где нет прочерка, означает, что автору не известны публикации, в которых говорится о введении данной меры близости в соответствующем пространстве. В тех случаях, когда ссылки имеются, они относятся в первую очередь к авторам рассматриваемых мер близости, а также к наиболее известным работам по их обоснованию (например, аксиоматическому) и применению в социологии.

<sup>7</sup> Или, возможно, верхней (нижней) оценки на верхней (нижней) полурешетке с максимальным (минимальным) элементом.

Таблица 1  
Меры близости

Обозначение	Вид меры близости в пространстве $Q$	Вид меры близости на решетке $\mathcal{L}$	Тип пространства $Q$			
			$\mathcal{S}^0$	$\mathcal{S}^F$	$\mathbb{R}_+^n$	$L_+$
$S_1$	$\frac{a+d}{a+b+c+d}$ ;	$\frac{1}{n} \left( n - \sum_i  x_i - y_i  \right)$	$\frac{+}{[150, 161]}$	+	-	-
$S_2$	$\frac{2(a+d)}{2(a+d)+b+c}$ ;	$\frac{2 \left( n - \sum_i  x_i - y_i  \right)}{2n - \sum_i  x_i - y_i }$	$\frac{+}{[152]}$	+	-	-
$S_3$	$\frac{a+d}{a+d+2(b+c)}$ ;	$\frac{n - \sum_i  x_i - y_i }{n + \sum_i  x_i - y_i }$	$\frac{+}{[140]}$	+	-	-
$S_4$	$\frac{d}{a+b+c+d}$ ;	$\frac{1}{n} \sum_i \min(x_i, y_i)$	$\frac{+}{[141, 145]}$	+	(+)	(+)
$S_5^*$	$\frac{d}{b+c+d}$ ;	$\frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$	$\frac{+}{[102, 116, 132]}$	+	$\frac{+}{[102]}$	-
$S_6^*$	$\frac{a}{a+b+c}$ ;	$\frac{n - \sum_i \max(x_i, y_i)}{n - \sum_i \min(x_i, y_i)}$	+	+	-	-

Таблица 1 (продолжение)

Обозначение	Вид меры близости в пространстве $\mathcal{Q}$	Вид меры близости на решетке $\mathcal{Z}$	Тип пространства $\mathcal{Q}$			
			$\mathcal{S}^p$	$\mathcal{S}^p Y$	$\mathbb{R}^n$	$L_+$
$S_7^*$	$\frac{d}{2(b+c)+d}; \frac{\sum_i \min(x_i, y_i)}{\sum_i [2 \max(x_i, y_i) - \min(x_i, y_i)]}$	$\frac{v(X \wedge Y)}{2d_v(X, Y) + v(X \wedge Y)}$	$^+_{[152]}$	+	+	+
$S_8^*$	$\frac{2d}{b+c+2d}; \frac{2 \sum_i \min(x_i, y_i)}{\sum_i (x_i + y_i)}$	$\frac{2v(X \wedge Y)}{v(X) + v(Y)}$	$^+_{[89, 94, 117, 153]}$	+	+	+
$S_9^{**}$	$\frac{d}{V(b+d)(c+d)}; \frac{\sum_i \min(x_i, y_i)}{\sqrt{\sum_i x_i \sum_i y_i}}$	$\frac{v(X \wedge Y)}{\sqrt{v(X)v(Y)}}$	$^+_{[91, 133]}$	+	+	+
$S_{10}^{**}$	$\frac{1}{2} \left( \frac{d}{b+d} + \frac{d}{c+d} \right); \frac{\sum_i \min(x_i, y_i)}{2 \sum_i y_i} + \frac{\sum_i \min(x_i, y_i)}{2 \sum_i x_i}$	$\frac{v(X \wedge Y)}{2} \left[ \frac{1}{v(X)} + \frac{1}{v(Y)} \right]$	$^+_{[124]}$	+	+	+

Таблица 1 (продолжение)

Обозначение	Вид меры близости в пространстве $\mathcal{Q}$	Вид меры близости на решетке $\mathcal{L}$	Тип пространства $\mathcal{Q}$			
			$d^p$	$d^p_F$	$\mathbb{R}^n$	$L_+$
$S_{11}$	$\frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}};$ $\sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i$		$\begin{pmatrix} + \\ [16, 36, \\ 152] \end{pmatrix}$	$(+)$	$\begin{pmatrix} + \\ [25, 78, \\ 106, 114, \\ 152] \end{pmatrix}$	$(+)$
$S_{12}$	$\sqrt{\left[ \sum_i x_i^2 - \frac{1}{n} \left( \sum_i x_i \right)^2 \right] \left[ \sum_i y_i^2 - \frac{1}{n} \left( \sum_i y_i \right)^2 \right]}$ $\frac{a - b - c + d}{a + b + c + d}; \quad \frac{1}{n} \left( n - 2 \sum_i  x_i - y_i  \right)$	$\frac{v(1) - 2d_v(X, Y)}{v(1)}$	$\begin{pmatrix} + \\ [105, 146] \end{pmatrix}$	$(+)$	$-$	$-$
$S_{13}$	$\frac{ad - bc}{\frac{1}{2} (a + b + c + d)(b + c) + ad - bc};$ $\sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i$		$\begin{pmatrix} + \\ [52, 53] \end{pmatrix}$	$(+)$	$(+)$	$(+)$
$D_1$	$\frac{1}{2} \left( \sum_i x_i + \sum_i y_i \right) - \frac{1}{n} \sum_i x_i \sum_i y_i$ $\frac{b+c}{a+b+c+d}; \quad \frac{1}{n} \sum_i  x_i - y_i $	$\frac{d_v(X, Y)}{v(1)}$	$\begin{pmatrix} + \\ [37, 40, \\ 79, 90, \\ 121, 138] \end{pmatrix}$	$\begin{pmatrix} + \\ [21, \\ 24, 39, \\ 137] \end{pmatrix}$	$\begin{pmatrix} + \\ [20, 130, \\ 157] \end{pmatrix}$	$\begin{pmatrix} + \\ [130] \end{pmatrix}$
$D_2$	$\sqrt{\frac{b+c}{a+b+c+d}}; \quad \sqrt{\frac{1}{n} \sum_i (x_i - y_i)^2}$		$\begin{pmatrix} + \\ [10, 79, \\ 80, 149] \end{pmatrix}$	$\begin{pmatrix} + \\ [21, \\ 39] \end{pmatrix}$	$\begin{pmatrix} + \\ [29, 108, \\ 135, 149, \\ 157] \end{pmatrix}$	$(+)$

Таблица 1 (продолжение)

Обозначение	Вид меры близости в пространстве $\mathcal{Q}$	Вид меры близости на решетке $\mathcal{L}$	Тип пространства $\mathcal{Q}$			
			$\mathcal{F}$	$\mathcal{F}^p$	$\mathbb{R}^n$	$L_+$
$D_3$	$\sup_i  x_i - y_i $	$\sup_i d_v(X^i, Y^i)$	$\begin{smallmatrix} + \\ [102] \end{smallmatrix}$	$+$	$\begin{smallmatrix} (+) \\ [102, 130] \end{smallmatrix}$	$\begin{smallmatrix} (+) \\ [130] \end{smallmatrix}$
$D_4^*$	$\frac{b+c}{b+c+d}; \quad \frac{\sum_i  x_i - y_i }{\sum_i \max(x_i, y_i)}$	$\frac{d_v(X, Y)}{v(X \vee Y)}$	$\begin{smallmatrix} + \\ [49, 50, \\ 98, 128, \\ 136, 148] \end{smallmatrix}$	$\begin{smallmatrix} + \\ [49, \\ 50] \end{smallmatrix}$	$\begin{smallmatrix} (+) \\ [49, 50] \end{smallmatrix}$	$\begin{smallmatrix} + \\ [49, \\ 50, \\ 128] \end{smallmatrix}$
$D_5^*$	$\frac{b+c}{a+b+c}; \quad \frac{\sum_i  x_i - y_i }{n - \sum_i \min(x_i, y_i)}$	$\frac{d_v(X, Y)}{v(X \vee Y)} = \frac{d_v(X, Y)}{v(1) - v(X \wedge Y)}$	$\begin{smallmatrix} + \\ [50], \end{smallmatrix}$	$\begin{smallmatrix} + \\ [50] \end{smallmatrix}$	$-$	$-$
$D_6$	$\frac{2(b+c)}{a+2(b+c)}; \quad \frac{2 \sum_i  x_i - y_i }{n + \sum_i  x_i - y_i }$	$\frac{2d_v(X, Y)}{v(1) + d_v(X, Y)}$	$\begin{smallmatrix} + \\ [49] \end{smallmatrix}$	$+$	$-$	$-$
$D_7^*$	$\frac{b+c}{b+c+2d}; \quad \frac{\sum_i  x_i - y_i }{\sum_i (x_i + y_i)}$	$\frac{d_v(X, Y)}{v(X) + v(Y)}$	$+$	$+$	$\begin{smallmatrix} + \\ [84] \end{smallmatrix}$	$+$
$D_8^{***}$	$\frac{1}{n} \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$		$+$	$\begin{smallmatrix} + \\ [72] \end{smallmatrix}$	$\begin{smallmatrix} (+) \\ \end{smallmatrix}$	$\begin{smallmatrix} (+) \\ \end{smallmatrix}$

Таблица 1 (окончание)

Обозначение	Вид меры близости в пространстве $\mathcal{Q}$	Вид меры близости на решетке $\mathcal{L}$	Тип пространства $\mathcal{Q}$			
			$\mathcal{P}$	$\mathcal{P}_F$	$\mathbb{R}^n_+$	$L_+$
$D_9$	$\frac{b+c}{a+b+c+d}; \quad \frac{1}{n} \sum_i (x_i - y_i)^2$		$\begin{smallmatrix} + \\ [30] \end{smallmatrix}$	$\begin{smallmatrix} + \\ [39] \end{smallmatrix}$	$\begin{smallmatrix} (+) \\ [12, 12^2] \end{smallmatrix}$	(+)
$D_{10}^{**}$	$\log_2 \frac{b+c+d}{d}; \quad \log_2 \frac{\sum_i \max(x_i, y_i)}{\sum_i \min(x_i, y_i)}$	$-\log_2 S_6(X, Y)$	$\begin{smallmatrix} (+) \\ [40] \end{smallmatrix}$	(+)	(+)	(+)
$D_{11}^{***}$	$\frac{1}{k} \sum_{i=1}^k \left[ \frac{\sum_{j=1}^{n_i}  x_j - y_j }{\sum_{j=1}^{n_i} \max(x_j, y_j)} \right]$	$\frac{1}{k} \sum_{i=1}^k D_4(X^i, Y^i)$	$\begin{smallmatrix} + \\ [49] \end{smallmatrix}$	$\begin{smallmatrix} + \\ [49] \end{smallmatrix}$	$\begin{smallmatrix} + \\ [53] \end{smallmatrix}$	+
$D_{12}^{***}$	$\frac{1}{n} \sum_i \frac{ x_i - y_i }{x_i + y_i}$	$\frac{1}{n} \sum_i D_7(X^i, Y^i)$	+	+	$\begin{smallmatrix} + \\ [125] \end{smallmatrix}$	+
$D_{13}^{***}$	$\frac{1}{n} \sum_i \frac{ x_i - y_i }{\max_{T \in Q} t_i - \min_{T \in Q} t_i}$	$\frac{1}{n} \sum_i \frac{d_v(X^i, Y^i)}{v(l^i) - v(o^i)}$	+	+	$\begin{smallmatrix} + \\ [101] \end{smallmatrix}$	+

\*  $S(0, 0) = S(1, 1) = 1$ ;  $D(0, 0) = D(1, 1) = 0$ .\*\*  $X \wedge Y \neq 0$ .\*\*\*  $D(0^i, 0^i) = 0$ .

Во многих случаях очень нелегко правильно определить приоритет в «изобретении» мер близости. Так, широко известные коэффициенты ранговой корреляции Кендалла и Спирмена впервые были предложены другими авторами, хотя оба упомянутых математика, безусловно, внесли решающий вклад в изучение и распространение этих показателей связи. Метрика несходства  $D_4$  (дополнение до единицы к хорошо известному в таксонометрии коэффициенту общности П. Жаккара  $S_5$ ), по-видимому, впервые была использована Э. Галантером [98], а затем И. Снитом [148] и изучалась в работе Е. Марчевского и Г. Штейнгауза [128]; более поздние упоминания о ней встречаются у Ч. Райского [136], П. Има [115] и Ю. Д. Апресяна [3]. Между тем в отечественной литературе эта мера близости иногда именуется «метрикой Таннмото» со ссылкой на работу [140], в которой в действительности была предложена квазиметрика  $D_{10}$  — логарифмическое преобразование коэффициента общности  $S_5$ . Из-за отсутствия достаточно полных и легко доступных публикаций по проблемам измерения близости расстояние  $D_4$  неоднократно «открывалось» различными исследователями, включая автора этой статьи.

Расстояние  $D_1$ , напротив, настолько общеизвестно, что установить приоритет в его изобретении невозможно. Эту меру близости называют иногда псевдометрикой Фреше—Никодима в пространстве  $\mathcal{P}$ , метрикой Гливенко [99] — на решетках, в функциональном пространстве и в пространстве  $R^n$  — расстоянием Минковского  $l_1$  [130], метрикой city—block или «Манхэттенским расстоянием» [45, с. 11], в пространстве булевых векторов — метрикой Хемминга [106]. Кроме того, метрика  $D_1$  именуется расстоянием Кемени [121] или Миркина—Черного [37], если речь идет соответственно о пространствах бинарных отношений линейного порядка или эквивалентности.

В большинстве случаев меры близости из табл. 1 могут быть представлены в виде функций от  $v(X)$ ,  $v(Y)$ ,  $v(X \wedge Y)$  или  $v(X \vee Y)$ . Это вполне понятно: близость объектов  $X$  и  $Y$  определяется наличием у них общих признаков, информативная ценность которых ( $v(X \cap Y)$ ) должна быть неким образом соотнесена с информацией об объектах, рассматриваемых порознь ( $v(X)$ ,  $v(Y)$ ) или в совокупности ( $v(X \cup Y)$ ).

Меры сходства  $S_1$  —  $S_3$ ,  $S_{12}$  могут быть представлены одним выражением

$$S(X, Y) = \frac{v(1) + \alpha d_v(X, Y)}{v(1) + \beta d_v(X, Y)};$$

варьируя параметры  $\alpha$  и  $\beta$ , можно получить тот или иной показатель сходства. Аналогичным образом выражение

$$\tilde{S}(X, Y) = \frac{v(X \wedge Y)}{\gamma v(X \wedge Y) + \delta v(X \vee Y)}$$

определяет  $S_5$ ,  $S_7$ ,  $S_8$ . Значения коэффициентов  $\gamma$ ,  $\delta$  выбираются в соответствии с тем, что исследователь считает наиболее важным

для определения близости — наличие совпадающих или, напротив, несовпадающих признаков.

Сходство объектов в принципе может быть измерено не только по наличию, но и по отсутствию у них одних и тех же признаков — «сходство по антипатиям» [44, с. 144]. Понятно, что меры близости, учитывающие лишь информацию о несовпадающих признаках ( $S_1 - S_3, S_{12}, D_1 - D_3, D_9, D_{10}$ ), инвариантны или, по определению Г. Бока [78], симметричны относительно выбора любой из этих альтернатив. Прочим же мерам сходства и несходства можно поставить в соответствие их «двойника», т. е.  $S(\bar{X}, \bar{Y})$  или  $D(\bar{X}, \bar{Y})$ ; пример тому — меры близости  $S_5$  и  $S_6, D_4$  и  $D_5$ . Можно также попытаться одновременно учесть как наличие, так и отсутствие одинаковых признаков ( $S_3, D_6$ ).

Нормировка мер близости  $S_1, S_4, D_1, D_2$  и  $D_9$  связана с ограничением сверху в  $S_1$  и  $D_1$ . Что же касается показателей сходства  $S_3, S_5 - S_{10}$  и несходства  $D_4 - D_8$ , то здесь процедура нормирования содержательно оправдана: информация о сходстве (различии) соотносится тем самым со всей имеющейся информацией об объектах. Заметим, что  $S_8 - S_{10}$  можно представить в виде

$$S(X, Y) = \frac{v(X \wedge Y)}{M[v(X), v(Y)]},$$

где  $M(A, B)$  — некоторое среднее величин  $A$  и  $B$ . В данном случае речь идет соответственно о среднем арифметическом, среднем геометрическом и среднем гармоническом.

Меры близости  $S_{11}, S_{12}$  и  $S_{13}$  приведены в табл. 1 в качестве наглядной иллюстрации к рассуждениям о связи между коэффициентами корреляции и мерами сходства. В пространстве матриц инцидентности отношений линейного порядка показатель  $S_{12}$ , очевидным образом связанный с мерами близости  $S_1$  и  $D_1$ , есть не что иное, как коэффициент ранговой корреляции Кендалла [65]. Коэффициент корреляции  $S_{11}$ , введенный в практику анализа данных Ф. Гальтоном и Э. Пирсоном, также часто применяется в качестве меры сходства булевых векторов. Если оценивается связь между двумя ранжировками  $X$  и  $Y$ , причем  $X = (R_X(1), \dots, R_X(n)) \in \mathbb{R}_+^n$ , где  $R_X(i)$  — ранг  $i$ -го объекта в ранжировке  $X$ , то  $S_{11}$  переходит в коэффициент ранговой корреляции Спирмена, связанный линейной зависимостью с мерой близости Кендалла  $D_9$  [122].

Обобщением понятия связи (корреляции) величин, измеренных в шкале не ниже порядковой, для множеств  $X, Y \in \mathcal{P}(E)$  является понятие соответствия, которое может быть полным или частичным. Необходимое условие полного соответствия — равномощность сравниваемых множеств. Полное соответствие естественно считать максимальным при  $X = Y$  и минимальным при  $X = \bar{Y}$ . Коэффициент  $S_{13}$  был предложен в работе [52] в качестве меры соответствия обычных множеств или, что то же самое, булевых векторов [53, § 4.4]. Интересно отметить, что при  $|X| = |Y|$  коэффициенты корреляции ( $S_{11}$ ) и соответствия ( $S_{13}$ ) совпадают,

в остальных случаях всегда  $|S_{11}| < |S_{13}|$ . Наконец, при  $|X| = |Y| = 1/2 |E|$   $S_{11} \equiv S_{12} \equiv S_{13}$ . Следовательно, измерение соответствия матриц инцидентности линейного порядка равносильно измерению ранговой корреляции по Кендаллу.

В табл. 1. не приведены многие «узкоспециализированные» меры близости. Расстояния в пространстве разбиений, предложенные различными авторами, чаще всего представимы в виде  $d_v(X, Y)$ , нормированных мер близости  $D_4, D_5$  или  $D_7$ . Функция  $d_v(X, Y)$  обладает свойствами псевдометрики, если  $v$  — положительная изотонная верхняя или нижняя оценка [81]. Аналогичный результат для частично упорядоченных множеств более общего вида см. в [131]. Перечень некоторых оценок, используемых при построении метрик на решетке разбиений, можно найти в работах [71, 73, 81]. Теорема 1 дает возможность конструировать произвольные псевдометрики типа  $d_v(X, Y)$ .

## Маргинальные и ассоциативные меры близости

Применение большинства мер близости из табл. 1 основывается на предположении если не равноценности, то, по крайней мере, соизмеримости информационного вклада различных признаков в измеряемое сходство (несходство) объектов  $X, Y$ . Однако отнюдь не представляется нереальной такая ситуация, когда множество  $E$  разложимо на  $k \leq n$  групп признаков существенно различной природы, например, измеренных в различных шкалах:

$$E = \bigcup_{i=1}^k E^i, \quad E^i \cap E^j = \emptyset \quad \text{при } i \neq j.$$

Тогда существует  $k$  частных расстояний между  $X$  и  $Y$ , каждое из которых может быть определено с помощью неких частных (маргинальных) мер близости [49; 53]. Все  $k$  маргинальных расстояний можно считать измеренными в одной и той же шкале — шкале отношений [40], так что в принципе можно попытаться оценить среднее расстояние (конечно, если это не лишено содержательного смысла). Будем искать его в форме ассоциативного среднего по Колмогорову:

$$\rho(X, Y) = F^{-1} \left\{ \sum_{i=1}^k \alpha_i F[\rho^i(X, Y)] \right\},$$

где  $\rho^i(X, Y) \doteq \rho(X^i, Y^i)$  — маргинальная мера близости;  $X^i = X \cap E^i$ ;  $\sum_i \alpha_i = 1$ .

А. И. Орлов [41] показал, что в шкале отношений допустимыми (в том смысле, в каком этот термин употребляется в теории измерений) средними по Колмогорову являются степенные средние ( $F(t) = t^p, p > 0$ ). В соответствии с этим результатом в работе

[49] было введено понятие ассоциативной меры близости порядка  $p$ :

$$\rho_A(X, Y) = \left\{ \sum_{i=1}^k \alpha_i [\rho^i(X, Y)]^p \right\}^{1/p}. \quad (4)$$

Здесь  $\rho^i$  могут в общем случае принадлежать различным классам мер близости.

Метрика  $D_{11}$  — ассоциативная  $D$ -метрика порядка 1; в пространстве  $\mathcal{P}$  она при  $k = n$  переходит в расстояние  $D_1$ , а в более общих случаях  $(\mathbb{R}_+^n, L_+)$  представляет собой метрику, инвариантную относительно масштаба измерения разнотипных признаков, предложенную в [53]. Меры близости  $D_1, D_2, D_8 - D_{10}, D_{12}, D_{13}$ , впрочем, также могут рассматриваться в качестве ассоциативных. Идея измерения близости в форме среднего арифметического неких частных близостей высказывалась также в работах Ю. А. Воронина [8], Б. П. Кононова [19] и других авторов.

## Допустимые и инвариантные меры близости

Анализ табл. 1 приводит к заключению, что совокупность известных мер близости не упорядочена и, возможно, избыточна, хотя и неясно, в какой именно степени. Попытки ограничения многообразия мер близости предпринимались неоднократно. Подход к решению этой задачи, основанный на традициях теории измерений, связан с требованием инвариантности меры близости относительно допустимых преобразований исходных данных. Ю. Н. Толстова [59, 61], обсуждая проблему корректности функции расстояния, приводит следующие варианты возможных ограничений на выбор меры близости:

а)  $\rho(X, Y) < \rho(X', Y')$  тогда и только тогда, когда

$$\rho(\varphi(X), \varphi(Y)) < \rho(\varphi(X'), \varphi(Y')); \quad (5)$$

$$б) \rho(X, Y) = \rho(\varphi(X), \varphi(Y)), \quad (6)$$

где  $\varphi(X)$  — результат применения допустимого преобразования к измерению  $X$ . Второе требование представляется более сильным, хотя при определенных условиях они эквивалентны [9, 59]. В. С. Высоцкий [9] предложил использовать понятия простой и сильной инвариантности; условия (5) — (6) совпадают с определением простой инвариантности по Высоцкому, если  $X$  и  $Y$  измерены в одной и той же шкале. В обзоре [63] в качестве адекватных мер близости предлагается использовать те и только те, что удовлетворяют условию (6). В табл. 2 перечислены некоторые меры близости, являющиеся адекватными именно в этом смысле по отношению к результатам измерений в наиболее часто употребляемых шкалах. Ясно, однако, что если измерения в номинальной или в

**Таблица 2**  
**Адекватные меры близости**

Тип шкалы	Группа допустимых преобразований $\Phi = \{\varphi\}$	Допустимые меры близости из табл. 1	Тип пространства $Q$
Номинальная	Взаимнооднозначные отображения	$S_1 - S_{13},$ $D_1 - D_{13}$	Пространство матриц инцидентности бинарных отношений $\mathcal{B} \subset \mathcal{P}$
Порядковая	Строго возрастающие непрерывные отображения	$S_1 - S_{13},$ $D_1 - D_{13}$	
Разностей	Преобразование сдвига $\varphi(t) = t + b; b > 0$	$S_1 - S_3, S_{11}, S_{12}$ $D_1 - D_3, D_6, D_9, D_{13}$	$\mathbb{R}_+^n, L_+$
Отношений	Преобразования подобия $\varphi(t) = at; a > 0$	$S_5, S_7 - S_{11},$ $D_4, D_7, D_9, D_{10} - D_{13}$	$\mathbb{R}_+^n$ $L_+$
Интервальная	Линейные преобразования $\varphi(t) = at + b; a > 0,$ $b > 0$	$S_{11}, D_{13}$	

порядковой шкале представлены в виде элементов пространства бинарных отношений, погруженного в  $\mathcal{P}$  (вектором  $(x_1, \dots, x_n)$ ) при этом оказывается матрица инцидентности отношения на множестве мощности  $N = \sqrt{n}$ , то все меры близости из табл. 1 адекватны. Проблема выбора меры близости в этом случае остается нерешенной.

### Аксиоматический подход к измерению близости: обзор основных результатов

Более эффективным инструментом выбора меры близости оказался так называемый аксиоматический метод, восходящий к работе Дж. Кемени [121]<sup>8</sup>. Автору удалось однозначно определить меру близости, удовлетворяющую нескольким требованиям (аксиомам), которые представляются вполне естественными в контексте изучавшейся им задачи отыскания результирующего упорядочения  $N$  объектов  $m$  экспертами. Такой мерой близости оказалась метрика  $d_p(X, Y)$  в пространстве матриц инцидентности соответствующих бинарных отношений.

В 1970 г. Б. Г. Миркин и Л. Б. Черный [37] предложили аксиоматику, близкую к системе аксиом Кемени, с тем чтобы обосновать выбор расстояния  $D_1$  для измерения близости между разбие-

<sup>8</sup> На русском языке этот результат был опубликован в 1972 г. в известной монографии Дж. Кемени и Дж. Снелла «Кибернетическое моделирование» (М., 1972).

ниями конечного множества. Проблема метризации решетки разбиений обсуждалась и ранее [81]. Расстояние  $D_1$  в пространстве матриц инцидентности отношений эквивалентности предлагалось, в частности, С. Ренье [138], однако в работе [37] эта мера близости впервые была определена с помощью ряда содержательных требований, включая понятие «находиться между» (betweenness property)<sup>9</sup>. Авторы распространили этот результат на пространство отношений квазипорядка [38], а в более поздних работах Б. Г. Миркина была предложена менее «жесткая» аксиоматическая конструкция, позволяющая вводить в пространстве разбиений метрики и квазиметрики, согласованные с понятием «между» [34, 36].

К. Богарт [79, 80] обосновал применение расстояния  $D_1$  для согласования отношений частичного порядка и предложил систему аксиом, однозначно определяющую евклидову метрику в пространстве асимметричных бинарных отношений. В дальнейшем эти результаты были обобщены на случай взвешенных (метризованных) отношений частичного порядка [157].

Во второй половине 70-х годов в нашей стране было опубликовано большое количество работ, в которых аксиоматически и, как правило, однозначно вводилась метрика  $D_1$  для измерения близости в пространстве толерантностей [39], взвешенных ранжирований [23], нечетких множеств [4] и отношений [24], векторов взвешенных инцидентий [48], метризованных отношений [27, 29], сверхтранзитивных графов [47], в морфологическом пространстве [69] и пр. Согласно нашим обозначениям, это соответствует метризация различных подпространств, вложенных в  $\mathcal{P}$ ,  $\mathcal{P}_F$ ,  $\mathbb{R}^n$  или  $\mathbb{R}^n$ . Аксиоматика Кемени была обобщена Б. Г. Миркиным [35] и Б. Г. Литваком [28] на случай произвольных бинарных отношений. А. И. Орлов [40] предложил простую систему аксиом, определяющую (с точностью до меры  $\nu$  на  $\sigma$ -алгебре с единицей  $E$ ) меру близости  $D_1$  — псевдометрику в пространстве  $\mathcal{P}$ .

Система аксиом, определяющие в качестве единственной меры близости евклидову метрику, предлагались, помимо уже упоминавшихся работ К. Богарта и Дж. Уикса [79, 80, 157], Г. М. Гамбаровым с соавторами [40] и Б. Г. Литваком [29]. Важно упомянуть и об аксиоматическом введении мер сходства Ю. А. Ворониним [8] и В. П. Козыревым [18].

Для измерения близости в пространствах, наделенных сложной внутренней структурой, предлагаются так называемые структурные меры близости [14, 28, 29, 73, 76, 81, 96, 131, 146]. Общая идея этого подхода применительно к задаче измерения близости между разбиениями конечного множества объектов подробно изложена в работе С. Бормана и Ф. Арабье [81]. Вводится понятие минималь-

<sup>9</sup> Элемент  $Y$  решетки  $\mathcal{L}$  находится между ее элементами  $X$  и  $Z$  (обозначение:  $[X, Y, Z]$ ) в том случае, если  $X \wedge Z \leq Y \leq X \vee Z$  [6]. Если же исходить из чисто «геометрических» представлений, то  $[X, Y, Z]$  можно определить как равенство в  $D_4$ . Для метрики  $D_1$  эти определения эквивалентны [6].

ной трансформации разбиения (например, перемещение элемента из одного класса разбиения в другой); два разбиения, которые могут быть получены одно из другого путем минимальной трансформации, называются соседними. Устанавливается соответствие между пространством разбиений и неориентированным графом, любые две вершины которого соединены ребром тогда и только тогда, когда соответствующие им разбиения — соседние. Расстояние между произвольными вершинами определяется как длина кратчайшего пути между ними; очевидно, такая мера близости есть метрика.

В работе [81] рассматриваются различные определения минимальной трансформации разбиения и порождаемые ими метрики [62]. Многие из них сводятся к мерам близости вида  $D_1$  и  $D_4$ , причем функция  $\nu$  может быть оценкой, верхней или нижней в зависимости от определения соседства на графе. Для других мер близости аналитического выражения не найдено, хотя способ их вычисления можно указать (о вычислительных проблемах измерения структурной близости см. [92]).

Структурные меры близости на частично упорядоченных множествах в пространствах бинарных отношений аксиоматически вводились Ж. П. Бартелеми [73], Б. Г. Литваком [28, 29], Т. А. Казанской [14] и другими авторами. Получены результаты, устанавливающие связь между кратчайшими путями на графах и псевдометриками семейства  $D_1$ , индуцируемыми верхними и нижними оценками на частично упорядоченных множествах [73, 74, 76, 129, 131].

«Геометрический» подход к измерению близости, однако, не всегда приводит к наилучшим результатам, ибо во многих случаях измерять следует не расстояние, а сходство (или несходство) между объектами.

Попытки использования расстояний  $D_1$  или  $D_2$ , равно как и многих других мер близости, в задачах автоматической классификации объектов, описываемых номинальными признаками, по ряду причин не представляются достаточно обоснованными [50; 62]. В 1981 г. автор этой статьи предложил систему аксиом, включающую условие (3) в качестве экспликации понятия несходства и определяющую  $D$ -метрику (метрику несходства) в пространствах  $\mathcal{P}$  и  $\mathcal{P}_F$  (с обобщением для  $\mathbb{R}_+^n$  и  $L_+$ ) [49]. В силу (1) — (2) это эквивалентно аксиоматическому введению коэффициента общности  $S_5$  для измерения сходства между объектами, описываемыми информативными признаками  $e_i \in E$ .

Определение близости как сходства является одной из интенсивно обсуждаемых проблем математической психологии и психометрии [87, 95, 102, 109, 123, 134, 139, 144, 147, 161]. В 1977 г. А. Тверски [155] предложил оригинальную систему аксиом, формализующих характерные черты сходства как психофизиологического понятия. В основе подхода автора лежит следующее утверждение. Сходство пары объектов обуславливается наличием у каждого из них некоторых характерных признаков или свойств (features),

причем величина сходства зависит от наличия как совпадающих, так и несовпадающих признаков. Пять аксиом А. Тверски определяют меру сходства объектов  $X$  и  $Y$  как функцию вида

$$S(X, Y) = \theta f(X \cap Y) - \alpha f(X \setminus Y) - \beta f(Y \setminus X), \quad (7)$$

где  $X \setminus Y$  — теоретико-множественная разность множеств признаков, соответствующих  $X$  и  $Y$ ;  $f$  — неотрицательная аддитивная функция множеств;  $\theta, \alpha, \beta \geq 0$ . Нетрудно показать, что модели сходства (7) соответствует модель несходства

$$D(X, Y) = \gamma f(X \setminus Y) + \delta f(Y \setminus X) - \eta f(X \cap Y),$$

откуда, учитывая аддитивность функции  $f$ , получаем

$$D(X, Y) = \gamma f(X) + \delta f(Y) - \varepsilon f(X \cap Y). \quad (8)$$

Меры близости (7) и (8) в общем случае не удовлетворяют всей совокупности требований  $S1-S3$  либо  $D1-D3$ . В то же время надлежащим выбором параметров  $\alpha, \beta, \theta, \gamma, \delta, \varepsilon$  можно получить некоторые показатели сходства или несходства из табл. 1. В этой же статье А. Тверски упоминает меру сходства вида

$$S(X, Y) = \frac{f(X \cap Y)}{f(X \cap Y) + \alpha f(X \setminus Y) + \beta f(Y \setminus X)}, \quad (9)$$

представляющую собой интересное обобщение мер близости  $S_5, S_7, S_8$ .

Работа [155] вызвала оживленную дискуссию о моделировании сходства в психологии [109, 123, 134], подробное освещение которой увело бы нас в сторону от рассматриваемой проблемы измерения близости в задачах анализа социологической информации. Отметим, впрочем, одну из последних работ А. Тверски и Дж. Гати [156], в которой на основании некоторых экспериментальных данных подвергается сомнению аддитивность ассоциативного расстояния относительно его маргинальных компонент вида (8) в том случае, когда последние отвечают существенно различным, несопоставимым признакам.

Аксиоматический метод, вообще говоря, может быть полезен лишь постольку, поскольку используемые аксиомы эксплицируют желательные и содержательно значимые свойства меры близости в реальных условиях конкретной задачи. Важно отметить, что аксиоматическая характеристика метрики  $D_1$ , предложенная Кемени, не является решением чисто формальной математической задачи, но естественным образом вытекает из существа самой проблемы согласования ранжировок. Далеко не обо всех исследованиях, выполненных вслед за этой работой и по аналогии с ней, можно сказать то же самое.

Особое положение занимают работы французских математиков Ж.-П. Бартелеми, Ж. Борде и Б. Монжарде [73—75, 82,

131], в которых аксиоматический подход к метризации частично упорядоченных множеств рассматривается с сугубо математической точки зрения, что позволяет достигнуть высокой степени обобщения получаемых результатов. О большом интересе, который вызывает проблема метризации частично упорядоченных множеств, можно судить уже по тому, что ее обсуждению был полностью посвящен один из выпусков журнала «Mathematiques et Sciences Humaines» [129].

## Аксиоматическая характеристика семейства мер близости

В большинстве упомянутых выше работ аксиоматический метод был использован для однозначного определения той или иной меры близости. Между тем представляется полезной разработка аксиоматической конструкции, эксплицирующей достаточно общие и естественные свойства, которыми обладает не единственная функция, но некий класс  $\mathcal{P}$  мер близости. Введением дополнительных условий, связанных с содержательным смыслом конкретной задачи анализа нечисловой информации, можно было бы определить искомую меру близости  $\rho \in \mathcal{P}$ . В работе [51] подобный подход был использован для характеристики класса псевдометрик в пространстве  $\mathcal{P}_F$ . Ниже мы приведем обобщение этих результатов на случай ограниченной дистрибутивной решетки  $\mathcal{L}$  с элементами  $X, Y, \dots$ , максимальным элементом 1 и минимальным элементом 0.

По аналогии с расстоянием на числовой прямой<sup>10</sup> потребуем, чтобы

$$A\ 1. \ \rho(X, Y) = \rho(X \vee Y, X \wedge Y) \geq 0.$$

Отсюда следует  $\rho(X, Y) = \rho(Y, X)$ .

Следующая аксиома, как и первая, отталкивается от простых геометрических представлений. Если  $X$  и  $Y$  лежат «позади»  $Z$  (т. е.  $X \leq Z, Y \leq Z$ , откуда  $X \vee Y \leq Z, X \wedge Y \leq Z$ ), то

$$A\ 2. \ \rho(X, Z) + \rho(Y, Z) = \rho(X \wedge Y, Z) + \rho(X \vee Y, Z).$$

Пусть, далее, для всех  $X, Y, Z$ , таких, что  $X \leq Y \leq Z$  ( $X \neq Z$ ), выполнено условие, которое мы назовем условием усиленной монотонности:

$$A\ 3. \ \frac{\rho(Z, Y)}{\rho(Z, X)} = \frac{v(Z) - v(Y)}{v(Z) - v(X)},$$

где  $v: \mathcal{L} \rightarrow \mathbb{R}$  — строго изотонная функция.

Требование монотонности в обычном смысле слова можно было бы сформулировать так:  $0 \leq \frac{\rho(Z, Y)}{\rho(Z, X)} \leq 1$ . Условие A 3 отличается

<sup>10</sup> Очевидно, в любом полностью упорядоченном множестве  $Q = \{a, b, \dots\}$  выполняется равенство  $\rho(a, b) = \rho(a \vee b, a \wedge b)$ .

от него прежде всего тем, что в явном виде вводит изотонную функцию  $v$ ; выбором этой функции мы, как это следует из приведенной ниже теоремы, определяем некое семейство мер близости. Заметим также, что из А 3 следует  $\rho(X, X) = 0$  и  $\rho(X, Y) > 0$  при  $X > Y$ , что, принимая во внимание А 1, дает  $\rho(X, Y) > 0$  при  $X \neq Y$ .

**Теорема 2.** 1) Меры близости, определяемые аксиоматикой А 1—А 3, удовлетворяют соотношению

$$\rho(X, Y) = \rho(X \vee Y, 0) \frac{d_v(X, Y)}{v(X \vee Y)}, \quad X \vee Y \neq 0, \quad (10)$$

причем функция  $v$  — положительная оценка на  $Z$ . 2) Для того чтобы мера близости вида (10) была метрикой, необходимо и достаточно, чтобы функция  $\psi(X) = \frac{v(X)}{\rho(X, 0)}$ ,  $X \neq 0$  была изотонной:

$$X \leq Y \Leftrightarrow \frac{v(X)}{\rho(X, 0)} \leq \frac{v(Y)}{\rho(Y, 0)}. \quad (11)$$

Теорема 2 (точнее, ее вторая часть) развивает результаты, полученные ранее Э. Галантером [98], Е. Марчевским и Г. Штейнгаузом [128], Ж. Борде [82] и автором [50]. Класс мер близости, удовлетворяющих соотношению (10), обозначим  $\mathbb{P}$  (класс  $\mathbb{P}$  не пуст: легко убедиться в том, что он содержит, в частности, расстояния  $D_1$  и  $D_4$ ). Меры близости  $\rho \in \mathbb{P}$  определены с точностью до расстояния  $\rho(X, 0)$  и, очевидно, достигают максимального значения при  $Y = \bar{X}$ , где  $\bar{X} = v^{-1}[v(1) - v(X)]$  (конечно, если такой элемент существует; для решеток  $\mathcal{P}$  и  $\mathcal{P}_F$  эта оговорка будет излишней), причем  $\max \rho(X, Y) = \rho(1, 0)$ . Не ограничивая общности, положим  $\rho(1, 0) \stackrel{\mathcal{P}}{=} 1$ .

**Следствие.** Итак, метрики  $D_1$  и  $D_4$  принадлежат классу  $\mathbb{P}$  и, следовательно, в известном смысле «близки» друг к другу. Различие же между ними, как будет далее показано, заключается в различии исходного определения сходимости. В целях большей наглядности перейдем от произвольной решетки  $L$  к пространству множеств  $\mathcal{P}$ . Аксиомы А 1—А 3 накладывают на функции  $\rho \in \mathbb{P}$  ограничения, свойственные расстоянию вообще и не зависящие от нашего понимания сходимости объектов  $X, Y, \dots$ . Если считать информативным только наличие признака  $e_i$  у объекта  $X$  ( $\chi_X(e_i) = 1$ ), то естественно будет принять условие (3), полагая од инаково несходными друг с другом любые объекты, не имеющие об щих признаков. Тогда  $\rho(X, \emptyset) = 1$  при  $X \neq \emptyset$  и  $\rho(X, Y) = \frac{d_v(X, Y)}{v(X \cup Y)} = D_4$ .

Может, впрочем, возникнуть и иная ситуация. Для чисто дихотомических номинальных признаков  $e_i$  отсутствие признака так же информативно, как и его наличие (примеры: пол, экстравертность — интравертность респондента и пр.). Здесь безразлично, что именно считать наличием признака  $e_i$  у объекта  $X$  ( $\chi_X(e_i) = 1$ ), а что — отсутствием ( $\chi_X(e_i) = 0$ ). Следовательно

но, мера близости должна быть инвариантной относительно замены признака его «отрицанием», иначе говоря,  $\rho(X, Y) = \rho(\bar{X}, \bar{Y})$ . Нетрудно показать, что подобное требование выделяет в классе  $\mathcal{P}$  семейство расстояний вида  $D_1$ .

Интересно отметить, что, исходя из соображений, продиктованных существом рассматриваемых нами конкретных задач измерения близости, мы в качестве мер близости получили метрики  $D_1$  и  $D_4$ , хотя выполнение условий  $D_4$  или (11) заранее не требовалось.

Классу мер близости  $\mathcal{P}$  соответствует двойственный ему класс  $\bar{\mathcal{P}}$  (аксиоматика, порождающая  $\bar{\mathcal{P}}$ , двойственна аксиомам  $A_1 - A_3$ ). Расстояние  $d_v(X, Y)$ , будучи двойственно-инвариантным, принадлежит обоим классам  $\mathcal{P}$  и  $\bar{\mathcal{P}}$ . «Двойником»  $D$ -метрики в классе  $\mathcal{P}$  является уже знакомое нам расстояние  $D_6$  ( $\bar{D}$ -метрика):

$$\bar{D}(X, Y) = D_4(\bar{X}, \bar{Y}) = \frac{d_v(X, Y)}{v(1) - v(X \wedge Y)}; \quad X \wedge Y \neq 1.$$

Более подробное изучение свойств метрики несходства  $D_4$  и ее обобщения для пространства  $L_+$  (метрики подобия) см. в [50]. Отметим, что неравенство треугольника для  $D$ -метрики вырождается в равенство тогда и только тогда, когда  $Y = X \vee Z$  (для  $\bar{D}$ -метрики соответственно при  $Y = X \wedge Z$ ).

## Близость мер близости

Интересные возможности открываются при изучении сходства самих мер близости. Если при использовании каких-либо мер близости в работе некоего алгоритма анализа данных всегда получаются достаточно близкие или совпадающие результаты, то такие меры близости разумно считать сходными или даже эквивалентными.

Наиболее простой способ сравнения функций близости  $\rho'$  и  $\rho''$  — это непосредственное сопоставление полученных с их помощью матриц близости  $\|\rho'_{ij}\|$ ,  $\|\rho''_{ij}\|$  между объектами  $X_1, \dots, X_N$ , принадлежащими некоторой совокупности  $O$ . Исходная совокупность объектов выбирается из определенных содержательных соображений (например, данные специального тестирования [103, 155, 156]) или генерируется на ЭВМ [54, 55, 71, 88, 142, 143]. В качестве оценки сходства  $\rho'$  и  $\rho''$  используется расстояние Минковского [119]

$$l_p(\rho', \rho'') = \left( \sum_{1 \leq i < j \leq N} |\rho'_{ij} - \rho''_{ij}|^p \right)^{1/p},$$

коэффициент корреляции [25, 71, 88, 151] или какой-нибудь другой показатель связи [36, 114, 142]. Поскольку обычное, «количественное» измерение близости производится в шкале отношений, для сопоставления  $\rho'$  и  $\rho''$  можно (см. табл. 2) использо-

вать коэффициент корреляции  $S_{11}$ ,  $D$ -метрику или связанные с ней меры  $S_5$  и  $S_{13}$  и т. д., расстоянием же Минковского  $l_p$ , напротив, пользоваться не следует.

Заключение об эквивалентности двух мер близости либо о наличии стохастической связи между ними может быть сформулировано в рамках проверки соответствующей гипотезы с помощью статистик, основанных на матрицах  $\|\rho'_{ij}\|$ ,  $\|\rho''_{ij}\|$  [112, 114]. Для сопоставления мер близости используются также методы многомерного шкалирования [71, 88].

Не вдаваясь в детальное изложение результатов, полученных различными авторами, можно констатировать, что многие меры сходства, связи, расстояния и пр. часто порождают весьма «похожие» (например, сильно коррелированные) матрицы близости. В ряде ситуаций можно говорить о наличии нескольких кластеров мер близости [55, 71, 142].

Между тем во многих прикладных задачах — в первую очередь, в социологических и биомедицинских исследованиях — разумно считать значения меры близости не «числовыми», а порядковыми [30, с. 84]. Это связано как с неточностью, «засоренностью» исходной информации, так и со специфическими характеристиками применяемых алгоритмов анализа данных. Н. Жарден [118] считает, что в большинстве случаев меры сходства и несходства, рассчитанные по описанию объектов некоторыми признаками, дают не количественную, а в лучшем случае порядковую информацию о близости. С другой стороны, для упорядочения  $N$  объектов достаточно порядковой информации об их близости [110]. То же самое верно и для обширного класса задач и алгоритмов многомерного шкалирования, кластер-анализа и др.

Если считать, что близость объектов должна или может измеряться в порядковой шкале, то равноценными следует считать любые меры близости, монотонно связанные друг с другом:  $\rho'$  и  $\rho''$  эквивалентны [78], если для любых  $X, Y, Z, T \in \mathcal{L}$

$$\rho'(X, Y) \leq \rho'(Z, T) \Leftrightarrow \rho''(X, Y) \leq \rho''(Z, T). \quad (12)$$

Меры близости, удовлетворяющие условию (12), иногда называют комонотонными [54, 55] или порядковоэквивалентными [57]. Попарно комонотонны, например, меры близости  $S_1, S_2, S_3, S_{12}, D_1$  и  $D_9$  из табл. 1 и многие другие показатели сходства, связи и т. д.

Наряду с исследованием необходимых и достаточных условий взаимной монотонности мер близости [54, 55, 57, 93, 126] можно попытаться выделить классы ситуаций, в которых функции, в общем случае не удовлетворяющие (12), приводят к порядковоэквивалентным матрицам близости<sup>11</sup>. Ниже мы рассмотрим результат именно такого типа, однако вначале нам понадобится ввести определение.

<sup>11</sup> Матрицы  $\|\rho'_{ij}\|$  и  $\|\rho''_{ij}\|$  мы будем называть порядковоэквивалентными, если ранжировки их элементов совпадают.

**Определение 4.** Функция  $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$  называется условно монотонной по аргументу  $x_i$ , если она монотонна по этому аргументу при фиксированных  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ .

Выше мы отмечали, что практически все меры близости на произвольной решетке  $\mathcal{L}$  представимы в виде функций от  $v(X), v(Y), v(X \wedge Y)$  (напомним, что для оценки  $v - v(X \vee Y) = v(X) + v(Y) - v(X \wedge Y)$ ). Таковы, в частности, меры схождения и несхождения  $S_1 - S_{10}, S_{12}, D_1, D_4 - D_7, D_{11}$  из табл. 1, все меры близости вида (7)–(9) и — при весьма незначительных требованиях к функции  $\rho(X, 0)$  — любые меры близости из класса  $\mathcal{P}$ . Заметим также, что все упомянутые функции имеют вид

$$\rho(X, Y) = f[v(X \wedge Y), v(X), v(Y)], \quad (13)$$

где  $f: \mathbb{R}^3 \rightarrow \mathbb{R}$  — функция, условно монотонная по первому аргументу. По типу условной монотонности функции  $f$  (возрастающая или убывающая) меры близости вида (13) делятся на два класса эквивалентности (меры схождения и меры несхождения).

**Лемма.** Пусть для любых  $X, Y, \dots$ , принадлежащих некоторому набору  $\mathcal{L}'$  элементов решетки  $\mathcal{L}$ , выполнено условие

$$v(X) \equiv \text{const}, \quad (14)$$

где  $v$  — оценка на  $\mathcal{L}$ . Тогда любые меры схождения вида (13) эквивалентны на  $\mathcal{L}'$ , т. е. для любой пары  $\rho', \rho''$  выполнено (12). Аналогичное утверждение верно и для показателей несхождения вида (13).

В реальных задачах анализа нечисловой информации условия леммы выполняются не так уж редко. В математической социологии и в теории группового выбора довольно часто возникает необходимость агрегации или аппроксимации турниров [75] (полных асимметричных бинарных отношений), матрицы инцидентности которых содержат фиксированное число ненулевых элементов. Для отношения линейного порядка, например, эта величина составляет  $1/2N(N-1)$ , где  $N$  — число упорядочиваемых объектов. Матрицы парных сравнений также могут рассматриваться в этом контексте. Другой пример: обработка анкет социологического опроса с вопросами закрытого типа и с фиксированным числом ответов, выбираемых респондентом. Так называемая экспертная классификация представляет собой частный случай подобного опроса, если классы описаны заранее. Наконец, в таких экспертных процедурах, как метод вероятностных оценок и метод решающих матриц [53, подразд. 1.3], эксперту предлагается распределить фиксированное число шансов или заданную сумму весов, что равносильно требованию (14). Удобным обобщением этой ситуации оказывается пространство функций плотности вероятностей  $g, h, \dots$ , вложенное в пространство  $L_+(E)$ . Выбрав в качестве оценки  $v$  интеграл по области определения, получаем  $v(g) = \int_E g dv \equiv 1$ .

С учетом леммы становится очевидным, что в любой из рассмотренных выше ситуаций целесообразно пользоваться наиболее

простыми мерами близости вида (13), а именно показателями  $S_1$ ,  $S_4$  или  $D_1$ .

Возвращаясь к более общему случаю, отметим, что отклонение от комонотонности может служить мерой близости функций  $\rho'$  и  $\rho''$ . Согласно лемме, с этой целью разумно будет использовать расстояние  $D_1$  — нормированное число пар  $(X, Y)$ ,  $(Z, T)$ , таких, что

$$|\rho'(X, Y) - \rho'(Z, T)| |\rho''(X, Y) - \rho''(Z, T)| < 0,$$

[55, 93, 126] либо дополнение этой меры до единицы, т. е.  $S_1$  [57].

## Анализ близости в практических приложениях: основные проблемы и возможные пути их решения

При обработке социологической информации измерение близости не всегда выделяется в отдельную, явно обозначенную процедуру. Так, находя медиану или иное среднее совокупности наблюдений  $X_1, \dots, X_m$ , мы тем самым автоматически вводим некую меру близости. Нетрудно показать, что практически любые средние, используемые при обработке реальных данных, могут быть определены как решение оптимизационной задачи

$$\sum_{i=1}^m \rho(X_i, X) \rightarrow \text{opt}, \quad X \in Q \quad (15)$$

где  $\rho$  — мера близости, а максимум либо минимум суммы близостей (15) ищется на некотором множестве  $Q$  допустимых значений переменной  $X$ . В частности, в пространстве  $\mathbb{R}$  (обычные «числовые» данные) выбор медианы в качестве среднего равносильен выбору расстояния  $D_1$  или соответственно показателей  $S_1 - S_3, S_{12}$ ; положив  $\rho = D_2$ , приходим к среднему арифметическому и т. д. Таким образом, теоретико-измерительные проблемы адекватности средних и мер близости тесно связаны [40, 42, 60].

Минимизация суммы расстояний в (15) для агрегации нечисловых данных предлагалась многими исследователями [32, 40, 75]. Для социологических приложений весьма полезным оказывается тот факт, что на решетке бинарных отношений  $\mathcal{L}_R$  решение задачи (15) при  $\rho = D_1$  принадлежит интервалу

$$[\bigvee_{A \subset \mathcal{L}_R} (\bigwedge_{i \in A} X_i), \bigvee_{B \subset \mathcal{L}_R} (\bigwedge_{i \in B} X_i)],$$

где  $A, B$  — сочетания, содержащие соответственно  $|A| = \left\lceil \frac{m+2}{2} \right\rceil$  и  $|B| = \left\lfloor \frac{m+1}{2} \right\rfloor$  элементов  $\mathcal{L}_R$ . Указанный интервал иногда определяют как медиану, а поиск бинарного отношения, среднего в смысле (15) относительно меры близости  $D_1$ , как процедуру (поиска) медианы (median procedure) [75]. Применение этой процедуры для согласования индивидуальных предпочтений в задачах социального выбора представляется наиболее обоснованным [159].

В различных методах классификации и структуризации данных обычно используются попарные близости объектов  $X_1, \dots, X_m$ , суммы близостей и иные функционалы, основанные на матрице близостей [2, 12, 17, 36, 40, 45, 60, 78]. В алгоритмах кластер-анализа, предложенных С. Ренье [138] и Б. Г. Миркин [32, 33], оптимальное разбиение ищется как решение задачи (15) относительно расстояния  $D_i$  на решетке разбиений исходной совокупности объектов.

Активное использование как самих эмпирических близостей, так и функций от них выдвигает ряд серьезных требований к выбору мер близости, адекватности алгоритмов анализа близости и обоснованности последующей интерпретации результатов. Некоторые из них рассматривались нами выше, исчерпывающее же обсуждение всех возникающих при этом вопросов вряд ли вообще возможно. Остановимся все же на одной из наиболее актуальных проблем.

Хорошо известно, что реальные данные далеко не всегда соответствуют той несложной модели, которую мы рассматривали выше (170). Так, множество признаков  $E$  зачастую избыточно вследствие понятного опасения исследователя упустить некие факторы, относительную важность которых трудно определить заранее. С другой стороны, некоторые существенные признаки могут быть все же пропущены. Сказанное справедливо не только для множества признаков, но отчасти и для множества рассматриваемых объектов. Наконец, в большинстве случаев доступные и включенные в рассмотрение признаки разнотипны — это и номинальные, и порядковые, и количественные переменные. Указанные обстоятельства подчеркивают стохастический характер исходных данных и приводят к выводу о необходимости статистической оценки измеряемого сходства, корреляции, расстояния и пр.

Для коэффициентов обычной и ранговой корреляции имеется развитое табличное обеспечение, позволяющее проверять гипотезу об отсутствии связи ( $H_0: r = 0$ , где  $r$  — некий коэффициент корреляции) как при большом, так и при малом объеме выборки [7, 16]. Статистический анализ близости связан с существенно большими сложностями, прежде всего из-за затруднений с самой формулировкой нуль-гипотезы. В пространстве  $\mathcal{P}$  можно предложить следующий подход. Поскольку набор номинальных признаков  $E$  выбирается априори, наблюдаемое совпадение их у объектов  $X, Y$  может быть обусловлено случайными причинами. Нуль-гипотеза состоит в том, что распределение признаков на каждом из объектов — равномерное, а проверять ее можно, например, по отклонению величины мощности пересечения  $|X \cap Y|$  от математического ожидания (при  $H_0$ ). Тогда, если мощность каждого из множеств  $X, Y$  фиксирована, величина  $|X \cap Y|$  имеет гипергеометрическое распределение с параметрами  $|X|, |Y|, n$ . Значимость величины  $|X \cap Y|$  или, что то же самое, отклонения от нуля величины  $|X \cap Y| - (1/n) |X| |Y| = \frac{1}{n} (ad - bc)$  при заданном уровне доверительной вероятности может быть проверена с помощью соответствующих таблиц [7, табл. 5.6].

Серьезное затруднение состоит, однако, в том, что величины  $|X|$  и  $|Y|$  далеко не всегда фиксированы, в особенности в задачах кластер-анализа. Рассматривая аналогичную ситуацию при проверке  $H_0: ad - bc = 0$  для таблиц сопряженности  $2 \times 2$ , М. Кендалл и А. Стьюарт рекомендуют пользоваться критерием, основанным на гипергеометрическом распределении, и в тех случаях, когда истинные значения параметров распределения неизвестны [16, подразд. 33.24]. Показано, что данный критерий является равномерно наиболее мощным несмещенным как при известных, так и при неизвестных параметрах  $|X|$ ,  $|Y|$ , что дает основание использовать его для проверки значимости близости  $X, Y \in \mathcal{P}$ , определенной с помощью показателей вида (13) (например,  $S_1 - S_{13}$ ,  $D_1, D_2, D_5 - D_7, D_9, D_{10}$ ) [52, 53].

При проверке значимости совокупной близости более чем двух объектов пользоваться рассмотренной выше схемой практически невозможно: число попарных близостей  $m$  объектов растет пропорционально  $0,5 m(m-1)$ , соответственно меняется и уровень значимости при проверке  $H_0$ . Приемлемой мерой совокупной бли-

зости оказываются показатели типа  $\frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \rho(X_i, X_j)$ : коэффициенты конкордации  $W$ , согласия  $U$  (для порядковой шкалы) и согласованности  $V$  (для булевых векторов). Коэффициент согласованности  $V$  предложен в [52, 53, подразд. 4.4] и определяется как отношение

$$V = \frac{\sum_{j=1}^n \left( \sum_{i=1}^m \chi_{X_i}(e_j) \right)^2 - \frac{1}{n} \left( \sum_{i=1}^m |X_i| \right)^2}{\sum_{r=1}^k \left\{ \left[ \sum_{s=r}^k N(s) \right]^2 (|X|^{(s)} - |X|^{(s-1)}) \right\} - \frac{1}{n} \left( \sum_{i=1}^m |X_i| \right)^2},$$

где  $|X|^{(0)} = 0$ ,  $|X|^{(1)} < |X|^{(2)} < \dots < |X|^{(k)}$  — мощности множеств  $X_i$ , занумерованные в порядке возрастания,  $k \leq m$ ,  $N(s)$  — число множеств  $X_i$  мощности  $s$ . При  $|X_i| \equiv \text{const}$

$$V = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m S_{13}(X_i, X_j).$$

Проверяется нуль-гипотеза о равномерном распределении генеральной совокупности объектов соответственно на множестве всех ранжировок ( $W, U$ ) или подмножеств  $E(V)$ . Подобную процедуру можно использовать, например, для проверки реальности кластера, полученного с помощью некоего алгоритма классификации [51].

В более общем случае (пространства  $L_+$  и  $L$ ) укажем на обнадеживающие результаты, полученные с помощью развиваемых А. И. Орловым методов статистики объектов нечисловой

природы. К настоящему моменту исследованы асимптотические свойства расстояний  $D_1$  и  $D_4$ , найдены несмещенные оценки параметров асимптотического распределения вектора попарных расстояний между функциями  $g, h, \dots, \in L$ , что позволяет строить соответствующие алгоритмы проверки согласованности, находить оценки для диаметра кластера и пр. Преимущественное внимание, уделяемое метрикам  $D_1$  и  $D_4$ , объясняется наличием содержательно интерпретируемого аксиоматического обоснования для выбора этих мер близости (см. А1 — А3 и следствие из теоремы 2).

В случае разнотипных данных ( $k$  групп данных) целесообразно применять набор из  $k$  маргинальных мер близости, каждая из которых инвариантна по отношению к допустимым в данной шкале преобразованиям. Если веса в формуле (4) не представляется возможным назначить априори, разумно будет перейти к ранжированным матрицам близости. В зависимости от результатов проверки конкордации  $k$  ранжировок попарных близостей следует либо построить среднюю в смысле (15) ранжировку ( $H_0$  отклонена), либо провести кластеринг ранжировок ( $H_0$  не отклонена) и привлечь дополнительную информацию для окончательного решения.

Рассматривая проблему измерения близости, мы убедились в том, что она не имеет простого и однозначного решения для всего многообразия задач анализа социологических данных. В зависимости от существа поставленной задачи, характера и объема доступной информации и т. д. исследователь должен самостоятельно и последовательно проанализировать как теоретико-измерительные, так и содержательные аспекты этой проблемы в своем конкретном случае. Окончательный выбор может и не свестись к какой-либо одной мере близости или к одному-единственному алгоритму анализа близостей. Однако разработанные к настоящему времени теоретические основы измерения близости и большой практический опыт, накопленный исследователями в различных областях науки (и прежде всего в социологии), позволяют существенно сузить область поиска и принять обоснованное решение.

### Литература

1. Аванесов В. С. Тесты в социологическом исследовании. М., 1982.
2. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. М., 1974.
3. Апресян Ю. Д. Алгоритм построения классов на матрице расстояний. — В кн.: Машинный перевод и прикладная лингвистика. М., 1966, вып. 9, с. 3—18.
4. Батыршин И. З. О некоторых свойствах мер невероятностной энтропии размытых множеств. — В кн.: Прикладной многомерный статистический анализ. М., 1978, с. 345—348.
5. Батыршин И. З. Кластеризация на основе размытых отношений сходства. — В кн.: Тез. III науч.-техн. семинара «Управление при наличии расплывчатых категорий». Пермь, 1980, с. 25—27.
6. Биркгоф Г. Теория структур. М., 1984.
7. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., 1983.

8. Воронин Ю. А. Введение мер сходства и связи для решения геолого-географических задач. — ДАН СССР, 1971, т. 199, № 5, с. 1011—1014.
9. Висоцкий В. С. Инвариантность коэффициентов связи. — В кн.: Прикладной многомерный статистический анализ. М., 1978, с. 348—351.
10. Гамбаров Г. М., Мандель И. Д., Рыбина И. А. О некоторых метриках, возникающих в задачах обработки данных. — Автоматика и телемеханика, 1980, № 12, с. 116—123.
11. Гретцер Г. Общая теория решеток. М., 1982.
12. Дюран Б., Одед П. Кластерный анализ. М., 1977.
13. Ефремович В. А. Инфинитезимальные пространства. — ДАН СССР, 1951, т. 76, № 3, с. 341—343.
14. Казанская Т. А. Измерение сходства произвольного бинарного отношения и бинарных отношений с заданными свойствами. — В кн.: Анализ нечисловых данных в системных исследованиях. М., 1982, с. 131—139.
15. Келли Дж. Л. Общая топология. М., 1981.
16. Кендалл М., Стьюарт А. Статистические выводы и связи. М., 1973.
17. Клигер С. А., Косолапов М. С., Толстова Ю. Н. Шкалирование при сборе и анализе социологической информации. М., 1978.
18. Козырев В. П. Аксиоматическое определение меры близости матриц некоторого класса. — В кн.: Исследования по дискретной математике. М., 1972, с. 200—213.
19. Кононов В. П. Аспекты способы сравнения объектов. — НТИ. Сер. 2, 1980, № 7, с. 24—29.
20. Коулмен Д. С. Системы социального обмена. — В кн.: Математика в социологии: Моделирование и обработка информации. М., 1977, с. 417—442.
21. Кофман А. Введение в теорию нечетких множеств. М., 1982.
22. Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер-анализом. — В кн.: Классификация и кластер. М., 1980, с. 20—41.
23. Кузьмин В. Б., Овчинников С. В. Априорные модели предпочтений. — В кн.: Информационное обеспечение в задачах управления научными подразделениями НИИ. М., 1976, с. 51—54.
24. Кузьмин В. Б., Овчинников С. В. Построение групповых решений в пространстве нечетких бинарных отношений. — В кн.: Экспертные оценки. М., 1979, с. 91—109. (Вопросы кибернетики; Вып. 58).
25. Куперштох В. Л., Полищук Л. И., Трофимов В. А. Коэффициенты корреляции связей между объектами. — В кн.: Модели агрегирования социально-экономической информации. Новосибирск, 1978, с. 17—34.
26. Куперштох В. Л., Трофимов В. А. Алгоритм анализа структуры матрицы связей. — Автоматика и телемеханика, 1975, № 11, с. 170—180.
27. Литвак Б. Г. Меры близости на метризованных отношениях. — В кн.: Прикладной многомерный статистический анализ. М., 1978, с. 78—93.
28. Литвак Б. Г. Меры близости и анализ нечисловой информации. — В кн.: Экспертные оценки. М., 1979, с. 73—86. (Вопросы кибернетики; Вып. 58).
29. Литвак Б. Г. Экспертная информация: Методы получения и анализа. М., 1982.
30. Мамула Д. В. Методы теории графов в алгоритмах кластер-анализа. — В кн.: Классификация и кластер. М., 1980, с. 83—111.
31. Меллер Ф., Капекки В. Роль энтропии в номинальной классификации. — В кн.: Математика в социологии: Моделирование и обработка информации. М., 1977, с. 301—338.
32. Миркин Б. Г. Проблема группового выбора. М., 1974.
33. Миркин Б. Г. Дискретные задачи классификации взаимосвязанных объектов: (Обзор). — В кн.: Вопросы анализа сложных систем. Новосибирск, 1974, с. 66—77.
34. Миркин Б. Г. Модели качественного анализа социально-экономической информации. — В кн.: Математика в социологии: Моделирование и обработка информации. М., 1977, с. 339—352.
35. Миркин Б. Г. Аксиоматическая характеристика расстояния в совокупном пространстве номинальных и ранговых признаков. — В кн.: Моделирование в экономических исследованиях. Новосибирск, 1978, с. 95—108.

36. Миркин Б. Г. Анализ качественных признаков и структур. М., 1980.
37. Миркин Б. Г., Черный Л. Б. Об измерении близости между разбиениями конечного множества объектов. — Автоматика и телемеханика, 1970, № 5, с. 120—127.
38. Миркин Б. Г., Черный Л. Б. Некоторые свойства пространства разбиений. — В кн.: Математический анализ экономических моделей. Новосибирск, 1972, ч. 3, с. 126—147.
39. Орлов А. И. Проблемы устойчивости и обоснованности решений в теории экспертных оценок. — В кн.: Статистические методы анализа экспертных оценок. М., 1977, с. 7—30. (Учен. зап. по статистике; Т. 29).
40. Орлов А. И. Устойчивость в социально-экономических моделях. М., 1979.
41. Орлов А. И. Связь между средними величинами и допустимыми преобразованиями шкалы. — Мат. заметки, 1981, т. 30, № 4, с. 561—568.
42. Орлов А. И. Статистика объектов нечисловой природы и обработка социологических данных. — В кн.: Математические методы в социологическом исследовании. М., 1981, с. 67—75.
43. Паниотто В. И. Об одном направлении в моделировании динамики структуры малых групп. — В кн.: Математические методы в социологическом исследовании. М., 1981, с. 294—300.
44. Парыцкий Г. Основы статистической информатики. М., 1981.
45. Райзин Дж. Вэн. Классификация и кластер. М., 1980.
46. Рао С. Р. Кластер-анализ в применении к изучению перемешивания рас в популяциях людей. — В кн.: Классификация и кластер. М., 1980, с. 148—167.
47. Рапопорт А. М. Измерение расстояний между сверхтранзитивными графами. — В кн.: Экспертные методы в системных исследованиях. М., 1979, с. 55—61.
48. Рапопорт А. М., Шнейдерман М. В. Анализ экспертных суждений, заданных в виде структур. — В кн.: Прикладной многомерный статистический анализ. М., 1978, с. 150—164.
49. Раушенбах Г. В. Об измерении близости между конечными множествами. — Вестн. АМН СССР, с. 27. Рук. деп. в ВИНТИ, N 4250—81.
50. Раушенбах Г. В. Меры близости в пространстве множеств. — В кн.: Алгоритмы анализа данных социально-экономических исследований. Новосибирск, 1982, с. 29—43.
51. Раушенбах Г. В. Об одной процедуре статистического анализа результатов автоматической классификации. — В кн.: Применение математических методов обработки и медико-биологических данных и ЭВМ в медицинской технике. М., 1984, с. 138—139.
52. Раушенбах Г. В. Статистический анализ результатов экспертной классификации. — В кн.: Всесоюз. конф. «Нечисловая статистика, экспертные оценки и смежные вопросы»: Тез. докл. М., 1984, с. 103—104.
53. Раушенбах Г. В., Филиппов О. В. Экспертные оценки в медицине. М., 1983.
54. Ростовцев П. С. К анализу сходства мер связи между качественными признаками. — В кн.: Проблемы анализа дискретной информации. — Новосибирск, 1976, ч. II, с. 85—91.
55. Ростовцев П. С. Некоторые инструменты анализа социологической информации. — В кн.: Методы анализа многомерной экономической информации. Новосибирск, 1981, с. 108—147.
56. Рыкина Р. В., Герчиков В. И., Новожилов Ю. В., Суховский М. П., Тапилина В. С. Опыт изучения сельского образа жизни на основе его типологии. — В кн.: Применение факторного и классификационного анализа для типологизации социальных явлений. Новосибирск, 1976, с. 69—117.
57. Сатаров Г. А. Сравнение двух алгоритмов шкалирования дихотомических данных. — В кн.: Математические методы в социологическом исследовании. М., 1981, с. 90—98.
58. Терехина А. Ю. Многомерный анализ субъективных данных о сходствах или различиях. М., 1978.

59. Типология и классификация в социологических исследованиях. М., 1982.
60. Типология потребления. М., 1978.
61. Толстова Ю. Н. Сопоставимость результатов классификации при использовании различных шкал.— Социол. исслед., 1978, № 3, с. 178—184.
62. Тюрин Ю. Н. Экспертная классификация.— В кн.: Экспертные оценки в системных исследованиях. М., 1979, с. 5—15.
63. Тюрин Ю. Н., Литвак Б. Г., Орлов А. И., Сатаров Г. А., Шмерлинг Д. С. Анализ нечисловой информации. М., 1981, с. 80.
64. Харват Ф., Кучера Я. К теории социальной зависимости.— В кн.: Математика в социологии: Моделирование и обработка информации. М., 1977, с. 170—200.
65. Черный Л. Б. О связи метода пространства разбиений с известными методами анализа данных.— В кн.: Вопросы анализа сложных систем. Новосибирск, 1974, с. 84—89.
66. Черный Л. Б. Об оценке близости объектов в терминах взаимосвязей признаков.— В кн.: Проблемы анализа дискретной информации. Новосибирск, 1975, с. 167—175.
67. Черный Л. Б. Обобщенные группировки.— В кн.: Модели агрегирования социально-экономической информации. Новосибирск, 1978, с. 139—151.
68. Шрейдер Ю. А. Равенство, сходство, порядок. М., 1971, с. 256.
69. Якимец В. Н. Исследование морфологического пространства вариантов системы.— В кн.: Модели и методы формирования и многокритериального выбора предпочтительных вариантов систем. М., 1981, с. 6—23.
70. Arabie P. Clustering representations of group overlap.— J. Math. Sociol., 1977, vol. 5, p. 113—128.
71. Arabie P., Boorman S. A. Multidimensional scaling of measures of distance between partitions.— J. Math. Psychol., 1973, vol. 10, p. 148—203.
72. Balakrishnan V., Sanghvi L. D. Distance between population on the basis of attribute data.— Biometrics, 1968, vol. 24, p. 859—865.
73. Barthelemy J. P. Proprietes metriques des ensembles ordonnes: Comparision et agregation des relations binaires: These d'Etat/Univ. d'Franche-Comte. Besancon, 1979. 169 p.
74. Barthelemy J. P. Caracterisations axiomatiques de la difference metrique entre des relations binaires.— Math. et sci. hum., 1979, N 67, p. 85—113.
75. Barthelemy J. P., Monjardet B. The median procedure in cluster analysis and social choice theory.— Math. Soc. Sci., 1981, vol. 1, p. 235—267.
76. Barthelemy J. P., Flament Cl., Monjardet B. Ordered sets and social sciences: Proc. of the symp. on ordered sets. 1982. 37 p. Prepr.
77. Bartholomew D. J. Stochastic models for social processes. N. Y.: Wiley, 1973.
78. Bock H. H. Automatische Klassifikation. Göttingen: Vandenhoeck and Ruprecht, 1974. 480 S.
79. Bogart K. P. Preference structures I: Distances between transitive preference relations.— J. Math. Sociol., 1973, vol. 3, p. 49—67.
80. Bogart K. P. Preference structures II: Distances between asymmetric relations.— SIAM J. Appl. Math., 1975, vol. 29, p. 254—262.
81. Boorman S. A., Arabie P. Structural measures and the method of sorting.— In: Multidimensional scaling. N. Y.: Seminar press, 1972, vol. 1, p. 225—249.
82. Bordes G. Metriques bornees definies par des valuations sur un demi-treillis.— Math. et sci. hum., 1976, N 56. p. 89—95.
83. Boyd J. P. Information distances for discrete structures.— In: Multidimensional scaling. N. Y.: Seminar press, 1972, vol. 1, p. 213—223.
84. Bray J. R., Curtis J. T. An ordination of the upland forest communities of S. Wisconsin.— Ecol. Monogr., 1957, vol. 27, p. 325—349.
85. Brumelle S. L., Gerchak J. The structure of indices of social mobility and inheritance.— J. Math. Sociol., 1982, vol. 8, p. 251—264.
86. Cormack R. M. A review of classification.— J. Roy. Statist. Soc. A, 1971, vol. 134, p. 321—353.
87. Cunningham J. P. Free trees and bidirectional trees as representations of

- psychological distance.— *J. Math. Psychol.*, 1978, vol. 17, p. 165—188.
88. *Cunningham J. P., Shepard R. N.* Monotone mappings of similarities into a general metric space.— *J. Math. Psychol.*, 1974, vol. 11, p. 355 — 363.
  89. *Czekanowski J.* Zarys metod statystycznych w zastosowaniu do antropologii. W-wa, 1913.
  90. *Czekanowski J.* «Coefficient of racial likeness» und «durchschnittliche Differenz».— *Anthropol. Anz.*, 1932, Bd. 9, S. 227—249.
  91. *Dale A. G., Dale N.* Some clumping experiments for associative document retrieval.— *Amer. Document.*, 1965, vol. 16, p. 5—9.
  92. *Day W. H. E.* The complexity of computing metric distances between partitions.— *Math. Soc. Sci.*, 1981, vol. 1, p. 269—287.
  93. *De La Vega W. F.* Techniques de classification automatique utilisant un indice de ressemblance.— *Rev. franç. sociol.*, 1967, vol. 8, p. 506 — 520.
  94. *Dice L. R.* Measures of the amount of ecological association between species.— *Ecology*, 1945, vol. 26, p. 297—302.
  95. *Eisler H., Ekman G.* A mechanism of subjective similarity.— *Acta psychol.*, 1959, vol. 16, p. 1—10.
  96. *Flament Cl.* Applications of graph theory to group structure. N. Y.: Prentice Hall, 1963.
  97. *Frechet M. R.* Sur quelques points de calcul fonctionnelle. P., 1906.
  98. *Galanter E. H.* An axiomatic and experimental study of sensory order and measure.— *Psychol. Rev.*, 1956, vol. 63, p. 16—28.
  99. *Glivenko V.* Theorie des structures. P.: Gauthier-Villars, 1938.
  100. *Goodall D. W.* Objective methods for classification of vegetation. I. The use of positive interspecific correlation.— *Austral. J. Bot.*, 1953, vol. 1, p. 39—63.
  101. *Gower J. C.* A general coefficient of similarity and some of its properties.— *Biometrics*, 1971, vol. 27, p. 857—872.
  102. *Gregson R. A. M.* Psychometric of similarity. N. Y.: Acad. Press, 1975.
  103. *Gregson R. A. M.* A comparative evaluation of seven similarity models.— *Brit. J. Math. and Statist. Psychol.*, 1976, vol. 29, p. 139—156.
  104. *Greiner R.* Über das Fehlersystem der Kollektivmasslehre.— *Ztschr. Math. und Phys.*, 1909, Bd. 57, S. 121, 225, 337.
  105. *Hamman U.* Merkmalbestand und Verwandtschaftsbeziehungen der Farinosae: Ein Beitrag zum System der Monokotyledonen.— *Willdenovia*, 1961, Bd. 2, S. 639—678.
  106. *Hamming R. M.* Error detecting and error correcting codes.— *Bell Syst. Techn. J.*, 1950, vol. 29, p. 147—160.
  107. *Hartigan J. A.* Representation of similarity matrices by trees.— *J. Amer. Statist. Assoc.*, 1967, vol. 62, p. 1140—1158.
  108. *Heincke F.* Naturgeschichte des Herings. I. Die Lokalformen und die Wanderungen des Herings in den europäischen Meeren.— *Abh. Dt. Seefischerei-Verein*, 1898, Bd. 2, i-CXXXVI, S. 1—223.
  109. *Holman E. W.* Monotonic models for asymmetric proximities.— *J. Math. Psychol.*, 1979, vol. 20, p. 1—15.
  110. *Hubert L. J.* Some applications of graph theory and related non-metric techniques to problems of approximate seriation: The case of symmetric proximity measures.— *Brit. J. Math. and Statist. Psychol.*, 1974, vol. 27, p. 133—153.
  111. *Hubert L. J.* Seriation using asymmetric proximity measures.— *Brit. J. Math. and Statist. Psychol.*, 1976, vol. 29, p. 32—53.
  112. *Hubert L. J.* Generalized proximity function comparisons.— *Brit. J. Math. and Statist. Psychol.*, 1978, vol. 31, p. 179—192.
  113. *Hubert L. J.* Analyzing proximity matrices: The assesment of internal variation in combinatorial structure.— *J. Math. Psychol.*, 1980, vol. 21, p. 247—267.
  114. *Hubert L. J., Golledge R. G.* A heuristic method for the comparison of related structures.— *J. Math. Psychol.*, 1981, vol. 23, p. 214—226.

115. *Ihm P.* Automatic classification in anthropology.— In: The use of computers in anthropology. Hague: Mouton and Co, 1965, p. 357—376.
116. *Jaccard P.* Gesetze der Pflanzenverteilung der alpinen Region.— *Flora*, 1902, Bd. 90, H. 3, S. 349—377.
117. *Jaccard P.* The distribution of the flora in the alpine zone.— *New Phytol.*, 1912, vol. 11, p. 37.
118. *Jardine N.* Discussion on Dr. Cormack's paper.— *J. Roy. Statist. Soc. A*, 1971, vol. 134, p. 356—357.
119. *Jardine C. J., Jardine N., Sibson R.* The structure and construction of taxonomic hierarchies.— *Math. Biosci.*, 1967, vol. 1, p. 173—179.
120. *Kahl J. A., Davis J. A.* A comparison of indexes of socioeconomic status.— *Amer. Sociol. Rev.*, 1955, vol. 20, p. 315—320.
121. *Kemeny J. G.* Mathematics without numbers.— *Daedalus*, 1959, vol. 88, p. 577—591.
122. *Kendall M. G.* Discrimination and classification.— In: *Multivariate analysis*. N. Y.: Acad. press, 1966, vol. 1, p. 165—185.
123. *Krumhansl C. L.* Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density.— *Psychol. Rev.*, 1978, vol. 85, p. 445—463.
124. *Kulezynski S.* Die Pflanzenassoziationen der Pinien.— *Bull. Intern. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat.*, 1927, B, p. 57—203.
125. *Lance G. N., Williams W. T.* Computer programs for hierarchical polythetic classification.— *Comput. J.*, 1966, vol. 9, p. 60—64.
126. *Lerman J. C.* Les bases de la classification automatique. P.: Gauthier-Villars, 1970.
127. *Leung Y.* A fuzzy set analysis of sociometric structure.— *J. Math. Sociol.*, 1980, vol. 7, p. 159—180.
128. *Marczewski E., Steinhaus H.* On a certain distance of sets and the corresponding distance of functions.— *Colloq. math.*, 1958, vol. 6, p. 319—327.
129. *Metriques et relations* (special issue).— *Math. et sci. hum.*, 1979, N 67.
130. *Minkowski H.* Gesammelte Anhandlungen. B.: Teubner, 1911. Bd. II.
131. *Monjardet B.* Metrics on partially ordered sets — a survey.— *Discrete Math.*, 1981, vol. 35, p. 173—184.
132. *Needham R. M.* A method of using computers in information classification.— In: *Information processing*. Amsterdam: North-Holland, 1962, p. 284—287.
133. *Ochiai A.* Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions.— *Bull. Jap. Soc. Sci. Fish.*, 1957, vol. 22, p. 526—530.
134. *Ortony A.* Beyond literal similarity.— *Psychol. Rev.*, 1979, vol. 86, N 3, p. 161—181.
135. *Penrose L. S.* Distance, size and shape.— *Ann. Eugenics*, 1954, vol. 18, p. 337—343.
136. *Rajski C.* A metric space of discrete probability distributions.— *Inform. and Contr.*, 1961, vol. 4, p. 371—377.
137. *Rao C. R.* The utilization of multiple measurements in problems of biological classification.— *J. Roy. Statist. Soc. B*, 1948, vol. 10, p. 159—193.
138. *Regnier S.* Sur quelques aspects mathématiques des problèmes de classification automatique. Rome, 1965. (ICC Bull.; N 4).
139. *Restle F.* A metric and an ordering on sets.— *Psychometrika*, 1959, vol. 24, p. 207—220.
140. *Rogers D. J., Tanimoto T. T.* A computer program for classifying plants.— *Science*, 1960, vol. 132, N 3434, p. 1115—1118.
141. *Russel R. F., Rao T. R.* On habitat and association of species of anophe-line larvae in South-Eastern Madras.— *J. Malor. Inst. India*, 1940, vol. 3, p. 153—178.
142. *Särndal C. E.* A comparative study of association measures.— *Psychometrika*, 1974, vol. 39, p. 165—187.
143. *Särndal C. E.* A Monte Carlo study of some asymmetric association measures.— *Brit. J. Math. and Statist. Psychol.*, 1976, vol. 29, p. 94—102.
144. *Sattath S., Tverski A.* Additive similarity trees.— *Psychometrika*, 1977, vol. 42, p. 319—345.

145. *Shepard R. N., Arabie P.* Additive clustering: Representation of similarities as combinations of discrete overlapping properties.— *Psychol. Rev.*, 1979, vol. 86, p. 87—123.
146. *Shipstone E. I.* Some variables affecting pattern conception.— *Psychol. Monogr.*, 1960, vol. 74, p. 1—40.
147. *Smith T. E., Yu W.* A prominence theory of content-sensitive choice behaviour.— *J. Math. Sociol.*, 1982, vol. 8, p. 225—249.
148. *Sheath P. H. A.* The application of computers to taxonomy.— *J. Gen. Microbiol.*, 1957, vol. 17, p. 201—226.
149. *Sokal R. R.* Distance as a measure of taxonomic similarity.— *Syst. Zool.*, 1961, vol. 10, p. 70—79.
150. *Sokal R. R., Michener C. D.* A statistical method for evaluating systematic relationships.— *Univ. Kans. Sci. Bull.*, 1958, Mar., N 20, p. 1409—1438.
151. *Sokal R. R., Rohlf F. J.* The comparison of dendrograms by objective methods.— *Taxon*, 1962, vol. 11, p. 33—40.
152. *Sokal R. R., Sneath P. H. A.* Principles of numerical taxonomy. (San Francisco: W. H. Freeman and Co, 1963.
153. *Sørensen T.* A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons.— *Biol. skr.*, 1948, vol. 5, p. 1—34.
154. *Theil H.* Statistical decomposition analysis. Amsterdam: North-Holland, 1972.
155. *Tversky A.* Features of similarity.— *Psychol. Rev.*, 1977, vol. 84, p. 327—352.
156. *Tversky A., Gati J.* Similarity, separability and the Triangle inequality.— *Psychol. Rev.*, 1982, vol. 89, p. 123—154.
157. *Weeks J. R., Bogart K. P.* Consensus signed digraphs.— *SIAM J. Appl. Math.*, 1979, vol. 36, p. 1—14.
158. *Winship C.* A distance model for sociometric structure.— *J. Math. Sociol.*, 1977, vol. 5, p. 21—39.
159. *Young H. P., Levenglick A.* A consistent extension of Condorset's election principle.— *SIAM J. Appl. Math.*, 1978, vol. 35, p. 285—300.
160. *Zadeh L. A.* Similarity relations and fuzzy orderings.— *Inform. Sci.*, 1971, vol. 3, p. 177—200.
161. *Zubin T.* A technique for measuring like-mindedness.— *J. Abnorm. Soc. Psychol.*, 1938, vol. 33, p. 508—516.

**П. С. РОСТОВЦЕВ**

## АЛГОРИТМЫ АНАЛИЗА СТРУКТУРЫ ПРЯМОУГОЛЬНЫХ МАТРИЦ «ПЯТНА» И «ПОЛОСЫ»

Предлагаемые автором методы позволяют отыскивать блоки матриц исходных данных, которые состоят из «похожих» друг на друга элементов. Анализируемые матрицы могут иметь достаточно произвольный вид (матрицы разного рода коэффициентов связи; матрицы типа объект — признак для признаков, измеренных по произвольным шкалам и т. д.). Это обуславливает возможность решения с помощью предлагаемых в статье методов довольно широкого круга задач (осуществление предварительного анализа структуры исходных данных с целью выдвижения содержательных гипотез; группировку как объектов, так и признаков; выделение информативных систем признаков и т. д.). Предложенные алгоритмы, в частности, позволяют решать задачи, подобные задачам, решаемым с помощью детерминационного анализа. Практически описанные в статье задачи можно сформулировать как задачи группировки.

Данные, представляющие собой прямоугольные числовые матрицы (таблицы), встречаются в самых разнообразных исследованиях в социологии, геологии, медицине и т. д. Нередко такие таблицы сложны и труднообозримы. В этом случае для их исследования необходимо привлекать ЭВМ.

В данной работе как средство автоматического анализа рассматриваются два алгоритма поиска блочной структуры прямоугольных матриц. Интерпретация структуры разбиения ведется преимущественно для матриц связи двух систем признаков  $X = \{x_1, \dots, x_m\}$  и  $Y = \{y_1, \dots, y_n\}$ , хотя не исключена и другая природа данных. Подобные задачи анализа структуры матриц с целью одновременной кластеризации множеств и строк и столбцов рассматривались в работах [9, 1], с целью выделенных существенных подсистем признаков [5], задачи анализа структуры квадратных матриц — в работе [8].

### Качество аппроксимации при заданном числе блоков

Пусть задана матрица  $A = \|a_{xy}\|$ ,  $x \in X$ ,  $y \in Y$ . Подмножество множества  $X \times Y$  будем называть блоком матрицы. Разбивкой матрицы назовем множество непересекающихся блоков

$$R = \{R_1, \dots, R_s\},$$

объединение которых совпадает с  $X \times Y$ .

Матрицу  $B = \|b_{xy}\|$ ,  $x \in X$ ,  $y \in Y$ , будем называть матрицей блочной структуры, если ее элементы постоянны на блоках:

$$b_{xy} = \lambda_t, \text{ если } (x, y) \in R_t.$$

Автоматический анализ структуры матрицы в соответствии с аппроксимационным подходом, изложенным в работе [6], состоит в поиске матрицы блочной структуры  $B$ , близкой к  $A$ , при условии, что ее разбиение принадлежит к некоторому классу разбиений  $E$ .

Близость матриц определяется функционалом

$$\Phi(A, R, \lambda) = (A - B, A - B),$$

где  $\lambda = (\lambda_1, \dots, \lambda_s)$ , операция  $(\dots)$  означает операцию скалярного произведения матриц, рассматриваемых как векторы в  $m \times n$ -мерном пространстве.

При фиксированном разбиении  $R$  минимум функционала  $\Phi$  по  $\lambda$  несложно найти:

$$\min_{\lambda} \Phi(A, R, \lambda) = (A, A) - \sum_{t=1}^s \left( \sum_{(x, y) \in R_t} a_{xy} \right)^2 / |R_t|,$$

где  $|\cdot|$  — мощность множества. В силу этого минимизация  $\Phi$  эквивалентна максимизации

$$f_1(R) = \sum_{t=1}^s \left( \sum_{(x,y) \in R_t} a_{xy} \right)^2 / |R_t|.$$

Для дальнейшего иногда удобнее использовать показатель, эквивалентный  $f_1$ :

$$f_1(R) = (1/\sigma^2) \sum_{t=1}^s \left( \sum_{(x,y) \in R_t} (a_{xy} - \bar{a}) \right)^2 / |R_t|,$$

$\sigma$  — среднее квадратическое отклонение элементов матрицы,  $\bar{a}$  — их средняя величина.

## Сложность разбиения

Увеличивая сложность разбиения  $s$ , в конечное число шагов можно добиться оптимальной величины целевого показателя  $\Phi(A, R, \lambda) = 0$  независимо от качества данных и алгоритма. Такая ситуация возникает, в частности, при  $s = m \times n$ . Такое разбиение описывает каждый элемент матрицы и практической ценности не имеет. Это означает, что оптимизация  $\Phi$  на широком классе разбиений  $E$ , включающем разбиения разной сложности, вообще говоря, не имеет смысла и следует рассматривать классы разбиений фиксированной сложности  $E_s$ , либо найти критерий выбора  $s$ .

Один из путей выбора  $s$  состоит в содержательном изучении и интерпретации разбиений разной сложности. Формальный путь выбора  $s$ , реализованный в алгоритмах, заключается в сравнении результатов агрегирования с результатами случайного агрегирования случайных данных, которое можно понимать как агрегирование в самой плохой ситуации<sup>1</sup>. Случайными данными будем считать матрицу, каждый элемент которой порожден независимо и случайно, а случайным алгоритмом — тот, который случайно выбирает разбиение. Результатом работы такого алгоритма является случайная величина целевого показателя  $v_s = f(R_s)$ . При слабых ограничениях на распределения элементов матрицы и характер ее разбиения величина  $v_s$  асимптотически по  $m \times n \rightarrow \infty$  распределена как  $\chi^2_{s-1}$  [7, 9].

Реальные результаты целесообразно сравнивать со средними случайными результатами. Для выбора сложности  $s$  последовательно получают разбиения  $R^s$  сложности  $s = 1, 2, \dots$  до

<sup>1</sup> Подробнее данный подход рассмотрен в работе [7]. При жестких условиях на алгоритм Хартигану удалось получить критерий остановки, использующий только «случайность» данных [9].

тех пор, пока приращение  $f(R^{s+1}) - f(R^s)$  не достигнет среднего случайного приращения, т. е. до тех пор, пока алгоритм работает «не хуже „случайного“». Несложно видеть, что случайное приращение равно  $M(\chi_s^2 - \chi_{s-1}^2) = 1$ .

### Алгоритм «пятна»

Этот алгоритм получает блоки в виде прямоугольников  $R_t = X_t \times Y_t$ ,  $X_t \in X$  и  $Y_t \in Y$ ,  $t = 1, \dots, s-1$ . Блок  $R_s$  представляет собой дополнение объединения всех предыдущих блоков:

$$R_s = (X \times Y) \setminus \bigcup_{t=1}^{s-1} R_t.$$

Типична интерпретация блоков  $R_t$  как «пятен», содержащих особенности взаимосвязи систем  $X$  и  $Y$ , а блока  $R_s$  — как «фона», на котором проявляются эти особенности. Значения  $\lambda_t$  характеризуют средний уровень взаимосвязи подсистем  $X_t$  и  $Y_t$ ,  $t = 1, \dots, s-1$ . При такой интерпретации существенными для анализа связи систем  $X$  и  $Y$  оказываются подсистемы  $\bigcup_{t=1}^{s-1} X_t$  и  $\bigcup_{t=1}^{s-1} Y_t$ .

Каждое подмножество  $X_t$  множества строк  $X$  определяет бинарное разбиение множества  $X$ :

$$Q_t = \{X_t, X \setminus X_t\}.$$

Это дает возможность определить, подобно тому, как это сделано в [9], классификацию строк как пересечение разбиение  $Q_t$ . Классы этой классификации составляют строки, похожие в рамках сделанного агрегирования информации. Аналогично классифицируются столбцы матрицы.

### [Реализация алгоритма «пятна»

При фиксированной сложности  $s$  алгоритм реализует локальную оптимизацию целевого показателя.

Пусть имеется произвольное разбиение матрицы  $R^s$  на «пятна»  $R_1^s, \dots, R_{s-1}^s$  и «фон»  $R_s^s$ . Обозначим  $I(i, t)$  проекцию «пятна»  $R_t^s$  на строку  $i \in X$ :

$$I(i, t) = \{(i, y) \mid y \in Y_t\}$$

и  $J(j, t)$  — проекцию  $R_t^s$  на столбец  $j \in Y$ :

$$J(j, t) = \{(x, j) \mid x \in X_t\}.$$

Локальное изменение «пятна»  $R_t^s$  по строкам состоит в переносе  $I = I(i, t)$  в «фон»  $R_s^s$ , если  $I \subset R_t^s$ , и присоединении  $I$  к  $R_t^s$ , если  $I(i, t) \subset R_s^s$ .

При этом приращение целевого показателя  $f_1$  имеет вид

$$\Delta f_1 = \left( \sum_{R_t^s} a_{xy} + \delta \sum_I a_{xy} \right)^2 / (|R_t^s| + \delta |I|) + \\ + \left( \sum_{R_s^s} a_{xy} - \delta \sum_I a_{xy} \right)^2 / (|R_s^s| - \delta |I|) - \\ - \left( \sum_{R_t^s} a_{xy} \right)^2 / |R_t^s| - \left( \sum_{R_s^s} a_{xy} \right)^2 / |R_s^s|,$$

где

$$\delta = \begin{cases} 1, & \text{если } I \text{ переносится из «фона»,} \\ -1, & \text{если } I \text{ переносится в «фон».} \end{cases}$$

Аналогичная формула выписывается для приращения  $f_1$  при изменении  $R^s$  по столбцам.

В ходе алгоритма последовательно рассматриваются все «пятна»  $R_t^s$ ,  $t = 1, \dots, s-1$ . Для каждого «пятна»  $R_t^s$  рассматриваются все строки  $i \in X$ , и, если  $I(i, t)$  не пересекается с «пятнами»  $R_{t_1}^s$ ,  $t \neq t_1$ , вычисляется величина  $\Delta f_1$ . Если  $\Delta f_1 > 0$ , осуществляется соответствующий перенос  $I$  и в дальнейшем используется измененное таким образом разбиение. В противном случае  $R^s$  не изменяется.

После просмотра строк просматриваются столбцы и т. д. до тех пор, пока  $R^s$  не перестанет изменяться.

Начальное разбиение  $R^{s+1}$  получается из  $R^s$  путем отделения в качестве нового пятна подстроки или подстолбца одного из блоков разбиения. Точнее,  $R_s^{s+1}$  полагается равным  $I(i, t)$ , где  $i \in X_t^s$ , или  $J(j, t)$ , где  $j \in Y_t^s$ . Для  $t = s$  полагается

$$I(i, t) = \{(i, y) \mid (i, y) \in R_s^s\}$$

и

$$J(j, t) = \{(x, j) \mid (x, j) \in R_s^s\}.$$

В качестве «фона»  $R_{s+1}^{s+1}$  берется  $R_s^s \setminus I(i, t)$  или  $R_s^s \setminus J(j, t)$ , если новое пятно выделилось из  $R_s^s$  и  $R_{s+1}^{s+1} = R_s^s$ , если этого не произошло.

При таком переходе от  $R^s$  к  $R^{s+1}$  величина приращения показателя  $f_1$  при выделении в качестве нового «пятна» подстроки имеет вид

$$\Delta f_1 = \left( \sum_{R_t^s \setminus I} a_{xy} \right)^2 / |R_t^s \setminus I| + \left( \sum_I a_{xy} \right)^2 / |I| - \left( \sum_{R_t^s} a_{xy} \right)^2 / |R_t^s|.$$

Аналогичная формула может быть записана, когда в качестве нового пятна выделяется подстолбец.

При таком формировании  $R^{s+1}$  параметры  $i$  ( $j$ ) и  $t$  выбирается таким образом, чтобы величина приращения  $\Delta f_1$  была максимальной.

В целом работа алгоритма начинается с получения из  $R^1 = \{\{X \times Y\}\}$  разбиения  $R^2$  и процесса локального улучшения  $R^2$ , затем из  $R^2$  получается  $R^3$  и т. д., до тех пор, пока не выполнится неравенство

$$f(R^{s+1}) - f(R^s) < 1,$$

если сложность определяется формальным критерием.

В алгоритме возможно использование метода вторых разностей [2], однако это приведет к значительному увеличению необходимых размеров оперативной памяти ЭВМ.

### Упорядочение столбцов и строк матрицы при выдаче результатов агрегирования

Для того, чтобы выявленная структура  $R^s$  была обозрима, необходимо упорядочить строки и столбцы матрицы, чтобы блоки по возможности образовывали связные области при распечатке. С этой целью перестановка совершается таким образом, чтобы все «пятна» были сосредоточены в левом верхнем углу матрицы так, чтобы одинаковые по структуре строки и столбцы были рядом.

Перестановка строк (и столбцов) проводится  $s - 1$  раз — по числу «пятен». Первой перестановкой в верхнюю часть таблицы ставятся строки  $x \in X_{s-1}$ , второй — строки  $x \in X_{s-2}$  и т. д., причем в каждой перестановке прежде всего переставляются нижние строки, а затем уже верхние. Таким образом, самыми верхними будут строки  $x \in X_1 \cap X_2$ , затем пойдут строки  $x \in X_1 \cap$

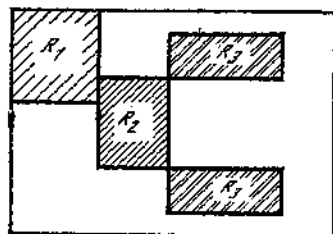


Рис. 1. Результаты работы алгоритма «пятна»

$\cap X_2$ , после них  $x \in X_1 \cap X_2$ , далее строки  $x \in X_1 \cup X_2$ .

Точно так же справа налево переставляются столбцы. Благодаря такой перестановке строк и столбцов первые два «пятна» примут вид касающихся друг друга прямоугольников (рис. 1).

Такое представление структуры оправдано тем, что обычно самый большой вклад в целевой показатель вносят первые «пятна», т. е. они наиболее существенны для анализа.

## Алгоритм «полосы»

По результатам работы этот алгоритм напоминает алгоритм лингвистического анализа [4]. Форма блоков, получаемых этим алгоритмом, также прямоугольная, однако на структуру разбиения накладываются ограничения и определяется она следующим образом.

Столбцы матрицы разбиваются на непересекающиеся классы  $Y_1, \dots, Y_L$  — «полосы». Для каждого класса  $Y_l$  строится разбиение строк  $Q_l = \{X_1^l, \dots, X_{k_l}^l\}$ . Блоками здесь считаются

$$R_{l,t} = X_t \times Y_l, \quad l = 1, \dots, L; \quad t = 1, \dots, k_l.$$

Одна из интерпретаций для матриц коэффициентов связи такова. Пусть признаки  $X$  детерминируют признаки системы  $Y$ . При определении блочной структуры решаются задачи разделения признаков системы  $Y$  на классы, похожие по структуре их детерминации, и выявления уровней детерминации каждого класса.

Помимо разбиения столбцов  $Y_1, \dots, Y_L$ , можно определить классификацию строк на классы, похожих в рамках полученной структуры, и эта классификация является пересечением разбиений  $Q_l$ ,  $l = 1, \bar{L}$ .

## Реализация алгоритма «полосы»

Программа, реализующая алгоритм, работает в трех режимах: когда задано заранее число полос  $L$  и число уровней, на которое разбиваются полосы (при задании этого режима можно взять только  $k_1 = k_2 = \dots, k_L$ );

когда заранее задано только число полос  $L$ , число уровней для каждой полосы определяется автоматически;

все параметры определяются автоматически.

Этот алгоритм также локально-оптимальный. Рассмотрим сначала первый режим работы алгоритма. Оптимизация разбиения осуществляется оптимизацией разбиения строк каждой «полосы» и перемещением столбцов из полосы в полосу.

Разбиение строк «полосы»  $Y_l$  на заданное число блоков состоит в вычислении  $\xi_x = \sum_{Y_l} a_{xy}$  и разбиении  $\{\xi_x\}$  на классы. При этом максимизируется

$$F_l = \sum_{i=1}^{R_l} \left( \sum_{X_i^l} \xi_x \right)^2 / |X_i^l|.$$

Для разбиения  $\{\xi_x\}$ ,  $x \in X$ , на классы существует точный алгоритм [3], однако нами применяется локальный алгоритм, поскольку он использует ранее полученные разбиения.

Приращение  $F_i$  при переносе  $i$  из класса  $X_i^l$  в класс  $X_{i_1}^l$ ,  $i_1 \neq i$ , имеет вид

$$\Delta F_i = \left( \sum_{X_i^l \setminus \{i\}} \xi_x \right)^2 / (|X_i^l| - 1) + \left( \sum_{X_{i_1}^l \cup \{i\}} \xi_x \right)^2 / (|X_{i_1}^l| + 1) - \\ - \left( \sum_{X_i^l} \xi_x \right)^2 / |X_i^l| - \left( \sum_{X_{i_1}^l} \xi_x \right)^2 / |X_{i_1}^l|.$$

В алгоритме просматриваются все  $i \in X$ , для каждого  $i$  в каждом классе  $X_i^l$  вычисляется величина приращения  $\Delta F_i$ ;  $i$  переносится в тот класс, который дает максимальное приращение. Цикл повторяется до тех пор, пока величина  $F_i$  растет.

При перемещении столбца  $j \in Y_i$  в «полосу»  $Y_{i_1}$ ,  $i_1 \neq i$ , изменение целевого показателя  $f_1$  имеет вид

$$\Delta f_1 = \sum_{t=1}^{k_i} \left[ \left( \sum_{X_i^l \setminus \{j\}} a_{xy} \right)^2 / (|R_{i,t}| - |X_i^l|) - \right. \\ \left. - \left( \sum_{R_{i,t}} a_{xy} \right)^2 / |R_{i,t}| \right] + \sum_{t=1}^{k_{i_1}} \left[ \left( \sum_{X_{i_1}^l \cup \{j\}} a_{xy} \right)^2 / (|R_{i_1,t}| + |X_{i_1}^l|) - \right. \\ \left. - \left( \sum_{R_{i_1,t}} a_{xy} \right)^2 / |R_{i_1,t}| \right].$$

Оптимизация перемещением столбцов проводится следующим образом. Последовательно рассматриваются все столбцы, и для каждого столбца по величине  $\Delta f_1$  определяется полоса  $Y_{i_1}$ , в которую его предпочтительней переместить. Если при этом  $\Delta f_1 > 0$ , такое перемещение производят. Эта процедура повторяется до тех пор, пока существуют положительные приращения.

Работа алгоритма в первом режиме состоит в последовательном улучшении начального разбиения. Сначала оптимально разбиваются все «полосы», затем работает процедура перемещения столбцов. Обе процедуры повторяются до тех пор, пока прирост целевого показателя положителен.

Для формирования начального разбиения выбирается  $L$  столбцов — центров «полос». За первый центр принимается столбец, ближайший в смысле евклидова расстояния к средней величине всех столбцов:

$$j_1 = \arg \min_j \sum_X (a_{xj} - \sum_Y a_{xy} / |Y|)^2.$$

В качестве  $l+1$  центра берется столбец, наиболее удаленный от всех предыдущих центров:

$$j_{l+1} = \arg \max_y \min_{k \leq l} \sum_x (a_{xy} - a_{xj_k})^2.$$

Далее столбцы распределяются по близости к центрам:

$$y \in Y_l \Leftrightarrow y = \arg \min_j \sum_X (a_{xj} - a_{xj_l})^2.$$

Таким образом, первая «полоса» собирается из «средних» столбцов, остальные «полосы» содержат особенности данных.

Во втором режиме работы алгоритма, кроме оптимального разбиения «полос» и перемещения столбцов, производится доразбиение «полос». В этой процедуре рассматривается каждая «полоса» и, если полоса  $Y_i$  была разбита на  $K_i$  уровней, то с использованием процедуры оптимального разбиения «полос» проверяется, не следует ли ее разбить на  $K_{i+1}$  уровней. Поскольку сложность

$R_{1,1}$	$R_{2,1}$	$R_{3,1}$	
	$R_{2,2}$		
$R_{1,2}$	$R_{2,1}$		
	$R_{2,2}$		

Рис. 2. Результаты работы алгоритма «полосы»

при этом возрастает на 1, приращение  $f$  в этом случае должно быть больше единицы. Следуя формальному критерию, «полосу» доразбивают, если  $\Delta f > 1$ .

В целом работа алгоритма во втором режиме начинается с доразбиения «полос» начального разбиения, в котором  $K_1 = K_2 = \dots = K_L = 1$ , затем происходит оптимизация перемещением столбцов. Далее алгоритм работает в первом режиме. После этого опять начинается доразбиение «полос» и цикл повторяется. Работа алгоритма заканчивается, когда на этапе доразбиения полос разбиение  $R$  не изменяется.

Работа алгоритма в третьем режиме состоит в последовательном получении во втором режиме разбиений  $R_i(L)$  с  $L$  «полосами»,  $L = 2, 3, \dots$ . Пусть разбиение  $R(L)$  имеет сложность  $S_L$ . Алгоритм прекращает работу, если

$$f(R(L+1)) - f(R(L)) \leq S_{L+1} - S_L,$$

поскольку при переходе от  $L$  к  $L+1$  прирост случайной величины целевого показателя равен  $S_{L+1} - S_L$ .

По окончании работы алгоритма производится упорядочение строк и столбцов матрицы. Столбцы первой полосы ставятся слева, за ними — столбцы второй полосы и т. д. Строки вначале упорядочиваются по уровням «полосы»  $Y_L$ , затем — по уровням «полосы»  $Y_{L-1}$  и т. д. В результате такого упорядочения первыми пойдут строки  $X_1^1 \cap X_1^2 \cap \dots \cap X_1^L$ , затем — строки  $X_1^1 \cap X_1^2 \cap \dots \cap X_2^L$  и т. д. (рис. 2).

## Иллюстрация применения алгоритмов

Содержание данного ниже примера отличается от анализа связи систем признаков, однако интерпретация результатов работы алгоритмов для него принципиально не отличается от изложенной выше схемы.

Для иллюстративного примера взяты данные анкетного обследования жителей Рубцовска, проведенного ИЭиОПП СО АН СССР в 1972 и 1980 гг. под руководством В. А. Артемова. По данным сформирована таблица  $5 \times 8$  элементов. Строкам таблицы соответствуют пять групп населения. Первые две группы составляет работающая молодежь до 29 лет: I — несемейные, II — семейные без детей; группа III — относительно плохо обеспеченные жильем семейная молодежь с детьми и работники 30—39 лет; IV — те же, но хорошо обеспеченные жильем; V — работники не моложе 40 лет. Столбцам таблицы соответствуют ответы на вопрос о намерениях в отношении проведения свободного времени: домашний труд (ДТ); пассивный отдых (ПО); семейное общение, занятия с детьми (СО); учеба (УЧ); потребление информации — телевизор, радио и пр., кроме чтения (ПИ); чтение (ЧТ); отдых на природе, в том числе работа в саду, сбор ягод, грибов, охота и пр. (ОП); любительские занятия — шитье, радиodelo и пр. (ЛЗ).

Элементом  $a_{ij}$  таблицы является изменение доли  $j$ -го ответа среди всех ответов в  $i$ -й группе населения, произошедшее с 1980 г. по сравнению с 1972 г. (табл. 1).

Таблица 1

Исходная таблица иллюстративного примера

Группа населения	Вид деятельности							
	ДТ	ПО	СО	УЧ	ПИ	ЧТ	ОП	ЛЗ
I	7	-5	1	-12	-2	-1	2	3
II	5	0	1	-7	-5	-4	4	1
III	1	-7	7	-5	-1	0	2	3
IV	3	-5	9	-5	-5	3	3	0
V	2	-6	3	-3	-2	1	5	1

Алгоритм «пятна» в автоматическом режиме выделил 5 пятен (табл. 2). Содержательный анализ структуры заключается в объяснении отклонений средних величин пятен.

Первое пятно составили столбцы ПО — пассивный отдых и ПИ — потребление информации. Доля намерений заниматься этими видами деятельности преимущественно уменьшилась. ПИ, как и ПО, относится к пассивным, нетворческим видам досуга, включает в основном проведение времени у телевизора.

Пятна 3, 4 составляют столбец УЧ — учеба. Отрицательные числа в таблице могут объясняться выравниванием материального положения рабочих и интеллигенции. Выделение в отдельное пятно величины изменения доли намерений заниматься учебой

Таблица 2

Таблица, структурированная алгоритмом «пятна» в автоматическом режиме определения сложности

Группа населения	Вид деятельности								№ пятна
	ПО	ПИ	ЧТ	УЧ	ЛЗ	ДТ	СО	ОП	
IV	-5	-5	3	-5	0	3	9	3	1 . 3 . 5
I	-5	-2	-1	-12	3	7	1	2	1 2 . 4 .
II	0	-5	-4	-7	1	5	1	4	1 2 3 . 5
V	-6	-2	1	-3	1	2	3	5	1 2 3 . 5
III	-7	-1	0	-5	3	1	7	2	1 2 3 . .
№ пятна	1	1	2	3	4	5	.	.	

Таблица 3

Структура таблицы, полученная алгоритмом «полосы» (информация о делении полосы на уровни отмечена символом \*)

Группа населения	Вид деятельности													
	ДТ	ОП	*	УЧ	*	ПО	*	СО	*	ПИ	*	ЧТ	ЛЗ	*
II	5	4	1	-7	1	0	1	1	2	-5	1	-4	1	1
III	1	2	1	-5	1	-7	2	7	1	-1	1	0	3	1
IV	3	3	1	-5	1	-5	2	9	1	-5	1	3	0	1
V	2	5	1	-3	1	-6	2	3	2	-2	1	1	1	1
I	7	2	1	-12	2	-5	2	1	2	-2	1	-1	3	1
№ полосы	1	1	*	2	*	3	*	4	*	5	*	6	6	*

I группы населения можно объяснить тем, что молодежь более чутко реагирует на изменения этих условий.

Для объяснения сравнительно небольшого уровня величины в пятнах 2, 5, по-видимому, необходимо провести дополнительные исследования.

Алгоритм «полосы» в автоматическом режиме выделил в матрице 6 «полос» (табл. 3).

Полоса I объединяет столбцы ДТ и ОП — домашний труд и отдых на природе. Эта полоса не разделилась на блоки по уровням. Положительные величины коэффициентов в этом блоке можно

объяснить ростом бытовых потребностей населения, а также ростом потребности общения с природой.

Полоса 2 содержит единственный столбец — УЧ — учебу, в которой выделено два уровня, один из которых состоит из единственной строки — группы I. Разбиение этого столбца соответствует пятнам 3 и 4 на табл. 2, которые мы уже интерпретировали.

Если на табл. 2 столбцы ПО и ПИ составляли пятно 1, то на табл. 3 эти столбцы разделились соответственно на полосы 3 и 5, причем в полосе 3 отдельным уровнем выделялся коэффициент, соответствующий группе II. Такое отличие для интерпретации требует, по-видимому, специального анализа.

Алгоритм «пятна» «не заметил» структуру столбца СО — семейное общение. Алгоритмом «полосы» СО выделилось в отдельную полосу, из которой в свою очередь отдельным уровнем выделились группы III и IV. Относительно высокий уровень коэффициентов в этом блоке можно объяснить повышением значимости занятий с детьми. Для остальных групп жителей СО не так актуально, поскольку они относятся к семьям, в которых либо нет детей, либо они уже выросли.

Алгоритм «полосы» «не заметил» структуры столбца ЧТ — чтение, которую «заметил» алгоритм «пятна». Здесь ЧТ соединилось в один блок — полосу с ЛЗ — любительскими занятиями.

Анализ иллюстративного примера показал, что предложенные инструменты позволяют выявить основные особенности данных, которые дают возможность поставить содержательные гипотезы, а также сформулировать задачи их объяснения. Таким образом, алгоритмы целесообразно использовать на промежуточных этапах исследования данных.

### Литература

1. Авен И. О., Киселева Н. Е., Мучник И. Б. Аппроксимация матрицы данных перекрывающимися блоками. — Автоматика и телемеханика, 1980, № 8, с. 122—132.
2. Бородин А. М. Метод вторых разностей для локальной оптимизации в экстремальных комбинаторных задачах. — Автоматика и телемеханика, 1976, № 10, с. 111—119.
3. Бородин С. М. Оптимизация группировки взаимосвязанных упорядоченных объектов. — Автоматика и телемеханика, 1980, с. 165—172.
4. Жуковская В. М., Мучник И. Б. Факторный анализ в социально-экономических исследованиях. М., 1976.
5. Миркин Б. Г., Ростовцев П. С. Метод выделения связанных групп признаков. — В кн.: Модели агрегирования социально-экономической информации. Новосибирск, 1978, с. 107—110.
6. Миркин Б. Г. Анализ качественных признаков и структур. М., 1980.
7. Ростовцев П. С. Определение сложности структуры агрегирования данных на основе статистических критериев. — В кн.: Алгоритмы анализа данных в социально-экономических исследованиях. Новосибирск, 1982, с. 105—128.
8. Трофимов В. А. Конечный метод решения задачи качественного факторного анализа. — В кн.: Методы анализа многомерной экономической информации. Новосибирск, 1981, с. 12—29.
9. Hartigan J. A. Direkt clustering of data matrix. — J. Am. Stat. Assoc. 67, N 337, p. 123—129.

## ЗАКЛЮЧЕНИЕ

Настоящий сборник — первая попытка систематизации идей и методов новой области анализа данных — анализа нечисловой информации (или, что то же самое, статистики объектов нечисловой природы), развивающейся во многом под влиянием запросов практики конкретных социологических исследований. Конечная цель работ в этом направлении, как уже говорилось, состоит в обеспечении социологических коллективов современными средствами обработки данных, теоретически обоснованными и удобными в реальной деятельности. Однако для достижения этой цели необходимо сделать еще многое.

I. Прежде всего должны быть решены проблемы в области статистики объектов нечисловой природы. Наличие их обусловлено молодостью этой ветви математики.

Более подробно хотелось бы остановиться на одной из основных нерешенных проблем. Дело в том, что хотя нечисловой характер социологических данных является их главной особенностью, отличающей их от тех данных, с которыми имеет дело исследователь, занимающийся естественными науками, эта особенность отнюдь не является единственной. Социологическим данным присуще еще одно существенное свойство, которое «накладывается» на нечисловой характер социологических данных. Имеется в виду статистический, вероятностный характер большинства интересующих социолога явлений. Истинно научное их изучение возможно лишь при использовании положений теории вероятностей и математической статистики. Иначе невозможны соотнесение выводов, полученных для выборки, с соответствующими утверждениями для генеральной совокупности; оценка влияния на содержательные выводы возможных ошибок наблюдения; проверка статистических гипотез и т. д.<sup>1</sup> Однако известные утверждения этой теории имеют силу лишь для числовых данных. Статистика нечисловой информации находится в самом начале своего развития. Отсутствие «вероятностного оформления» многих нечисловых математических методов снижает их научную значимость. К сожалению, это касается и ряда методов, предлагаемых в статьях сборника.

При применении почти всех методов анализа статистических данных возникают следующие две важные проблемы.

<sup>1</sup> Ср. с высказанным Ю. Н. Гаврильцом положением о том, что в статистических исследованиях «надо опираться грамотно» (*Гаврилец Ю. Н. Целевые функции социально-экономического планирования. М., 1983, с. 14*).

а) Все реальные данные известны лишь с некоторыми ошибками. Анализу причин их появления с оценками величин ошибок в конкретных областях посвящены многочисленные специальные исследования. Реальный смысл имеют лишь те выводы математической обработки данных, которые не меняются при колебаниях исходных данных в границах возможных ошибок. Значит, все методы анализа данных должны быть изучены на устойчивость и снабжены рекомендациями по выделению устойчивых выводов из совокупности всех результатов расчетов<sup>2</sup>.

б. Обычно при социологическом исследовании изучают лишь часть интересующей специалистов совокупности. Результаты конкретных обследований бывает желательно перенести на более широкую совокупность, чем множество изученных объектов. Такой перенос возможен лишь на основе определенных теоретических предположений, в соответствии с которыми конкретное обследование рассматривается как выборочное, а его результаты используются для получения результатов о соответствующей генеральной совокупности. Другими словами, перенос выводов с выборки на всю совокупность опирается на теоретические представления, формулируемые в терминах определенных вероятностных моделей действительности. Этот перенос осуществляется с помощью методов математической статистики, позволяющих указать точность получаемых выводов (точность выражается обычно с помощью доверительных интервалов).

Таким образом, проблема устойчивости касается точности выводов по обследованной совокупности, а проблема переноса — точности выводов по всей интересующей исследователя совокупности. Знать оценку точности необходимо для принятия решения при сравнении результатов социологических исследований, проведенных в разных регионах, в разное время и т. д. Пока точность неизвестна, метод анализа данных можно рассматривать лишь как поисковый, а не как научно обоснованный. Доказать с его помощью ничего нельзя, но можно угадать полезные соотношения, сформулировать содержательные гипотезы, которые затем необходимо проверить с помощью научно обоснованных методов математической статистики. Методы, не позволяющие указать точность получаемых выводов, можно сравнить с «временками», на месте которых должны быть построены здания вероятностно-статистических моделей, для которых проверить адекватность и оценить точность выводов можно строго обоснованными методами.

Для того чтобы показать, к чему может привести использование методов, не удовлетворяющих указанному условию, а также возможные пути разработки методов указания требующейся точности, рассмотрим соответствующие аспекты детерминационного анализа и примеры из статьи В. А. Малахова и др.<sup>3</sup>

<sup>2</sup> В связи с задачами классификации подобные вопросы рассматриваются в книге «Прикладная статистика» (М., 1983, с. 166—179).

<sup>3</sup> Приведенные результаты принадлежат А. И. Орлову.

Как говорилось выше, при переносе результатов изучения выборки на генеральную совокупность необходимо указывать точность полученных результатов. Скажем, вместо формулировки типа «удовлетворенность жизнью составила в такой-то группе 70%» писать: «...составила  $70 \pm 10\%$ ». Тогда будет ясно, что если в другой группе «удовлетворенность жизнью» составила 65%, то по данным конкретного обследования удовлетворенность в этих группах нельзя считать различной.

Пусть с помощью детерминационного анализа найдена эмпирическая закономерность  $(x_\alpha, y_s)$  с интенсивностью  $I = I(x_\alpha, y_s)$  и емкостью  $C = C(x_\alpha, y_s)$ . С какой точностью можно использовать числа  $I$  и  $C$  как значения соответствующих характеристик генеральной совокупности? Близкий вопрос: насколько могут измениться значения  $I$  и  $C$  при повторении обследования на другой, но аналогичной совокупности респондентов?

Детерминационный анализ может стать научно обоснованным методом, будучи дополненным правилами определения точности результатов. Для этого нужна соответствующая вероятностно-статистическая теория. Пока ее нет. Здесь же приведем несколько предварительных замечаний. Для определенности обсудим  $I$ , для  $C$  рассуждения аналогичны.

Точность  $I$  определяется ее средним квадратическим отклонением, которое обозначим  $\sigma(I)$ . Целесообразно указывать точность в виде

$$I \pm 2\sigma(I) \quad (1)$$

(это соответствует 95%-й доверительной вероятности, которой будем придерживаться и в дальнейшем).

В ходе конкретного анализа сравниваются интенсивности  $I_1 = I(x_\alpha, y_s)$  и  $I_2 = I(x_\beta, y_s)$  для двух комбинаций  $(x_\alpha, y_s)$  и  $(x_\beta, y_s)$ . Различие можно считать значимым только в случае

$$|I_1 - I_2| > 2\sqrt{\sigma^2(I_1) + \sigma^2(I_2)}. \quad (2)$$

Если это условие не выполнено, то данные конкретного обследования не дают оснований считать, что интенсивности различны.

Величина  $\sigma(I)$  может зависеть от способа организации выборки. Рассмотрим сначала случай, когда выборка — квотная по комбинации  $x_\alpha$ , т. е.  $N_{x_\alpha}$  задано при планировании выборки. Математическая теория в этом случае наиболее проста:  $N_{x_\alpha y_s}$  — биномиальная случайная величина, и, как известно, в качестве  $\sigma(I)$  следует использовать

$$\sigma(I) = \sqrt{\frac{I(1-I)}{N_{x_\alpha}}}. \quad (3)$$

При нескольких тысячах респондентов величина  $N_{x_\alpha}$  для представляющих интерес комбинаций  $x_\alpha$  вряд ли превышает несколько сотен. Рассчитанная с помощью формул (1) и (3) табл. 1 показывает, что  $I$  можно определить лишь с точностью 5—10%.

**Таблица 1**  
**Точность разметки интенсивности**

$I$	50%			80%		
$N_{x\alpha}$ $I + 2\sigma(I)$	25 $50 \pm 20\%$	100 $50 \pm 10\%$	400 $50 \pm 5\%$	25 $80 \pm 16\%$	100 $80 \pm 8\%$	400 $80 \pm 4\%$

Какую разницу между интенсивностями можно уловить? Пусть  $0,5 \leq I_1, I_2 \leq 0,8$ , а  $Nx_\alpha = Nx_\beta = 100$ . Тогда из (2) и (3) следует, что разность  $I_1 - I_2$ , меньшая 11%, не является значимой, т. е. может объясняться чисто случайными причинами.

Формулы (2) и (3) относятся к случаю, когда  $I$  не очень близко к 1 (скажем,  $I \leq 0,9$ ). Рассмотрим противоположную ситуацию, когда  $I \approx 1$ . Что можно сказать об интенсивности  $I_0$  в генеральной совокупности? Нетрудно рассчитать нижнюю границу  $I_{\text{нижн}}$ , такую, что  $I_0 \geq I_{\text{нижн}}$  с принятой нами доверительной вероятностью 0,95. В табл. 2 приведены значения  $I_{\text{нижн}}$  для случая полной детерминации  $I = 1$  и для случая, когда единодушие нарушает только один респондент — только для него значение признака  $y$  не равно  $y_s$ .

**Таблица 2**

**Нижняя граница  $I_{\text{нижн}}$  интенсивности  $I_0$  в генеральной совокупности**

$N_{x\alpha}$	5	10	20	30	40
$I = 1$	0,55	0,74	0,86	0,90	0,92
$I = 1 - \frac{1}{N_{x\alpha}}$	0,31	0,55	0,73	0,80	0,84

Приведенные численные примеры (табл. 1 и 2) показывают, что при переносе результатов детерминационного анализа выборки на генеральную совокупность нужна определенная осторожность.

Перейдем от квотной выборки к выборке объема  $N$ , полученной путем простого случайного отбора. Тогда вектор  $(N_{x_\alpha y_s}, N_{x_\alpha} - N_{x_\alpha y_s})$  имеет мультиномиальное распределение с некоторыми параметрами  $(p_1, p_2)$ . Нами показано, что случайная величина  $I$  асимптотически нормальна с параметрами

$$E(I) \approx \frac{p_1}{p_1 + p_2}, \quad D(I) \approx \frac{p_1 p_2}{N(p_1 + p_2)^3}. \quad (4)$$

Подставляя в (4) вместо  $p_1$  и  $p_2$  их оценки

$$p_1 = \frac{N_{x_\alpha y_s}}{N}, \quad p_2 = \frac{N_{x_\alpha} - N_{x_\alpha y_s}}{N},$$

получаем для среднего квадратического отклонения  $\sigma(I)$  оценку (3).

Рассмотренные выше подходы к оценке точности определения основывались на предположении, что  $x_\alpha$  и  $y_s$  заданы. Однако реально они получаются в результате работы некоторого алгоритма. Как и для других статистических задач<sup>4</sup>, это существенно усложняет теоретическое исследование. Приведем в связи с этим постановку одной нерешенной задачи, связанной с выявлением «ложных детерминаций».

Предположим, что  $y$  не зависит от  $x$ , т. е.

$$p(x_\alpha y_s) = p(x_\alpha) p(y_s)$$

для любых  $x_\alpha, y_s$ , где  $p(\cdot)$  — вероятность появления в выборке комбинации значений, указанной в скобках. Как распределена величина

$$\xi_s = \max_{x_\alpha} \frac{N_{x_\alpha y_s}}{N_{x_\alpha}}$$

Ясно, что  $\xi_s \geq p(y_s)$ . Насколько больше? Пусть  $t_s$  — такое число, что

$$p(\xi_s \leq t_s) = 0,95.$$

Тогда при обработке реальных данных следует рассматривать только детерминации с  $I > t_s$ , все остальные могут быть ложными, т. е. объясняться игрой случая. Нерешенная задача: как вычислить  $t_s$ ?

Детерминационный анализ должен и будет обоснован методами математической статистики. Как уже отмечалось, подобного обоснования не существует для большинства методов, предлагаемых в статьях сборника.

Это можно сказать, например, о теории монотонных систем, описанной в статье Е. Н. Кузнецова и других, о методах группировки, предлагаемых в статье Б. Г. Миркина, о способах анализа матрицы данных из статьи П. С. Ростовцева. Почти не имеют вероятностно-статистического обоснования методы многомерного шкалирования. В частности, неизвестна точность получаемых решений, описываемая в математической статистике в терминах доверительных интервалов (в многомерном случае — доверительных множеств). Это в полной мере можно отнести к методам, о которых идет речь в статье Г. А. Сатарова. Некоторые предварительные результаты в направлении ликвидации этого пробела приведены в статье А. И. Орлова. Одно из достоинств подхода к многомерному шкалированию, предлагаемого в статье В. Т. Перекреста, состоит в том, что в принципе этот подход позволяет дать требующуюся оценку точности.

<sup>4</sup> См.: Загоруйко Н. Г., Орлов А. И. Некоторые нерешенные математические задачи прикладной статистики. — В кн.: Современные проблемы кибернетики: (Прикладная статистика). М., 1981, с. 53—63.

II. Необходима разработка программного обеспечения (дающего возможность реализации на ЭВМ как отдельных методов, так и их комплексного применения) для ЭВМ различных классов, включающего диалоговые системы на естественном языке. Необходимо подчеркнуть важность эффективной организации распространения пакетов прикладных программ и обучения пользователей как теоретическим основам внедряемых методов, так и их применению в конкретных задачах.

III. Нужно разработать серии методических рекомендаций по поводу того, в каких областях конкретных социологических исследований целесообразно использовать те или иные методы анализа нечисловой информации. Для ответов на соответствующие вопросы необходимо сочетание глубокого проникновения в содержательную область с пониманием возможностей математических методов.

IV. Необходимо накопить достаточно большой опыт применения математических методов в социологических исследованиях. В соответствии с накопленным опытом способы решения проблем, изложенных в п. I—III, должны регулярно корректироваться.

Перечисленные задачи могут быть решены лишь в процессе постоянного взаимодействия социологов, математиков, программистов. Запросы конкретных областей применения приводят к постановкам новых теоретических задач и требуют модификации программного обеспечения. В то же время знание социологами возможностей известных математических методов может привести к расширению области применения последних.

Подготовка метода к его активному внедрению в практику является сложным, многоступенчатым процессом. Почти все статьи сборника отражают лишь его начало. Наивно надеяться, что отдельные лица (или даже коллективы) полностью «подготовят к употреблению» методы, которым посвящен сборник. Но в нашей стране сейчас работает немало специалистов по применению математических методов в социологии, совместными усилиями которых эта цель может быть достигнута.

*В. Г. Андреева,  
А. И. Орлов,  
Ю. Н. Толстова*

# СОДЕРЖАНИЕ

Предисловие . . . . .	3
-----------------------	---

## I

### ОБЩИЕ ВОПРОСЫ АНАЛИЗА НЕЧИСЛОВОЙ ИНФОРМАЦИИ

В. Г. АНДРЕЕНКОВ, Ю. Н. ТОЛСТОВА. Особенности применения математических методов в социологических исследованиях . . . . .	7
Е. Н. КУЗНЕЦОВ, И. Б. МУЧНИК, Л. В. ШВАРЦЕР. Монотонные системы и их свойства . . . . .	29
А. И. ОРЛОВ. Общий взгляд на статистику объектов нечисловой природы. . . . .	58
С. А. ПЕТРОВСКИЙ. Прогнозирование и анализ на проблемных сетях. . . . .	92

## II

### ПРОБЛЕМЫ ИЗМЕРЕНИЯ

В. Т. ПЕРЕКРЕСТ. Функциональный подход в метрическом многомерном шкалировании . . . . .	113
Г. А. САТАРОВ. Многомерное шкалирование и другие методы при комплексном анализе данных . . . . .	132

## III

### МЕТОДЫ КЛАССИФИКАЦИИ

Г. С. ЛЕОВ, Т. М. ПЕСТУНОВА. Группировка объектов в пространстве разнотипных признаков . . . . .	141
В. А. МАЛАХОВ, А. А. ВЕСЕЛОВ, В. Н. ДЕЗА, А. Я. ПОДРАБИ-НОВИЧ. Опыт применения детерминационного анализа в исследовании влияния социально-демографических факторов . . . . .	150
Б. Г. МИРКИН. О математическом аппарате метода группировок в современных социально-экономических исследованиях . . . . .	161
Г. В. РАУШЕНБАХ. Меры близости и сходства в социологии . . . . .	169
П. С. РОСТОВЦЕВ. Алгоритмы анализа структуры прямоугольных матриц «пятна» и «полосы» . . . . .	203
Заключение . . . . .	215

**АНАЛИЗ  
НЕЧИСЛОВОЙ  
ИНФОРМАЦИИ  
В СОЦИОЛОГИЧЕСКИХ  
ИССЛЕДОВАНИЯХ**

Утверждено к печати  
Институтом социологических исследований  
АН СССР

Редакторы издательства В. В. Онуфриев, Т. В. Саркисова

Художник А. А. Куценко

Художественный редактор Н. Н. Фильчагина

Технические редакторы Н. П. Переверза, М. Ю. Соловьева

Корректоры А. Б. Васильев, Н. Б. Габасова

ИБ № 29703

Сдано в набор 9.04.85. Подписано к печати 22.07.85.

Т-14853. Формат 60×90<sup>1</sup>/<sub>16</sub>. Бумага книжно-журнальная импортная

Гарнитура обыкновенная. Печать высокая

Усл. печ. л. 14. Усл. кр. отт. 14,125. Уч.-изд. л. 15,4. Тираж 3250 экз.

Тип. зак. 1389 Цена 1 р. 60 к.

Ордена Трудового Красного Знамени

издательство «Наука»

117864 ГСП-7, Москва В-485, Профсоюзная ул., 90.

2-я типография издательства «Наука»

121099, Москва, Г-99, Шубинский пер., 6

## **В ИЗДАТЕЛЬСТВЕ «НАУКА» ГОТОВЯТСЯ К ПЕЧАТИ:**

---

### **РАЗВИТИЕ СОЦИАЛЬНОЙ СТРУКТУРЫ ОБЩЕСТВА В СССР: АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ**

17 л.— 1 р. 70 к.

В монографии анализируется социальная структура советского общества в условиях развитого социализма. Определена социальная политика партии и государства в сфере экономики, демографии, других сфер жизнедеятельности, рассматриваются вопросы дальнейшего укрепления единства советского народа как новой исторической общности людей. Анализ позволяет сформулировать актуальные проблемы социологических исследований.

Для социологов, философов.

---

### **СОЦИАЛЬНЫЕ И ЭКОНОМИЧЕСКИЕ ПРОБЛЕМЫ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ НАУКИ**

16 л.— 1 р. 60 к.

Книга написана по материалам двух научных симпозиумов по проблемам эффективности науки, проведенных в 1981 и 1982 гг. В ней анализируются актуальные, требующие решения вопросы развития науки, предложения по совершенствованию ее организации и управления. Рассматриваются социальные аспекты развития, вопросы формирования личности ученого и деятельности научных коллективов.

Для научных и практических работников, занимающихся вопросами организации научной деятельности.

Для получения книг почтой заказы просим направлять по адресу: 117192, Москва, Мичуринский проспект, 12, магазин «Книга — почтой»; Центральной конторы «Академкнига»; 197345 Ленинград, Петрозаводская ул., 7, магазин «Книга — почтой» Северо-Западной конторы «Академкнига» или в ближайший магазин «Академкнига», имеющий отдел «Книга — почтой».

- |        |  |        |  |
|--------|--|--------|--|
| 480091 | Алма-Ата, ул. Фурманова, 91/97 («Книга — почтой»);           | 197345 | Ленинград, Петрозаводская ул., 7 («Книга — почтой»);                 |
| 370005 | Баку, ул. Джапаридзе, 13 («Книга — почтой»);                 | 220012 | Минск, Ленинский проспект, 72 («Книга — почтой»);                    |
| 232800 | Вильнюс, ул. Университетского, 4;                            | 103009 | Москва, ул. Горького, 19а;   |
| 690088 | Владивосток, Океанский проспект, 140;                        | 117312 | Москва, ул. Вавилова, 55/7;  |
| 320093 | Днепропетровск, проспект Гагарина, 24 («Книга — почтой»);    | 117192 | Москва, Мичуринский проспект, 12 («Книга — почтой»);                 |
| 734001 | Душанбе, проспект Ленина, 95 («Книга — почтой»);             | 630076 | Новосибирск, Красный проспект, 51;                                   |
| 375002 | Ереван, ул. Туманяна, 31;                                    | 630090 | Новосибирск, Академгородок, Морской проспект, 22 («Книга — почтой»); |
| 684033 | Иркутск, ул. Лермонтова, 289 («Книга — почтой»);             | 142284 | Протвино, Московской обл., «Академкнига»;                            |
| 420043 | Казань, ул. Достоевского, 53;                                | 142292 | Пушино, Московской обл., МР, «В», 1;                                 |
| 252030 | Киев, ул. Ленина, 42;  | 620151 | Свердловск, ул. Мамин-Сибиряка, 137 («Книга — почтой»);              |
| 252142 | Киев, проспект Вернадского, 79;                              | 700029 | Ташкент, ул. Ленина, 73;   |
| 252030 | Киев, ул. Пирогова, 4 («Книга — почтой»);                    | 700100 | Ташкент, ул. Шота Руставели, 43;                                     |
| 277012 | Кишинев, проспект Ленина, 148 («Книга — почтой»);            | 700187 | Ташкент, ул. Дружбы народов, 6 («Книга — почтой»);                   |
| 343900 | Краматорск, Донецкой обл., ул. Марата, 1 («Книга — почтой»); | 634050 | Томск, наб. реки Ушайки, 18;   |
| 660049 | Красноярск, проспект Мира, 84;                               | 450059 | Уфа, ул. Р. Зорге, 10 («Книга — почтой»);                            |
| 443002 | Куйбышев, проспект Ленина, 2 («Книга — почтой»);             | 450025 | Уфа, ул. Коммунистическая, 49;                                       |
| 191104 | Ленинград, Литейный проспект, 57;                            | 720001 | Фрунзе, бульвар Дзержинского, 42 («Книга — почтой»);                 |
| 199164 | Ленинград, Таможенный пер., 2;                               | 310078 | Харьков, ул. Чернышевского, 87 («Книга — почтой»).                   |
| 196034 | Ленинград, В/О, 9 линия, 16;                                 |        |  |